MEASURING VISION-LANGUAGE STEM SKILLS OF NEURAL MODELS

Jianhao Shen[1,2*], Ye Yuan[1,2,3*], Srbuhi Mirzoyan[1,2,3], Ming Zhang[1,2,3†], Chenguang Wang[4†]
[1]School of Computer Science, Peking University
[2]National Key Laboratory for Multimedia Information Processing, Peking University
[3]Peking University-Anker Embodied AI Lab
[4]Washington University in St. Louis
{jhshen,yuanye_pku,mzhang_cs}@pku.edu.cn, srbuhimirzoyan@stu.pku.edu.cn
chenguangwang@wustl.edu

ABSTRACT

We introduce a new challenge to test the STEM skills of neural models. The problems in the real world often require solutions, combining knowledge from STEM (science, technology, engineering, and math). Unlike existing datasets, our dataset requires the understanding of multimodal vision-language information of STEM. Our dataset features one of the largest and most comprehensive datasets for the challenge. It includes 448 skills and $1,073,146$ questions spanning all STEM subjects. Compared to existing datasets that often focus on examining expert-level ability, our dataset includes fundamental skills and questions designed based on the K-12 curriculum. We also add state-of-the-art foundation models such as CLIP and GPT-3.5-Turbo to our benchmark. Results show that the recent model advances only help master a very limited number of lower grade-level skills (2.5% in the third grade) in our dataset. In fact, these models are still well below (averaging 54.7%) the performance of elementary students, not to mention near expert-level performance. To understand and increase the performance on our dataset, we teach the models on a training split of our dataset. Even though we observe improved performance, the model performance remains relatively low compared to average elementary students. To solve STEM problems, we will need novel algorithmic innovations from the community.

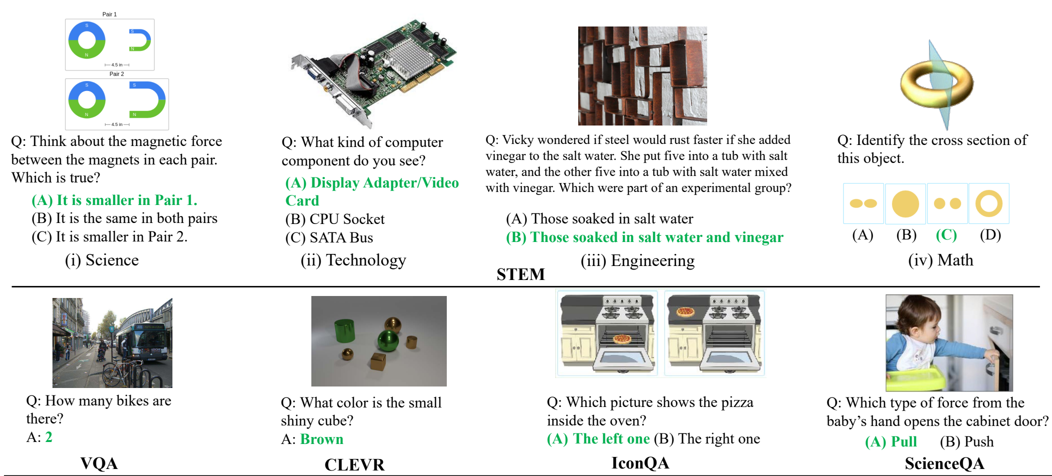
1 INTRODUCTION

STEM, namely, science, technology, engineering, and math, is the basis of solving a wide set of real-world problems. This helps solve hard problems to better understand the world and universe, such as modeling gravitational waves and protein structures, proving mathematics theorem, designing new principles for quantum computing, and engineering the James Webb telescope. Mirroring real-world scenarios, understanding multimodal vision-language information is vital to a great variety of STEM skills. For example, we are asked to compute the magnetic force given a diagram in physics. Geometry problems often require mathematical reasoning based on diagrams.

The challenges of the real world often require solutions that combine knowledge from STEM. Existing vision-language benchmarks, however, often concentrate on evaluating one of the STEM subjects. For example, IconQA (Lu et al., 2021b) and Geometry3K (Lu et al., 2021a) focus on evaluating mathematics understanding, while ScienceQA (Lu et al., 2022) examines science related skills. Other multimodal datasets such as VQA (Antol et al., 2015) and CLEVR (Johnson et al., 2017) are not specifically designed for STEM. Another set of benchmarks often includes textual STEM skill sets, where images are converted to LaTeX or formal languages (Hendrycks et al., 2021a;b).

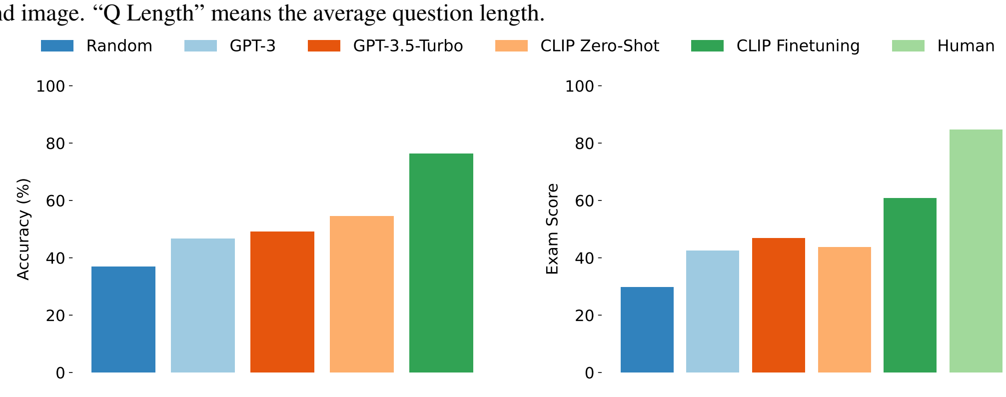
In this paper, we create a new challenge to test the STEM skills of neural models. We collect a large-scale multimodal dataset, called `STEM`, consisting of 448 skills and $1,073,146$ questions

[*] Equal contribution.
[†] Corresponding authors.



