MITSUBISHI ELECTRIC RESEARCH LABORATORIES
https://www.merl.com

Evaluating Large Vision-and-Language Models on Children's Mathematical Olympiads

Cherian, Anoop; Peng, Kuan-Chuan; Lohit, Suhas; Matthiesen, Joanna; Smith, Kevin; Tenenbaum, Joshua B.

TR2024-160 November 22, 2024

Abstract

Recent years have seen a significant progress in the general-purpose problem solving abilities of large vision and language models (LVLMs), such as ChatGPT, Gemini, etc.; some of these breakthroughs even seem to enable AI models to outperform human abilities in varied tasks that demand higher-order cognitive skills. Are the current large AI models indeed capable of generalized problem solving as humans do? A systematic analysis of AI capabilities for joint vision and text reasoning, however, is missing in the current scientific literature. In this paper, we make an effort towards filling this gap, by evaluating state-of-the-art LVLMs on their mathematical and algorithmic reasoning abilities using visuo-linguistic problems from children's Olympiads. Specifically, we consider problems from the Mathematical Kangaroo (MK) Olympiad, which is a popular international competition targeted at children from grades 1-12, that tests children's deeper mathematical abilities using puzzles that are appropriately gauged to their age and skills. Using the puzzles from MK, we created a dataset, dubbed SMART-840, consisting of 840 problems from years 2020-2024. With our dataset, we analyze LVLMs power on mathematical reasoning; their responses on our puzzles offer a direct way to compare against that of children. Our results show that modern LVLMs do demonstrate increasingly powerful reasoning skills in solving problems for higher grades, but lack the foundations to correctly answer problems designed for younger children. Further analysis shows that there is no significant correlation between the reasoning capabilities of AI models and that of young children, and their capabilities appear to be based on a different type of reasoning than the cumulative knowledge that underlies children's mathematics and logic skills.

Advances in Neural Information Processing Systems (NeurIPS) 2024

Mitsubishi Electric Research Laboratories, Inc.
201 Broadway, Cambridge, Massachusetts 02139

Evaluating Large Vision-and-Language Models on Children's Mathematical Olympiads

Anoop Cherian[1] **Kuan-Chuan Peng**[1] **Suhas Lohit**[1] **Joanna Matthiesen**[2]
Kevin Smith[3] **Joshua B. Tenenbaum**[3]
[1]Mitsubishi Electric Research Labs, Cambridge, MA, [2]Math Kangaroo USA NFP
[3]Massachusetts Institute of Technology, Cambribdge, MA

Abstract

Recent years have seen a significant progress in the general-purpose problem solving abilities of large vision and language models (LVLMs), such as ChatGPT, Gemini, etc.; some of these breakthroughs even seem to enable AI models to outperform human abilities in varied tasks that demand higher-order cognitive skills. *Are the current large AI models indeed capable of generalized problem solving as humans do?* A systematic analysis of AI capabilities for joint vision and text reasoning, however, is missing in the current scientific literature. In this paper, we make an effort towards filling this gap, by evaluating state-of-the-art LVLMs on their mathematical and algorithmic reasoning abilities using visuo-linguistic problems from children's Olympiads. Specifically, we consider problems from the Mathematical Kangaroo (MK) Olympiad, which is a popular international competition targeted at children from grades 1-12, that tests children's deeper mathematical abilities using puzzles that are appropriately gauged to their age and skills. Using the puzzles from MK, we created a dataset, dubbed *SMART-840*, consisting of 840 problems from years 2020-2024. With our dataset, we analyze LVLMs power on mathematical reasoning; their responses on our puzzles offer a direct way to compare against that of children. Our results show that modern LVLMs do demonstrate increasingly powerful reasoning skills in solving problems for higher grades, but lack the foundations to correctly answer problems designed for younger children. Further analysis shows that there is no significant correlation between the reasoning capabilities of AI models and that of young children, and their capabilities appear to be based on a different type of reasoning than the cumulative knowledge that underlies children's mathematics and logic skills.

1 Introduction

> "*Mathematics is not about numbers, equations, computations, or algorithms: it is about understanding.*"
>
> *William Paul Thurston*

Recent multimodal artificial intelligence frameworks incorporating large vision and language models (LVLMs), such as GPT-4o, DALL-E, Gemini, *etc*., are seen to demonstrate outstanding reasoning capabilities [43], seemingly flustering our established measures of machine intelligence [11, 19, 23, 24, 35]. These scaled up Transformer models [41] trained on internet-scale datasets using purportedly simplistic training losses such as mask predictions, suddenly appear to have emergent abilities rivaling expert human intellect even on tasks demanding higher-level cognition. Such superior accomplishments naturally raises several questions: Are these models indeed capable of having core knowledge and generalizing it towards deriving innovative methods for problem solving? Are

38th Conference on Neural Information Processing Systems (NeurIPS 2024) Track on Datasets and Benchmarks.

they equipped with the faculties to reason like children or are they exploiting implicit biases in their web-scale training datasets towards generating responses that are seemingly correct? Where do AI models rank in their generalized intellectual capacities against humans?

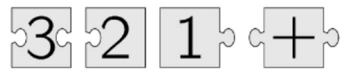
There have been several recent studies that attempt to answer the above questions through novel datasets, tasks, and benchmarks, *e.g.*, SMART-101 [10], MATHVISTA [32], Math-Vision [42], MathOdyssey [17], MathScape [45], *etc*. While, all these datasets and tasks evaluate varied facets of the generative and reasoning abilities of LVLMs, they typically compare the performance of an LVLM against prior state-of-the-art (SOTA) AI models. While, some of these tasks even include human performances, these evaluations use relatively few human subjects, and do not include the diversity, demographics, background, and other subjective attributes that could influence the solution scheme, making the comparison of AI models to human performances to have significant room for speculation. Notably, there appears to be a lack of a systematic study that benchmarks the capabilities of current SOTA AI models against human cognition on the respective tasks at scale.

Contrary to current AI models that are potentially trained on web-scale data at once, humans develop their problem solving abilities over a period of development towards adulthood, and the type and nature of problems that they can typically solve at different stages of their growth vary significantly. For example, a first grader may be able to solve problems related to tracing a given curved path, however a 12-th grader is expected to solve problems related to finding the intersection points of curves. On the one hand, this incremental nature of building knowledge is essential to the development of solid human problem solving [7, 16, 18]. On the other hand, this cumulative knowledge gathering also enforces an order to the way cognitive foundations are established in rational agents, *e.g.*, a 12-th grader is implicitly assumed to have the knowledge to solve problems that a first grader may or may not be able to solve. If we want artificial generalist models that think and reason like intelligent humans, we should expect those models to reliably demonstrate more primitive concepts, in order to build up to reason about more complex problems.

Guided by this insight, we make a first attempt towards systematically comparing the performance of AI models against children's abilities over the period of their growth. Similar to previous and contemporary studies, such as SMART-101 [10], MATHVISTA [32], and Math-Vision [42], we base our approach on the analysis of the problem-solving skills of LVLMs on mathematical and algorithmic reasoning problems selected from Mathematical Olympiads. In contrast to exams typically given in schools, that test the overall grasp of taught subject matter, Olympiads often incorporate problems that explore deeper understanding of concepts, critical thinking abilities, innovative ways of looking at data, and deriving connections across knowledge for solutions. Among many such math Olympiads (such as IMO, AMC, *etc*., that are aimed at higher grade students), one that we use to base our benchmark experiments in this paper is the international Mathematical Kangaroo (MK) Olympiad [1], which is an international competition held since 1998 in 80 countries with nearly 45,000 student participants from the USA in 2024. Further, MK: i) offers competitions for all grades of children, consisting of age appropriate problems, thus allowing comparisons of the performance of AI models against children from varied age groups and skill levels, ii) uses a multiple choice answer format allowing for easy and objective evaluations that are directly comparable to children's responses, iii) do not demand the participants to have advanced mathematical background (*e.g.*, calculus, analysis, *etc*.) for deriving the solutions, and iv) offers both text only and image-text problems, thus offering multimodal LVLM evaluation. These aspects make MK Olympiads suitable for our study, thus allowing for a clear gauging of AI's reasoning abilities against the extent of developing human intellect.

To conduct our benchmarking experiments, we created an MK Olympiad dataset, dubbed SMART-840, by collating all the math competition problems from years 2000-2024 for all grades from 1-12th. Our dataset offers a range of problem difficulties (some targeted at 1-2nd graders, some at 11-12th graders), and can further subdivide them by how children actually perform on those problems. Thus, for human-like mathematical reasoning we should expect: (a) better performance on earlier grade problems, and (b) a correlation with problem difficulty within grades. An example problem is provided in Figure 1. We also collect various performance statistics of children on all the exams from MK that are useful in our analysis, including the children's answer responses and time taken in deriving the solutions, among many others. Next, we shortlist several SOTA LVLMs, including closed-source models such as GPT-4o [37], Gemini [13], and Claude-3 [3], and open-source models including LLaVa-Next [25], and XGEN-MM-Phi3 [39], among others, that are popular on their generalist abilities. Our analysis reveals several interesting aspects of the AI models: namely i)

Q: When you put the 4 puzzle pieces together correctly, they form a rectangle with a calculation on it. What is the result of this calculation?
A: 6 B: 15 C: 18 D: 24 E: 33



GPT-4o: To solve this puzzle, we need to arrange the 4 pieces (3, 2, 1, and +) to form a correct mathematical equation. Let's try arranging them and checking the validity of each possible equation: 1. 3 + 2 = 1 2. 3 + 1 = 2 3. 2 + 1 = 3 Let's check each possibility: 1. "3 + 2 = 1" is incorrect because 3 + 2 equals 5, not 1. 2. "3 + 1 = 2" is incorrect because 3 + 1 equals 4, not 2. 3. "2 + 1 = 3" is correct because 2 + 1 equals 3. Thus, we have: 2 + 1 = 3 So the correct equation formed by the puzzle pieces is: "2 + 1 = 3" Hence, the result of this calculation is: C: 18

Gemini-Pro: 1. **Identify the Pieces:** The image shows four puzzle pieces, each with a part of a mathematical equation. 2. **Visualize the Rectangle:** Imagine fitting these pieces together to form a rectangle. The edges will help us connect the parts of the equation. 3. **Mentally Arrange:** Piece by piece, try to mentally arrange them. Notice that some pieces have parts of numbers, and others have operators (+, -). 4. **Form the Equation:** By carefully arranging the pieces, you can form the equation: ``` 18 + 6 ``` 5. **Calculate:** The result of the calculation 18 + 6 is 24. **D: 24**

Figure 1: A 3rd grader puzzle from our SMART-840 dataset and the LVLM responses (both incorrect).

their performance on the SMART-840 dataset appears to be significantly weaker than the average performance of children, ii) the performance gap is larger (nearly 30-40%) on the tests for younger children and improves to 10-20% for higher-schoolers, iii) there appears to be a lack of any significant correlation between the difficulty of a puzzle to children against that to an AI model, thus making it difficult to judge if a model could answer a given test problem, iv) there appears to be significant variability in the responses of current AI models when repeating or changing prompts, and v) recent LVLMs seem to demonstrate increasingly superior skills in solving *text-only* problems, outperforming children. We believe our analysis brings new insights into various aspects for testing AI models against human cognition, as well as sheds light into the issues around reliability of current LVLMs for solving math problems.

