

542 **Appendix**

543 **A Ethics Statement**

544 All questions of MATH-V are from publicly available sources. Although we do not commercialize
 545 MATH-V, we should like to demonstrate that we are far from the boundary for action or infringement.

546 **Legal compliance.** MATH-V problems are created by the the Mathematical Association of America
 547 (MAA), Kangaroo of Mathematics in Austria (KMA) and United Kingdom Maths Trust (UKMT). For
 548 decades, the MAA has not protected its problem IP even from separate organizations which sell MAA
 549 problems, such as AoPS. Courts have ruled that this implies the IP rights are permanently forfeited.
 550 Both of the KMA and UKMT are leading charities that advances the education of young people in
 551 mathematics, and MATH-V only contains free test questions published on their official website.

552 For MATH-V, we abide by Fair Use §107: “the fair use of a copyrighted work, including such use by
 553 ... scholarship, or research, is not an infringement of copyright”, where fair use is determined by
 554 “the purpose and character of the use, including whether such use is of a commercial nature or is for
 555 nonprofit educational purposes” and “the effect of the use upon the potential market for or value of
 556 the copyrighted work.”

557 **Dataset intended usage and license.** We document the MATH-V dataset within the paper and note
 558 that the MATH-V and code for results reproduction will be publicly available. The MATH-V dataset is
 559 not supposed, though the risk exists, to be used to train models for cheating. We intend for researchers
 560 to use this dataset to better evaluate LMMs’ mathematical reasoning capabilities and consequently
 561 facilitate future studies in this area. We bear all responsibility in case of violation of rights. The
 562 MATH-V data and our open-source code are under an MIT license.

563 **B Dataset Statistics**

Statistic	Number
Total questions	3,040
- multiple-choice questions	1,532 (50.4%)
- Free-form questions	1,508 (49.6%)
- Questions in testmini	304 (10.0%)
Multiple-choice questions	1,532
- Proportion of answers A	299(19.5%)
- Proportion of answers B	311(20.3%)
- Proportion of answers C	320(20.9%)
- Proportion of answers D	326(21.2%)
- Proportion of answers E	278(18.1%)
Unique number of images	3,472
Unique number of answers	370
Levels	5
Subjects	16
Source competitions	19
- Math Kangaroo	1,499
- UK (Grey,Pink,Junior,...)	535
- AMC (8,10,12), AIME	759
Maximum question length	280
Minimum question length	5
Average question length	42.3

Table 4: Key statistics of MATH-V. The unit of question length is words.

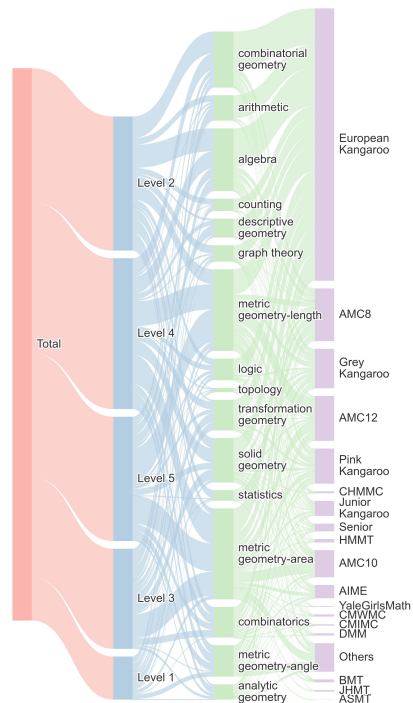


Figure 6: Levels, subjects and sources distribution of MATH-V.

564 **C More Results on Text-only LLMs**

Table 5: Performance of text-only LLMs with image captions.

Model	Overall	Alg	AnaG	Ari	CombG	Comb	Cnt	DescG	GrphT	Log	Angle	Area	Len	SolG	Star	Topo	TransG
LLMs (Text + Image Caption)																	
ChatGPT-3.5 (CoT)	9.74	10.7	20.0	18.6	10.1	7.7	17.9	16.4	10.0	13.5	6.4	5.8	6.5	9.4	12.1	4.4	10.7
GPT-4 (CoT)	13.10	16.5	20.2	34.3	10.4	17.9	19.4	7.7	11.1	10.1	9.8	9.6	9.1	13.5	13.8	8.7	12.5
MathCoder-7B	6.6	4.9	15.5	12.1	9.1	6.0	7.5	11.5	5.6	7.6	2.3	4.0	4.9	5.3	8.6	13.0	11.3
WizardMath-7B	4.5	3.5	7.1	5.0	8.4	0.6	10.5	10.5	4.4	5.0	1.7	3.6	2.7	4.5	3.5	13.0	5.4
MAmmoTH-7B-Mistral	12.4	11.6	16.7	15.7	11.4	10.1	11.9	14.4	14.4	8.4	8.7	12.8	11.4	11.5	19.0	21.7	16.6

565 **D Data Contamination**

566 We have also considered the issue of data contamination:

567 Firstly, we primarily collect questions from the official websites of various competitions. Each file
 568 typically contains 15 to 30 questions and is kept separate from the answers or solutions. Given the
 569 concerns surrounding data contamination in large language models (LLMs), we prioritize selecting
 570 questions that do not have readily available answers, such as those found on separate documents or
 571 pages.

572 Secondly, the three main series of closed-source models, GPT-4, Gemini, and Claude, all evaluate
 573 their models using several contests in their technical reports, including the AMC, which is a primary
 574 source of our problems. Therefore, we assume that these models adhere to the common practice of
 575 not using leaked data for training. However, as these are closed-source models, this assumption is our
 576 best estimation.

577 Thirdly, among open-source models, the most widely used base model is LLaVA, which utilizes
 578 a subset of the CC-3M dataset for pretraining and LAION/CC/SBU BLIP captions for instruction
 579 tuning. We compared our dataset with the training dataset and found no overlap between the two.

580 **E Testmini**

581 In this section, we show the effectiveness of the testmini subset in mirroring the performance trends
 582 observed in the full test set across a diverse range of mathematical subjects. The detailed results
 583 are presented in Table 6. We find that the model performances, when evaluated on both the entire
 584 set and testmini, exhibit close parallelism in trend. For instance, the performance delta between the
 585 full test set and testmini for the Qwen-VL-Plus model is a mere 0.19%, indicating a high degree of
 586 similarity in model behavior across the two datasets. Similarly, the GPT4V model also exhibits a
 587 negligible performance difference of -0.39%, further corroborating the testmini subset’s capability
 588 to accurately reflect the full test set’s characteristics. Furthermore, even for LLaVA-v1.5-13B and
 589 Share-GPT4V-7B, two models with the largest performance variations, where differences of 1.98%
 590 and 1.97% are observed respectively, the deviation remains within 2%, underscoring the testmini
 591 subset’s reliability for performance evaluation. These findings suggest that models’ performance on
 592 the testmini subset should be indicative of their performance on the full test set, thereby providing a
 593 reliable means for preliminary evaluation and comparison of models across various mathematical
 594 subjects.

595 **F Evaluation Details**

596 We conduct all experiments on A100 80G GPUs.

597 **F.1 Prompts for Response Generation**

598 In this section, we present details of the prompts used for the evaluation of response generation
 599 models. The prompts are designed to ensure that the models generate responses that are consistent

Model	Prompt
GPT4-CoT (Text Only)	Please solve the problem step by step and put your answer in one " <input type="checkbox"/> ". If it is a multiple choice question, only one letter is allowed in the " <input type="checkbox"/> ".
ChatGPT 3.5-CoT (Text + Image Caption)	Please first solve the problem step by step, then put your final answer or a single letter (if it is a multiple choice question) in one " <input type="checkbox"/> ".
GPT4-CoT (Text + Image Caption)	Please first solve the problem step by step, then put your final answer or a single letter (if it is a multiple choice question) in one " <input type="checkbox"/> ".
LLaVa-v1.5-7B	Answer the question using a single word or phrase.
SPHINX (V2)	Answer the question using a single word or phrase.
ShareGPT4V-7B	Answer the question using a single word or phrase.
LLaVa-v1.5-13B	Answer the question using a single word or phrase.
ShareGPT4V-13B	Answer the question using a single word or phrase.
SPHINX-MoE	Answer the question using a single word or phrase.
InternVL-Chat-V1-2-Plus	Please solve the problem and put your answer in one " <input type="checkbox"/> ". If it is a multiple choice question, only one letter letter is allowed in the " <input type="checkbox"/> ".
InternLM-XComposer2-VL	Please solve the problem and put your answer in one " <input type="checkbox"/> ". If it is a multiple choice question, only one letter letter is allowed in the " <input type="checkbox"/> ".
InternLM-XComposer2-VL-CoT	Please solve the problem step by step and put your answer in one " <input type="checkbox"/> ". If it is a multiple choice question, only one letter is allowed in the " <input type="checkbox"/> ".
Qwen-VL-Plus	Answer the question using a single word or phrase.
Qwen-VL-Max	Answer the question using a single word or phrase.
Qwen-VL-Max-CoT	Please solve the problem step by step and put your answer in one " <input type="checkbox"/> ". If it is a multiple choice question, only one letter is allowed in the " <input type="checkbox"/> ".
Gemini-1.5 Pro	Answer the question using a single letter or number or word.
Gemini Pro	Answer the question using a single letter or number or word.
Gemini Pro-CoT	Please first solve the problem step by step, then put your final answer or a single letter (if it is a multiple choice question) in one " <input type="checkbox"/> ".
GPT4-V/o/turbo	Please solve the problem and put your answer in one " <input type="checkbox"/> ". If it is a multiple choice question, only one letter is allowed in the " <input type="checkbox"/> ".
GPT4V-CoT	Please solve the problem step by step and put your answer in one " <input type="checkbox"/> ". If it is a multiple choice question, only one letter is allowed in the " <input type="checkbox"/> ".

Table 7: The prompts used in the evaluation for response generation.

Model	Source	URL
ChatGPT 3.5	gpt-3.5-turbo-0125	https://platform.openai.com/
GPT4	gpt-4-0125-preview	https://platform.openai.com/
GPT4V	gpt-4-1106-vision-preview	https://platform.openai.com/
GPT-4o	gpt-4o-2024-05-13	https://platform.openai.com/
GPT-4-turbo	gpt-4-turbo-2024-04-09	https://platform.openai.com/
Gemini Pro	gemini-pro-vision	https://ai.google.dev/
Gemini-v1.5 Pro	gemini-1.5-pro	https://ai.google.dev/
InternLM-XComposer2-VL	local checkpoint	https://huggingface.co/internlm/internlm-xcomposer2-vl-7b
InternVL-Chat-v1.2-Plus	local checkpoint	https://huggingface.co/OpenGVLab/InternVL-Chat-V1-2-Plus
Qwen-VL-Max	qwen-vl-max	https://help.aliyun.com/zh/dashscope/developer-reference/vl-plus-quick-start
Qwen-VL-Plus	qwen-vl-plus	https://help.aliyun.com/zh/dashscope/developer-reference/vl-plus-quick-start
SPHINX-MoE	local checkpoint	https://huggingface.co/Alpha-VLLM/LLaMA2-Accessory/tree/main/finetune/mm/SPHINX/SPHINX-MoE
SPHINX (V2)	local checkpoint	https://huggingface.co/Alpha-VLLM/LLaMA2-Accessory/tree/main/finetune/mm/SPHINX/SPHINX-v2-1k
LLaVa-v1.5-7B	local checkpoint	https://huggingface.co/liuhaotian/llava-v1.5-7b
LLaVa-v1.5-13B	local checkpoint	https://huggingface.co/liuhaotian/llava-v1.5-13b
ShareGPT4V-7B	local checkpoint	https://huggingface.co/Lin-Chen/ShareGPT4V-7B
ShareGPT4V-13B	local checkpoint	https://huggingface.co/Lin-Chen/ShareGPT4V-13B

Table 8: The source of the models used in the evaluation.

610 **G Comparison Details**

611 **G.1 Comparing with MathVista**

Template	Proportion	Examples
IconQA[51]		
Object Counting	106/400	What number is shown?
Perimeter Calculation	51/400	What is the perimeter of the square?
Length Measurement with a given reference	150/400	Move the ruler to measure the length of the line to the nearest centimeter. The line is about _ centimeters long.
Time Inquiry based on a clock	67/400	What time is shown?
CLEVR-Math[52]		
Subtract ... How many objects are left?	400/400	Subtract all metallic cylinders. Subtract all gray things. How many objects are left?
FigureQA[53]		
Is X the ...?	176/400	Is Khaki the minimum/maximum/ low median/high median/smoothest/ roughest?
Does X have the ...?	86/400	Does Brown have the maximum/ minimum area under the curve? (the highest/lowest value?)
Is X ... Y?	117/400	Is Hot Pink less/greater than Black?
Does X intersect Y?	21/400	Does Slate intersect Light Salmon?
Super-CLEVEL[54]		
Are there ... X than Y?	146/400	Are there fewer/more tiny gray cars than small red sedans?
Is the number of X ... than Y?	232/400	Is the number of regular buss less/greater than the number of brown rubber minivans?
Others	22/400	How many cyan double buss have the same size as the matte aeroplane?
KVQA[55]		
What is the age gap between the ...?	310/382	What is the age gap between these two people in image?
How many people in the image were born ...?	72/382	How many people in the image were born after the end of World War II?
PlotQA[56]		
In how many years, is the X greater than Y?	57/105	In how many years, is the population in the largest city greater than 26 %?
What is the difference between ...?	31/105	What is the difference between two consecutive major ticks on the Y-axis?
Others	17/105	What is the average rating of statistical capacity in Bolivia per year?

Table 9: Some source datasets of MathVista featuring similar questions. The table shows the template, proportion within the whole dataset and typical examples of them.

612 MathVista contains numerous similar questions with restricted question variability across a variety of
 613 source datasets. After eliminating questions with identical stem text, only 4,740 questions remain.
 614 For instance, among the 6,141 questions in MathVista, the question "What is the age gap between
 615 these two people in the image?" appears 270 times. Some images of this question are shown in

616 Figure 7. Furthermore, there are many template questions within its collection that merely alter a few
 617 words. For instance, questions containing "What is the age gap between" appear 310 times. Another
 618 example is questions like "Subtract all ... objects. How many objects are left?". Figure 8 illustrates
 619 this type of question involving very similar geometry recognition, which occurs 400 times. In Table 9,
 620 we present detailed statistics on similar questions.

621 We also provide some qualitative comparisons illustrating the different question lengths between
 622 MathVista and our dataset. Figure 9 presents the 3 main types of questions featuring an abstract
 623 scene in MathVista, constituting over 90% of the total questions, as discussed in Section. 1. We also
 624 showcase representative questions in MathVista that involve a function plot in Figure 10, which pose
 625 similar, straightforward, and concise inquiries regarding the depicted function graph. For comparison,
 626 we present examples of our MATH-V on analytic geometry in Figure 11, which feature longer questions
 627 and incorporate more complex function concepts, such as mutual symmetry and iterative functions.
 628 Moreover, Figure 12 and Figure 13 exhibit topology and graph theory-related questions in our dataset,
 629 two categories absent in MathVista, which require intricate visual recognition and mathematical
 reasoning.

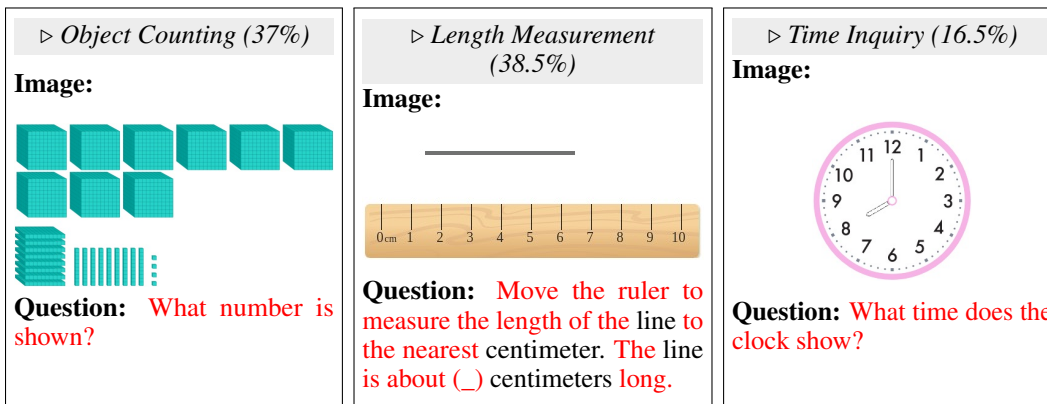


Figure 9: Three main types of questions among the 375 questions with an abstract scene in MathVista: Object Counting, Length Measurement, and Time Inquiry. Among the 200 sampled questions, the percentage of each category is in the figure. Text marked in red indicates that it belongs to a fixed template and is used in other similar questions.

630

631 G.2 Comparing with MMMU

632 In this section, we showcase some examples of different subjects of MMMU demanding sophisticated
 633 college-level domain knowledge. Figure 14, Figure 15, and Figure 16 present the subject of Graph
 634 Theory, Group Theory, and Operation Research, respectively. Moreover, typical examples of Topology
 635 and Graph Theory in our MATH-V are displayed in Figure 13 and Figure 13. Although both sets of
 636 subjects are intricate mathematical disciplines, they exhibit notable differences in the nature of their
 637 posed questions. Questions in the MMMU dataset often involve advanced college-level mathematical
 638 concepts such as Kruskal's algorithm, alternating group structures, and DFS Spanning Tree. In
 639 contrast, the questions in our dataset predominantly comprise puzzles that are readily solvable by
 640 middle and high school students.

641 H More Dataset Details

642 H.1 Question Distribution

643 All questions in MATH-V are written in English. As illustrated in Appendix B, the average number of
 644 words in a question is 42.3, with a maximum of 280 words and a minimum of 5 words. Figure 17
 645 further elucidates the distribution of word counts, highlighting the diverse patterns of questions.



Figure 7: Some images with the question "What is the age gap between these two people in the image?" from MathVista. Questions containing "What is the age gap between" appear 310 times.



Figure 8: Problem involving very similar geometry recognition from MathVista. Questions like "Subtract all ... objects. How many objects are left?" appear 400 times.

<p>▷ <i>Function symmetry, etc.</i></p> <p>Image:</p> <p>Question1: Is this an odd function? (A) Yes (B) No</p> <p>Question2: How many zeros does this function have?</p> <p>Question3: What is the value of $f(0)$?</p> <p>Question4: Is this function differentiable at each point?</p> <p>Question5: $f(2)$ is ____ zero.</p> <p>Question6: The derivative of this function at $x=0$ is ____</p> <p>Question7: What is this function most likely be? (A) a polynomial (B) a trigonometric function (C) an exponential function (D) a logarithmic function</p>	<p>▷ <i>Function monotonicity, etc.</i></p> <p>Image:</p> <p>Question1: Which function is monotonic in range $[0, \pi]$? (A) the red one (B) the blue one (C) both (D) none of them</p> <p>Question2: At $x=1$, the value of the red curve is ____ that of the blue curve.</p> <p>Question3: Which function is even?</p> <p>Question4: Which function is odd?</p> <p>Question5: Which function is periodic?</p> <p>Question6: What is the value of the red curve at $x=0$?</p> <p>Question7: What is the value of the blue curve at $x=0$?</p>	<p>▷ <i>Function roots, etc.</i></p> <p>Image:</p> <p>Question1: How many zeros does this function have?</p> <p>Question2: What is the zero of this function?</p> <p>Question3: The derivative of y at $x=6$ is ____ that at $x=8$.</p> <p>Question4: What is the value of y at $x=2$?</p> <p>Question5: What is the value of y at $x=1$?</p> <p>Question6: If this function invertible?</p> <p>Question7: What is this function most likely be? A a polynomial B a trigonometric function C an exponential function D a logarithmic function</p>
---	--	--

Figure 10: Typical examples of 400 questions with a function plot of MathVista. Each image has approximately seven questions about function symmetry, monotonicity, roots, periodicity, values, etc.

646 We can see that the median word count and the average word count are close, indicating that the
 647 distribution is not skewed. The distribution of word counts is near a normal distribution, with more
 648 than 98.75% of questions having a word count of less than 100.

649 H.2 Division of Levels

650 In this section, we detail the division of MATH-V's difficulty levels. Our difficulty classification is
 651 divided into two steps. The first step involves roughly dividing the difficulty levels based on the
 652 grades of the contestants as specified in the competition. The second step entails manually checking
 653 and correcting any unreasonable divisions.

654 **Step 1.** The existing competition's grade requirements span grades 1 to 12. Because the questions
 655 for grades 1 and 2 are considered too simple, we classify them separately as level 1. The questions
 656 for grades 3, 4, 5, and 6 are similar in difficulty, so we classify them as level 2. Grades 7 and 8 are
 657 categorized as level 3, grades 9 and 10 as level 4, and grades 11 and above as level 5.

658 **Step 2.** We manually review the questions in each level and adjust the classification according to the
 659 actual difficulty of the questions. This step is necessary because, within the same exam, the difficulty
 660 range of the questions can be quite broad. For instance, the most difficult question in grade 4 might
 661 be more challenging than the easiest question in grade 5. In such cases, we make manual adjustments.

