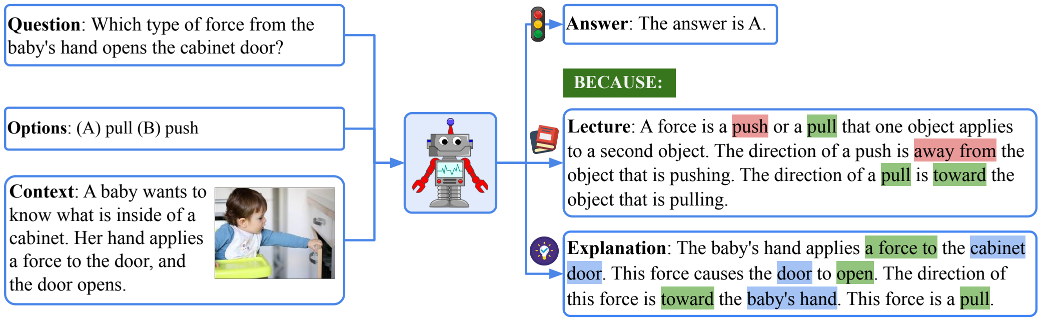
Learn to Explain: Multimodal Reasoning via Thought Chains for Science Question Answering

Anonymous submission ID: 4596

Abstract

When answering a question, humans utilize the information available across different modalities to synthesize a consistent and complete *chain of thought* (CoT). This process is normally a black box in the case of deep learning models like large-scale language models. Recently, science question benchmarks have been used to diagnose the multi-hop reasoning ability and interpretability of an AI system. However, existing datasets fail to provide annotations for the answers, or are restricted to the textual-only modality, small scales, and limited domain diversity. To this end, we present Science Question Answering ($\mathbb{SQA}$), a new benchmark that consists of $\sim$21k multimodal multiple choice questions with a diverse set of science topics and annotations of their answers with corresponding lectures and explanations. We further design language models to learn to generate lectures and explanations as the *chain of thought* (CoT) to mimic the multi-hop reasoning process when answering $\mathbb{SQA}$ questions. $\mathbb{SQA}$ demonstrates the utility of CoT in language models, as CoT improves the question answering performance by 1.20% in few-shot GPT-3 and 3.99% in fine-tuned UnifiedQA. We also explore the upper bound for models to leverage explanations by feeding those in the input; we observe that it improves the few-shot performance of GPT-3 by 18.96%. Our analysis further shows that language models, similar to humans, benefit from explanations to learn from fewer data and achieve the same performance with just 40% of the data.



Question: Which type of force from the baby's hand opens the cabinet door?

Options: (A) pull (B) push

Context: A baby wants to know what is inside of a cabinet. Her hand applies a force to the door, and the door opens.

Answer: The answer is A.

Lecture: A force is a push or a pull that one object applies to a second object. The direction of a push is away from the object that is pushing. The direction of a pull is toward the object that is pulling.

Explanation: The baby's hand applies a force to the cabinet door. This force causes the door to open. The direction of this force is toward the baby's hand. This force is a pull.

Figure 1: We construct the $\mathbb{SQA}$ dataset where a data example consists of multimodal question answering information and the grounded lecture and explanation. We study if models can generate a reasonable explanation to reveal the chain-of-thought reasoning when answering an $\mathbb{SQA}$ question.

1 Introduction

A long-standing goal of AI systems is to act reliably and learn complex tasks efficiently like human beings. In the process of reliable decision making, humans follow an explicit *chain-of-thought* (CoT) reasoning process that is typically expressed as an explanation. However, machine learning models are trained mostly using a large number of input-output examples to perform a specific task. These black-box models only generate the final decision without reliably revealing the underlying reasoning

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

process. Not surprisingly, it is unclear if they understand the task and can generalize even though they perform well on the benchmark. On the other hand, humans are able to learn from instructions or explanations from past experience and generalize them to novel and unseen problems. This helps them learn more quickly with fewer data. In this work, we explore if machines can be endowed with such reasoning abilities in the context of science-based question answering.

Recently, science problem solving benchmarks [17] have been used to diagnose the multi-hop reasoning ability and interpretability of AI systems. To answer science questions, a model needs to not only understand multimodal contents but also extract external knowledge to arrive at the correct answer. Since these tasks require domain-specific knowledge and explicit multi-hop reasoning, a model would be not interpretable if it fails to provide explanations to reveal the reasoning process. However, current science question datasets [17, 16, 46] mostly lack annotated explanations for the answers. To address this issue, other science datasets annotate the explanations, but they are restricted to the textual only modality and limited to small data scales [12, 7, 33] or a small set of topics [19, 13]. Therefore, we collect Science Question Answering (𝕊ℚ𝔸), a large-scale multi-choice dataset that contains multimodal science questions with explanations and features rich domain diversity.

𝕊ℚ𝔸 is collected from elementary and high school science curricula, and contains 21,208 examples along with lectures and explanations. Different from existing datasets [16, 17, 46], 𝕊ℚ𝔸 has richer domain diversity from three different subjects: natural science, social science, and language science. A typical 𝕊ℚ𝔸 example consists of a question, multiple choices, visual and textual contexts, a correct answer, as well as a lecture and an explanation. The lecture and explanation provide general external knowledge and specific reasons, respectively, for arriving at the correct answer.

Consider the thoughts one person might have when answering the question in Figure 1. One first recalls the knowledge regarding the definition of a force learned from textbooks: "*A force is a push or a pull that ... The direction of a* ***push*** *is ... The direction of a* ***pull*** *is ...*", then forms a line of reasoning: "*The baby's* ***hand*** *applies a force to the cabinet* ***door****. → This force causes the* ***door*** *to* ***open****. → The direction of this force is* ***toward*** *the baby's* ***hand****.*", and finally arrives at the correct answer: "*This force is a* ***pull****.*". Following [36], we formulate the 𝕊ℚ𝔸 task to output a natural explanation alongside the predicted answer. In this paper, we train language models to generate lectures and explanations as the *chain of thought* (CoT) to mimic the multi-hop reasoning process to answer 𝕊ℚ𝔸 questions.