Before going further, we emphasize below the primary contributions of this paper.

1. We provide a first of its kind benchmarking of the performance of large vision-and-language models against the mathematical and algorithmic reasoning abilities of young children using data from professional math Olympiads.
2. We gauge the reasoning abilities of AI against the cumulative knowledge building progression of children over their growth.
3. Our experiments compare SOTA AI models on both text-only and vision-and-text math problems, analyzing the performances across multiple dimensions.

2 Related Works

General LVLM benchmarks: Several benchmarks now exist that test different capabilities of LVLMs. These include MMBench [26] which contains thousands of questions in a VQA format about both perception (1844 questions) and reasoning (1104) where the models select an answer from a given set of options. It also uses a "circular evaluation" strategy to ensure that the models are robust to the ordering of answer options. Although logical and relational reasoning are part of the dataset, this benchmark does not test particularly for different types of mathematical reasoning capabilities. MMMU [44] is another popular benchmark that contains about 12.5K multimodal questions covering six different disciplines of study, but only at the college level and tests expert-level knowledge of LVLMs. In contrast, we are interested in understanding abilities of LVLMs that children demonstrate. A few benchmarks have been designed to test specific capabilities of LVLMs like ScienceQA [29] for scientific understanding and reasoning, VisIT-Bench [6] for instruction following, Bongard Problems [8, 20, 21], Raven's Progressive Matrices [5], Abstraction and Reasoning Corpus [11] for abstract visual reasoning, OCRBench [27], and TextVQA [40] for text recognition, *etc*.

Benchmarks for mathematical reasoning: MathVista [32] is a recent benchmark for mathematical reasoning based on puzzles that involve images, while also measuring performance of LVLMs for different types of mathematical reasoning (logical, arithmetic, geometric, *etc.*) and different types of context images (natural images, line plots, scientific figures, *etc.*). GSM-8k [12] is a similar dataset containing about 8.5K math word problems, but only has text inputs and outputs, no images are involved. In GSM-Symbolic [36], the fragility in logical reasoning of LLMs is explored. The main difference of the proposed SMART-840 benchmark against these prior works is that the these datasets neither separate puzzles based on hardness (*e.g.*, ease of solving by children at different school levels), which our proposed benchmark explicitly addresses, nor are they supported by human performances at scale. TabMWP [30] is a benchmark with about 38k grade-school problems but is limited to just tabular math word problems. SMART-101 [10] is the most closely related benchmark to ours, that provides programmatically generated 2000 variations for each of 101 puzzles from just 1st and 2nd grade Math Kangaroo puzzles. These variations can be used to train larger models than using just a small number of puzzles. In contrast to this dataset, SMART-840 contains 840 puzzles from all grades 1-12 and is designed to benchmark the zero-shot mathematical reasoning abilities of AI models on a wide range of problem solving skills, along with important information on performance statistics from test-takers. We note that we are not the first to use MK performance statistics for research, in fact in [2, 4, 33, 34] MK exams were used to study the development of mathematical competencies in children, however to the best of our knowledge these tests have not been used to benchmark AI models.

3 Benchmarking Approach and Experiments

We first elucidate our data collection process, followed by the details of the LVLMs that we select to benchmark in this study. The subsequent subsections evaluate the performances of LVLMs on various aspects of the Olympiads deriving correlations with the performances of children.

3.1 Mathematical Kangaroo Olympiad

As alluded to above, most of the Mathematical Olympiads (such as International Math Olympiad, AMC-8, AHSME, *etc.*) are targeted at middle or high-school students, while Math Kangaroo is one Olympiad that conducts competitions for K-12 grades, making it a compelling source for this study. Started in France in 1991, the competition has been organized in the USA every year since 1998 and currently takes place in over 100 countries. Typically, there is a single exam for grades $\{n, n+1\}$, for $n \in \{1, 3, 5, 7, 9, 11\}$, thus there are a total of six exams in a year and children of both grades n and $n+1$ participate in the same exam.

Each exam consists of either 24 questions (for grades 1–4) or 30 questions for all higher grades, and is in a multiple choice format with 5 candidate answers, of which only one option is the correct answer. The questions can be purely text-based or can contain both text and an image, interpretation of both jointly is then usually important for solving the problem. Each question is attributed weights in $\{3, 4, 5\}$, where the lower points are given to problems that are typically deemed "easier" for that grade (*e.g.*, single step reasoning problems for grade 1) while higher points are attributed to problems that need multi-level reasoning, enumeration of solution possibilities, *etc.* that typically involve deeper (but age appropriate) problem solving skills. The participant is given 75 minutes to complete an exam and the performance is computed as the weighted sum of correct responses.

3.2 Data Collection

For this study, we envisage a data collection methodology that is fair, balanced, and offers an unbiased benchmarking of AI models against children's performance on MK tests. To this end, we decided to use all the questions from MK exams without any omissions so that there is no selection bias in our evaluations. As the statistical data we desire on the performance of children is unavailable for MK competitions prior to year 2020, in this study we consider only MK competitions from 2020-2024, that amounts to 840 problems in our dataset, dubbed SMART-840, and consisting of 240 questions all together from grades 1–4 and 600 questions from grades 5–12, evenly split between pairs of grades as described above. Figures 2(a), 2(b), and 2(c) show the distribution of the number of children who participated across years from all grades, which adds to nearly 30K students per year. The participant number is highest for grades 1–8 and then drops to less than 1000 for grades 9–12 (Figure 2(b)) perhaps because higher-grade children have other Olympiad options, *e.g.*, AMC, IMO,

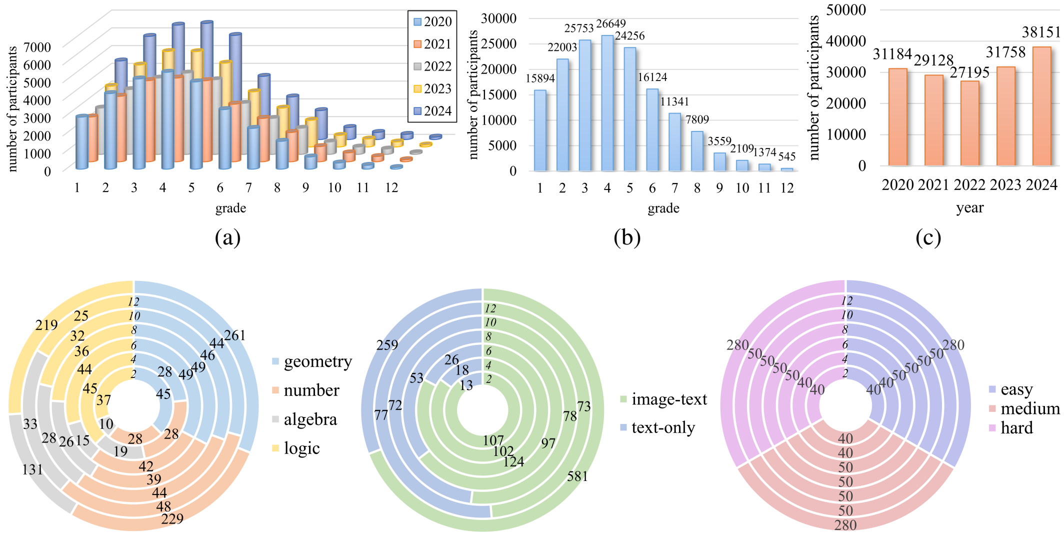


Figure 2: Figure 2(a) plots the distributions of children participating in MK Olympiads per year over 2020–2024 for grades 1–12. Figure 2(b) plots the total number of participants per grade during 2020–2024. Figure 2(c) plots the total number of participants each year over all grades (1-12). Figure 2(d) shows the number of puzzles and its portion for each category. Figure 2(e) shows the statistics of image-text and text-only puzzles. Figure 2(f) shows the statistics of puzzle difficulty (defined by their attributed weights).

etc. Nevertheless, we see that the number of participants put together from the last five years still produce a substantially large sample set for our analysis.

For creating the SMART-840 dataset, we downloaded publicly available[1] question papers (which are image embedded PDF documents), followed by running optical character recognition software for extracting the text of the puzzles, and manually cropping the associated image parts. Each such extracted puzzle in the dataset was manually inspected for errors in its text and puzzle images. MK also provides a segregation of each puzzle into one of the four categories, namely (i) geometry, (ii) logic, (iii) algebra, and (iv) numbers. In Figure 2(d), we present the overall statistics of problem distribution in the SMART-840 dataset. We see that *geometric* puzzles capture nearly 31% of all the puzzles in our set, while the split is about equal between *logic* (26%) and *numbers* (27%), and *algebra* based problems are about 15.5%. In Figure 2(e), we plot the distribution of the number of problems that need both text and image reasoning (∼69%) against those that only have text questions. Figures 2(d) and 2(e) also show the split across grades. We see that for higher grades (>8), the number of text-only problems are higher: about 52% in grades 11-12 against <20% in grades 1-4.

3.3 Selected Large Vision-and-Language Models

We compare the performance of seven popular and SOTA LVLMs on the SMART-840 dataset. Specifically, we consider i) GPT-4o [37], ii) Gemini-Pro [13], and iii) Claude-3 [3], that are popular for their abilities in solving challenging math and visual reasoning problems. Thus, we believe it is a useful exercise to understand how they perform on children's grade problems. Alongside these SOTA LVLMs, we also consider other AI models that are popular, such as GPT-4v which is the first vision-and-language version of the GPT series, ii) Gemini-Flash that is well-known for its faster response time, and to recent open-source LVLMs such as XGen-MM-Phi3-Instruct-v1 [39], LlaVa-Next [25], InternVL-Chat-V1-2 [9], and InternLM-XComposer-2.5 [14].

[1]Note that each test paper involves a small cost for download.

