Are Deep Neural Networks SMARTer than Second Graders?

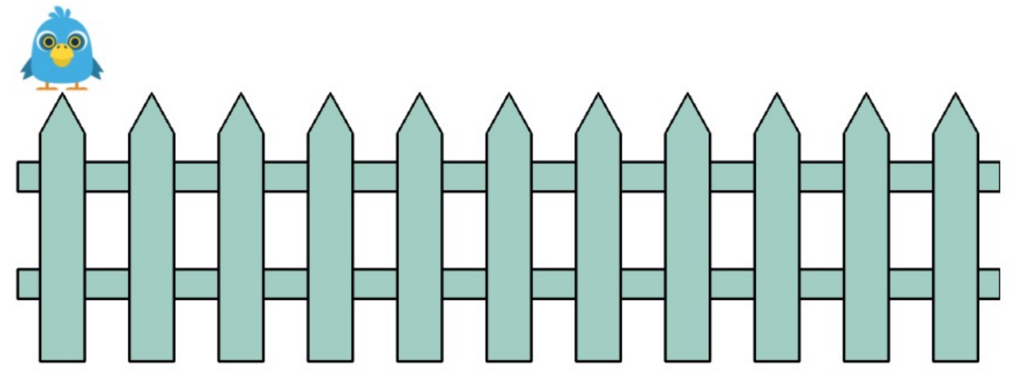
MITSUBISHI ELECTRIC RESEARCH LABORATORIES, INC
Massachusetts Institute of Technology

Anoop Cherian[1] Kuan-Chuan Peng[1] Suhas Lohit[1] Kevin Smith[2] Joshua B. Tenenbaum[2]
[1]Mitsubishi Electric Research Labs (MERL), Cambridge, MA [2]Massachusetts Institute of Technology (MIT), Cambridge MA
{cherian, kpeng, slohit}@merl.com} {k2smith, jbt}@mit.edu

JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

Problem:

Can a state-of-the-art (deep) machine learning model solve the simple puzzle below?

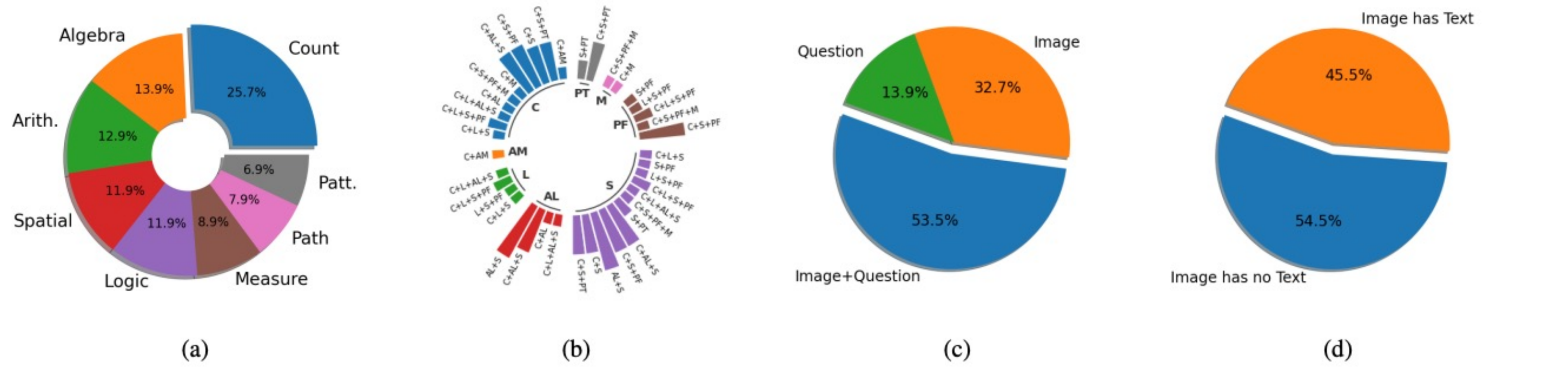


Question: *Bird Bobbie jumps on a fence from the post on the left end to the other end. Each jump takes him 4 seconds. He makes 4 jumps ahead and then 1 jump back. Then he again makes 4 jumps ahead and 1 jump back, and so on. In how many seconds can Bobbie get from one end to the other end?*
Answer Options: A: 64 B: 48 C: 56 D: 68 E: 72

Contributions:

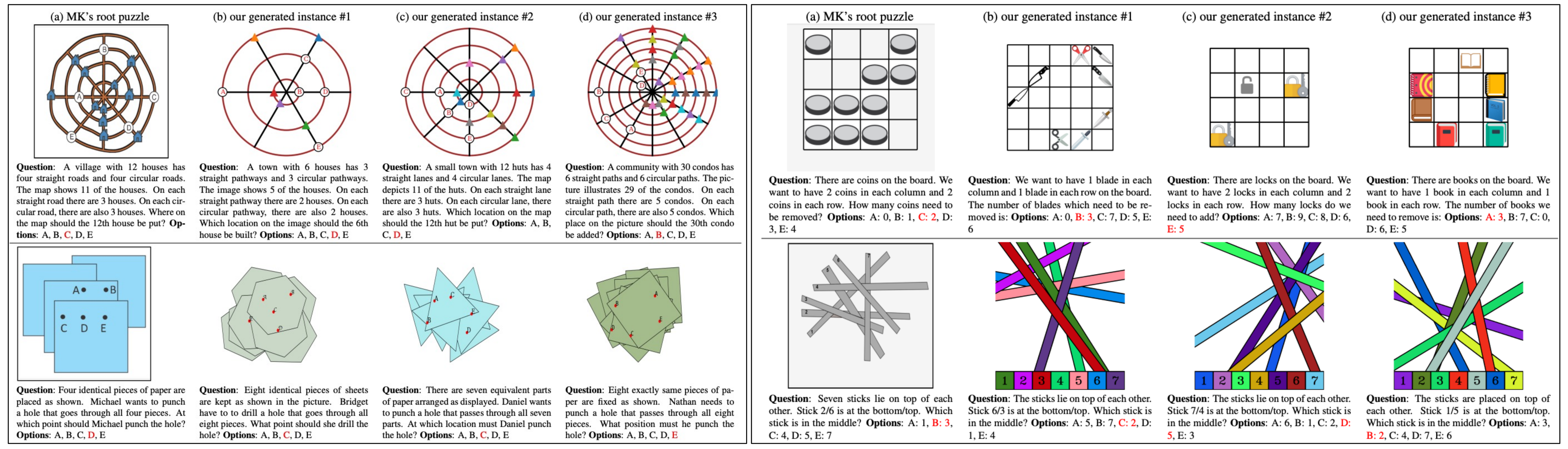
(i) We introduce **SMART: Simple Multi-modal Algorithmic Reasoning Task** for evaluating the abstraction, deduction, and generalization abilities of neural networks in solving visuo-linguistic puzzles designed specifically for first/second grade children.
(ii) To ensure the puzzles are solvable by kids, we take them from the **Math Kangaroo (MK) Olympiad** intended for second graders.
(iii) We introduce the **SMART-101 dataset** built from 101 unique MK puzzles to evaluate the progress in multimodal artificial general intelligence.
(iv) We propose **programmatic augmentation** to replicate each MK puzzle to arbitrary number of instances for training large machine learning models, so that the models learn the 'solution algorithm'.
(v) We analyze the **generalization performances** of state-of-the-art vision and language pretrained models and show that they are not better than second grader performances (yet).

SMART-101 Statistics:



We plot the distributions of: (a) 8 primary algorithmic skills needed to solve the 101 puzzles, (b) compositional reasoning skills, (c) puzzles that need image and/or question reasoning, (d) puzzles that need methods to read text within images (e.g., needing OCR abilities).

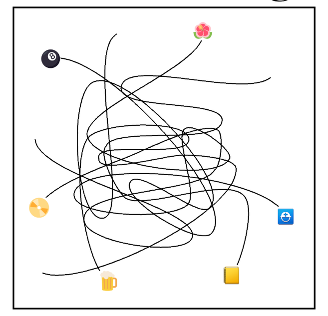
SMART Programmatic Puzzle Augmentations:



We use computer programs to replicate each puzzle; the arguments of these programs can be randomly sampled to produce various augmentations of the respective puzzle; e.g., change question, change appearances, etc. while keeping the underlying solution algorithm the same. We can control the difficulty of each puzzle as shown above using this method. Thus, when trained the expectation is that the model must learn the 'algorithm'.

