T-SciQ: Teaching Multimodal Chain-of-Thought Reasoning via Large Language Model Signals for Science Question Answering

Lei Wang
Singapore Managemen University
Singapore, Singapore
lei.wang.2019@phdcs.smu.edu.sg

Yi Hu
University of Electronic Science and Technology of China
Chengdu, China
yihu0118@gmail.com

Jiabang He
University of Electronic Science and Technology of China
Chengdu, China
JiaBangH@outlook.com

Xing Xu
University of Electronic Science and Technology of China
Chengdu, China
xing.xu@uestc.edu.cn

Ning Liu
Beijing Forestry University
Beijing, China
liuning0928@bjfu.edu.cn

Hui Liu
Beijing Rongda Technology Co., Ltd.
Beijing, China
ryuki122382@gmail.com

Heng Tao Shen
University of Electronic Science and Technology of China
Chengdu, China
shenhengtao@hotmail.com

ABSTRACT

Large Language Models (LLMs) have recently demonstrated exceptional performance in various Natural Language Processing (NLP) tasks. They have also shown the ability to perform chain-of-thought (CoT) reasoning to solve complex problems. Recent studies have explored CoT reasoning in complex multimodal scenarios, such as the science question answering task, by fine-tuning multimodal models with high-quality human-annotated CoT rationales. However, collecting high-quality COT rationales is usually time-consuming and costly. Besides, the annotated rationales are hardly accurate due to the redundant information involved or the essential information missed. To address these issues, we propose a novel method termed *T-SciQ* that aims at teaching science question answering with LLM signals. The T-SciQ approach generates high-quality CoT rationales as teaching signals and is advanced to train much smaller models to perform CoT reasoning in complex modalities. Additionally, we introduce a novel data mixing strategy to produce more effective teaching data samples for simple and complex science question answer problems. Extensive experimental results show that our T-SciQ method achieves a new state-of-the-art performance on the ScienceQA benchmark, with an accuracy of 96.18%. Moreover, our approach outperforms the most powerful fine-tuned baseline by 4.5%.

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-1-4503-XXXX-X/18/06…$15.00
https://doi.org/XXXXXXX.XXXXXXX

CCS CONCEPTS

• **Theory of computation** → **Semantics and reasoning**; • **Applied computing** → *Education.*

KEYWORDS

Large Language Models, Fine-tuning, Multimodal Chain-of-Thought

ACM Reference Format:
Lei Wang, Yi Hu, Jiabang He, Xing Xu, Ning Liu, Hui Liu, and Heng Tao Shen. 2023. T-SciQ: Teaching Multimodal Chain-of-Thought Reasoning via Large Language Model Signals for Science Question Answering. In *Proceedings of ACM Conference (Conference'17).* ACM, New York, NY, USA, 19 pages. https://doi.org/XXXXXXX.XXXXXXX

ACKNOWLEDGMENTS

To Robert, for the bagels and explaining CMYK and color spaces.

1 INTRODUCTION

Scientific problem solving has recently been employed to evaluate the multi-hop reasoning capability and interpretability of AI systems [6, 12, 28]. However, these datasets [11, 12] suffer from limited scale. To address this issue, Lu et al. [22] introduces a large-scale science question-answering dataset across broad topics and skills called ScienceQA. This dataset consists of 21,208 multimodal data examples associated with questions, context, images, options, lectures, and explanations. An example is shown in Figure 1, illustrating that a model must comprehend multimodal inputs and incorporate external knowledge to answer scientific questions.

Recently, Large Language Models (LLMs) have shown exceptional performance in various Natural Language Processing (NLP) tasks [2, 30]. Specifically, they have demonstrated the chain-of-thought (CoT) ability to solve complex reasoning problems they

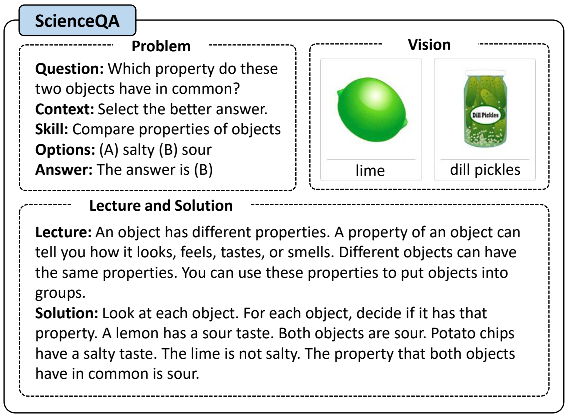


Figure 1: Example of ScienceQA dataset. Each data example includes input from multiple modalities, encompassing a question, context, images, skill information, and options. Context and images may not be present. Human annotations include the ground truth answer, lecture, and solution.

have not encountered before by conditioning on a few demonstration examples or using inherent knowledge in LLMs, without additional training [17, 36, 41]. However, the existing research on CoT reasoning is mainly limited to the language modality [7, 24, 34, 43], with little attention paid to multimodal scenarios, such as science question answering. To address this limitation, a common approach is to use caption models to translate visual information into the language modality and prompt LLMs to perform CoT reasoning. However, the use of caption generation models in scientific problems may result in significant information loss when meeting highly complex images. To overcome this challenge, Zhang et al. [42] proposed a method called Multimodal-CoT that incorporates both language and vision modalities into a two-stage fine-tuning process, which separates rationale generation and answer inference.

The Multimodal-CoT method has a significant disadvantage because it relies on the human-annotated CoT rationale to fine-tune the model. While incorporating human-annotated CoT signals is helpful for training models to facilitate CoT reasoning ability, it has three fundamental limitations. Firstly, the human annotation of CoT reasoning can be time-consuming, particularly for complex tasks like Science Question Answering, which necessitates extensive expert knowledge and background to create a rationale to deduce the final answer. Secondly, human-annotated rationales often include redundant information that may not be necessary to derive the correct answer [29]. Finally, the annotated rationale may lack essential information to derive the final answer due to the limited expertise of human annotators.

To address these issues, we propose a novel approach named *T-SciQ* in this paper to solve the complex science question answering task. As the general flowchart shown in Figure 2, the proposed T-SciQ method utilizes the CoT reasoning capabilities of Large Language Models (LLMs) to teach small multimodal models. Specifically, we introduce a zero-shot prompting method that allows LLMs to generate QA-CoT samples as teaching data samples. To improve the quality of the rationale, the ground truth answer for each training data sample is used as a hint to guide the LLM to generate high-quality CoT rationale as teaching signals. The generated teaching signals are then used to fine-tune smaller student models using the Multimodal-CoT [42] two-stage fine-tuning framework that includes rationale generation teaching and answer inference teaching. Both stages use the same model architecture. During inference, the model trained in the first stage generates rationales for the test sets based on the input. The generated rationales are subsequently used in the second stage to infer answers.

The proposed QA-CoT has shown promise in addressing the limitations of human-annotated CoT rationales. However, applying QA-CoT teaching to highly complex problems remains challenging, particularly in science question answering tasks, where some questions are exceedingly intricate. To overcome this challenge, we propose a 3-step zero-shot prompting approach that enables LLMs to use planning-based CoT rationales as teaching signals. These rationales decompose complex problems into simpler subproblems that are easier to solve, and we refer to these teaching samples as QA-PCoT samples. To leverage the benefits of both QA-CoT for simple problems and QA-PCoT for complex problems, we further introduce a data mixture strategy to obtain a new dataset called the T-SciQ teaching dataset. To build the T-SciQ dataset, we utilize the validation set to determine whether PCoT or CoT is more appropriate for data samples of a given skill.

We empirically evaluate T-SciQ over GPT-3.5 on the ScienceQA benchmark [22]. Experimental results show that our method surpasses the previous state-of-the-art by a large margin. Specifically, Mutimodal-T-SciQ$_{\text{Large}}$ outperforms the most powerful fine-tuned baseline, Mutimodal-CoT$_{\text{Large}}$, which is trained by annotated chain-of-thought signals, by 4.5% (91.68% $\rightarrow$ 96.18%), the strongest instruction-tuning based multimodal baseline, LLaVa [21], by 5.26% (90.92% $\rightarrow$ 96.18%), the best GPT-4 [23] based few-shot baseline, Chameleon, by 9.64% (86.54% $\rightarrow$ 96.18%), and human performance by 7.78% (88.40% $\rightarrow$ 96.18%). The contributions of this work can be summarized as follows:

- We propose a novel framework for generating high-quality CoT rationale and training smaller models to perform CoT reasoning in complex modalities.
- We introduce a new data mixing strategy to produce effective teaching data samples for simple and complex science question answer problems.
- Our method achieves a new state-of-the-art performance on the ScienceQA benchmark, surpassing all previous state-of-the-art models by a large margin.

2 RELATED WORK

Chain-of-Thought with In-context Learning. Recently, in order to take advantage of the complex reasoning ability of large language models, Chain-of-Thought prompting [37] has been widely used. CoT prompting can achieve better results in multi-step reasoning problems by prompting large language models to generate intermediate reasoning processes before reaching the final answer. Subsequently, a lot of work has been proposed to further improve the performance of CoT prompting from different aspects, including improving the quality of demonstrations [7, 24, 27, 41] and improving the quality of reasoning chains [5, 14, 20, 34, 35, 43]. Zero-shot

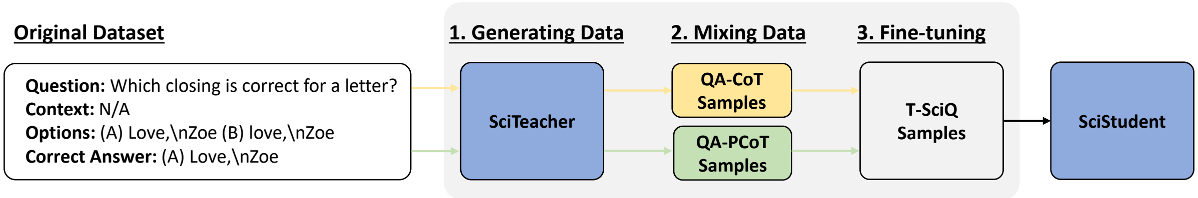


Figure 2: Key steps of our T-SciQ approach. T-SciQ consists of three stages: (i) generating data; (ii) mixing data; and (iii) fine-tuning. Firstly, we use two different prompting format, namely CoT prompting and PCoT prompting, to generate lectures and solutions of ScienceQA. Secondly, we mix QA-CoT and QA-PCoT data based on the results on the validation set. Finally, we fine-tune multimodal models on the mixed data.

