SceMQA: A Scientific College Entrance Level Multimodal Question Answering Benchmark

Zhenwen Liang[1], **Kehan Guo**[1], **Gang Liu**[1], **Taicheng Guo**[1], **Yujun Zhou**[1], **Tianyu Yang**[1], **Jiajun Jiao**[2], **Renjie Pi**[3], **Jipeng Zhang**[3], and **Xiangliang Zhang** [✉1]

[1]University of Notre Dame, {zliang6,xzhang33}@nd.edu
[2]New York University
[3]Hong Kong University of Science and Technology

Abstract

The paper introduces SceMQA, a novel benchmark for scientific multimodal question answering at the college entrance level. It addresses a critical educational phase often overlooked in existing benchmarks, spanning high school to pre-college levels. SceMQA focuses on core science subjects including Mathematics, Physics, Chemistry, and Biology. It features a blend of multiple-choice and free-response formats, ensuring a comprehensive evaluation of AI models' abilities. Additionally, our benchmark provides specific knowledge points for each problem and detailed explanations for each answer. SceMQA also uniquely presents problems with identical contexts but varied questions to facilitate a more thorough and accurate assessment of reasoning capabilities. In the experiment, we evaluate both open-source and close-source state-of-the-art Multimodal Large Language Models (MLLMs), across various experimental settings. The results show that further research and development are needed in developing more capable MLLM, as highlighted by only 50% to 60% accuracy achieved by the strongest models.

1 Introduction

In recent years, the evolution of large language models (LLMs) has marked a significant milestone in artificial intelligence. Initially, these models excelled in diverse natural language processing tasks (Brown et al., 2020; Ouyang et al., 2022; Touvron et al., 2023a,b; OpenAI, 2023; Google, 2023), but their utility has since increasingly expanded, transforming them into incredible agents for various downstream tasks such as reasoning and planning (Li et al., 2023; Wu et al., 2023b; Park et al., 2023; Guo et al.). Notably, LLMs have shown proficiency in tasks that typically pose significant challenges to even highly skilled humans, such as tackling intricate mathematical problems (Lu et al., 2023; Romera-Paredes et al., 2023) and accelerating scientific discoveries (Birhane et al., 2023). This evolution demonstrates the versatility of LLMs and their potential to revolutionize areas traditionally dominated by human expertise.

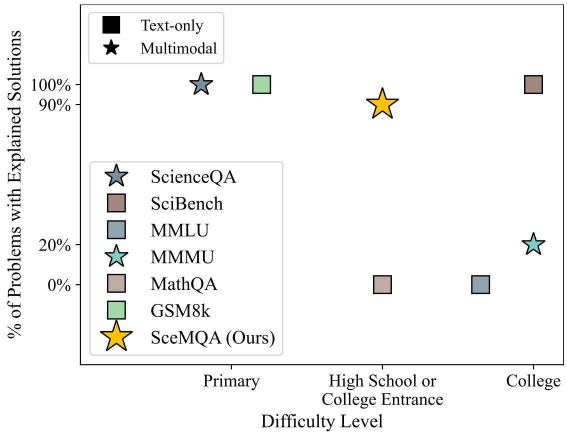

Figure 1: The comparison between SceMQA and other existing benchmarks. Y-axis is the percentage of problems that have detailed solution explanations. Most problems (over 90%) in SceMQA has detailed explanations to solutions except for some straightforward problems. More comparison can be found in Table 1.

Alongside, the rapid development of vision-based LLMs has garnered considerable attention within the AI community, especially with the release of platforms like OpenAI's GPT4-V (OpenAI, 2023) and Google's Gemini Ultra (Google, 2023). These models have demonstrated exceptional abilities in tasks requiring advanced reasoning and planning, often surpassing existing benchmarks and approaching human-level performance. This progress has spurred researchers to create more sophisticated and challenging benchmarks for Multimodal LLMs (MLLMs), one of the most representative is the science domain, which is a long-standing focus for humans. For example, the MathVista benchmark (Lu et al., 2023), comprising 6,141 problems, demands a high level of visual understanding and

Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pages 109–119
August 11-16, 2024 ©2024 Association for Computational Linguistics

mathematical reasoning. Additionally, the Massive Multi-discipline Multimodal Understanding and Reasoning Benchmark (MMMU) (Yue et al., 2023a) poses college-level multimodal reasoning challenges. Currently, even the most advanced models achieve only about 50% accuracy on these benchmarks. The importance of such benchmarks lies in their role as vital tools for assessing and pushing the boundaries of AI capabilities. By presenting AI models with tasks that mimic complex, real-world scenarios, benchmarks provide a clear measure of progress and highlight areas for future development.

However, in the science domain, a critical observation in multimodal reasoning benchmarks is the disparity in the levels of difficulty. Prior benchmarks like ScienceQA (Lu et al., 2022) primarily focused on elementary and middle-school levels, while MMMU leaps to a college-level challenge. This leaves a significant educational phase in human learning – the high school, or college entrance level – relatively unaddressed. In fact, learning progressively in difficulty levels is not only important for humans, but also can facilitate AI systems including LLMs via curriculum learning (Bengio et al., 2009) and progressive training (Xu et al., 2023; Mitra et al., 2023). Therefore, we fill this gap by introducing a novel benchmark named **S**cience **c**ollege **e**ntrance level **M**ultimodal **Q**uestion **A**nswering (SceMQA), designed for this critical educational stage, with four key subjects: Mathematics, Physics, Chemistry, and Biology.