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

(a) Comparison between STEM and existing datasets. Upper: examples of STEM and other datasets. Lower: key statistics of STEM and other datasets. "#Questions", "#Images", "#Answers", "#Skills" denote the number of questions, images, answers, skills. "Multimodal" indicates whether every question of a dataset contains both text and image. "Q Length" means the average question length.



(i) Average accuracy of all subjects.

(ii) Average exam scores of all subjects.

(b) Neural model performance on STEM dataset.

Figure 1: Summary of our dataset and results.

spanning across all four STEM subjects. STEM provides the largest set of both skills and questions among existing datasets. Figure 1(a) shows the comparison of its key statistics with other datasets. The dataset consists of multi-choice questions, and Figure 1(a) shows an example for each subject. STEM is multimodal as we exclude a question if both the question and its answers are text. Each question consists of a question text with an optional image context. The corresponding answers to the question are either in text (Figure 1(a)(i)) or image (Figure 1(a)(iv)). The design of skills in STEM is important: we focus on fundamental skills based on the K-12 curriculum. This enables us to present a diverse and comprehensive STEM skill set. More importantly, this facilitates the understanding of neural models from different perspectives such as at skill level. We use IXL Learning (Learning, 2019) as our main data source to create STEM as it aligns best with our design principle.

The STEM dataset is challenging. Although our dataset focuses on the fundamentals of STEM, its multimodal nature makes it very difficult for modern neural models. Different from previous multimodal benchmarks, we include foundation models such as the state-of-the-art vision-language model, CLIP (Radford et al., 2021), and the large language model, GPT-3.5-Turbo (Ouyang et al.,

Table 1: STEM dataset statistics.

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

2022). While these models are able to advance the model performance compared to the near random-chance performance of previous neural models, they still drop the performance by averaging 54.7% compared to that of average elementary students. For example, the models are only capable of understanding 2.5% third-grade skills. Notably, our model results are evaluated quantitatively under the same real-world exam environment as humans. Instead of manual evaluation which is expensive, we simulate the conditions of IXL's online exams and use their scoring system to grade the model results. Compared to accuracy, this score (Bashkov et al., 2021) aims to measure humans' true understanding of skills by integrating the learning progress into the final score calculation. While the majority of existing benchmarks do not yet provide detailed meta information for analysis, the design of STEM supports deep performance analysis at different granularities, e.g., at a particular subject, skill, or grade level. For example, we show that basic math skills are still challenging for existing models. This is often due to the models failing to parse the images that are of great importance to mastering multimodal skills (e.g., geometry). To understand and increase the model performance on STEM, we teach models on a large-scale training split of STEM. However, the model performance still remains relatively low compared to general elementary students, not to mention near expert-level performance.

Our contributions are as follows. (i) We create a new dataset, called STEM, to benchmark the multimodal STEM skills of neural models. STEM is the largest dataset among existing datasets. Its design focuses on fundamental skills in the K-12 curriculum. This enables diverse and comprehensive tests across all STEM subjects. To facilitate future research, we also contribute a large-scale training set in STEM. STEM is challenging and useful to help advance models to solve more real-world problems. (ii) We benchmark a wide set of neural models including foundation models such as GPT-3.5-Turbo and CLIP on STEM. The meta information in STEM (e.g., skills and grades) supports a deeper understanding of model performance, and helps point out important shortcomings of existing models. (iii) We show current neural model performances are still far behind that of average elementary students in terms of STEM problem solving. We conclude important insights that suggest new algorithmic advancements from the community are necessary for understanding STEM skills.

2 The STEM Benchmark

2.1 Dataset

We create a massive dataset, called STEM to test the STEM problem solving abilities. Unlike existing benchmarks, STEM features a large-scale multimodal dataset covering all STEM subjects spanning science, technology, engineering, and mathematics. We split the dataset into a train set, a validation set, and a test set for model development and evaluation. The overall dataset statistics are included in Table 1. More details of STEM dataset are described in the appendix.

Attributes Our dataset includes the following key attributes to support deep analysis of model performances. (i) **Subjects.** There are four subjects in STEM, namely science, technology, engineering, and math. We follow this high-level concept to create our dataset. (ii) **Skills.** We design skills according to the U.S. National Education and California Common Core Content Standards. This design also aligns with the skill categorization of our data resources (details are below) and closely follows recent studies (Hendrycks et al., 2021b; Lu et al., 2021b). (iii) **Grades.** We use the grade information of our dataset resources in STEM. STEM does not contain grade information for the technology subset as its raw data does not provide the grade-level information. (iv) **Questions.** Each question in STEM is a multi-choice question and is multimodal. We exclude a question if both the question and its answers are text. Each question belongs to a particular skill, hence a subject.

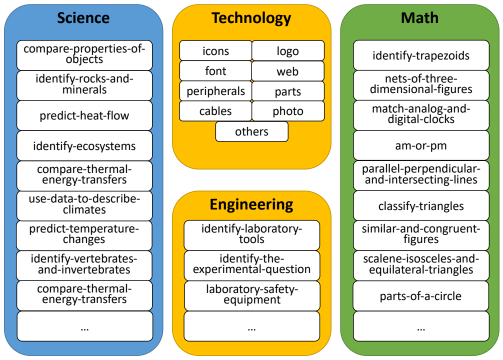


Figure 2: A summary of skills.

