

# What Is the True Performance of Large Multimodal Models in Visual Context-Based Mathematical Reasoning? An Analysis of Multiple Datasets and Future Research Directions

1<sup>st</sup> Jieun Park, Sungeun Park  
*AI Lab*  
*Tutorus Labs Inc.*  
Daejeon, Republic of Korea  
geeg51, parkse@tutoruslabs.com

2<sup>nd</sup> Hyungbae Jeon  
*AI Lab*  
*Tutorus Labs Inc.*  
Daejeon, Republic of Korea  
hbjeon@tutoruslabs.com

3<sup>rd</sup> Joon-Ho Lim  
*AI Lab*  
*Tutorus Labs Inc.*  
Daejeon, Republic of Korea  
jhlim@tutoruslabs.com

**Abstract**—In this paper, we study the capability of visual context-based mathematical reasoning within the rapidly evolving field of Large Multimodal Models (LMMs). Achieving visual context-based mathematical reasoning requires cognitive skills similar to those used in human reasoning, as it involves interpreting diverse visual elements and converting them into mathematical representations. Previous research has explored various methodologies to address this challenge, but these approaches tend to work only under specific conditions and are often constrained by the limitations of available datasets. As a result, this paper offers a comprehensive analysis of datasets related to visual context-based mathematical reasoning and evaluates the effectiveness of LMMs using these datasets. We aim to identify the limitations of existing techniques and suggest future research directions.

**Index Terms**—Visual context-based mathematical reasoning, mathematical reasoning, mathematical reasoning, Large Multimodal Models

## I. INTRODUCTION

In recent years, the significant advancement of Large Multimodal Models (LMMs) [1]–[3] has spurred numerous innovations across various application domains. Among them, visual context-based mathematical reasoning is notably critical, as it addresses the challenge of comprehending and processing intricate visual information prevalent in our daily experiences. For example, in education, presenting math problems visually can help students understand them more intuitively. It can also help them progress in science and other areas of everyday life. [4]

The ability to perform visual context-based mathematical reasoning requires skills beyond simple computations. It necessitates the interpretation of diverse visual elements, such as shape, position, color, and size of objects, and their conversion into mathematical representations. This complex process is similar to human cognitive abilities, and reasoning at a similar level remains a challenge for LMMs. Therefore, continued research is essential to fulfill this need. [5]

Prior studies have explored various approaches to visual context-based mathematical reasoning. [6], [7] However, many

such studies demonstrate performance efficacy only under specific conditions or are hindered by dataset limitations. For example, models optimized for a specific type of problem may not be able to cope with other types of problems, which hinders the generalization ability of the model. In addition, the variety and quality of the datasets used to train the models can also be a limitation, which can hinder their practical application.

Therefore, this paper aims to build on the achievements of previous studies, comprehensively investigate and analyze datasets that perform mathematical reasoning within visual contexts, and analyze the performance of LMMs based on them. By doing so, we hope to contribute to the advancement of this field by clarifying the limitations of current techniques and suggesting future research directions.

### Key research questions include:

1. What are the characteristics, advantages, and limitations of various mathematical reasoning datasets within visual contexts?
2. How do LMMs perform across these datasets?
3. What are the limitations and challenges faced by the datasets and models, and how can they be overcome?

The structure of this paper is as follows. First, we survey and categorize various datasets related to visual context-based mathematical reasoning. Then, the performance of LMMs on these datasets is evaluated and analyzed based on the existing literature. Finally, we discuss the limitations of the models and future research directions with suggestions to overcome them.

## II. MULTIMODAL MATH REASONING DATASET

To study visual context-based mathematical reasoning, it is essential to have high-quality datasets that contain a variety of cases. These datasets provide the basic material for training and evaluating models, and also form the basis for predicting performance in real-world applications. Without appropriate datasets, the reliability of a model's performance cannot be assured, and the direction of research may falter.