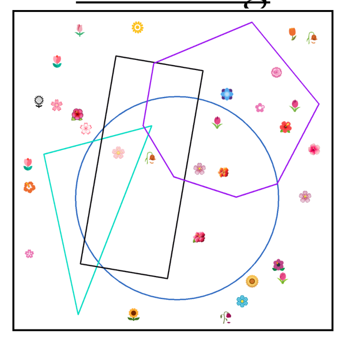
SMART Puzzle Categories:

Path Tracing



Question: Which object is linked to the hat?
Options: A: flower
B: disk C: book
D: drink E: ball

Counting

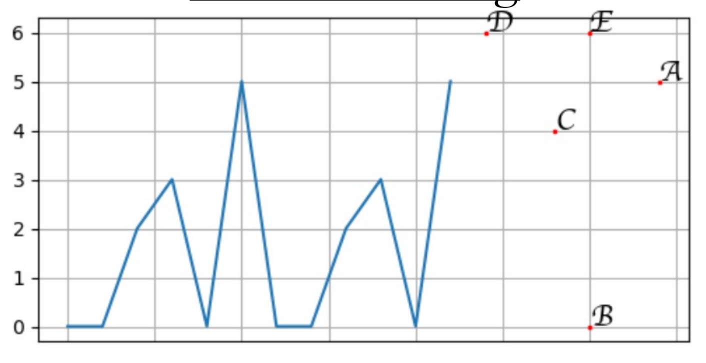


Question: All the flowers which are inside the circle but outside the triangle simultaneously are picked up. The number of flowers which are picked up is:
Options: A: 7 B: 2 C: 6 D: 3 E: 5

Logic

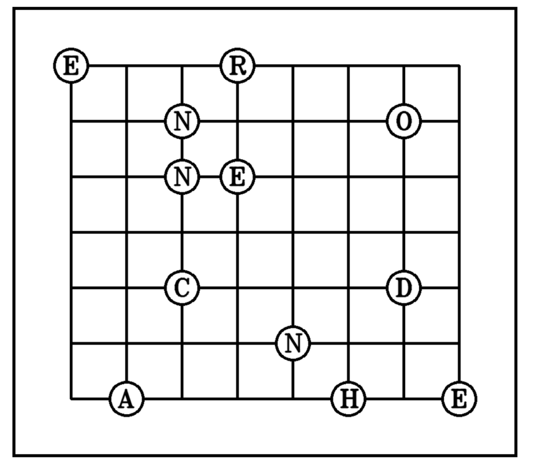
Question: Charles has 6 toy objects: a helmet, a clamp, a mop, teddies, a photo album, and a sketch pad. He puts each toy on a different line of the shelf. He puts photo album as shown. The helmet ahead of mop and sketch pad behind teddies. Clamp ahead of mop. On which line can the helmet not be placed?
Options: A: 1 B: 6 C: 5 D: 4 E: 3

Pattern Finding



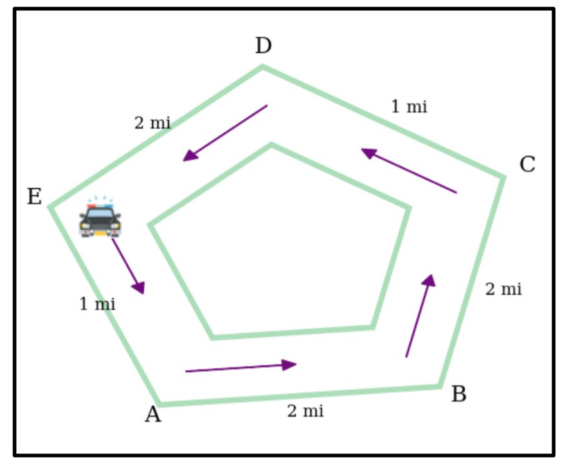
Question: Lawrence built a pattern twice, as demonstrated in the figure. Which position will he get to when he builds the next pattern.
Options: A B C D E

Measurement



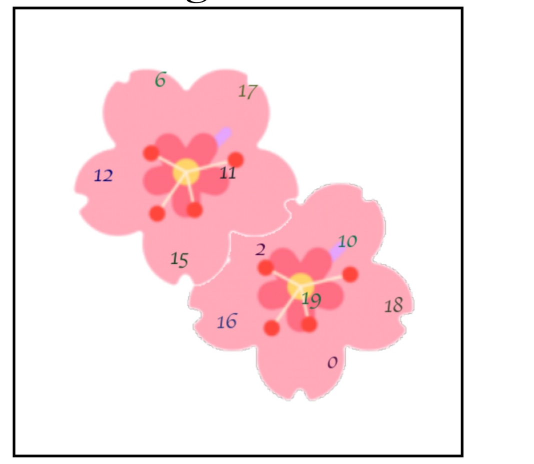
Question: We want to walk from N to E along the lines and pick up the letters NONADHERENCE in the correct order. The length of the shortest walk in units is (the length of each grid is 1 unit):?
Options: A: 50 B: 44 C: 46 D: 47 E: 45