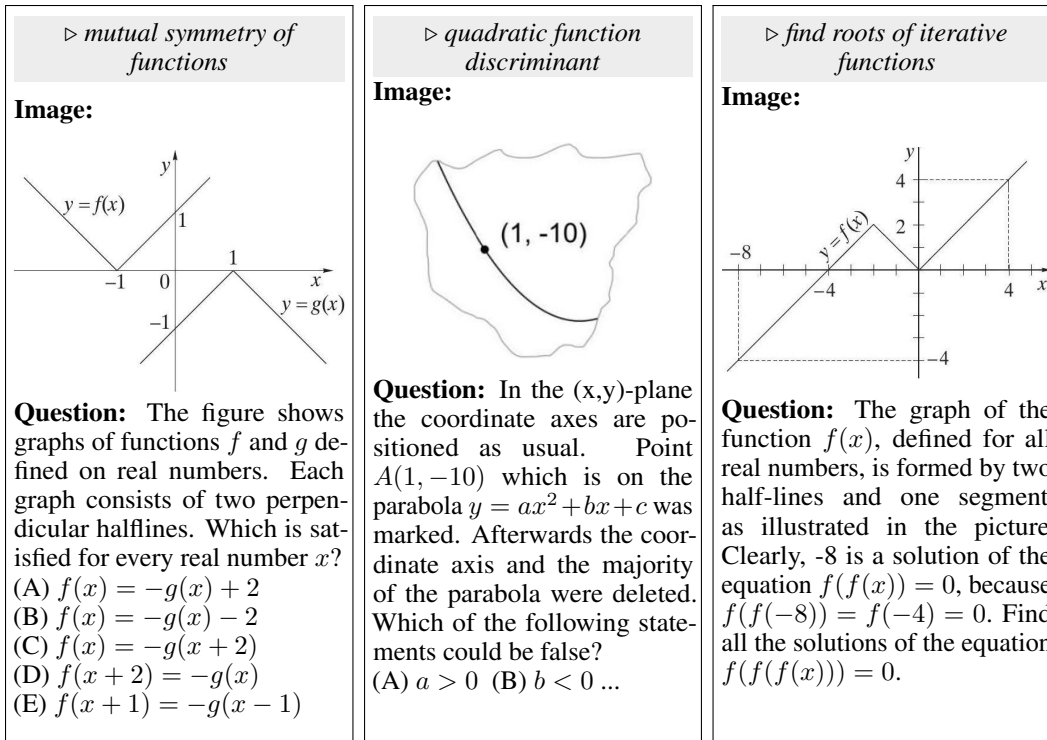


Figure 11: Three examples from MATH-V on analytic geometry, which are about mutual symmetry of functions, quadratic function discriminant, and roots of iterative functions. We omit some choices for brevity.

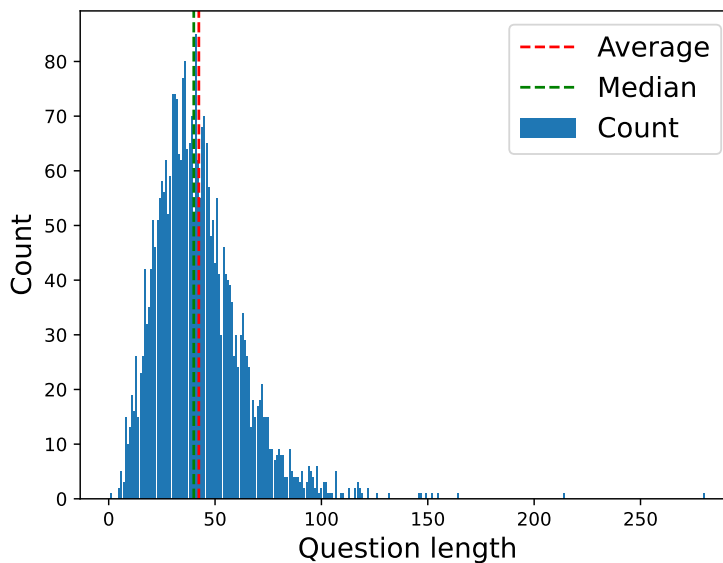



Figure 17: The distribution of the number of words per question in MATH-V.

662 Figure 18.a shows a comparison of the level distribution between our MATH-V and the MATH dataset.
 663 We observe that both datasets have the fewest number of questions in level 1 and a similar number of
 664 questions in the other levels. Figure 18.b illustrates the models' accuracy across different levels. It is


▷ *String Knot Probability Puzzle*

Image:

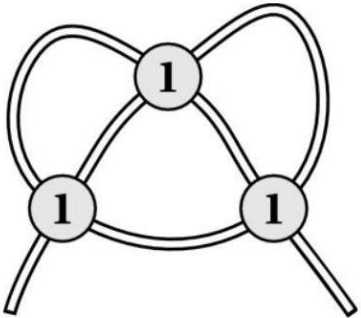
<image1>



<image2>



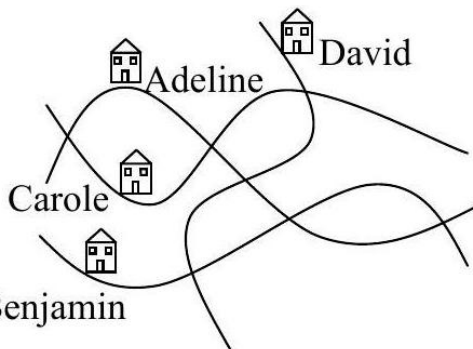
<image3>



Question: A piece of string is lying on the table. It is partially covered by three coins as seen in the figure. Under each coin the string is equally likely to pass over itself like this: <image1> or like this: <image2>. What is the probability that the string is knotted after its ends are pulled?

▷ *Village Map Sketching Challenge*

Image:

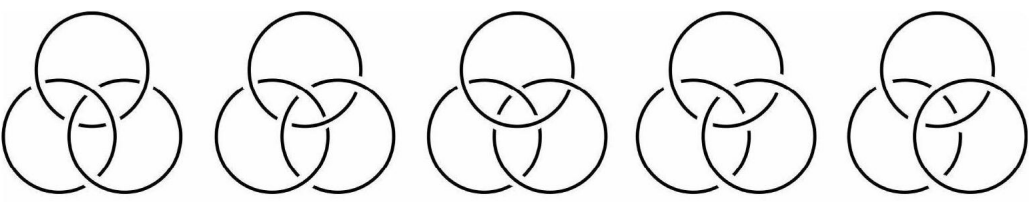


Question: During a rough sailing trip, Jacques tried to sketch a map of his village. He managed to draw the four streets, the seven places where they cross and the houses of his friends. The houses are marked on the correct streets, and the intersections are correct, however, in reality, Arrow Street, Nail Street and Ruler Street are all absolutely straight. The fourth street is Curvy Street. Who lives on Curvy Street?

(A) Adeline
 (B) Benjamin
 (C) Carole
 (D) David
 (E) It is impossible to tell without a better map

▷ *Borromaic Rings*

Image:



A) B) C) D) E)

Question: The "Borromaic Rings" have an extraordinary property. Although no two are interlocked, they are strongly connected within each other. If one ring is cut through, the other two fall apart. Which of the following diagrams shows the picture of "Borromaic Rings"?

Figure 12: Three examples from MATH-V on topology. We omit some choices for brevity.

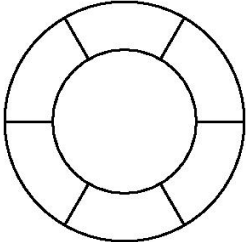
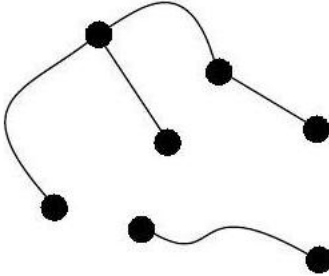
<p style="text-align: center;">▷ Adjacent Coloring Puzzle</p> <p>Image:</p> <table border="1" style="width: 100%; border-collapse: collapse; text-align: center;"> <tr><td>A</td><td>B</td><td></td><td></td><td></td></tr> <tr><td>C</td><td>D</td><td></td><td></td><td></td></tr> <tr><td></td><td></td><td>B</td><td></td><td></td></tr> <tr><td></td><td></td><td></td><td></td><td style="background-color: #cccccc;"></td></tr> <tr><td>B</td><td></td><td></td><td></td><td></td></tr> </table> <p>Question: In the diagram we want to colour the fields with the colours A, B, C D so that adjacent fields are always in different colours. (Even fields that share only one corner, count as adjacent.) Some fields have already been coloured in. In which colour can the grey field be coloured in?</p>	A	B				C	D						B								B					<p style="text-align: center;">▷ Coloring Options</p> <p>Image:</p>  <p>Question: The figure below shows a ring made of six small sections which you are to paint on a wall. You have four paint colors available and will paint each of the six sections a solid color. Find the number of ways you can choose to paint each of the six sections if no two adjacent section can be painted with the same color.</p>	<p style="text-align: center;">▷ Optimal Line Additions</p> <p>Image:</p>  <p>Question: The picture shows seven points and the connections between them. What is the least number of connecting lines that could be added to the picture so that each of the seven points has the same number of connections with other points? (Connecting lines are allowed to cross each other.)</p>
A	B																										
C	D																										
		B																									
B																											

Figure 13: Three examples from MATH-V on graph theory. We omit some choices for brevity.

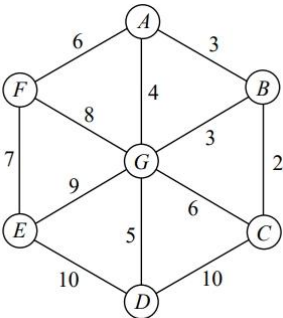
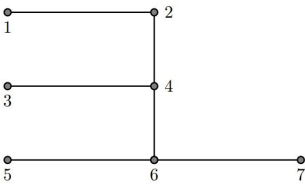
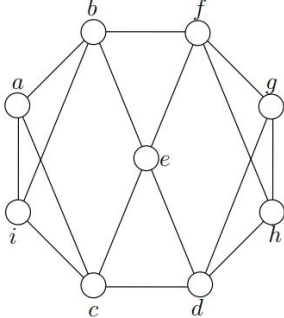
<p style="text-align: center;">▷ Kruskal's Algorithm</p> <p>Image:</p>  <p>Question: Using Kruskal's algorithm find a minimal spanning tree of <image 1>. Show the list of chosen edges only in alphabet order. Is 'AB, AF, BC, BG, DG, EF' the correct answer? (A) Yes (B) No</p>	<p style="text-align: center;">▷ Prüfer Code</p> <p>Image:</p>  <p>Question: Give the Prüfer code of <image 1> (A) 2,4,4,6,6 (B) 2,3,3,6,6 (C) 2,2,4,6,6 (D) 2,4,4,6,7</p>	<p style="text-align: center;">▷ DFS Spanning Tree</p> <p>Image:</p>  <p>Question: Apply DFS Spanning Tree Algorithm to <image 1> with vertex pre-ordering a, b, c, e, i, h, g, d, f. Show the list of chosen edges only in alphabet order. (A) ab,be,ci,cd,ce,dh,fg,gh (B) ab,ai,be,bf,cd,ci,fg,fh</p>
---	---	--

Figure 14: Three examples from the Graph Theory subfield in the Math subset of MMMU.

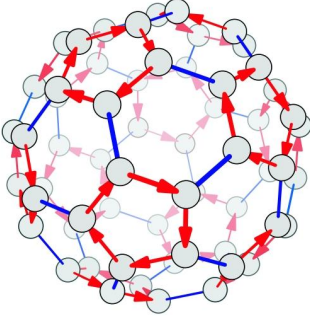
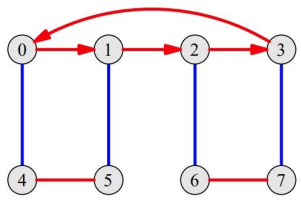
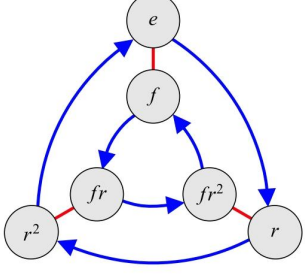
<p>▷ Alternating Group</p> <p>Image:</p>  <p>Question: Is <image 1> an Alternating group? (A) Yes (B) No</p>	<p>▷ Cayley Diagram</p> <p>Image:</p>  <p>Question: Is <image 1> a Cayley diagram? (A) Yes (B) No</p>	<p>▷ Symmetric Group</p> <p>Image:</p>  <p>Question: Is <image 1> a Symmetric group? (A) Yes (B) No</p>
---	--	--

Figure 15: Three examples from the Group Theory subfield in the Math subset of MMMU.

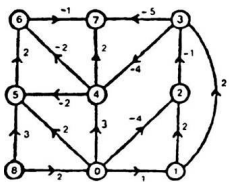
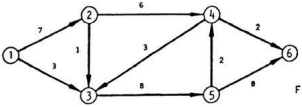
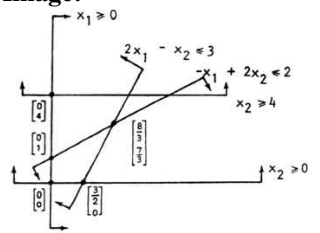
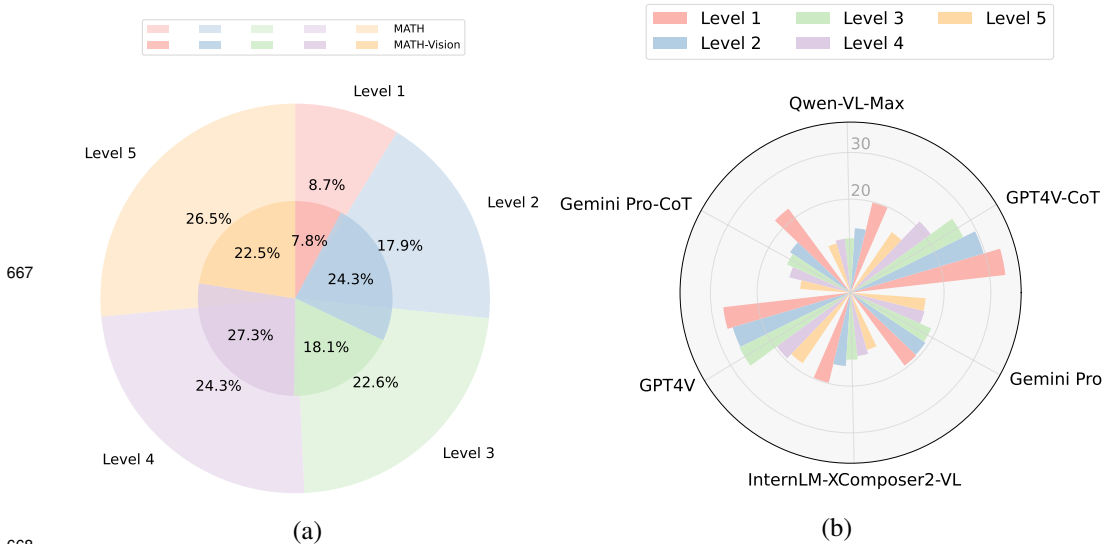
<p>▷ Graph Minimum Path</p> <p>Image:</p>  <p>Question: Find the minimum path from v_0 to v_7 in the graph G of figure 1. Notice that it has no circuit whose length is negative. <image 1> (A) $(v_0, v_1, v_2, v_3, v_4, v_5, v_6, v_7)$ with length 10 (B) $(v_0, v_2, v_3, v_4, v_6, v_7)$ with length -12 (C) $(v_0, v_1, v_3, v_5, v_6, v_7)$ with length 8 (D) $(v_0, v_2, v_4, v_5, v_6, v_7)$ with length 2</p>	<p>▷ Ford and Fulkerson Algorithm</p> <p>Image:</p>  <p>Question: Consider the network shown in Fig. 1. The problem is to maximize the flow from node 1 to node 6 given the capacities shown on the arcs. Solve by Ford and Fulkerson algorithm. <image 1> (A) 5 units (B) 7 units (C) 9 units (D) 11 units</p>	<p>▷ Linear Programming</p> <p>Image:</p>  <p>Question: Consider the following problem: Minimize $2x_1 + 3x_2$ Subject to $x_1 + 2x_2 \leq 2, 2x_1 - x_2 \leq 3, x_2 \geq 4, x_1, x_2 \geq 0$. Solve by the graphical approach. <image 1> (A) The problem has a unique feasible solution. (B) The problem has multiple feasible solutions. (C) The problem is infeasible. (D) The problem is unbounded.</p>
--	--	---

Figure 16: Three examples from the Operation Research subfield in the Math subset of MMMU.

665 evident that the accuracy of the models decreases as the difficulty level increases, which aligns with
 666 the expectation that the difficulty of the questions escalates with the level.



668 Figure 18: (a) Comparison of the level distribution between our MATH-V and the MATH dataset [33].
 669 (b) Accuracy on MATH-V on six typical LMMs across 5 difficulty levels.

669 H.3 Images by Subjects

670 In this section, we demonstrate some images from the problems in our dataset. Figure 20, Figure 21,
 671 Figure 22, Figure 23, Figure 24, Figure 25, Figure 26, Figure 27, Figure 28, Figure 29, Figure 30,
 672 Figure 31, Figure 32, Figure 33, Figure 34 and Figure 35 show images from the problems under
 673 the category of Algebra, Analytic Geometry, Arithmetic, Combinatorial Geometry, Combinatorics,
 674 Counting, Descriptive Geometry, Graph Theory, Logic, Metric Geometry - Angle, Metric Geometry -
 675 Area, Metric Geometry - Length, Solid Geometry, Statistics, Topology and Transformation Geometry,
 676 respectively.

677 We can see that the images are diverse among different categories, and the images in the same
 678 category have similar visual patterns. For example, the images in the category of Statistics are mainly
 679 various types of charts and tables, while the images in the category of Arithmetic and Counting,
 680 which are categories mainly for primary school students, are more colorful and have more cartoon
 681 characters or real-world objects, and, conversely, the images in the category of Analytic Geometry
 682 are more complex and abstract.

683 H.4 Introduction of Subjects

684 **Algebra.** Algebra is a branch of mathematics that deals with symbols and the rules for manipulating
 685 these symbols. It is a unifying thread of almost all of mathematics and includes everything from
 686 solving simple equations to studying abstractions such as groups, rings, and fields. Algebra allows
 687 us to express relationships and solve problems where we don't initially know the values of certain
 688 quantities. It serves as the foundation for more advanced topics in mathematics and sciences, enabling
 689 the formulation and solving of equations and inequalities.

690 **Analytic Geometry.** Analytic Geometry, also known as coordinate geometry, merges algebra and
 691 geometry to discuss shapes and figures using a coordinate system. This branch of mathematics allows
 692 the representation of geometric shapes in a numerical way by translating them into equations and vice
 693 versa. It enables the study of geometric properties and relationships by means of algebraic equations,
 694 thus providing a powerful tool for solving geometric problems.

695 **Arithmetic.** Arithmetic is the most basic branch of mathematics, dealing with the study of numbers
696 and the traditional operations on them: addition, subtraction, multiplication, and division. It forms
697 the foundation upon which more advanced mathematical concepts are built. Arithmetic is used
698 in everyday life for counting, measuring, and performing calculations. It is the first type of math
699 that students are introduced to and remains fundamental for understanding more complex areas of
700 mathematics.

701 **Combinatorial Geometry.** Combinatorial Geometry is a field of mathematics that studies geometric
702 objects and their arrangement, with a particular focus on discrete and combinatorial aspects. It
703 involves the investigation of geometric structures and their properties under various combinatorial
704 processes, such as counting, arrangement, and partitioning. This branch often intersects with areas
705 like graph theory and combinatorics, exploring problems related to the configuration of points, lines,
706 and other geometric figures.

707 **Combinatorics.** Combinatorics is a branch of mathematics concerned with counting, arrangement,
708 and combination of objects. It deals with questions of how many different ways such objects can
709 be chosen, arranged, and combined under certain conditions. Combinatorics has applications in
710 numerous areas, including computer science, optimization, and probability theory. It is fundamental
711 in developing analytical skills and understanding complex mathematical concepts.

712 **Counting.** Counting is the most basic and fundamental concept in mathematics, involving the
713 enumeration of objects in a set. It serves as the foundation for more advanced mathematical concepts,
714 such as arithmetic and combinatorics. Counting is not only about determining the total number
715 of items in a collection but also includes understanding and applying principles like one-to-one
716 correspondence, the counting principle, permutations, and combinations.

717 **Descriptive Geometry.** Descriptive Geometry is the branch of geometry that allows the representation
718 of three-dimensional objects in two dimensions, through the use of projections or perspective drawings.
719 It is a method of visually communicating complex geometric shapes and is fundamental in fields
720 such as engineering, architecture, and design. Descriptive geometry provides tools and techniques for
721 solving spatial problems and for understanding the spatial relationships between objects.

722 **Graph Theory.** Graph Theory is a branch of mathematics that studies graphs, which are mathematical
723 structures used to model pairwise relations between objects. A graph is made up of vertices (or
724 nodes) and edges (or arcs) that connect them. Graph theory is used to study networks and their
725 properties, such as paths, cycles, and connectivity, and has applications in computer science, biology,
726 transportation, social sciences, and more.

727 **Logic.** Logic is the foundation of mathematical reasoning, focusing on the study of argument validity,
728 structure, and form. It involves the analysis of propositions and their relationships to determine the
729 truth values of statements. Logic is used to construct and analyze mathematical proofs, ensuring that
730 conclusions follow logically from premises. It encompasses various systems, such as propositional
731 logic, predicate logic, and modal logic, each with its own rules and applications.

732 **Metric Geometry - Angle.** Metric Geometry concerning angles focuses on the properties and
733 measurements of angles within various geometric figures. It involves understanding the types of
734 angles (acute, right, obtuse, and straight), their relationships, and how they can be measured and
735 used to determine other geometric properties. This aspect of metric geometry is crucial in solving
736 problems related to shapes, patterns, and designs in both two and three dimensions.

737 **Metric Geometry - Area.** Metric Geometry in the context of area deals with the measurement of
738 the surface enclosed within geometric figures. It involves calculating the extent of two-dimensional
739 shapes, such as squares, rectangles, triangles, and circles, using various formulas. Understanding area
740 is essential in numerous practical applications, including architecture, land surveying, and crafting,
741 where the determination of space is necessary.

