Measuring Vision-Language STEM Skills of Neural Models

Anonymous ACL submission

Abstract

We introduce a new challenge to test the STEM skills of neural models. Unlike existing datasets, our dataset requires the understanding of multimodal vision-language information. Our dataset includes 448 skills and $1,073,146$ questions spanning all STEM (science, technology, engineering, math) subjects. The dataset is considered one of the largest and most comprehensive datasets for the test. Compared to existing datasets that often focus on expert-level ability testing, our test includes fundamental skills and questions designed according to the K-12 curriculum. We also add state-of-the-art models such as CLIP and GPT-3 to our test. Results show that the recent model advances only help master a very small portion of the low grade-level skills (2.5% in the third grade) in our dataset. In fact, these models significantly underperform elementary students by on average 54.7%, not to mention to meet expert-level performance. To understand and increase the performance on our dataset, we teach the models using a training split of our dataset. Even though we are able to obtain improved performance, our results show that the model performance remains relatively low compared to average elementary students. To solve STEM problems, we will need novel algorithmic innovations from the broader research community. The code and dataset are available at `https://anonymous.4open.science/r/STEM` and will be made publicly available.

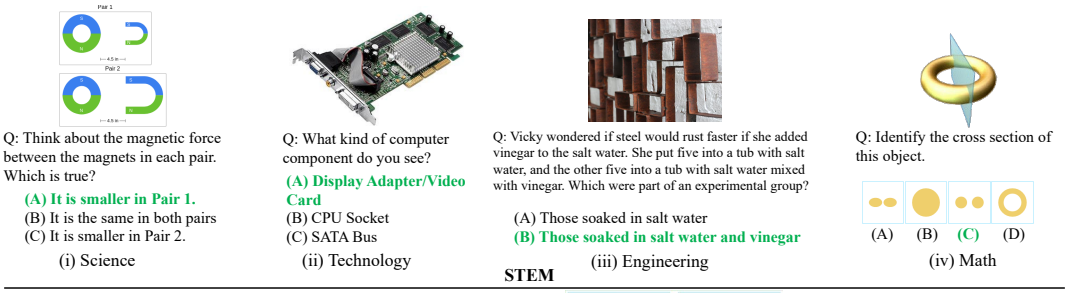
1 Introduction

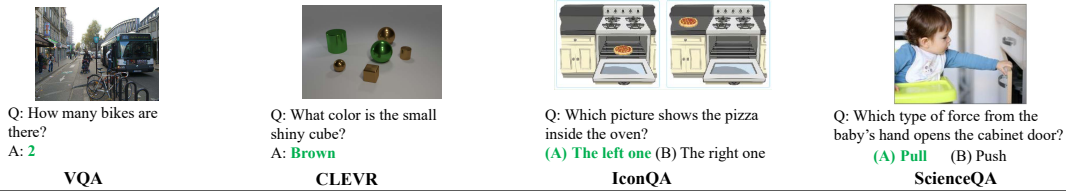
STEM, namely, science, technology, engineering, and math, is the basis of solving a wide set of real-world problems. This helps explore scientific hard problems to better understand the world and universe, such as modeling gravitational waves and protein structures, proving mathematics theorem, designing 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.

To understand the multimodal STEM problem solving ability of neural networks, existing vision-language benchmarks often concentrate on evaluating one of the STEM subjects. For example, IconQA (Lu et al., 2021b) and Geometry3K (Lu et al., 2021a) only evaluate the 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 only 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 spanning across all four STEM skills. `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 includes multi-choice questions, and Figure 1(a) shows an example for each subject. A question in `STEM` is multimodal since each question consists of a question text with an optional image context. And the corresponding answers to the question are either in the 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 have a diverse and comprehensive STEM skill set. More importantly, this facilitates the understanding of neural models. We use IXL Learning (Learning, 2019) as our main data source to build `STEM` as it aligns best with our design principle.

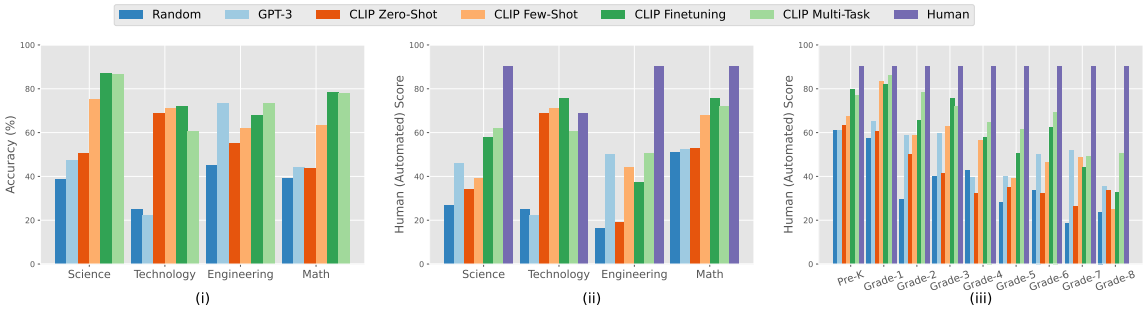
The `STEM` dataset is challenging. Even though





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 all questions of a dataset include text and image. "Q Length" means the average question length.



(b) Neural model performance on STEM dataset. (i): accuracy on all subjects. (ii): human scores on all subjects. (iii) average human scores on each grade of all subjects.

Figure 1: Summary of our dataset and results.

our dataset includes only fundamentals of STEM, we find its multimodal nature makes it very difficult for contemporary neural models. Different from previous multimodal benchmarks, we include both the state-of-the-art multimodal vision-language model, CLIP (Radford et al., 2021), and the language model, GPT-3 (Chen et al., 2020a). While these models are able to advance the model performance compared to the near random-chance performance of previous neural models, their model performances drop by 54.7% compared to that of average elementary students. For example, the models are only capable of 2.5% third grade skills. Notably, our human comparison is based on a quantitative evaluation. Instead of manual evaluation which is expensive, we utilize IXL's online scoring system to produce the human scores at scale. Compared to accuracy, this score is considered the best available score to measure humans' true understanding of a skill by considering the learning progress. While the majority of existing benchmarks do not yet provide the meta information for analysis, the design of STEM supports deep analysis at different granularities, e.g., a particular subject, skill, or grade-level question that a model is capable of. For example, we show that math is particularly challenging for modern neural networks. This is due to these models often failing to deeply understand the images that are of great importance to skills (e.g., geometry). To understand and increase the model performance on STEM, we teach models using a large-scale training split of STEM. However, the model performance remains relatively low compared to general elementary students, not to

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

Table 1: STEM dataset statistics.

mention meets 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 following the K-12 curriculum that enables diverse and comprehensive skill 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 state-of-the-art GPT-3 and CLIP on STEM. The meta information in by STEM (e.g., skills, 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 STEM problem solving. We suggest that we should not get either too excited or too worried about the current status of mainstream models. We conclude important insights that suggest new algorithmic advancements from the broader research 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 collected $1,073,146$ multi-choice questions in total. Example questions are shown in Figure 1(a). Moreover, STEM provides a comprehensive STEM skill set containing 448 skills across the subjects. Figure 2 shows example STEM skills in our dataset. We split the dataset into a train set, a validation set, and a test set for model development and evaluation (Details are in Sec. 2.3). The overall dataset statistics are included in Table 1. More details of STEM dataset such as the complete skill set and examples are described in Appendix K.

Science Science includes branches of domain knowledge that examine 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. The science portion includes skills examining basics of science such as identifying properties of an object, calculating density (Appendix K shows a complete skill set). For example, to test the skill of comparing magnitudes of magnetic forces, Figure 1(a)(i) is an example question. We automatically collect questions from IXL Science (details are in Appendix H). Its skills and questions are designed based on U.S. National Education and California Common Core Content Standards. 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. This results in $186,740$ questions and 82 skills (Table 1) from second grade to eighth grade.

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 basics of programming languages (Appendix K). We automatically extract the questions from Triviaplaza Computer [1] to build this dataset. The dataset includes questions for tech interviews. After the same data processing procedure with the science subset, we in total have $8,566$ questions and 9 skills (Table 1). To the best of our knowledge, STEM provides the first technology problem set for the multimodal test.

Engineering Engineering examines the fluid intelligence of engineering practices. This subset includes a skill set that covers fundamental engineering practices ranging from solving problems using magnets to exploring design about traveling to other planets (Appendix K). Figure 1(a)(iii) illustrates an example. The dataset is constructed based on the engineering portion of IXL. After data processing, this subset contains $18,981$ questions and 6 skills (Table 1). The skills and questions range from third grade to eighth grade. To our best knowledge, this subset is considered an early exploration testing multimodal engineering practical knowledge.

[1] https://www.triviaplaza.com

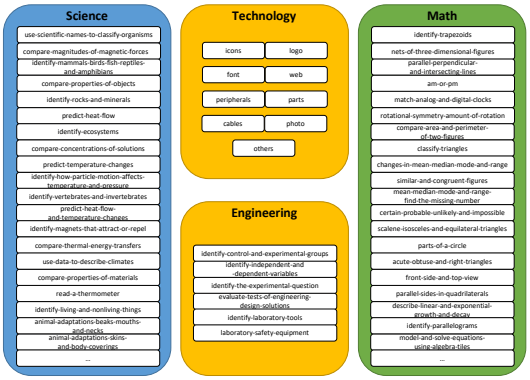


Figure 2: A summary of STEM skills.

Mathematics Mathematics requires procedural knowledge and reasoning abilities. Specifically, solving math also tests algebra generalization abilities. For example, the addition of numbers obeys the same rules everywhere. It includes fundamental math skills such as addition, algebra, comparing, counting, geometry and spatial reasoning (Appendix K). An example is shown in Figure 1(a)(iv). We automatically collect the questions from IXL Math. We conduct the same data processing. In addition, to encode mathematical expressions, we use LaTeX to avoid unusual symbols or cumbersome formal languages. After this, we obtain $858,859$ questions and 351 skills (Table 1) from pre-K to eighth grade. Compared to previous math datasets (Hendrycks et al., 2021b; Saxton et al., 2019; Zheng et al., 2022), our math subset focuses on large-scale fundamental multimodal skills and questions.

Comparison with Existing Datasets STEM is the first large-scale mulitmodal STEM dataset. As shown in Table 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. Compared to previous datasets, 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).

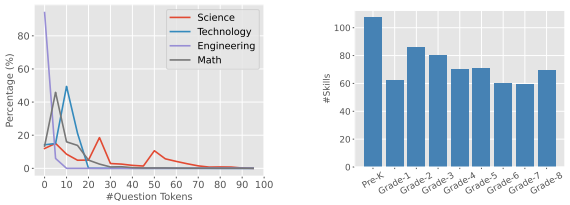
2.2 Analysis

To provide more insights into our dataset, we conduct the below analysis regarding the unique perspectives of STEM, namely skills, grades, and questions.

Skills The design of STEM emphasizes diverse skills spanning all STEM subjects. Figure 2 presents a brief summary of the skills (Appendix K provides a complete skill set). 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 (in particular vision and language) is crucial to master 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. The skills that are challenging for models differ from the ones that are challenging for humans. For example, it is easier for models to identify a logo of a computer related brand than for humans.

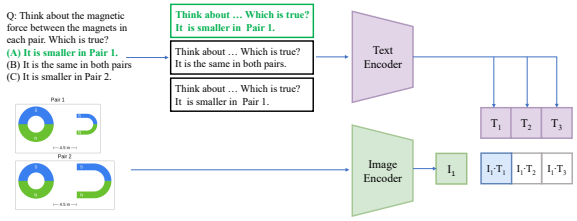
Grades STEM is designed with a comprehensive K-12 curriculum to examine fundamentals of STEM. This not only provides grade-level difficulties for the skills and questions, but also leads to another unique feature of testing on STEM: we are able to obtain the performance of models at each grade. The majority of existing datasets aim to compare models with human experts e.g., solving competition-level questions (Hendrycks et al., 2021b; Zheng et al., 2022). By contrast, STEM enables us to monitor the progress of modern neural models. For example, not mentioning human experts, models are just competitive with first grade students in understanding certain STEM skills, while they are also not as good as pre-K students. The grade statistics are shown in Figure 3(b).

Questions and Answers STEM provides the largest question set among existing datasets (Figure 1(a)). As shown in Figure 1(a), STEM contains multi-choice questions (Appendix K provides a question example for each skill). The question contains a textual question with an optional image context. Answers come in two formats. All answer options are in text (Figure 1(a)(i)) or in

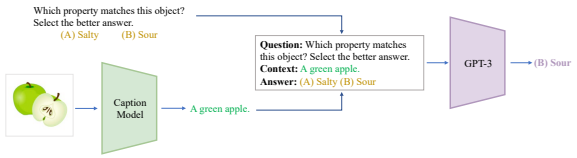


(a) Question length distribution. (b) #Skills per grade.

Figure 3: Data analysis of STEM.



(a) CLIP.



(b) GPT-3.

Figure 4: Zero-shot model setups.

image (Figure 1(a)(iv)). We further analyze the questions from the following aspects. (i) Question length. Figure 3(a) depicts the distribution of question lengths. We can see all subjects generally follow a long-tail distribution, while math distribution is most steep and science distribution is flatter. Heuristically, longer questions are more difficult to solve. (ii) Question type. We categorize the questions based on the first word of the question text as shown in the appendix. STEM mostly includes factoid questions that start with words such as "which" and "what". We show the word cloud of the question text in Appendix A. We can see the most common words like "block" and "numbers". This indicates the questions require joint reasoning of the text and images. (iii) The number of answers. STEM has averaging 2.8 answer options for each question. The distribution is presented in the appendix. In practice, the more answer options one question has, the more difficult the question is.

2.3 Models

We benchmark both state-of-the-art multimodal (vision-language) models (e.g., CLIP) and language models (e.g., GPT-3) on the STEM.

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 and Johnson, 2021) for the zero-shot evaluation of multimodal models. Multimodal models generally include two modules: image encoder and text encoder. CLIP is the state-of-the-art multimodal model. 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) **Few-Shot.** We use CLIP to benchmark the multimodal few-shot results as it is currently state-of-the-art. For this setting, we follow (Radford et al., 2021)'s few-shot linear probe setup. For the k-shot setup, we randomly select k questions for each skill from the training set (Table 1) as a meta training set. For each STEM subject, we train the model on the meta training set and select the best model on the validation set. At test time, the evaluation is the same with the zero-shot setup.

(iii) **Finetuning.** We also finetune CLIP. For each subject, we use the entire training set as shown in Table 1. The remaining setup is the same as the few-shot setting.

(iv) **Multi-Task.** Under this setting, we train CLIP on a mixture of training sets to produce a single model for all subjects.

Compared to previous work on multimodal benchmarks, our benchmark includes the state-of-the-art CLIP.

Language Models

(i) **Zero-Shot.** We use GloVe (Pennington et al., 2014), UnifiedQA (Khashabi et al., 2020) and GPT-3 (Chen et al., 2020a) zero-shot for the language model evaluation. We formalize the task as a question answering task. For GPT-3, we use the OpenAI API "text-davinci-002" corresponding to the best-performed GPT-3. We convert images to visual context text based on a captioning model following Lu et al. (2022). Figure 4(b) shows an example. Other models follow the same zero-shot setup.

Appendix B and C provide additional details of the models and evaluation setups.

2.4 Metrics and Human Performance

We report accuracy on the test set of each subject. We also average accuracy over all subject test

Model		Science	Technology	Engineering	Math	Average
Random Guesses		38.6	25.0	44.9	39.1	36.9
Language Models						
GloVe (2014)		38.0	25.2	48.1	39.0	37.6
UnifiedQA$_{\text{Small}}$ (2020)		39.6	27.2	58.0	39.6	41.1
UnifiedQA$_{\text{Base}}$ (2020)		42.6	28.8	55.4	40.0	41.7
GPT-3 (2020a)		47.1	22.1	73.5	44.0	46.7
Vision-Language Models						
Virtex (2021)		37.5	24.0	48.1	38.9	37.1
12-in-1 (2020)		39.4	27.5	44.2	41.9	38.3
ViLBERT (2019)		39.0	32.1	44.2	42.7	39.5
UNITER (2020b)		50.8	34.6	55.1	43.2	45.9
CLIP (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

Table 2: Zero-shot results on STEM.

Method		Science	Technology	Engineering	Math	Average
CLIP	Zero-Shot	50.3	68.7	55.1	43.6	54.4
	Few-Shot	75.2	70.9	61.9	63.2	67.8
	Finetuning	87.0	71.9	67.7	78.4	76.3
	Multi-Task	86.3	60.4	73.4	77.7	74.5

Table 3: Results of CLIP with different training schemes.

sets. In addition, we introduce two kinds of human scores. (i) Human (automated) 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 a human understands a STEM skill. SmartScore is computed online via IXL scoring system. We use the model outputs to mimic human behaviors to obtain the score. A SmartScore higher than 90.0 means a mastered skill. For technology, we use the average human accuracy available at Triviaplaza. The average accuracy is 68.6. (ii) Human (manual) score. We sampled 80 questions from our test sets (20 questions for each subject) and collected the responses from seven high-quality crowd workers in top universities. The average human accuracy is 83.0. All evaluation scores are higher the better.

3 Experiments

In this section, we illustrate the performance of a wide set of neural models and the human performance on STEM. The results show that state-of-the-art models like CLIP and GPT-3 significantly underperform the performance of general elementary students. The details of the experimental setup, more results and analysis are described in Appendix B.

3.1 Main Results

Zero-Shot The results are shown in Table 2. We first evaluate language models to test if models with text understanding are capable of skills in STEM. The performance of GloVe is close to random-chance accuracy, meaning that STEM cannot be solved by simply matching the semantic similarity between questions and answers. UnifiedQA is slightly better than GloVe with an improvement of 4.1% on average. GPT-3 performs the best among these language models, reaching 46.7% accuracy on average. GPT-3 achieves high accuracy in engineering, mainly because many engineering problems are about experimental settings (see Figure 1(a)(iii)) and GPT-3 is able to understand these textual descriptions. However, GPT-3 does not achieve competitive results in other subjects, implying that vision understanding is also important.

Next, we evaluate vision-language models. We observe that the performance of Virtex, 12-in-1, and ViLBERT is close to the performance of random guesses. They have a limited understanding of STEM subjects. UNITER and CLIP surpass random-chance accuracy by a large margin. Specifically, CLIP-RN50x64 achieves the highest average accuracy on STEM, with 18.0% improvements over random and 8.2% over GPT-3, showing that CLIP has a basic understanding of multimodal STEM skills and vision understanding ability is helpful. Among all subjects, the improvement in math is the smallest, with only 5.2% over random. This implies that math is the most challenging subject for current neural models.

