

POLYMATH: A CHALLENGING MULTI-MODAL MATHEMATICAL REASONING BENCHMARK

Anonymous authors

Paper under double-blind review

ABSTRACT

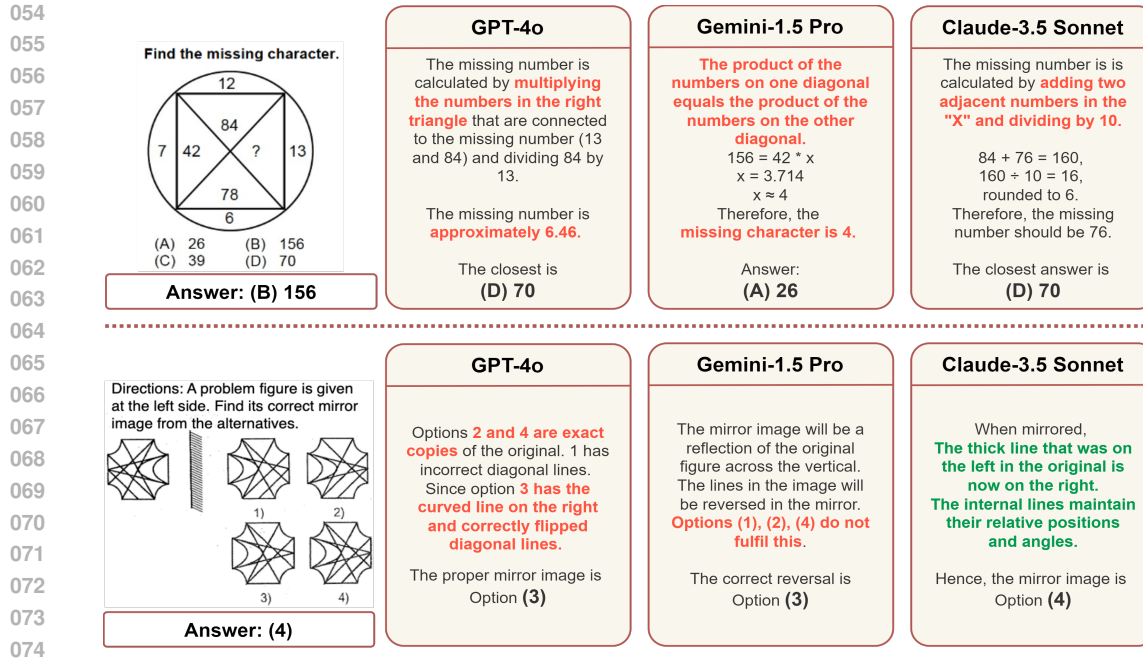
Multi-modal Large Language Models (MLLMs) exhibit impressive problem-solving abilities in various domains, but their visual comprehension and abstract reasoning skills remain under-evaluated. To this end, we present POLYMATH, a challenging benchmark aimed at evaluating the general cognitive reasoning 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 15 MLLMs using four diverse prompting strategies, including Chain-of-Thought and Step-Back. The best scores achieved on POLYMATH are $\sim 41\%$, $\sim 36\%$, and $\sim 27\%$, obtained by Claude-3.5 Sonnet, GPT-4o and Gemini-1.5 Pro 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. Finally, we evaluate the OpenAI o1 models and find that their performance only matches the human baseline, highlighting the difficulty of the benchmark. 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) (Brown et al., 2020; Jiang et al., 2024; Touvron et al., 2023a; Achiam et al., 2023) and Multi-modal Large Language Models (MLLMs) (OpenAI, 2023c; Team et al., 2023; Su et al., 2023; Chen et al., 2023b) have rapidly become a pivotal area of research. MLLMs with robust reasoning capabilities in visual contexts can solve complex educational problems (Seo et al., 2015; Wang et al., 2017), support analysts with logical queries on statistical data (Wu et al., 2023; Yang et al., 2023), and contribute to advanced research areas such as theorem proving and scientific discovery (Taylor et al., 2022; Dong et al., 2023; Trinh et al., 2024). 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 (Wang et al., 2023b; Lu et al., 2023a). Existing evaluation benchmarks (Fu et al., 2023a; Liu et al., 2023d; Li et al., 2023b; Fu et al., 2023b; Sun et al., 2024) 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.

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*

¹<https://anonymous.4open.science/r/PolyMATH-052D>



108
109
110
111
112
113
114
115
116
117
118
119
120
121
122
123
124
125
126
127
128
129
130
131
132
133
134
135
136
137
138
139
140
141
142
143
144
145
146
147
148
149
150
151
152
153
154
155
156
157
158
159
160
161

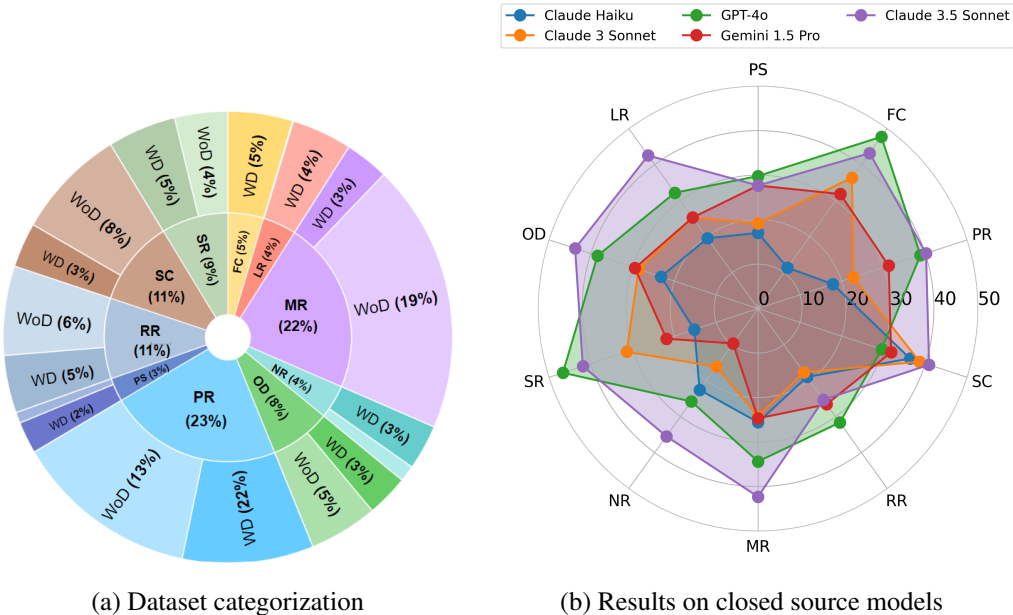


Figure 2: An overview of POLYMATH’s distribution and difficulty (a) exhibits the per-category split of the 5000 questions in the dataset, along with the split of *with diagram* (WD) and *without diagram* (WoD) for that category ; (b) Compares the per-category performance of various MLLMs.

2 RELATED WORK

The development of MLLMs builds on the progress of LLMs (Touvron et al., 2023a;b; OpenAI, 2023a; Jiang et al., 2024) and large vision models (Kirillov et al., 2023; Zhang et al., 2023d;c;e). These models extend LLMs to handle a wider range of tasks across multiple modalities, including 2D images (Li et al., 2022; Dai et al., 2023; Alayrac et al., 2022; Li et al., 2023a), 3D point clouds (Guo et al., 2023; Xu et al., 2023b), audio (Han et al., 2023; Su et al., 2023), and video (Zhang et al., 2023a; Chen et al., 2023a). Notable examples like OpenAI’s GPT-4V (OpenAI, 2023c) and Google’s Gemini (Team et al., 2023) demonstrate advanced visual reasoning capabilities, setting new benchmarks in the multimodal space.

As MLLMs rapidly advance (Li et al., 2023c), there is a growing need for benchmarks that evaluate mathematical problem-solving in visual contexts. Existing benchmarks, such as GeoQA (Chen et al., 2021a), VQA (Goyal et al., 2017), and UniGeo (Chen et al., 2022a), focus mostly on geometric problems. Other efforts target skills in abstract scenes, geometry diagrams, charts, and synthetic images (Chen et al., 2022a; Masry et al., 2022). Recent datasets also assess external knowledge, commonsense reasoning, and scientific or medical understanding (Zhang et al., 2023g). MathVista (Lu et al., 2023a) expands multimodal math tasks, while MMMU (Yue et al., 2023a) focuses on college-level problems. Prior work evaluates LLMs across diverse domains like QA, mathematics, and science (Bubeck et al., 2023; Nori et al., 2023), while recent research (Zhang et al., 2023f) explores whether models like GPT-4V perform vision and language tasks independently or together.

Existing extensive benchmarks (Fu et al., 2023a; Liu et al., 2023d; Li et al., 2023b; Xu et al., 2023a) 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

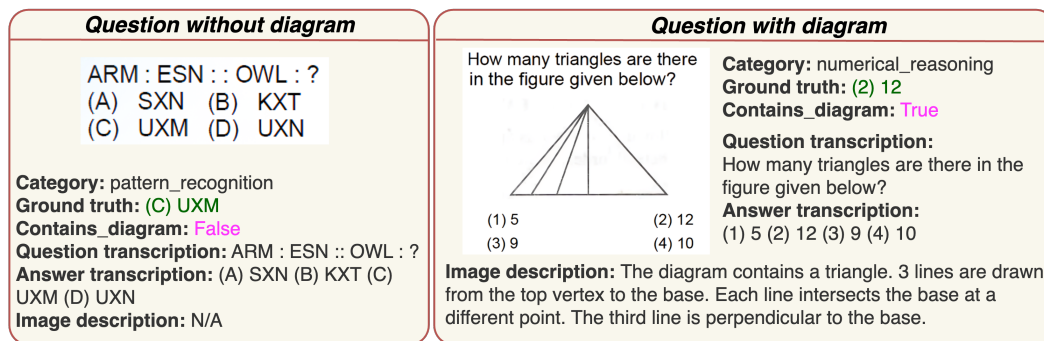


Figure 3: Examples of *with diagram* and *without diagram* questions. In addition to the question image, POLYMATH includes the metadata shown above. Question *without diagram* is not present in *test-img* while both kinds of questions will be present in *testmini*.

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 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 1.

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 (Fig 3), 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.

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 1: 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.

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 collection, we identify 10 distinct question categories. We enumerate these categories along with their definitions in Table 1. We further distinguish between questions *with diagram* and *without diagram*. The *with diagram* questions are designed around the information presented in the diagrams (Fig 3). We show examples of *with diagram* and *without diagram* questions for each category in §F. 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 (Fig 3). Results on text-only characterization of questions in our dataset can be found in §4.3.

Category	PS	FC	PR	SC	RR	MR	NR	SR	OD	LR	Overall
<i>Full dataset</i>											
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
<i>testmini</i>											
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
<i>test-img</i>											
Total Questions	60	122	248	84	108	82	85	129	79	3	1000

Table 2: 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.

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 1), 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.

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 in this work 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. Owing to the imbalance of *with diagram* questions across categories, we use a random sampling strategy to create *test-img*.² For data distribution, see Table 2. Further details on data collection and annotation are available in §C.

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. Additional experimental results and qualitative examples are present in §D and H.

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) (OpenAI, 2024a), OpenAI O1 (o1-preview-2024-09-12, o1-mini-2024-09-12) (OpenAI, 2024b), Gemini-1.5 Pro (gemini-1.5-pro-002) (Team et al., 2023), Claude-3.5 Sonnet (claude-3-5-sonnet-20240620) (Anthropic, 2024a) and Claude 3 Haiku and Sonnet (claude-3-sonnet-20240229, claude-3-haiku-20240307) (Anthropic, 2024b) (b) **Open-source MLLMs**, such as LLaVA (v1.5-13B, v1.6-Mistral-7B, v1.6-Vicuna-13B) (Liu et al., 2023a), LLaVA-v1.6-34B (Liu et al., 2024), G-LLaVA (7B, 13B) (Gao et al., 2023a), ShareGPT4V (7B, 13B) (Chen et al., 2023c) &

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

Category	PS	FC	PR	SC	RR	MR	NR	SR	OD	LR	Overall
<i>Baseline</i>											
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
<i>Zero Shot Inference</i>											
Claude Haiku	17.02	11.36	17.86	36.36	18.99	25.55	22.58	15.24	23.21	19.54	20.80
Claude-3 Sonnet	19.15	36.36	22.77	38.64	17.72	24.23	16.13	31.43	28.57	25.29	25.40
GPT-4o	29.79	47.73	38.84	29.55	31.65	34.36	25.81	46.67	38.39	32.18	36.60
Gemini-1.5 Pro	27.66	31.82	31.25	31.82	26.58	24.67	9.68	21.90	29.46	25.29	26.90
Claude-3.5 Sonnet	27.66	43.18	40.18	40.91	25.32	42.29	35.48	41.90	43.75	42.53	39.70
<i>Few Shot Inference</i>											
Claude Haiku	19.35	12.77	18.06	36.61	19.05	25.89	22.73	16.09	24.05	20.45	22.40
Claude-3 Sonnet	19.35	19.15	25.99	25.89	32.38	30.36	29.55	26.44	31.65	52.27	28.90
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-1.5 Pro	19.35	29.79	25.11	16.96	29.52	30.80	20.45	29.89	32.91	38.64	27.40
Claude-3.5 Sonnet	32.26	44.68	40.53	41.96	26.67	42.41	36.36	42.53	46.84	52.27	40.60
<i>Chain-of-Thought Prompting Inference</i>											
Claude Haiku	19.15	15.91	21.88	20.45	26.58	25.55	19.35	21.90	25.00	28.74	23.50
Claude-3 Sonnet	23.40	34.09	30.80	40.91	27.85	31.72	22.58	33.33	22.32	26.44	29.70
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-1.5 Pro	27.66	34.09	39.29	22.73	27.85	30.84	35.48	30.48	31.25	26.44	31.90
Claude-3.5 Sonnet	31.91	43.18	41.52	45.45	27.85	43.17	48.39	38.10	45.54	44.83	41.20
<i>Step Back Prompting Inference</i>											
Claude Haiku	12.77	20.45	23.66	15.91	27.85	26.87	19.35	14.29	20.54	20.69	22.00
Claude-3 Sonnet	27.66	43.18	36.16	27.27	24.05	28.63	22.58	29.52	35.71	33.33	31.60
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-1.5 Pro	31.91	38.64	38.84	25.00	29.11	31.28	32.26	31.43	32.14	27.59	32.70
Claude-3.5 Sonnet	34.04	43.18	41.96	47.73	29.11	43.61	48.39	38.10	46.43	45.98	41.90

Table 3: Results of closed-source MLLMs on the *testmini* split of POLYMATH. We report model results using the following prompting strategies: zero shot inference, few short 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.)

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.

Qwen2-VL-2B-Instruct (Wang et al., 2024b) (c) Text Based LLMs Reka Flash (Ormazabal et al., 2024), Llama-3 (70B) (AI@Meta, 2024), Mistral Large (AI, 2024). We conduct experiments on all open-source models using six NVIDIA A100 GPUs. Hyperparameters are available in §D.