Apart from the difficulty level, our benchmark also has a detailed annotation granularity. Firstly, most problems (over 90%) in SceMQA has detailed explanations to solutions except for some straightforward problems. Besides, each problem is associated with a specific knowledge component, facilitating detailed knowledge tracing for models. Moreover, SceMQA uniquely features problems with the same context but different questions. This design is informed by prior research indicating that without diverse question types for each narrative context, models might resort to learning shallow heuristics or patterns rather than developing a deep, semantic understanding (Patel et al., 2021; Yang et al., 2022). This approach ensures a more comprehensive and precise evaluation of reasoning capabilities. In Figure 1, we compare the difficulty level, annotation granularity, and covered modality among existing benchmarks.

2 Related Work

Multimodal Question Answering Multimodal Question Answering (QA) has been a focal area in AI research. The Visual Question Answering (VQA) benchmark (Antol et al., 2015), established in 2015, pioneered free-form, open-ended visual QA, necessitating intricate image comprehension and reasoning. ChartQA (Masry et al., 2022) emphasized complex reasoning about charts, merging visual and logical thought processes. VisIT-Bench (Bitton et al., 2023) tested vision-language models across real-world tasks, ranging from simple recognition to advanced creative generation.

Multimodal LLMs In addition to notable models like GPT4-V and Google Gemini, various open-source Multimodal LLMs (MLLMs) have emerged. MiniGPT-4 (Zhu et al., 2023) improved vision-language understanding by syncing a visual encoder with a language LLM. LLaVAR (Zhang et al., 2023b) combined OCR with text-only GPT-4 for enhanced visual instruction tuning in text-rich image contexts. mPLUG-Owl (Ye et al., 2023) proposed a modular framework for equipping LLMs with multimodal capabilities, focusing on image-text alignment. InstructBLIP (Dai et al., 2023) excelled in vision-language instruction tuning, demonstrating remarkable zero-shot performance in diverse tasks. For a more detailed summary of related studies, please refer to these surveys (Wu et al., 2023a; Yin et al., 2023).

Science Question Answering Various benchmarks have been developed for specific scientific subjects, including MATH (Hendrycks et al., 2021b), MathVista (Lu et al., 2023), chemistry (Guo et al., 2023), etc. More comprehensive science QA benchmarks like ScienceQA (Lu et al., 2022), C-EVAL (Huang et al., 2023), AGIEVAL (Zhong et al., 2023), MMMU (Yue et al., 2023a), and SciBench (Wang et al., 2023b) have recently been introduced, providing a broader scope of assessment.

3 Our Benchmark SceMQA

Our benchmark is designed to bridge a significant gap in existing multimodal benchmarks, which typically span from elementary to college levels, and overlook the crucial high school/college entrance stages. This educational phase is crucial in the human learning process. Although existing benchmarks (Zhong et al., 2023; Zhang et al., 2023a)

	Problem Format	# Problems Per Subject	Problem Modality	Solution Explanation*	Difficulty Level
MMLU	MC	279	T	No	College
SciBench	FR	232	T	Yes	College
ScienceQA	MC	816	T+I	Yes	Primary
MathVista	MC + FR	-	T+I	No	Unspecified
MMMU	MC + FR	385	T+I	No	College
SceMQA (Ours)	**MC + FR**	261	**T+I**	**Yes**	**College Entrance**

Table 1: A comparative overview of various benchmarks. The first column indicates the problem types inside the benchmark, with "MC" representing multiple choice and "FR" indicating free-response formats. The second column shows the average number of problems per subject. The third column describes the problem modality, where "I" stands for image-based and "T" for text-based problems. (*) The fourth column categorizes benchmarks based on whether over 90% of problems are annotated with solutions explanations. The final column presents the difficulty level. All superior and unique features of our benchmark are highlighted.

incorporate problems at this level, they predominantly feature text-only questions. A comparative analysis of our dataset against existing benchmarks is detailed in Table 1. Although our benchmark appears smaller in total problem count, it focuses specifically on the science domain, offering a substantial average number of problems per subject. Furthermore, it excels in quality, as evidenced by the high proportion of problems accompanied by detailed explanations. The collection and annotation protocol is located in Section A.3. Example problems in our benchmark are shown in the Appendix (Figure 5).

	Multiple Choice	**Free Response**
Total Questions	845	200
Unique Images	632	118
Max Question Length	1816	1906
Max Answer Length	1124	2614
Average Question Length	452	410
Average Answer Length	297	330

Table 2: SceMQA Statistics.

SceMQA has in total 1,045 problems, with an average of 261 problems per subject. Details can be found in Table 2. This set of problems ensures a thorough evaluation across all included subjects.

4 Experimental Examination of SceMQA

In this section, we evaluate the state-of-the-art MLLMs on SceMQA by firstly reporting their answer accuracy across various settings. Additionally, we conduct a detailed *error analysis* (Section 4.3) and show an *accuracy distribution across knowledge categories* (Section A.1), which provide significant insights to identify the current MLLMs' limitations and demonstrate the value of our benchmark in exploring them. We will move those important experiments to the main body of our paper when we have more space upon paper acceptance.

