PolyMATH: A Challenging Multi-modal Mathematical Reasoning Benchmark

Himanshu Gupta[1*] **Shreyas Verma**[2*] **Ujjwala Anantheswaran**[1*] **Kevin Scaria**[1*]
Mihir Parmar[1] **Swaroop Mishra**[1] **Chitta Baral**[1]
[1]Arizona State University [2]Georgia Institute of Technology
{hgupta35, kscaria}asu.edu

Abstract

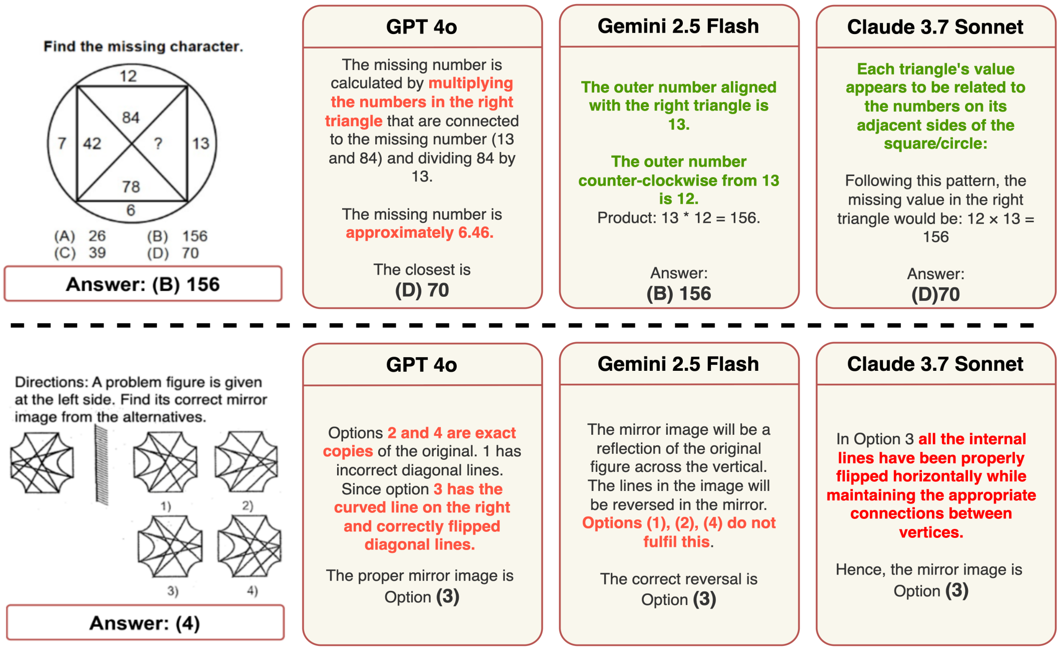
Multi-modal Large Language Models (MLLMs) exhibit impressive problem-solving abilities in various domains, but their visual comprehension and abstract reasoning skills remain under-evaluated. To this end, we present PolyMATH, a challenging benchmark aimed at evaluating the general cognitive reasoning abilities of MLLMs. PolyMATH comprises 5,000 manually collected high-quality images of cognitive textual and visual challenges across 10 distinct categories, including pattern recognition, spatial reasoning, and relative reasoning. We conducted a comprehensive, and quantitative evaluation of 12 MLLMs using four diverse prompting strategies, including Chain-of-Thought and Step-Back. The best scores achieved on PolyMATH are $\sim$ 54%, $\sim$ 36%, and $\sim$ 57%, obtained by Claude-3.7 Sonnet, GPT-4o and Gemini-2.5 Flash respectively - highlighting the logical and visual complexity of these questions. A further fine-grained error analysis reveals that these models struggle to understand spatial relations and perform drawn-out, high-level reasoning. This is further strengthened by our ablation study estimating MLLM performance when given textual descriptions in place of diagrams. As evidenced by $\sim$ 4% improvement over textual descriptions as opposed to actual images, we discover that models do not truly comprehend visual diagrams and the spatial information therein, and are thus prone to logical errors. The results on PolyMATH highlight the room for improvement in multi-modal reasoning and provide unique insights to guide the development of future MLLMs [1].

1 Introduction

Large Language Models (LLMs) [1–4] and Multi-modal Large Language Models (MLLMs) [5–8] have rapidly become a pivotal area of research. MLLMs with robust reasoning capabilities in visual contexts can solve complex educational problems [9, 10], support analysts with logical queries on statistical data [11, 12], and contribute to advanced research areas such as theorem proving and scientific discovery [13–15]. Despite their impressive performance in various assessments of human-like intelligence, these models still exhibit notable shortcomings on tasks requiring cognitive and logical reasoning, such as commonsense numerical reasoning, scientific problem-solving, and abstract puzzles [16, 17]. Existing evaluation benchmarks [18–22] have focused primarily on specific concrete domains. While general-purpose visual question-answering (VQA) datasets capture some elements of mathematical reasoning, a systematic investigation into abstract and general cognitive reasoning which are essential for tasks like visual puzzles remains an underexplored frontier.

[1] Codebase: https://github.com/kevinscaria/PolyMATH
Dataset: https://huggingface.co/datasets/him1411/polymath
*Equal Contribution

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: Foundations of Reasoning in Language Models.



Find the missing character.