Implementation Details All reported results are based on the *testmini* subset of the dataset. To establish a baseline for comparison, 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 (four engineering graduates and two PhDs), serving as a human performance baseline. We evaluate the benchmark using various prompting methods,

Category	PS	FC	PR	SC	RR	MR	NR	SR	OD	LR	Overall
<i>MLLM Inference on Diagrams (Multi-modal)</i>											
Claude-3 Haiku	16.67	15.57	18.55	22.62	25.93	19.51	31.76	17.83	21.52	33.33	20.60
Claude-3 Sonnet	21.67	23.77	22.98	17.86	20.37	24.39	32.94	22.48	26.58	66.67	23.60
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-1.5 Pro	11.67	23.77	22.58	27.38	28.70	25.61	10.59	18.60	29.11	66.67	22.50
Claude-3.5 Sonnet	31.67	27.87	25.00	19.05	28.70	25.61	25.88	22.48	31.65	100.00	26.20
<i>MLLM Inference on Diagram Descriptions (Text-only)</i>											
Claude-3 Haiku	30.00	25.41	18.55	19.05	25.93	28.05	27.06	26.36	30.38	100.00	24.60
Claude-3 Sonnet	30.00	32.79	25.40	22.62	26.85	36.59	37.65	26.36	31.65	100.00	29.30
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-1.5 Pro	25.00	26.23	25.00	27.38	21.30	28.05	16.47	19.38	22.78	33.33	23.60
Claude-3.5 Sonnet	38.33	30.33	26.61	23.81	37.96	35.37	34.12	28.68	36.71	100.00	31.40
<i>LLM Inference on Diagram Descriptions (Text-only)</i>											
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: Results of visual comprehension ablation study *test-img* split of POLYMATH. We use MLLMs and conduct multi-modal inference on questions containing diagrams, and then use the same MLLMs to infer on the same questions, but with a detailed text description in place of the diagram. For each inference setting, the highest and lowest scores achieved by a model per category are highlighted. Additionally, we report the performance of text-only LLMs on the textual representation of these questions.

including zero shot, few shot (2-shot), Chain-of-Thought (Wei et al., 2022b), and Step Back prompting (Zheng et al., 2024). 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. Details about these prompts are provided in §E. 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), Claude-3.5 Sonnet performed best with these advanced prompts, achieving up to 41.90% accuracy in Step Back Prompting, compared to 39.70% in zero shot. GPT-4o followed closely, especially in FC and PS questions, showing strong performance with zero shot and Step Back Prompting. Gemini-1.5 Pro performed moderately across all categories but lacked dominance in any specific area, while Claude Haiku being the smallest of the closed sourced MLLMs, consistently underperformed across all prompting strategies. In terms of prompting strategies, Chain-of-Thought and Step Back Prompting enhanced the performance of top models like Claude-3.5 Sonnet and GPT-4o, allowing them to excel in tasks requiring structured reasoning and re-evaluation. Both strategies led to marked improvements over zero shot prompting, particularly in categories like SR, PR, and LR.

Open Source Models Table 4 showcases the results of popular open-source MLLMs. LLaVA-v1.6-Mistral-7B model achieved the highest overall score of 15.2%, demonstrating remarkable performance across several categories. Notably, 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.

Error Name	Definition	Gemini	GPT	Claude
Incomplete (IC)	Model generated incomplete solution, or output hit token limit	6.36	5.08	0.42
Logical Flaw (LF)	Reasoning step violated established logical rules or real-world principles (such as equality or cardinality)	58.05	52.54	57.20
Memory Flaw (MF)	Model forgets information provided in the question or earlier in the solution	11.86	9.75	11.44
Spatial Misunderstanding (SM)	Model misunderstands spatial relations or “misreads” specific details of given image.	16.10	24.58	16.53
Calculation Error (CE)	Model commits a mathematical error, or substitutes the wrong value in an equation.	2.97	1.27	6.36
Misalignment (MG)	Model reasons correctly, but concludes the answer incorrectly (eg. identifying the pattern but selecting the wrong option)	4.66	6.78	8.05

Table 6: The types of errors found in model reasoning patterns. The errors are defined to be mutually distinct and leave very little room for ambiguity. We also report the frequency of these errors for each model (Gemini-1.5 Pro, Claude-3.5 Sonnet, GPT-4o) over the 236 questions analysed.

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. This strategy aimed to prevent them from gaining additional information from another question from same category. They were asked to provide only the final answer without detailed reasoning. Therefore, we do not report the Chain-of-Thought evaluation results for human performance, alongside the ‘Random Chance’ baseline.

4.3 EXPERIMENTAL ANALYSIS

MLLMs Rely More on Image Descriptions than Image To evaluate the visual reasoning capabilities of closed-source models, we conducted inference on the *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-3.5 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 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 approximately $\sim 3 - 4\%$, indicating that the models struggle with visualizing 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 significantly 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 total of 236 samples where all three state of the art MLLMs (Claude-3.5 Sonnet, GPT-4o and Gemini-1.5 Pro) gave incorrect answers on *testmini*. Based on the manual inspection of the responses, we identified 7 types of errors that MLLMs make (Table 6). The total error distribution of all three models is present in Table 11. Qualitative examples for category-wise errors are present in §H. 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 ($\sim 25\%$). Figure 4 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 visual information (which models fall short of) is integral to solving these questions. Additionally, in questions involving extrapolation over multiple weakly connected data points, models came to conclusions that contradicted earlier data, pointing to a lack of information retention. Finally, we find that models fell into the same fallacious reasoning patterns as one another - for example, making the assumption that a pattern holds across each row, when the correct reasoning involves a pattern replicated across columns. The category with the highest % of shared errors was PR, where we observed that GPT4-o, Gemini-1.5 Pro, and Claude-3.5 Sonnet 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. For more details, see §G.

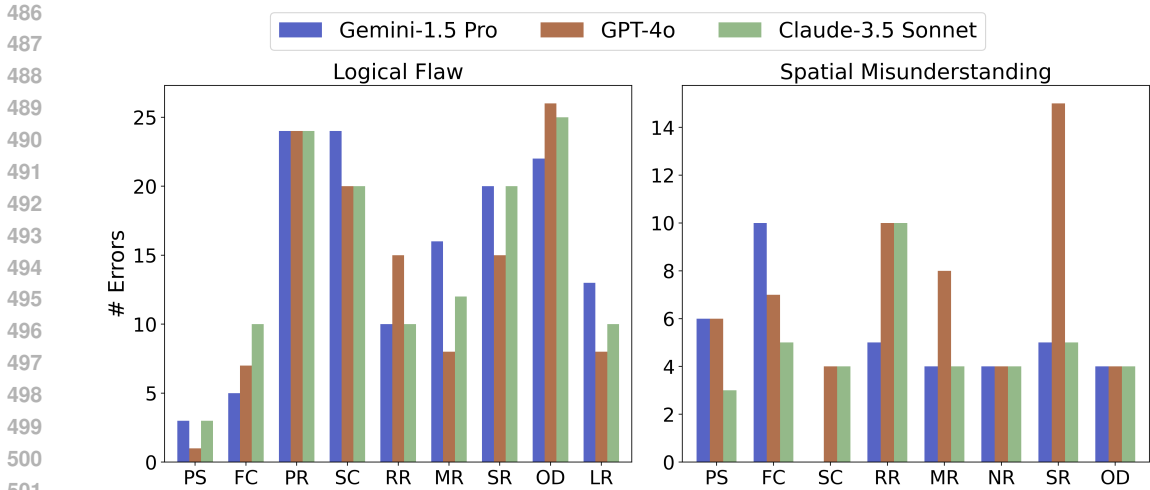


Figure 4: Frequency of Logical Flow (LF) and Spatial Misunderstanding (SM) errors across different question categories. We report per-model figures to enable a comparison of model abilities. They are most prevalent in the OD, PR, and SC categories of questions, owing to the amount of logical leaps and visual reasoning required by these questions.

Category	PS	FC	PR	SC	RR	MR	NR	SR	OD	LR	Overall
# Instances	4	0	125	79	58	196	14	34	41	44	595
Human Eval	100	-	61.60	69.62	82.76	64.29	71.43	79.41	82.93	59.09	68.40
o1-mini	0.00	-	58.40	30.38	91.38	64.80	71.43	44.12	63.41	40.91	58.15
o1-preview	0.00	-	75.20	50.63	81.03	70.41	57.14	44.12	73.17	56.82	66.72

Table 7: Results of OpenAI o1-mini and o1-preview on the *without diagram* (text-only) samples from the *testmini* split. We observe that while overall, human cognitive abilities have a slight edge over o1 models, over certain categories (PR, MR), o1 models outperform human performance.

Evaluation of OpenAI o1 models To understand the capabilities of OpenAI’s latest text-only reasoning models (o1-preview and o1-mini), we evaluate these models on 595 questions of *testmini* that do not contain diagrams. We also present human baseline scores on the without diagram section of *testmini*. The results of our study are presented in Table 7. o1-preview (~ 67%) has scores that are competitive with human performance (~ 68%), while o1-mini (~ 58%) lags behind the human baseline by 10%.

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-3.5 Sonnet models. However, a substantial gap of ~ 24% still exists between Claude-3.5 Sonnet, 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

- 540
541
542 Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman,
543 Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report.
544 *arXiv preprint arXiv:2303.08774*, 2023.
- 545 Mistral AI. Au large, Apr 2024. URL <https://mistral.ai/news/mistral-large/>.
- 546
547 AI@Meta. Llama 3 model card, 2024. URL [https://github.com/meta-llama/llama3/
548 blob/main/MODEL_CARD.md](https://github.com/meta-llama/llama3/blob/main/MODEL_CARD.md).
- 549
550 Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel
551 Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language
552 model for few-shot learning. *Advances in Neural Information Processing Systems*, 35:23716–
553 23736, 2022.
- 554
555 Aida Amini, Saadia Gabriel, Shanchuan Lin, Rik Koncel-Kedziorski, Yejin Choi, and Hannaneh
556 Hajishirzi. Mathqa: Towards interpretable math word problem solving with operation-based
557 formalisms. In *Proceedings of the 2019 Conference of the North American Chapter of the
558 Association for Computational Linguistics: Human Language Technologies (NAACL)*, pp. 2357–
2367, 2019.
- 559
560 Anthropic. Claude 2, 2023. URL <https://www.anthropic.com/index/claude-2>.
- 561
562 Anthropic. Claude 3.5 sonnet model card addendum, 2024a. URL [https://api.
563 semanticscholar.org/CorpusID:270667923](https://api.semanticscholar.org/CorpusID:270667923).
- 564
565 Anthropic. The claude 3 model family: Opus, sonnet, haiku, 2024b. URL [https://api.
566 semanticscholar.org/CorpusID:268232499](https://api.semanticscholar.org/CorpusID:268232499).
- 567
568 Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick,
569 and Devi Parikh. VQA: Visual question answering. In *Proceedings of the IEEE international
570 conference on computer vision*, pp. 2425–2433, 2015.
- 571
572 Anas Awadalla, Irena Gao, Josh Gardner, Jack Hessel, Yusuf Hanafy, Wanrong Zhu, Kalyani Marathe,
573 Yonatan Bitton, Samir Gadre, Shiori Sagawa, et al. OpenFlamingo: An open-source framework for
574 training large autoregressive vision-language models. *arXiv preprint arXiv:2308.01390*, 2023.
- 575
576 Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou,
577 and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization,
578 text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 2023.
- 579
580 Yonatan Bitton, Hritik Bansal, Jack Hessel, Rulin Shao, Wanrong Zhu, Anas Awadalla, Josh Gardner,
581 Rohan Taori, and Ludwig Schimdt. VisIT-Bench: A benchmark for vision-language instruction
582 following inspired by real-world use. *arXiv preprint arXiv:2308.06595*, 2023.
- 583
584 Nitzan Bitton-Guetta, Yonatan Bitton, Jack Hessel, Ludwig Schmidt, Yuval Elovici, Gabriel
585 Stanovsky, and Roy Schwartz. Breaking common sense: WHOOPS! A vision-and-language
586 benchmark of synthetic and compositional images. *arXiv preprint arXiv:2303.07274*, 2023.
- 587
588 Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal,
589 Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are
590 few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.
- 591
592 Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar,
593 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.
- 594
595 Jie Cao and Jing Xiao. An augmented benchmark dataset for geometric question answering through
596 dual parallel text encoding. In *Proceedings of the 29th International Conference on Computational
597 Linguistics*, pp. 1511–1520, 2022.
- 598
599 Shuaichen Chang, David Palzer, Jialin Li, Eric Fosler-Lussier, and Ningchuan Xiao. MapQA: A
600 dataset for question answering on choropleth maps. *arXiv preprint arXiv:2211.08545*, 2022.

- 594 Guo Chen, Yin-Dong Zheng, Jiahao Wang, Jilan Xu, Yifei Huang, Junting Pan, Yi Wang, Yali Wang,
595 Yu Qiao, Tong Lu, et al. Videollm: Modeling video sequence with large language models. *arXiv*
596 *preprint arXiv:2305.13292*, 2023a.
- 597 Jiaqi Chen, Jianheng Tang, Jinghui Qin, Xiaodan Liang, Lingbo Liu, Eric P. Xing, and Liang Lin.
598 Geoqa: A geometric question answering benchmark towards multimodal numerical reasoning.
599 *ArXiv*, abs/2105.14517, 2021a. URL [https://api.semanticscholar.org/CorpusID:
600 235253782](https://api.semanticscholar.org/CorpusID:235253782).
- 601 Jiaqi Chen, Tong Li, Jinghui Qin, Pan Lu, Liang Lin, Chongyu Chen, and Xiaodan Liang. UniGeo:
602 Unifying geometry logical reasoning via reformulating mathematical expression. In *Proceedings*
603 *of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 3313–3323,
604 2022a.
- 605 Jun Chen, Deyao Zhu¹ Xiaoqian Shen¹ Xiang Li, Zechun Liu² Pengchuan Zhang, Raghuraman
606 Krishnamoorthi² Vikas Chandra² Yunyang Xiong, and Mohamed Elhoseiny. Minigt-v2: Large
607 language model as a unified interface for vision-language multi-task learning. *arXiv preprint*
608 *arXiv:2310.09478*, 2023b.
- 609 Lin Chen, Jinsong Li, Xiao wen Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua
610 Lin. Sharegpt4v: Improving large multi-modal models with better captions. *ArXiv*, abs/2311.12793,
611 2023c. URL <https://api.semanticscholar.org/CorpusID:265308687>.
- 612 Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared
613 Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. Evaluating large
614 language models trained on code. *arXiv preprint arXiv:2107.03374*, 2021b.
- 615 Wenhu Chen, Xueguang Ma, Xinyi Wang, and William W Cohen. Program of thoughts prompt-
616 ing: Disentangling computation from reasoning for numerical reasoning tasks. *arXiv preprint*
617 *arXiv:2211.12588*, 2022b.
- 618 Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser,
619 Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. Training verifiers to solve
620 math word problems. *arXiv preprint arXiv:2110.14168*, 2021.
- 621 Adam Dahlgren Lindström and Savitha Sam Abraham. CLEVR-Math: A dataset for compositional
622 language, visual and mathematical reasoning. In *16th International Workshop on Neural-Symbolic*
623 *Learning and Reasoning, NeSy 2022, Windsor, UK, september 28-30, 2022.*, volume 3212. CEUR-
624 WS, 2022.
- 625 Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang,
626 Boyang Li, Pascale Fung, and Steven Hoi. InstructBLIP: Towards general-purpose vision-language
627 models with instruction tuning, 2023.
- 628 Qingxiu Dong, Li Dong, Ke Xu, Guangyan Zhou, Yaru Hao, Zhifang Sui, and Furu Wei. Large
629 language model for science: A study on P vs. NP. *arXiv preprint arXiv:2309.05689*, 2023.
- 630 Xiaoyi Dong, Pan Zhang, Yuhang Zang, Yuhang Cao, Bin Wang, Linke Ouyang, Xilin Wei, Songyang
631 Zhang, Haodong Duan, Maosong Cao, et al. Internlm-xcomposer2: Mastering free-form text-image
632 composition and comprehension in vision-language large model. *arXiv preprint arXiv:2401.16420*,
633 2024.
- 634 Simon Frieder, Luca Pinchetti, Ryan-Rhys Griffiths, Tommaso Salvatori, Thomas Lukasiewicz,
635 Philipp Christian Petersen, Alexis Chevalier, and Julius Berner. Mathematical capabilities of
636 chatgpt. In *37th Conference on Neural Information Processing Systems (NeurIPS 2023) Track on*
637 *Datasets and Benchmarks*, 2023.
- 638 Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu
639 Zheng, Ke Li, Xing Sun, Yunsheng Wu, and Rongrong Ji. Mme: A comprehensive evaluation
640 benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*, 2023a.
- 641 Chaoyou Fu, Renrui Zhang, Haojia Lin, Zihan Wang, Timin Gao, Yongdong Luo, Yubo Huang,
642 Zhengye Zhang, Longtian Qiu, Gaoxiang Ye, et al. A challenger to gpt-4v? early explorations of
643 gemini in visual expertise. *arXiv preprint arXiv:2312.12436*, 2023b.
- 644

