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

A Dataset Analysis

A.1 Data Collection

Questions in the SCIENCEQA dataset are sourced from open resources managed by IXL Learning, an online learning platform curated by experts in the field of K-12 education. The dataset includes problems that align with *California Common Core Content Standards*. To construct SCIENCEQA, we downloaded the original science problems and then extracted individual components (e.g. questions, hints, images, options, answers, lectures, and solutions) from them based on heuristic rules.

We manually removed invalid questions, such as questions that have only one choice, questions that contain faulty data, and questions that are duplicated, to comply with *fair use* and *transformative use* of the law. If there were multiple correct answers that applied, we kept only one correct answer. Also, we shuffled the answer options of each question to ensure the choices do not follow any specific pattern. To make the dataset easy to use, we then used semi-automated scripts to reformat the lectures and solutions. Therefore, special structures in the texts, such as tables and lists, are easily distinguishable from simple text passages. Similar to ImageNet, ReClor, and PMR datasets, SCIENCEQA is available for non-commercial research purposes only and the copyright belongs to the original authors. To ensure data quality, we developed a data exploration tool to review examples in the collected dataset, and incorrect annotations were further manually revised by experts. The tool can be accessed at <https://scienceqa.github.io/explore.html>.

A.2 Question Statistics

Figure 9 (a) is a word cloud showing the most frequently appeared words in the question texts. Stopping words that do not contain any semantic meaning, such as “*what*” or “*and*”, are removed to give us a clearer view of the semantic range of SCIENCEQA. The diagram shows that SCIENCEQA covers a wide range of topics, with words from different topics showing up across the cloud.

Figures 9 (b) (c) (d) show the word clouds for each of the three subjects. We can observe from the word clouds that the words are well-matched to the subject themes. In natural science questions, words such as “*trait*”, “*magnet*”, and “*force*” appear frequently. Words such as “*capital*” and “*state*” show up frequently in social science questions, whereas words such as “*dictionary*” and “*page*” are common in language science questions.

A.3 Choice Statistics

Table 8 shows the number of questions with each number of different choices. Questions have a minimum of two options and a maximum of five options. Figure 10 shows the distribution of choice length in SCIENCEQA. Most choices are short, containing up to five words. However, the distribution has a long tail where about 5% of the choices contain more than 15 words. Hence, it requires models to have a high level of text understanding to address diversely distributed choices.

Choice number	Size	Percent
2	11,045	52.08%
3	5,078	23.94%
4	4,893	23.07%
5	192	0.91%

Table 8: Choice number distribution.



Figure 9: Word cloud distributions of question texts in different subjects.

A.4 Subject Statistics

Figure 11 shows the question length distribution of each subject. The three subjects all feature long-tail distributions in terms of the number of question words. On average, social science questions are the shortest, while language science questions are the longest. Language science questions are distributed more evenly than other questions across different numbers of words. These features imply that the SCIENCEQA dataset is rich in compositional diversity.

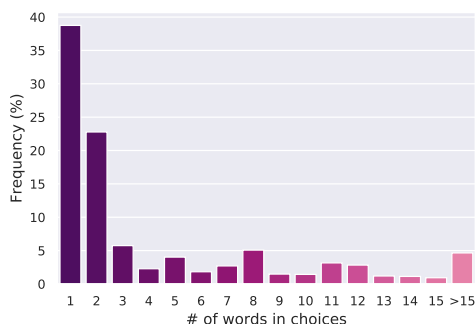


Figure 10: Choice length distribution.

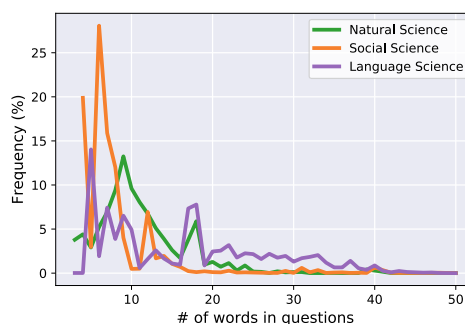


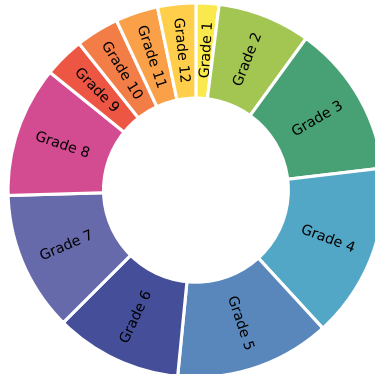
Figure 11: Question distributions of diff. subjects.