3.4 Grade-wise Performance Comparisons

In this experiment, we compare the performance of the LVLMs listed above against the performance of children on our SMART-840 dataset. For the human performance, we report the percentage of average correct response rate, which we denote as *accuracy* going forward, and is computed by: i) finding the ratio of the total number of correct children's responses on a problem to the total number of attempts, and (ii) averaging this ratio across all problems in the grade set. For the LVLMs, we use the API interface to query the model using a suitable hand-crafted prompt. Specifically, we found the following prompt to work well for all closed-source LVLMs: `"Solve this question with explanation of the intermediate steps. Your response should end with one of the selected answer options from A1, B2, C3, D4, or E5."` which is accompanied by the text for the problem question and the image data.[2] For AI models, we report their accuracy as the (percentage) of problems correctly answered to the total number of problems in the set.

In Table 1, we present results comparing the performances of LVLMs against children on the entire SMART-840 dataset. First, we report a *random* baseline to benchmark all our results, which is computed by randomly sampling a response from a probability distribution over all the human responses across the answer options for a problem. As is clear, all the answer options in the problems are equally likely, and thus the random performance is close to one-fifth. Next, for each LVLM, we queried (at least 2 times) each of the problems in SMART-840 using the prompt described above. Note that for LVLM evaluation we also consider two additional possibilities, namely: (i) if a response is not in the expected format as demanded in the prompt, and if we are unable to automatically extract a valid response, we consider the response to be invalid in general (except in experiments when we manually validate the responses), and (ii) in many cases, an LVLM decides not to solve a problem (*e.g.*, it mistakes the provided puzzle image to contain security issues), in which case as well, we declare that problem as unsolved by the respective model. We manually inspected all the output responses of GPT-4o (reported as GPT-4o (M) in the table) to ensure the our prompt is suitable, and the model produces responses that are reasonable, grounded in the problem specification (and are not due to issues such as network failures, response parsing failures, etc.) and its solution attempt is reasonable (but not necessarily correct). All problems where the solution was unreasonable (even if the *selected option* is correct), we manually marked them as a *failed response*.

We see from Table 1 that GPT-4o demonstrates an accuracy of 42.5% on average across all the grades, followed by Claude-3-Opus at 38% and Gemini-Pro at nearly 32%. The more recent Claude-3-Sonnet model in fact outperforms the performances of all the earlier models, with an average accuracy of 49.7%, while outperforming GPT-4o in grades 1–6 and 9–10. There are several intriguing aspects in the performance of LVLMs that we can witness in Table 1.

i) **Performance gap:** The performance of AI models are below that of children across the grades and interestingly this gap is consistent in all the models we experimented. Specifically, the best accuracy of LVLMs are in the range of 40-50% while the children's average performance is consistently near 60% or above. Note that we report children's performances for each grade separately, where kids of a pair of grades take the same exam. Unsurprisingly, we find that children of the higher grades perform significantly better than those of lower grades (although this gap reduces as problem solving abilities mature towards higher grades) suggesting a cumulative set of core problem solving skills that children build over their growth period.

ii) **Performance trend:** In Table 1, we see yet another consistent trend of LVLMs, *i.e.*, being better at solving problems of higher grades (8-12) than at lower grades (*e.g.*GPT-4o, Gemini-Pro, etc.) or similar performances in solving both higher- and lower-grader problems (*e.g.*Claude-3 models), which is surprising given the complexity of solutions increases with grades. This trend was also seen in [10] where the authors compared the performance of LLMs on second grader problems. We see that while GPT-4o shows this increasing trend with an accuracy of 40% at grades 1-2 towards nearly 50% for grades 11-12, the trend is more striking for other LVLMs such as Gemini-Pro that varies by about 25% (for grade 1-2) to 40% for grades 11-12. We find that Claude-3 models produce a reasonably consistent performance around 40% albeit having a different trend: dip in the performance for middle grades than lower or higher grades.

[2]The format of A1, B2, *etc*. allow us to uniquely parse the LVLM response to automatically validate it.

Model \ Grade	1	2	3	4	5	6	7	8	9	10	11	12	Mean
Human	58.8	67.6	62.3	70.1	59.1	65.4	59.7	64.3	64.2	69.3	64.9	65.6	**64.2**
Random	20.1		20.2		20.1		20.2		20.3		20.1		20.1
GPT-4o	41.6 (7.1)		38.6 (1.7)		35.1 (0.8)		47.1 (0.8)		41.3 (2.0)		50 (4.0)		42.4
GPT-4o (M)	42.5		36.7		36.0		46.7		43.3		50.0		42.5
GPT-4v	39.2 (0.6)		38.3 (0.6)		29.3 (3.3)		35.3 (1.9)		38.7 (1.9)		43.3 (3.7)		37.4
Gemini-Pro	25.8 (3.5)		27.5 (0.6)		25.3 (3.3)		30.7 (1.8)		39.3 (3.7)		41.3 (2.8)		31.7
Gemini-Flash	19.2 (0.6)		29.2 (10.4)		22.0 (8.4)		30.7 (9.7)		38.7 (13.7)		36.7 (4.3)		29.4
Claude-3 Opus	38.3 (5.3)		33.3 (5.8)		31.3 (6.6)		40.7 (10.4)		42.0 (5.6)		44.0 (2.8)		38.3
Claude-3 Sonnet	51.6 (0)		47.9 (2.9)		38.6 (0.9)		44.9 (3.3)		46.7 (0.0)		49.7 (4.1)		49.7
XGEN-MM-Phi3-v1 (5B)	7.5		9.1		5.3		8.0		10.0		8.0		8.0
InternVL-Chat-V1.2 (40B)	16.7		25		17.3		14.6		15.3		16.7		17.6
InternLM-XComposer2 (7B)	22.5		14.2		18.6		24.2		18.1		16.9		19.1
LlaVa-NEXT (34B)	15.0		9.0		20.1		14.6		18.7		16.0		15.6

Table 1: Accuracy (%) of correct responses of children in the respective grades against the accuracy of LVLMs when the agent is asked to provide explanation of their responses. GPT-4o (M) denotes the performance of GPT-4o after manual validation of GPT-4o responses. We report the standard deviation in brackets. The last block shows the performances of recent open-source LVLMs, their size (in brackets).

We find from Table 1 that open source LVLMs such as XGEN-MM, InternLM-XComposer2-VL and InternVL-Chat-V1.2 models perform poorly in comparison. Note that these models show nearly 60% performances on the Math-Vista leader board [31]. We also compare against the recently released LlaVa-NEXT (34B) model which received 47% on the Math-Vista leader board. We find that on SMART-840, these models either selected incorrect answer options or many-a-times did not follow the instruction, thereby producing invalid outputs. As can be noted from the table, while the performance of recent open-source LVLMs are still significantly below that of closed-source LVLMs. We further note that there is nearly a 40% gap in the performances of these models between Math-Vista and SMART-840, suggesting that the mathematical reasoning skills needed to solve SMART-840 are substantially different from existing math datasets publicly available.

Further, we find that the performance of GPT-4v is inferior to that of GPT-4o, which is expected given the latter being a more advanced version of GPT-4v. Further, the accuracy of faster LVLMs such as Gemini-Flash is below that of its advanced counterpart. Thus, in our subsequent study, we only consider the best performing LVLMs, namely GPT-4o, Gemini-Pro, and Claude-3 Sonnet. The recent OpenAI o1-preview model is text-only at this time, thus making it incomparable to other LVLMs in this study; however we report its performance on a text-only subset of our dataset in Table 3.

iii) **Variance:** We ran the LVLMs on each problem at least twice[3] and the variance was computed on the differences in the accuracies; (e.g., we ran GPT-4o models about 5 times, while Gemini was run only twice). From Table 1, we find that there is substantial variance in the performance of SOTA LVLMs. For example, the standard deviation for GPT-4o is nearly 7% in solving 1-2 grade problems, while there is a reducing trend in the magnitude of this deviation for higher grades, the reliability in the responses are still questionable. The standard deviation is worse for Claude-3-Opus, where it is nearly 5% across grades, even reaching 10% for grades 7-8. Interestingly, for grades 11-12, the deviation appears more stable at nearly 3-4%, while the performance is also the best.

4 Analysis of Results

In this section, we take a deeper look at our results in Table 1 to gain insights into how the AI model responses correlate with those of children. Even though a model is expected to perform as well as an adult, given that their performances are below that of children, it is imperative to ask if they atleast behave like high-performing children in their responses? Specifically, we seek to answer the question: *Are problems that are hard for children also hard for AI?* To answer this, we conduct different types of correlation analysis, presented below.

Difficulty Index of a problem [15, 38] is the ratio of the number of correct responses for a test problem to the total number of solution attempts. This index provides a score between 0 and 1 for

[3]The number of runs was constrained by the cost involved.

each problem, where 0 implies none of the children were able to solve it (hard problem). In Table 2 (Diff-I), we report the Pearson's correlations coefficient between the difficulty index and the responses by LVLMs. We see that there is in general only weak correlation between model and human accuracy, and this correlation mostly occurs at the higher grade levels. This suggests that all LVLMs in general find a different set of problems to be difficult than children do.

Discriminative Index [15, 22, 28] measures the use of *knowledge* by a test taker. To compute this score, we split the student population into two groups, the *good learners* that correspond to the top-20% of participants who score the highest, and *bad learners* that constitute the bottom-20%. Next, we compute the difficulty index for each of these sets separately, and define discriminative index as the difference between the two difficulty indices. Thus, the value of discriminative index is in $[-1, 1]$, where 1 corresponds to a test problem where all the good learners produced correct responses while all the bad learners made a mistake – *i.e.*, a problem that can separate the good learns from the bad. To understand if an AI model is a good learner or bad learner, we propose to compute the Pearson's correlation between the discriminative index of children's performances against that of the models. The result of this analysis is provided in Table 2 (Disc.-I). Surprisingly, we find a negative trend across all grades, suggesting that an AI model finds it easier to solve problems that are less discriminative, and whose answer options are plausibly discernible without substantial reasoning.

Weight Correlation measures the interaction between the hardness of a puzzle as attributed by MK (through its weight) against the response. Notably, we convert this weight in $\{3, 4, 5\}$ to the corresponding difficulty score of $\{1.0, 0.66, 0.33\}$ for each problem, and compute the Pearsons correlation to the AI responses, which are 1 if the answer option selected to the problem is correct and 0 otherwise. We find a slightly stronger positive correlation on this experiment in Table 2 (Weight-C.), suggesting the AI is able to solve problems that (the adult creator) thought are of the easier kind.

Entropy Correlation measures the correlation between the entropy of the distribution of children's selected answer choices against AI responses. As entropy is higher for problems which are hard or their options confusing for children, a positive correlation would suggest AI is similarly confounded. However, the trend in Table 2 shows the reverse, with slightly stronger negative correlations, suggesting that AI models are apparently not much confused on problems children find indecisive.

