Dataset Leaderboard Paper

Measuring Vision-Language STEM Skills of Neural Models

ICLR 2024

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

School of Computer Science, Peking University

National Key Laboratory for Multimedia Information Processing, Peking University

Peking University-Anker Embodied AI Lab

Washington University in St. Louis

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

* Equal Contribution

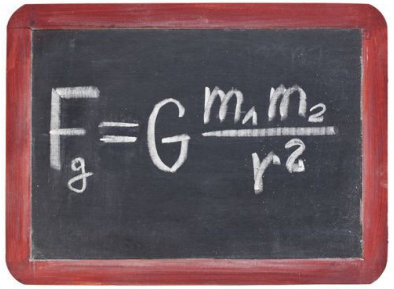
♤ Corresponding Authors

Understanding Vision-Language STEM is Important

Imagine if we want to launch a Falcon rocket…

Understanding Vision-Language STEM is Important

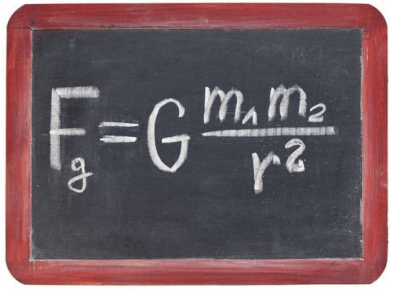
Imagine if we want to launch a Falcon rocket…



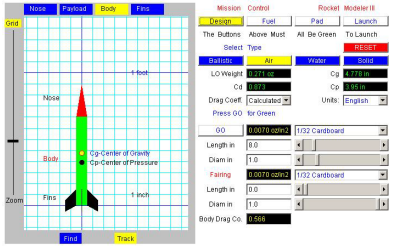
Science

Understanding Vision-Language STEM is Important

Imagine if we want to launch a Falcon rocket…



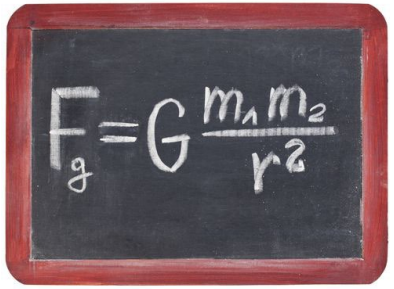
Science



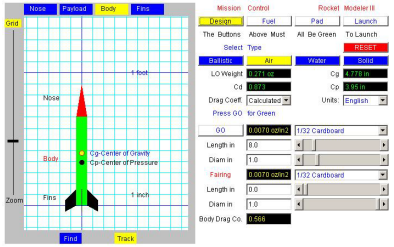
Technology

Understanding Vision-Language STEM is Important

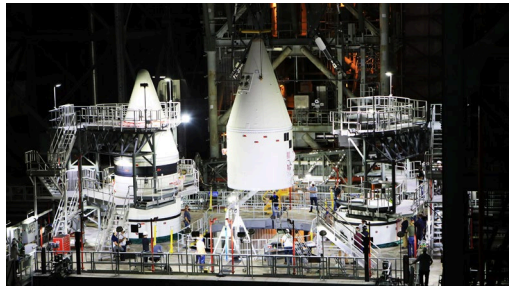
Imagine if we want to launch a Falcon rocket…



