PolyMATH: A Challenging Multi-modal Mathematical Reasoning Benchmark

Anonymous ACL submission

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].

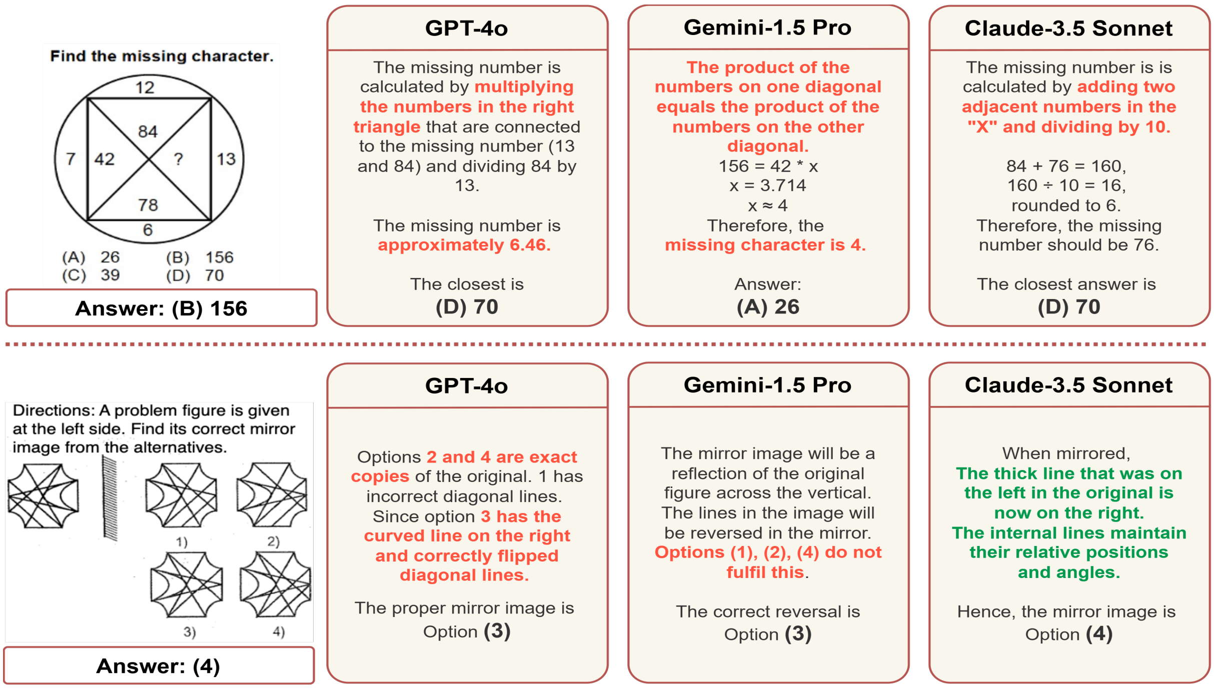
1 Introduction

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, 2023b; 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 reasoning*, *pattern recognition*, and *numerical reasoning*. and (2) we cover a diverse array of visual contexts, including images with venn diagrams, spatially-related layouts, as well as geometric figures. Poly-

---

[1] https://anonymous.4open.science/r/PolyMATH-052D



Find the missing character.

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

Answer: (B) 156

GPT-4o

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

The missing number is approximately 6.46.

The closest is
(D) 70

Gemini-1.5 Pro

The product of the numbers on one diagonal equals the product of the numbers on the other diagonal.
156 = 42 * x
x = 3.714
x ≈ 4
Therefore, the missing character is 4.

Answer:
(A) 26

Claude-3.5 Sonnet

The missing number is is calculated by adding two adjacent numbers in the "X" and dividing by 10.

84 + 76 = 160,
160 ÷ 10 = 16,
rounded to 6.
Therefore, the missing number should be 76.

The closest answer is
(D) 70

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

Answer: (4)

GPT-4o

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

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

Gemini-1.5 Pro

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

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

Claude-3.5 Sonnet

When mirrored, The thick line that was on the left in the original is now on the right. The internal lines maintain their relative positions and angles.

Hence, the mirror image is Option **(4)**

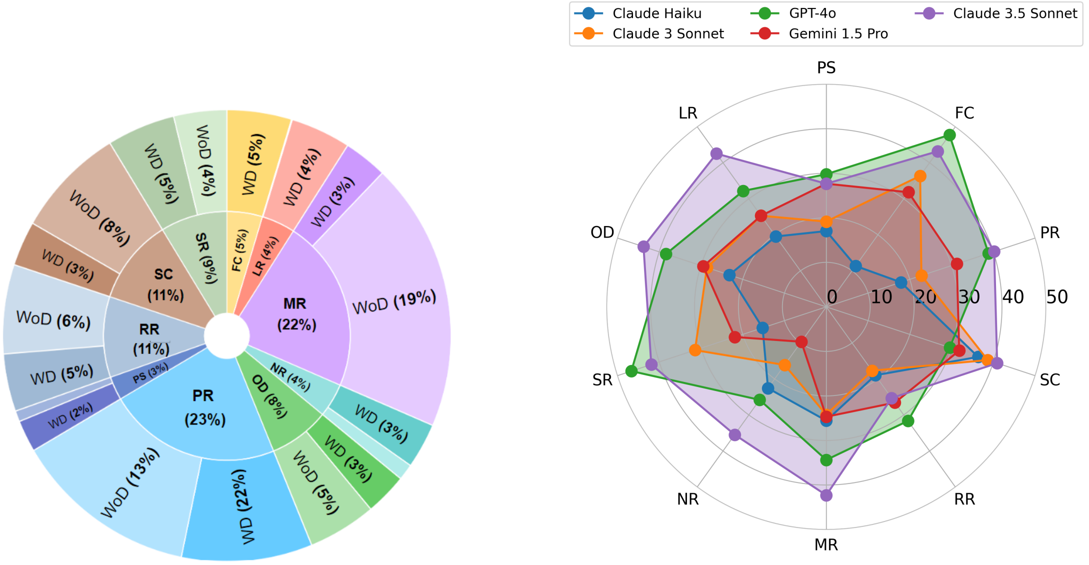
Figure 1: Examples of the reasoning patterns employed by MLLMs when faced with questions involving visual information. In the top row, models fail to perceive the relationship between adjacent semicircles; in the bottom row, models fail to comprehend fine details in the answer images.

MATH is a meticulously curated dataset of 5000 multimodal reasoning problems newly acquired from a publicly available source (Table 1). The problems of the original source have been crafted and rigorously reviewed by expert annotators, and require diverse fine-grained problem-solving capabilities. Additionally, we provide detailed textual representations of diagrams of the samples. As denoted in fig. 1, these problems are designed to assess the logical reasoning abilities of the average high school student over text and diagrams. We observe that MLLMs fail to demonstrate the cognitive reasoning skills required to solve these questions.

We conduct extensive experiments on PolyMATH with state-of-the-art (SOTA) closed-source MLLMs like the Claude family (3.5 Sonnet, 3 Sonnet, 3 Haiku), Gemini-1.5 Pro, and GPT-4o, and 9 open-source MLLMs like LLaVA (34B) and ShareGPT4V. We evaluate them via zero shot, few shot, Chain-of-Thought (Wei et al., 2022b) and step back prompting (Zheng et al., 2024). We show that PolyMATH is a challenging benchmark, with human performance (established by qualified human annotators with graduate degrees) reaching only 66.3% accuracy. The most powerful model we evaluate, Claude-3.5 Sonnet, achieves the best score of 41.90% followed by GPT-4o, which attains 36.50%. The best open source models like LLaVA-v1.6 Mistral (7B) and ShareGPT4V (13B) achieves the accuracy of 15.20% and 12.80% respectively. We additionally create a diagram only subset (*test-img*) of the benchmark to gauge the gap in visual reasoning abilities between the multimodal models and average human capability. We find that the performance of these models drops further to 26.20% for Claude-3.5 Sonnet and 22.50% by Gemini-1.5 Pro when evaluated on *test-img* only. In contrast with human cognitive patterns, when given text descriptions in place of the diagram in these questions, model accuracy improves by ∼4-7%. We also conduct an error analysis on Claude-3.5 Sonnet, Gemini-1.5 Pro and GPT-4o, and find that the most common errors stem from misunderstanding diagrams (∼ 60%), misidentifying logical patterns (∼ 25%), and forgetting relational information (∼ 12%). Finally, we evaluate OpenAI o1 models (OpenAI, 2024b) on without diagram questions of the benchmark and observe 66.72% accuracy (o1-preview), 2% points below than the human baseline.

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 mod-



(a) Dataset categorization

(b) Results on closed source models

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

els 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, 2023b) 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 geometric 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 works evaluate LLMs across 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 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

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

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

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 6.

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 4), 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 populated.

3.2 Dataset categorization

We develop a categorization schema that catalogues questions on basis of the information provided and the type of reasoning assessed by the question. Based on the continuous human evaluation during collection, we identify 10 distinct question categories. We enumerate these categories along with their definitions in Table 6. We further distinguish between questions *with diagram* and *without diagram*. The *with diagram* questions are designed around the information presented in the diagrams (Fig 4). 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 multimodal 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 4). Results on text-only characterization of questions in our dataset can be found in §4.3.

3.4 Quality Assurance

Following the collection and annotation process, we conduct a comprehensive quality check. We discard samples that are [1] of low resolution, [2] outside the scope of the categories (Table 6), or [3] missing vital information. We also discard samples with 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 were obtained on this *testmini* subset of POLYMATH. We also create a *test-img* question set, consisting solely of 1000 *with diagram* questions, aimed at faster, focused assessment of models' visual comprehension. We use a random sampling strategy to create *test-img* due to diagram imbalance. [2] For data distribution, see Table 1. Further details on data collection and annotation are available in §B.1.

4 Experiments

We conduct a systematic evaluation of existing MLLMs on POLYMATH. We first introduce the experimental setup in this section. Then we present our findings followed by multiple dataset analysis experiments. Additional experimental results and qualitative examples are present in §C and E.3.

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, 2024b) and Claude 3 Haiku and Sonnet (`claude-3-sonnet-20240229`, `claude-3-haiku-20240307`) (Anthropic, 2024a) (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) & 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 §C.

Implementation Details All reported results are on the $testmini$ subset. As a comparative baseline, we simulate random chance by selecting a random option for multiple-choice questions over 1000 trials. Additionally, the problems in POLYMATH were independently solved by the paper's authors (four engineering graduates and two PhDs), serving as a human performance baseline. We evaluate the benchmark using various prompting methods, including zero shot, few shot (2-shot), Chain-of-Thought (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 §C.2. 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.

---

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

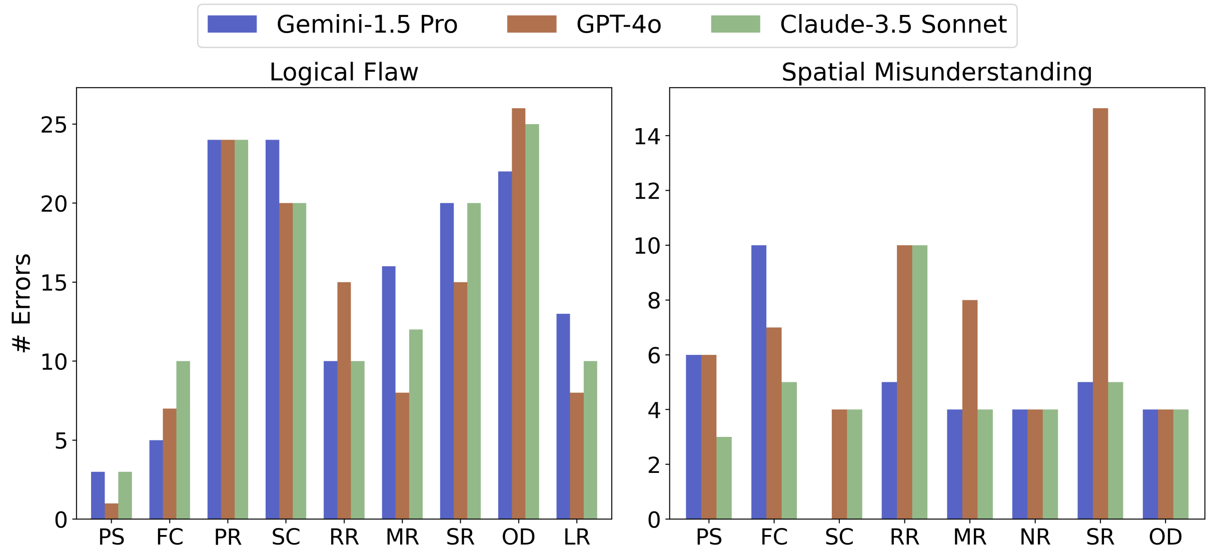


Figure 3: Frequency of LF and SM errors across different question categories. We report per-model figures to enable a comparison of model abilities. They are most prevalent in the OD, PR, and SC categories of questions, owing to the amount of logical leaps and visual reasoning required by these questions.

Category	PS	FC	PR	SC	RR	MR	NR	SR	OD	LR	**Overall**
Baseline											
Random chance	9.68	4.26	6.61	9.82	9.52	9.82	15.91	6.90	7.59	9.09	8.60
Human eval	51.08	70.57	61.82	69.35	69.84	76.64	58.71	62.64	64.98	51.14	66.62
Zero Shot Inference											
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
Chain-of Thought Prompting Inference											
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
Step Back Prompting Inference											
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 2: Closed-source MLLM results on *testmini* using zero-shot, CoT, and step-back prompting. Highest and lowest scores per strategy are highlighted including random chance and human baselines (avg of six runs).

