Are Deep Neural Networks SMARTer than Second Graders?





Anoop Cherian[1] Kuan-Chuan Peng[1] Suhas Lohit[1] Kevin Smith[2] Josh Tenenbaum[2]

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

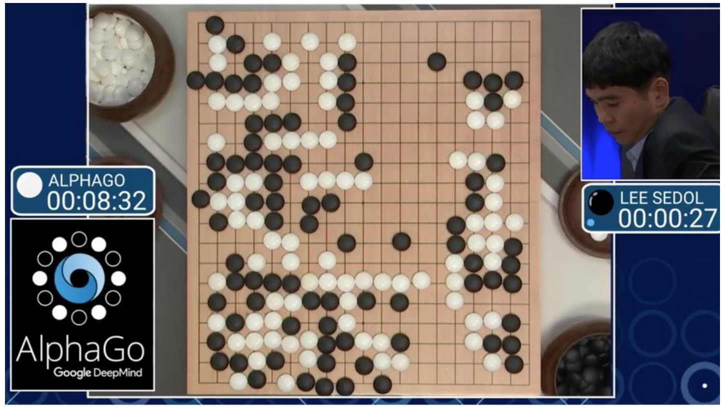
[2]Massachusetts Institute of Technology (MIT), Cambridge, MA

CVPR, 2023

Poster ID: WED-AM-248

In the recent times, …

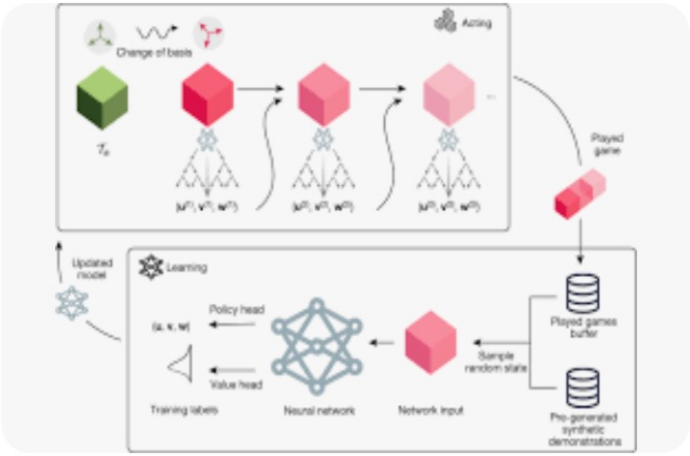
ChatGPT



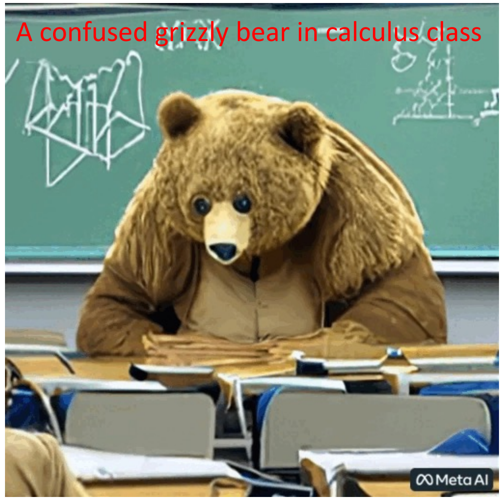
AlphaGo

Imagen

Solving University Level Math …"
Drori et al., PNAS, 2022



Neural Algorithmic Inference,
Fawzi et al., Nature, 2022



Make-a-Video

Are we there (yet) in achieving artificial general intelligence?

Prompt: “Mickey mouse goes for a vacation in Potato Land”
Created using MidJourney

Key Questions

1. How well do deep neural models perform on tasks that need broad skills?
2. Do they transfer knowledge to solve new problems?
3. How fluid are they in acquiring new skills?
4. How effective are they in the use of language for *algorithmic* reasoning?

Key Questions

1. How well do deep neural models perform on tasks that need broad skills?
2. Do they transfer knowledge to solve new problems?
3. How fluid are they in acquiring new skills? and
4. How effective are they in the use of language for *algorithmic* reasoning?

How should we go about answering the above questions?
Where should we start?

Key Questions

1. How well do deep neural models perform on tasks that need broad skills?
2. Do they transfer knowledge to solve new problems?
3. How fluid are they in acquiring new skills? and
4. How effective are they in the use of language for *algorithmic* reasoning?

How should we go about answering the above questions?
Where should we start?

Are state-of-the-art deep neural networks capable of emulating the thinking process of even young children?

Our Contributions

Are state-of-the-art deep neural networks capable of emulating the thinking process of even young children?

We propose **SMART: Simple Multi-modal Algorithmic Reasoning Task**

The task is to design deep neural networks that have foundational skills to effectively *analyze, interpret, and solve* simple algorithmic reasoning puzzles and generalize to new problems.

Our Contributions

Are state-of-the-art deep neural networks capable of emulating the thinking process of even young children?

We propose **SMART: Simple Multi-modal Algorithmic Reasoning Task**

- The task is to design deep neural networks that have foundational skills to effectively *analyze, interpret, and solve* simple algorithmic reasoning puzzles and generalize to new problems.

- We propose the **SMART-101 dataset** consisting of:
  101 distinct vision & language children's puzzles.

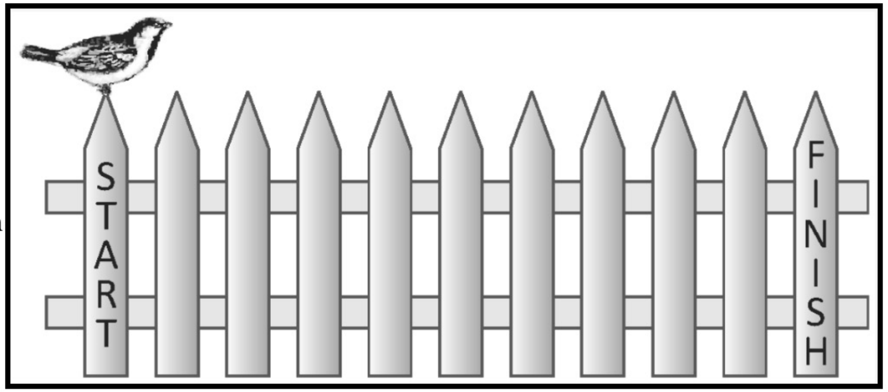
- Using SMART-101 dataset, we show that current large language models and visual processing pipelines do not generalize well to tasks that need broad reasoning skills.