(A) 26 (B) 156
(C) 39 (D) 70

Answer: (B) 156

GPT 4o

The missing number is calculated by multiplying the numbers in the right triangle that are connected to the missing number (13 and 84) and dividing 84 by 13.

The missing number is approximately 6.46.

The closest is **(D) 70**

Gemini 2.5 Flash

The outer number aligned with the right triangle is 13.

The outer number counter-clockwise from 13 is 12.
Product: 13 * 12 = 156.

Answer:
(B) 156

Claude 3.7 Sonnet

Each triangle's value appears to be related to the numbers on its adjacent sides of the square/circle:

Following this pattern, the missing value in the right triangle would be: 12 × 13 = 156

Answer:
(D)70

Directions: A problem figure is given at the left side. Find its correct mirror image from the alternatives.

Answer: (4)

GPT 4o

Options 2 and 4 are exact copies of the original. 1 has incorrect diagonal lines. Since option 3 has the curved line on the right and correctly flipped diagonal lines.

The proper mirror image is Option **(3)**

Gemini 2.5 Flash

The mirror image will be a reflection of the original figure across the vertical. The lines in the image will be reversed in the mirror. Options (1), (2), (4) do not fulfil this.

The correct reversal is Option **(3)**

Claude 3.7 Sonnet

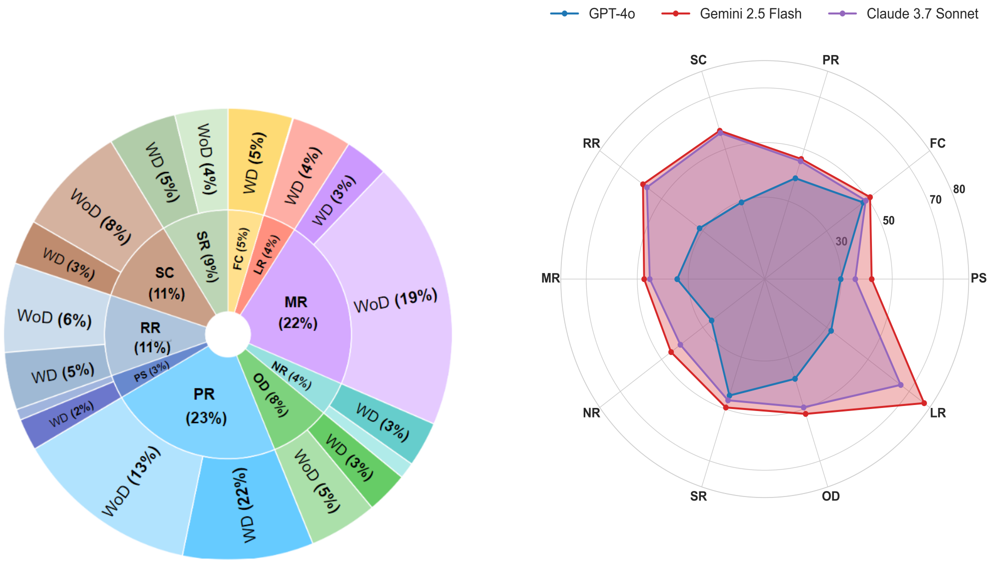
In Option 3 all the internal lines have been properly flipped horizontally while maintaining the appropriate connections between vertices.

Hence, the mirror image is Option **(3)**

Figure 1: Examples of the reasoning patterns employed by MLLMs when faced with questions involving visual information. In the top row, models fail to perceive the relationship between adjacent semicircles; in the bottom row, models fail to comprehend fine details in the answer images.

In this paper, we present POLYMATH, a benchmark specifically crafted to evaluate the complex multi-modal cognitive reasoning capabilities of MLLMs. We propose a task taxonomy to guide the development of POLYMATH: (1) we identify ten distinct reasoning skills, including *spatial reasoning*, *pattern recognition*, and *numerical reasoning*. and (2) we cover a diverse array of visual contexts, including images with venn diagrams, spatially-related layouts, as well as geometric figures. POLYMATH is a meticulously curated dataset of 5000 multimodal reasoning problems newly acquired from a publicly available source (Table 1). The problems of the original source have been crafted and rigorously reviewed by expert annotators, and require diverse fine-grained problem-solving capabilities. Additionally, we provide detailed textual representations of diagrams of the samples. As denoted in fig. 1, these problems are designed to assess the logical reasoning abilities of the average high school student over text and diagrams. We observe that MLLMs fail to demonstrate the cognitive reasoning skills required to solve these questions.

We conduct extensive experiments on POLYMATH with state-of-the-art (SOTA) closed-source MLLMs like the Claude 3.7 Sonnet, Gemini-2.5 Flash, and GPT-4o, and 9 open-source MLLMs like LLaVA (34B) and ShareGPT4V. We evaluate them via zero shot, few shot, Chain-of-Thought [23] and step back prompting [24]. We show that POLYMATH is a challenging benchmark, with human performance (established by qualified human annotators with graduate degrees) reaching only 66.3% accuracy. The most powerful model we evaluate, Gemini-2.5 Flash, achieves the best score of 57.00% followed by Claude 3.7 Sonnet, which attains 53.90%. The best open source models like LLaVA-v1.6 Mistral (7B) and ShareGPT4V (13B) achieves the accuracy of 15.20% and 12.80% respectively. We additionally create a diagram only subset (*test-img*) of the benchmark to gauge the gap in visual reasoning abilities between the multi-modal models and average human capability. We find that the performance of these models drops further to 26.20% for Claude-3.7 Sonnet and 32.50% by Gemini-2.5 Flash when evaluated on *test-img* only. In contrast with human cognitive patterns, when given text descriptions in place of the diagram in these questions, model accuracy improves by $\sim$4-7%. We also conduct an error analysis on Claude-3.7 Sonnet, Gemini-2.5 Flash and GPT-4o, and find that the most common errors stem from misunderstanding diagrams ($\sim 60\%$), misidentifying logical patterns ($\sim 25\%$), and forgetting relational information ($\sim 12\%$).