- 648 Lingyue Fu, Huacan Chai, Shuang Luo, Kounianhua Du, Weiming Zhang, Longteng Fan, Jiayi Lei,
649 Renting Rui, Jianghao Lin, Yuchen Fang, et al. CodeApex: A bilingual programming evaluation
650 benchmark for large language models. *arXiv preprint arXiv:2309.01940*, 2023c.
- 651
- 652 Jiahui Gao, Renjie Pi, Jipeng Zhang, Jiacheng Ye, Wanjun Zhong, Yufei Wang, Lanqing Hong,
653 Jianhua Han, Hang Xu, Zhenguo Li, et al. G-llava: Solving geometric problem with multi-modal
654 large language model. *arXiv preprint arXiv:2312.11370*, 2023a.
- 655
- 656 Peng Gao, Jiaming Han, Renrui Zhang, Ziyi Lin, Shijie Geng, Aojun Zhou, Wei Zhang, Pan Lu,
657 Conghui He, Xiangyu Yue, Hongsheng Li, and Yu Qiao. LLaMA-Adapter V2: Parameter-efficient
658 visual instruction model. *arXiv preprint arXiv:2304.15010*, 2023b.
- 659
- 660 Peng Gao, Renrui Zhang, Chris Liu, Longtian Qiu, Siyuan Huang, Weifeng Lin, Shitian Zhao, Shijie
661 Geng, Ziyi Lin, Peng Jin, et al. Sphinx-x: Scaling data and parameters for a family of multi-modal
662 large language models. *arXiv preprint arXiv:2402.05935*, 2024.
- 663
- 664 Google. Bard, 2023. URL <https://bard.google.com/>.
- 665
- 666 Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the V
667 in VQA matter: Elevating the role of image understanding in visual question answering. In
668 *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6904–6913,
2017.
- 669
- 670 Ziyu Guo, Renrui Zhang, Xiangyang Zhu, Yiwen Tang, Xianzheng Ma, Jiaming Han, Kexin Chen,
671 Peng Gao, Xianzhi Li, Hongsheng Li, et al. Point-bind & point-llm: Aligning point cloud
672 with multi-modality for 3d understanding, generation, and instruction following. *arXiv preprint
673 arXiv:2309.00615*, 2023.
- 674
- 675 Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and
676 Jeffrey P Bigham. VizWiz grand challenge: Answering visual questions from blind people. In
677 *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3608–3617,
2018.
- 678
- 679 Jiaming Han, Renrui Zhang, Wenqi Shao, Peng Gao, Peng Xu, Han Xiao, Kaipeng Zhang, Chris Liu,
680 Song Wen, Ziyu Guo, et al. Imagebind-llm: Multi-modality instruction tuning. *arXiv preprint
681 arXiv:2309.03905*, 2023.
- 682
- 683 Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob
684 Steinhardt. Measuring massive multitask language understanding. *Proceedings of the International
685 Conference on Learning Representations (ICLR)*, 2021a.
- 686
- 687 Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song,
688 and Jacob Steinhardt. Measuring mathematical problem solving with the math dataset. *NeurIPS*,
2021b.
- 689
- 690 Yining Hong, Haoyu Zhen, Peihao Chen, Shuhong Zheng, Yilun Du, Zhenfang Chen, and Chuang
691 Gan. 3d-llm: Injecting the 3d world into large language models. *Advances in Neural Information
692 Processing Systems*, 36, 2024.
- 693
- 694 Wenlong Huang, Pieter Abbeel, Deepak Pathak, and Igor Mordatch. Language models as zero-shot
695 planners: Extracting actionable knowledge for embodied agents. In *International Conference on
696 Machine Learning*, pp. 9118–9147. PMLR, 2022.
- 697
- 698 Yuzhen Huang, Yuzhuo Bai, Zhihao Zhu, Junlei Zhang, Jinghan Zhang, Tangjun Su, Junteng Liu,
699 Chuancheng Lv, Yikai Zhang, Jiayi Lei, et al. C-eval: A multi-level multi-discipline chinese
700 evaluation suite for foundation models. *arXiv preprint arXiv:2305.08322*, 2023.
- 701
- 702 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.

- 702 Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris
703 Bamford, Devendra Singh Chaplot, Diego de Las Casas, Emma Bou Hanna, Florian Bressand,
704 Gianna Lengyel, Guillaume Bour, Guillaume Lample, L elio Renard Lavaud, Lucile Saulnier, Marie-
705 Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le
706 Scao, Th eophile Gervet, Thibaut Lavril, Thomas Wang, Timoth e Lacroix, and William El Sayed.
707 Mixtral of experts. *Arxiv 2401.04088*, 2024.
- 708
709 Kushal Kafle, Brian Price, Scott Cohen, and Christopher Kanan. DVQA: Understanding data
710 visualizations via question answering. In *Proceedings of the IEEE conference on computer vision
711 and pattern recognition*, pp. 5648–5656, 2018.
- 712
713 Samira Ebrahimi Kahou, Vincent Michalski, Adam Atkinson,  kos K ad ar, Adam Trischler, and
714 Yoshua Bengio. FigureQA: An annotated figure dataset for visual reasoning. *arXiv preprint
715 arXiv:1710.07300*, 2017.
- 716
717 Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi.
718 A diagram is worth a dozen images. In *Computer Vision–ECCV 2016: 14th European Conference,
719 Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*, pp. 235–251.
720 Springer, 2016.
- 721
722 Aniruddha Kembhavi, Minjoon Seo, Dustin Schwenk, Jonghyun Choi, Ali Farhadi, and Hannaneh
723 Hajishirzi. Are you smarter than a sixth grader? Textbook question answering for multimodal
724 machine comprehension. In *Proceedings of the IEEE Conference on Computer Vision and Pattern
725 recognition*, pp. 4999–5007, 2017.
- 726
727 Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete
728 Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. *arXiv preprint
729 arXiv:2304.02643*, 2023.
- 730
731 Jason J Lau, Soumya Gayen, Asma Ben Abacha, and Dina Demner-Fushman. A dataset of clinically
732 generated visual questions and answers about radiology images. *Scientific data*, 5(1):1–10, 2018.
- 733
734 Kenton Lee, Mandar Joshi, Iulia Raluca Turc, Hexiang Hu, Fangyu Liu, Julian Martin Eisenschlos,
735 Urvashi Khandelwal, Peter Shaw, Ming-Wei Chang, and Kristina Toutanova. Pix2Struct: Screen-
736 shot parsing as pretraining for visual language understanding. In *International Conference on
737 Machine Learning*, pp. 18893–18912. PMLR, 2023.
- 738
739 Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Fanyi Pu, Jingkang Yang, Chunyuan Li, and
740 Ziwei Liu. Mimic-it: Multi-modal in-context instruction tuning. *arXiv preprint arXiv:2306.05425*,
741 2023a.
- 742
743 Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. Seed-bench: Bench-
744 marking multimodal llms with generative comprehension. *ArXiv*, abs/2307.16125, 2023b.
- 745
746 Chunyuan Li, Zhe Gan, Zhengyuan Yang, Jianwei Yang, Linjie Li, Lijuan Wang, and Jianfeng Gao.
747 Multimodal foundation models: From specialists to general-purpose assistants. *arXiv preprint
748 arXiv:2309.10020*, 2023c.
- 749
750 Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-
751 training for unified vision-language understanding and generation. In *International Conference on
752 Machine Learning*, pp. 12888–12900. PMLR, 2022.
- 753
754 Zhuowan Li, Xingrui Wang, Elias Stengel-Eskin, Adam Kortylewski, Wufei Ma, Benjamin
755 Van Durme, and Alan L Yuille. Super-CLEVR: A virtual benchmark to diagnose domain ro-
756 bustness in visual reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision
757 and Pattern Recognition*, pp. 14963–14973, 2023d.
- 758
759 Thomas Liao, Rohan Taori, Inioluwa Deborah Raji, and Ludwig Schmidt. Are we learning yet? A
760 meta review of evaluation failures across machine learning. In *Thirty-fifth Conference on Neural
761 Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021.

- 756 Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr
757 Dollár, and C Lawrence Zitnick. Microsoft COCO: Common objects in context. In *Computer
758 Vision—ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014,
759 Proceedings, Part V 13*, pp. 740–755. Springer, 2014.
- 760
761 Ziyi Lin, Chris Liu, Renrui Zhang, Peng Gao, Longtian Qiu, Han Xiao, Han Qiu, Chen Lin, Wenqi
762 Shao, Keqin Chen, et al. Sphinx: The joint mixing of weights, tasks, and visual embeddings for
763 multi-modal large language models. *arXiv preprint arXiv:2311.07575*, 2023.
- 764
765 Fangyu Liu, Francesco Piccinno, Syrine Krichene, Chenxi Pang, Kenton Lee, Mandar Joshi, Yasemin
766 Altun, Nigel Collier, and Julian Martin Eisenschlos. MatCha: Enhancing visual language pretrain-
767 ing with math reasoning and chart derendering. *arXiv preprint arXiv:2212.09662*, 2022.
- 768
769 Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction
770 tuning, 2023a.
- 771
772 Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *arXiv
773 preprint arXiv:2304.08485*, 2023b.
- 774
775 Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee.
776 Llava-next: Improved reasoning, ocr, and world knowledge, January 2024. URL [https://
777 llava-vl.github.io/blog/2024-01-30-llava-next/](https://llava-vl.github.io/blog/2024-01-30-llava-next/).
- 778
779 Xiao Liu, Hao Yu, Hanchen Zhang, Yifan Xu, Xuanyu Lei, Hanyu Lai, Yu Gu, Hangliang Ding,
780 Kaiwen Men, Kejuan Yang, et al. AgentBench: Evaluating LLMs as agents. *arXiv preprint
781 arXiv:2308.03688*, 2023c.
- 782
783 Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi
784 Wang, Conghui He, Ziwei Liu, et al. MMBench: Is your multi-modal model an all-around player?
785 *arXiv preprint arXiv:2307.06281*, 2023d.
- 786
787 Yuliang Liu, Zhang Li, Hongliang Li, Wenwen Yu, Mingxin Huang, Dezhi Peng, Mingyu Liu, Mingrui
788 Chen, Chunyuan Li, Lianwen Jin, et al. On the hidden mystery of OCR in large multimodal models.
789 *arXiv preprint arXiv:2305.07895*, 2023e.
- 790
791 Pan Lu, Ran Gong, Shibiao Jiang, Liang Qiu, Siyuan Huang, Xiaodan Liang, and Song-Chun Zhu.
792 Inter-GPS: Interpretable geometry problem solving with formal language and symbolic reasoning.
793 In *The 59th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2021a.
- 794
795 Pan Lu, Liang Qiu, Jiaqi Chen, Tony Xia, Yizhou Zhao, Wei Zhang, Zhou Yu, Xiaodan Liang,
796 and Song-Chun Zhu. IconQA: A new benchmark for abstract diagram understanding and visual
797 language reasoning. In *The 35th Conference on Neural Information Processing Systems (NeurIPS)
798 Track on Datasets and Benchmarks*, 2021b.
- 799
800 Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord,
801 Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for
802 science question answering. In *The 36th Conference on Neural Information Processing Systems
803 (NeurIPS)*, 2022.
- 804
805 Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chun yue Li, Hannaneh Hajishirzi, Hao Cheng,
806 Kai-Wei Chang, Michel Galley, and Jianfeng Gao. Mathvista: Evaluating math reasoning in visual
807 contexts with gpt-4v, bard, and other large multimodal models. *ArXiv*, abs/2310.02255, 2023a.
- 808
809 Pan Lu, Liang Qiu, Wenhao Yu, Sean Welleck, and Kai-Wei Chang. A survey of deep learning
810 for mathematical reasoning. In *The 61st Annual Meeting of the Association for Computational
811 Linguistics (ACL)*, 2023b.
- 812
813 Haipeng Luo, Qingfeng Sun, Can Xu, Pu Zhao, Jianguang Lou, Chongyang Tao, Xiubo Geng,
814 Qingwei Lin, Shifeng Chen, and Dongmei Zhang. Wizardmath: Empowering mathematical
815 reasoning for large language models via reinforced evol-instruct. *arXiv preprint arXiv:2308.09583*,
816 2023.

- 810 Ahmed Masry, Xuan Long Do, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. ChartQA: A benchmark
811 for question answering about charts with visual and logical reasoning. In *Findings of the Association*
812 *for Computational Linguistics: ACL 2022*, pp. 2263–2279, 2022.
- 813
814 Ahmed Masry, Parsa Kavehzadeh, Xuan Long Do, Enamul Hoque, and Shafiq Joty. UniChart: A
815 universal vision-language pretrained model for chart comprehension and reasoning. *arXiv preprint*
816 *arXiv:2305.14761*, 2023.
- 817 Minesh Mathew, Viraj Bagal, Rubèn Tito, Dimosthenis Karatzas, Ernest Valveny, and CV Jawahar.
818 InfographicsVQA. In *Proceedings of the IEEE/CVF Winter Conference on Applications of*
819 *Computer Vision*, pp. 1697–1706, 2022.
- 820
821 Nitesh Methani, Pritha Ganguly, Mitesh M Khapra, and Pratyush Kumar. PlotQA: Reasoning over
822 scientific plots. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer*
823 *Vision*, pp. 1527–1536, 2020.
- 824 Swaroop Mishra, Matthew Finlayson, Pan Lu, Leonard Tang, Sean Welleck, Chitta Baral, Tanmay
825 Rajpurohit, Oyvind Tafjord, Ashish Sabharwal, Peter Clark, and Ashwin Kalyan. LILA: A unified
826 benchmark for mathematical reasoning. In *The 2022 Conference on Empirical Methods in Natural*
827 *Language Processing (EMNLP)*, 2022.
- 828
829 Shaghayegh Mobasher, Ghazal Zamaninejad, Maryam Hashemi, Melika Nobakhtian, and Sauleh
830 Eetemadi. ParsVQA-Caps: A benchmark for visual question answering and image captioning in
831 persian. *people*, 101:404, 2022.
- 832 Harsha Nori, Nicholas King, Scott Mayer McKinney, Dean Carignan, and Eric Horvitz. Capabilities
833 of GPT-4 on medical challenge problems. *arXiv preprint arXiv:2303.13375*, 2023.
- 834
835 OpenAI. Chatgpt. <https://chat.openai.com>, 2023a.
- 836
837 OpenAI. OpenAI’s GPT-4. <https://openai.com/research/gpt-4>, 2023b.
- 838
839 OpenAI. GPT-4V(ision) system card, 2023c. URL <https://openai.com/research/gpt-4v-system-card>.
- 840
841 OpenAI. Gpt 4o models, 2024a. URL <https://openai.com/index/hello-gpt-4o/>.
- 842
843 OpenAI. Open ai o1 models, 2024b. URL <https://openai.com/index/introducing-openai-o1-preview/>.
- 844
845 Aitor Ormazabal, Che Zheng, Cyprien de Masson d’Autume, Dani Yogatama, Deyu Fu, Donovan
846 Ong, Eric Chen, Eugenie Lamprecht, Hai Pham, Isaac Ong, Kaloyan Aleksiev, Lei Li, Matthew
847 Henderson, Max Bain, Mikel Artetxe, Nishant Relan, Piotr Padlewski, Qi Liu, Ren Chen, Samuel
848 Phua, Yazheng Yang, Yi Tay, Yuqi Wang, Zhongkai Zhu, and Zhihui Xie. Reka core, flash, and
849 edge: A series of powerful multimodal language models, 2024.
- 850
851 Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal,
852 Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever.
853 Learning transferable visual models from natural language supervision. In *International Conference*
854 *on Machine Learning*, 2021. URL <https://api.semanticscholar.org/CorpusID:231591445>.
- 855
856 Subhro Roy and Dan Roth. Solving general arithmetic word problems. *ArXiv*, abs/1608.01413, 2016.
857 URL <https://api.semanticscholar.org/CorpusID:560565>.
- 858
859 Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi
860 Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. LAION-5B:
861 An open large-scale dataset for training next generation image-text models. *Advances in Neural*
862 *Information Processing Systems*, 35:25278–25294, 2022.
- 863
864 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.

