What Makes Multimodal Chain-of-Thought Matter in Complex Multimodal Reasoning Tasks?

By

Kaiwei Cen
Master of Science in Artificial Intelligence

Faculty of Science

Dr. Bruno Martins, University of Lisbon
Advisor

Dr. Denis Paperno, Utrecht University
Committee Member

Dr. Dong Nguyen, Utrecht University
Committee Member

July 29th, 2024

Date

What Makes Multimodal Chain-of-Thought Matter in Complex Multimodal Reasoning Tasks?

By

Kaiwei Cen

Advisor: Dr. Bruno Martins, University of Lisbon
and
Dr. Denis Paperno, Utrecht University

An abstract of
A thesis submitted to the
Faculty of Science
in partial fulfillment of the requirements for the degree of
Master of Science in Artificial Intelligence
in Faculty of Science
2024

Abstract

What Makes Multimodal Chain-of-Thought Matter in Complex Multimodal Reasoning Tasks?
By Kaiwei Cen

Multimodal chain-of-thought (M-CoT) reasoning has been increasingly applied to Vision and Language Models (VLM) in multimodal reasoning tasks, to improve their reasoning abilities. However, compared to the research that demonstrates the effectiveness of M-CoT, the explanation of why this strategy can improve the performance of VLMs still remains underexplored. In this work, we test whether M-CoT can improve the performance of VLMs on multimodal reasoning tasks, following zero-shot setting. We analyze the most likely patterns of M-CoT that contribute to improving the performance of VLMs, to find out why they can benefit the model's performance. We specifically designed different experiments to explore what is important to M-CoT's success. Our study shows that M-CoT can improve the accuracy of InstructBLIP 7B by 10.71%, and InstructBLIP 13B by 14.88%, on the ScienceQA benchmark. The M-CoT rationales that can improve the performance of InstructBLIP have information about the image and commonsense knowledge, which might help the model to perform better reasoning and answer the question more accurately. Whether the textual part of M-CoT is relevant to the question is important to the improvement of results with VLMs. The validity of the reasoning chain in the textual part of M-CoT can significantly affect the performance of VLMs on multimodal reasoning tasks. VLMs might rely on the conclusion of the textual part of the M-CoT rationale to make their decisions.

Acknowledgments

Thanks for the subversion of Dr. Bruno and Dr. Denis on this project. Thanks for the support, encouragement, and company from my friends Yue Pang, Huiting Zeng, and Siyu Wu. Thanks for the support from my wonderful grandmother. Thanks for the support from my family. Thanks for the guidance and company from my study advisor Sara O'Keeffe, and my psychologist Qingyu Lan. Thanks to the great philosophers in the East and the West such as Zen Buddhism and Mao, psychologists such as Dr. Jordan Peterson and Dr. Gabor Mate, and sociologists such as Ueno Chizuko and Simone de Beauvoir. Their wisdom across time and space, inspires me to face all the difficulties in this journey. Thanks to the information from Dr. Kees. Thanks to INESC-ID for providing accessible GPU resources. Also, thanks to Dr. Jelle and the Data Management Group for kindly providing the computing resources for this project. All these things help me to fight my demons and finally have a stronger mind and more independent soul.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

The development of Artificial Intelligence (AI) is gradually changing people's lives. For example, people can nowadays ask any kind of questions on the ChatGPT through their phones, finding correct answers to inquiries such as "What is the departure time of my flight to Los Angeles?", and "When will Apple release the next generation of iPhone?", more efficiently when compared to searching on internet search engines, or contacting the related individual. Besides, the use of ChatGPT can help people increase their productivity at work or study. For instance, people can ask ChatGPT to generate an email automatically, instead of writing by themselves. ChatGPT is built upon Large Language Models (LLM), which refer to models with billions of parameters. LLMs can understand the instructions humans give, from the way and the corpus they are trained on. The way people give instructions to ChatGPT is called "prompting". One of the most effective prompting strategies is to hint the model to solve the question step by step, for example, through sentences like "let's think step by step," or providing with several demonstrations that contain the reasoning steps to reach the final answer, before asking the actual question. The AI research community

called this specific prompting method Chain-of-Thought Prompting (CoT). By using CoT prompting, LLMs can have better performance across different reasoning tasks, as CoT can divide the problem into a set of simpler sub-problems, and guiding the model to solve them one by one.

1.2 Thesis statement

With scaling, Large Language Models (LLMs) are demonstrated to have emergent abilities that improve their performance in solving complex tasks [Zhao et al., 2023]. One of the emergent abilities is the ability to reason step by step [Zhao et al., 2023], which is shown by the Chain-of-thought (CoT) [Wei et al., 2023] prompting strategy. Chain-of-thought prompting has demonstrated its effectiveness in boosting the reasoning ability of LLMs via in-context learning strategies [Wei et al., 2023, Kojima et al., 2023] or fine-tuning [Ho et al., 2023].