Our experiments show that current multimodal methods [49, 1, 20, 9, 24, 31] fail to achieve satisfactory performance on 𝕊ℚ𝔸 and do not generate correct explanations. However, we find that CoT can help large language models not only in the few-shot learning setting but also in the fine-tuning setting. When combined with CoT to generate the lecture and explanation, the fine-tuned UnifiedQA [18] achieves an improvement of 3.99% as opposed to not using CoT in the fine-tuning stage. The few-shot GPT-3 model [5] via chain-of-thought prompting can obtain 75.17% on 𝕊ℚ𝔸 with an improvement of 1.20% compared to the few-shot GPT-3 without CoT. Prompted with CoT, GPT-3 can generate reasonable explanations as evaluated by automated metrics, and promisingly, 65.2% of explanations meet the gold standard of human evaluations. We also investigate the upper bound for models to harness explanations by including them in the input. We find that doing so improves GPT-3's few-shot performance by 18.96%, suggesting that explanations do aid models and are currently underutilized in the CoT framework. Further analysis shows that, like humans, language models benefit from explanations to learn with less data: UnifiedQA with CoT obtains the same results as UnifiedQA without CoT with only 40% of the training data.

To sum up, our contributions are three-fold: (a) To bridge the gap in existing datasets in the scientific domain, we build Science Question Answering (𝕊ℚ𝔸), a new dataset containing 21,208 multimodal science questions with rich domain diversity. To the best of our knowledge, 𝕊ℚ𝔸 is the first large-scale multimodal dataset that annotates lectures and explanations for the answers. (b) We show that CoT benefits large language models in both few-shot and finetuning learning by improving model performance and reliability via generating explanations. (c) We further explore the upper bound of GPT-3 and show that CoT helps language models learn from fewer data.

2 Related Work

Visual question answering. Since the task of visual question answering (VQA) was first proposed in [2], there have been plenty of VQA datasets [50, 52, 22, 10, 14, 11] conducted to facilitate the research work. Although our 𝕊ℚ𝔸 dataset shares some features with VQA, there are several main differences between them. First, 𝕊ℚ𝔸 is more challenging than existing VQA datasets because it contains multimodal contexts and diverse topics in the scientific domain. In addition, most answers are annotated with lectures and explanations, which makes $\mathbb{SQA}$ a suitable dataset for multi-model question answering and multi-hop reasoning for AI systems. Inspired by the recent remarkable performance achieved for VQA [9, 24, 8], in this paper, we further extensively benchmark $\mathbb{SQA}$ with a wide range of attention-based [1, 30, 20, 9] and Transformer-based [28, 24, 25, 8] methods.

Datasets for science problems. Science problem solving is a challenging task that requires an AI system not only to understand the multimodal information from the science curriculum but also to reason about how to answer the domain-specific questions. Current science problem datasets such as AI2D [16], DVQA [15], VLQA [46], and FOODWEDS [23] have contributed to multimodal reasoning in the scientific domain. These datasets, however, lack annotated explanations for the answers to reveal the reasoning steps. Some other datasets annotate the answers in the forms of supporting facts [33, 19], entailment trees [7], explanation graphs [12], reasoning chains [13]. However, these datasets are restricted to the single text modality with small data scales and limited topics. Instead, our $\mathbb{SQA}$ annotates the answers with grounded lectures and explanations. Besides, $\mathbb{SQA}$ features a richer domain diversity across 3 subjects, 26 topics, 127 categories, and 379 skills.

Learning from explanations and few-shot Learning. Explanations help humans understand a task better, and there have been several attempts to show the same for models. For examples, the learning from instruction paradigm [35, 38, 47, 34] where the task level explanation is provided in the form of instruction to improve model performance significantly. An example of learning from explanations in the scientific domain is proposed in [45] where the model interprets demonstrative solutions to solve geometry problems. Recently, there has been a surge of interest in few-shot learning, where language models learn a specific task from a few examples [40, 3]. For instance, [37, 48] find that explanations in the format of the chain of thought can improve the reasoning ability of language models in few-shot learning. In this paper, we show that the chain of thought boosts the performance of large language models like UnifiedQA [18] if the models generate explanations along with the answer in a fine-tuned way. Furthermore, a few-shot GPT-3 model via chain-of-thought prompting is able to improve the reasoning performance on $\mathbb{SQA}$ and generate reasonable explanations.

3 Dataset

We collect $\mathbb{SQA}$, which is a multimodal multiple-choice science question dataset containing 21,208 examples. An example in $\mathbb{SQA}$ is shown in Figure 1. Given the science question and multimodal contexts, the task is to select the correct answer from multiple options. Different from existing datasets [44, 16, 46, 29, 23], $\mathbb{SQA}$ covers diverse topics across three subjects: natural science, social science, and language science. Moreover, most questions are annotated with grounded lectures and detailed explanations. The lecture provides general knowledge that introduces the background information for solving problems of a similar class. The explanation reveals a specific reason for the answer. To effectively answer the questions, a model often needs to be able to understand the multimodal content in the input and extract external knowledge, similar to how humans do. More importantly, the goal of $\mathbb{SQA}$ is to aid development of a reliable model that is capable of generating a coherent chain of thought when arriving at the correct answer to reveal the multi-step reasoning process. For data collection details, see Appendix A.1.

	#Q	#I	AvgQ	MaxQ	Grades	Science subjects	Contexts	Images	Lecture	Explanation
Geometry3K [29]	3,002	2,342	10.1	46	6-12	natural (geometry)	image	diagram	✘	✘
AI2D [16]	4,563	4,903	9.8	64	1-6	natural	image	diagram	✘	✘
FOODWEBS [23]	≈5,000	≈5,00	-	-	8	natural (foodweb only)	image	diagram	✘	✘
ARC [6]	7,787	0	**20.4**	128	3-9	natural	✘	✘	✘	✘
VLQA [46]	9,267	10,209	15.0	-	-	natural	image, text	natural, diagram	✘	✘
TQA [17]	**26,260**	3,455	9.2	57	6-8	natural	image, text	diagram	✔	✘
WorldTree [12]	1,680	0	-	-	3-5	natural	✘	✘	✘	✔
OpenBookQA [33]	5,957	0	10.6	68	1-6	natural	✘	✘	✘	✔
QASC [19]	9,980	0	8.0	25	1-9	natural	✘	✘	✘	✔
SQA (ours)	21,208	**10,332**	12.1	**141**	**1-12**	natural, social, language	image, text	natural, diagram	✔	✔

Table 1: Statistics for the $\mathbb{SQA}$ dataset and comparisons with existing datasets.