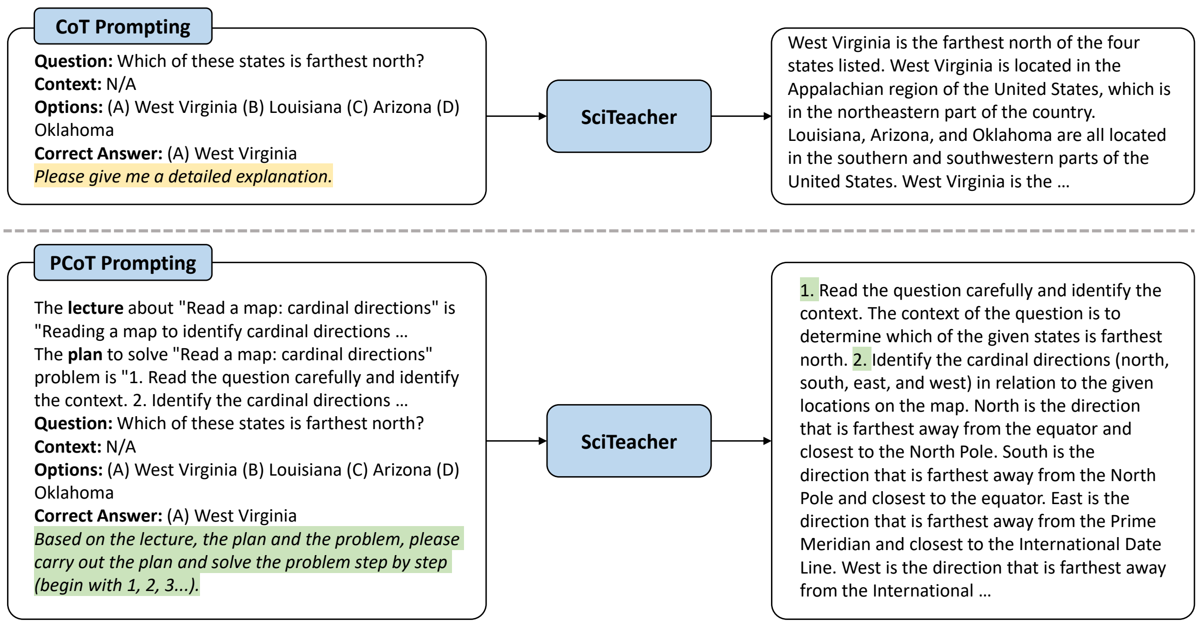


Figure 3: Examples of generating solutions by CoT prompting and PCoT prompting. The lecture and plan in PCoT prompting need to be generated firstly for each skill.

CoT [17] elicited reasoning step by appending a prompt like "Let's think step by step" to the test question. However, the performance of Zero-shot CoT generally does not perform as well as Few-shot CoT. Therefore, most recent studies focused on how to improve the performance of Few-shot CoT. Iterative Prompting [33] proposed a multi-step and iterative context-aware prompter, which learns to dynamically synthesize prompts conditioned on the current step's contexts. PoT Prompting [5] introduced program-of-thoughts, which wrote a program as a rationale and invoked the reasoning ability of LLMs by executing the generated program. Auto-CoT [41] designed an automatic CoT prompting method, which obtained k representative examples by partitioning questions of a given dataset into k clusters. It then followed the process of Zero-shot CoT to generate reasoning chain for each selected example to construct demonstrations. Chameleon [23] proposed a plug-and-play compositional reasoning framework which utilized large language models, off-the-shelf vision models, web search engines, Python functions, and rule-based modules to obtain high quality prompting. With the help of various tools, it got promising results on ScienceQA [22] dataset and TabMWP [24] dataset.

Our work is different from the above works by focusing on mixing single-step reasoning chain and multi-step reasoning chain based on the prediction results of the validation set.

Chain-of-Thought with Fine-tuning. In recent studies, CoT reasoning was elicited using fine-tuned language models. Ho et al. [10] proposed Fine-tune-CoT, which leveraged the capabilities of large language models to generate reasoning samples and teach smaller models via fine-tuning. Based on its diverse reasoning, Fine-tune-CoT made the performance of fine-tuned small language models in arithmetic reasoning, commonsense reasoning and symbolic reasoning similar to that of large language models with in-context learning. Lu et al. [22] performed fine-tuning on the encoder-decoder T5 model using a large-scale multimodal dataset annotated with CoT and image captions. However, a reduction in performance was observed when CoT was used to infer answers by generating the reasoning chain before the answer. This is because of the loss of information when converting images into image captions. In order to solve this problem, Multimodal-CoT [42] used two-stage training and visual features rather than image captions to achieve state of the art results on ScienceQA dataset.

However, Multimodal-CoT [42] used homogeneous human annotated lectures and solutions, which may have redundant information and affect the prediction results of the model. Our work exploits two styles of reasoning chains generated by large language models, which can further improve performance.

3 OUR T-SCIQ APPROACH

3.1 Overview

In this section, we present a novel approach called T-SciQ, which is a fine-tuning strategy that utilizes a large language teacher model (LLM) named SciTeacher to generate teaching data and subsequently improve the performance of a smaller student model (SciStudent). The proposed T-SciQ strategy comprises three key components: generating teaching data, mixing teaching data, and fine-tuning, as depicted in Figure 2. To generate the teaching data, we leverage SciTeacher to produce chain-of-thought rationale (CoT), which enables us to obtain Question-Answer-CoT (QA-CoT) samples, and planning-based chain-of-thought rationale (PCoT), which enables us to obtain Question-Answer-PCoT (QA-PCoT) samples. By combining the strengths of QA-CoT and QA-PCoT samples, we can create T-SciQ samples, which we then use to fine-tune the smaller student models. In the following, we provide a detailed description of generating teaching data process, mixing teaching data, and fine-tuning.

3.2 Generating Teaching Data

As depicted in Figure 3, we produce two types of data samples for teaching, each with a unique teaching rationale. Specifically, the first type is the QA-CoT sample with a generated CoT rationale. In contrast, the second type is the QA-PCoT sample equipped with a generated PCoT rationale.

QA-CoT Sample Generation. Although using human-annotated chain-of-thought signals is valuable for training models to enable chain-of-thought reasoning ability, it has three inherent limitations. Firstly, human annotation of chain-of-thought reasoning can be time-consuming, particularly for complex tasks such as Science Question Answering [22], which require extensive expert knowledge and background to write a rationale to derive the final answer. Secondly, human-annotated rationales often contain redundant information that may not be necessary to infer the correct answer [29]. Finally, the annotated rationale may lack essential information due to human annotators' restricted expertise.

To address these issues, we introduce a zero-shot prompting method that allows LLMs to generate high-quality chains of thought rationales. We achieve this by converting the input training data example X into a prompt, utilizing a straightforward template that reads as follows: "Question: $[X_q]$. Context: $[X_c]$. Options: $[X_o]$. Correct Answer: $[A]$. $[Instruct]$". Here, the $[X_q]$ slot is used to contain the input question, the $[X_c]$ slot is used to contain the input context, the $[X_o]$ slot contains the possible options, the $[A]$ slot is used to contain the correct answer that can work as a hint to guide the LLM to generate a more reliable rationale, and the $[Instruct]$ slot contains instructions that guide LLMs on how to perform the task. Note that the context may not be included for some data examples, in which case the context slot is replaced with "N/A". For the purpose of generating rationales, we use the instruction "*Please give me a detailed explanation.*" Subsequently, we feed the updated prompt to LLMs, which outputs a reasoning process for each given training data example to obtain QA-CoT data set $D_{\text{QA-CoT}}$.

QA-PCoT Sample Generation. Although using QA-CoT samples can address the limitations associated with human-annotated CoT, addressing highly complex problems remains a challenge. In the domain of Science Question Answering, many questions are exceedingly complex. To overcome this challenge and obtain appropriate teaching chain-of-thought rationale, we introduce a 3-step zero-shot prompting approach that enables Language Models to decompose complex problems into simpler subproblems that are easier to solve.

Step 1: Lecture Generation. The lecture template used to generate a lecture for a particular skill is formulated as follows: "Skill: $[S]$. QA pairs: $[X_q, A]$ ... $[Instruct]$." In this prompt, $[Instruct]$ is as follows: "*based on the problems above, please give a general lecture on the* $[S]$ *type of question in one sentence.*". The lecture generated for this skill will be used in the second step of prompting, facilitating the LLM's generation of a plan for this skill. Examples of lecture generation can be found in the supplementary material.

Step 2: Plan Generation. The template used to generate a plan for a specific skill based on a lecture is formulated as follows: "Skill: $[S]$. Lecture: $[L]$. QA pairs: $[X_q, A]$ ... $[Instruct]$.". In this prompt, $[Instruct]$ is written as follows: "*Based on the lecture above and these problems, let's understand these problems and devise a general and brief plan step by step to solve these problems (begin with 1, 2, 3...)*". The plan generated using this template will be utilized in the third stage of prompting to assist the LLM in constructing a plan-based chain-of-thought rationale for each training example. Examples of plan generation can be found in the supplementary material.

Step 3: Rationale Generation. The lecture and plan generated by the first two prompts are used to generate a plan-based chain-of-thought rationale for each training example. The rationale generation template is formulated as follows: "Skill: $[S]$. Lecture: $[L]$. Plan: $[P]$. QA pairs: $[X_q, A]$ ... $[Instruct]$.". In this prompt, $[Instruct]$ is written as follows: "Based on the lecture, the plan and the problem, please carry out the plan and solve the problem step by step (begin with 1, 2, 3...)". This prompting method provides a strategy for problem-solving by leveraging the expertise of the lecture and plan, and utilizing a structured reasoning framework to solve each problem. Examples of rational generation can be found in the supplementary material.