(a) Dataset categorization

(b) Results on closed source models

Figure 2: An overview of PolyMath's distribution and difficulty (a) exhibits the per-category split of the 5000 questions in the dataset, along with the split of *with diagram* (WD) and *without diagram* (WoD) for that category ; (b) Compares the per-category performance of various MLLMs.

2 Related Work

The development of MLLMs builds on the progress of LLMs [3, 25, 26, 2] and large vision models [27–30]. These models extend LLMs to handle a wider range of tasks across multiple modalities, including 2D images [31–34], 3D point clouds [35, 36], audio [37, 7], and video [38, 39]. Notable examples like OpenAI's GPT-4V [5] and Google's Gemini [6] demonstrate advanced visual reasoning capabilities, setting new benchmarks in the multimodal space.

As MLLMs rapidly advance [40], there is a growing need for benchmarks that evaluate mathematical problem-solving in visual contexts. Existing benchmarks, such as GeoQA [41], VQA [42], and UniGeo [43], focus mostly on geometric problems. Other efforts target skills in abstract scenes, geometry diagrams, charts, and synthetic images [43, 44]. Recent datasets also assess external knowledge, commonsense reasoning, and scientific or medical understanding [45]. MathVista [17] expands multimodal math tasks, while MMMU [46] focuses on college-level problems. Prior work evaluates LLMs across diverse domains like QA, mathematics, and science [47, 48], while recent research [49] explores whether models like GPT-4V perform vision and language tasks independently or together.

Existing extensive benchmarks [18–20, 50] primarily focus on concrete, real-world problems within specific domains. These benchmarks often include comparatively simple diagram interpretation questions involving plots or mathematical questions related to geometry, which primarily evaluate models' abilities to parse information from a single image and solve problems using well-established logical principles and formulae. However, they do not sufficiently test models' capabilities in abstract visual reasoning, including spatial recognition, visual logic and puzzle solving, and pattern recognition. This limitation represents a notable gap, as visual puzzle tasks require logical leaps that differ fundamentally from reasoning patterns over textual or linguistic problems. Moreover, spatial reasoning questions assess models' abilities to internalize and manipulate configurations in 3D space, as well as reason over spatial information and infer implicit relationships based on positional data. This category of questions aligns closely with human cognition and reasoning abilities, and evaluating model performance against human baselines on these questions reveals the substantial gap in reasoning abilities that models must bridge to approach human-comparable reasoning capability. Our proposed dataset aims to address this gap by challenging and comprehensively evaluating previously underexplored model skills in categories where their performance still lags significantly behind human reasoning baselines. Additionally, we provide a detailed analysis of the strengths

Category	PS	FC	PR	SC	RR	MR	NR	SR	OD	LR	Overall
Full dataset											
Questions with Diag.	114	233	472	160	206	157	162	246	151	3	1904
Questions w/o Diag.	39	0	664	398	319	964	58	191	246	217	3096
Total Questions	153	233	1136	558	525	1121	220	437	397	220	5000
testmini											
Questions with Diag.	27	47	102	33	47	28	30	53	38	0	405
Questions w/o Diag.	4	0	125	79	58	196	14	34	41	44	595
Total Questions	31	47	227	112	105	224	44	87	79	44	1000
test-img											
Total Questions	60	122	248	84	108	82	85	129	79	3	1000

Table 1: An overview of the per-category distribution of questions in the *test*, *testmini*, and *test-img* splits of POLYMATH. *testmini* and *test-img* are 1000-instance subsets, aimed at faster and image-focused evaluations respectively. We also report the frequency of *with diagram* and *without diagram* questions for each category.

and weaknesses of these models across a wide range of categories and skills, shedding light on specific reasoning errors and their frequency of occurrence across categories and in comparison to one another.

3 Curating POLYMATH

POLYMATH is curated mainly from questions directed at students taking the National Talent Search Examination, a nationwide competitive exam held by the National Council of Educational Research and Training of India. These questions and their solutions are created by experts in their fields and rigorously peer-reviewed, and thus contain minimal errors. These questions aim to assess Scholastic Aptitude (SAT), or the ability to recall domain-specific scientific and mathematical knowledge, as well as Mental Ability (MAT), or the ability to think logically and apply a range of analytical skills. We catalog the skills assessed by each sample along the categorization schema defined in Table 2.

3.1 Collection Pipeline

To guarantee high-quality data, we manually collected image snippets and engineered a streamlined, automated framework for curation and annotation. Continuous human reviews were conducted throughout the process, ensuring quality and preventing error propagation.