Science



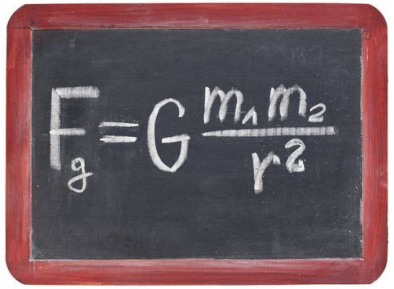
Technology



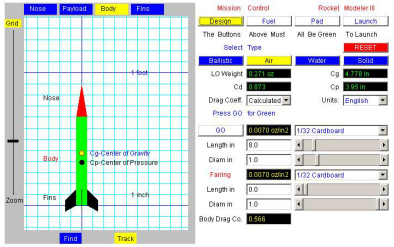
Engineering

Understanding Vision-Language STEM is Important

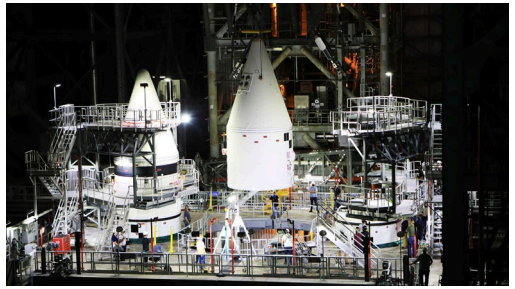
Imagine if we want to launch a Falcon rocket…



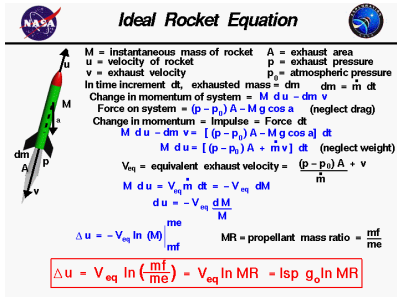
Science



Technology



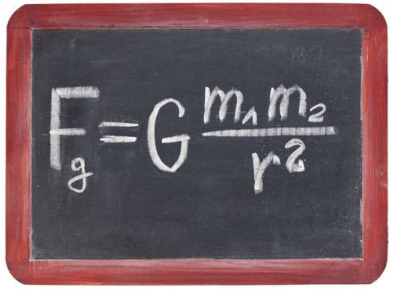
Engineering



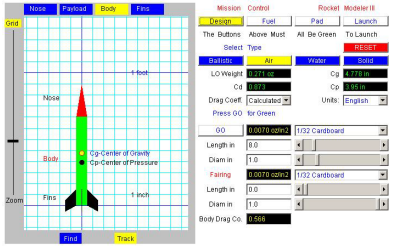
Math

Understanding Vision-Language STEM is Important

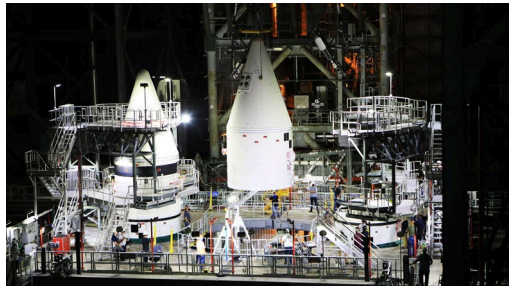
Imagine if we want to launch a Falcon rocket…



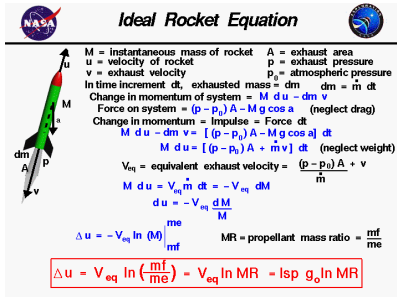
Science



Technology



Engineering



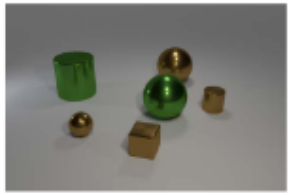
Math

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

Challenges for current benchmarks

Q: How many bikes are there?
A: 2

VQA



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

CLEVR

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

IconQA

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

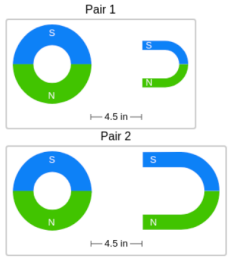
ScienceQA

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 (2021a)	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

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

Our STEM Benchmark

Examples of 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

(iii) Engineering

Q: Identify the cross section of this object.

(A) (B) **(C)** (D)

(iv) Math

Our STEM benchmark contains vision-language multi-choice questions across all STEM subjects, i.e., science, technology, engineering and math. These questions cover K-12 curriculum of U.S. national standard.