4.2 Results

Closed Source Models Across various prompting strategies (Table 2), 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-

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 3: 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.

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, in categories like SR, PR, and LR.

Open Source Models Table 3 showcases the results of open-source MLLMs. LLaVA-v1.6-Mistral-7B model achieved the highest overall score of 15.2%. It excelled in OD (22.78%), SR (19.54%), RR (18.1%), and MR (15.18%) indicating its proficiency in generating precise, coherent, and relevant responses, even for out-of-distribution samples. The ShareGPT4V (13B) model exhibited the second-highest overall score of 12.8%, with outstanding performance in the PR (13.66%), SC (12.5%), RR (15.24%), MR (10.71%), SR (12.64%), and OD (17.72%) categories. Other models, such as LlaVA-v1.6-Vicuna 13B, LlaVA-1.5 (13B), G-LLaVA (13B), and LlaVA-v1.6 (34B), exhibited varying levels of success across the different categories, highlighting their individual strengths and weaknesses in handling the diverse reasoning aspects tested by the dataset.

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

4.3 Experimental Analysis

MLLMs Rely More on Image Descriptions than Image To evaluate the visual reasoning capabilities, we used *test-img* subset, which contains questions with diagrams. Additionally, we generated a text-only version of *test-img* by replacing all diagrams with detailed textual descriptions. Both experiments were carried out in a zero shot setting. Our analysis reveals three key findings. First, we observed a noticeable decline in performance on *test-img*, particularly for models like GPT-4o and Claude-3.5 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 $\sim 4\%$, indicating that the models struggle with diagrams and benefit from textual representations. Finally, we evaluated popular text-only LLMs such as LLaMA-3 (70B), Reka Flash, and Mistral Large on the text-description version of *test-img*. Their scores ($\sim 15\%$) were lower than those of the MLLMs ($\sim 27\%$), underscoring the advantage of multi-modal models in handling visually-grounded tasks.

A Closer Look at Model Errors We analysed 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 10). The total error distribution of all three models is present in Table 11. Qualitative examples for category-wise errors are present in §E.3. 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 3 shows the category-wise distribution of the two types of error. These errors were most prevalent in OD, PR, and SC category of questions, as making uncommon logical leaps and fully comprehending visuals is integral to solving these. Furthermore,

Category	PS	FC	PR	SC	RR	MR	NR	SR	OD	LR	**Overall**
MLLM Inference on Diagrams (Multi-modal)											
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
MLLM Inference on Diagram Descriptions (Text-only)											
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
LLM Inference on Diagram Descriptions (Text-only)											
Mistral Large	15.00	13.11	11.29	15.48	18.52	13.41	9.41	17.83	25.32	33.33	14.90
Reka Flash	16.67	13.93	12.10	16.67	19.44	14.63	9.41	18.60	26.58	33.33	15.80
Llama-3 (70B)	16.67	13.93	11.69	16.67	19.44	14.63	10.59	18.60	26.58	33.33	15.80

Table 4: Visual comprehension ablation results on *test-img*. We compare [1] multi-modal inference with diagrams and [2] unimodal inference using text descriptions. Highest and lowest scores per category are highlighted. Unimodal LLM performance on text-only questions is also reported.

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 5: 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.

in questions involving extrapolation over multiple weakly connected data points, models came to conclusions that contradicted earlier data, indicating a lack of information retention. Finally, we found that models fell into identical fallacious reasoning patterns, e.g. assuming that a pattern holds across each row when a pattern is replicated across columns. The category with the highest % of shared errors was PR, where we observed that GPT, Gemini, and Claude followed the same incorrect reasoning structure on nearly 80% of the analysed samples. Thus, despite their differences, in practice we see that MLLMs share the same strengths and shortcomings. For more details, see §E.

Evaluation of OpenAI o1 models To understand the capabilities of recent text-only reasoning models (o1-preview and o1-mini), we evaluate these models on 595 text-only questions. We also present human baseline scores on these questions. These results are presented in Table 5. o1-preview ($\sim 67\%$) scores competitively with human performance ($\sim 68\%$), while o1-mini ($\sim 58\%$) 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 $\sim 24\%$ 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.

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.

References

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

Mistral AI. 2024. Au large.

AI@Meta. 2024. Llama 3 model card.

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

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

Anthropic. 2023. Claude 2.

Anthropic. 2024a. The claude 3 model family: Opus, sonnet, haiku.

Anthropic. 2024b. Claude 3.5 sonnet model card addendum.

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

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

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Xiaoyi Dong, Pan Zhang, Yuhang Zang, Yuhang Cao, Bin Wang, Linke Ouyang, Xilin Wei, Songyang Zhang, Haodong Duan, Maosong Cao, et al. 2024. Internlm-xcomposer2: Mastering free-form text-image composition and comprehension in vision-language large model. *arXiv preprint arXiv:2401.16420*.

Simon Frieder, Luca Pinchetti, Ryan-Rhys Griffiths, Tommaso Salvatori, Thomas Lukasiewicz, Philipp Christian Petersen, Alexis Chevalier, and Julius Berner. 2023. Mathematical capabilities of chatgpt. In *37th Conference on Neural Information Processing Systems (NeurIPS 2023) Track on Datasets and Benchmarks*.

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

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

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

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

Peng Gao, Jiaming Han, Renrui Zhang, Ziyi Lin, Shijie Geng, Aojun Zhou, Wei Zhang, Pan Lu, Conghui He, Xiangyu Yue, Hongsheng Li, and Yu Qiao. 2023b. LLaMA-Adapter V2: Parameter-efficient visual instruction model. *arXiv preprint arXiv:2304.15010*.

Peng Gao, Renrui Zhang, Chris Liu, Longtian Qiu, Siyuan Huang, Weifeng Lin, Shitian Zhao, Shijie Geng, Ziyi Lin, Peng Jin, et al. 2024. Sphinx-x: Scaling data and parameters for a family of multi-modal large language models. *arXiv preprint arXiv:2402.05935*.

Google. 2023. Bard.

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

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

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

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

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

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021b. Measuring mathematical problem solving with the math dataset. *NeurIPS*.

Yining Hong, Haoyu Zhen, Peihao Chen, Shuhong Zheng, Yilun Du, Zhenfang Chen, and Chuang Gan. 2024. 3d-llm: Injecting the 3d world into large language models. *Advances in Neural Information Processing Systems*, 36.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Ziyi Lin, Chris Liu, Renrui Zhang, Peng Gao, Longtian Qiu, Han Xiao, Han Qiu, Chen Lin, Wenqi Shao, Keqin Chen, et al. 2023. Sphinx: The joint mixing of weights, tasks, and visual embeddings for multi-modal large language models. *arXiv preprint arXiv:2311.07575*.

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

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

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

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

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

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

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

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

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

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

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

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

Haipeng Luo, Qingfeng Sun, Can Xu, Pu Zhao, Jianguang Lou, Chongyang Tao, Xiubo Geng, Qingwei Lin, Shifeng Chen, and Dongmei Zhang. 2023. Wizardmath: Empowering mathematical reasoning for large language models via reinforced evol-instruct. *arXiv preprint arXiv:2308.09583*.

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

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

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

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

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

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

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

OpenAI. 2023a. Chatgpt. `https://chat.openai.com`.

OpenAI. 2023b. GPT-4V(ision) system card.

OpenAI. 2023c. OpenAI's GPT-4. `https://openai.com/research/gpt-4`.

OpenAI. 2024a. Gpt 4o models.

OpenAI. 2024b. Open ai o1 models.

Aitor Ormazabal, Che Zheng, Cyprien de Masson d'Autume, Dani Yogatama, Deyu Fu, Donovan Ong, Eric Chen, Eugenie Lamprecht, Hai Pham, Isaac Ong, Kaloyan Aleksiev, Lei Li, Matthew Henderson, Max Bain, Mikel Artetxe, Nishant Relan, Piotr Padlewski, Qi Liu, Ren Chen, Samuel Phua, Yazheng Yang, Yi Tay, Yuqi Wang, Zhongkai Zhu, and Zhihui Xie. 2024. Reka core, flash, and edge: A series of powerful multimodal language models. *Preprint*, arXiv:2404.12387.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*.

Subhro Roy and Dan Roth. 2016. Solving general arithmetic word problems. *ArXiv*, abs/1608.01413.

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

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

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

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

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

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

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

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

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

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

John Chong Min Tan and Mehul Motani. 2023. Large language model (llm) as a system of multiple expert agents: An approach to solve the abstraction and reasoning corpus (arc) challenge. *arXiv preprint arXiv:2310.05146*.

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

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

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

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

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

Ke Wang, Houxing Ren, Aojun Zhou, Zimu Lu, Sichun Luo, Weikang Shi, Renrui Zhang, Linqi Song, Mingjie Zhan, and Hongsheng Li. 2024a. Mathcoder: Seamless code integration in LLMs for enhanced mathematical reasoning. In *The Twelfth International Conference on Learning Representations*.

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

Ruocheng Wang, Eric Zelikman, Gabriel Poesia, Yewen Pu, Nick Haber, and Noah D Goodman. 2023a. Hypothesis search: Inductive reasoning with language models. *arXiv preprint arXiv:2309.05660*.

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

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

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

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

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

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

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

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

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

Qinghao Ye, Haiyang Xu, Jiabo Ye, Ming Yan, Anwen Hu, Haowei Liu, Qi Qian, Ji Zhang, Fei Huang, and Jingren Zhou. 2023b. mplug-owl2: Revolutionizing multi-modal large language model with modality collaboration. *Preprint*, arXiv:2311.04257.

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

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

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

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

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

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

Renrui Zhang, Jiaming Han, Aojun Zhou, Xiangfei Hu, Shilin Yan, Pan Lu, Hongsheng Li, Peng Gao, and Yu Qiao. 2024. LLaMA-adapter: Efficient fine-tuning of large language models with zero-initialized attention. In *The Twelfth International Conference on Learning Representations*.

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

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

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

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

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

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

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

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

Aojun Zhou, Ke Wang, Zimu Lu, Weikang Shi, Sichun Luo, Zipeng Qin, Shaoqing Lu, Anya Jia, Linqi Song, Mingjie Zhan, et al. 2023. Solving challenging math word problems using gpt-4 code interpreter with code-based self-verification. *arXiv preprint arXiv:2308.07921*.

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

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

Appendix

Appendix Overview

- Section A: Extended Related Work
- Section B: Dataset creation
- Section C: Additional Experimental Details
- Section D: Extended Analysis
- Section E: Qualitative Error Analysis

A 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 and 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.

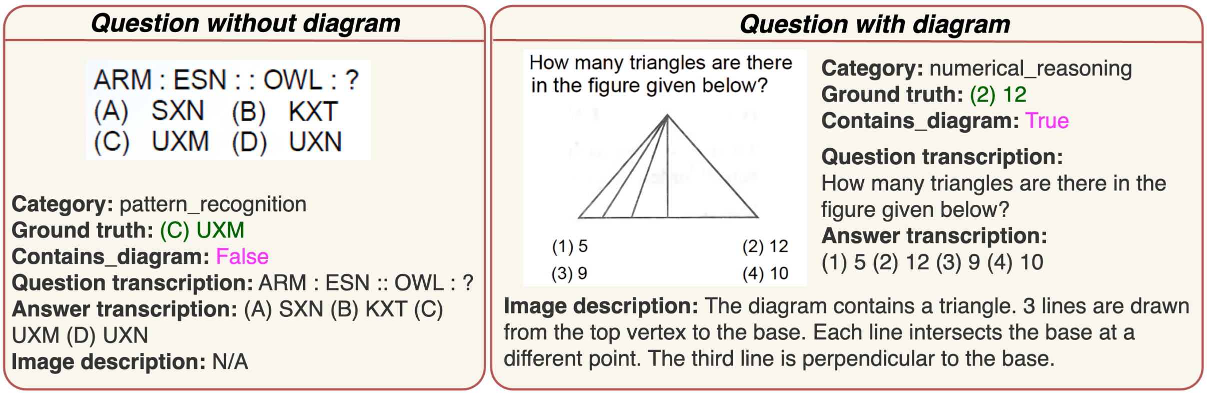
Despite substantial mathematical reasoning encapsulated in visual modalities, most existing benchmarks (Amini et al., 2019; Cobbe et al., 2021; Mishra et al., 2022; Frieder et al., 2023; Lu et al., 2023b) are textual only. Moreover, some datasets exhibit performance saturation, with GPT-4 achieving 92.0% 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 attempts like GeoQA (Chen et al., 2021a), while MathVista (Lu et al., 2023a) and MMMU (Yue et al., 2023a) incorporated various multimodal tasks and college-level questions, respectively.

MLLMs, building upon LLMs (Touvron et al., 2023a,b; OpenAI, 2023a; Jiang et al., 2024; Brown et al., 2020) and large vision models (Radford et al., 2021; Kirillov et al., 2023; Zhang et al., 2023d,c,e), have become increasingly prominent. They extend LLMs to diverse tasks and 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; Hong et al., 2024), audio (Han et al., 2023; Su et al., 2023), and video (Zhang et al., 2023a; Chen et al., 2023a). Noteworthy examples like OpenAI's GPT-4V (OpenAI, 2023b) and Google's Gemini (Team et al., 2023) exhibit exceptional visual reasoning capabilities, setting new benchmarks in multi-modal performance.

