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

Lei Wang[1,2], **Yi Hu**[3], **Jiabang He**[3], **Xing Xu**[3], **Ning Liu**[1*], **Hui Liu**[4], **Heng Tao Shen**[3]

[1] School of Information Science and Technology, Beijing Forestry University, China
[2] Singapore Management University, Singapore
[3] School of Computer Science and Engineering, University of Electronic Science and Technology of China
[4] Beijing Rongda Technology Co., Ltd., China
demolwang@gmail.com, yihu0118@gmail.com, JiaBangH@outlook.com
xing.xu@uestc.edu.cn, liuning0928@bjfu.edu.cn, ryuki122382@gmail.com, 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 external 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%. The code is publicly available at https://github.com/T-SciQ/T-SciQ.

Introduction

Scientific problem solving has recently been employed to evaluate the multi-hop reasoning capability and interpretability of AI systems (Kembhavi et al. 2017; Sampat, Yang, and Baral 2020; Dalvi et al. 2021). However, these datasets (Kembhavi et al. 2017; Jansen et al. 2018) suffer from limited scale. To address this issue, Lu et al. (2022a) 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.

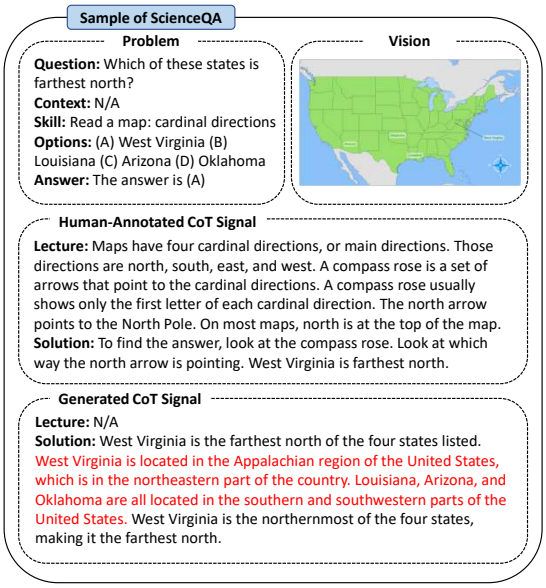


Figure 1. The input of a ScienceQA data example includes a question, context, image, skill, and options. Annotations include the ground truth answer and CoT rationale (lecture and solution). Compared to annotated CoT, LLM-generated CoT includes greater amounts of essential external knowledge.

Recently, Large Language Models (LLMs) have shown exceptional performance in various Natural Language Processing (NLP) tasks (Brown et al. 2020; Thoppilan et al. 2022). Specifically, they have demonstrated the chain-of-thought (CoT) ability to solve complex reasoning problems by using a few demonstration examples without additional training (Wei et al. 2022a; Kojima et al. 2022; Zhang et al. 2022). However, the existing research on CoT reasoning is mainly limited to the language modality (Wang et al. 2022a; Zhou et al. 2022; Lu et al. 2022b; Fu et al. 2022), with little attention paid to multimodal scenarios, such as science question answering. To address this issue, a common approach

*Corresponding Author

is to use caption models to translate visual information into the language modality and prompt LLMs to perform CoT reasoning (Lu et al. 2022a). However, the use of caption generation models in scientific problems may result in significant information loss when meeting highly complex images. To overcome this issue, Zhang et al. (2023b) proposed a framework called Multimodal-CoT that models both language and visual 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 two fundamental limitations. First, the human annotation of CoT reasoning is time-consuming (Nye et al. 2021; Cobbe et al. 2021), particularly for complex tasks like ScienceQA, which necessitates extensive expert knowledge to create a reasoning process for the answer. Second, as shown in Figure 1, the annotated rationale may lack essential external 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* to solve the ScienceQA task. The proposed T-SciQ framework in Figure 2 consists of three stages: generating teaching data, mixing teaching data, and fine-tuning. For teaching data generation, we use a simple zero-shot instruction and a hint of the correct answer to generate a CoT rationale for a QA data example to obtain a QA-CoT sample. Although the model taught by QA-CoT samples excels at tackling simple problems, it still struggles with highly complicated problems. To overcome this challenge, we follow the zero-shot plan-and-solve prompting (Wang et al. 2023) to generate plan-based CoT (PCoT) rationales, which decompose complex problems into simpler subproblems to solve, to obtain QA-PCoT teaching samples.

To this end, we construct a new teaching dataset called T-SciQ by mixing QA-CoT and QA-PCoT datasets to combine the strengths of both teaching signals. Specifically, we use the validation set to determine whether the PCoT teaching signal or CoT teaching signal is more appropriate for each data example in a given skill. Then, we fine-tune the student model with teaching data. We follow the Multimodal-CoT (Zhang et al. 2023b) to build our student model, which consists of two-stage: rationale generation teaching and answer inference teaching. During inference, the model trained in the first stage generates rationales for the test data. The generated rationales are subsequently used in the second stage to infer answers. Experiment results on the ScienceQA benchmark show that our method surpasses the previous state-of-the-art approaches by a large margin.