SMART: Simple Multi-modal Algorithmic Reasoning Task

- The puzzles in the SMART-101 dataset are taken from Math Kangaroo (MK) USA Olympiad
  - ➢ An annual math competition for school kids from 1—12th grade.
- We selected puzzles from 2012-2021 competitions designed for children of first and second grades (6 – 8 age group)
- We used only “simple” puzzles from the MK competitions.

Math Kangaroo USA
since 1998
International Competition in Mathematics

SMART: Simple Multi-modal Algorithmic Reasoning Task



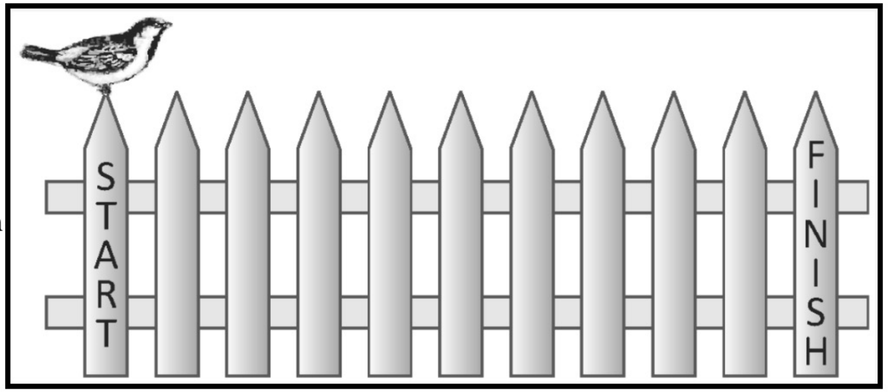
Question: *Sparoow Jack jumps on a fence from one post to another. Each jump takes him 1 second. He makes 4 jumps ahead and then 1 jump back. Then he again makes 4 jumps ahead and 1 back, and so on. In how many seconds does Jack get from start to finish?*

Answer Options: A: 10 B: 11 C: 12 D: 13 E: 14

Original MK Puzzle

Programmatic Augmentation of Puzzles

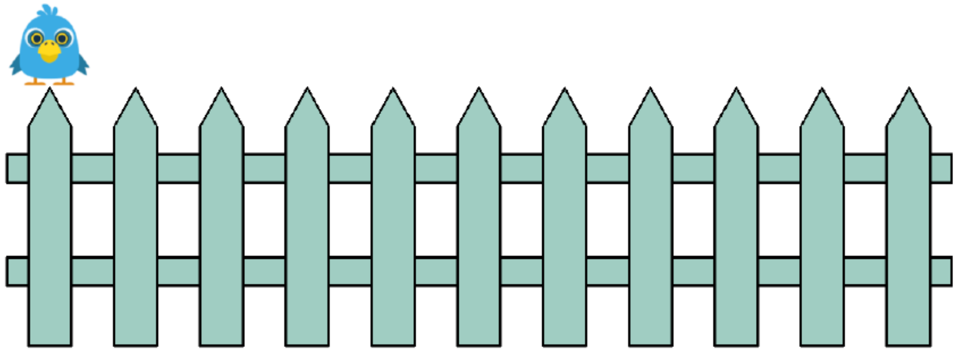
Original MK Puzzle



Question: *Sparoow Jack jumps on a fence from one post to another. Each jump takes him 1 second. He makes 4 jumps ahead and then 1 jump back. Then he again makes 4 jumps ahead and 1 back, and so on. In how many seconds does Jack get from start to finish?*

Answer Options: A: 10 B: 11 C: 12 D: 13 E: 14

Programmatically Created Puzzle



Question: *Bird Bobbie jumps on a fence from the post on the left end to the other end. Each jump takes him 4 seconds. He makes 4 jumps ahead and then 1 jump back. Then he again makes 4 jumps ahead and 1 jump back, and so on. In how many seconds can Bobbie get from one end to the other end?*

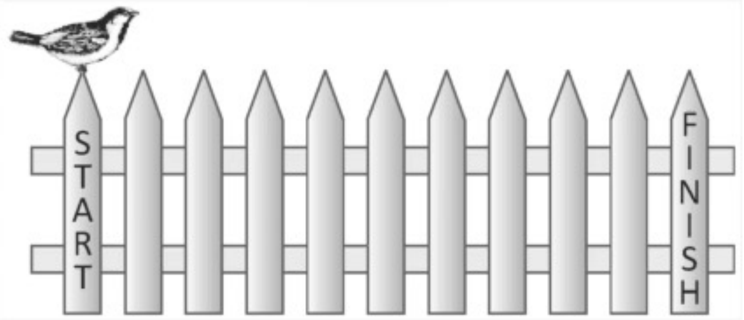
Answer Options: A: 64 B: 48 C: 56 D: 68 E: 72

An example puzzle from the SMART-101 dataset produced using our programmatic augmentation method.

Programmatic Puzzle Generation/Augmentation

MK Question:
Sparoow Jack jumps on a fence from one post to another. Each jump takes him 1 second. He makes 4 jumps ahead and then 1 jump back. Then he again makes 4 jumps ahead and 1 back, and so on. In how many seconds does Jack get from start to finish?

MK Puzzle

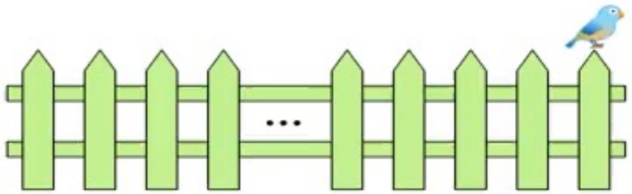


Augmented Question:

A bird jumps from the post on one end to the other end on a fence, and there are 27 posts in total. He needs 1 second for each jump. He makes 8 jumps ahead and then 5 jumps back. Then he again makes 8 jumps ahead and 5 jumps back, and so on. In how many seconds can the bird get from one end to the other end?