3.3 Mixing Teaching Data

The QA-PCoT dataset is effective for teaching problem-solving skills for complex problems. However, for simpler problems, there is no need to decompose them. In contrast, the QA-CoT dataset

Algorithm 1 The Process of Mixing Teaching Data

Input: QA-CoT data set $D_{\text{QA-CoT}}$, QA-PCoT data set $D_{\text{QA-PCoT}}$
Output: T-SciQ data set $D_{\text{T-SciQ}}$
Use $D_{\text{QA-CoT}}$ and $D_{\text{QA-PCoT}}$ to train two models, respectively.
Obtain predictions $P_{\text{QA-CoT}}$ and $P_{\text{QA-PCoT}}$ of the two models on the validation dataset.
$D_{\text{T-SciQ}} \leftarrow \emptyset$
for each skill in validation set **do**
    $D_{\text{T-SciQ}} \leftarrow F(P_{\text{QA-CoT}}, P_{\text{QA-PCoT}})$
end for
procedure F($P_{\text{QA-CoT}}$, $P_{\text{QA-PCoT}}$)
    Count prediction errors $N_{\text{QA-CoT}}$ and $N_{\text{QA-PCoT}}$
    if $N_{\text{QA-PCoT}} > N_{\text{QA-CoT}}$ **then**
        Return $D_{\text{QA-CoT}}$
    else
        Return $D_{\text{QA-PCoT}}$
    end if
end procedure

can be used to teach problem-solving skills for simple problems. To leverage the strengths of both datasets, QA-CoT dataset and QA-PCoT dataset, we propose a data mixture strategy for obtaining a new dataset, T-SciQ teaching dataset, by selecting rationales from PCoT or CoT for data samples of a particular skill. To accomplish this, we utilize the validation set to determine whether PCoT or CoT is more appropriate for data samples of a given skill. For data samples of a particular skill, if the number of errors for QA-PCoT validation samples is lower than that of QA-CoT validation samples, we select PCoT rationale as the teaching rationale. Otherwise, we select CoT rationale for the data samples of this skill. The entire process is illustrated in Algorithm 1. The T-SciQ teaching samples are used to perform fine-tuning of the small student model, resulting in our final SciStudent model.

3.4 Fine-Tuning

Our teaching is based on the Multimodal-CoT [42] two-stage fine-tuning framework, which consists of two stages: rationale generation teaching and answer inference teaching. Both stages use the same model architecture but differ in their input X and output Y. During inference, the model trained in the first stage generates rationales for the test sets based on the input X, which are subsequently utilized in the second stage to infer answers.

Rationale Generation Teaching. In this stage, we define the input as $X = \{X_{\text{la}}^1, X_{\text{v}}\}$, where X_{la}^1 represents the language input and X_{v} represents the vision input. A concrete example of X could be the concatenation of a question, context, and answer options. The rationale generation model $F_r(X)$ can be trained to predict the target R. This rationale may either be CoT rationale from QA-CoT sample or generated PCoT rationale from QA-PCoT sample. Formally, the probability of generating rationale R can be formulated as follows:

$$p(R|X_{\text{la}^1}, X_{\text{v}}) = \prod_{i=1}^{N} p_{\theta_r}\left(R_i \mid X_{\text{la}}^1, X_{\text{v}}, R_{<i}\right), \tag{1}$$

where θ_r represents learnable parameters in the rationale generation teaching stage.

Answer Inference Teaching. In the second stage, we construct the language input X_{la}^2 by appending the teaching rationale R to the original language input X_{la}^1. The new input X' is then fed to the answer inference model to infer the final answer $A = F_a(X')$, where $X' = \{X_{\text{la}}^2, X_{\text{v}}\}$. Formally, the probability of generating answer A can be formulated as follows:

$$p(A|X_{\text{la}^2}, X_{\text{v}}) = \prod_{i=1}^{N} p_{\theta_a}\left(A_i \mid X_{\text{la}}^2, X_{\text{v}}, R_{<i}\right), \tag{2}$$

where θ_a represents learnable parameters in the answer inference teaching stage. Note that rationale generation and answer inference share the same model but differ in the input and output.

Model Architecture. Our model architecture follows the model architecture of Multimodal-CoT, which employs a Transformer model [32] for encoding language and a vision Transformer for encoding visual information. The gated fusion mechanism, proposed in [18, 38], is used to effectively integrate the language and vision representations. Finally, a Transformer decoder is used to generate the target output.

4 EXPERIMENT

4.1 Experimental Setup

Dataset. We evaluate our proposed method on the **ScienceQA** [22] dataset, a multimodal multiple-choice science question dataset comprising 21,208 examples. ScienceQA encompasses a wide range of topics across three distinct subjects: natural science, social science, and language science. The dataset comprises 26 topics, 127 categories, and 379 skills that are relevant to these three subjects. We employ the official split provided by ScienceQA, which divides the dataset into training, validation, and test sets with a ratio of 60:20:20. The training, validation, and test sets contain 12,726, 4,241, and 4,241 examples, respectively. The dataset also includes annotations for each data example, which consists of grounded lectures and detailed explanations. In this work, we extract our training signals from large language models instead of using ground truth signals annotated by humans.

Baselines. We provide a comparison of our proposed method with two types of baseline methods: fine-tuned baselines and in-context learning baselines.

We include several early visual question-answering models in the fine-tuned baselines, including `MCAN` [39], `Top-Down` [1], `BAN` [15], `DFAF` [8]. These VQA baselines use the question, context, and answer choices as textual input and the image as the visual input. They predict a score distribution over the answer candidates using a linear classifier. In addition, we include pre-trained text-to-text and multimodal models such as `ViLT` [16], `Patch-TRM` [25], and `VisualBERT` [19], `UnifiedQA`[13], `MM-COT` [42]. These methods use large-scale pre-trained models as backbone models and incorporate additional modules to handle multimodal signals if necessary. We also include recent LLM-based multimodal fine-tuned baselines such as `LLaMa-Adapter` [40] and `LLaVA` [21]. LLM-based multimodal fine-tuned models use strong open-access LLMs such as `LLaMa` [31] as backbone models and incorporate a vision encoder to model visual information.

Table 1: Main results (%) on the test set of ScienceQA. There are totally 8 classes of questions, namely natural science (NAT), social science (SOC), language science (LAN), text context (TXT), image context (IMG), no context (NO), grades 1-6 (G1-6), and grades 7-12 (G7-12). The best results are boldfaced. The improvements are shown in blue.

Model	Size	NAT	SOC	LAN	TXT	IMG	NO	G1-6	G7-12	Avg
Human	-	90.23	84.97	87.48	89.60	87.50	88.10	91.59	82.42	88.40
MCAN [39] (2019)	95M	56.08	46.23	58.09	59.43	51.17	55.40	51.65	59.72	54.54
Top-Down [1] (2018)	70M	59.50	54.33	61.82	62.90	54.88	59.79	57.27	62.16	59.02
BAN [15] (2018)	112M	60.88	46.57	66.64	62.61	52.60	65.51	56.83	63.94	59.37
DFAF [8] (2019)	74M	64.03	48.82	63.55	65.88	54.49	64.11	57.12	67.17	60.72
ViLT [16] (2021)	113M	60.48	63.89	60.27	63.20	61.38	57.00	60.72	61.90	61.14
Patch-TRM [25] (2021)	90M	65.19	46.79	65.55	66.96	55.28	64.95	58.04	67.50	61.42
VisualBERT [19] (2019)	111M	59.33	69.18	61.18	62.71	62.17	58.54	62.96	59.92	61.87
UnifiedQA$_{\text{Base}}$ [13] (2020)	223M	68.16	69.18	74.91	63.78	61.38	77.84	72.98	65.00	70.12
LLaMa-Adapter [40] (2023)	>7B	84.37	88.30	84.36	83.72	80.32	86.90	85.83	84.05	85.19
LLaVA [21] (2023)	>7B	90.36	95.95	88.00	89.49	88.00	90.66	90.93	90.90	90.92
GPT-3.5 [4] (2020)	>175B	74.64	69.74	76.00	74.44	67.28	77.42	76.80	68.89	73.97
GPT-3.5 w/ CoT [22] (2022)	>175B	75.44	70.87	78.09	74.68	67.43	79.93	78.23	69.68	75.17
ChatGPT w/ CoT [23] (2023)	>175B	78.82	70.98	83.18	77.37	67.92	86.13	80.72	74.03	78.31
GPT-4 w/ CoT [23] (2023)	>175B	84.06	73.45	87.36	81.87	70.75	90.73	84.69	79.10	82.69
Chameleon [23] (2023)	>175B	89.83	74.13	89.82	88.27	77.64	92.13	88.03	83.72	86.54
UnifiedQA-CoT$_{\text{Base}}$ [22] (2022)	223M	71.00	76.04	78.91	66.42	66.53	81.81	77.06	68.82	74.11
UnifiedQA-T-SciQ$_{\text{Base}}$ (Ours)	223M	76.56	88.99	80.45	72.90	73.84	83.47	81.09	75.19	79.41
Improvement	-	+5.56	+12.95	+1.54	+6.48	+7.31	+1.66	+4.03	+6.37	+5.30
Mutimodal-CoT$_{\text{Base}}$ [42] (2023)	223M	87.52	77.17	85.82	87.88	82.90	86.83	84.65	85.37	84.91
Mutimodal-T-SciQ$_{\text{Base}}$ (Ours)	223M	91.52	91.45	92.45	91.94	90.33	92.26	92.11	91.10	91.75
Improvement	-	+4.00	+14.28	+6.63	+4.06	+7.43	+5.43	+7.46	+5.73	+6.84
Mutimodal-CoT$_{\text{Large}}$ [42] (2023)	738M	95.91	82.00	90.82	95.26	88.80	92.89	92.44	90.31	91.68
Mutimodal-T-SciQ$_{\text{Large}}$ (Ours)	738M	**96.89**	**95.16**	**95.55**	**96.53**	**94.70**	**96.79**	**96.44**	**95.72**	**96.18**
Improvement	-	+0.98	+13.16	+4.73	+1.27	+5.90	+3.90	+4.00	+5.41	+4.50

Table 2: Ablation study of the impact of different signals provided by LLMs across all topics.

Model	NAT	SOC	LAN	TXT	IMG	NO	G1-6	G7-12	Avg
Multimodal-T-SciQ$_{\text{Base}}$ (Mixing)	91.52	91.45	92.45	91.94	90.33	92.26	92.11	91.10	91.75
Multimodal-T-SciQ$_{\text{Base}}$ only w/ QA-CoT	87.83	83.46	84.27	87.59	84.38	86.06	85.10	87.61	85.99
Multimodal-T-SciQ$_{\text{Base}}$ only w/ QA-PCoT	87.30	89.43	90.45	88.27	86.51	89.69	89.06	87.67	88.56
Mutimodal-CoT$_{\text{Base}}$	87.52	77.17	85.82	87.88	82.90	86.83	84.65	85.37	84.91
Multimodal-T-SciQ$_{\text{Large}}$ (Mixing)	96.89	95.16	95.55	96.53	94.70	96.79	96.44	95.72	96.18
Multimodal-T-SciQ$_{\text{Large}}$ only w/ QA-CoT	94.58	91.56	92.64	93.89	91.22	94.29	93.43	93.47	93.44
Multimodal-T-SciQ$_{\text{Large}}$ only w/ QA-PCoT	94.01	93.70	94.64	93.30	90.98	96.31	94.75	92.95	94.11
Mutimodal-CoT$_{\text{Large}}$	95.91	82.00	90.82	95.26	88.80	92.89	92.44	90.31	91.68

For in-context learning baselines, we compare to the widely-used method, a chain of thought (COT) prompting [36], where each in-context demonstration example comprises a question text, options, correct answer text, and reasoning process composed of annotated lecture and detailed explanation. Specifically, we compare our model's performance to COT over different API-based OpenAI LLMs, such as GPT-3.5 of the text-davinci-002 version (`GPT-3.5 w/ COT`), ChatGPT (`ChatGPT w/ COT`), GPT-4 (`GPT-4 w/ COT`) and `Chameleon` [23]. Additionally, we compared our model's performance with the standard few-shot prompting approach using GPT-3.5 of the text-davinci-002 version (`GPT-3.5`).