Our main contributions are summarized as follows: 1) We propose a novel framework for generating high-quality CoT rationale and training student models to perform CoT reasoning for the ScienceQA task; 2) We introduce a data mixing strategy to produce effective teaching data samples for simple and complex problems; 3) Our method achieves a new state-of-the-art performance on the ScienceQA benchmark, surpassing all previous models by a large margin.

Related Work

Chain-of-Thought Prompting. Recently, to solve complex reasoning tasks, Wei et al. (2022b) propose CoT prompting 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 CoT prompting from different aspects, including improving the quality of demonstrations (Rubin, Herzig, and Berant 2021; Zhang et al. 2022; Fu et al. 2022; Lu et al. 2022b; He et al. 2023) and improving the quality of reasoning chains (Zhou et al. 2022; Khot et al. 2022; Chen et al. 2022; Wang et al. 2022b,a; Li et al. 2022b; Tian et al. 2023). Zero-shot CoT (Kojima et al. 2022) elicited reasoning step by appending a prompt like "*Let's think step by step*" to the test question. Chameleon (Lu et al. 2023) proposed a plug-and-play compositional reasoning framework to utilize multiple modules to obtain high quality prompting. Our work mainly focuses on mixing different teaching CoT rationales for different problems.

LLMs as Teachers. In recent studies, CoT reasoning is elicited in small models using fine-tuned language models. Magister et al. (2022) benefit smaller models through CoT distillation. Huang et al. (2022) show that LLMs can enhance reasoning using self-generated solutions from unlabeled data. Ho, Schmid, and Yun (2022) propose Fine-tune-CoT to leverage the capabilities of LLMs to generate reasoning samples and teach smaller models via fine-tuning. Distilling step-by-step (Hsieh et al. 2023) improves small model performance using LLM rationales with less data. Multimodal-CoT (Zhang et al. 2023b) uses two-stage fine-tuning with annotated CoT rationales and visual features to achieve state-of-the-art results on the ScienceQA benchmark. Our work exploits generating two types of teaching data from LLMs and mixing teaching data. We discover that this simple method highly improves student performance in complex multi-modality tasks, which has not yet been recognized in previous studies on fine-tuning with CoT reasoning (Hsieh et al. 2023; Ho, Schmid, and Yun 2022; Huang et al. 2022; Magister et al. 2022; Fu et al. 2023; Hu et al. 2023).

Our T-SciQ Approach

Overview

This section presents the proposed fine-tuning strategy T-SciQ, which utilizes a LLM named SciTeacher to generate teaching data and improve the performance of a smaller student model (SciStudent) by generated teaching data. The proposed T-SciQ strategy comprises three 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 CoT rationales to obtain Question-Answer-CoT (QA-CoT) samples, and planning-based CoT rationale (PCoT) to obtain Question-Answer-PCoT (QA-PCoT) samples. To combine the strengths of both datasets, we create a new teaching dataset called T-SciQ by mixing QA-CoT and QA-PCoT datasets. Specifically, we use the validation set to determine whether the PCoT teaching signal or CoT teaching signal is more appropriate for each data example in a given skill. We then use T-SciQ teaching samples to fine-tune

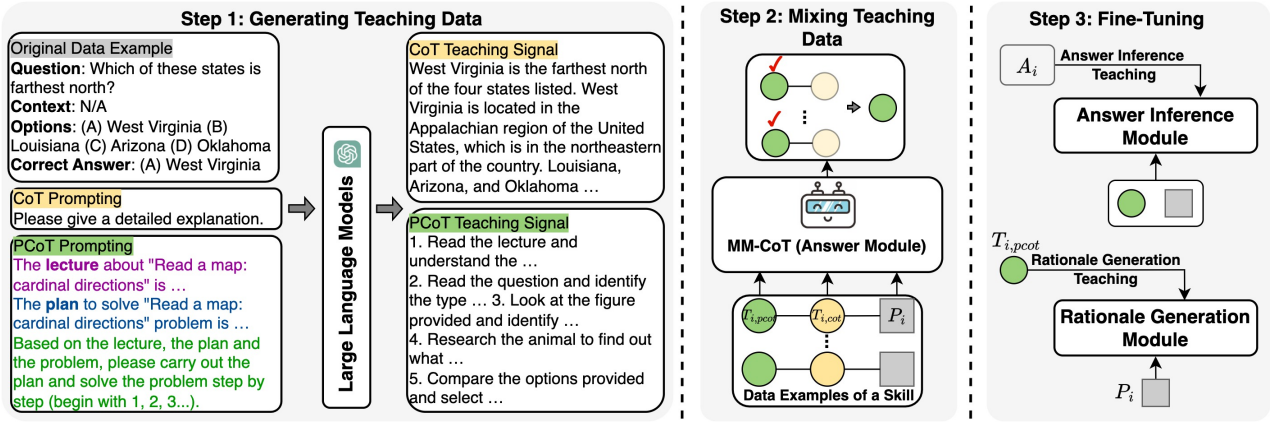


Figure 2. Key steps of our T-SciQ approach. T-SciQ consists of three stages: (i) generating teaching data; (ii) mixing teaching data; and (iii) fine-tuning.

the smaller student models. In the following, we provide a detailed description of these three components.