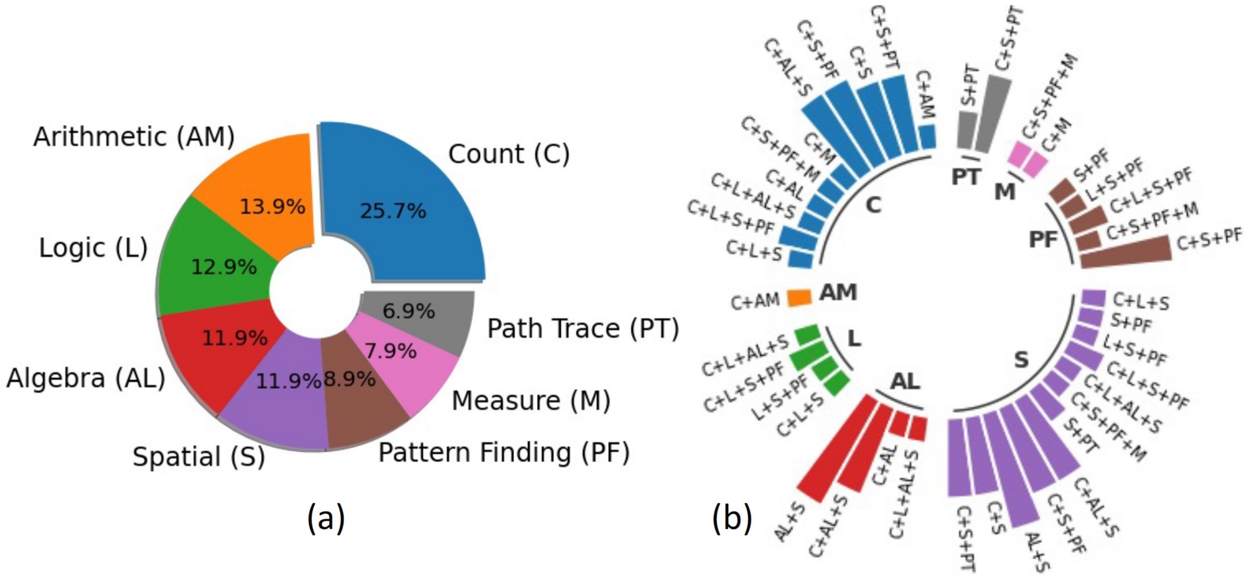
Options:
A: 90 B: 86 C: 83 D: 84 E: 87

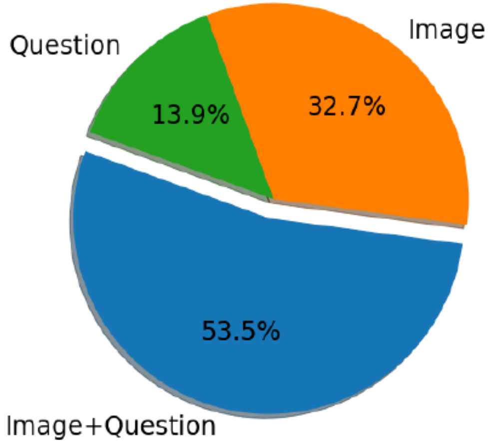
Augmented Puzzle

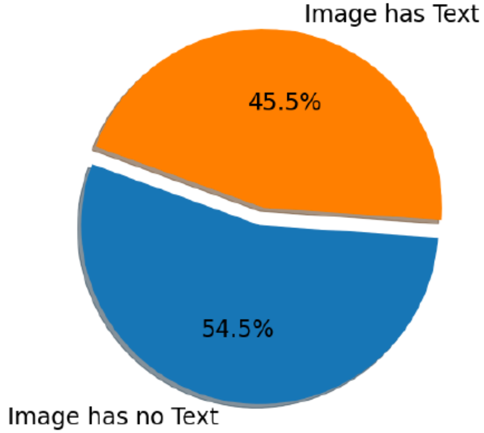


Programmatic Puzzle Generation/Augmentation

Dataset	Involve language	Dataset size	Task nature
Bongard-LOGO [45]	✗	12K	few-shot concepts, abstract shape reasoning
Bongard-HOI [32]	✗	53K	few-shot concepts, human-object interaction
ARC [12]	✗	800	generate image based on abstract rules
Machine Number Sense [66]	✗	280K	solving arithmetic problems
RAVEN [64]	✗	70K	finding next image in sequence
Image riddles [5]	✓(fixed question)	3333	finding common linguistic descriptions
VLQA [54]	✓(variable questions)	9267	spatio-temporal reasoning, info lookup, mathematical, logical, causality, analogy, *etc.*
PororoQA [34]	✓(variable questions)	8913	reason from cartoon videos about action, person, abstract, detail, location, *etc.*
CLEVR [33]	✓(variable questions)	100K	exist, count, query attributes, compare integers/attribute
SMART-101 (ours)	✓(variable questions)	200K	8 predominant algorithmic skills and their compositions (see Figure 2)

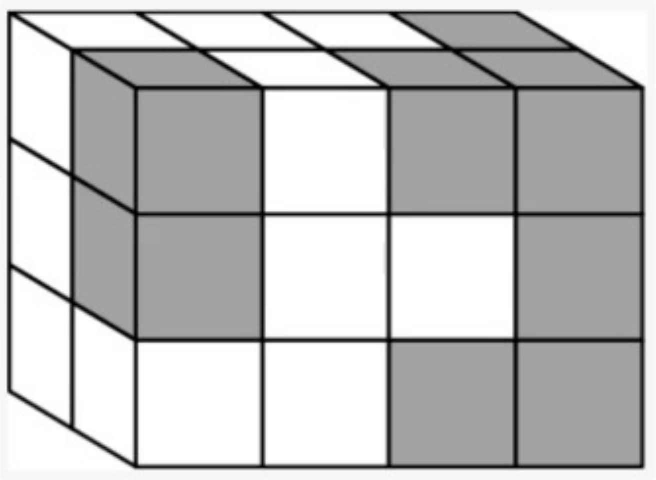






Augmented Puzzles: Spatial Reasoning Class

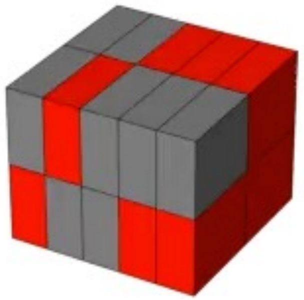
MK Puzzle



MK Question:
Mary made a shape using some white cubes and 14 gray cubes. many of these gray cubes cannot be seen in the picture?

Augmented Question:

Augmented Puzzle



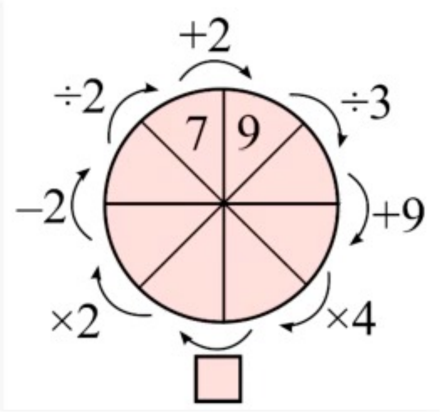
Melissa created a shape using some red blocks and 12 gray blocks.
How many of these gray blocks are not visible in the image?

Options:
A: 4
B: 0
C: 7
D: 9
E: 5

Augmented Puzzles: Arithmetic Class

MK Question:
What should you put in the square on the bottom to get a correct diagram?

