MATHVISTA: Evaluating Mathematical Reasoning of Foundation Models in Visual Contexts

Pan Lu[1,3]**, Hritik Bansal**[1]**, Tony Xia**[1]**, Jiacheng Liu**[2]**, Chunyuan Li**[3]**,**
Hannaneh Hajishirzi[2]**, Hao Cheng**[3]**, Kai-Wei Chang**[1]**, Michel Galley**[3]**, Jianfeng Gao**[3]
[1]UCLA, [2]University of Washington, [3]Microsoft Research, Redmond
https://mathvista.github.io

Abstract

Although Large Language Models (LLMs) and Large Multimodal Models (LMMs) exhibit impressive skills in various domains, their ability for mathematical reasoning within visual contexts has not been formally examined. Equipping LLMs and LMMs with this capability is vital for general-purpose AI assistants and showcases promising potential in education, data analysis, and scientific discovery. To bridge this gap, we present MATHVISTA, a benchmark designed to amalgamate challenges from diverse mathematical and visual tasks. We first taxonomize the key task types, reasoning skills, and visual contexts from the literature to guide our selection from 28 existing math-focused and visual question answering datasets. Then, we construct three new datasets, IQTest, FunctionQA, and PaperQA, to accommodate for missing types of visual contexts. The problems featured often require deep visual understanding beyond OCR or image captioning, and compositional reasoning with rich domain-specific tools, thus posing a notable challenge to existing models. We conduct a comprehensive evaluation of 11 prominent open-source and proprietary foundation models (LLMs, LLMs augmented with tools, and LMMs). The best-performing model, Multimodal Bard, achieves only 58% of human performance (34.8% vs 60.3%), indicating ample room for further improvement. Given this significant gap, MATHVISTA fuels future research in the development of general-purpose AI agents capable of tackling mathematically intensive and visually rich real-world tasks.

1 Introduction

The latest version is available at arXiv.

Mathematical reasoning is a cornerstone of human intelligence, demanding rigorous reasoning and problem-solving skills, along with domain-specific knowledge and multistep reasoning (Lightman et al., 2023). This complexity is not only observed in textual scenarios but also significantly in visual contexts. For instance, when assessing a child's mathematical and reasoning capabilities, problems are often designed to encompass visual contexts in addition to arithmetic calculations (Stipek & Iver, 1989; Pollitt et al., 2020). Concurrently, AI agents with strong mathematical reasoning capabilities within visual contexts have various real-world applications, such as solving complex problems in educational disciplines (Seo et al., 2015; Wang et al., 2017), helping analysts with logical queries about statistical data (Wu et al., 2023; Yang et al., 2023), and assisting in theorem proving and scientific discovery in advanced research fields (Taylor et al., 2022; Dong et al., 2023).

Numerous datasets have been curated to assess the mathematical reasoning abilities of AI systems, with most presented purely in text form. Some datasets such as ChartQA (Lu et al., 2021a; Dahlgren Lindström & Abraham, 2022; Masry et al., 2022) have explored mathematical reasoning in vision-language settings. However, these datasets tend to either focus on specific tasks, like math word problems, or particular visual contexts, such as geometry problems or bar charts. General-purpose
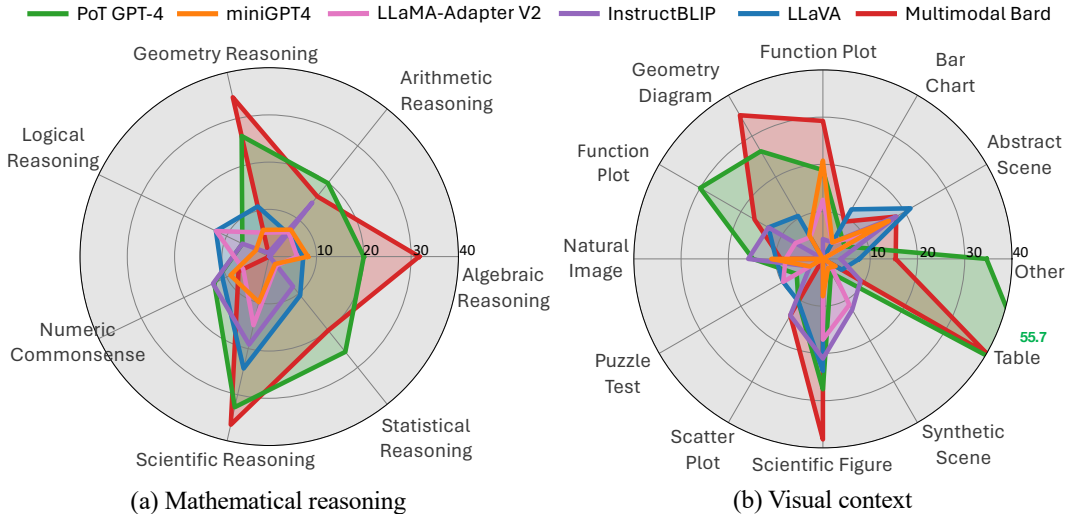
Figure 1: Accuracy scores of one leading LLM (PoT GPT-4) and five primary LMMs on our proposed MATHVISTA across mathematical reasoning and visual context types. Scores are adjusted by subtracting the random chance score and normalized to a range of $0 - 100$. PoT refers to program-of-thought prompting, and PoT GPT-4 is a textual LLM augmented with the caption and OCR text.

visual question answering (VQA) datasets on natural scenes contain only a small portion of questions necessitating mathematical reasoning, leaving a comprehensive investigation of vision-language reasoning within a mathematical framework largely unexplored. Large Language Models (LLMs) (OpenAI, 2022, 2023a) and Large Multimodal Models (LMMs) (Google, 2023; OpenAI, 2023b) have exhibited impressive capabilities across various domains. Recently, some studies have aimed to augment existing LLMs with mathematical and scientific reasoning capabilities in visual contexts via external tools (Lu et al., 2023a). However, the ability of these foundational models to perform mathematical reasoning within visual contexts has not been formally examined.

To address this, we introduce MATHVISTA, a consolidated **Math**ematical reasoning benchmark within **Vis**ual contexts. We propose a taxonomy as the guideline for constructing MATHVISTA: (1) we identify seven mathematical reasoning types: *algebraic reasoning*, *arithmetic reasoning*, *geometry reasoning*, *logical reasoning*, *numeric common sense*, *scientific reasoning*, and *statistical reasoning*; (2) we focus on five primary tasks: *figure question answering* (FQA), *geometry problem solving* (GPS), *math word problem* (MWP), *textbook question answering* (TQA), and *visual question answering* (VQA); and (3) we aim at encompassing a diverse array of visual contexts, including natural images, geometry diagrams, abstract scenes, synthetic scenes, as well as various figures, charts, and plots. In addition, we create three new datasets, IQTest, FunctionQA, and PaperQA, which address the missing visual domains and are tailored to evaluate logical reasoning on puzzle test figures, algebraic reasoning over functional plots, and scientific reasoning with academic paper figures, respectively. MATHVISTA also incorporates 9 math-targeted question answering (MathQA) datasets and 19 VQA datasets from the literature, which significantly enrich the diversity and complexity of visual perception and mathematical reasoning challenges within our benchmark. MATHVISTA consists of 6,141 examples, with 736 of them being newly annotated (Table 1). To facilitate fine-grained metrics, examples are further enriched with metadata—covering question, answer, task, grade level, visual context, and reasoning skills. More details on data collection can be found in §2, §B, and §C.

We conduct thorough experiments on MATHVISTA to evaluate the reasoning abilities of 11 foundational models known for their leading performance in mathematical and multimodal reasoning. This ensemble includes three LLMs, seven open-source LMMs, and one proprietary LMM. For LLMs, we examine zero-shot and few-shot settings using two prompting strategies: chain-of-thought (CoT) (Wei et al., 2022b) and program-of-thought (PoT) (Chen et al., 2022b). We also explore their augmented variants with image captions and OCR text from off-the-shelf visual models. Additionally, we establish a human performance baseline by engaging qualified human annotators with a high school diploma or higher. Given that MATHVISTA features advanced topics such as college curricula and scientific reasoning, it proves to be a challenging benchmark, with human performance reaching only 60.3% accuracy.
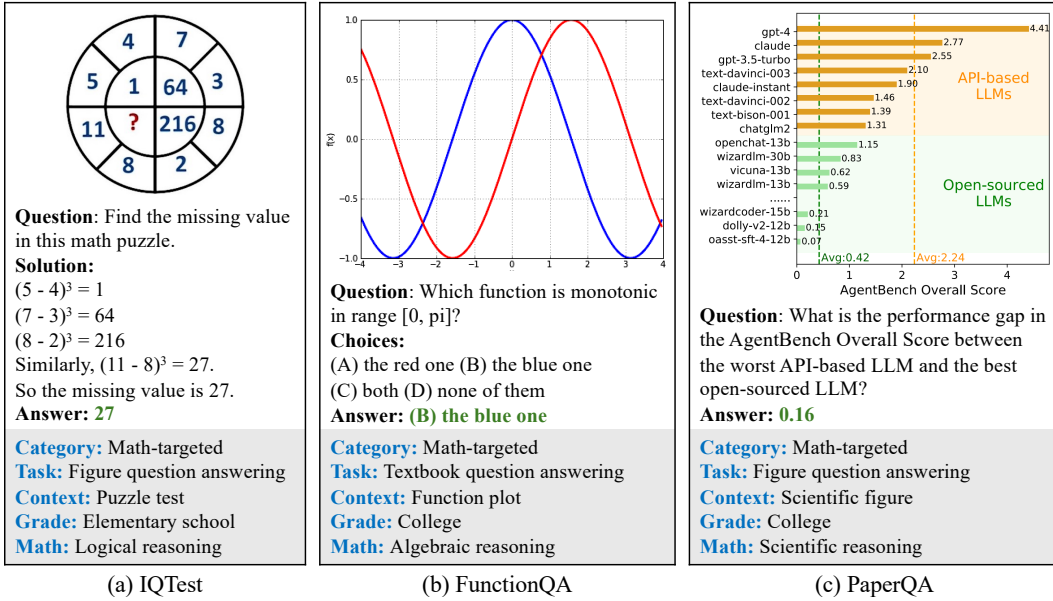
(a) IQTest

Question: Find the missing value in this math puzzle.
Solution:
$(5 - 4)^3 = 1$
$(7 - 3)^3 = 64$
$(8 - 2)^3 = 216$
Similarly, $(11 - 8)^3 = 27$.
So the missing value is 27.
Answer: 27

Category: Math-targeted
Task: Figure question answering
Context: Puzzle test
Grade: Elementary school
Math: Logical reasoning

(b) FunctionQA

Question: Which function is monotonic in range [0, pi]?
Choices:
(A) the red one (B) the blue one
(C) both (D) none of them
Answer: (B) the blue one

Category: Math-targeted
Task: Textbook question answering
Context: Function plot
Grade: College
Math: Algebraic reasoning

(c) PaperQA

Question: What is the performance gap in the AgentBench Overall Score between the worst API-based LLM and the best open-sourced LLM?
Answer: 0.16

Category: Math-targeted
Task: Figure question answering
Context: Scientific figure
Grade: College
Math: Scientific reasoning

Figure 2: Examples of our newly annotated datasets: IQTest, FunctionQA, and PaperQA.

Statistic	Number
Total questions	6,141
- multiple-choice questions	3,392 (55.2%)
- Free-form questions	2,749 (44.8%)
- Questions with annotations	5,261 (85.6%)
- Questions newly annotated	736 (12.0%)
Unique number of images	5,487
Unique number of questions	4,746
Unique number of answers	1,464
Source datasets	31
- Existing VQA datasets	19
- Existing MathQA datasets	9
- Our newly annotated datasets	3
Visual context (image) classes	19
Maximum question length	213
Maximum answer length	27
Maximum choice number	8
Average question length	15.6
Average answer length	1.2
Average choice number	3.4

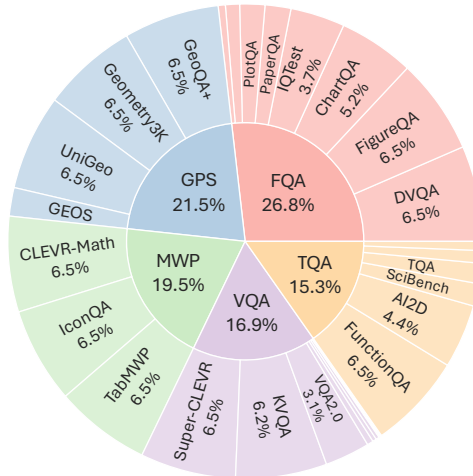Table 1: Key statistics of MATHVISTA.

Figure 3: Source dataset distribution of MATHVISTA. FQA: figure question answering, GPS: geometry problem solving, MWP: math word problem, TQA: textbook question answering, VQA: visual question answering.

Our results show that CoT GPT-4, the best-performing LLM without visual tool augmentations, achieves an overall accuracy of 29.2%, while the best performing LMM, Multimodal Bard, achieves 34.8% (§3). Notably, Multimodal Bard attains only 58% of human performance (34.8% vs 60.3%), indicating substantial room for improvement. When augmented with Bard image captions and detected OCR text, PoT GPT-4 obtains 33.9%, closely matching Multimodal Bard (§F.1). Key results are highlighted in Figure 1. Further analysis indicates that model failures arise from incorrect calculations and hallucinations caused by visual perception and textual reasoning (§F.2).

2 MATHVISTA Benchmark

2.1 Data Analysis

The main statistics of MATHVISTA are presented in Table 1. There are two types of questions: multiple-choice and free-form. Answers to free-form questions are categorized as integers, floating numbers, or lists. The large unique number of images, questions, and answers ensures pattern diversity

in MATHVISTA. MATHVISTA is derived from 31 source datasets, including three newly annotated datasets to address the missing types of mathematical reasoning over specific visual contexts. Dataset examples in Table 4 (§B.2) highlight the richness of mathematical reasoning involved. Examples in §B.3 demonstrate the diverse visual contexts present in MATHVISTA. Further details on data analysis are available in §D.

3   Experiments Results

3.1   Evaluation Protocols

Recent LLMs and LMMs have been instructed to generate long responses in conventional settings instead of short text. Therefore, we propose a new strategy for benchmarking MATHVISTA, unlike using human-designed or template matching rules (Lu et al., 2022). The evaluation process consists of three stages: *response generation*, *answer extraction*, and *score calculation*. Initially, the baselines generate responses given the input query, which incorporates the task description, the question, choices, and metadata, using the template defined in Table 9 (§E.3). Next, the short answer text is extracted from the detailed response. Then, we propose an answer extractor (§E.2) based on LLMs such as GPT-4, inspired by its remarkable ability for text processing (Wei et al., 2022b). A preliminary study of 200 examples shows that GPT-4 can extract the answer text with more than 99.5% accuracy. Finally, the extracted answer is normalized to a required answer format (e.g., an option letter or an integer), and the target metric scores are computed. Taking advantage of the fact that the instances in MATHVISTA are either multiple-choice questions for textual answers or free-form questions for numerical answers, accuracy scores are used as metrics for deterministic evaluation.

3.2   Experimental Setup

We evaluate the models on MATHVISTA under three setups: (a) *Text-Only LLMs* including ChatGPT (OpenAI, 2022), GPT-4 (OpenAI, 2023a), and Claude-2 (Anthropic, 2023) in zero-shot and two-shot settings with Chain-of-Thought (CoT) (Wei et al., 2022b) and Program-of-Thought (PoT) (Chen et al., 2022b), (b) *Augmented-LLMs* where the LLMs are provided with additional visual information including the generated image captions from Multimodal Bard (Google, 2023) and the detected OCR text from EasyOCR (JaidedAI, 2020), (c) *LMMs* that include open models such as IDEFICS-9B (Laurençon et al., 2023), mPLUG-OWL-LLaMA-7B (Ye et al., 2023), miniGPT-4-LLaMA-2-7B (Zhu et al., 2023a), LLaMA-Adapter-V2-7B (Zhang et al., 2023a), InstructBLIP-Vicuna-7B (Dai et al., 2023), LLaVA-LLaMA-2-13B (Liu et al., 2023a), LLaVAR Zhang et al. (2023c), and closed model Bard. We provide the prompts for LLMs and the hyperparameters used for LMMs in §E.

3.3   Experimental Results

We compare the performance of several models, including Text-only LLMs, Augmented LLMs, and LMMs on MATHVISTA in Table 2. We include the random chance (i.e., one of the four options in multiple-choice questions, and empty in the free-form questions) and frequency guess (Appendix §E.1) as naive baselines. Additionally, we established a human performance baseline using Amazon Mechanical Turk. Eligible human annotators must have a satisfactory annotating history, successfully pass qualification examples, and possess a high school degree or higher. We asked each annotator to complete five questions within 20 minutes. Further details can be found in §E.6.

