

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

Anoop Cherian¹, Kuan-Chuan Peng¹, Suhas Lohit¹,
Joanna Matthiesen², Kevin Smith³, and Joshua B. Tenenbaum³

¹Mitsubishi Electric Research Labs (MERL), Cambridge, MA

²Math Kangaroo USA, ³Massachusetts Institute of Technology (MIT), Cambridge, MA

Vancouver, NeurIPS Datasets & Benchmarks, 2024

In the recent times, ...



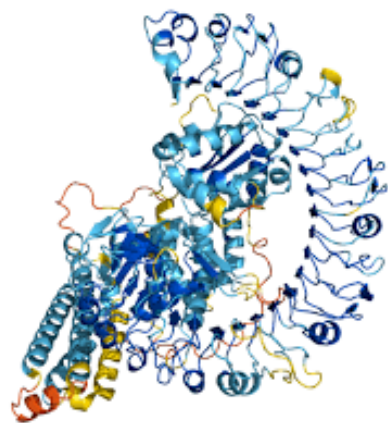
GPT Models



Alpha Geometry, Nature, 2024



DALL.E 3, OpenAI



AlphaFold, Nature, 2021



FunSearch, Nature, 2024



Emu-video, Meta, 2024

Motivation: Key Questions

1. Are the current large AI models indeed capable of generalized problem solving as humans do?
2. How well do AI models perform on tasks that need broad skills?
3. Do they transfer knowledge to solve new problems?
4. How effective are they in solving vision-and-language problems that even children can perform well on?



Human Cognitive Development and AI Expectations

- The type and nature of problems that humans can solve vary significantly at different stages of their growth
- Humans learn over the years through cumulative knowledge gathering
- We should expect AI models to know primitive concepts to build up to reason.

How will we do such comparative analysis?

Our Approach

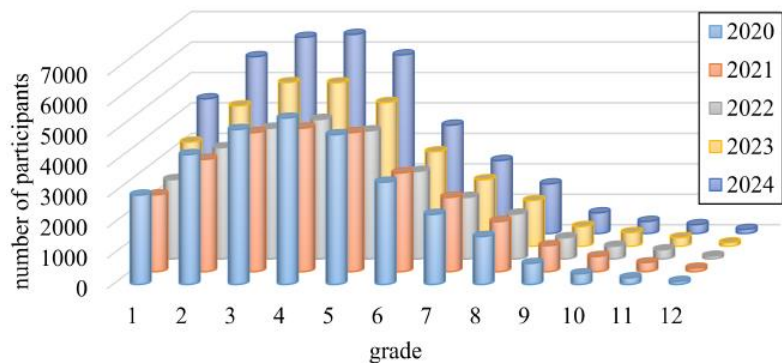
- **Key Idea: Compare human and AI performance on tasks that can be compared one-to-one on.**
- Which task?
 - Mathematical and algorithmic reasoning for solving Olympiad problems

Our Approach

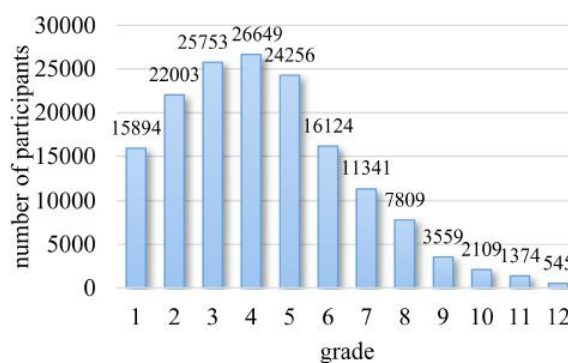
- We consider problems from the Mathematical Kangaroo (MK) Olympiad
 - ❖ A popular international competition targeted at children from grades 1-12
 - ❖ Each exam tests children's deeper mathematical abilities using puzzles that are appropriately gauged to their age and skills.
- Using the puzzles from MK, we created a dataset: *SMART-840*,
 - ❖ Our dataset consists of 840 problems from years 2020-2024 for grades 1-12
 - ❖ MK also has recorded the children's performances for each of these exams.

Our idea: To **analyze** LVLMs power on mathematical reasoning and compare their performances **directly** to that of children on the corresponding puzzles.