TABLE I  
FORMAT OF MATHEMATICAL DATASETS FOR VISUAL REASONING

Dataset	Question	Image	Choices	Unit	Answer	Question Type	Answer Type	Query	PID	Grade
MathVista	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
MathVerse	✓	✓	✓		✓	✓		✓	✓	
MathV360K	✓	✓	✓		✓			✓	✓	
MathVision	✓	✓	✓		✓				✓	
We-Math	✓	✓	✓		✓				✓	

TABLE II  
OVERVIEW OF MATHEMATICAL DATASETS FOR VISUAL REASONING

Dataset	Data Count	Level	Note	Format	Submitted
MathVista	6.1K	E, H, C	Question Language : English, Chinese, Persian	parquet	2023.10
MathVerse	4.7K	H	Six distinct versions : TextDominant, TextLite, Vision Intensive, Vision Dominant Vision Only, Text Only	parquet	2024.03
MathV360K	339K	-	Incorporates 24 existing datasets	json	2024.06
MathVision	3.3K	E, M, H	Problems are put into difficulty levels 1-5	parquet	2024.02
We-Math	1.7K	E, M, H	Decomposed problem into 2 step or 3 step by concept	json	2024.07

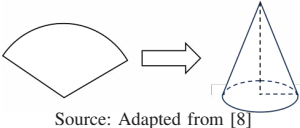
(E: Elementary M: Middle School H: High school C:College)  
The Data Count in the table is based on the dataset available online on HuggingFace.

Such datasets should be designed to include a range of visual scenarios and complex mathematical problems. This helps to broaden the applicability of models and improve their ability to solve realistic problems. Among the many datasets available, we will focus specifically on mathematical datasets. The selected data consists of five different datasets.

#### A. MathVista

The MathVista [8] consists of 6,141 samples collected from 31 different datasets. It significantly enhances the diversity and complexity of visual recognition and mathematical reasoning tasks by integrating the MathQA [9] dataset with 19 VQA [10] datasets. MathVista dataset integrates a total of 28 existing multimodal datasets. The dataset is divided into 5,140 samples for training and 1,000 samples for testing.

TABLE III  
MATHVISTA QA EXAMPLE

Question	Answer	Image
Use a sector paper sheet with a central angle of 120.0 and a radius of 6.0 to roll into a conical bottomless paper cap (as shown in the picture), then the bottom perimeter of the paper cap is ()	$4\pi\text{cm}$	

The dataset identifies seven types of mathematical reasoning: algebraic reasoning, arithmetic reasoning, geometry reasoning, logical reasoning, numerical reasoning, scientific reasoning and statistical reasoning. It focuses on five primary tasks of figure question answering (FQA), geometry problem solving (GPS), mathematical word problem solving (MWP), textbook question answering (TQA) and visual question answering (VQA). In the testmini dataset, which contains 1,000 data points, there are two types of questions: free-response

and multiple-choice, which make up 46% and 54% of the dataset respectively. In the larger test dataset, which contains 5,140 data points, free-form questions account for 44.5%, while multiple-choice questions account for 55.5%. The image and question-answer data in Table. III belong to the VQA type of data within the MathVista dataset. MathVista includes items, answers, and images, as well as metadata that includes question type, answer type, task category, grade level, visual context, and the reasoning skills required. This comprehensive information about the data facilitates detailed evaluation, demonstrating the depth and usefulness of the dataset in educational contexts.

#### B. MathVerse

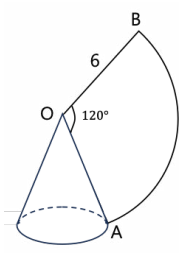
The MathVerse [11] is a collection of 2,612 math problems with diagrams were collected. To avoid limiting performance, they collected problems mainly high school level problems, requiring moderate mathematical knowledge, and excluded college level topics such as calculus and graph theory.

They classified the content of the text in the problem into three categories: Descriptive Information, Implicit Property, and Essential Condition. This classified text was then progressively condensed, with information being increasingly integrated into images, resulting in six detailed versions.

Each problem is divided by humans into six versions : Text Dominant, Text Lite, Text Only, Vision Intensive, Vision Dominant, and Vision Only, resulting in a total of 15,000 data. By using these six different versions of the approach, it is possible to comprehensively assess how well visual diagrams can be understood for mathematical reasoning and to what extent they can be effectively utilized.