We find that the Multimodal Bard is currently the best model on MATHVISTA with 34.8%, while humans perform 60.3%. This highlights that there is a significant scope for further improvements on our benchmark. Among text-only LLMs, all models perform better than the random baseline, and 2-shot GPT-4 with chain-of-thought prompting achieves 29.2%. The inability of the text-only LLMs to perform well indicates that our dataset requires the models to reason over the visual contexts for good performance. With access to image captions and detected OCR text, we find that augmented LLMs perform better than text-only LLMs on MATHVISTA. Specifically, the best performing method is 2-shot GPT-4 with program-of-thought prompting with 33.9%.

As shown in Table 2, on the LMM side, we find that the performance of the open-source models (IDEFICS to LLaVA) achieves underwhelming performance on MATHVISTA. This can be attributed to their lack of math reasoning capabilities, text recognition (useful for math word problems), shape detection (useful for geometrical problems), and chart understanding. Notably, these models utilize

Model	Input	ALL	FQA	GPS	MWP	TQA	VQA	ALG	ARI	GEO	LOG	NUM	SCI	STA
Heuristics baselines														
Random chance	-	17.9	18.2	21.6	3.8	19.6	26.3	21.7	14.7	20.1	13.5	8.3	17.2	16.3
Frequent guess	-	26.3	22.7	34.1	20.4	31.0	24.6	33.1	18.7	31.4	24.3	19.4	32.0	20.9
Large Language Models (LLMs)														
Zero-shot ChatGPT	Q only	23.5	21.9	26.9	9.1	38.6	23.5	27.7	15.9	25.7	21.6	9.9	41.5	20.5
Zero-shot GPT-4	Q only	26.1	22.3	37.0	7.0	39.2	27.4	33.6	17.4	35.6	16.2	9.2	45.8	19.5
Zero-shot Claude-2	Q only	26.4	21.9	34.1	13.4	36.1	29.1	32.8	20.4	33.3	13.5	12.1	36.4	20.5
2-shot CoT Claude-2	Q only	24.4	18.6	29.8	9.7	33.5	34.1	29.2	19.0	28.0	5.4	13.9	36.9	18.9
2-shot CoT ChatGPT	Q only	26.8	20.1	36.5	8.6	44.9	28.5	35.6	17.0	33.5	21.6	14.6	45.9	17.9
2-shot CoT GPT-4	Q only	29.2	20.1	44.7	8.6	46.2	31.3	41.6	19.3	41.0	18.9	13.9	47.5	18.9
2-shot PoT ChatGPT	Q only	25.1	19.0	30.8	16.1	38.0	25.7	29.9	19.8	29.3	24.3	19.4	38.5	16.9
2-shot PoT GPT-4	Q only	26.0	20.1	33.2	8.1	44.9	28.5	32.7	16.7	31.0	24.3	13.2	48.4	18.3
Augmented Large Language Models (Augmented-LLMs)														
2-shot CoT Claude-2	Q, I_c, I_t	33.2	26.0	31.7	35.5	48.1	30.2	32.4	32.3	33.0	16.2	17.4	54.9	36.2
2-shot CoT ChatGPT	Q, I_c, I_t	33.2	27.5	29.3	36.0	49.4	29.1	31.0	32.9	31.0	16.2	17.4	50.8	37.2
2-shot CoT GPT-4	Q, I_c, I_t	33.2	27.9	31.7	31.2	51.9	28.5	33.5	30.9	32.2	13.5	12.5	58.2	37.6
2-shot PoT ChatGPT	Q, I_c, I_t	26.8	24.5	26.4	23.7	33.5	27.9	27.8	26.1	28.0	18.9	13.2	33.6	29.9
2-shot PoT GPT-4	Q, I_c, I_t	33.9	30.1	39.4	30.6	39.9	31.3	37.4	31.7	41.0	18.9	20.1	44.3	37.9
Large Multimodal Models (LMMs)														
IDEFICS-9B-Instruct	Q, I	19.8	21.6	21.1	6.5	25.9	24.0	22.1	15.0	19.8	18.9	9.9	24.6	18.1
mPLUG-Owl-LLaMA-7B	Q, I	22.2	22.7	23.6	10.2	27.2	27.9	23.6	19.2	23.9	13.5	12.7	26.3	21.4
miniGPT4-LLaMA-2-7B	Q, I	23.1	18.6	26.0	13.4	30.4	30.2	28.1	21.0	24.7	16.2	16.7	25.4	17.9
LLaMA-Adapter-V2-7B	Q, I	23.9	21.2	25.5	11.3	32.3	31.8	26.3	20.4	24.3	24.3	13.9	29.5	18.3
LLaVAR	Q, I	25.2	21.9	25.0	16.7	34.8	30.7	24.2	22.1	23.0	13.5	15.3	42.6	21.9
InstructBLIP-Vicuna-7B	Q, I	25.3	23.1	20.7	18.3	32.3	35.2	21.8	27.1	20.7	18.9	20.4	33.0	23.1
LLaVA-LLaMA-2-13B	Q, I	26.1	26.8	29.3	16.1	32.3	26.3	27.3	20.1	28.8	24.3	18.3	37.3	25.1
Multimodal Bard	Q, I	34.8	26.0	47.1	29.6	48.7	26.8	46.5	28.6	47.8	13.5	14.9	47.5	33.0
Human														
Human performance	Q, I	60.3	59.7	48.4	73.0	63.2	55.9	50.9	59.2	51.4	40.7	53.8	64.9	63.9

Table 2: Accuracy scores on the *testmini* subset of MATHVISTA. Input: Q: question, I: image, I_c: image caption, I_t: OCR text detected in the image. ALL: overall accuracy. Task types: FQA: figure question answering, GPS: geometry problem solving, MWP: math word problem, TQA: textbook question answering, VQA: visual question answering. Mathematical reasoning types: ALG: algebraic reasoning, ARI: arithmetic reasoning, GEO: geometry reasoning, LOG: logical reasoning, NUM: numeric commonsense, SCI: scientific reasoning, STA: statistical reasoning. The highest scores among models in each section and overall are highlighted in blue and red, respectively.

different model architectures for processing the vision (e.g., OpenCLIP, CLIP, Vit-G) and language (e.g., LLaMA-1, LLaMA-2), different alignment strategies (e.g., MLP projection in LLaVA, Q-former in InstructBLIP, visual abstractor in mPLUGOwl), and instruction tuning data (e.g., 150K instruction-response pairs from LLaVA data, 3,500 instruction-response pairs from miniGPT-4). While fine-tuned with instruction-following data from text-rich images, LLaVAR does not perform well, indicating that strong text recognition abilities do not guarantee high performance on MATHVISTA, which requires comprehensive visual perception and mathematical reasoning. This underscores that there are immense possibilities for innovations in model, data, or training objectives to improve the zero-shot performance of LMMs on MATHVISTA.

4 Conclusion

In this work, we introduce MATHVISTA, a comprehensive and robust benchmark for evaluating mathematical reasoning capabilities within visual contexts in foundation models. MATHVISTA comprises 6,141 examples, sourced from three of our newly created datasets and 28 existing ones. Our extensive experiments on LLMs and LMMs reveal that the best-performing model, Multimodal Bard, achieves only 58% of human performance, underscoring the challenges in integrating mathematical reasoning with visual comprehension. Further analysis highlights two future directions: (1) proposing general-purpose LMMs with enhanced visual perception and mathematical reasoning; and (2) developing augmented LLMs powered by external tools to improve visual perception and domain-specific reasoning abilities. Another direction is evaluating the integrity of model explanations at scale. Our analysis, through human evaluation, represents a first step towards this goal.

References

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35:23716–23736, 2022. 12

Aida Amini, Saadia Gabriel, Shanchuan Lin, Rik Koncel-Kedziorski, Yejin Choi, and Hannaneh Hajishirzi. Mathqa: Towards interpretable math word problem solving with operation-based formalisms. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL)*, pp. 2357–2367, 2019. 12

Anthropic. Claude 2, 2023. URL https://www.anthropic.com/index/claude-2. 4, 12

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. VQA: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pp. 2425–2433, 2015. 12, 18

Anas Awadalla, Irena Gao, Josh Gardner, Jack Hessel, Yusuf Hanafy, Wanrong Zhu, Kalyani Marathe, Yonatan Bitton, Samir Gadre, Shiori Sagawa, et al. OpenFlamingo: An open-source framework for training large autoregressive vision-language models. *arXiv preprint arXiv:2308.01390*, 2023. 12

Yonatan Bitton, Hritik Bansal, Jack Hessel, Rulin Shao, Wanrong Zhu, Anas Awadalla, Josh Gardner, Rohan Taori, and Ludwig Schimdt. VisIT-Bench: A benchmark for vision-language instruction following inspired by real-world use. *arXiv preprint arXiv:2308.06595*, 2023. 12

Nitzan Bitton-Guetta, Yonatan Bitton, Jack Hessel, Ludwig Schmidt, Yuval Elovici, Gabriel Stanovsky, and Roy Schwartz. Breaking common sense: WHOOPS! A vision-and-language benchmark of synthetic and compositional images. *arXiv preprint arXiv:2303.07274*, 2023. 12

Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021. 12

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020. 12

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*, 2023. 12

Jie Cao and Jing Xiao. An augmented benchmark dataset for geometric question answering through dual parallel text encoding. In *Proceedings of the 29th International Conference on Computational Linguistics*, pp. 1511–1520, 2022. 12, 18

Shuaichen Chang, David Palzer, Jialin Li, Eric Fosler-Lussier, and Ningchuan Xiao. MapQA: A dataset for question answering on choropleth maps. *arXiv preprint arXiv:2211.08545*, 2022. 12, 18

Jiaqi Chen, Tong Li, Jinghui Qin, Pan Lu, Liang Lin, Chongyu Chen, and Xiaodan Liang. UniGeo: Unifying geometry logical reasoning via reformulating mathematical expression. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 3313–3323, 2022a. 12, 18

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*, 2021. 12

Wenhu Chen, Xueguang Ma, Xinyi Wang, and William W Cohen. Program of thoughts prompting: Disentangling computation from reasoning for numerical reasoning tasks. *arXiv preprint arXiv:2211.12588*, 2022b. 2, 4

Wenhu Chen, Ming Yin, Max Ku, Elaine Wan, Xueguang Ma, Jianyu Xu, Tony Xia, Xinyi Wang, and Pan Lu. TheoremQA: A theorem-driven question answering dataset. *arXiv preprint arXiv:2305.12524*, 2023. 18

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021. 12

Adam Dahlgren Lindström and Savitha Sam Abraham. CLEVR-Math: A dataset for compositional language, visual and mathematical reasoning. In *16th International Workshop on Neural-Symbolic Learning and Reasoning, NeSy 2022, Windsor, UK, september 28-30, 2022.*, volume 3212. CEUR-WS, 2022. 1, 12, 18

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. InstructBLIP: Towards general-purpose vision-language models with instruction tuning, 2023. 4, 12

Qingxiu Dong, Li Dong, Ke Xu, Guangyan Zhou, Yaru Hao, Zhifang Sui, and Furu Wei. Large language model for science: A study on P vs. NP. *arXiv preprint arXiv:2309.05689*, 2023. 1

Lingyue Fu, Huacan Chai, Shuang Luo, Kounianhua Du, Weiming Zhang, Longteng Fan, Jiayi Lei, Renting Rui, Jianghao Lin, Yuchen Fang, et al. CodeApex: A bilingual programming evaluation benchmark for large language models. *arXiv preprint arXiv:2309.01940*, 2023. 12

Google. Bard, 2023. URL https://bard.google.com/. 2, 4, 12

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the V in VQA matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6904–6913, 2017. 12, 18

Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P Bigham. VizWiz grand challenge: Answering visual questions from blind people. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3608–3617, 2018. 12, 18

Wenlong Huang, Pieter Abbeel, Deepak Pathak, and Igor Mordatch. Language models as zero-shot planners: Extracting actionable knowledge for embodied agents. In *International Conference on Machine Learning*, pp. 9118–9147. PMLR, 2022. 12

Yuzhen Huang, Yuzhuo Bai, Zhihao Zhu, Junlei Zhang, Jinghan Zhang, Tangjun Su, Junteng Liu, Chuancheng Lv, Yikai Zhang, Jiayi Lei, et al. C-eval: A multi-level multi-discipline chinese evaluation suite for foundation models. *arXiv preprint arXiv:2305.08322*, 2023. 12

JaidedAI. EasyOCR: Ready-to-use OCR, 2020. URL https://github.com/JaidedAI/EasyOCR. 4

Anya Ji, Noriyuki Kojima, Noah Rush, Alane Suhr, Wai Keen Vong, Robert D Hawkins, and Yoav Artzi. Abstract visual reasoning with tangram shapes. *arXiv preprint arXiv:2211.16492*, 2022. 12

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38, 2023. 29

Kushal Kafle, Brian Price, Scott Cohen, and Christopher Kanan. DVQA: Understanding data visualizations via question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5648–5656, 2018. 12, 18

Samira Ebrahimi Kahou, Vincent Michalski, Adam Atkinson, Ákos Kádár, Adam Trischler, and Yoshua Bengio. FigureQA: An annotated figure dataset for visual reasoning. *arXiv preprint arXiv:1710.07300*, 2017. 12, 18

Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. A diagram is worth a dozen images. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*, pp. 235–251. Springer, 2016. 12, 18

Aniruddha Kembhavi, Minjoon Seo, Dustin Schwenk, Jonghyun Choi, Ali Farhadi, and Hannaneh Hajishirzi. Are you smarter than a sixth grader? Textbook question answering for multimodal machine comprehension. In *Proceedings of the IEEE Conference on Computer Vision and Pattern recognition*, pp. 4999–5007, 2017. 12, 18

Jason J Lau, Soumya Gayen, Asma Ben Abacha, and Dina Demner-Fushman. A dataset of clinically generated visual questions and answers about radiology images. *Scientific data*, 5(1):1–10, 2018. 12, 18

Hugo Laurençon, Lucile Saulnier, Léo Tronchon, Stas Bekman, Amanpreet Singh, Anton Lozhkov, Thomas Wang, Siddharth Karamcheti, Alexander M. Rush, Douwe Kiela, Matthieu Cord, and Victor Sanh. OBELICS: An open web-scale filtered dataset of interleaved image-text documents, 2023. 4

Kenton Lee, Mandar Joshi, Iulia Raluca Turc, Hexiang Hu, Fangyu Liu, Julian Martin Eisenschlos, Urvashi Khandelwal, Peter Shaw, Ming-Wei Chang, and Kristina Toutanova. Pix2Struct: Screenshot parsing as pretraining for visual language understanding. In *International Conference on Machine Learning*, pp. 18893–18912. PMLR, 2023. 12

Zhuowan Li, Xingrui Wang, Elias Stengel-Eskin, Adam Kortylewski, Wufei Ma, Benjamin Van Durme, and Alan L Yuille. Super-CLEVR: A virtual benchmark to diagnose domain robustness in visual reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14963–14973, 2023. 12, 18

Thomas Liao, Rohan Taori, Inioluwa Deborah Raji, and Ludwig Schmidt. Are we learning yet? A meta review of evaluation failures across machine learning. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021. 12

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. *arXiv preprint arXiv:2305.20050*, 2023. 1

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pp. 740–755. Springer, 2014. 12

Fangyu Liu, Francesco Piccinno, Syrine Krichene, Chenxi Pang, Kenton Lee, Mandar Joshi, Yasemin Altun, Nigel Collier, and Julian Martin Eisenschlos. MatCha: Enhancing visual language pretraining with math reasoning and chart derendering. *arXiv preprint arXiv:2212.09662*, 2022. 12

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*, 2023a. 4, 12

Xiao Liu, Hao Yu, Hanchen Zhang, Yifan Xu, Xuanyu Lei, Hanyu Lai, Yu Gu, Hangliang Ding, Kaiwen Men, Kejuan Yang, et al. AgentBench: Evaluating LLMs as agents. *arXiv preprint arXiv:2308.03688*, 2023b. 12

Yuliang Liu, Zhang Li, Hongliang Li, Wenwen Yu, Mingxin Huang, Dezhi Peng, Mingyu Liu, Mingrui Chen, Chunyuan Li, Lianwen Jin, et al. On the hidden mystery of OCR in large multimodal models. *arXiv preprint arXiv:2305.07895*, 2023c. 12

Pan Lu, Ran Gong, Shibiao Jiang, Liang Qiu, Siyuan Huang, Xiaodan Liang, and Song-Chun Zhu. Inter-GPS: Interpretable geometry problem solving with formal language and symbolic reasoning. In *The 59th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2021a. 1, 12, 18