Few-Shot We use the CLIP ViT-L/14@336px model in these settings by default. We use CLIP for this reference unless specified otherwise. The additional experimental details are in Appendix B. The 16-shot results are shown in Table 3. We observe that CLIP gains much improvement in all subjects after few-shot learning. This implies that CLIP has already stored STEM related knowledge and a few samples are able to trigger such knowledge. We also show performance varies when the number of samples of each skill changes (Figure 6). The overall performance improves with more samples, but 1-shot and 2-shot in technology are worse than zero-shot. Since there are only 9 skills in technology, 1-shot and 2-shot learning in technology might lead to overfitting.

Finetuning The results are shown in Table 3. It is encouraging as finetuning on science and math leads to great improvement larger than 30%. However, it only brings a 12.6% improvement in engineering and a 3.2% improvement in technology.

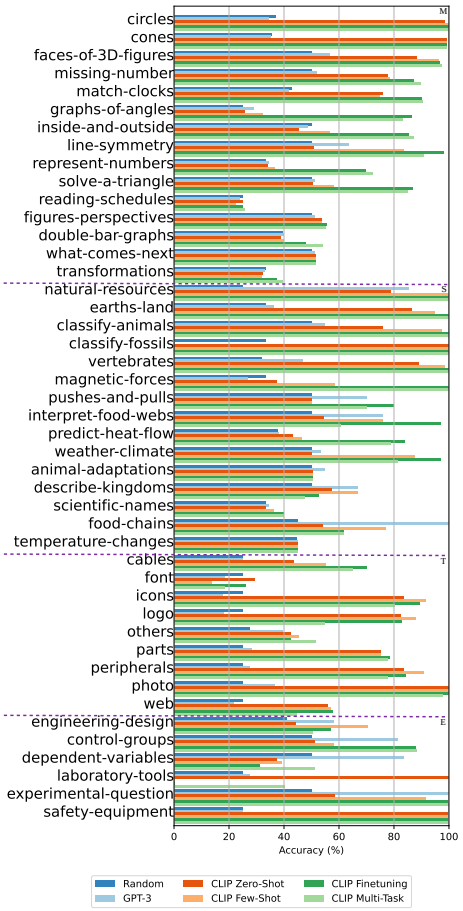


Figure 5: Results categorized by sampled skills of each subject. M: math. S: science. T: technology. E: engineering. Full results are in Appendix I.

One reason is that the amount of data for science and math is sufficiently larger than that for engineering and technology. We further discuss the neural model performance compared with humans in Sec. 3.3.

Multi-Task We show the results in Table 3. Multi-task learning improves in engineering but performs worse in other subjects compared with individual finetuned models. The reason for the great drop in technology is mainly because its data is much less than other subjects. Multi-task training actually improves performance in engineering. This implies that data from one subject may be beneficial for another when the knowledge is transferable. For example, science shares many common topics with engineering like chemical experiments.

3.2 Results Analysis

Skills As STEM provides detailed information about skills, analyzing models' performance on each skill helps us understand them better. We show the accuracy of GPT-3 and CLIP on part of

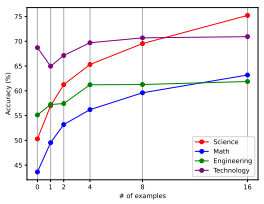


Figure 6: Result of few-shot CLIP.

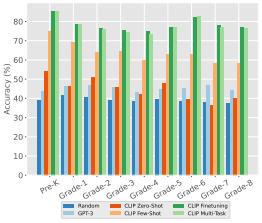


Figure 7: Average accuracy on each grade.

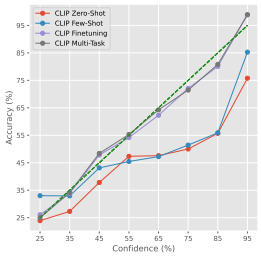


Figure 8: CLIP calibration results.

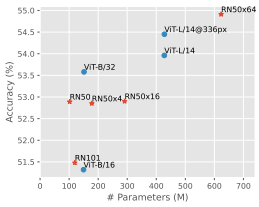


Figure 9: Zero-shot CLIP model scaling results.

skills in Figure 5. The complete results are in Appendix I. On most skills, the accuracy of CLIP zero-shot is higher than random-chance accuracy but lower than 80%. After finetuning, the accuracy is able to surpass 80%. This implies that CLIP zero-shot has a basic understanding of each skill but does not master them yet. As for GPT-3, it achieves high accuracy on skills like food chains and experimental questions, showing that GPT-3 has some specific knowledge about STEM skills.

Grades Skills in higher grades are more difficult for humans. We compare the model accuracy in each grade to see if the same trend exists for neural models. We show the accuracy of each grade in Figure 7. There is no obvious trend in performance drop as the increase in grade levels, implies the learning curve for neural models may be different from humans. This is because neural models are trained on all grade levels of data together while humans learn from low grades to high grades.

Questions and Answers We analyze how the model performance varies with the length of questions and the number of answers. Results are in Appendix D. Generally speaking, it is more difficult to answer questions with more tokens and more answers.

Calibration A trustworthy model should be calibrated, meaning that its confidence approximately matches the actual probability of the prediction

being correct (Guo et al., 2017). We show the relationship between the confidence of CLIP and the corresponding accuracy in Figure 8. We use the softmax probability as the confidence. We observe that CLIP zero-shot and few-shot are not well calibrated and they are overconfident about their predictions. After full training, CLIP is more calibrated, meaning that the output probability is a good estimation of the actual accuracy.

Scaling Laws Figure 9 shows the average zero-shot performance of CLIP with different 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 model performance on STEM.

3.3 Comparison with Human

We compare GPT-3 and CLIP with humans to better understand the performance of neural models.

Automated Evaluation We use human (automated) scores (Sec. 2.4) to evaluate neural models. Figure 1(b)(ii) shows the human (automated) scores of models and humans in each subject. Figure 1(b)(iii) shows the scores on each grade. All neural model performances are far behind that of elementary students (a score of 90). In technology, CLIP finetuning and multi-task achieve competitive performance. This is mainly because most technology skills in STEM are about specific empirical knowledge. After finetuning the models are fully adapted to the empirical knowledge. There is still a large performance gap between general neural models and average elementary students even in understanding the fundamental skills in STEM.

Manual Evaluation We also use the human (manual) scores (Sec. 2.4) to evaluate neural models. The overall result is similar to automated evaluation. The general models fall far behind humans, and finetuning is effective in technology. For manual evaluation, finetuned models also achieve competitive performance in science, implying that models also adapt to science empirical knowledge after finetuning. We list more details in Appendix J.2.

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 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). 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.

Pretraining techniques help achieve state-of-the-art performance in many NLP tasks, and have also been applied to vision-language models (Lu et al., 2019; Krishna et al., 2017; Chen et al., 2020b; Desai and Johnson, 2021; Lu et al., 2020), among which CLIP (Radford et al., 2021) is one of the state-of-the-arts. It leverages a large-scale of paired image-text data on the Internet. Other similar models include GLIP (Li et al., 2022) and GLIDE (Nichol et al., 2022). 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 test purpose. It consists of $1,073,146$ multi-choice questions and 448 skills spanning all STEM subjects. STEM focuses on fundamentals of STEM based on the K-12 curriculum. We also include state-of-the-art GPT-3 and CLIP in the test. The benchmark results suggest that current neural model performances are still far behind that of elementary students. STEM poses unique challenges for the broader community to develop fundamental algorithmic advancements. We hope our benchmark will foster future research in multimodal understanding.

6 Limitations

For the limitations of our benchmark, even though the dataset is overall large-scale, the size of the engineering and technology subset is relatively small compared to the other two subjects. Another limitation is that the benchmark only includes multi-choice questions. Other question types, e.g., open questions or fill-in-the-blank are not part of the benchmark yet. Based on the error analysis, algorithmic innovations are required to advance the performance of modern neural networks on our benchmark. The dataset also does not provide step-by-step solutions or explanations yet, which can potentially help improve the performance. Therefore, future work could include dataset expansion and new model developments.

References

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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. 2022. Grounded language-image pre-training. In *CVPR*, pages 10955–10965.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

A More Statistics on STEM

We show the number of answers, number of skills and questions distribution in each grade in Figure 10. We also show the word cloud of our STEM in Figure 11.

Skill Comparison We compare the skills of STEM with other related datasets in Figure 12.

B Experimental Details

For the zero-shot setting, we evaluate all models on the test set. For the few-shot, finetuning, and multi-task setting, we train CLIP-ViT-L/14@336px on the corresponding train set, tune hyperparameters on the valid set, and finally evaluate on the test set. We use AdamW for optimization and tune hyperparameters as follows: batch size is chosen from {16, 32, 64, 128}, and set to 16 for few-shot learning, 128 for finetuning and multi-task learning after hyperparameter tuning. The learning rate is chosen between [5e-6, 5e-5] and set to 1e-5 for all training. We set the warm-up ratio to 0.1 and set weight decay as 0.2. We set the maximum of training samples to 100k for finetuning, 200k for multitask training, and 10 epochs for few-shot training, all with early stop on the valid set.

C Benchmarking Models

In this section, we introduce our benchmarking models in details.

C.1 Models

CLIP (Radford et al., 2021) CLIP is pretrained on a sufficiently large dataset of 400 million text-image pairs across the Internet. It uses a Transformer as the text encoder, and has several variants of image encoder, including ResNet (RN) backbones and Vision Transformers (ViT) (Dosovitskiy et al., 2020). CLIP aligns the text and image representation by training on in-batch contrastive loss, and is able to zero-shot transfer to downstream vision language tasks. To align with CLIP pretraining, we formulate question answering as matching text and images. We use the cosine similarity between the text and image embeddings as the matching function, the same as the original zero-shot image-text retrieval settings in CLIP (Radford et al., 2021).

ViLBERT and 12-in-1 (Lu et al., 2019, 2020) ViLBERT adopts two parallel streams to process image regions and text segments separately, with co-attentional transformer layers connecting them. There is also a multi-task version called 12-in-1 (Lu et al., 2020) that trains 12 different tasks with individual task-specific heads sharing 1 "trunk" ViLBERT model. Its multi-modal alignment prediction serves as the matching score.

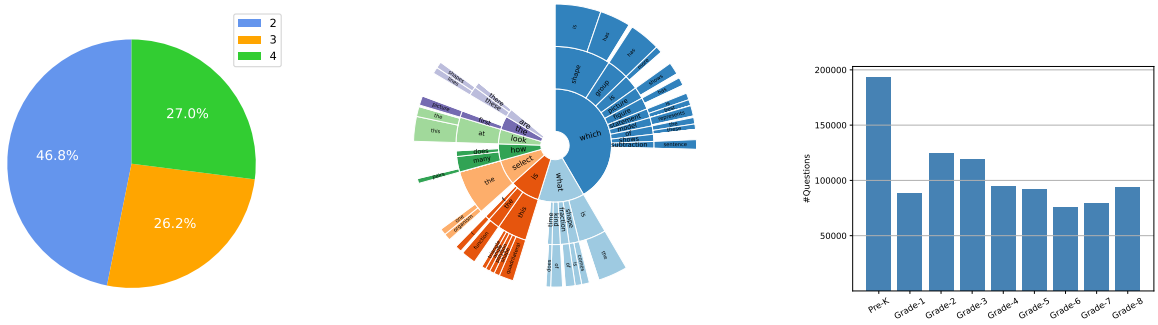
UNITER (Chen et al., 2020b) UNITER consists of an Image Embedder with Faster R-CNN (Anderson et al., 2018), a Text Embedder with Transformer (Vaswani et al., 2017), as well as a multi-layer Transformer to get cross-modality representation. During inference on STEM, the matching score function is the same as CLIP, i.e., the cosine similarity between the text and image embeddings (Chen et al., 2020b).

Virtex (Desai and Johnson, 2021) Virtex first extracts visual features with ResNet-50 (He et al., 2016) backbone. The visual features are then fed into a text head, which consists of two unidirectional Transformers, to predict captions. We extract the image feature with the image encoder, then feed text into the textual head and use the sum of bidirectional generation logits as the matching score.

C.2 Evaluation Details

We benchmark both state-of-the-art multimodal (vision-language) models (e.g., CLIP) and language models (e.g., GPT-3) on STEM.

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 and Johnson, 2021) for the zero-shot evaluation of multimodal models. Multimodal models generally include two modules: image encoder and text encoder. CLIP is the state-of-the-art multimodal model. For zero-shot CLIP, we follow its original setup in Radford et al. (2021). Figure 4(a) illustrates an example. The input to the text encoder is the concatenation of the question text and an answer option. The input to the image encoder is the image context. The output is the cosine similarity scores between the text embeddings and image embedding. Then the answer option with the largest similarity score serves as an answer. For questions with image answer options, the input to the image encoder will also add the image answer options.



(a) #Answers distribution. (b) #Question distribution in STEM. (c) #Questions per grade.

Figure 10: Data analysis of STEM.

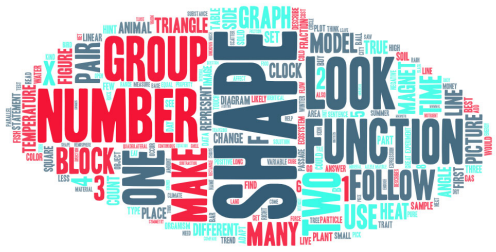


Figure 11: Word cloud of question texts in STEM.

(ii) **Few-Shot.** We use CLIP to benchmark the multimodal few-shot results as it is currently the state-of-the-art. For this setting, we follow (Radford et al., 2021)'s few-shot linear probe setup, where a softmax output layer is added to CLIP architecture. The logits are the similarity scores based on the text and image embeddings. For k-shot setup, we randomly select k questions for each skill from the training set (Table 1) as a meta training set. For each STEM subject, we train the model on the meta training set and select the best model on the validation set. At test time, the evaluation is the same as the zero-shot setup.

(iii) **Finetuning.** We also finetune CLIP. For each subject, we use the entire training set as shown in Table 1. The remaining setup is the same as the few-shot setting.

(iv) **Multi-Task.** Under this setting, we train CLIP on the mixture of training sets to produce a single model for all subjects.

Compared to previous work on multimodal benchmarks, our benchmark includes the state-of-the-art CLIP.

Language Models

(i) **Zero-Shot.** We use GloVe (Pennington et al., 2014), UnifiedQA (Khashabi et al., 2020) and GPT-3 (Chen et al., 2020a) zero-shot for the language model evaluation. We formalize the task as a question answering task. For GPT-3, we use the OpenAI API "text-davinci-002" corresponding to the best-performed GPT-3. The input to GPT-3 is the concatenation of the question text, the context text, and multiple answer options. The output is to predict a final answer from answer options. For images, we follow Lu et al. (2022) to convert them to visual context text based on a captioning model consisting of ViT (Dosovitskiy et al., 2020) and GPT-2 (Radford et al., 2019). Figure 4 (b) shows an example. For UnifiedQA, we use both its base and small versions. Its zero-shot setup is the same as that of GPT-3. For GloVe, we use the similarity between the embedding of the concatenation of the question and context text, and the embedding of each answer option to decide the final answer. The answer option with the largest similarity score is the answer output. We use average pooling based on the 300-dimensional word vectors to obtain embeddings. The images are also converted to text using the same method as that of GPT-3.

D Detailed Experimental Analysis

Few-Shot In the few-shot setting, we sample different number of samples in each grade to see how the learning performance varies. Specifically, we sample 1, 2, 4, 8 and 16 samples per skill and train CLIP on the sampled data. The results are shown in Figure 6.

Question Lengths Figure 15 shows how the question length affects model accuracy. For GPT-3 and Clip zero-shot, the accuracy decreases slightly as the question becomes longer. For tuned models, the same trend holds for questions less than 70 to-

Subject	IconQA	ScienceQA	STEM
Science	0	167	82
Technology	0	0	9
Engineering	0	0	6
Math	13	0	351
Total	13	167	448

(a) Statistics comparison.

IconQA	STEM	
Counting	Count to 10, Count shapes in rows, Count sides and corners ...	
Geometry	Classify triangles, Identify symmetry, Identify shapes ...	
Time	Match times, Identify A.M./P.M., Read a calendar ...	
...	...	
Not cover	*Science*	Compare concentrations of solutions ...
	Technology	Identify peripherals ...
	Engineering	Identify laboratory tools ...
	Math	Linear and exponential functions ...

(b) Skill comparison between STEM and IconQA.

Figure 12: Comparison between our STEM dataset and existing datasets (IconQA and ScienceQA).

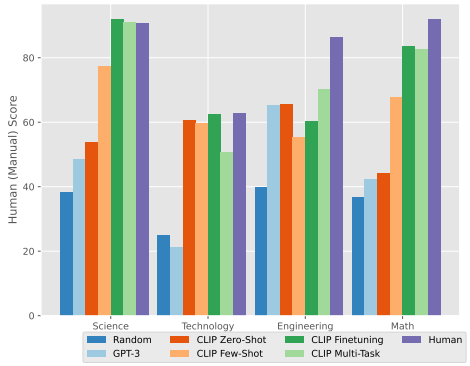


Figure 13: Human (manual) scores on sampled STEM.

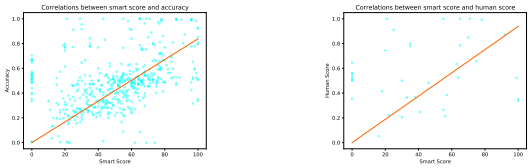


Figure 14: The correlation graphs of human (automated) scores with model accuracy (left) and human (manual) scores (right).

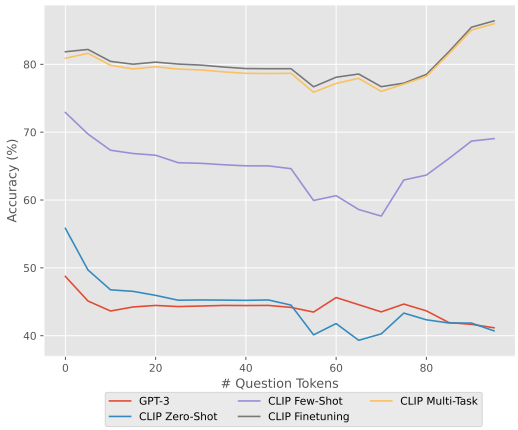


Figure 15: Results on questions with different lengths.