Evaluation Metrics. As ScienceQA is a benchmark for multiple-choice question answering, the *accuracy* of the answer is evaluated by comparing the ground truth option with the final prediction generated by the evaluated model.

Implementation Details. By default, we utilize the GPT-3.5 text-davinci-003 version as the teacher model for our approach unless otherwise specified. To validate the generalizability of our method, we incorporate three distinct student models, namely UnifiedQA$_{\text{Base}}$ w/ CoT [22], Mutimodal-CoT$_{\text{Base}}$ [22], and Mutimodal-CoT$_{\text{Large}}$ [42]. These models are chosen due to their strong performances achieved by fine-tuning with annotated reasoning signals.

Table 3: Accuracy (%) of Mutimodal-T-SciQ$_{\text{Base}}$ using different vision features.

Method	T-SciQ		
	QA-CoT	QA-PCoT	T-SciQ
Language Only	84.44	85.38	87.24
w/ CLIP	86.18	87.41	90.90
w/ DETR	85.99	88.56	91.75
w/ ResNet	86.06	87.69	91.44

To ensure fairness of comparison and effectiveness of our proposed method, we only replace the training signals generated by our approach with annotated signals while maintaining the same settings as the original paper for Mutimodal-CoT if not otherwise specified. These student models are 200× smaller than their teacher models.

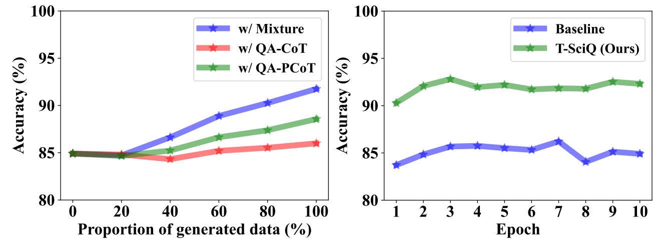
4.2 Main Results

4.2.1 T-SciQ versus Baselines. Table 1 details the performance accuracy of student models trained using the proposed T-SciQ signals compared to the fine-tuned and in-context baselines. Mutimodal-T-SciQ$_{\text{Large}}$, which is the model architecture of Mutimodal-CoT$_{\text{Large}}$ fine-tuned with planning and chain-of-thought reasoning signals from the LLM (text-davinci-003 by default), attains an accuracy of 96.18% and consistently outperforms all state-of-the-art methods by a large margin for all topics across all subjects. Specifically, Mutimodal-T-SciQ$_{\text{Large}}$ outperforms the most powerful fine-tuned baseline, Mutimodal-CoT$_{\text{Large}}$, which is trained by annotated chain-of-thought signals, by 4.5% (91.68% → 96.18%), the strongest instruction-tuning based multimodal baseline, LLaVa, by 5.26% (90.92% → 96.18%), the best GPT-4 based few-shot baseline, Chameleon, by 9.64% (86.54% → 96.18%), and human performance by 7.78% (88.40% → 96.18%). This significant improvement of our proposed method suggests that higher-quality teaching signals of planning and reasoning provided by LLMs have resulted in better planning and chain-of-thought reasoning ability in student models smaller than 1B.

4.2.2 T-SciQ with Different Base Student Models. Instead of using the model architecture of Mutimodal-CoT$_{\text{Large}}$ as the base student model, we evaluate different base student models fine-tuned with planning and chain-of-thought signals of T-SciQ: the variant UnifiedQA-T-SciQ$_{\text{Base}}$ and Mutimodal-T-SciQ$_{\text{Base}}$. The relative performance ranking between the base student model with annotated chain-of-thought reasoning signals and the one with planning and chain-of-thought reasoning signals of T-SciQ remains unchanged. Specifically, UnifiedQA-T-SciQ$_{\text{Base}}$ outperforms UnifiedQA$_{\text{Base}}$ w/ CoT by 5.3% (74.11% → 79.41%), and Mutimodal-T-SciQ$_{\text{Base}}$ outperforms Mutimodal-CoT$_{\text{Base}}$ by 6.84% (84.91% → 91.75%). T-SciQ still achieves the best performance with different base student models. These encouraging results indicate the generalizability of the effectiveness of T-SciQ.

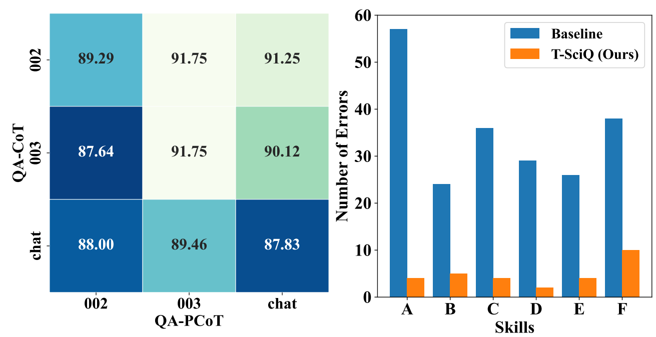
4.3 Further Analysis

4.3.1 Effect of Different Signals of T-SciQ. Our approach incorporates two distinct components for teaching signals: QA-CoT and



(a) (b)

Figure 4: Further analysis on (a) the effect of Mutimodal-T-SciQ$_{\text{Base}}$ trained with different proportion of generated data and (b) accuracy curve of the baseline Mutimodal-CoT$_{\text{Base}}$ and our Mutimodal-T-SciQ$_{\text{Base}}$ across epochs.

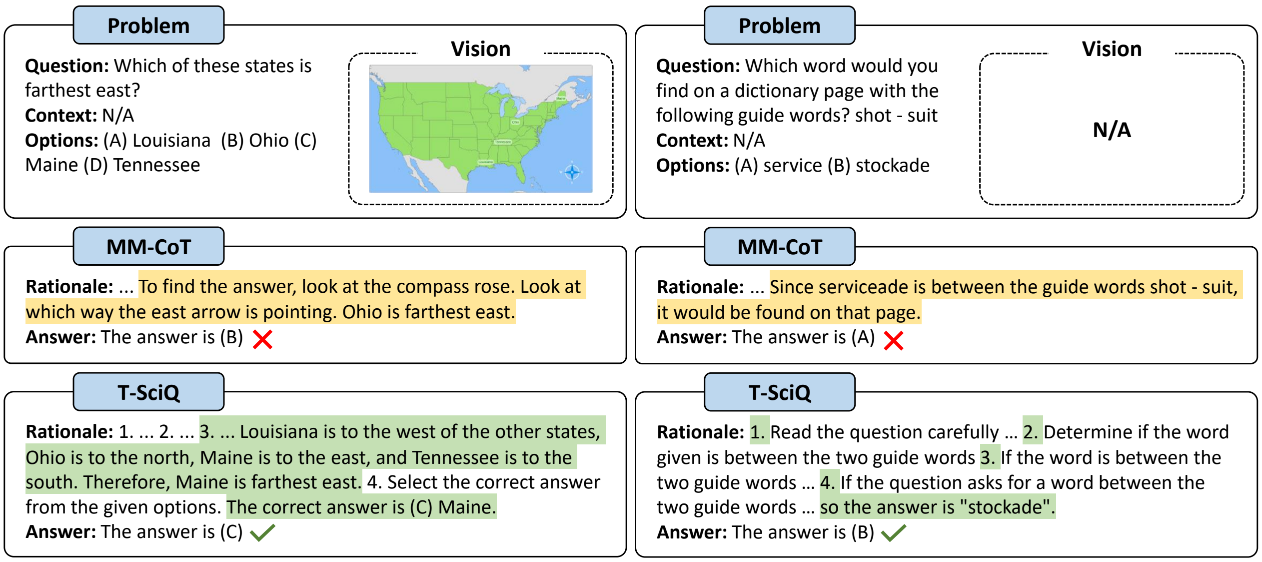


(a) (b)

Figure 5: Further analysis on (a) accuracy (%) of Mutimodal-T-SciQ$_{\text{Base}}$ with teaching signals provided by different base LLMs and (b) error analysis of prediction for specific skills.

QA-PCoT. We early show that combining these two signals (i.e., Mutimodal-T-SciQ) yields significantly better results than using only human-annotated CoT signals (i.e., Mutimodal-CoT) when teaching student models. In this section, we aim to evaluate the individual impact of each teaching signal by examining the performance of Mutimodal-T-SciQ$_{\text{Base}}$ and Mutimodal-T-SciQ$_{\text{Large}}$ when either QA-CoT or QA-PCoT signal is removed. As demonstrated in Table 2, we can observe that a significant decrease in answering accuracy when either of the teaching signals was removed. These findings indicate the effectiveness of both proposed teaching signals. This is because 1) student models, which are taught by QA-CoT signals, can incorporate a more extensive range of knowledge from the open world rather than solely relying on the knowledge of annotators and 2) student models, which are taught by QA-PCoT signals, can decompose complex problems into several simpler sub-problems.

4.3.2 Impact of Vision Features. The choice of vision features can significantly affect the performance of models for ScienceQA. Thus, we conduct an evaluation of three widely-used vision features, which are CLIP [26], DETR [3], and ResNet [9]. Both CLIP and DETR can provide patch-level features, and DETR is designed for object detection. As for ResNet features, we use ResNet-50 to derive



(a) (b)

Figure 6: Examples of MM-CoT (baseline) and the model trained with T-SciQ (ours) signals for generating rationales and predicting answers. To solve these examples, commonsense knowledge such as geographic knowledge (a) and multi-step reasoning (b) are required.

vision features. Table 3 details the results of comparing these three vision features. Our findings suggest that incorporating vision features yields superior performance than relying on language-only baselines. Notably, DETR consistently outperforms the other two features in most cases, and hence, we adopt it as the default vision feature in our main experiments.

4.3.3 Proportion of Generated Data in Training Data. To further compare the T-SciQ signals produced by LLMs and the annotated CoT signals, we experiment with manipulating the proportion of these two signals within the training data. We vary the proportion of T-SciQ signals from 0% to 100%. As demonstrated in Figure 4a, the increasing proportion of training data with T-SciQ signals increases performance.