3.1 Comparisons with Existing Datasets

Table 1 shows a comparison of $\mathbb{SQA}$ and other science problem datasets. As shown in the table, $\mathbb{SQA}$ is much larger than most other datasets. $\mathbb{SQA}$ also has the largest set of images, spans across

Statistic	Number
Total questions	21,208
Questions with text context	10,220 (48.2%)
Questions with image context	10,332 (48.7%)
* Image of natural format	≈2,960 (14.0%)
* Image of diagram format	≈7,372 (34.8%)
Questions with both contexts	6,532 (30.8%)
Questions with a lecture	17,798 (83.9%)
Questions with a explanation	19,202 (90.5%)
Different questions	9,122
Different lectures	261
Topic classes	26
Category classes	127
Skill classes	379
Average question length	12.11
Average choice length	4.40
Average lecture length	125.06
Average explanation length	47.66

Table 2: Main statistics in 𝕊ℚ𝔸.

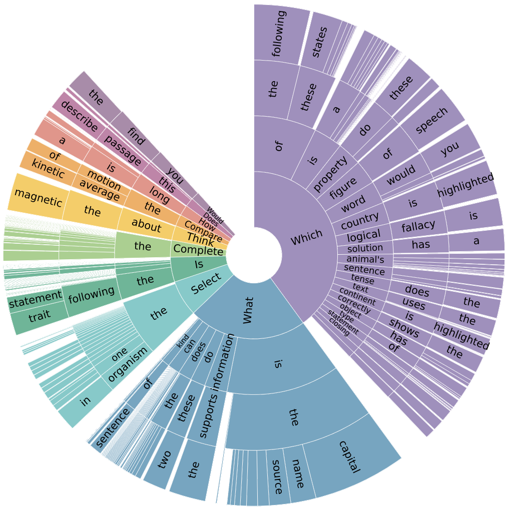
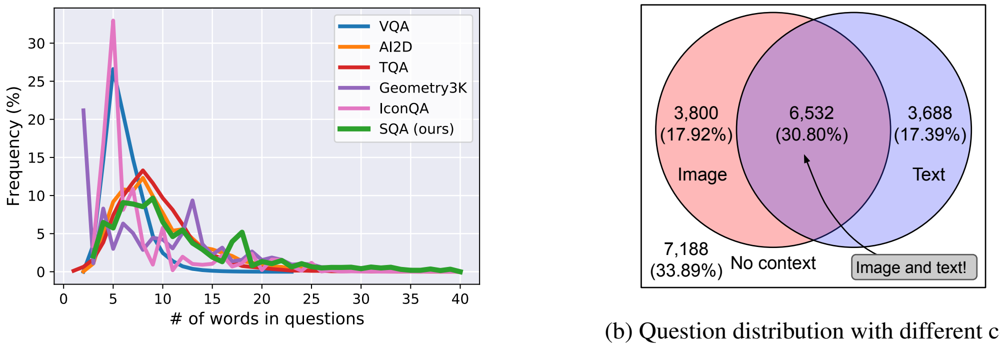


Figure 2: Question distribution in 𝕊ℚ𝔸.

all 12 grades, contains the longest questions, and has the most diverse input sources. As opposed to limiting the subject to only natural science, 𝕊ℚ𝔸 also includes social science and language science, largely adding to the domain diversity of the dataset. Furthermore, most of the questions in 𝕊ℚ𝔸 are annotated with textual lectures (83.9%) and explanations (90.5%), which reveal the reasoning path to the correct answer. To the best of our knowledge, 𝕊ℚ𝔸 is the first large-scale multimodal science question dataset that annotates the answers with detailed lectures and explanations.

3.2 Data Analysis

Key statistics. We randomly split the dataset into training, validation, and test splits with a ratio of 60:20:20. Each split has 12,726, 4,241, and 4,241 examples, respectively. Table 2 shows the main statistics of 𝕊ℚ𝔸. 𝕊ℚ𝔸 has a large set of unique questions, totaling up to 9,122. Out of the 21,208 questions in 𝕊ℚ𝔸, 10,332 (48.7%) have an image context, 10,220 (48.2%) have a text context, and 6,532 (30.8%) have both. 83.9% of the questions are annotated with a lecture, while 91.3% of the questions feature an explanation. The cross-combination of these information sources diversifies the problem scenario: sometimes the model is given a lot of information from multiple sources, while at other times, the only source of information is the question itself. This level of complexity is very common in grade-level science exams.



(a) Question length distribution of VQA and science datasets. 𝕊ℚ𝔸 is distributed more evenly in terms of the number of question words than other datasets.

(b) Question distribution with different context formats. 66.11% of the questions in 𝕊ℚ𝔸 have either an image or text context, while 30.80% of the questions have both.

Figure 3: Question length distribution of different datasets (a) and context distribution in 𝕊ℚ𝔸 (b).

Question analysis. 𝕊ℚ𝔸 has a diverse set of science questions. Figure 2 shows a distribution of the first four words in the question text. A large number of question lengths and formats highlight the diversity of 𝕊ℚ𝔸. The question lengths range from 3 words to 141 words, and the questions in 𝕊ℚ𝔸 have an average length of 12.11 words. The question length distribution is visualized against

