# **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

### **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 <u>need broad skills</u>?
- 3. Humans learn over the years through <u>cumulative knowledge</u> gathering. Do AI models demonstrate such accumulation of knowledge?
- 4. Do AI models and humans have similar core competencies?
- 5. How <u>correlated</u> are their reasoning and problem-solving abilities?

### 2. Approach

Compare humans and AI on tasks that allow direct 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 <u>children</u> from grades 1-12.
  - Each exam tests children's <u>deeper mathematical abilities</u> 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.



Massachusetts **Institute of Technology** 





#### 5. Al vs Humans: Grade-level Performance

Grade Model	1	2	3	4	5	6	7	8	9	10	11	12
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
Random	20.1		20.2		20.1		20.2		20.3		20.1	
GPT-40	41.6 (7.1)		38.6 (1.7)		35.1 (0.8)		47.1 (0.8)		41.3 (2.0)		50 (4.0)	
GPT-40 (M)	42.5		36.7		36.0		46.7		43.3		50.0	
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)	
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)	
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)	
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)	
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)	
XGEN-MM-Phi3-v1 (5B)	7.:	5	9	.1	5	.3	8	.0	10	).0	8	.0
InternVL-Chat-V1.2 (40B)	L-Chat-V1.2 (40B) 16.7		25		17.3		14.6		15.3		16.7	
InternLM-XComposer2 (7B)	M-XComposer2 (7B) 22.5		14.2		18.6		24.2		18.1		16.9	
LlaVa-NEXT (34B)		.0	9.0		20.1		14.6		18.7		16.0	

What does this mean? Al performs better in higher-grade questions than those of lower-grade