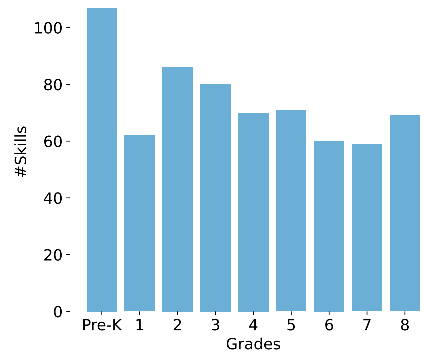


Figure 3: #Skills per grade.

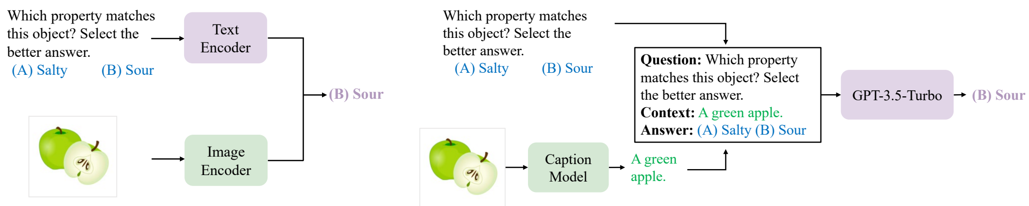
Science Science includes branches of domain knowledge focusing on testing reasoning abilities. Subject areas include biology, chemistry, physics and so on. Science tests specific domain knowledge, e.g., physics tests understanding of fundamental physics principles. It includes skills examining basics of science such as identifying properties of an object or calculating density. For example, to test the skill of comparing magnitudes of magnetic forces, an example question in Figure 1(a)(i) will be asked. We collect questions from IXL Science. Its skills and questions are designed based on U.S. National Education and California Common Core Content Standards. It includes questions from second grade to eighth grade. We also processed the data such as deduplicating questions and randomly shuffling the order of answers to each question. We exclude a question if both the question and its answers are text.

Technology Technology includes principles that test the knowledge of empirical methods. This subject mainly includes computer science. An example is included in Figure 1(a)(ii). It includes fundamental skills such as identifying parts of a computer or the basics of programming languages. We collect the questions from Triviaplaza Computer, which includes questions for tech interviews. To the best of our knowledge, `STEM` provides the first technology problem set for the multimodal test.

Engineering This engineering subset includes a skill set that covers fundamental engineering practices ranging from solving problems using magnets to exploring the design of spaceships. Figure 1(a)(iii) illustrates an example. The dataset is constructed based on the engineering portion of IXL. The skills and questions are ranging from third grade to eighth grade. To our knowledge, this subset is considered an early exploration on testing multimodal practical knowledge in engineering.

Mathematics Mathematics often requires reasoning and abstract knowledge. For example, solving math tests algebra generalization abilities. For example, the addition of numbers obeys the same rules everywhere. This subset includes fundamental math skills such as addition, algebra, comparison, counting, geometry and spatial reasoning. An example is shown in Figure 1(a)(iv). The questions are from IXL Math spanning from pre-K to eighth grade. To encode mathematical expressions, we use LaTeX to avoid unusual symbols or cumbersome formal languages.

Comparison with Existing Datasets `STEM` is the first large-scale mulitmodal STEM dataset. As shown in Figure 1(a), `STEM` provides the largest number of questions and skills among existing STEM related datasets. Compared to the previous largest multimodal STEM datasets, `STEM` is about 10 times larger in terms of the number of questions. `STEM` offers the most thorough fundamental skill and question set ranging from pre-K to eighth grade. Compared to datasets of a particular subject, `STEM` covers all STEM subjects and is at least competitive in terms of the number of questions and skills. For example, `STEM`'s math subset has 27 times more skills compared to the recent math benchmark (Lu et al., 2021b).



(a) CLIP. (b) GPT-3.5-Turbo.

Figure 4: Zero-shot model setups.

2.2 Analysis

To provide more insights into our dataset, we conduct the below analysis with a focus on the unique perspectives of STEM including skills and grades. Other dataset details such as question analysis are shown in the appendix.

Skills The design of STEM emphasizes diverse skills spanning all STEM subjects. Figure 2 presents a brief summary of the skills (a complete skill set is included in the appendix). STEM contains the largest skill set among existing datasets (Figure 1(a)). Each skill contains $2,395$ questions on average. A large number of new skills are introduced to STEM that are not yet covered by existing datasets, e.g., skills in technology and engineering. Besides, understanding multimodal information is crucial to these skills. For example, solving the geometry problem in Figure 1(a)(iv) is challenging since both the image and text contribute to the problem solving. Through this design, STEM helps to recognize important shortcomings of machine learning models by referring to difficult skills for these models.

Grades STEM is designed with a comprehensive K-12 curriculum to examine fundamentals of STEM. This leads to another unique feature of testing on STEM: we are able to obtain the grade-level performance of models. The majority of existing datasets aim to compare models with human experts e.g., solving competition-level questions (Hendrycks et al., 2021b; Zheng et al., 2022). However, thanks to the grade-level information provided by STEM we find that models are only competitive with first graders in understanding certain STEM skills. Figure 3 shows the total number of skills per grade of all subjects.

2.3 Models

We benchmark both state-of-the-art and foundation models on STEM including: multimodal (vision-language) models such as CLIP and language models such as GPT-3.5-Turbo.