Generating Teaching Data

We produce two types of data samples for teaching: QA-CoT sample with a generated CoT rationale and QA-PCoT sample equipped with a generated PCoT rationale.

QA-CoT Sample Generation. Although using human-annotated CoT signals is valuable for training models to elicit CoT reasoning ability, it has two inherent limitations: time-consuming and lack of external essential information due to human annotators' restricted expertise.

To address these issues, we introduce a zero-shot prompting to generate high-quality CoT rationales from LLMs. 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 for the input question, the $[X_c]$ slot is for the input context, the $[X_o]$ slot contains the possible options, the $[A]$ slot is for the correct answer that can work as a hint to guide LLMs to generate a more reliable rationale, and the $[Instruct]$ slot contains instructions, i.e., "*Please give me a detailed explanation.*", to guide LLMs 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". Subsequently, we feed the filled prompt to LLMs to output a reasoning process for a given training data example to obtain QA-CoT data $D_{\text{QA-CoT}}$.

QA-PCoT Sample Generation. Although using QA-CoT samples can address issues of human-annotated CoT, addressing highly complex problems remains a challenge. To overcome this challenge and obtain appropriate teaching CoT rationale, we introduce a 3-step zero-shot prompting to decompose complex problems into simpler subproblems.

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.*". Note that many QA examples need the same skill to be solved.

Step 2: Plan Generation. The template used to generate a plan for a specific skill based on the generated 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...)*".

Step 3: Rationale Generation. The lecture and plan generated by the first two prompts are used to generate a plan-based CoT rationale for each training example. The rationale generation template is formulated as follows: "Skill: $[S]$. Lecture: $[L]$. Plan: $[P]$. QA pair: $[X_q, A]$. $[Instruct]$.". 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...)". Examples of this three-step prompting can be found in the supplementary material.

Mixing Teaching Data

The QA-PCoT dataset is effective for teaching problem-solving skills for complex problems, while simpler problems don't require decomposition. In contrast, the QA-CoT dataset is suitable for teaching problem-solving skills for simple problems. To combine the strengths of both datasets, we create a new teaching dataset called T-SciQ by mixing QA-CoT and QA-PCoT datasets. We introduce a new approach that uses the validation set to determine whether the PCoT teaching signal or CoT teaching signal is more appropriate for a data example in a given skill.

Given a ScienceQA problem P_i with the language input $X_{i,\text{la}}$ and the visual input $X_{i,\text{v}}$, our objective is to let an answer generation model F_a^s help identify the optimal teaching signal $T_{i,k}$ from the possible choices T_i, i.e., CoT teaching signal $T_{i,\text{cot}}$ or PCoT teaching signal $T_{i,\text{pcot}}$, thereby maxi-

mizing the answer accuracy of the validation set. The answer generation module F_a^s is similar to the one described in Multimodal-CoT (Zhang et al. 2023b). The generated answer $\hat{A}_i$ is produced by $F_a^s(X_{i,\text{la}}, X_{i,\text{v}}, T_{i,k})$, and the number of errors is obtained by comparing the generated answer $\hat{A}_i$ and the label A_i. If the number of errors for validation samples with PCoT in a skill is lower than that of validation samples with CoT in a skill, we select PCoT rationale as the teaching rationale for all training data examples in this skill. Otherwise, we select CoT rationale. The obtained teaching samples are then used to fine-tune the student model. To train the answer generation module, we utilize a subset of training data examples, each of which is associated with the human-annotated teaching signal from the original ScienceQA dataset.

Fine-Tuning

Our teaching follows the Multimodal-CoT (Zhang et al. 2023b) two-stage fine-tuning framework: rationale generation teaching and answer inference teaching.

Rationale Generation Teaching. In this stage, the rationale generation model $F_r(P_i)$ is trained to predict the teaching signal T_i for a given problem P_i, where T_i either be CoT rationale or PCoT rationale. The input of $F_r(P_i)$ consists of $X^1_{i,\text{la}}$ and $X_{i,\text{v}}$, where $X^1_{i,\text{la}}$ represents the language input and $X_{i,\text{v}}$ represents the visual input. Formally, the probability of generating rationale T_i can be formulated as follows:

$$p(T_i|X_{i,\text{la}^1}, X_{i,\text{v}}) = \prod_{j=1}^{N_{T_i}} p_{\theta_r}\left(T_{i,j} \mid X^1_{i,\text{la}}, X_{i,\text{v}}, T_{i,<j}\right), \tag{1}$$

where θ_r represents learnable parameters of the rationale generation model F_r and N_{T_i} is the length of T_i.

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

$$p(A_i|X^2_{i,\text{la}}, X_{i,\text{v}}) = \prod_{j=1}^{N_{A_i}} p_{\theta_a}\left(A_i \mid X^2_{i,\text{la}}, X_{i,\text{v}}, A_{i,<j}\right), \tag{2}$$

where θ_a represents learnable parameters in the answer inference teaching stage.

Model Architecture We utilize the Multimodal-CoT (Zhang et al. 2023b) model architecture as our default, which employs a Transformer model (Vaswani et al. 2017) for encoding language and a vision Transformer for encoding visual information. The gated fusion mechanism, proposed in (Li et al. 2022a), is used to effectively integrate the language and vision representations. Finally, a Transformer decoder is used to generate the target output. Note that rationale generation and answer inference share the same model but differ in the input and output.

