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

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

https://smartdataset.github.io/smart840

MITSUBISHI ELECTRIC RESEARCH LABORATORIES, INC

Math Kangaroo USA since 1998 — International Competition in Mathematics

Massachusetts Institute of Technology

NEURAL INFORMATION PROCESSING SYSTEMS

1. AI vs Human Cognition: Key Questions

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.

1. Are the current large AI models indeed capable of generalized problem solving as humans do?
2. Can they perform well on tasks that need broad skills?
3. Humans learn over the years through cumulative knowledge gathering. Do AI models demonstrate such accumulation of knowledge?
4. Do AI models and humans have similar core competencies?
5. How correlated are their reasoning and problem-solving abilities?

2. Approach

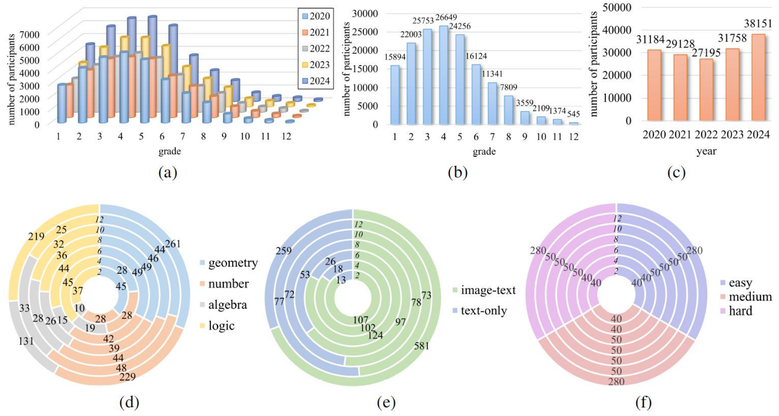
Compare human and AI on tasks that allow direct one-to-one comparison.

Our idea: To **analyze** LVLMs capabilities in mathematical and algorithmic reasoning using problems from Mathematical Olympiads with high human participation and compare their performances **directly** to that of human performance on the corresponding problems.

3. Math Kangaroo Olympiad & SMART-840 Dataset

- We consider problems from the **Math Kangaroo (MK) Olympiad**
  - A popular international math competition targeted at children from grades 1-12
  - Each exam tests children's deeper mathematical abilities using multiple choice vision-and-language 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 children's performances for each of these exams.

4. SMART-840 Dataset: Statistics & Examples

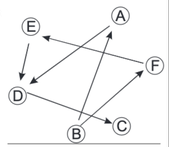


Statistics of human participation in MK exams and the distributions of puzzle attributes in the SMART-840 dataset.

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

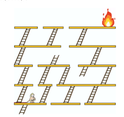


Year 2020, Grade-1 & 2, Difficulty: hard

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

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?

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

Answer Options:

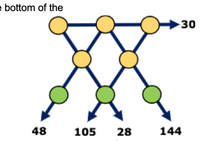
A	3
B	4
C	5
D	6
E	10

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

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

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

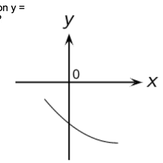
1		6	3
	2	2	8
	7		4
		7	

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

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

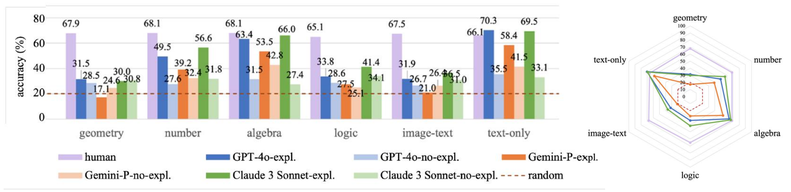


Year 2020, Grade-11&12, Difficulty: hard

5. AI vs Humans: Grade-level Performance

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

6. AI vs Humans: Category-wise Performance



7. AI vs Humans: Problem Solving Correlation

	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

8. 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