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

Despite comprehensive benchmarks (Fu et al., 2023a; Liu et al., 2023d; Li et al., 2023b; Xu et al., 2023a) for general visual instruction-following scenarios, the specific potential of MLLMs for visual mathematical problem-solving remains under-explored. Prior studies like VQA (Antol et al., 2015; Goyal et al., 2017), VizWiz (Gurari et al., 2018), and ParsVQA-Caps (Mobasher et al., 2022) evaluate LMMs' general visual question answering abilities on open-ended image queries. Additionally, works have assessed LMMs' specific skills beyond natural scenes, such as abstract shapes (Antol et al., 2015; Lu et al., 2021b; Ji et al., 2022), geometry diagrams (Seo et al., 2015; Lu et al., 2021a; Chen et al., 2022a; Cao and Xiao, 2022), charts (Methani et al., 2020; Masry et al., 2022; Kahou et al., 2017; Chang et al., 2022; Kafle et al., 2018), documents (Singh et al., 2019; Mathew et al., 2022; Liu et al., 2023e), synthetic im-



Category: pattern_recognition
Ground truth: (C) UXM
Contains_diagram: False
Question transcription: ARM : ESN :: OWL : ?
Answer transcription: (A) SXN (B) KXT (C) UXM (D) UXN
Image description: N/A

Category: numerical_reasoning
Ground truth: (2) 12
Contains_diagram: True
Question transcription: How many triangles are there in the figure given below?
Answer transcription: (1) 5 (2) 12 (3) 9 (4) 10
Image description: The diagram contains a triangle. 3 lines are drawn from the top vertex to the base. Each line intersects the base at a different point. The third line is perpendicular to the base.

Figure 4: 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*.

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 6: 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.

ages (Dahlgren Lindström and Abraham, 2022; Li et al., 2023d; Bitton-Guetta et al., 2023), external knowledge (Schwenk et al., 2022; Shah et al., 2019), commonsense reasoning (Zellers et al., 2019; Yin et al., 2021), scientific knowledge (Lu et al., 2022; Kembhavi et al., 2017, 2016), and medical understanding (Zhang et al., 2023g; Lau et al., 2018).

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

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

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

B Dataset creation

B.1 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 tools[3]. 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

---

[3] https://opencv.org/

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.

- LLM-based categorization, followed by human verification, was performed to obtain question categories. Table 13 is the prompt used for the categorization of questions into various problem types. Figures 15, 16, 17, 18, 19, 20, 21, 22, 23, 24 demonstrate examples from each question category defined in Table 6.

C Additional experiment details

C.1 Hyperparameters

The experimental hyperparameters are enumerated in Table 7. Furthermore, Table 8 provides the source repositories and model cards for the various models used in our experiments. Table 9 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.

C.2 Prompts for inference

The various prompts are detailed in this section. 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 4.

D Extended Analysis

D.1 Additional inference results

In this section, we show inference results from additional experiments to further illustrate model performance on POLYMATH. The results of open-source models on *test-mini* is shown in Table 9. We also document model performance on the full POLYMATH dataset (5000 questions) in Table 20. Additionally, we create a 153-sample set for each category to form a category-balanced subset of POLYMATH, on which we show model performance in Table 21.

D.2 Reliance on diagram descriptions

In order to quantify the maximum performance gain achievable by providing diagram description to MLLMs, we conduct an additional experiment where we provide diagram description, question, and diagram for all questions in *test-img*. The results are shown in Table 22.

E Error Analysis

E.1 Methodology

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.

E.2 Quantitative Analysis

We define the 6 types of errors found in model reasoning patterns and their frequency of occurrence in Table 10. Table 11 provides a detailed quantitative analysis of error type frequency per question category. Additionally, we analyse error patterns for the most-performant and least-performant open source models in Tables 24 and 23 respectively.

E.3 Qualitative Analysis

This section presents examples of the qualitative error analysis that was carried out. Figures 15, 16, 17, 18, 19, 20, 21, 22, 23 and 24 contains examples of failures by three proprietary models viz. Gemini-1.5 Pro, GPT-4o, and Claude-3.5 Sonnet across all categories.

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 7: Hyperparameters used in the experiments

Model	**Release Time**	**Source**
GPT-4o (OpenAI, 2024a)	2023-03	`https://platform.openai.com/`
Claude 3 family (Anthropic, 2024b,a)	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 8: 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**
Chain of Thought Inference											
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
Step Back Inference											
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 9: Results of open-source MLLMs on the *testmini* split of POLYMATH. We report model results using Chain-of-Thought, and Step Back prompting methods.

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 10: 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.

Error Type	**PS**	**FC**	**PR**	**SC**	**RR**	**MR**	**NR**	**SR**	**OD**	**LR**	**Overall**
Gemini-1.5 Pro											
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
GPT-4o											
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
Claude-3.5 Sonnet											
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%.

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.

Mathematical reasoning: this question purely requires calculations of a mathematical nature. This includes solving a straightforward equation.

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.

Sequence completion: given a sequence of numbers or figures, this question involves finding the sequentially next element in a series.

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. Th goal is to complete the figure and identify the element missing from a marked position.

Odd one out: given a set of elements, identify the element that is not like the others.

Spatial reasoning: questions involving reasoning observationally and visualizing the question in order to arrive at the answer.

Perspective shift: Questions where a figure is given and you are instructed to morph it according to the instructions (flip, mirror image, rotate, etc)

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.

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.

Logical reasoning: Questions involving simple logical reasoning such as entailment and contradiction.

Now, observe the following question.

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.

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

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.

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

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."

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.

Output a JSON with the keys Q1, Q2, Q3, Q4, Q5, Q6 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 16: Prompt for Chain-of-Thought inference

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."
Step back category prompt:
Mathematical Reasoning: "Q1: What is the relation of all given variables to one another? How is each variable related to the missing value?
Q2: Which are the mathematical operations involved in solving a question like this?"
Pattern Recognition: "Q1: What is the pattern being followed in this question? Provide an example.
Q2: Which are the elements in this question that follow this pattern?"
Sequence Completion: "Q1: What is a numerical sequence?
Q2: What is the relationship between previous and subsequent elements in a sequence? What is the relationship between elements in the sequence present in this question?"

Figure Completion: "Q1: How do you approach a figure completion problem?
Q2: What is the information you have and the missing information? What are their spatial relationships to one another?"

Odd one out: "Q1: How do you identify an odd element out of a set?
Q2: Describe the elements in this set. Now ,what do almost all of these elements have in common?"

Spatial Reasoning: "Q1: What are the spatial manipulations that occur in this question? Eg. unfolding, folding, 2D to 3D reconstruction, etc.
Q2: Given the original question image, how can you visualize the resulting image after the manipulations mentioned in the question? Explain in detail."

Perspective Shift: "Q1: What are the attributes of an image that is flipped, rotated, or its mirror image? What differentiates the result of these operations from the original image?
Q2: Which of these operations apply in this image, and in what order?"

Numerical Reasoning: "Q1: What is the information you are given? What do you need to find out? How can you arrive at this number? Q2: What are the main points of concern in solving such a question? How can you ensure that you do not under or over estimate the final number?"

Relative Reasoning: "Q1: What is the information you are given? What are the relationships of the given data points to one another? What is the information you need to discover? Which data points are directly or indirectly related to the missing variable? Explain in detail.
Q2: What principles of relational logic do you need to apply to this question?"

Logical Reasoning: "Q1: what are the principle of logical reasoning involved in solving this question? Q2: What is the information provided in this question? What is the objective of this question?"

Meta Prompt: Step back category prompt. Q3: Based on the above information, provide a step-by-step solution to the question in the image. Q4: What is the answer to this question? Remember, the answer must be present in the given list of answer options 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

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.

Model	**FC**	**LR**	**MR**	**NR**	**OD**	**PR**	**PS**	**RR**	**SC**	**SR**	**Overall**
Gemini-1.5 Pro	24.03	37.27	34.61	30.00	36.27	27.11	16.34	30.29	29.39	32.49	30.68
GPT-4o	22.32	55.91	34.61	40.91	47.86	27.82	19.61	32.00	25.81	47.37	34.16
Claude-3 sonnet	21.46	43.64	25.51	32.73	33.25	25.09	22.88	33.14	27.60	27.23	28.06
Claude Haiku	19.31	24.09	20.87	28.64	30.23	23.50	23.53	20.57	25.09	22.43	23.28
Claude-3.5 Sonnet	29.18	79.09	35.59	50.91	53.65	27.82	45.75	32.76	31.18	51.03	38.42

Table 20: Results on the entire POLYMATH dataset

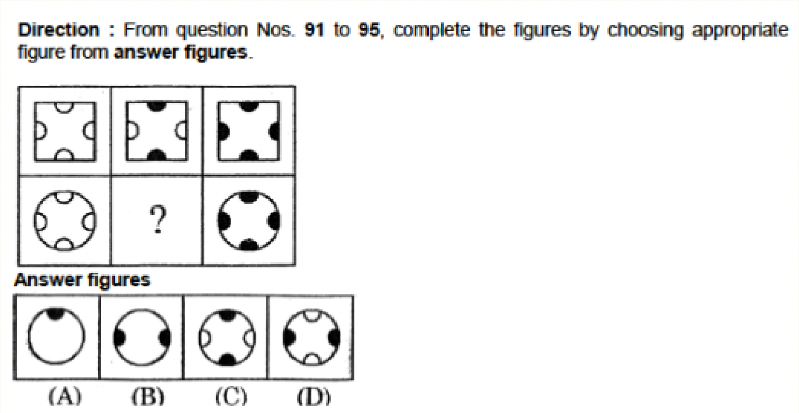
Model	**FC**	**LR**	**MR**	**NR**	**OD**	**PR**	**PS**	**RR**	**SC**	**SR**	**Overall**
Gemini 1.5 Pro	27.45	33.99	35.95	28.76	32.68	23.53	13.07	33.33	29.39	33.99	29.21
GPT 4o	19.61	58.82	32.68	43.14	45.75	30.07	22.88	28.76	24.18	48.37	35.42
Claude 3 sonnet	19.61	45.75	22.22	35.95	35.95	22.22	19.61	36.60	26.14	28.76	29.28
Claude Haiku	22.88	20.92	19.61	30.07	28.76	24.84	22.22	21.57	26.14	20.92	23.79
Claude 3.5 sonnet	26.14	82.35	38.56	47.71	55.56	26.80	49.02	29.41	30.07	52.29	43.79

Table 21: Results on a 153-sample set of each category, showing model scores on a balanced distribution across question categories.

	FC	**LR**	**MR**	**NR**	**OD**	**PR**	**PS**	**RR**	**SC**	**SR**	**Overall**
Claude Haiku	23.40	25.00	24.55	18.18	35.44	29.52	48.39	28.57	40.18	34.48	30.00
GPT 4o	42.55	52.27	40.18	45.45	35.44	46.26	70.97	33.33	53.57	60.92	45.60
Gemini 1.5 Pro	44.68	54.55	41.96	59.09	56.96	28.63	29.03	42.86	38.39	37.93	40.50
Claude-3 Sonnet	42.55	47.73	36.16	50.00	62.03	29.52	29.03	42.86	38.39	37.93	39.00
Claude-3.5 Sonnet	46.81	56.82	42.86	59.09	37.97	47.14	83.87	36.19	58.04	63.22	49.00

Table 22: Results of experiment setting combining diagram description, along with diagram image and question.