Time-taken Correlation analyzes the dependency between how much time children (on average) used in answering problems – and thus potentially capturing the problem hardness – to whether AI models also find those problems challenging. To this end, we aggregated the duration children spent on each problem, followed by separating the duration into two sets on their median. We marked all problems above the median as hard and the rest as easy. Next, we computed the Pearsons correlation between the responses of AI models against this hardness. Our results in Table 2 (Time-C.) shows again a weak negative correlation trend, suggesting that the model finds it easier to solve problems that take longer for children – a surprising result!

Category-Level Performances: As alluded to above, SMART-840 dataset consists of problems in four different categories (as per their creators), each involving entirely different skill sets and knowledge background for their solutions. For the performances reported in Table 1, in Figure 3, we present the results of humans and LVLMs on the four problem categories, namely (i) geometry, ii) numbers, iii) algebra, and iv) logic. While children perform consistently well on all these categories, we find that AI models falter significantly in geometry and logic, with their best performances at about half of that of humans while they perform reasonably well on numbers and algebra. We further analyze the performance of LVLMs on problems involving both image and text (e.g., geometry problems) and text-only puzzles. The results show that it is indeed the image-text problems that the models struggle with and we see a strong similarity between performances on geometry and logic problems with image-text problems. Interestingly, we also find that on text-only puzzles (which are about 30%) in our dataset, GPT-4o-Expl.[4] shows better performances than the average human performance, while other LVLMs (with suffix "-expl.") are also performing reasonably well.

To analyze this further, in Table 3, we report the performance of LVLMs on image-text puzzles and text-only puzzles separately for each grade pair. We can make several observations, namely: i) human performance is consistently between 60-70%, irrespective of text-only or image-text problems, ii) the performance of LVLMs are significantly higher (nearly double) on text-only problems than on image-text problems, however the trend remains the same, *i.e.*, LVLMs appear to find lower-grader

[4]Which uses a prompt to explain its reasoning.

	Model \ Grade	1	2	3	4	5	6	7	8	9	10	11	12
Diff-I	GPT-4o	0.14	0.16	0.15	0.17	-0.09	-0.05	0.12	0.13	0.22	0.22	0.20	0.26
	Gemini-P	0.23	0.27	-0.05	-0.06	0.01	-0.01	0.05	0.06	0.21	0.19	0.20	0.16
	Claude-3	0.11	0.13	0.09	0.11	0.08	0.06	0.14	0.15	0.16	0.16	0.25	0.18
Disc-I	GPT-4o	-0.07	-0.15	0.07	-0.01	0.07	-0.01	-0.09	-0.08	-0.14	-0.18	-0.11	-0.13
	Gemini-P	-0.05	-0.25	-0.04	-0.05	-0.01	-0.01	0.01	0.03	-0.18	-0.18	-0.15	-0.13
	Claude-3	-0.02	-0.14	0.17	0.06	-0.04	-0.09	-0.07	-0.09	-0.16	-0.11	-0.09	-0.16
Weight-C.	GPT-4o	-0.04	-0.04	-0.02	-0.02	-0.00	-0.00	0.08	0.08	0.13	0.13	0.15	0.15
	Gemini-P	0.05	0.05	-0.07	-0.07	0.00	0.00	0.02	0.02	0.27	0.27	0.30	0.30
	Claude-3	-0.10	-0.10	-0.02	-0.02	0.00	0.00	0.15	0.15	0.18	0.18	0.30	0.30
Entropy-C.	GPT-4o	-0.18	-0.18	-0.15	-0.15	0.10	0.10	-0.14	-0.14	-0.23	-0.23	-0.24	-0.24
	Gemini-P	-0.26	-0.26	0.03	0.03	-0.01	-0.01	-0.08	-0.08	-0.23	-0.23	-0.19	-0.19
	Claude-3	-0.12	-0.12	-0.06	-0.06	-0.02	-0.02	-0.15	-0.15	-0.18	-0.18	-0.24	-0.24
Time-C.	GPT-4o	-0.08	-0.12	-0.14	-0.10	0.03	-0.03	0.08	0.03	-0.09	-0.07	-0.17	-0.09
	Gemini-P	-0.06	-0.17	-0.06	-0.06	-0.03	-0.06	0.03	0.03	-0.20	-0.12	-0.27	-0.19
	Claude-3	0.14	0.10	-0.07	-0.07	-0.04	-0.01	-0.01	-0.07	-0.09	-0.07	-0.16	-0.13

Table 2: Pearson's correlation coefficient (ρ) between the difficulty index (top) and discriminative index (bottom) of the SMART-840 problems against the responses generated by LVLMs. The green/red cell color indicates positive/negative ρ, where darker cells represent larger absolute values.

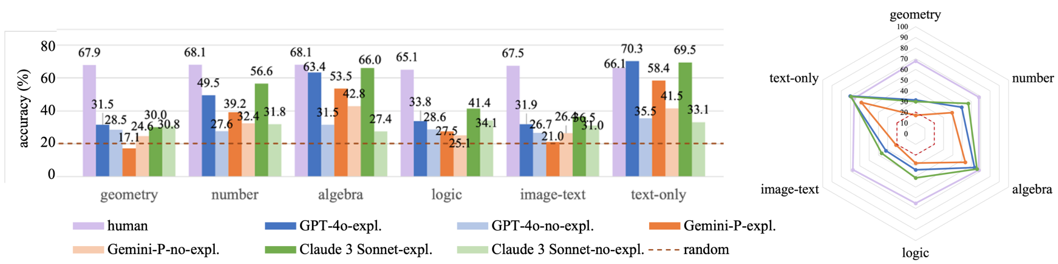


Figure 3: Comparison of the average accuracy (%) of humans and LVLMs on each category of the Olympiad problems with the corresponding radar plot.

problems equally difficult as that for higher-grader problems, iii) some of the very recent powerful LVLMs, such as OpenAI's o1-preview and Claude-3 models appear to excel in the performances on lower grade text-only problems, especially grades 1-4. Specifically, the performance of o1-preview has substantially improved from that of GPT-4o on text-only problems, however their trend as noted above appears similar, that there are dips in their performances for lower-grader problems compared to higher grader ones.

Importance of Reasoning with Explanation: In this experiment, we changed the LVLM prompt to: `"Solve this question. You should provide a response without any explanation. Your response should end with one of the selected answer options from A1, B2, C3, D4, or E5."`. In Figure 3, we show this result for all the LVLMs (suffixed "-no-expl.") on the six categories. We see a trend of a dip in performance among all the models, specifically GPT-4o drops from 49.5% to 17.6% on the highly-performing 'number' category, and from 63.4% to 31.5% on algebra. The drop is also substantial on text-only problems. The trend is similar on other LVLMs (e.g., Claude-3), however slightly lower in Gemini-Pro.

5 Discussion and Conclusions

This paper tackles the important problem of understanding the reasoning abilities of LVLMs. Our analysis using the proposed SMART-840 dataset reveals several intriguing results: i) there is a lack of any significant correlation between the perceived complexity in solving puzzles by children and by AI models; instead there are surprising negative correlations, ii) there is a significant trend among LVLMs in performing low on younger grade problems and progressively get better at higher grades, which is counter intuitive. While, one may attribute this observation to the availability of better training data or increased number of text puzzles, it is still unsettling that AI models struggle to perform

Model \ Grade	1 & 2	3 & 4	5 & 6	7 & 8	9 & 10	11 & 12	1 & 2	3 & 4	5 & 6	7 & 8	9 & 10	11 & 12
	Image-text problems only						Text-only problems					
% dataset	89.17	85.00	82.67	64.67	52.0	48.67	10.83	15.00	17.33	35.33	48.0	51.33
	Accuracy (%) on image-text problems						Accuracy (%) on text-only problems					
Human	67.98	69.91	65.59	64.76	71.22	65.24	64.51	70.89	64.7	63.51	67.23	65.94
GPT-4o	36.45	30.39	31.45	38.14	32.05	32.88	92.31	72.22	57.69	62.26	55.56	66.23
Gemini-Pro	20.56	19.61	21.77	21.65	21.79	20.55	69.23	72.22	42.31	47.17	58.33	61.04
Claude-Sonnet	45.79	40.20	35.48	35.05	32.05	30.14	100.00	77.78	57.69	56.60	62.50	62.34
OpenAI o1-preview	-	-	-	-	-	-	100.00	100.00	84.6	92.4	84.7	90.9

Table 3: Top row, left and right blocks show the % of problems in SMART-840 that belong to image-text or text-only categories per grade. Lower block show the separated performances of LVLMs on image-text and text-only problems against humans.

Model \ Grade	1	2	3	4	5	6	7	8	9	10	11	12
GPT-4o	49/57.4	49/30.4	61/26.8	60/13.2	70/44.3	70/29.0	57/46.4	58/39.7	70/23.8	70/16.8	66/21.1	50/29.2
Gemini-P	78/7.3	78/2.3	69/14.9	68/6.6	75/35.6	75/21.9	80/14.4	81/10.0	79/11.8	77/7.7	51/43.7	34/56.9
Claude-3-O	69/20.6	69/6.7	81/2.5	80/1.1	86/18.3	86/9.3	65/34.3	66/28.9	85/7.1	82/4.0	53/39.8	36/54.6
Claude-3-S	41/71.1	41/45.7	48/49.2	47/30.2	87/16.8	87/8.2	78/16.5	79/12.9	38/72.7	38/59.0	56/34.8	40/50.0

Table 4: National Rank (↓) / percentile (↑) ranking of LVLMs against children's performance on MK 2024 Olympiad based on the test scores computed from the model response.

even on puzzles involving simple geometry and logic, that accentuates the lack of understanding between language and multimodal content. Further, while there is a substantial gap between the best of LVLMs and the worst, or random baselines, 40% for GPT4o vs 20% for random or 25% for Gemini-Pro, this is only a 20% difference; in contrast the gap between even the best LVLMs and human adult level performance in reasoning is much greater. We ought to point this out.

Our results suggest some ways that LVLMs, even the most advanced ones, may not really be reasoning in the ways that humans do. For humans, reasoning is an ability to think that goes beyond just similarity to training examples. But here we are seeing signs that similarity to the large mass of training examples appears to be what is driving performance across all levels of these problems. Of course, we do not fully know what is in the training corpus SOTA LVLMs used. But it may include many Olympiad problems than there are math kangaroo grade 1-2 style problems. Yet for people, and not for frontier LVLMs, the MK grade 1-2 problems are far easier than the Olympiad style grade 11-12 problems. This suggests both that human reasoning is based on a different set of core competencies, which the early grade problems test, and which a pure machine learning approach to training reasoning is not really picking up on.