MK Puzzle

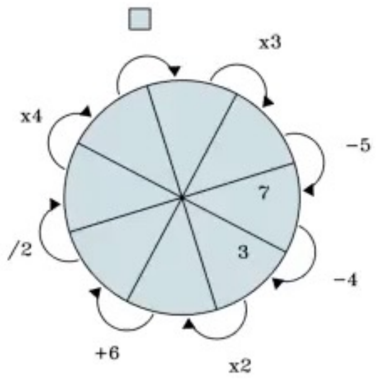


Augmented Question:

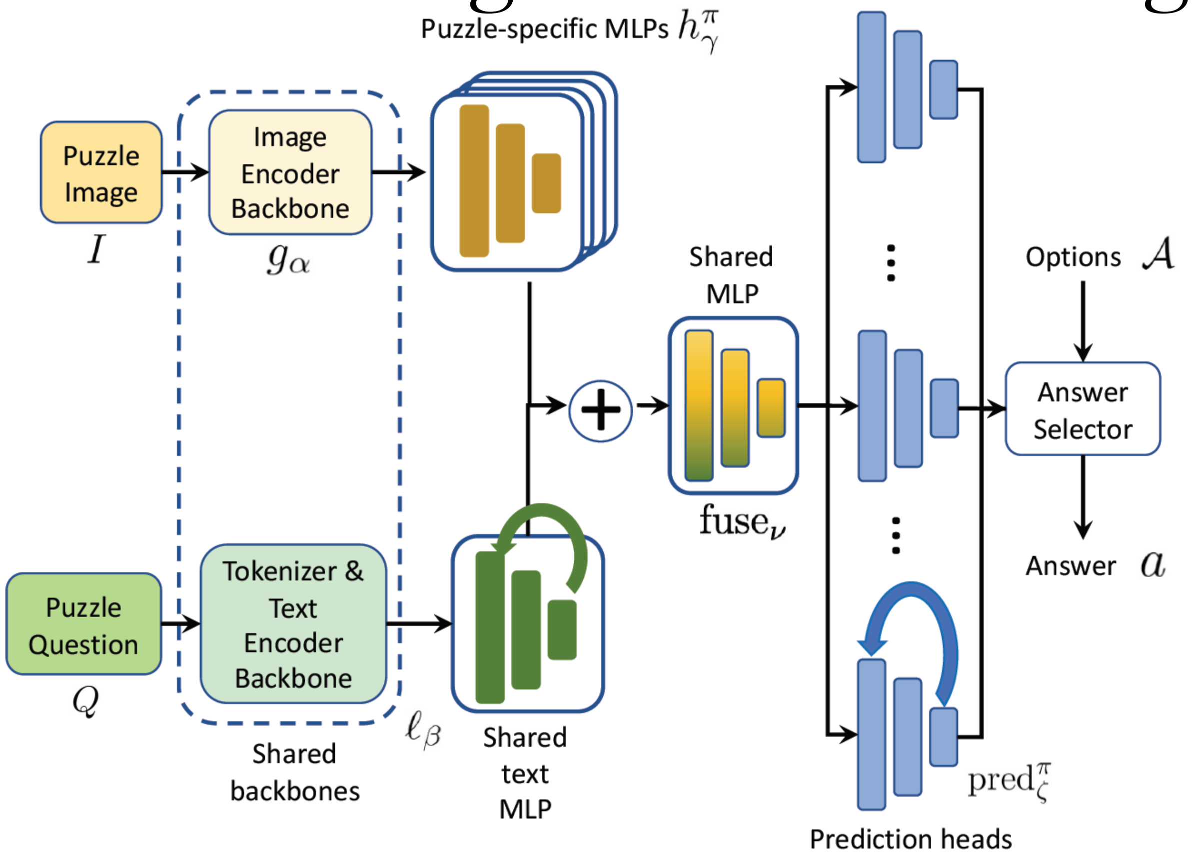
What should you put in the square to get a correct diagram?

Options:
A: -6
B: +4
C: +9
D: /6
E: x5

Augmented Puzzle



SMART: Meta-Learning based Reasoning Baseline



Training

$$\min_{\Theta} \mathbb{E}_{\pi\sim\mathcal{R}} \mathbb{E}_{(I,Q,a)\sim\mathcal{P}_\pi} \text{loss}_\pi\left(f_\theta^\pi(I,Q) - a\right)$$

Inference

$$\hat{a} = \arg\max_{\alpha\in\mathcal{A}} \text{sim}_\pi\left(f_\theta^\pi(I,Q), \alpha\right)$$

Experiments and Results

SMART101 - Data Splits

1. **Puzzle Split (PS) : Extreme Generalization**
   - Split 101 puzzles to 77 for training, 3 for validation, and 21 for test.
   - We use 300 instances in the 21 puzzles in the test set to report performance
2. **Few-shot Split (FS) : m-shot Generalization**
   - Use m=10 instances of the 21 puzzles during training, along with the instances in the 77 puzzle-training set in PS
3. **Instance Split (IS): Supervised Learning**
   - Supervised training setting
   - Use 80% of all instances from every puzzle for training, 5% used in validation, and 15% for test.
4. **Answer Split (AS) : Answer Generalization**
   - We remove the median answer from all correct answers for every puzzle instance from training set and test only on these removed questions.

SMART101 – Evaluation Metrics

S_{acc}: Solution Accuracy

- Out of the discretized answer range, what is the accuracy in selecting the correct answer?

O_{acc}: Option Selection Accuracy

- Among five candidate answers, what is the accuracy in selecting the correct answer?

Example: Let's say a model produced an answer 8, but the correct answer is 9. If 8 is not in the set of candidate answers, but the closest to 8 is 9, and thus selected 9, then correct option will be selected even if the wrong answer is produced. In this case, its O_{acc} = 100%, while its S_{acc} = 0%.