Experiment

Experimental Setup

Dataset. We evaluate our proposed method on the **ScienceQA** (Lu et al. 2022a) dataset, a latest 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 3:1:1, i.e., $12,726$, $4,241$, and $4,241$ examples, respectively. The dataset includes annotated reasoning chains for each data example. In this work, we extract our training signals from large language models instead of using human annotated signals.

Baselines. We provide a comparison of our proposed method with extensive baseline methods. Specifically, we have several early VQA models, including `MCAN` (Yu et al. 2019), `Top-Down` (Anderson et al. 2018), `BAN` (Kim, Jun, and Zhang 2018), `DFAF` (Gao et al. 2019). 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` (Kim, Son, and Kim 2021), `Patch-TRM` (Lu et al. 2021), and `VisualBERT` (Li et al. 2019), `UnifiedQA` (Khashabi et al. 2020), `MM-COT` (Zhang et al. 2023b). These methods use 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` (Zhang et al. 2023a) and `LLaVA` (Liu et al. 2023). They use a strong open-access LLM such as `LLaMa` (Touvron et al. 2023) as the base model and incorporate a vision module to model visual information. We also include widely-used in-context learning baselines: the chain of thought (CoT) prompting (Wei et al. 2022a), where each in-context demonstration example comprises the input question and output annotated reasoning process. We compare to the CoT baselines over different API-based OpenAI LLMs (OpenAI 2022, 2023), such as GPT-3.5 (`GPT-3.5 w/ COT`), ChatGPT (`ChatGPT w/ COT`), GPT-4 (`GPT-4 w/ COT`), and `Chameleon` (Lu et al. 2023). Additionally, we also compare to the standard few-shot prompting approach using GPT-3.5 (`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 of text-davinci-003 version as the teacher model for our approach unless otherwise specified. To validate the generalizability of our method, we experiment with three distinct student models, namely UnifiedQA$_{\texttt{Base}}$ w/ CoT (Lu et al. 2022a), Mutimodal-CoT$_{\texttt{Base}}$(Lu et al. 2022a), and Mutimodal-CoT$_{\texttt{Large}}$(Zhang et al. 2023b). These models are chosen due to their strong performances achieved by fine-

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 (Yu et al. 2019)	95M	56.08	46.23	58.09	59.43	51.17	55.40	51.65	59.72	54.54
Top-Down (Anderson et al. 2018)	70M	59.50	54.33	61.82	62.90	54.88	59.79	57.27	62.16	59.02
BAN (Kim, Jun, and Zhang 2018)	112M	60.88	46.57	66.64	62.61	52.60	65.51	56.83	63.94	59.37
DFAF (Gao et al. 2019)	74M	64.03	48.82	63.55	65.88	54.49	64.11	57.12	67.17	60.72
ViLT (Kim, Son, and Kim 2021)	113M	60.48	63.89	60.27	63.20	61.38	57.00	60.72	61.90	61.14
Patch-TRM (Lu et al. 2021)	90M	65.19	46.79	65.55	66.96	55.28	64.95	58.04	67.50	61.42
VisualBERT (Li et al. 2019)	111M	59.33	69.18	61.18	62.71	62.17	58.54	62.96	59.92	61.87
UnifiedQA$_{\texttt{Base}}$ (Khashabi et al. 2020)	223M	68.16	69.18	74.91	63.78	61.38	77.84	72.98	65.00	70.12
LLaMa-Adapter (Zhang et al. 2023a)	>7B	84.37	88.30	84.36	83.72	80.32	86.90	85.83	84.05	85.19
LLaVA (Liu et al. 2023)	>7B	90.36	95.95	88.00	89.49	88.00	90.66	90.93	90.90	90.92
GPT-3.5 (Chen et al. 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 (Lu et al. 2022a)	>175B	75.44	70.87	78.09	74.68	67.43	79.93	78.23	69.68	75.17
ChatGPT w/ CoT (Lu et al. 2023)	>175B	78.82	70.98	83.18	77.37	67.92	86.13	80.72	74.03	78.31
GPT-4 w/ CoT (Lu et al. 2023)	>175B	84.06	73.45	87.36	81.87	70.75	90.73	84.69	79.10	82.69
Chameleon (Lu et al. 2023)	>175B	89.83	74.13	89.82	88.27	77.64	92.13	88.03	83.72	86.54
UnifiedQA-CoT$_{\texttt{Base}}$ (Lu et al. 2022a)	223M	71.00	76.04	78.91	66.42	66.53	81.81	77.06	68.82	74.11
UnifiedQA-T-SciQ$_{\texttt{Base}}$ (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$_{\texttt{Base}}$ (Zhang et al. 2023b)	223M	87.52	77.17	85.82	87.88	82.90	86.83	84.65	85.37	84.91
Mutimodal-T-SciQ$_{\texttt{Base}}$ (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$_{\texttt{Large}}$ (Zhang et al. 2023b)	738M	95.91	82.00	90.82	95.26	88.80	92.89	92.44	90.31	91.68
Mutimodal-T-SciQ$_{\texttt{Large}}$ (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 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.