Before concluding, we present in Table 4, the national rank and percentile of the three SOTA models on the scores they received for 2024 MK Olympiad (when compared against children). We see that AI models are substantially below children in ranking, with GPT-4o best on grade 7-8 in rank in 50's and Gemini-Pro at 34 for grade 12, and the recent Claude-3 Sonnet model outperforming its previous Claude-3 Opus model as well as other models across other grades, however yet their best national rank is more than 30. These scores are based on the percentiles received from MK. The table shows that there is a large gap to fill for LVLMs against children's problem solving skills.

Limitations: In comparison to recent and popular datasets used in the recent AI mathematical reasoning benchmarks, our SMART-840 is smaller in size. However. we ought to emphasize that this small size is by necessity. Our intended use is to see how well general purpose language models (including those presumably trained with a fair amount of mathematical and visual reasoning data in their training sets) come towards capturing this spectrum of mathematical problem solving and reasoning. Thus, we do not intend to compare models trained/fine-tuned on this dataset, as that would invalidate any comparisons to human performances, where humans are assumed not to have seen the problems when taking the tests. Our goal is to bring out the disparity in machine cognition with respect to humans via studying the zero-shot performance of LVLMs with respect to human cognition on our task. Given the goal of this study is not to train LVLMs to excel on this task, instead is to evaluate the zero-shot performance, and given that our results clearly demonstrate a discrepancy between human and LVLM performance, we believe our dataset does help bring out the failure modes of state-of-the-art LVLMs, and point to directions that would need improvements.

Acknowledgements: Authors thank Preethi Ann Cyril for helping with data curating and Math Kangaroo USA for providing the performance data.

References

[1] Math Kangaroo USA, NFP Inc. https://mathkangaroo.org/mks/, 2012–2024. 2

[2] Lukas Andritsch, Evita Hauke, and Jakob Kelz. How to create and solve: Analysis of items from the mathematical kangaroo from two perspectives. *Engaging young students in mathematics through competitions—World perspectives and practices*, 2:117–136, 2020. 4

[3] Anthropic. Claude-3 Opus. `https://www.anthropic.com/claude-opus`, 2023. Accessed: 2024-05-29. 2, 5

[4] Mark Applebaum and Roza Leikin. Girls' performance in the kangaroo contest. *Including the Highly Gifted and Creative Students–Current Ideas and Future Directions*, pp. 87, 2019. 4

[5] Yaniv Benny, Niv Pekar, and Lior Wolf. Scale-localized abstract reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12557–12565, 2021. 3

[6] Yonatan Bitton, Hritik Bansal, Jack Hessel, Rulin Shao, Wanrong Zhu, Anas Awadalla, Josh Gardner, Rohan Taori, and Ludwig Schimdt. Visit-bench: A benchmark for vision-language instruction following inspired by real-world use. *arXiv preprint arXiv:2308.06595*, 2023. 3

[7] Elizabeth L Bjork, Robert A Bjork, et al. Making things hard on yourself, but in a good way: Creating desirable difficulties to enhance learning. *Psychology and the real world: Essays illustrating fundamental contributions to society*, 2(59-68), 2011. 2

[8] MM Bongard. The recognition problem. tech. rep. 1968. 3

[9] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. InternVL: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24185–24198, 2024. 5

[10] Anoop Cherian, Kuan-Chuan Peng, Suhas Lohit, Kevin A. Smith, and Joshua B. Tenenbaum. Are deep neural networks smarter than second graders? In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2023. 2, 4, 6

[11] François Chollet. On the measure of intelligence. *arXiv preprint arXiv:1911.01547*, 2019. 1, 3

[12] Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021. 4

[13] Google DeepMind. Gemini pro. `https://www.deepmind.com/gemini-pro`, 2023. Accessed: 2024-05-29. 2, 5

[14] Xiaoyi Dong, Pan Zhang, Yuhang Zang, Yuhang Cao, Bin Wang, Linke Ouyang, Xilin Wei, Songyang Zhang, Haodong Duan, Maosong Cao, et al. Internlm-xcomposer2: Mastering free-form text-image composition and comprehension in vision-language large model. *arXiv preprint arXiv:2401.16420*, 2024. 5

[15] Robert L. Ebel and David A. Frisbie. *Essentials of Educational Measurement*. Prentice Hall, 1991. 7, 8

[16] K Anders Ericsson, Ralf T Krampe, and Clemens Tesch-Römer. The role of deliberate practice in the acquisition of expert performance. *Psychological review*, 100(3):363, 1993. 2

[17] Meng Fang, Xiangpeng Wan, Fei Lu, Fei Xing, and Kai Zou. Mathodyssey: Benchmarking mathematical problem-solving skills in large language models using odyssey math data. *arXiv preprint arXiv:2406.18321*, 2024. 2

[18] John Hattie and Helen Timperley. The power of feedback. *Review of educational research*, 77(1):81–112, 2007. 2

[19] José Hernández-Orallo. *The measure of all minds: evaluating natural and artificial intelligence*. Cambridge University Press, 2017. 1

[20] Huaizu Jiang, Xiaojian Ma, Weili Nie, Zhiding Yu, Yuke Zhu, and Anima Anandkumar. Bongard-HOI: Benchmarking few-shot visual reasoning for human-object interactions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. 3

[21] Huaizu Jiang, Xiaojian Ma, Weili Nie, Zhiding Yu, Yuke Zhu, and Anima Anandkumar. Bongard-hoi: Benchmarking few-shot visual reasoning for human-object interactions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 19056–19065, 2022. 3

[22] Truman Lee Kelley. The selection of upper and lower groups for the validation of test items. *Journal of Educational Psychology*, 30(1):17, 1939. 8

[23] Brenden M Lake, Tomer D Ullman, Joshua B Tenenbaum, and Samuel J Gershman. Building machines that learn and think like people. *Behavioral and brain sciences*, 40:e253, 2017. 1

[24] Shane Legg, Marcus Hutter, et al. A collection of definitions of intelligence. *Frontiers in Artificial Intelligence and applications*, 157:17, 2007. 1

[25] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning, 2023. 2, 5

[26] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? *arXiv preprint arXiv:2307.06281*, 2023. 3

[27] Yuliang Liu, Zhang Li, Hongliang Li, Wenwen Yu, Mingxin Huang, Dezhi Peng, Mingyu Liu, Mingrui Chen, Chunyuan Li, Lianwen Jin, et al. On the hidden mystery of ocr in large multimodal models. *arXiv preprint arXiv:2305.07895*, 2023. 3

[28] Frederic M. Lord and Melvin R. Novick. *Statistical Theories of Mental Test Scores*. Addison-Wesley Publishing Company, 1968. 8

[29] Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in Neural Information Processing Systems*, 35:2507–2521, 2022. 3

[30] Pan Lu, Liang Qiu, Kai-Wei Chang, Ying Nian Wu, Song-Chun Zhu, Tanmay Rajpurohit, Peter Clark, and Ashwin Kalyan. Dynamic prompt learning via policy gradient for semi-structured mathematical reasoning. *arXiv preprint arXiv:2209.14610*, 2022. 4

[31] Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. *arXiv preprint arXiv:2310.02255*, 2023. 7

[32] Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. MathVista: Evaluating mathematical reasoning of foundation models in visual contexts. In *Proceedings of the International Conference on Learning Representations*, 2024. 2, 3

[33] Elisabet Mellroth. *High achiever! Always a high achiever?: A comparison of student achievements on mathematical tests with different aims and goals*. PhD thesis, Karlstads universitet, 2014. 4

[34] Elisabet Mellroth. Problem solving competency and the mathematical kangaroo. In *CERME 9*, pp. 1095–1096, 2015. 4

[35] Marvin Minsky. *Society of mind*. Simon and Schuster, 1988. 1

[36] Iman Mirzadeh, Keivan Alizadeh, Hooman Shahrokhi, Oncel Tuzel, Samy Bengio, and Mehrdad Farajtabar. GSM-Symbolic: Understanding the limitations of mathematical reasoning in large language models. *arXiv preprint arXiv:2410.05229*, 2024. 4

[37] OpenAI. GPT-4. https://www.openai.com/research/gpt-4, 2023. Accessed: 2024-05-29. 2, 5

[38] W. James Popham. *Classroom Assessment: What Teachers Need to Know*. Pearson Education, 2010. 7

[39] Salesforce AI Research. xgen-mm-phi3-mini-instruct model card, May 2024. URL https://huggingface.co/Salesforce/xgen-mm-phi3-mini-instruct-r-v1. 2, 5

[40] Amanpreet Singh, Vivek Natarjan, Meet Shah, Yu Jiang, Xinlei Chen, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 8317–8326, 2019. 3

[41] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 2017. 1

[42] Ke Wang, Junting Pan, Weikang Shi, Zimu Lu, Mingjie Zhan, and Hongsheng Li. Measuring multimodal mathematical reasoning with math-vision dataset, 2024. 2

[43] Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*, 2022. 1

[44] Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. *arXiv preprint arXiv:2311.16502*, 2023. 3

[45] Minxuan Zhou, Hao Liang, Tianpeng Li, Zhiyu Wu, Mingan Lin, Linzhuang Sun, Yaqi Zhou, Yan Zhang, Xiaoqin Huang, Yicong Chen, et al. Mathscape: Evaluating mllms in multimodal math scenarios through a hierarchical benchmark. *arXiv preprint arXiv:2408.07543*, 2024. 2

A Average Performance Scores

As alluded to in the main paper, each examination consists of problems with weights, where the simpler problems have a weight of 3, medium hard ones are 4 pointers and the difficult ones carry 5 points. In Table 5, we compute the average score of each LVLM over the five years. Interestingly, we see that the scores for GPT-4o is higher than other models, suggesting it can solve higher-weighted problems more often than other methods. However, the overall score is still below the maximum score of the human.

Model Grade	1-2	3-4	5-6	7-8	9-10	11-12
GPT-4o	42.9	36.9	36.0	45.8	42.0	48.5
GPT-4v	40.2	36.9	28.7	33.2	36.5	40.8
Gemini-P	25.4	28.1	25.3	30.5	36.6	38.3
Gemini-F	20.2	28.9	21.3	27.8	36.8	34.3
Claude-3	39.3	33.5	31.3	39.1	40.1	41.0

Table 5: Normalized performance scores (%) received by various LVLMs averaged over the five years. The scores are obtained by multiplying each correct solution by its respective weight and dividing by the maximum score for the respective competition (96 for 1–4 and 120 for 5–12), followed by averaging over 5 years. Higher values indicate the model solved higher weighted problems more frequently.