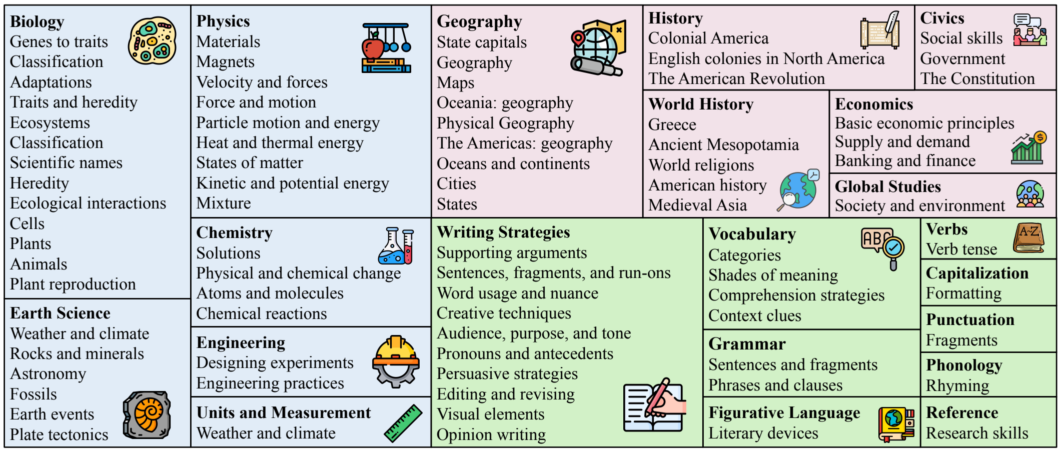


Figure 4: Domain diversity in 𝕊ℚ𝔸. Each color corresponds to one subject: natural science, social science, and language science. For visual clarity, only the most frequent classes are shown.

other VQA datasets in Figure 3 (a). As shown in the diagram, 𝕊ℚ𝔸's distribution is flatter than other datasets, spanning more evenly across different question lengths.

Context analysis. Figure 3 (b) shows the number and percentage of questions with either an image context, a text context, or both. There are a total of 7,803 unique image contexts and 4,651 unique text contexts. 66.11% of the questions have at least one type of context information. The image context is in the format of diagrams or natural images, which visualize the critical scenario necessary for question answering or simply illustrate the question for better understanding. Similarly, the textual context can provide either semantically rich information or a simple hint to the question. Therefore, models need to be flexible and general to understand these diverse types of contexts.

Domain analysis. Each 𝕊ℚ𝔸 question belongs to one of the three subjects: natural science, language science, and social science. With each subject, they are categorized first by the topic (*Biology*, *Physics*, *Chemistry*, etc.), then by the category (*Plants*, *Cells*, *Animals*, etc.), and finally by the specific skill (*Classify fruits and vegetables as plant parts*, *Identify countries of Africa*, etc.). 𝕊ℚ𝔸 has a total of 26 topics, 127 categories, and 379 skills. The treemap in Figure 4 visualizes the different subjects, topics, and categories and shows that 𝕊ℚ𝔸 questions are very diverse, spanning a wide range of domains.

4 Baselines and Chain-of-Thought Models

In this section, we establish various baselines and develop two chain-of-thought models on 𝕊ℚ𝔸.

4.1 Baselines

Heuristic baselines. The first heuristic baseline is *random chance*: we randomly select one from the multiple options. Each trial is completed on the whole test set, and we take three different trials for an average result. The second heuristic baseline is *human performance*. We post the task to Amazon Mechanical Turk and ask workers to answer 𝕊ℚ𝔸 questions. Only workers who obtain a high school or higher degree and pass the qualification examples are qualified for the study. Each worker needs to answer a set of 10 test questions, and each question is answered by three different workers. For more details of the human performance study, see Appendix B.2.

Zero-shot and few-shot baselines. We establish the zero-shot baselines on top of UnifiedQA [18] and GPT-3 [5]. The zero-shot setup follows the format of QCM→A where the input is the concatenation of tokens of the question text (Q), the context text (C), and multiple options (M), while the output is to predict the answer (A) from the option set. We extract the caption from the captioning model based on ViT [8] and GPT-2 [41] for the image as the visual context. In the few-shot setting, we follow the standard prompting [4] where in-context examples from the training set are concatenated before the test instance. These in-context examples serve as an instruction for the language model to adjust to the specific task in 𝕊ℚ𝔸.

Fine-tuning baselines. We first consider the fine-tuning baselines from VQA models [1, 20, 49, 9, 21, 31, 24] proposed in recent years. These VQA baselines take the question, the context, and choices as the textual input, take the image as the visual input, and predict the score distribution over choice candidates via a linear classifier. In addition, we build the fine-tuning baseline on top of the large language model UnifiedQA [18]. UnifiedQA takes the textual information as the input and outputs the answer option. Similarly, the image is converted into a caption that provides the visual semantics for the language model.

4.2 Language Models with the Chain of Thought

A chain of thought refers to a coherent flow of sentences that reveals the premises and conclusion of a reasoning problem [48]. A chain of thought clearly decomposes a multi-hop reasoning task into intermediate steps instead of solving the task in a black-box way. The chain of thought can be the step-by-step thought process [48] before arriving at the final answer or explanations [36] that come after the answer. The annotated lectures and explanations in $\mathbb{SQA}$ serve as *demonstrations* of the chain of thought that mimics the multi-step reasoning steps of human beings. In this paper, we study if large language models can generate reasonable explanations as the chain of thought to reveal the thought process when answering $\mathbb{SQA}$ questions. Further, we explore how the chain of thought can improve the reasoning ability of language models on $\mathbb{SQA}$ in both few-shot and fine-tuning learning.

UnifiedQA with the chain of thought. UnifiedQA [18] is a state of the art model for multi-option question answering. The original architecture of UnifiedQA takes the question and options as the input and outputs a short phrase as the final answer. We make a format modification to develop UnifiedQA with the chain of thought (CoT) i.e. UnifiedQA is fine-tuned to generate a long sequence of text which consists of the answer followed by the lecture and explanation.

GPT-3 via chain-of-thought prompting. Recent research work [5] has shown that GPT-3 [5] can perform various tasks when provided in-context examples in a standard prompt. Take multi-option question answering as an example, the standard prompt [32, 51, 27] builds instructions using in-context examples with components of the question text, options, and the correct answer text. This style of few-shot learning enables the GPT-3 model to answer specific questions without parameter updates. Different from standard prompting, we build GPT-3 via chain-of-thought (CoT) prompting, as shown in Figure 5. To be specific, for each test problem t, we map the prompt instruction $I : \{I_i\}_n, I_t$ into a textual format where $\{I_i\}_n$ refers to the instruction set of n-shot in-context examples from the training set, while I_t denotes the test instruction. Instead of the way where the explanation comes before the answer [48], we feed the instruction I into the encoder-decoder model GPT-3 to generate the answer a followed by the lecture $lect$ and explanation exp: $M : \{I_i\}_n, I_t \rightarrow a, lect, exp$.