Model	Avg
Multimodal-T-SciQ$_{\texttt{Base}}$ (Mixing)	91.75
Multimodal-T-SciQ$_{\texttt{Base}}$ only w/ QA-CoT	85.99
Multimodal-T-SciQ$_{\texttt{Base}}$ only w/ QA-PCoT	88.56
Mutimodal-CoT$_{\texttt{Base}}$	84.91
Multimodal-T-SciQ$_{\texttt{Large}}$ (Mixing)	96.18
Multimodal-T-SciQ$_{\texttt{Large}}$ only w/ QA-CoT	93.44
Multimodal-T-SciQ$_{\texttt{Large}}$ only w/ QA-PCoT	94.11
Mutimodal-CoT$_{\texttt{Large}}$	91.68

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

tuning with annotated reasoning signals. 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. These student models are 200× smaller than their teacher models.

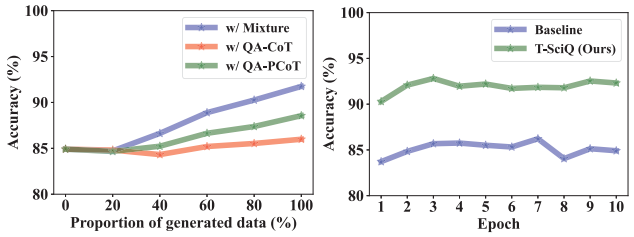
Main Results

T-SciQ *v.s.* Baselines. Table 1 details the performance accuracy of baselines and student models trained using the proposed T-SciQ signals. Mutimodal-T-SciQ$_{\texttt{Large}}$, which is the model architecture of Mutimodal-CoT$_{\texttt{Large}}$ fine-tuned with mixed teacher signals, 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$_{\texttt{Large}}$ outperforms the most powerful fine-tuning baseline, Mutimodal-CoT$_{\texttt{Large}}$, 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 elicit better planning and chain-of-thought reasoning ability in student models smaller than 1B.

T-SciQ with Different Base Student Models. Instead of only using the model architecture of Mutimodal-CoT$_{\texttt{Large}}$ as the base student model, we evaluate different base student models fine-tuned with mixed teaching signals: the variant UnifiedQA-T-SciQ$_{\texttt{Base}}$ and Mutimodal-T-SciQ$_{\texttt{Base}}$. The relative performance ranking between the base student model with annotated CoT signals and the one with mixing teacher signals remains unchanged. Specifically, UnifiedQA-T-SciQ$_{\texttt{Base}}$ outperforms UnifiedQA$_{\texttt{Base}}$ w/ CoT by 5.3%

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

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



(a) (b)

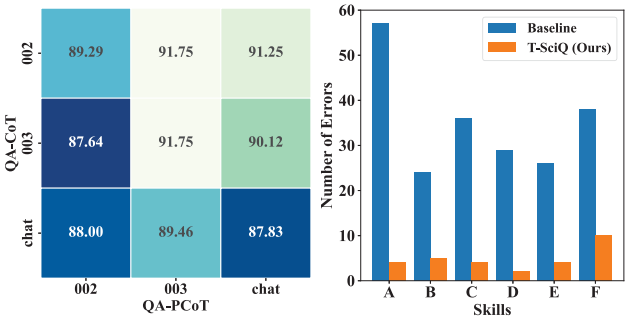
Figure 3. Further analysis on (a) the effect of Mutimodal-T-SciQ$_{\text{Base}}$ trained with different proportion of generated data and (b) accuracy curve of the baseline Mutimodal-CoT$_{\text{Base}}$ and our Mutimodal-T-SciQ$_{\text{Base}}$ across epochs.

(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 proposed teaching signals.

Further Analysis

Effect of Different Signals of T-SciQ. Our approach incorporates two distinct components for teaching signals: QA-CoT and 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 impact of each teaching signals by testing 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 a significant decrease in answering accuracy when either teaching signal was removed. These findings indicate the effectiveness of both proposed teaching signals. This is because 1) student models 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 taught by QA-PCoT signals can decompose complex problems into several simpler sub-problems.

Impact of visual Features. The choice of visual features can significantly affect the performance of models on ScienceQA. Thus, we conduct an evaluation of three widely-used visual features, which are CLIP (Radford et al. 2021), DETR (Carion et al. 2020), and ResNet (He et al. 2016). Both



(a) (b)

Figure 4. 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.

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 visual features. Table 3 shows the results of comparing these three visual features. Our findings suggest that incorporating visual 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 visual feature in our main experiments.

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 3a, the increasing proportion of training data with T-SciQ signals increases performance.

Performance Change with Epoch. Figure 3b 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. It indicates that our method exhibits relatively higher accuracy at the initial training stages.