Pan Lu, Liang Qiu, Jiaqi Chen, Tony Xia, Yizhou Zhao, Wei Zhang, Zhou Yu, Xiaodan Liang, and Song-Chun Zhu. IconQA: A new benchmark for abstract diagram understanding and visual language reasoning. In *The 35th Conference on Neural Information Processing Systems (NeurIPS) Track on Datasets and Benchmarks*, 2021b. 12, 18

Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *The 36th Conference on Neural Information Processing Systems (NeurIPS)*, 2022. 4, 12, 18

Pan Lu, Baolin Peng, Hao Cheng, Michel Galley, Kai-Wei Chang, Ying Nian Wu, Song-Chun Zhu, and Jianfeng Gao. Chameleon: Plug-and-play compositional reasoning with large language models. In *The 37th Conference on Neural Information Processing Systems (NeurIPS)*, 2023a. 2

Pan Lu, Liang Qiu, Kai-Wei Chang, Ying Nian Wu, Song-Chun Zhu, Tanmay Rajpurohit, Peter Clark, and Ashwin Kalyan. Dynamic prompt learning via policy gradient for semi-structured mathematical reasoning. In *International Conference on Learning Representations (ICLR)*, 2023b. 18

Pan Lu, Liang Qiu, Wenhao Yu, Sean Welleck, and Kai-Wei Chang. A survey of deep learning for mathematical reasoning. In *The 61st Annual Meeting of the Association for Computational Linguistics (ACL)*, 2023c. 12, 29

Ahmed Masry, Xuan Long Do, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. ChartQA: A benchmark for question answering about charts with visual and logical reasoning. In *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 2263–2279, 2022. 1, 12, 18

Ahmed Masry, Parsa Kavehzadeh, Xuan Long Do, Enamul Hoque, and Shafiq Joty. UniChart: A universal vision-language pretrained model for chart comprehension and reasoning. *arXiv preprint arXiv:2305.14761*, 2023. 12

Minesh Mathew, Viraj Bagal, Rubèn Tito, Dimosthenis Karatzas, Ernest Valveny, and CV Jawahar. InfographicsVQA. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 1697–1706, 2022. 12, 18

Nitesh Methani, Pritha Ganguly, Mitesh M Khapra, and Pratyush Kumar. PlotQA: Reasoning over scientific plots. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 1527–1536, 2020. 12, 18

Swaroop Mishra, Matthew Finlayson, Pan Lu, Leonard Tang, Sean Welleck, Chitta Baral, Tanmay Rajpurohit, Oyvind Tafjord, Ashish Sabharwal, Peter Clark, and Ashwin Kalyan. LILA: A unified benchmark for mathematical reasoning. In *The 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2022. 12

Shaghayegh Mobasher, Ghazal Zamaninejad, Maryam Hashemi, Melika Nobakhtian, and Sauleh Eetemadi. ParsVQA-Caps: A benchmark for visual question answering and image captioning in persian. *people*, 101:404, 2022. 12

Harsha Nori, Nicholas King, Scott Mayer McKinney, Dean Carignan, and Eric Horvitz. Capabilities of GPT-4 on medical challenge problems. *arXiv preprint arXiv:2303.13375*, 2023. 12

OpenAI. Chatgpt, 2022. URL https://openai.com/blog/chatgpt. 2, 4, 12

OpenAI. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023a. 2, 4, 12

OpenAI. GPT-4V(ision) system card, 2023b. URL https://openai.com/research/gpt-4v-system-card. 2

Rachel Pollitt, Caroline Cohrssen, and Wee Tiong Seah. Assessing spatial reasoning during play: Educator observations, assessment and curriculum planning. *Mathematics Education Research Journal*, 32(2):331–363, 2020. 1

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. LAION-5B: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294, 2022. 12

Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. A-OKVQA: A benchmark for visual question answering using world knowledge. In *European Conference on Computer Vision*, pp. 146–162. Springer, 2022. 12, 18

Minjoon Seo, Hannaneh Hajishirzi, Ali Farhadi, Oren Etzioni, and Clint Malcolm. Solving geometry problems: Combining text and diagram interpretation. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pp. 1466–1476, 2015. 1, 12, 18

Sanket Shah, Anand Mishra, Naganand Yadati, and Partha Pratim Talukdar. KVQA: Knowledge-aware visual question answering. In *Proceedings of the AAAI conference on artificial intelligence*, pp. 8876–8884, 2019. 12, 18

Wenqi Shao, Yutao Hu, Peng Gao, Meng Lei, Kaipeng Zhang, Fanqing Meng, Peng Xu, Siyuan Huang, Hongsheng Li, Yu Qiao, et al. Tiny LVLM-eHub: Early multimodal experiments with bard. *arXiv preprint arXiv:2308.03729*, 2023. 12

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual Captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2556–2565, 2018. 12

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards VQA models that can read. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8317–8326, 2019. 12, 18

Deborah Stipek and Douglas Mac Iver. Developmental change in children's assessment of intellectual competence. *Child development*, pp. 521–538, 1989. 1

Liangtai Sun, Yang Han, Zihan Zhao, Da Ma, Zhennan Shen, Baocai Chen, Lu Chen, and Kai Yu. SciEval: A multi-level large language model evaluation benchmark for scientific research. *arXiv preprint arXiv:2308.13149*, 2023. 12

Ross Taylor, Marcin Kardas, Guillem Cucurull, Thomas Scialom, Anthony Hartshorn, Elvis Saravia, Andrew Poulton, Viktor Kerkez, and Robert Stojnic. Galactica: A large language model for science. *arXiv preprint arXiv:2211.09085*, 2022. 1

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. LLaMA: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023. 12

Xiaoxuan Wang, Ziniu Hu, Pan Lu, Yanqiao Zhu, Jieyu Zhang, Satyen Subramaniam, Arjun R Loomba, Shichang Zhang, Yizhou Sun, and Wei Wang. SciBench: Evaluating college-level scientific problem-solving abilities of large language models. *arXiv preprint arXiv:2307.10635*, 2023. 12, 18

Yan Wang, Xiaojiang Liu, and Shuming Shi. Deep neural solver for math word problems. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 845–854, 2017. 1

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*, 2022a. 12

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Chi, Quoc Le, and Denny Zhou. Chain of thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*, 2022b. 2, 4

Shijie Wu, Ozan Irsoy, Steven Lu, Vadim Dabravolski, Mark Dredze, Sebastian Gehrmann, Prabhanjan Kambadur, David Rosenberg, and Gideon Mann. BloombergGPT: A large language model for finance. *arXiv preprint arXiv:2303.17564*, 2023. 1

Peng Xu, Wenqi Shao, Kaipeng Zhang, Peng Gao, Shuo Liu, Meng Lei, Fanqing Meng, Siyuan Huang, Yu Qiao, and Ping Luo. LVLM-eHub: A comprehensive evaluation benchmark for large vision-language models. *arXiv preprint arXiv:2306.09265*, 2023. 12

Hongyang Yang, Xiao-Yang Liu, and Christina Dan Wang. FinGPT: Open-source financial large language models. *arXiv preprint arXiv:2306.06031*, 2023. 1

Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, et al. mPlug-Owl: Modularization empowers large language models with multimodality. *arXiv preprint arXiv:2304.14178*, 2023. 4, 12

Da Yin, Liunian Harold Li, Ziniu Hu, Nanyun Peng, and Kai-Wei Chang. Broaden the vision: Geo-diverse visual commonsense reasoning. *arXiv preprint arXiv:2109.06860*, 2021. 12

Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. MM-Vet: Evaluating large multimodal models for integrated capabilities. *arXiv preprint arXiv:2308.02490*, 2023. 12

Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. From recognition to cognition: Visual commonsense reasoning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6720–6731, 2019. 12

Renrui Zhang, Jiaming Han, Aojun Zhou, Xiangfei Hu, Shilin Yan, Pan Lu, Hongsheng Li, Peng Gao, and Qiao Yu. LLaMA-Adapter: Efficient fine-tuning of language models with zero-init attention. *arXiv preprint arXiv:2303.16199*, 2023a. 4

Xiaoman Zhang, Chaoyi Wu, Ziheng Zhao, Weixiong Lin, Ya Zhang, Yanfeng Wang, and Weidi Xie. PMC-VQA: Visual instruction tuning for medical visual question answering. *arXiv preprint arXiv:2305.10415*, 2023b. 12, 18

Yanzhe Zhang, Ruiyi Zhang, Jiuxiang Gu, Yufan Zhou, Nedim Lipka, Diyi Yang, and Tong Sun. LLaVAR: Enhanced visual instruction tuning for text-rich image understanding. *arXiv preprint arXiv:2306.17107*, 2023c. 4, 12

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. MiniGPT-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023a. 4, 12

Wanrong Zhu, Jack Hessel, Anas Awadalla, Samir Yitzhak Gadre, Jesse Dodge, Alex Fang, Youngjae Yu, Ludwig Schmidt, William Yang Wang, and Yejin Choi. Multimodal C4: An open, billion-scale corpus of images interleaved with text. *arXiv preprint arXiv:2304.06939*, 2023b. 12

A Detailed Related Work

Mathematical reasoning benchmarks. Recently, numerous benchmarks (Amini et al., 2019; Cobbe et al., 2021; Mishra et al., 2022) have been proposed to evaluate the mathematical reasoning capabilities of Large Language Models (LLMs). However, most of these are textual only (Lu et al., 2023c), despite a substantial amount of mathematical information and reasoning being encapsulated in visual modalities. Meanwhile, some datasets exhibit performance saturation; for instance, GPT-4 achieves 92.0% accuracy on GSM-8K (Cobbe et al., 2021), a dataset of grade-school mathematics questions. On the other hand, the recent rapid advancement of Large Multimodal Models (LMMs) necessitates the establishment of robust multimodal benchmarks. However, current multimodal reasoning benchmarks provide limited coverage of rigorous and scientific domains (Antol et al., 2015; Kembhavi et al., 2016; Kahou et al., 2017; Mathew et al., 2022), which are key components for creating general-purpose AI assistants. To bridge this gap, it is crucial to develop a robust math reasoning dataset that integrates visual contexts.

Vision-language reasoning benchmarks. High-quality evaluation datasets and benchmarks are a cornerstone for assessing the progress of machine learning models to solve real-world tasks Liao et al. (2021). Prior studies such as VQA (Antol et al., 2015; Goyal et al., 2017), VizWiz (Gurari et al., 2018), and ParsVQA-Caps (Mobasher et al., 2022) assess the general-purpose visual question answering abilities of the LMMs, with or without task-specific training, on open-ended questions about images. In addition, there are several works that focus on evaluating specific skills of the LMMs beyond natural scenes, such as abstract scenes and shapes) (Antol et al., 2015; Lu et al., 2021b; Ji et al., 2022), geometry diagrams (Seo et al., 2015; Lu et al., 2021a; Chen et al., 2022a; Cao & Xiao, 2022), figures and charts (Methani et al., 2020; Masry et al., 2022; Kahou et al., 2017; Chang et al., 2022; Kafle et al., 2018), documents (text in images) (Singh et al., 2019; Mathew et al., 2022; Liu et al., 2023c), or synthetic images (Dahlgren Lindström & Abraham, 2022; Li et al., 2023; Bitton-Guetta et al., 2023). Besides, there has been significant progress on developing datasets to judge LMMs on skills that require external knowledge (Schwenk et al., 2022; Shah et al., 2019), common sense reasoning (Zellers et al., 2019; Yin et al., 2021), scientific-knowledge (Lu et al., 2022; Kembhavi et al., 2017, 2016), medical understanding (Zhang et al., 2023b; Lau et al., 2018). In this work, we create new datasets (IQTest, FunctionQA, PaperQA) and subsequently design a benchmark for holistic evaluation of the math reasoning capabilities of the LMMs.

Generative foundation models and their evaluation. Recently, there has been a surge of generative foundation models (Bommasani et al., 2021) that are trained on web-scale data, such as GPT-3, ChatGPT, GPT-4, Claude, LLaMA (Brown et al., 2020; OpenAI, 2022, 2023a; Anthropic, 2023; Touvron et al., 2023), with the ability to solve a wide range of downstream tasks (Wei et al., 2022a) without any task-specific finetuning. Prior work has focused on evaluating their abilities to respond to the queries from various disciplines, grounded in text, such as QA, math, medicine, coding and science (Bubeck et al., 2023; Nori et al., 2023; Chen et al., 2021; Fu et al., 2023; Sun et al., 2023; Wang et al., 2023; Huang et al., 2023, 2022; Liu et al., 2023b). Prior work, such as PixStruct (Lee et al., 2023), MatCha (Liu et al., 2022), and UniChart (Masry et al., 2023), has focused on developing specialized pretraining recipe for improved math and chart reasoning in visual contexts.

On the vision-language side, there are several generative foundation models such as LLaVA, miniGPT4, InstructBLIP, Flamingo, Multimodal Bard (Liu et al., 2023a; Zhu et al., 2023a; Dai et al., 2023; Alayrac et al., 2022; Awadalla et al., 2023; Google, 2023) that are trained on vast amount of paired (Schuhmann et al., 2022; Sharma et al., 2018; Lin et al., 2014) and interleaved image-text data (Zhu et al., 2023b). In addition, there has been recent development on specialized versions of these LMMs for document understanding where visual contexts require text recognition, math understanding being one of them (Zhang et al., 2023c; Ye et al., 2023). In recent times, there have been several works, such as Visit-Bench, LVLM-eHub, MMBench (Bitton et al., 2023; Yu et al., 2023; Xu et al., 2023; Shao et al., 2023), that assess their instruction-following and reasoning capabilities.

As the generative foundation models become more relevant to real-world applications, unlike prior work, we propose MATHVISTA to benchmark their capabilities of math reasoning (logical, arithmetic, statistical) on a diverse set of visual contexts (word problems in images, natural scenes, geometrical shapes, and plots).

B Data Collection Guidelines

B.1 Mathematical Reasoning Definition

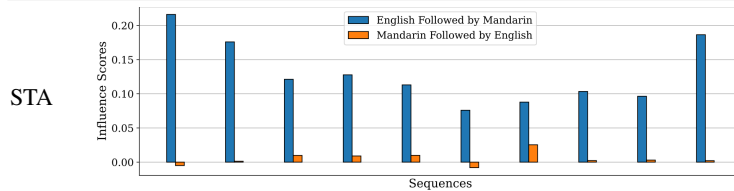Seven mathematical reasoning types are defined in Table 3.

Math Reasoning	Description
Arithmetic Reasoning (34.1%)	It covers the *fundamental operations* such as addition, subtraction, multiplication, division, and understanding of n*umber properties*. It may also include the ability to interpret numerical data in different forms.
Statistical Reasoning (30.5%)	It focuses on *data interpretation* and *analysis*, including measures (mean, median, mode), dispersion metrics (standard deviation, range), probability concepts, regression, correlation, and data inferences. It also identifies trends, outliers, and patterns.
Algebraic Reasoning (28.5%)	It encompasses understanding *variables*, *equations*, and the manipulation of *expressions* with polynomials and exponents. It also covers solving simple to complex equations, and grasping functions, their properties, and graphical depictions.
Geometry Reasoning (23.3%)	It emphasizes *spatial* understanding, analysis of 2D and 3D *figures*, and reasoning about their *shapes, sizes, and relationships*. It includes symmetry, congruency, similarity, area, volume, and transformations.
Numeric common sense (14.0%)	It involves intuitive understanding of *daily numerical concepts*, including understanding time differences, numerical judgment, and estimates. It covers temporal reasoning, spatial numeric assessments, and practical uses like budgeting and time reading.
Scientific Reasoning (10.7%)	It deals with the application of mathematical concepts in *scientific contexts*. This includes scientific notations, formula use, understanding rates, proportions, and percentages in practical situations, and problem-solving in scientific inquiries.
Logical Reasoning (3.8%)	It focuses on *critical thinking* and *deduction* from provided information, including pattern recognition, sequence understanding, predictions, and statement evaluation. Key components include premises, conclusions, and the use of abstract reasoning.

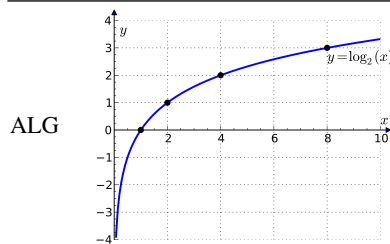Table 3: Definitions and proportions of seven mathematical reasoning categories in MATHVISTA.