4.1 Experimental Settings

We choose InstructBLIP (Dai et al., 2023), MiniGPT4 (Zhu et al., 2023) and LLaVa (Liu et al., 2023a) as the open-source MLLM solvers for SceMQA. As for close-sourced models, we focus on three of the most representative MLLMs currently available: Google Bard, Gemini Pro and GPT4-V. Furthermore, we test GPT4-V and Gemini Pro under three distinct settings: zero-shot, few-shot, and text-only. In the zero-shot setting, the models are provided with the problem without any prior examples. The few-shot setting involves giving the models a small number of example problems and solutions to "learn" from, before attempting the new problems. We use hand-crafted text-only problems as examples since it is not flexible to insert multiple images in one API call. The text-only setting is a unique approach under zero-shot where only the textual content of the problem is provided to the model, without any images. All the prompts in our experiments, along with detailed descriptions of each setting, will be available for public view after the paper is accepted.

For the evaluation metric, we have chosen to use exact-match-based accuracy, which is consistent with several prior studies (Lu et al., 2023; Yue et al., 2023a) in this domain. This metric is particularly suitable for our benchmark as both the multiple-choice and free-response problems have definitive, singular correct answers. In the multiple-choice format, this involves selecting the correct option out of the presented choices. For the free-response format, it requires generating an accurate and precise answer, be it a numerical value, a yes/no response, or a specific term for fill-in-the-blank questions. Empirically we use rule-based answer exaction for

Open-sourced models										
Model	Multiple Choice					Free Response				
	Math	Physics	Chemistry	Biology	Overall	Math	Physics	Chemistry	Biology	Overall
InstructBLIP-7B	16.98	21.86	20.30	22.75	20.48	6.00	6.00	0.00	38.00	12.50
InstructBLIP-13B	19.34	19.53	17.33	28.91	21.31	8.00	12.00	4.00	30.00	13.50
MiniGPT4-7B	18.87	20.93	25.25	22.75	21.90	4.00	0.00	2.00	20.00	6.50
MiniGPT4-13B	27.39	20.93	27.23	35.55	27.74	2.00	4.00	8.00	14.00	7.00
LLaVA1.5-7B	25.94	25.12	21.78	36.97	27.50	10.00	4.00	2.00	26.00	10.50
LLaVA1.5-13B	31.13	28.37	26.24	38.86	31.19	12.00	4.00	4.00	32.00	13.00
Yi-VL-6B	43.87	26.98	28.79	48.37	37.14	2.00	2.00	2.00	16.00	5.50
Deepseek-VL-Chat-7B	24.53	21.86	26.26	34.42	26.79	6.00	10.00	6.00	34.00	14.00
InternLM-XComposer2-7B	29.25	26.98	31.82	33.95	30.48	8.00	4.00	10.00	30.00	13.00
Qwen-VL-chat	25.47	23.72	22.22	34.42	26.55	4.00	0.00	0.00	24.00	7.00

Close-sourced models											
Model	Setting	Multiple Choice					Free Response				
		Math	Physics	Chemistry	Biology	Overall	Math	Physics	Chemistry	Biology	Overall
Google Bard	Text-only	43.40	40.93	24.75	54.88	41.31	14.00	12.00	22.00	34.00	20.50
Gemini Pro	Text-only	21.70	19.53	32.51	46.51	30.06	8.00	6.00	8.00	38.00	15.00
	Few-shot	36.79	30.23	37.44	48.84	38.34	18.00	12.00	12.00	36.00	19.50
	Zero-shot	37.26	30.70	42.36	54.42	41.18	20.00	12.00	18.00	36.00	21.50
GPT4-V	Text-only	35.38	47.91	58.13	63.72	51.24	12.00	24.00	28.00	22.00	21.50
	Few-shot	54.72	53.95	58.62	67.44	58.70	30.00	24.00	30.00	48.00	33.00
	Zero-shot	**55.19**	**55.81**	**60.10**	**72.09**	**60.83**	**36.00**	**24.00**	**36.00**	**48.00**	**36.00**

Table 3: Accuracy of examining GPT4-V and Gemini Pro across different settings on Multiple Choice and Free Response problems in SceMQA.

multiple choice questions, and GPT4 as evaluators for free response questions.

4.2 Accuracy for Solving SceMQA

The performance of examined MLLMs on SceMQA is presented in Table 3. Foremost, in all evaluated scenarios, the zero-shot GPT4-V consistently outperforms other models. Despite this, the challenge posed by the benchmark remains significant for even the most advanced MLLMs, including GPT4-V and Google Gemini. This parity shows the challenging nature of our benchmark and the necessity for further improving MLLMs' reasoning capabilities. It can be also observed that the performance of open-sourced models are significantly inferior to close-sourced ones. We have looked into the error cases and found that the both instruction-following and reasoning abilities of open-sourced models are not very satisfactory, leaving a huge room for improvement.

Additionally, in the few-shot setting, we noticed an intriguing trend: it underperforms the zero-shot setting. We hypothesize that the few-shot examples, while providing guidance on scientific reasoning, do not enhance the models' ability to interpret scientific images. This could inadvertently lead the models to prioritize logical reasoning over critical image interpretation. Also, we can see a significantly lower performance in the text-only setting. This highlights the indispensability of visual information in solving the problems in our benchmark.

Another notable finding is the variation in performance across different subjects. The models perform better in Chemistry and Biology compared to Math and Physics. We infer that this is because Math and Physics often require precise calculations for correct answers, while Chemistry and Biology tend to focus more on conceptual understanding. This pattern suggests that the integration of external computational tools, such as calculators or Python programs, might be beneficial in improving performance on our benchmark, particularly in subjects with extensive calculations like Math and Physics.