#### 6. Al vs Humans: Category-wise Performance



#### What does this mean? Asking AI to explain the answer improves its performance!

## 7. Al vs Humans: Problem Solving Correlation

Model \ Grade	1	2	3	4	5	6	7	8	9	10	11	12
GPT-40	0.14	0.16	0.15	0.17	-0.09	-0.05	0.12	0.13	0.22	0.22	0.20	0.2
Gemini-P	0.23	0.27	-0.05	-0.06	0.01	-0.01	0.05	0.06	0.21	0.19	0.20	0.1
Claude-3	0.11	0.13	0.09	0.11	0.08	0.06	0.14	0.15	0.16	0.16	0.25	0.1
GPT-40	-0.07	-0.15	0.07	-0.01	0.07	-0.01	-0.09	-0.08	-0.14	-0.18	-0.11	-0.1
Gemini-P	-0.05	-0.25	-0.04	-0.05	-0.01	-0.01	0.01	0.03	-0.18	-0.18	-0.15	-0.1
Claude-3	-0.02	-0.14	0.17	0.06	-0.04	-0.09	-0.07	-0.09	-0.16	-0.11	-0.09	-0.1
GPT-40	-0.08	-0.12	-0.14	-0.10	0.03	-0.03	0.08	0.03	-0.09	-0.07	-0.17	-0.0
Gemini-P	-0.06	-0.17	-0.06	-0.06	-0.03	-0.06	0.03	0.03	-0.20	-0.12	-0.27	-0.1
Claude-3	0.14	0.10	-0.07	-0.07	-0.04	-0.01	-0.01	-0.07	-0.09	-0.07	-0.16	-0.1
GPT-40	-0.04	-0.04	-0.02	-0.02	-0.00	-0.00	0.08	0.08	0.13	0.13	0.15	0.1
Gemini-P	0.05	0.05	-0.07	-0.07	0.00	0.00	0.02	0.02	0.27	0.27	0.30	0.3
Claude-3	-0.10	-0.10	-0.02	-0.02	0.00	0.00	0.15	0.15	0.18	0.18	0.30	0.3
GPT-40	-0.18	-0.18	-0.15	-0.15	0.10	0.10	-0.14	-0.14	-0.23	-0.23	-0.24	-0.2
Gemini-P	-0.26	-0.26	0.03	0.03	-0.01	-0.01	-0.08	-0.08	-0.23	-0.23	-0.19	-0.1
Claude-3	-0.12	-0.12	-0.06	-0.06	-0.02	-0.02	-0.15	-0.15	-0.18	-0.18	-0.24	-0.2
	Model \ Grade GPT-40 Gemini-P Claude-3 GPT-40 Gemini-P Claude-3 GPT-40 Gemini-P Claude-3 GPT-40 Gemini-P Claude-3 GPT-40 Gemini-P Claude-3	Model \ Grade 1   GPT-40 0.14   Gemini-P 0.23   Claude-3 0.11   GPT-40 -0.07   Gemini-P -0.05   Claude-3 -0.02   GPT-40 -0.08   GPT-40 -0.08   GPT-40 -0.08   Gemini-P -0.06   Claude-3 0.14   GPT-40 -0.08   Gemini-P -0.06   Claude-3 0.14   GPT-40 -0.04   Gemini-P 0.05   Claude-3 -0.101   GPT-40 -0.18   Gemini-P -0.18   Gemini-P -0.26   Claude-3 -0.12	Model \ Grade12GPT-400.140.16Gemini-P0.230.27Claude-30.110.13GPT-40-0.07-0.15Gemini-P-0.05-0.25Claude-3-0.02-0.14GPT-40-0.08-0.12Gemini-P-0.06-0.17Claude-30.140.10GPT-40-0.04-0.04GPT-40-0.040.05GPT-40-0.04-0.04GPT-40-0.10-0.10GPT-40-0.10-0.10GPT-40-0.18-0.18GPT-40-0.18-0.18GPT-40-0.12-0.26Claude-3-0.12-0.12	Model \ Grade123GPT-400.140.160.15Gemini-P0.230.27-0.05Claude-30.110.130.09GPT-40-0.07-0.150.07Gemini-P-0.05-0.25-0.04Claude-3-0.02-0.140.17GPT-40-0.08-0.12-0.14GPT-40-0.08-0.12-0.14GPT-40-0.06-0.17-0.06Claude-30.140.10-0.07GPT-40-0.04-0.04-0.02GPT-40-0.04-0.04-0.02Gemini-P0.050.05-0.07Claude-3-0.18-0.18-0.15GPT-40-0.18-0.18-0.15GPT-40-0.18-0.260.03Claude-3-0.12-0.0260.03Claude-3-0.12-0.12-0.06	