Quantitative Comparisons: Generalization Experiments

Puzzle Category →	Count	Arithmetic	Logic	Path Trace	Algebra	Measure	Spatial	Pattern Finding	Average
Puzzle Split (PS) – Extreme Generalization Experiments									
Avg. 2^{nd} Grader Performance	**72.8**	**81.3**	**82.2**	**81.1**	**64.5**	**90.4**	**74.8**	**88.6**	**77.1**
Greedy (baseline)	19.1/21.4	14.0/21.4	18.5/21.1	21.8/21.1	13.5/21.5	23.1/20.9	18.2/21.2	21.4/21.4	17.7/21.3
Uniform (baseline)	7.74/20.0	8.00/20.0	7.61/20.0	18.9/20.0	6.94/20.0	5.62/20.0	14.2/20.0	20.0/20.0	11.20/20.0
MAE + BERT	7.2/12.0	3.3/23.1	10.4/34.1	9.6/22.0	7.3/14.7	3.7/15.2	8.5/16.5	2.6/16.4	7.21/19.1
SimSiam + BERT	6.4/18.4	4.8/20.9	7.7/41.4	2.5/22.2	4.2/25.3	7.9/20.5	11.8/22.2	0.2/17.2	6.41/23.9
Swin_T + BERT	810.5/17.3	4.7/24.7	5.6/29.3	11.4/21.5	6.5/16.8	10.3/23.3	11.9/16.3	17.3/19.1	9.25/20.1
ViT-16 + BERT	9.41/22.7	5.77/26.8	6.95/25.1	4.72/18.7	5.57/15.1	8.68/21.3	11.6/21.5	18.9/19.7	8.51/21.6
CLIP	9.1/15.7	1.4/18.5	7.4/30.6	14.2/21.4	7.5/18.6	8.9/22.2	12.4/18.4	19.0/19.6	11.9/24.1
FLAVA	8.3/20.2	4.0/22.2	8.1/31.3	9.5/20.3	3.1/22.2	19.0/32.0	9.7/18.1	20.9/21.2	7.21/19.0
R50 + BERT (FT + Cls.)	10.9/18.3	6.96/15.8	12.8/20.8	19.6/19.7	7.95/15.1	16.9/26.7	13.4/17.7	0.0/21.2	11.7/18.9
R50 + BERT (FT + Reg.)	12.0/22.8	5.08/21.3	4.24/16.2	18.4/18.4	4.89/22.2	15.1/25.9	11.9/17.9	19.0/19.0	8.21/19.7

Second grader accuracy is computed over the responses from nearly 3000 kids who took the MK Olympiad in 2020-2021

Massachusetts Institute of Technology

JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

Quantitative Comparisons: State-of-the-Art Experiments

Puzzle Category →	Count	Arithmetic	Logic	Path Trace	Algebra	Measure	Spatial	Pattern Finding	Average
Puzzle Split (PS) – Extreme Generalization Experiments									
Avg. 2^{nd} Grader Performance	**72.8**	**81.3**	**82.2**	**81.1**	**64.5**	**90.4**	**74.8**	**88.6**	**77.1**
Greedy (baseline)	19.1/21.4	14.0/21.4	18.5/21.1	21.8/21.1	13.5/21.5	23.1/20.9	18.2/21.2	21.4/21.4	17.7/21.3
Uniform (baseline)	7.74/20.0	8.00/20.0	7.61/20.0	18.9/20.0	6.94/20.0	5.62/20.0	14.2/20.0	20.0/20.0	11.20/20.0
MAE + BERT	7.2/12.0	3.3/23.1	10.4/34.1	9.6/22.0	7.3/14.7	3.7/15.2	8.5/16.5	2.6/16.4	7.21/19.1
SimSiam + BERT	6.4/18.4	4.8/20.9	7.7/41.4	2.5/22.2	4.2/25.3	7.9/20.5	11.8/22.2	0.2/17.2	6.41/23.9
Swin_T + BERT	810.5/17.3	4.7/24.7	5.6/29.3	11.4/21.5	6.5/16.8	10.3/23.3	11.9/16.3	17.3/19.1	9.25/20.1
ViT-16 + BERT	9.41/22.7	5.77/26.8	6.95/25.1	4.72/18.7	5.57/15.1	8.68/21.3	11.6/21.5	18.9/19.7	8.51/21.6
CLIP	9.1/15.7	1.4/18.5	7.4/30.6	14.2/21.4	7.5/18.6	8.9/22.2	12.4/18.4	19.0/19.6	11.9/24.1
FLAVA	8.3/20.2	4.0/22.2	8.1/31.3	9.5/20.3	3.1/22.2	19.0/32.0	9.7/18.1	20.9/21.2	7.21/19.0
R50 + BERT (FT + Cls.)	10.9/18.3	6.96/15.8	12.8/20.8	19.6/19.7	7.95/15.1	16.9/26.7	13.4/17.7	0.0/21.2	11.7/18.9
R50 + BERT (FT + Reg.)	12.0/22.8	5.08/21.3	4.24/16.2	18.4/18.4	4.89/22.2	15.1/25.9	11.9/17.9	19.0/19.0	8.21/19.7
Few-Shot Split (FS) Experiments									
R50 + BERT (Cls.)	23.9/37.3	32.7/41.2	32.7/40.7	22.1/22.2	10.2/27.5	17.4/32.8	28.4/35.3	33.9/33.9	24.4/33.4
R50 + BERT (Reg.)	19.8/33.6	13.9/26.3	18.2/26.9	18.7/18.7	10.3/24.4	11.6/25.8	20.8/29.8	21.9/22.3	16.7/26.5