4.3 Error Analysis

To delve deeper into the shortcomings of state-of-the-art MLLMs, we conducted a comprehensive error analysis. We randomly selected 150 instances of errors made by GPT4-V on the SceMQA dataset and enlisted two human experts for a detailed examination. These experts categorized each error into one of six categories: *Image Perceptual Errors,*

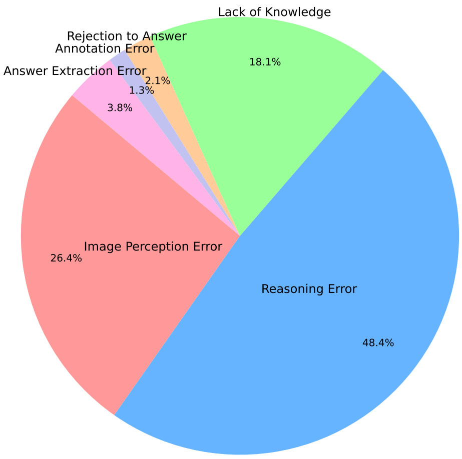

Figure 2: Distribution of GPT4-V's error types across 100 samples.

Reasoning Errors, Lack of Knowledge, Rejection to Answer, Annotation Error, and *Answer Extraction Error*. The inter-rater reliability, assessed using the Kappa agreement score, was found to be greater than 0.5, indicating a moderate level of agreement between the annotators. We then averaged their annotations to determine the proportion of each error type, as depicted in Figure 2. The top-3 error types are shown in Figure 3 and analyzed below:

Reasoning Error The most prevalent error type is categorized under *Reasoning Error*. It occurs when the model correctly processes image-based information but fails to construct an accurate reasoning chain to arrive at the correct answer. Common mistakes include omitting necessary steps or making incorrect calculations. And we find these errors evenly spread in four subjects in SceMQA, underscoring the need for further development in the reasoning abilities of MLLMs. Drawing on insights from studies on LLMs, approaches such as prompting engineering (Wei et al., 2022) or supervised fine-tuning (Yu et al., 2023; Yue et al., 2023b) might prove beneficial.

Image Perception Error This occurs when the model misinterprets visual information—such as incorrectly reading numbers or coordinates, or failing to differentiate between points in a geometric diagram. This type of error happens more often in the math subject because many math problems require precise diagram or table perception, which suggests that the image perception capabilities of current MLLMs require significant enhancement for precision and interpretation. Incorporation of external tools like OCR, as suggested in studies like (Liu et al., 2023b), could potentially improve the model's understanding of visual content.

Lack of Knowledge This type of error arises when the model fails to correctly identify or apply relevant knowledge concepts, such as misusing formulas or misinterpreting theorems. These errors occur more in physics, chemistry and biology, which are indicative of gaps in the model's learned knowledge base, suggesting that enriching the training datasets of foundation models with diverse and domain-specific knowledge is essential to enhance their expertise in those domains.

Rejection to Answer and Annotation Error Interestingly, a smaller portion of errors were categorized as *Rejection to Answer* and *Annotation Error*. *Rejection to Answer* occurs when the model refuses to provide an answer, possibly due to uncertainty or inability to comprehend the query. *Annotation Error*, on the other hand, arises from inaccuracies or inconsistencies in the dataset's annotations, leading to confusion for the model. These categories highlight the importance of robust dataset design and also the need for models to handle ambiguous or complex instructions and questions effectively.

Through this detailed error analysis, we have identified specific patterns and weaknesses of MLLMs' performance on scientific problems. These findings provide valuable insights and directions for future research aimed at enhancing the capabilities of MLLMs. Addressing these identified issues could lead to significant improvements in the application of MLLMs in educational and research contexts, particularly in the domain of science.

5 Conclusion

In this paper, we introduced SceMQA, a novel multimodal question answering dataset tailored for the college entrance level, including key scientific subjects: mathematics, physics, chemistry, and biology. A standout feature of SceMQA is its high annotation granularity, with over 90% problems accompanied by detailed explanations and associated with specific knowledge points. We conduct extensive experiments including accuracy comparison, error analysis, and category accuracy distribution, employing state-of-the-art MLLMs and highlighting the opportunities and obstacles for multimodal AI models in scientific reasoning.

Limitation

Model Comparison Our SceMQA is evaluated on a small number of state-of-the-art MLLMs due to limited computational resources. We plan to evaluate a wider range of models in the future. We will include both open-source models, such as Qwen-VL (Bai et al., 2023) and CogVLM (Wang et al., 2023a), and closed-source ones like Claude. This comprehensive comparison will provide deeper insights into the capabilities and limitations of those AI models in multimodal scientific reasoning.

Data Scope We will enhance both the depth and breadth of our dataset. In terms of depth, we plan to incorporate more diverse problems within each scientific subject. This will involve adding more complex and varied question types. As for breadth, we aim to extend the range of subjects covered by our dataset beyond the traditional sciences, including more disciplines that are encountered in the human cognitive process.

References

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*.

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. 2009. Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning*, pages 41–48.

Abeba Birhane, Atoosa Kasirzadeh, David Leslie, and Sandra Wachter. 2023. Science in the age of large language models. *Nature Reviews Physics*, pages 1–4.