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)

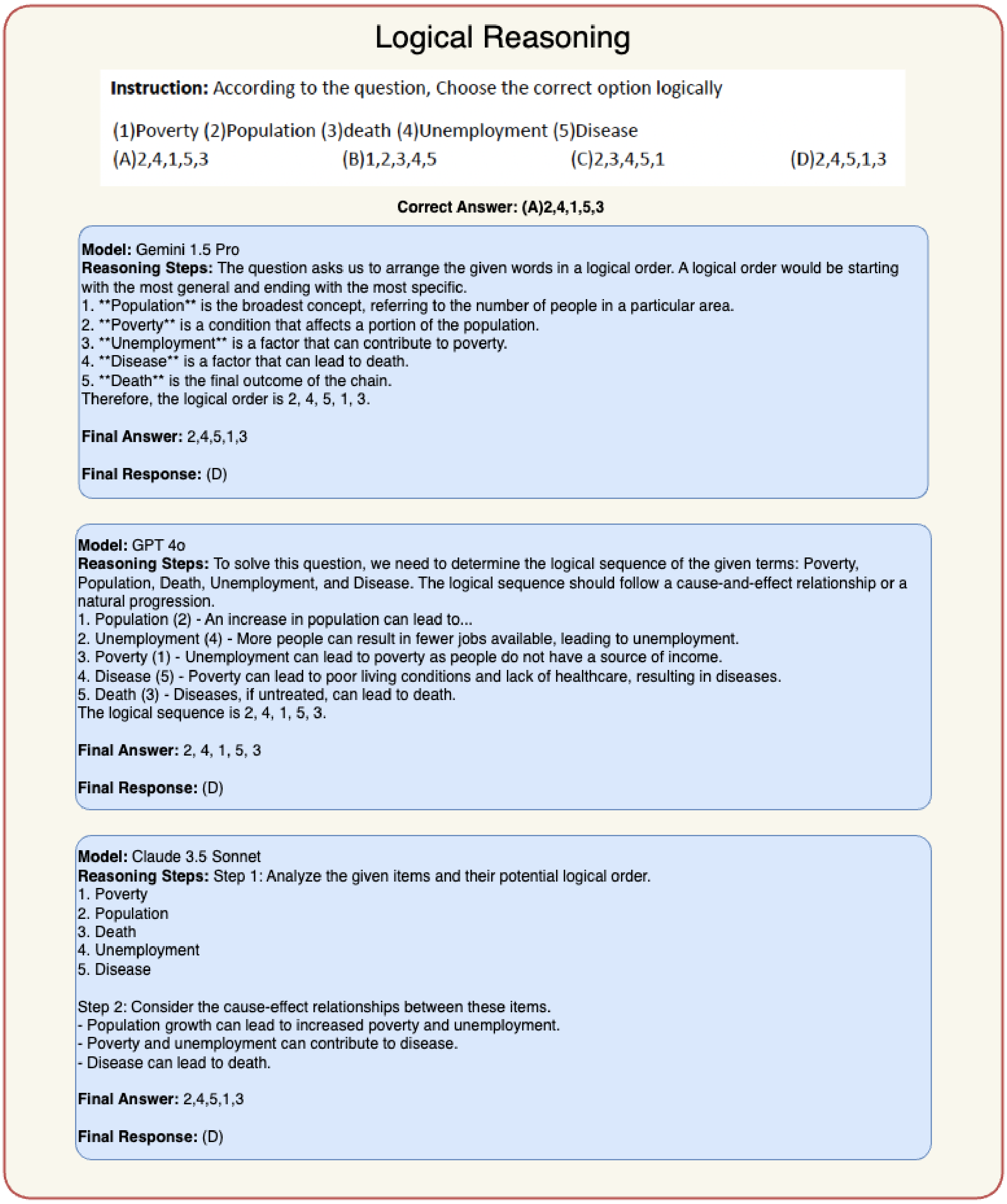
Model: Claude 3.5 Sonnet
Reasoning Steps: 1. Observe the pattern in the given figures:
- The top row shows an increasing number of black segments.
- The bottom left figure has four white segments.
- The bottom right figure has four black segments.

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 5: Erroneous model reasoning patterns observed on an FC question



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)

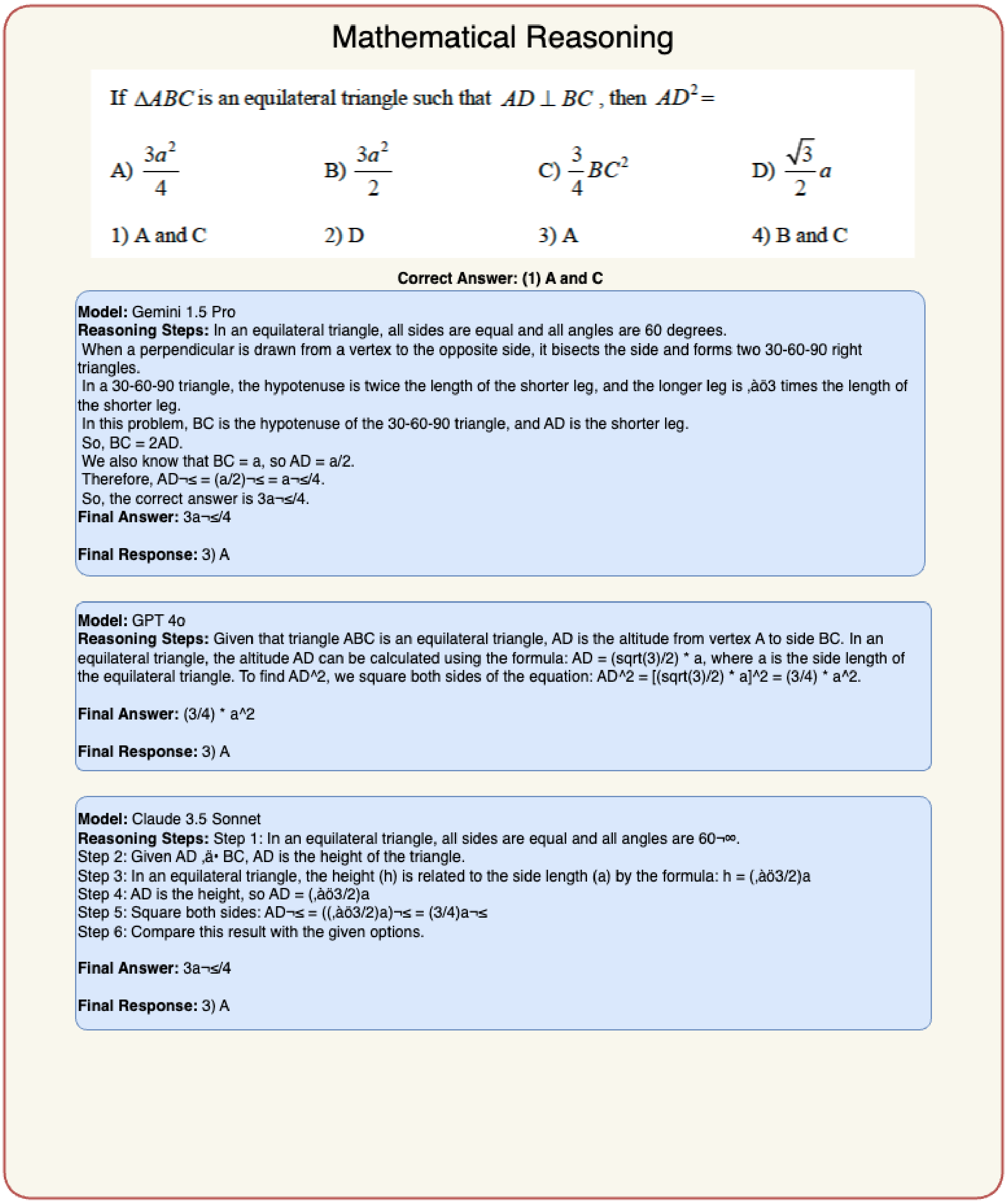
Model: Claude 3.5 Sonnet
Reasoning Steps: Step 1: Analyze the given items and their potential logical order.
1. Poverty
2. Population
3. Death
4. Unemployment
5. Disease

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 6: Erroneous model reasoning patterns observed on an LR question



Mathematical Reasoning

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

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 ‚àö3 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¬≤ = (a/2)¬≤ = a¬≤/4.
So, the correct answer is 3a¬≤/4.
Final Answer: 3a¬≤/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

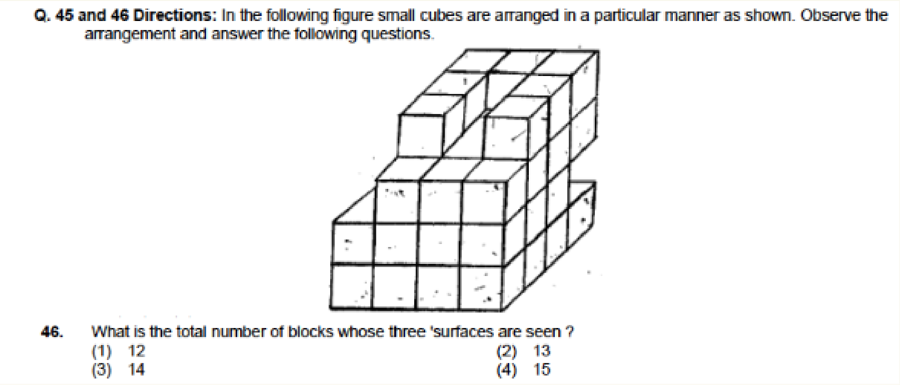
Model: Claude 3.5 Sonnet
Reasoning Steps: Step 1: In an equilateral triangle, all sides are equal and all angles are 60¬∞.
Step 2: Given AD ‚ä• BC, AD is the height of the triangle.
Step 3: In an equilateral triangle, the height (h) is related to the side length (a) by the formula: h = (‚àö3/2)a
Step 4: AD is the height, so AD = (‚àö3/2)a
Step 5: Square both sides: AD¬≤ = ((‚àö3/2)a)¬≤ = (3/4)a¬≤
Step 6: Compare this result with the given options.

Final Answer: 3a¬≤/4

Final Response: 3) A

Figure 7: Erroneous model reasoning patterns observed on an MR question

Numerical Reasoning



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

Correct Answer: (2) (13)

Model: Gemini 1.5 Pro
Reasoning Steps: The question asks to count the number of blocks whose three surfaces are seen. Looking at the image, we see that the blocks on the topmost layer (4 blocks in total) are the ones with three surfaces visible, so the answer is 4.

Final Answer: 4

Final Response: (4) 15

Model: GPT 4o
Reasoning Steps: To determine the number of blocks whose three surfaces are seen, we need to identify the corner blocks of the structure. These blocks are located at the vertices of the visible faces of the cube. In the given figure, the structure is a 3x3x3 cube with some blocks removed. We need to count the corner blocks that have three visible surfaces.

Step-by-step solution:
1. Identify the corner blocks on the top layer: There are 4 corner blocks on the top layer.
2. Identify the corner blocks on the bottom layer: There are 4 corner blocks on the bottom layer.
3. Identify the corner blocks on the middle layer: There are 4 corner blocks on the middle layer.
4. Sum the number of corner blocks: 4 (top) + 4 (bottom) + 4 (middle) = 12 blocks.

Final Answer: 12

Final Response: (1) 12

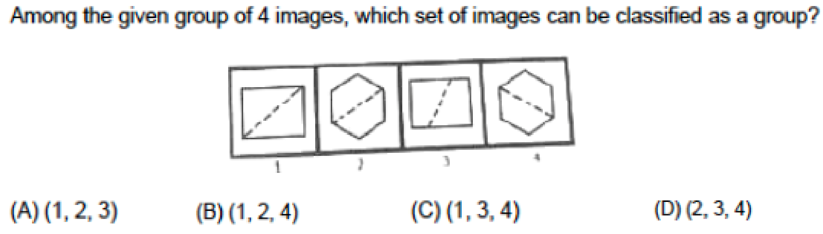
Model: Claude 3.5 Sonnet
Reasoning Steps: Step 1: Understand what 'three surfaces are seen' means. It refers to blocks that are visible from three sides: top, front, and side.
Step 2: Identify the blocks that meet this criteria. These will be the blocks at the corners of the structure that are not covered by other blocks.
Step 3: Count these blocks systematically:
- Top layer: 4 corner blocks
- Middle layer: 2 corner blocks (front right and back left)
- Bottom layer: 1 corner block (front right)\
Step 4: Sum up the total number of blocks with three visible surfaces: 4 + 2 + 1 = 7

Final Answer: 7

Final Response: (1)

Figure 8: Erroneous model reasoning patterns observed on an NR question

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)

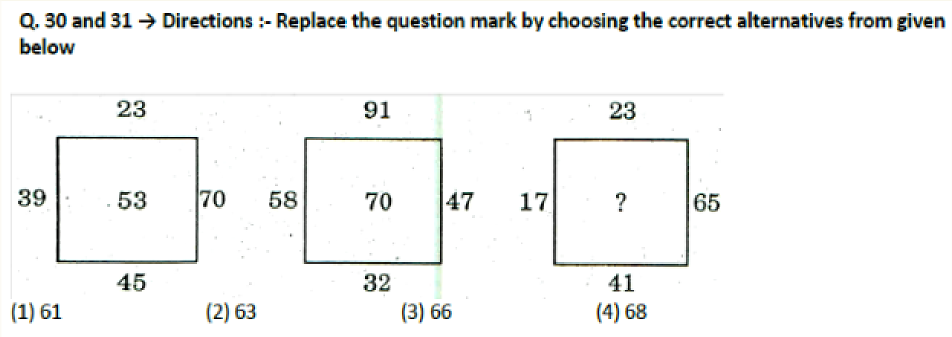
Model: Claude 3.5 Sonnet
Reasoning Steps: Step 1: Examine each image carefully.
Step 2: Identify the common characteristics in the images.
Step 3: Image 1 shows a rectangle with a diagonal line.
Step 4: Image 2 shows a rectangle with no additional lines.
Step 5: Image 3 shows a square with a diagonal 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 9: Erroneous model reasoning patterns observed on an OD question