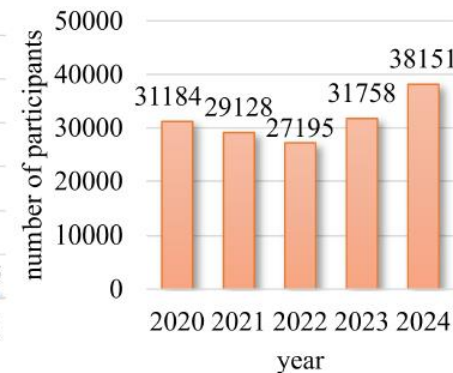
SMART-840 Data and Children's Participation Statistics



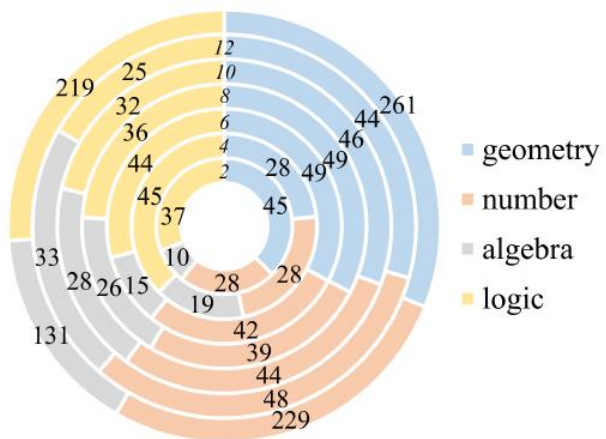
(a)



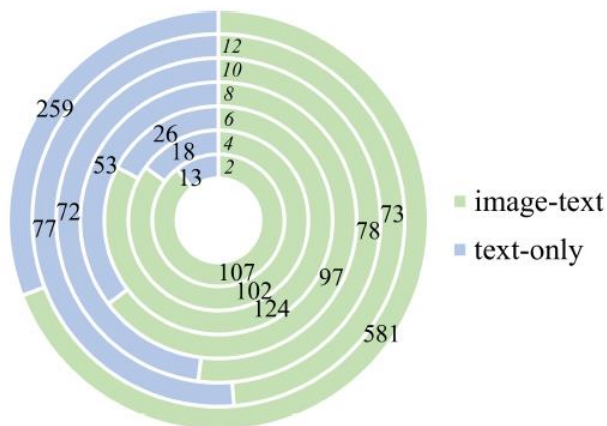
(b)



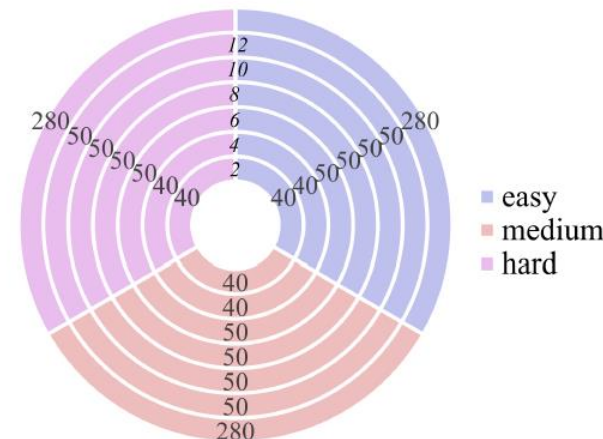
(c)



(d)



(e)



(f)

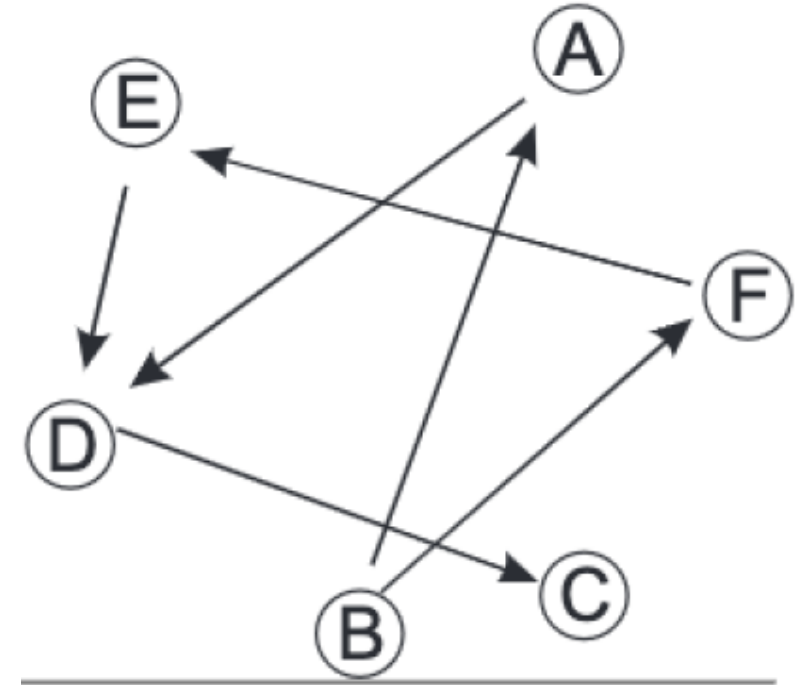
Example Problems in the SMART-840 Dataset

Question

An arrow pointing from one person to another means that the first person is taller than the second. For example, person B is taller than person A. Who is the shortest?