Yonatan Bitton, Hritik Bansal, Jack Hessel, Rulin Shao, Wanrong Zhu, Anas Awadalla, Josh Gardner, Rohan Taori, and Ludwig Schimdt. 2023. Visit-bench: A benchmark for vision-language instruction following inspired by real-world use. *Advances in Neural Information Processing Systems*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33:1877–1901.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. 2023. Instructblip: Towards general-purpose vision-language models with instruction tuning. *arXiv preprint arXiv:2305.06500*.

Google. 2023. Introducing gemini: our largest and most capable ai model.

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

Taicheng Guo, Kehan Guo, Zhengwen Liang, Zhichun Guo, Nitesh V Chawla, Olaf Wiest, Xiangliang Zhang, et al. 2023. What indeed can gpt models do in chemistry? a comprehensive benchmark on eight tasks. *arXiv preprint arXiv:2305.18365*.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021a. Measuring massive multitask language understanding. *Proceedings of the International Conference on Learning Representations (ICLR)*.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021b. Measuring mathematical problem solving with the math dataset. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

Namgyu Ho, Laura Schmid, and Se-Young Yun. 2022. Large language models are reasoning teachers. *arXiv preprint arXiv:2212.10071*.

Cheng-Yu Hsieh, Chun-Liang Li, Chih-Kuan Yeh, Hootan Nakhost, Yasuhisa Fujii, Alexander Ratner, Ranjay Krishna, Chen-Yu Lee, and Tomas Pfister. 2023. Distilling step-by-step! outperforming larger language models with less training data and smaller model sizes. *arXiv preprint arXiv:2305.02301*.

Yuzhen Huang, Yuzhuo Bai, Zhihao Zhu, Junlei Zhang, Jinghan Zhang, Tangjun Su, Junteng Liu, Chuancheng Lv, Yikai Zhang, Jiayi Lei, et al. 2023. C-eval: A multi-level multi-discipline chinese evaluation suite for foundation models. *arXiv preprint arXiv:2305.08322*.

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

Guohao Li, Hasan Abed Al Kader Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. 2023. Camel: Communicative agents for" mind" exploration of large language model society. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Zhenwen Liang, Dian Yu, Xiaoman Pan, Wenlin Yao, Qingkai Zeng, Xiangliang Zhang, and Dong Yu. 2024a. Mint: Boosting generalization in mathematical reasoning via multi-view fine-tuning. *COLING-LREC*.

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

Zhenwen Liang, Wenhao Yu, Tanmay Rajpurohit, Peter Clark, Xiangliang Zhang, and Ashwin Kaylan. 2023. Let gpt be a math tutor: Teaching math word problem solvers with customized exercise generation. *EMNLP*.

Zhenwen Liang and Xiangliang Zhang. 2021. Solving math word problems with teacher supervision. In *IJCAI*, pages 3522–3528.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023a. Visual instruction tuning. *Advances in Neural Information Processing Systems*.

Yuliang Liu, Zhang Li, Hongliang Li, Wenwen Yu, Mingxin Huang, Dezhi Peng, Mingyu Liu, Mingrui Chen, Chunyuan Li, Lianwen Jin, et al. 2023b. On the hidden mystery of ocr in large multimodal models. *arXiv preprint arXiv:2305.07895*.

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

Pan Lu, Swaroop Mishra, Tony 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. In *Advances in Neural Information Processing Systems*.

Ahmed Masry, Xuan Long Do, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. 2022. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2263–2279.

Arindam Mitra, Luciano Del Corro, Shweti Mahajan, Andres Codas, Clarisse Simoes, Sahaj Agarwal, Xuxi Chen, Anastasia Razdaibiedina, Erik Jones, Kriti Aggarwal, et al. 2023. Orca 2: Teaching small language models how to reason. *arXiv preprint arXiv:2311.11045*.

OpenAI. 2023. GPT-4 Technical Report.

OpenAI. 2023. GPT-4V(ision) System Card. `https://cdn.openai.com/papers/GPTV_System_Card.pdf`.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, pages 1–22.

Arkil Patel, Satwik Bhattamishra, and Navin Goyal. 2021. Are nlp models really able to solve simple math word problems? In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2080–2094.

Bernardino Romera-Paredes, Mohammadamin Barekatain, Alexander Novikov, Matej Balog, M Pawan Kumar, Emilien Dupont, Francisco JR Ruiz, Jordan S Ellenberg, Pengming Wang, Omar Fawzi, et al. 2023. Mathematical discoveries from program search with large language models. *Nature*, pages 1–3.

Kumar Shridhar, Jakub Macina, Mennatallah El-Assady, Tanmay Sinha, Manu Kapur, and Mrinmaya Sachan. 2022. Automatic generation of socratic subquestions for teaching math word problems. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4136–4149, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

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*.

Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Xixuan Song, et al. 2023a. Cogvlm: Visual expert for pretrained language models. *arXiv preprint arXiv:2311.03079*.

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

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.

Jiayang Wu, Wensheng Gan, Zefeng Chen, Shicheng Wan, and Philip S Yu. 2023a. Multimodal large language models: A survey. *arXiv preprint arXiv:2311.13165*.

Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Shaokun Zhang, Erkang Zhu, Beibin Li, Li Jiang, Xiaoyun Zhang, and Chi Wang. 2023b. Autogen: Enabling next-gen llm applications via multi-agent conversation framework. *arXiv preprint arXiv:2308.08155*.