- 864 Minjoon Seo, Hannaneh Hajishirzi, Ali Farhadi, Oren Etzioni, and Clint Malcolm. Solving geometry
865 problems: Combining text and diagram interpretation. In *Proceedings of the 2015 conference on*
866 *empirical methods in natural language processing*, pp. 1466–1476, 2015.
- 867
- 868 Sanket Shah, Anand Mishra, Naganand Yadati, and Partha Pratim Talukdar. KVQA: Knowledge-
869 aware visual question answering. In *Proceedings of the AAAI conference on artificial intelligence*,
870 pp. 8876–8884, 2019.
- 871 Wenqi Shao, Yutao Hu, Peng Gao, Meng Lei, Kaipeng Zhang, Fanqing Meng, Peng Xu, Siyuan
872 Huang, Hongsheng Li, Yu Qiao, et al. Tiny lvm-ehub: Early multimodal experiments with bard.
873 *arXiv preprint arXiv:2308.03729*, 2023.
- 874
- 875 Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual Captions: A cleaned,
876 hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th*
877 *Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp.
878 2556–2565, 2018.
- 879 Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh,
880 and Marcus Rohrbach. Towards VQA models that can read. In *Proceedings of the IEEE/CVF*
881 *conference on computer vision and pattern recognition*, pp. 8317–8326, 2019.
- 882
- 883 Yixuan Su, Tian Lan, Huayang Li, Jialu Xu, Yan Wang, and Deng Cai. Pandagpt: One model to
884 instruction-follow them all. *arXiv preprint arXiv:2305.16355*, 2023.
- 885 Keqiang Sun, Junting Pan, Yuying Ge, Hao Li, Haodong Duan, Xiaoshi Wu, Renrui Zhang, Aojun
886 Zhou, Zipeng Qin, Yi Wang, et al. Journeydb: A benchmark for generative image understanding.
887 *Advances in Neural Information Processing Systems*, 36, 2024.
- 888
- 889 Liangtai Sun, Yang Han, Zihan Zhao, Da Ma, Zhennan Shen, Baocai Chen, Lu Chen, and Kai Yu.
890 SciEval: A multi-level large language model evaluation benchmark for scientific research. *arXiv*
891 *preprint arXiv:2308.13149*, 2023.
- 892
- 893 John Chong Min Tan and Mehul Motani. Large language model (llm) as a system of multiple expert
894 agents: An approach to solve the abstraction and reasoning corpus (arc) challenge. *arXiv preprint*
arXiv:2310.05146, 2023.
- 895
- 896 Ross Taylor, Marcin Kardas, Guillem Cucurull, Thomas Scialom, Anthony Hartshorn, Elvis Saravia,
897 Andrew Poulton, Viktor Kerkez, and Robert Stojnic. Galactica: A large language model for science.
898 *arXiv preprint arXiv:2211.09085*, 2022.
- 899
- 900 Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu
901 Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. Gemini: a family of highly capable
multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.
- 902
- 903 Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée
904 Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. LLaMA: Open and
efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023a.
- 905
- 906 Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay
907 Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation
908 and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023b.
- 909
- 910 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.
- 911
- 912 Ke Wang, Houxing Ren, Aojun Zhou, Zimu Lu, Sichun Luo, Weikang Shi, Renrui Zhang, Linqi Song,
913 Mingjie Zhan, and Hongsheng Li. Mathcoder: Seamless code integration in LLMs for enhanced
914 mathematical reasoning. In *The Twelfth International Conference on Learning Representations*,
915 2024a. URL <https://openreview.net/forum?id=z8TW0ttBPp>.
- 916
- 917 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*, 2024b.

- 918 Ruo Cheng Wang, Eric Zelikman, Gabriel Poesia, Yewen Pu, Nick Haber, and Noah D Goodman.
919 Hypothesis search: Inductive reasoning with language models. *arXiv preprint arXiv:2309.05660*,
920 2023a.
- 921 Xiaoxuan Wang, Ziniu Hu, Pan Lu, Yanqiao Zhu, Jieyu Zhang, Satyen Subramaniam, Arjun R
922 Loomba, Shichang Zhang, Yizhou Sun, and Wei Wang. SciBench: Evaluating college-level
923 scientific problem-solving abilities of large language models. *arXiv preprint arXiv:2307.10635*,
924 2023b.
- 925 Yan Wang, Xiaojiang Liu, and Shuming Shi. Deep neural solver for math word problems. In
926 *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*
927 *(EMNLP)*, pp. 845–854, 2017.
- 928 Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama,
929 Maarten Bosma, Denny Zhou, Donald Metzler, et al. Emergent abilities of large language models.
930 *arXiv preprint arXiv:2206.07682*, 2022a.
- 931 Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Chi, Quoc Le, and Denny
932 Zhou. Chain of thought prompting elicits reasoning in large language models. *arXiv preprint*
933 *arXiv:2201.11903*, 2022b.
- 934 Shijie Wu, Ozan Irsoy, Steven Lu, Vadim Dabravolski, Mark Dredze, Sebastian Gehrmann, Prabhan-
935 jan Kambadur, David Rosenberg, and Gideon Mann. BloombergGPT: A large language model for
936 finance. *arXiv preprint arXiv:2303.17564*, 2023.
- 937 Peng Xu, Wenqi Shao, Kaipeng Zhang, Peng Gao, Shuo Liu, Meng Lei, Fanqing Meng, Siyuan
938 Huang, Yu Qiao, and Ping Luo. LVLm-eHub: A comprehensive evaluation benchmark for large
939 vision-language models. *arXiv preprint arXiv:2306.09265*, 2023a.
- 940 Runsen Xu, Xiaolong Wang, Tai Wang, Yilun Chen, Jiangmiao Pang, and Dahua Lin. PointLLM:
941 Empowering large language models to understand point clouds. *arXiv preprint arXiv:2308.16911*,
942 2023b.
- 943 Hongyang Yang, Xiao-Yang Liu, and Christina Dan Wang. FinGPT: Open-source financial large
944 language models. *arXiv preprint arXiv:2306.06031*, 2023.
- 945 Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu,
946 Pengcheng Shi, Yaya Shi, et al. mPlug-Owl: Modularization empowers large language models
947 with multimodality. *arXiv preprint arXiv:2304.14178*, 2023a.
- 948 Qinghao Ye, Haiyang Xu, Jiabo Ye, Ming Yan, Anwen Hu, Haowei Liu, Qi Qian, Ji Zhang, Fei
949 Huang, and Jingren Zhou. mplug-owl2: Revolutionizing multi-modal large language model with
950 modality collaboration, 2023b.
- 951 Da Yin, Liunian Harold Li, Ziniu Hu, Nanyun Peng, and Kai-Wei Chang. Broaden the vision:
952 Geo-diverse visual commonsense reasoning. *arXiv preprint arXiv:2109.06860*, 2021.
- 953 Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang,
954 and Lijuan Wang. MM-Vet: Evaluating large multimodal models for integrated capabilities. *arXiv*
955 *preprint arXiv:2308.02490*, 2023.
- 956 Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu
957 Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin,
958 Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhao Chen.
959 Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert
960 agi. *arXiv preprint arXiv:2311.16502*, 2023a.
- 961 Xiang Yue, Xingwei Qu, Ge Zhang, Yao Fu, Wenhao Huang, Huan Sun, Yu Su, and Wenhao Chen.
962 Mammoth: Building math generalist models through hybrid instruction tuning. *arXiv preprint*
963 *arXiv:2309.05653*, 2023b.
- 964 Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. From recognition to cognition: Visual
965 commonsense reasoning. In *Proceedings of the IEEE/CVF conference on computer vision and*
966 *pattern recognition*, pp. 6720–6731, 2019.

- 972 Hang Zhang, Xin Li, and Lidong Bing. Video-llama: An instruction-tuned audio-visual language
973 model for video understanding. *arXiv preprint arXiv:2306.02858*, 2023a.
- 974
- 975 Renrui Zhang, Jiaming Han, Aojun Zhou, Xiangfei Hu, Shilin Yan, Pan Lu, Hongsheng Li, Peng Gao,
976 and Qiao Yu. LLaMA-Adapter: Efficient fine-tuning of language models with zero-init attention.
977 *arXiv preprint arXiv:2303.16199*, 2023b.
- 978 Renrui Zhang, Xiangfei Hu, Bohao Li, Siyuan Huang, Hanqiu Deng, Hongsheng Li, Yu Qiao, and
979 Peng Gao. Prompt, generate, then cache: Cascade of foundation models makes strong few-shot
980 learners. *CVPR 2023*, 2023c.
- 981
- 982 Renrui Zhang, Zhengkai Jiang, Ziyu Guo, Shilin Yan, Junting Pan, Hao Dong, Peng Gao, and
983 Hongsheng Li. Personalize segment anything model with one shot. *ICLR 2024*, 2023d.
- 984 Renrui Zhang, Liuhui Wang, Yu Qiao, Peng Gao, and Hongsheng Li. Learning 3d representations
985 from 2d pre-trained models via image-to-point masked autoencoders. *CVPR 2023*, 2023e.
- 986
- 987 Renrui Zhang, Jiaming Han, Aojun Zhou, Xiangfei Hu, Shilin Yan, Pan Lu, Hongsheng Li, Peng Gao,
988 and Yu Qiao. LLaMA-adapter: Efficient fine-tuning of large language models with zero-initialized
989 attention. In *The Twelfth International Conference on Learning Representations*, 2024. URL
990 <https://openreview.net/forum?id=d4UiXAHN2W>.
- 991 Xiang Zhang, Senyu Li, Zijun Wu, and Ning Shi. Lost in translation: When gpt-4v (ision) can't see
992 eye to eye with text. a vision-language-consistency analysis of vllms and beyond. *arXiv preprint*
993 *arXiv:2310.12520*, 2023f.
- 994 Xiaoman Zhang, Chaoyi Wu, Ziheng Zhao, Weixiong Lin, Ya Zhang, Yanfeng Wang, and Weidi
995 Xie. PMC-VQA: Visual instruction tuning for medical visual question answering. *arXiv preprint*
996 *arXiv:2305.10415*, 2023g.
- 997
- 998 Yanzhe Zhang, Ruiyi Zhang, Jiuxiang Gu, Yufan Zhou, Nedim Lipka, Diyi Yang, and Tong Sun.
999 LLaVAR: Enhanced visual instruction tuning for text-rich image understanding. *arXiv preprint*
1000 *arXiv:2306.17107*, 2023h.
- 1001 Huaixiu Steven Zheng, Swaroop Mishra, Xinyun Chen, Heng-Tze Cheng, Ed H Chi, Quoc V Le, and
1002 Denny Zhou. Take a step back: Evoking reasoning via abstraction in large language models. In
1003 *The Twelfth International Conference on Learning Representations*, 2024.
- 1004
- 1005 Aojun Zhou, Ke Wang, Zimu Lu, Weikang Shi, Sichun Luo, Zipeng Qin, Shaoqing Lu, Anya Jia,
1006 Linqi Song, Mingjie Zhan, et al. Solving challenging math word problems using gpt-4 code
1007 interpreter with code-based self-verification. *arXiv preprint arXiv:2308.07921*, 2023.
- 1008 Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. MiniGPT-4: En-
1009 hancing vision-language understanding with advanced large language models. *arXiv preprint*
1010 *arXiv:2304.10592*, 2023a.
- 1011
- 1012 Wanrong Zhu, Jack Hessel, Anas Awadalla, Samir Yitzhak Gadre, Jesse Dodge, Alex Fang, Youngjae
1013 Yu, Ludwig Schmidt, William Yang Wang, and Yejin Choi. Multimodal C4: An open, billion-scale
1014 corpus of images interleaved with text. *arXiv preprint arXiv:2304.06939*, 2023b.
- 1015
- 1016
- 1017
- 1018
- 1019
- 1020
- 1021
- 1022
- 1023
- 1024
- 1025

APPENDIX

APPENDIX OVERVIEW

- Section **A**: Limitation and Future Work
- Section **B**: Extended Related Work
- Section **C**: Data Collection Pipeline Details
- Section **D**: Additional Experimental Details
- Section **E**: Prompts for Dataset Curation and Experiments
- Section **F**: Dataset Examples
- Section **H**: Qualitative Error Analysis

A LIMITATION AND FUTURE WORK

Our benchmark, POLYMATH, makes key contributions by integrating mathematical and visual tasks. While we have made progress in evaluating model performance, we recognize certain limitations. One limitation is dataset coverage. Although POLYMATH covers a wide range of tasks and visual contexts, some mathematical problems and visual types may be underrepresented. Additionally, focusing on mathematical reasoning within visual contexts, especially in domains like competitive high-school-level questions involving problems in spatial and logical reasoning, requires a more labor-intensive data collection process than text-only or general-purpose datasets. Consequently, the scalability and generalizability of our benchmark to other areas remain challenging. Annotations were performed by the authors meticulously, however, due to the diversity of questions and images appearing in these sources, the annotations lack a consistent format.

In future iterations, our benchmark will aim to cover a wider range of problems and visual contexts, with unified and comprehensive annotations. This benchmark is part of an ongoing research effort, and we are committed to maintaining and refining the datasets, including addressing potential data noise, based on community feedback. Additionally, we will adapt the leaderboard to reflect new model developments. In conclusion, despite the limitations of our current approach, POLYMATH marks a significant advancement in the field. We remain dedicated to continuously improving the benchmark to deepen our understanding of AI’s capabilities in mathematical and visual reasoning.

B EXTENDED RELATED WORK

High-quality evaluation datasets and benchmarks are crucial for assessing the progress of machine learning models in solving real-world tasks (Liao et al., 2021). Mathematical reasoning benchmarks have emerged as a significant focus area, posing challenges for large foundational models like Large Language Models (LLMs) and Multi-modal Large Language Models (MLLMs). Initial datasets addressed basic algebraic (Hendrycks et al., 2021b) and arithmetic (Roy & Roth, 2016) word problems with limited scope. Subsequent efforts, including MATH (Hendrycks et al., 2021b), GSM8K (Cobbe et al., 2021), MMLU (Hendrycks et al., 2021a), and others (Zhou et al., 2023; Yue et al., 2023b; Wang et al., 2024a; Gao et al., 2023a; Luo et al., 2023), expanded the range and quality of textual mathematical problems, establishing robust benchmarks for LLM evaluation.

Despite substantial mathematical reasoning encapsulated in visual modalities, most existing benchmarks (Amini et al., 2019; Cobbe et al., 2021; Mishra et al., 2022; Frieder et al., 2023; Lu et al., 2023b) are textual only. Moreover, some datasets exhibit performance saturation, with GPT-4 achieving 92.0% accuracy on GSM-8K (Cobbe et al., 2021), a grade-school mathematics dataset. The rapid advancement of Large Multimodal Models (LMMs) necessitates robust multimodal benchmarks, as current benchmarks (Antol et al., 2015; Kembhavi et al., 2016; Kahou et al., 2017; Mathew et al., 2022) provide limited coverage of rigorous scientific domains crucial for general-purpose AI assistants.

While these benchmarks assess text-only mathematical reasoning, the rapid progress of MLLMs necessitates high-quality benchmarks for evaluating visual mathematical problem-solving. Prior

1080 attempts like GeoQA (Chen et al., 2021a), while MathVista (Lu et al., 2023a) and MMMU (Yue et al.,
1081 2023a) incorporated various multimodal tasks and college-level questions, respectively.

1082 MLLMs, building upon LLMs (Touvron et al., 2023a;b; OpenAI, 2023a; Jiang et al., 2024; Brown
1083 et al., 2020) and large vision models (Radford et al., 2021; Kirillov et al., 2023; Zhang et al.,
1084 2023d;c;e), have become increasingly prominent. They extend LLMs to diverse tasks and modalities,
1085 including 2D images (Li et al., 2022; Dai et al., 2023; Alayrac et al., 2022; Li et al., 2023a), 3D
1086 point clouds (Guo et al., 2023; Xu et al., 2023b; Hong et al., 2024), audio (Han et al., 2023; Su et al.,
1087 2023), and video (Zhang et al., 2023a; Chen et al., 2023a). Noteworthy examples like OpenAI’s GPT-
1088 4V (OpenAI, 2023c) and Google’s Gemini (Team et al., 2023) exhibit exceptional visual reasoning
1089 capabilities, setting new benchmarks in multi-modal performance.