This paper will focus on the Vision Intensive version, one of the six types of MathVerse. As of now, there are 4.7K publicly available datasets on Hugging Face, with free-form and multiple-choice questions accounting for 44.7% and 55.3%, respectively.

TABLE IV  
MATHVERSE VISION DOMINANT QA EXAMPLE

Question	Answer	Image
As shown in the figure, If OA and OB are overlapped to form a cone side, the diameter of the bottom of the cone is () Choices: A:2cm B:4cm C:3cm D:5cm	B	 <p>Source: Adapted from [11]</p>

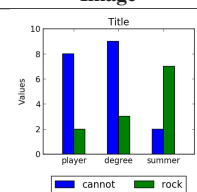
MathVerse consists of items, answers, images, as well as information such as question type, answer type, query-cot, and problem-version. The questions and answers in Table. IV belong to the GeoQA [12] dataset. The sentence "There is a sector with a central angle of 120.0 and a radius of 6.0" in Table. IV does not exist in the Vision Dominant version. Instead, this information is displayed in the image of Table. IV

### C. MathV360K

The MathV360K [7] was created by collecting 40,000 high-quality images and question-answer pairs from 24 open-source multimodal question-answer datasets and synthesizing 320,000 new pairs. This dataset is used as an instructional dataset for LMMs.

Based on the image dataset, they classified images into two categories: Image Clarity and Image Comprehension Complexity. In Image Clarity, label 0 signifies low-quality images, while label 1 denotes high-quality images. For Image Comprehension Complexity, lower scores (closer to 0) imply images that are easier to understand, while higher scores (nearing 3) represent images with more challenging contextual comprehension. The tasks are categorized into FQA, MWP, GPS, TQA, and VQA, with visual contexts including bar charts, tables, geometry diagrams, scientific figures, natural images, and more.

TABLE V  
MATHV360K QA EXAMPLE

Question	Answer	Image
What is the value of the largest individual bar in the whole chart?	The answer is 9	 <p>Source: Adapted from [7]</p>
What is the difference in value between the 'cannot' and 'rock' categories for 'degree'?	The answer is 6	
What is the total value of the 'player' category for both 'cannot' and 'rock'?	The answer is 10	

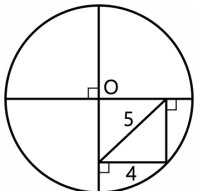
Despite the potential for additional questions, the visual information of the images was not fully utilized. To address this, GPT-4V [13] with Few-Shot Learning was employed to

generate additional questions for each image and augmented the original question by mining the image.

However, MathV360K does not provide information on the difficulty level or type of the problems, apart from the questions, answers, and choices.

The question and answer data in Table. V is from the DVQA [14] dataset within the collected MathV360K data.

TABLE VI  
MATHVISION QA EXAMPLE

Question	Answer	Image
The point O is the center of the circle in the picture. What is the diameter of the circle?	10	 <p>Source: Adapted from [15]</p>

### D. MathVision

The MathVision [15] contains 3,040 data points collected from 16 different area of mathematics and it is categorized into five levels of difficulty. This dataset contains high quality mathematical problems within visual contexts, carefully curated from actual mathematical competitions. Table. VI illustrates cases from MathVision. The subject pertains to metric geometry - length, and the problem is in a free-form format. The test dataset, comprising 3,040 data points, includes both free-form and multiple-choice questions, with each type representing 49.6% and 50.4% of the dataset, respectively. In contrast, the testmini dataset, containing 304 data points, features free-form questions at 37.5% and multiple-choice questions at 62.5%. Information regarding the types and levels of each problem is a unique feature of MathVision.

MathVista comprises many similar problems from different source datasets, resulting in limited problem diversity. After eliminating duplicate problems with identical stem text, only 4,740 unique problems remain. Additionally, this collection includes many template problems with only a few words altered. There are three main types of problems featuring abstract scenes in MathVista, which account for over 90% of the total problems. For instance, a typical question in MathVista related to function graphs poses simple and concise questions about the depicted function graph. In contrast, MathVision includes more complex function concepts, such as symmetry and periodic functions, and has longer questions.