Pattern Recognition



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

(1) 61
(2) 63
(3) 66
(4) 68

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

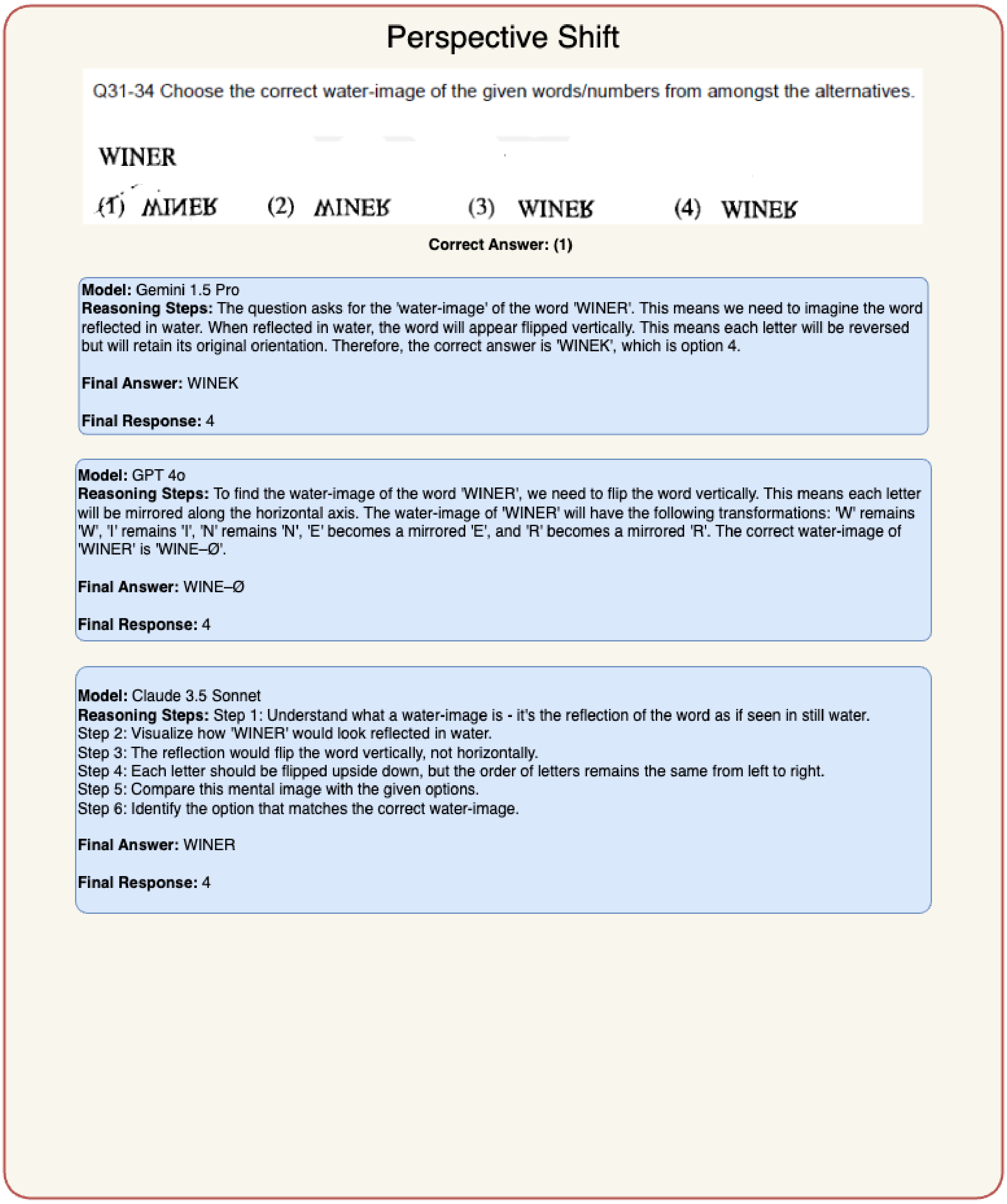
Final Response: 1

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 10: Erroneous model reasoning patterns observed on a PR question


Perspective Shift
WINER
(1)
(2)
(3)
(4)

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

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

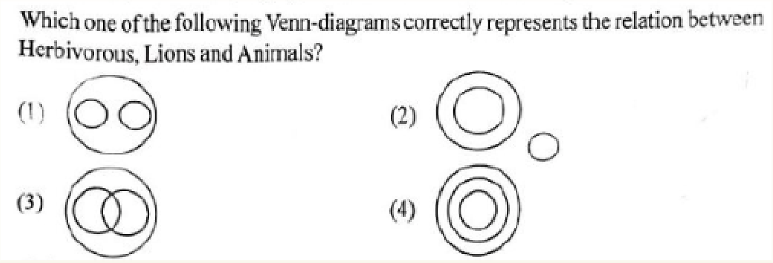
Model: Claude 3.5 Sonnet
Reasoning Steps: Step 1: Understand what a water-image is - it's the reflection of the word as if seen in still water.
Step 2: Visualize how 'WINER' would look reflected in water.
Step 3: The reflection would flip the word vertically, not horizontally.
Step 4: Each letter should 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 11: Erroneous model reasoning patterns observed on a PS question

Relative Reasoning


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

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.

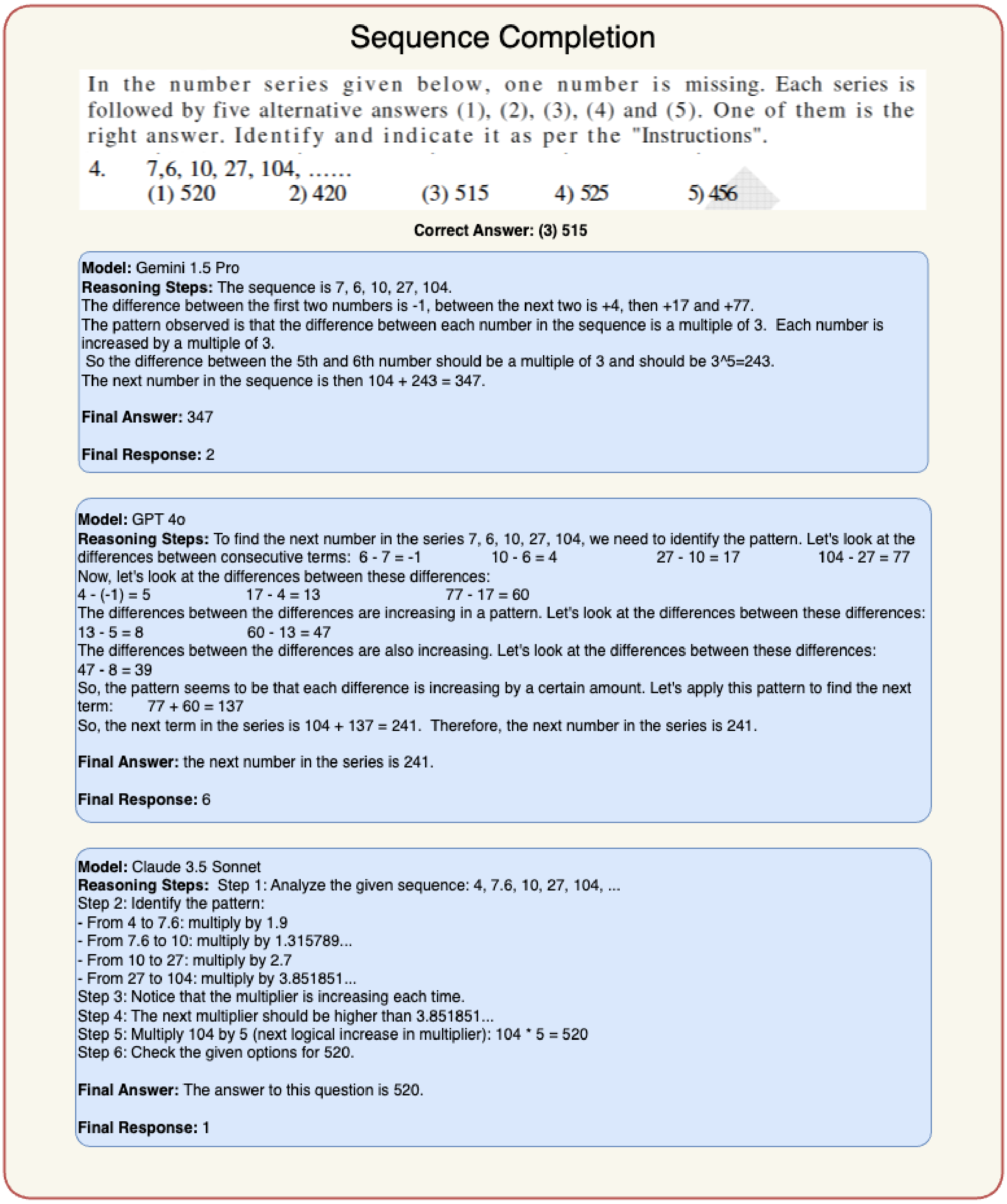
Step 2: Analyze each Venn diagram option:
(1) Two separate circles: This represents two distinct groups with no overlap, which is incorrect.
(2) One circle inside another, with a separate small circle: This is incorrect as it suggests one group is entirely contained within another, while the third is 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 12: Erroneous model reasoning patterns observed on an RR question



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

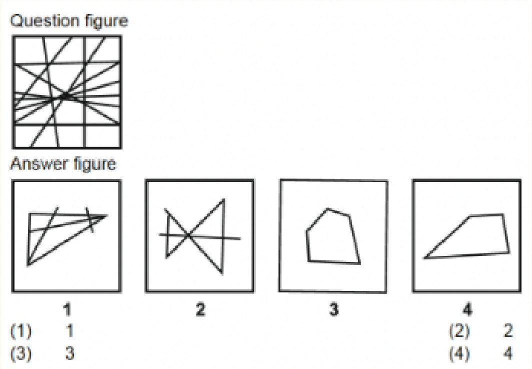
Model: Claude 3.5 Sonnet
Reasoning Steps: Step 1: Analyze the given sequence: 4, 7.6, 10, 27, 104, ...
Step 2: Identify the pattern:
- From 4 to 7.6: multiply by 1.9
- From 7.6 to 10: multiply by 1.315789...
- From 10 to 27: multiply by 2.7
- From 27 to 104: multiply by 3.851851...
Step 3: Notice that 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 13: Erroneous model reasoning patterns observed on an SC question

Spatial Reasoning



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 14: Erroneous model reasoning patterns observed on an SR question

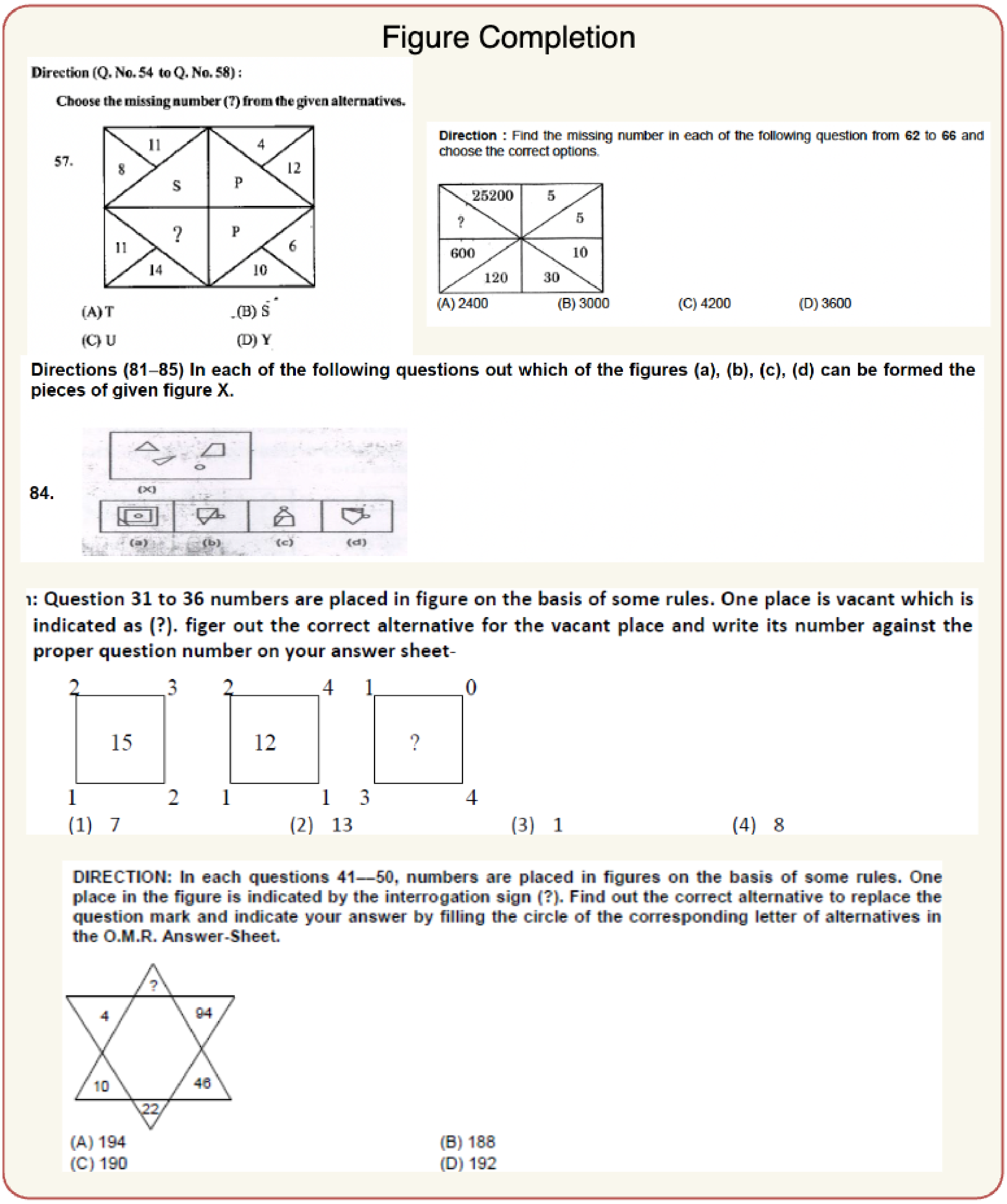


Figure Completion

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.