1090 However, their closed-source nature hinders broader application and development of MLLMs. Concur-
1091 rently, open-source MLLMs like LLaMA-Adapter (Zhang et al., 2024; Gao et al., 2023b), LLaVA (Liu
1092 et al., 2023b; 2024; 2023a), MiniGPT-4 (Zhu et al., 2023a; Chen et al., 2023b), mPLUG-Owl (Ye et al.,
1093 2023b), Qwen-VL (Bai et al., 2023), InternLM-XComposer (Dong et al., 2024), and SPHINX (Lin
1094 et al., 2023; Gao et al., 2024) have been explored, leveraging CLIP (Radford et al., 2021) for image
1095 encoding and LLaMA (Touvron et al., 2023a) for multi-modal instruction tuning, advancing MLLMs’
1096 visual understanding and generalization.

1097 Despite comprehensive benchmarks (Fu et al., 2023a; Liu et al., 2023d; Li et al., 2023b; Xu et al.,
1098 2023a) for general visual instruction-following scenarios, the specific potential of MLLMs for visual
1099 mathematical problem-solving remains under-explored. Prior studies like VQA (Antol et al., 2015;
1100 Goyal et al., 2017), VizWiz (Gurari et al., 2018), and ParsVQA-Caps (Mobasher et al., 2022) evaluate
1101 LLMs’ general visual question answering abilities on open-ended image queries. Additionally, works
1102 have assessed LLMs’ specific skills beyond natural scenes, such as abstract shapes (Antol et al., 2015;
1103 Lu et al., 2021b; Ji et al., 2022), geometry diagrams (Seo et al., 2015; Lu et al., 2021a; Chen et al.,
1104 2022a; Cao & Xiao, 2022), charts (Methani et al., 2020; Masry et al., 2022; Kahou et al., 2017; Chang
1105 et al., 2022; Kafle et al., 2018), documents (Singh et al., 2019; Mathew et al., 2022; Liu et al., 2023e),
1106 synthetic images (Dahlgren Lindström & Abraham, 2022; Li et al., 2023d; Bitton-Guetta et al., 2023),
1107 external knowledge (Schwenk et al., 2022; Shah et al., 2019), commonsense reasoning (Zellers et al.,
1108 2019; Yin et al., 2021), scientific knowledge (Lu et al., 2022; Kembhavi et al., 2017; 2016), and
1109 medical understanding (Zhang et al., 2023g; Lau et al., 2018).

1110 Generative foundation models like GPT-3 (Brown et al., 2020), GPT-4 (OpenAI, 2023b), Claude (An-
1111 thropic, 2023), LLaMA (Touvron et al., 2023a), and LLaMA-Adapter (Zhang et al., 2023b) can
1112 solve various downstream tasks (Wei et al., 2022a) without task-specific fine-tuning. Prior work has
1113 evaluated their text-based abilities in QA, math, medicine, coding, and science (Bubeck et al., 2023;
1114 Nori et al., 2023; Chen et al., 2021b; Fu et al., 2023c; Sun et al., 2023; Wang et al., 2023b; Huang et al.,
1115 2023; 2022; Liu et al., 2023c; Zhang et al., 2023b). Some work focused on specialized pretraining for
1116 improved visual math and chart reasoning, like PixStruct (Lee et al., 2023), MatCha (Liu et al., 2022),
1117 and UniChart (Masry et al., 2023). On the vision-language front, models like LLaVA (Liu et al.,
1118 2023b), miniGPT4 (Zhu et al., 2023a), InstructBLIP (Dai et al., 2023), Flamingo (Alayrac et al., 2022;
1119 Awadalla et al., 2023), LLaMA-Adapter V2 (Gao et al., 2023b), and Multimodal Bard (Google, 2023)
1120 leverage paired (Schuhmann et al., 2022; Sharma et al., 2018; Lin et al., 2014) and interleaved (Zhu
1121 et al., 2023b) image-text data. Additionally, specialized versions like LLaVAR (Zhang et al., 2023h;
1122 Ye et al., 2023a) emphasize document understanding and math comprehension. Recent works like
1123 Visit-Bench (Bitton et al., 2023), LVLM-eHub (Yu et al., 2023), MMBench (Liu et al., 2023d; Xu
1124 et al., 2023a; Shao et al., 2023) assess these models’ instruction-following and reasoning capabilities.

1124 Large language models (LLMs) have demonstrated remarkable reasoning abilities, further enhanced
1125 by approaches like chain-of-thought (CoT) (Wei et al., 2022b), program-of-thought (PoT) (Chen
1126 et al., 2022b), and inductive reasoning (Wang et al., 2023a; Tan & Motani, 2023). The feasibility of
1127 using LLMs to solve the Abstraction and Reasoning Corpus (ARC) challenge has been verified using
1128 zero-shot, few-shot, and context-grounded prompting (Tan & Motani, 2023).

1129 OpenAI’s GPT-4V, the multimodal version of GPT-4, exhibits promising performance in vision-
1130 language reasoning. However, a fine-grained study of its strengths and limitations is still lacking.
1131 Recent work (Zhang et al., 2023f) explores whether large multimodal models (LMMs) like GPT-4V
1132 execute vision and language tasks consistently or independently, contributing pioneering efforts in
1133 this field.

C DATA COLLECTION PIPELINE DETAILS

Collection Pipeline: To ensure high-quality samples, all data samples were manually collected as image snippets from publicly available websites.

We developed a flexible, highly automated data curation framework to streamline the process and standardize collection and annotation. Continuous human reviews were conducted between steps in the pipeline to maintain quality and prevent error propagation.

- Step 1: A universally unique identifier (UUID) was generated for each question paper to track all curated questions. This step also updated a shared record containing details of the paper and the annotator’s alias, enabling efficient assignment of questions for peer review.
- Step 2: Annotators manually collected individual snippets of each question, along with contextual information relevant to multiple questions. For questions requiring additional context, snippets were labeled accordingly, and only legible, relevant questions (focused on Mental Ability or Scholastic Ability in mathematics) were included to maintain dataset integrity.
- Step 3: An image-merging script automatically identified and merged split question images or context snippets (based on the naming convention) using open-source image processing tools³. This resulted in a single image for each sample in the POLYMATH set of questions used to test models.
- Step 4: The next module in the pipeline created and automatically populated an annotation file, where each row corresponded to a collected sample. Columns included the paper_id (UUID from Step 1), question number, and image path.
- Step 5: Using an answer key or solution set, LLM-powered transcription extracted the ground truth answers for each question. Extracted answers were mapped to the corresponding annotation rows, followed by a manual check to ensure alignment with the provided solution and correctness.

D ADDITIONAL EXPERIMENT DETAILS

Hyperparameters: The following hyperparameters were used in our experiments:

Model	Hyperparameters
Gemini-1.5 Pro	temperature: 1, top_p: 0.95, top_k: 64, max_output_tokens: 8192, response_mime_type: text/plain
GPT-4o	top_p: 0.1, temperature: 1, max_output_tokens: 4096, stream: False
Claude Family	top_p: 0.1, temperature: 1, max_output_tokens: 4096, stream: False
Open Source Models	max_new_tokens: 3600, temperature: 0.7, top_p: 0.3, num_beams: 1

Table 8: Hyperparameters used in the experiments

Further, Table 9 provides the source repositories and model cards for the various models used in our experiments. Table 10 shows the performance of open-source models across categories using two additional prompting strategies: *Chain-of-Thought* and *Step-back*. Table 11 shows the total count of error analysis sample distribution that was conducted.

³<https://opencv.org/>

Model	Release Time	Source
GPT-4o OpenAI (2024a)	2023-03	https://platform.openai.com/
Claude 3 family Anthropic (2024a;b)	2023-03	https://www.anthropic.com/news/claude-3-family
Gemini-1.5 Pro Team et al. (2023)	2023-12	https://ai.google.dev/
LLaVA-1.5 Liu et al. (2023a)	2023-10	https://huggingface.co/liuhaotian/llava-v1.5-13b
G-LLaVA Gao et al. (2023a)	2023-12	https://github.com/pipilurj/G-LLaVA/tree/main
ShareGPT4V Chen et al. (2023c)	2023-11	https://github.com/ShareGPT4Omni/ShareGPT4V/blob/master/docs/ModelZoo.md#sharegpt4v-models
LLaVA-NeXT Liu et al. (2024)	2024-01	https://github.com/LLaVA-VL/LLaVA-NeXT
Qwen2-VL Wang et al. (2024b)	2024-01	https://huggingface.co/Qwen/Qwen2-VL-2B-Instruct

Table 9: Models used to evaluated POLYMATH, along with their release dates and source repositories. We use both open-source and closed-source models for a comprehensive evaluation.

Category	PS	FC	PR	SC	RR	MR	NR	SR	OOO	LR	Overall
<i>Chain of Thought Inference</i>											
Qwen2 VL 2B Instruct	12.90	2.13	6.61	0.89	9.52	3.57	6.82	5.75	10.13	4.55	5.70
Llava v1.6 Mistral 7B	12.90	8.51	15.86	15.18	20.00	15.63	11.36	21.84	25.32	15.91	16.80
G-LLaVA 7B	16.13	0.00	9.69	4.46	5.71	8.04	4.55	5.75	3.80	9.09	7.00
ShareGPT4V 7B	9.68	19.15	16.74	14.29	8.57	12.05	13.64	12.64	8.86	13.64	13.20
Llava v1.6 Vicuna 13B	16.13	17.02	9.25	9.82	14.29	6.25	18.18	9.20	15.19	9.09	10.60
Llava v1.5 13B	6.45	17.02	8.37	12.50	8.57	7.14	11.36	9.20	12.66	15.91	9.80
ShareGPT4V 13B	12.90	19.15	14.10	13.39	16.19	11.61	11.36	14.94	18.99	11.36	14.10
G-LLaVA 13B	16.13	2.13	11.45	6.25	8.57	10.27	2.27	6.90	6.33	9.09	8.70
Llava v1.6 34B	12.90	25.53	10.13	0.89	7.62	10.71	15.91	10.34	16.46	9.09	10.5
<i>Step Back Inference</i>											
Qwen2 VL 2B Instruct	16.13	4.26	7.05	1.79	10.48	4.02	9.09	6.90	11.39	6.82	6.70
Llava v1.6 Mistral 7b	16.13	6.38	16.74	14.29	20.95	14.29	13.64	21.84	26.58	18.18	17.00
G-LLaVA 7B	12.90	0.00	9.25	3.57	5.71	7.59	2.27	4.60	3.80	6.82	7.30
ShareGPT4V 7B	16.13	23.40	16.30	15.18	10.48	11.61	15.91	10.34	6.33	15.91	13.50
Llava v1.6 Vicuna 13B	19.35	14.89	10.13	8.04	13.33	6.70	20.45	10.34	16.46	11.36	11.00
Llava 1.5 13B	12.90	14.89	8.37	13.39	7.62	7.59	13.64	8.05	13.92	20.45	10.30
ShareGPT4V 13B	9.68	17.02	13.66	15.18	18.10	12.05	13.64	12.64	17.72	15.91	14.30
G-LLaVA 13B	19.35	4.26	11.89	7.14	9.52	10.71	4.55	8.05	7.59	11.36	9.70
Llava v1.6 34B	16.13	27.66	10.57	1.79	8.57	11.16	18.18	11.49	17.72	11.36	11.50

Table 10: Results of open-source MLLMs on the *testmini* split of POLYMATH. We report model results using Chain-of-Thought, and Step Back prompting methods.

E PROMPTS FOR DATASET CURATION AND EXPERIMENTS

The various prompts are detailed in this section. Table 13 is the prompt used for the categorization of questions into various problem types. Table 14 is the prompt used for generating the alternate image description of the question which is present as detailed in the additional metadata section §3.3. Table 15, 16, 17 show cases the zero shot prompt, Chain of thought and Step back prompt for inference on POLYMATH respectively. Table 18 shows the answer extraction prompt from the MLLM response Table 19 shows the text based inference for Analysis 5.

Error Type	PS	FC	PR	SC	RR	MR	NR	SR	OD	LR	Overall
<i>Gemini-1.5 Pro</i>											
Calculation Error (CE)	1	0	0	0	0	5	1	0	0	0	7
Incomplete (IC)	1	0	0	4	5	4	1	0	0	0	15
Logical Flaw (LF)	3	5	24	24	10	16	0	20	22	13	137
Memory Flaw (MF)	0	2	6	0	10	1	4	5	0	0	28
Misalignment (MG)	3	0	0	4	0	0	0	0	4	0	11
Spatial Misunderstanding (SM)	6	10	0	0	5	4	4	5	4	0	38
Overall Errors	14	17	30	32	30	30	10	30	30	13	236
<i>GPT-4o</i>											
Calculation Error (CE)	1	0	0	0	0	1	1	0	0	0	3
Incomplete (IC)	0	3	0	4	0	4	1	0	0	0	12
Logical Flaw (LF)	1	7	24	20	15	8	0	15	26	8	124
Memory Flaw (MF)	0	0	6	0	5	8	4	0	0	0	23
Misalignment (MG)	6	0	0	4	0	1	0	0	0	5	16
Spatial Misunderstanding (SM)	6	7	0	4	10	8	4	15	4	0	58
Overall Errors	14	17	30	32	30	30	10	30	30	13	236
<i>Claude-3.5 Sonnet</i>											
Calculation Error (CE)	1	0	0	0	0	12	1	0	1	0	15
Incomplete (IC)	0	0	0	0	0	1	0	0	0	0	1
Logical Flaw (LF)	3	10	24	20	10	12	1	20	25	10	135
Memory Flaw (MF)	1	0	6	0	10	1	4	5	0	0	27
Misalignment (MG)	6	2	0	8	0	0	0	0	0	3	19
Spatial Misunderstanding (SM)	3	5	0	4	10	4	4	5	4	0	39
Overall Errors	14	17	30	32	30	30	10	30	30	13	236

Table 11: Type of errors made by Gemini-1.5 Pro, GPT4-o, and Claude-3.5 Sonnet over various question categories.

Category	PS	FC	PR	SC	RR	MR	NR	SR	OOO	LR	Overall
Human 1	45.16	80.85	52.86	69.64	74.29	67.86	52.27	60.92	72.15	40.91	63.10
Human 2	41.94	53.19	45.81	80.36	84.76	85.71	75.00	77.01	75.95	40.91	69.10
Human 3	67.74	63.83	86.78	54.46	61.90	80.80	72.73	44.83	79.75	40.91	70.70
Human 4	64.52	78.72	85.90	47.32	43.81	80.80	47.73	68.97	56.96	56.82	68.30
Human 5	45.16	87.23	45.81	79.46	80.00	75.00	54.55	60.92	51.90	75.00	65.10
Human 6	41.94	59.57	53.74	84.82	74.29	69.64	50.00	63.22	53.16	52.27	63.40

Table 12: Per-category accuracy scores achieved by six human evaluators. The average human accuracy over all categories is 66.62%.

F DATASET EXAMPLES

Figures 5, 6, 7, 8, 9, 10, 11, 12, 13, 14 demonstrate examples from each question category defined in Table 1.

G MORE DETAILS ON ERROR ANALYSIS

We leveraged 2 authors of this work to act as error evaluators independently and in parallel. Each evaluator has a graduate degree in Computer Science and experience in similar puzzle-solving. Owing to the clear and mutually-exclusive definitions of error types, there is little ambiguity in identifying the error type of the incorrect responses. Our measure of inter-evaluator agreement is Cohen’s Kappa (K), found to be 0.9 - indicating near-unanimous agreement. For questions where there was disagreement in evaluations, a consensus was reached after discussion.