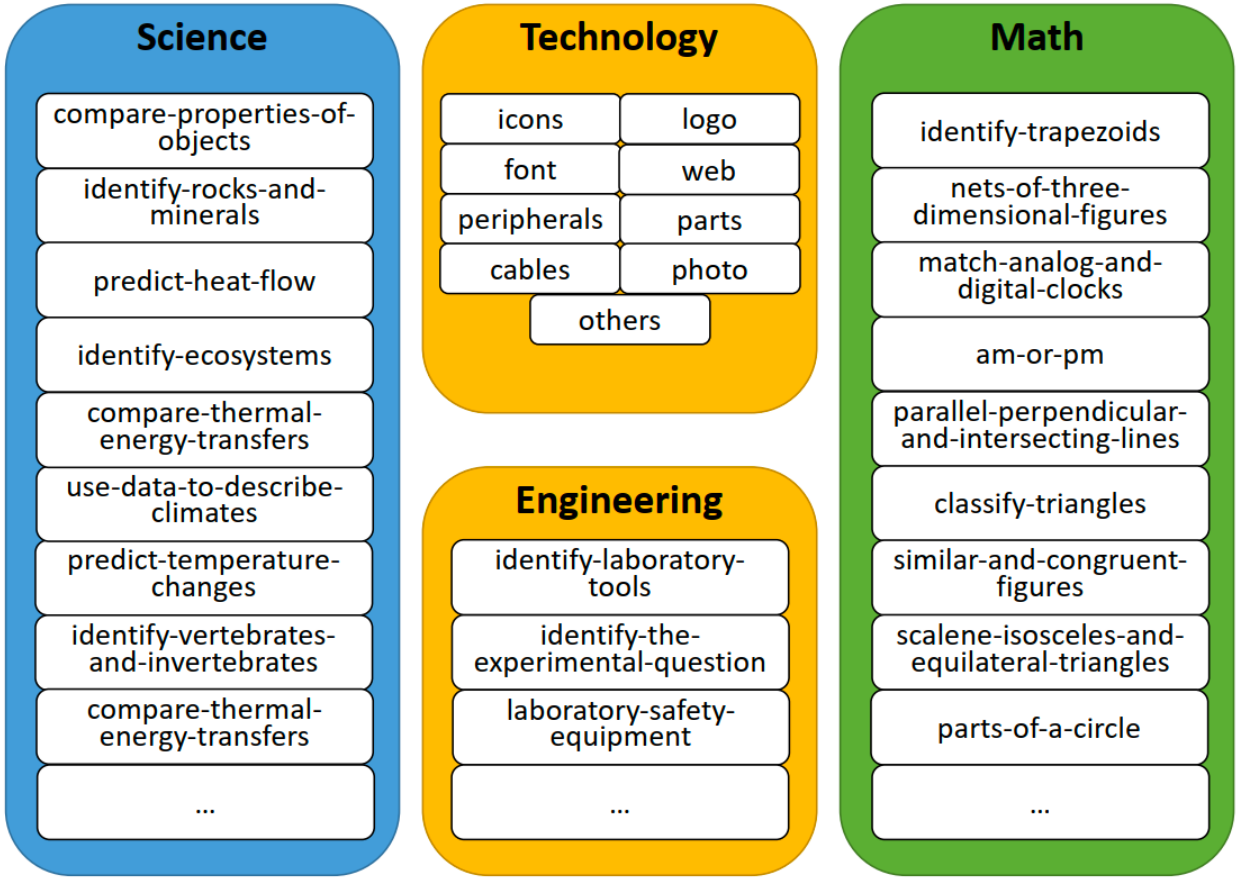
Our STEM Benchmark

Basic statistics of the STEM benchmark

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

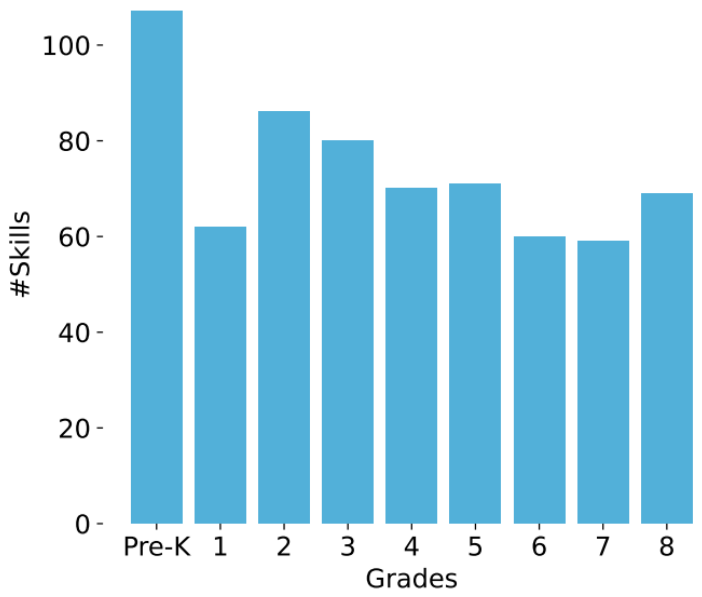
Our STEM benchmark is the largest multimodal STEM dataset in terms of number of skills and questions.