4.3.4 Performance Change with Epoch. Figure 4b shows the performance trends of the baseline Mutimodal-CoT$_{\text{Base}}$ and our proposed Mutimodal-T-SciQ$_{\text{Base}}$ across different training epochs. Notably, our method consistently outperforms the baseline across all epochs. We adopt a two-stage training approach similar to the baseline Mutimodal-CoT$_{\text{Base}}$, where we first train the explanation generation module and then train the answer prediction. Hence, like the baseline, our method exhibits relatively higher accuracy at the initial training stages.

4.3.5 Effect of Teaching Signals Provided by Different Base LLMs. We use the GPT-3.5 model by default, specifically the text-davinci-003 version, to generate teaching signals in the main experiment. However, other powerful LLMs can also provide such signals, such as the earlier version of GPT-3.5, text-davinci-002, and the recently popular ChatGPT model. This study explores the effectiveness of a mixture of QA-CoT signals from text-davinci-002, text-davinci-003, or ChatGPT, and QA-PCoT signals from the same API-based models. We conduct this experiment using the Multimodal-T-SciQ$_{\text{Base}}$. Figure 5a shows the comparison of the performance of nine different mixture strategies. Our results show that even the worst strategy, which involves a mixture of QA-CoT signal from text-davinci-003 and QA-PCoT signal from text-davince-002, outperforms annotated CoT signal by a significant margin. It indicates that regardless of the mixture strategy used, LLMs can provide signals with more useful knowledge from the open world. Additionally, planning-based CoT can help decompose complex problems into simpler ones.

4.3.6 Error Analysis. To better understand the model's behavior trained using our proposed T-SciQ signals and to facilitate future studies, we analyze the number of errors that occur in six selected skills. Figure 5b shows the error analysis of prediction for specific skills. In paticular, skills A, B, C, D, E and F represent "Using guide words", "Comparing properties of objects", "Reading a map: cardinal directions", "Identifying oceans and continents", "How is temperature related to thermal energy?", and "Identifying the Thirteen Colonies", respectively. We can observe that training with T-SciQ signals can significantly reduce the number of errors. Examples of skills such as "Using guide words", "Comparing properties of objects", "Identifying oceans and continents", and "How is temperature related to thermal energy?" require multi-step complex reasoning that can be provided by the model trained using T-SciQ signals. On the other hand, examples of skills such as "Reading a map: cardinal directions" and "Identifying the Thirteen Colonies" require a great deal of common sense and factual knowledge from the open world, which can also be generated by the model trained using T-SciQ signals.

4.3.7 Case Study. The case study of the comparison of T-SciQ and Multimodal-CoT on the ScienceQA benchmark is illustrated in Figure 6. In Figure 6a, cases that require geographic knowledge from the open world are displayed. Human-annotated CoT rationale may lack essential information from the open world. On the other hand, T-SciQ samples generated by LLMs include knowledge from the open world. Figure 6b shows the case in the absence of image input requiring multi-step reasoning. The output of Multimodal-CoT is incorrect since the word "service" is not between the guide words of "shot" and "suit". Our model can provide the correct answer by decomposing the final question into several intermediate questions to answer. These cases demonstrate that the model trained with the proposed T-SciQ signals is well-suited to handle science question answering problems that require open knowledge and decomposition.

5 CONCLUSION

This paper introduces a new approach named T-SciQ that utilizes large language models' chain-of-thought (CoT) reasoning capabilities to teach small multimodal models for complex science question answering tasks. Our zero-shot prompting method generates QA-CoT samples as teaching data. We also present a 3-step zero-shot prompting approach using planning-based CoT for highly complex problems. Furthermore, our data mixture strategy combines CoT and planning-based CoT to create a new T-SciQ teaching dataset. Empirical evaluation on ScienceQA shows significant improvement over previous state-of-the-art baselines. Our method overcomes the limitations of human-annotated CoT, providing a promising approach for complex science question answering. Future work includes exploring extensive LLMs and parameter-efficient fine-tuning with LLM teachers.

REFERENCES

[1] 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 *Proceedings of the IEEE conference on computer vision and pattern recognition*. 6077–6086.

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

[3] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. 2020. End-to-End Object Detection with Transformers. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I*. 213–229.

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

[5] Wenhu Chen, Xueguang Ma, Xinyi Wang, and William W Cohen. 2022. Program of thoughts prompting: Disentangling computation from reasoning for numerical reasoning tasks. *arXiv preprint arXiv:2211.12588* (2022).

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

[7] Yao Fu, Hao Peng, Ashish Sabharwal, Peter Clark, and Tushar Khot. 2022. Complexity-based prompting for multi-step reasoning. *arXiv preprint arXiv:2210.00720* (2022).

[8] Peng Gao, Zhengkai Jiang, Haoxuan You, Pan Lu, Steven CH Hoi, Xiaogang Wang, and Hongsheng Li. 2019. Dynamic fusion with intra-and inter-modality attention flow for visual question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 6639–6648.

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*. IEEE Computer Society, 770–778. https://doi.org/10.1109/CVPR.2016.90

[10] Namgyu Ho, Laura Schmid, and Se-Young Yun. 2022. Large Language Models Are Reasoning Teachers. *arXiv preprint arXiv:2212.10071* (2022).

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

[12] Aniruddha Kembhavi, Minjoon Seo, Dustin Schwenk, Jonghyun Choi, Ali Farhadi, and Hannaneh Hajishirzi. 2017. 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)*. 4999–5007.

[13] 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. *arXiv preprint arXiv:2005.00700* (2020).

[14] Tushar Khot, Harsh Trivedi, Matthew Finlayson, Yao Fu, Kyle Richardson, Peter Clark, and Ashish Sabharwal. 2022. Decomposed prompting: A modular approach for solving complex tasks. *arXiv preprint arXiv:2210.02406* (2022).

[15] Jin-Hwa Kim, Jaehyun Jun, and Byoung-Tak Zhang. 2018. Bilinear attention networks. *Advances in neural information processing systems* 31 (2018).

[16] Wonjae Kim, Bokyung Son, and Ildoo Kim. 2021. Vilt: Vision-and-language transformer without convolution or region supervision. In *International Conference on Machine Learning*. PMLR, 5583–5594.

[17] Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *arXiv preprint arXiv:2205.11916* (2022).

[18] Bei Li, Chuanhao Lv, Zefan Zhou, Tao Zhou, Tong Xiao, Anxiang Ma, and Jingbo Zhu. 2022. On Vision Features in Multimodal Machine Translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 6327–6337.

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

[20] Yifei Li, Zeqi Lin, Shizhuo Zhang, Qiang Fu, Bei Chen, Jian-Guang Lou, and Weizhu Chen. 2022. On the advance of making language models better reasoners. *arXiv preprint arXiv:2206.02336* (2022).

[21] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *arXiv preprint arXiv:2304.08485* (2023).

[22] Pan Lu, Swaroop Mishra, Tanglin 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. *Advances in Neural Information Processing Systems* 35 (2022), 2507–2521.

[23] Pan Lu, Baolin Peng, Hao Cheng, Michel Galley, Kai-Wei Chang, Ying Nian Wu, Song-Chun Zhu, and Jianfeng Gao. 2023. Chameleon: Plug-and-Play Compositional Reasoning with Large Language Models. *arXiv preprint arXiv:2304.09842* (2023).

[24] Pan Lu, Liang Qiu, Kai-Wei Chang, Ying Nian Wu, Song-Chun Zhu, Tanmay Rajpurohit, Peter Clark, and Ashwin Kalyan. 2022. Dynamic prompt learning via policy gradient for semi-structured mathematical reasoning. *arXiv preprint arXiv:2209.14610* (2022).

[25] Pan Lu, Liang Qiu, Jiaqi Chen, Tony Xia, Yizhou Zhao, Wei Zhang, Zhou Yu, Xiaodan Liang, and Song-Chun Zhu. 2021. Iconqa: A new benchmark for abstract diagram understanding and visual language reasoning. *arXiv preprint arXiv:2110.13214* (2021).

[26] 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 *International Conference on Machine Learning*. PMLR, 8748–8763.

[27] Ohad Rubin, Jonathan Herzig, and Jonathan Berant. 2021. Learning to retrieve prompts for in-context learning. *arXiv preprint arXiv:2112.08633* (2021).

[28] Shailaja Keyur Sampat, Yezhou Yang, and Chitta Baral. 2020. Visuo-Lingustic Question Answering (VLQA) Challenge. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings (EMNLP)*. 4606–4616.

[29] Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed Chi, Nathanael Schärli, and Denny Zhou. 2023. Large language models can be easily distracted by irrelevant context. *arXiv preprint arXiv:2302.00093* (2023).

[30] Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, et al. 2022. Lamda: Language models for dialog applications. *arXiv preprint arXiv:2201.08239* (2022).

[31] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971* (2023).

[32] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach,

Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett (Eds.). 5998–6008.
[33] Boshi Wang, Xiang Deng, and Huan Sun. 2022. Iteratively prompt pre-trained language models for chain of thought. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. 2714–2730.
[34] Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, and Denny Zhou. 2022. Rationale-augmented ensembles in language models. *arXiv preprint arXiv:2207.00747* (2022).
[35] Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, and Denny Zhou. 2022. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171* (2022).
[36] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Chi, Quoc Le, and Denny Zhou. 2022. Chain of Thought Prompting Elicits Reasoning in Large Language Models. *ArXiv preprint* abs/2201.11903 (2022).
[37] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Chi, Quoc Le, and Denny Zhou. 2022. Chain of thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903* (2022).
[38] Zhiyong Wu, Lingpeng Kong, Wei Bi, Xiang Li, and Ben Kao. 2021. Good for Misconceived Reasons: An Empirical Revisiting on the Need for Visual Context in Multimodal Machine Translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Association for Computational Linguistics, Online, 6153–6166. https://doi.org/10.18653/v1/2021.acl-long.480
[39] Zhou Yu, Jun Yu, Yuhao Cui, Dacheng Tao, and Qi Tian. 2019. Deep modular co-attention networks for visual question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 6281–6290.
[40] Renrui Zhang, Jiaming Han, Aojun Zhou, Xiangfei Hu, Shilin Yan, Pan Lu, Hongsheng Li, Peng Gao, and Yu Qiao. 2023. Llama-adapter: Efficient fine-tuning of language models with zero-init attention. *arXiv preprint arXiv:2303.16199* (2023).
[41] Zhuosheng Zhang, Aston Zhang, Mu Li, and Alex Smola. 2022. Automatic chain of thought prompting in large language models. *arXiv preprint arXiv:2210.03493* (2022).
[42] Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. 2023. Multimodal chain-of-thought reasoning in language models. *arXiv preprint arXiv:2302.00923* (2023).
[43] Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Olivier Bousquet, Quoc Le, and Ed Chi. 2022. Least-to-most prompting enables complex reasoning in large language models. *arXiv preprint arXiv:2205.10625* (2022).