Answer Options:

A	Person A
B	Person B
C	Person C
D	Person D
E	Person E



Year 2020, Grade-1 & 2, Difficulty: hard

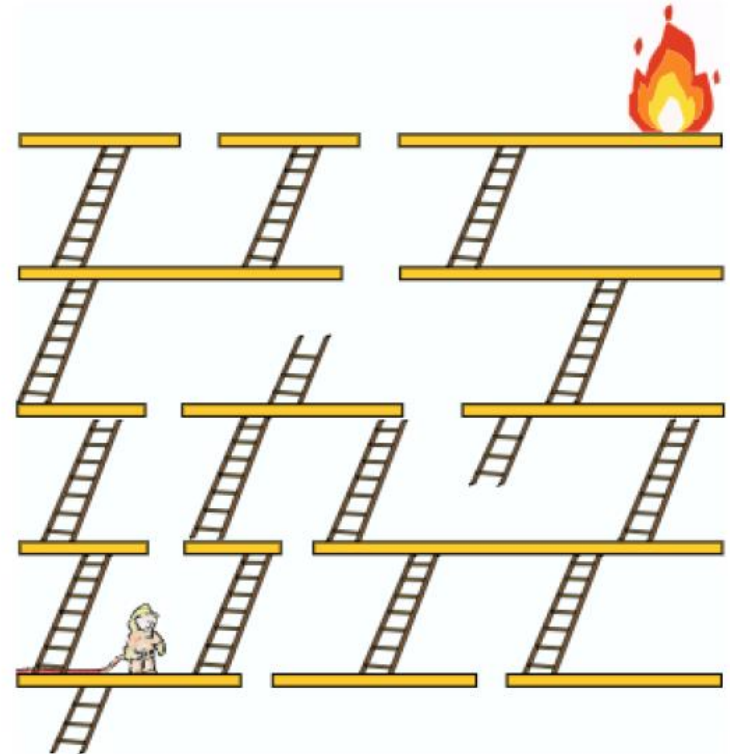
Example Problems in the SMART-840 Dataset

Question

What is the smallest number of ladders the firefighter must use to reach the fire without jumping between platforms?

Answer Options:

A	4
B	5
C	6
D	7
E	8



Year 2024, Grade-3&4, Difficulty: Easy

Example Problems in the SMART-840 Dataset

Question

Martin has three cards with numbers written on both sides. The card with number 1 on one side has number 4 on the opposite side, the card with 2 on has 5 on the opposite side and the card with 3 on has 6 on the opposite side. Martin randomly places all three cards on the table and adds up the three numbers he sees. How many different sums can Martin get?

Answer Options:

A	3
B	4
C	5
D	6
E	10

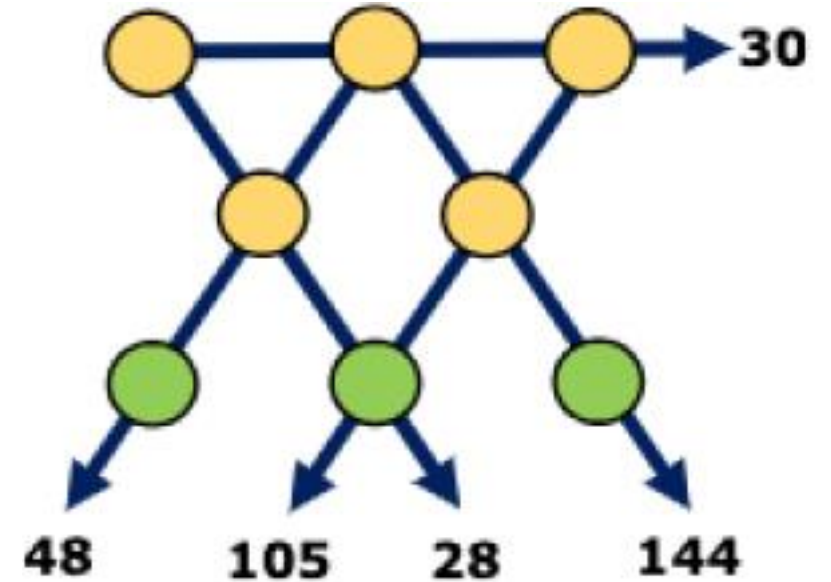
	Front	Back
Card 1	1	4
Card 2	2	5
Card 3	3	6