The MathVision dataset also comprises questions on topology and graph theory, two categories absent in MathVista, requiring complex visual recognition and mathematical reasoning.

### E. We-Math

The We-Math [16] contains a total of 6.5K items and covers five domains, including 67 hierarchical knowledge concepts

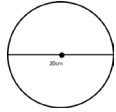

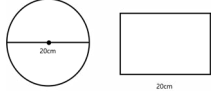
TABLE VII  
PERFORMANCE METRICS BY MODEL AND DATASET

Model Type	Dataset	MathVista	MathVision	MathVerse	We-Math
	Model				
-	Human	60.3	68.82	64.9	-
close-LMMs	GPT-4o	63.8 (+3.5)	30.39 (-38.43)	-	42.86
	GPT-4V	49.9 (-10.4)	22.76 (-46.06)	33.6 (-31.3)	31.05
	Gemini 1.5 Pro	52.1 (-8.2)	19.24 (-49.58)	-	26.38
	Qwen-VL-Max	-	15.59 (-53.23)	23.3 (-41.6)	10.48
	Qwen-VL-Plus	43.3 (-17)	10.72 (-58.1)	13.0 (-51.9)	-
open-LMMs	Math-LLaVA-13B	46.6 (-13.7)	15.69 (-53.13)	-	-
	SPHINX-MoE	42.3 (-18)	-	12.5 (-52.4)	-
	SPHINX-Plus	36.8 (-23.5)	-	13.5 (-51.4)	-

The numbers in parentheses within the table represent the performance difference between each LMMs and human. A red '+' symbol indicates that a model outperforms humans, while a blue '-' symbol indicates underperformance.

and areas such as Measurement, Position and Direction, and Transformation and Motion of Figures. Depending on the knowledge concepts required to solve the problems, complex problems are decomposed into sub-problems, resulting in changes to the questions and images in the sub-problems from the original problems. There are currently 1.7K publicly available datasets on Hugging Face. The We-Math introduces a new 4-dimensional metric called Insufficient Knowledge (IK), Inappropriate Generalization (IG), Complete Mastery (CM), and Rote Memorization (RM) to hierarchically evaluate problems inherent in the reasoning process.

TABLE VIII  
WE-MATH 2STEPS QUESTION AND ANSWER

Step	Question	Answer	Image
1	What is the area of the circle in the diagram below (in square centimeters)? ( $\pi = 3.14$ )	A	 Source: Adapted from [16]
2	In the diagram below, the area of the rectangle is 314 square centimeters. The width of the rectangle is ( ) cm. ( $\pi = 3.14$ )	C	 Source: Adapted from [16]
Multi	The areas of the two shapes below are equal. The width of the rectangle is ( ) cm. ( $\pi = 3.14$ )	C	 Source: Adapted from [16]

The We-Math is composed of multiple-choice questions, where participants choose an answer from options A, B, C, and D and includes knowledge concept descriptions for solving problems. The following Table. VIII illustrates the transformation of a multi-step question into a two-step format. Based on the knowledge concepts required to solve the problem, it guides the student to first calculate the area of the circle and then determine the width of the rectangle based on the given area condition.

We have organised an overview of the mathematics datasets

previously examined in the Table. I and Table. II. The information presented in the table is derived from datasets available on Hugging Face. It provides details such as the size of each dataset, difficulty level, features, format, and submission date of the paper. MathV360K integrates 24 datasets, resulting in a variety of difficulty levels. Consequently, specific difficulty markings are not provided.

### III. PERFORMANCE ANALYSIS

In this chapter, we compare the performance of different LMMs on each dataset and analyze the performance differences based on dataset characteristics. The main models used in this study consist of closed-source LMMs GPT-4o [1], GPT-4V [13], Gemini 1.5 Pro [17], [18], Qwen-VL-Plus [19], Qwen-VL-Max [19], and open-source LMMs Math-LLaVA [7], SPHINX-MoE [20] and SPHINX-Plus [20].