B.2 Mathematical Reasoning Examples

Math	Examples

ARI

silk scraps	$9.08/lb
denim scraps	$8.47/lb
canvas scraps	$8.17/lb
felt scraps	$7.29/lb
faux fur scraps	$11.79/lb
lace scraps	$6.37/lb

Question: Karen bought 4 pounds of silk scraps and 4 pounds of canvas scraps. How much did she spend? (Unit: $)
Solution:
Find the cost of the silk scraps. Multiply: $9.08 \times 4 = \$36.32$
Find the cost of the canvas scraps. Multiply: $8.17 \times 4 = \$32.68$
Now find the total cost by adding: $\$36.32 + \$32.68 = \$69$
She spent $69.
Answer: 69

STA

Question: How many sequences have negative Influence Scores?
Answer: 2

ALG

Question: The derivative of y at $x = 6$ is _____ that at $x = 8$.
Choices: (A) larger than (B) equal to (C) smaller than
Answer: (A) larger than

Question: How many zeros does this function have?
Answer: 1

Question: What is the value of y at $x = 1$?
Answer: 0

GEO
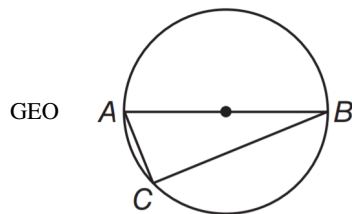
Question: $\overline{AB}$ is a diameter, $AC = 8$ inches, and $BC = 15$ inches. Find the radius of the circle.
Diagram logic forms:
```
PointLiesOnLine(D, Line(B, A))
PointLiesOnCircle(B, Circle(D, radius))
PointLiesOnCircle(A, Circle(D, radius))
PointLiesOnCircle(C, Circle(D, radius))
```
Answer: (C) 8.5

NUM
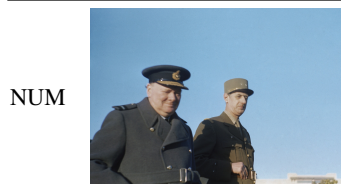
Question: What is the age gap between these two people in image? (unit: years)
Named entities: Winston Churchill, Charles de Gaulle
Wiki caption: Winston Churchill and General de Gaulle at Marrakesh, January 1944
Answer: 16

SCI
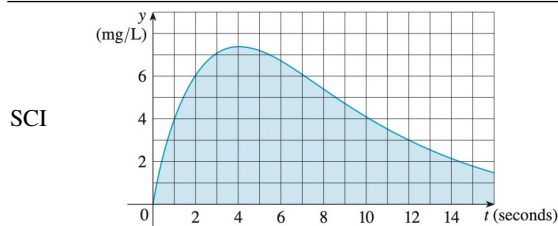
Question: The graph of the concentration function $c(t)$ is shown after a 7-mg injection of dye into a heart. Use Simpson's Rule to estimate the cardiac output.
Answer: 5.77

LOG

Question: Find the value of the square in the figure.
Solution:
Circle + Square = 5, Triangle + Triangle = 8,
Triangle = 4.
Circle + Triangle = 7, Circle = 3.
Therefore Square = 2
Answer: 2

Table 4: Examples of seven mathematical reasoning categories in MATHVISTA.
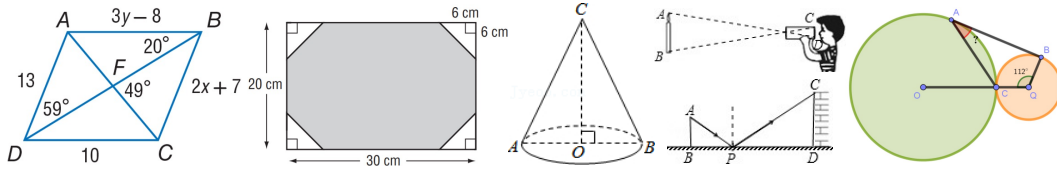
B.3 Visual Context Types

Figure 4: Examples of the visual context for the *geometry diagram* type.
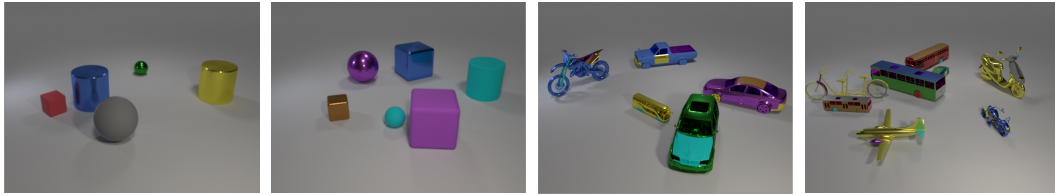

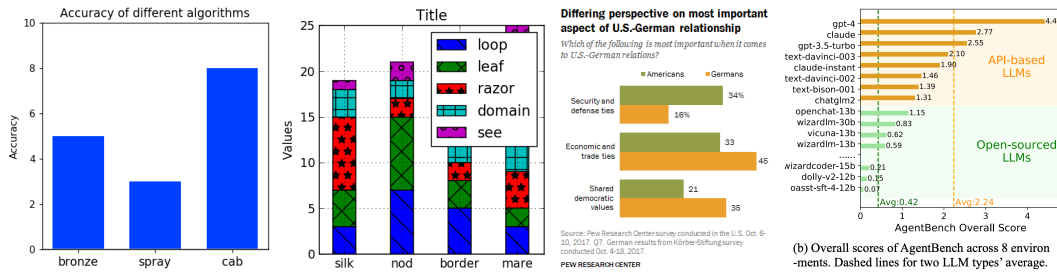Figure 5: Examples of the visual context for the *synthetic scene* type.

Figure 6: Examples of the visual context for the *bar chart* type.

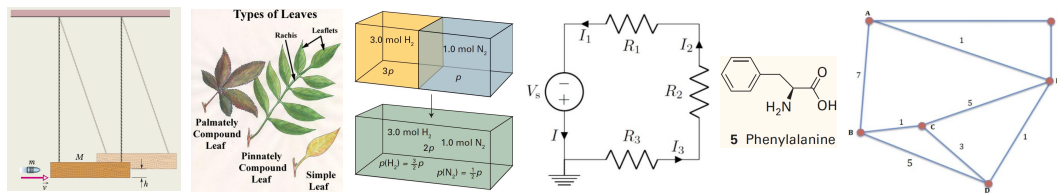Figure 7: Examples of the visual context for the *natural image* type.

Figure 8: Examples of the visual context for the *scientific figure* type.

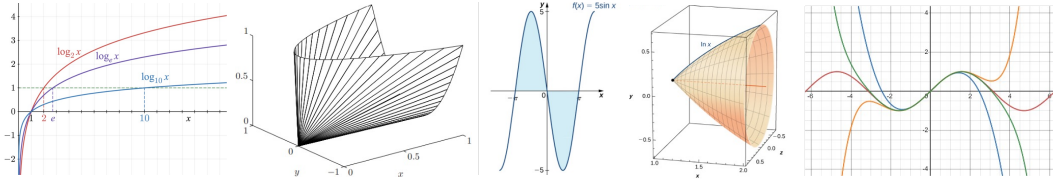Figure 9: Examples of the visual context for the *table* type.

Figure 10: Examples of the visual context for the *function plot* type.
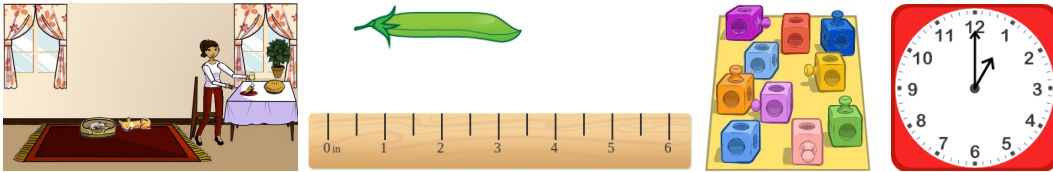

Figure 11: Examples of the visual context for the *abstract scene* type.
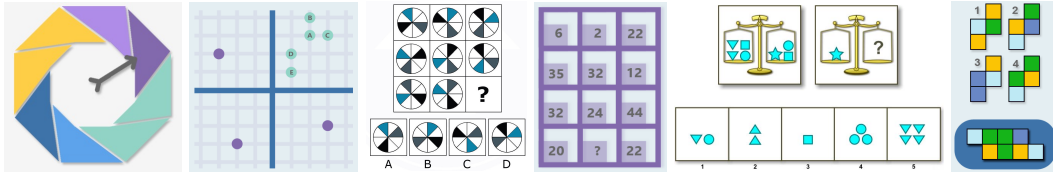

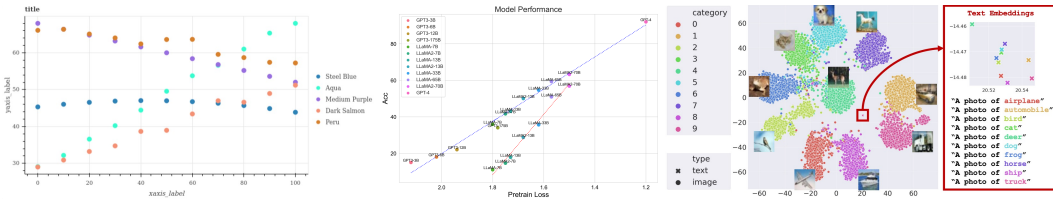Figure 12: Examples of the visual context for the *puzzle test* type.

Figure 13: Examples of the visual context for the *scatter plot* type.

Figure 14: Examples of the visual context for the *line plot* type.

Figure 15: Examples of the visual context for the *pie chart* type.

Figure 16: Examples of the visual context for the *document image* type.
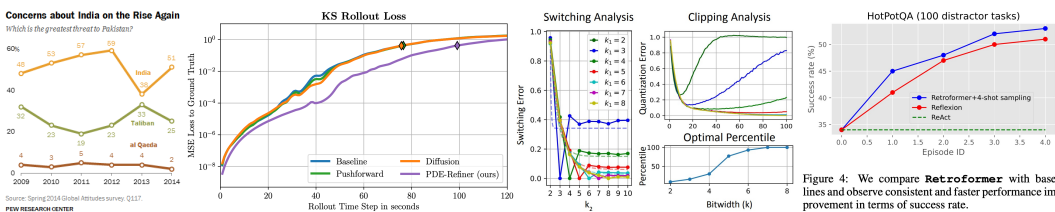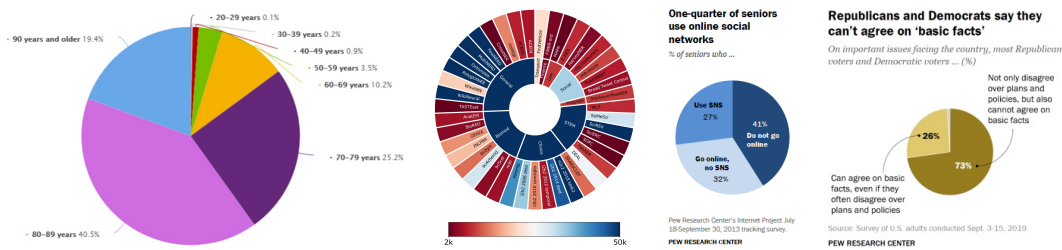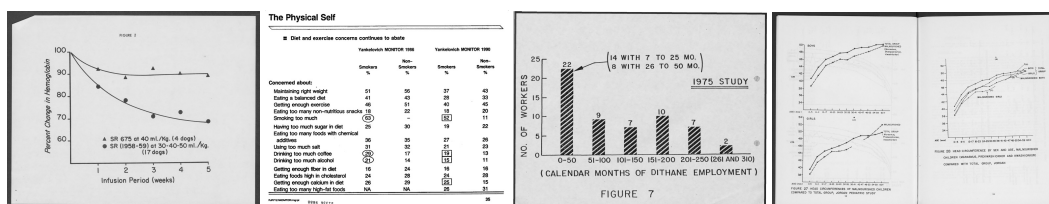

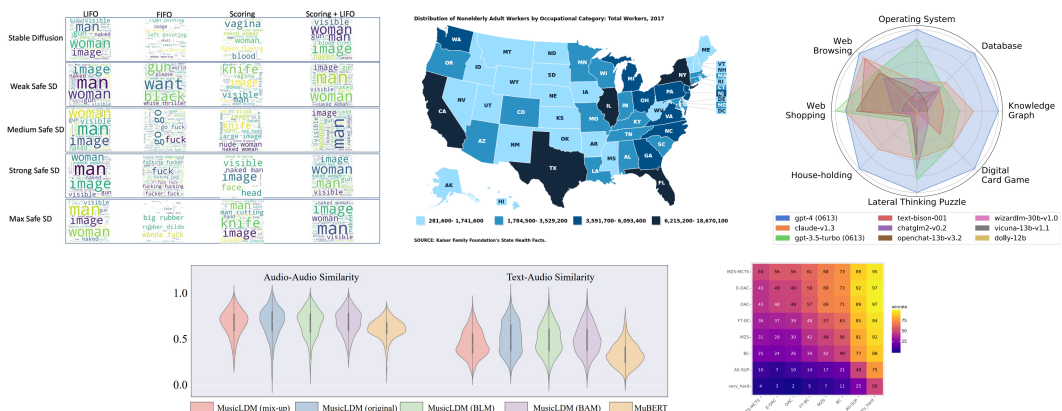Figure 17: Examples of the visual context for the *medical image* type.

Figure 18: Examples of the visual context for *other* types, including word cloud, map chart, radar chart, violin plot, and heatmap chart.

B.4 Source Dataset Summary

The source datasets are summarized in Table 5.

Dataset	Category	Task	Context	Math Skill
IQTest (Ours)	Math-Targeted	FQA	Puzzle Test	Logical, Arithmetic
PaperQA (Ours)	Math-Targeted	FQA	Charts and Plots	Scientific
FunctionQA (Ours)	Math-Targeted	TQA	Function Plot	Algebraic
Geometry3K (2021a)	Math-Targeted	GPS	Geometry Diagram	Geometry, Algebraic
GeoQA+ (2022)	Math-Targeted	GPS	Geometry Diagram	Geometry, Algebraic
GEOS (2015)	Math-Targeted	GPS	Geometry Diagram	Geometry, Algebraic
UniGeo (2022a)	Math-Targeted	GPS	Geometry Diagram	Geometry, Algebraic
CLEVR-Math (2022)	Math-Targeted	MWP	Synthetic Scene	Arithmetic
IconQA (2021b)	Math-Targeted	MWP	Abstract Scene	Arithmetic
TabMWP (2023b)	Math-Targeted	MWP	Table	Statistical, Arithmetic
SciBench (2023)	Math-Targeted	TQA	Scientific Figure	Scientific
TheoremQA (2023)	Math-Targeted	TQA	Scientific Figure	Scientific
ChartQA (2022)	General VQA	FQA	Charts and Plots	Statistical
FigureQA (2017)	General VQA	FQA	Charts and Plots	Statistical
DVQA (2018)	General VQA	FQA	Bar Chart	Statistical
MapQA (2022)	General VQA	FQA	Map Chart	Statistical
PlotQA (2020)	General VQA	FQA	Scatter Plot	Statistical
DocVQA (2022)	General VQA	FQA	Document Image	Statistical
AI2D (2016)	General VQA	TQA	Scientific Figure	Scientific
ScienceQA (2022)	General VQA	TQA	Scientific Figure	Scientific
TQA (2017)	General VQA	TQA	Scientific Figure	Scientific
A-OKVQA (2022)	General VQA	VQA	Natural Image	Arithmetic, Numeric
KVQA (2019)	General VQA	VQA	Natural Image	Arithmetic, Numeric
ParsVQA-Caps (2022)	General VQA	VQA	Natural Image	Arithmetic, Numeric
TextVQA (2019)	General VQA	VQA	Natural Image	Arithmetic, Numeric
VizWiz (2018)	General VQA	VQA	Natural Image	Arithmetic, Numeric
VQA2.0 (2017)	General VQA	VQA	Natural Image	Arithmetic, Numeric
PMC-VQA (2023b)	General VQA	VQA	Medical Image	Scientific
VQA-RAD (2018)	General VQA	VQA	Medical Image	Scientific
Super-CLEVR (2023)	General VQA	VQA	Synthetic Scene	Arithmetic
VQA-AS (2015)	General VQA	VQA	Abstract Scene	Arithmetic

Table 5: Summary of the 31 different source datasets in MATHVISTA. Among these, FunctionQA, IQTest, and PaperQA are our newly annotated datasets. The table provides details on their category, task, visual context, and primary mathematical reasoning skill types.

C   Data Collection Details

C.1   Collection Guidelines

As discussed previously, there is a notable gap in existing benchmarks, which primarily evaluate mathematical reasoning in textual contexts, overlooking the intrinsic visual nature of many mathematical problems. Our dataset, MATHVISTA, is therefore motivated to bridge this gap, offering a robust evaluation benchmark for mathematical reasoning intertwined with visual understanding, thus pushing AI assistants towards general-purpose capabilities. Our benchmark adheres to the following collection guidelines: (1) it covers multiple tasks and topics to mirror real-world applications; (2) it incorporates diverse visual contexts and mathematical skills to foster a well-rounded evaluation; (3) it offers varying levels of challenge to effectively probe and uncover the potential limitations of current models; and (4) it provides robust evaluation settings for deterministic evaluations.