742 **Metric Geometry - Length.** Metric Geometry concerning length involves the measurement of
743 distance between points in geometric figures. It is the foundation for determining the perimeter of
744 shapes, the distance around a figure, and the concept of metric spaces in more advanced mathematics.

745 Length measurements are fundamental in various real-world applications, such as construction,
746 fabrication, and navigation.

747 **Solid Geometry.** Solid Geometry is the branch of mathematics that deals with the study of three-
748 dimensional figures, including spheres, cubes, cylinders, and pyramids, among others. It involves
749 understanding the properties, measurements, and relationships of these figures, such as volume,
750 surface area, and the angles between intersecting planes. Solid geometry is essential in fields such as
751 engineering, architecture, and physics, where three-dimensional models are crucial.

752 **Statistics.** Statistics is the branch of mathematics that deals with collecting, analyzing, interpreting,
753 and presenting data. It involves the use of theories of probability to make inferences about populations
754 from samples. Statistics is applied in a wide range of disciplines, including economics, social sciences,
755 biology, and engineering, to make decisions or predictions based on data analysis.

756 **Topology.** Topology is a branch of mathematics concerned with the properties of space that are
757 preserved under continuous transformations, such as stretching and bending, but not tearing or gluing.
758 It is often referred to as "rubber-sheet geometry" because of its focus on the intrinsic properties of
759 spaces. Topology has applications in various areas, including quantum physics, computer science,
760 and biology, and deals with concepts like continuity, compactness, and connectedness.

761 **Transformation Geometry.** Transformation Geometry involves the study of geometric figures
762 and the changes they undergo through transformations such as translation, rotation, reflection, and
763 dilation. It focuses on understanding how these transformations affect the properties and relations
764 of shapes and figures. Transformation geometry is fundamental in understanding the symmetry,
765 congruence, and similarity of geometric objects, and it has applications in areas such as computer
766 graphics, robotics, and pattern recognition.

767 I Error Examples of GPT-4V

768 In this section, we present detailed examples of errors made by GPT-4V. We categorize these errors
769 into six types: reasoning error, vision recognition error, knowledge error, question misunderstood
770 error, refusal to answer, and calculation error.

771 I.1 Reasoning Error (42.2%)

772 Reasoning errors occur when GPT-4V fails to apply logical steps correctly. In Figure 36, the model
773 incorrectly substitutes an unknown variable with a known value, leading to an incorrect conclusion.
774 Figure 37 shows the model making an illogical hypothesis about the relationships of train wagon
775 numbers.

776 I.2 Vision Recognition Error (31.9%)

777 Vision recognition errors highlight the model's struggles with interpreting visual data. Figure 38
778 depicts an instance where GPT-4V fails to read information from a simple bar chart. In Figure 39,
779 the error occurs because the model mistakenly believes that 3, 6, and y are on the same side of the
780 triangle. Figure 40 shows the model experiencing severe hallucinations, as there is no balance in the
781 image at all.

782 I.3 Knowledge Error (15.1%)

783 Knowledge errors are instances where GPT-4V lacks the necessary information to answer a question
784 and provides information that is outdated or incorrect. Figure 41 and Figure 42 both showcase
785 situations where the model did not know how to solve the problems.

786 **I.4 Question Misunderstood Error (6.9%)**

787 These errors occur when GPT-4V misinterprets the user’s query. Figure 43 demonstrates a failure
788 to grasp the query’s intent, resulting in an incorrect answer. Similarly, in Figure 44, the model
789 mistakenly believes that the key below corresponds to the lock above, even though the question does
790 not provide this information.

791 **I.5 Reject to Answer (2.6%)**

792 Sometimes, GPT-4V refuses to provide an answer, which is categorized as a Reject to answer error.
793 Figure 45 and Figure 46 illustrate instances where the model either cannot generate a response or
794 chooses not to, citing various reasons.

795 **I.6 Calculation Error (1.3%)**

796 Calculation errors are straightforward mistakes in numerical or symbolic calculations. The model
797 only makes errors when performing symbolic calculations in our analysis. Both Figure 47 and
798 Figure 48 show a simple symbolic computational mistake, undermining the model’s reliability in
799 symbolic calculations.

800 **J Examples of GPT-4 with Image Captions**

801 In this section, we provide some qualitative examples of GPT-4 with image captions, in comparison
802 to other models like GPT-4V, Gemini Pro, and Qwen-VL-Max.

803 **J.1 Correct Examples**

804 Some typical correct examples (Figure 49, Figure 50, Figure 51, Figure 52, Figure 53, Figure 54,
805 Figure 55) of GPT-4 with image captions on subjects like algebra, arithmetic, analytic geometry,
806 counting, and combinatorics. Image captions are either accurate or have some small hallucinations.

807 **J.2 Wrong Examples**

808 Some typical wrong examples (Figure 56, Figure 57, Figure 58, Figure 59, Figure 60, Figure 61) of
809 GPT-4 with image captions on other subjects. Image captions are usually inaccurate to describe the
810 image.

811 **K Details of Human Annotators**

812 We hire human annotators for three tasks: 1. Verify the alignment of text questions and images; 2.
813 Categorize the problems into 16 different subjects; 3. Evaluate human performance on MATH-V.

814 **K.1 Alignment of Text Questions and Images**

815 We hire ten annotators in stage one, and all of them are senior college students, all majoring in
816 science-related disciplines. We pay them \$0.35 per question, and they were asked to judge whether
817 the text of the question matched the corresponding image and whether the image was complete.
818 When the image is incomplete or incorrectly positioned, they are asked to re-capture the image from
819 the original file.

820 **K.2 Subjects Classification**

821 For task two, we have 10 annotators and three authors, to categorize the problems into 16 different
822 subjects independently in three groups with our specially designed annotation tool, as illustrated
823 in Figure 19. We ask them to read our definitions of the 16 subjects in advance as shown in

824 Appendix H.4, and will randomly check their classification of certain topics. If there are mistakes in
 825 classification, we will promptly correct them and inform them. We pay \$0.45 per question. Notably,
 826 we also utilize existing LMMs (such as Gemini and GPT-4V) for subject classification. Detailed
 827 instructions to annotators and prompts to models can be found in Table 10. Finally, we will combine
 828 the classification results of the three groups with the classification results returned by LMMs to give
 829 the final classification.

<i>▷ Instructions for annotators on how to label questions into 16 categories</i>
<p>We are constructing a dataset about mathematical problems in visual contexts, and we have collected quantities of questions that require categorization into different mathematical subjects. These categorizations are logic, algebra, counting, arithmetic, combinatorics, graph theory, topology, statistics, combinatorial geometry, solid geometry, metric geometry - angle, metric geometry - area, metric geometry - length, analytic geometry, descriptive geometry, and transformation geometry. Here are the detailed definitions of them: ... (definitions in Appendix H.4)</p> <p>If you are unsure about how to categorize the problem into these subjects, choose 'pending' instead.</p>
<i>▷ The prompt for models to get subject advice</i>
<p>What branch of mathematics does the problem belong to? Choose from the following: logic, algebra, counting, arithmetic, combinatorics, graph theory, topology, statistics, combinatorial geometry, solid geometry, metric geometry - angle, metric geometry - area, metric geometry - length, analytic geometry, descriptive geometry, transformation geometry, pending.</p> <p>You are not supposed to solve the problem. Wrap your final answer, a word or a short phrase, in "□".</p>

Table 10: Instructions for annotators on how to label questions into 16 categories and the prompt for models to get subject advice.

830 **K.3 Human Performance**

831 To obtain average human performance on our benchmark, we recruit 100 annotators possessing a
 832 high school degree or higher. These annotators had not previously participated in the competitions
 833 where our data originated and were seeing the questions in our test set for the first time. We divided
 834 the 3040 questions of MATH-V into 100 subsets, each containing 30 or 31 questions. Subsequently,
 835 we randomly allocated two subsets from this pool to each student. Participants were tasked with
 836 completing both subsets within a 70-minute time frame, spread across two separate days. Note that
 837 this time limit closely aligns with that of the Kangaroo competition, where participants are assigned a
 838 similar duration for a test containing 30 problems. Each participant receives a one-time payment of
 839 \$40 as compensation for their contribution.

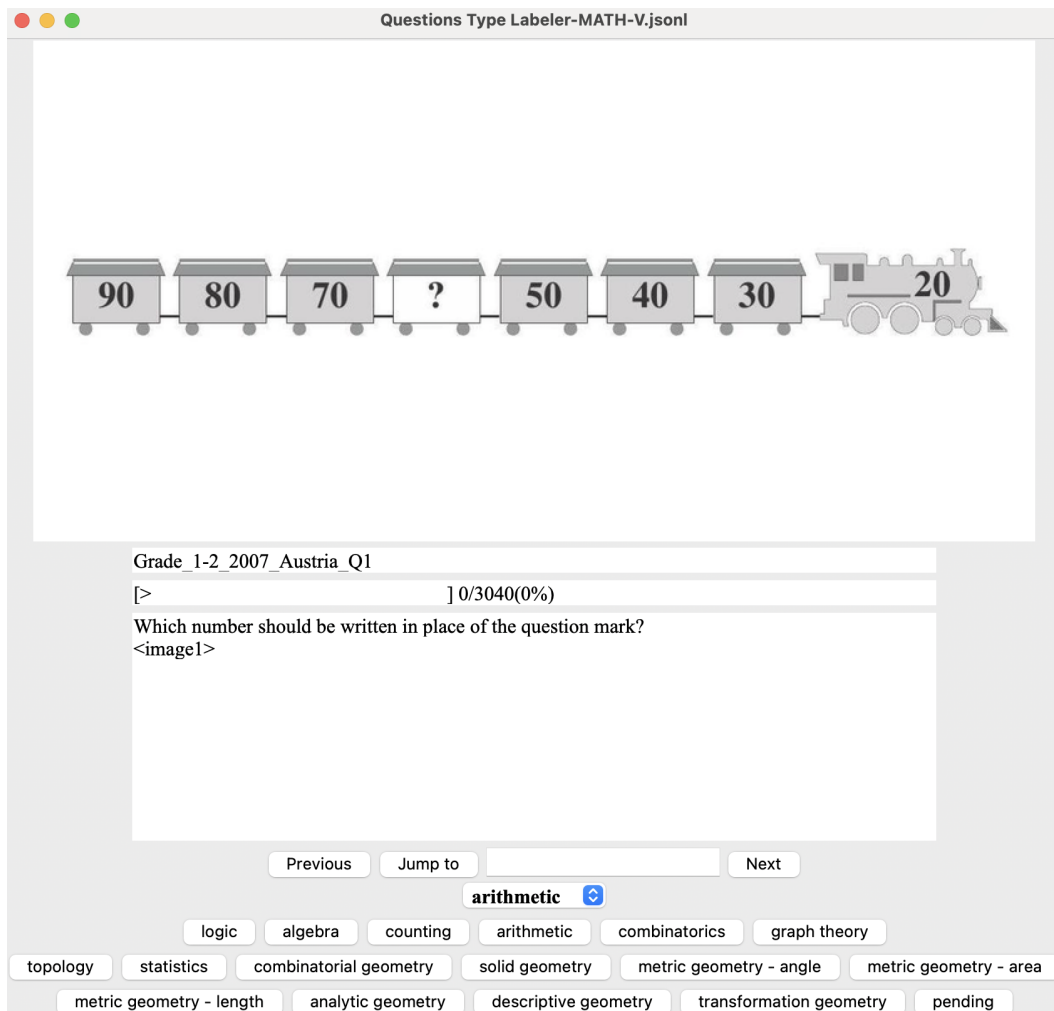


Figure 19: GUI for categorizing the problems into 16 different subjects.

Algebra

$\triangle \triangle \triangle \diamond$

\triangle

$\bullet \diamond$

A	B	C	1				25
	A	B					
+		A			X		
	3	0	0	17			81

Exchange Rates

1 turkey = 5 ducks
1 goose + 2 hens = 3 ducks
4 hens = 1 goose

44

59

38

$\triangle = \lambda$ $\diamond = \mu$ $\bullet = \nu$
 $\square = \rho$ $\circ = \sigma$ $\circ = \tau$

1	3	5	...	97	99
4	8	12		196	

$\begin{matrix} 1 & 2 \\ \diagdown & / \\ 3 & 4 & + & 5 & 6 \end{matrix}$

\times			7
	J	K	56
	M	36	N
	P	27	R
6	18	S	42

0	0	0
0	0	0
0	0	0

	18	
	47	
13		?

	18	32	48
	12	16	30

	24		
	26		
	40		
	27	20	?

?

Multiply by 0.5

Multiply by 1/3

Square the number

Add 1

50

Figure 20: Some images from Algebra.

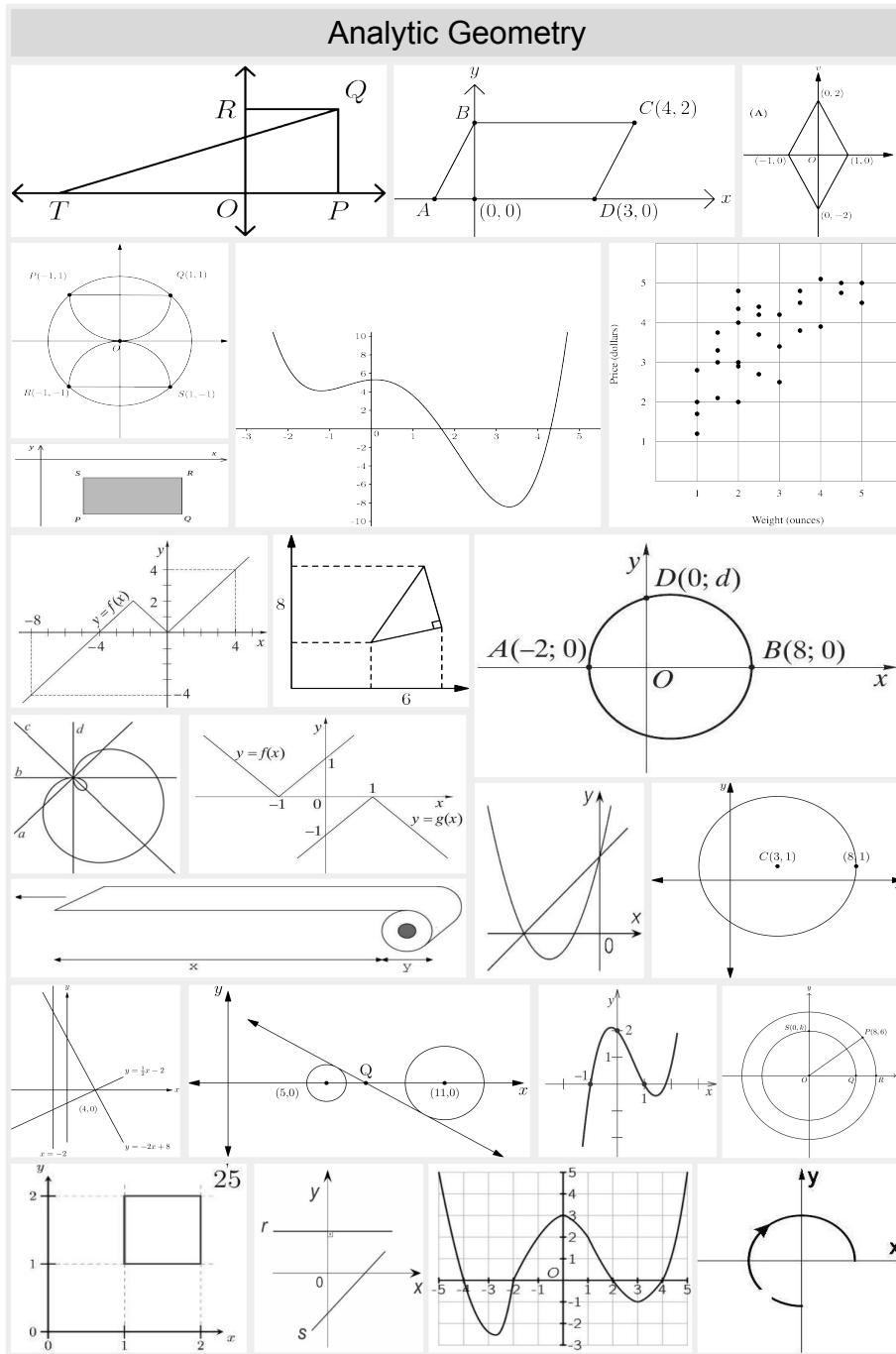


Figure 21: Some images from Analytic Geometry.

Arithmetic

$N \rightarrow$

if N is even $\rightarrow \frac{N}{2}$

if N is odd $\rightarrow 3N + 1$

Azaleas	×	×				
Irises	×	×	×			
Roses	×	×	×	×	×	
Gerberas	×	×	×	×	×	×

- ↑ Seat 1 to 20
- Seat 21 to 40
- ← Seat 41 to 60
- ⇐ Seat 61 to 80
- ⇐ Seat 81 to 100

Figure 22: Some images from Arithmetic.

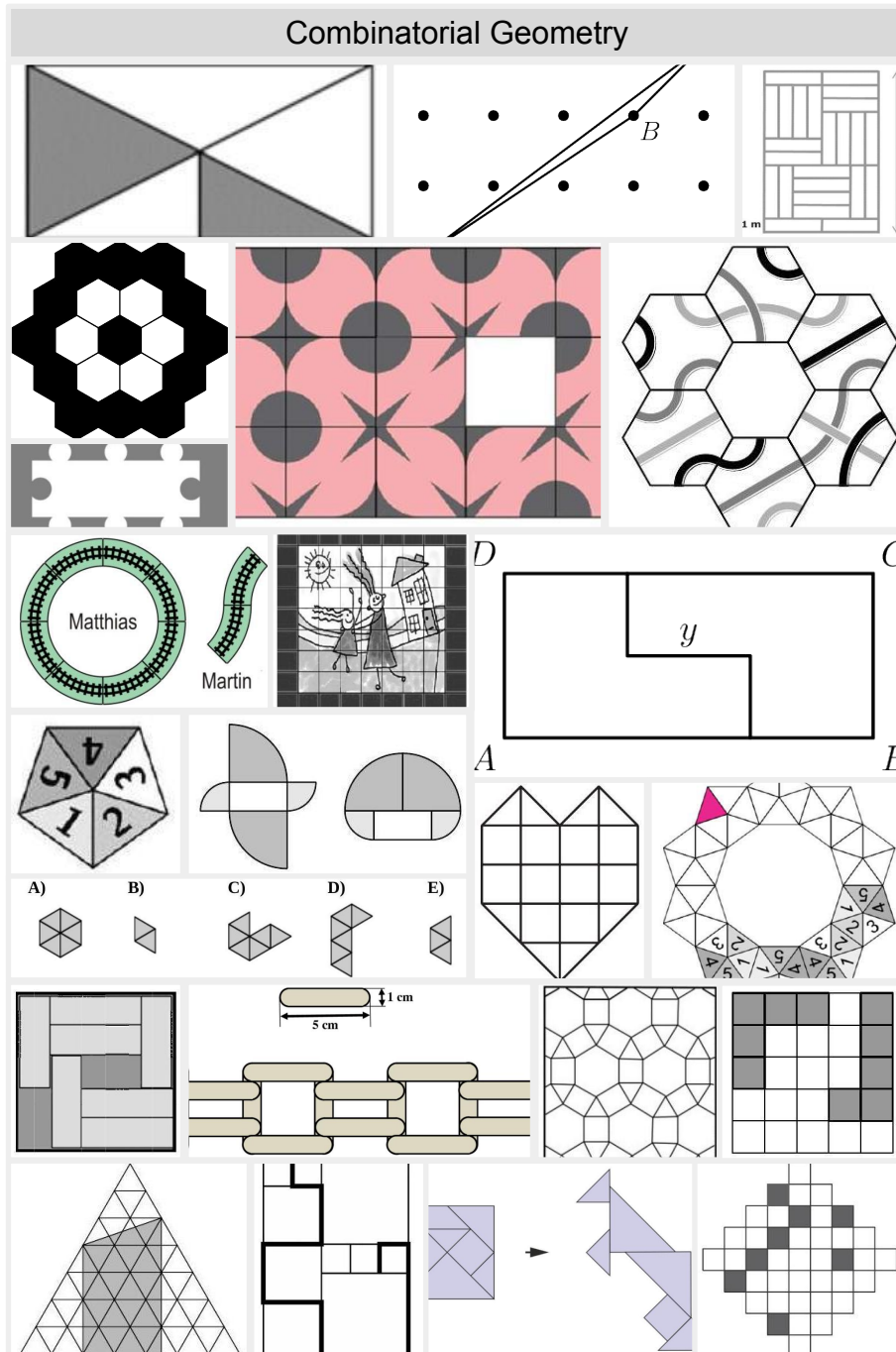


Figure 23: Some images from Combinatorial Geometry.