Arithmetic



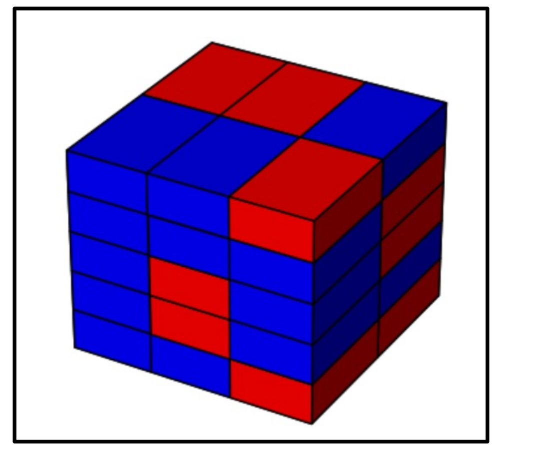
Question: Albert the vehicle moves along the road. He starts at stop E and traverses the path of the arrows presented in the cartoon. The vehicle moves a total of 24 miles. Where does he end up?
Options: A B C D E

Algebra



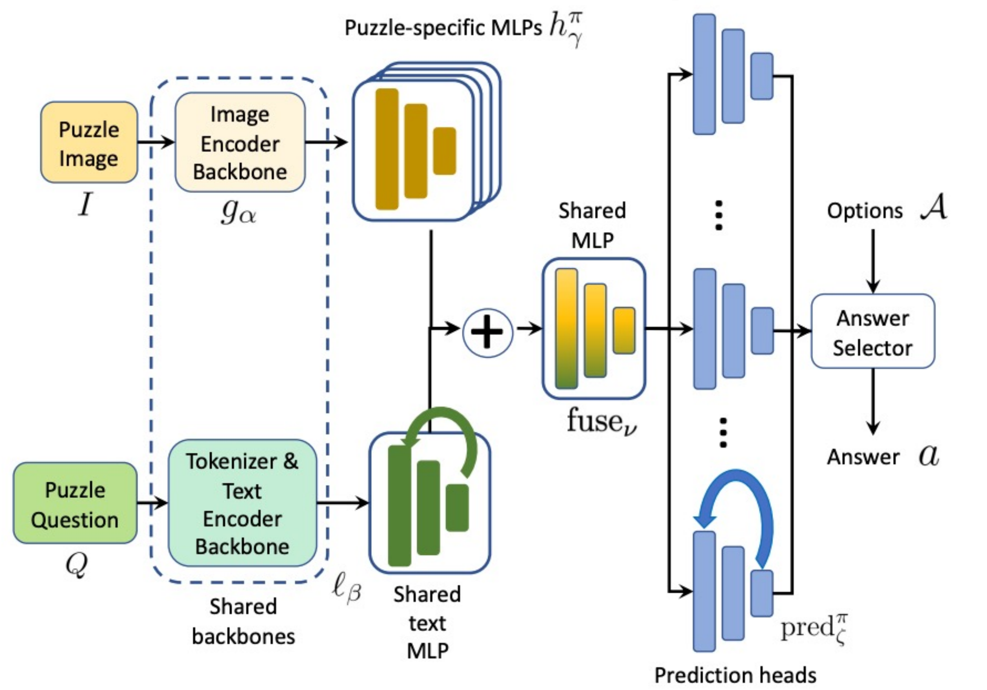
Question: Unique values are written on a pair of flowers. One value on a flower is concealed. The sums of the values on the two flowers are identical. What is the concealed value?
Options: A: 19 B: 8 C: 6 D: 11 E: 5

Spatial Reasoning



Question: Sandra made a structure using some red bricks and 21 blue bricks. How many of these blue bricks are not visible in the figure?
Options: A: 10 B: 11 C: 5. D: 1. E: 8

SMART Meta-Learning Reasoning Architecture:



Our architecture allows evaluating varied image and backbone networks for solving the SMART puzzles.

We evaluate ViTs, ResNets, MAE, etc. for image and BERT/GPT/GloVe for language backbones. We also compare to vision-and-language models such as CLIP & FLAVA.

Experiments and Results:

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, $m = 10$									
R50 + BERT (Cls.)	17.3/28.0	11.2/25.8	18.0/37.6	19.2/19.2	7.9/21.9	14.8/31.2	18.7/25.8	17.8/17.8	15.2/25.3
R50 + BERT (Reg.)	13.3/25.2	8.3/24.7	11.2/23.3	17.3/18.6	6.6/18.9	19.5/34.2	18.5/26.4	21.1/21.1	13.6/23.3
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

We show performances on SMART-101 under various generalization settings.

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**

Comparisons on large language models using a text subset of SMART 101.