The taxonomy for this work is introduced as follows: We identify seven types of mathematical reasoning: *algebraic reasoning*, *arithmetic reasoning*, *geometry reasoning*, *logical reasoning*, *numeric common sense*, *scientific reasoning*, and *statistical reasoning*, with detailed definitions provided in §B.1 and examples shown in §B.2. We focus on five primary tasks: *figure question answering* (FQA), which centers around statistical reasoning over multiple charts and plots; *geometry problem solving* (GPS), which deals with geometrical topics; *math word problem* (MWP), which involves arithmetic reasoning in everyday scenarios; *textbook question answering* (TQA), which usually entails knowledge-intensive reasoning on scientific topics and figures; and *visual question answering* (VQA). Furthermore, our objective is to account for a diverse array of visual contexts, including natural images, geometry diagrams, abstract scenes, synthetic scenes, multiple charts and plots, scientific figures, tables, function plots, puzzle test figures, and more, with examples shown in §B.3.

C.2   Data Collection

Collection of MathQA datasets.   We collected nine MathQA datasets in multimodal settings, including four for GPS, two for MWP with visual contexts of synthetic scenes, abstract diagrams, and tables, and two for TQA on college curricula (see §B.4). Annotations such as solutions, programs, parsing results, and grounded theorems are also collected, providing demonstration examples for LLMs. Each source dataset is limited to up to 400 examples to ensure a balanced representation of each source in our final compiled benchmark. In total, we collected 2,666 examples.

Review and collection of VQA datasets.   Many existing VQA datasets feature instances requiring mathematical reasoning abilities, such as arithmetic operations or numeric common sense. Incorporating these datasets enhances problem diversity in terms of tasks, domains, visual contexts, and reasoning skills involved. We reviewed more than 70 datasets, collecting 19 of them that contain math-related instances and are publicly available, as listed in §B.4. Since these datasets are not originally math-targeted, we initially designed heuristic rules to automatically select examples likely to involve mathematical reasoning from a large pool of candidates. Examples with numeric answers or those containing quantity words (as listed in §C.5) in the questions were selected. This automatic filtration yielded 4,949 VQA-format examples, though some false positive examples remained. Therefore, we engaged three expert annotators to manually label these examples to determine if they involve mathematical reasoning (more details in § C.6). Utilizing majority voting and limiting each source dataset to 400 examples, we finalized a collection of 2,739 examples.

Collection of three new datasets.   While the source datasets we collected encompass multiple visual contexts and mathematical reasoning abilities, certain scenarios remain unaddressed: logical reasoning on puzzle test diagrams, statistical reasoning on functional plots, and scientific reasoning on academic figures. To address these gaps, we introduced three new datasets: IQTest, FunctionQA, and PaperQA, with examples illustrated in Figure 2. IQTest comprises 228 examples requiring inductive reasoning, abstract thinking, pattern prediction, and calculations, sourced from puzzle test figures on online learning platforms. FunctionQA, with 400 examples, emphasizes subtle visual perceptions of functional plots and algebraic reasoning concerning variables, expressions, equations, and functions. PaperQA is a novel dataset featuring questions derived from informative academic illustrations, including tables, figures, and charts from online education resources, with 107 examples

sourced from papers released in August 2023 on Huggingface[1]. To ensure data quality, all questions were manually annotated by graduate students in STEM fields and further refined through a rigorous review process. The GUI of the annotation tool is shown in Figure 20 in §C.7.

C.3   Metadata Annotation

Fine-grained metadata facilitates a comprehensive analysis of models' reasoning capabilities across various aspects. To this end, we annotate the examples in MATHVISTA with information including question type, answer type, language, source, category, task, grade level, and visual context, which can be accurately obtained from the details provided in the source datasets. MATHVISTA features seven different types of mathematical reasoning abilities, as categorized in Table 3 (§B.1). Coarse labels of mathematical reasoning can be automatically obtained from the details of the source datasets. To verify the quality of automatic annotation, expert annotators manually label the mathematical reasoning categories from seven candidates for 1,000 examples, using the annotation tool illustrated in §C.8. The results show that 94.1% of the examples from automatic and human annotations have the exact same set of reasoning types, while 98.79% of the individual labels are identical, indicating that the automatic annotation for the labeling of mathematical reasoning is highly accurate.

C.4   Data Preparation and Release

MATHVISTA consists of 6,141 examples, divided into two subsets: *testmini* and *test*. *testmini* contains 1,000 examples, intended for model development validation or for those with limited computing resources. The *test* set features the remaining 5,141 examples for standard evaluation. Notably, the answer labels for *test* will not be publicly released to prevent data contamination, and we will maintain an online evaluation platform. To ensure that each source dataset is well represented in *testmini* and to maintain a distribution in *testmini* closely resembling the whole set, we adopted this sampling strategy: (1) first, randomly sample questions with a threshold number of 4 for each source dataset; (2) then, randomly sample the remaining questions for each source dataset on its proportion in the entire set. The KL Divergence and Total Variation (TV) distance between the *testmini* set and the entire set are 0.008 and 0.035, respectively, suggesting that *testmini* is close to the distribution of the whole set. We also conducted several quality checks to address any unidentified errors.

C.5   Automatic Selection of Mathematical Problems

most, least, fewest more, less, fewer, largest, smallest, greatest, larger, smaller, greater, highest, lowest, higher, lower, increase, decrease, minimum, maximum, max, min, mean, average, median, total, sum, add, subtract, difference, quotient, gap, half, double, twice, triple, square, cube, root, approximate, approximation, triangle, rectangle, circle, square, cube, sphere, cylinder, cone, pyramid, multiply, divide, percentage, percent, ratio, proportion, fraction, rate

Table 6: Dictionary of quantity words used for the automatic selection of questions likely to involve mathematical reasoning.

C.6   Human Labeling of Mathematical Problems

We developed an annotation tool, as illustrated in Figure 19, to enable expert annotators to label problems that involve mathematical reasoning. Annotators were trained using detailed instructions, as shown in Figure 7, along with a variety of examples—positive ones that involve mathematical reasoning and negative ones that do not. We provided three labeling options:

- *Yes* - This indicates that the problem involves mathematical reasoning.

- *No* - This indicates that the problem does not involve mathematical reasoning.

- *Unsure* - This option should be selected if it is uncertain whether the problem involves mathematical reasoning. (Annotators are advised to use this option sparingly.)

---

[1] https://huggingface.co/papers

Figure 19: GUI for labeling if a problem involves mathematical reasoning.

We are compiling a dataset that incorporates image context and involves mathematical reasoning (MathQA in visual contexts). We have gathered a set of examples in which some involve mathematical reasoning, while others do not.

In our task, a question can be classified as a mathematical problem if it

- Involves numbers or symbols in the question text or the image context, AND requires further operations or transformations to be performed on them to reach a solution.
- Involves more complex forms of mathematical reasoning, including logical reasoning, abstract thought, and understanding of patterns.

Based on the definition above, a problem is classified as a negative example (NOT involving mathematical reasoning) if it:

- Does not involve any numbers or quantity words, OR
- Involves only counting, reading, or recognizing numbers, OR
- Relies solely on factual information, such as recalling years and dates.

Table 7: Instructions for human annotators to identify if a problem involves mathematical reasoning.

They may leave comments if they find anything incorrect or offensive for removal at a later stage.

In our study, we employed the Fleiss Kappa score to conduct an inter-annotator agreement analysis among three annotators tasked with labeling examples based on mathematical reasoning. The Fleiss Kappa score is a statistical measure used to evaluate the reliability of agreement between multiple raters, providing a quantifiable metric to assess the consistency across different annotators. A score of 1 indicates perfect agreement, while a score of 0 suggests no agreement beyond what would be expected by chance. Our analysis yielded a Fleiss Kappa score of 0.775, indicating a substantial level of consistency among the annotators. This high degree of agreement underscores the reliability of our annotation process and affirms the quality of the labeled data generated for our study.

C.7 Annotating Three New Datasets

C.8 Human Labeling of Mathematical Reasoning

Figure 20: GUI for annotating our new source datasets.

Figure 21: GUI for labeling mathematical reasoning skills.

D More Dataset Analysis

Question distribution. Apart from English questions, MATHVISTA contains 6.57% non-English questions, including languages such as Chinese and Persian. The multilingual feature necessitates that models be capable of understanding and processing multiple languages to ensure accurate results across the dataset. As illustrated in Table 3, the average number of words in English questions within MATHVISTA is 15.58, while the maximum number of words in a question reaches 213.

Figure 22 further elucidates the distribution of word counts, highlighting the diverse patterns of questions. MATHVISTA features two types of questions: multiple-choice questions and free-form questions. For multiple-choice questions, the average number of choices is 3.4, while the maximum number of choices is 8. In the case of free-form questions, answers can be integers, floating-point numbers, or lists, which can be converted into a standard format. The standard settings in question and answer types facilitate consistent accuracy evaluation for existing models.
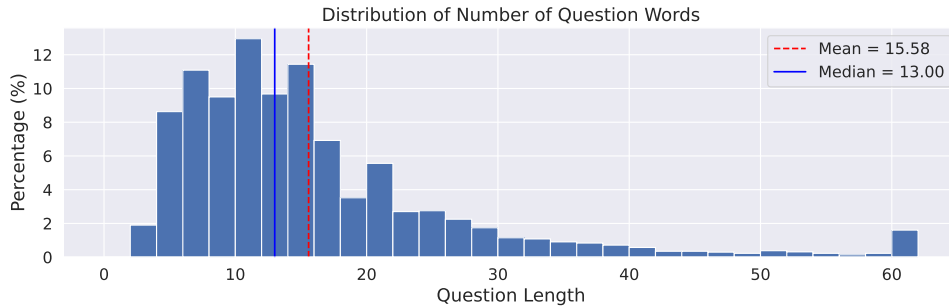
Figure 22: The distribution of the number of words per question in MATHVISTA. Questions with a length greater than 60 are categorized as 61 for visualization simplicity.

Dataset category and task type. Source datasets in MATHVISTA can be categorized into two types: math-targeted VQA datasets, which are originally proposed for assessing mathematical reasoning, and general VQA datasets, which address visual reasoning in everyday scenarios. The distribution proportions of these two categories (55.4% vs. 44.6%, as illustrated in Figure 23) within MATHVISTA enable a balanced examination of mathematical reasoning in both domain-specific and general-purpose applications. The distribution of the five tasks contained within MATHVISTA is visualized in Figure 24. The relatively balanced distribution of these tasks enhances the benchmarking robustness that our dataset provides.
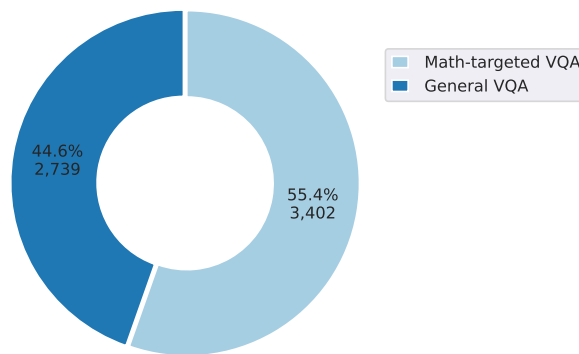
Figure 23: Category distribution of problems within MATHVISTA.

Grade level. The datasets within MATHVISTA are categorized into four distinct grade levels: *daily life*, *elementary school*, *high school*, and *college*, each representing a different level of reasoning complexity and contextual application. The daily life category encompasses questions grounded in everyday scenarios requiring basic reasoning and arithmetic skills. The elementary school category aligns with the typical mathematical curriculum of elementary education, introducing basic topics such as arithmetic operations and introductory geometry. High school level questions delve into more
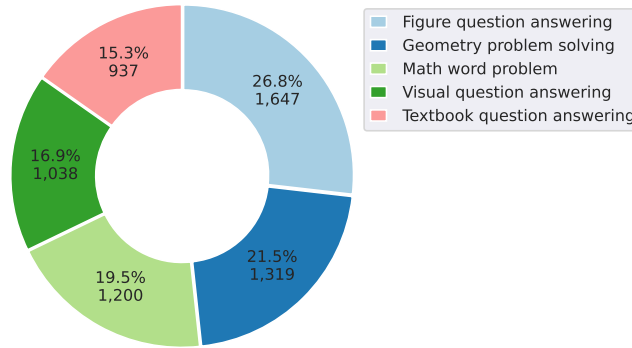
Figure 24: Task type distribution of problems within MATHVISTA.

complex mathematical concepts such as algebra, geometry, and introductory calculus. The college category encapsulates the highest level of complexity, featuring questions on advanced mathematical and scientific concepts like calculus, linear algebra, and physics.

The distribution of questions across these grade levels is visualized in Figure 25. This structured categorization enriches the diversity of MATHVISTA, providing a meaningful framework for evaluating and benchmarking the mathematical and visual reasoning capabilities of various models across different educational contexts, thereby assessing their practical utility and educational relevance.
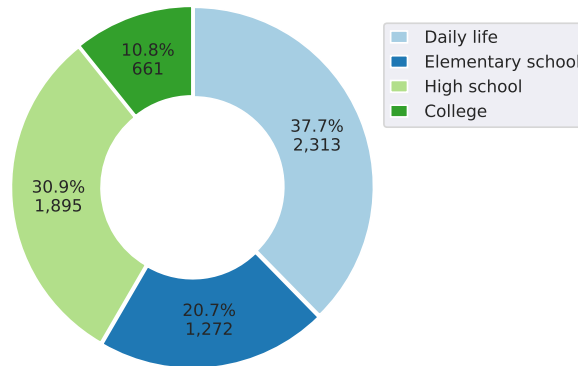
Figure 25: Distribution of questions across different grade levels within MATHVISTA.

Visual context. The datasets within MATHVISTA encompass over 10 different visual contexts (with the distribution shown in Figure 26), crucial for evaluating models' ability to interpret and reason across diverse visual information. Common visual contexts include geometry diagrams, synthetic scenes, bar charts, natural images, and scientific figures as illustrated in Figure 5 to Figure 16. Less frequent, yet equally important visual contexts such as medical images, word clouds, map charts, radar charts, violin plots, and heatmap charts are depicted in Figure 17 and Figure 18. These visual contexts, ranging from common to specialized representations, challenge the models to decode and reason with varying visual information, contributing to a more robust and comprehensive evaluation. The diversity in visual contexts enriches MATHVISTA, enhancing the benchmarking robustness and providing a solid foundation for understanding the practical utility and domain-specific performance of various models across different domains and applications.

Mathematical reasoning ability. The datasets within MATHVISTA encompass a spectrum of seven distinct mathematical reasoning types, facilitating a thorough evaluation of models' mathematical reasoning capabilities. Figure 27 illustrates the portion of each reasoning type involved in the problems, with arithmetic being the most frequent and logical reasoning being the least frequent. This distribution reflects the varying degrees of mathematical reasoning required across different problems. Figure 28 further delineates the distribution of reasoning types, showcasing a mean of 1.45. The sparse distribution observed aids in the precise analysis of each type's performance by
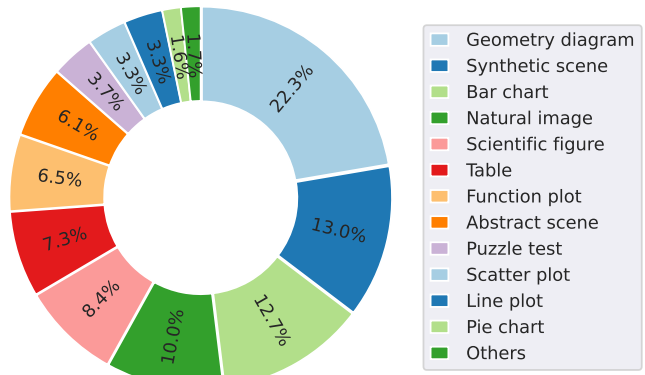
Figure 26: Visual context distribution within MATHVISTA.

the models, providing a nuanced understanding of their strengths and weaknesses across different mathematical reasoning domains. This structured representation of mathematical reasoning types within MATHVISTA not only enriches the dataset but also significantly contributes to a more robust and comprehensive evaluation of models, aiding in the identification of areas for improvement and the development of more proficient mathematical reasoning models.
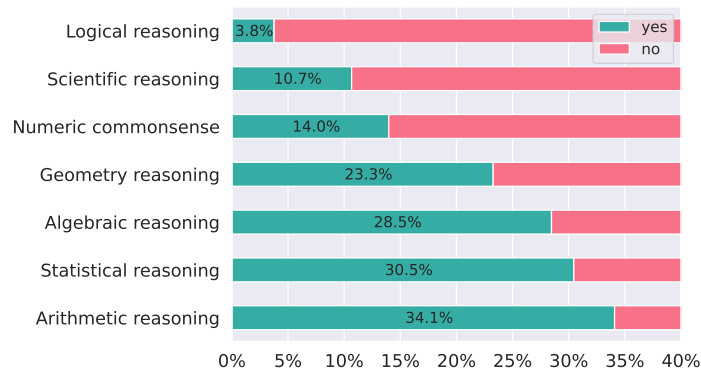
Figure 27: Portion of each mathematical reasoning type involved in the problems of MATHVISTA.