1296	<p>You are given a question designed to test a student on mathematical or logical reasoning. These questions can be categorized based on the skills and techniques used to solve them. These are the categories of questions.</p> <p>Mathematical reasoning: this question purely requires calculations of a mathematical nature. This includes solving a straightforward equation.</p> <p>Pattern recognition: 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.</p> <p>Sequence completion: given a sequence of numbers or figures, this question involves finding the sequentially next element in a series.</p> <p>Figure completion: You are given a figure 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.</p> <p>Odd one out: given a set of elements, identify the element that is not like the others.</p> <p>Spatial reasoning: questions involving reasoning observationally and visualizing the question in order to arrive at the answer.</p> <p>Perspective shift: Questions where a figure is given and you are instructed to morph it according to the instructions (flip, mirror image, rotate, etc)</p> <p>Numerical reasoning: questions involving counting the number of elements mentioned. The elements may be part of a single figure or conform to a specified pattern, but solving these questions requires counting.</p> <p>Relative reasoning: the question contains distinct data points, and solving the questions requires understanding the relationships between all data points and extrapolating relationships that are not explicitly mentioned. Questions involving venn diagrams, family relations, or relative positions given a reference point fall into this category.</p> <p>Logical reasoning: Questions involving simple logical reasoning such as entailment and contradiction.</p> <p>Now, observe the following question.</p> <p>Using the categorization schema explained above, classify this question into a category. Provide a detailed explanation. Output a JSON with the key "question" containing a transcript of the question, "category" containing the classification category, and "explanation" containing the reasoning for assigning the question to this category, and "contains diagram" which should be True or False depending on whether there is a diagram provided in the question.</p>
1297	
1298	
1299	
1300	
1301	
1302	
1303	
1304	
1305	
1306	
1307	
1308	
1309	
1310	
1311	
1312	
1313	
1314	
1315	
1316	
1317	
1318	
1319	
1320	
1321	
1322	
1323	
1324	
1325	
1326	
1327	
1328	
1329	
1330	
1331	
1332	
1333	
1334	
1335	
1336	
1337	
1338	
1339	

Table 13: Prompt used for categorization of question of image.

1340
1341
1342
1343
1344
1345
1346
1347
1348
1349

H QUALITATIVE ERROR ANALYSIS

This section presents examples of the qualitative error analysis that was carried out. Figures 5, 6, 7, 8, 9, 10, 11, 12, 13 and 14 contains examples of failures by three proprietary models viz. Gemini-1.5 Pro, GPT-4o, and Claude-3.5 Sonnet across all categories.

1350
1351
1352
1353
1354
1355
1356
1357
1358
1359
1360
1361
1362
1363
1364
1365
1366
1367
1368
1369
1370
1371
1372
1373
1374
1375
1376

You are given a mathematical question involving a diagram. You are an accessibility reader for the blind. Output a detailed text description describing the diagram.

Example description: "description": "The diagram contains a circle, triangle, and rectangle overlapping. The circle is the topmost figure, the triangle is figure with the lowest base. The rectangle top cuts through the circle and triangle, while its lower side only passes through the triangle. The portion of the circle that does not overlap with any other figure contains the number 10. The intersection between circle and triangle contains the number 12. The intersection of only the circle and rectangle contains the number 5. The area where all 3 figures intersect contains 20. The area of the rectangle that interacts with no other figure contains 14. The area of the intersection between only the rectangle and triangle contains 17. Finally, the area of the triangle does not intersect with any other figures contains the number 16. Outside these figures are text labels and arrows. The arrow labeled Teacher points to the circle. The arrow labeled Doctor points to the rectangle. The arrow labeled Musician points to the triangle."

Now, generate a similarly comprehensive text description for the diagram in this question.

Image:image

Remember, the description must be detailed enough that the user can recreate the diagram exactly as shown based on the description alone. Do not add any information or make assumptions that are not explicitly mentioned in the image.

Output a JSON with the key "description" whose value is the generated description. Output only the JSON. Go!

Table 14: Prompt used to generated detailed textual description of diagrams.

1377
1378
1379
1380
1381
1382
1383
1384
1385
1386
1387
1388
1389
1390
1391
1392
1393
1394
1395
1396
1397
1398
1399
1400
1401
1402
1403

Common Prefix: "You are given a question to solve below:

This question requires skills and reasoning related to category. Definition: category definition.

This question has a list of options : answer range.

Your output must be a valid JSON."

Zeroshot Prompt: "Q1: Provide a step by step solution to this question.

Q2: What is the answer to this question? Remember, the answer must be present in the given list of answer options

Q3: Which is the option from answer range that corresponds to the answer above? Output only the option and nothing else.

Output a JSON with the keys Q1, Q2, Q3 with their answers."

Common postfix: "Remember, your output must be a valid JSON in this format: 'Q1':<answer>,'Q2':<answer>,'Q3':<answer> If your JSON is incomplete, incorrectly delimited or badly formatted, you will be destroyed. Output the valid JSON and nothing else. Go!"

Table 15: Prompt for zero shot inference

1404
1405
1406
1407
1408
1409
1410
1411
1412
1413
1414
1415
1416
1417
1418
1419
1420
1421
1422
1423
1424
1425
1426
1427
1428
1429
1430
1431
1432
1433
1434
1435
1436
1437
1438
1439
1440
1441
1442
1443
1444
1445
1446
1447
1448
1449
1450
1451
1452
1453
1454
1455
1456
1457

<p>Common Prefix: "You are given a question to solve below: This question requires skills and reasoning related to category. Definition: category definition. This question has a list of options : answer range. Your output must be a valid JSON."</p> <p>CoT Prompt: Now answer the following questions. Q1: What is the list of variables and their values provided in the questions? Q2: What is the variable that needs to be solved for? Q3: What information that is not present in the question, can you infer from the given variables? Q4: Provide a step-by-step solution with reasoning to obtain the answer to this question. Provide the solution at each step. Q5: What is the answer to this question? Remember, the answer must be present in the given list of answer options. Q6: Which is the option from answer range that corresponds to the answer above? Output only the option and nothing else.</p> <p>Output a JSON with the keys Q1, Q2, Q3, Q4, Q5, Q6 with their answers.</p> <p>Common postfix: "Remember, your output must be a valid JSON in this format: 'Q1':<answer>,'Q2':<answer>,'Q3':<answer> If your JSON is incomplete, incorrectly delimited or badly formatted, you will be destroyed. Output the valid JSON and nothing else. Go!"</p>

Table 16: Prompt for Chain-of-Thought inference

1458	
1459	Common Prefix: "You are given a question to solve below:
1460	This question requires skills and reasoning related to category. Definition: category definition.
1461	This question has a list of options : answer range.
1462	Your output must be a valid JSON."
1463	
1464	Step back category prompt:
1465	
1466	Mathematical Reasoning: "Q1: What is the relation of all given variables to one another? How is
1467	each variable related to the missing value?
1468	Q2: Which are the mathematical operations involved in solving a question like this?"
1469	
1470	Pattern Recognition: "Q1: What is the pattern being followed in this question? Provide an example.
1471	Q2: Which are the elements in this question that follow this pattern?"
1472	
1473	Sequence Completion: "Q1: What is a numerical sequence?
1474	Q2: What is the relationship between previous and subsequent elements in a sequence? What is the
1475	relationship between elements in the sequence present in this question?"
1476	
1477	Figure Completion: "Q1: How do you approach a figure completion problem?
1478	Q2: What is the information you have and the missing information? What are their spatial
1479	relationships to one another?"
1480	
1481	Odd one out: "Q1: How do you identify an odd element out of a set?
1482	Q2: Describe the elements in this set. Now ,what do almost all of these elements have in common?"
1483	
1484	Spatial Reasoning: "Q1: What are the spatial manipulations that occur in this question? Eg.
1485	unfolding, folding, 2D to 3D reconstruction, etc.
1486	Q2: Given the original question image, how can you visualize the resulting image after the
1487	manipulations mentioned in the question? Explain in detail."
1488	
1489	Perspective Shift: "Q1: What are the attributes of an image that is flipped, rotated, or its mirror
1490	image? What differentiates the result of these operations from the original image?
1491	Q2: Which of these operations apply in this image, and in what order?"
1492	
1493	Numerical Reasoning: "Q1: What is the information you are given? What do you need to find out?
1494	How can you arrive at this number?
1495	Q2: What are the main points of concern in solving such a question? How can you ensure that you
1496	do not under or over estimate the final number?"
1497	
1498	Relative Reasoning: "Q1: What is the information you are given? What are the relationships of the
1499	given data points to one another? What is the information you need to discover? Which data points
1500	are directly or indirectly related to the missing variable? Explain in detail.
1501	Q2: What principles of relational logic do you need to apply to this question?"
1502	
1503	Logical Reasoning: "Q1: what are the principle of logical reasoning involved in solving this
1504	question?
1505	Q2: What is the information provided in this question? What is the objective of this question?"
1506	
1507	Meta Prompt: Step back category prompt
1508	Q3: Based on the above information, provide a step-by-step solution to the question in the image.
1509	Q4: What is the answer to this question? Remember, the answer must be present in the given list of
1510	answer options
1511	Q5: Which is the option from answer range that corresponds to the answer above? Output only the
	option and nothing else.
	Output a JSON with the keys Q1, Q2, Q3, Q4, Q5 with their answers.

Table 17: Per-category and meta-prompts for Step Back prompt inference

1512
1513
1514
1515
1516
1517
1518
1519
1520
1521
1522
1523
1524
1525
1526
1527
1528
1529
1530
1531
1532
1533
1534
1535
1536
1537
1538
1539
1540
1541
1542
1543
1544
1545
1546
1547
1548
1549
1550
1551
1552
1553
1554
1555
1556
1557
1558
1559
1560
1561
1562
1563
1564
1565

You are given a mathematical question with a list of multiple choice answers. You are an accessibility reader for the blind. Transcribe the textual part of the question, and the list of answer options provided.
Example: 'question': 'How many triangles are present in this diagram?', 'answer list': '(A) 23 (B) 21 (C) 29 (D) 34'
Now, generate a question and answer list transcript for the question in the image.
Output a JSON with the keys "question" and "answer list" as described. Output only the JSON. Go!

Table 18: Prompt to transcribe list of answer options from question image

You are given a question to solve below:

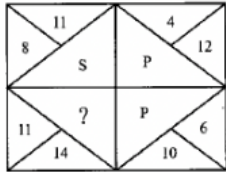
This question requires skills and reasoning related to category. This question contains a diagram that is crucial to solving the question whose textual description as been provided. Definition: category definition. Problem: extracted question. Diagram: image description extracted answer list
Q1: Provide a step by step solution to this question.
Q2: What is the answer to this question? Remember, the answer must be present in the given list of answer options
Q3: Which is the option from answer range that corresponds to the answer above? Output only the option and nothing else.
Output a JSON with the keys Q1, Q2, Q3 with their answers.
Remember, your output must be a valid JSON in this format: 'Q1': <answer>, 'Q2': <answer>, 'Q3': <answer> If your JSON is incomplete, incorrectly delimited or badly formatted, you will be destroyed. Output the valid JSON and nothing else. Go!

Table 19: Prompt for text-only inference.

1566
1567
1568
1569
1570
1571
1572
1573
1574
1575
1576
1577
1578
1579
1580
1581
1582
1583
1584
1585
1586
1587
1588
1589
1590
1591
1592
1593
1594
1595
1596
1597
1598
1599
1600
1601
1602
1603
1604
1605
1606
1607
1608
1609
1610
1611
1612
1613
1614
1615
1616
1617
1618
1619

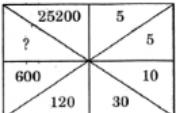
Figure Completion

Direction (Q. No. 54 to Q. No. 58) :
Choose the missing number (?) from the given alternatives.

57. 

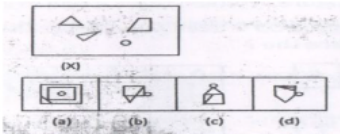
(A) T (B) S
(C) U (D) Y

Direction : Find the missing number in each of the following question from 62 to 66 and choose the correct options.

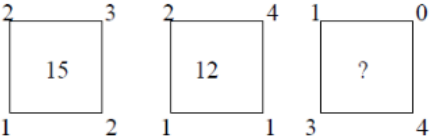


(A) 2400 (B) 3000 (C) 4200 (D) 3600

Directions (81–85) In each of the following questions out which of the figures (a), (b), (c), (d) can be formed the pieces of given figure X.

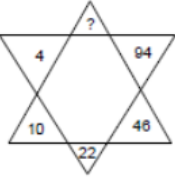
84. 

Question 31 to 36 numbers are placed in figure on the basis of some rules. One place is vacant which is indicated as (?). figer out the correct alternative for the vacant place and write its number against the proper question number on your answer sheet-



(1) 7 (2) 13 (3) 1 (4) 8

DIRECTION: In each questions 41–50, numbers are placed in figures on the basis of some rules. One place in the figure is indicated by the interrogation sign (?). Find out the correct alternative to replace the question mark and indicate your answer by filling the circle of the corresponding letter of alternatives in the O.M.R. Answer-Sheet.



(A) 194 (B) 188
(C) 190 (D) 192

Figure 5: Questions belonging to the *figure_completion* (FC) category

1620
1621
1622
1623
1624
1625
1626
1627
1628
1629
1630
1631
1632
1633
1634
1635
1636
1637
1638
1639
1640
1641
1642
1643
1644
1645
1646
1647
1648
1649
1650
1651
1652
1653
1654
1655
1656
1657
1658
1659
1660
1661
1662
1663
1664
1665
1666
1667
1668
1669
1670
1671
1672
1673

Logical Reasoning

If the fish were birds, what would be the sea be?
 (1) Forest (2) Sky (3) Nest (4) Island.

Questions 38-40 : Three words in bold letters are given in each questions, which have something in common among themselves. Out of the four given alternatives, choose the most appropriate description about these three words

39. Newspaper : Hoarding : Television
 (1) Press (2) Media (3) Broadcast (4) Rumour

“Cards marked with numbers 13, 14, 15,.....,60 are placed in a box and mixed thoroughly. One card is drawn at random from the box.” Read the information carefully and match the following.

i) The probability of the number that is on the card drawn is divisible by 5.	p) $\frac{1}{4}$
ii) The probability of the number that is one the card drawn is a prime.	q) $\frac{36}{48}$
iii) The probability of the number that is on the card drawn is a multiple of 19.	r) $\frac{5}{24}$
iv) The probability of the number that is on the card drawn is a composite number.	s) $\frac{1}{16}$

1) $p \rightarrow iv, q \rightarrow iii, r \rightarrow ii, s \rightarrow i$ 2) $p \rightarrow iii, q \rightarrow ii, r \rightarrow iv, s \rightarrow i$
 3) $p \rightarrow i, q \rightarrow ii, r \rightarrow iii, s \rightarrow iv$ 4) $p \rightarrow ii, q \rightarrow iv, r \rightarrow i, s \rightarrow iii$

Questions 74-77 : In each of the questions given below, there are two statements labelled as **Assertions (A)** and **Reason (R)**. Mark your answer as per the options provided below the question.

75. **Assertion (A)** :
 Vaccines prevents disease.
Reason (R)
 Vaccine must be given to children.
 (1) Both (A) and (R) are true and (R) is the correct explanation of (A)
 (2) Both (A) and (R) are true but (R) is not the correct explanation of (A)
 (3) (A) is true but (R) is false
 (4) (A) is false but (R) is true

10 November, 1981 was Tuesday. What was the day on 11 November, 1581 ?
 (A) Tuesday (B) Wednesday
 (C) Friday (D) Saturday

Figure 6: Questions belonging to the *logical_reasoning* (LR) category

1674
1675
1676
1677
1678
1679
1680
1681
1682
1683
1684
1685
1686
1687
1688
1689
1690
1691
1692
1693
1694
1695
1696
1697
1698
1699
1700
1701
1702
1703
1704
1705
1706
1707
1708
1709
1710
1711
1712
1713
1714
1715
1716
1717
1718
1719
1720
1721
1722
1723
1724
1725
1726
1727

Mathematical Reasoning

If $x = 2 + 2^{1/3} + 2^{2/3}$ then $x^3 - 6x^2 + 6x = \dots\dots$

(A) 2 (B) 1
(C) 4 (D) 3

What is the co-efficient of $(x+y)^2$ in the expansion of x^2y^2 ?