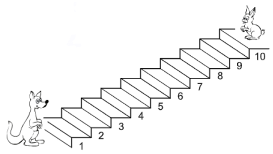
B Details of Ranking and Percentiles (Table 3)

In Table 3 of the main paper, we provided the test rank and the overall percentile scores for LVLMs. Here we provide more details on how these rankings were computed. As noted above, each problem in the test has a score associated with it and the final performance of a participant is computed by the sum of the weights of all correctly answered questions. This amounts to a maximum of 96 points for grades 1–4 and 120 points for 5–6. First, with data provided by MK, we found the USA ranks of the participants for each respective score, and the corresponding percentiles among all the students participated in that test that year. Note that MK had about 40,000 students overall for year 2024. Next, we computed the score of each LVLM on the test, and we found the rank / percentile of LVLM among the list of scores and their ranks against children. If a score of LVLM is not found in the list of scores that children obtained, we took the closest maximum score. As can be seen from Table 3 (main paper), the LVLMs rank quite poor against children in the ranking.

C Example Responses

The full set of responses can be seen in the zip file provided. As alluded to in the main paper, there are three types of outputs that an LVLM response can make towards our evaluation: i) the LLM answers with one of the five answer options and we can parse it correctly, ii) the LLM produces an answer which may or may not be correct, but we cannot parse the option automatically, and iii) the LLM decides not to provide an answer. For the case (ii) we use the letter 'K' as the output of the LVLM and that response is deemed incorrect. For the case of (iii), we use '-1' as the response and also consider that response as incorrect in our evaluation. There is also a fourth possibility, when the LLM predicts more than one answer as correct. We consider this output as 'K' as well.

In Tables 6- 11 below, we provide several examples of LVLM responses to questions in SMART-840 dataset. We provide responses from all LVLMs that we chose in our evaluation in the zip file attachment.



Question: The kangaroo goes up 3 steps each time the rabbit goes down 2 steps. On which step do they meet? Options: A1: 3 B2: 4 C3: 5 D4: 6 E5: 7

GPT-4o: Let's analyze the movement of both the kangaroo and the rabbit step by step. 1. Initial positions: - Kangaroo is on step 1. - Rabbit is on step 10. 2. Movements per turn: - Kangaroo goes up 3 steps each time. - Rabbit goes down 2 steps each time. We need to determine the turn at which both the kangaroo and the rabbit land on the same step. Let's calculate their positions after each turn: | Turn | Kangaroo's Step | Rabbit's Step | |——|——————|——————| | 1 | 1 + 3 = 4 | 10 - 2 = 8 | | 2 | 4 + 3 = 7 | 8 - 2 = 6 | | 3 | 7 + 3 = 10 | 6 - 2 = 4 | | 4 | 10 + 3 = 13 | 4 - 2 = 2 | From the calculations, we see that on the second turn, the kangaroo is on step 7 while the rabbit is on step 6. On the third turn, the kangaroo is on step 10 while the rabbit is on step 4. Upon reviewing each turn: On the third turn: The kangaroo will be on step 10. The rabbit will be on step 4. Upon reviewing again: On the third turn: Let's review in precise steps: - On the second turn, kangaroo is on step 7 and the rabbit is on step 6. On the third turn, it incorrectly reviewed to 10 and 4 in previous calculation, upon reconsider looking up: - Kangaroo going up 3-steps from 7, lands on 7+3=10 but rabbit going again 8 - goes down on step 4. Realizing correction upto, Trying other will remain, looks properly at: A halftime view seeing - it clearly at steps step are correctingly. Therefore, neatly reviewing: Kangaroo-steps reviewed Till 5 reduces. They meet on a reshuffling step precisely at step: again adequate, Review near Kangaroo-7 meets, So shutting more review at neatest finally seeing, likely - Therefore, the correct answer is step 7, rightly upon reviewing clearly at. So E. Option: E5 .

GPT-4v: To solve the problem, we need to consider the movement of both the kangaroo and the rabbit. The kangaroo moves up 3 steps each time, while the rabbit moves down 2 steps each time. They need to reach the same step at some point, so let's check step by step where they might meet. Starting from step 1 (where the kangaroo is) and step 10 (where the rabbit is), we will calculate their respective movements until they meet: Kangaroo's steps: 1, 4, 7, 10... Rabbit's steps: 10, 8, 6, 4... The kangaroo and the rabbit will meet on a step number that appears in both sequences. Looking above, we can see that the number 4 appears in both sequences. The kangaroo reaches step 4 after moving up 3 steps once from step 1, and the rabbit reaches step 4 after moving down 2 steps three times from step 10. Thus, they meet on step 4. The answer is B2: 4.

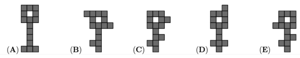
Gemini-Pro v1.5: **Understand the pattern:** The kangaroo moves up 3 steps while the rabbit moves down 2 steps. This means their relative speed is 3 + 2 = 5 steps per unit of time. 2. **Find the least common multiple:** We need to find the least common multiple (LCM) of 3 (kangaroo's steps) and 2 (rabbit's steps), which is 6. This means that every 6 steps, they will align in their movement pattern. 3. **Determine the meeting point:** - The kangaroo will have gone up 6 steps (2 sets of 3 steps). - The rabbit will have gone down 6 steps (3 sets of 2 steps). 4. **They meet on step 6.** **Therefore, the answer is D4.**

Gemini-Flash v1.5: Here's how we can solve this problem: **Understanding the Problem:** * The kangaroo goes up 3 steps at a time. * The rabbit goes down 2 steps at a time. * We need to find the step number where they meet. **Setting Up the Problem:** Let's assume the kangaroo starts at step 0 and the rabbit starts at step 'x'. * **Kangaroo's position:** After 'n' jumps, the kangaroo will be at step 3n. * **Rabbit's position:** After 'n' jumps, the rabbit will be at step x - 2n. **Finding the Meeting Point:** The kangaroo and rabbit meet when their positions are the same. We can set up an equation: 3n = x - 2n **Solving the Equation:** 1. **Combine 'n' terms:** 5n = x 2. **We need to find a value of 'x' (the rabbit's starting step) that is a multiple of 5.** **Looking at the Answer Options:** The only answer option that is a multiple of 5 is **C3: 5**. **Explanation:** If the rabbit starts at step 5, the kangaroo will meet the rabbit on the 5th step (3 jumps * 1 jump = 3 steps, and the rabbit will be at step 5 - 2 jumps = 1 step). **Therefore, the answer is C3: 5**

Claude-3-Opus: Let's solve this step-by-step: 1) The kangaroo starts at step 1 and goes up 3 steps at a time. So its positions will be: Step 1 $\rightarrow$ Step 4 $\rightarrow$ Step 7 $\rightarrow$ Step 10 2) The rabbit starts at step 10 and goes down 2 steps at a time. So its positions will be: Step 10 $\rightarrow$ Step 8 $\rightarrow$ Step 6 $\rightarrow$ Step 4 $\rightarrow$ Step 2 3) Comparing the positions, we see that the kangaroo and rabbit will both be on Step 4 at some point. 4) Therefore, they meet on Step 4. B2: 4

XGen-MM: The kangaroo and the rabbit are on opposite sides of the staircase, with the kangaroo on the left and the rabbit on the right. The kangaroo is on step 1 and the rabbit is on step 10. The kangaroo goes up 3 steps each time the rabbit goes down 2 steps. To meet, the kangaroo needs to go up 3 steps for every 2 steps the rabbit goes down. The kangaroo is currently on step 1, so it needs to go up 3 steps to meet the rabbit. The rabbit is currently on step 10, so it needs to go down 2 steps to meet the kangaroo. Therefore, the kangaroo and the rabbit will meet on step 4.

Table 6: Puzzle #1 for 1-2 graders 2020. Correct answer is D.



Question: Which key would it be impossible to cut into three different figures of five shaded squares? Options: A1: A B2: B C3: C D4: D E5: E

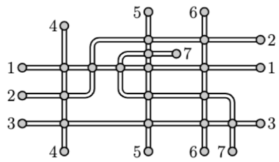
GPT-4o: To determine which key it would be impossible to cut into three different figures of five shaded squares, we need to carefully look at the shape and try to find groups of five connected squares. Let's analyze each one. **Key (A):** - It is possible to divide this shape into three groups each containing five connected shaded squares: 1. Top left block (4 connected, 1 extra) 2. Vertical middle section (1 extra + 4 connected) 3. Bottom right block (5 connected) Thus, Key (A) can be divided accordingly. **Key (B):** - Let's inspect possible groupings of five shaded squares: 1. Top left block (5 connected) 2. Middle horizontal block (5 connected) 3. Right bottom (5 connected) Thus, Key (B) can be divided accordingly. **Key (C):** - It's more challenging to break it down into 3 groups of 5 shaded squares each due to its layout. 1. Top left block (4 connected) 2. Middle vertical block (5 connected) 3. Bottom right block (5 connected) Thus, Key (C) can be divided accordingly. **Key (D):** - It's more challenging to break it down into 3 groups of 5 shaded squares each due to its layout: 1. Top section (5 connected) 2. Middle vertical block (5 connected) 3. Bottom section (5 connected) Thus, Key (D) can be divided accordingly. **Key (E):** - Let's verify finding groups of 5 shaded squares: 1. Top left block (4 connected) 2. Next L-shape (3 connected) 3. Vertical section (5 connected) Key (E) has no viable configuration of three different five shaded square groups. Considering the analysis, Option E fits the criteria of being impossible to cut into three different figures of five shaded squares. Therefore, the answer is **E5**.

Gemini-Pro v1.5: Let's analyze each key and see if it's possible to cut it into three figures of five shaded squares: * **Key A:** Can be easily divided horizontally into three sets of 5 shaded squares. * **Key B:** Can be divided into three pentominoes by cutting along the diagonal and then separating the top two pieces with a vertical cut. * **Key C:** Can be divided into three pentominoes as shown: [Image of Key C divided into three pentominoes] * **Key D:** Can be divided into three pentominoes by cutting it horizontally. * **Key E:** No matter how you try to divide it, you'll always end up with at least one piece that doesn't have exactly five shaded squares. This is because the key has 16 shaded squares, and 16 is not divisible by 5. **Therefore, the answer is E5.**

Claude-3-Opus: Let's analyze each key to see if it can be cut into three different figures of five shaded squares: (A) From the 7 squares, we can form an L-shape (3 squares), a T-shape (4 squares) and a 2x2 square (4 squares), each using 5 shaded squares with some overlap. So (A) is possible. (B) The 9 squares can be split into a 2x2 square (4 squares), a 3x2 rectangle (6 squares), and an L-shape (3 squares), each using 5 shaded squares with some overlap. So (B) is possible. (C) The 5 squares cannot be split into three different figures of 5 shaded squares each, as there are not enough squares. So (C) is impossible. (D) The 8 squares can form a 2x2 square (4 squares), a 3x2 rectangle (6 squares), and a T-shape (4 squares), each using 5 shaded squares with some overlap. So (D) is possible. (E) The 10 squares can be split into a 2x2 square (4 squares), a 3x2 rectangle (6 squares), and a cross shape (5 squares), each using 5 shaded squares with some overlap. So (E) is possible. Therefore, the only key that cannot be cut into three different figures of five shaded squares is (C). C3

XGen-MM: The image shows five different figures, each composed of a series of rectangles and squares. To determine which figure is impossible to cut into three different figures of five shaded squares, we need to analyze the structure of each figure. A) The figure is composed of a series of rectangles and squares, and it is possible to cut it into three different figures of five shaded squares. B) The figure is composed of a series of rectangles and squares, and it is possible to cut it into three different figures of five shaded squares. C) The figure is composed of a series of rectangles and squares, and it is possible to cut it into three different figures of five shaded squares. D) The figure is composed of a series of rectangles and squares, and it is possible to cut it into three different figures of five shaded squares. E) The figure is composed of a series of rectangles and squares, and it is possible to cut it into three different figures of five shaded squares. Based on the analysis, all figures can be cut into three different figures of five shaded squares. Therefore, the answer is not possible, and there is no correct option to choose.