ı: 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-

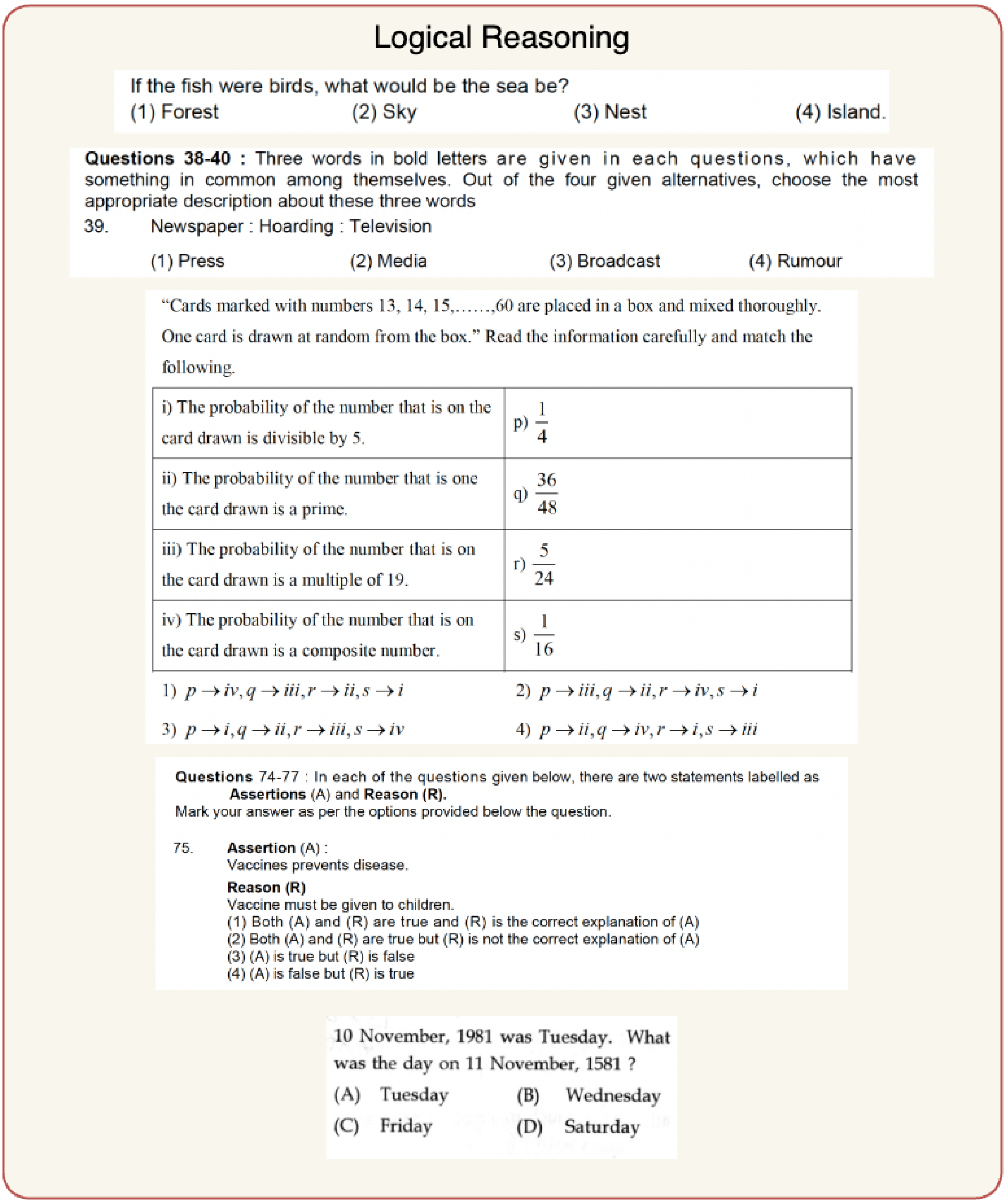
(1) 7 (2) 13 (3) 1 (4) 8

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 15: Questions belonging to the *figure_completion* (FC) category



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.

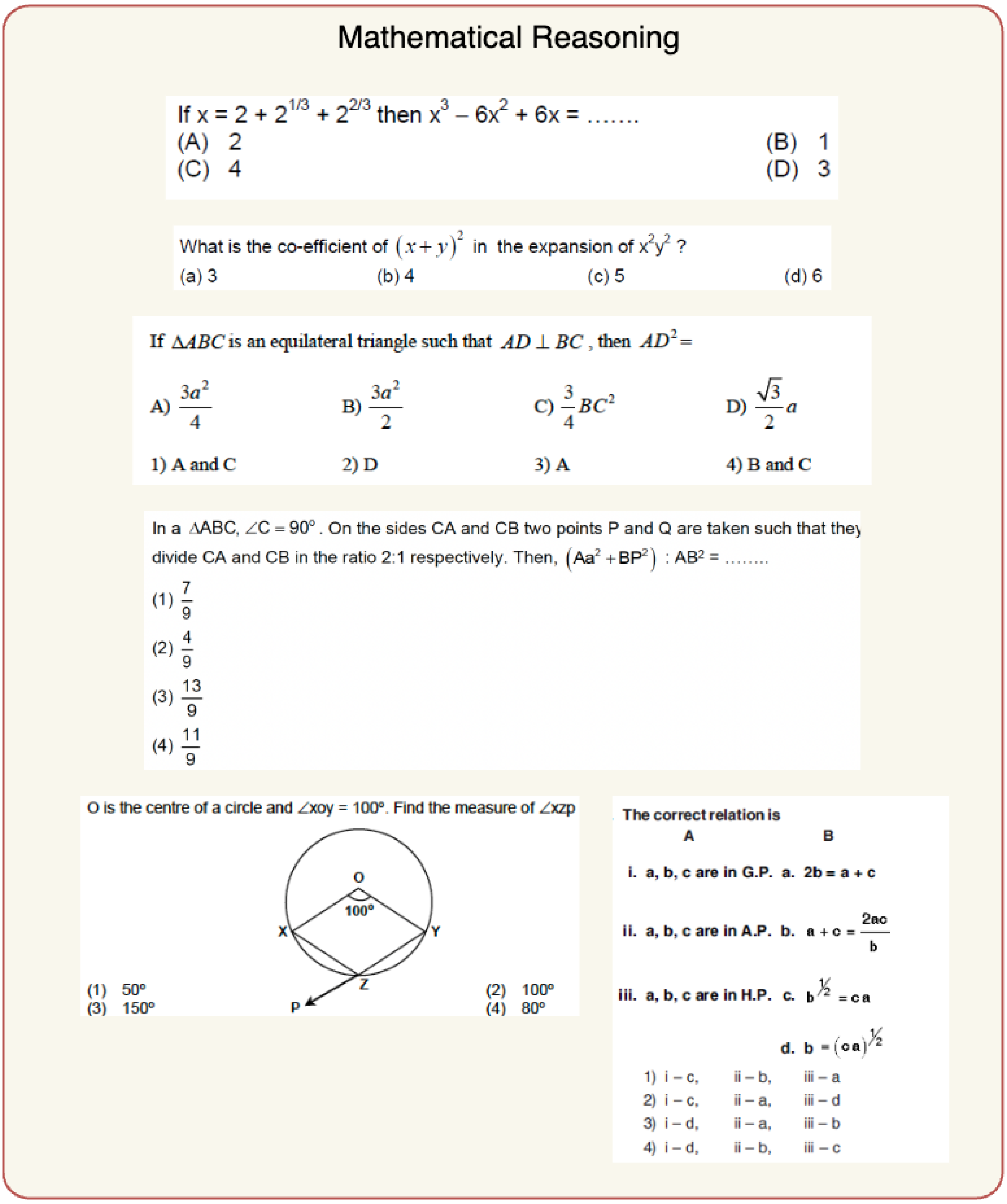
75. **Assertion** (A) :
Vaccines prevents disease.
Reason (R)
Vaccine must be given to children.
(1) Both (A) and (R) are true and (R) is the correct explanation of (A)
(2) Both (A) and (R) are true but (R) is not the correct explanation of (A)
(3) (A) is true but (R) is false
(4) (A) is false but (R)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

10 November, 1981 was Tuesday. What was the day on 11 November, 1581 ?

(A) Tuesday (B) Wednesday

(C) Friday (D) Saturday

Figure 16: Questions belonging to the *logical_reasoning* (LR) category



Mathematical Reasoning

If $x = 2 + 2^{1/3} + 2^{2/3}$ then $x^3 - 6x^2 + 6x =$ …….

(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, $(Aa^2 + BP^2)$: $AB^2 =$ ……..

(1) $\frac{7}{9}$

(2) $\frac{4}{9}$

(3) $\frac{13}{9}$

(4) $\frac{11}{9}$

O is the centre of a circle and ∠xoy = 100°. Find the measure of ∠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.	$b = (ca)^{1/2}$

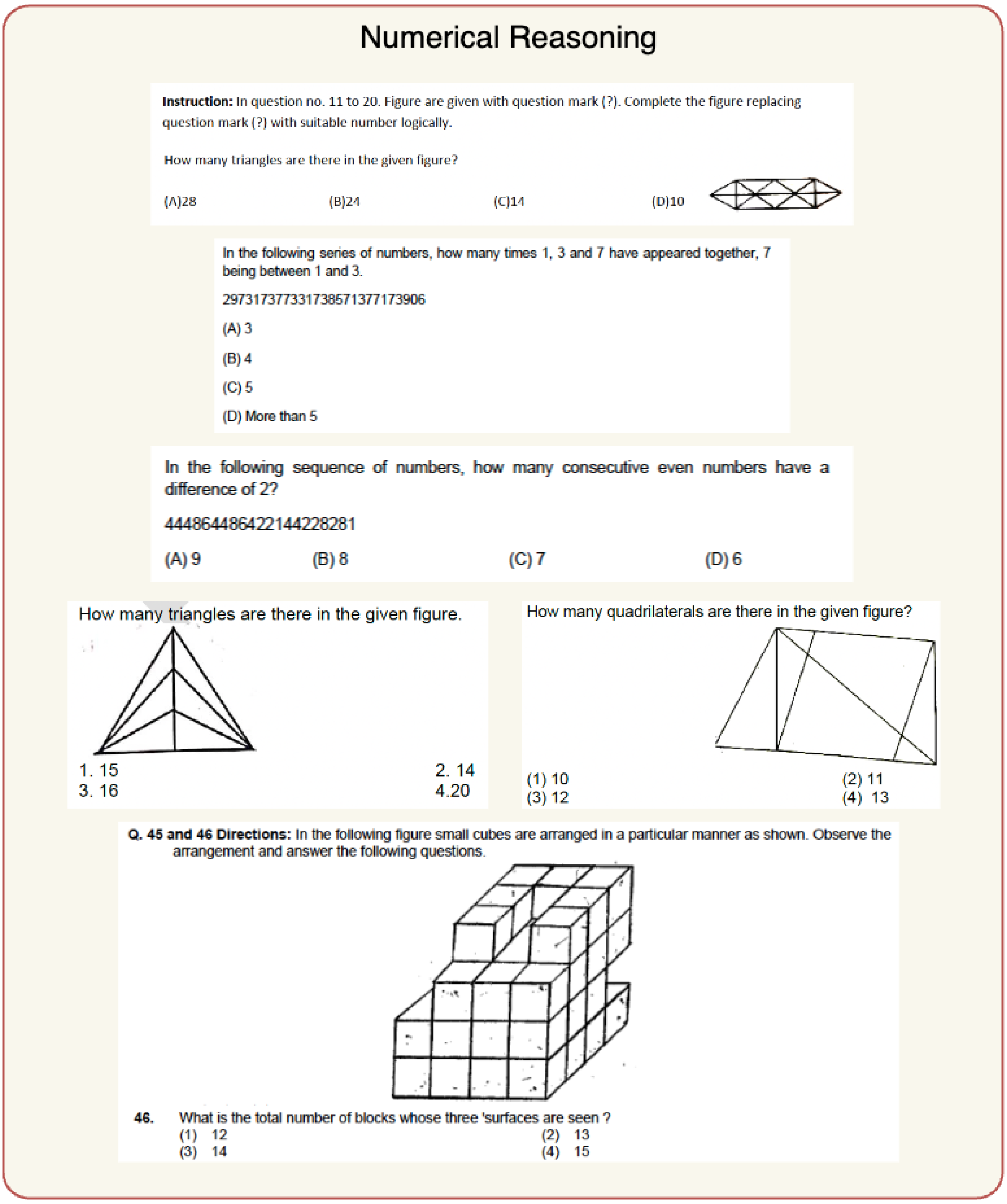
1) i – c, ii – b, iii – a

2) i – c, ii – a, iii – d

3) i – d, ii – a, iii – b

4) i – d, ii – b, iii – c

Figure 17: Questions belonging to the *mathematical_reasoning* (MR) category



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.