For the comparative analysis of each model's performance, we used four main datasets: MathVerse, MathVista, MathVision, and We-Math. This datasets have in common that they are English-based and designed to solve mathematical problems presented in a visual context. Table. VII reports the performance of the LLMs on each dataset. We will explain the features of each dataset and the performance of the corresponding LMMs based on Table. VII. (MathV360K is omitted because it is designed for instruction fine-tuning).

**MathVista** The MathVista dataset consists of a variety of math problems, geometric diagrams, and text-rich images. The main model, GPT-4o, achieved a performance level of 63.8%, slightly outperforming human performance of 60.3%. The performance of the other close-LMMs is 52.1% for Gemini 1.5 Pro, 49.9% for GPT-4V, 43.3% for Qwen-VL-Plus. The performance of the open-LMMs is 46.6% for Math-LLaVA-13B, 42.3% for SPHINX-MoE, and 36.8% for SPHINX-Plus. MathVista has a relatively straightforward structure for the models to solve, featuring many simple and repetitive question patterns. GPT-4V performed well on geometry problem-solving and algebraic reasoning but did not adequately explain math in a visual context. This limitation can be attributed to inaccuracies in image caption generation and a lack of geometric and mathematical reasoning capabilities.

**MathVision** The highest performance on the MathVision dataset was achieved by GPT-4o with 30.39%, followed by

GPT-4V with 22.76%, Gemini 1.5 Pro with 19.24%, MathLLaVA-13B with 15.69%, Qwen-VL-Max with 15.59% and Qwen-VL-Plus with 10.72%. The dataset consisted of simple problems at the elementary level, but the image-centered reasoning problems posed challenges to the models. Based on GPT-4V, the error types are categorized as follows: reasoning errors(42.2%), visual recognition errors(31.9%), knowledge errors(15.1%), question misunderstood error(6.9%), reject to answer(2.6%), and calculation errors(1.3%). [15] Reasoning errors account for nearly half of all errors, suggesting the limitations of the model’s logical processing capabilities. Visual recognition errors indicate that the model is struggling to interpret visual data accurately.

**MathVerse** In the MathVerse dataset, GPT-4V performed the best with 33.6% performance. Qwen-VL-Max and SPHINX-Plus achieved lower performances at 23.3% and 13.5%, respectively, while SPHINX-MoE and Qwen-VL-Plus attained performances of 12.5% and 13.0%, respectively. This dataset focuses on interpreting mathematical diagrams. The baseline error types for GPT-4V are visual perception errors(42.4%), reasoning errors(36.4%), knowledge errors(12.1%), and calculation errors(9.1%) [11]. Visual perception errors are particularly prominent in advanced diagram interpretation, highlighting the need for improved performance of visual encoders. On the other hand, knowledge errors and calculation errors are relatively low, suggesting that GPT-4V’s understanding of the mathematical concepts used in MathVerse is relatively good.

**We-Math** We-Math is a dataset that decomposes complex problems into 67 knowledge concepts based on textbook knowledge. The performance of the models is as follows: 42.86% for GPT-4o, 31.05% for GPT-4V, 26.38% for Gemini 1.5 Pro, and 10.48% for Qwen-VL-Max. We-Math uses matrices such as IK, IG, CM, and RM to evaluate the reasoning process of LMMs. GPT-4o showed an advantage in handling multiple knowledge concepts, while other LMMs showed limitations in solving multiple knowledge concepts in complex problems. GPT-4o frequently made knowledge errors in more than 45 out of 67 knowledge concepts, and visual errors in about 30 knowledge concepts, particularly in the understanding of certain concepts, such as angles. This indicates that GPT-4o needs to be strengthened with fine-grained measurement capabilities.

In this chapter, we deeply analyzed datasets for performing mathematical reasoning in a visual context and evaluated the performance of LMMs on each dataset. We found that model performance varies depending on the characteristics of each dataset. In the next chapter, we will discuss potential causes of these performance discrepancies and propose improvements to both models and datasets to mitigate these differences in future research.