Vision-Language Models

(i) **Zero-Shot.** We use CLIP (Radford et al., 2021), ViLBERT (Lu et al., 2019), 12-in-1 (Lu et al., 2020), UNITER (Chen et al., 2020b), and Virtex (Desai & Johnson, 2021) for the zero-shot evaluation of multimodal models. Multimodal models generally include two modules: an image encoder and a text encoder. CLIP is considered one of the state-of-the-art multimodal models. For zero-shot CLIP, we follow its original setup in Radford et al. (2021). Figure 4(a) illustrates an example. Other models follow the same zero-shot setup.

(ii) **Finetuning.** To test the learning ability of the models, we also finetune CLIP. We follow the linear probe setup presented in Radford et al. (2021). For each subject, we train the model on its entire training set as shown in Table 1 and select the best model on the validation set. At test time, the evaluation is the same as the zero-shot setup.

Language Models

(i) **Zero-Shot.** We use GloVe (Pennington et al., 2014), UnifiedQA (Khashabi et al., 2020), GPT-3 (Chen et al., 2020a) and GPT-3.5-Turbo (Ouyang et al., 2022) zero-shot for the language model evaluation. We formalize the task as a question answering task. We use the OpenAI API "text-davinci-002" and "gpt-3.5-turbo" corresponding to the best-performing GPT-3 and GPT-3.5-Turbo

Table 2: Results on STEM dataset. All evaluation scores are higher the better.

Model		Science	Technology	Engineering	Math	Average
Random Guesses		38.6	25.0	44.9	39.1	36.9
Language Models						
GloVe (Pennington et al., 2014)		38.0	25.2	48.1	39.0	37.6
UnifiedQA$_{\text{Small}}$ (Khashabi et al., 2020)		39.6	27.2	58.0	39.6	41.1
UnifiedQA$_{\text{Base}}$ (Khashabi et al., 2020)		42.6	28.8	55.4	40.0	41.7
GPT-3 (Brown et al., 2020)		47.1	22.1	73.5	44.0	46.7
GPT-3.5-Turbo		50.1	26.3	74.6	45.0	49.0
Vision-Language Models						
Virtex (Desai & Johnson, 2021)		37.5	24.0	48.1	38.9	37.1
12-in-1 (Lu et al., 2020)		39.4	27.5	44.2	41.9	38.3
ViLBERT (Lu et al., 2019)		39.0	32.1	44.2	42.7	39.5
UNITER (Chen et al., 2020b)		50.8	34.6	55.1	43.2	45.9
CLIP (Radford et al., 2021)	RN50	47.8	64.4	55.8	43.6	52.9
	RN101	50.3	65.3	46.7	43.7	51.5
	RN50x4	48.8	69.2	49.4	44.1	52.9
	RN50x16	49.8	66.1	51.4	44.3	52.9
	RN50x64	50.9	70.0	55.5	43.2	54.9
	ViT-B/32	48.3	63.7	59.5	42.8	53.6
	ViT-B/16	48.6	65.9	47.2	43.6	51.3
	ViT-L/14	49.8	68.6	54.3	43.1	54.0
	ViT-L/14@336px	50.3	68.7	55.1	43.6	54.4
	+Finetuning	87.0	71.9	67.7	78.4	76.3

respectively. We convert images to visual context text based on a captioning model following Lu et al. (2022). Figure 4(b) shows an example. All language models follow the same setup.

2.4 Metrics and Human Performance

We report accuracy on the test set of each subject. We use accuracy as the evaluation metric since all questions in our dataset are multiple-choice questions. We also compute macro average accuracy across the test sets of all subjects. Unlike the micro evaluation setting, this score relieves data or class imbalance issues. In addition, we focus on two kinds of evaluations for human performance comparison purposes. (i) Exam score. In particular, for science, engineering, and math, we use the IXL SmartScore (Learning, 2019). Different from accuracy, SmartScore considers the progress of learning and is designed to measure how well humans understand a STEM skill (Bashkov et al., 2021). It starts at 0, increases as students answer questions correctly, and decreases if questions are answered incorrectly. We simulate the conditions of its real online exams. The final score is graded by IXL's SmartScore system. According to IXL (IXL, b;a), a score higher than 90.0 is considered excellent for a mastered skill. Therefore, we use this score as a reference to human performance. For technology, we use the average human accuracy available at Triviaplaza. The average accuracy is 68.6. (ii) Accuracy. We sampled 80 questions from our test sets (20 questions for each subject) and collected the responses from seven university students. They attained an average accuracy of 83.0 on all subjects. All evaluation scores are higher the better.

3 Experiments

In this section, we show the performance of a wide set of neural models as well as humans on STEM. The results show that state-of-the-art foundation models like CLIP and GPT-3.5-Turbo still underperform general elementary students. The details of the experimental setup, additional results and analysis are described in the appendix.

3.1 Main Results

Zero-Shot The results are shown in Table 2. We first test language models to see whether models that only understand text are proficient at the multimodal skills in STEM. GloVe has near random-chance accuracy. This means that STEM cannot be solved by simply matching the text semantic similarity between questions and answers. UnifiedQA does slightly better than GloVe with an improvement of averaging 4.1% points. GPT-3.5-Turbo performs the best among these language models, reaching 49.0% accuracy on average. Both foundation models (GPT-3.5-Turbo and GPT-3) perform well in engineering. This is mainly because engineering practices are mainly described in the text (see Figure 1(a)(iii)). Recent advancements in large language models help dramatically improve text understanding capabilities. However, large language models still struggle in other subjects. This implies that the understanding of both vision and language information is essential to STEM skills.