Nowadays, towards the goal of developing Artificial General Intelligence (AGI), LLMs are integrated with vision encoders (also called Vision Large Language Models, VLMs) to have the ability to perceive and reason about multimodal inputs [Yin et al., 2023]. Inspired by the success of CoT prompting in LLMs, more and more studies have explored the application of CoT in the multimodal scenario [Lu et al., 2022, Zhang et al., 2023].

However, most of the previous studies about multimodal CoT have focused on: (1) the design of the chain in the prompt to make it effective in vision and language reasoning tasks [Rose et al., 2024, Himakunthala et al., 2023], and (2) leveraging CoT rationales to boost the reasoning abilities of VLMs in the training process [Lu et al., 2022, Zhang et al., 2023]. Compared to the amount of research on explaining the effectiveness of CoT via both theoretical [Prystawski et al., 2023] and empirical [Madaan and Yazdanbakhsh, 2022, Wang et al., 2023a] methods, there is less attention on exploring

whether CoT can replicate its success in the vision and language scenario, analyzing how VLMs leverage CoT in multimodal reasoning tasks. Zheng et al. [2023] conducted a detailed analysis of applying CoT on the GPT-3 model in multimodal reasoning tasks. Wu et al. [2023] explored whether GPT-4V [Yang et al., 2023] can have a large performance improvement by applying Multimodal CoT (M-CoT). Although M-CoT has demonstrated its effectiveness in GPT-4V, whether M-CoT can improve the performance of VLMs that are based on smaller LLMs compared to GPT-4V remains an open question, as well as why M-CoT matters and how VLMs leverage M-CoT.

Through this study, we aim to answer the following research questions: ***Can M-CoT improve the performance of VLMs in the complex multimodal reasoning task? If yes, which aspect of M-CoT contributes to the performance gain?*** To achieve this, we first test whether M-CoT can improve the performance of InstructBLIP [Dai et al., 2023] on the ScienceQA [Lu et al., 2022] benchmark in zero-shot setting. Then, we explore the important parts of M-CoT. Inspired by Wang et al. [2023a], we speculate that two specific components might be the most important aspect of M-CoT, namely: (1) the relevance between the textual part of M-CoT rationale with the question, and (2) the validity of the reasoning part in the textual part of M-CoT rationale. We design specific ablation experiments to test these two aspects.

1.3 Contributions

The main contributions of our study are as follows:

(1) We find that M-CoT can largely improve the performance of VLMs in complex multimodal reasoning tasks, and the improvement is proportional to the model scale. Compared with a simple zero-shot result, M-CoT can improve the accuracy of In-

structBLIP 7B by 10.71% and InstructBLIP 13B by 14.88%. M-CoT is effective in improving the accuracy across different subjects and difficulty levels of science questions in the ScienceQA dataset.

(2) We observe that the information about the image and the commonsense knowledge provided in the prompt both contribute to boosting the reasoning ability of the model to answer challenging questions, such as questions that need to extract complex information from the images and require commonsense knowledge.

(3) Our experiments also demonstrate that whether the textual part of M-CoT is relevant to the question is important to the improvement of M-CoT. Besides, we find that whether the reasoning chain of the textual part of M-CoT is logically valid can affect the performance of VLMs on complex reasoning tasks. We hypothesize that VLMs rely on the conclusion of the textual part of the M-CoT rationale to select the answer choice.

Chapter 2

Background

The important AI concepts that are associated to our project

This section introduces important concepts related to our project, namely Large Language Models (LLMs), Vision and Large Language Models (VLMs), few-shot prompting, zero-shot prompting, and chain-of-thought prompting.

2.1 Large Language Models

Large Language Models (LLM) refer to language models that have hundreds of billions of parameters and are trained on massive text corpus, such as GPT-3 [Brown et al., 2020], PaLM [Chowdhery et al., 2022], and LLaMA [Touvron et al., 2023]. LLMs have a strong capacity for language understanding and task-solving via text generation. [Zhao et al., 2023].

LLMs are mainly built upon the Transformer architecture [Vaswani et al., 2023], where multi-head attention layers are stacked in a very deep neural network. Existing LLMs use similar Transformer architectures and pre-training objectives as smaller language

models such as GPT-2. The difference between LLMs and smaller language models is that LLMs significantly extend their model size, data size, and total compute (orders of magnification). Extensive research has shown that scaling can improve model capacity by a large margin [Chowdhery et al., 2022, Brown et al., 2020, Radford et al., 2019]. [Zhao et al., 2023].

In this section, we first discuss the emergent abilities of LLMs related to our project, and then we briefly introduce GPT-3, which is a representative example.