Model \ Grade1234GPT-400.140.160.150.17Gemini-P0.230.27-0.05-0.06Claude-30.110.130.090.11GPT-40-0.07-0.150.07-0.01Gemini-P-0.05-0.25-0.04-0.05Claude-3-0.02-0.140.170.06GPT-40-0.08-0.12-0.14-0.10Gemini-P-0.06-0.17-0.06-0.07GPT-40-0.04-0.01-0.07-0.07GPT-40-0.04-0.04-0.02-0.02GPT-40-0.04-0.04-0.02-0.02GPT-40-0.10-0.05-0.07-0.07Claude-3-0.10-0.10-0.02-0.02GPT-40-0.18-0.18-0.15-0.15Gemini-P-0.26-0.260.030.03GPT-40-0.18-0.26-0.260.03GPT-40-0.18-0.26-0.260.03GPT-40-0.18-0.26-0.260.03Gemini-P-0.26-0.260.030.03Gemini-P-0.26-0.260.030.03Gemini-P-0.26-0.260.030.03Gemini-P-0.26-0.260.030.03Gemini-P-0.26-0.260.030.03Gemini-P-0.26-0.260.06-0.06	Model \ Grade12345GPT-400.140.160.150.17-0.09Gemini-P0.230.27-0.05-0.060.01Claude-30.110.130.090.110.08GPT-40-0.07-0.150.07-0.010.07Gemini-P-0.05-0.25-0.04-0.05-0.01Claude-3-0.02-0.140.170.06-0.04GPT-40-0.08-0.12-0.14-0.100.03Gemini-P-0.06-0.17-0.06-0.04-0.03Gemini-P0.06-0.17-0.06-0.07-0.04GPT-40-0.04-0.04-0.07-0.07-0.04GPT-40-0.04-0.04-0.02-0.02-0.00Gemini-P0.050.05-0.07-0.070.00Gemini-P-0.10-0.10-0.02-0.020.00GPT-40-0.18-0.18-0.15-0.150.10GPT-40-0.18-0.18-0.15-0.150.10GPT-40-0.26-0.260.030.03-0.01GPT-40-0.18-0.18-0.15-0.150.10Gemini-P-0.26-0.260.030.03-0.01Gemini-P-0.26-0.260.030.03-0.01Gemini-P-0.26-0.260.030.03-0.01Gemini-P-0.26-0.260.030.03-0.01 <td>Model \ Grade123456GPT-400.140.160.150.17-0.09-0.05Gemini-P0.230.27-0.05-0.060.01-0.01Claude-30.110.130.090.110.080.06GPT-40-0.07-0.150.07-0.010.07-0.01Gemini-P-0.05-0.25-0.04-0.05-0.01-0.01Claude-3-0.02-0.140.170.06-0.04-0.09GPT-40-0.08-0.12-0.14-0.100.03-0.03GPT-40-0.08-0.12-0.14-0.100.03-0.01GPT-40-0.08-0.17-0.06-0.06-0.03-0.06Gemini-P-0.06-0.17-0.06-0.07-0.04-0.01GPT-40-0.140.10-0.07-0.07-0.00-0.00Gemini-P0.050.05-0.07-0.070.000.00Gemini-P-0.18-0.18-0.15-0.150.100.10GPT-40-0.18-0.16-0.05-0.01-0.01-0.01GPT-40-0.18-0.15-0.150.100.10Gemini-P-0.26-0.260.030.03-0.01-0.01GPT-40-0.12-0.12-0.16-0.06-0.02-0.01GPT-40-0.18-0.18-0.15-0.150.10-0.11Gemini-P-0.26<!--</td--><td>Model \ Grade1234567GPT-400.140.160.150.17-0.09-0.050.12Gemini-P0.230.27-0.05-0.060.01-0.010.05Claude-30.110.130.090.110.080.060.14GPT-40-0.07-0.150.07-0.010.07-0.01-0.09Gemini-P-0.05-0.25-0.04-0.05-0.01-0.010.01Claude-3-0.02-0.140.170.06-0.04-0.09-0.07GPT-40-0.08-0.12-0.14-0.100.03-0.030.08Gemini-P-0.06-0.17-0.06-0.06-0.03-0.060.03GPT-40-0.04-0.04-0.07-0.07-0.04-0.01-0.01GPT-40-0.04-0.04-0.07-0.07-0.04-0.01-0.01GPT-40-0.04-0.04-0.02-0.02-0.000.000.02Gemini-P0.05-0.05-0.07-0.000.000.02GPT-40-0.18-0.18-0.15-0.150.100.10-0.14Gemini-P0.05-0.06-0.050.000.000.01-0.08Gemini-P-0.26-0.260.030.03-0.01-0.01-0.08Gemini-P-0.26-0.260.06-0.06-0.02-0.02-0.02-0.15<td>Model \ Grade12345678GPT-400.140.160.150.17-0.09-0.050.120.13Gemini-P0.230.27-0.05-0.060.01-0.010.050.06Claude-30.110.130.090.110.080.060.140.15GPT-40-0.07-0.150.07-0.010.07-0.01-0.09-0.08Gemini-P-0.02-0.140.170.06-0.04-0.09-0.07-0.09Claude-3-0.08-0.12-0.14-0.100.03-0.030.080.03GPT-40-0.08-0.12-0.14-0.100.03-0.040.030.03GPT-40-0.08-0.12-0.14-0.100.03-0.060.030.03Gemini-P-0.06-0.17-0.06-0.06-0.04-0.01-0.01-0.07GPT-40-0.04-0.04-0.02-0.02-0.00-0.000.030.03Gemini-P0.050.05-0.07-0.070.000.000.020.02GPT-40-0.18-0.18-0.15-0.150.100.10-0.14-0.14Gemini-P0.05-0.26-0.02-0.020.000.000.150.15GPT-40-0.18-0.18-0.15-0.150.100.10-0.14-0.14Gemini-P-0.16-0.18-0.16</td><td>Model \ Grade   1   2   3   4   5   6   7   8   9     GPT-40   0.14   0.16   0.15   0.17   -0.09   -0.05   0.12   0.13   0.22     Gemini-P   0.23   0.27   -0.05   -0.06   0.01   -0.01   0.05   0.06   0.11     Claude-3   0.11   0.13   0.09   0.11   0.08   0.06   0.14   0.15   0.16     GPT-40   -0.07   -0.15   0.07   -0.01   0.07   -0.01   0.07   -0.01   0.01   0.09   -0.08   -0.14     Gemini-P   -0.05   -0.25   -0.04   -0.05   -0.01   -0.01   0.01   0.03   -0.07   -0.08   -0.18     Gemini-P   -0.02   -0.14   0.10   0.03   -0.03   0.08   0.03   -0.09   -0.07   -0.09   -0.07   -0.09   -0.01   -0.07   -0.09   -0.01   -0.01   -0.01   -0.01   -</td><td>Model \ Grade12345678910GPT-400.140.160.150.17-0.09-0.050.120.130.220.22Gemini-P0.230.27-0.05-0.060.01-0.010.050.060.210.19Claude-30.110.130.090.110.080.060.140.150.160.16GPT-40-0.07-0.150.07-0.010.07-0.010.010.03-0.08-0.18Gemini-P-0.05-0.25-0.04-0.05-0.01-0.010.010.03-0.18-0.18Claude-3-0.02-0.140.170.06-0.04-0.09-0.07-0.09-0.16-0.11GPT-40-0.08-0.12-0.14-0.100.03-0.030.080.03-0.09-0.07Gemini-P-0.06-0.17-0.06-0.07-0.04-0.09-0.01-0.07-0.09-0.07Gemini-P-0.06-0.07-0.07-0.04-0.01-0.01-0.07-0.09-0.01-0.07-0.09-0.07GPT-40-0.04-0.02-0.02-0.00-0.000.080.080.03-0.130.130.13Gemini-P-0.05-0.07-0.07-0.000.000.020.020.270.27Glaude-3-0.18-0.18-0.15-0.150.100.00<td< td=""><td>Model \ Grade   1   2   3   4   5   6   7   8   9   10   11     GPT-40   0.14   0.16   0.15   0.17   -0.09   -0.05   0.12   0.13   0.22   0.22   0.20     Gemini-P   0.23   0.27   -0.05   -0.06   0.01   -0.01   0.05   0.06   0.21   0.19   0.20     Claude-3   0.11   0.13   0.09   0.11   0.08   0.06   0.14   0.15   0.16   0.16   0.21   0.19   0.20     GPT-40   -0.07   -0.15   0.07   -0.01   0.07   -0.01   0.03   0.01   