#### IV. INSIGHTS

In this chapter, we analyze the performance differences of LMMs on different datasets. Our goal is to identify the reasons for two main issues that emerge from this: (1) the

performance differences for each dataset on the same LMMs, and (2) the difference in performance of each LMMs and human performance. We propose five main factors as the causes of these issues.

In this chapter, we conduct a comprehensive analysis of the performance of LMMs on various datasets. Our goal is to identify two main issues: (1) the performance differences for each dataset on the same LMMs, and (2) the difference in performance of each LMMs and human performance. We propose five main factors as the causes of these issues.

**The complexity of the dataset exerts both direct and indirect influences on performance.** MathVista’s experimental results show that LMMs are more adept at coping with college-level problems compared to elementary-level problems. This is related to the learning styles of elementary school students and the limitations of age-specific educational materials. Elementary education materials are mainly abstract and characterised by relatively limited data. As a result, We-Math, which is mainly composed of elementary school-level data, performs relatively poorly, while MathVerse, which targets high school-level problems, outperforms We-Math. These results suggest that the lower the level of the dataset, the more likely the model will perform poorly.

**The potential for contamination of the dataset is also an important consideration.** When evaluating the datasets contained in the training data, LMMs perform well. In the case of Open-LMMs, the training data is explicitly stated, so the potential for data contamination can be assessed, but in the case of Close-LMMs, the training data is often not explicitly disclosed, and the potential for data contamination exists. Examination of models with advanced visual context-based mathematical reasoning, such as GPT-4V and GPT-4o, reveals that GPT-4V was released in September 2023 and GPT-4o in May 2024, with training data sourced up until April 2023 and October 2023, respectively. This suggests the possibility that MathVista, released before these models, might have been included in the LMMs’ training data, thereby posing a risk of data contamination. Conversely, datasets such as MathVision, MathVerse, and We-Math, released after the models, imply a lower risk of contamination. Notably, MathVista outperforms other datasets by approximately 26% on average. These results show that there is potential for dataset contamination in the LMMs.

**The type of dataset can also influence LLMs performance.** In the performance analysis of MathVerse, higher accuracy was observed in text-intensive datasets compared to vision-intensive datasets. This indicates that the LMMs performs better on text-intensive datasets than on image-intensive datasets. Datasets containing various types of questions, such as MathVista, which includes Math Word Problems and Vision Question Answer, tend to show better performance than those focused solely on Vision Question Answer (We-Math, MathVerse, MathVision). These results suggest that Vision Question Answer types of problems tend to perform worse than other types, which is also related to the model’s poor image recognition performance.

**The ability of LMMs to recognise visual data is currently limited.** Performance analysis of GPT-4V on MathVision and MathVerse shows that visual errors constitute over 30% of total errors. The LMMs frequently fail to accurately process visual information. This results in diagram interpretation errors, which are particularly noticeable in problems that require advanced diagram analysis. The observation that text-centric datasets demonstrate higher accuracy than image-centric ones in MathVerse further corroborates the LMMs' limited capability in processing visual data. This deficiency partially accounts for the lower performance of LMMs compared to human counterparts.

**Deficiencies in the reasoning abilities required to solve complex mathematical problems detrimentally impact overall performance.** For GPT-4V, visual errors comprise more than 35% of total errors in MathVision and MathVerse datasets. This indicates that LMMs' reasoning capabilities remain underdeveloped, contributing significantly to the substantial gap between human and model performance.

In this chapter, we have identified the main factors responsible for the performance differences of LMMs on different datasets and the performance gap between humans and LMMs. Future research should focus on addressing these factors to improve the overall performance of LMMs.