Question: `question` : I_i^{ques}
Options: (A) `option` : I_{i1}^{opt} (B) `option` : I_{i2}^{opt} (C) `option` : I_{i3}^{opt}
Context: `context` : I_i^{cont}
Answer: The answer is `answer` : I_i^{a}. BECAUSE: `lecture` : I_i^{lect} `explanation` : I_i^{exp}

Question: `question` : I_t^{ques}
Options: (A) `option` : I_{t1}^{opt} (B) `option` : I_{t2}^{opt} (C) `option` : I_{t3}^{opt} (D) `option` : I_{t4}^{opt}
Context: `context` : I_t^{cont}
Answer:

Figure 5: Prompt instruction encoding for the text example t for GPT-3 (CoT). The prompt above consists of a 1-shot training example I_i and a test example I_t.

5 Experiments

5.1 Experimental Setup

Evaluation metrics. The heuristics and VQA baselines treat our $\mathbb{SQA}$ task as a multi-class classification problem with multiple options and are evaluated with the accuracy metrics. UnifiedQA and GPT-3 treat $\mathbb{SQA}$ as a text generation problem. So the most similar option is selected as the final prediction to evaluate the question answering accuracy. The generated lectures and explanations are evaluated by automatic metrics [39, 26, 43] and human scores by annotators.

Model	Learning	Format	NAT	SOC	LAN	TXT	IMG	G1-6	G7-12	Avg
Random chance	-	M→A	40.28	29.25	47.45	40.08	33.66	39.35	40.67	39.83
Q only [1]	train set	Q→A	41.34	27.22	47.00	41.79	35.15	39.28	40.87	39.85
C_I only [1]	train set	C_I→A	41.34	29.25	45.45	42.33	36.09	39.21	41.07	39.87
Q+M only [1]	train set	QM→A	52.66	51.86	60.18	55.57	50.37	52.53	57.88	54.44
Q+C_T+M only [1]	train set	QC_TM→A	57.28	49.04	61.36	60.46	52.80	54.44	60.51	56.61
Q+C_I+M only [1]	train set	QC_IM→A	58.97	53.77	60.45	62.85	54.49	56.72	61.04	58.26
MCAN [49]	train set	QCM→A	56.08	46.23	58.09	59.43	51.17	51.65	59.72	54.54
Top-Down [1]	train set	QCM→A	59.50	54.33	61.82	62.90	54.88	57.27	62.16	59.02
BAN [20]	train set	QCM→A	60.88	46.57	66.64	62.61	52.60	56.83	63.94	59.37
DFAF [9]	train set	QCM→A	64.03	48.82	63.55	65.88	54.49	57.12	67.17	60.72
ViLT [21]	train set	QCM→A	60.48	63.89	60.27	63.20	61.38	60.72	61.90	61.14
Patch-TRM [31]	train set	QCM→A	65.19	46.79	65.55	66.96	55.28	58.04	67.50	61.42
VisualBERT [24, 25]	train set	QCM→A	59.33	69.18	61.18	62.71	62.17	62.96	59.92	61.87
UnifiedQA$_{\text{SMALL}}$ [42]	zero-shot	QCM→A	47.78	40.49	46.00	50.24	44.12	45.56	46.21	45.79
UnifiedQA$_{\text{BASE}}$ [42]	zero-shot	QCM→A	50.13	44.54	48.18	53.08	48.09	47.58	50.03	48.46
UnifiedQA$_{\text{SMALL}}$ [42]	train set	QCM→A	53.77	58.04	61.09	52.10	51.51	58.22	53.59	56.57
UnifiedQA$_{\text{BASE}}$ [42]	train set	QCM→A	68.16	69.18	74.91	63.78	61.38	72.98	65.00	70.12
UnifiedQA$_{\text{BASE}}$ (CoT)	train set	QCM→AE	70.60	74.02	78.36	65.69	64.80	75.48	69.48	$73.33_{3.21\uparrow}$
UnifiedQA$_{\text{BASE}}$ (CoT)	train set	QCM→ALE	71.00	76.04	78.91	66.42	66.53	77.06	68.82	$74.11_{3.99\uparrow}$
GPT-3 [5]	zero-shot	QCM→A	75.04	66.59	78.00	74.24	65.74	76.36	**69.87**	74.04
GPT-3 [5]	2-shot	QCM→A	74.64	69.74	76.00	74.44	67.28	76.80	68.89	73.97
GPT-3 (CoT)	2-shot	QCM→AE	**76.60**	65.92	77.55	**75.51**	66.09	**78.49**	67.63	$74.61_{0.64\uparrow}$
GPT-3 (CoT)	2-shot	QCM→ALE	75.44	**70.87**	**78.09**	74.68	**67.43**	78.23	69.68	$\mathbf{75.17}_{1.20\uparrow}$
Human	-	QCM→A	90.23	84.97	87.48	89.60	87.50	91.59	82.42	88.40

Table 3: Evaluation of various baselines over different classes in accuracy (%). Model names: Q = question, M = multiple options, C = context, C_T = text context, C_I = image context, CoT = chain of thought. Format names: A = answer, AE = answer with explanation, ALE = answer with lecture and explanation. Question classes: NAT = natural science, SOC = social science, LAN = language science, TXT = text context, IMG = image context, G1-6 = grades 1-6, G7-12 = grades 7-12.

Implementation details. The VQA baselines are trained for a maximum number of 50 epochs with a learning rate of $5e{-}5$. We fine-tune the UnifiedQA for $50k$ iterations and evaluate every $1k$ iteration. The training process will be stopped following the early stopping strategy with a patience period of three evaluations. For GPT-3, we use the `text-davinci-002` engine, which is the most capable model version suggested in the official documentation. More details can be found in Appendix B.1.

5.2 Results for Question Answering

Table 3 demonstrates the empirical results for Science Question Answering.

VQA baselines. We feed the VQA baseline models with the input of QCM format to predict answers A. Out of all the VQA models we benchmarked, VisualBERT [24, 25] performs the best on average (61.87%). Interestingly, Patch-TRM [31] beats VisualBERT in natural science (NAT) and language science (LAN), and it also performs better in higher-grade questions (67.50% *v.s.* 59.92%). However, in the subject of social science (SOC), VisualBERT outperforms Patch-TRM by a large margin (+22.39%). Such drastic changes in performance might imply that current VQA models are not generalized to process the challenging questions in $\mathbb{SQA}$.