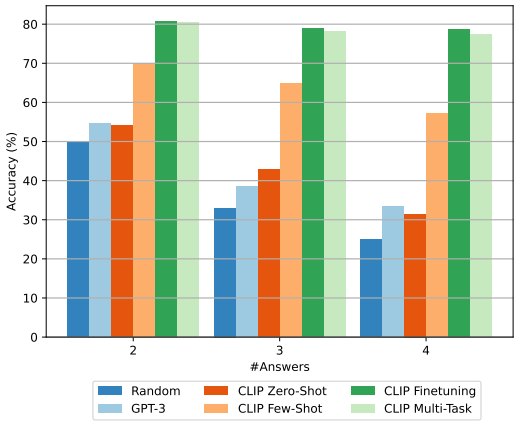


Figure 16: Results on questions with different number of answers.

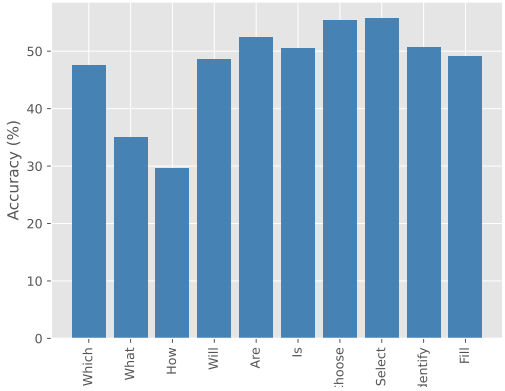


Figure 17: Zero-shot CLIP performance on different question types.

kens, but the accuracy starts to increase for longer questions. We think this may be caused by some bias in longer questions and the tuned models learn such bias and achieve higher accuracy. Since there are only a small proportion of questions that are longer than 70 tokens, such bias will not affect the whole dataset much.

Number of Answers We also analyze how model performance changes with the number of answers. The results are shown in Figure 16. We find that for GPT-3, CLIP zero-shot, and few-shot, the accuracy drops as the number of answers increases, but the accuracy of CLIP finetuning and multi-task does not drop. This implies that models after full training are actually solving the problem rather than guessing, so the number of choices does not affect the performance much.

Question Type We mark the types of problems as the first word in the question or request of each problem. In Figure 17 we show the accuracy of the top 10 frequent types. Questions starting with "What" and "How" have relatively low accuracy, as these questions are more difficult to answer.

Comparison with Humans In table 4 we show the numerical results of models compared with humans measured by two human scores, and Figure 13 shows the comparison on the human (manual) scores.

Correlation Between Human Scores and Accuracy We evaluate human (automated) scores' correlation with model accuracy and human (manual) scores (Figure 14). They in general positively correlated to each other. Even though human (automated) score is different from accuracy, it overall captures accuracy as an important factor.

E Case Study

We show some cases in Figure 18 to demonstrate what kinds of skills are easy or difficult for CLIP. We categorize skills into three groups: 1) **Easy**: CLIP zero-shot has high accuracy on these skills; 2) **Medium**: CLIP zero-shot has low accuracy on these skills but CLIP finetuning has high accuracy; 3) **Hard**: both CLIP zero-shot and finetuning have low accuracy on these skills. The easy skills mainly ask about the names of objects like shapes or animals. The medium skills involve some abstract concepts like symmetry and the direction of force. The hard skills require complex logical reasoning, such as finding patterns or inferring the function of animal adaption.

F Error Analysis

To better understand the errors made by CLIP zero-shot, we sample 25 error cases of CLIP zero-shot on math and science. We manually check the reasons for these errors. For math, 36% errors are caused by a lack of mathematical commonsense, such as area formulas and symmetry. Other errors include failure of calculation (24%), counting objects (16%), reading tables or graphs (12%, e.g., graphs of functions), and transformation (12%, e.g., rotation of a 3D object). For science, comparison causes the most errors with a ratio of 40%. Most of these questions only require a straightforward comparison like the distance between two pairs of magnets. However, CLIP fails on such basic problems. This indicates that it is not good at comparing objects and properties yet. Lacking science commonsense also leads to a good number of errors (32%), followed by identifying directions (20%, e.g., the directions of push and pull, towards and away) and reading tables or graphs (8%). More information of the errors is included in Appendix F.

G Zero-Shot Prompt Sensitivity

We study the effect of prompts on CLIP zero-shot. We design 5 types of prompts and demonstrate them with an example problem. The example question is "Which property matches this object?" and the answer is "Rough". Examples of different prompt types and the corresponding accuracies are shown in Table 6. We observe that "Q+A results in the best performance on average but the difference is only marginal, meaning that CLIP zero-shot is not very sensitive to the format of prompts.

H Dataset Collection Details

We design the STEM to comprehensively evaluate vision language understanding abilities on STEM subjects. This requires to cover a broad and diverse range of topics with text and images appearing at the same time. We search for online open sources and choose the following three platforms: *IXL*[2], *ProProfs Quizzes*[3], and *Triviaplaza*[4]. *IXL* is an online learning website designed for K-12 education. The learning skills are fully aligned with U.S. state educational standards[5]. *ProProfs Quizzes* and *Triviaplaza* are online quiz websites with more than 100k quizzes and 50 million quiz takers. They provide a variety of topics and have been widely used for employee skill assessment.

We collect science and math problems from the *IXL* website. As for engineering, we use the skills

[2] https://www.ixl.com/
[3] https://www.proprofs.com/quiz-school
[4] https://www.triviaplaza.com/
[5] https://www.ixl.com/standards

Method		Human (Automated) Score				Human (Manual) Score			
		Science	Engineering	Math	Technology	Science	Technology	Engineering	Math
Human		90.0	90.0	90.0	68.6	90.7	62.9	86.4	92.1
Random		26.7	16.1	51.1	25.0	38.3	25.0	40.0	36.8
GPT-3		45.7	50.2	51.4	22.1	48.4	21.3	65.2	42.4
CLIP	Zero-Shot	33.9	19.0	52.9	68.7	53.8	60.7	65.5	44.3
	Few-Shot	39.1	43.9	67.6	70.9	77.3	59.7	55.5	67.8
	Finetuning	57.8	37.4	75.7	71.9	91.9	62.6	60.3	83.5
	Multi-Task	61.9	50.3	72.0	60.4	90.9	50.6	70.2	82.5

Table 4: Comparison between models and humans.

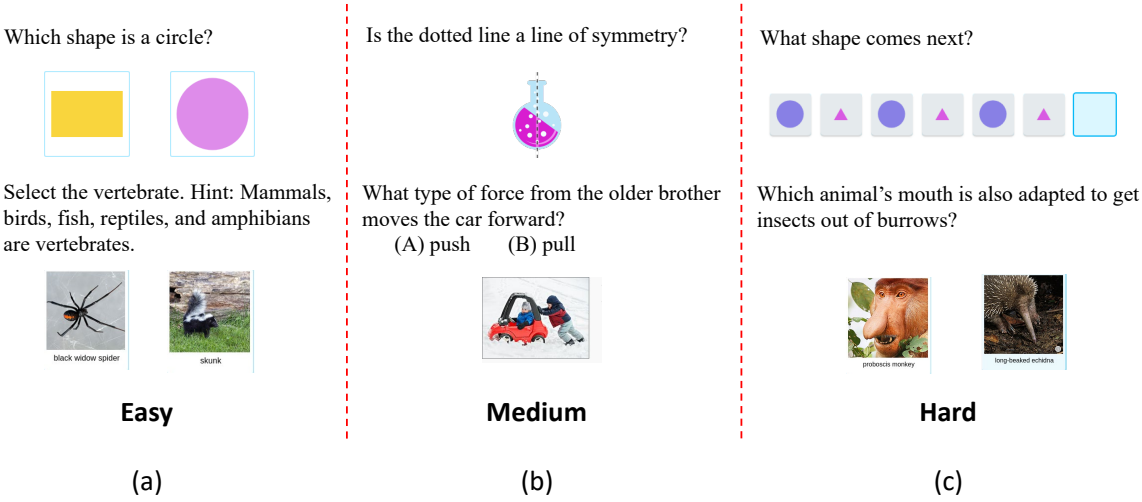


Figure 18: Examples of different difficulty levels for CLIP.

Subject	Reason	Ratio (%)
Math	Commonsense	36
	Numerical calculation	24
	Counting	16
	Read table/graph	12
	Transformation	12
Science	Comparison	40
	Commonsense	32
	Direction	20
	Read table/graph	8

Table 5: Ratio of different errors.

from the "Science and engineering practices" topic in science. Figure 19 shows the collecting procedure of our science, math, and engineering dataset, with the following steps:

- **Select target skills.** As the target of STEM is to collect a multimodal STEM dataset, we go through all science, engineering, and math skills in the *IXL* website and select those with at least one image in either question context or answers. We only keep multi-choice problems and in the future, we may consider other types of problems like fill-in-the-blank.
- **Collect problems.** We build automatic scripts to mimic a virtual user practicing on *IXL*, and crawl problems it encountered, including the natural language description, images, choices, and the correct answer. For each skill, we collect at most 2,000 problems.
- **Data cleaning and deduplication.** Due to network latency and fluctuation, some collected images are corrupted. We remove these samples by comparing the image sizes. Also, there are some duplicated problems in each skill. We compare each sample by hashing and remove all duplication.
- **(Math dataset only) Transform the formulas.** There are many formulas embedded in math problems that are not represented in text. We use the Mathpix[6] OCR API to convert these math formulas into the latex format.

Finally, we manually go through *ProProfs Quizzes* and *Triviaplaza* and select the categories

[6] https://mathpix.com/

Prompt Format	Example	Science	Technology	Engineering	Math	Average
Q+A	Which property matches this object? Rough.	50.3	68.7	55.1	43.6	54.4
A+Q	Rough. Which property matches this object?	50.0	66.0	49.6	43.2	52.2
Q "Choose the best answer:" A	Which property matches this object? **Choose the best answer:** Rough.	50.1	70.7	49.7	44.2	53.7
"Answer the question:" Q + A	**Answer the question:** Which property matches this object? Rough.	49.4	67.6	51.0	43.6	52.9
A "best answers the question" Q	Rough **best answers the question:** Which property matches this object?	49.7	69.5	50.8	43.8	53.4

Table 6: Examples for different prompts and their zero-shot accuracy.

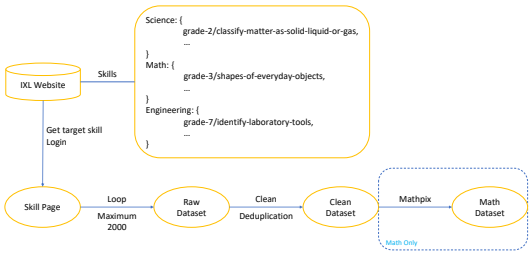


Figure 19: Procedure for collecting data from the *IXL* website.

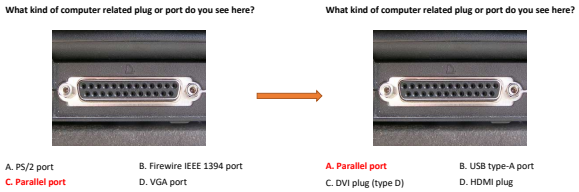


Figure 20: Technology dataset augmentation.

with pictures. Moreover, we augment the dataset by sampling from all choices in a category. For example, we generate more problems by replacing the choices, as shown in Figure 20.

I Performance on Skills

We show the accuracy of random guesses, CLIP zero-shot and finetuning on all 448 skills in Figure 21 to 26. We can see that the zero-shot performance is generally better than random guesses on most skills and achieves near 100% on some skills (e.g., "circles" and "cones"). After finetuning, accuracy improves on most skills and becomes near 100% on many skills.

J Human Score Details

J.1 Human (Automated) Score

We test human (automated) scores on all skills in engineering and technology, and randomly choose 40 skills from math, and 30 skills from science due to technical and time constrains. We compare neural models with humans using the human (automated) score, and the resutls are shonw in Table 4. The detailed scores and skills are listed in Table 7.

J.2 Human (Manual) Score

We randomly sample 20 problems for each subject and ask 7 Ph.D. students to answer these questions, and calculate the average accuracy for each subject as the human (manual) score. To evaluate neural models according to the human (manual) score, we use the corresponding skill accuracy for each sampled problem as the models' score on this problem and average all accuracy together as the final score. We do not evaluate models on these sampled data directly since the small number of samples will lead to a large variance, and skill accuracy can avoid such variance. The comparison results are shown in Table 4. All sampled problems for human (manual) score are listed in Table 8 to 13.

K Summary of Skills

We list all skills in STEM in Table 14 to 16 and show some examples in Table 17 to 23.

Subject	Grade/Skill	Random	Zero-shot	Finetune
Science	grade-2/classify-matter-as-solid-liquid-or-gas	28	40	100
	grade-2/identify-animals-with-and-without-backbones	0	70	70
	grade-2/identify-mammals-birds-fish-reptiles-and-amphibians	0	0	18
	grade-2/identify-materials-in-objects	21	40	100
	grade-2/identify-properties-of-an-object	35	65	65
	grade-3/compare-strengths-of-magnetic-forces	0	18	63
	grade-3/describe-ecosystems	65	50	100
	grade-3/find-evidence-of-changes-to-earths-surface	17	38	100
	grade-3/identify-ecosystems	35	100	100
	grade-3/identify-minerals-using-properties	35	11	35
	grade-4/compare-properties-of-objects	10	17	20
	grade-4/describe-ecosystems	74	100	100
	grade-4/identify-minerals-using-properties	35	16	35
	grade-4/use-evidence-to-classify-mammals-birds-fish-reptiles-and-amphibians	26	35	35
	grade-5/animal-adaptations-beaks-mouths-and-necks	17	27	35
	grade-5/classify-elementary-substances-and-compounds-using-models	75	75	75
	grade-5/compare-ancient-and-modern-organisms-use-observations-to-support-a-hypothesis	32	32	50
	grade-5/identify-directions-of-forces	0	26	35
	grade-5/identify-the-photosynthetic-organism	0	0	100
	grade-5/predict-temperature-changes	0	22	0
	grade-5/use-evidence-to-classify-animals	35	35	35
	grade-5/use-evidence-to-classify-mammals-birds-fish-reptiles-and-amphibians	18	35	35
	grade-5/weather-and-climate-around-the-world	60	36	60
	grade-6/compare-concentrations-of-solutions	15	11	100
	grade-6/describe-the-effects-of-gene-mutations-on-organisms	52	13	69
	grade-6/diffusion-across-membranes	50	25	50
	grade-7/describe-the-effects-of-gene-mutations-on-organisms	42	13	69
	grade-8/classify-symbiotic-relationships	25	36	45
	grade-8/diffusion-across-membranes	0	18	35
	grade-8/moss-and-fern-life-cycles	0	12	0
Engineer	grade-6/evaluate-tests-of-engineering-design-solutions	0	0	100
	grade-6/identify-control-and-experimental-groups	0	0	0
	grade-6/identify-independent-and-dependent-variables	0	0	100
	grade-6/identify-the-experimental-question	30	30	30
	grade-7/evaluate-tests-of-engineering-design-solutions	0	0	0
	grade-7/identify-control-and-experimental-groups	0	0	40
	grade-7/identify-independent-and-dependent-variables	0	0	30
	grade-7/identify-the-experimental-question	40	0	40
	grade-8/identify-control-and-experimental-groups	0	0	0
	grade-8/identify-the-experimental-question	60	0	40
	grade-5/identify-laboratory-tools	21	42	31
	grade-6/identify-laboratory-tools	21	21	21
	grade-6/laboratory-safety-equipment	24	65	52
	grade-7/identify-laboratory-tools	10	28	21
	grade-7/laboratory-safety-equipment	9	58	52
	grade-8/identify-laboratory-tools	49	21	21
	grade-8/laboratory-safety-equipment	9	58	58
Math	algebra-2/factor-quadratics-using-algebra-tiles	40	51	55
	algebra-2/outliers-in-scatter-plots	55	47	97
	calculus/determine-continuity-using-graphs	36	63	80
	calculus/find-limits-at-vertical-asymptotes-using-graphs	60	65	85
	grade-1/subtraction-sentences-up-to-10-which-model-matches	50	30	99
	grade-2/identify-halves-thirds-and-fourths	65	75	97
	grade-2/identify-lines-of-symmetry	70	64	99
	grade-2/interpret-bar-graphs-ii	14	23	12
	grade-2/ordinal-numbers-up-to-10th	32	61	28
	grade-3/compare-fractions-in-recipes	55	50	68
	grade-3/identify-parallelograms	51	64	70
	grade-3/is-it-a-polygon	71	60	98
	grade-3/parallel-sides-in-quadrilaterals	29	66	45
	grade-4/nets-of-three-dimensional-figures	68	40	99
	grade-5/nets-of-three-dimensional-figures	53	40	99
	grade-6/changes-in-mean-median-mode-and-range	38	14	15
	grade-6/classify-triangles	47	38	45
	grade-6/identify-polyhedra	75	75	75
	grade-6/mean-median-mode-and-range-find-the-missing-number	55	41	99
	grade-6/model-and-solve-equations-using-algebra-tiles	36	36	57
	grade-6/rational-numbers-find-the-sign	31	78	99
	grade-6/rotational-symmetry	62	56	78
	grade-6/similar-and-congruent-figures	34	33	46
	grade-6/which-figure-is-being-described	36	27	86
	grade-7/rational-numbers-find-the-sign	47	58	99
	grade-8/rotational-symmetry-amount-of-rotation	47	32	63
	kindergarten/count-on-ten-frames-up-to-10	15	2	49
	kindergarten/fewer-and-more-up-to-20	80	62	97
	kindergarten/subtraction-sentences-up-to-5-which-model-matches	41	30	96
	pre-k/addition-sentences-up-to-10-which-model-matches	60	55	96
	pre-k/count-on-ten-frames-up-to-3	84	50	51
	pre-k/fewer-and-more-compare-by-matching	63	52	90
	pre-k/one-less-with-pictures-up-to-10	61	37	66
	pre-k/one-more-with-pictures-up-to-5	48	36	75
	pre-k/shapes-of-everyday-objects	67	96	96
	pre-k/spheres	67	96	96
	pre-k/triangles	57	75	75
	pre-k/what-comes-next	75	56	70
	pre-k/ordinal-numbers-up-to-tenth	27	84	82
	kindergarten/are-there-enough	40	99	96

Table 7: Human (automated) scores for each skill.