Figure 28: Distribution of the number of mathematical reasoning types within MATHVISTA.

E Details on the Setup

E.1 Frequent Guess

We employ a strategy where the most frequent answers in the *testmini* set are utilized as predictions for various question and answer types. For multiple-choice questions, the most frequent option is selected based on the number of available options. For instance, option B is chosen for questions with two options, aligning with the answer distribution in *testmini*. Similarly, for questions requiring an answer type of integer, a floating number with one decimal place, a floating number with two decimal places, or a list, we use 2, 1.2, 0.21, and $[0, 2, 0, 2, 1, 7, 1, 2, 0, 3, 0, 6]$ respectively, in accordance with the answer distribution observed in *testmini*.

E.2 Prompt for Answer Extraction

The prompt used to instruct GPT-4 for answer extraction is illustrated in Table 8.

| Element	Prompt
Task description	Please read the following example. Then extract the answer from the model response and type it at the end of the prompt.
Example 1	**Hint:** Please answer the question requiring an integer answer and provide the final value, e.g., 1, 2, 3, at the end. **Question:** Which number is missing? **Model response:** The number missing in the sequence is 14. **Extracted answer: 14**
Example 2	**Hint:** Please answer the question requiring a floating-point number with one decimal place and provide the final value, e.g., 1.2, 1.3, 1.4, at the end. **Question:** What is the fraction of females facing the camera? **Model response:** The fraction of females facing the camera is 0.6, which means that six out of ten females in the group are facing the camera. **Extracted answer: 0.6**
Example 3	**Hint:** Please answer the question requiring a floating-point number with two decimal places and provide the final value, e.g., 1.23, 1.34, 1.45, at the end. **Question:** How much money does Luca need to buy a sour apple candy and a butterscotch candy? (Unit: $) **Model response:** Luca needs $1.45 to buy a sour apple candy and a butterscotch candy. **Extracted answer: 1.45**
Example 4	**Hint:** Please answer the question requiring a Python list as an answer and provide the final list, e.g., [1, 2, 3], [1.2, 1.3, 1.4], at the end. **Question:** Between which two years does the line graph saw its maximum peak? **Model response:** The line graph saw its maximum peak between 2007 and 2008. **Extracted answer: [2007, 2008]**
Example 5	**Hint:** Please answer the question and provide the correct option letter, e.g., A, B, C, D, at the end. **Question:** What fraction of the shape is blue? **Choices:** (A) 3/11 (B) 8/11 (C) 6/11 (D) 3/5 **Model response:** The correct answer is (B) 8/11. **Extracted answer: B**

Table 8: Task description along with five examples used to prompt GPT-4 for answer extraction.

E.3 Prompts for Response Generation

Question type	Answer type	Task instruction
multiple-choice	Text	Please answer the question and provide the correct option letter, e.g., A, B, C, D, at the end.
Free-form	Integer	Please answer the question requiring an integer answer and provide the final value, e.g., 1, 2, 3, at the end.
Free-form	Float (1)	Please answer the question requiring a floating-point number with one decimal place and provide the final value, e.g., 1.2, 1.3, 1.4, at the end.
Free-form	Float (2)	Please answer the question requiring a floating-point number with two decimal places and provide the final value, e.g., 1.23, 1.34, 1.45, at the end.
Free-form	List	Please answer the question requiring a Python list as an answer and provide the final list, e.g., [1, 2, 3], [1.2, 1.3, 1.4], at the end.

Table 9: The task instructions for different question and answer types in answer extraction. Here, Float (1) refers to a floating-point number with one decimal place, and Float (2) refers to a floating-point number with two decimal places.

E.4 Prompt for Caption Generation

We instruct Multimodal Bard to generate a detailed description for an input image, aiming to augment current LLMs with visual understanding capabilities. The prompt is shown in Table 10.

Describe the fine-grained content of the image or figure, including scenes, objects, relationships, and any text present.

Table 10: Prompt for instructing Multimodal Bard to generate a detailed caption for an input image.

E.5 Model Hyperparameters

The hyperparameters for the experiments in §3.2 are set to their default values unless specified otherwise. Table 11 and Table 12 detail specific generation parameters for the various large language models (LLMs) and large multimodal models (LMMs) we evaluated, respectively.

Model	Generation Setup
Claude-2	model = `claude-2`, temperature = 0, max_tokens = 1024
ChatGPT	model = `gpt-3.5-turbo`, temperature = 0, max_tokens = 1024
GPT-4	model = `gpt-4-0613`, temperature = 0, max_tokens = 1024

Table 11: Generating parameters for various LMMs.

E.6 Human Performance

We conducted a study to evaluate human performance on the *testmini* subset of the MATHVISTA, utilizing Amazon Mechanical Turk (AMT). Each question from the *testmini* subset was assigned to five annotators, all of whom have a history of completing more than 5,000 HIT tasks and boast an acceptance score higher than 0.99, to ensure the quality of the results. The study comprised five test questions and two qualification questions, which were to be answered within a 20-minute timeframe. The qualification questions consisted of elementary math word problems requiring basic arithmetic operations (e.g., addition and subtraction). Only annotators who successfully answered the qualification questions were deemed eligible for the study, and their responses were included in the final analysis. Additionally, annotators were requested to provide information regarding their highest level of educational attainment. We retained the results exclusively from annotators who had achieved

Model	Generation Setup
IDEFICS-9B-Instruct	max_new_tokens = 256, temperature = 1.0
mPLUG-Owl-LLaMA-7B	do_sample = True, top-k = 5, max_length = 512
miniGPT4-LLaMA-2-7B	num_beams = 1, temperature = 1.0, max_new_tokens = 300, max_length = 1000
LLaMA-Adapter-V2-7B	max_gen_len = 256, temperature = 0.1, top_p= 0.75
LLaVAR	do_sample = True, temperature = 0.2, max_new_tokens = 1024
InstructBLIP-Vicuna-7B	do_sample = False, num_beams = 5, max_length = 256, min_length = 1, top_p = 0.9, repetition_penalty = 1.0, temperature = 1
LLaVA-LLaMA-2-13B	do_sample = True, temperature = 0.2, max_new_tokens = 1024

Table 12: Generating parameters for various LMMs.

a high school diploma or higher, as 30.9% of the problems in MATHVISTA are of high-school level difficulty and 10.8% correspond to college-level curricula.

E.7 Multimodal Bard Assessment Task

A screenshot of our AMT worker interface, utilized for the Multimodal Bard assessment task, is provided in Figure 29. The workers were compensated at a rate of $18 per hour.
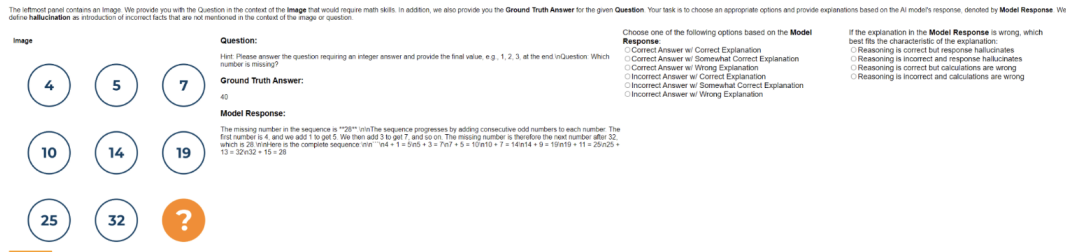

Figure 29: Screenshot of the Multimodal Bard assessment task interface.

F   More Experimental Results

F.1   Fine-grained Results

We also report fine-grained scores for a comprehensive study of the capabilities of existing models across different tasks (Table 2), mathematical reasoning abilities (Table 2), grade levels (Figure 33 in §F.4), and visual context types (Figure 1, Figure 34 in §F.5). Multimodal Bard excels in the geometry problem solving task with an accuracy of 47.1%, closely approaching human accuracy at 48.4%. CoT GPT-4 performs well on the textbook question-answering task, achieving a gain of 32.3% over random chance, and on questions involving scientific reasoning, with a gain of 41.0%, showcasing its superiority in domain-specific knowledge. PoT GPT-4 is the best performing model in the categories of figure question answering, statistical reasoning, and table context, thanks to its ability to generate high-quality codes for precise mathematical reasoning. We perform an ablation study on the augmented LLMs and present the results in Table 35 (see §F.6). The gap in the performance of the Augmented LLMs can be attributed to poor image captions, which may not adequately describe the math in visual contexts, the inability of the OCR to detect shapes useful for geometrical reasoning, and the lack of mathematical reasoning capabilities.
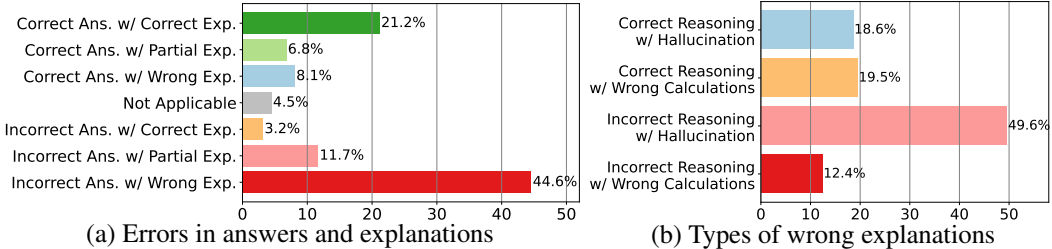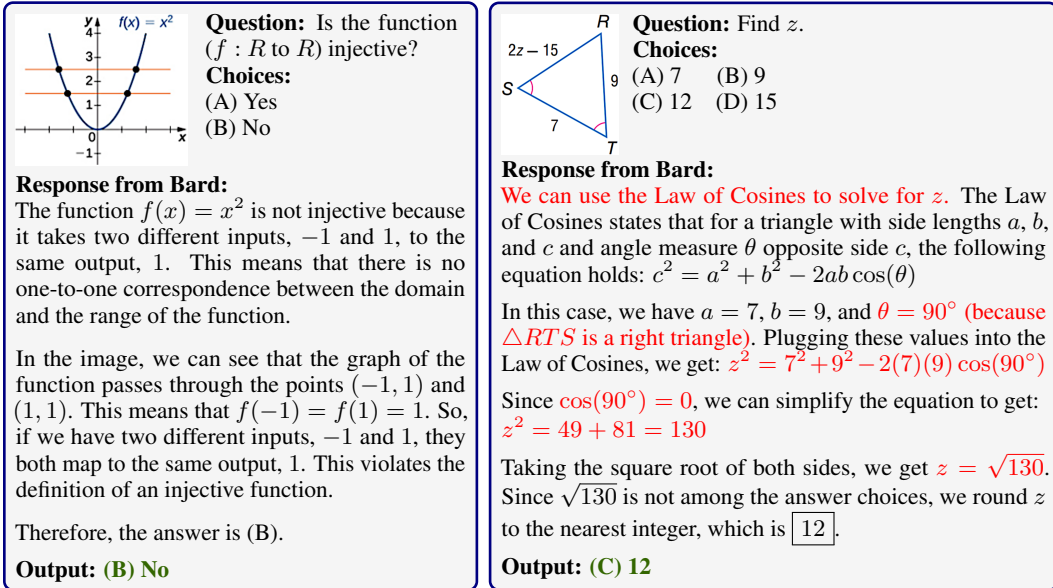
F.2   Qualitative Analysis

Figure 30: Error analysis of Bard results: (a) presents errors in answers and explanations; (b) delves into the details of wrong explanations. Notations: "Answer" is "Ans.", "Explanation" is "Exp.", "Partially Correct" is "Partial", and "Not applicable" refers to unanswerable or indeterminate cases.

Success and failure analysis of Multimodal Bard.   In §3, we observe that Multimodal Bard achieves the highest average accuracy on MATHVISTA. Here, we analyze its predictions through human evaluation to understand its mode of success and failure. To do so, we ask the human workers, from Amazon Mechanical Turk (AMT), to study Bard's predictions given the math question, its associated image, and the ground truth from MATHVISTA dataset for 250 instances. Specifically, workers were instructed to decide whether the predictions contained the correct answer with the correct explanation. If the workers find that the model's explanation is incorrect, they had to choose whether the wrong explanation was due to various failure modes such as incorrect reasoning with *hallucination* or wrong calculations. In our setup, we define hallucination as an introduction of incorrect facts, in the model explanation, that is not mentioned in the context of the image or question (e.g., in Figure 36 and Figure 37). More details can be found in §E.7.

We present the distribution of the quality of Bard's predictions, judged by the human annotators, in Figure 30 (a). We find that 44.6% of the Bard's predictions had incorrect answers with incorrect explanations. Interestingly, we observe that Bard responds with partial (6.8%) or completely (8.1%) incorrect explanations despite giving the correct answer to the input image and question, highlighting its failure to reach the correct answer for the wrong reasons. In Figure 30 (b), we present the distribution over possible reasons when Bard provides incorrect explanations. Notably, we find that 49.6% of its responses contain hallucinations. Our analysis highlights that hallucination is a major source of errors in the generative foundation models (Lu et al., 2023c; Ji et al., 2023). We also observe that the model responds with correct reasoning but either hallucinates (18.6%) or performs wrong calculations (19.5%) leaving an overall impression of being a wrong explanation.

Qualitative examples of Multimodal Bard.   We also present a few qualitative examples of Bard's predictions. In Figure 31 (a), we find that Bard generates the correct answer with the correct explanation, including detecting the correct function (i.e., $f(x) = x^2$) and analyzing its properties (i.e., injective) to answer the question. However, in Figure 31 (b), we observe that the model provides the correct answer (i.e., 12) but with an incorrect explanation (i.e., using the law of cosines when

Question: Is the function ($f : R$ to R) injective?
Choices:
(A) Yes
(B) No

Response from Bard:
The function $f(x) = x^2$ is not injective because it takes two different inputs, -1 and 1, to the same output, 1. This means that there is no one-to-one correspondence between the domain and the range of the function.

In the image, we can see that the graph of the function passes through the points $(-1, 1)$ and $(1, 1)$. This means that $f(-1) = f(1) = 1$. So, if we have two different inputs, -1 and 1, they both map to the same output, 1. This violates the definition of an injective function.

Therefore, the answer is (B).

Output: (B) No

Question: Find z.
Choices:
(A) 7 (B) 9
(C) 12 (D) 15

Response from Bard:
We can use the Law of Cosines to solve for z. The Law of Cosines states that for a triangle with side lengths a, b, and c and angle measure θ opposite side c, the following equation holds: $c^2 = a^2 + b^2 - 2ab\cos(\theta)$

In this case, we have $a = 7$, $b = 9$, and $\theta = 90°$ (because $\triangle RTS$ is a right triangle). Plugging these values into the Law of Cosines, we get: $z^2 = 7^2 + 9^2 - 2(7)(9)\cos(90°)$

Since $\cos(90°) = 0$, we can simplify the equation to get: $z^2 = 49 + 81 = 130$

Taking the square root of both sides, we get $z = \sqrt{130}$. Since $\sqrt{130}$ is not among the answer choices, we round z to the nearest integer, which is $\boxed{12}$.

Output: (C) 12

(a) **Correct answer** and **explanation** (b) **Correct answer** but **wrong explanation**

Figure 31: Two examples from Bard. In (b), Bard does not correctly identify the geometry symbols and relationships. The accurate correct should identify the isosceles triangle and apply its properties.

the question requires an understanding of the properties of isosceles triangles). We present more examples in §F.8. Overall, our analysis of Bard highlights its modes of failure in detail, which could guide future foundation model design to address these issues.

Qualitative examples of Augmented GPT-4. Augmented with external visual models, CoT GPT-4 and PoT GPT-4 are able to achieve comparable performance with Multimodal Bard. As shown in Figure 32 (a), provided with the accurate OCR text detected in the image, PoT GPT-4 accurately understands the structural information of the image and generates a code snippet to perform precise statistical reasoning. In Figure 32 (b), the caption provides some accurate descriptions of the image (e.g., $f(x) = c$) along with hallucination (e.g., $y = 3$, the line passes through $(0, 3)$) caused by the external Bard model. Although CoT GPT-4 predicts the correct answer given the partially correct information, the qualities of visual information augmented by external models have an impact on the accurate visual perception and thus the final mathematical reasoning performance. Examples in §F.9 show failure cases due to hallucination caused by external visual models.