Quantitative Comparisons: Supervised Experiments

Puzzle Category →	Count	Arithmetic	Logic	Path Trace	Algebra	Measure	Spatial	Pattern Finding	Average
Instance Split (IS) – Supervised Learning Experiments									
Greedy (baseline)	21.7/22.6	8.97/21.5	18.5/21.0	22.7/21.2	10.2/21.1	12.8/21.1	22.3/21.3	20.6/21.3	17.3/21.6
Uniform (baseline)	9.41/20.0	3.65/20.0	7.91/20.0	11.1/20.0	5.01/20.0	3.63/20.0	15.5/20.0	16.7/20.0	8.41/20.0
Swin-T + Emb.	23.1/35.1	33.7/41.0	20.3/28.8	16.7/18.6	17.7/29.5	26.3/34.3	24.5/29.1	17.5/26.5	22.5/30.8
Swin-B + Emb.	22.0/34.0	29.4/36.5	17.7/26.1	16.7/17.0	17.1/30.2	25.0/34.2	26.2/30.7	21.5/29.6	21.6/29.9
Cross-Transformer + Emb.	20.5/30.4	6.3/15.3	15.5/22.9	15.1/15.6	8.7/23.9	10.7/18.2	21.7/24.7	19.0/27.3	14.7/22.8
ViT-16 + Emb.	25.6/36.4	39.7/47.1	21.2/30.8	15.5/16.3	20.1/33.8	39.4/40.8	29.0/33.0	20.3/29.6	25.9/33.5
MAE + Emb.	25.4/36.7	34.2/43.2	21.6/31.5	16.4/16.7	20.0/33.3	32.0/39.7	28.2/32.9	18.6/26.6	24.5/33.0
SimSiam + Emb.	44.9/56.1	35.1/43.5	45.7/50.8	25.0/26.6	23.4/35.1	64.7/73.5	55.0/57.2	42.8/49.1	39.5/47.0
R18 + Emb.	44.0/54.0	8.8/19.8	41.1/47.6	24.5/26.7	13.7/26.5	30.9/40.2	43.3/45.5	29.5/34.8	29.4/37.4
R50 + Emb.	46.6/57.8	38.0/45.9	43.2/50.1	24.6/26.4	23.3/35.1	56.9/67.4	57.9/58.6	44.8/51.0	39.8/47.5
R50 + GloVe	46.0/56.3	39.2/48.5	53.9/56.4	26.7/28.9	21.5/32.4	58.9/68.5	48.5/50.4	43.3/47.8	40.0/47.2
R50 + GPT2	47.0/57.9	44.8/53.1	55.1/58.6	26.1/28.4	27.2/39.3	61.0/71.3	49.0/50.2	42.5/48.4	42.1/49.6
R50 + BERT	48.5/59.3	46.1/54.9	56.7/60.2	26.5/28.4	28.5/39.7	65.6/75.4	44.3/46.2	39.9/45.3	42.8/50.2
CLIP	41.3/52.9	18.2/29.3	33.3/41.1	19.8/21.9	12.9/24.9	27.8/42.8	32.2/36.2	29.9/36.1	27.3/36.4
FLAVA	47.7/58.1	20.2/29.7	41.4/47.1	25.4/27.1	19.6/31.2	30.5/41.9	33.2/35.7	38.3/44.2	32.3/40.2
Answer Split (AS) – Answer Generalization Experiments									
R50 + BERT (FT + Cls.)	0.1/23.8	1.5/13.2	0.0/16.8	0.0/1.6	0.4/17.3	0.0/21.1	0.0/6.0	0.0/15.0	0.19/10.2
R50 + BERT (FT + Reg.)	12.0/28.4	10.4/25.7	19.6/30.8	9.5/10.6	3.64/18.3	9.42/28.6	14.1/21.1	25.5/30.9	16.3/23.4

Supervised Experiments Using Transformers

Puzzle Category →	Count	Arithmetic	Logic	Path Trace	Algebra	Measure	Spatial	Pattern Finding	Average
Instance Split (IS) – Supervised Learning Experiments									
Greedy (baseline)	21.7/22.6	8.97/21.5	18.5/21.0	22.7/21.2	10.2/21.1	12.8/21.1	22.3/21.3	20.6/21.3	17.3/21.6
Uniform (baseline)	9.41/20.0	3.65/20.0	7.91/20.0	11.1/20.0	5.01/20.0	3.63/20.0	15.5/20.0	16.7/20.0	8.41/20.0
Swin-T + Emb.	23.1/35.1	33.7/41.0	20.3/28.8	16.7/18.6	17.7/29.5	26.3/34.3	24.5/29.1	17.5/26.5	22.5/30.8
Swin-B + Emb.	22.0/34.0	29.4/36.5	17.7/26.1	16.7/17.0	17.1/30.2	25.0/34.2	26.2/30.7	21.5/29.6	21.6/29.9
Cross-Transformer + Emb.	20.5/30.4	6.3/15.3	15.5/22.9	15.1/15.6	8.7/23.9	10.7/18.2	21.7/24.7	19.0/27.3	14.7/22.8
ViT-16 + Emb.	25.6/36.4	39.7/47.1	21.2/30.8	15.5/16.3	20.1/33.8	39.4/40.8	29.0/33.0	20.3/29.6	25.9/33.5
MAE + Emb.	25.4/36.7	34.2/43.2	21.6/31.5	16.4/16.7	20.0/33.3	32.0/39.7	28.2/32.9	18.6/26.6	24.5/33.0
SimSiam + Emb.	44.9/56.1	35.1/43.5	45.7/50.8	25.0/26.6	23.4/35.1	64.7/73.5	55.0/57.2	42.8/49.1	39.5/47.0
R18 + Emb.	44.0/54.0	8.8/19.8	41.1/47.6	24.5/26.7	13.7/26.5	30.9/40.2	43.3/45.5	29.5/34.8	29.4/37.4
R50 + Emb.	46.6/57.8	38.0/45.9	43.2/50.1	24.6/26.4	23.3/35.1	56.9/67.4	57.9/58.6	44.8/51.0	39.8/47.5
R50 + GloVe	46.0/56.3	39.2/48.5	53.9/56.4	26.7/28.9	21.5/32.4	58.9/68.5	48.5/50.4	43.3/47.8	40.0/47.2
R50 + GPT2	47.0/57.9	44.8/53.1	55.1/58.6	26.1/28.4	27.2/39.3	61.0/71.3	49.0/50.2	42.5/48.4	42.1/49.6
R50 + BERT	48.5/59.3	46.1/54.9	56.7/60.2	26.5/28.4	28.5/39.7	65.6/75.4	44.3/46.2	39.9/45.3	42.8/50.2
CLIP	41.3/52.9	18.2/29.3	33.3/41.1	19.8/21.9	12.9/24.9	27.8/42.8	32.2/36.2	29.9/36.1	27.3/36.4
FLAVA	47.7/58.1	20.2/29.7	41.4/47.1	25.4/27.1	19.6/31.2	30.5/41.9	33.2/35.7	38.3/44.2	32.3/40.2
Answer Split (AS) – Answer Generalization Experiments									
R50 + BERT (FT + Cls.)	0.1/23.8	1.5/13.2	0.0/16.8	0.0/1.6	0.4/17.3	0.0/21.1	0.0/6.0	0.0/15.0	0.19/10.2
R50 + BERT (FT + Reg.)	12.0/28.4	10.4/25.7	19.6/30.8	9.5/10.6	3.64/18.3	9.42/28.6	14.1/21.1	25.5/30.9	16.3/23.4