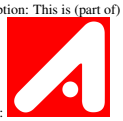
Subject: Technology Description: This is a(n old) logo of which famous app or program? Picture:  Choices: [Microsoft Office Outlook, Microsoft Office OneDrive, OfficeSuite Pro, Opera,] Answer index: 3	Subject: Technology Description: What kind of computer component do you see here? Picture:  Choices: [TV Tuner Card, PC Card, Motherboard, Modem Card,] Answer index: 2
Subject: Technology Description: This is (part of) a (former) logo of which computer related brand? Picture:  Choices: [ASRock, Amiga Inc., Arctic, ATI Technologies,] Answer index: 3	Subject: Technology Description: This is (part of) a (former) logo of which computer related brand? Picture:  Choices: [Fujitsu, Samsung, Iiyama, Brother,] Answer index: 2
Subject: Technology Description: This is (part of) a (former) logo of which computer related brand? Picture:  Choices: [Xiaomi, Cisco, Intel, Wii,] Answer index: 3	Subject: Technology Description: What kind of computer component do you see here? Picture:  Choices: [Display Adapter/Video Card, PC Card, Power Supply Unit, Hard Disk Drive,] Answer index: 3
Subject: Technology Description: What meaning or function is usually associated with this web interface symbol? Picture:  Choices: [Paste, Search, Tip/Idea, Calendar/Event,] Answer index: 2	Subject: Technology Description: This is a(n old) logo of which famous app or program? Picture:  Choices: [YouTube Music, Beats Music, MX Player, YouTube,] Answer index: 2
Subject: Technology Description: What kind of computer related plug or port do you see here? Picture:  Choices: [USB type-C plug, DVI plug (type D), HDMI plug, 3.5mm Audio Cable plug,] Answer index: 0	Subject: Technology Description: Identify this font type Picture: ActionQuiz Choices: [Lucida MT, News Gothic MT, Fixedsys, Courier New,] Answer index: 2
Subject: Technology Description: Identify this font type Picture: ActionQuiz Choices: [Commercial Script BT, Brush Script MT, Vivaldi D, ShelleyVolante BT,] Answer index: 3	Subject: Technology Description: Identify this font type Picture: ActionQuiz Choices: [Garamond, Times New Roman, Courier New, Georgia,] Answer index: 3
Subject: Technology Description: This is (part of) a (former) logo of which computer related brand? Picture:  Choices: [BenQ, Lexmark, Creative Technology, Lenovo,] Answer index: 2	Subject: Technology Description: Identify this font type Picture: ActionQuiz Choices: [Webdings, Courier, Impact, System,] Answer index: 3
Subject: Technology Description: Identify this font type Picture: ActionQuiz Choices: [Serifa BT, Stylus ITC, Calisto MT, Tempus Sans ITC,] Answer index: 0	Subject: Technology Description: What meaning or function is usually associated with this web interface symbol? Picture:  Choices: [Pin/Make something sticky, Storage for deleted files, Options/Settings, Print (preview),] Answer index: 1
Subject: Technology Description: What meaning or function is usually associated with this web interface symbol? Picture:  Choices: [Zoom in, Help, Like something, Link select,] Answer index: 3	Subject: Technology Description: What meaning or function is usually associated with this web interface symbol? Picture:  Choices: [Apply, Options/Settings, Reload/Refresh, Download,] Answer index: 2
Subject: Technology Description: What type of video game console do you see here? Picture:  Choices: [Mattel Intellivision, Sega Master System, Magnavox Odyssey 2, Atari 5200,] Answer index: 3	Subject: Technology Description: What meaning or function is usually associated with this web interface symbol? Picture:  Choices: [Find, Delete, Attachment, Calendar/Event,] Answer index: 0

Table 8: Human (manual) evaluation problem set (part 1).

Subject: Engineer Description: Select the gloves. Picture: None  Choices: Answer index: 0	Subject: Engineer Description: iption: In this experiment, which were part of an experimental group? The passage below describes an experiment. Lucy and Erik were taking a snowboarding class. During the class, their instructor said they would go faster if they applied wax to the undersides of their snowboards. After the class, Lucy applied a thin layer of wax to the underside of a snowboard and rode the board straight down a hill. Then, she removed the wax and rode the snowboard straight down the hill again. Erik timed how long each ride took. Lucy repeated these rides on four other snowboards, alternating whether she first rode with or without wax.  Picture: Choices: [the snowboards with wax removed, the snowboards with wax added,] Answer index: 1
Subject: Engineer Description: Select the test tube. Picture: None  Choices: Answer index: 2	Subject: Engineer Description: Select the funnel. Picture: None  Choices: Answer index: 2
Subject: Engineer Description: Select the round-bottom flask. Picture: None  Choices: Answer index: 1	Subject: Engineer Description: iption: In this experiment, which were part of an experimental group? The passage below describes an experiment. Kimberly grew roses for a flower shop. One day, she noticed tumor-like growths on her rose stems. She could tell that the plants had crown gall disease, which is caused by a type of bacteria. She knew that allicin, a chemical in garlic, can kill bacteria. Kimberly wondered if spraying her plants with garlic juice would prevent more tumors from forming on her plants. Once a day, Kimberly sprayed garlic juice on ten infected plants and left another 10 infected plants unsprayed. After one month, she compared the number of new tumors on plants in the two groups.  Picture: Choices: [the roses sprayed with garlic juice, the roses that were not sprayed,] Answer index: 0
Subject: Engineer Description: iption: Which of the following could Kendra's test show? Wind turbines use wind power to produce electricity. Kendra was a materials engineer who designed wind turbines. She wanted to design a new turbine that would produce 10The passage below describes how the engineering-design process was used to test a solution to a problem. Read the passage. Then answer the question below.  Picture: Choices: [how much the new turbine would weigh, whether the new turbine could produce 10% more electricity, if the new turbine could turn easily,] Answer index: 1	Subject: Engineer Description: iption: In this experiment, which were part of an experimental group? The passage below describes an experiment. Isaac and his friend Belle flew nylon kites on the beach. They wondered if putting a tail on a kite would affect how well the kite flew. Isaac flew a kite that did not have a tail for five minutes. Then, he attached a four-foot-long tail and flew the kite for five more minutes. Isaac repeated this with three similar kites, alternating whether he started the kite with or without a tail. During each flight, Belle counted the number of times the kite crashed to the ground.  Picture: Choices: [the kites without tails, the kites with tails,] Answer index: 1
Subject: Engineer Description: iption: Identify the question that Bryant and Lamar's experiment can best answer. The passage below describes an experiment. Read the passage and then follow the instructions below. Bryant placed a ping pong ball in a catapult, pulled the catapult's arm back to a 45° angle, and launched the ball. Then, Bryant launched another ping pong ball, this time pulling the catapult's arm back to a 30° angle. With each launch, his friend Lamar measured the distance between the catapult and the place where the ball hit the ground. Bryant and Lamar repeated the launches with ping pong balls in four more identical catapults. They compared the distances the balls traveled when launched from a 45° angle to the distances the balls traveled when launched from a 30° angle.  Picture: Choices: [Do ping pong balls stop rolling along the ground sooner after being launched from a 30° angle or a 45° angle?, Do ping pong balls travel farther when launched from a 30° angle compared to a 45° angle?,] Answer index: 1	Subject: Engineer Description: Select the Erlenmeyer flask. Picture: None  Choices: Answer index: 3

Table 9: Human (manual) evaluation problem set (part 2).

Subject: Engineer Description: iption: Which of the following could Ivan's test show? Ivan was a landscape architect who was hired to design a new city park. The city council wanted the park to have space for outdoor concerts and to have at least 20The passage below describes how the engineering-design process was used to test a solution to a problem. Read the passage. Then answer the question below.  Picture: Choices: [if at least 20Answer index: 1	Subject: Engineer Description: Select the beaker. Picture: None  Choices: Answer index: 3
Subject: Engineer Description: iption: Identify the question that Zeke's experiment can best answer. The passage below describes an experiment. Read the passage and then follow the instructions below. Zeke divided 40 unripe bananas evenly among eight paper bags and sealed the bags. He poked 20 small holes in four of the bags and left the other four without holes. He kept the bags at room temperature for three days. Then, Zeke opened the bags and counted the number of brown spots on each banana. He compared the average number of brown spots on bananas from bags with holes to the average number of brown spots on bananas from bags without holes.  Picture: Choices: [Do bananas develop more brown spots if they are kept in bags with holes compared to bags without holes?, Do bananas develop more brown spots when they are kept at room temperature compared to in a cold refrigerator?,] Answer index: 0	Subject: Engineer Description: iption: Hint: An independent variable is a variable whose effect you are investigating. A dependent variable is a variable that you measure. Which of the following was an independent variable in this experiment? The passage below describes an experiment. Read the passage and think about the variables that are described. Tyler designed an electric circuit to test how well different types of metal conduct electricity. The circuit included a battery, a light bulb, wires, and clips that could be attached to a sheet of metal. If the metal conducted electricity poorly, the light bulb would appear dim. If the metal conducted electricity well, the light bulb would appear bright. Tyler collected nine equally sized sheets of metal: three sheets of copper, three sheets of iron, and three sheets of aluminum. He used the clips to attach each metal sheet, one sheet at a time, to the circuit. For each sheet, Tyler used a light meter to measure how much light the bulb produced.  Picture: Choices: [the amount of light produced by the light bulb, the type of metal sheet used in the circuit,] Answer index: 1
Subject: Engineer Description: iption: Identify the question that Devon's experiment can best answer. The passage below describes an experiment. Read the passage and then follow the instructions below. Devon poured four ounces of water into each of six glasses. Devon dissolved one tablespoon of salt in each of three glasses, and did not add salt to the other three. Then, Devon placed an egg in one glass and observed if the egg floated. She removed the egg and dried it. She repeated the process with the other five glasses, recording each time if the egg floated. Devon repeated this test with two more eggs and counted the number of times the eggs floated in fresh water compared to salty water.  Picture: Choices: [Does the amount of water in a glass affect whether eggs sink or float in the water?, Are eggs more likely to float in fresh water or salty water?,] Answer index: 1	Subject: Engineer Description: iption: Which of the following could Luke's test show? Luke had a cookie recipe that made soft, thick cookies. But he preferred crunchy cookies. Luke read that using different types of sugar affects how firm the cookies are. His recipe used both white and brown sugar, so he decided to see if the cookies would be crunchy if he didn't use any brown sugar. Luke baked a batch of cookies using his recipe, but he left out the brown sugar and doubled the amount of white sugar. He baked the cookies for the same amount of time as in his original recipe. After the cookies finished baking and cooling, he tried one to find out how firm it was. The passage below describes how the engineering-design process was used to test a solution to a problem. Read the passage. Then answer the question below.  Picture: Choices: [if cookies made with only white sugar were soft, if baking cookies for longer made them more crunchy, if cookies made with double the amount of brown sugar were crunchy,] Answer index: 0
Subject: Engineer Description: iption: Identify the question that Myra's experiment can best answer. The passage below describes an experiment. Read the passage and then follow the instructions below. Myra glued lids onto 16 cardboard shoe boxes of equal size. She painted eight of the boxes black and eight of the boxes white. Myra made a small hole in the side of each box and then stuck a thermometer partially into each hole so she could measure the temperatures inside the boxes. She placed the boxes in direct sunlight in her backyard. Two hours later, she measured the temperature inside each box. Myra compared the average temperature inside the black boxes to the average temperature inside the white boxes.  Picture: Choices: [Do the temperatures inside boxes depend on the sizes of the boxes?, Do the insides of white boxes get hotter than the insides of black boxes when the boxes are left in the sun?,] Answer index: 1	Subject: Engineer Description: iption: Which of the following could Zoe and Evelyn's test show? Zoe and Evelyn were making batches of concrete for a construction project. To make the concrete, they mixed together dry cement powder, gravel, and water. Then, they checked if each batch was firm enough using a test called a slump test. They poured some of the fresh concrete into an upside-down metal cone. They left the concrete in the metal cone for 30 seconds. Then, they lifted the cone to see if the concrete stayed in a cone shape or if it collapsed. If the concrete in a batch collapsed, they would know the batch should not be used. The passage below describes how the engineering-design process was used to test a solution to a problem. Read the passage. Then answer the question below.  Picture: Choices: [if the concrete from each batch took the same amount of time to dry, if a new batch of concrete was firm enough to use,] Answer index: 1
Subject: Engineer Description: iption: Identify the question that Belle's experiment can best answer. The passage below describes an experiment. Read the passage and then follow the instructions below. Belle planted 25 tomato seeds one-half inch below the soil surface in each of six pots. Belle added an equal amount of fertilizer to three of the six pots. She placed the pots in a plant growth chamber where all the seeds experienced the same temperature, amount of light, and humidity level. After two weeks, Belle counted the number of seedlings that grew in each pot. She compared the number of seedlings in the pots with fertilizer to the number of seedlings in the pots without fertilizer.  Picture: Choices: [Do more tomato seedlings grow when they are planted in soil with fertilizer compared to soil without fertilizer?, Does the humidity level where tomato seeds are planted affect the number of tomato seedlings that grow?,] Answer index: 0	Subject: Engineer Description: iption: In this experiment, which were part of a control group? The passage below describes an experiment. After a severe winter storm, Sandeep's driveway was covered with ice. He read that salt makes ice melt at a lower temperature. Before covering his entire driveway with salt, he wanted to know if adding salt could actually help melt ice in the freezing outdoor temperatures. Sandeep weighed twenty ice cubes. He sprinkled salt on half of the ice cubes and left the other half unsalted. He placed all the ice cubes outside. One hour later, Sandeep quickly dried each ice cube and reweighed it to see how much it had melted.  Picture: Choices: [the salted ice cubes, the unsalted ice cubes,] Answer index: 1

Table 10: Human (manual) evaluation problem set (part 3).

Subject: Science
Description: Select the gray mineral.
Picture: None

Choices:
Answer index: 1

Subject: Science
Description: Select the one substance that is not a mineral.
Picture: None

Choices:
Answer index: 0

Subject: Science
Description: iption: This organism is a spot-fin porcupinefish. Its scientific name is Diodon hystrix. Select the organism in the same genus as the spot-fin porcupinefish.
Picture:

Choices:
Answer index: 1

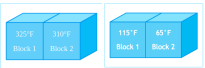
Subject: Science
Description: Select the liquid.
Picture: None

Choices:
Answer index: 3

Subject: Science
Description: iption: Fish are a group of animals with similar traits. The following traits can be used to identify fish: They have fins, not limbs. They make eggs with no shells. Observe the animals and read the descriptions. Select the one animal that has all of the fish traits listed above. Brown pelicans live along the west coast of North America. They dive underwater to catch fish in their beaks. Brown pelicans keep their eggs warm by standing on the shells with their large, webbed feet. Salmon lay eggs with no shells at the bottom of freshwater streams. Salmon use their powerful fins to swim. They can even jump up small waterfalls!
Picture: None
Choices:
Answer index: 0

Subject: Science
Description: iption: Two identical blocks are heated to different temperatures. The blocks are placed so that they touch each other. Heat can flow from one block to another but cannot escape from the blocks. Later, the temperature of each block is measured again. Which pair of temperatures is possible?
Picture:

Choices:

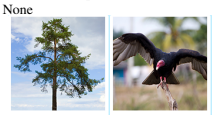


Answer index: 0

Subject: Science
Description: iption: Two solid blocks are heated to the temperatures shown. The blocks are placed so they touch. Which diagram shows the direction heat will flow?
Picture: None
Choices:

Answer index: 0

Subject: Science
Description: Select the plant.
Picture: None



Choices:
Answer index: 0

Subject: Science
Description: iption: Use the data to answer the question below. Is the following statement about our solar system true or false? Of the four smallest planets, two are made mainly of gas.
Picture:

Choices: [false, true,]
Answer index: 0

Subject: Science
Description: iption: Think about the magnetic force between the magnets in each pair. Which of the following statements is true? The images below show two pairs of magnets. The magnets in different pairs do not affect each other. All the magnets shown are made of the same material, but some of them are different sizes.
Picture:

Choices: [The magnitude of the magnetic force is smaller in Pair 2., The magnitude of the magnetic force is the same in both pairs., The magnitude of the magnetic force is smaller in Pair 1.,]
Answer index: 0

Table 11: Human (manual) evaluation problem set (part 4).

Subject: Science
Description: iption: Think about the magnetic force between the magnets in each pair. Which of the following statements is true? The images below show two pairs of magnets. The magnets in different pairs do not affect each other. All the magnets shown are made of the same material.
Picture: 
Choices: [The magnetic force is stronger in Pair 2., The magnetic force is stronger in Pair 1., The strength of the magnetic force is the same in both pairs.,]
Answer index: 0

Subject: Science
Description: Select the gas.
Picture: None

hammer

bowling ball

air inside a tire

door
Choices:
Answer index: 2

Subject: Science
Description: Select the plant.
Picture: None

Orchids can grow flowers.

Elephants eat plants.

Peregrine falcons walk and fly.

Manta rays swim underwater.
Choices:
Answer index: 0

Subject: Science
Description: iption: Which property matches this object? Select the better answer.
Picture: 

Choices: [soft, smooth,]
Answer index: 1

Subject: Science
Description: iption: Select the animal that does not have a backbone. Hint: Insects, spiders, and worms do not have backbones.
Picture: None

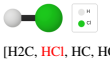
bull ant

harvest mouse
Choices:
Answer index: 0

Subject: Science
Description: iption: The diagram below is a model of two solutions. Each green ball represents one particle of solute. Which solution has a higher concentration of green particles?
Picture: 
Choices: [neither; their concentrations are the same, Solution A, Solution B,]
Answer index: 2

Subject: Science
Description: iption: Two solid blocks are at different temperatures. The blocks are touching. Which picture shows how heat will move?
Picture: None

Choices:
Answer index: 1

Subject: Science
Description: Select the chemical formula for this molecule.
Picture: 
Choices: [H2C, HCl, HC, HCl2,]
Answer index: 1

Subject: Science
Description: iption: Which statement best describes the climate of Bangor? Hint: Summers in the Northern Hemisphere occur in June, July, and August. Winters in the Northern Hemisphere occur in December, January, and February. Bangor, Maine, is a city in the United States. It has a warm summer continental climate.

Picture:
Choices: [Summers have higher temperatures and slightly more precipitation than winters., On average, On average,]
Answer index: 1

Subject: Science
Description: Select the temperature shown by this thermometer.

Picture:
Choices: [13°F, 61°F, 56°F,]
Answer index: 2

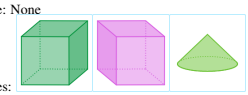
Table 12: Human (manual) evaluation problem set (part 5).