297317377331738571377173906

(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 2. 14
3. 16 4.20

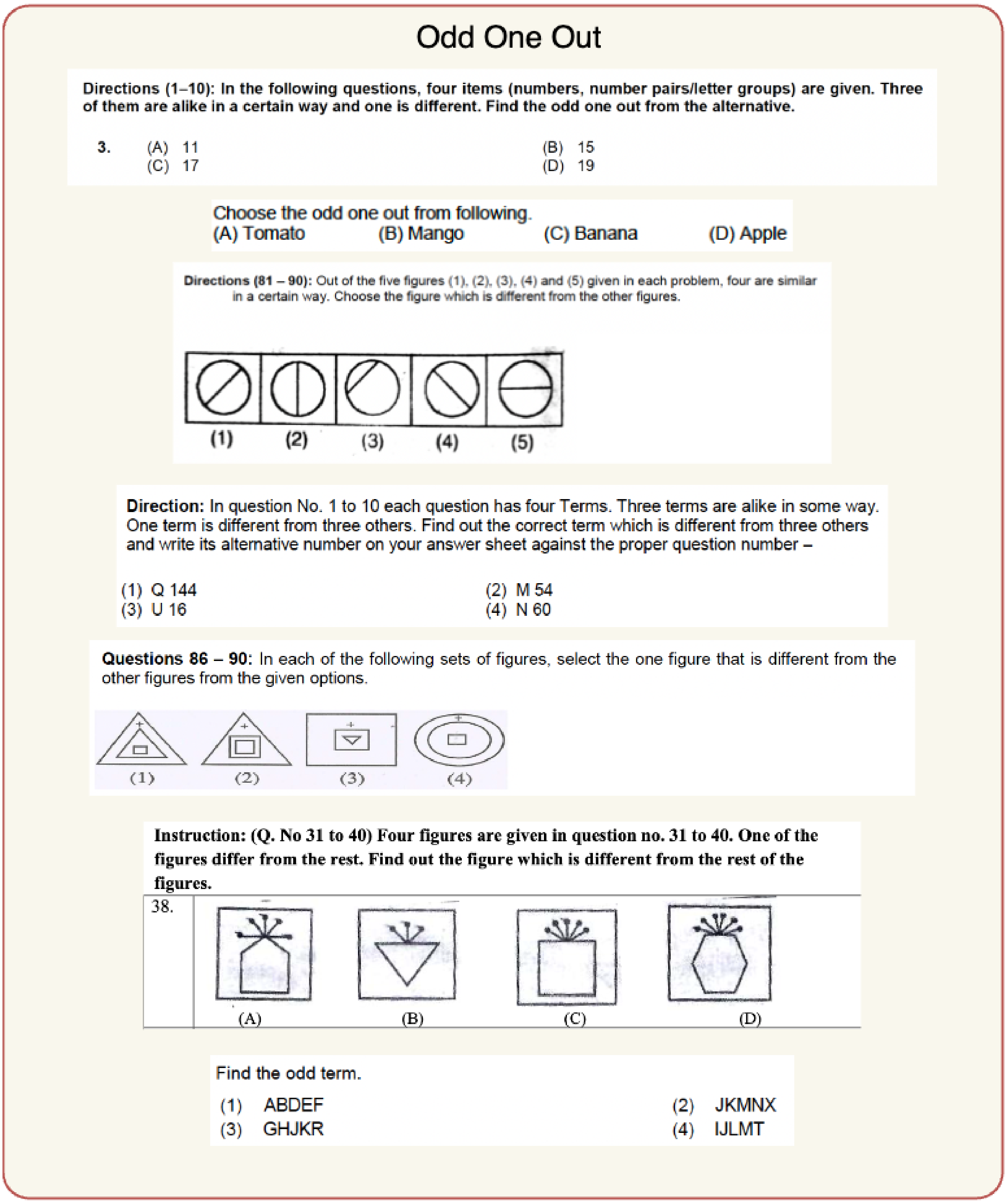
How many quadrilaterals are there in the given figure?

(1) 10 (2) 11
(3) 12 (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 18: Questions belonging to the *numerical_reasoning* (NR) category



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.

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

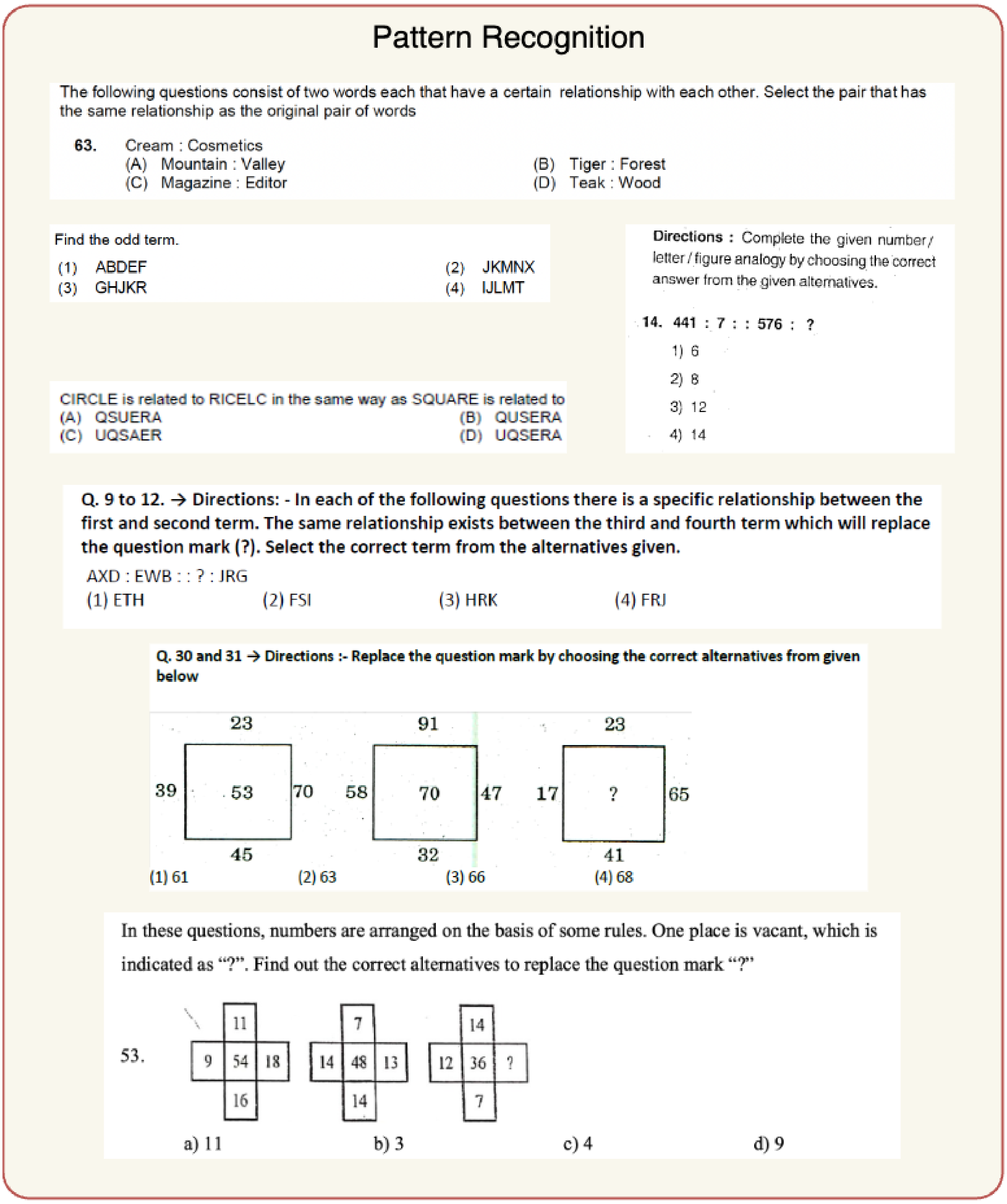
Questions 86 – 90: In each of the following sets of figures, select the one figure that is different from the other figures from the given options.

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.

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

Figure 19: Questions belonging to the *odd_one_out* (OD) category



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

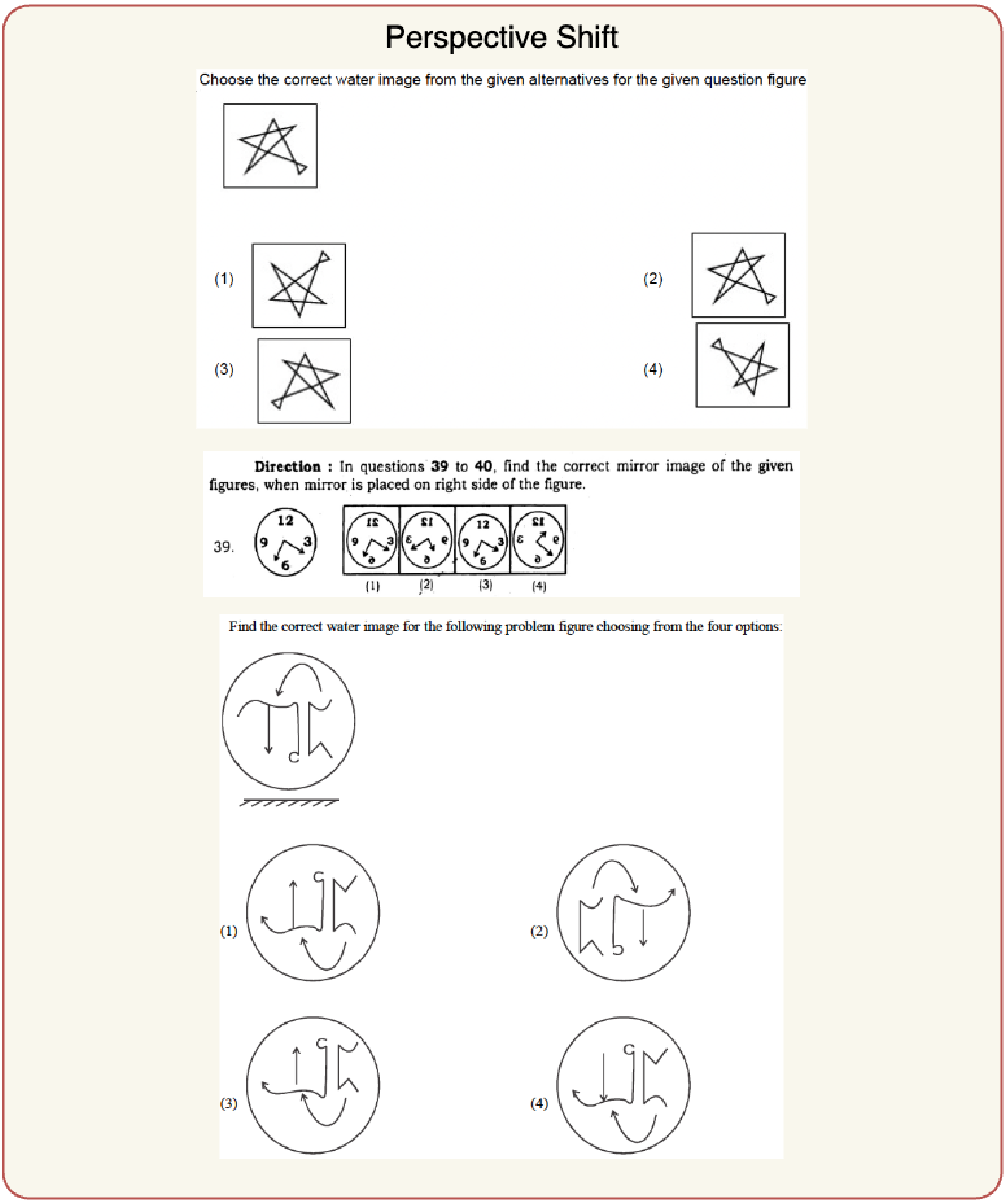
(1) 61 (2) 63 (3) 66 (4) 68

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.