- **Step 1**: We generate a universally unique identifier (UUID) for a given question paper to identify all the questions curated from it.
- **Step 2**: Annotators manually collected separate snippets of each question and their associated contextual information (including disconnected pieces) that apply to multiple questions.
- **Step 3**: An image merging script automatically identified and merged question images (in case the question gets split by pages) with their relevant context images.
- **Step 4**: We used an LLM to transcribe the questions and their ground truth answers. We also generate additional metadata, including category (§3.2), whether it contains a diagram , and image description (§3.3). A manual check was performed to ensure the quality of the generated metadata.
- **Step 5**: An annotation file, where each row corresponds to a question, is automatically created and populated.

3.2 Dataset categorization

We develop a categorization schema that catalogues questions on basis of the information provided and the type of reasoning assessed by the question. Based on the continuous human evaluation during

Category name	**Definition**	**Avg len**	**Max len**
Perspective Shift (PS)	A figure is given and the solver is instructed to morph it according to the instructions (flip, mirror image, rotate, etc.)	18.60	59
Figure Completion (FC)	A figure is given with an arrangement of numbers or characters such that their relationship to one another based on their position in the figure is consistent. The goal is to complete the figure and identify the element missing from a marked position.	23.97	364
Pattern Recognition (PR)	This requires the understanding of a one-to-one relationship or pattern and replicating that pattern. For example, given the relationship between a and b, determining the equivalent of b to c. Questions involving substituting characters and operations in a pre-defined pattern fall into this category.	31.98	391.4
Sequence Completion (SC)	Given a sequence of numbers or figures, this question involves finding the sequentially next element in a series.	30.22	227
Relative Reasoning (RR)	The question contains distinct data points and their relationship with one another. The solver must extrapolate relationships that may not be explicitly mentioned to answer the questions. Questions involving Venn diagrams, family relations, or relative positions given a reference point fall into this category.	27.22	137
Mathematical Reasoning (MR)	This question entails calculations of a mathematical nature, such as solving a given equation.	25.61	156
Numerical Reasoning (NR)	Questions involving counting the number of elements mentioned. The elements may be part of a single figure or conform to a specified pattern.	15.63	65
Spatial Reasoning	These questions require the solver to visualize the context and reason observationally to arrive at the answer.	27.67	78
Odd One Out (OD)	Given a set of elements, identify the element that is not like the others.	26.64	214
Logical Reasoning (LR)	Questions involving simple logical reasoning such as entailment and contradiction.	34.68	144
Overall		**27.68**	**391.4**

Table 2: An overview of our question categorization schema. Questions are categorized on the basis of the information provided in the question and the reasoning skills assessed.

collection, we identify 10 distinct question categories. We enumerate these categories along with their definitions in Table 2. We further distinguish between questions *with diagram* and *without diagram*. The overall per-category distribution, along with the *with diagram* and *without diagram* split, is visualized in Figure 2.

3.3 Additional metadata

The complexity of collected question images and the heavy presence of diagram-based reasoning tasks makes POLYMATH a challenging multi-modal benchmark. To make POLYMATH usable for both text and vision model evaluations, we provide transcriptions of questions and answers. To further facilitate text-based evaluation, we generate detailed, human-vetted text descriptions of attached diagrams such that a human could visualize the image based on this description Results on text-only characterization of questions in our dataset can be found in §4.3.

3.4 Quality Assurance

Following the collection and annotation process, we conduct a comprehensive quality check. We discard samples that are [1] of low resolution, [2] outside the scope of the categories (Table 2), or [3] missing vital information. We also discard samples with noticeable watermarks and other visual noise that renders the sample illegible. Our subject-expert annotators rectify incorrectly-extracted ground truth answers. Concurrently, we verify that the questions belong to their assigned categories, and correct any observed misalignments therein.

Category	PS	FC	PR	SC	RR	MR	NR	SR	OD	LR	Overall
Baseline											
Random chance	9.68	4.26	6.61	9.82	9.52	9.82	15.91	6.90	7.59	9.09	8.60
Human eval	51.08	70.57	61.82	69.35	69.84	76.64	58.71	62.64	64.98	51.14	66.62
Zero Shot Inference											
GPT-4o	29.79	47.73	38.84	29.55	31.65	34.36	25.81	44.83	38.39	32.18	36.60
Gemini-2.5 Flash	41.94	51.06	46.26	57.14	59.05	47.32	45.45	49.43	51.90	77.27	51.20
Claude-3.7 Sonnet	35.48	48.94	45.37	56.25	57.14	45.09	40.91	46.67	49.37	65.91	48.60
Few Shot Inference											
GPT-4o	29.03	14.89	33.48	38.39	40.00	40.18	18.18	36.78	21.52	50.00	34.60
Gemini-2.5 Flash	48.39	59.57	47.58	60.71	61.90	49.11	52.27	51.72	54.43	84.09	54.20
Claude-3.7 Sonnet	41.94	53.19	46.26	58.93	59.05	46.88	47.73	49.43	51.90	75.00	51.40
Chain-of-Thought Prompting Inference											
GPT-4o	21.28	54.55	41.96	25.00	27.85	29.96	9.68	40.95	41.07	33.33	35.00
Gemini-2.5 Flash	51.61	65.96	48.02	64.29	64.76	49.55	59.09	57.47	58.23	93.18	57.00
Claude-3.7 Sonnet	54.84	55.32	46.70	61.61	63.81	47.77	50.00	55.17	54.43	77.27	53.90
Step Back Prompting Inference											
GPT-4o	12.77	45.45	42.41	27.27	31.65	34.80	16.13	41.90	41.07	37.93	36.50
Gemini-2.5 Flash	48.39	59.57	47.58	62.50	64.76	48.21	54.55	55.17	58.23	88.64	55.40
Claude-3.7 Sonnet	48.39	48.94	45.37	60.71	63.81	47.77	45.45	55.17	53.16	75.00	51.90