Language models. We evaluate whether large-scale pretraining on text can help language models learn scientific knowledge and thus perform better on the $\mathbb{SQA}$ task. For this purpose, we have tried two of the state-of-the-art pre-trained language models: UnifiedQA and GPT-3.

(i) **UnifiedQA.** The results show that without any supervised fine-tuning (zero-shot), UnifiedQA cannot beat any VQA baseline model, while the pretraining does help the model obtain some scientific knowledge to outperform the random baseline. By training the T5 model (UnifiedQA backbone) with the answer labels in $\mathbb{SQA}$, UnifiedQA$_{\text{BASE}}$ reports an accuracy of 70.12% on average. By further teaching the model to generate the answer along with lecture and explanation, the developed language

Question: What do these two changes have in common?\n
bleaching clothes\n
a piece of apple turning brown\n
Options: (A) Both are only physical changes. (B) Both are caused by cooling. (C) Both are chemical changes. (D) Both are caused by heating.\n
Answer: **The answer is C. BECAUSE:** A **physical change** is a change to a substance that does not change what the substance is. A **physical change** only changes the substance's physical properties, such as its size, shape, or color.\n
A **chemical change** is a change to a substance that changes what the substance is. A **chemical change** changes the substance's chemical properties, such as its chemical structure.\n
The changes that you listed are both chemical changes.\n
When you **bleach clothes**, the chemical structure of the fabric changes. This change makes the fabric less likely to hold color.\n
When **a piece of fruit turns brown**, the chemical structure of the fruit changes. This change makes the fruit taste different.

Figure 6: One example of the predicted answer along with the chain of thought from GPT-3 (CoT).

model with chain-of-thought (UnifiedQA$_{\text{BASE}}$ (CoT)) brings additional improvements of +3.21% (QCM→AE) and +3.99% (QCM→ALE). These results show that generating the chain of thought along with the answer benefits the reasoning ability of language models.

(ii) **GPT-3.** The positive effect of pretraining is also proved by the surprisingly good results from GPT-3 in the same zero-shot setting as UnifiedQA. Without any fine-tuning, GPT-3 already reaches almost the best performance we can get. Interestingly, prompting the GPT-3 with two training examples with just answers results in a negligible difference. However, if we prompt GPT-3 with chain-of-thought prompting (QCM→ALE), we obtain the state-of-the-art results on $\mathbb{SQA}$ so far (75.17%) on $\mathbb{SQA}$.

Human performance. Humans outperform all benchmarks consistently across question classes, context types, and grades, *e.g.,* a 20.07% gap for questions with the image context (IMG) between humans and our best performing model. The gap is to be filled by future research on multimodal reasoning for scientific question answering.

5.3 Results for Generated Explanations

One prediction example of GPT-3 (CoT) is visualized in Figure 6. We can see that GPT-3 (CoT) predicts the correct answer and generates a reasonable lecture and explanation to mimic the human thought process. We further report automatic metrics (BLEU-1 [39], ROUGE-L [39], and (sentence) Similarity [43]) to evaluate the generated lectures and explanations, as shown in Table 4. The Similarity metric computes the cosine-similarity of semantic embeddings between two sentences based on the Sentence-BERT network [43]. The results show that UnifiedQA$_{\text{BASE}}$ (CoT) generates the most similar explanations to the given ones. However, it's commonly agreed that automatic evaluation of generated texts only provides a partial view and has to be complemented by a human study. By asking annotators to rate the relevance, correctness, and completeness of generated explanations, we find that the explanations generated by GPT-3 (CoT) conform best to human judgment.

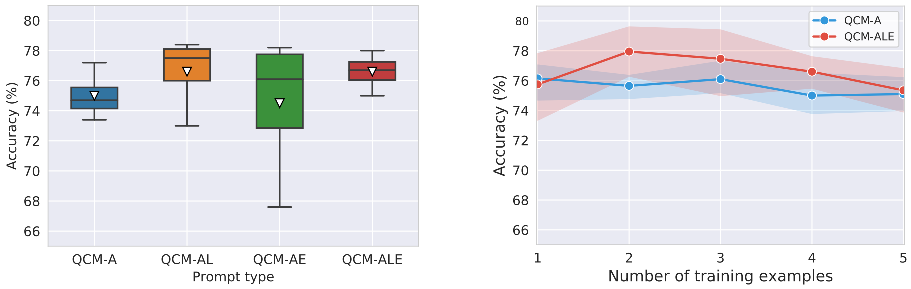
Model	Format	BLEU-1	ROUGE-L	Similarity	Relevant	Correct	Complete	Gold
UnifiedQA$_{\text{BASE}}$ (CoT)	QCM→ALE	**0.397**	**0.714**	**0.811**	80.4%	76.6%	76.1%	56.9%
GPT-3 (CoT)	QCM→AE	0.234	0.351	0.561	76.9%	73.0%	70.5%	52.5%
GPT-3 (CoT)	QCM→ALE	0.192	0.323	0.595	**88.5%**	**78.8%**	**84.5%**	**65.2%**

Table 4: Automatic metrics (BLEU-1, ROUGE-L, Similarity) and human evaluation of generated explanations. Note that a gold explanation refers to one that is relevant, correct, and complete.

5.4 Analysis

Blind studies. Blind studies are conducted on top of the modification of the full model, Top-Down [1]. The results achieved in blind studies of Q only and C_I only are close to random chance, showing that the $\mathbb{SQA}$ dataset is robust and reliable in distribution. The performance drops in Q+M only, Q+C_T+M only, and Q+C_I+M only indicate that all input components provide critical information for answering $\mathbb{SQA}$ questions.

Prompt types. We study the effect of prompt types and visualize the comparison in Figure 7 (a). It shows that prompting the GPT-3 model with both lecture and explanation (QCM→ALE) results in the highest accuracy on average and the smallest variance. In contrast, prompting with only the explanation (QCM→AE) gives the largest variance, resulting in a less stable model.



(a) Acc. v.s. different prompts with 4-shot examples.

(b) Acc. v.s. different # of training examples.