0.03   0.01   0.03</td></td<></td></td></td>	Model \ Grade123456GPT-400.140.160.150.17-0.09-0.05Gemini-P0.230.27-0.05-0.060.01-0.01Claude-30.110.130.090.110.080.06GPT-40-0.07-0.150.07-0.010.07-0.01Gemini-P-0.05-0.25-0.04-0.05-0.01-0.01Claude-3-0.02-0.140.170.06-0.04-0.09GPT-40-0.08-0.12-0.14-0.100.03-0.03GPT-40-0.08-0.12-0.14-0.100.03-0.01GPT-40-0.08-0.17-0.06-0.06-0.03-0.06Gemini-P-0.06-0.17-0.06-0.07-0.04-0.01GPT-40-0.140.10-0.07-0.07-0.00-0.00Gemini-P0.050.05-0.07-0.070.000.00Gemini-P-0.18-0.18-0.15-0.150.100.10GPT-40-0.18-0.16-0.05-0.01-0.01-0.01GPT-40-0.18-0.15-0.150.100.10Gemini-P-0.26-0.260.030.03-0.01-0.01GPT-40-0.12-0.12-0.16-0.06-0.02-0.01GPT-40-0.18-0.18-0.15-0.150.10-0.11Gemini-P-0.26 </td <td>Model \ Grade1234567GPT-400.140.160.150.17-0.09-0.050.12Gemini-P0.230.27-0.05-0.060.01-0.010.05Claude-30.110.130.090.110.080.060.14GPT-40-0.07-0.150.07-0.010.07-0.01-0.09Gemini-P-0.05-0.25-0.04-0.05-0.01-0.010.01Claude-3-0.02-0.140.170.06-0.04-0.09-0.07GPT-40-0.08-0.12-0.14-0.100.03-0.030.08Gemini-P-0.06-0.17-0.06-0.06-0.03-0.060.03GPT-40-0.04-0.04-0.07-0.07-0.04-0.01-0.01GPT-40-0.04-0.04-0.07-0.07-0.04-0.01-0.01GPT-40-0.04-0.04-0.02-0.02-0.000.000.02Gemini-P0.05-0.05-0.07-0.000.000.02GPT-40-0.18-0.18-0.15-0.150.100.10-0.14Gemini-P0.05-0.06-0.050.000.000.01-0.08Gemini-P-0.26-0.260.030.03-0.01-0.01-0.08Gemini-P-0.26-0.260.06-0.06-0.02-0.02-0.02-0.15<td>Model \ Grade12345678GPT-400.140.160.150.17-0.09-0.050.120.13Gemini-P0.230.27-0.05-0.060.01-0.010.050.06Claude-30.110.130.090.110.080.060.140.15GPT-40-0.07-0.150.07-0.010.07-0.01-0.09-0.08Gemini-P-0.02-0.140.170.06-0.04-0.09-0.07-0.09Claude-3-0.08-0.12-0.14-0.100.03-0.030.080.03GPT-40-0.08-0.12-0.14-0.100.03-0.040.030.03GPT-40-0.08-0.12-0.14-0.100.03-0.060.030.03Gemini-P-0.06-0.17-0.06-0.06-0.04-0.01-0.01-0.07GPT-40-0.04-0.04-0.02-0.02-0.00-0.000.030.03Gemini-P0.050.05-0.07-0.070.000.000.020.02GPT-40-0.18-0.18-0.15-0.150.100.10-0.14-0.14Gemini-P0.05-0.26-0.02-0.020.000.000.150.15GPT-40-0.18-0.18-0.15-0.150.100.10-0.14-0.14Gemini-P-0.16-0.18-0.16</td><td>Model \ Grade   1   2   3   4   5   6   7   8   9     GPT-40   0.14   0.16   0.15   0.17   -0.09   -0.05   0.12   0.13   0.22     Gemini-P   0.23   0.27   -0.05   -0.06   0.01   -0.01   0.05   0.06   0.11     Claude-3   0.11   0.13   0.09   0.11   0.08   0.06   0.14   0.15   0.16     GPT-40   -0.07   -0.15   0.07   -0.01   0.07   -0.01   0.07   -0.01   0.01   0.09   -0.08   -0.14     Gemini-P   -0.05   -0.25   -0.04   -0.05   -0.01   -0.01   0.01   0.03   -0.07   -0.08   -0.18     