Table 3: Results of closed-source LLMs on the *testmini* split of POLYMATH. We report model results using the following prompting strategies: zero-shot inference, few-shot inference, Chain-of-Thought, and Step Back prompting. For each prompting setting, the highest and lowest scores achieved by a model per category are highlighted. In addition to model accuracy, we report a Random chance baseline (i.e. the accuracy of a model that randomly selects an option without visibility into the question, and a Human eval baseline, where we report the average scores of six human evaluators.)

3.5 Division of the *testmini* Subset.

The final iteration of POLYMATH comprises 5000 questions. To enable faster model validation, we extract a 1000-instance subset, *testmini*, using stratified sampling over all categories. All quantitative results reported were obtained on this *testmini* subset of POLYMATH. We also create a *test-img* question set, consisting solely of 1000 *with diagram* questions, aimed at faster, focused assessment of models' visual comprehension. We use a random sampling strategy to create *test-img* due to diagram imbalance. [2] For data distribution, see Table 1.

4 Experiments

We conduct a systematic evaluation of existing MLLMs on POLYMATH. We first introduce the experimental setup in this section. Then we present our findings followed by multiple dataset analysis experiments.

4.1 Experimental Setup

Evaluation Models: We examine the performance of foundation models across two distinct categories on POLYMATH: (a) **Closed-source MLLMs**, represented by models like GPT-4o (`gpt-4o-2024-05-13`) Gemini-2.5 Flash (`Gemini-2.5-flash-002`) [6], Claude-3.7 Sonnet (`claude-3-7-sonnet`) (b) **Open-source MLLMs**, such as LLaVA (v1.5-13B, v1.6-Mistral-7B, v1.6-Vicuna-13B) [51], LLaVA-v1.6-34B [52], G-LLaVA (7B, 13B) [53], ShareGPT4V (7B, 13B) [54] & Qwen2-VL-2B-Instruct [55] (c) **Text Based LLMs** Reka Flash [56], Llama-3 (70B) [57], Mistral Large [58]. We conduct experiments on open-source models using six NVIDIA A100 GPUs.

Implementation Details All reported results are on the $testmini$ subset. As a comparative baseline, we simulate random chance by selecting a random option for multiple-choice questions over 1000 trials. Additionally, the problems in POLYMATH were independently solved by the paper's authors

[2] All datasets (*test*, *testmini* and *test-img*) will be publicly released

Model	PS	FC	PR	SC	RR	MR	NR	SR	OD	LR	Overall
Qwen2 VL (2B) Instruct	9.38	2.13	6.17	6.25	8.57	3.57	4.55	4.60	8.86	2.27	5.60
LLaVA-v1.6 Mistral (7B)	6.45	4.26	14.98	14.29	18.10	15.18	9.09	19.54	22.78	13.64	15.20
G-LLaVA (7B)	12.90	0.00	9.25	3.57	5.71	7.59	2.27	4.60	3.80	6.82	6.30
ShareGPT4V (7B)	6.45	10.64	16.30	13.39	7.62	11.61	11.36	11.49	10.13	11.36	12.10
LLaVA-v1.6 Vicuna (13B)	12.90	12.77	8.37	8.04	13.33	5.80	15.91	6.90	13.92	4.55	9.10
LLaVA 1.5 (13B)	3.23	14.89	7.49	11.61	7.62	6.70	9.09	8.05	11.39	13.64	8.70
ShareGPT4V (13B)	9.68	17.02	13.66	12.50	15.24	10.71	9.09	12.64	17.72	6.82	12.80
G-LLaVA (13B)	13.67	2.33	11.12	5.69	7.98	10.23	1.07	6.70	5.76	7.98	8.26
LLaVA-v1.6 (34B)	9.68	25.33	9.69	12.50	6.67	10.71	13.64	10.34	15.19	9.09	11.30

Table 4: Results of open-source MLLMs on the *testmini* split of POLYMATH. We report model results using zero shot inference. The highest and lowest scores achieved by a model in each category are highlighted.

(four engineering graduates and two PhDs), serving as a human performance baseline. We evaluate the benchmark using various prompting methods, including zero shot, few shot (2-shot), Chain-of-Thought [23], and Step Back prompting [24]. For multiple-choice questions, we use exact match for answer comparison. The model inference prompts are structured to elicit a step-by-step solution, the final answer, and the corresponding option. As part of our analysis, we conducted three additional experiments: (1) analyzing model performance on the *test-img* split, (2) converting the questions from *test-img* into text, along with the transformation of diagrams into descriptions, and (3) evaluating OpenAI o1 models on questions without diagrams.