Subject: Math
Description: iption: This table shows Jason's January budget. What could Jason do to balance his budget?
Picture: 
Choices: [increase income from shoveling snow to 40, *spend*15 less at Pizza Palace, spend only $40 at the arcade, spend $20 more on video games,]
Answer index: 2

Subject: Math
Description: In solving this triangle, which law must you use first?
Picture: 
Choices: [Law of Cosines, Law of Sines,]
Answer index: 1

Subject: Math
Description: Is this angle acute, right, obtuse, or straight?
Picture:
Choices: [straight, obtuse, acute, right,]
Answer index: 1

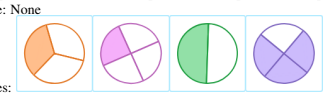
Subject: Math
Description: iption: Look at this cube: If the side lengths are tripled, then which of the following statements about its volume will be true?
Picture: 
Choices: [The ratio of the new volume to the old volume will be 81:1., The ratio of the new volume to the old volume will be 1:8., The ratio of the new volume to the old volume will be 3:1., The ratio of the new volume to the old volume will be 27:1.,]
Answer index: 3

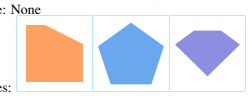
Subject: Math
Description: Which shape is a cone?
Picture: None
Choices: 
Answer index: 2

Subject: Math
Description: Is the function f(x) continuous on the open interval (3,7)?
Picture: 
Choices: [no, yes,]
Answer index: 1

Subject: Math
Description: iption: Use the diagram to help you answer the question below. Which of the following is a rational number but not an integer?
Picture: 
Choices: [–123, 83, 194, 6.53,]
Answer index: 3

Subject: Math
Description: Is this polygon a trapezoid?
Picture:
Choices: [no, yes,]
Answer index: 1

Subject: Math
Description: Look at the colored part of each shape. Which shape shows one-third?
Picture: None
Choices: 
Answer index: 0

Subject: Math
Description: Which shape has 5 equal sides?
Picture: None
Choices: 
Answer index: 1

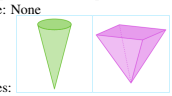
Subject: Math
Description: iption: Identify the cross section of this object. Assume objects are perpendicular if they appear so.
Picture: 
Choices:
Answer index: 0

Subject: Math
Description: Are there more circles or triangles?
Picture: 
Choices: [circles, triangles,]
Answer index: 0

Subject: Math
Description: iption: Look at this shape: Which image shows a reflection?
Picture:
Choices: [C, A, B,]
Answer index: 2

Subject: Math
Description: Is the function f(x) continuous?
Picture: 
Choices: [no, yes,]
Answer index: 0

Subject: Math
Description: An ice cream sundae costs 1 dollar and 41 cents. Do you have enough money to buy it?
Picture: 
Choices: [yes, no,]
Answer index: 0

Subject: Math
Description: Which shape has a triangle as a face?
Picture: None
Choices: 
Answer index: 1

Subject: Math
Description: iption: Look at this figure: What is the shape of its bases?
Picture: 
Choices: [decagon, octagon, rectangle, circle,]
Answer index: 2

Subject: Math
Description: The graph below shows a function. Is its inverse also a function?
Picture: 
Choices: [yes, no,]
Answer index: 0

Subject: Math
Description: What is the range of this exponential function?
Picture: 
Choices: 
Answer index: 0

Subject: Math
Description: iption: Look at this graph: Is this relation a function?
Picture: 
Choices: [no, yes,]
Answer index: 1

Table 13: Human (manual) evaluation problem set (part 6).

Subject	Grade	Skills
Science	grade-2	classify-fruits-and-vegetables-as-plant-parts, classify-matter-as-solid-liquid-or-gas, classify-matter-as-solid-or-liquid, classify-rocks-and-minerals-by-color-and-shape, compare-properties-of-materia ls, compare-properties-of-objects, compare-temperatures-on-thermometers, find-evidence-of-changes-to-earths-surface, identify-animals-with-and-without-backbones, identify-earth-s-land-features, identi fy-living-and-nonliving-things, identify-magnets-that-attract-or-repel, identify-mammals-birds-fish-reptiles-and-amphibians, identify-materials-in-objects, identify-plants-and-animals, identify-proper ties-of-an-object, identify-pushes-and-pulls, identify-solids-and-liquids, identify-solids-liquids-and-gases, identifying-mixtures, natural-resources, predict-heat-flow, read-a-thermometer
	grade-3	animal-adaptations-beaks-mouths-and-necks, animal-adaptations-feet-and-limbs, animal-adaptations-skins-and-body-coverings, classify-fruits-and-vegetables-as-plant-parts, classify-matter-as-solid-liqui d-or-gas, classify-rocks-and-minerals-by-color-shape-and-texture, classify-rocks-as-igneous-sedimentary-or-metamorphic, compare-ancient-and-modern-organisms-use-observations-to-support-a-hypothesis, c ompare-properties-of-materials, compare-properties-of-objects, compare-strengths-of-magnetic-forces, compare-temperatures-on-thermometers, find-evidence-of-changes-to-earths-surface, how-do-balanced-a nd-unbalanced-forces-affect-motion, identify-earth-s-land-features, identify-ecosystems, identify-living-and-nonliving-things, identify-magnets-that-attract-or-repel, identify-mammals-birds-fish-repti les-and-amphibians, identify-materials-in-objects, identify-minerals-using-properties, identify-plants-and-animals, identify-properties-of-an-object, identify-pushes-and-pulls, identify-rocks-using-pr operties, identify-roles-in-food-chains, identify-solids-liquids-and-gases, identify-vertebrates-and-invertebrates, interpret-food-webs, natural-resources, predict-heat-flow, predict-temperature-chang es, read-a-thermometer, use-climate-data-to-make-predictions, use-data-to-describe-u-s-climates, use-data-to-describe-world-climates, weather-and-climate-around-the-world
	grade-4	animal-adaptations-beaks-mouths-and-necks, animal-adaptations-feet-and-limbs, animal-adaptations-skins-and-body-coverings, classify-fruits-and-vegetables-as-plant-parts, classify-rocks-as-igneous-sedi mentary-or-metamorphic, compare-amplitudes-and-wavelengths-of-waves, compare-ancient-and-modern-organisms-use-observations-to-support-a-hypothesis, compare-properties-of-materials, compare-properties- of-objects, compare-strengths-of-magnetic-forces, compare-temperatures-on-thermometers, describe-classify-and-compare-kingdoms, evaluate-natural-energy-sources, how-do-balanced-and-unbalanced-forces-a ffect-motion, identify-and-classify-fossils, identify-and-sort-solids-liquids-and-gases, identify-common-and-scientific-names, identify-directions-of-forces, identify-earths-land-features-using-photog raphs, identify-earths-land-features-using-satellite-images, identify-ecosystems, identify-living-and-nonliving-things, identify-magnets-that-attract-or-repel, identify-mammals-birds-fish-reptiles-and -amphibians, identify-minerals-using-properties, identify-phases-of-the-moon, identify-rocks-using-properties, identify-roles-in-food-chains, identify-vertebrates-and-invertebrates, interpret-food-web s, origins-of-scientific-names, predict-heat-flow, predict-temperature-changes, read-a-thermometer, use-climate-data-to-make-predictions, use-data-to-describe-climates, use-evidence-to-classify-animal s, use-evidence-to-classify-mammals-birds-fish-reptiles-and-amphibians, use-scientific-names-to-classify-organisms, weather-and-climate-around-the-world
	grade-5	animal-adaptations-beaks-mouths-and-necks, animal-adaptations-feet-and-limbs, animal-adaptations-skins-and-body-coverings, classify-elementary-substances-and-compounds-using-models, classify-fruits-an d-vegetables-as-plant-parts, classify-rocks-as-igneous-sedimentary-or-metamorphic, compare-amplitudes-and-wavelengths-of-waves, compare-ancient-and-modern-organisms-use-observations-to-support-a-hypot hesis, compare-magnitudes-of-magnetic-forces, compare-properties-of-objects, describe-classify-and-compare-kingdoms, evaluate-natural-energy-sources, flowering-plant-and-conifer-life-cycles, how-do-ba lanced-and-unbalanced-forces-affect-motion, identify-and-classify-fossils, identify-common-and-scientific-names, identify-directions-of-forces, identify-earths-land-features-using-photographs, identif y-earths-land-features-using-satellite-images, identify-ecosystems, identify-magnets-that-attract-or-repel, identify-mammals-birds-fish-reptiles-and-amphibians, identify-phases-of-the-moon, identify-r ocks-and-minerals, identify-roles-in-food-chains, identify-the-photosynthetic-organism, identify-vertebrates-and-invertebrates, match-chemical-formulas-to-ball-and-stick-models, moss-and-fern-life-cyc les, origins-of-scientific-names, predict-heat-flow, predict-temperature-changes, use-data-to-describe-climates, use-evidence-to-classify-animals, use-evidence-to-classify-mammals-birds-fish-reptiles- and-amphibians, use-scientific-names-to-classify-organisms, weather-and-climate-around-the-world
	grade-6	analyze-data-to-compare-properties-of-planets, classify-elementary-substances-and-compounds-using-models, classify-rocks-as-igneous-sedimentary-or-metamorphic, classify-symbiotic-relationships, compar e-ages-of-fossils-in-a-rock-sequence, compare-amplitudes-wavelengths-and-frequencies-of-waves, compare-concentrations-of-solutions, compare-magnitudes-of-magnetic-forces, compare-thermal-energy-transf ers, describe-populations-communities-and-ecosystems, describe-tectonic-plate-boundaries-around-the-world, describe-the-effects-of-gene-mutations-on-organisms, diffusion-across-membranes, flowering-pl ant-and-conifer-life-cycles, identify-and-compare-air-masses, identify-common-and-scientific-names, identify-earths-land-features-using-photographs, identify-earths-land-features-using-satellite-image s, identify-ecosystems, identify-elementary-substances-and-compounds-using-models, identify-how-particle-motion-affects-temperature-and-pressure, identify-phases-of-the-moon, identify-rocks-and-minera ls, identify-the-photosynthetic-organism, match-chemical-formulas-to-ball-and-stick-models, moss-and-fern-life-cycles, origins-of-scientific-names, predict-heat-flow-and-temperature-changes, use-data- to-describe-climates, use-scientific-names-to-classify-organisms, weather-and-climate-around-the-world
	grade-7	analyze-data-to-compare-properties-of-planets, angiosperm-and-conifer-life-cycles, classify-elementary-substances-and-compounds-using-models, classify-rocks-as-igneous-sedimentary-or-metamorphic, clas sify-symbiotic-relationships, compare-ages-of-fossils-in-a-rock-sequence, compare-amplitudes-wavelengths-and-frequencies-of-waves, compare-concentrations-of-solutions, compare-magnitudes-of-magnetic-f orces, compare-thermal-energy-transfers, describe-populations-communities-and-ecosystems, describe-tectonic-plate-boundaries-around-the-world, describe-the-effects-of-gene-mutations-on-organisms, diff usion-across-membranes, identify-and-compare-air-masses, identify-chemical-formulas-for-ball-and-stick-models, identify-common-and-scientific-names, identify-ecosystems, identify-how-particle-motion-a ffects-temperature-and-pressure, identify-phases-of-the-moon, identify-rocks-and-minerals, identify-the-photosynthetic-organism, moss-and-fern-life-cycles, origins-of-scientific-names, predict-heat-fl ow-and-temperature-changes, use-data-to-describe-climates, use-scientific-names-to-classify-organisms
	grade-8	analyze-data-to-compare-properties-of-planets, angiosperm-and-conifer-life-cycles, classify-elementary-substances-and-compounds-using-models, classify-symbiotic-relationships, compare-ages-of-fossils- in-a-rock-sequence, compare-amplitudes-wavelengths-and-frequencies-of-waves, compare-concentrations-of-solutions, compare-magnitudes-of-magnetic-forces, compare-thermal-energy-transfers, describe-popu lations-communities-and-ecosystems, describe-tectonic-plate-boundaries-around-the-world, describe-the-effects-of-gene-mutations-on-organisms, diffusion-across-membranes, identify-and-compare-air-masse s, identify-chemical-formulas-for-ball-and-stick-models, identify-common-and-scientific-names, identify-ecosystems, identify-how-particle-motion-affects-temperature-and-pressure, identify-phases-of-th e-moon, identify-rocks-and-minerals, identify-the-photosynthetic-organism, moss-and-fern-life-cycles, origins-of-scientific-names, predict-heat-flow-and-temperature-changes, use-data-to-describe-clima tes, use-punnett-squares-to-calculate-probabilities-of-offspring-types, use-punnett-squares-to-calculate-ratios-of-offspring-types, use-scientific-names-to-classify-organisms
Technology	-	cables, font, icons, logo, parts, peripherals, photo, web, others
Engineering	grade-5	identify-laboratory-tools
	grade-6	evaluate-tests-of-engineering-design-solutions, identify-control-and-experimental-groups, identify-independent-and-dependent-variables, identify-laboratory-tools, identify-the-experimental-question, l aboratory-safety-equipment
	grade-7	evaluate-tests-of-engineering-design-solutions, identify-control-and-experimental-groups, identify-independent-and-dependent-variables, identify-laboratory-tools, identify-the-experimental-question, l aboratory-safety-equipment
	grade-8	identify-control-and-experimental-groups, identify-laboratory-tools, identify-the-experimental-question, laboratory-safety-equipment

Table 14: Full skill summary(part 1), including science, technology and engineering skills.

Subject	Grade	Skills
	algebra-1	compare-linear-functions-graphs-and-equations, compare-linear-functions-tables-graphs-and-equations, describe-linear-and-exponential-growth-and-decay, domain-and-range-of-absolute-value-functions-grap hs, domain-and-range-of-exponential-functions-graphs, domain-and-range-of-square-root-functions-graphs, factor-quadratics-using-algebra-tiles, identify-direct-variation-and-inverse-variation, identify -functions, identify-functions-vertical-line-test, identify-linear-and-exponential-functions-from-graphs, identify-linear-and-exponential-functions-from-tables, identify-linear-functions-from-graphs-a nd-equations, identify-linear-functions-from-tables, identify-linear-quadratic-and-exponential-functions-from-graphs, identify-linear-quadratic-and-exponential-functions-from-tables, identify-properti onal-relationships, interpret-a-scatter-plot, interpret-the-slope-and-y-intercept-of-a-linear-function, linear-functions-over-unit-intervals, match-exponential-functions-and-graphs-ii, model-and-solve -linear-equations-using-algebra-tiles, multiply-two-binomials-using-algebra-tiles, perimeter-and-area-changes-in-scale, perimeter-area-and-volume-changes-in-scale, special-right-triangles, surface-are a-and-volume-changes-in-scale, write-compound-inequalities-from-graphs
	algebra-2	classify-variation, describe-linear-and-exponential-growth-and-decay, domain-and-range-of-absolute-value-functions-graphs, domain-and-range-of-exponential-and-logarithmic-functions, domain-and-range-o f-radical-functions, factor-quadratics-using-algebra-tiles, find-inverse-functions-and-relations, find-solutions-using-a-table, graphs-of-angles, identify-the-direction-a-parabola-opens, linear-functi ons-over-unit-intervals, match-exponential-functions-and-graphs, outliers-in-scatter-plots, solve-a-triangle
	calculus	describe-linear-and-exponential-growth-and-decay, determine-continuity-on-an-interval-using-graphs, determine-continuity-using-graphs, determine-one-sided-continuity-using-graphs, domain-and-range, do main-and-range-of-exponential-and-logarithmic-functions, find-inverse-functions-and-relations, find-limits-at-vertical-asymptotes-using-graphs, identify-functions, identify-graphs-of-continuous-functi ons

Table 15: Full skill summary(part 2), including math skills for algebra-{1,2} and calculus.