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

Figure 20: Questions belonging to the *pattern_recognition* (PR) category



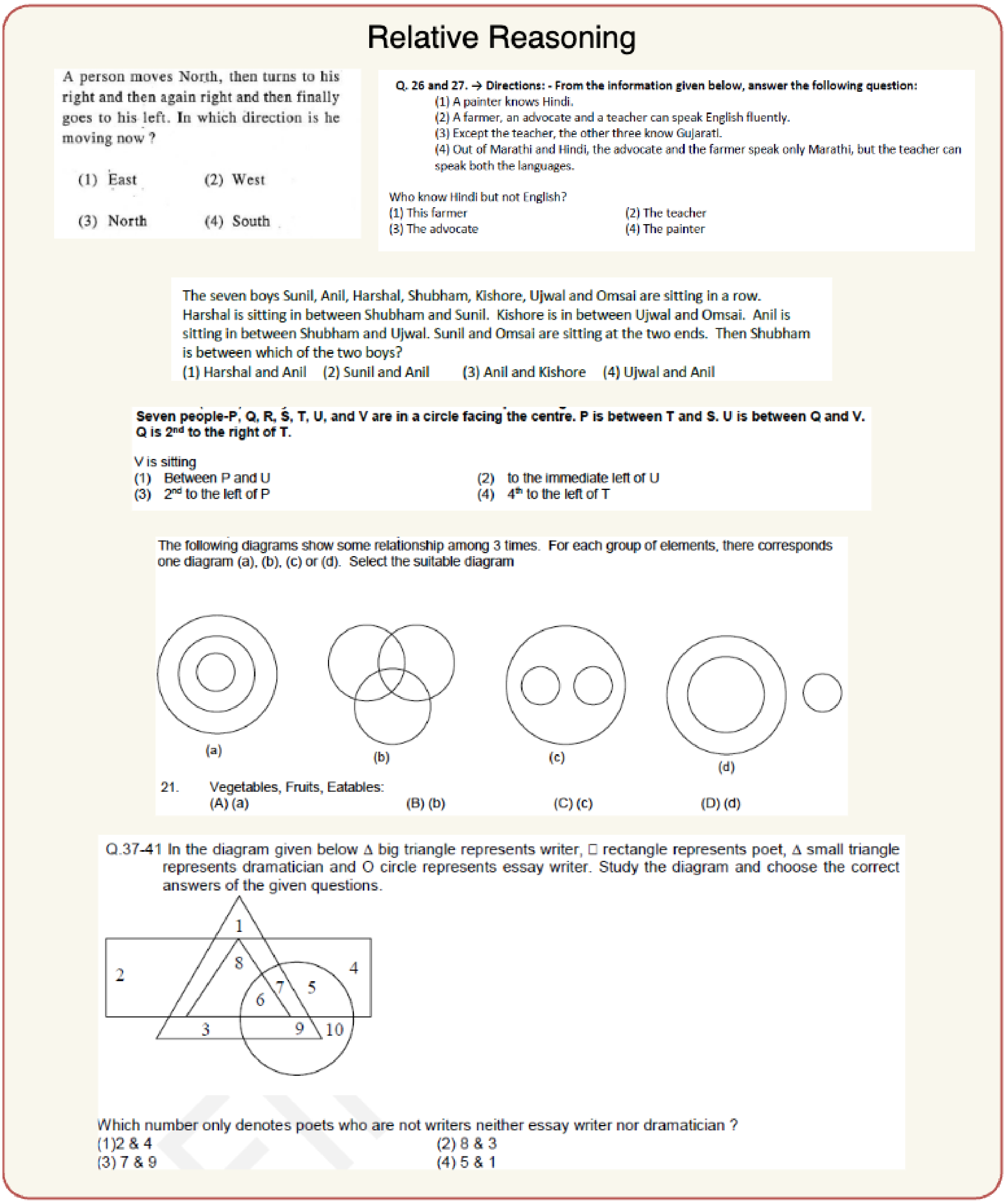
Perspective Shift

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

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.

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

Figure 21: Questions belonging to the *perspective_shift* (PS) category



Relative Reasoning

A person moves North, then turns to his right and then again right and then finally goes to his left. In which direction is he moving now ?

(1) East (2) West

(3) North (4) South

Q. 26 and 27. → Directions: - From the information given below, answer the following question:

(1) A painter knows Hindi.
(2) A farmer, an advocate and a teacher can speak English fluently.
(3) Except the teacher, the other three know Gujarati.
(4) Out of Marathi and Hindi, the advocate and the farmer speak only Marathi, but the teacher can speak both the languages.

Who know Hindi but not English?
(1) This farmer (2) The teacher
(3) The advocate (4) The painter

The seven boys Sunil, Anil, Harshal, Shubham, Kishore, Ujwal and Omsai are sitting in a row. Harshal is sitting in between Shubham and Sunil. Kishore is in between Ujwal and Omsai. Anil is sitting in between Shubham and Ujwal. Sunil and Omsai are sitting at the two ends. Then Shubham is between which of the two boys?
(1) Harshal and Anil (2) Sunil and Anil (3) Anil and Kishore (4) Ujwal and Anil

Seven people-P, Q, R, S, T, U, and V are in a circle facing the centre. P is between T and S. U is between Q and V. Q is 2nd to the right of T.

V is sitting
(1) Between P and U (2) to the immediate left of U
(3) 2nd to the left of P (4) 4th to the left of T

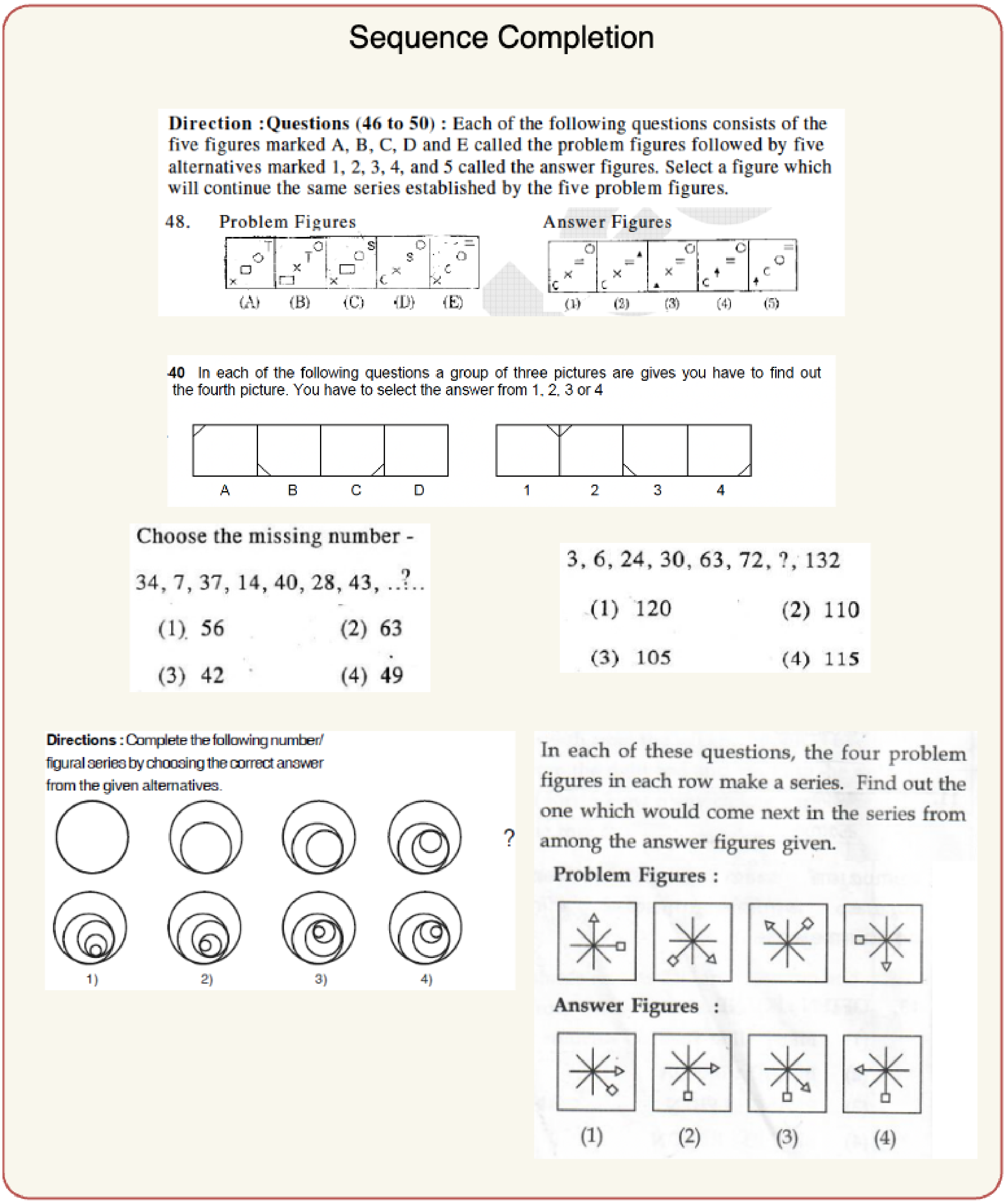
The following diagrams show some relationship among 3 times. For each group of elements, there corresponds one diagram (a), (b), (c) or (d). Select the suitable diagram

21. Vegetables, Fruits, Eatables:
(A) (a) (B) (b) (C) (c) (D) (d)

Q.37-41 In the diagram given below Δ big triangle represents writer, □ rectangle represents poet, Δ small triangle represents dramatician and O circle represents essay writer. Study the diagram and choose the correct answers of the given questions.

Which number only denotes poets who are not writers neither essay writer nor dramatician ?
(1)2 & 4 (2) 8 & 3
(3) 7 & 9 (4) 5 & 1

Figure 22: Questions belonging to the *relative_reasoning* (RR) category



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

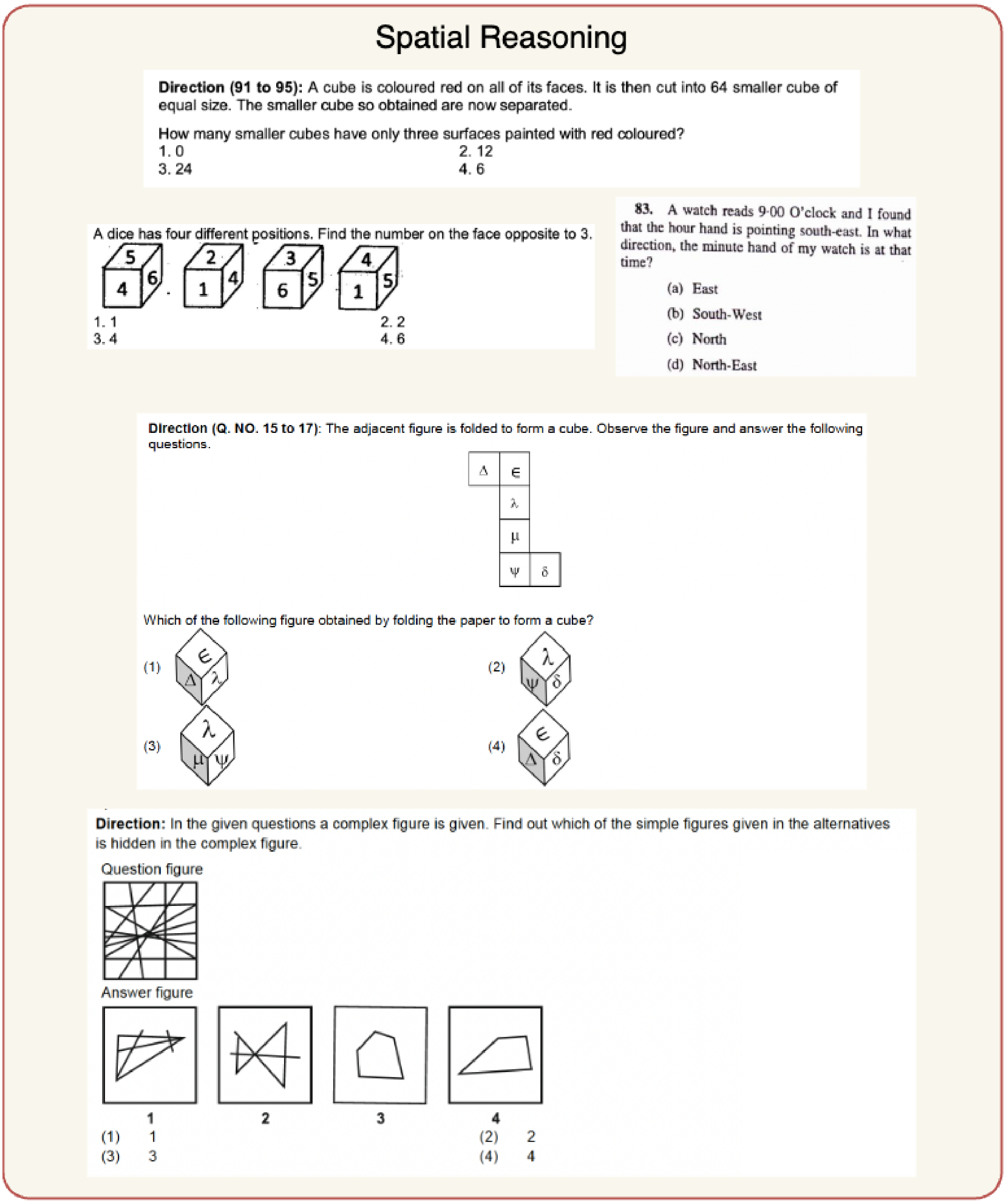
Directions : Complete the following number/ figural series by choosing the correct answer from the given alternatives.

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 23: Questions belonging to the *sequence_completion* (SC) category



Spatial Reasoning

Direction (91 to 95): A cube is coloured red on all of its faces. It is then cut into 64 smaller cube of equal size. The smaller cube so obtained are now separated.

How many smaller cubes have only three surfaces painted with red coloured?

1. 0 2. 12
3. 24 4. 6

A dice has four different positions. Find the number on the face opposite to 3.

1. 1 2. 2
3. 4 4. 6

83. A watch reads 9-00 O'clock and I found that the hour hand is pointing south-east. In what direction, the minute hand of my watch is at that time?

(a) East
(b) South-West
(c) North
(d) North-East

Direction (Q. NO. 15 to 17): The adjacent figure is folded to form a cube. Observe the figure and answer the following questions.

Which of the following figure obtained by folding the paper to form a cube?

(1) (2)
(3) (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.

Question figure

Answer figure

1 2 3 4

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

Figure 24: Questions belonging to the *spatial_reasoning* (SR) category

Row Labels	**Percent**
calc error	3.81
incomplete	0.00
logical flaw	63.55
memory	2.11
misalign	0.84
spatial	29.66

Table 23: Qwen2 VL (2B) Instruct (5) - Least performant open source model

Row Labels	**Percent**
calc error	2.54
incomplete	0.84
logical flaw	60.59
memory	4.23
misalign	3.38
spatial	28.38

Table 24: LLaVA-v1.6 Mistral (7B) (15) - Best performing open source model