Figure 7: Accuracy of GPT-3 (CoT) cross different prompt types (a) and # of training examples (b).

Number of in-context examples. In Figure 7 (b), we further investigate how different numbers of training examples encoded in prompts can affect the prediction accuracy. The QCM→ALE prompt type outperforms or performs comparably the QCM→A type with all numbers of examples. And we observe the peak performance of QCM→ALE with 2 training examples being prompted. After that, the accuracy goes down as more training examples are added to the model.

Dynamic sampling. In Table 5, instead of random sampling, we try to dynamically select the in-context examples to prompt with the same class as the test sample. However, slight differences in prediction accuracy are observed when comparing them to simple random sampling.

Prompt type	**Sampling**	**Acc. (%)**
QCM→ALE	Dynamic (same topic)	75.15
QCM→ALE	Dynamic (same category)	74.58
QCM→ALE	Dynamic (same skill)	75.10

Table 5: Dynamic sampling for GPT-3 (CoT).

Upper bound. We search the upper bound of the GPT-3 accuracy by feeding the gold lecture and explanation in the test prompt. As reported in Table 6, QCME*→A outperforms the QCM→ALE baseline by 18.86% and QCMLE*→A outperforms QCM→ALE by 18.96%, indicating a potential improvement direction by generating correct explanations before answering science questions.

Prompt type	**Sampling**	**Acc. (%)**
QCM→ALE	Random	75.17
QCML*→A	Random	73.59
QCME*→A	Random	94.03$_{18.86\uparrow}$
QCMLE*→A	Random	**94.13**$_{18.96\uparrow}$

Table 6: Upper bound of GPT-3 (CoT).

CoT learns with fewer data. To study if the chain of thought helps language models learn more efficiently, we report the accuracies of UnifiedQA and UnifiedQA (CoT) fine-tuned on different sizes of the training set in Figure 8. UnifiedQA (CoT) benefits the language models by learning the coherent reasoning path when answering 𝕊ℚ𝔸 questions, resulting in similar accuracy with fewer training examples.

Error analysis. GPT-3 via chain-of-chain prompting obtains promising results but still fails to answer a wide range of challenging questions in 𝕊ℚ𝔸. See examples of failure cases in Appendix B.4. The failure cases can be classified into two types: (a) the model fails to understand the multimodal inputs and lacks domain-specific knowledge to arrive at the correct answer; (b) the model generates the wrong chain of thought with irrelevant, incorrect, or incomplete information.

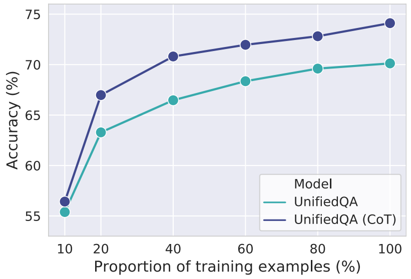


Figure 8: UnifiedQA (CoT) learns efficiently with fewer training examples.

6 Discussion and Conclusion

In this paper, we propose Science Question Answering 𝕊ℚ𝔸, a dataset that features 21,208 multi-option questions with multimodal contexts from the science curriculum. To the best of our knowledge, 𝕊ℚ𝔸 is the first large-scale multimodal science dataset where most questions are annotated with corresponding lectures and explanations. We establish various baselines, including recent VQA models and large language models on 𝕊ℚ𝔸. We further study if language models can generate reasonable explanations and then benefit the reasoning ability. Experiments show that the UnifiedQA with the chain of thought can achieve an improvement of 3.99% and few-shot GPT-3 via chain-of-thought (CoT) prompting can obtain a satisfactory accuracy of 75.17% on 𝕊ℚ𝔸. 65.2% of the generated explanations from GPT-3 (CoT) meet the gold standard by the human evaluations.

References

[1] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[2] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision (CVPR)*, pages 2425–2433, 2015.

[3] Jonathan Bragg, Arman Cohan, Kyle Lo, and Iz Beltagy. Flex: Unifying evaluation for few-shot nlp. *Advances in Neural Information Processing Systems (NeurIPS)*, 34, 2021.

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

[5] Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey E Hinton. Big self-supervised models are strong semi-supervised learners. *Advances in neural information processing systems (NeurIPS)*, 33:22243–22255, 2020.

[6] Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*, 2018.

[7] Bhavana Dalvi, Peter Jansen, Oyvind Tafjord, Zhengnan Xie, Hannah Smith, Leighanna Pipatanangkura, and Peter Clark. Explaining answers with entailment trees. *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2021.

[8] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *The International Conference on Learning Representations (ICLR)*, 2021.

[9] Peng Gao, Zhengkai Jiang, Haoxuan You, Pan Lu, Steven CH Hoi, Xiaogang Wang, and Hongsheng Li. Dynamic fusion with intra-and inter-modality attention flow for visual question answering. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6639–6648, 2019.

[10] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the V in VQA matter: Elevating the role of image understanding in Visual Question Answering. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[11] Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6700–6709, 2019.

[12] Peter A Jansen, Elizabeth Wainwright, Steven Marmorstein, and Clayton T Morrison. Worldtree: A corpus of explanation graphs for elementary science questions supporting multi-hop inference. *arXiv preprint arXiv:1802.03052*, 2018.

[13] Harsh Jhamtani and Peter Clark. Learning to explain: Datasets and models for identifying valid reasoning chains in multihop question-answering. *arXiv preprint arXiv:2010.03274*, 2020.

[14] Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2901–2910, 2017.

[15] Kushal Kafle, Brian Price, Scott Cohen, and Christopher Kanan. Dvqa: Understanding data visualizations via question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5648–5656, 2018.

[16] Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Min Joon Seo, Hannaneh Hajishirzi, and Ali Farhadi. A diagram is worth a dozen images. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2016.

[17] Aniruddha Kembhavi, Minjoon Seo, Dustin Schwenk, Jonghyun Choi, Ali Farhadi, and Hannaneh Hajishirzi. Are you smarter than a sixth grader? textbook question answering for multimodal machine comprehension. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4999–5007, 2017.

[18] Daniel Khashabi, Sewon Min, Tushar Khot, Ashish Sabharwal, Oyvind Tafjord, Peter Clark, and Hannaneh Hajishirzi. Unifiedqa: Crossing format boundaries with a single qa system. In *Findings of the Association for Computational Linguistics (EMNLP)*, pages 1896–1907, 2020.