Canwen Xu, Corby Rosset, Luciano Del Corro, Shweti Mahajan, Julian McAuley, Jennifer Neville, Ahmed Hassan Awadallah, and Nikhil Rao. 2023. Contrastive post-training large language models on data curriculum. *arXiv preprint arXiv:2310.02263*.

Zhicheng Yang, Jinghui Qin, Jiaqi Chen, and Xiaodan Liang. 2022. Unbiased math word problems benchmark for mitigating solving bias. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1401–1408, Seattle, United States. Association for Computational Linguistics.

Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, et al. 2023. mplug-owl: Modularization empowers large language models with multimodality. *arXiv preprint arXiv:2304.14178*.

Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. 2023. A survey on multimodal large language models. *arXiv preprint arXiv:2306.13549*.

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

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. 2023a. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. *arXiv preprint arXiv:2311.16502*.

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

Qiyuan Zhang, Lei Wang, Sicheng Yu, Shuohang Wang, Yang Wang, Jing Jiang, and Ee-Peng Lim. 2021. Noahqa: Numerical reasoning with interpretable graph question answering dataset. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4147–4161.

Xiaotian Zhang, Chunyang Li, Yi Zong, Zhengyu Ying, Liang He, and Xipeng Qiu. 2023a. Evaluating the performance of large language models on gaokao benchmark. *arXiv preprint arXiv:2305.12474*.

Yanzhe Zhang, Ruiyi Zhang, Jiuxiang Gu, Yufan Zhou, Nedim Lipka, Diyi Yang, and Tong Sun. 2023b. Llavar: Enhanced visual instruction tuning for text-rich image understanding. *arXiv preprint arXiv:2306.17107*.

Wanjun Zhong, Ruixiang Cui, Yiduo Guo, Yaobo Liang, Shuai Lu, Yanlin Wang, Amin Saied, Weizhu Chen, and Nan Duan. 2023. Agieval: A human-centric benchmark for evaluating foundation models. *arXiv preprint arXiv:2304.06364*.

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*.

A Appendix

A.1 Accuracy across Knowledge Points

In SceMQA, each problem is associated with a specific knowledge point. The individual accuracy on those knowledge points can be found in Figure 7 and 8. We can observe that the model generally performs better in chemistry and biology than in math and physics. Also, the worst-performed categories of knowledge points are generally related to image understanding (e.g., limits and continuity, optics) or calculation (e.g., one-variable data analysis, integration), which indicate the weaknesses of current MLLMs to some extent.

A.2 Features of SceMQA

To evaluate the difficulty of the problems in our benchmark, we utilize GPT-4 to respond to the questions within our dataset, as well as those from both a primary level and a college level benchmark. Figure 4 demonstrates the moderate difficulty level of our benchmark, positioning between the existing benchmark on primary and college levels. The example problems in SceMQA are located in Figure 5, with the following features:

Science Subjects Focusing on the core science subjects such as mathematics, physics, biology, and chemistry, our benchmark aligns with both existing text-only benchmarks, such as SciBench (Wang et al., 2023b), and major human exams like the GaoKao (i.e., Chinese national college entrance

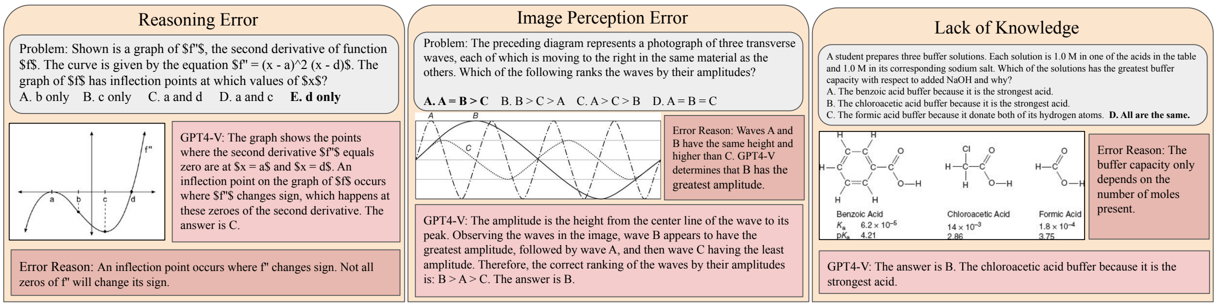

Figure 3: Example of errors made by GPT4-V on SceMQA.

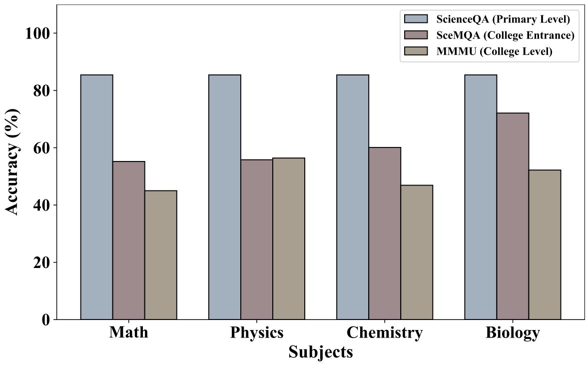

Figure 4: Comparison of GPT-4 performance across different benchmarks, illustrating the accuracy percentages achieved by GPT-4 in different subjects.

exam). To effectively address these problems, AI models must demonstrate a robust understanding of images, tables, and diagrams, coupled with deep domain knowledge to recall necessary formulae, theorems, and other elements for advanced reasoning. This presents a suitable challenge for current AI systems, testing their limits in areas typically reserved for advanced human cognition.