(a) 3 (b) 4 (c) 5 (d) 6

If ΔABC is an equilateral triangle such that $AD \perp BC$, then $AD^2 =$

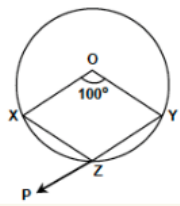
A) $\frac{3a^2}{4}$ B) $\frac{3a^2}{2}$ C) $\frac{3}{4}BC^2$ D) $\frac{\sqrt{3}}{2}a$

1) A and C 2) D 3) A 4) B and C

In a ΔABC , $\angle C = 90^\circ$. On the sides CA and CB two points P and Q are taken such that they divide CA and CB in the ratio 2:1 respectively. Then, $(Aa^2 + BP^2) : AB^2 = \dots\dots$

(1) $\frac{7}{9}$
(2) $\frac{4}{9}$
(3) $\frac{13}{9}$
(4) $\frac{11}{9}$

O is the centre of a circle and $\angle xoy = 100^\circ$. Find the measure of $\angle xzp$



(1) 50° (2) 100°
(3) 150° (4) 80°

The correct relation is

A B

i. a, b, c are in G.P. a. $2b = a + c$

ii. a, b, c are in A.P. b. $a + c = \frac{2ac}{b}$

iii. a, b, c are in H.P. c. $b^{1/2} = ca$

d. $b = (ca)^{1/2}$

1) i - c, ii - b, iii - a
2) i - c, ii - a, iii - d
3) i - d, ii - a, iii - b
4) i - d, ii - b, iii - c

Figure 7: Questions belonging to the *mathematical_reasoning* (MR) category


1728
1729
1730
1731
1732
1733
1734
1735
1736
1737
1738
1739
1740
1741
1742
1743
1744
1745
1746
1747
1748
1749
1750
1751
1752
1753
1754
1755
1756
1757
1758
1759
1760
1761
1762
1763
1764
1765
1766
1767
1768
1769
1770
1771
1772
1773
1774
1775
1776
1777
1778
1779
1780
1781

Numerical Reasoning

Instruction: In question no. 11 to 20. Figure are given with question mark (?). Complete the figure replacing question mark (?) with suitable number logically.

How many triangles are there in the given figure?

(A)28 (B)24 (C)14 (D)10



In the following series of numbers, how many times 1, 3 and 7 have appeared together, 7 being between 1 and 3.

297317377331738671377173906

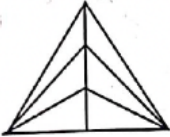
(A) 3
(B) 4
(C) 5
(D) More than 5

In the following sequence of numbers, how many consecutive even numbers have a difference of 2?

444864486422144228281

(A) 9 (B) 8 (C) 7 (D) 6

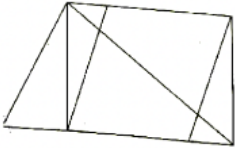
How many triangles are there in the given figure.



1. 15
3. 16

2. 14
4. 20

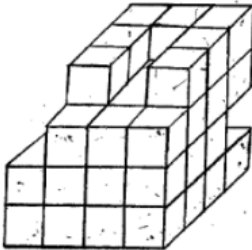
How many quadrilaterals are there in the given figure?



(1) 10
(3) 12

(2) 11
(4) 13

Q. 45 and 46 Directions: In the following figure small cubes are arranged in a particular manner as shown. Observe the arrangement and answer the following questions.



46. What is the total number of blocks whose three 'surfaces are seen ?

(1) 12 (2) 13
(3) 14 (4) 15

Figure 8: Questions belonging to the *numerical_reasoning* (NR) category

1782
1783
1784
1785
1786
1787
1788
1789
1790
1791
1792
1793
1794
1795
1796
1797
1798
1799
1800
1801
1802
1803
1804
1805
1806
1807
1808
1809
1810
1811
1812
1813
1814
1815
1816
1817
1818
1819
1820
1821
1822
1823
1824
1825
1826
1827
1828
1829
1830
1831
1832
1833
1834
1835

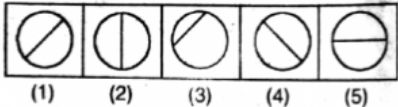
Odd One Out

Directions (1–10): In the following questions, four items (numbers, number pairs/letter groups) are given. Three of them are alike in a certain way and one is different. Find the odd one out from the alternative.

3. (A) 11 (B) 15
(C) 17 (D) 19

Choose the odd one out from following.
(A) Tomato (B) Mango (C) Banana (D) Apple

Directions (81 – 90): Out of the five figures (1), (2), (3), (4) and (5) given in each problem, four are similar in a certain way. Choose the figure which is different from the other figures.




(1) (2) (3) (4) (5)

Direction: In question No. 1 to 10 each question has four Terms. Three terms are alike in some way. One term is different from three others. Find out the correct term which is different from three others and write its alternative number on your answer sheet against the proper question number –

(1) Q 144 (2) M 54
(3) U 16 (4) N 60

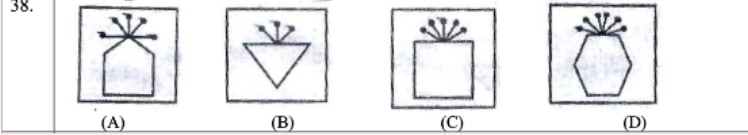
Questions 86 – 90: In each of the following sets of figures, select the one figure that is different from the other figures from the given options.



(1) (2) (3) (4)

Instruction: (Q. No 31 to 40) Four figures are given in question no. 31 to 40. One of the figures differ from the rest. Find out the figure which is different from the rest of the figures.

38.



(A) (B) (C) (D)

Find the odd term.

(1) ABDEF (2) JKMNX
(3) GHJKR (4) IJLMT

Figure 9: Questions belonging to the *odd_one_out* (OD) category

1836
1837
1838
1839
1840
1841
1842
1843
1844
1845
1846
1847
1848
1849
1850
1851
1852
1853
1854
1855
1856
1857
1858
1859
1860
1861
1862
1863
1864
1865
1866
1867
1868
1869
1870
1871
1872
1873
1874
1875
1876
1877
1878
1879
1880
1881
1882
1883
1884
1885
1886
1887
1888
1889

Pattern Recognition

The following questions consist of two words each that have a certain relationship with each other. Select the pair that has the same relationship as the original pair of words

63. Cream : Cosmetics
 (A) Mountain : Valley (B) Tiger : Forest
 (C) Magazine : Editor (D) Teak : Wood

Find the odd term.
 (1) ABDEF (2) JKMNX
 (3) GHJKR (4) IJLMT

Directions : Complete the given number / letter / figure analogy by choosing the correct answer from the given alternatives.

14. 441 : 7 : : 576 : ?
 1) 6
 2) 8
 3) 12
 4) 14

CIRCLE is related to RICELC in the same way as SQUARE is related to
 (A) QSUERA (B) QUSERA
 (C) UQSAER (D) UQSERA

Q. 9 to 12. → **Directions:** - In each of the following questions there is a specific relationship between the first and second term. The same relationship exists between the third and fourth term which will replace the question mark (?). Select the correct term from the alternatives given.

AXD : EWB :: ? : JRG
 (1) ETH (2) FSI (3) HRK (4) FRJ

Q. 30 and 31 → **Directions :-** Replace the question mark by choosing the correct alternatives from given below

<table style="width: 100%; text-align: center;"> <tr><td>23</td></tr> <tr><td>39</td></tr> <tr><td>53</td></tr> <tr><td>70</td></tr> <tr><td>45</td></tr> </table>	23	39	53	70	45	<table style="width: 100%; text-align: center;"> <tr><td>91</td></tr> <tr><td>58</td></tr> <tr><td>70</td></tr> <tr><td>47</td></tr> <tr><td>32</td></tr> </table>	91	58	70	47	32	<table style="width: 100%; text-align: center;"> <tr><td>23</td></tr> <tr><td>17</td></tr> <tr><td>?</td></tr> <tr><td>65</td></tr> <tr><td>41</td></tr> </table>	23	17	?	65	41
23																	
39																	
53																	
70																	
45																	
91																	
58																	
70																	
47																	
32																	
23																	
17																	
?																	
65																	
41																	
(1) 61	(2) 63	(3) 66	(4) 68														

In these questions, numbers are arranged on the basis of some rules. One place is vacant, which is indicated as "?". Find out the correct alternatives to replace the question mark "?"

53.

11		
9	54	18
16		

7		
14	48	13
14		

14		
12	36	?
7		


a) 11 b) 3 c) 4 d) 9

Figure 10: Questions belonging to the *pattern_recognition* (PR) category


1890
1891
1892
1893
1894
1895
1896
1897
1898
1899
1900
1901
1902
1903
1904
1905
1906
1907
1908
1909
1910
1911
1912
1913
1914
1915
1916
1917
1918
1919
1920
1921
1922
1923
1924
1925
1926
1927
1928
1929
1930
1931
1932
1933
1934
1935
1936
1937
1938
1939
1940
1941
1942
1943

Perspective Shift


Choose the correct water image from the given alternatives for the given question figure




(1)




(2)








(3)




(4)



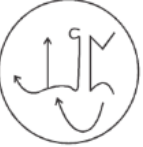
Direction : In questions 39 to 40, find the correct mirror image of the given figures, when mirror is placed on right side of the figure.

39. 






Find the correct water image for the following problem figure choosing from the four options:




(1)



(2)



(3)



(4)




Figure 11: Questions belonging to the *perspective_shift* (PS) category

1998
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
2020
2021
2022
2023
2024
2025
2026
2027
2028
2029
2030
2031
2032
2033
2034
2035
2036
2037
2038
2039
2040
2041
2042
2043
2044
2045
2046
2047
2048
2049
2050
2051

Sequence Completion

Direction : Questions (46 to 50) : Each of the following questions consists of the five figures marked A, B, C, D and E called the problem figures followed by five alternatives marked 1, 2, 3, 4, and 5 called the answer figures. Select a figure which will continue the same series established by the five problem figures.

48. Problem Figures **Answer Figures**

40 In each of the following questions a group of three pictures are gives you have to find out the fourth picture. You have to select the answer from 1, 2, 3 or 4

Choose the missing number -

34, 7, 37, 14, 40, 28, 43, ...?

(1) 56 (2) 63
(3) 42 (4) 49

3, 6, 24, 30, 63, 72, ?, 132

(1) 120 (2) 110
(3) 105 (4) 115

Directions : Complete the following number/figural series by choosing the correct answer from the given alternatives.

In each of these questions, the four problem figures in each row make a series. Find out the one which would come next in the series from among the answer figures given.

Problem Figures :

Answer Figures :

Figure 13: Questions belonging to the *sequence_completion* (SC) category

2106
2107
2108
2109
2110
2111
2112
2113
2114
2115
2116
2117
2118
2119
2120
2121
2122
2123
2124
2125
2126
2127
2128
2129
2130
2131
2132
2133
2134
2135
2136
2137
2138
2139
2140
2141
2142
2143
2144
2145
2146
2147
2148
2149
2150
2151
2152
2153
2154
2155
2156
2157
2158
2159

Figure Completion

Direction : From question Nos. 91 to 95, complete the figures by choosing appropriate figure from **answer figures**.

Correct Answer: (C)

Model: Gemini 1.5 Pro
Reasoning Steps: The sequence of figures shows that the number of black circles is increasing from the first figure to the third figure. The first figure has 1 black circle, the second figure has 2 black circles, and the third figure has 3 black circles. Therefore, the missing figure should have 4 black circles..
Final Answer: A circle with 4 black circles on it
Final Response: (D)

Model: GPT 4o
Reasoning Steps: Step 1: Observe the given figures in the top row. Each figure has a pattern of black and white segments. Step 2: Notice the pattern in the first row: the black segments are rotating clockwise. Step 3: Apply the same pattern to the second row. The first figure in the second row has black segments at the top and bottom. Step 4: The second figure in the second row should have black segments rotated 90 degrees clockwise from the first figure in the second row. Step 5: The correct figure should have black segments on the left and right sides.
Final Answer: The figure with black segments on the left and right sides.
Final Response: (B)

Model: Claude 3.5 Sonnet
Reasoning Steps: 1. Observe the pattern in the given figures:
 - The top row shows an increasing number of black segments.
 - The bottom left figure has four white segments.
 - The bottom right figure has four black segments.
 2. The missing figure should have an intermediate number of black segments between the bottom left and bottom right.
 3. Looking at the answer options, we need a figure with two black segments and two white segments.
 4. Option B matches this description, with two black segments on top and two white segments on the bottom.
Final Answer: The figure with two black segments on top and two white segments on the bottom
Final Response: (B)

Figure 15: Erroneous model reasoning patterns observed on an FC question

2160
2161
2162
2163
2164
2165
2166
2167
2168
2169
2170
2171
2172
2173
2174
2175
2176
2177
2178
2179
2180
2181
2182
2183
2184
2185
2186
2187
2188
2189
2190
2191
2192
2193
2194
2195
2196
2197
2198
2199
2200
2201
2202
2203
2204
2205
2206
2207
2208
2209
2210
2211
2212
2213

Logical Reasoning

Instruction: According to the question, Choose the correct option logically

(1)Poverty (2)Population (3)death (4)Unemployment (5)Disease

(A)2,4,1,5,3 (B)1,2,3,4,5 (C)2,3,4,5,1 (D)2,4,5,1,3

Correct Answer: (A)2,4,1,5,3

Model: Gemini 1.5 Pro

Reasoning Steps: The question asks us to arrange the given words in a logical order. A logical order would be starting with the most general and ending with the most specific.

1. **Population** is the broadest concept, referring to the number of people in a particular area.
2. **Poverty** is a condition that affects a portion of the population.
3. **Unemployment** is a factor that can contribute to poverty.
4. **Disease** is a factor that can lead to death.
5. **Death** is the final outcome of the chain.

Therefore, the logical order is 2, 4, 5, 1, 3.

Final Answer: 2,4,5,1,3

Final Response: (D)

Model: GPT 4o

Reasoning Steps: To solve this question, we need to determine the logical sequence of the given terms: Poverty, Population, Death, Unemployment, and Disease. The logical sequence should follow a cause-and-effect relationship or a natural progression.

1. Population (2) - An increase in population can lead to...
2. Unemployment (4) - More people can result in fewer jobs available, leading to unemployment.
3. Poverty (1) - Unemployment can lead to poverty as people do not have a source of income.
4. Disease (5) - Poverty can lead to poor living conditions and lack of healthcare, resulting in diseases.
5. Death (3) - Diseases, if untreated, can lead to death.

The logical sequence is 2, 4, 1, 5, 3.

Final Answer: 2, 4, 1, 5, 3

Final Response: (D)

Model: Claude 3.5 Sonnet

Reasoning Steps: Step 1: Analyze the given items and their potential logical order.

1. Poverty
2. Population
3. Death
4. Unemployment
5. Disease

Step 2: Consider the cause-effect relationships between these items.

- Population growth can lead to increased poverty and unemployment.
- Poverty and unemployment can contribute to disease.
- Disease can lead to death.