A.5 Grade Statistics

The grade distribution is shown in Figure 12. The majority of questions come from the middle level curriculum (i.e., from grade 3 to grade 8) while around 10% are taken from the high school curriculum (i.e., from grade 9 to grade 12). These high school level questions are close to or at the difficulty level of the U.S. standardized tests for college admissions. Machine algorithms need to master a large amount of scientific knowledge and perform complex reasoning in order to perform well on SCIENCEQA.

Grades	Number	Percent
Grade 1	95	0.45%
Grade 2	1,678	7.91%
Grade 3	3,032	14.3%
Grade 4	3,544	16.71%
Grade 5	3,086	14.55%
Grade 6	2,450	11.55%
Grade 7	2,749	12.96%
Grade 8	2,546	12.0%
Grade 9	491	2.32%
Grade 10	558	2.63%
Grade 11	539	2.54%
Grade 12	440	2.07%

(a) Grade distribution statistics.



(b) Grade distribution visualization.

Figure 12: SCIENCEQA questions and their corresponding grades.

B Experiments

B.1 Experimental Details

Below are details on the experiments:

- **Fine-tuning on the dataset.** Fine-tuning baselines (VQA baselines and UnifiedQA) are trained on the training set, developed on the validation set, and evaluated on the test set.
- **Input sizes:** For VQA baselines, we set the maximum number of input words or tokens as 100.
- **Batch sizes.** We use batches of 64 and 4 for VQA baselines and fine-tuned UnifiedQA, respectively.
- **Newline character.** For language models, the newline separators (`\n`) in the text are replaced with `\n` when encoding the inputs because `\n` is normally used as a stop symbol, following the original works [4, 19].
- **Captioning model.** We use the tool² to generate captions for the images in the dataset. The maximum length of generated captions is 16, the number of beams is 4, and the maximum number of output tokens is 512.
- **Compute resources.** We use two GeForce RTX 3090 GPUs for fine-tuning VQA baselines and UnifiedQA on the dataset.
- **Questions without any context.** For questions without any context, the context text is replaced with an empty string.
- **GPT-3:** Following default settings, we choose temperature, frequency penalty and presence penalty as 0.0, and top probability as 1.0. All experiments for GPT-3 are run via the online API. Experiments in Figure 7 are repeated four times with in-context examples listed in Table 9. Experiments in Table 3, 5, 6, and 7 are conducted using examples with the trial ID of 1.

Trial IDs	Random seeds	In-context example IDs
1	3	6493, 16241, 14954, 3598, 10088
2	5	17099, 6960, 20290, 9780, 18898
3	7	8836, 4144, 10781, 17852, 1363
4	9	12701, 16832, 10180, 7289, 3801

Table 9: Training example candidates used in four trials for GPT-3 (CoT).

B.2 Human Performance Study

In order to understand how humans perform on SCIENCEQA questions, we used Amazon Mechanical Turk (AMT) to crowdsource answers to the test set. The interface of instructions and one example

²<https://huggingface.co/nlpconnect/vit-gpt2-image-captioning>

Overview

Thank you for helping us with our research!

- You will be answering up to **15 multiple choice questions** from grade school science curriculum. You will have up to **20 minutes**.
- Each question will have **2 - 5 choices**. Only one is the correct answer. For some questions, there will be text or image (or both) context information that will help you answer the question.
- If a particular question seems ambiguous (no correct answer/more than one correct answer/etc.), please choose the answer that makes the most sense to you.
- We will kindly ask you to provide us with your highest achieved degree. This will help us with data analysis in our research.

What's your highest achieved degree?

☐ Others
☒ High school degree
☐ Bachelor's degree
☐ Master's degree or higher

(a) Instructions to answer the SCIENCEQA questions.

Instruction: Select the choice that best answers the question.

Which of these organisms contains matter that was once part of the lichen?

Below is a food web from a tundra ecosystem in Nunavut, a territory in Northern Canada. A food web models how the matter eaten by organisms moves through an ecosystem. The arrows in a food web represent how matter moves between organisms in an ecosystem.

☐ bilberry
☐ mushroom

(b) One test question example.

Figure 13: Interfaces of instructions and one test question example for AMT workers.