MITSUBISHI ELECTRIC
Changes for the Better

Massachusetts Institute of Technology

JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

Supervised Experiments: Large Language Models

Puzzle Category →	Count	Arithmetic	Logic	Path Trace	Algebra	Measure	Spatial	Pattern Finding	Average
Instance Split (IS) – Supervised Learning Experiments									
Greedy (baseline)	21.7/22.6	8.97/21.5	18.5/21.0	22.7/21.2	10.2/21.1	12.8/21.1	22.3/21.3	20.6/21.3	17.3/21.6
Uniform (baseline)	9.41/20.0	3.65/20.0	7.91/20.0	11.1/20.0	5.01/20.0	3.63/20.0	15.5/20.0	16.7/20.0	8.41/20.0
Swin-T + Emb.	23.1/35.1	33.7/41.0	20.3/28.8	16.7/18.6	17.7/29.5	26.3/34.3	24.5/29.1	17.5/26.5	22.5/30.8
Swin-B + Emb.	22.0/34.0	29.4/36.5	17.7/26.1	16.7/17.0	17.1/30.2	25.0/34.2	26.2/30.7	21.5/29.6	21.6/29.9
Cross-Transformer + Emb.	20.5/30.4	6.3/15.3	15.5/22.9	15.1/15.6	8.7/23.9	10.7/18.2	21.7/24.7	19.0/27.3	14.7/22.8
ViT-16 + Emb.	25.6/36.4	39.7/47.1	21.2/30.8	15.5/16.3	20.1/33.8	39.4/40.8	29.0/33.0	20.3/29.6	25.9/33.5
MAE + Emb.	25.4/36.7	34.2/43.2	21.6/31.5	16.4/16.7	20.0/33.3	32.0/39.7	28.2/32.9	18.6/26.6	24.5/33.0
SimSiam + Emb.	44.9/56.1	35.1/43.5	45.7/50.8	25.0/26.6	23.4/35.1	64.7/73.5	55.0/57.2	42.8/49.1	39.5/47.0
R18 + Emb.	44.0/54.0	8.8/19.8	41.1/47.6	24.5/26.7	13.7/26.5	30.9/40.2	43.3/45.5	29.5/34.8	29.4/37.4
R50 + Emb.	46.6/57.8	38.0/45.9	43.2/50.1	24.6/26.4	23.3/35.1	56.9/67.4	57.9/58.6	44.8/51.0	39.8/47.5
R50 + GloVe	46.0/56.3	39.2/48.5	53.9/56.4	26.7/28.9	21.5/32.4	58.9/68.5	48.5/50.4	43.3/47.8	40.0/47.2
R50 + GPT2	47.0/57.9	44.8/53.1	55.1/58.6	26.1/28.4	27.2/39.3	61.0/71.3	49.0/50.2	42.5/48.4	42.1/49.6
R50 + BERT	48.5/59.3	46.1/54.9	56.7/60.2	26.5/28.4	28.5/39.7	65.6/75.4	44.3/46.2	39.9/45.3	42.8/50.2
CLIP	41.3/52.9	18.2/29.3	33.3/41.1	19.8/21.9	12.9/24.9	27.8/42.8	32.2/36.2	29.9/36.1	27.3/36.4
FLAVA	47.7/58.1	20.2/29.7	41.4/47.1	25.4/27.1	19.6/31.2	30.5/41.9	33.2/35.7	38.3/44.2	32.3/40.2
Answer Split (AS) – Answer Generalization Experiments									
R50 + BERT (FT + Cls.)	0.1/23.8	1.5/13.2	0.0/16.8	0.0/1.6	0.4/17.3	0.0/21.1	0.0/6.0	0.0/15.0	0.19/10.2
R50 + BERT (FT + Reg.)	12.0/28.4	10.4/25.7	19.6/30.8	9.5/10.6	3.64/18.3	9.42/28.6	14.1/21.1	25.5/30.9	16.3/23.4

Supervised Experiments: Vision-Language Foundation Models

Puzzle Category →	Count	Arithmetic	Logic	Path Trace	Algebra	Measure	Spatial	Pattern Finding	Average
Instance Split (IS) – Supervised Learning Experiments									
Greedy (baseline)	21.7/22.6	8.97/21.5	18.5/21.0	22.7/21.2	10.2/21.1	12.8/21.1	22.3/21.3	20.6/21.3	17.3/21.6
Uniform (baseline)	9.41/20.0	3.65/20.0	7.91/20.0	11.1/20.0	5.01/20.0	3.63/20.0	15.5/20.0	16.7/20.0	8.41/20.0
Swin-T + Emb.	23.1/35.1	33.7/41.0	20.3/28.8	16.7/18.6	17.7/29.5	26.3/34.3	24.5/29.1	17.5/26.5	22.5/30.8
Swin-B + Emb.	22.0/34.0	29.4/36.5	17.7/26.1	16.7/17.0	17.1/30.2	25.0/34.2	26.2/30.7	21.5/29.6	21.6/29.9
Cross-Transformer + Emb.	20.5/30.4	6.3/15.3	15.5/22.9	15.1/15.6	8.7/23.9	10.7/18.2	21.7/24.7	19.0/27.3	14.7/22.8
ViT-16 + Emb.	25.6/36.4	39.7/47.1	21.2/30.8	15.5/16.3	20.1/33.8	39.4/40.8	29.0/33.0	20.3/29.6	25.9/33.5
MAE + Emb.	25.4/36.7	34.2/43.2	21.6/31.5	16.4/16.7	20.0/33.3	32.0/39.7	28.2/32.9	18.6/26.6	24.5/33.0
SimSiam + Emb.	44.9/56.1	35.1/43.5	45.7/50.8	25.0/26.6	23.4/35.1	64.7/73.5	55.0/57.2	42.8/49.1	39.5/47.0
R18 + Emb.	44.0/54.0	8.8/19.8	41.1/47.6	24.5/26.7	13.7/26.5	30.9/40.2	43.3/45.5	29.5/34.8	29.4/37.4
R50 + Emb.	46.6/57.8	38.0/45.9	43.2/50.1	24.6/26.4	23.3/35.1	56.9/67.4	57.9/58.6	44.8/51.0	39.8/47.5
R50 + GloVe	46.0/56.3	39.2/48.5	53.9/56.4	26.7/28.9	21.5/32.4	58.9/68.5	48.5/50.4	43.3/47.8	40.0/47.2
R50 + GPT2	47.0/57.9	44.8/53.1	55.1/58.6	26.1/28.4	27.2/39.3	61.0/71.3	49.0/50.2	42.5/48.4	42.1/49.6
R50 + BERT	48.5/59.3	46.1/54.9	56.7/60.2	26.5/28.4	28.5/39.7	65.6/75.4	44.3/46.2	39.9/45.3	42.8/50.2
CLIP	41.3/52.9	18.2/29.3	33.3/41.1	19.8/21.9	12.9/24.9	27.8/42.8	32.2/36.2	29.9/36.1	27.3/36.4
FLAVA	47.7/58.1	20.2/29.7	41.4/47.1	25.4/27.1	19.6/31.2	30.5/41.9	33.2/35.7	38.3/44.2	32.3/40.2
Answer Split (AS) – Answer Generalization Experiments									
R50 + BERT (FT + Cls.)	0.1/23.8	1.5/13.2	0.0/16.8	0.0/1.6	0.4/17.3	0.0/21.1	0.0/6.0	0.0/15.0	0.19/10.2
R50 + BERT (FT + Reg.)	12.0/28.4	10.4/25.7	19.6/30.8	9.5/10.6	3.64/18.3	9.42/28.6	14.1/21.1	25.5/30.9	16.3/23.4