Next, we examine vision-language models. We find that the performance of Virtex, 12-in-1, and ViLBERT is nearing the performance of random guesses. These models capture very limited knowledge of STEM subjects. On the other hand, UNITER and CLIP show significant improvements over the random-chance accuracy. Specifically, CLIP-RN50x64 achieves the best result on STEM. It achieves 18.0% points improvements over random guesses. Notably, CLIP-RN50x64 outperforms GPT-3.5-Turbo by 5.9% points. This shows that CLIP has a basic understanding of multimodal STEM skills. Its vision understanding ability certainly contributes to this performance. Among all subjects, we see only marginal improvements in math. This applies to all foundation models. In addition, the result implies that math is the most challenging subject for current neural models. Novel algorithm advancements that can enable strong reasoning ability are necessary to solve math problems.

Finetuning The results are shown in Table 2. It is encouraging as finetuning CLIP ViT-L/14@336px is able to significantly boost the performance on science and math by averaging 30% points over its zero-shot setting. The performance improvements on other subjects are 7.9% points, which is much smaller. While having a large amount of training data helps to some extent, the finetuning performance is still far behind that of an average elementary student (the human-level performance is presented in Sec. 3.3). This indicates that more fundamental advancements are required to solve STEM questions in the STEM dataset. For simplicity, we use CLIP to represent CLIP ViT-L/14@336px in the rest of this section.

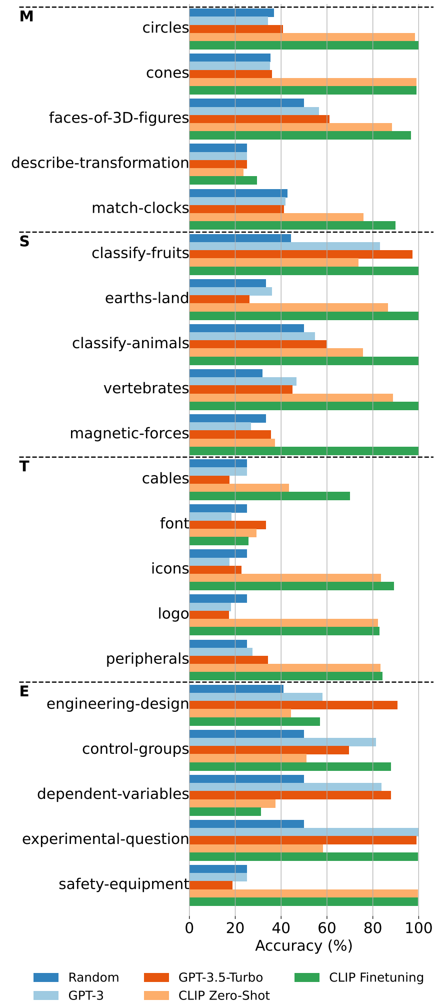


Figure 5: Results categorized by sampled skills of each subject. M: math. S: science. T: technology. E: engineering. Full results are in the appendix.

3.2 Results Analysis

Skills As STEM provides massive skills, analyzing models' performance at the skill level helps understand models better. We show the performance of foundation models (GPT-3, GPT-3.5-Turbo, and CLIP) on an uncurated set of skills of each subject in Figure 5. We find that these foundation models are able to perform well zero-shot on skills focusing on identifying common objects (e.g., classifying fruits). However, zero-shot and finetuned foundation models all fail in challenging skills that require abstract knowledge and complex reasoning (e.g., describing transformation).

Grades Intuitively, questions for higher graders are more difficult than those for lower graders. We illustrate the grade-level model performance to investigate if the same trend holds for neural models as well. We show the exam scores of models along each grade in Figure 6. Surprisingly, there is no obvious performance drop as the increase in grade levels. This implies the learning curve for neural models may be different from that of humans. A reason is that neural models are trained on data including all grade-level questions simultaneously while humans

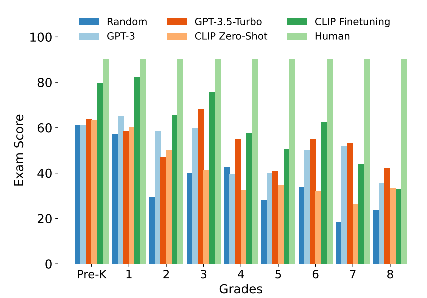


Figure 6: Average grade-level exam scores.

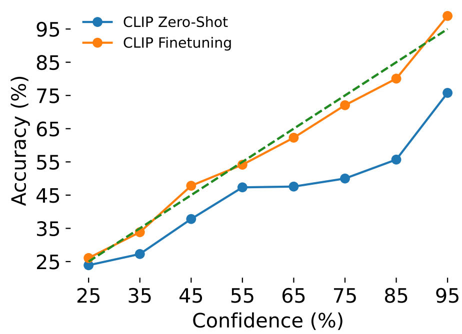


Figure 7: CLIP calibration results.

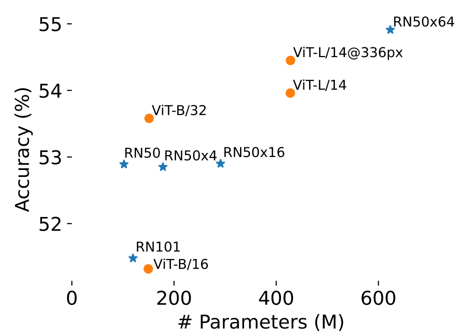
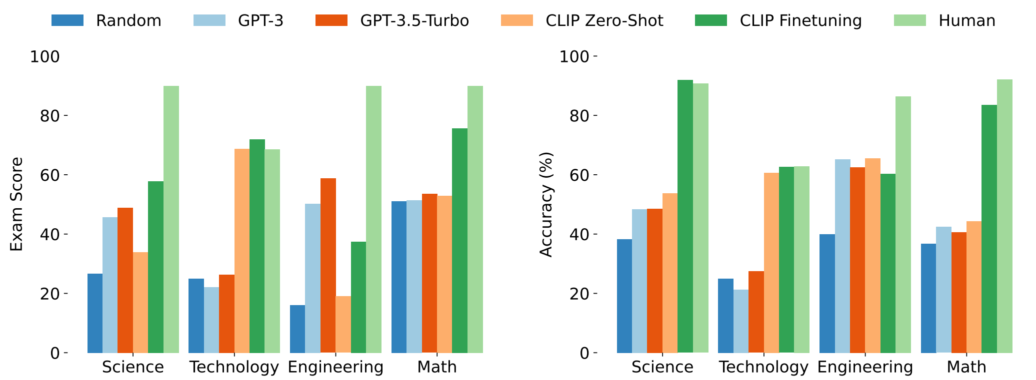


Figure 8: Zero-shot CLIP model scaling results.