Overview

Thank you for helping us with our research!

- You will be given **10 multiple choice questions**. For each question, you will also be given a **model proposal**. The model proposal is a solution along with its explanation generated by a machine learning algorithm.
- Your task is to determine whether the machine learning proposal is **relevant**, **correct**, and **complete**.
- A proposal is **relevant** if it talks about topics related to the question.
- A proposal is **correct** if it picks the correct answer and has a correct explanation.
- A proposal is **complete** if the explanation fully explains the answer.
- If it's not clear whether a proposal is relevant, correct, or complete, pick the best option by your judgment.
- We will kindly ask you to provide us with your highest achieved degree. This will help us with data analysis in our research.

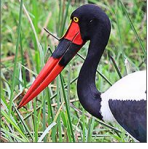
Figure 14: Interface of instructions for AMT workers to evaluate the explanations generated from UnifiedQA (CoT) and GPT-3 (CoT).

of a test question is shown in Figure 13. A total of 4,241 test questions were shuffled and split into 425 batches, with each batch having 10 questions (excluding the last one). For each batch, we also randomly added five training questions as exam examples. Each set of 15 questions was then assigned to 3 AMT workers. Only workers who correctly answer 4 out of the 5 exam examples or more are qualified for the human performance study. In other words, workers who failed to pass the qualified exam were eliminated from the analysis. For each set of 15 questions, we provided the worker with \$0.5 per HIT task. At the rate of 3 questions per minute, this amounts to \$6.0 per hour.

B.3 Human Evaluation of Generated Explanations

We also evaluated the quality of predictions from GPT-3 (CoT) and UnifiedQA (CoT) by asking AMT workers to rate the model-generated explanations. The interface is shown in Figure 14. Each sample's question text, contexts, choices, and answers were presented, along with the corresponding explanation generated by language models. The workers were asked to decide whether the proposed explanation is *relevant* (is related to the question), *correct* (gives a correct answer and explanation), and *complete* (fully explains the answer). Prediction outputs that contain textual explanations were grouped into batches of 10, each assigned to 3 workers for evaluation. For each batch, we provided

Question: Which animal's neck is also adapted for hunting prey while keeping the rest of its body still?
Context: Saddle-billed storks live near wetlands and lakes. They eat mostly fish. The 's neck helps it grab fish while keeping the rest of its body still. If the stork had to move its body, it might scare the fish away.
 Figure: saddle-billed stork. A black and white bird is standing in the grass.
Options: (A) northern pintail (B) black-headed heron
Answer: The answer is B. BECAUSE: The northern pintail is a duck. It does not hunt for fish.
 The black-headed heron is a type of heron. It lives near wetlands and lakes. It eats mostly fish.
 The heron's neck helps it grab fish while keeping the rest of its body still. If the heron had to move its body, it might scare the fish away.



(Subject: Natural science, Topic: Biology, Category: Adaptations, Skill: Animal adaptations: beaks, mouths, and necks)

(a) A natural science example with the correct answer and a gold explanation.

Question: Suppose Mason decides to bake blueberry muffins. Which result would be a cost?
Context: Mason is deciding whether to bake blueberry muffins or cranberry muffins. He wants the muffins to be tasty. But he also wants to make a lot of muffins.
Options: (A) Mason will get to eat blueberry muffins. He thinks blueberry muffins are tastier than cranberry muffins. (B) Mason will give up some muffins. He could have made more cranberry muffins than blueberry muffins.
Answer: The answer is B. BECAUSE: A cost is anything that you have to give up to get something else.
 In this case, Mason will have to give up some muffins. He could have made more cranberry muffins than blueberry muffins. So, the cost of making blueberry muffins is the opportunity cost of not making cranberry muffins.

(Subject: Social science, Topic: economics, Category: Basic economic principles, Skill: Costs and benefits)

(b) A social science example with a correct answer and a gold explanation.

Question: Which sentence uses a metaphor?
Context: N/A
Options: (A) The cat's silver eyes were two shiny coins. (B) The cat's silver eyes were like two shiny coins.
Answer: The answer is A. BECAUSE: A metaphor is a figure of speech that uses one thing to describe another. A metaphor is a type of analogy.
 In a metaphor, one thing is said to be another thing.
 For example, you might say, "My dad is a bear."
 This means that your dad is big and strong, not that he is actually a bear!
 The sentence "The cat's silver eyes were two shiny coins" is a metaphor.
 The sentence "The cat's silver eyes were like two shiny coins" is not a metaphor. It is a simile.
 A simile is a figure of speech that uses the words "like" or "as" to compare two things.