Year 2023, Grade-5&6, Difficulty: Hard

Example Problems in the SMART-840 Dataset

Question

The numbers 1 to 8 are placed, once each, in the circles shown. The numbers by the arrows show the products of the three numbers in the circles on that straight line. What is the sum of the numbers in the three circles at the bottom of the figure?



Answer Options:

A	11
B	12
C	15
D	17
E	19

Year 2022, Grade-7&8, Difficulty: Medium

Example Problems in the SMART-840 Dataset

Question

Anna wants to write a number in each of the squares of the grid so that the sum of the four numbers in each row and the sum of the four numbers in each column are the same. She has already written some numbers, as shown. What number does she write in the shaded square?

Answer Options:

A	5
B	6
C	7
D	8
E	9

1		6	3
	2	2	8
	7		4
		7	

Year 2020, Grade-9 &10, Difficulty: hard

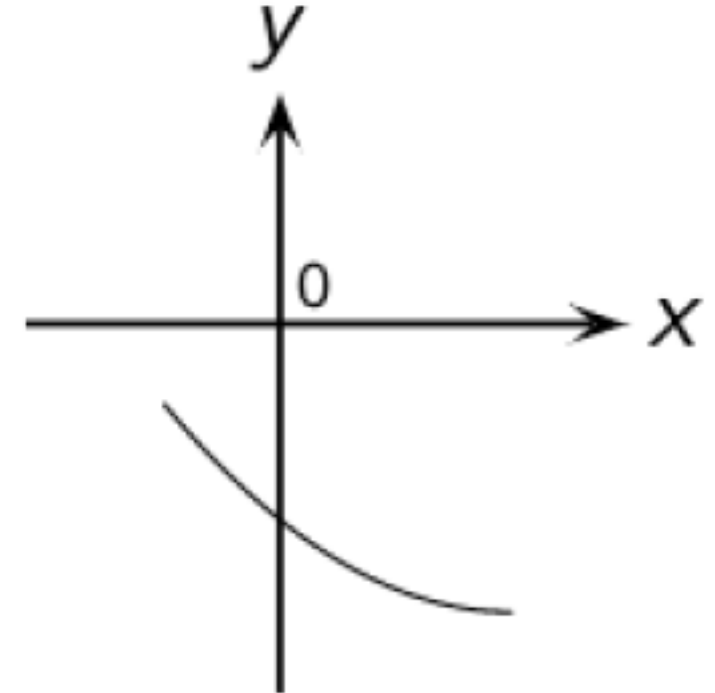
Example Problems in the SMART-840 Dataset

Question

The figure shows a section of the parabola with equation $y = ax^2 + bx + c$. Which of the following numbers is positive?

Answer Options:

A	c
B	$b+c$
C	ac
D	be
E	ab

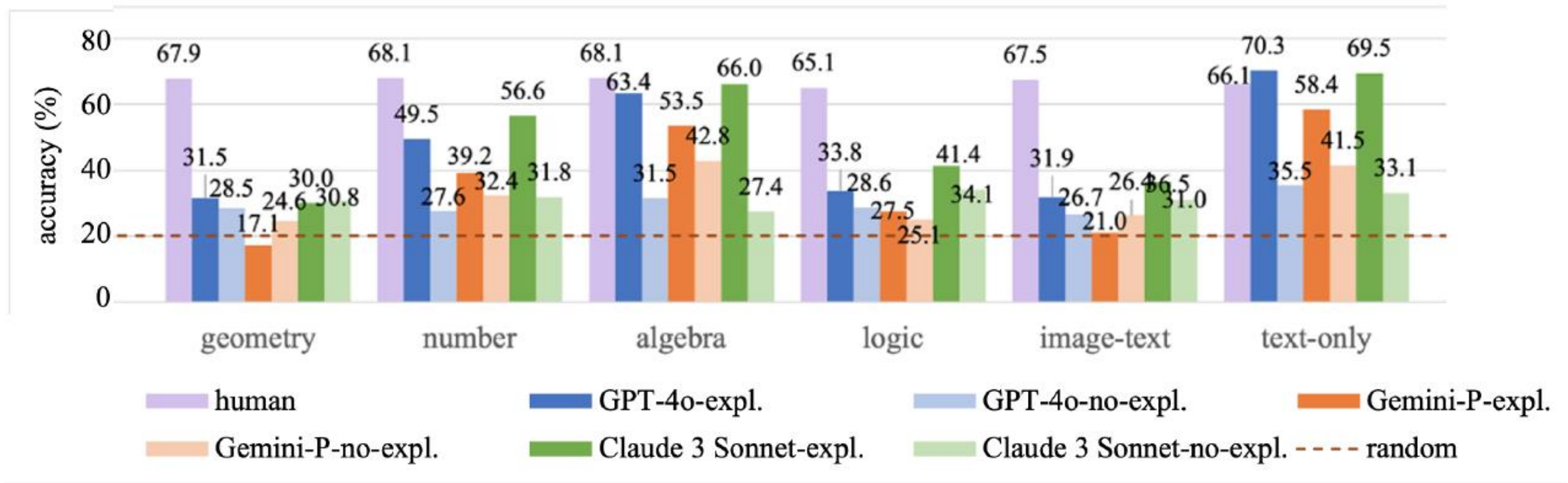


Year 2020, Grade-11&12, Difficulty: hard

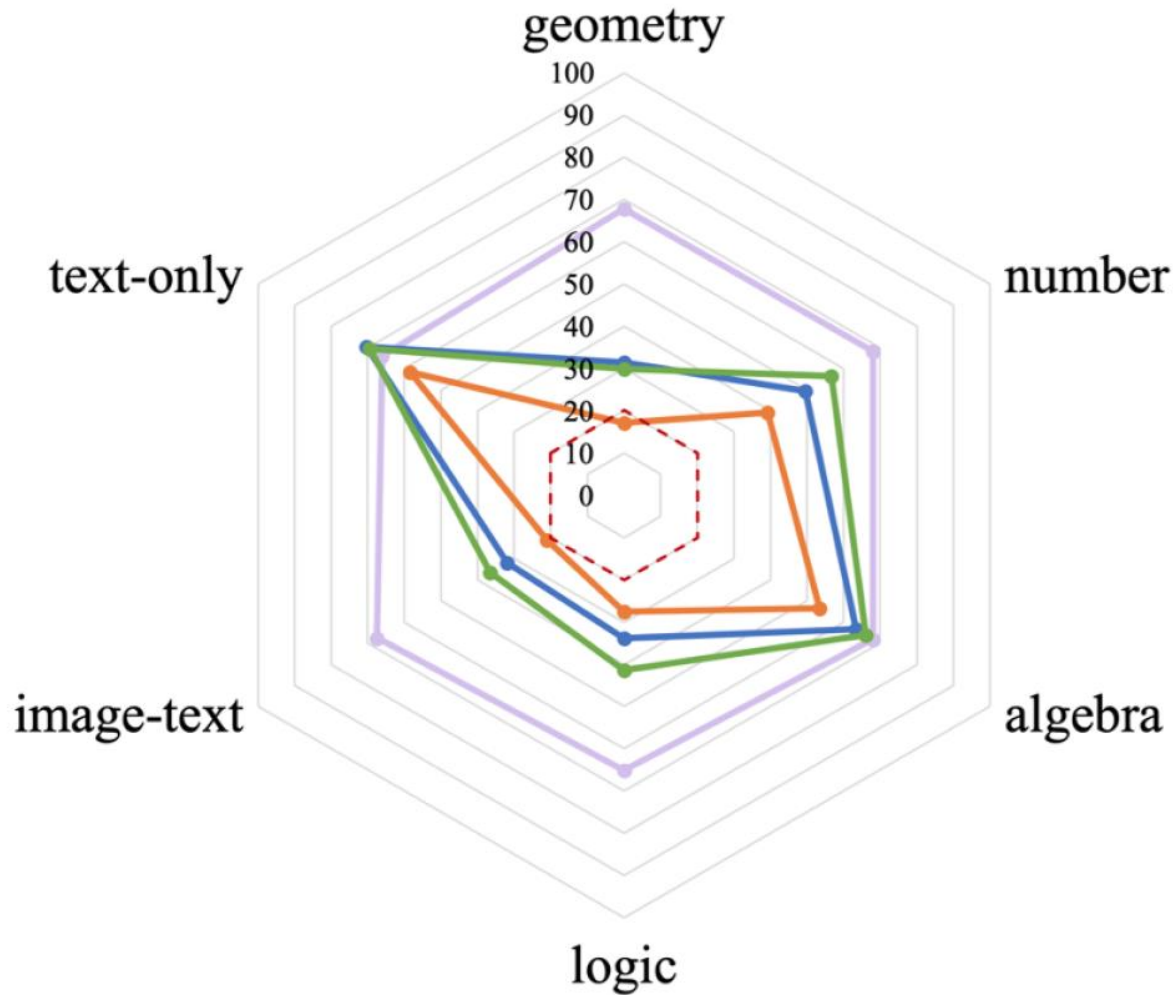
Performances of LVLMs on SMART-840

Grade \ Model	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	20.2	20.1	20.1	20.2	20.2	20.3	20.1	20.1	20.1	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						

Performances of LVLMs on SMART-840



Performances of LVLMs on SMART-840



- human
- GPT-4o-expl.
- Gemini-P-expl.
- Claude 3 Sonnet-expl.

Performances of LVLMs on SMART-840

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
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
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

Diff-I: Difficulty Index

Disc-I: Discriminative Index

Time-C: Correlation on the difficulty of questions based on the **time taken** to solve them

Weight-C: Correlation on the difficulty of questions based on their **number of points**

Entropy-C: Correlation on the **distribution of answer selections** by humans

Responses from LVMs

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?

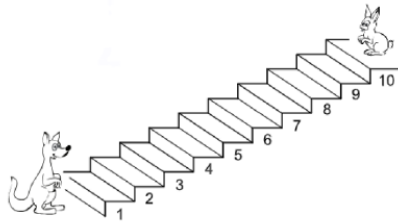
A: 6 B: 15 C: 18 D: 24 E: 33



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****

Responses from LVMs

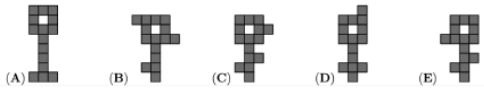


Question: The kangaroo goes up 3 steps each time the rabbit goes down 2 steps. On which step do they meet?
Options: A1: 3 B2: 4 C3: 5 D4: 6 E5: 7

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 correctly. 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 .

Correct answer is D

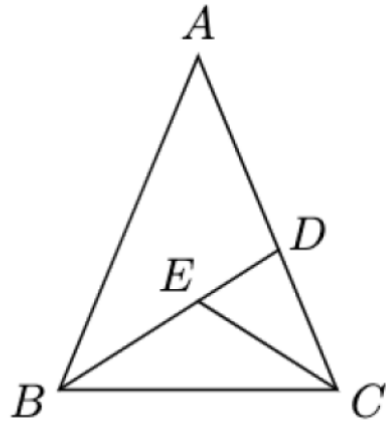
Responses from LVMs



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

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