4.2 Results

Closed Source Models Across various prompting strategies (Table 3), Gemini-2.5 Flash performed best with these advanced prompts, achieving up to 57.00% accuracy in Step Back Prompting, compared to 54.20% in few shot. Claude-3.7 Sonnet followed closely, especially in FC and PS questions, showing strong performance with zero shot and Step Back Prompting. GPT-4o Flash performed moderately across all categories but lacked dominance in any specific area. In terms of prompting strategies, Chain-of-Thought and Step Back Prompting enhanced the performance of top models like Claude-3.7 Sonnet and Gemini-2.5 Flash, allowing them to excel in tasks requiring structured reasoning and re-evaluation. Both strategies led to marked improvements over zero shot prompting, in categories like SR, PR, and LR.

Open Source Models Table 4 showcases the results of open-source MLLMs. LLaVA-v1.6-Mistral-7B model achieved the highest overall score of 15.2%. It excelled in OD (22.78%), SR (19.54%), RR (18.1%), and MR (15.18%) indicating its proficiency in generating precise, coherent, and relevant responses, even for out-of-distribution samples. The ShareGPT4V (13B) model exhibited the second-highest overall score of 12.8%, with outstanding performance in the PR (13.66%), SC (12.5%), RR (15.24%), MR (10.71%), SR (12.64%), and OD (17.72%) categories. Other models, such as LlaVA-v1.6-Vicuna 13B, LlaVA-1.5 (13B), G-LLaVA (13B), and LlaVA-v1.6 (34B), exhibited varying levels of success across the different categories, highlighting their individual strengths and weaknesses in handling the diverse reasoning aspects tested by the dataset.

Human Evaluation To ascertain the difficulty of the dataset, we asked six graduate students specifically for the evaluation of human performance on POLYMATH. We assigned questions from a specific problem category to each student. They were asked to provide only the final answer without detailed reasoning, simulating zero-shot inference.

4.3 Experimental Analysis

MLLMs Rely More on Image Descriptions than Image To evaluate the visual reasoning capabilities, we used *test-img* subset, which contains questions with diagrams. Additionally, we generated a text-only version of *test-img* by replacing all diagrams with detailed textual descriptions. Both experiments were carried out in a zero shot setting. Our analysis reveals three key findings. First, we observed a noticeable decline in performance on *test-img*, particularly for models like GPT-4o and Claude-3.7 Sonnet, compared to their results on the *testmini* subset. This suggests that both models perform well on questions without diagrams, and their decreased accuracy on *test-img* is largely due

Category	PS	FC	PR	SC	RR	MR	NR	SR	OD	LR	Overall
MLLM Inference on Diagrams (Multi-modal)											
GPT-4o	20.00	20.49	22.18	19.05	23.15	20.73	20.00	17.05	34.18	66.67	21.80
Gemini-2.5 Flash	26.67	34.43	27.42	36.90	39.81	35.37	22.35	28.68	41.77	100.00	32.10
Claude-3.7 Sonnet	43.34	33.61	27.42	29.76	37.03	31.71	34.12	26.36	37.98	100.00	26.20
MLLM Inference on Diagram Descriptions (Text-only)											
GPT-4o	26.67	28.69	29.44	23.81	31.48	34.15	30.59	29.46	27.85	33.33	29.30
Gemini-2.5 Flash	38.33	34.43	27.82	35.71	28.71	41.46	25.88	26.36	34.17	100.00	31.50
Claude-3.7 Sonnet	43.34	33.61	29.43	29.76	40.74	42.69	42.36	32.56	45.57	100.00	33.50
LLM Inference on Diagram Descriptions (Text-only)											
Mistral Large	15.00	13.11	11.29	15.48	18.52	13.41	9.41	17.83	25.32	33.33	14.90
Reka Flash	16.67	13.93	12.10	16.67	19.44	14.63	9.41	18.60	26.58	33.33	15.80
Llama-3 (70B)	16.67	13.93	11.69	16.67	19.44	14.63	10.59	18.60	26.58	33.33	15.80

Table 5: Visual comprehension ablation results on *test-img*. We compare [1] multi-modal inference with diagrams and [2] unimodal inference using text descriptions. Highest and lowest scores per category are highlighted. Unimodal LLM performance on text-only questions is also reported.

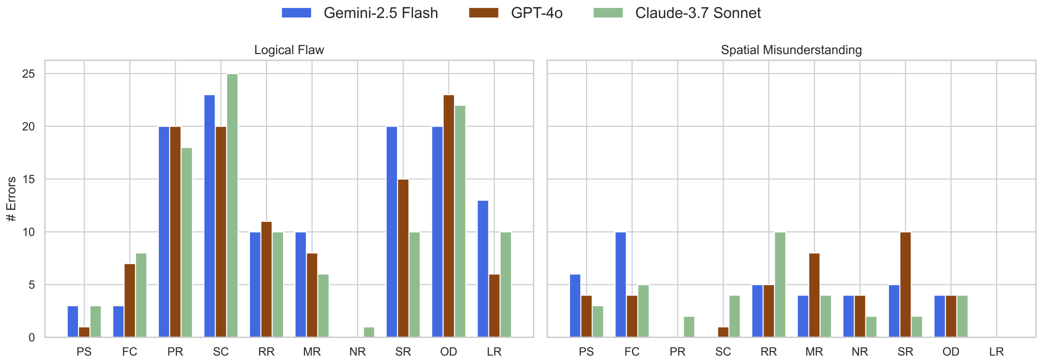


Figure 3: Frequency of LF and SM errors across different question categories. We report per-model figures to enable a comparison of model abilities. They are most prevalent in the OD, PR, and SC categories of questions, owing to the amount of logical leaps and visual reasoning required by these questions.

to the presence of diagram-based problems. Second, when we replaced the diagrams in *test-img* with text descriptions, the performance of all models improved by $\sim 6\%$, indicating that the models struggle with diagrams and benefit from textual representations. Finally, we evaluated popular text-only LLMs such as LLaMA-3 (70B), Reka Flash, and Mistral Large on the text-description version of *test-img*. Their scores ($\sim 15\%$) were lower than those of the MLLMs ($\sim 27\%$), underscoring the advantage of multi-modal models in handling visually-grounded tasks.