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 useful 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 above API-based models. We conduct this experiment using the Multimodal-T-SciQ$_{\text{Base}}$. Figure 4a 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

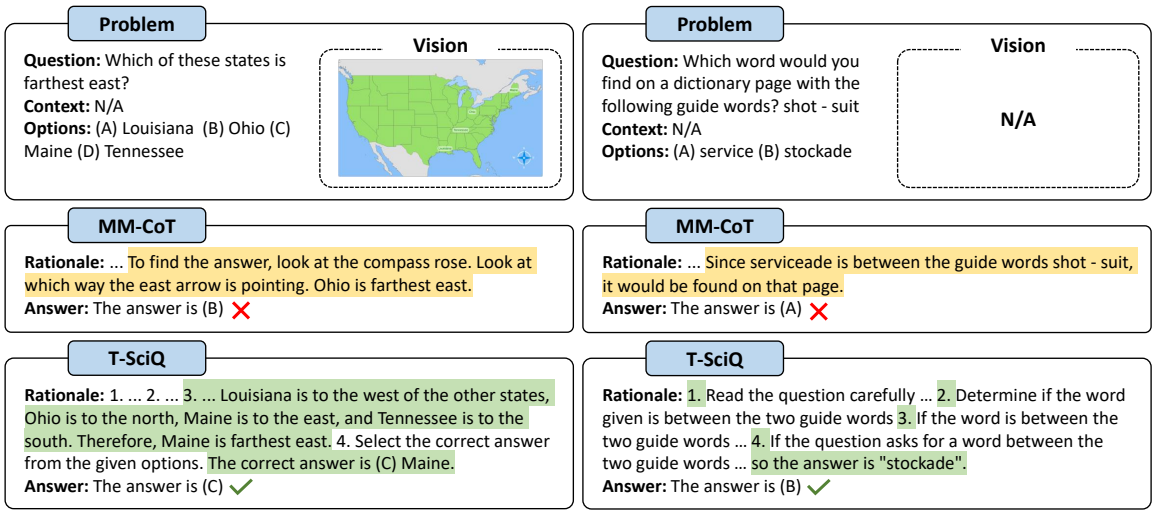


Figure 5. Examples of MM-CoT (baseline) and the model trained with T-SciQ (ours) signals for generating rationales and predicting answers. To solve these examples, commonsense knowledge such as geographic knowledge (a) and multi-step reasoning (b) are required.

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.

Error Analysis. To better understand the model's behavior trained using our proposed T-SciQ signals, we analyze six selected skills shown in Figure 4b. It shows the error analysis of prediction for six specific skills (A-F), i.e., "*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 "*Identifying oceans and continents*" require multi-step complex reasoning that T-SciQ teaching signals can teach. On the other hand, examples of skills such as "*Reading a map: cardinal directions*" require common sense and factual knowledge from the open world, which T-SciQ signals can also provide.

Case Study. The case study compares T-SciQ and Multimodal-CoT on the ScienceQA benchmark (Figure 5). Figure 5a shows cases needing geographic knowledge. Human-annotated CoT may lack open-world information, while T-SciQ includes it. Figure 5b shows a multi-step reasoning case without image input. Multimodal-CoT errors while our model decomposes and answers correctly. These highlight that T-SciQ is well-suited to handle problems that require open knowledge and decomposition.

Comparison on Other NLP Reasoning Datasets. To verify the versatility of our teaching approach, we additionally assess our approach on six reasoning tasks, following Reason-Teacher (Ho, Schmid, and Yun 2022): arithmetic (Aqua (Ling et al. 2017)), symbolic (Coin Flip (Wei et al. 2022b)), commonsense (CommonSenseQA (CSQA) (Talmor et al. 2018), StrategyQA (Geva et al. 2021)) reasoning, and logic (Date Understanding, Tracking Shuffled Objects) (Geva et al. 2021). In Table 4, we compare T-SciQ to diverse reasoning teaching signals introduced by Reason-Teacher. The results show that our T-SciQ surpasses Reason-Teacher by a large margin in 5 out of 6 datasets. It performs equally well in the remaining dataset, Coin Flip. These results indicate that higher-quality teaching signals of planning and reasoning can lead to a remarkable improvement in small student models across different scenarios.

Method	Aqua	Date	Shuffled Objects	Coin Flip	CS QA	Strategy QA
Reason-Teacher	24.02	60.36	64.44	98.67	56.76	55.02
T-SciQ	74.80	89.29	70.28	98.67	70.76	76.74

Table 4. Accuracy (%) on other six reasoning datasets.

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 plan-based CoT for highly complex problems. Furthermore, our data mixture strategy combines CoT and plan-based CoT to create a new T-SciQ teaching dataset. 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.

Acknowledgments

This work was supported in part by Central University Basic Research Business Fee Special Fund Project No. BLX202139, the National Natural Science Foundation of China under Grant No. 62222203, the New Cornerstone Science Foundation through the XPLORER PRIZE, and the Science and Technology Innovation Committee of Shenzhen Municipality Foundation (No. JCYJ20210324132203007).

References

Anderson, P.; He, X.; Buehler, C.; Teney, D.; Johnson, M.; Gould, S.; and Zhang, L. 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.

Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33: 1877–1901.

Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; and Zagoruyko, S. 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.

Chen, T.; Kornblith, S.; Swersky, K.; Norouzi, M.; and Hinton, G. E. 2020. Big self-supervised models are strong semi-supervised learners. *Advances in neural information processing systems*, 33: 22243–22255.

Chen, W.; Ma, X.; Wang, X.; and Cohen, W. W. 2022. Program of thoughts prompting: Disentangling computation from reasoning for numerical reasoning tasks. *arXiv preprint arXiv:2211.12588*.