(Subject: Language science, Topic: Figurative language, Category: Literary devices, Skill: Identify similes and metaphors)

(c) A language science example with a correct answer and a gold explanation.

Figure 15: Three examples with a correct answer and a gold explanation predicted by GPT-3 (CoT).

the workers with a monetary compensation of \$0.3. Finally, the human scores for each explanation were determined by taking a majority vote.


B.4 Case Study and Limitations

Figure 15 shows three examples with correct answers and gold explanations predicted by GPT-3 via *chain-of-thought* prompting (CoT). We can see that GPT-3 (CoT) not only predicts the correct answers but also generates reasonable explanations, which follow the multi-hop reasoning process of human beings. This suggests that large language models like GPT-3 have great promise for implementing high-level reasoning abilities.

Figure 16 visualizes three more examples with predictions from GPT-3 (CoT). In these examples, GPT-3 (CoT) is able to predict the correct answers but fails to generate gold explanations. For example, GPT-3 (CoT) generates an *irrelevant* explanation because the context text does not include fine-grained visual information in the image (Figure 16a). In the example shown in Figure 16b, GPT-3 (CoT) fails to predict the coherent thought chains, where there are an *incorrect* example and an *incorrect* statement for a *chemical change*. The third example is given in Figure 16c, where the generated explanation is just a repetition of the input question and the output answer, instead of following the *complete* thought chain to arrive at the final answer.

Four failure examples with wrong predicted answers are listed in Figure 17. We extract the image captions and feed them to the large language model as the visual content input. However, these captions lack fine-grained semantics and usually do not work well for diagrams, which results in

Question: Which is the main persuasive appeal used in this ad?
Context: A black dog standing in front of a fence.
Options: (A) logos (reason) (B) ethos (character) (C) pathos (emotion)



Answer: The answer is C. **BECAUSE:** The ad is trying to make you feel a certain way about the product. It is not trying to persuade you with logic or with an appeal to character.

(Subject: Language science, Topic: Writing strategies, Category: Persuasive strategies, Skill: Identify appeals to ethos, pathos, and logos in advertisements)

(a) An example with a correct answer but an irrelevant explanation.

Question: What do these two changes have in common?
 mixing sand and gravel
 crushing a mineral into powder
Context: N/A
Options: (A) Both are chemical changes. (B) Both are only physical changes. (C) Both are caused by cooling. (D) Both are caused by heating.

Answer: The answer is B. **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.
 For example, when you cut a mineral, you change its shape. But the mineral is still the same substance. So, cutting a mineral is a physical change.
 A chemical change is a change to a substance that changes what the substance is. A chemical change changes the substance's chemical properties.
 For example, when you mix sand and gravel, you make concrete. Concrete is a new substance with different properties from sand and gravel. So, mixing sand and gravel is a chemical change.
 Crushing a mineral into powder is a physical change. The mineral is still the same substance.

(Subject: Natural science, Topic: Chemistry, Category: Physical and chemical change, Skill: Compare physical and chemical changes)

(b) An example with a correct answer but an incorrect explanation.

Question: Is the following trait inherited or acquired?
 Marshall has five fingers on each hand.
Context: Hint: Marshall was born with five fingers on each hand.
Options: (A) acquired (B) inherited

Answer: The answer is B. **BECAUSE:** Marshall was born with five fingers on each hand. This trait was inherited from Marshall's parents.

(Subject: Natural science, Topic: Biology, Category: Heredity, Skill: Identify inherited and acquired traits)

(c) An example with a correct answer but an incomplete explanation.

Figure 16: Three examples with predictions from GPT-3 (CoT). The answers are correct but the generated explanations are irrelevant, incorrect, or incomplete.

two failure cases shown in Figure 17a and 17b. Moreover, there exist challenges for large language models to reason about the questions that require them to understand complex and uncommon domain knowledge. For example, GPT-3 (CoT) cannot understand accurately the terminology of *personification* in language science (Figure 17c) and a series of complex *chemical changes* happen in the formation process of dinosaur fossils (Figure 17d).