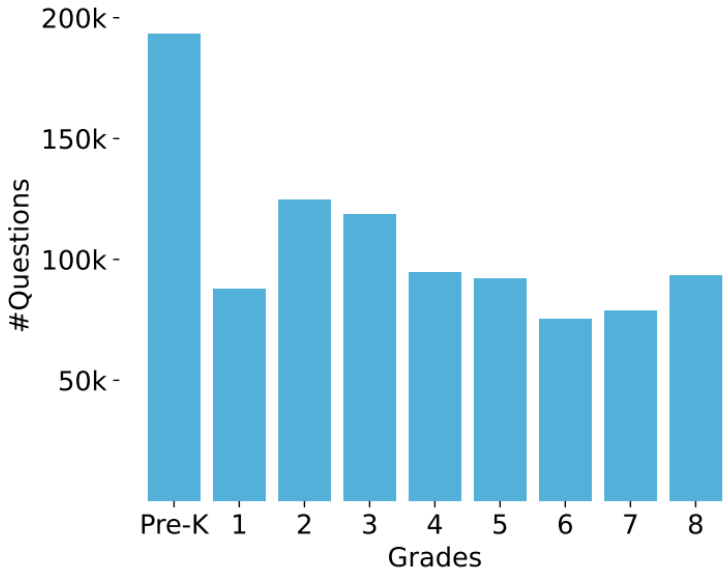
Skills in our STEM Benchmark



Our STEM benchmark covers a large number of fundamental STEM skills.

Our STEM Benchmark





We show the distribution of #Skills and #Questions over the K-12 grades.

Our STEM Benchmark

- Data source:

IXL

ProProfs Quiz Maker

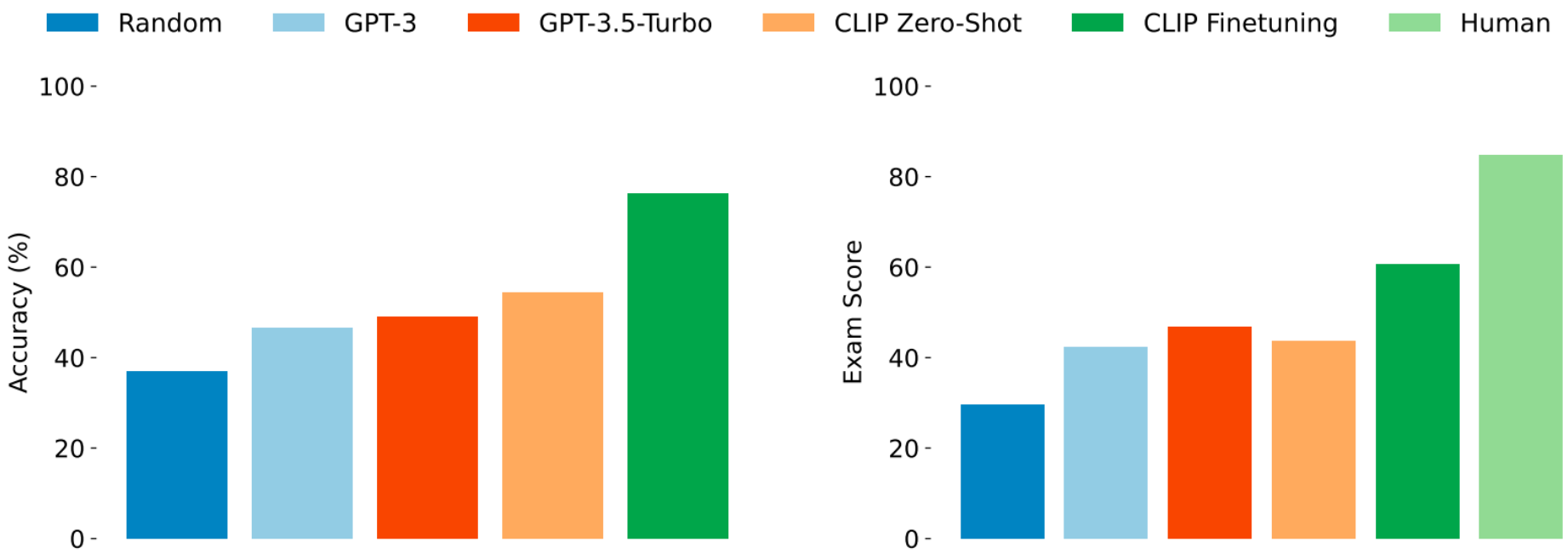
Triviaplaza

Human Performance Evaluation

- Exam Score:
  - Consider the progress of learning
  - Measure how well humans understand a STEM skill.
  - A score higher than 90.0 is considered excellent for a mastered skill.
- Accuracy: sampled 80 questions, human got 83.0% accuracy.

We enable model performance comparison to millions of elementary students.