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 $\angle BAC$ in the isosceles triangle ABC with given conditions, follow these steps: 1. Since $AB = AC$, triangle ABC is isosceles, which implies that $\angle CAB = \angle ACB$. 2. The point D divides the AB side in half, so $AD = DB$. Also, $CD = CE$ and $EC = EB$ as provided. Let's denote: - $\angle BAC = 2\alpha$ (since $AB = AC$, so $\angle BAC$ is twice the angle we are trying to find) - $\angle ABC = \angle ACB = \beta$ (since triangle ABC is isosceles) From the triangle properties, we can establish: [$2\alpha + 2\beta = 180$ degrees] [$\alpha + \beta = 90$ 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 $\angle ADE = \angle EDA = x$ - Let $\angle CDE = \angle CED = y$ Given each triangle is isosceles, sum of angles in triangle ADE and CDE : [$2x + \alpha = 180$ 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: [$\alpha + \beta + \alpha = 180$ 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 $\angle BAC$: [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 ($\angle BAC = 36$ degrees Option E5 solved value : Result conclusion, E5: 36.

What if LVLMs competed against Children on MK Olympiads?

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 (\downarrow) / percentile (\uparrow) ranking of LVLMs against children’s performance on MK 2024 Olympiad based on the test scores computed from the model response.

Analysis and Conclusions

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 better at higher grades, which is counter intuitive.

iii) Further, while there is a substantial gap between the best of LVLMs and the worst

Analysis and Conclusions

- AI models may not really be reasoning in the ways that humans do.
- Our analysis suggests signs that similarity to the large mass of training examples is perhaps driving AI performance
- **Human reasoning is based on a different set of core competencies than of AI models**

Thank you!

See the project page for details:
<https://smardataset.github.io/smart840>