Measuring Vision-Language STEM Skills of Neural Models

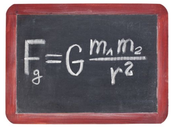
Jianhao Shen*, Ye Yuan*, Srbuhi Mirzoyan, Ming Zhang ♤, Chenguang Wang ♤

{jhshen, yuanye_pku, mzhang_cs}@pku.edu.cn, srbuhimirzoyan@stu.pku.edu.cn, chenguangwang@wustl.edu

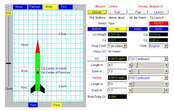
Paper HFDataset Leaderboard

➢ STEM is the basis of solving a wide set of real-world problems

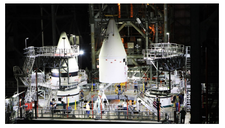
If we want to launch a Falcon rocket...



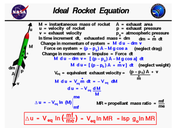
Science



Technology



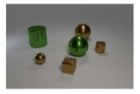
Engineering



Math

➢ Challenges for neural models

Q: How many bikes are there?
A: 2



Q: What color is the small shiny cube?
A: Brown

Q: Which picture shows the pizza inside the oven?
(A) The left one (B) The right one

Q: Which type of force from the baby's hand opens the cabinet door?
(A) Pull (B) Push

	VQA		CLEVR		IconQA				ScienceQA		
Dataset	#Questions	#Images	Multimodal	Q Length	#Answers	#Skills	Subjects	Grades	Image Type	Answer Type	Difficulty
VQA (2015)	614,163	204,721	✓	6.1	-	-	-	-	Natural	Text	-
CLEVR (2017)	999,968	100,000	✓	18.4	-	-	-	-	Natural	Text&Number	-
MATH (2021b)	12,500	-	✗	64.8	-	7	Math	9~12	-	Number	Advanced
MMLU (2021a)	15,908	-	✗	52.6	4	-	STEM	-	-	Multi-choice	Advanced
Geometry3K (2021)	3,002	2,342	✓	10.1	4	-	Math	6~12	Diagram	Multi-choice	Medium
IconQA (2021b)	107,439	96,817	✓	8.4	2-5	13	Math	Pre-K~3	Icon	Multi-choice&Others	Fundamental
ScienceQA (2022)	21,208	10,332	✗	12.1	2-5	379	Science	1~12	Natural&Diagram	Multi-choice	Medium
STEM (ours)	1,073,146	1,911,728	✓	17.4	2-4	448	STEM	Pre-K~8	Natural&Diagram	Multi-choice	Fundamental

#1: Existing datasets focus on examining expert-level ability
#2: There is no multimodal and unified STEM benchmark

➢ Our STEM benchmark

Q: Think about the magnetic force between the magnets in each pair. Which is true?
(A) It is smaller in Pair 1.
(B) It is the same in both pairs
(C) It is smaller in Pair 2.
(i) Science

Q: What kind of computer component do you see?
(A) Display Adapter/Video Card
(B) CPU Socket
(C) SATA Bus
(ii) Technology

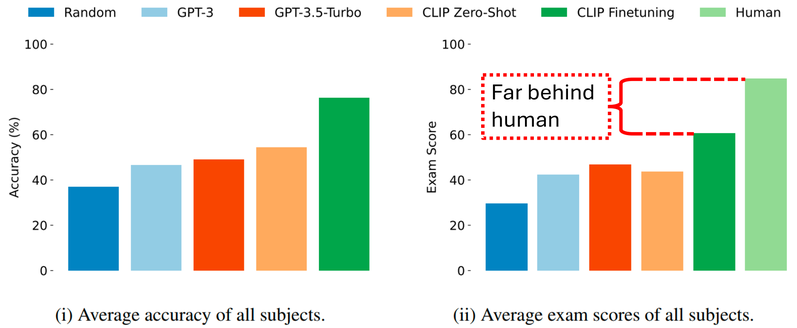
Q: Vicky wondered if steel would rust faster if she added vinegar to the salt water. She put five into a tub with salt water, and the other five into a tub with salt water mixed with vinegar. Which were part of an experimental group?
(A) Those soaked in salt water
(B) Those soaked in salt water and vinegar
STEM (iii) Engineering

Q: Identify the cross section of this object.
(A) (B) (C) (D)
(iv) Math

Subject	#Skills	#Questions	Average #A	#Train	#Valid	#Test
Science	82	186,740	2.8	112,120	37,343	37,277
Technology	9	8,566	4.0	5,140	1,713	1,713
Engineering	6	18,981	2.5	12,055	3,440	3,486
Math	351	858,859	2.8	515,482	171,776	171,601
Total	448	1,073,146	2.8	644,797	214,272	214,077

STEM Benckmark has the largest multimodal STEM dataset in terms of number of skills and questions

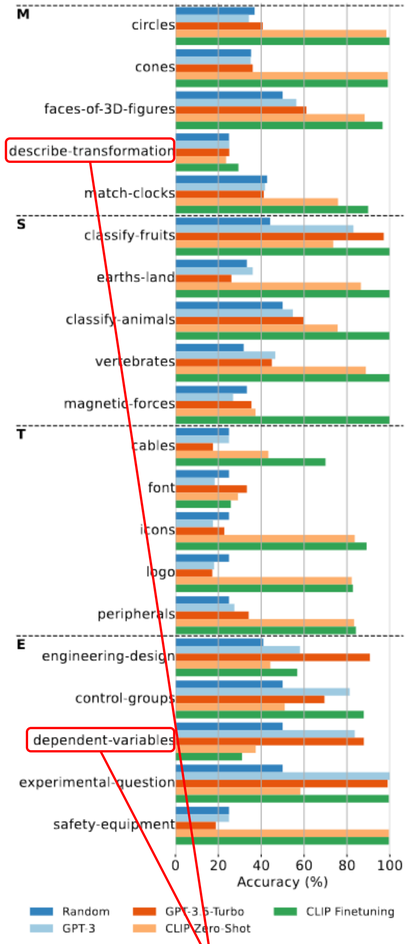
➢ Main results



(i) Average accuracy of all subjects. (ii) Average exam scores of all subjects.

Models took real-world exams and performed far behind millions of elementary students

➢ Skill analysis



Challenging skills: abstract knowledge & complex reasoning