F.3 Results on the Test Set

Table 13 reports the accuracy scores of two heuristic baselines, two leading augmented LLMs (CoT GPT-4, PoT GPT-4), and one leading LMM (LLaVA-LLaMA-2-13B) on the *test* subset. The minor differences between scores on the *test* subset and the *testmini* subset, as shown in Table 2, suggest that *testmini* effectively mirrors the *test* subset, serving as a valuable evaluation subset for model development, especially for those who have limited computing resources.

F.4 Scores Across Different Grade Levels

Table 33 displays the average accuracy scores across different grade levels (elementary school, high school, and college) for the leading foundation models. Humans exhibit the highest performance on questions at the elementary school level, while they fare the worst on college-level questions within MATHVISTA. Foundation model baselines exhibit varying performance behaviors: they achieve better accuracy scores on high school level questions compared to the other two categories. Furthermore, the smallest performance gap between Multimodal Bard, the best performing LMM baseline, and humans is observed on the high school level questions. This might be attributed to the training regimen of the models, which perhaps aligns well with the high school curriculum, enabling better performance. On the college curriculum, CoT GPT-4, augmented by visual inputs, outperforms

Laps driven	
Day	**Number of laps**
Thursday	53
Friday	53
Saturday	53
Sunday	55
Monday	55

Question: A race car driver kept track of how many laps he drove in the past 5 days. What is the mode of the numbers?

Detected OCR text:
([55, 9], Laps driven),
([21, 34], Day),
([81, 37], Number of laps),
([1, 64], Thursday), ([129, 67], 53),
([3, 99], Friday), ([129, 99], 53),
([1, 126], Saturday), ([129, 129], 53),
([3, 161], Sunday), ([129, 161], 55),
([1, 188], Monday), ([129, 191], 55)

Response from GPT-4:

```
def mode(laps):
    return max(set(laps),
    key=laps.count)

laps = [53, 53, 53, 55, 55]
print(mode(laps))
```

Output: 53

(a) **Correct answer** and **code**

Question: What is this function most likely be?
Choices:
(A) polynomial (B) trigonometric
(C) exponential (D) logarithmic

Caption from Bard:
The image shows a graph of a function $f(x) = c$ on a graph paper. The x axis is the horizontal axis and the y axis is the vertical axis. The graph is a horizontal line that passes through the point $(0, c)$. ... The text "$y = 3$" means that the y-value of all points on the graph is 3. The text "$(0, c)$" means that the point $(0, c)$ is on the graph. The text "$f(x) = c$" means that the function $f(x)$ is equal to c for all values of x. ... The graph of a constant function is a horizontal line. In this case, the horizontal line passes through the point $(0, 3)$, so the constant value of the function is 3. ...

Response from GPT-4:
The function shown in the image is a constant function, which is a type of polynomial function. Thus, the answer is A.

Output: (A) polynomial

(b) **Correct answer** with **partially correct outputs**

Figure 32: Two examples from GPT-4. GPT-4 depends on the qualities of the generated caption and detected OCR texts. In (b), some information is incorrect, even though the final answer is correct.

Model	Input	ALL	FQA	GPS	MWP	TQA	VQA	ALG	ARI	GEO	LOG	NUM	SCI	STA
Random chance	-	17.86	15.46	24.12	4.54	23.36	24.33	25.84	13.85	22.69	13.40	8.82	15.76	14.28
Frequent guess	-	23.48	20.97	27.18	16.27	26.06	28.87	28.29	20.86	25.71	11.86	19.61	20.45	20.08
2-shot CoT GPT-4	Q, I_c, I_t	30.50	27.21	35.91	21.30	43.13	28.17	35.72	25.17	35.80	24.74	15.41	47.28	31.29
2-shot PoT GPT-4	Q, I_c, I_t	31.74	27.58	37.35	23.87	43.00	30.27	37.15	27.93	37.48	22.68	15.83	44.47	31.87
LLaVA-LLaMA-2-13B	Q, I	25.40	22.86	24.57	18.15	35.82	29.69	26.93	22.47	24.45	19.07	19.05	34.71	21.61

Table 13: Accuracy scores on the *test* subset of MATHVISTA. Input: Q: question, I: image, I_c: image caption, I_t: OCR texts detected from the image. ALL: overall accuracy. Task types: FQA: figure question answering, GPS: geometry problem solving, MWP: math word problem, TQA: textbook question answering, VQA: visual question answering. Mathematical reasoning types: ALG: algebraic reasoning, ARI: arithmetic reasoning, GEO: geometry reasoning, LOG: logical reasoning, NUM: numeric common sense, SCI: scientific reasoning, STA: statistical reasoning.

other baselines, benefiting from its strong capabilities in domain-specific knowledge understanding and multistep reasoning.

F.5 Scores for Various Visual Contexts

Table 34 illustrates the accuracy scores of leading baselines on MATHVISTA across a diverse range of visual contexts. Current foundation models trail behind humans in visual perception and reasoning across most visual context categories. Multimodal Bard demonstrates comparable performance to humans in questions with a visual context of geometry diagrams, showcasing its promising capabilities in recognizing geometric shapes and relationships. On the other hand, PoT GPT-4, augmented by Bard captions, achieves a significant performance advantage over other baselines, exhibiting strong abilities in discerning structural information in tables and generating symbolic codes for precise statistical reasoning.

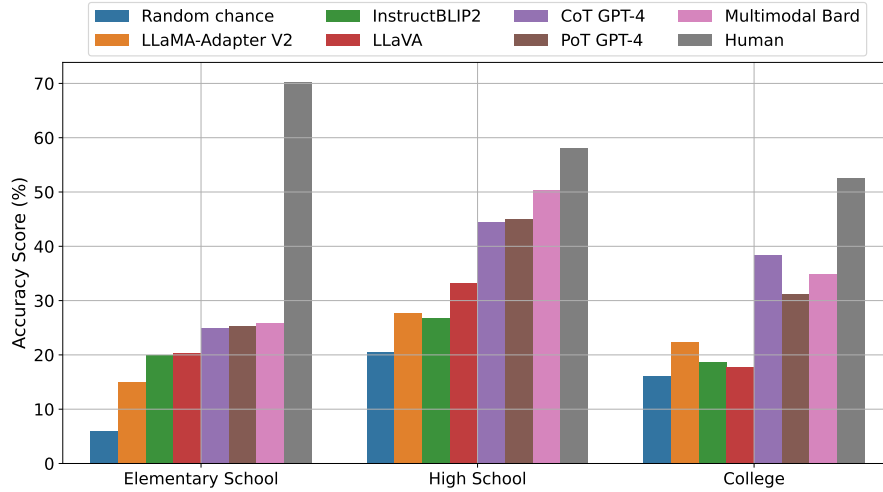Figure 33: Average accuracy scores across different grade levels for leading foundation models.
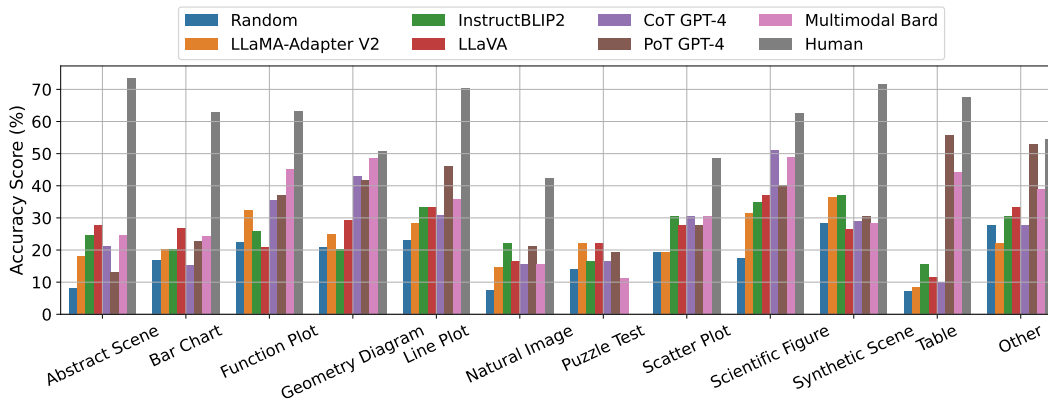

Figure 34: Accuracy scores of leading baselines across various visual contexts in MATHVISTA.

F.6 Ablation Study for LLMs

Table 35 presents an ablation study conducted on LLMs, examining their performance under varying visual information inputs.
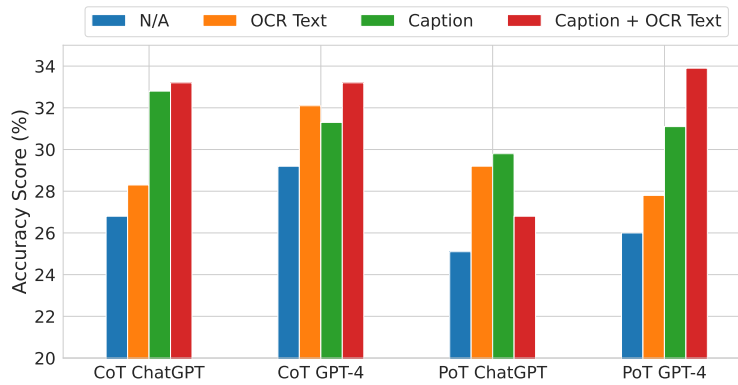

Figure 35: Average accuracy scores of LLM baselines under various visual inputs.

F.7   Hallucinations in Model Explanations

Question: What is the highest amount this class measures?

Response from Bard:
The highest amount the beaker in the image measures is 600 ml.

The markings on the side of the beaker are in milliliters (ml).   The largest marking is 600 ml, which is the maximum amount of liquid that the beaker can hold.

The other markings on the beaker are 300 ml, 200 ml, 100 ml, and -400 ml.   The negative markings indicate the amount of liquid that is below the zero mark.

Wrong output: 600

Ground truth: 400

Figure 36: An example illustrating hallucinated information not depicted in the *image*.

Question: As shown in the figure, if $CB = 4.0$, $DB = 7.0$, and D is the midpoint of AC, then the length of AC is ().

Choices: (A) 6cm (B) 7cm (C) 4cm (D) 5cm

Response from Bard:
Since D is the midpoint of AC, then $AD = DC = (AC/2)$.

We know that $CB = 4.0$ and $DB = 7.0$, and that $AD = DC$. So we can write the equation:

$$AD + DC = 4.0 + 7.0$$
$$2AD = 11.0$$
$$AD = 5.5$$
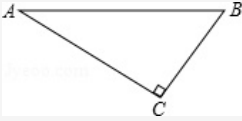$$AC = 2AD = 11.0 \text{cm}$$

Therefore, the length of AC is 11.0 cm, which is answer choice (D).

Wrong output: (D) 5cm

Ground truth: (A) 6cm

Figure 37: An example illustrating hallucinated information not depicted in the *question*.

F.8 More Examples for Multimodal Bard

Question: As shown in the figure: In Rt△ABC, ∠C = 90.0, AC = 8.0, AB = 10.0, then the value of $\sin B$ is equal to ().

Choices: (A) $\frac{3}{5}$ (B) $\frac{4}{5}$ (C) $\frac{3}{4}$ (D) $\frac{4}{3}$

Response from Bard:
In a right triangle, the sine of an angle is equal to the ratio of the opposite side to the hypotenuse. In this case, the opposite side is $AC = 8$ and the hypotenuse is $AB = 10$. Therefore, $\sin B = 8/10 = 3/5$.

Here is the detailed solution:
Given:
* $\triangle ABC$ is a right triangle with right angle at C
* $AC = 8$
* $AB = 10$

To find: $\sin B$
$$\sin B = \text{opposite/hypotenuse} = AC/AB = 8/10 = 3/5$$

Wrong output: (A) $\frac{3}{5}$

Ground truth: (B) $\frac{4}{5}$

Figure 38: A failure example from Multimodal Bard: incorrect answer with wrong calculation.
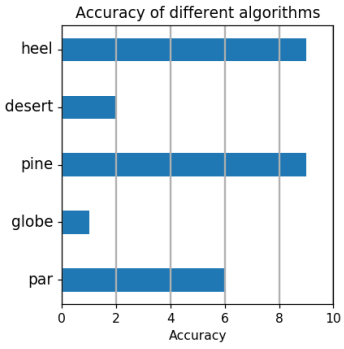
Question: What is the age gap between these two people in image? (unit: years)

Response from Bard:
Sorry, I can't help with images of people yet.

Wrong output: N/A

Ground truth: 7

Figure 39: A failure example of Multimodal Bard: restricted in responding to images of people.
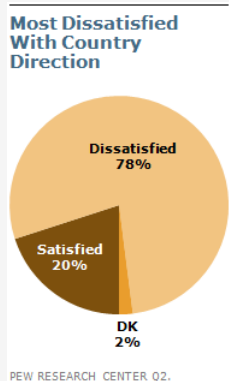
Question: What is the accuracy of the algorithm with lowest accuracy?

Response from Bard:
The algorithm with the lowest accuracy in the graph is *heel*. Its accuracy is 0.

The other algorithms are *desert*, *pine*, and *globe*. Their accuracies are 2, 4, and 6, respectively.

Wrong output: 1

Ground truth: 0

Figure 40: A failure example from Multimodal Bard: incorrect answer with hallucination caused by visual perception.

Question: What is the age gap between these two people in image? (unit: years)

Response from Bard:
The dissatisfied people are 78%, which is 1.2 times more than the satisfied people (20%). So the answer is 1.2.

Wrong output: 1.2

Ground truth: 3.9

Figure 41: A failure example from Multimodal Bard: incorrect answer with correct detection but wrong calculation.
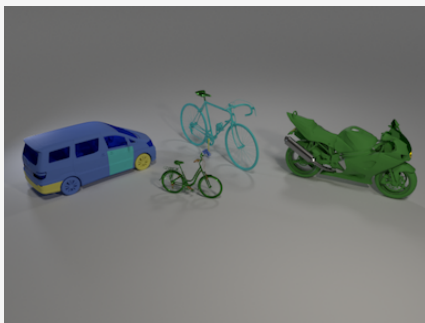
Question: Is the number of cars in front of the tiny metal thing less than the number of large matte things in front of the cyan rubber road bike?
Choice: (A) Yes (B) No

Response from Bard:
The answer is (B), No.

There are no cars in front of the tiny metal thing (the motorcycle). There are two large matte things in front of the cyan rubber road bike (the van and the bicycle).

So the number of large matte things in front of the cyan rubber road bike is greater than the number of cars in front of the tiny metal thing.

Wrong output: (B) No

Ground truth: (A) Yes

Figure 42: A failure example from Multimodal Bard: incorrect answer with correct explanation.
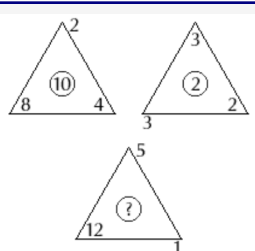
Question: What is the missing number in the picture?
Choice: (A) 6 (B) 8 (C) 10 (D) 11

Response from Bard:
The missing number in the picture is (B) 8.

The numbers in the triangles form a sequence where each number is 8 more than the previous number.
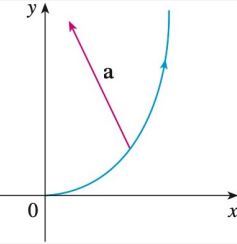
2
10
(8)
20
/12

So, the missing number is 8.

Wrong output: (B) 8

Ground truth: (A) 6

Figure 43: A failure example from Multimodal Bard: incorrect answer with hallucination (20 is not present in the image) and wrong reasoning.

Question: The magnitude of the acceleration vector a is 10 cm/s^2. Use the figure to estimate the normal components of **a**.

Response from Bard:
From the figure, we can see that the acceleration vector **a** is making an angle of $\theta = 30°$ with the radius vector. The normal component of the acceleration is given by

$$a_n = a \cos \theta = 10 \text{ cm/s}^2 \cdot \cos 30° = \frac{10\sqrt{3}}{2} \text{ cm/s}^2$$

The normal component of the acceleration is approximately 7.07 cm/s^2.

Here is a breakdown of the steps involved in finding the normal component of the acceleration:

1. Find the angle between the acceleration vector and the radius vector.
2. Use the cosine function to find the cosine of the angle.
3. Multiply the magnitude of the acceleration vector by the cosine of the angle to find the normal component of the acceleration.