Cobbe, K.; Kosaraju, V.; Bavarian, M.; Chen, M.; Jun, H.; Kaiser, L.; Plappert, M.; Tworek, J.; Hilton, J.; Nakano, R.; et al. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Dalvi, B.; Jansen, P.; Tafjord, O.; Xie, Z.; Smith, H.; Pipatanangkura, L.; and Clark, P. 2021. Explaining answers with entailment trees. *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Fu, Y.; Peng, H.; Ou, L.; Sabharwal, A.; and Khot, T. 2023. Specializing Smaller Language Models towards Multi-Step Reasoning. *arXiv preprint arXiv:2301.12726*.

Fu, Y.; Peng, H.; Sabharwal, A.; Clark, P.; and Khot, T. 2022. Complexity-based prompting for multi-step reasoning. *arXiv preprint arXiv:2210.00720*.

Gao, P.; Jiang, Z.; You, H.; Lu, P.; Hoi, S. C.; Wang, X.; and Li, H. 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.

Geva, M.; Khashabi, D.; Segal, E.; Khot, T.; Roth, D.; and Berant, J. 2021. Did aristotle use a laptop? a question answering benchmark with implicit reasoning strategies. *Transactions of the Association for Computational Linguistics*, 9: 346–361.

He, J.; Wang, L.; Hu, Y.; Liu, N.; Liu, H.; Xu, X.; and Shen, H. T. 2023. ICL-D3IE: In-Context Learning with Diverse Demonstrations Updating for Document Information Extraction. In *IEEE/CVF International Conference on Computer Vision*, 19428–19437.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 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*, 770–778. IEEE Computer Society.

Ho, N.; Schmid, L.; and Yun, S.-Y. 2022. Large Language Models Are Reasoning Teachers. *arXiv preprint arXiv:2212.10071*.

Hsieh, C.-Y.; Li, C.-L.; Yeh, C.-K.; Nakhost, H.; Fujii, Y.; Ratner, A.; Krishna, R.; Lee, C.-Y.; and Pfister, T. 2023. Distilling step-by-step! outperforming larger language models with less training data and smaller model sizes. *arXiv preprint arXiv:2305.02301*.

Hu, Z.; Lan, Y.; Wang, L.; Xu, W.; Lim, E.-P.; Lee, R. K.-W.; Bing, L.; and Poria, S. 2023. LLM-Adapters: An Adapter Family for Parameter-Efficient Fine-Tuning of Large Language Models. *arXiv preprint arXiv:2304.01933*.

Huang, J.; Gu, S. S.; Hou, L.; Wu, Y.; Wang, X.; Yu, H.; and Han, J. 2022. Large language models can self-improve. *arXiv preprint arXiv:2210.11610*.

Jansen, P. A.; Wainwright, E.; Marmorstein, S.; and Morrison, C. T. 2018. Worldtree: A corpus of explanation graphs for elementary science questions supporting multi-hop inference. *arXiv preprint arXiv:1802.03052*.

Kembhavi, A.; Seo, M.; Schwenk, D.; Choi, J.; Farhadi, A.; and Hajishirzi, H. 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.

Khashabi, D.; Min, S.; Khot, T.; Sabharwal, A.; Tafjord, O.; Clark, P.; and Hajishirzi, H. 2020. Unifiedqa: Crossing format boundaries with a single qa system. *arXiv preprint arXiv:2005.00700*.

Khot, T.; Trivedi, H.; Finlayson, M.; Fu, Y.; Richardson, K.; Clark, P.; and Sabharwal, A. 2022. Decomposed prompting: A modular approach for solving complex tasks. *arXiv preprint arXiv:2210.02406*.

Kim, J.-H.; Jun, J.; and Zhang, B.-T. 2018. Bilinear attention networks. *Advances in neural information processing systems*, 31.

Kim, W.; Son, B.; and Kim, I. 2021. Vilt: Vision-and-language transformer without convolution or region supervision. In *International Conference on Machine Learning*, 5583–5594. PMLR.

Kojima, T.; Gu, S. S.; Reid, M.; Matsuo, Y.; and Iwasawa, Y. 2022. Large language models are zero-shot reasoners. *arXiv preprint arXiv:2205.11916*.

Li, B.; Lv, C.; Zhou, Z.; Zhou, T.; Xiao, T.; Ma, A.; and Zhu, J. 2022a. 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.

Li, L. H.; Yatskar, M.; Yin, D.; Hsieh, C.-J.; and Chang, K.-W. 2019. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*.

Li, Y.; Lin, Z.; Zhang, S.; Fu, Q.; Chen, B.; Lou, J.-G.; and Chen, W. 2022b. On the advance of making language models better reasoners. *arXiv preprint arXiv:2206.02336*.

Ling, W.; Yogatama, D.; Dyer, C.; and Blunsom, P. 2017. Program induction by rationale generation: Learning to solve and explain algebraic word problems. *arXiv preprint arXiv:1705.04146*.

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2023. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*.

Lu, P.; Mishra, S.; Xia, T.; Qiu, L.; Chang, K.-W.; Zhu, S.-C.; Tafjord, O.; Clark, P.; and Kalyan, A. 2022a. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in Neural Information Processing Systems*, 35: 2507–2521.