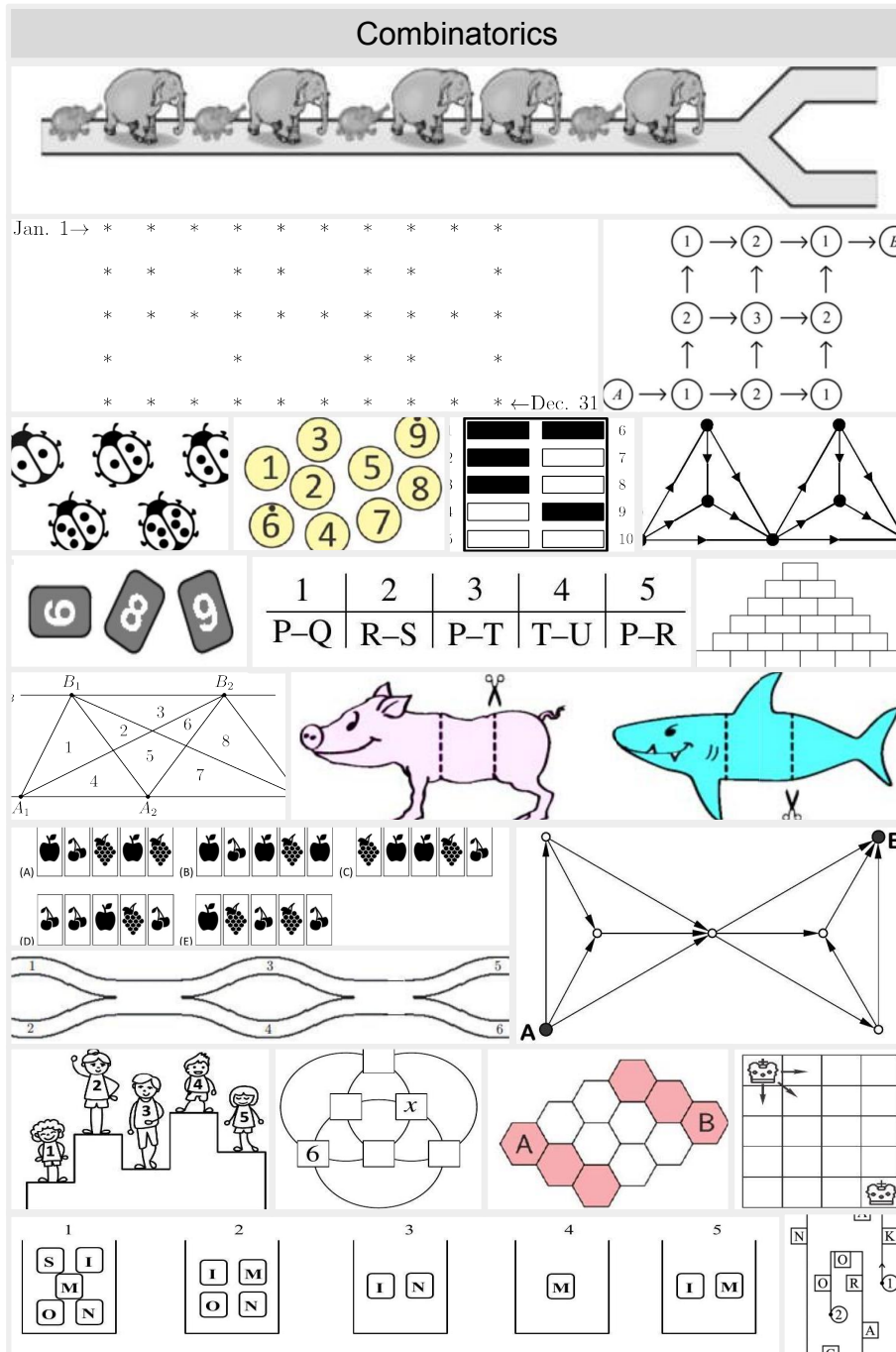


Figure 24: Some images from Combinatorics.

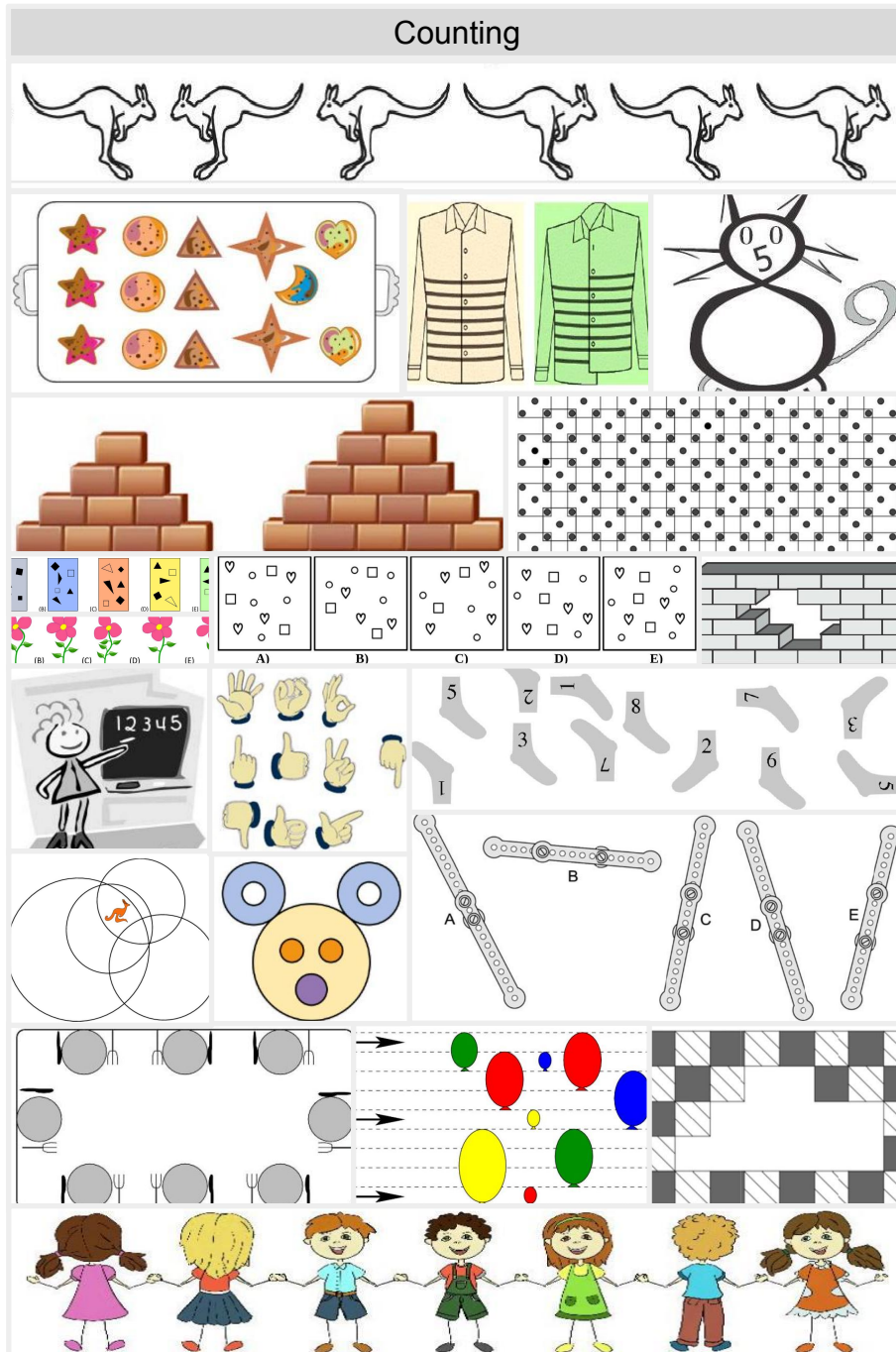


Figure 25: Some images from Counting.

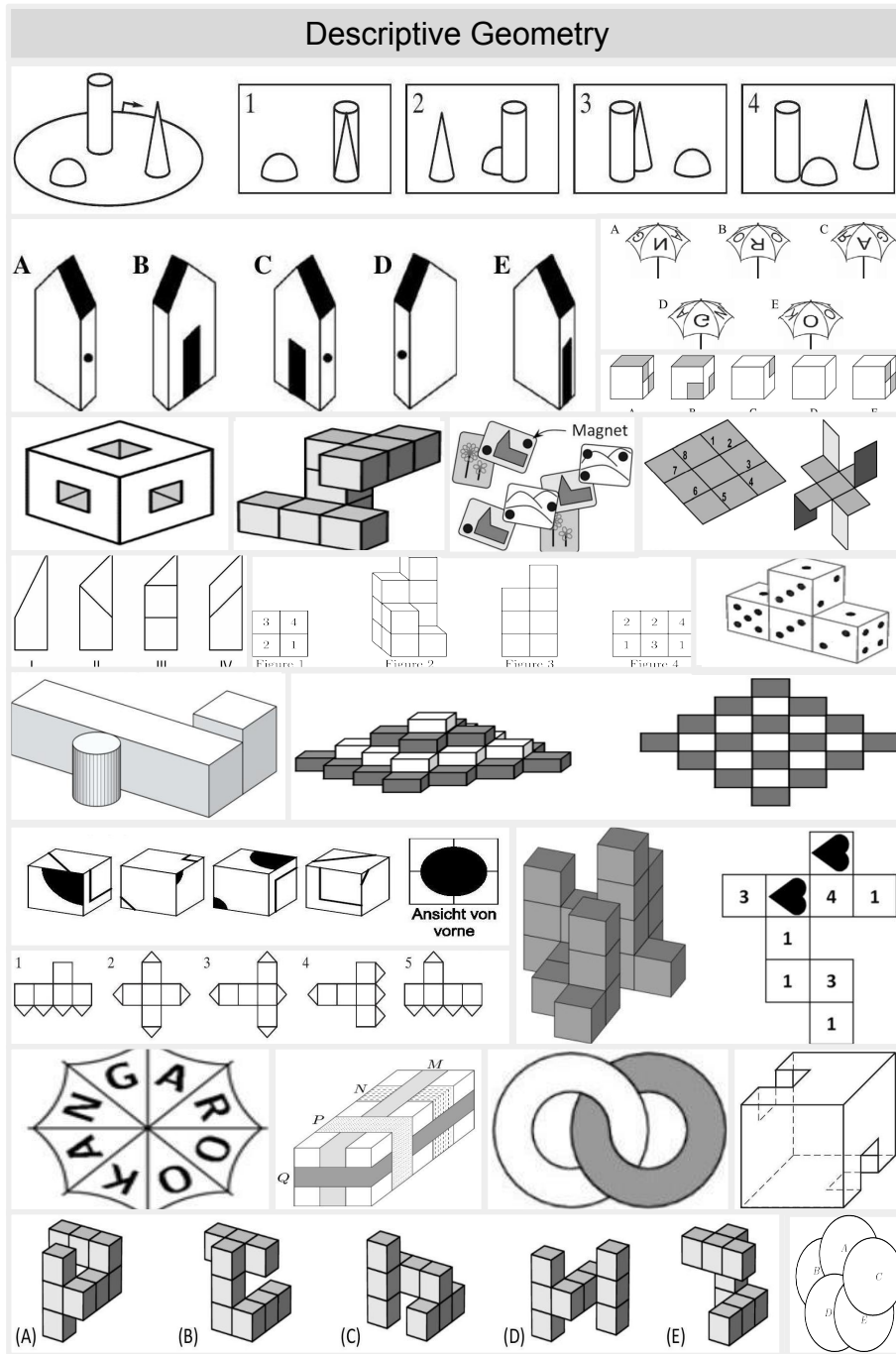


Figure 26: Some images from Descriptive Geometry.

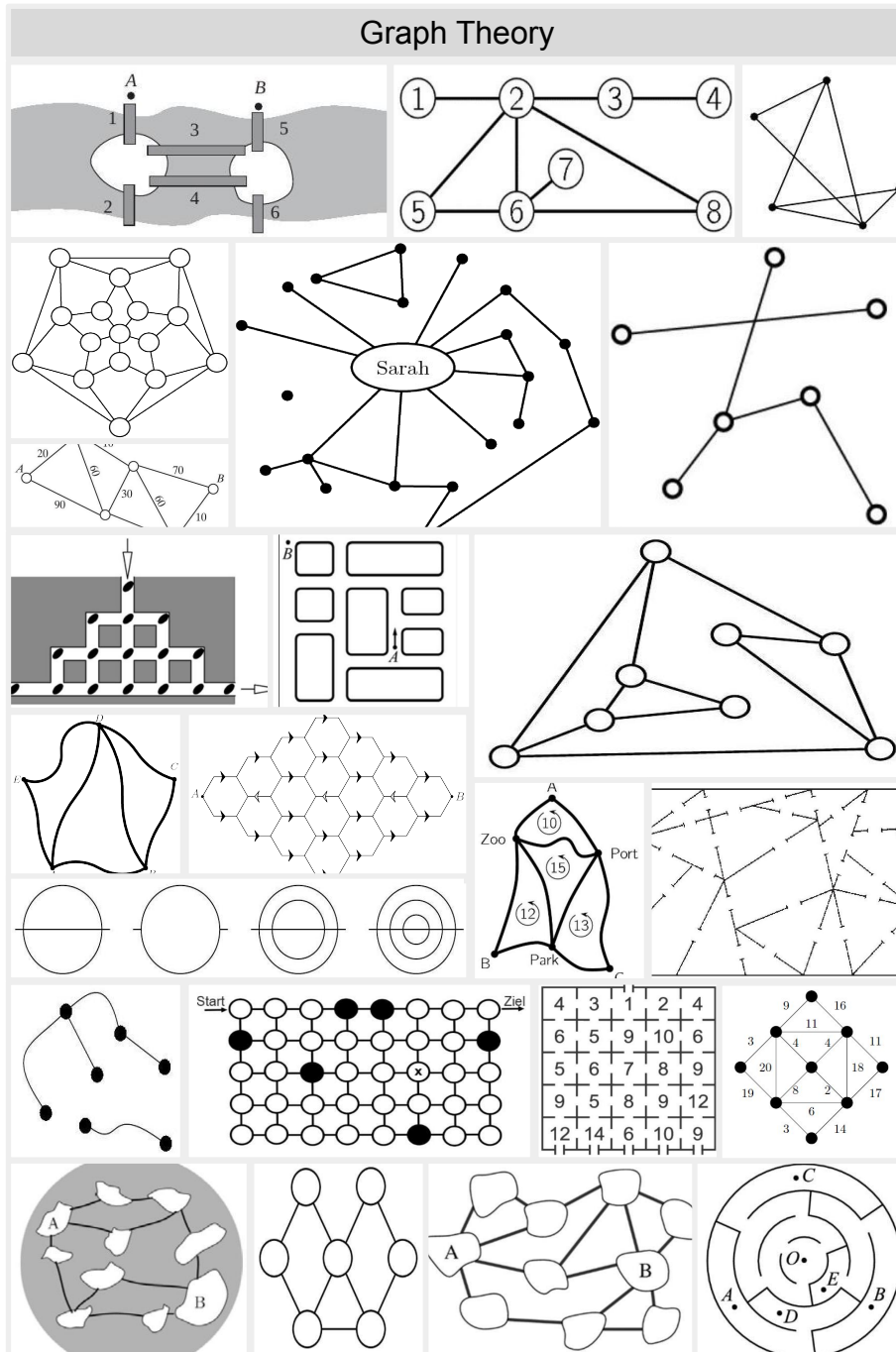


Figure 27: Some images from Graph Theory.

Logic

4	1	3	2
---	---	---	---

 Two digits are correct but in the wrong places.

9	8	2	6
---	---	---	---

 One digit is correct and in the right place.

5	0	7	9
---	---	---	---

 Two digits are correct with one of them being in the right place and the other one in the wrong place.

2	7	4	1
---	---	---	---

 One digit is correct but in the wrong place.

7	6	4	2
---	---	---	---

 None of the digits is correct.

green green

blue x blue

E

K

6

4

7

	a	a	
2			
		2	
1			

If it's blue, it's round.
 If it's square, it's red.
 It's either blue or yellow.
 If it's yellow, it's square.
 It's either square or round.

E

A

B

C

D

♠		?	
♣	♠		
	♦		
	♥		

P	K
T	M
H	M

T	M
---	---

T	K
M	P

K	T
---	---

M

1

3

4

7

1	2	6
3		

If it's blue, it's round.
 If it's square, it's red.
 It's either blue or yellow.
 If it's yellow, it's square.
 It's either square or round.

Figure 28: Some images from Logic.

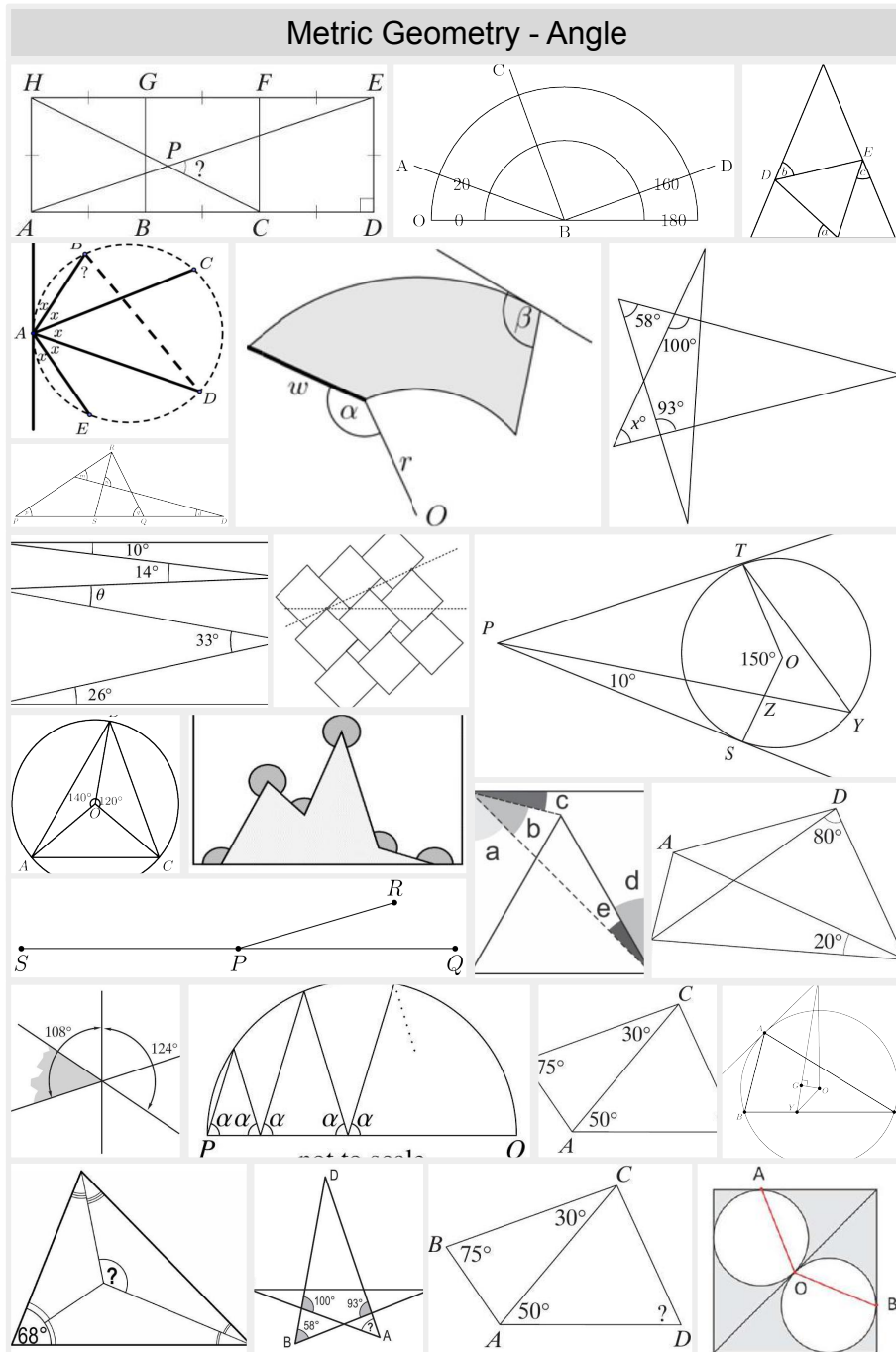


Figure 29: Some images from Metric Geometry - Angle.

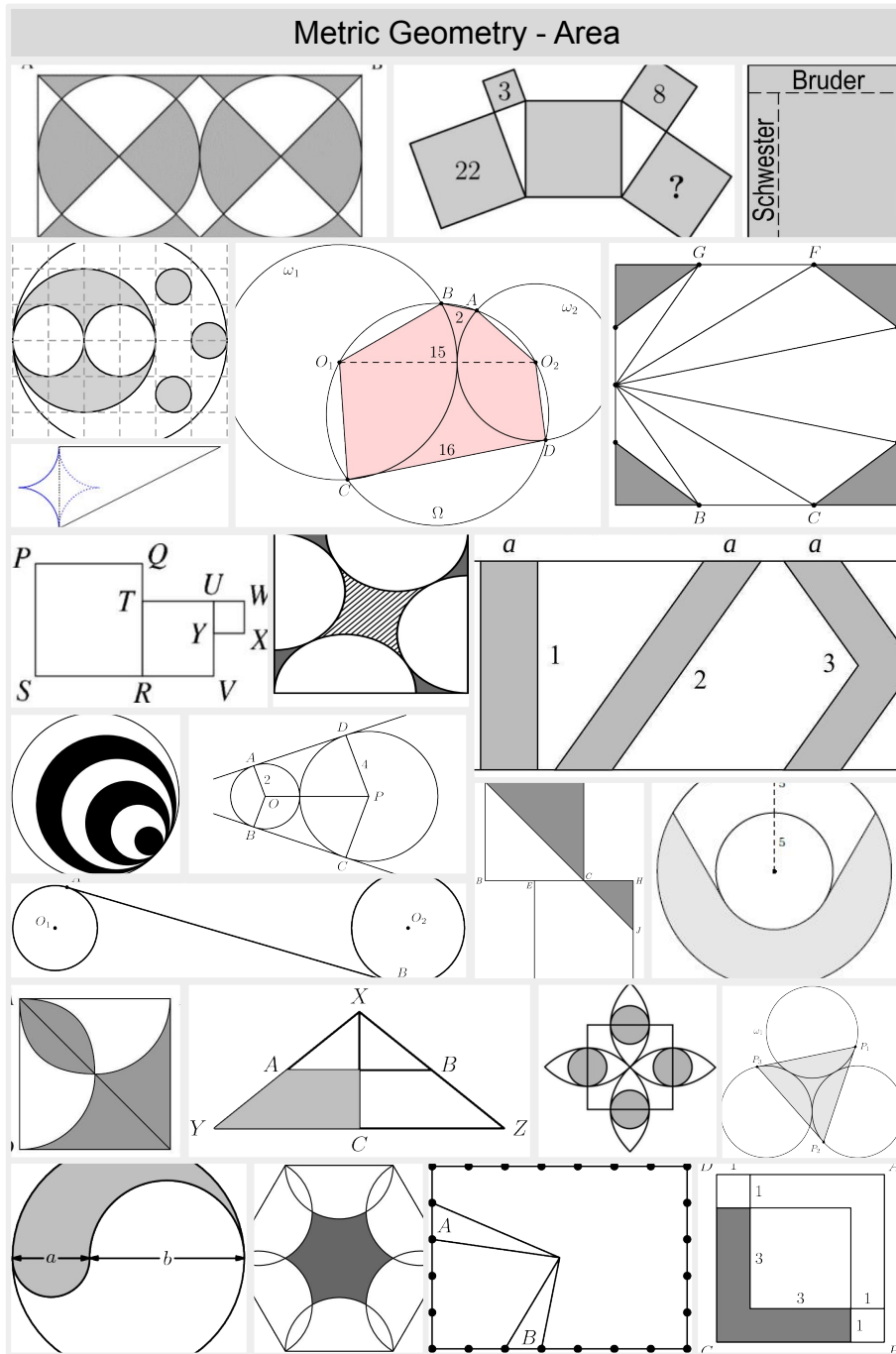


Figure 30: Some images from Metric Geometry - Area.