(a) Exam scores on each subject.

(b) Accuracies of a real-world test on a subset of STEM.

Figure 9: Comparison between models and humans.

gradually learn from lower to higher grade-level questions. Also, the average exam scores of elementary grades (grades 1-6) equals 40.8, which is 54.7% lower than human reference (i.e., 90).

Calibration A trustworthy model should be calibrated. This means that its confidence should approximately match the actual probability of the prediction being correct (Guo et al., 2017a). However modern neural networks are often not well calibrated (Nguyen et al., 2015; Guo et al., 2017b). We show the relationship between the confidence of CLIP and the corresponding accuracy in Figure 7. We use the softmax probability as the confidence. We observe that the zero-shot CLIP model is not well calibrated. In fact, it is overconfident about its predictions and is only loosely related to its actual accuracy. After finetuning, CLIP is more calibrated. The results suggest that further improving calibration on STEM is another promising direction.

Scaling Laws Figure 8 shows the average accuracy of zero-shot CLIP with different model sizes. As expected, the performance improves as models grow larger. But the performance also saturates. This implies that other than increasing model scales, new advancements in model design or training schema are required to improve the performance on STEM.

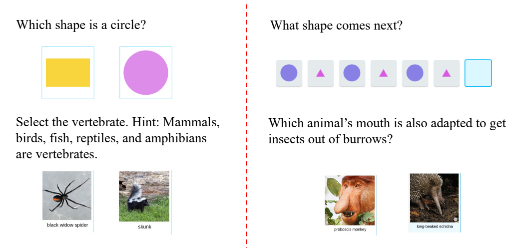
3.3 Comparison with Human

In this section, we explore whether the best-performing foundation models namely CLIP, GPT-3, and GPT-3.5-Turbo are nearing human-level performance.

Figure 9(a) shows the exam scores (Sec. 2.4) of models and humans on each subject. A score of 90 means a student is proficient in the subject. The zero-shot performances of all tested neural models are well below that bar. In technology, CLIP finetuning achieves human-level performance. This is mainly because most technology skills are about specific empirical knowledge, which is learnable for neural models after finetuning. Overall, there is still a large performance gap between general neural models and average elementary students even in understanding the fundamental skills in STEM. In addition, the offline real-world test-takers (Sec. 2.4) produce similar outputs with the above online setup on a subset of questions in the STEM. The results are shown in Figure 9(b).

3.4 Case Study

We show examples of GPT-3.5-Turbo predictions in Figure 10. We show an example of correct and incorrect predictions respectively. For the correct ones, the corresponding skills are mainly about the basics, such as names of objects (e.g., shapes or animals). The incorrect predictions are mainly due to the complex nature of skills. These skills are often about abstract concepts such as symmetry and the direction of force. They are also more relevant to logical reasoning, such as finding patterns or inferring the function of animal adaption.



(a) Correct predictions **(b) Incorrect predictions**

Figure 10: Examples of GPT-3.5-Turbo predictions.

4 Related Work

There are various types of vision-language tasks, such as reference resolution (Kazemzadeh et al., 2014), image captioning or tagging (Thomee et al., 2016; Sharma et al., 2018), image-text retrieval (Lin et al., 2014; Plummer et al., 2015), visual question answering (Antol et al., 2015; Goyal et al., 2017; Zhang et al., 2016; Zhu et al., 2016), and visual reasoning (Suhr et al., 2017; Johnson et al., 2017). Our STEM differs from the previous datasets in that it covers diverse fundamentals of STEM and requires both multimodal understanding and domain knowledge in STEM. This makes STEM a natural testbed to evaluate the real-world problem solving abilities of machine learning models.

Existing STEM related benchmarks do not cover all STEM skills for multimodal understanding. There are benchmarks targeting math (Saxton et al., 2019; Hendrycks et al., 2021b; Zheng et al., 2022; Lu et al., 2021a;b; Xiong et al., 2023b). PIQA (Bisk et al., 2020) is a benchmark for physical commonsense understanding. ScienceQA (Lu et al., 2022) is a multimodal dataset for general science. MMLU (Hendrycks et al., 2021a) contains 57 tasks including STEM but is only restricted to single text modality. Our STEM is the first to include all STEM subjects for vision-language understanding.