Solution Explanation We have meticulously annotated every problem in SceMQA. Almost all solutions (> 90%) are accompanied by detailed, human-verified explanations except for some straightforward solutions, as shown in Figure 5. These explanations are useful for identifying errors in model predictions and could also be instrumental in future supervised fine-tuning (SFT) (Ho et al., 2022; Hsieh et al., 2023) and few-shot prompting methodologies (Wei et al., 2022).

Identified Knowledge Category Additionally, each problem is associated with specific knowledge components within its subject, also shown in Figure 5. The availability of these components aids in building a knowledge state for the evaluated models, facilitating knowledge tracing and understanding the depth of the model's capabilities.

Question Variation Furthermore, our benchmark features a variety of questions based on the same image and context, as shown in Figure 6. Solving such kind of question sets has been demonstrated to be challenging for AI models (Liang and Zhang, 2021), where they usually fail to detect subtle differences among various questions related to the same context (Patel et al., 2021). This one-context multiple-questions setting can not only test the depth of understanding and reasoning capabilities of these AI models (Patel et al., 2021; Yang et al., 2022) but also have the potential to support advancements in Socratic learning (Shridhar et al., 2022) and interpretable reasoning (Zhang et al., 2021).

A.3 Data Collection Protocol

The data for SceMQA was meticulously sourced from publicly available online materials tailored for college entrance level tests in four key subjects: math (including calculus and statistics), biology, physics, and chemistry. In selecting these questions, our team of annotators strictly adhered to the licensing regulations of the source websites, ensuring no copyrighted material was included. This adherence to legal and ethical standards was a priority throughout the data collection process.

For the curation of SceMQA, we specify its intended use to ensure compatibility with the original access conditions. The dataset is designed for academic research and educational technology development. It is not intended for commercial use or outside of research contexts, especially considering that the data is derived from educational resources accessed for research purposes. This specification helps maintain ethical standards and respects the original access conditions of the sourced materials. We also asked annotators to carefully check

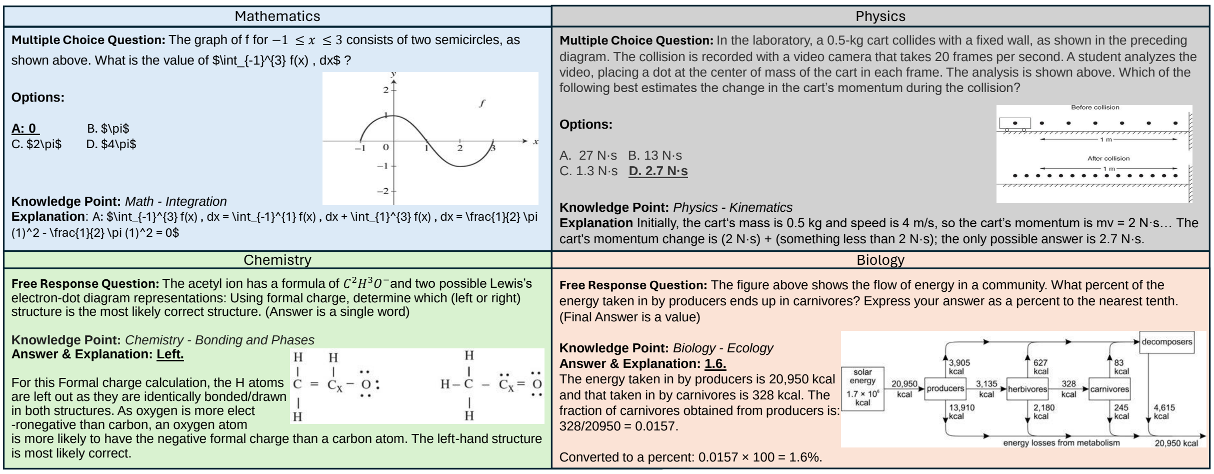

Figure 5: Example problems in SceMQA, which contains four scientific subjects - math, physics, chemistry and biology in two formats - multiple choice and free response.

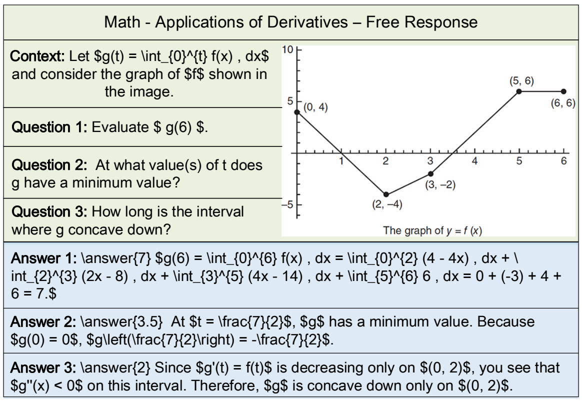

Figure 6: SceMQA contains multiple questions under the same context.

whether the data that was collected contained any personal identifier or offensive content and remove them if necessary.

Each problem within our dataset contains one image that is essential for solving the corresponding question, aligning with the multimodal nature of SceMQA. The problems are presented in two formats: multiple-choice and free-response. The multiple-choice questions offer 4 to 5 options, denoted by uppercase letters, a format consistent with other established benchmarks. Following previous studies (Hendrycks et al., 2021a; Lewkowycz et al., 2022), we transform all mathematical expressions into latex codes, making them easy to process for LLMs, as shown in Figure 5 and 6.