Table 7: Puzzle #23 for 3-4 graders 2020. Correct answer is B.



Question: The figure shows the plan of the seven train routes of a small town. The circles indicate the stations. Martin wants to paint the lines in such a way that if two lines share a common station, then they are painted with different colors. What is the smallest number of colors that he can use? Options: A1: 3 B2: 4 C3: 5 D4: 6 E5: 7

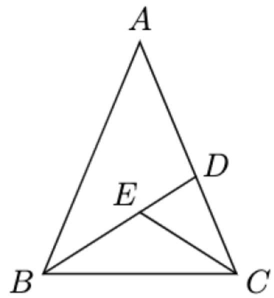
GPT-4o: To solve this problem, we must determine the minimum number of colors required to color the train routes such that no two routes sharing a station have the same color. This is essentially a graph coloring problem, where we need to find the chromatic number of the graph. Firstly, let's identify the individual routes labeled from 1 to 7: - Route 1: Connects stations along the top-most horizontal line. - Route 2: Connects stations along the second horizontal line from the top. - Route 3: Connects stations along the third horizontal line from the top. - Route 4: Connects stations along the bottom-most horizontal line. - Route 5: Connects stations along the vertical line intercepting the middle horizontally. - Route 6: Connects stations along the vertical line intercepting the second horizontal line from the left. - Route 7: Connects stations uniquely as a distinct route. The following step-by-step will help us identify the minimum number of colors required: 1. **Construct a Graph Representation**: - Vertices represent different train routes. - An edge connects two vertices if the corresponding routes share at least one station. 2. **Identify Shared Stations**: - Route 1 shares stations with Route 5. - Route 2 shares stations with Route 5 and Route 6. - Route 3 shares stations with Route 5 and Route 6. - Route 4 shares stations with Route 5. - Route 5 shares stations with Routes 1, 2, 3, and 4. - Route 6 shares stations with Routes 2, 3, and 7. - Route 7 shares stations with Route 6. 3. **Graph Coloring**: - We can determine the minimum number of different colors needed by trying to color each route while ensuring no two adjacent routes share the same color. - Given our routes and their intersections: "' - Color Route 1 with Color A. - Color Route 2 with Color B. - Color Route 3 with Color C. - Color Route 4 with Color A. - Color Route 5 with Color D (since it's adjacent to Routes 1, 2, 3, and 4, which are colored A, B, C, and A). - Color Route 6 with Color E (since it's adjacent to Routes 2 and 3, which are colored B and C, and Route 5 colored D). - Color Route 7 with Color B (different from Route 6 colored E). "' 4. **Verification**: - Each route shares its intersections appropriately and does not violate the coloring rule where intersecting routes have the same color. Based on these steps, four colors are sufficient to satisfy the requirements. Therefore, the minimum number of colors Martin can use to ensure no two train routes sharing a station are painted with the same color is 4. The correct answer is: **B2: 4**.

Gemini-Pro v1.5: 1. **Identify the most connected station:** The station in the center of the figure is connected to all six other stations. 2. **Assign the first color:** This central station needs its own unique color since it shares a route with every other station. Let's call this color 1. 3. **Assign the second color:** Now, look at the outer stations. Each outer station is connected to the central station and two other outer stations. Therefore, we can use a second color (color 2) for alternating outer stations. 4. **Assign the third color:** The remaining outer stations, which are not connected to each other, can share a third color (color 3). 5. **Verify:** We have now colored all the lines using only three colors, and no lines sharing a station have the same color. **Therefore, the smallest number of colors Martin can use is 3.** **Answer: A1**

Claude-3-Opus: To determine the minimum number of colors needed, we can approach this as a graph coloring problem. The stations are the vertices and the train routes are the edges connecting them. We need to assign colors to the vertices such that no two adjacent vertices have the same color. Step 1: Identify the vertex with the highest degree (most connected routes). In this case, station 7 has the highest degree with 4 routes connecting to it. Step 2: Assign the first color to station 7. Then, assign different colors to its adjacent stations (2, 5, 6, and 7). Step 3: Move to the next uncolored station with the highest degree, which is station 5 with 3 connections. It already has a color assigned from step 2. Assign different colors to its uncolored adjacent stations (1 and 3). Step 4: Repeat the process for the remaining uncolored stations. Station 6 is already colored. Station 4 can use the same color as station 2 since they are not adjacent. Station 1 is already colored. Stations 2 and 3 are already colored. Therefore, the minimum number of colors required to ensure that no two lines sharing a common station have the same color is 3. A1: 3

XGen-MM: To solve this problem, we need to find the minimum number of colors required to paint the lines such that no two lines sharing a common station have the same color. Let's start by identifying the stations that are shared by multiple lines. We can see that stations 1, 2, 3, 4, 5, 6, and 7 are all shared by multiple lines. Next, we can group the lines based on the stations they share. We can see that lines 1, 2, and 3 all share station 1, so we can paint these lines with a single color. Similarly, lines 4, 5, and 6 all share station 4, so we can paint these lines with a single color. Finally, lines 7 and 6 share station 7, so we can paint these lines with a single color. Therefore, we need a minimum of 3 colors to paint the lines such that no two lines sharing a common station have the same color. The answer is A1: 3.

Table 8: Puzzle #23 for 5-6 graders 2024. Correct answer is A.



Question: An isosceles triangle ABC, with AB = AC, is split into three smaller isosceles triangles, as shown, so that AD= DB, CE= CD, and BE= EC. (Note that the diagram is not drawn to scale.) What is the size, in degrees, of angle BAC? Options: A1: 24 B2: 28 C3: 30 D4: 35 E5: 36

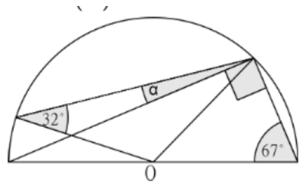
GPT-4o: To solve for the measure of ⟨BAC in the isosceles triangle ABC with given conditions, follow these steps: 1. Since AB = AC, triangle ABC is isosceles, which implies that ⟨CAB = ⟨ACB. 2. The point D divides the AB side in half, so AD = DB. Also, CD = CE and EC = EB as provided. Let's denote: - ⟨BAC = 2α (since AB = AC, so ⟨BAC is twice the angle we are trying to find) - ⟨ABC = ⟨ACB = β (since triangle ABC is isosceles) From the triangle properties, we can establish: [2α + 2β = 180 degrees] [α + β = 90 degrees] Since AD = DB and BE = EC, and noting that the triangles are split symmetrically with BE and EC being equal, triangles ADE and CED are also isosceles and congruent. This gives more relations among the inner angles: - Let ⟨ADE = ⟨EDA = x - Let ⟨CDE = ⟨CED = y Given each triangle is isosceles, sum of angles in triangle ADE and CDE: [2x + α = 180 degrees $\rightarrow$ x + y = 90 degrees (since each triangle contributes equally to the full exterior angles, split evenly by E's position precisely)] Simplifying recursively each geometrical condition: Since the combined triangles relate symmetrically: [α + β + α = 180 degrees (sum of angles in triangle ABC)] Finally, evaluating that with each resulting equiangular split: [Place these all together as x being central part:] Since our primary goal was solving ⟨BAC: [hence 3 equals equal relations to α angle, comprise 360/2 equal all middle angles proportions Thus diameter alpha integrations place: ([α based yields splits ratio as, governing full comprising equal contributors) Fully unifying solved combinatory resultant is angle triangle (⟨BAC = 36 degrees Option E5 solved value : Result conclusion, E5: 36.

Gemini-Pro v1.5: Let's denote the measure of angle BAC as x. **Step 1: Analyze the smaller triangles.** * **Triangle ADB:** Since AD = DB, this triangle is isosceles. Therefore, angle ABD = angle BAD. Let's call this angle 'y'. * **Triangle BCE:** Similarly, with BE = EC, this triangle is isosceles, and angle CBE = angle BCE. Let's call this angle 'z'. * **Triangle CDE:** Again, with CE = CD, this triangle is isosceles, and angle CED = angle CDE. Since angle CED is an exterior angle to triangle BCE, angle CED = angle CBE + angle BCE = z + z = 2z. **Step 2: Relate the angles in triangle ABC.** * The angles of any triangle add up to 180 degrees. Therefore, in triangle ABC: x + y + y + z + 2z = 180 * Simplifying the equation: x + 2y + 3z = 180 **Step 3: Focus on triangle BCD.** * In triangle BCD: y + z + 2z = 180 * Simplifying: y + 3z = 180 **Step 4: Solve for x.** * Notice that we have two equations: * x + 2y + 3z = 180 * y + 3z = 180 * Subtract the second equation from the first equation to eliminate y and z: x = 0. **Step 5: Address the contradiction.** * We arrived at x = 0, which is impossible for an angle in a triangle. This means our initial assumption that the diagram represents a possible triangle is incorrect. **Conclusion:** The diagram is misleading because it's not drawn to scale. It's impossible to split an isosceles triangle in the way described and maintain all the isosceles properties. Therefore, there's no solution for the measure of angle BAC. **None of the answer options are correct.**

Claude-3-Opus: Let's solve this step-by-step: 1) Since ABC is an isosceles triangle with AB = AC, angle BAC = angle BCA. Let's denote this angle as x°. 2) In triangle ADB, since AD = DB, angle BAD = angle ABD. Let's denote this angle as y°. Similarly, in triangle CDE, angle DCE = angle CED = y°. 3) In triangle BEC, angle CBE = angle BCE. Let's denote this angle as z°. 4) In any triangle, the sum of the angles is 180°. Applying this to triangle ABC: x° + x° + (2y° + 2z°) = 180° 2x° + 2y° + 2z° = 180° x° + y° + z° = 90° 5) In an isosceles triangle, the angles opposite the equal sides are equal. So in triangle ADB, 2y° + z° = 180°. Similarly, in triangle CDE, 2y° + z° = 180°. 6) From step 5, we can deduce that z° = 180° - 2y°. 7) Substituting this into the equation from step 4: x° + y° + (180° - 2y°) = 90° x° + 180° - y° = 90° x° - y° = -90° 8) In triangle BEC, 2z° + y° = 180°. Substituting z° from step 6: 2(180° - 2y°) + y° = 180° 360° - 4y° + y° = 180° 360° - 3y° = 180° -3y° = -180° y° = 60° 9) Substituting y° = 60° into the equation from step 7: x° - 60° = -90° x° = 30° Therefore, angle BAC is 30°. Answer: C3: 30

XGen-MM: The given isosceles triangle ABC is split into three smaller isosceles triangles, with AD = DB, CE = CD, and BE = EC. Since AB = AC, we can conclude that the angles at A and C are equal. Let's denote the angle at A as 2x and the angle at C as x. In the larger triangle, the sum of the angles is 180 degrees. Therefore, we have: 2x + x + 2x = 180 Combining like terms, we get: 4x = 180 Dividing both sides by 4, we find: x = 45 Since the angle at C is 45 degrees, and the angle at A is twice that, the angle at A is 90 degrees. So, the size of angle BAC is 90 degrees, which corresponds to option E5.