B.5 Broader Impacts

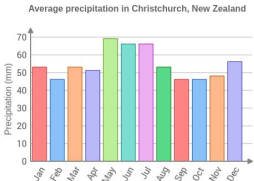
Societal impact. The SCIENCEQA dataset collects science questions sourced from textbooks and is proposed to diagnose the multimodal understanding and multi-hop reasoning abilities of AI systems. Due to the nature of data sources, SCIENCEQA does not contain any user usage data or personally sensitive information such as gender and race. After careful examination of our dataset, to our best knowledge, we have not found any improper content, such as pornographic information, racial remarks, or harmful social bias. We adhere to the goal of AI for the common good, and any antisocial data points will be removed from the dataset based on feedback.

Potential usage. The proposed SCIENCEQA dataset and designed methods in this paper are beneficial to both follow-up research work and real-world applications. SCIENCEQA provides a useful benchmark for multi-modal learning, multi-hop reasoning, and general artificial intelligence. Besides,

Question: Which month is the wettest on average in Christchurch?

Context: Use the graph to answer the question below. A series of photographs showing a variety of colorful objects.

Options: (A) April (B) December (C) May



Answer: The answer is B. BECAUSE: The graph shows the average monthly rainfall in Christchurch from 1971 to 2000. The wettest month on average is December, with an average rainfall of 154.8 mm. The second wettest month is May, with an average rainfall of 153.1 mm. The third wettest month is April, with an average rainfall of 151.4 mm.

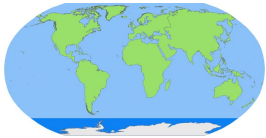
(Subject: Natural science, Topic: Earth science, Category: Weather and climate, Skill: Use climate data to make predictions)

(a) An example with a **wrong** answer and a **wrong** explanation.

Question: Which ocean is highlighted?

Context: A painting of a penguin on a blue background.

Options: (A) the Indian Ocean (B) the Arctic Ocean (C) the Pacific Ocean (D) the Southern Ocean



Answer: The answer is C. BECAUSE: The Pacific Ocean is the largest ocean. It covers more than one-third of the Earth's surface.

(Subject: Social science, Topic: Geography, Category: Physical Geography, Skill: Oceans and continents)

(b) An example with a **wrong** answer and a **wrong** explanation.

Question: Complete the sentence so that it uses personification.

At night, when the machines were () the factory seemed to be an entirely different place.

Context: N/A

Options: (A) off (B) asleep

Answer: The answer is A. BECAUSE: Personification is a figure of speech that gives human qualities to nonhuman things. For example, the sentence "The wind was howling" uses personification. It gives the human quality of sound to the nonhuman thing wind. The sentence "The sun was smiling" uses personification. It gives the human quality of expression to the nonhuman thing sun. In the sentence "At night, when the machines were off, the factory seemed to be an entirely different place," the machines are given the human quality of sound.

(Subject: Language science, Topic: Writing strategies, Category: Creative techniques, Skill: Use personification)

(c) An example with a **wrong** answer and a **wrong** explanation.

Question: Complete the sentence.

A dinosaur bone turning into rock over millions of years is a ().

Context: N/A

Options: (A) chemical change (B) physical change

Answer: The answer is B. BECAUSE: A physical change is a change to a substance that does not change what the substance is made of. A chemical change is a change to a substance that does change what the substance is made of. In this sentence, the word 'change' is a verb. It is an action word. The word 'change' is also a noun. It is a thing. The word 'change' can be a physical change or a chemical change. A physical change is a change to a substance that does not change what the substance is made of. For example, when water freezes, it changes from a liquid to a solid. But it is still made of water molecules. So, freezing is a physical change. A chemical change is a change to a substance that does change what the substance is made of. For example, when water evaporates, it changes from a liquid to a gas. But it is still made of water molecules. So, evaporating is a physical change. In this sentence, the word 'change' is a physical change. A dinosaur bone turning into rock over millions of years is a physical change.

(Subject: Natural science, Topic: Chemistry, Category: Physical and chemical change, Skill: Compare physical and chemical changes)

(d) An example with a **wrong** answer and a **wrong** explanation.

Figure 17: Four failure examples with predictions from GPT-3 (CoT). The answers are **wrong**, and the generated explanations fail to follow the right chain-of-thought reasoning process.

SCIENCEQA will contribute to the development of K-12 education applications such as tutoring systems. Furthermore, the designed methods with the chain of thought investigate the ability of large language models to mimic the human mind process when reasoning about a challenging task.