Final Answer: 2,4,5,1,3

Final Response: (D)

Figure 16: Erroneous model reasoning patterns observed on an LR question

2214
2215
2216
2217
2218
2219
2220
2221
2222
2223
2224
2225
2226
2227
2228
2229
2230
2231
2232
2233
2234
2235
2236
2237
2238
2239
2240
2241
2242
2243
2244
2245
2246
2247
2248
2249
2250
2251
2252
2253
2254
2255
2256
2257
2258
2259
2260
2261
2262
2263
2264
2265
2266
2267

Mathematical Reasoning

If $\triangle ABC$ is an equilateral triangle such that $AD \perp BC$, then $AD^2 =$

A) $\frac{3a^2}{4}$ B) $\frac{3a^2}{2}$ C) $\frac{3}{4}BC^2$ D) $\frac{\sqrt{3}}{2}a$

1) A and C 2) D 3) A 4) B and C

Correct Answer: (1) A and C

Model: Gemini 1.5 Pro
Reasoning Steps: In an equilateral triangle, all sides are equal and all angles are 60 degrees. When a perpendicular is drawn from a vertex to the opposite side, it bisects the side and forms two 30-60-90 right triangles. In a 30-60-90 triangle, the hypotenuse is twice the length of the shorter leg, and the longer leg is $\sqrt{3}$ times the length of the shorter leg. In this problem, BC is the hypotenuse of the 30-60-90 triangle, and AD is the shorter leg. So, $BC = 2AD$. We also know that $BC = a$, so $AD = a/2$. Therefore, $AD^2 = (a/2)^2 = a^2/4$. So, the correct answer is $3a^2/4$.
Final Answer: $3a^2/4$
Final Response: 3) A

Model: GPT 4o
Reasoning Steps: Given that triangle ABC is an equilateral triangle, AD is the altitude from vertex A to side BC. In an equilateral triangle, the altitude AD can be calculated using the formula: $AD = (\sqrt{3}/2) * a$, where a is the side length of the equilateral triangle. To find AD^2 , we square both sides of the equation: $AD^2 = [(\sqrt{3}/2) * a]^2 = (3/4) * a^2$.
Final Answer: $(3/4) * a^2$
Final Response: 3) A

Model: Claude 3.5 Sonnet
Reasoning Steps: Step 1: In an equilateral triangle, all sides are equal and all angles are 60°. Step 2: Given $AD \perp BC$, AD is the height of the triangle. Step 3: In an equilateral triangle, the height (h) is related to the side length (a) by the formula: $h = (\sqrt{3}/2)a$. Step 4: AD is the height, so $AD = (\sqrt{3}/2)a$. Step 5: Square both sides: $AD^2 = ((\sqrt{3}/2)a)^2 = (3/4)a^2$. Step 6: Compare this result with the given options.
Final Answer: $3a^2/4$
Final Response: 3) A

Figure 17: Erroneous model reasoning patterns observed on an MR question

2322
2323
2324
2325
2326
2327
2328
2329
2330
2331
2332
2333
2334
2335
2336
2337
2338
2339
2340
2341
2342
2343
2344
2345
2346
2347
2348
2349
2350
2351
2352
2353
2354
2355
2356
2357
2358
2359
2360
2361
2362
2363
2364
2365
2366
2367
2368
2369
2370
2371
2372
2373
2374
2375

Odd One Out

Among the given group of 4 images, which set of images can be classified as a group?

(A) (1, 2, 3) (B) (1, 2, 4) (C) (1, 3, 4) (D) (2, 3, 4)

Correct Answer: (B) (1,2,4)

Model: Gemini 1.5 Pro
Reasoning Steps: The question asks us to identify the group of images that can be classified together. We can observe that images 1, 3, and 4 have a diagonal line in the shape. Image 2 does not have a diagonal line, it has a horizontal line. Therefore, image 2 is the odd one out.
Final Answer: (2, 3, 4)
Final Response: (D)

Model: GPT 4o
Reasoning Steps: To solve this problem, we need to identify the common characteristic among the images and find the one that does not fit. Upon examining the images, we notice that images 1, 3, and 4 have a diagonal line inside a square. Image 2, however, has a diagonal line inside a pentagon. Therefore, image 2 is the odd one out. The set of images that can be classified as a group are 1, 3, and 4.
Final Answer: (1, 3, 4)
Final Response: (C)

Model: Claude 3.5 Sonnet
Reasoning Steps: Step 1: Examine each image carefully.
Step 2: Identify the common characteristics in the images.
Step 3: Image 1 shows a rectangle with a diagonal line.
Step 4: Image 2 shows a rectangle with no additional lines.
Step 5: Image 3 shows a square with a diagonal line.
Step 6: Image 4 shows a rectangle with a horizontal line.
Step 7: Notice that images 1, 3, and 4 all contain an additional line within the shape.
Step 8: Image 2 is the only one without an additional line.
Step 9: Therefore, images 1, 3, and 4 can be classified as a group based on the presence of an additional line.
Final Answer: Images 1, 3, and 4 can be classified as a group
Final Response: (C)

Figure 19: Erroneous model reasoning patterns observed on an OD question

2376
2377
2378
2379
2380
2381
2382
2383
2384
2385
2386
2387
2388
2389
2390
2391
2392
2393
2394
2395
2396
2397
2398
2399
2400
2401
2402
2403
2404
2405
2406
2407
2408
2409
2410
2411
2412
2413
2414
2415
2416
2417
2418
2419
2420
2421
2422
2423
2424
2425
2426
2427
2428
2429

Pattern Recognition

Q. 30 and 31 → Directions :- Replace the question mark by choosing the correct alternatives from given below

Correct Answer: (3) 66

Model: Gemini 1.5 Pro
Reasoning Steps: The difference between the top and bottom number in the first square is 14 (23-39), the second square is 19 (91-70), and the third square is 12 (23-41). The difference between the left and right numbers in the first square is 17 (70-53), the second square is 12 (70-58), and the third square is 16 (65-49). The missing number in the third square is 49.

Final Answer: Figure 1
Final Response: 1

Model: GPT 4o
Reasoning Steps: To solve this problem, we need to identify which of the given simple figures (options 1-4) is hidden within the complex figure. We will compare each option with the complex figure to see if it can be found within it.

1. Option 1: This figure has a combination of intersecting lines forming a triangular shape. Upon close inspection, this pattern can be found in the complex figure.
2. Option 2: This figure has a different set of intersecting lines forming a star-like shape. This pattern is not present in the complex figure.
3. Option 3: This figure is a simple polygon with five sides. This shape is not present in the complex figure.
4. Option 4: This figure is a simple quadrilateral. This shape is not present in the complex figure.

Therefore, the correct answer is option 1.

Final Answer: Option 1
Final Response: 1

Model: Claude 3.5 Sonnet
Reasoning Steps: Step 1: Examine the complex figure carefully.
 Step 2: Look at each of the simple figures in the answer options.
 Step 3: For each simple figure, try to find its shape within the complex figure.
 Step 4: Option 1 can be found in the bottom left corner of the complex figure.
 Step 5: Option 2 can be found in the center of the complex figure.
 Step 6: Options 3 and 4 cannot be clearly identified within the complex figure.
 Step 7: Determine that option 2 is the most clearly identifiable simple figure within the complex figure.

Final Answer: The simple figure that is hidden in the complex figure is option 2.
Final Response: 2


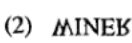
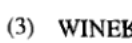
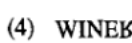
Figure 20: Erroneous model reasoning patterns observed on a PR question

2430
2431
2432
2433
2434
2435
2436
2437
2438
2439
2440
2441
2442
2443
2444
2445
2446
2447
2448
2449
2450
2451
2452
2453
2454
2455
2456
2457
2458
2459
2460
2461
2462
2463
2464
2465
2466
2467
2468
2469
2470
2471
2472
2473
2474
2475
2476
2477
2478
2479
2480
2481
2482
2483

Perspective Shift

Q31-34 Choose the correct water-image of the given words/numbers from amongst the alternatives.

WINER

(1)  (2)  (3)  (4) 

Correct Answer: (1)

Model: Gemini 1.5 Pro
Reasoning Steps: The question asks for the 'water-image' of the word 'WINER'. This means we need to imagine the word reflected in water. When reflected in water, the word will appear flipped vertically. This means each letter will be reversed but will retain its original orientation. Therefore, the correct answer is 'WINEK', which is option 4.
Final Answer: WINEK
Final Response: 4

Model: GPT 4o
Reasoning Steps: To find the water-image of the word 'WINER', we need to flip the word vertically. This means each letter will be mirrored along the horizontal axis. The water-image of 'WINER' will have the following transformations: 'W' remains 'W', 'I' remains 'I', 'N' remains 'N', 'E' becomes a mirrored 'E', and 'R' becomes a mirrored 'R'. The correct water-image of 'WINER' is 'WINE-Ø'.
Final Answer: WINE-Ø
Final Response: 4


Model: Claude 3.5 Sonnet
Reasoning Steps: Step 1: Understand what a water-image is - it's the reflection of the word as if seen in still water.
Step 2: Visualize how 'WINER' would look reflected in water.
Step 3: The reflection would flip the word vertically, not horizontally.
Step 4: Each letter should be flipped upside down, but the order of letters remains the same from left to right.
Step 5: Compare this mental image with the given options.
Step 6: Identify the option that matches the correct water-image.
Final Answer: WINER
Final Response: 4


Figure 21: Erroneous model reasoning patterns observed on a PS question


2484
2485
2486
2487
2488
2489
2490
2491
2492
2493
2494
2495
2496
2497
2498
2499
2500
2501
2502
2503
2504
2505
2506
2507
2508
2509
2510
2511
2512
2513
2514
2515
2516
2517
2518
2519
2520
2521
2522
2523
2524
2525
2526
2527
2528
2529
2530
2531
2532
2533
2534
2535
2536
2537


Relative Reasoning

Which one of the following Venn-diagrams correctly represents the relation between Herbivorous, Lions and Animals?

(1) 

(2) 

(3) 

(4) 

Correct Answer: (1)

Model: Gemini 1.5 Pro
Reasoning Steps: All herbivorous animals are animals. Lions are animals. However, lions are not herbivores. Thus the Venn diagram must show Lions as a subset of Animals, and Herbivores as a subset of Animals. There must not be any overlap between Lions and Herbivores because they are distinct groups. Option 4 correctly represents the above relationships.
Final Answer: Option 4 correctly represents the relationship between Herbivorous, Lions, and Animals.
Final Response: D

Model: GPT 4o
Reasoning Steps: To solve this question, we need to understand the relationship between herbivorous animals, lions, and animals in general. All lions are animals, and all herbivorous animals are also animals. However, lions are not herbivorous animals. Therefore, the correct Venn diagram should show two separate circles for herbivorous animals and lions, both within a larger circle representing animals. This relationship is best represented by option (2), where there are two separate circles within a larger circle.
Final Answer: Option (2) correctly represents the relationship between herbivorous animals, lions, and animals.
Final Response: B

Model: Claude 3.5 Sonnet
Reasoning Steps: Step 1: Understand the relationships between Herbivorous, Lions, and Animals.
 - All herbivores are animals.
 - All lions are animals.
 - Lions are not herbivores.
 - There are animals that are neither herbivores nor lions.
 Step 2: Analyze each Venn diagram option:
 (1) Two separate circles: This represents two distinct groups with no overlap, which is incorrect.
 (2) One circle inside another, with a separate small circle: This is incorrect as it suggests one group is entirely contained within another, while the third is separate.
 (3) Two overlapping circles: This is incorrect as it doesn't account for the fact that both lions and herbivores are subsets of animals.
 (4) Two concentric circles with a smaller circle inside the inner circle: This correctly represents the relationship.
 - The outermost circle represents all Animals.
 - The middle circle represents Herbivores, which are a subset of Animals.
 - The innermost circle represents Lions, which are also a subset of Animals but separate from Herbivores.
 Step 3: Choose the correct diagram based on the analysis.
Final Answer: The correct Venn diagram is the one with two concentric circles and a smaller circle inside the inner circle.
Final Response: D

Figure 22: Erroneous model reasoning patterns observed on an RR question

2538
2539
2540
2541
2542
2543
2544
2545
2546
2547
2548
2549
2550
2551
2552
2553
2554
2555
2556
2557
2558
2559
2560
2561
2562
2563
2564
2565
2566
2567
2568
2569
2570
2571
2572
2573
2574
2575
2576
2577
2578
2579
2580
2581
2582
2583
2584
2585
2586
2587
2588
2589
2590
2591

Sequence Completion

In the number series given below, one number is missing. Each series is followed by five alternative answers (1), (2), (3), (4) and (5). One of them is the right answer. Identify and indicate it as per the "Instructions".

4. 7, 6, 10, 27, 104,

(1) 520 2) 420 (3) 515 4) 525 5) 456

Correct Answer: (3) 515

Model: Gemini 1.5 Pro

Reasoning Steps: The sequence is 7, 6, 10, 27, 104.
The difference between the first two numbers is -1, between the next two is +4, then +17 and +77.
The pattern observed is that the difference between each number in the sequence is a multiple of 3. Each number is increased by a multiple of 3.
So the difference between the 5th and 6th number should be a multiple of 3 and should be $3^5=243$.
The next number in the sequence is then $104 + 243 = 347$.

Final Answer: 347

Final Response: 2

Model: GPT 4o

Reasoning Steps: To find the next number in the series 7, 6, 10, 27, 104, we need to identify the pattern. Let's look at the differences between consecutive terms: $6 - 7 = -1$ $10 - 6 = 4$ $27 - 10 = 17$ $104 - 27 = 77$
Now, let's look at the differences between these differences:
 $4 - (-1) = 5$ $17 - 4 = 13$ $77 - 17 = 60$
The differences between the differences are increasing in a pattern. Let's look at the differences between these differences:
 $13 - 5 = 8$ $60 - 13 = 47$
The differences between the differences are also increasing. Let's look at the differences between these differences:
 $47 - 8 = 39$
So, the pattern seems to be that each difference is increasing by a certain amount. Let's apply this pattern to find the next term: $77 + 60 = 137$
So, the next term in the series is $104 + 137 = 241$. Therefore, the next number in the series is 241.

Final Answer: the next number in the series is 241.

Final Response: 6

Model: Claude 3.5 Sonnet

Reasoning Steps: Step 1: Analyze the given sequence: 4, 7.6, 10, 27, 104, ...
Step 2: Identify the pattern:
- From 4 to 7.6: multiply by 1.9
- From 7.6 to 10: multiply by 1.315789...
- From 10 to 27: multiply by 2.7
- From 27 to 104: multiply by 3.851851...
Step 3: Notice that the multiplier is increasing each time.
Step 4: The next multiplier should be higher than 3.851851...
Step 5: Multiply 104 by 5 (next logical increase in multiplier): $104 * 5 = 520$
Step 6: Check the given options for 520.

Final Answer: The answer to this question is 520.

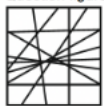
Final Response: 1

Figure 23: Erroneous model reasoning patterns observed on an SC question


2592
2593
2594
2595
2596
2597
2598
2599
2600
2601
2602
2603
2604
2605
2606
2607
2608
2609
2610
2611
2612
2613
2614
2615
2616
2617
2618
2619
2620
2621
2622
2623
2624
2625
2626
2627
2628
2629
2630
2631
2632
2633
2634
2635
2636
2637
2638
2639
2640
2641
2642
2643
2644
2645

Spatial Reasoning

Question figure




Answer figure




1


(1) 1
(3) 3



2



3



4

(2) 2
(4) 4

Direction: In the given questions a complex figure is given. Find out which of the simple figures given in the alternatives is hidden in the complex figure.

Correct Answer: (4) 4

Model: Gemini 1.5 Pro
Reasoning Steps: The question asks to find a simple figure present within the complex figure. We need to look at each simple figure provided and try to visualize if it is present within the complex figure. Comparing the figures, we can see that Figure 1 is present in the complex figure, as the lines in Figure 1 are present in the complex figure and are arranged in the same way.
Final Answer: Figure 1
Final Response: 1

Model: GPT 4o
Reasoning Steps: To solve this problem, we need to identify which of the given simple figures (options 1-4) is hidden within the complex figure. We will compare each option with the complex figure to see if it can be found within it.
 1. Option 1: This figure has a combination of intersecting lines forming a triangular shape. Upon close inspection, this pattern can be found in the complex figure.
 2. Option 2: This figure has a different set of intersecting lines forming a star-like shape. This pattern is not present in the complex figure.
 3. Option 3: This figure is a simple polygon with five sides. This shape is not present in the complex figure.
 4. Option 4: This figure is a simple quadrilateral. This shape is not present in the complex figure.
 Therefore, the correct answer is option 1.
Final Answer: Option 1
Final Response: 1

Model: Claude 3.5 Sonnet
Reasoning Steps: Step 1: Examine the complex figure carefully.
 Step 2: Look at each of the simple figures in the answer options.
 Step 3: For each simple figure, try to find its shape within the complex figure.
 Step 4: Option 1 can be found in the bottom left corner of the complex figure.
 Step 5: Option 2 can be found in the center of the complex figure.
 Step 6: Options 3 and 4 cannot be clearly identified within the complex figure.
 Step 7: Determine that option 2 is the most clearly identifiable simple figure within the complex figure.
Final Answer: The simple figure that is hidden in the complex figure is option 2.
Final Response: 2

Figure 24: Erroneous model reasoning patterns observed on an SR question