Gemini-P   -0.02   -0.14   0.10   0.03   -0.03   0.08   0.03   -0.09   -0.07   -0.09   -0.07   -0.09   -0.01   -0.07   -0.09   -0.01   -0.01   -0.01   -0.01   -</td><td>Model \ Grade12345678910GPT-400.140.160.150.17-0.09-0.050.120.130.220.22Gemini-P0.230.27-0.05-0.060.01-0.010.050.060.210.19Claude-30.110.130.090.110.080.060.140.150.160.16GPT-40-0.07-0.150.07-0.010.07-0.010.010.03-0.08-0.18Gemini-P-0.05-0.25-0.04-0.05-0.01-0.010.010.03-0.18-0.18Claude-3-0.02-0.140.170.06-0.04-0.09-0.07-0.09-0.16-0.11GPT-40-0.08-0.12-0.14-0.100.03-0.030.080.03-0.09-0.07Gemini-P-0.06-0.17-0.06-0.07-0.04-0.09-0.01-0.07-0.09-0.07Gemini-P-0.06-0.07-0.07-0.04-0.01-0.01-0.07-0.09-0.01-0.07-0.09-0.07GPT-40-0.04-0.02-0.02-0.00-0.000.080.080.03-0.130.130.13Gemini-P-0.05-0.07-0.07-0.000.000.020.020.270.27Glaude-3-0.18-0.18-0.15-0.150.100.00<td< td=""><td>Model \ Grade   1   2   3   4   5   6   7   8   9   10   11     GPT-40   0.14   0.16   0.15   0.17   -0.09   -0.05   0.12   0.13   0.22   0.22   0.20     Gemini-P   0.23   0.27   -0.05   -0.06   0.01   -0.01   0.05   0.06   0.21   0.19   0.20     Claude-3   0.11   0.13   0.09   0.11   0.08   0.06   0.14   0.15   0.16   0.16   0.21   0.19   0.20     GPT-40   -0.07   -0.15   0.07   -0.01   0.07   -0.01   0.03   0.01   0.03   0.01   0.03</td></td<></td></td>	Model \ Grade1234567GPT-400.140.160.150.17-0.09-0.050.12Gemini-P0.230.27-0.05-0.060.01-0.010.05Claude-30.110.130.090.110.080.060.14GPT-40-0.07-0.150.07-0.010.07-0.01-0.09Gemini-P-0.05-0.25-0.04-0.05-0.01-0.010.01Claude-3-0.02-0.140.170.06-0.04-0.09-0.07GPT-40-0.08-0.12-0.14-0.100.03-0.030.08Gemini-P-0.06-0.17-0.06-0.06-0.03-0.060.03GPT-40-0.04-0.04-0.07-0.07-0.04-0.01-0.01GPT-40-0.04-0.04-0.07-0.07-0.04-0.01-0.01GPT-40-0.04-0.04-0.02-0.02-0.000.000.02Gemini-P0.05-0.05-0.07-0.000.000.02GPT-40-0.18-0.18-0.15-0.150.100.10-0.14Gemini-P0.05-0.06-0.050.000.000.01-0.08Gemini-P-0.26-0.260.030.03-0.01-0.01-0.08Gemini-P-0.26-0.260.06-0.06-0.02-0.02-0.02-0.15 <td>Model \ Grade12345678GPT-400.140.160.150.17-0.09-0.050.120.13Gemini-P0.230.27-0.05-0.060.01-0.010.050.06Claude-30.110.130.090.110.080.060.140.15GPT-40-0.07-0.150.07-0.010.07-0.01-0.09-0.08Gemini-P-0.02-0.140.170.06-0.04-0.09-0.07-0.09Claude-3-0.08-0.12-0.14-0.100.03-0.030.080.03GPT-40-0.08-0.12-0.14-0.100.03-0.040.030.03GPT-40-0.08-0.12-0.14-0.100.03-0.060.030.03Gemini-P-0.06-0.17-0.06-0.06-0.04-0.01-0.01-0.07GPT-40-0.04-0.04-0.02-0.02-0.00-0.000.030.03Gemini-P0.050.05-0.07-0.070.000.000.020.02GPT-40-0.18-0.18-0.15-0.150.100.10-0.14-0.14Gemini-P0.05-0.26-0.02-0.020.000.000.150.15GPT-40-0.18-0.18-0.15-0.150.100.10-0.14-0.14Gemini-P-0.16-0.18-0.16</td> <td>Model \ Grade   1   2   3   4   5   6   7   8   9     GPT-40   0.14   0.16   0.15   0.17   -0.09   -0.05   0.12   0.13   0.22     Gemini-P   0.23   0.27   -0.05   -0.06   0.01   -0.01   0.05   0.06   0.11     Claude-3   0.11   0.13   0.09   0.11   0.08   0.06   0.14   0.15   0.16     GPT-40   -0.07   -0.15   0.07   -0.01   0.07   -0.01   0.07   -0.01   0.01   0.09   -0.08   -0.14     Gemini-P   -0.05   -0.25   -0.04   -0.05   -0.01   -0.01   0.01   0.03   -0.07   -0.08   -0.18     Gemini-P   -0.02   -0.14   0.10   0.03   -0.03   0.08   0.03   -0.09   -0.07   -0.09   -0.07   -0.09   -0.01   -0.07   -0.09   -0.01   -0.01   -0.01   -0.01   -</td> <td>Model \ Grade12345678910GPT-400.140.160.150.17-0.09-0.050.120.130.220.22Gemini-P0.230.27-0.05-0.060.01-0.010.050.060.210.19Claude-30.110.130.090.110.080.060.140.150.160.16GPT-40-0.07-0.150.07-0.010.07-0.010.010.03-0.08-0.18Gemini-P-0.05-0.25-0.04-0.05-0.01-0.010.010.03-0.18-0.18Claude-3-0.02-0.140.170.06-0.04-0.09-0.07-0.09-0.16-0.11GPT-40-0.08-0.12-0.14-0.100.03-0.030.080.03-0.09-0.07Gemini-P-0.06-0.17-0.06-0.07-0.04-0.09-0.01-0.07-0.09-0.07Gemini-P-0.06-0.07-0.07-0.04-0.01-0.01-0.07-0.09-0.01-0.07-0.09-0.07GPT-40-0.04-0.02-0.02-0.00-0.000.080.080.03-0.130.130.13Gemini-P-0.05-0.07-0.07-0.000.000.020.020.270.27Glaude-3-0.18-0.18-0.15-0.150.100.00<td< td=""><td>Model \ Grade   1   2   3   4   5   6   7   8   9   10   11     GPT-40   0.14   0.16   0.15   0.17   -0.09   -0.05   0.12   0.13   0.22   0.22   0.20     Gemini-P   0.23   0.27   -0.05   -0.06   0.01   -0.01   0.05   0.06   0.21   0.19   0.20     Claude-3   0.11   0.13   0.09   0.11   0.08   0.06   0.14   0.15   0.16   0.16   0.21   0.19   0.20     GPT-40   -0.07   -0.15   0.07   -0.01   0.07   -0.01   0.03   0.01   0.03   0.01   0.03</td></td<></td>	Model \ Grade12345678GPT-400.140.160.150.17-0.09-0.050.120.13Gemini-P0.230.27-0.05-0.060.01-0.010.050.06Claude-30.110.130.090.110.080.060.140.15GPT-40-0.07-0.150.07-0.010.07-0.01-0.09-0.08Gemini-P-0.02-0.140.170.06-0.04-0.09-0.07-0.09Claude-3-0.08-0.12-0.14-0.100.03-0.030.080.03GPT-40-0.08-0.12-0.14-0.100.03-0.040.030.03GPT-40-0.08-0.12-0.14-0.100.03-0.060.030.03Gemini-P-0.06-0.17-0.06-0.06-0.04-0.01-0.01-0.07GPT-40-0.04-0.04-0.02-0.02-0.00-0.000.030.03Gemini-P0.050.05-0.07-0.070.000.000.020.02GPT-40-0.18-0.18-0.15-0.150.100.10-0.14-0.14Gemini-P0.05-0.26-0.02-0.020.000.000.150.15GPT-40-0.18-0.18-0.15-0.150.100.10-0.14-0.14Gemini-P-0.16-0.18-0.16	Model \ Grade   1   2   3   4   5   6   7   8   9     GPT-40   0.14   0.16   0.15   0.17   -0.09   -0.05   0.12   0.13   0.22     Gemini-P   0.23   0.27   -0.05   -0.06   0.01   -0.01   0.05   0.06   0.11     Claude-3   0.11   0.13   0.09   0.11   0.08   0.06   0.14   0.15   0.16     GPT-40   -0.07   -0.15   0.07   -0.01   0.07   -0.01   0.07   -0.01   0.01   0.09   -0.08   -0.14     Gemini-P   -0.05   -0.25   -0.04   -0.05   -0.01   -0.01   0.01   0.03   -0.07   -0.08   -0.18     Gemini-P   -0.02   -0.14   0.10   0.03   -0.03   0.08   0.03   -0.09   -0.07   -0.09   -0.07   -0.09   -0.01   -0.07   -0.09   -0.01   -0.01   -0.01   -0.01   -	Model \ Grade12345678910GPT-400.140.160.150.17-0.09-0.050.120.130.220.22Gemini-P0.230.27-0.05-0.060.01-0.010.050.060.210.19Claude-30.110.130.090.110.080.060.140.150.160.16GPT-40-0.07-0.150.07-0.010.07-0.010.010.03-0.08-0.18Gemini-P-0.05-0.25-0.04-0.05-0.01-0.010.010.03-0.18-0.18Claude-3-0.02-0.140.170.06-0.04-0.09-0.07-0.09-0.16-0.11GPT-40-0.08-0.12-0.14-0.100.03-0.030.080.03-0.09-0.07Gemini-P-0.06-0.17-0.06-0.07-0.04-0.09-0.01-0.07-0.09-0.07Gemini-P-0.06-0.07-0.07-0.04-0.01-0.01-0.07-0.09-0.01-0.07-0.09-0.07GPT-40-0.04-0.02-0.02-0.00-0.000.080.080.03-0.130.130.13Gemini-P-0.05-0.07-0.07-0.000.000.020.020.270.27Glaude-3-0.18-0.18-0.15-0.150.100.00 <td< td=""><td>Model \ Grade   1   2   3   4   5   6   7   8   9   10   11     GPT-40   0.14   0.16   0.15   0.17   -0.09   -0.05   0.12   0.13   0.22   0.22   0.20     Gemini-P   0.23   0.27   -0.05   -0.06   0.01   -0.01   0.05   0.06   0.21   0.19   0.20     Claude-3   0.11   0.13   0.09   0.11   0.08   0.06   0.14   0.15   0.16   0.16   0.21   0.19   0.20     GPT-40   -0.07   -0.15   0.07   -0.01   0.07   -0.01   0.03   0.01   0.03   0.01   0.03</td></td<>	Model \ Grade   1   2   3   4   5   6   7   8   9   10   11     GPT-40   0.14   0.16   0.15   0.17   -0.09   -0.05   0.12   0.13   0.22   0.22   0.20     Gemini-P   0.23   0.27   -0.05   -0.06   0.01   -0.01   0.05   0.06   0.21   0.19   0.20     Claude-3   0.11   0.13   0.09   0.11   0.08   0.06   0.14   0.15   0.16   0.16   0.21   0.19   0.20     GPT-40   -0.07   -0.15   0.07   -0.01   0.07   -0.01   0.03   0.01   0.03   0.01   0.03

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 (score or points), Entropy-C: Correlation on the distribution of answer selections by humans

What does this mean? Human and AI performance have little to no correlations!

#### 8. Conclusions

- Al models may not 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.