[19] Tushar Khot, Peter Clark, Michal Guerquin, Peter Alexander Jansen, and Ashish Sabharwal. Qasc: A dataset for question answering via sentence composition. *ArXiv*, abs/1910.11473, 2020.

[20] Jin-Hwa Kim, Jaehyun Jun, and Byoung-Tak Zhang. Bilinear attention networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 1571–1581, 2018.

[21] Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, pages 5583–5594, 2021.

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

[23] Jayant Krishnamurthy, Oyvind Tafjord, and Aniruddha Kembhavi. Semantic parsing to probabilistic programs for situated question answering. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 160–170, 2016.

[24] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*, 2019.

[25] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. What does bert with vision look at? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 5265–5275, 2020.

[26] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004.

[27] Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. What makes good in-context examples for gpt-3? *arXiv preprint arXiv:2101.06804*, 2021.

[28] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 13–23, 2019.

[29] Pan Lu, Ran Gong, Shibiao Jiang, Liang Qiu, Siyuan Huang, Xiaodan Liang, and Song-Chun Zhu. Inter-gps: Interpretable geometry problem solving with formal language and symbolic reasoning. In *The 59th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2021.

[30] Pan Lu, Hongsheng Li, Wei Zhang, Jianyong Wang, and Xiaogang Wang. Co-attending free-form regions and detections with multi-modal multiplicative feature embedding for visual question answering. In *The AAAI Conference on Artificial Intelligence (AAAI)*, 2018.

[31] Pan Lu, Liang Qiu, Jiaqi Chen, Tony Xia, Yizhou Zhao, Wei Zhang, Zhou Yu, Xiaodan Liang, and Song-Chun Zhu. Iconqa: A new benchmark for abstract diagram understanding and visual language reasoning. In *The 35th Conference on Neural Information Processing Systems (NeurIPS) Track on Datasets and Benchmarks*, 2021.

[32] Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. *arXiv preprint arXiv:2104.08786*, 2021.

[33] Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. Can a suit of armor conduct electricity? a new dataset for open book question answering. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2018.

[34] Swaroop Mishra, Daniel Khashabi, Chitta Baral, Yejin Choi, and Hannaneh Hajishirzi. Reframing instructional prompts to gptk's language. *ACL Findings*, 2021.

[35] Swaroop Mishra, Daniel Khashabi, Chitta Baral, and Hannaneh Hajishirzi. Cross-task generalization via natural language crowdsourcing instructions. *The 59th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2021.

[36] Sharan Narang, Colin Raffel, Katherine Lee, Adam Roberts, Noah Fiedel, and Karishma Malkan. Wt5?! training text-to-text models to explain their predictions. *arXiv preprint arXiv:2004.14546*, 2020.

[37] Maxwell Nye, Anders Johan Andreassen, Guy Gur-Ari, Henryk Michalewski, Jacob Austin, David Bieber, David Dohan, Aitor Lewkowycz, Maarten Bosma, David Luan, et al. Show your work: Scratchpads for intermediate computation with language models. *arXiv preprint arXiv:2112.00114*, 2021.

[38] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *arXiv preprint arXiv:2203.02155*, 2022.

[39] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics (ACL)*, pages 311–318, 2002.

[40] Ethan Perez, Douwe Kiela, and Kyunghyun Cho. True few-shot learning with language models. *Advances in Neural Information Processing Systems (NeurIPS)*, 34, 2021.

[41] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

[42] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research (JMLR)*, 21:1–67, 2020.

[43] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, 11 2019.

[44] Mrinmaya Sachan, Kumar Dubey, and Eric Xing. From textbooks to knowledge: A case study in harvesting axiomatic knowledge from textbooks to solve geometry problems. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 773–784, 2017.

[45] Mrinmaya Sachan and Eric Xing. Learning to solve geometry problems from natural language demonstrations in textbooks. In *Proceedings of the 6th Joint Conference on Lexical and Computational Semantics (* SEM 2017)*, pages 251–261, 2017.

[46] Shailaja Keyur Sampat, Yezhou Yang, and Chitta Baral. Visuo-lingustic question answering (vlqa) challenge. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings (EMNLP)*, pages 4606–4616, 2020.

[47] Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. Finetuned language models are zero-shot learners. *The International Conference on Learning Representations (ICLR)*, 2021.

[48] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Chi, Quoc Le, and Denny Zhou. Chain of thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*, 2022.

[49] Zhou Yu, Jun Yu, Yuhao Cui, Dacheng Tao, and Qi Tian. Deep modular co-attention networks for visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6281–6290, 2019.

[50] Peng Zhang, Yash Goyal, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Yin and Yang: Balancing and answering binary visual questions. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[51] Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. Calibrate before use: Improving few-shot performance of language models. In *International Conference on Machine Learning (ICML)*, pages 12697–12706. PMLR, 2021.

[52] Yuke Zhu, Oliver Groth, Michael Bernstein, and Li Fei-Fei. Visual7w: Grounded question answering in images. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] Yes, we did the error analysis in Section 5.4 and discussed the limitations of the work in Appendix B.4.
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] We discussed the broader impacts in Appendix B.5.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A]
   (b) Did you include complete proofs of all theoretical results? [N/A]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] We included 100 data examples and the data visualizer tool in the supplemental material. The whole dataset and code will be available at `https://sqa.github.io`.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] See Section 5.1 and Appendix B.1 for experimental details.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] We reported the error bars for GPT-3 (CoT) experiments in Figure 7, where each experiment was repeated four times.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] We discussed compute resources in Appendix B.1.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes] We collected the $\mathbb{SQA}$ dataset from `https://www.ixl.com/`. The copyright belongs to IXL.
   (b) Did you mention the license of the assets? [Yes] $\mathbb{SQA}$ is under the CC BY-NC-SA 4.0 license and is used for non-commercial research purposes.
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes] We included data examples and a visualizer tool in the supplemental material. The dataset will be available at `https://sqa.github.io`.
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [Yes] The collected data does not contain personally identifiable information or offensive content.
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [Yes] We included screenshots of the instructions in Appendix B.2 and B.3.
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [Yes] We included the monetary compensation details in Appendix B.2 and B.3.