Main Results

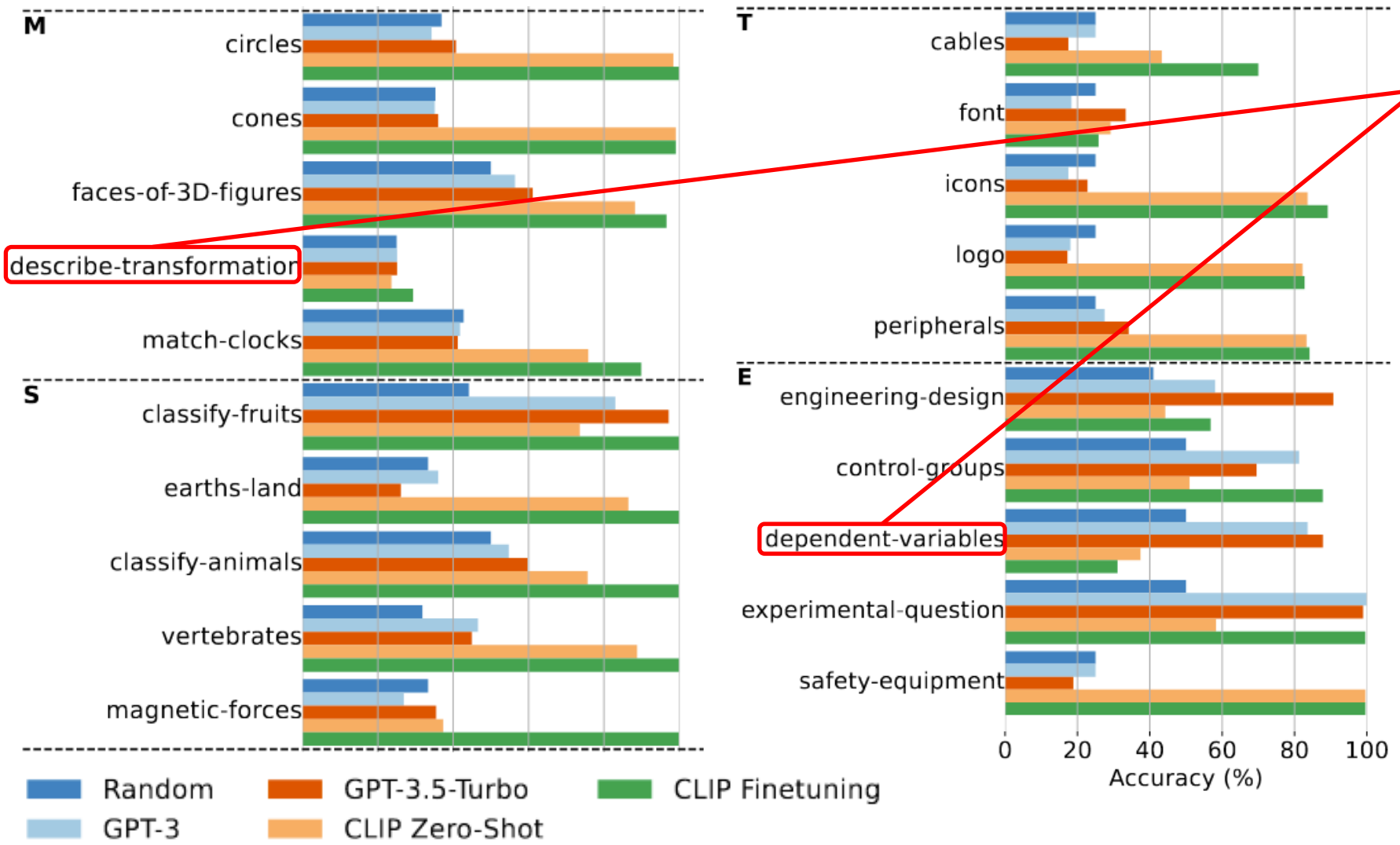


(i) Average accuracy of all subjects.

(ii) Average exam scores of all subjects.

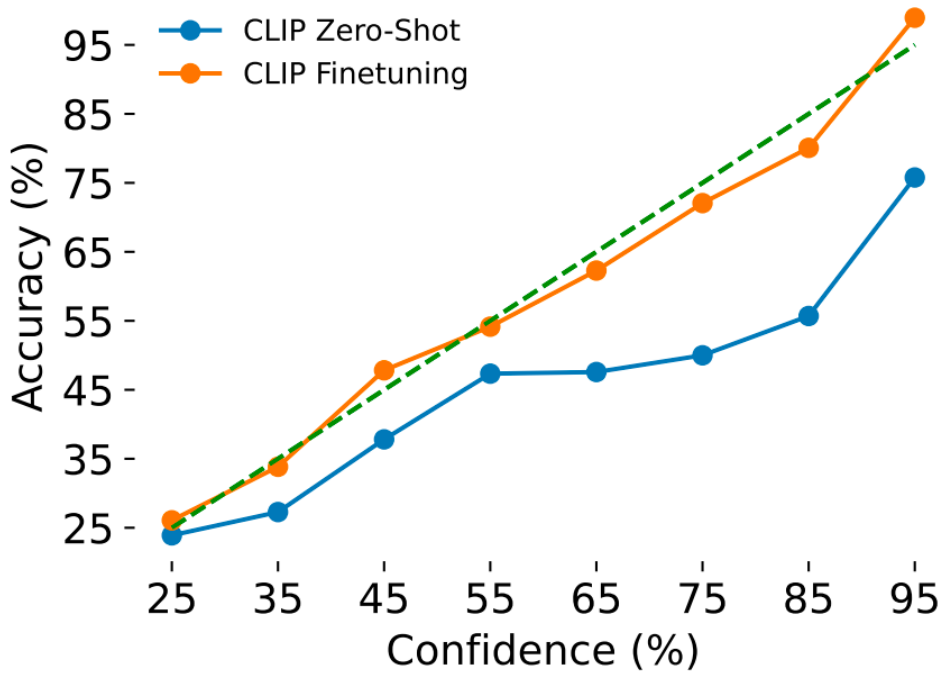
Model performance is still far behind that of average elementary students.