A MORE DETAILED ANALYSIS

In addition to exploring the effects of different visual features, we also tested the performance of different backbones. Table 4 shows the results of UnifiedQA and FLAN-T5 on our generated datasets. When using various backbone networks, the model trained by QA-CoT data is the worst among the three generated data types. However, it is also better than the manually annotated data, which indicates that the manually annotated data has certain limitations, such as redundant information, single style, etc. Furthermore, when using the Mixture dataset of QA-CoT and QA-PCoT, the fact that all four backbones achieved the best performances illustrates the effectiveness and generality of our strategy.

Table 4: Accuracy (%) of using different backbones.

Method	Size	Annotated CoT	T-SciQ		
			QA-CoT	QA-PCoT	Mixture
UnifiedQA$_{\text{Base}}$	223M	84.91	85.99	88.56	91.75
UnifiedQA$_{\text{Large}}$	738M	91.68	93.44	93.54	96.18
FLAN-T5$_{\text{Base}}$	248M	85.85	86.87	89.04	92.33
FLAN-T5$_{\text{Large}}$	783M	93.02	93.54	95.68	**96.49**

B MORE CASES ANALYSIS

To investigate the impact of different teaching signals, we conducted an evaluation of model predictions trained on various types of data, including both generated teaching data and manually annotated data. In particular, we compared the test examples produced by the model trained with manually annotated data (MM-CoT) and the QA-CoT teaching data (T-SciQ (QA-CoT)) in Figure 7. We also displayed prediction cases of models trained on manually annotated data (MM-CoT) and QA-PCoT teaching data (T-SciQ (QA-PCoT)) in Figure 8. Furthermore, we compared the prediction cases of models trained using our proposed teaching data, QA-CoT and QA-PCoT, as depicted in Figure 9 and Figure 10. From the figures, it is evident that the model trained with QA-PCoT data performs better in solving multi-step reasoning problems, while the one trained with QA-CoT is more proficient at solving straightforward problems.

C DATA GENERATION PROCESS OF PCOT

In this section, we several show cases of how to generate PCoT teaching signals. To obtain appropriate planning-based chain-of-thought rationale, we introduce a 3-step zero-shot prompting approach that enables Language Models to decompose complex problems into simpler subproblems that are easier to solve.

The lecture template used to generate a lecture for a particular skill is formulated as follows: "Skill: $[S]$. QA pairs: $[X_q, A]$ ... $[Instruct]$." In this prompt, $[Instruct]$ is as follows: "*based on the problems above, please give a general lecture on the* $[S]$ *type of question in one sentence.*". The lecture generated for this skill will be used in the second step of prompting, facilitating the LLM's generation of a plan for this skill. Examples of lecture generation are shown in Table 6.

The template used to generate a plan for a specific skill based on a lecture is formulated as follows: "Skill: $[S]$. Lecture: $[L]$. QA pairs: $[X_q, A]$ ... $[Instruct]$.". In this prompt, $[Instruct]$ is written as follows: "*Based on the lecture above and these problems, let's understand these problems and devise a general and brief plan step by step to solve these problems (begin with 1, 2, 3...)*". The plan generated using this template will be utilized in the third stage of prompting to assist the LLM in constructing a plan-based chain-of-thought rationale for each training example. Examples are shown in Table 7.

The lecture and plan generated by the first two prompts are used to generate a plan-based chain-of-thought rationale for each training example. The rationale generation template is formulated as follows: "Skill: $[S]$. Lecture: $[L]$. Plan: $[P]$. QA pairs: $[X_q, A]$ ... $[Instruct]$.". In this prompt, $[Instruct]$ is written as follows: "Based on the lecture, the plan and the problem, please carry out the plan and solve the problem step by step (begin with 1, 2, 3...)". This prompting method provides a strategy for problem-solving by leveraging the expertise of the lecture and plan, and utilizing a structured reasoning framework to solve each problem. Examples are shown in Table 8.

Problem

Question: Which of these states is farthest east?
Context: N/A
Options: (A) Nebraska (B) Louisiana (C) Idaho (D) Arizona
Answer: The answer is (B)

Vision

MM-CoT

Rationale: Maps have four cardinal directions, or main directions. Those directions are north, south, east, and west. A compass rose is a set of arrows that point to the cardinal directions. A compass rose usually shows only the first letter of each cardinal direction. The north arrow points to the North Pole. On most maps, north is at the top of the map. To find the answer, look at the compass rose. Look at which way the east arrow is pointing. Idaho is farthest east..
Answer: The answer is (C) ✗

T-SciQ (QA-CoT)

Rationale: Louisiana is the farthest east of the four states listed. Nebraska is located in the Midwestern United States, Idaho is located in the Northwestern United States, and Arizona is located in the Southwestern United States. Louisiana is located in the Southeastern United States, making it the farthest east of the four states.
Answer: The answer is (B) ✓

(a) There is some redundancy in manually annotating data.

Problem

Question: When was Benjamin Franklin born?
Context: N/A
Options: (A) 1856 (B) 1656 (C) 1776 (D) 1706
Answer: The answer is (D)

Vision

N/A

MM-CoT

Rationale: Benjamin Franklin was the the first of Independence. 1776.., 17he was have been born in 17. Declaration was56 gave still short. was notve been bornly years old. 1776.., 17 17 year was 1776.
Answer: The answer is (C) ✗

T-SciQ (QA-CoT)

Rationale: Benjamin Franklin was born on January 17, 1706 in Boston, Massachusetts. He was one of the Founding Fathers of the United States and a leading author, printer, political theorist, politician, freemason, postmaster, scientist, inventor, civic activist, statesman, and diplomat. He was a major figure in the American Enlightenment and the history of the United States. He is best known for his discoveries and theories regarding electricity, as well as his famous experiments with a kite and a key. He also wrote the famous Poor Richard's Almanack and was a major contributor to the Declaration of Independence and the Constitution of the United States. Therefore, the correct answer is (D) 1706.
Answer: The answer is (D) ✓

(b) The QA-CoT data can be used to solve problems with commonsense knowledge.

Figure 7: Examples of MM-CoT (baseline) and T-SciQ (QA-CoT) (ours) for generating rationales and predicting answers.

Problem

Question: Which word would you find on a dictionary page with the following guide words? shot - suit
Context: N/A
Options: (A) service (B) stockade
Answer: The answer is (B)

Vision

N/A

MM-CoT

Rationale: Guide words appear on each page of a dictionary. They tell you the first word and last word on the page. The other words on the page come between the guide words in alphabetical order. To put words in alphabetical order, put them in order by their first letters. If the first letters are the same, look at the second letters. If the second letters are the same, look at the third letters, and so on. If one word is shorter, and there are no more letters to compare, then the shorter word comes first in alphabetical order. For example, be comes before bed. Put the words in alphabetical order. Since serviceade is between the guide words shot - suit, it would be found on that page.
Answer: The answer is (A) ✗

T-SciQ (QA-PCoT)

Rationale: 1. Read the question carefully and identify the guide words given. The guide words given are "shot" and "suit". 2. Determine if the word given is between the two guide words. The word given is "stockade". 3. If the word is between the two guide words, answer "yes". If the word is not between the two guide words, answer "no". The word "stockade" is between the two guide words, so the answer is "yes". 4. If the question asks for a word between the two guide words, look for a word that fits the context of the question. The context of the question is not given, so the answer is "stockade".
Answer: The answer is (B) ✓

(a) T-SciQ (QA-PCoT) can solve problems by executing the plan in multiple steps.

Problem

Question: Based on the map, which of the following areas did the Mongol Empire control?
Context: The Mongol Empire controlled most of Asia and some parts of Eastern Europe from around 1210 to 1375. Look at the map of the Mongol Empire. Then answer the question below.
Options: (A) South Asia (B) Southeast Asia (C) the Middle East
Answer: The answer is (C)

Vision

MM-CoT

Rationale: Look at at the map. Mon indicate the map show that areas of the area. of areas on that area the by the Mongol Empire. The Mon shows that the Mongol Empire controlled most of of Asia Asia, Southeast Asia, and Middle East, and Africaberia. ThenThe Mongol Empire controlled not control all Asia or any Asia.
Answer: The answer is (B) ✗

T-SciQ (QA-PCoT)

Rationale: 1. Read the lecture and understand the context of the question. 2. Look at the map of the Mongol Empire. 3. Identify the areas that the Mongol Empire controlled. 4. Compare the areas identified in the map with the options given in the question. 5. Select the option that matches the areas identified in the map. Answer: The Mongol Empire controlled most of Asia and some parts of Eastern Europe from around 1210 to 1375. Looking at the map of the Mongol Empire, it can be seen that the Mongol Empire controlled the Middle East, which is option (C). Therefore, the correct answer is (C) the Middle East.
Answer: The answer is (C) ✓

(b) The QA-PCoT data can be used to solve problems with commonsense knowledge.

Figure 8: Examples of MM-CoT (baseline) and T-SciQ (QA-PCoT) (ours) for generating rationales and predicting answers.

Problem

Question: Which of the following contains a vague pronoun reference?
Context: N/A
Options: (A) Abby was surprised when the article said that at least one-third of American adults get less than seven hours of sleep each night. (B) Abby was surprised when it said that at least one-third of American adults get less than seven hours of sleep each night.
Answer: The answer is (B)

Vision

N/A

T-SciQ(CoT)

Rationale: Option (A) contains a vague pronoun reference because the pronoun "it" is not clearly defined. It is not clear who "it" is referring to, this sentence. Option (A) does not contain a vague pronoun reference because the pro clearly defined.
Answer: The answer is (A)

T-SciQ (QA-PCoT)

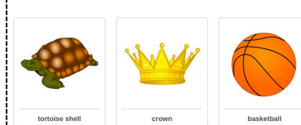
Rationale: Vague pronoun references occur when a pronoun is used to refer to a noun without providing enough context to determine which noun the pronoun is referring to. 1. Read the question and the context carefully. 2. Identify the pronoun in the sentence. The pronoun in the sentence is "it". 3. Determine the noun that the pronoun is referring to. The pronoun "it" is referring to the article. 4. Compare the two options and decide which one contains a vague pronoun reference. Option B contains a vague pronoun reference because it pronoun "it" is not clear enough to determine which noun the is referring to. Option A does not contain a vague pronoun reference because the pronoun "the article" is specific enough to determine which noun it is referring to. Therefore, the correct answer is Option B
Answer: The answer is (B)

(a) T-SciQ (QA-PCoT) can solve problems by executing the plan in multiple steps.