Subject	Grade	Skills
Math	grade-1	addition-sentences-up-to-10-what-does-the-model-show, addition-sentences-up-to-10-which-model-matches, addition-sentences-using-number-lines-sums-up-to-20, am-or-pm, certain-probable-unlikely-and-impossible, compare-clocks, compare-money-amounts, compare-objects-length-and-height, compare-sides-and-corners, compare-size-weight-and-capacity, compare-vertices-edges-and-faces, comparing-review, count-sides-and-corners, count-to-fill-a-ten-frame, cubes-and-rectangular-prisms, equal-sides, estimate-to-the-nearest-ten, even-or-odd, find-the-next-shape-in-a-growing-pattern, find-the-next-shape-in-a-pattern, flip-turn-and-slide, holds-more-or-less, identify-faces-of-three-dimensional-shapes, identify-fourths, identify-halves, identify-halves-and-fourths, identify-halves-thirds-and-fourths, identify-shapes-traced-from-solids, identify-thirds, interpret-bar-graphs-ii, light-and-heavy, match-analog-and-digital-clocks, match-analog-clocks-and-times, match-digital-clocks-and-times, more-less-and-equally-likely, name-the-three-dimensional-shape, name-the-two-dimensional-shape, names-and-values-of-all-coins, names-and-values-of-common-coins, open-and-closed-shapes, ordinal-numbers, purchases-do-you-have-enough-money, read-a-calendar, read-a-calendar-ii, rhombuses, select-three-dimensional-shapes, select-two-dimensional-shapes, shapes-of-everyday-objects, simple-fractions-what-fraction-does-the-shape-show, square-corners, subtraction-sentences-up-to-10-which-model-matches, subtraction-sentences-using-number-lines-up-to-10, subtraction-sentences-using-number-lines-up-to-20, symmetry, time-and-clocks-word-problems, times-of-everyday-events, two-dimensional-and-three-dimensional-shapes, which-bar-graph-is-correct, which-picture-graph-is-correct, which-table-is-correct, which-tally-chart-is-correct, wide-and-narrow
	grade-2	am-or-pm, certain-probable-unlikely-and-impossible, choose-the-appropriate-measuring-tool, compare-clocks, compare-sides-and-vertices, compare-vertices-edges-and-faces, correct-amount-of-change, cubes, equal-sides, equivalent-amounts-of-money-up-to-1-dollar, estimate-to-the-nearest-ten, even-or-odd, find-the-next-shape-in-a-growing-pattern, find-the-next-shape-in-a-repeating-pattern, flip-turn-and-slide, fractions-of-a-group, fractions-of-a-whole-modeling-word-problems, greatest-and-least-word-problems-up-to-100, greatest-and-least-word-problems-up-to-1000, how-much-more-to-make-a-dollar, identify-faces-of-three-dimensional-shapes, identify-fourths, identify-halves, identify-halves-thirds-and-fourths, identify-lines-of-symmetry, identify-multiplication-sentences-for-equal-groups, identify-repeated-addition-in-arrays-sums-to-10, identify-repeated-addition-in-arrays-sums-to-25, identify-shapes-traced-from-solids, identify-the-fraction, identify-thirds, interpret-bar-graphs-ii, interpret-tally-charts, match-addition-sentences-and-models-sums-to-10, match-analog-and-digital-clocks, match-analog-clocks-and-times, match-digital-clocks-and-times, more-less-and-equally-likely, name-the-three-dimensional-shape, name-the-two-dimensional-shape, names-and-values-of-all-coins, names-and-values-of-common-coins, ordinal-numbers-up-to-10th, place-value-models-up-to-hundreds, place-value-tens-and-ones, place-value-up-to-hundreds, place-value-up-to-thousands, purchases-do-you-have-enough-money-up-to-1-dollar, purchases-do-you-have-enough-money-up-to-5-dollars, read-a-calendar, read-a-calendar-ii, read-a-thermometer, select-figures-with-a-given-area, select-three-dimensional-shapes, shapes-of-everyday-objects, skip-counting-stories, symmetry, which-bar-graph-is-correct, which-picture-shows-more-up-to-5-dollars, which-shape-illustrates-the-fraction, which-table-is-correct, which-tally-chart-is-correct, write-subtraction-sentences-to-describe-pictures-up-to-18, write-subtraction-sentences-to-describe-pictures-up-to-two-digits
	grade-3	acute-obtuse-and-right-triangles, am-or-pm, angles-greater-than-less-than-or-equal-to-a-right-angle, certain-probable-unlikely-and-impossible, choose-the-appropriate-measuring-tool, compare-area-and-perimeter-of-two-figures, compare-fractions-in-recipes, compare-fractions-using-models, compare-fractions-using-number-lines, coordinate-planes-as-maps, correct-amount-of-change, division-input-output-tables-find-the-rule, find-the-next-shape-in-a-pattern, fractions-of-a-group-denominators-2-3-4-6-8, fractions-of-a-group-unit-fractions, identify-equivalent-fractions-on-number-lines, identify-faces-of-three-dimensional-shapes, identify-multiplication-expressions-for-arrays, identify-multiplication-expressions-for-equal-groups, identify-parallelograms, identify-rhombuses, identify-three-dimensional-shapes, identify-trapezoids, identify-two-dimensional-shapes, identify-unit-fractions-on-number-lines, interpret-line-graphs, is-it-a-polygon, lines-line-segments-and-rays, match-analog-and-digital-clocks, match-clocks-and-times, match-fractions-to-models-halves-thirds-and-fourths, match-mixed-numbers-to-models, multiplication-input-output-tables-find-the-rule, open-and-closed-shapes, parallel-perpendicular-and-intersecting-lines, parallel-sides-in-quadrilaterals, purchases-do-you-have-enough-money-up-to-10-dollars, read-a-calendar, read-a-thermometer, reading-schedules, reflection-rotation-and-translation, scalene-isosceles-and-equilateral-triangles, select-figures-with-a-given-area, select-fractions-equivalent-to-whole-numbers-using-models, shapes-of-everyday-objects, symmetry, which-picture-shows-more
	grade-4	acute-obtuse-and-right-triangles, acute-right-obtuse-and-straight-angles, angles-as-fractions-of-a-circle, angles-of-90-180-270-and-360-degrees, classify-triangles, compare-area-and-perimeter-of-two-figures, compare-decimals-using-models, compare-fractions-in-recipes, compare-fractions-using-models, compare-fractions-with-like-numerators-or-denominators-using-models, decompose-fractions-into-unit-fractions-using-models, elapsed-time, estimate-angle-measurements, find-the-next-shape-in-a-pattern, fractions-of-a-whole-word-problems, identify-equivalent-fractions-using-number-lines, identify-faces-of-three-dimensional-figures, identify-lines-of-symmetry, identify-parallel-perpendicular-and-intersecting-lines, identify-parallelograms, identify-rhombuses, identify-three-dimensional-figures, identify-trapezoids, interpret-bar-graphs, interpret-stem-and-leaf-plots, is-it-a-polygon, measure-angles-with-a-protractor, multiplication-input-output-tables-find-the-rule, multiply-fractions-by-whole-numbers-using-models, multiply-unit-fractions-by-whole-numbers-using-models, nets-of-three-dimensional-figures, parallel-perpendicular-and-intersecting-lines, parallel-sides-in-quadrilaterals, points-lines-line-segments-rays-and-angles, properties-of-three-dimensional-figures, rotational-symmetry, scalene-isosceles-and-equilateral-triangles, sides-and-angles-of-quadrilaterals, transportation-schedules, what-decimal-number-is-illustrated
	grade-5	acute-obtuse-and-right-triangles, adjust-a-budget, angles-of-90-180-270-and-360-degrees, classify-triangles, compare-decimals-using-grids, compare-fractions-and-mixed-numbers, compare-patterns, fractions-of-a-whole-word-problems, identify-parallelograms, identify-rhombuses, identify-three-dimensional-figures, identify-trapezoids, interpret-bar-graphs, is-it-a-polygon, line-symmetry, mean-find-the-missing-number, median-find-the-missing-number, multiplication-input-output-tables-find-the-rule, multiply-unit-fractions-by-whole-numbers-using-models, multiplying-fractions-by-whole-numbers-choose-the-model, nets-of-three-dimensional-figures, parallel-perpendicular-and-intersecting-lines, parallel-sides-in-quadrilaterals, parts-of-a-circle, points-lines-line-segments-rays-and-angles, range-find-the-missing-number, reflection-rotation-and-translation, regular-and-irregular-polygons, rotational-symmetry, rotational-symmetry-amount-of-rotation, scalene-isosceles-and-equilateral-triangles, three-dimensional-figures-viewed-from-different-perspectives, types-of-angles, understanding-probability
	grade-6	absolute-value-and-integers-word-problems, changes-in-mean-median-mode-and-range, classify-rational-numbers-using-a-diagram, classify-triangles, compare-and-order-rational-numbers-using-number-lines, compare-area-and-perimeter-of-two-figures, compare-checking-accounts, front-side-and-top-view, identify-complementary-supplementary-vertical-adjacent-and-congruent-angles, identify-equivalent-expressions-using-strip-models, identify-polyhedra, identify-trapezoids, interpret-bar-graphs, interpret-double-bar-graphs, interpret-graphs-of-proportional-relationships, interpret-histograms, line-symmetry, mean-median-mode-and-range-find-the-missing-number, model-and-solve-equations-using-algebra-tiles, nets-of-three-dimensional-figures, occupations-education-and-income, quadrants, rational-numbers-find-the-sign, reflection-rotation-and-translation, rotational-symmetry, rotational-symmetry-amount-of-rotation, similar-and-congruent-figures, understanding-area-of-a-triangle, understanding-area-of-trapezoids, understanding-percents-strip-models, which-figure-is-being-described, which-is-the-better-coupon, which-model-represents-the-ratio
	grade-7	apply-addition-and-subtraction-rules, apply-multiplication-and-division-rules, bases-of-three-dimensional-figures, changes-in-mean-median-mode-and-range, classify-quadrilaterals, classify-rational-numbers-using-a-diagram, compare-and-order-integers, cross-sections-of-three-dimensional-figures, describe-a-sequence-of-transformations, front-side-and-top-view, identify-alternate-interior-and-alternate-exterior-angles, identify-complementary-supplementary-vertical-and-adjacent-angles, identify-equivalent-linear-expressions-using-algebra-tiles, identify-linear-and-nonlinear-functions, identify-reflections-rotations-and-translations, identify-trapezoids, identify-trends-with-scatter-plots, interpret-circle-graphs, interpret-graphs-of-proportional-relationships, line-symmetry, make-predictions-with-scatter-plots, mean-median-mode-and-range-find-the-missing-number, model-and-solve-equations-using-algebra-tiles, nets-of-three-dimensional-figures, parallel-perpendicular-and-intersecting-lines, parts-of-a-circle, perimeter-and-area-changes-in-scale, rational-numbers-find-the-sign, rotational-symmetry, rotational-symmetry-amount-of-rotation, similar-and-congruent-figures, simplify-expressions-by-combining-like-terms-with-algebra-tiles, transversals-of-parallel-lines-name-angle-pairs, which-is-the-better-coupon
	grade-8	angle-angle-criterion-for-similar-triangles, apply-addition-and-subtraction-rules, apply-addition-subtraction-multiplication-and-division-rules, apply-multiplication-and-division-rules, base-plans, changes-in-mean-median-mode-and-range, classify-quadrilaterals, compare-and-order-integers, compare-linear-functions-graphs-and-equations, compare-linear-functions-tables-graphs-and-equations, congruent-triangles-sss-sas-and-asa, describe-a-sequence-of-transformations, front-side-and-top-view, identify-alternate-interior-and-alternate-exterior-angles, identify-complementary-supplementary-vertical-adjacent-and-congruent-angles, identify-congruent-figures, identify-functions-graphs, identify-linear-and-nonlinear-functions-graphs-and-equations, identify-linear-and-nonlinear-functions-tables, identify-lines-of-best-fit, identify-reflections-rotations-and-translations, identify-similar-triangles, identify-trapezoids, identify-trends-with-scatter-plots, interpret-graphs-of-proportional-relationships, interpret-the-slope-and-y-intercept-of-a-linear-function, irrational-numbers-on-number-lines, line-symmetry, make-predictions-with-scatter-plots, mean-median-mode-and-range-find-the-missing-number, model-and-solve-equations-using-algebra-tiles, multiply-polynomials-using-algebra-tiles, nets-of-three-dimensional-figures, parts-of-a-circle, parts-of-three-dimensional-figures, perimeter-and-area-changes-in-scale, quadrants-and-axes, rotational-symmetry, rotational-symmetry-amount-of-rotation, similar-and-congruent-figures, transversals-of-parallel-lines-name-angle-pairs
	kindergarten	addition-sentences-up-to-10-what-does-the-model-show, addition-sentences-up-to-10-which-model-matches, addition-sentences-up-to-5-what-does-the-model-show, addition-sentences-up-to-5-which-model-matches, am-or-pm, are-there-enough, circles, classify-shapes-by-color, coin-names-penny-through-quarter, compare-sides-and-corners, compare-size-weight-and-capacity, compare-two-groups-of-coins-pennies-through-dimes, cones, count-corners, count-cubes-up-to-10, count-cubes-up-to-5, count-dots-0-to-5, count-dots-up-to-10, count-money-pennies-and-nickels, count-money-pennies-through-dimes, count-on-ten-frames-up-to-10, count-pictures-up-to-10, count-pictures-up-to-3, count-pictures-up-to-5, count-scattered-shapes-up-to-10, count-scattered-shapes-up-to-5, count-shapes-in-rings-up-to-10, count-shapes-in-rows-up-to-10, count-shapes-in-rows-up-to-5, count-shapes-up-to-3, count-sides, count-sides-and-corners, count-to-100, count-to-fill-a-ten-frame, cubes, curved-parts, cylinders, different, equal-sides, fewer-and-more-compare-by-counting, fewer-and-more-compare-by-matching, fewer-and-more-compare-in-a-mixed-group, fewer-and-more-up-to-20, fewer-more-and-same, flat-and-solid-shapes, hexagons, holds-more-or-less, identify-halves-thirds-fourths, identify-pictures-with-symmetry, identify-shapes-traced-from-solids, inside-and-outside, introduction-to-symmetry, light-and-heavy, match-analog-and-digital-clocks, match-analog-clocks-and-times, match-digital-clocks-and-times, more-or-less-likely, name-the-three-dimensional-shape, name-the-two-dimensional-shape, one-less-with-pictures-up-to-10, one-less-with-pictures-up-to-5, one-more-and-one-less-with-pictures-up-to-10, one-more-with-pictures-up-to-10, one-more-with-pictures-up-to-5, ordinal-numbers-up-to-fifth, ordinal-numbers-up-to-tenth, rectangles, represent-numbers-up-to-10, represent-numbers-up-to-20, represent-numbers-with-pictures-up-to-3, represent-numbers-with-pictures-up-to-5, represent-numbers-with-shapes-up-to-3, represent-numbers-with-shapes-up-to-5, select-three-dimensional-shapes, select-two-dimensional-shapes, shapes-of-everyday-objects, spheres, square-corners, squares, subtraction-sentences-up-to-10-what-does-the-model-show, subtraction-sentences-up-to-10-which-model-matches, subtraction-sentences-up-to-5-what-does-the-model-show, subtraction-sentences-up-to-5-which-model-matches, take-apart-10-words, take-apart-numbers-up-to-10-words, take-apart-numbers-up-to-5-words, tall-and-short, times-of-everyday-events, triangles, wide-and-narrow
	pre-k	addition-sentences-up-to-10-what-does-the-model-show, addition-sentences-up-to-10-which-model-matches, addition-sentences-up-to-5-what-does-the-model-show, addition-sentences-up-to-5-which-model-matches, are-there-enough, circles, circles-squares-and-triangles, circles-squares-triangles-and-rectangles, classify-shapes-by-color, compare-size-weight-and-capacity, cones, count-corners, count-cubes-up-to-10, count-cubes-up-to-5, count-dots-up-to-10, count-dots-up-to-3, count-dots-up-to-5, count-on-ten-frames-up-to-10, count-on-ten-frames-up-to-3, count-on-ten-frames-up-to-5, count-pennies, count-pictures-up-to-10, count-pictures-up-to-3, count-pictures-up-to-5, count-scattered-shapes-up-to-10, count-scattered-shapes-up-to-5, count-shapes-in-rings-up-to-10, count-shapes-in-rows-up-to-10, count-shapes-in-rows-up-to-5, count-shapes-up-to-3, count-sides, count-sides-and-corners, cubes, cylinders, different, dimes-and-quarters, fewer, fewer-and-more-compare-by-counting, fewer-and-more-compare-by-matching, fewer-and-more-compare-in-a-mixed-group, fewer-more-and-same, flat-and-solid-shapes, holds-more-or-less, identify-shapes-traced-from-solids, inside-and-outside, light-and-heavy, more, name-the-shape, name-the-solid-shape, one-less-with-pictures-up-to-10, one-less-with-pictures-up-to-5, one-more-with-pictures-up-to-10, one-more-with-pictures-up-to-5, ordinal-numbers-up-to-fifth, ordinal-numbers-up-to-tenth, pennies-and-nickels, pennies-nickels-dimes-and-quarters, rectangles, represent-numbers-up-to-10, represent-numbers-up-to-20, represent-numbers-with-pictures-up-to-3, represent-numbers-with-pictures-up-to-5, represent-numbers-with-shapes-up-to-3, represent-numbers-with-shapes-up-to-5, select-solid-shapes, shapes-of-everyday-objects, spheres, squares, subtraction-sentences-up-to-10-which-model-matches, subtraction-sentences-up-to-5-which-model-matches, tall-and-short, tally-marks-up-to-10, triangles, what-comes-next, wide-and-narrow
	precalculus	determine-continuity-on-an-interval-using-graphs, determine-continuity-using-graphs, determine-one-sided-continuity-using-graphs, find-limits-at-vertical-asymptotes-using-graphs, identify-graphs-of-continuous-functions, outliers-in-scatter-plots, solve-a-triangle

Table 16: Full skill summary(part 3), including math skills for grade 1-8 and pre-k, kindergarten and pre-calculus.

Subject: Engineer
Skill: evaluate-tests-of-engineering-design-solutions
Description: Which of the following could Eliana's test show?Eliana was taking part in her school's engineering competition. To win the competition, she needed to build the popsicle-stick bridge that would hold the most weight. She could use only 200 popsicle sticks. She had two different design ideas. She had to pick one of the designs to use in the competition. To test which design was strongest, Eliana built two prototypes, each with 200 popsicle sticks. She then added 1 kg weights to each prototype until one of them broke. The passage below describes how the engineering-design process was used to test a solution to a problem. Read the passage. Then answer the question below.

Picture: 

Choices: [how much weight a bridge built with 300 popsicle sticks could hold, which design could hold more weight,]
Answer index: 1

Subject: Engineer
Skill: identify-control-and-experimental-groups
Description: In this experiment, which were part of a control group?The passage below describes an experiment. Madelyn has a bubble machine and wants to know how to make the bubbles last longer. She read that bubbles burst when the liquid that makes up the bubbles evaporates. Madelyn knew that when liquids are warmer, they evaporate faster. So, she wondered if she could make her bubbles last longer by cooling the bubble solution. Madelyn cooled six bottles of bubble solution to 30°F below room temperature. She left another six bottles of bubble solution at room temperature. Then, she measured how long bubbles made from the solution in each bottle lasted.

Picture: 

Choices: [the bottles that were cooled down, the bottles that were at room temperature,]
Answer index: 1

Subject: Engineer
Skill: identify-independent-and-dependent-variables
Description: Hint: An independent variable is a variable whose effect you are investigating. A dependent variable is a variable that you measure.Which of the following was a dependent variable in this experiment? The passage below describes an experiment. Read the passage and think about the variables that are described. Giardia is a microscopic parasite that lives in water and can infect humans. Dr. Roth designed a drinking straw that contained a filter to remove Giardia from water. Dr. Roth wanted to know if a longer filtering straw would remove more Giardia. Dr. Roth made six filtering straws: three that were five inches long and three that were ten inches long. She prepared six one-liter batches of water, each containing 10,000 Giardia. Then, Dr. Roth passed one batch of water through each straw. After each batch passed through the straw, she used a microscope to count the number of Giardia that remained in a small sample of the water.

Picture: 

Choices: [the number of Giardia that remained in the water, the length of the filtering straw,]
Answer index: 0

Subject: Engineer
Skill: identify-laboratory-tools
Description: Select the round-bottom flask.
Picture: None

Choices: 

Answer index: 1

Subject: Engineer
Skill: identify-the-experimental-question
Description: Identify the question that Jeffrey's experiment can best answer.The passage below describes an experiment. Read the passage and then follow the instructions below. Jeffrey mixed bacteria into a nutrient-rich liquid where the bacteria could grow. He poured four ounces of the mixture into each of ten glass flasks. In five of the ten flasks, he also added one teaspoon of cinnamon. He allowed the bacteria in the flasks to grow overnight in a 37°C room. Then, Jeffrey used a microscope to count the number of bacteria in a small sample from each flask. He compared the amount of bacteria in the liquid with cinnamon to the amount of bacteria in the liquid without cinnamon.

Picture: 

Choices: [Do more bacteria grow in liquid with cinnamon than in liquid without cinnamon?, Does temperature affect how much bacteria can grow in liquid?,]
Answer index: 0

Subject: Engineer
Skill: laboratory-safety-equipment
Description: Select the apron.
Picture: None

Choices: 

Answer index: 3

Subject: Math
Skill: absolute-value-and-integers-word-problems
Description: Debbie likes watching the show Engineering Marvels. In last night's episode, the engineering team visited a tall skyscraper and a deep mine. A banner at the bottom of the screen showed the elevation of each location the team visited.Which location is closer to sea level?

Picture: 

Choices: [bottom of the mine, top of the skyscraper,]
Answer index: 0

Subject: Math
Skill: acute-obtuse-and-right-triangles
Description: What kind of triangle is this?

Picture: 

Choices: [obtuse, right, acute,]
Answer index: 0

Subject: Math
Skill: acute-right-obtuse-and-straight-angles
Description: Is this angle acute, right, obtuse, or straight?

Picture: 

Choices: [straight, obtuse, acute, right,]
Answer index: 1

Subject: Math
Skill: addition-sentences-up-to-10-what-does-the-model-show
Description: Which addition sentence does the picture show?