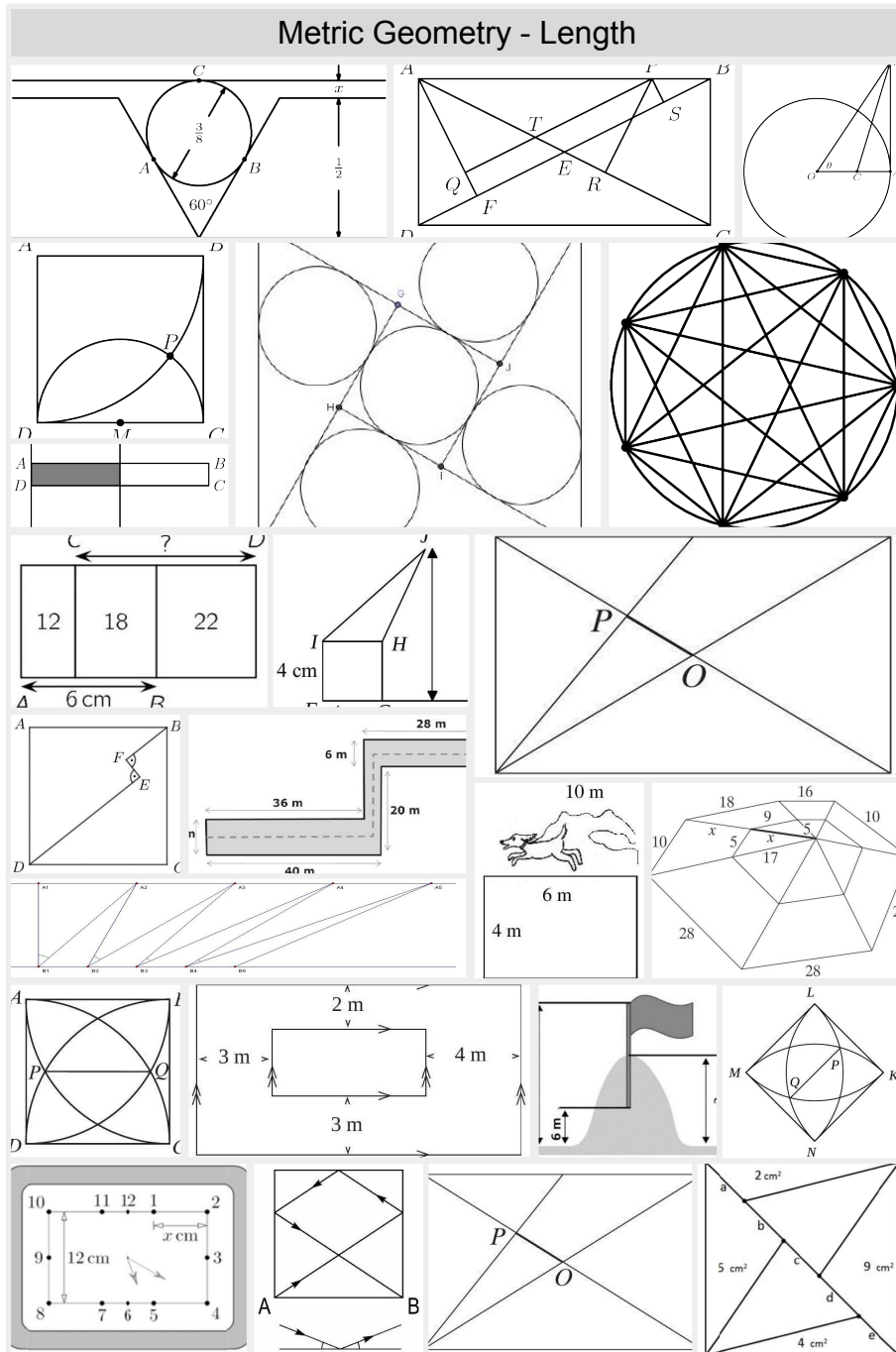


Figure 31: Some images from Metric Geometry - Length.

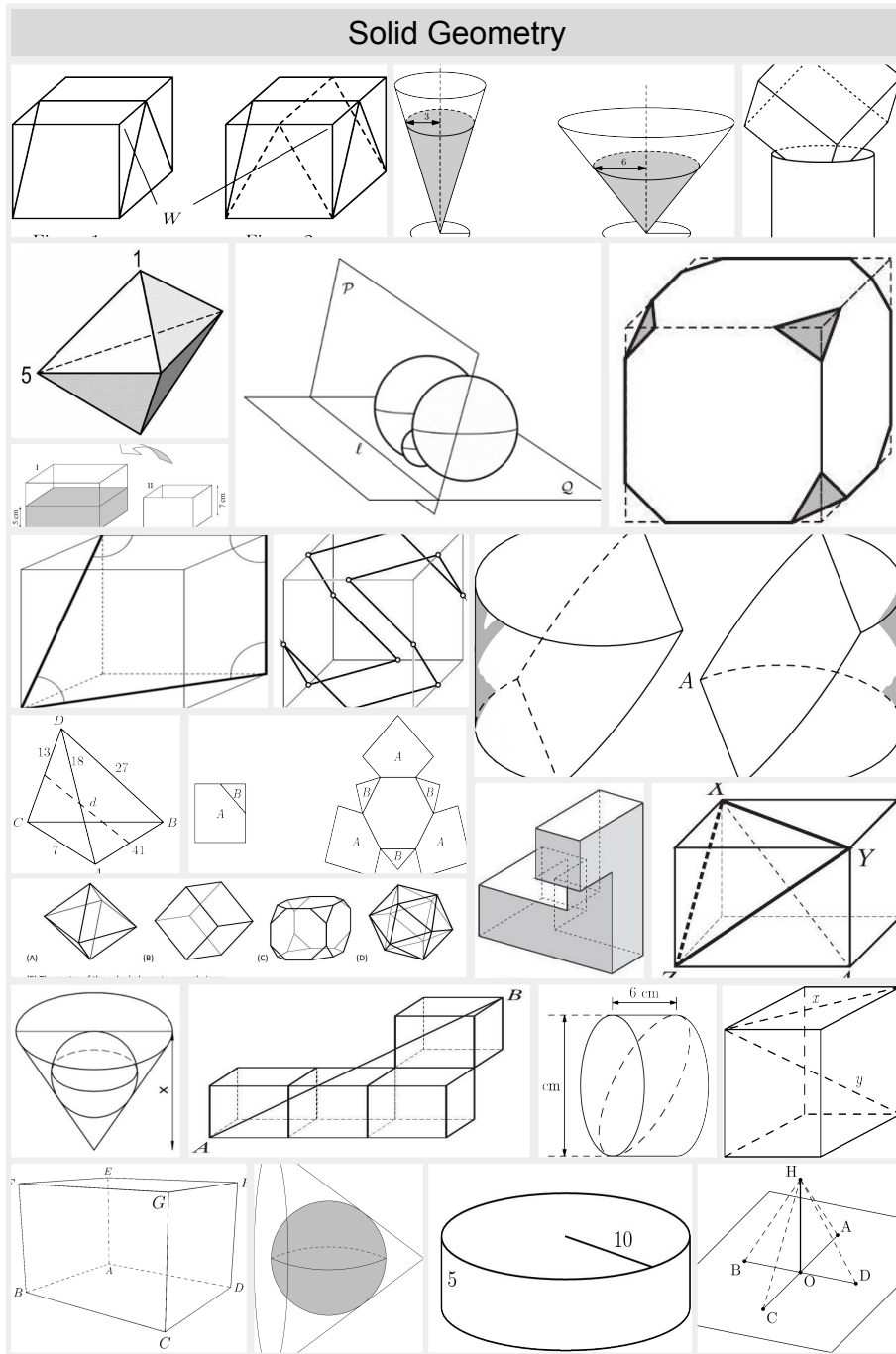


Figure 32: Some images from Solid Geometry.

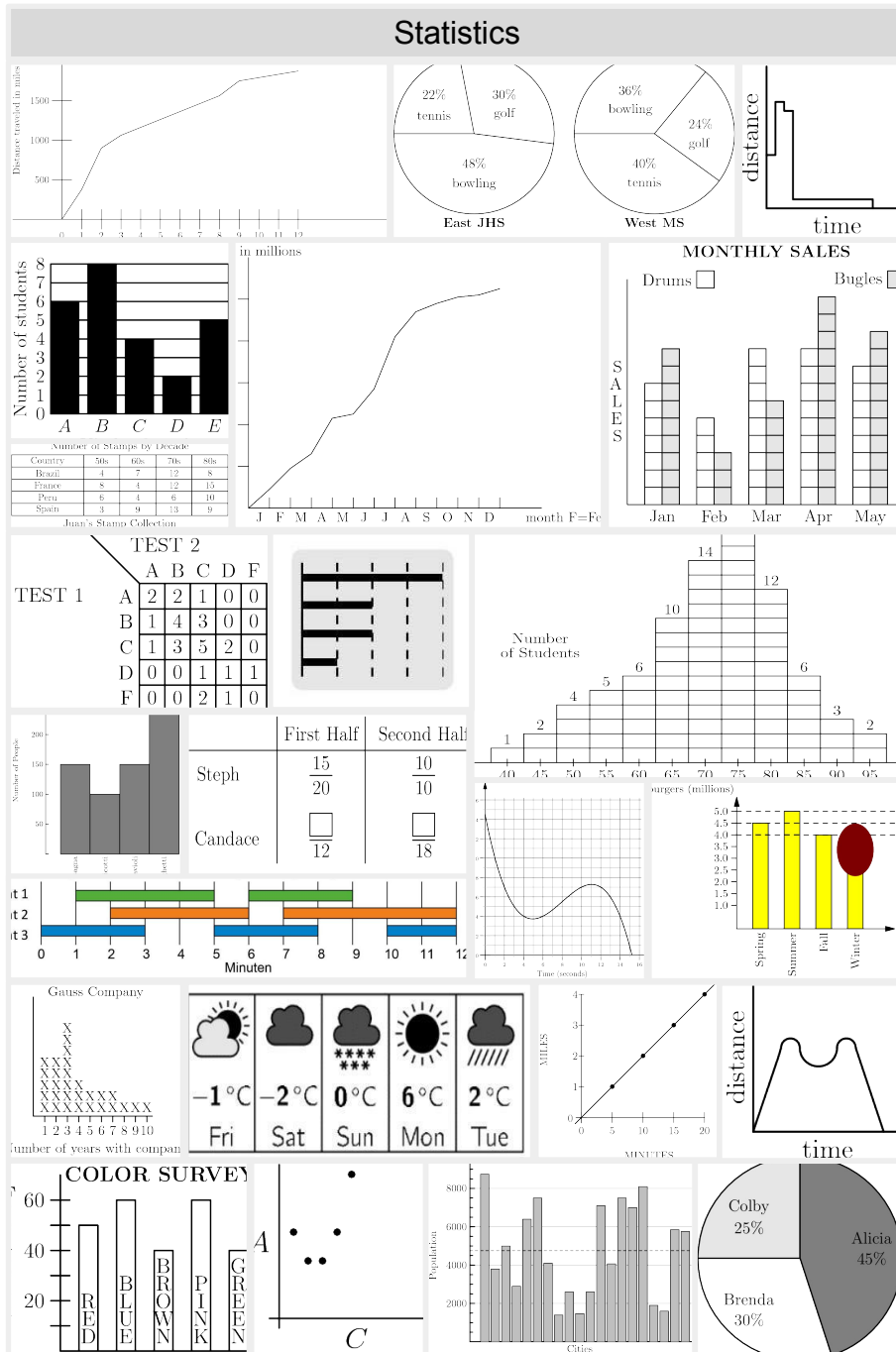


Figure 33: Some images from Statistics.

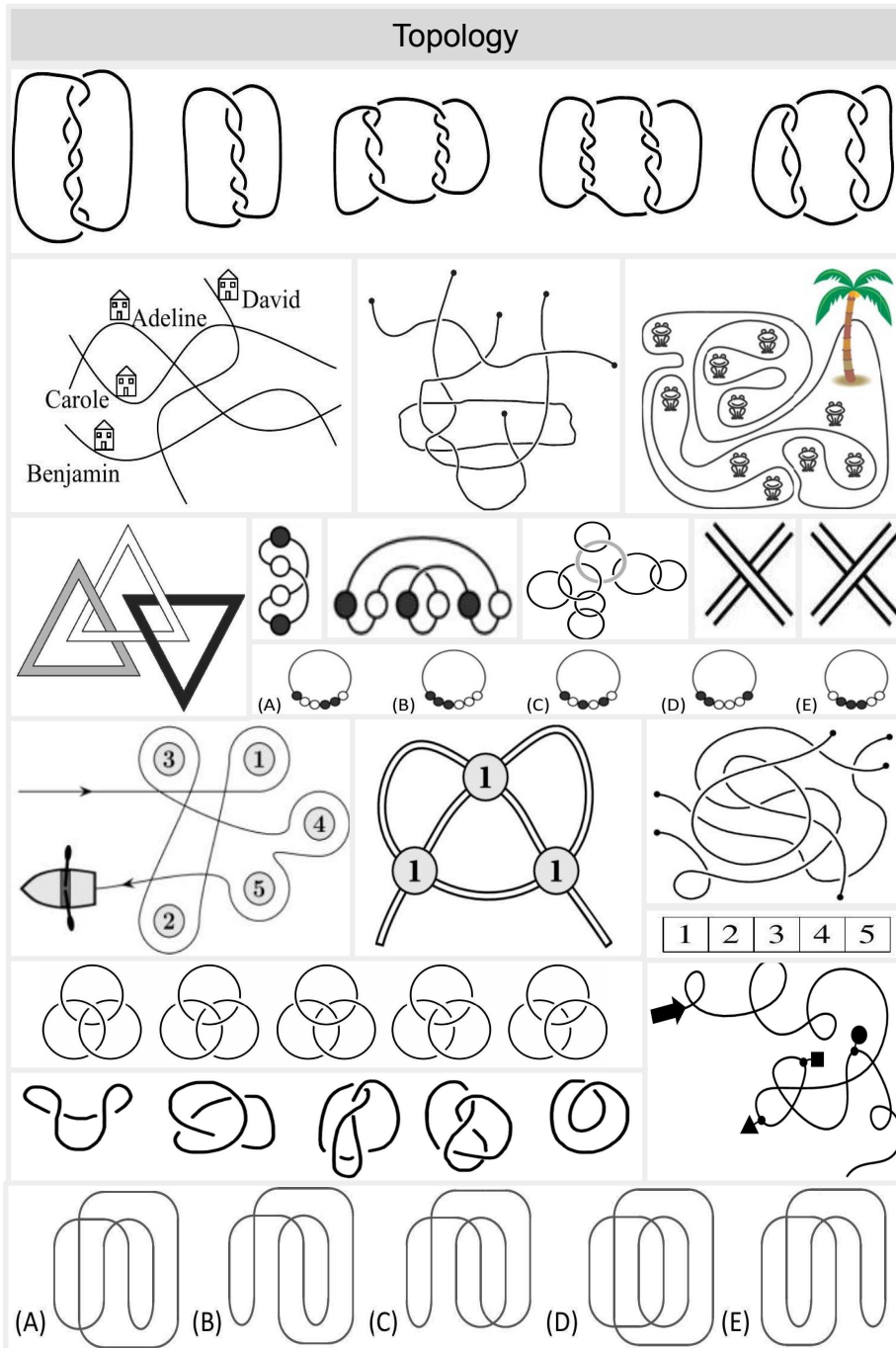


Figure 34: Some images from Topology.

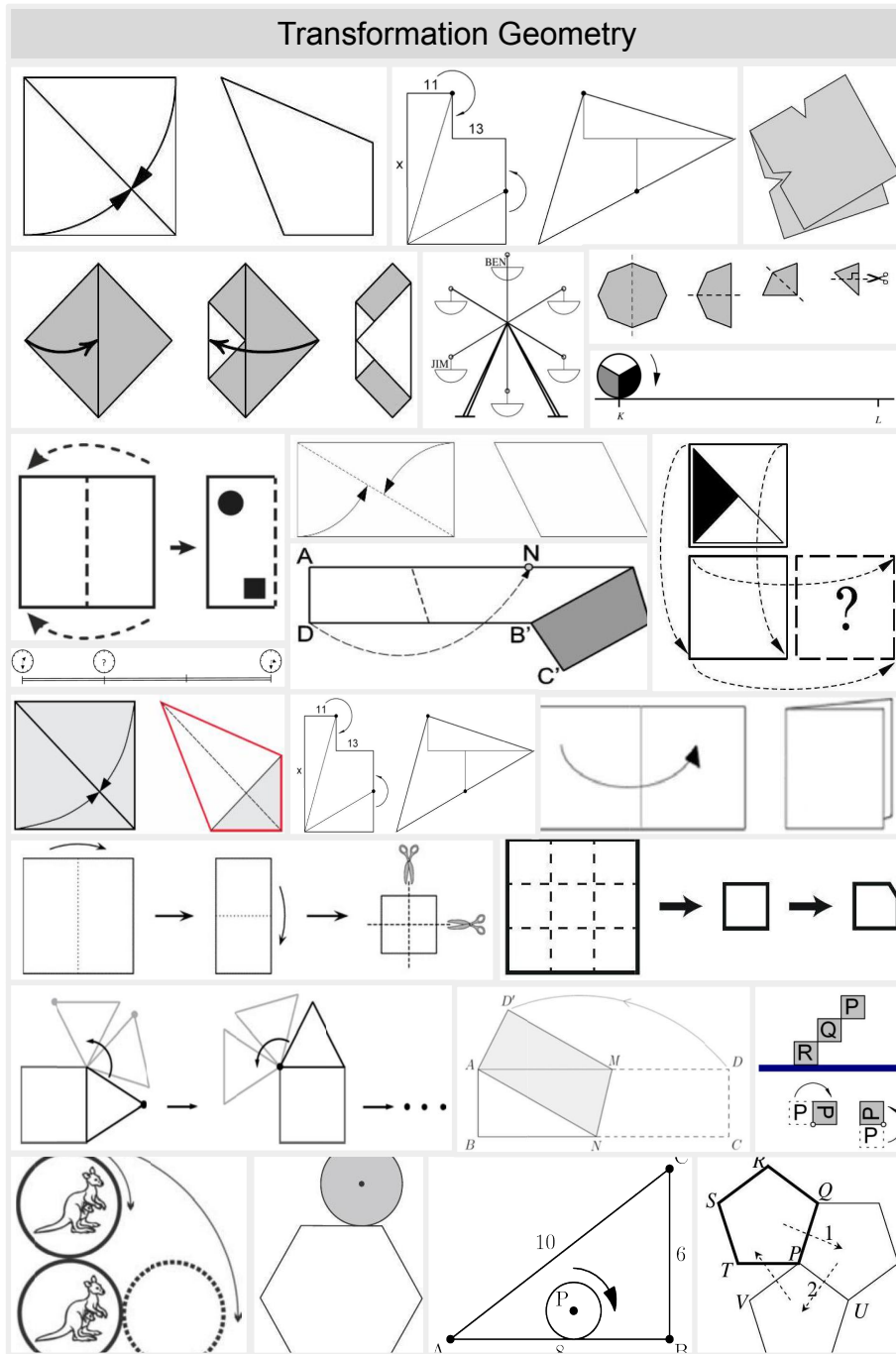


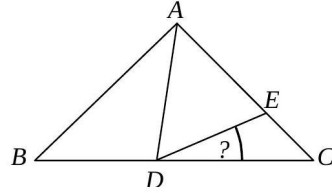
Figure 35: Some images from Transformation Geometry.

Reasoning Error

Question: In triangle ABC , $AB = AC$, $AE = AD$ and angle $BAD = 30^\circ$. What is the size of angle CDE ?

- (A) 10°
- (B) 15°
- (C) 20°
- (D) 25°
- (E) 30°

Ground Truth: (B) 15°



GPT4V Answer:

First, let's review the given information:

- Triangle ABC is isosceles with $AB = AC$. - $AE = AD$. - $\angle BAD = 30^\circ$.