Problem

Question: Which property do these three objects have in common?
Context: Select the best answer.
Options: (A) shiny (B) slippery (C) opaque
Answer: The answer is (C)

Vision



T-SciQ(CoT)

Rationale: The three objects have the property of being shiny in common. Opaque objects are those that do not allow light to pass through them. Examples means that they light hits an opaque object, it is not able to pass through it. can reflectivereflected back of it. This of opaque objects include glass, metal, and, and plastic. All
Answer: The answer is (A)

T-SciQ (QA-PCoT)

Rationale: Comparing properties of objects involves selecting the best or better answer from a given set of options that have a common property. 1. Read the question and context carefully. 2. Identify the common property among the given options. 3. The common property among the given options is that they are all physical properties. objects. 4. Select the best answer that has the common property. 5. The best answer that has the common property is (C) opaque. 6. Check your answer to make sure it is correct. 7. Yes, (C) opaque is the correct answer.
Answer: The answer is (C)

(b) T-SciQ (QA-PCoT) can solve problems by executing the plan in multiple steps.

Figure 9: Examples of T-SciQ (QA-CoT) (ours) and T-SciQ (QA-PCoT) (ours) for generating rationales and predicting answers.

Problem

Question: What is the probability that a muskmelon plant produced by this cross will be homozygous recessive for the fruit taste gene?
Context: In a group of muskmelon plants, some individuals have sour fruit and others have sweet fruit. In this group, the gene for the fruit taste trait has two alleles. The allele for sweet fruit (f) is recessive to the allele for sour fruit (F). This Punnett square shows a cross between two muskmelon plants.
Options: (A) 2/4 (B) 1/4 (C) 0/4 (D) 3/4 (E) 4/4
Answer: The answer is (C)

Vision

	F	F
F	FF	FF
F	FF	FF

T-SciQ(QA-PCoT)

Rationale: Punnett squares can be used to calculate the probability of different offspring types based on the alleles of the parents. 1. Read the context of the problem carefully and identify the alleles of the parents. The alleles of the parents are F (sour fruit) and f (sweet fruit). 2. Draw a Punnett square to represent the cross between the two parents. F f F FF Ff f Ff ff 3. Count the number of offspring with each genotype and calculate the probability of each genotype. FF: 0 offspring, probability 0/4 Ff: 2 offspring, probability 2/4 ff: 2 offspring, probability 2/4 4. Compare the probabilities of the different genotypes to determine the probability of the desired genotype. The desired genotype is homozygous recessive for the fruit taste gene (ff). The probability of this genotype is 2/4."
Answer: The answer is (A) ✗

T-SciQ (QA-CoT)

Rationale: Since the allele for sweet fruit (f) is recessive, the only way for a muskmelon plant to be homozygous recessive for the fruit taste gene is if both alleles are f. However, this cross will not produce any muskmelon plants with two f alleles, so the probability that a muskmelon plant produced by this cross will be homozygous recessive for the fruit taste gene is 0/4, or 0%.
Answer: The answer is (C) ✓

(a) T-SciQ (QA-PCoT) may generate error rationale when vision features are complicated.

Problem

Question: Which is this organism's common name?
Context: This organism is a goldfish. It is also called Carassius auratus.
Options: (A) Carassius auratus (B) goldfish
Answer: The answer is (B)

Vision

T-SciQ(QA-PCoT)

Rationale: In this type of question, the goal is to identify the common name of an organism given its scientific name, or vice versa. 1. Read the question and context carefully. 2. Identify the type of question. This is a question that requires us to identify the common name of an organism given its scientific name. 3. Identify the organism's common name or scientific name. The organism's scientific name is Carassius auratus. 4. Select the correct answer from the given options. The correct answer is (A) goldfish.
Answer: The answer is (A) ✗

T-SciQ (QA-CoT)

Rationale: The correct answer is (B) goldfish. This scientific name of goldfish is Carassius auratus, but its common name is gold goldfish. Common scientific name is used to identify a species, while the common name is used to refer to organism in everyday language.
Answer: The answer is (B) ✓

(b) T-SciQ (QA-CoT) is better than T-SciQ (QA-PCoT) on some simple problems.

Figure 10: Examples of T-SciQ (QA-CoT) (ours) and T-SciQ (QA-PCoT) (ours) for generating rationales and predicting answers.

Table 5: Data generation process of CoT prompting.

Question: Which of these states is farthest north?
Context: N/A
Options: (A) West Virginia (B) Louisiana (C) Arizona (D) Oklahoma
Correct Answer: (A) West Virginia

Please give a detailed explanation.

Solution: West Virginia is the farthest north of the four states listed. West Virginia is located in the Appalachian region of the United States, which is in the northeastern part of the country. Louisiana, Arizona, and Oklahoma are all located in the southern and southwestern parts of the United States. West Virginia is the northernmost of the four states, making it the farthest north.

Question: Identify the question that Tom and Justin's experiment can best answer.
Context: The passage below describes an experiment. ... 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. Figure: a catapult for launching ping pong balls.
Options: (A) Do ping pong balls stop rolling along the ground sooner after being launched from a 30° angle or a 45° angle? (B) Do ping pong balls travel farther when launched from a 30° angle compared to a 45° angle?
Correct Answer: (B) Do ping pong balls travel farther when launched from a 30° angle compared to a 45° angle?

Please give a detailed explanation.

Solution: The question that Tom and Justin's experiment can best answer is (B) Do ping pong balls travel farther when launched from a 30° angle compared to a 45° angle? This is because the experiment is designed to compare the distances the balls traveled when launched from a 45° angle to the distances the balls traveled when launched from a 30° angle. By measuring the distances between the catapult and the place where the ball hit the ground, Tom and Justin can determine which angle produces the greatest distance.

Question: What does the verbal irony in this text suggest? According to Mr. Herrera's kids, his snoring is as quiet as a jackhammer.
Context: N/A
Options: (A) The snoring is loud. (B) The snoring occurs in bursts.
Correct Answer: (A) The snoring is loud.

Please give a detailed explanation.

Solution: Verbal irony is when someone says something that is the opposite of what they mean. In this text, Mr. Herrera's kids are saying that his snoring is as quiet as a jackhammer, which implies that it is actually very loud. This suggests that the snoring is loud, which is option (A).

Question: Which of the following could Wendy's test show?
Context: People can use the engineering-design process to develop solutions to problems. One step in the process is testing if a potential solution meets the requirements of the design. ... She read that giving the bacteria more nutrients could affect the amount of insulin they produced. So, Wendy gave extra nutrients to some of the bacteria. Then, she measured how much insulin those bacteria produced compared to bacteria that did not get extra nutrients. Figure: studying bacteria in a laboratory.
Options: (A) whether producing more insulin would help the bacteria grow faster (B) whether different types of bacteria would need different nutrients to produce insulin (C) whether she added enough nutrients to help the bacteria produce 20% more insulin
Correct Answer: (C) whether she added enough nutrients to help the bacteria produce 20% more insulin

Please give a detailed explanation.

Solution: The passage describes how Wendy used the engineering-design process to test a solution to a problem. She wanted to increase the amount of insulin that the bacteria produced by 20%, so she gave extra nutrients to some of the bacteria and then measured how much insulin those bacteria produced compared to bacteria that did not get extra nutrients. This testing would show whether she added enough nutrients to help the bacteria produce 20% more insulin. Option A is incorrect because the passage does not mention anything about the bacteria growing faster. Option B is incorrect because the passage does not mention anything about different types of bacteria needing different nutrients.

Question: Does this passage describe the weather or the climate?
Context: Figure: Chicago. Chicago is known as The Windy City. But on average, the wind there only blows at about 10 miles per hour. Hint: Weather is what the atmosphere is like at a certain place and time. Climate is the pattern of weather in a certain place.
Options: (A) weather (B) climate
Correct Answer: (B) climate

Please give a detailed explanation.

Solution: The passage describes the average wind speed in Chicago, which is a pattern of weather in the city. This pattern of weather is part of the climate of Chicago, so the correct answer is (B) climate. Climate is the average weather conditions in a certain area over a long period of time. It is determined by factors such as temperature, precipitation, humidity, and wind speed. Weather, on the other hand, is the atmospheric conditions at a certain place and time. It is determined by the same factors as climate, but it is more localized and can change quickly.

Table 6: The process of generating lectures.

Here are some problems about "Read a map: cardinal directions"

Question: Which of these states is farthest north?
Context: N/A
Options: (A) West Virginia (B) Louisiana (C) Arizona (D) Oklahoma
Question: Which of these states is farthest north?
Context: N/A
Options: (A) Alabama (B) Idaho (C) South Carolina (D) Oklahoma
Question: Which of these states is farthest north?
Context: N/A
Options: (A) Maine (B) South Carolina (C) Kansas (D) Delaware
Question: Which of these states is farthest east?
Context: N/A
Options: (A) Montana (B) New Mexico (C) Arkansas (D) Nebraska
Question: Which of these states is farthest north?
Context: N/A
Options: (A) West Virginia (B) Mississippi (C) Nebraska (D) Oklahoma

Based on the problems above, please give a general lecture on the "Read a map: cardinal directions" type of question in one sentence.

Lecture: Reading a map to identify cardinal directions involves determining which direction a given location is in relation to other locations on the map.

Here are some problems about "Interpret figures of speech"

Question: What does the verbal irony in this text suggest?According to Mr. Herrera's kids, his snoring is as quiet as a jackhammer.
Context: N/A
Options: (A) The snoring is loud. (B) The snoring occurs in bursts.
Question: What does the metaphor in this text suggest? All religions, arts, and sciences are branches of the same tree. ž014Albert Einstein
Context: N/A
Options: (A) All religions, arts, and sciences are distant from one another. (B) All religions, arts, and sciences are related.
Question: What does the verbal irony in this text suggest? This is the best day of my life, Mr. Hogan mumbled after his car broke down on the way to an important job interview.
Context: N/A
Options: (A) Mr. Hogan was having a bad day. (B) Mr. Hogan liked fixing cars.
Question: What does the allusion in this text suggest?Tyler seems to have the Midas touch. Without any special experience or training, he launched a thriving business and then established a well-respected charity.
Context: N/A
Options: (A) Tyler is successful at all that he does. (B) Tyler has a hands-on approach to his work.
Question: What does the idiom in this text suggest? Speak of the devil! Whitney declared when Charlie strolled into the room.
Context: N/A
Options: (A) Whitney had just been speaking about Charlie. (B) Whitney didn't trust Charlie.