Picture:

Choices: [4+3=7, 5+2=7,]
Answer index: 0

Subject: Math
Skill: addition-sentences-up-to-10-which-model-matches
Description: Which shows 8+1=9?
Picture: None

Choices: 

Answer index: 0

Subject: Math
Skill: addition-sentences-up-to-5-what-does-the-model-show
Description: Which addition sentence does the picture show?

Picture:

Choices: [4+1=5, 3+1=4,]
Answer index: 0

Subject: Math
Skill: addition-sentences-up-to-5-which-model-matches
Description: Which shows 2+1=3?
Picture: None

Choices: 

Answer index: 0

Subject: Math
Skill: addition-sentences-using-number-lines-sums-up-to-20
Description: Which addition sentence does this model show?

Picture: 

Choices: [3+3=6, 5+6=11, 5+4=9, 5+3=8,]
Answer index: 3

Subject: Math
Skill: adjust-a-budget
Description: This table shows Angie's February budget.What could Angie do to balance her budget?

Picture: 

Choices: [spend only $60 at the mall, and teach a baking class for $30, decorate more custom cookies to earn another $35, and spend only $60 at the mall, spend $20 more on the baking kit, and teach a baking class for $30, decorate more custom cookies to earn another $35, and spend $20 more on the baking kit,]
Answer index: 1

Subject: Math
Skill: am-or-pm
Description: Farmer Keenan is getting up to go milk his cows. It is just before sunrise. His watch shows:What time is it?

Picture: 

Choices: [4:30 P.M., 4:30 A.M.,]
Answer index: 1

Subject: Math
Skill: angle-angle-criterion-for-similar-triangles
Description: FGH and JKL are shown below.Which statement is true?

Picture: 

Choices: [FGH is similar to JKL., FGH is not similar to JKL., There is not enough information to determine whether the triangles are similar.,]
Answer index: 2

Subject: Math
Skill: angles-as-fractions-of-a-circle
Description: What fraction of the circle does this angle cut out?

Picture: 

Choices: [1/4, 3/4, 1 whole, 1/2,]
Answer index: 0

Table 17: Question examples for each skill (part 1).

Subject: Math Skill: count-money-pennies-and-nickels Description: How much money is there? Picture:  Choices: [6¢, 11¢, 16¢,] Answer index: 1	Subject: Math Skill: count-money-pennies-through-dimes Description: How much money is there? Picture:  Choices: [18¢, 16¢, 19¢,] Answer index: 0
Subject: Math Skill: count-on-ten-frames-up-to-10 Description: How many dots are on the frame? Picture:  Choices: [4, 9, 2, 8, 1, 7, 5, 10, 6, 3,] Answer index: 1	Subject: Math Skill: count-on-ten-frames-up-to-3 Description: How many dots are on the frame? Picture:  Choices: [3, 2, 1,] Answer index: 1
Subject: Math Skill: count-on-ten-frames-up-to-5 Description: How many squares are on the frame? Picture: Choices: [4, 5, 1, 2, 3,] Answer index: 1	Subject: Math Skill: count-pennies Description: How much money is there? Picture:  Choices: [7¢, 6¢, 8¢,] Answer index: 2
Subject: Math Skill: count-pictures-up-to-10 Description: How many parrots are there? Picture:  Choices: [1, 4, 3, 6, 9, 5, 7, 10, 8, 2,] Answer index: 4	Subject: Math Skill: count-pictures-up-to-3 Description: How many butterflies are there? Picture: Choices: [2, 1, 3,] Answer index: 2
Subject: Math Skill: count-pictures-up-to-5 Description: How many snowmen are there? Picture: Choices: [2, 5, 1, 4, 3,] Answer index: 4	Subject: Math Skill: count-scattered-shapes-up-to-10 Description: How many shapes are there? Picture:  Choices: [2, 3, 8, 1, 9, 10, 4, 6, 7, 5,] Answer index: 5
Subject: Math Skill: count-scattered-shapes-up-to-5 Description: How many rectangles are there? Picture:  Choices: [2, 4, 5, 3, 1,] Answer index: 2	Subject: Math Skill: count-shapes-in-rings-up-to-10 Description: How many triangles are there? Picture:  Choices: [9, 3, 5, 2, 7, 6, 8, 1, 4, 10,] Answer index: 4
Subject: Math Skill: count-shapes-in-rows-up-to-10 Description: How many hearts are there? Picture:  Choices: [5, 1, 7, 9, 8, 10, 3, 4, 2, 6,] Answer index: 3	Subject: Math Skill: count-shapes-in-rows-up-to-5 Description: How many shapes are there? Picture: Choices: [3, 2, 1, 4, 5,] Answer index: 0
Subject: Math Skill: count-shapes-up-to-3 Description: How many triangles are there? Picture: Choices: [2, 1, 3,] Answer index: 0	Subject: Math Skill: count-sides Description: Which shape has 4 sides? Picture: None Choices:  Answer index: 1
Subject: Math Skill: count-sides-and-corners Description: Which shape has 5 corners? Picture: None Choices:  Answer index: 1	Subject: Math Skill: count-to-100 Description: How many dots are there? Picture:  Choices: [46, 49, 44,] Answer index: 0

Table 18: Question examples for each skill (part 2).

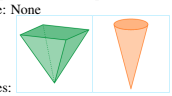
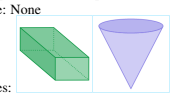
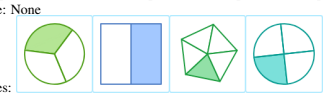
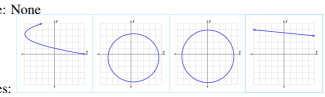
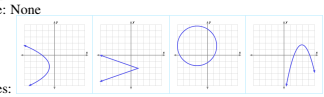
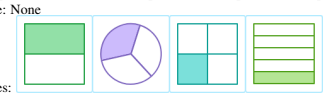
Subject: Math Skill: identify-alternate-interior-and-alternate-exterior-angles Description: line{RT} and line{UW} are parallel lines.Which angles are alternate interior angles? Picture:  Choices: [angle{TSV} and angle{UVS}, angle{TSV} and angle{TSQ}, angle{TSV} and angle{RSV}, angle{TSV} and angle{WVS},] Answer index: 0	Subject: Math Skill: identify-complementary-supplementary-vertical-adjacent-and-congruent-angles Description: Which angle is vertical to angle{3}? Picture:  Choices: [angle{6}, angle{5}, angle{4}, angle{2},] Answer index: 0
Subject: Math Skill: identify-complementary-supplementary-vertical-and-adjacent-angles Description: Which angles are adjacent to each other? Picture:  Choices: [angle{1}angle{3} and angle{7}, angle{1}angle{5} and angle{1}angle{4}, angle{8} and angle{4}, angle{1}angle{0} and angle{4},] Answer index: 1	Subject: Math Skill: identify-congruent-figures Description: Are these shapes congruent? Picture:  Choices: [no, yes,] Answer index: 1
Subject: Math Skill: identify-direct-variation-and-inverse-variation Description: Which equation shows direct variation? Picture: None Choices: $y=\frac{x}{-34}$ $y=\frac{x}{-34}$ Answer index: 1	Subject: Math Skill: identify-equivalent-expressions-using-strip-models Description: This model represents the expression x+x+1+1.Which expression is equivalent to x+x+1+1? Picture: Choices: [4x, 2x+3, 3x+1, 2x+2,] Answer index: 3
Subject: Math Skill: identify-equivalent-fractions-on-number-lines Description: Is 1/2 equivalent to 1/3 ? Picture: Choices: [no, yes,] Answer index: 0	Subject: Math Skill: identify-equivalent-fractions-using-number-lines Description: Is 2/3 equivalent to 4/6 ? Picture: Choices: [yes, no,] Answer index: 0
Subject: Math Skill: identify-equivalent-linear-expressions-using-algebra-tiles Description: These tiles represent the expression 3x+5x.Which expression is equivalent to 3x+5x? Picture:  Choices: [x+8, 2x, 8x, 8x+2,] Answer index: 2	Subject: Math Skill: identify-faces-of-three-dimensional-figures Description: Which shape has a circle as a face? Picture: None Choices:  Answer index: 1
Subject: Math Skill: identify-faces-of-three-dimensional-shapes Description: Which shape has a circle as a face? Picture: None Choices:  Answer index: 1	Subject: Math Skill: identify-fourths Description: Look at the colored part of each shape. Which shape shows one-fourth? Picture: None Choices:  Answer index: 3
Subject: Math Skill: identify-functions Description: Look at this graph:Is this relation a function? Picture:  Choices: [yes, no,] Answer index: 1	Subject: Math Skill: identify-functions-graphs Description: Which of these relations is a function? Picture: None Choices:  Answer index: 3
Subject: Math Skill: identify-functions-vertical-line-test Description: Which of these relations is a function? Picture: None Choices:  Answer index: 3	Subject: Math Skill: identify-graphs-of-continuous-functions Description: Is the function f(x) continuous? Picture:  Choices: [yes, no,] Answer index: 0
Subject: Math Skill: identify-halves Description: Look at the colored part of each shape. Which shape shows one-half? Picture: None Choices:  Answer index: 0	Subject: Math Skill: identify-halves-and-fourths Description: Which figure shows fourths? Picture: None Choices:  Answer index: 1

Table 19: Question examples for each skill (part 3).

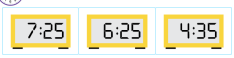
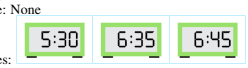
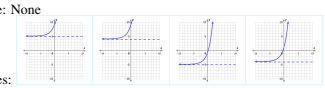
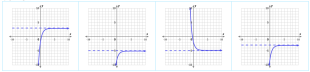
Subject: Math Skill: lines-line-segments-and-rays Description: What is this? Picture:  Choices: [line, line segment, ray,] Answer index: 1	Subject: Math Skill: make-predictions-with-scatter-plots Description: Based on the scatter plot below, which is a better prediction for x when y = 46? Picture:  Choices: [50, 98,] Answer index: 0
Subject: Math Skill: match-addition-sentences-and-models-sums-to-10 Description: Which shows 2+2=4? Picture: None Choices:  Answer index: 0	Subject: Math Skill: match-analog-and-digital-clocks Description: Look at the analog clock:Which digital clock shows the same time? Picture:  Choices:  Answer index: 0
Subject: Math Skill: match-analog-clocks-and-times Description: What time does the clock show? Picture:  Choices: [5:00, 4:30,] Answer index: 0	Subject: Math Skill: match-clocks-and-times Description: What time does the clock show? Picture:  Choices: [eight fifty, seven fifty, nine forty,] Answer index: 1
Subject: Math Skill: match-digital-clocks-and-times Description: Which clock shows six thirty-five? Picture: None Choices:  Answer index: 1	Subject: Math Skill: match-exponential-functions-and-graphs Description: formula_desc.png Picture: None Choices:  Answer index: 0
Subject: Math Skill: match-exponential-functions-and-graphs-ii Description: formula_desc.png Picture: None Choices:  Answer index: 2	Subject: Math Skill: match-fractions-to-models-halves-thirds-and-fourths Description: Look at the colored part of each shape. Which shape shows one-fourth? Picture: None Choices:  Answer index: 2
Subject: Math Skill: match-mixed-numbers-to-models Description: Which mixed number is shown? Picture: Choices: [3 3/8, 4 2/8, 3 2/8, 3 5/8,] Answer index: 0	Subject: Math Skill: mean-find-the-missing-number Description: Susan has the following data:If the mean is 25, which number could r be? Picture: Choices: [29, 38,] Answer index: 0
Subject: Math Skill: mean-median-mode-and-range-find-the-missing-number Description: Jayla has the following data:If the mean is 14, which number could s be? Picture: Choices: [11, 3,] Answer index: 0	Subject: Math Skill: measure-angles-with-a-protractor Description: Is this angle acute, right, or obtuse? Picture:  Choices: [right, obtuse, acute,] Answer index: 2
Subject: Math Skill: median-find-the-missing-number Description: Danny has the following data:If the median is 97, which number could c be? Picture: Choices: [98, 47,] Answer index: 0	Subject: Math Skill: model-and-solve-equations-using-algebra-tiles Description: Which equation does this set of algebra tiles represent? Picture: Choices: [– 4x–1= – 9, – 8x–1= – 9, 8x–1= – 9, – x–1= – 10,] Answer index: 3
Subject: Math Skill: model-and-solve-linear-equations-using-algebra-tiles Description: Which equation does this set of algebra tiles represent? Picture: Choices: [3x=27, 3x=24, 2x=26, 2x=24,] Answer index: 1	Subject: Math Skill: more Description: Which group has more? Picture: None Choices:  Answer index: 0

Table 20: Question examples for each skill (part 4).

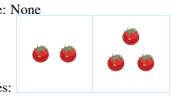
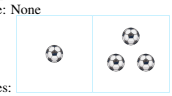
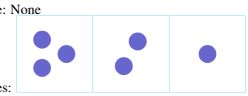
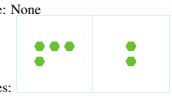
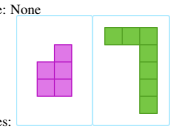
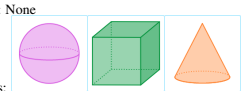
Subject: Math Skill: reflection-rotation-and-translation Description: How has this figure been transformed?It has been... Picture:  Choices: [translated, reflected, rotated,] Answer index: 1	Subject: Math Skill: regular-and-irregular-polygons Description: Is this shape a regular polygon? Picture: Choices: [yes, no,] Answer index: 1
Subject: Math Skill: represent-numbers-up-to-10 Description: Which group has 6 triangles? Picture: None Choices:  Answer index: 0	Subject: Math Skill: represent-numbers-up-to-20 Description: Which picture shows 8 dots? Picture: None Choices: Answer index: 0
Subject: Math Skill: represent-numbers-with-pictures-up-to-3 Description: Which shows 2? Picture: None Choices:  Answer index: 0	Subject: Math Skill: represent-numbers-with-pictures-up-to-5 Description: Which shows 1? Picture: None Choices:  Answer index: 0
Subject: Math Skill: represent-numbers-with-shapes-up-to-3 Description: Which group has 3 circles? Picture: None Choices:  Answer index: 0	Subject: Math Skill: represent-numbers-with-shapes-up-to-5 Description: Which group has 4 hexagons? Picture: None Choices:  Answer index: 0
Subject: Math Skill: rhombuses Description: Which shape is a rhombus? Picture: None Choices:  Answer index: 0	Subject: Math Skill: rotational-symmetry Description: Does this picture have rotational symmetry? Picture:  Choices: [no, yes,] Answer index: 0
Subject: Math Skill: rotational-symmetry-amount-of-rotation Description: This image has rotational symmetry. What is the smallest fraction of a full turn you need to rotate the image for it to look the same? Picture:  Choices: [1 2 of a full turn, 1 6 of a full turn, 1 4 of a full turn, 1 3 of a full turn,] Answer index: 0	Subject: Math Skill: scalene-isosceles-and-equilateral-triangles Description: Is this triangle scalene? Picture:  Choices: [yes, no,] Answer index: 1
Subject: Math Skill: select-figures-with-a-given-area Description: Which shape has an area of 7 square units? The shapes are made of unit squares. Picture: None Choices:  Answer index: 1	Subject: Math Skill: select-fractions-equivalent-to-whole-numbers-using-models Description: Count the equal parts. What fraction does this picture show? Picture: Choices: [2/4, 4/8, 8/2, 2/8,] Answer index: 2
Subject: Math Skill: select-solid-shapes Description: Which shape is a cone? Picture: None Choices:  Answer index: 2	Subject: Math Skill: select-three-dimensional-shapes Description: Which shape is a rectangular prism? Picture: None Choices:  Answer index: 2
Subject: Math Skill: select-two-dimensional-shapes Description: Which shape is a hexagon? Picture: None Choices:  Answer index: 2	Subject: Math Skill: shapes-of-everyday-objects Description: Which is shaped like a cylinder? Picture: None Choices:  Answer index: 1

Table 21: Question examples for each skill (part 5).

Subject: Science
Skill: animal-adaptations-feet-and-limbs
Description: Star-nosed moles are found in many parts of North America. They live in burrows. The moles eat earthworms and nuts, which they find in the soil. The feet of the star-nosed mole are adapted for digging.Which animal's feet are also adapted for digging?
Picture: 
Choices:
Answer index: 0

Subject: Science
Skill: animal-adaptations-skins-and-body-coverings
Description: Emerald tree boas live in the forests of South America. The tree boa is adapted to be camouflaged among green leaves.Which animal is also adapted to be camouflaged among green leaves?
Picture: 
Choices:
Answer index: 1

Subject: Science
Skill: classify-elementary-substances-and-compounds-using-models
Description: Complete the statement.Nitrogen is The model below represents a molecule of nitrogen. Nitrogen gas makes up nearly 80Picture:
Choices: [an elementary substance, a compound,]
Answer index: 0

Subject: Science
Skill: classify-fruits-and-vegetables-as-plant-parts
Description: People use lettuce plants for food. We usually eat the part of this plant that makes most of the food for the plant.Hint: A plant's leaves make food. A plant's seeds can grow into a new plant. Which part of the lettuce plant do we usually eat?
Picture: 
Choices: [the leaves, the seeds,]
Answer index: 0

Subject: Science
Skill: classify-matter-as-solid-liquid-or-gas
Description: Is the water from a faucet a solid, a liquid, or a gas?
Picture: 
Choices: [a solid, a liquid, a gas,]
Answer index: 1

Subject: Science
Skill: classify-matter-as-solid-or-liquid
Description: Is a coin a solid or a liquid?
Picture: 
Choices: [a liquid, a solid,]
Answer index: 1

Subject: Science
Skill: classify-rocks-and-minerals-by-color-and-shape
Description: Select the black mineral.
Picture: None
Choices: 
Answer index: 0

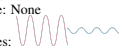
Subject: Science
Skill: classify-rocks-and-minerals-by-color-shape-and-texture
Description: Select the brown rock.
Picture: None
Choices: 
Answer index: 1

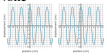
Subject: Science
Skill: classify-rocks-as-igneous-sedimentary-or-metamorphic
Description: Diorite is a type of rock. When melted rock cools below the earth's surface, it can form diorite. Diorite is usually made of large mineral grains.What type of rock is diorite?
Picture: 
Choices: [sedimentary, igneous,]
Answer index: 1

Subject: Science
Skill: classify-symbiotic-relationships
Description: Which type of relationship is formed when an Alcon blue caterpillar lives in a Myrmica ant nest?Read the passage. Then answer the question. Alcon blue butterflies spend the first part of their lives as caterpillars that live with Myrmica ants. When a caterpillar lives with the ants, it mimics, or pretends to be, an ant. The caterpillar can mimic the ants by copying their smell. The caterpillar can also make noises that make it sound like a queen ant. Queen ants receive more food and better protection than any other ants in the nest. So, when the caterpillar mimics an ant, the ants feed and protect the caterpillar instead of other ants in the nest.
Picture: 
Choices: [mutualistic, commensal, parasitic,]
Answer index: 2

Subject: Science
Skill: compare-ages-of-fossils-in-a-rock-sequence
Description: This diagram shows fossils in an undisturbed sedimentary rock sequence.Which of the following fossils is older? Select the more likely answer.
Picture: 

Choices:
Answer index: 1

Subject: Science
Skill: compare-amplitudes-and-wavelengths-of-waves
Description: Select the wave with the greater amplitude.
Picture: None
Choices: 
Answer index: 0

Subject: Science
Skill: compare-amplitudes-wavelengths-and-frequencies-of-waves
Description: Select the graph of the wave with the greater amplitude.The graphs below describe two waves. The waves are traveling at the same speed.
Picture: None
Choices: 
Answer index: 1

Subject: Science
Skill: compare-ancient-and-modern-organisms-use-observations-to-support-a-hypothesis
Description: Which statement supports the following hypothesis?The American lobster and Homarus hakelensis have similar adaptations to survive underwater. The American lobster and Homarus hakelensis have similar adaptations to survive underwater.
Picture: 
Choices: [Homarus hakelensis used its claws to find food underwater, The American lobster uses its claws to find food underwater, The American lobster uses its claws to find food underwater,]
Answer index: 2

Subject: Science
Skill: compare-concentrations-of-solutions
Description: The diagram below is a model of two solutions. Each green ball represents one particle of solute.Which solution has a higher concentration of green particles?
Picture: 
Choices: [Solution B, Solution A, neither; their concentrations are the same,]
Answer index: 0

Subject: Science
Skill: compare-magnitudes-of-magnetic-forces
Description: Think about the magnetic force between the magnets in each pair. Which of the following statements is true?The images below show two pairs of magnets. The magnets in different pairs do not affect each other. All the magnets shown are made of the same material, but some of them are different shapes.
Picture: 
Choices: [The magnitude of the magnetic force is smaller in Pair 2., The magnitude of the magnetic force is the same in both pairs., The magnitude of the magnetic force is smaller in Pair 1.,]
Answer index: 1

Subject: Science
Skill: compare-properties-of-materials
Description: Which is harder?
Picture: None
Choices: 

Answer index: 1

Subject: Science
Skill: compare-properties-of-objects
Description: Which property do these four objects have in common?Select the best answer.
Picture:
Choices: [sticky, sour, soft,]
Answer index: 2

Table 22: Question examples for each skill (part 6).