Given that $AB = AC$, we know that $\angle ABC = \angle ACB$ because the base angles of an isosceles triangle are equal.

Since $\angle BAD = 30^\circ$ and $AE = AD$, we have another isosceles triangle, $\triangle ADE$, and $\angle ADE = \angle AED$. Let's call this unknown angle θ .

Because $AE = AD$, triangle ADE is isosceles, which makes $\angle ADE = \angle AED = \theta$. Also, the sum of angles in a triangle is 180° , so in $\triangle ADE$, we have:

$$\angle DAE + \angle ADE + \angle AED = 180^\circ$$

$$30^\circ + \theta + \theta = 180^\circ$$

$$2\theta = 180^\circ - 30^\circ$$

$$2\theta = 150^\circ$$

$$\theta = 75^\circ$$

So, both $\angle ADE$ and $\angle AED$ are 75° .

Now we move to $\triangle CDE$. We are looking for $\angle CDE$. The angle $\angle CDE$ is part of $\angle CDA$, which is an exterior angle to $\triangle ADE$ and equals the sum of the opposite interior angles, which we have found to be $\angle ADE + \angle AED$.

Therefore:

$$\angle CDA = \angle ADE + \angle AED$$

$$\angle CDA = 75^\circ + 75^\circ$$

$$\angle CDA = 150^\circ$$

Since $\angle CDA$ is a straight line and is part of $\angle CDE$, the following relationship holds:

$$\angle CDA = \angle CDE + \angle ADE$$

$$150^\circ = \angle CDE + 75^\circ$$

$$\angle CDE = 150^\circ - 75^\circ$$

$$\angle CDE = 75^\circ$$

This is not amongst the given options, indicating that there's an error.

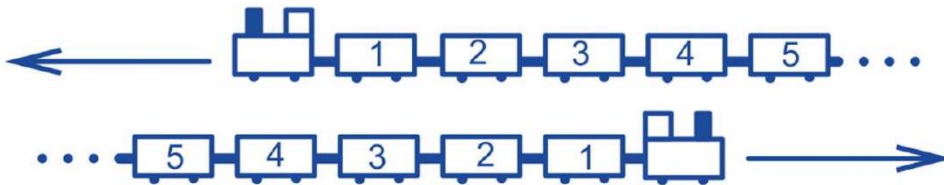
[Truncated Due to Length]

Figure 36: An example of reasoning error.

Reasoning Error

Question:

Two equal trains, each with 31 numbered wagons, travel in opposite directions. When the wagon number 7 of a train is side by side with the wagon number 12 of the other train, which wagon is side by side with the wagon number 11 ?



Ground Truth: 8

GPT4V Answer:

To solve this problem, we'll use the information that we have about the wagons that are currently side by side. We are told that wagon number 7 of one train is side by side with wagon number 12 of the other train. We can use this to establish a pattern or a relationship between the numbers of the wagons that are side by side.

If wagon 7 is next to 12, we can infer that as one train's wagon number increases, the other train's corresponding wagon number decreases if the trains are moving in opposite directions. The difference between the wagon numbers is 5 ($12 - 7 = 5$).

Now, if we want to find out which wagon would be side by side with wagon number 11, we can use the established difference to determine the opposite wagon number. We subtract 5 from 11 to find the matching wagon number on the other train:

$$11 - 5 = 6$$

Therefore, the wagon number 6 of the one train will be side by side with wagon number 11 of the other train. So, the solution is:

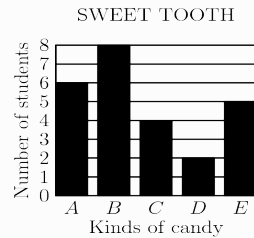
6

Figure 37: An example of reasoning error.

Vision Recognition Error

Question: The students in Mrs. Sawyer's class were asked to do a taste test of five kinds of candy. Each student chose one kind of candy. A bar graph of their preferences is shown. What percent of her class chose candy E?

Ground Truth: 2



GPT4V Answer:

To find the percentage of the class that chose candy E, we first need to determine the number of students who chose candy E by looking at the bar graph. Then, we calculate the total number of students who participated in the taste test by adding the number of students who chose each kind of candy. Finally, we can find the percentage of students who chose candy E by dividing the number of students who chose candy E by the total number of students and then multiplying by 100.

From the bar graph, the number of students who chose candy E can be counted directly:
Candy E chosen by = 2 students

Now let's count the total number of students who participated:

Candy A chosen by = 7 students Candy B chosen by = 8 students Candy C chosen by = 5 students
Candy D chosen by = 3 students Candy E chosen by = 2 students (as identified earlier)

Total number of students = $7 + 8 + 5 + 3 + 2 = 25$ students

Now we determine the percentage of students who chose candy E:

$$\text{Percentage choosing candy E} = \left(\frac{\text{Number choosing candy E}}{\text{Total number of students}} \right) \times 100$$

$$\text{Percentage choosing candy E} = \left(\frac{2}{25} \right) \times 100$$

$$\text{Percentage choosing candy E} = \frac{2}{25} \times 100$$

$$\text{Percentage choosing candy E} = 0.08 \times 100$$

$$\text{Percentage choosing candy E} = 8\%$$

Therefore, the percentage of the class that chose candy E is:

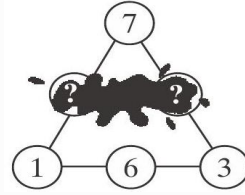
8%

Figure 38: An example of vision recognition error.

Vision Recognition Error

Question: The sums of the all the three numbers on each side of the triangle are equal. Two numbers happened to be stained with ink. How much is the sum of these two numbers?

Ground Truth: 2



GPT4V Answer:

To solve this problem, we need to realize that the sum of the numbers on each side of the triangle is the same. We will call this sum S .

Let's denote the two unknown numbers, which are obscured by ink, as x and y .

We can write the sum on each side of the triangle using these variables. For the side with the numbers 7, x , and 1, the sum is $7 + x + 1$. **For the side with the numbers 3, 6, and y , the sum is $3 + 6 + y$.**

Because the sums are equal, we have:

$$7 + x + 1 = 3 + 6 + y$$

$$8 + x = 9 + y$$

We want to find the sum of x and y . Let's add y to both sides of the equation to move terms involving y to the left side and terms involving x to the right side:

$$y + 9 + y = 8 + x + y$$

$$2y + 9 = 8 + x + y$$

Now subtract y from both sides:

$$y + 9 = 8 + x$$

We know that the other side of the triangle is $1 + 6 + 3 = 10$, so the sum of the numbers on each side must be 10, which means:

$$y + 9 = 10$$

$$y = 10 - 9$$

$$y = 1$$

Now we substitute $y = 1$ into the sum for the other sides to find x :

$$x + 8 = 10$$

$$x = 10 - 8$$

$$x = 2$$

So, the two numbers are $x = 2$ and $y = 1$, and their sum is:

$$x + y = 2 + 1 = 3$$

Thus, the sum of the two numbers that are stained with ink is 3, and so we have:

3

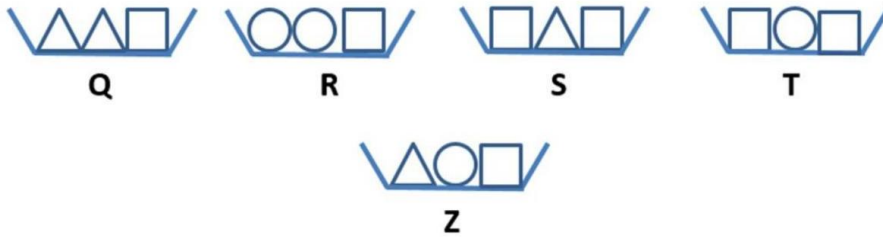
Figure 39: An example of vision recognition error.

Vision Recognition Error

Question:

Karin wants to place five bowls on a table so that they are ordered according to their weight. She has already placed the bowls *Q*, *R*, *S* and *T* in order, where *Q* is lightest and *T* is heaviest. Where does she have to place bowl *Z*?

- (A) to the left of bowl *Q*
- (B) between bowls *Q* and *R*
- (C) between bowls *R* and *S*
- (D) between bowls *S* and *T*
- (E) to the right of bowl *T*



Ground Truth: (B)

GPT4V Answer:

The image depicts a series of bowls marked *Q*, *R*, *S*, *T*, and *Z*, with balance scales suggesting their comparative weights. To determine where bowl *Z* should be placed in the sequence, we examine the scale readings:

- Bowl *Q* is lighter than all other bowls since it's shown on the higher side of all the scales it's compared with.
- Bowl *R* is heavier than *Q* but lighter than the other bowls since it is on the lower side when compared with *Q* and on the higher side compared to *S* and *T*.
- Bowl *S* is heavier than *Q* and *R* but lighter than *T*, as displayed by being on the lower side of the scales versus *Q* and *R*, and the higher side versus *T*.
- Bowl *T* is the heaviest as it is on the lower side of every comparison.

Bowl *Z* is shown compared only to bowl *S*. *Z* is on the lower side, suggesting it is heavier than *S*. Given that there's no additional comparison provided, we cannot confirm *Z*'s weight relative to *T*; however, *Z* must be placed between *S* and *T*, as *Z* is heavier than *S* and there is no information to assume it is heavier than *T*.

Therefore, the correct placement for bowl *Z* is between bowls *S* and *T*:

D

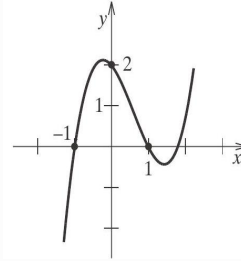
Figure 40: An example of vision recognition error. Model is experiencing severe hallucinations in this example

Knowledge Error

Question: We see in the diagram at the right a piece of the graphic of the function

$$f(x) = ax^3 + bx^2 + cx + d.$$

What is the value of b ?



Ground Truth: -2

GPT4V Answer:

To determine the value of b in the cubic function $f(x) = ax^3 + bx^2 + cx + d$, we need to look at the properties of the function's graph provided in the diagram.

From the graph, we can identify the coordinates of the two local extrema (maximum and minimum). At these points, the first derivative $f'(x)$ of the function must be zero since the slope of the tangent line at these points is zero.

The first derivative of the cubic function $f(x)$ is given by:

$$f'(x) = 3ax^2 + 2bx + c.$$

We have two extrema points visible on the graph, at $x = -1$ and $x = 1$. Let's set up the derivative to be equal to zero at these points:

For $x = -1$:

$$0 = 3a(-1)^2 + 2b(-1) + c$$

$$0 = 3a - 2b + c$$

$$2b = 3a + c \quad (1)$$

For $x = 1$:

$$0 = 3a(1)^2 + 2b(1) + c$$

$$0 = 3a + 2b + c \quad (2)$$

Let's solve this system of equations (1) and (2) for b .

Subtract (1) from (2):

$$3a + 2b + c - (3a - 2b + c) = 0$$

$$3a + 2b + c - 3a + 2b - c = 0$$

$$4b = 0$$

$$b = 0$$

Therefore, the value of b is zero.

$$\boxed{b = 0}$$

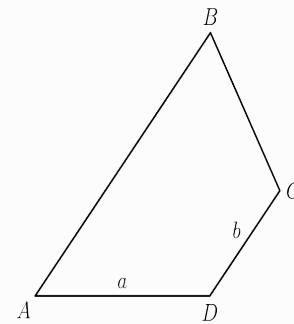
Figure 41: An example of knowledge error.

Knowledge Error

Question: In the accompanying figure, segments AB and CD are parallel, the measure of angle D is twice the measure of angle B , and the measures of segments AB and CD are a and b respectively. Then the measure of AB is equal to

- (A) $\frac{1}{2}a + 2b$
- (B) $\frac{3}{2}b + \frac{3}{4}a$
- (C) $2a - b$
- (D) $4b - \frac{1}{2}a$
- (E) $a + b$

Ground Truth: -2



GPT4V Answer:

The image you've provided appears to be a quadrilateral with sides AB and CD denoted as being parallel and segments AB labeled as a , and CD as b . However, without additional information such as angle measures or the relationship between the sides and angles, it is impossible to determine a definitive relationship between a and b based on the given options (A) through (E).

Typically, in problems like this, you might apply properties of parallel lines and transversals or properties of similar triangles if any triangle relationships can be determined. But the given information is insufficient to apply these properties directly.

Given the description of the problem, it seems to be expecting the application of a geometric principle or theorem which is not evident from the image alone. It's possible that there is either additional information that has not been provided or there may be an error in the problem's depiction.

With the information currently at hand, it is not feasible to reach a solution, and I cannot provide an answer in the requested format . If you can provide further details or clarify the conditions of the problem, I would be happy to reattempt to solve it.

Figure 42: An example of knowledge error.

Question Misunderstood Error

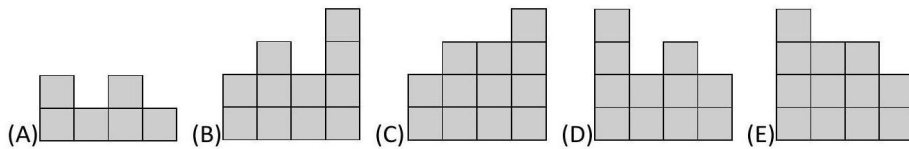
Question:

Johann stacks 1×1 cubes on the squares of a 4×4 grid. The diagram <image1> shows how many cubes were piled on top of each other on each square of the grid. What will Johann see if he looks from behind (hinten) at the tower?

<image1>:

<i>HINTEN</i>			
4	2	3	2
3	3	1	2
2	1	3	1
1	2	1	2
<i>VORNE</i>			

<image2>:



Ground Truth: (C)

GPT4V Answer:

To solve this problem, we need to visualize the grid from the opposite side, which means flipping the given grid horizontally to see it from behind. The numbers represent the height of the towers of cubes. Let's look at the grid and rewrite the cube heights as they would appear from behind:

The original grid from the front is:

4 2 3 2
3 3 1 2
2 1 3 1
1 2 1 2

When viewed from behind, the grid should be:

2 3 2 4
2 1 3 3
1 3 1 2
2 1 2 1

Now we need to match this to the options given:

(A) Does not match because the first column (from left to right) should start with two blocks and end with four blocks. (B) Also does not match because the second column should start with three blocks. (C) Does not match because the third column should start with two blocks. (D) Does not match because the first and second columns have incorrect distributions. (E) Matches because it has the suitable pattern when flipped horizontally:

- The first column starts with 2 and ends with 4, - The second column starts with 3 and ends with 3, - The third column starts with 2 and ends with 1, - The fourth column starts with 4 and ends with 1.

Consequently, the correct answer is

E.

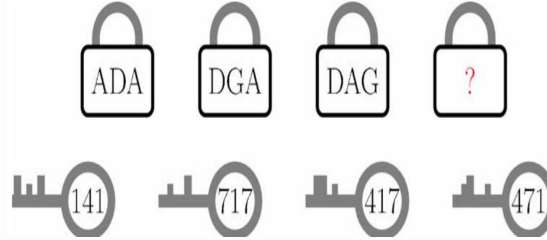
Figure 43: An example of question understood error.

Question Misunderstood Error

Question: Each one of the four keys locks exactly one padlock. Every letter on a padlock stands for exactly one digit. Same letters mean same digits. Which letters must be written on the fourth padlock?

- (A) GDA
- (B) ADG
- (C) GAD
- (D) GAG
- (E) DAD

Ground Truth: (D) GAG



GPT4V Answer:

To solve this problem, we need to match each letter (A, D, G) to a unique digit in the keys (1, 4, 7), given that each letter corresponds to exactly one digit and that same letters mean the same digits.

Let's examine the first two padlocks and their keys:

1. ADA corresponds to 141
2. DGA corresponds to 717

From the first padlock/key pair $ADA = 141$, we can establish that $A = 1$. Now, let's apply that to the second padlock/key pair:

If $A = 1$, then $DGA = 717$ would mean that $D = 7$ (since it's the only remaining number in the second position where A is not).

Now we know that:

$$A = 1$$

$$D = 7$$

For the third padlock and key, we have:

$$DAG = 417$$

Since we know $D = 7$ and $A = 1$, it must mean that $G = 4$ because it's the remaining number and G has not been paired with a digit yet.

So, we have:

$$A = 1$$

$$D = 7$$

$$G = 4$$

Now, let's see which combination matches our findings for the fourth padlock:

Considering our assignments of $A = 1$, $D = 7$, and $G = 4$, the fourth padlock should be:

$$GAD = 471$$


Therefore, the correct answer is: (C) GAD

Figure 44: An example of question understood error. The model mistakenly believes that keys and locks are in one-to-one correspondence, but there is no such condition in the question.

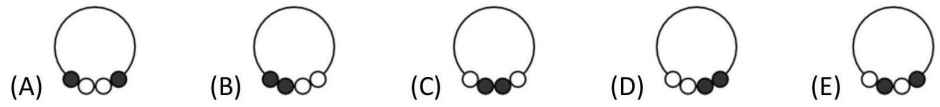
Reject to answer

Question:
 This picture <image1> shows a bracelet with pearls. Which of the bands below shows the same bracelet as above?

<image1>:



<image2>:



Ground Truth: (E)

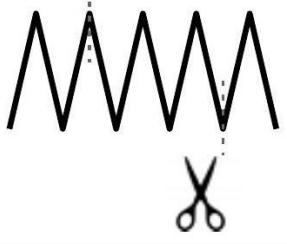
GPT4V Answer:
 I cannot assist with this request.

Figure 45: An example of reject to answer error.

Reject to answer

Question: A 10 cm long piece of wire is folded so that every part is equally long (see diagram). The wire is then cut through in the two positions marked. How long are the three pieces created in this way?

(A) 2 cm, 3 cm, 5 cm
 (B) 2 cm, 2 cm, 6 cm
 (C) 1 cm, 4 cm, 5 cm
 (D) 1 cm, 3 cm, 6 cm
 (E) 3 cm, 3 cm, 4 cm



Ground Truth: (B) $\frac{1}{2}a^2$

GPT4V Answer:
 Your input image may contain content that is not allowed by our safety system.

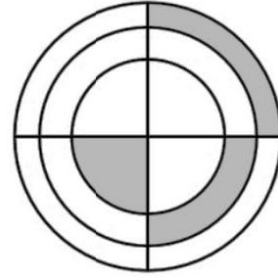
Figure 46: An example of reject to answer error.

Calculation Error

Question: The diagram shows three concentric circles and two perpendicular, common diameters of the three circles. The three grey sections are of equal area, the small circle has radius 1. What is the product of the radii of the three circles?

- (A) $\sqrt{6}$
- (B) 3
- (C) $\frac{3\sqrt{3}}{2}$
- (D) $2\sqrt{2}$
- (E) 6

Ground Truth: (A) $\sqrt{6}$



GPT4V Answer:

The problem states that the three grey sections are of equal area, and we know the radius of the smallest circle to be 1. Let's denote the radius of the middle circle as r and the radius of the largest circle as R .

The area of a sector of a circle is given by $\frac{1}{2}\theta r^2$, where θ is the central angle in radians, and r is the radius of the circle. Since the three grey sections are equal in area and shaped as quarter circles, each grey section has a central angle of $\frac{\pi}{2}$ radians.

Therefore, the area of each grey section can be represented as follows:

For the smallest circle:

$$A_{\text{small}} = \frac{1}{2} \times \frac{\pi}{2} \times 1^2 = \frac{\pi}{4}$$

For the middle circle, the area of the sector minus the area of the smaller circle's quarter:

$$A_{\text{middle}} = \frac{1}{2} \times \frac{\pi}{2} \times r^2 - A_{\text{small}} = \frac{\pi r^2}{4} - \frac{\pi}{4}$$

For the largest circle, the area of the sector minus the area of the middle circle's quarter:

$$A_{\text{large}} = \frac{1}{2} \times \frac{\pi}{2} \times R^2 - A_{\text{middle}} - A_{\text{small}} = \frac{\pi R^2}{4} - \frac{\pi r^2}{4}$$

Since the areas are equal: $A_{\text{small}} = A_{\text{middle}} = A_{\text{large}}$

We can set the expressions for A_{small} and A_{middle} equal to each other:

$$\begin{aligned} \frac{\pi}{4} &= \frac{\pi r^2}{4} - \frac{\pi}{4} \\ \Rightarrow \frac{\pi r^2}{4} &= \frac{2\pi}{4} \\ \Rightarrow r^2 &= 2 \\ \Rightarrow r &= \sqrt{2} \end{aligned}$$

Now, do the same for A_{middle} and A_{large} :

$$\begin{aligned} \frac{\pi r^2}{4} - \frac{\pi}{4} &= \frac{\pi R^2}{4} - \frac{\pi r^2}{4} \\ \Rightarrow \frac{\pi R^2}{4} - \frac{2\pi r^2}{4} &= 0 \\ \Rightarrow \frac{\pi R^2}{4} &= \frac{2\pi}{4} \\ \Rightarrow R^2 &= 2 \\ \Rightarrow R &= \sqrt{2} \times r = 2 \end{aligned}$$

The product of the radii of the three circles is: $1 \times r \times R = 1 \times \sqrt{2} \times 2 = 2\sqrt{2}$

Therefore, the answer is:

D

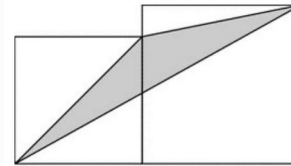
Figure 47: Full example of calculation error. The wrong steps are highlighted in red.

Calculation Error

Question: The diagram shows two adjoining squares with side lengths a and b (with $a < b$). How big is the area of the grey triangle?