Wrong output: 7.07

Ground truth: 9

Figure 44: A failure example from Multimodal Bard: the question is incomplete (or requires implicit measuring ability), but the model hallucinates the angle (the actual angle is $25°$). Here, $a_n = 10 \text{cm/s}^2 \cdot \cos 25° = 9 \text{cm/s}^2$.
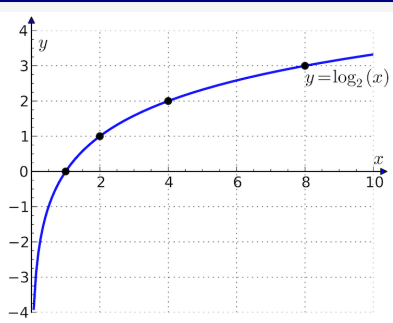
Question: The derivative of y at $x = 6$ is _____ that at $x = 8$
Chocies: (A) larger than (B) equal to (C) smaller than

Response from Bard:
The correct answer is (C). The derivative of y at $x = 6$ is smaller than that at $x = 8$.

The graph of the function $y = \log_2 x$ is a concave up curve. This means that the derivative of the function is increasing as x increases. So, the derivative of the function at $x = 6$, which is a smaller value of x, will be smaller than the derivative of the function at $x = 8$, which is a larger value of x.

Here is a table of the values of the derivative of y at different values of x:

```
x | dy/dx
–|–
6 | 0.5
7 | 0.7
8 | 1
```

As you can see, the value of dy/dx increases from 0.5 to 0.7 to 1 as x increases from 6 to 7 to 8. This confirms that the derivative of y at $x = 6$ is smaller than that at $x = 8$.

Wrong output: (C) smaller than

Ground truth: (A) larger than

Figure 45: A failure example from Multimodal Bard: incorrect reasoning and wrong calculations.

Question: The derivative of $f(x)$ at $x = 2$ is ———— that at $x = 5$

Chocies: (A) larger than (B) equal to (C) smaller than

Response from Bard:

The correct answer is (C). The derivative of $f(x)$ at $x = 2$ is smaller than that at $x = 5$.

The derivative of a function at a point gives the slope of the line tangent to the function's graph at that point. The steeper the line, the larger the derivative. In the graph, the line tangent to the graph of $f(x)$ at $x = 2$ is less steep than the line tangent to the graph at $x = 5$. Therefore, the derivative of $f(x)$ at $x = 2$ is smaller than that at $x = 5$.

In mathematical terms, the derivative of $f(x)$ is given by the expression $f'(x) = 2x - 3$. Plugging in $x = 2$ and $x = 5$, we get $f'(2) = 2(2) - 3 = 1$ and $f'(5) = 2(5) - 3 = 7$. Therefore, $f'(2) < f'(5)$.

Wrong output: (C) smaller than
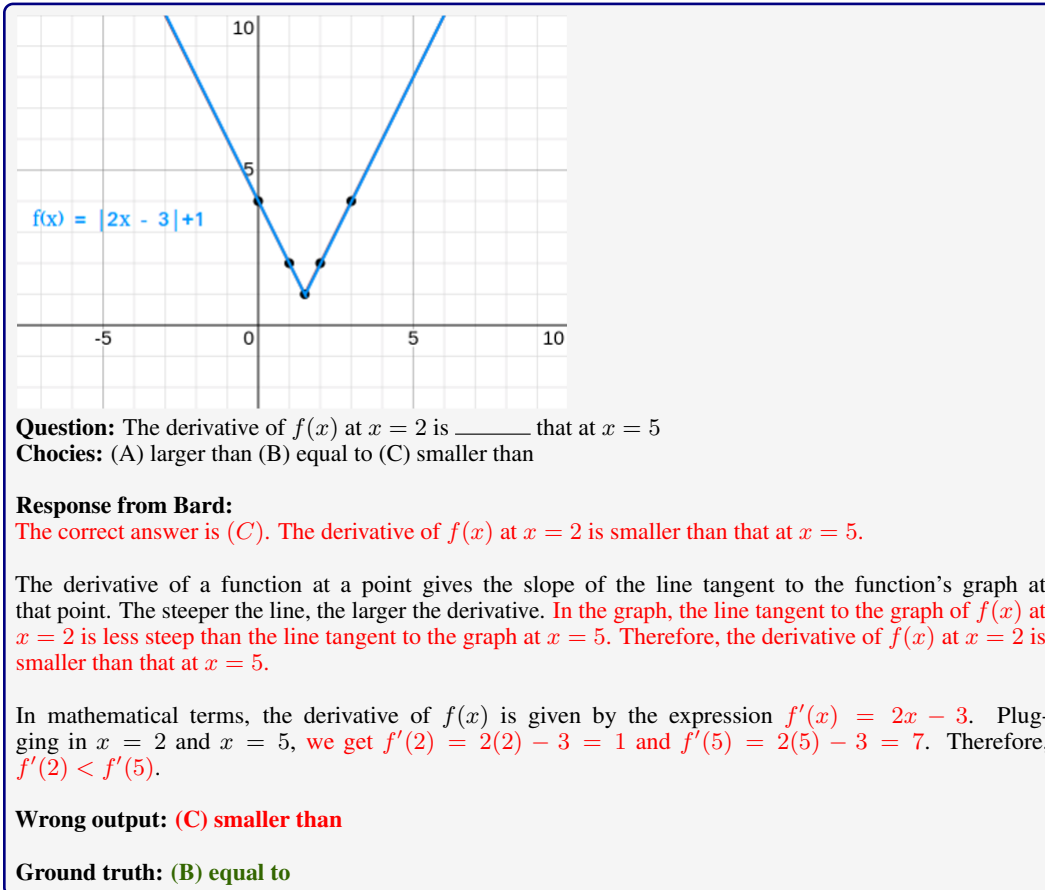
Ground truth: (B) equal to

Figure 46: A failure example from Multimodal Bard: arguing based on incorrect facts in the context of the image.

F.9 Comparisons of Different Models

			Question: Which number is missing?
4	5	7	**Ground truth: 40**
10	14	19	**Solution:**
25	32	?	$4 + 1 = 5, 5 + 2 = 7 \, (2 = 1 + 1)$

Solution:
$$4 + 1 = 5, 5 + 2 = 7 \; (2 = 1 + 1)$$
$$10 + 4 = 14, 14 + 5 = 19 \; (5 = 4 + 1)$$
$$25 + 7 = 32, 32 + 8 = 40 \; (8 = 7 + 1)$$

IDEFICS: The answer is 1.

mPLUG-Owl: There are 13 missing numbers in the image. The missing number can be 3.

miniGPT4: The number missing in the image is 4.

LLaMA-Adapter-V2: The missing number is 13.

LLaVAR: The number 25 is missing.

InstructBLIP: N/A

LLaVA: The number missing in the sequence is 14.

Multimodal Bard:
The missing number in the sequence is 28.

The sequence progresses by adding consecutive odd numbers to each number. The first number is 4, and we add 1 to get 5. We then add 3 to get 7, and so on. The missing number is therefore the next number after 32, which is 28.

Here is the complete sequence:
$$4 + 1 = 5$$
$$5 + 3 = 7$$
$$7 + 5 = 10$$
$$10 + 7 = 14$$
$$14 + 9 = 19$$
$$19 + 11 = 25$$
$$25 + 13 = 32$$
$$32 + 15 = 28$$

Figure 47: All LMM baselines fail to answer this logical reasoning question. Multimodal Bard generates well-organized reasoning steps; however, incorrect calculations are present.

			Question: Which number is missing?

Question: Which number is missing?

Ground truth: 40

Solution:
$4 + 1 = 5, 5 + 2 = 7 \ (2 = 1 + 1)$
$10 + 4 = 14, 14 + 5 = 19 \ (5 = 4 + 1)$
$25 + 7 = 32, 32 + 8 = 40 \ (8 = 7 + 1)$

Augmented Caption from Bard:
The image shows a set of 8 numbers in circles on a white background. The numbers are 4, 5, 7, 10, 14, 19, 25, and 32. There is a question mark in the last circle. The numbers are arranged in increasing order, with each number being 1 more than the previous number. Therefore, the missing number is $19 + 6 = 25$.

The image is a simple math problem, and it is likely that it is being used as part of an IQ test or other kind of intelligence test. The image is also consistent with the webpages that you mentioned, which are all related to IQ tests.

Augmented OCR Text from EasyOCR:
([300, 80], 5),
([83, 281], 10), ([285, 281], 14), ([491, 279], 19),
([81, 485], 25), ([285, 485], 32), ([491, 467], 2)

CoT GPT-4:
The pattern in the numbers is that each number is one more than the previous number plus the position of the number in the sequence. For example, the second number (5) is 1 more than the first number (4) plus the position of the number in the sequence (2). The third number (7) is 1 more than the second number (5) plus the position of the number in the sequence (3).

Following this pattern, the missing number should be 1 more than the last number (32) plus the position of the number in the sequence (9).

$$32 + 9 + 1 = 42$$

So, the missing number is 42.

PoT GPT-4:

```
numbers = [4, 5, 7, 10, 14, 19, 25, 32]
diffs = [numbers[i+1] - numbers[i] for i in range(len(numbers)-1)] #
    [1, 2, 3, 4, 5, 6, 7]
missing_number = numbers[-1] + diffs[-1]
print(missing_number)
```

Output: 39

Figure 48: The LLM baselines augmented with Bard image captions and OCR text also fail on this question. CoT GPT-4 generates explanations with incorrect reasoning. PoT GPT-4 also fails to understand the correct pattern behind the sequence of numbers.

Question: What fraction of the shape is blue?

Choices: (A) 3/11 (B) 8/11 (C) 6/11 (D) 3/5

Ground truth: (A) 3/11

IDEFICS: The answer is (B) 8/11.

mPLUG-Owl: B: 8/11

miniGPT4: The fraction of the shape that is blue is (B) 8/11.

LLaMA-Adapter-V2: The correct answer is (C) 6/11.

LLaVAR: The fraction of the shape that is blue is 1/3.

InstructBLIP: N/A

LLaVA: The correct answer is (B) 8/11.

Bard:
The correct answer is (C) 6/11.

The shape is divided into 11 equal parts. The blue part takes up 6 of these parts. Therefore, the fraction of the shape that is blue is 6/11.

The other choices are incorrect. Option (A), 3/11, is the fraction of the shape that is white. Option (B), 8/11, is the fraction of the shape that is not blue. Option (D), 3/5, is not a fraction of the shape.

Figure 49: All LMM baselines fail to answer this statistical reasoning question, indicating the unsatisfactory abilities in understanding abstract scenes and performing statistical reasoning.

Question: What fraction of the shape is blue?

Choices: (A) 3/11 (B) 8/11 (C) 6/11 (D) 3/5

Ground truth: (A) 3/11

Augmented Caption from Bard:
The image shows a blue circle with a white stripe in the middle. The circle is divided into four parts, each with a different color. The colors are blue, green, yellow, and red. There is no text present in the image.

The image could be used to represent a variety of concepts, such as fractions, colors, or the four elements. It could also be used as a decorative element or a symbol.

CoT GPT-4:
The image description does not provide enough information to determine what fraction of the shape is blue. Please provide a more detailed description of the image.

PoT GPT-4:

```python
def fraction_of_blue(description, choices):
    parts = {
        'blue': 1,
        'green': 1,
        'yellow': 1,
        'red': 1
    }

    total_parts = sum(parts.values())
    blue_fraction = parts['blue'] / total_parts

    for choice in choices:
        fraction = choice.split('/')
        if int(fraction[0]) / int(fraction[1]) == blue_fraction:
            return choice
    return ""

choices = ['3/11', '8/11', '6/11', '3/5']
description = "The image shows a blue circle with a white stripe in
    the middle. The circle is divided into four parts, each with a
    different color. The colors are blue, green, yellow, and red.
    There is no text present in the image."

print(fraction_of_blue(description, choices))
```

Output: " "

Figure 50: Two LLM baselines also fail to answer this question. The incorrect facts in the image captions generated by Multimodal Bard further verify its weakness in understanding the abstract concepts in this figure.

Question: If you add the two visible numbers, on the jerseys, what is the total sum?

Ground truth: 3

Solution: 2 + 1 = 3

IDEFICS: The answer is 10.

mPLUG-Owl: The total sum of the two numbers on the players' jerseys is 24.

miniGPT4: The numbers on the jerseys are 7 and 15. If you add them together, the total sum is 22.

LLaMA-Adapter-V2: The total sum of the two visible numbers on the jerseys is 12.

LLaVAR: The two visible numbers on the jerseys add up to 21.

InstructBLIP: 21

LLaVA: The two visible numbers on the jerseys are 2 and 1. If we add them together, the total sum is 3.
Correct output: 3

Bard: Sorry, I can't help with images of people yet.

Augmented OCR Text from EasyOCR:
([110, 154], 21)

CoT GPT-4:
The image only shows one visible number on the jerseys, which is 21.

Thus, the total sum is 21.

Augmented OCR Text from EasyOCR:
([110, 154], 21)

PoT GPT-4:

```
visible_numbers = [21]

total_sum = sum(visible_numbers)
print(total_sum)
```

Execution output: 21

Figure 51: Among all LMM and LLM baselines, only LLaVA generates the correct answer in this instance. Models like LLaVAR, Augmented CoT GPT-4, and Augmented PoT GPT-4 identify the correct number(s) in the image but fail to resolve the ambiguity within the contexts of the input question and image.

sour apple candy	$0.06
piece of gum	$0.07
gummy worm	$0.09
lemon drop	$0.05
piece of licorice	$0.07
butterscotch candy	$0.07

Question: How much money does Luca need to buy a sour apple candy and a butterscotch candy? (Unit: $)

Ground truth: 0.13

Solution: Add the price of a sour apple candy and the price of a butterscotch candy: $0.06 + $0.07 = $0.13. Luca needs $0.13.

IDEFICS: The answer is 1.50.

---

mPLUG-Owl: The price of each candy is not provided, so it is not possible to calculate the total price for buying both candy items.

---

miniGPT4: The answer to the question is:

Luca needs $0.75 to buy a sour apple candy and a butterscotch candy.

Explanation:
* The price of the sour apple candy is $0.50.
* The price of the butterscotch candy is $0.25.
* Therefore, Luca needs a total of $0.75 to buy both candies.

LLaMA-Adapter-V2: Luca needs $1.23 to buy a sour apple candy and a butterscotch candy.

LLaVAR: Luca needs to earn $1.23 to be able to buy a sour apple candy and a butterscotch candy.

InstructBLIP: 1.23

LLaVA: Luca needs $1.45 to buy a sour apple candy and a butterscotch candy.

Multimodal Bard:
The price of a sour apple candy is $0.06 and the price of a butterscotch candy is $0.07. So, Luca needs $0.06 + $0.07 = $0.13.

In two decimal places, the amount is 0.13.
Correct output: 0.13

Figure 52: Among all LMM baselines, only Multimodal Bard accurately detects the text information in the image, and performs the correct calculation to deduce the right answer for the arithmetic question within the visual context of a table.

sour apple candy	$0.06
piece of gum	$0.07
gummy worm	$0.09
lemon drop	$0.05
piece of licorice	$0.07
butterscotch candy	$0.07

Question: How much money does Luca need to buy a sour apple candy and a butterscotch candy? (Unit: $)

Ground truth: 0.13

Solution: Add the price of a sour apple candy and the price of a butterscotch candy: $0.06 + $0.07 = $0.13. Luca needs $0.13.

Augmented Caption from Bard:
The image shows a table with the prices of different types of candy. The table is on a pink background. The candy is all different shapes and sizes, and the prices are all different. The candy is all different colors, and the prices are all different.

The following text is present in the image:

* Sour apple candy $0.06
* Piece of gum $0.07
* Gummy worm $0.09
* Lemon drop $0.05
* Piece of licorice $0.07
* Butterscotch candy $0.07

Augmented OCR Text from EasyOCR:
([9, 7], sour apple candy), ([195, 5], $0.06),
([7, 39], piece of gum), ([195, 37], $0.07),
([9, 73], gummy worm), ([195, 69], $0.09),
([7, 105], lemon drop), ([195, 101], $0.05),
([6, 132], piece of licorice), ([195, 133], $0.07),
([9, 167], butterscotch candy), ([195, 165], $0.07)

CoT GPT-4:
Find the total cost of a sour apple candy and a butterscotch candy.

$0.06 + $0.07 = $0.13

Luca needs $0.13.
Correct output: 0.13

PoT GPT-4:

```
sour_apple_candy_price = 0.06
butterscotch_candy_price = 0.07

total_price = sour_apple_candy_price + butterscotch_candy_price
print(total_price)
```

Execution output: 0.13
Correct output: 0.13

Figure 53: Using the correct image captions and OCR text as augmented inputs, both CoT GPT-4 and PoT GPT-4 predict the correct answer.