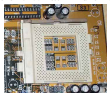
Subject: Science Skill: use-data-to-describe-world-climates Description: Which statement best describes the climate of Santa Fe?Hint: Summers in the Northern Hemisphere occur in June, July, and August. Winters in the Northern Hemisphere occur in December, January, and February. Santa Fe, New Mexico, is a city in the United States. It has a semiarid climate. Picture:  Choices: [Winters have much lower temperatures than summers., Winters have less precipitation than summers.,] Answer index: 0	Subject: Science Skill: use-evidence-to-classify-animals Description: Placental mammals are a group of animals with similar traits. The following traits can be used to identify placental mammals:They give birth to live offspring. They have fur or hair. Observe the animals and read the descriptions. Select the one animal that has all of the placental mammal traits listed above. Sea otters have very thick fur. Their fur helps keep them warm in cold water. Female sea otters give birth to live offspring in the water. Red salamanders do not have lungs! They can breathe through their moist, smooth skin. Adult red salamanders live near rivers or ponds. They lay eggs with no shells under rocks or logs. The baby red salamanders live underwater. Picture: None Choices: Answer index: 0
Subject: Science Skill: use-evidence-to-classify-mammals-birds-fish-reptiles-and-amphibians Description: Fish are a group of animals with similar traits. The following traits can be used to identify fish:They have fins, not limbs. They make eggs with no shells. Observe the animals and read the descriptions. Select the one animal that has all of the fish traits listed above. Thresher sharks hatch from eggs with no shells. They have a long tail and fins. They can use their tail to hit and stun their prey. Thresher sharks live in salt water. Greater flameback woodpeckers have feathers and two wings. They use their strong beaks to make holes in trees. The woodpeckers use these holes as nests for their eggs, which have white shells. Picture: None Choices: Answer index: 0	Subject: Science Skill: use-punnett-squares-to-calculate-probabilities-of-offspring-types Description: In a group of dachshund dogs, some individuals have rough fur and others have soft fur. In this group, the gene for the fur texture trait has two alleles. The allele for rough fur (F) is dominant over the allele for soft fur (f).This Punnett square shows a cross between two dachshund dogs. What is the probability that a dachshund dog produced by this cross will be homozygous dominant for the fur texture gene? Picture:  Choices: [3/4, 0/4, 2/4, 1/4, 4/4,] Answer index: 2
Subject: Science Skill: use-punnett-squares-to-calculate-ratios-of-offspring-types Description: In a group of Syrian hamsters, some individuals have short fur and others have long fur. In this group, the gene for the fur length trait has two alleles. The allele for short fur (F) is dominant over the allele for long fur (f).This Punnett square shows a cross between two Syrian hamsters. What is the expected ratio of offspring with long fur to offspring with short fur? Choose the most likely ratio. Picture:  Choices: [3:1, 1:3, 0:4, 2:2, 4:0,] Answer index: 2	Subject: Science Skill: use-scientific-names-to-classify-organisms Description: This organism is a mantled howler. Its scientific name is Alouatta palliata.Select the organism in the same species as the mantled howler. Picture:  Choices: Answer index: 1
Subject: Science Skill: weather-and-climate-around-the-world Description: Does this passage describe the weather or the climate?Hint: Weather is what the atmosphere is like at a certain place and time. Climate is the pattern of weather in a certain place. A cloud forest is a mountain ecosystem that is home to a wide variety of species. The skies were mostly clear last week over this cloud forest, which is in Ecuador. Picture:  Choices: [weather, climate,] Answer index: 0	Subject: Technology Skill: cables Description: What kind of computer related plug or port do you see here? Picture:  Choices: [USB type-A port, HDMI plug, VGA port, USB type-C plug,] Answer index: 3
Subject: Technology Skill: font Description: Identify this font type Picture: ActionQuiz Choices: [Times Ancient Roman, Matisse ITC, Human521 BT, Bookman Old Style,] Answer index: 2	Subject: Technology Skill: icons Description: This is a(n old) logo of which famous app or program? Picture:  Choices: [Acrobat Reader, Google Pay, Microsoft Office PowerPoint, GoFundMe,] Answer index: 2
Subject: Technology Skill: logo Description: This is (part of) a (former) logo of which computer related brand? Picture:  Choices: [Imation, Cisco, Nintendo, Verbatim,] Answer index: 3	Subject: Technology Skill: others Description: What is the function of this key? Picture:  Choices: [Copy, Undo, Delete, Paste,] Answer index: 1
Subject: Technology Skill: parts Description: What kind of computer component do you see here? Picture:  Choices: [Power Supply Unit, Computer Fan, CPU Socket, Molex Connector,] Answer index: 2	Subject: Technology Skill: peripherals Description: What kind of computer peripheral do you see here? Picture:  Choices: [Floppy Disk, DVD Spindle, Tablet, Bluetooth Headset,] Answer index: 3
Subject: Technology Skill: photo Description: What type of video game console do you see here? Picture:  Choices: [Nintendo Wii, Microsoft Xbox One, Microsoft Xbox, Mattel Intellivision,] Answer index: 0	Subject: Technology Skill: web Description: What meaning or function is usually associated with this web interface symbol? Picture:  Choices: [Storage for deleted files, Computer games, Reload/Refresh, Send e-mail,] Answer index: 2

Table 23: Question examples for each skill (part 7).

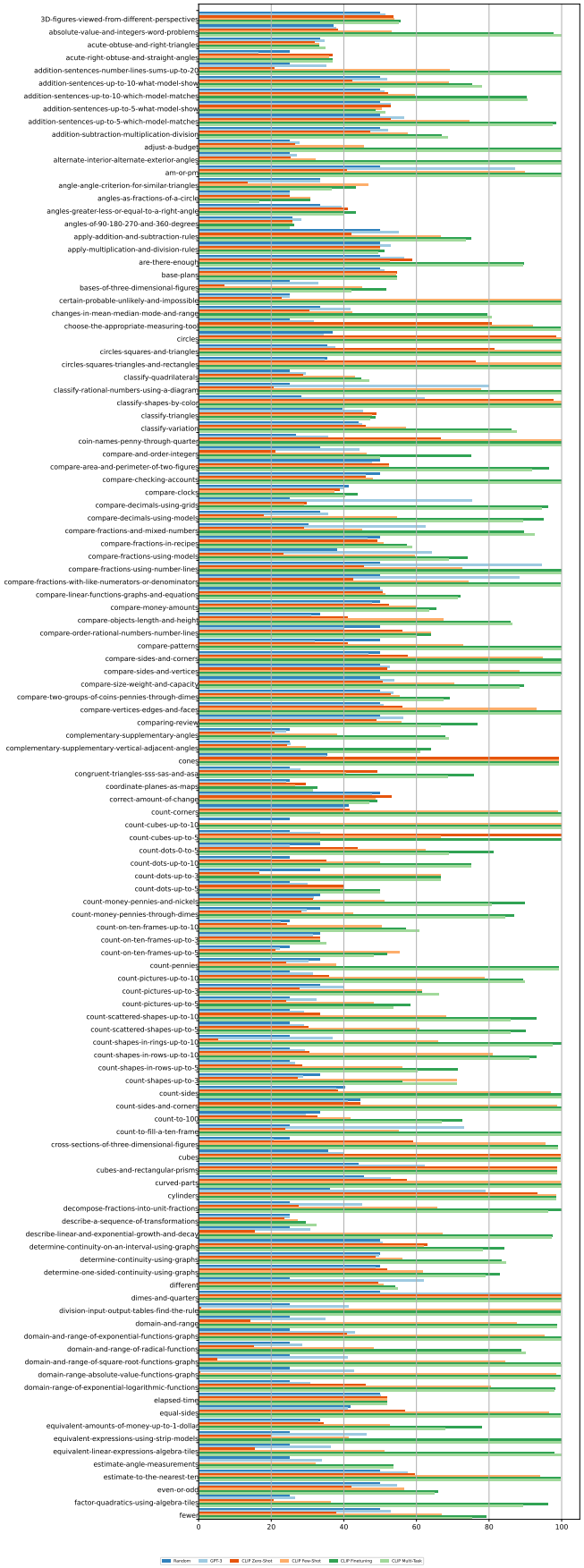


Figure 21: Accuracy per skill on math (part 1).

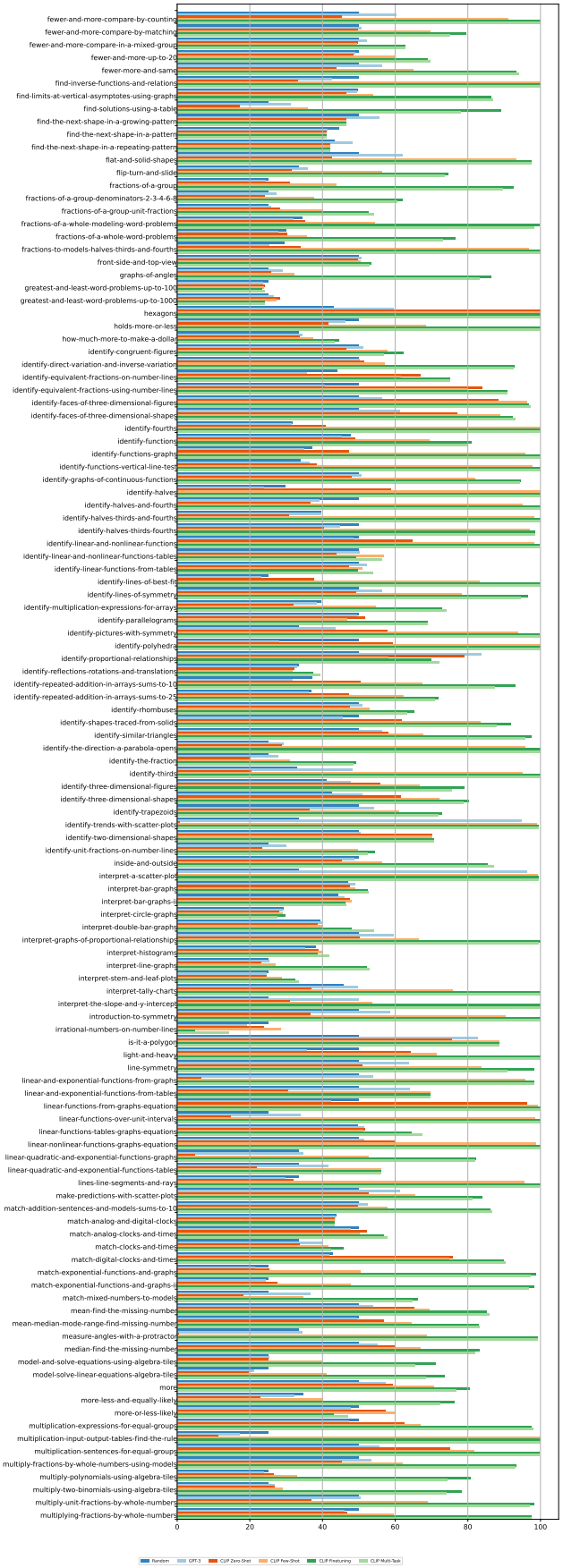


Figure 22: Accuracy per skill on math (part 2).

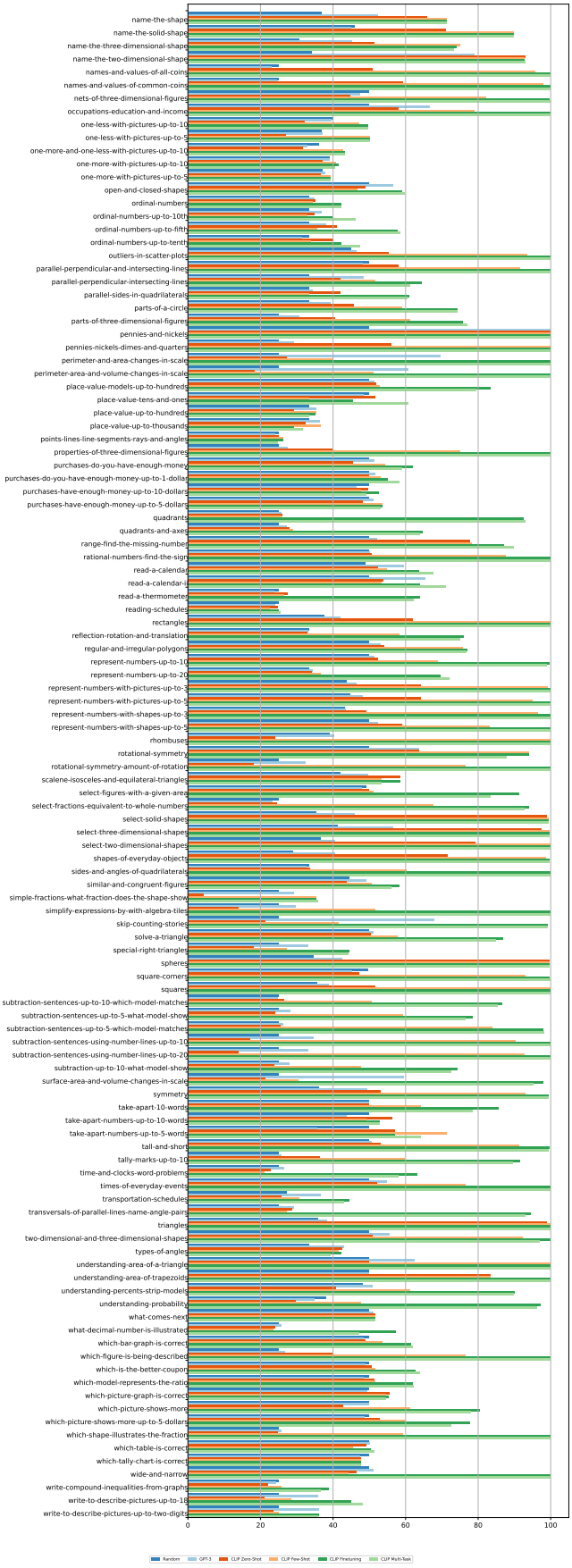


Figure 23: Accuracy per skill on math (part 3).

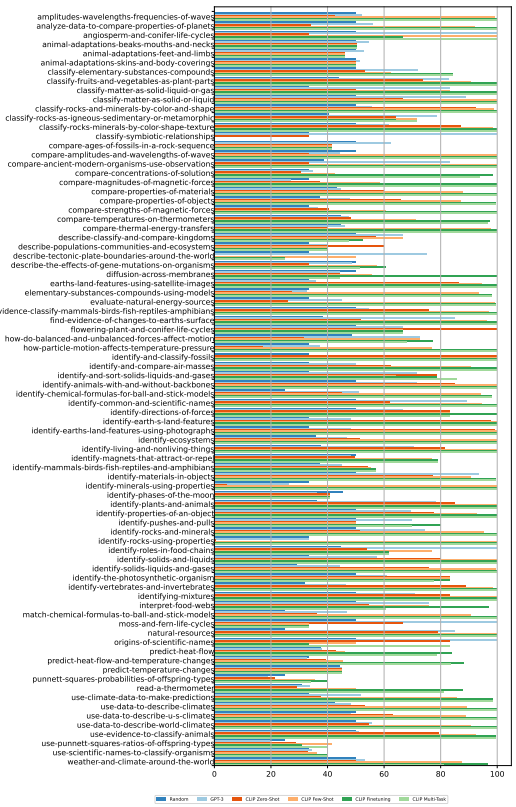


Figure 24: Accuracy per skill on science.

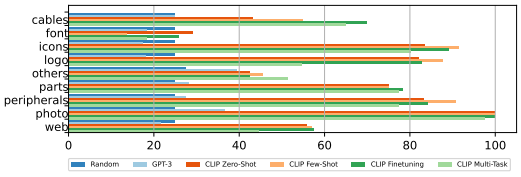


Figure 25: Accuracy per skill on technology.

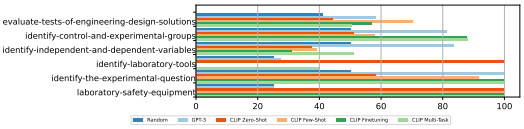


Figure 26: Accuracy per skill on engineering.