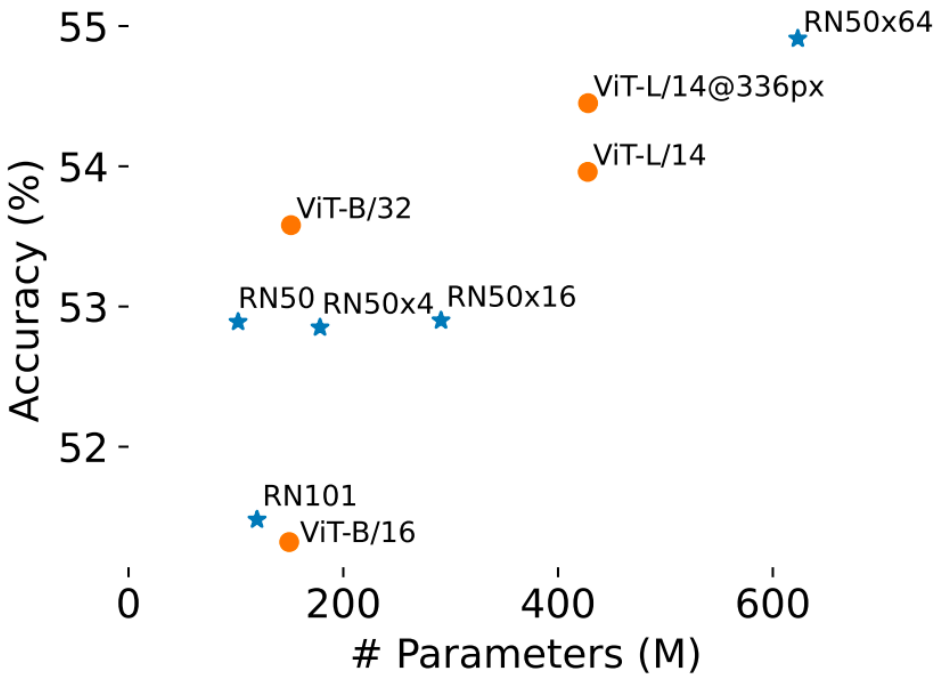
Skill Analysis



Challenging skills: abstract knowledge & complex reasoning

Calibration & model scaling





Fine-tuned CLIP model is well-calibrated.

Larger models are better.

Case study

Which shape is a circle?

Select the vertebrate. Hint: Mammals, birds, fish, reptiles, and amphibians are vertebrates.

(a) Correct predictions

What shape comes next?

Which animal's mouth is also adapted to get insects out of burrows?

(b) Incorrect predictions

Complex reasoning and abstract concepts are hard for neural models.

Conclusion

Our benchmark has the largest vision-language STEM dataset, which includes multimodal STEM subjects with 1,073,146 multi-choice questions and 448 skills.

STEM Benchmark

Conclusion

Our benchmark has the largest vision-language STEM dataset, which includes multimodal STEM subjects with 1,073,146 multi-choice questions and 448 skills.

We collected fundamental STEM skills from K-12 curriculum grades.

STEM Benchmark

Conclusion

Our benchmark has the largest vision-language STEM dataset, which includes multimodal STEM subjects with 1,073,146 multi-choice questions and 448 skills.

We collected fundamental STEM skills from K-12 curriculum grades.

STEM Benchmark

We evaluated foundation models including GPT-3.5-Turbo & CLIP.

Conclusion

Our benchmark has the largest vision-language STEM dataset, which includes multimodal STEM subjects with 1,073,146 multi-choice questions and 448 skills.

We collected fundamental STEM skills from K-12 curriculum grades.

STEM Benchmark

We evaluated foundation models including GPT-3.5-Turbo & CLIP.

We use the exam score to compare the model with millions of K-12 students results. The model performance is far behind human.

Thanks!

Dataset

Leaderboard

Paper

- Dataset: https://huggingface.co/datasets/stemdataset/STEM
- GitHub: https://github.com/stemdataset/STEM
- Leaderboard: https://huggingface.co/spaces/stemdataset/stem-leaderboard
- Arxiv: https://arxiv.org/abs/2402.17205