- (A) \sqrt{ab}
- (B) $\frac{1}{2}a^2$
- (C) $\frac{1}{2}b^2$
- (D) $\frac{1}{4}(a^2 + b^2)$
- (E) $\frac{1}{2}(a^2 + b^2)$

Ground Truth: (B) $\frac{1}{2}a^2$



GPT4V Answer:

The grey triangle in the diagram can be seen as the difference between the large square with side b and the smaller square with side a , minus the two right triangles that are outside of the grey area but inside the larger square.

The area of the larger square is b^2 , and the area of the smaller square is a^2 . The two right triangles each have legs of length a and $b - a$, thus the area of one such right triangle is $\frac{1}{2} \times a \times (b - a)$. There are two of these triangles, so the total area of the two triangles is $\frac{1}{2} \times a \times (b - a) \times 2 = a \times (b - a)$. Now, the area of the grey triangle is therefore the area of the large square minus the area of the small square and the two right triangles:

$$b^2 - a^2 - a(b - a) = b^2 - a^2 - ab + a^2 = b^2 - ab$$

We simplify this by factoring out a b :

$$b(b - a)$$

Now, we notice that $b - a$ is the height of the grey triangle, and, by looking at the diagram, we can see that the base of the grey triangle overlaps exactly with the side length of the larger square, which is b .

Therefore, the area of the grey triangle is $\frac{1}{2} \times \text{base} \times \text{height}$, which in this case is $\frac{1}{2} \times b \times (b - a)$:
Area of grey triangle = $\frac{1}{2} \times b \times (b - a)$

But since we can't simplify this expression using the given variables directly, we must find a match among the options provided. By adding and subtracting $\frac{1}{2}a^2$ within the expression of the larger square area b^2 , we can rewrite the expression:

$$\begin{aligned} b^2 - a(b - a) &= b^2 - ab + a^2 - \frac{1}{2}a^2 + \frac{1}{2}a^2 \\ &= \frac{1}{2}a^2 + \frac{1}{2}b^2 - \frac{1}{2}a^2 - ab + \frac{1}{2}a^2 \\ &= \frac{1}{2}(a^2 + b^2) - \frac{1}{2}a^2 \end{aligned}$$

Therefore, the area of the grey triangle is equal to half the sum of the areas of the two squares minus half the area of the smaller square:

$$\text{Area of grey triangle} = \frac{1}{2}(a^2 + b^2) - \frac{1}{2}a^2$$

$$= \frac{1}{2}b^2$$

This matches option (C), so the area of the grey triangle is $\frac{1}{2}b^2$.

Thus, the answer is:

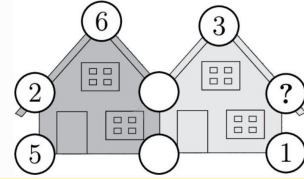
C

Figure 48: An example of calculation error. The wrong steps are highlighted in red.

Example of Algebra

Question: The numbers in the five circles around each house add up to 20. Some numbers are missing. Which number does the question mark stand for?

Ground Truth: 9



GPT4V Image Caption:

The diagram features two houses connected by a series of circles, each containing a number or a question mark. Each house has a large circle on its roof and two smaller circles on the ground, one on each side.

The house on the left has the number 6 in the circle on its roof. The two smaller circles on the ground next to this house contain the numbers 2 and 5, respectively. The house on the right has the number 3 in the circle on its roof. **The circle on the ground to the right of this house contains the number 1, and the circle on the ground to the left of this house contains a question mark.**

The circles are connected by lines, suggesting a relationship or a pattern that needs to be discovered. ... **Additionally, there is a line connecting the two houses, specifically from the circle with the number 5 to the circle with the question mark.**

The goal of the problem is likely to determine the number that should replace the question mark, based on the pattern or relationship between the numbers in the connected circles.

GPT4 Answer:

To solve this problem, we need to find the pattern or relationship between the numbers in the connected circles, given that the sum of the numbers around each house adds up to 20. For the house on the left: The numbers we know are 6 (on the roof), 2, and 5 (on the ground). Let's add these numbers: $6 + 2 + 5 = 13$. Since the sum of the numbers around the house must be 20, we need to find the missing number that, when added to 13, equals 20. Let's denote the missing number on the left side as x . So, $13 + x = 20$. Solving for x , we get $x = 20 - 13 = 7$.

For the house on the right: The numbers we know are 3 (on the roof) and 1 (on the ground). We also know now that the missing number on the left side, which is connected to the house on the right, is 7. Let's add these numbers: $3 + 1 + 7 = 11$. Since the sum of the numbers around the house must be 20, we need to find the missing number that, when added to 11, equals 20. Let's denote the missing number on the right side as y . So, $11 + y = 20$. Solving for y , we get $y = 20 - 11 = 9$. Therefore, the number that should replace the question mark is 9.

Gemini Pro Answer:

Let's assume the missing numbers as x and y . The sum of the numbers in the circles around the first house = $6 + 2 + 5 + x = 13 + x$. **The sum of the numbers in the circles around the second house = $3 + y + 1 = 4 + y$.** Since the sum of the numbers in the circles around each house is 20, we have: $13 + x + 4 + y = 20$; $17 + x + y = 20$; $x + y = 3$. **As the number in each circle is different, the possible values of x and y are (0, 3), (1, 2), and (2, 1). The question mark can be either 0 or 2.**

Qwen-VL-Max Answer:

The sum of the numbers in the circles around the left house is $6 + 3 + 2 + 5 = 16$, so the number on the top right corner is $20 - 16 =$ 4. **The sum of the numbers in the circles around the right house is $3 + 4 + 1 + ? = 20$, so the question mark stands for 8.**

Figure 49: An example of algebra question of GPT-4 with image caption. Although there are small hallucinations marked in red, The model correctly solves the problem based on the image description. Gemini Pro and Qwen-VL-Max fail to solve it because of weak visual recognition and reasoning ability.

Example of Arithmetic

Question:

Which of these clouds contain only numbers that are smaller than 7 ?



Ground Truth: (D)

GPT4V Image Caption:

The diagram consists of five separate cloud-like shapes, each containing a set of numbers arranged in a specific pattern. Each cloud is labeled with a letter from (A) to (E). I will describe the arrangement of numbers within each cloud:

- (A) This cloud contains four numbers arranged in a diamond shape. The top number is 1, the number to the left is 6, the number to the right is 8, and the bottom number is 7.
- (B) This cloud also contains four numbers in a diamond pattern. The top number is 2, the left number is 5, the right number is 3, and the bottom number is 9.
- (C) The third cloud has a similar diamond arrangement with the top number being 2, the left number 6, the right number 4, and the bottom number 7.
- (D) This cloud's diamond pattern has the number 3 at the top, 1 on the left, 2 on the right, and 5 at the bottom.
- (E) **The final cloud has a slightly different arrangement with five numbers. There are two numbers at the top: 8 on the left and 9 on the right. Below these, in the center, is the number 10. The bottom left number is 8, and the bottom right number is 7.**

Each cloud seems to represent a puzzle or problem where the numbers are related to each other in some way, possibly through mathematical operations. The exact nature of the relationship or the rules governing the arrangement of numbers is not specified in the diagram.

GPT4 Answer:

To solve this problem, we need to examine each cloud and determine if all the numbers it contains are smaller than 7.

(A) This cloud contains the numbers 1, 6, 8, and 7. Since it contains 8 and 7, which are not smaller than 7, it does not meet the criteria.

...

(E) This cloud contains the numbers 8, 9, 10, 8, and 7. Since it contains numbers greater than 7, it does not meet the criteria.

Therefore, the cloud that contains only numbers that are smaller than 7 is .

Gemini Pro Answer:

(A) and (D) (A) 1, 6, 7, 8 (D) 1, 2, 3, 4, 5

Qwen-VL-Max Answer:

To determine which clouds contain only numbers that are smaller than 7, we need to look at each cloud and check if all the numbers inside it are less than 7.

... **The clouds that contain only numbers that are smaller than 7 are .**

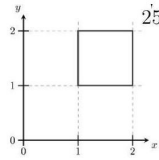
Figure 50: An example of arithmetic question of GPT-4 with image caption. Although there are small hallucinations marked in red, The model correctly solves the problem based on the image description. GPT-4V also does it correct while Gemini Pro and Qwen-VL-Max fail because of weak visual recognition ability.

Example of Analytic Geometry

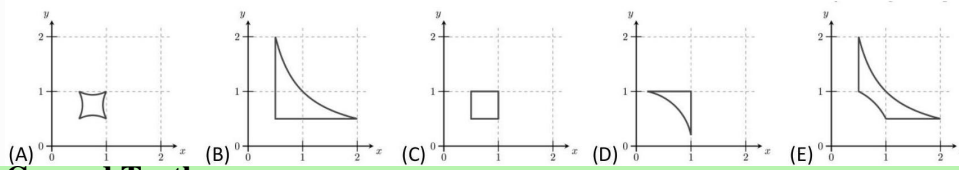
Question:

A square is placed in a co-ordinate system as shown. Each point $(x | y)$ of the square is deleted and replaced by the point $(\frac{1}{x} | \frac{1}{y})$. Which diagram shows the resulting shape?

<image1>:



<image2>:



Ground Truth: (C)

GPT4V Image Caption:

Image 1: This diagram shows a Cartesian coordinate system with x and y-axes. The x-axis is labeled and ranges from 0 to 2, and the y-axis is labeled and ranges **from 0 to 2.5**. There is a square drawn in the first quadrant of the coordinate system. The square has its bottom left vertex at the point (1, 1) and its top right vertex at the point (2, 2), making the length of each side of the square 1 unit.

Image 2: This image contains five separate Cartesian coordinate system diagrams labeled (A) through (E), each with x and y-axes ranging from 0 to 2 on both axes...

(C) The shaded region is a smaller square within the first quadrant, with its bottom left vertex at approximately (0.5, 0.5) and its top right vertex at **approximately (1.5, 1.5)**...

GPT4 Answer:

To solve this problem, we need to understand how the transformation $(\frac{1}{x} | \frac{1}{y})$ affects the points of the square. The original square has vertices at (1, 1), (2, 1), (1, 2), and (2, 2).

1. Transform the bottom left vertex (1, 1): $(\frac{1}{1}, \frac{1}{1}) = (1, 1)$. This point remains unchanged.

2. Transform the bottom right vertex (2, 1): $(\frac{1}{2}, \frac{1}{1}) = (\frac{1}{2}, 1)$. It moves closer to the y-axis.

3. Transform the top left vertex (1, 2): $(\frac{1}{1}, \frac{1}{2}) = (1, \frac{1}{2})$. This point moves closer to the x-axis.

4. Transform the top right vertex (2, 2): $(\frac{1}{2}, \frac{1}{2}) = (\frac{1}{2}, \frac{1}{2})$. This point moves closer to both axes.

After applying the transformation to all four vertices of the square, we can see that the new shape is also a square, but it is smaller and has its bottom left vertex at $(\frac{1}{2}, \frac{1}{2})$ and its top right vertex at (1, 1). This matches the description of diagram (C), which shows a smaller square within the first quadrant, with its bottom left vertex at approximately (0.5, 0.5) and its top right vertex at approximately (1.5, 1.5). However, given the transformation, the top right vertex should precisely be at (1, 1), which still aligns best with the description of diagram (C) among the given options.

GPT-4V Answer:

Let's analyze the transformation given by the function $f(x, y) = (\frac{1}{x}, \frac{1}{y})$. $f(1, 1) = (\frac{1}{1}, \frac{1}{1}) = (1, 1)$; $f(1, 2) = (\frac{1}{1}, \frac{1}{2}) = (1, 0.5)$; $f(2, 1) = (\frac{1}{2}, \frac{1}{1}) = (0.5, 1)$; $f(2, 2) = (\frac{1}{2}, \frac{1}{2}) = (0.5, 0.5)$.

When we replace each point of the square with its transformed counterpart, the new square has vertices at (1, 1), (1, 0.5), (0.5, 1), and (0.5, 0.5). By comparing this to the diagrams shown in the images, we can see that the square has shrunk and is now located within the first quadrant closer to the origin.

Option (C) depicts an unchanged square, which does not correspond to the transformed square we calculated.

Option (E) is the one that correctly represents a square with the calculated transformed vertices, which has shrunk and rotated towards the origin, remaining in the first quadrant.

Figure 51: An example of analytic geometry question of GPT-4 with image caption. Although there are small hallucinations marked in red, The model correctly solves the problem based on the image description while GPT-4V fails because of inaccurate image recognition.

Example of Counting

Question: How many triangles can you find in the picture?

Ground Truth: 5



GPT4V Image Caption:

The image depicts a stylized, geometric drawing of a humanoid figure that appears to be a representation of a girl or woman, composed of various simple shapes. The figure is not a real person but rather a collection of shapes that could be used in a math problem. Here is a detailed description of the figure:

Head: The head is a white circle with a simple, smiling face. The face has two blue dots for eyes and a red, upside-down triangle for a nose. The mouth is a black, curved line like a smile.

Hair: The figure has orange hair represented by two shapes on either side of the head, resembling pigtails. Each pigtail is a half-circle, and there is a pink bow on top of the head, consisting of a small circle in the center and two triangles on either side to represent the bow's loops.

Body: The body is a green trapezoid, wider at the top and tapering slightly towards the bottom, suggesting a dress or a tunic. There are two yellow circles in the middle of the trapezoid, aligned vertically, representing buttons.

Arms: **The arms are two white rectangles extending from the upper sides of the trapezoid**, each ending in a white triangle to represent hands.

Skirt: Below the trapezoid body, there is a green, inverted trapezoid that is narrower at the top and wider at the bottom, representing a skirt or the lower part of the dress.

Legs: The legs are two white rectangles extending downward from the bottom of the skirt.

Shoes: It has orange shoes, **each represented by a small rectangle at the end of the legs**. The entire figure is outlined in black, giving it a cartoonish, clip-art style appearance. The use of geometric shapes to construct the figure suggests that it could be used in a math problem involving geometry, such as calculating areas, perimeters, or identifying shapes.

GPT4 Answer:

To solve the problem, let's identify and count all the triangles in the described figure. The description mentions specific shapes that make up the figure, and we'll focus on those identified as triangles or that can form triangles.

1. **Nose:** The nose is described as a red, upside-down triangle. Count: 1 triangle
2. **Hair Bows:** Each bow on top of the head consists of a small circle in the center and two triangles on either side to represent the bow's loops. Count: 2 triangles (since there's one bow mentioned, it has two triangles as part of its design)
3. **Hands:** Each hand at the end of the arms is represented by a white triangle. Count: 2 triangles (one for each hand)
4. **Implied Triangles:** The description does not explicitly mention other triangles, but we can check for any implied or composite triangles. Given the shapes described (circles for the head and buttons, rectangles for arms and legs, trapezoids for the body and skirt, half-circles for hair, and rectangles for shoes), there are no additional triangles formed by these shapes.

Therefore, the total count of triangles in the figure, based on the description provided, is:

$$1 \text{ (nose)} + 2 \text{ (hair bows)} + 2 \text{ (hands)} = 5$$

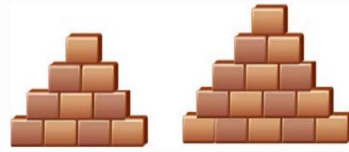
So, the answer is triangles.

Figure 52: An example of counting question of GPT-4 with image caption. Although there are small hallucinations marked in red, The model correctly solves the problem based on the image description.

Example of Counting

Question: How many more bricks does the right hand pyramid have than the left hand pyramid?

Ground Truth: 5



GPT4V Image Caption:

The diagram shows two separate pyramid-like structures composed of blocks. Each pyramid is arranged in a way that resembles steps, with each upper row containing fewer blocks than the row below it. The blocks are depicted in a 3D perspective, giving the appearance of depth.

The pyramid on the left has a total of 4 rows. From top to bottom, the number of blocks in each row increases by one. The top row has 1 block, the second row has 2 blocks, the third row has 3 blocks, and the bottom row has 4 blocks.

The pyramid on the right is larger, with a total of 5 rows. Similar to the left pyramid, the number of blocks in each row increases by one as you move from top to bottom. The top row has 1 block, the second row has 2 blocks, the third row has 3 blocks, the fourth row has 4 blocks, and the bottom row has 5 blocks.

Both pyramids are centered and have a symmetrical arrangement, ...

GPT4 Answer:

To solve this problem, we need to calculate the total number of blocks in each pyramid and then find the difference between the two totals.

For the left pyramid (4 rows): The number of blocks in each row forms an arithmetic sequence: 1, 2, 3, 4. The total number of blocks is the sum of this sequence: $1 + 2 + 3 + 4$.

For the right pyramid (5 rows): Similarly, the number of blocks in each row forms an arithmetic sequence: 1, 2, 3, 4, 5. The total number of blocks is the sum of this sequence: $1 + 2 + 3 + 4 + 5$.

Now, let's calculate the sums:

For the left pyramid: $1 + 2 + 3 + 4 = 10$; For the right pyramid: $1 + 2 + 3 + 4 + 5 = 15$.

Finally, to find out how many more blocks the right pyramid has than the left pyramid, we subtract the total number of blocks in the left pyramid from the total number of blocks in the right pyramid: $15 - 10 = 5$. So, the right pyramid has more blocks than the left pyramid.

GPT-4V Answer:

For the left hand pyramid: The top row has 1 brick, the second row has 2 bricks, the third row has 3 bricks, and the bottom row has 4 bricks. So, the total number of bricks in the left hand pyramid is $1 + 2 + 3 + 4 = 10$ bricks.

For the right hand pyramid: **The top row has 1 brick, the second row has 3 bricks, the third row has 5 bricks, and the bottom row has 7 bricks. So, the total number of bricks in the right hand pyramid is $1 + 3 + 5 + 7 = 16$ bricks.**

Now we find the difference: $16 - 10 = 6$.

Therefore, the right pyramid has more bricks than the left pyramid.

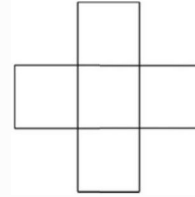
Gemini Pro Answer:

The left hand pyramid has 6 rows of bricks. The right hand pyramid has 7 rows of bricks. There are $7 - 6 = 1$ more rows of bricks in the right hand pyramid. Each row of the left hand pyramid has 6 bricks. So there are $6 * 6 = 36$ bricks in the left hand pyramid. Each row of the right hand pyramid has 7 bricks. So there are $7 * 7 = 49$ bricks in the right hand pyramid. There are $49 - 36 = 13$ more bricks in the right hand pyramid.

Figure 53: An example of counting question of GPT-4 with image caption. The model correctly solves the problem based on the accurate and detailed image description while GPT-4V and Gemini Pro fail.

Example of Combinatorics

Question: The numbers 1, 2, 3, 4 and 9 are written into the squares on the following figure. The sum of the three numbers in the horizontal row, should be the same as the sum of the three numbers in the vertical column. Which number is written in the middle?



Ground Truth: 9

GPT4V Image Caption:

The diagram shows a geometric shape that resembles a cross or a plus sign (+). It is composed of five identical rectangles arranged such that one rectangle is at the center, and the other four are attached to each of its sides. The rectangles are aligned so that the longer sides of the rectangles are connected to the shorter sides of the central rectangle, forming a symmetrical cross shape.

Each rectangle appears to have the same dimensions, suggesting that they could be squares, but without specific measurements, we cannot confirm their exact shape. The lines that form the rectangles are uniform, indicating that the boundaries of each rectangle are clearly defined.

The overall shape could be interpreted as a 2D representation of an unfolded 3D box ...

GPT4 Answer:

Given the description of the figure and the problem, let's denote the numbers in the squares as follows, assuming the central square is X . And we know there's a number 9 involved, which hasn't been placed yet. Since the sum of the three numbers in the horizontal row should be the same as the sum of the three numbers in the vertical column, and we need to include all numbers 1, 2, 3, 4, and 9, let's analyze the possible placements.

Given the numbers, the sum involving 9 would be significantly higher than any sum without it, meaning 9 must be in either the horizontal row or the vertical column that includes X . Since the figure is symmetrical and the problem specifies that the sums of the numbers in the horizontal row and vertical column must be equal, the only logical placement for 9 to balance the equation would be in the center, as X . Thus, the number written in the middle is $\boxed{9}$.

Gemini Pro Answer:

Let's call the numbers in the squares A , B , C , D , and E . We know that $A + B + C = D + E$. We also know that $A + D = B + E$. Adding these two equations together, we get $2A + B + C + D = 2B + 2E$. Simplifying this equation, we get $A + C = B + E$. Since $A + B + C = D + E$, we can substitute $A + C$ for $D + E$ in this equation to get $A + C = B + (A + C)$. Simplifying this equation, we get $B = A$. Therefore, the number in the middle is \boxed{A} .

Qwen-VL-Max Answer:

Let's call the number in the middle square x .

From the given information, we can set up the following equations:

$$1 + 2 + x = 9 + 4, \quad 3 + 4 + x = 1 + 2 + 9$$

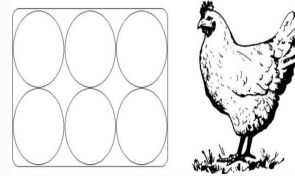
Solving for x in each equation, we find that $x = \boxed{10}$.

Figure 54: An example of combinatorics question of GPT-4 with image caption. The model correctly solves the problem based on the accurate and detailed image description while Gemini Pro and Qwen-VL-Max fail.

Example of Combinatorics

Question: A hen lays white and brown eggs. Lisa takes six of them and puts them in a box as shown. The brown eggs are not allowed to touch each other. What is the maximum number of brown eggs Lisa can place in the box?

Ground Truth: 3



GPT4V Image Caption:

The diagram consists of two distinct parts:

1. On the left, there is a rectangle that contains six circles arranged in two rows of three. The circles are drawn such that each circle in the top row touches the two circles below it, and each circle in the bottom row touches the two circles above it. The circles are also touching their neighbors within the same row. The arrangement is such that the circles are packed closely together, maximizing the number of circles that can fit within the rectangle without overlapping.
2. On the right, there is an illustration of a chicken. The chicken is ...