## V. CONCLUSION

In this study, we undertake a thorough analysis of multiple datasets and evaluate the performance of LMMs concerning visual context-based mathematical reasoning. The results reveal that model performance is markedly affected by dataset attributes and the training methodology employed. Notable deficiencies of LMMs are linked to inaccuracies in visual perception and cognitive reasoning, underscoring the need for more resilient vision encoders and the augmentation of logical reasoning faculties. To tackle these challenges, it is imperative to enhance the diversity and quality of training datasets and to adopt sophisticated training methodologies. Future research should prioritize the following areas:

**Enhancement of Vision Encoders:** Developing more advanced vision encoders capable of accurately interpreting visual data is essential. Such progress would enable more precise processing of intricate visual data, including diagrams and graphs.

**Strengthening MultiModal Reasoning Capabilities:** To diminish reasoning errors, the exploration of novel algorithms and training methodologies designed to improve multimodal processing and reasoning capabilities is essential.

**Development of Granular Performance Metrics:** Formulating detailed evaluation criteria and methodologies for assessing mathematical reasoning will allow for the precise identification and remediation of model deficiencies.

**Diversification and Quality Enhancement of Datasets:** Creating high-quality datasets that encompass a wide range of visual information and complex mathematical problems will enable models to learn and generalize across a broader spectrum of scenarios.

This research provides critical insights to scholars in the domain by delineating the current state and potential improvements for datasets and LMMs employed in visual context-based mathematical reasoning. With sustained research and enhancements in dataset quality, LLMs-based mathematical reasoning models can be widely applied in education, science, and other fields.

## ACKNOWLEDGMENT

This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (RS-2023-00262158, Development of AI Analytics-Generation-Coaching Copilot Technology for Augmented Teachers' Competency-Customized Education)

## REFERENCES

- [1] Open AI, "Hello GPT-4o," <https://openai.com/index/hello-gpt-4o/>, 2024
- [2] Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui
- [3] Google. Bard, URL <https://bard.google.com/>, 2023.
- [4] SEO, Minjoon, et al. Solving geometry problems: Combining text and diagram interpretation. In: Proceedings of the 2015 conference on empirical methods in natural language processing. 2015. p. 1466-1476.
- [5] YUE, Xiang, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024. p. 9556-9567.
- [6] ZHANG, Renrui, et al. Mathverse: Does your multi-modal llm truly see the diagrams in visual math problems?. arXiv preprint arXiv:2403.14624, 2024.
- [7] SHI, Wenhao, et al. Math-LLaVA: Bootstrapping Mathematical Reasoning for Multimodal Large Language Models. arXiv preprint arXiv:2406.17294, 2024.
- [8] LU, Pan, et al. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. arXiv preprint arXiv:2310.02255, 2023.
- [9] AMINI, Aida, et al. Mathqa: Towards interpretable math word problem solving with operation-based formalisms. arXiv preprint arXiv:1905.13319, 2019.
- [10] ANTOL, Stanislaw, et al. Vqa: Visual question answering. In: Proceedings of the IEEE international conference on computer vision. 2015. p. 2425-2433.
- [11] ZHANG, Renrui, et al. Mathverse: Does your multi-modal llm truly see the diagrams in visual math problems?. arXiv preprint arXiv:2403.14624, 2024.
- [12] CHEN, Jiaqi, et al. GeoQA: A geometric question answering benchmark towards multimodal numerical reasoning. arXiv preprint arXiv:2105.14517, 2021.
- [13] OpenAI. Gpt-4v (ision) system card. Citekey: gptvision, 2023.
- [14] KAFLE, Kushal, et al. Dvqa: Understanding data visualizations via question answering. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2018. p. 5648-5656.
- [15] WANG, Ke, et al. Measuring multimodal mathematical reasoning with math-vision dataset. arXiv preprint arXiv:2402.14804, 2024.
- [16] QIAO, Runqi, et al. We-Math: Does Your Large Multimodal Model Achieve Human-like Mathematical Reasoning?. arXiv preprint arXiv:2407.01284, 2024.
- [17] REID, Machel, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. arXiv preprint arXiv:2403.05530, 2024.
- [18] Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui
- [19] BAI, Jinze, et al. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. 2023.
- [20] LIN, Ziyi, et al. Sphinx: The joint mixing of weights, tasks, and visual embeddings for multi-modal large language models. arXiv preprint arXiv:2311.07575, 2023.