Pretrained foundation models help achieve state-of-the-art performance in both NLP and computer vision tasks. Pretrained language models (Radford et al., 2018; 2019; Devlin et al., 2019), especially the recent large language models (Chen et al., 2020a; Wang et al., 2020; 2022a; Ouyang et al., 2022; Crispino et al., 2023; OpenAI, 2023; Chowdhery et al., 2022) have significantly advanced the performance in general natural language understanding tasks. Based on these models, various techniques (Shen et al., 2022a;b; Imani et al., 2023; Jiang et al., 2023; Wang et al., 2023; Xiong et al., 2023a; Pan et al., 2024b;a) have been developed to address specific challenges in a domain such as math. We focus on testing the basic STEM ability of state-of-the-art models in a zero-shot setting and identifying room for improvement by referring to our finetuning results. CLIP (Radford et al., 2021) is one of the state-of-the-art pretrained vision-language models (Lu et al., 2019; Krishna et al., 2017; Chen et al., 2020b; Desai & Johnson, 2021; Lu et al., 2020). Other similar models include GLIP (Li et al., 2022b), GLIDE (Nichol et al., 2022), OFA (Wang et al., 2022b), and BLIP (Li et al., 2022a; 2023). We use CLIP in our test while the majority of existing benchmarks have not explored it yet.

5 Conclusion

We introduce STEM, a new challenge to examine the STEM skills of neural models. STEM is the largest multimodal benchmark for this challenge. It consists of a large number of multi-choice questions and skills spanning all STEM subjects. STEM focuses on fundamentals of STEM based on the K-12 curriculum. We also include state-of-the-art foundation models such as GPT-3.5-Turbo and CLIP for evaluations. The benchmark results suggest that current neural model performances are still far behind that of elementary students. STEM poses unique challenges for the research community to develop fundamental algorithmic advancements. We hope our benchmark will foster future research in multimodal understanding.

Ethics Statement

We hereby acknowledge that all of the co-authors of this work are aware of the provided *ICLR Code of Ethics* and honor the code of conduct. We collected data from several sources, and we cited the data creators. The copyright belongs to the original data owners. The STEM dataset is under the CC BY-NC-SA 4.0 license (Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International) and is used for non-commercial research purposes. The collected data does not contain any personally identifiable information or offensive content. Our dataset is mainly built upon instances from real-world exam data. Therefore it was less likely to contain sensitive data. We evaluate foundation models, for which the risks and potential harms are discussed (Brown et al., 2020; Radford et al., 2021).

Acknowledgements

This paper is partially supported by the National Key Research and Development Program of China with Grant No. 2023YFC3341203 as well as the National Natural Science Foundation of China with Grant No.62276002.

References

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *CVPR*, pp. 6077–6086, 2018.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. VQA: visual question answering. In *ICCV*, pp. 2425–2433, 2015.

Bozhidar M Bashkov, Kate Mattison, and Lara Hochstein. Ixl design principles. 2021.

Yonatan Bisk, Rowan Zellers, Ronan LeBras, Jianfeng Gao, and Yejin Choi. PIQA: reasoning about physical commonsense in natural language. In *AAAI*, pp. 7432–7439, 2020.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey E. Hinton. Big self-supervised models are strong semi-supervised learners. In *NeurIPS*, 2020a.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In *ECCV*, pp. 104–120, 2020b.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. Palm: Scaling language modeling with pathways. *CoRR*, abs/2204.02311, 2022.

Nicholas Crispino, Kyle Montgomery, Fankun Zeng, Dawn Song, and Chenguang Wang. Agent instructs large language models to be general zero-shot reasoners. *arXiv preprint arXiv:2310.03710*, 2023.

Karan Desai and Justin Johnson. Virtex: Learning visual representations from textual annotations. In *CVPR*, pp. 11162–11173, 2021.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio (eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies,*, pp. 4171–4186, 2019.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2020.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the V in VQA matter: Elevating the role of image understanding in visual question answering. In *CVPR*, pp. 6325–6334, 2017.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *ICML*, pp. 1321–1330, 2017a.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International conference on machine learning*, pp. 1321–1330. PMLR, 2017b.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pp. 770–778, 2016.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *ICLR*, 2021a.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the math dataset. *NeurIPS*, 2021b.

Shima Imani, Liang Du, and Harsh Shrivastava. Mathprompter: Mathematical reasoning using large language models. *CoRR*, abs/2303.05398, 2023.

IXL. Understanding the ixl smartscore. `https://blog.ixl.com/wp-content/uploads/2014/11/SmartScore-guide.pdf`, a.

IXL. How does the smartscore work? `https://www.ixl.com/help-center/article/1272663/how_does_the_smartscore_work`, b.

Albert Qiaochu Jiang, Sean Welleck, Jin Peng Zhou, Timothée Lacroix, Jiacheng Liu, Wenda Li, Mateja Jamnik, Guillaume Lample, and Yuhuai Wu. Draft, sketch, and prove: Guiding formal theorem provers with informal proofs. In *ICLR*, 2023.

Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C. Lawrence Zitnick, and Ross B. Girshick. CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning. In *CVPR*, pp. 1988–1997, 2017.

Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara L. Berg. Referitgame: Referring to objects in photographs of natural scenes. In *ACL*, pp. 787–798, 2014.

Daniel Khashabi, Sewon Min, Tushar Khot, Ashish Sabharwal, Oyvind Tafjord, Peter Clark, and Hannaneh Hajishirzi. Unifiedqa: Crossing format boundaries with a single QA system. In *Findings of EMNLP*, pp. 1896–1907, 2020.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *IJCV*, pp. 32–73, 2017.

IXL Learning. The impact of ixl math and ixl ela on student achievement in grades pre-k to 12 (pp. 1–27), 2019.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pp. 12888–12900. PMLR, 2022a.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023.

Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, Kai-Wei Chang, and Jianfeng Gao. Grounded language-image pre-training. In *CVPR*, pp. 10955–10965, 2022b.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, pp. 740–755, 2014.

Chengwu Liu, Jianhao Shen, Huajian Xin, Zhengying Liu, Ye Yuan, Haiming Wang, Wei Ju, Chuanyang Zheng, Yichun Yin, Lin Li, Ming Zhang, and Qun Liu. Fimo: A challenge formal dataset for automated theorem proving, 2023.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *NeurIPS*, pp. 13–23, 2019.