A Closer Look at Model Errors We analysed 203 samples where all three state-of-the-art MLLMs (Claude-3.7 Sonnet, GPT-4o and Gemini-2.5 Flash) gave incorrect answers on *testmini*. Based on the manual inspection of the responses, we identified 7 types of errors that MLLMs make. The most common error on this dataset was Logical Flaw (LF), occurring in nearly $\sim 60\%$ of incorrect samples. Spatial Misunderstanding (SM), which involves a lack of understanding of diagram structure and content, was a close second ($\sim 20\%$). Figure 3 shows the category-wise distribution of the two types of error. These errors were most prevalent in OD, PR, and SC category of questions, as making uncommon logical leaps and fully comprehending visuals is integral to solving these. Furthermore, in questions involving extrapolation over multiple weakly connected data points, models came to conclusions that contradicted earlier data, indicating a lack of information retention. Finally, we found that models fell into identical fallacious reasoning patterns, e.g. assuming that a pattern holds across each row when a pattern is replicated across columns. The category with the highest % of shared errors was PR, where we observed that GPT, Gemini, and Claude followed the same incorrect reasoning structure on nearly 80% of the analysed samples. Thus, despite their differences, in practice we see that MLLMs share the same strengths and shortcomings.

5 Conclusion

In this work, we introduce POLYMATH, a benchmark designed to systematically analyze the mathematical reasoning capabilities of state-of-the-art models in visually complex scenarios. Our evaluation of 14 prominent foundation models highlights that significant advancements have been made, especially with the GPT-4o and Claude-3.7 Sonnet models. However, a substantial gap of $\sim 10\%$ still exists between Gemini-2.5 Flash, the best-performing model, and human performance. This disparity sets a clear direction for future research, emphasizing the need for models that can seamlessly integrate mathematical reasoning with visual comprehension. Moreover, our analysis of model reasoning errors and experiments on samples containing diagrams and their textual representations offer valuable insights for future investigations.

References

[1] 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.

[2] Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mixtral of experts. *Arxiv 2401.04088*, 2024.

[3] 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.

[4] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[5] OpenAI. GPT-4V(ision) system card, 2023.

[6] Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

[7] Yixuan Su, Tian Lan, Huayang Li, Jialu Xu, Yan Wang, and Deng Cai. Pandagpt: One model to instruction-follow them all. *arXiv preprint arXiv:2305.16355*, 2023.

[8] Jun Chen, Deyao Zhu1 Xiaoqian Shen1 Xiang Li, Zechun Liu2 Pengchuan Zhang, Raghuraman Krishnamoorthi2 Vikas Chandra2 Yunyang Xiong, and Mohamed Elhoseiny. Minigpt-v2: Large language model as a unified interface for vision-language multi-task learning. *arXiv preprint arXiv:2310.09478*, 2023.

[9] 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*, pages 1466–1476, 2015.

[10] 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)*, pages 845–854, 2017.

[11] 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.

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

[13] 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.

[14] 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.

[15] Trieu H Trinh, Yuhuai Wu, Quoc V Le, He He, and Thang Luong. Solving olympiad geometry without human demonstrations. *Nature*, 625(7995):476–482, 2024.

[16] 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.

[17] Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chun yue Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. Mathvista: Evaluating math reasoning in visual contexts with gpt-4v, bard, and other large multimodal models. *ArXiv*, abs/2310.02255, 2023.

[18] Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, Yunsheng Wu, and Rongrong Ji. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*, 2023.

[19] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. MMBench: Is your multi-modal model an all-around player? *arXiv preprint arXiv:2307.06281*, 2023.

[20] Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. Seed-bench: Benchmarking multimodal llms with generative comprehension. *ArXiv*, abs/2307.16125, 2023.

[21] Chaoyou Fu, Renrui Zhang, Haojia Lin, Zihan Wang, Timin Gao, Yongdong Luo, Yubo Huang, Zhengye Zhang, Longtian Qiu, Gaoxiang Ye, et al. A challenger to gpt-4v? early explorations of gemini in visual expertise. *arXiv preprint arXiv:2312.12436*, 2023.

[22] Keqiang Sun, Junting Pan, Yuying Ge, Hao Li, Haodong Duan, Xiaoshi Wu, Renrui Zhang, Aojun Zhou, Zipeng Qin, Yi Wang, et al. Journeydb: A benchmark for generative image understanding. *Advances in Neural Information Processing Systems*, 36, 2024.

[23] 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*, 2022.

[24] Huaixiu Steven Zheng, Swaroop Mishra, Xinyun Chen, Heng-Tze Cheng, Ed H Chi, Quoc V Le, and Denny Zhou. Take a step back: Evoking reasoning via abstraction in large language models. In *The Twelfth International Conference on Learning Representations*, 2024.