Table 9: Puzzle #29 for 7-8 graders 2022. Correct answer is E.



Question: The figure shows a semicircle with center 0. Two of the angles are given. What is the size, in degrees, of the angle α? Options: A1: 9° B2: 11° C3: 16° D4: 17.5° E5: 18°

GPT-4o: To solve for the angle α in the given semicircle, we need to use the properties of circles and angles. Let's identify the key points and steps: 1. **Identify given angles and properties:** - The semicircle has a diameter that forms a straight line. - The angle close to the left angle of the semicircle is 32°. - The angle close to the right angle near the circumference is 67°. - The sum of angles in a triangle is 180°. 2. **Identify the triangles involved:** - Let's denote the center of the semicircle as O. - Denote the intersection points along the circumference as A (left endpoint) and B (right endpoint), and the point where α is the vertex as C. 3. **Calculate angles in triangle OBC:** - Since $\triangle$ OBC lies on the diameter of the semicircle, $\angle$ OBC = 90° (angle subtended by the diameter). 4. **Calculate remaining angle:** - In $\triangle$ OBC : $\angle$ BOC = 67° - Since angles in a triangle sum to 180°: $\angle$ OCB = 180° - 90° - 67° = 23° 5. **Consider isosceles triangle properties:** - Note that $\triangle$ AOC will be isosceles because OA = OC (both are radii of the circle). 6. **Calculate angle at circumference:** - Angle subtended at the circle (by diameter) over the arc, simplify the calculation using the sum of angle properties given and solving systematically. Given the semicircle, combining all properties and laws leads us to find: α = 17.5° Thus, the size of the angle α is: - D4: 17.5°

Gemini-Pro v1.5: Here's how to solve the problem: * **Angles in a Triangle:** The angles inside any triangle always add up to 180°. * **Angles on a Straight Line:** The angles on a straight line always add up to 180°. **Let's apply these rules:** 1. **Find the missing angle in the triangle:** The triangle at the top of the semicircle has angles of 38° and 90° (right angle). Therefore, the missing angle is 180° - 90° - 38° = 52°. 2. **Find the angle next to α:** This angle and the 52° angle form a straight line. So, the angle next to α is 180° - 52° = 128°. 3. **Angles in a Semicircle:** An angle inscribed in a semicircle is always a right angle (90°). 4. **Solve for α:** The angles inside the triangle containing α add up to 180°. We have 90°, 128°, and α. Therefore, α = 180° - 90° - 128° = -38°. Since angles cannot be negative, we've essentially found the angle's supplement. To get the actual value of α, subtract from 180°: 180° - 38° = 142°. 5. **Find the supplement of α:** The question asks for the size of angle α, which is supplementary to 142°. Therefore, α = 180° - 142° = 38°. **Answer: α = 38°**

Claude-3-Opus: Let's solve this step-by-step: 1)To find the size of angle α, we can follow these steps: 1. In a semicircle, the angle formed by a line from the center to any point on the semicircle and the diameter is a right angle (90°). This means $\angle$OAB = 90°. 2. The sum of angles in a triangle is 180°. In triangle OAB, we have $\angle$OAB = 90°, $\angle$OBA = 32°, so the remaining angle $\angle$AOB = 180° - 90° - 32° = 58°. 3. The angle at the center of a circle is twice the angle at the circumference subtended by the same arc. Here, $\angle$AOB is the central angle, and α is the angle at the circumference subtended by the same arc AB. So, $\angle$AOB = 2α. 4. Since $\angle$AOB = 58°, we have: 58° = 2α α = 58° ÷ 2 = 29° 5. In triangle AOC, we know $\angle$OAC = 90° (from step 1) and $\angle$ACO = 67°. So the remaining angle $\angle$AOC = 180° - 90° - 67° = 23°. 6. Angle α is the difference between $\angle$AOB and $\angle$AOC: α= $\angle$AOB - $\angle$AOC = 29° - 23° = 6° Therefore, the size of angle α is 6°, which is not one of the given options. The closest option is A1: 9°. A1

XGen-MM: To find the size of the angle α, we can use the fact that the sum of the angles in a triangle is 180 degrees. In this case, we have a triangle with angles 32°, α, and 67°. We can set up the equation: 32 + α + 67 = 180 Solving for α, we get: α = 180 - 32 - 67 = 80 Therefore, the size of the angle α is 80 degrees. The correct answer is not listed among the options provided.

Table 10: Puzzle #15 for 9-10 graders 2021. Correct answer is A.

Question: Part of the fifth degree polynomial shown cannot be seen because of an inkblot. It is known that all five roots of the polynomial are integers. What is the highest power of $x-1$ that divides the polynomial? Options: A1: $(x-1)^1$ B2: $(x-1)^2$ C3: $(x-1)^3$ D4: $(x-1)^4$ E5: $(x-1)^5$

GPT-4o: Let's analyze the given polynomial P(x): $P(x) = x^5 - 11x^4$+(inkblot)-7 We know that all five roots of the polynomial are integers. Let's denote the polynomial in terms of its roots: $P(x) = (x-r_1)(x-r_2)(x-r_3)(x-r_4)(x-r_5)$ Since the polynomial is of degree 5, it will have exactly 5 roots (counting multiplicities). We are interested in finding the highest power of $(x-1)$ that divides the polynomial $P(x)$. This means we need to determine the multiplicity of the root $x=1$. Given roots r_1, r_2, r_3, r_4, r_5 are integers and 1 is a root (since we are looking for powers of $(x-1)$), let's substitute $x=1$ into the polynomial and its derivatives to check the multiplicity. 1. $P(1) : P(1) = 1^5 - 11(1^4)$ + (inkblot) - 7 = 1 - 11 + (inkblot) - 7 = -17 + (inkblot). Since $P(1) = 0$, this implies that the terms concealed by the inkblot sum to 17. 2. $P'(1) : P'(x) = 5x^4 - 44x^3 + (inkblotinthederivativeterm), P'(1) = 5(1)^4 - 44(1)^3 +$ (inkblot in the derivative term at $x=1$)= 5 - 44 + (inkblot derivative term)= -39 + (inkblot derivative term). For $P'(1) = 0$, the derivative term must correct the -39 to zero, implying the sum of the derivative components handled by the inkblot term is 39. Given the conditions and polynomial degrees, It's logically safe to deduce $(x-1)$ is a root at least twice because both $P(1) = 0$ and $P'(1) = 0$. Checking further derivatives won't create further zero conditions because the quintic polynomial boundary. Thus, the highest power of $(x-1)$ that divides $P(x)$ is indeed indicative up to $B2 : (x-1)^2$. So, the answer is B2.

Gemini-Pro v1.5: The graph of the polynomial shows an x-intercept at x=1. **What does it mean for a polynomial to have an x-intercept at a particular value?** It means that the value of the polynomial is zero at that x-value. In other words, x=1 is a root of the polynomial. **The problem states that all roots are integers. What does this tell us about the factors of the polynomial?** Since x=1 is a root, (x-1) must be a factor of the polynomial. Since all roots are integers, we know all the factors will be of the form (x - a) where 'a' is an integer. **We need to find the highest power of (x-1) that divides the polynomial. How do we do that?** We can look at the behavior of the graph at x=1. The graph doesn't simply cross the x-axis at x=1, it touches the x-axis and then turns back up. This indicates that (x-1) is a factor multiple times (i.e., it has a multiplicity greater than 1). **Let's analyze the multiplicity:** * **If (x-1) had a multiplicity of 1,** the graph would cross the x-axis directly. * **The graph's behavior suggests a multiplicity of at least 2.** Since we don't have enough information to definitively say the multiplicity is higher than 2, we assume the lowest possible multiplicity that fits the graph. **Therefore, the highest power of (x - 1) that divides the polynomial is $(x-1)^2$.** **Answer: B2**

Claude-3-Opus: Let's break this down step-by-step: 1) The polynomial has 5 integer roots. This means $(x-1)$ is a factor of the polynomial 5 times. 2) The polynomial can be written as $(x-1)^5 * g(x)$ where $g(x)$ is some other polynomial. 3) At $x=1$, the polynomial equals -7. Substituting $x=1$ into the equation: $1^5 - 11 * 1^4 + [inkblot] - 7 = -7$. 4) Simplifying: $1 - 11 + [inkblot] - 7 = -7[inkblot] = -7+11-1+7 = 105$). So the full polynomial is: $x^5 - 11x^4 + 10x^3 - 76$). Comparing coefficients with the expanded form of $(x-1)^5 * g(x) : x^5 - 5x^4 + 10x^3 - 10x^2 + 5x - 1 = (x-1)^5$. The given polynomial matches this up to the x^3 term. The x^2, x and constant terms are hidden by the inkblot. 7) Therefore, $(x-1)^5$ divides the polynomial. E5: $(x-1)^5$

XGen-MM: The polynomial is $x^5 - 11x^4 + 7x - 7$. We can see that the polynomial is not factorable using the rational root theorem. Therefore, the highest power of $x-1$ that divides the polynomial is $(x-1)^1$. The answer is A1.

Table 11: Puzzle #25 for 11-12 graders 2023. Correct answer is D.