Jiasen Lu, Vedanuj Goswami, Marcus Rohrbach, Devi Parikh, and Stefan Lee. 12-in-1: Multi-task vision and language representation learning. In *CVPR*, 2020.

Pan Lu, Ran Gong, Shibiao Jiang, Liang Qiu, Siyuan Huang, Xiaodan Liang, and Song-Chun Zhu. Inter-GPS: Interpretable geometry problem solving with formal language and symbolic reasoning. In *ACL-IJCNLP*, pp. 6774–6786, 2021a.

Pan Lu, Liang Qiu, Jiaqi Chen, Tony Xia, Yizhou Zhao, Wei Zhang, Zhou Yu, Xiaodan Liang, and Song-Chun Zhu. Iconqa: A new benchmark for abstract diagram understanding and visual language reasoning. In *NeurIPS*, 2021b.

Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *NeurIPS*, 2022.

Anh Nguyen, Jason Yosinski, and Jeff Clune. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 427–436, 2015.

Alexander Quinn Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. GLIDE: towards photorealistic image generation and editing with text-guided diffusion models. In *ICML*, pp. 16784–16804, 2022.

OpenAI. GPT-4 technical report. *CoRR*, abs/2303.08774, 2023.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35: 27730–27744, 2022.

Yu Pan, Ye Yuan, Yichun Yin, Jiaxin Shi, Zenglin Xu, Ming Zhang, Lifeng Shang, Xin Jiang, and Qun Liu. Preparing lessons for progressive training on language models. *arXiv preprint arXiv:2401.09192*, 2024a.

Yu Pan, Ye Yuan, Yichun Yin, Zenglin Xu, Lifeng Shang, Xin Jiang, and Qun Liu. Reusing pretrained models by multi-linear operators for efficient training. *Advances in Neural Information Processing Systems*, 36, 2024b.

Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. Kosmos-2: Grounding multimodal large language models to the world, 2023.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. Glove: Global vectors for word representation. In *EMNLP*, pp. 1532–1543, 2014.

Bryan A. Plummer, Liwei Wang, Chris M. Cervantes, Juan C. Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *ICCV*, pp. 2641–2649, 2015.

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. 2018.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 2019.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, pp. 8748–8763, 2021.

David Saxton, Edward Grefenstette, Felix Hill, and Pushmeet Kohli. Analysing mathematical reasoning abilities of neural models. In *ICLR*, 2019.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *ACL*, pp. 2556–2565, 2018.

Da Shen, Xinyun Chen, Chenguang Wang, Koushik Sen, and Dawn Song. Benchmarking language models for code syntax understanding. In *EMNLP*, 2022a.

Jianhao Shen, Chenguang Wang, Ye Yuan, Jiawei Han, Heng Ji, Koushik Sen, Ming Zhang, and Dawn Song. Palt: Parameter-lite transfer of language models for knowledge graph completion. In *EMNLP*, 2022b.

Alane Suhr, Mike Lewis, James Yeh, and Yoav Artzi. A corpus of natural language for visual reasoning. In *ACL*, pp. 217–223, 2017.

Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. Eva-clip: Improved training techniques for clip at scale, 2023.

Bart Thomee, David A. Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. YFCC100M: the new data in multimedia research. *Commun. ACM*, pp. 64–73, 2016.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *NeurIPS*, 2017.

Chenguang Wang, Xiao Liu, and Dawn Song. Language models are open knowledge graphs. *arXiv preprint arXiv:2010.11967*, 2020.

Chenguang Wang, Xiao Liu, Zui Chen, Haoyun Hong, Jie Tang, and Dawn Song. Deepstruct: Pretraining of language models for structure prediction. In *ACL*, 2022a.

Haiming Wang, Ye Yuan, Zhengying Liu, Jianhao Shen, Yichun Yin, Jing Xiong, Enze Xie, Han Shi, Yujun Li, Lin Li, Jian Yin, Zhenguo Li, and Xiaodan Liang. DT-solver: Automated theorem proving with dynamic-tree sampling guided by proof-level value function. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12632–12646, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.706. URL `https://aclanthology.org/2023.acl-long.706`.

Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In *International Conference on Machine Learning*, pp. 23318–23340. PMLR, 2022b.

Jing Xiong, Zixuan Li, Chuanyang Zheng, Zhijiang Guo, Yichun Yin, Enze Xie, Zhicheng Yang, Qingxing Cao, Haiming Wang, Xiongwei Han, Jing Tang, Chengming Li, and Xiaodan Liang. Dq-lore: Dual queries with low rank approximation re-ranking for in-context learning, 2023a.

Jing Xiong, Jianhao Shen, Ye Yuan, Haiming Wang, Yichun Yin, Zhengying Liu, Lin Li, Zhijiang Guo, Qingxing Cao, Yinya Huang, Chuanyang Zheng, Xiaodan Liang, Ming Zhang, and Qun Liu. TRIGO: Benchmarking formal mathematical proof reduction for generative language models. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 11594–11632, Singapore, December 2023b. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.711. URL `https://aclanthology.org/2023.emnlp-main.711`.

Peng Zhang, Yash Goyal, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Yin and yang: Balancing and answering binary visual questions. In *CVPR*, pp. 5014–5022, 2016.

Kunhao Zheng, Jesse Michael Han, and Stanislas Polu. minif2f: a cross-system benchmark for formal olympiad-level mathematics. In *ICLR*, 2022.

Yuke Zhu, Oliver Groth, Michael S. Bernstein, and Li Fei-Fei. Visual7w: Grounded question answering in images. In *CVPR*, pp. 4995–5004, 2016.