Quantitative Comparisons: Answer Generalization

Puzzle Category →	Count	Arithmetic	Logic	Path Trace	Algebra	Measure	Spatial	Pattern Finding	Average
Instance Split (IS) – Supervised Learning Experiments									
Greedy (baseline)	21.7/22.6	8.97/21.5	18.5/21.0	22.7/21.2	10.2/21.1	12.8/21.1	22.3/21.3	20.6/21.3	17.3/21.6
Uniform (baseline)	9.41/20.0	3.65/20.0	7.91/20.0	11.1/20.0	5.01/20.0	3.63/20.0	15.5/20.0	16.7/20.0	8.41/20.0
Swin-T + Emb.	23.1/35.1	33.7/41.0	20.3/28.8	16.7/18.6	17.7/29.5	26.3/34.3	24.5/29.1	17.5/26.5	22.5/30.8
Swin-B + Emb.	22.0/34.0	29.4/36.5	17.7/26.1	16.7/17.0	17.1/30.2	25.0/34.2	26.2/30.7	21.5/29.6	21.6/29.9
Cross-Transformer + Emb.	20.5/30.4	6.3/15.3	15.5/22.9	15.1/15.6	8.7/23.9	10.7/18.2	21.7/24.7	19.0/27.3	14.7/22.8
ViT-16 + Emb.	25.6/36.4	39.7/47.1	21.2/30.8	15.5/16.3	20.1/33.8	39.4/40.8	29.0/33.0	20.3/29.6	25.9/33.5
MAE + Emb.	25.4/36.7	34.2/43.2	21.6/31.5	16.4/16.7	20.0/33.3	32.0/39.7	28.2/32.9	18.6/26.6	24.5/33.0
SimSiam + Emb.	44.9/56.1	35.1/43.5	45.7/50.8	25.0/26.6	23.4/35.1	64.7/73.5	55.0/57.2	42.8/49.1	39.5/47.0
R18 + Emb.	44.0/54.0	8.8/19.8	41.1/47.6	24.5/26.7	13.7/26.5	30.9/40.2	43.3/45.5	29.5/34.8	29.4/37.4
R50 + Emb.	46.6/57.8	38.0/45.9	43.2/50.1	24.6/26.4	23.3/35.1	56.9/67.4	57.9/58.6	44.8/51.0	39.8/47.5
R50 + GloVe	46.0/56.3	39.2/48.5	53.9/56.4	26.7/28.9	21.5/32.4	58.9/68.5	48.5/50.4	43.3/47.8	40.0/47.2
R50 + GPT2	47.0/57.9	44.8/53.1	55.1/58.6	26.1/28.4	27.2/39.3	61.0/71.3	49.0/50.2	42.5/48.4	42.1/49.6
R50 + BERT	48.5/59.3	46.1/54.9	56.7/60.2	26.5/28.4	28.5/39.7	65.6/75.4	44.3/46.2	39.9/45.3	42.8/50.2
CLIP	41.3/52.9	18.2/29.3	33.3/41.1	19.8/21.9	12.9/24.9	27.8/42.8	32.2/36.2	29.9/36.1	27.3/36.4
FLAVA	47.7/58.1	20.2/29.7	41.4/47.1	25.4/27.1	19.6/31.2	30.5/41.9	33.2/35.7	38.3/44.2	32.3/40.2
Answer Split (AS) – Answer Generalization Experiments									
R50 + BERT (FT + Cls.)	0.1/23.8	1.5/13.2	0.0/16.8	0.0/1.6	0.4/17.3	0.0/21.1	0.0/6.0	0.0/15.0	0.19/10.2
R50 + BERT (FT + Reg.)	12.0/28.4	10.4/25.7	19.6/30.8	9.5/10.6	3.64/18.3	9.42/28.6	14.1/21.1	25.5/30.9	16.3/23.4

Regression generalizes better, perhaps via answer interpolation.

Comparisons to ChatGPT, GPT-4, Bing, and Bard

puzzle ID	7	9	30	38	47	71	88	89	90	91	93	mean
Category	AL	S	AM	AM	AM	AM	AM	C	AL	L	M	
Human	NA	NA	NA	NA	NA	60.4	NA	NA	NA	NA	NA	60.4
Bard [1]	0.0	20.0	0.0	50.0	0.0	0.0	0.0	10.0	10.0	20.0	30.0	12.7
ChatGPT3.5 [3]	70.0	10.0	0.0	20.0	0.0	40.0	70.0	10.0	30.0	60.0	90.0	36.4
BGPT4-C [2]	20.0	0.0	100.0	90.0	10.0	0.0	100.0	0.0	10.0	20.0	30.0	26.4
BGPT4-B [2]	30.0	0.0	0.0	0.0	0.0	40.0	0.0	0.0	0.0	0.0	100.0	15.5
BGPT4-P [2]	100.0	0.0	100.0	70.0	0.0	90.0	0.0	0.0	0.0	0.0	30.0	35.5
PS split	NA	NA	NA	NA	NA	4.65	NA	NA	NA	25.5	NA	15.1
IS split	98.0	14.0	100.0	64.6	93.7	56.7	21.3	55.7	51.3	26.3	34.0	**55.9**

Puzzle Category →	Count	Arithmetic	Logic	Path Trace	Algebra	Measure	Spatial	Pattern Finding	Average
Puzzle Split (PS) – Extreme Generalization Experiments									
Avg. 2^{nd} Grader Performance	**72.8**	**81.3**	**82.2**	**[illegible]1.1**	**64.5**	**90.4**	**74.8**	**88.6**	**77.1**

BGPT4-C = Bing + GPT-4 + Creative variant
BGPT4-B = Bing + GPT-4 + Balanced variant
BGPT4-P = Bing + GPT-4 + Precise variant

Conclusions

- Are deep neural networks SMARTer than second graders?
  >> Not yet

- However, the recent large language models (e.g., ChatGPT) appear to showcase convincing out-of-domain generalization.

- Nevertheless, there appears to be gaps in its reasoning and perhaps a more systematic adherence to the development of foundational learning and reasoning skills is important.

Thank you!

Anthony Vetro
MERL

Tim K. Marks
MERL

Joanna Matthissen
Math Kangaroo USA

Mike Jones
MERL

Moitreya Chatterjee
MERL

Ayden Anoop

SMART-101 dataset is public at: https://zenodo.org/record/7775984

For questions, contact **cherian@merl.com**

MITSUBISHI ELECTRIC

Changes for the Better