The free-response section includes calculation-based problems where answers are numerical values. This format is particularly advantageous for evaluation purposes, as the correctness of model-generated answers can be straightforwardly determined by checking the final numerical value. This approach is in line with other benchmarks like GSM8k, SciBench, and MMMU. Besides calculations, our benchmark diversifies with other free-response types like Yes-or-No and fill-in-the-blank questions. These formats not only broaden the range of question types but also maintain ease of evaluation through exact matching. Given these characteristics, accuracy will be the primary metric for assessing performance on our benchmark.

In terms of data features, each problem was thoroughly reviewed by annotators to ensure it aligned with the intended high school and pre-college difficulty level. Moreover, every problem is accompanied by a clear explanation of the answer and is tagged with the main knowledge point from predefined knowledge sets. These annotations and categorizations have been verified by domain experts, ensuring that each problem accurately reflects the intended educational content and difficulty.

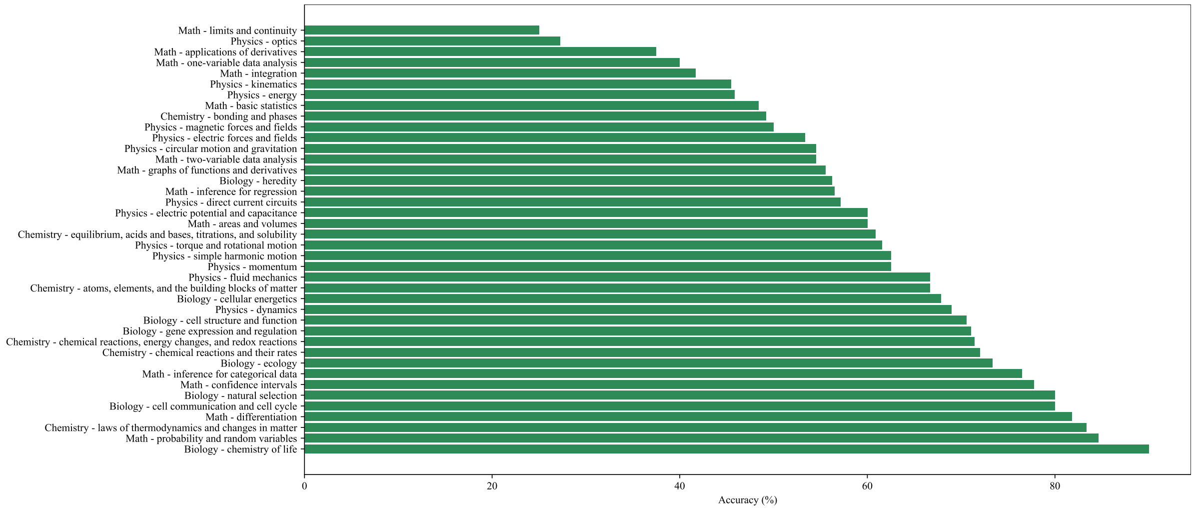

Figure 7: Accuracy distribution of GPT4-V on the knowledge points of SceMQA.

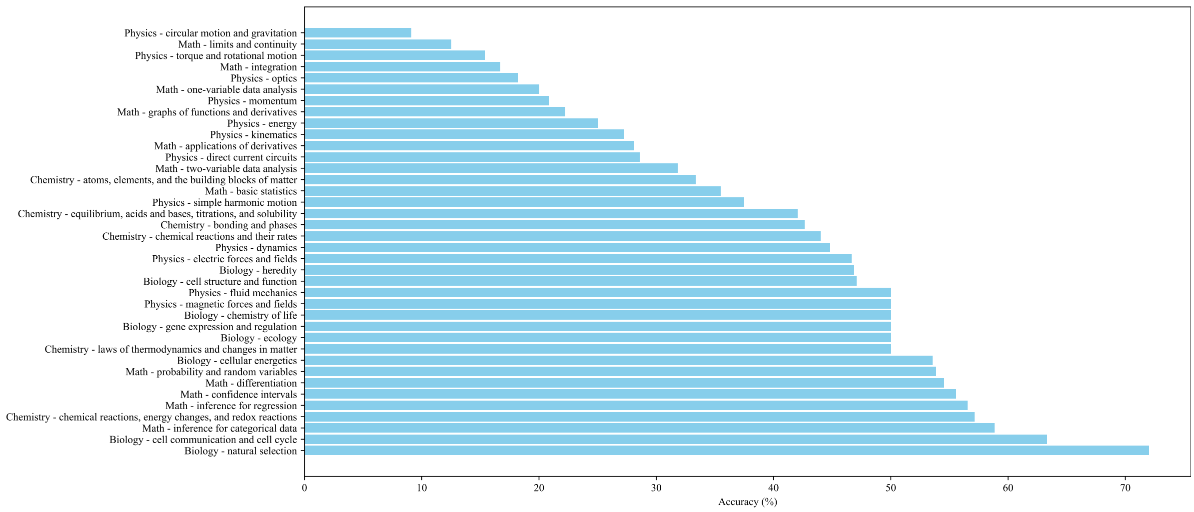

Figure 8: Accuracy distribution of Google Gemini on the knowledge points of SceMQA.