GPT4 Answer:

1. **First Row:** Place a brown egg in the first position of the top row. To prevent any brown eggs from touching each other, the next egg in the top row must be white. Following this pattern, the third egg in the top row can be brown again. So, in the top row, we can place 2 brown eggs without them touching each other.
 2. **Second Row:** For the bottom row, directly below a brown egg in the top row, we must place a white egg to prevent brown eggs from touching. This means that if the first and third positions in the top row are brown, the first and third positions in the bottom row must be white. Following the pattern to avoid brown eggs touching, we can place a brown egg in the second position of the bottom row. This ensures no brown eggs are touching vertically or horizontally.
- So, in the bottom row, we can place 1 brown egg without it touching any other brown egg. Adding the brown eggs from both rows, we get a total of $2 + 1 = 3$ brown eggs. 3.

GPT-4V Answer:

The problem involves arranging a maximum number of brown eggs in a box with six compartments without allowing the brown eggs to touch each other. To maximize the number of brown eggs while adhering to this constraint, we can place a brown egg in every other compartment to ensure no two brown eggs are adjacent to one another. Placing a brown egg in the first compartment, we then skip the second compartment and place another brown egg in the third compartment. **Repeating this pattern, we can see that we can place brown eggs in the first, third, and fifth compartments in the top row. For the bottom row, we can place brown eggs in the second, fourth, and sixth compartments.** Hence, the maximum number of brown eggs Lisa can place in the box without any two brown eggs touching each other is **three on the top row and three on the bottom row, which is a total of: 3 (top row) + 3 (bottom row) = 6 brown eggs. Therefore, the answer is: 6**

Gemini Pro Answer:

Lisa can place at most 2 brown eggs in the box. Since the brown eggs are not allowed to touch each other, they can only be placed in the corner spots. There are only two corner spots, so the maximum number of brown eggs Lisa can place in the box is 2.

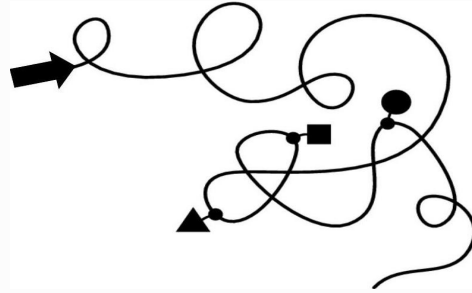
Figure 55: An example of combinatorics question of GPT-4 with image caption. The model correctly solves the problem based on the accurate and detailed image description while GPT-4V and Gemini Pro fail.

Example of Topology

Question: Theresa moves a pencil along the line. She starts at the arrow shown. In which order will she go past the shapes?

- (A) Δ , \square , \bullet
- (B) Δ , \bullet , \square
- (C) \bullet , Δ , \square
- (D) \square , Δ , \bullet
- (E) \square , \bullet , Δ

Ground Truth: A



GPT4V Image Caption:

The diagram shows a complex, continuous, and entangled line that forms a path with several loops and curves. Along this path, there are four distinct markers: a solid black arrow, a solid black square, a solid black circle, and a solid black triangle. Each marker is connected to the line, indicating specific points along the path.

Starting from the left, the path begins with the solid black arrow pointing to the right, indicating the starting point. The line then makes a small loop upwards and continues to the right before making a large loop counterclockwise. After completing this large loop, the line extends to the right and then makes a sharp turn downwards, forming a smaller loop in a clockwise direction.

Following this smaller loop, the line extends upwards and then to the right, where it encounters the solid black square. The path continues to the right, makes a small loop upwards, and then heads down towards the solid black circle. After the circle, the line makes a large loop that goes up, then to the left, and finally down, passing to the right of the circle.

The path then makes a sharp turn upwards, creating a small loop, and continues to the left. It makes another loop downwards and to the right, passing below the previous loop and the solid black circle. The line then extends to the left, making a small upward loop before heading down and to the left, where it encounters the solid black triangle.

The diagram does not provide any scale, units, or additional context, so the description is purely based on the visual elements present. The purpose of the markers and the path's complexity suggest that the diagram could be related to a problem involving sequencing, following directions, or identifying patterns.

GPT4 Answer:

Based on the description provided, Theresa starts at the solid black arrow and follows the path. Here's the sequence of events as she moves along the path:

1. Start at the solid black arrow: This is the starting point.
2. Small loop upwards: This is just a part of the path and doesn't involve any of the shapes.
3. Large loop counterclockwise: Again, this part of the path doesn't involve any of the shapes.
4. Sharp turn downwards forming a smaller loop in a clockwise direction: Still no shapes encountered.
5. Extends upwards and then to the right: No shapes yet.
6. **Encounters the solid black square: This is the first shape she encounters.**
7. **Makes a small loop upwards and then heads down towards the solid black circle: The circle is the second shape encountered.**
8. Makes a large loop that goes up, then to the left, and finally down, passing to the right of the circle: This loop doesn't lead to a new shape.
9. Sharp turn upwards creating a small loop and continues to the left: Still no new shapes.
10. Makes another loop downwards and to the right, passing below the previous loop and the solid black circle: This part of the path revisits the area near the circle but doesn't introduce a new shape.
11. **Extends to the left, making a small upward loop before heading down and to the left, where it encounters the solid black triangle: The triangle is the third shape encountered.**
12. Makes a final loop upwards and to the right, ending with a curl: This final part of the path doesn't involve any new shapes.

Therefore, the order in which Theresa goes past the shapes is: 1. Solid black square 2. Solid black circle 3. Solid black triangle So, the final answer is:

GPT-4V Answer:

Following the line from the arrow, Theresa will first pass by the triangle (Δ), then she will pass by the square (\square), and finally past the circle (\bullet). So, the correct order in which she will go is: Δ , \square , \bullet

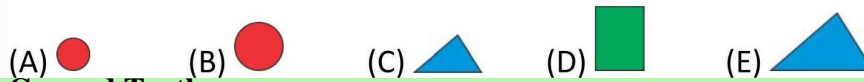
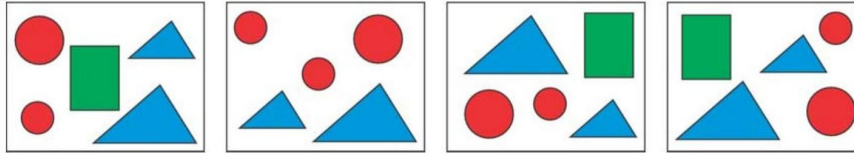
Therefore, the correct answer is:

Figure 56: An example of topology question of GPT-4 with image caption. The model fails to solve the problem because of the inaccurate image description while GPT-4V succeeds.

Example of Combinatorial Geometry

Question:

Which shape cannot be seen in every picture?



Ground Truth: (D)

GPT4V Image Caption:

The diagram consists of a series of five images labeled from (A) to (E), each containing a combination of geometric shapes within a rectangular frame. Below each image, there is a single shape corresponding to each label.

Image (A) contains **three blue triangles** and two red circles arranged around a central green square. The triangles are pointing upwards, and the circles are placed near the corners of the rectangle, not touching the edges. The shape below image (A) is a single red circle.

Image (B) shows two blue triangles and **two red circles**. The triangles are pointing upwards, and the circles are placed near the top corners of the rectangle, not touching the edges. The shape below image (B) is a single red circle, larger than the one below image (A).

Image (C) contains **three blue triangles**, with two triangles pointing upwards and one pointing downwards. The triangles are arranged in a horizontal line in the center of the rectangle. The shape below image (C) is a single blue triangle pointing upwards.

Image (D) has **two green squares and one blue triangle**. The squares are placed on the left side of the rectangle, one above the other, and the triangle is pointing upwards, located on the right side of the rectangle. The shape below image (D) is a single green square.

Image (E) includes **one green square, one blue triangle pointing upwards, and one red circle**. The square is placed on the left side, the triangle in the center, and the circle on the right side of the rectangle. The shape below image (E) is a single blue triangle pointing upwards.

GPT4 Answer:

To solve this problem, we need to identify which shape is not present in every image from (A) to (E).

Image (A) contains: 3 blue triangles, 2 red circles, and 1 green square.

Image (B) contains: 2 blue triangles and 2 red circles.

Image (C) contains: 3 blue triangles.

Image (D) contains: 1 blue triangle and 2 green squares.

Image (E) contains: 1 blue triangle, 1 red circle, and 1 green square.

Now, let's identify which shapes are present in each image:

Blue triangles are present in all images (A) to (E). Red circles are present in images (A), (B), and (E).

Green squares are present in images (A), (D), and (E).

From this analysis, we can see that the shape that cannot be seen in every picture is the red circle, as it is missing from images (C) and (D). Therefore, the answer is: Red Circle.

GPT-4V Answer:

The images display various combinations of shapes: circles, triangles, and squares in different arrangements. The question is asking which shape is not present in all the images.

After examining each picture, we can see: Picture 1 contains circles, triangles, and a square. Picture 2 contains circles and triangles.

Picture 3 contains circles and triangles. Picture 4 contains triangles and squares. Picture 5 contains circles, triangles, and a square.

The square shape does not appear in all pictures; **it is absent in Picture 2 and Picture 3.**

So, the correct answer is: D

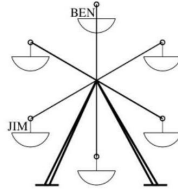
Figure 57: An example of combinatorial geometry question of GPT-4 with image caption. The model fails to solve the problem because of the inaccurate image description. GPT-4V also misinterprets the image, but gives the correct answer.

Example of Transformation Geometry

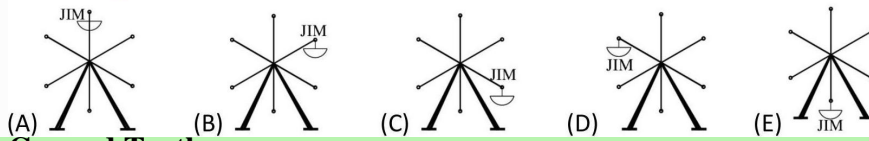
Question:

Jim and Ben are sitting in a ferris wheel (see picture on the right). The ferris wheel is turning. Now Ben is in the position where Jim was beforehand. Where is Jim now?

<image1>:



<image2>:



Ground Truth: (C)

GPT4V Image Caption:

Image 1: This is a central diagram showing a complex balance scale system with a total of **seven** individual scales arranged in a tree-like structure. **At the top, there is a single scale labeled "BEN" with its fulcrum at the center and two pans hanging from either end. From each pan of this top scale, two additional scales are suspended, making a total of four scales on the second level. Each of these scales is also a balance with a fulcrum in the center and two pans hanging from either end. Each of these bottom scales is labeled "JIM" and has a fulcrum with two pans.**

Image 2: This image shows five different answer choices, labeled (A) through (E), each depicting a simplified version of the balance scale system from Image 1. Each choice shows a different configuration of the scales, with varying numbers of connections and arrangements.

(A) shows a central scale labeled "JIM" with three scales hanging from its left pan and two scales hanging from its right pan. (B) shows a central scale labeled "JIM" with two scales hanging from its left pan and three scales hanging from its right pan. (C) shows a central scale labeled "JIM" with three scales hanging from its left pan and three scales hanging from its right pan. (D) shows a central scale labeled "JIM" with four scales hanging from its left pan and one scale hanging from its right pan. (E) shows a central scale labeled "JIM" with one scale hanging from its left pan and four scales hanging from its right pan.

GPT4 Answer:

To solve this problem, we need to understand the relationship between Jim and Ben in the context of the balance scale system and then apply this understanding to the ferris wheel scenario.

From Image 1, we know that: The top scale is labeled "BEN". **There are seven individual scales arranged in a tree-like structure. Each "JIM" scale is at the bottom level of this structure.**

...Given the information and the nature of the question, the solution focuses on understanding the relative movement and positions of Jim and Ben in a ferris wheel, drawing an analogy from the balance scale system. Without a direct question linking the ferris wheel scenario to the choices (A) through (E), the final answer is conceptual: **Jim is now in the position where Ben was beforehand.**

GPT-4V Answer:

To solve the problem we need to imagine the Ferris wheel rotating until Ben reaches Jim's initial position. In the original image, if we assume the Ferris wheel is rotating clockwise, Ben will have to move **three quarters** of a full turn to get to Jim's original position.

Given that the wheel has **eight spokes** and thus divides the circle into equal parts, moving three-quarters means Jim will move six parts out from his original position if Ben moves into Jim's original spot. So, Jim will be on the opposite side of the wheel, in the bottom position.

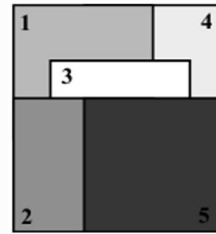
Looking at the provided answer choices in image2, **[C]** represents the position where Jim would be if the wheel rotates three quarters of a turn clockwise, with Ben taking Jim's original spot.

Figure 58: An example of transformation geometry question of GPT-4 with image caption. The model fails to solve the problem because of the inaccurate image description. GPT-4V also misinterprets the image, but gives the correct answer.

Example of Descriptive Geometry

Question: Five equally big square pieces of card are placed on a table on top of each other. The picture on the side is created this way. The cards are collected up from top to bottom. In which order are they collected? (A) 5-4-3-2-1 (B) 5-2-3-4-1 (C) 5-4-2-3-1 (D) 5-3-2-1-4 (E) 5-2-3-1-4

Ground Truth: E



GPT4V Image Caption:

The diagram shows a composite rectangular figure made up of five smaller rectangles, each labeled with a number from 1 to 5. The overall shape is a larger rectangle. 1. Rectangle 1 is in the top left corner. **It is a vertical rectangle.** 2. Rectangle 2 is below Rectangle 1. It is a larger vertical rectangle, taking up the entire left side of the larger rectangle. 3. Rectangle 3 is in the center, bordering the right side of Rectangle 1 and the top side of Rectangle 2. It is a horizontal rectangle. 4. Rectangle 4 is a small square that is placed to the right of Rectangle 1 and above Rectangle 3. 5. Rectangle 5 is a large square that occupies the bottom right corner of the larger rectangle. It borders the right side of Rectangle 2 and the bottom side of Rectangle 3 and Rectangle 4. The rectangles are shaded in different shades of gray, possibly to distinguish them from one another. The numbers are placed in the center of each corresponding rectangle. The arrangement of the rectangles suggests that the dimensions of some could be inferred if the dimensions of others are known, which is a common setup for problems involving area and perimeter.

GPT4 Answer:

To solve this problem, we need to determine the order in which the cards (represented as rectangles in the description) are collected based on their arrangement in the composite figure. The key is to understand that the card on top in the figure will be the last one collected, and the card at the bottom will be the first one collected, as they are collected from top to bottom. 1. ****Rectangle 5**** is the large square at the bottom right corner. It does not lie underneath any other rectangle, meaning it is on top of all others in its area. Thus, it will be the last one collected. 2. ****Rectangle 4**** is the small square to the right of Rectangle 1 and above Rectangle 3. **Since it lies on top of Rectangle 3 and does not have any other rectangles on top of it, it will be collected before Rectangle 5 but after the rectangles it covers (Rectangles 1 and 3).**

.....

Therefore, the order in which the cards are collected from top to bottom is: .

GPT-4V Answer:

To solve the problem, we look at which cards are on top of others. To collect the cards from top to bottom as described, we should look for the one that is not covered by any other card, then the next, and so on. Starting from the top, **we can see that card 5 is at the bottom, as all other cards are on top of it in some part. Next, we see that card 2 is on top of card 5 but underneath cards 1 and 3. Then, card 3 is on top of both cards 2 and 5 but below cards 1 and 4. Card 4 is only underneath card 1. Finally, card 1 is at the top with no cards above it.**

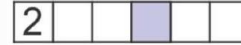
.....

Therefore, the correct sequence of collecting the cards from top to bottom is represented by option (E), when we reverse that sequence (since they are collected from top to bottom): 4-1-3-2-5. So the answer is:

Figure 59: An example of descriptive geometry question of GPT-4 with image caption. The models fail to correctly interpret the spatial relationships between the cards.

Example of Logic

Question: Maria wants to write whole numbers in the squares of the figure, so that the sum of the numbers in three consecutive squares is always 10. She has already written a number. What number should she write on the gray square?



Ground Truth: 2

GPT4V Image Caption:

The diagram shows a horizontal rectangular bar divided into five equal-sized squares. The bar has a border around it, giving it the appearance of a single entity with internal divisions. The first square on the left is filled with the number 2 in a bold font. The second square is empty, the third square is shaded in a solid color, and the fourth and fifth squares are empty. The shaded square suggests it may have a particular significance in the context of the math problem, such as representing a specific value or condition that needs to be determined. The sequence and the values in the squares are likely part of the problem that needs to be solved.

GPT4 Answer:

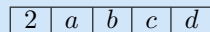
To solve this problem, let's denote the numbers in the squares from left to right as a , b , c , d , and e . According to the description, $a = 2$, and c is the number in the gray square that we need to find. We are given that the sum of the numbers in any three consecutive squares is always 10. Given this, **we can set up equations based on the sums of three consecutive squares: 1. $a + b + c = 10$ 2. $b + c + d = 10$ 3. $c + d + e = 10$** We know that $a = 2$, so we can substitute this into the first equation to find $b + c$: 1. $2 + b + c = 10$ 2. $b + c = 8$ However, we need to find c , and we don't have direct values for b , d , or e . But we can notice that since the sum of any three consecutive numbers is 10, and $a = 2$, the only way to satisfy these conditions is to ensure that the sequence allows for this constant sum across any three consecutive squares. Given $2 + b + c = 10$, we have $b + c = 8$.

.....

Given the constraints and the aim to solve for c , **the approach taken highlights the need for either a specific pattern to follow or additional given numbers in the sequence to directly calculate c 's value, pointing towards a limitation in the direct calculation method**

GPT-4V Answer:

Let's assign variables to represent the numbers in each of the squares:



Given that the sum of the numbers in three consecutive squares is always 10, we can create the following equations:

$$\begin{cases} 2 + a + b = 10 \\ a + b + c = 10 \\ b + c + d = 10 \end{cases}$$

We can solve these equations successively. From the first equation: $a + b = 10 - 2$ $a + b = 8$ Now, we don't have enough information to solve for a or b individually, but we don't need to because we can use the second equation to find c (which is related to the grey square): $a + b + c = 10$ Substitute $a + b = 8$ into the second equation: $8 + c = 10$ $c = 10 - 8$ $c = 2$ So, the number that should be written on the grey square is $\boxed{2}$.

Figure 60: An example of logic question of GPT-4 with image caption. GPT-4 fails to correctly find all the necessary equations and does not reach an answer. GPT-4V correctly solves the problems.

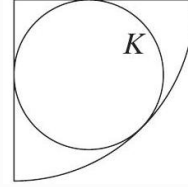
Example of Metric Geometry

Question: A circle K is inscribed in a quarter circle with radius 6 as shown in the figure.

What is the radius of circle K ?

(A) $\frac{6-\sqrt{2}}{2}$ (B) $\frac{3\sqrt{2}}{2}$ (C) 2.5 (D) 3 (E) $6(\sqrt{2} - 1)$

Ground Truth: E



GPT4V Image Caption:

The diagram shows a large circle with a smaller circle inside it. The smaller circle is tangent to the larger circle at a single point on the larger circle's interior. The smaller circle is labeled with the letter K. The space between the two circles forms a crescent shape. The larger circle appears to be a boundary that encloses the smaller circle, and there is no indication of any dimensions, angles, or other geometric properties.

GPT4 Answer:

To solve this problem, let's denote the radius of the smaller circle (circle K) as r . Since circle K is inscribed within a quarter circle of radius 6, we can visualize or draw a diagram where the quarter circle's radius forms a right-angled triangle with the sides being the radius of the quarter circle, the radius of circle K , and the distance from the point where circle K touches the quarter circle's arc to the corner of the quarter circle (the center of the quarter circle). The distance from the center of circle K to the corner of the quarter circle is r , and the distance from the corner of the quarter circle to the point where circle K touches the quarter circle's arc is also r . Therefore, the hypotenuse of the right-angled triangle formed (which is also the radius of the quarter circle) is 6 units. Using the Pythagorean theorem, we can set up the equation: $r + r = 2r$ (this is the base and the height of the right-angled triangle, both equal to the diameter of circle K since it touches both the arc and the two straight sides of the quarter circle). **The hypotenuse (the radius of the quarter circle) is 6 units, so: $(2r)^2 + (2r)^2 = 6^2$**

.....

Therefore, the radius of circle K is $\frac{3\sqrt{2}}{2}$ units.

GPT-4V Answer:

To solve this problem, we need to use geometric relationships within the quarter circle and the inscribed circle K . Consider a quarter circle of radius 6. If we inscribe a circle K inside it, such that K is tangent to the two sides and the arc of the quarter circle, there will be two points of tangency on the sides of the quarter circle which are at a distance equal to the radius of K from the corners of the quarter circle. Let's denote the radius of circle K as r . Since circle K is tangent to the two sides, we know that the distance from the center of K to each side is equal to r , the radius of K . This means there will be a square of side r formed between the center of circle K and the two sides of the quarter circle. The remaining section between the edge of this square and the quarter circle will be a right-angled isosceles triangle with sides of length $6 - r$ (since the quarter circle has a radius of 6). The hypotenuse of this right-angled isosceles triangle will be the radius of the quarter circle, which is 6. Using the Pythagorean theorem, we get: $(6 - r)^2 + (6 - r)^2 = 6^2$

.....

so our answer is: E

Figure 61: An example of metric geometry question of GPT-4 with image caption. GPT-4 fails to find the correct relationship between the variables. GPT-4V correctly solves the problems.