[25] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

[26] OpenAI. Chatgpt. https://chat.openai.com, 2023.

[27] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023.

[28] Renrui Zhang, Zhengkai Jiang, Ziyu Guo, Shilin Yan, Junting Pan, Hao Dong, Peng Gao, and Hongsheng Li. Personalize segment anything model with one shot. *ICLR 2024*, 2023.

[29] Renrui Zhang, Xiangfei Hu, Bohao Li, Siyuan Huang, Hanqiu Deng, Hongsheng Li, Yu Qiao, and Peng Gao. Prompt, generate, then cache: Cascade of foundation models makes strong few-shot learners. *CVPR 2023*, 2023.

[30] Renrui Zhang, Liuhui Wang, Yu Qiao, Peng Gao, and Hongsheng Li. Learning 3d representations from 2d pre-trained models via image-to-point masked autoencoders. *CVPR 2023*, 2023.

[31] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pages 12888–12900. PMLR, 2022.

[32] 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.

[33] 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.

[34] Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Fanyi Pu, Jingkang Yang, Chunyuan Li, and Ziwei Liu. Mimic-it: Multi-modal in-context instruction tuning. *arXiv preprint arXiv:2306.05425*, 2023.

[35] Ziyu Guo, Renrui Zhang, Xiangyang Zhu, Yiwen Tang, Xianzheng Ma, Jiaming Han, Kexin Chen, Peng Gao, Xianzhi Li, Hongsheng Li, et al. Point-bind & point-llm: Aligning point cloud with multi-modality for 3d understanding, generation, and instruction following. *arXiv preprint arXiv:2309.00615*, 2023.

[36] Runsen Xu, Xiaolong Wang, Tai Wang, Yilun Chen, Jiangmiao Pang, and Dahua Lin. Pointllm: Empowering large language models to understand point clouds. *arXiv preprint arXiv:2308.16911*, 2023.

[37] Jiaming Han, Renrui Zhang, Wenqi Shao, Peng Gao, Peng Xu, Han Xiao, Kaipeng Zhang, Chris Liu, Song Wen, Ziyu Guo, et al. Imagebind-llm: Multi-modality instruction tuning. *arXiv preprint arXiv:2309.03905*, 2023.

[38] Hang Zhang, Xin Li, and Lidong Bing. Video-llama: An instruction-tuned audio-visual language model for video understanding. *arXiv preprint arXiv:2306.02858*, 2023.

[39] Guo Chen, Yin-Dong Zheng, Jiahao Wang, Jilan Xu, Yifei Huang, Junting Pan, Yi Wang, Yali Wang, Yu Qiao, Tong Lu, et al. Videollm: Modeling video sequence with large language models. *arXiv preprint arXiv:2305.13292*, 2023.

[40] Chunyuan Li, Zhe Gan, Zhengyuan Yang, Jianwei Yang, Linjie Li, Lijuan Wang, and Jianfeng Gao. Multimodal foundation models: From specialists to general-purpose assistants. *arXiv preprint arXiv:2309.10020*, 2023.

[41] Jiaqi Chen, Jianheng Tang, Jinghui Qin, Xiaodan Liang, Lingbo Liu, Eric P. Xing, and Liang Lin. Geoqa: A geometric question answering benchmark towards multimodal numerical reasoning. *ArXiv*, abs/2105.14517, 2021.

[42] 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*, pages 6904–6913, 2017.

[43] 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*, pages 3313–3323, 2022.

[44] 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*, pages 2263–2279, 2022.

[45] 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*, 2023.

[46] Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. *arXiv preprint arXiv:2311.16502*, 2023.

[47] 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.

[48] 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.

[49] Xiang Zhang, Senyu Li, Zijun Wu, and Ning Shi. Lost in translation: When gpt-4v (ision) can't see eye to eye with text. a vision-language-consistency analysis of vllms and beyond. *arXiv preprint arXiv:2310.12520*, 2023.

[50] 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.

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

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

[53] Jiahui Gao, Renjie Pi, Jipeng Zhang, Jiacheng Ye, Wanjun Zhong, Yufei Wang, Lanqing Hong, Jianhua Han, Hang Xu, Zhenguo Li, et al. G-llava: Solving geometric problem with multi-modal large language model. *arXiv preprint arXiv:2312.11370*, 2023.

[54] Lin Chen, Jinsong Li, Xiao wen Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. Sharegpt4v: Improving large multi-modal models with better captions. *ArXiv*, abs/2311.12793, 2023.

[55] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.

[56] Aitor Ormazabal, Che Zheng, Cyprien de Masson d'Autume, Dani Yogatama, Deyu Fu, Donovan Ong, Eric Chen, Eugenie Lamprecht, Hai Pham, Isaac Ong, Kaloyan Aleksiev, Lei Li, Matthew Henderson, Max Bain, Mikel Artetxe, Nishant Relan, Piotr Padlewski, Qi Liu, Ren Chen, Samuel Phua, Yazheng Yang, Yi Tay, Yuqi Wang, Zhongkai Zhu, and Zhihui Xie. Reka core, flash, and edge: A series of powerful multimodal language models, 2024.

[57] AI@Meta. Llama 3 model card, 2024.

[58] Mistral AI. Au large, Apr 2024.