Lu, P.; Peng, B.; Cheng, H.; Galley, M.; Chang, K.-W.; Wu, Y. N.; Zhu, S.-C.; and Gao, J. 2023. Chameleon: Plug-and-Play Compositional Reasoning with Large Language Models. *arXiv preprint arXiv:2304.09842*.

Lu, P.; Qiu, L.; Chang, K.-W.; Wu, Y. N.; Zhu, S.-C.; Rajpurohit, T.; Clark, P.; and Kalyan, A. 2022b. Dynamic prompt learning via policy gradient for semi-structured mathematical reasoning. *arXiv preprint arXiv:2209.14610*.

Lu, P.; Qiu, L.; Chen, J.; Xia, T.; Zhao, Y.; Zhang, W.; Yu, Z.; Liang, X.; and Zhu, S.-C. 2021. Iconqa: A new benchmark for abstract diagram understanding and visual language reasoning. *arXiv preprint arXiv:2110.13214*.

Magister, L. C.; Mallinson, J.; Adamek, J.; Malmi, E.; and Severyn, A. 2022. Teaching small language models to reason. *arXiv preprint arXiv:2212.08410*.

Nye, M.; Andreassen, A. J.; Gur-Ari, G.; Michalewski, H.; Austin, J.; Bieber, D.; Dohan, D.; Lewkowycz, A.; Bosma, M.; Luan, D.; et al. 2021. Show your work: Scratchpads for intermediate computation with language models. *arXiv preprint arXiv:2112.00114*.

OpenAI. 2022. Introducing chatgpt. https://openai.com/blog/chatgpt. Accessed: 2022-11-30.

OpenAI. 2023. GPT-4 Technical Report. *arXiv preprint arXiv:2303.08774*.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 8748–8763. PMLR.

Rubin, O.; Herzig, J.; and Berant, J. 2021. Learning to retrieve prompts for in-context learning. *arXiv preprint arXiv:2112.08633*.

Sampat, S. K.; Yang, Y.; and Baral, C. 2020. Visuo-Lingustic Question Answering (VLQA) Challenge. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings (EMNLP)*, 4606–4616.

Talmor, A.; Herzig, J.; Lourie, N.; and Berant, J. 2018. Commonsenseqa: A question answering challenge targeting commonsense knowledge. *arXiv preprint arXiv:1811.00937*.

Thoppilan, R.; De Freitas, D.; Hall, J.; Shazeer, N.; Kulshreshtha, A.; Cheng, H.-T.; Jin, A.; Bos, T.; Baker, L.; Du, Y.; et al. 2022. Lamda: Language models for dialog applications. *arXiv preprint arXiv:2201.08239*.

Tian, Q.; Zhu, H.; Wang, L.; Li, Y.; and Lan, Y. 2023. R^3 Prompting: Review, Rephrase and Resolve for Chain-of-Thought Reasoning in Large Language Models under Noisy Context. *arXiv preprint arXiv:2310.16535*.

Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.-A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; and Polosukhin, I. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems 30*, 5998–6008.

Wang, L.; Xu, W.; Lan, Y.; Hu, Z.; Lan, Y.; Lee, R. K.-W.; and Lim, E.-P. 2023. Plan-and-solve prompting: Improving zero-shot chain-of-thought reasoning by large language models. *arXiv preprint arXiv:2305.04091*.

Wang, X.; Wei, J.; Schuurmans, D.; Le, Q.; Chi, E.; and Zhou, D. 2022a. Rationale-augmented ensembles in language models. *arXiv preprint arXiv:2207.00747*.

Wang, X.; Wei, J.; Schuurmans, D.; Le, Q.; Chi, E.; and Zhou, D. 2022b. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*.

Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Chi, E.; Le, Q.; and Zhou, D. 2022a. Chain of Thought Prompting Elicits Reasoning in Large Language Models. *ArXiv preprint*, abs/2201.11903.

Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Chi, E.; Le, Q.; and Zhou, D. 2022b. Chain of thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*.

Yu, Z.; Yu, J.; Cui, Y.; Tao, D.; and Tian, Q. 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.

Zhang, R.; Han, J.; Zhou, A.; Hu, X.; Yan, S.; Lu, P.; Li, H.; Gao, P.; and Qiao, Y. 2023a. Llama-adapter: Efficient fine-tuning of language models with zero-init attention. *arXiv preprint arXiv:2303.16199*.

Zhang, Z.; Zhang, A.; Li, M.; and Smola, A. 2022. Automatic chain of thought prompting in large language models. *arXiv preprint arXiv:2210.03493*.

Zhang, Z.; Zhang, A.; Li, M.; Zhao, H.; Karypis, G.; and Smola, A. 2023b. Multimodal chain-of-thought reasoning in language models. *arXiv preprint arXiv:2302.00923*.

Zhou, D.; Schärli, N.; Hou, L.; Wei, J.; Scales, N.; Wang, X.; Schuurmans, D.; Bousquet, O.; Le, Q.; and Chi, E. 2022. Least-to-most prompting enables complex reasoning in large language models. *arXiv preprint arXiv:2205.10625*.