Based on the problems above, please give a general lecture on the "Interpret figures of speech" type of question in one sentence.

Lecture: Interpret figures of speech questions require the reader to identify the figurative language used in a text and determine the meaning or connotation it conveys.

Here are some problems about "Animal adaptations: beaks, mouths, and necks"

Question: Which animal's mouth is also adapted for bottom feeding?
Context: Sturgeons eat invertebrates, plants, and small fish. They are bottom feeders. Bottom feeders find their food at the bottom of rivers, lakes, and the ocean. The 's mouth is located on the underside of its head and points downward. Its mouth is adapted for bottom feeding. Figure: sturgeon.
Options: (A) discus (B) armored catfish
Question: Which bird's beak is also adapted to tear through meat?
Context: Red-tailed hawks eat fish, mammals, and other birds. The shape of the 's beak is adapted to tear through meat. Figure: red-tailed hawk.
Options: (A) sand martin (B) Cape vulture
Question: Which animal's mouth is also adapted for bottom feeding?
Context: Armored catfish eat plants and small
invertebrates. They are bottom feeders. Bottom feeders find their food at the bottom of rivers, lakes, and the ocean. The catfish's mouth is located on the underside of its head and points downward. Its mouth is adapted for bottom feeding. Figure: armored catfish.
Options: (A) clown triggerfish (B) sturgeon
Question: Which bird's beak is also adapted to get nectar out of long flowers?
Context: Green violetears live in the forests of Central and South America. The shape of the 's beak is adapted to get nectar out of long flowers. Figure: green violetear.
Options: (A) ground hornbill (B) violet sabrewing
Question: ...

Based on the problems above, please give a general lecture on the "Animal adaptations: beaks, mouths, and necks" type of question in one sentence.

Lecture: Animal adaptations such as beaks, mouths, and necks are specialized features that enable animals to feed on different types of food, such as meat, insects, nuts, and plant matter.

Table 7: The process of generating plans.

Here are some problems about "Read a map: cardinal directions"

The lecture about "Read a map: cardinal directions" is "Reading a map to identify cardinal directions involves determining which direction a given location is in relation to other locations on the map."

Question: Which of these states is farthest north?
Context: N/A
Options: (A) West Virginia (B) Louisiana (C) Arizona (D) Oklahoma
Question: Which of these states is farthest north?
Context: N/A
Options: (A) Alabama (B) Idaho (C) South Carolina (D) Oklahoma
Question: Which of these states is farthest north?
Context: N/A
Options: (A) Maine (B) South Carolina (C) Kansas (D) Delaware
Question: ...

Based on the lecture above and these problems, let's understand these problems and devise a general and brief plan step by step to solve these problems (begin with 1, 2, 3...).

Plan: 1. Read the question carefully and identify the context. 2. Identify the cardinal directions (north, south, east, and west) in relation to the given locations on the map. 3. Compare the locations and determine which one is farthest in the given direction. 4. Select the correct answer from the given options.

Here are some problems about "Interpret figures of speech"

The lecture about "Interpret figures of speech" is "Interpret figures of speech questions require the reader to identify the figurative language used in a text and determine the meaning or connotation it conveys."

Question: What does the verbal irony in this text suggest?According to Mr. Herrera's kids, his snoring is as quiet as a jackhammer.
Context: N/A
Options: (A) The snoring is loud. (B) The snoring occurs in bursts.
Question: What does the metaphor in this text suggest? All religions, arts, and sciences are branches of the same tree. ž014Albert Einstein
Context: N/A
Options: (A) All religions, arts, and sciences are distant from one another. (B) All religions, arts, and sciences are related.
Question: What does the verbal irony in this text suggest? This is the best day of my life, Mr. Hogan mumbled after his car broke down on the way to an important job interview.
Context: N/A
Options: (A) Mr. Hogan was having a bad day. (B) Mr. Hogan liked fixing cars.
Question: ...

Based on the lecture above and these problems, let's understand these problems and devise a general and brief plan step by step to solve these problems (begin with 1, 2, 3...).

Plan: 1. Read the text carefully and identify the figure of speech used. 2. Analyze the context of the text to determine the meaning or connotation of the figure of speech. 3. Choose the option that best matches the meaning or connotation of the figure of speech.

Here are some problems about "Animal adaptations: beaks, mouths, and necks"

The lecture about "Animal adaptations: beaks, mouths, and necks" is "Animal adaptations such as beaks, mouths, and necks are specialized features that enable animals to feed on different types of food, such as meat, insects, nuts, and plant matter."

Question: Which animal's mouth is also adapted for bottom feeding?
Context: Sturgeons eat invertebrates, plants, and small fish. They are bottom feeders. Bottom feeders find their food at the bottom of rivers, lakes, and the ocean. The 's mouth is located on the underside of its head and points downward. Its mouth is adapted for bottom feeding. Figure: sturgeon.
Options: (A) discus (B) armored catfish
Question: Which bird's beak is also adapted to tear through meat?
Context: Red-tailed hawks eat fish, mammals, and other birds. The shape of the 's beak is adapted to tear through meat. Figure: red-tailed hawk.
Options: (A) sand martin (B) Cape vulture
Question: Which animal's mouth is also adapted for bottom feeding?
Context: Armored catfish eat plants and small
invertebrates. They are bottom feeders. Bottom feeders find their food at the bottom of rivers, lakes, and the ocean. The catfish's mouth is located on the underside of its head and points downward. Its mouth is adapted for bottom feeding. Figure: armored catfish.
Options: (A) clown triggerfish (B) sturgeon
Question: ...

Based on the lecture above and these problems, let's understand these problems and devise a general and brief plan step by step to solve these problems (begin with 1, 2, 3...).

Plan: 1. Read the lecture and understand the context of the questions. 2. Read the question and identify the type of adaptation being asked about. 3. Look at the figure provided and identify the animal being discussed. 4. Research the animal to find out what type of adaptation it has. 5. Compare the options provided and select the one that best matches the adaptation.

Table 8: Data generation of PCoT prompting.

Here are is a problem about "Read a map: cardinal directions"

The lecture about "Read a map: cardinal directions" is "Reading a map to identify cardinal directions involves determining which direction a given location is in relation to other locations on the map."

The plan to solve "Read a map: cardinal directions" problem is "1. Read the question carefully and identify the context. 2. Identify the cardinal directions (north, south, east, and west) in relation to the given locations on the map. 3. Compare the locations and determine which one is farthest in the given direction. 4. Select the correct answer from the given options."

Question: Which of these states is farthest north?
Context: N/A
Options: (A) West Virginia (B) Louisiana (C) Arizona (D) Oklahoma
Correct Answer: (A) West Virginia

Based on the lecture, the plan and the problem, please carry out the plan and solve the problem step by step (begin with 1, 2, 3...).

Solution: 1. Read the question carefully and identify the context. The context of the question is to determine which of the given states is farthest north. 2. Identify the cardinal directions (north, south, east, and west) in relation to the given locations on the map. North is the direction that is farthest away from the equator and closest to the North Pole. South is the direction that is farthest away from the North Pole and closest to the equator. East is the direction that is farthest away from the Prime Meridian and closest to the International Date Line. West is the direction that is farthest away from the International Date Line and closest to the Prime Meridian. 3. Compare the locations and determine which one is farthest in the given direction. West Virginia is farthest north, followed by Louisiana, Arizona, and Oklahoma. 4. Select the correct answer from the given options. The correct answer is (A) West Virginia.

Here is a problem about "Interpret figures of speech"

The lecture about "Interpret figures of speech" is "Interpret figures of speech questions require the reader to identify the figurative language used in a text and determine the meaning or connotation it conveys."

The plan to solve "Interpret figures of speech" problem is "1. Read the text carefully and identify the figure of speech used. 2. Analyze the context of the text to determine the meaning or connotation of the figure of speech. 3. Choose the option that best matches the meaning or connotation of the figure of speech."

Question: What does the verbal irony in this text suggest?According to Mr. Herrera's kids, his snoring is as quiet as a jackhammer.
Context: N/A
Options: (A) The snoring is loud. (B) The snoring occurs in bursts.
Correct Answer: (A) The snoring is loud.

Based on the lecture, the plan and the problem, please carry out the plan and solve the problem step by step (begin with 1, 2, 3...).

Solution: 1. Read the text carefully and identify the figure of speech used. The figure of speech used in the text is verbal irony. 2. Analyze the context of the text to determine the meaning or connotation of the figure of speech. The context of the text suggests that the verbal irony is used to convey the idea that Mr. Herrera's snoring is actually very loud, despite what his kids say. 3. Choose the option that best matches the meaning or connotation of the figure of speech. The option that best matches the meaning or connotation of the figure of speech is (A) The snoring is loud.

Here is a problem about "Animal adaptations: beaks, mouths, and necks"

The lecture about "Animal adaptations: beaks, mouths, and necks" is "Animal adaptations such as beaks, mouths, and necks are specialized features that enable animals to feed on different types of food, such as meat, insects, nuts, and plant matter."

The plan to solve "Animal adaptations: beaks, mouths, and necks" problem is "1. Read the lecture and understand the context of the questions. 2. Read the question and identify the type of adaptation being asked about. 3. Look at the figure provided and identify the animal being discussed. 4. Research the animal to find out what type of adaptation it has. 5. Compare the options provided and select the one that best matches the adaptation."
Question: Which animal's mouth is also adapted for bottom feeding?
Context: Sturgeons eat invertebrates, plants, and small fish. They are bottom feeders. Bottom feeders find their food at the bottom of rivers, lakes, and the ocean. The 's mouth is located on the underside of its head and points downward. Its mouth is adapted for bottom feeding. Figure: sturgeon.
Options: (A) discus (B) armored catfish
Correct Answer: (B) armored catfish

Based on the lecture, the plan and the problem, please carry out the plan and solve the problem step by step (begin with 1, 2, 3...).

Solution: 1. Read the lecture and understand the context of the questions. The lecture states that animal adaptations such as beaks, mouths, and necks are specialized features that enable animals to feed on different types of food. 2. Read the question and identify the type of adaptation being asked about. The question is asking about an adaptation related to bottom feeding. 3. Look at the figure provided and identify the animal being discussed. The figure provided is a sturgeon. 4. Research the animal to find out what type of adaptation it has. Researching the sturgeon reveals that its mouth is located on the underside of its head and points downward. This adaptation is specifically adapted for bottom feeding. 5. Compare the options provided and select the one that best matches the adaptation. The options provided are (A) discus and (B) armored catfish. The armored catfish has a mouth that is adapted for bottom feeding, making it the correct answer. Therefore, the correct answer is (B) armored catfish.