2.1.1 Emergent Abilities of LLMs

Wei et al. [2022b] define emergent abilities of LLMs as "the ability is not present in small models but arises in large models", which is one of the most prominent features that can distinguish LLMs from previous Pre-trained Language Models (PLMs). When a new ability emerges, a notable characteristic is that the performance rises by a large margin compared to the random baseline when the scale is up to a enough level. LLMs have three typical emergent abilities: in-context learning, instruction following, and step-by-step reasoning. [Zhao et al., 2023]. Here, we briefly discuss in-context learning and instruction following. Step-by-step reasoning will be discussed in detail in Section 2.4.

In-context Learning

The In-Context Learning (ICL) ability is formally defined in the GPT-3 paper [Brown et al., 2020]: assuming the language model is provided with a task description and/or several task demonstrations, it can generate the desired output for test instances by completing the word sequences in the given input, without updating parameters through gradient computations or requiring additional training. Among the GPT series, the 175B GPT-3 is demonstrated to have a strong ICL ability, while GPT-1 [Radford and Narasimhan, 2018] and GPT-2 [Radford et al., 2019] did not. [Zhao

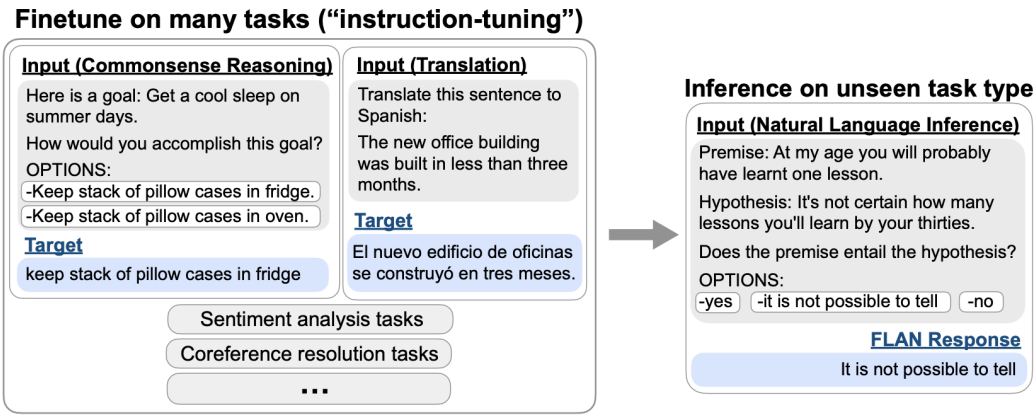


Figure 2.1: An overview of instruction tuning, as shown by Wei et al. [2022a].

et al., 2023].

Instruction Following

By fine-tuning on different tasks that are formatted via natural language instructions (a process called instruction tuning), LLMs can perform well on unseen tasks that are also formatted via natural language instructions [Sanh et al., 2022, Ouyang et al., 2022, Wei et al., 2022a]. With instruction tuning, LLMs can follow the task instructions of unseen tasks without explicit examples. Therefore, the generalization ability of LLMs is improved. As an early example, the instruction-tuned LaMDA-PT model [Thoppilan et al., 2022] could significantly outperform the untuned model on unseen tasks when the model scale reaches 68B, but not for 8B or smaller models [Wei et al., 2022a]. [Zhao et al., 2023]. Figure 2.1 presents an overview of instruction tuning.

2.1.2 GPT-3

GPT-3 [Brown et al., 2020] was released in 2020 with a model scale of 175B. GPT-3's paper formally introduced ICL, which utilizes LLMs in a few-shot or zero-shot way. ICL can teach LLMs to understand the tasks in the form of natural language text. With ICL, the pre-training and utilization of LLMs converge to the same language

modeling paradigm: pre-training requires the model to predict the text sequences conditioned on the context, while ICL requires the model to predict the correct task solution, which can be formatted as text sequences, conditioned on the given task demonstrations and descriptions. GPT-3 performs excellently on various NLP tasks and specially designed tasks requiring reasoning abilities and domain adaption. GPT-3 can be viewed as a remarkable capstone in the evolutionary process from PLMs to LLMs. It has empirically demonstrated that scaling the model to a significant size can largely improve its capacity. [Zhao et al., 2023].

2.2 Large Vision and Language Models

A large Vision and Language Model (VLM) is a kind of Multimodal Large Language Model (MLLM), which refers to an LLM-based model that has the ability to receive and reason with multimodal information [Yin et al., 2023]. The typical architecture of VLMs involves a vision encoder that is used to extract image features, a part that is designed to bridge the visual and textual modalities, and an LLM that is used to generate the desired output. Here, we focus on discussing the way of bridging different modalities in VLMs, which is regarded as one of the key questions to implementing VLMs, and the learning paradigm of CoT in VLMs, which is related to our project.

2.2.1 Modality Bridging

How to bridge the vision and language modalities effectively is one of the most important questions in the multimodal machine learning area, and the key to transferring LLM's successes to VLMs. There are broadly two ways of bridging text and vision modalities: (1) introducing a learnable interface between the vision encoder and the LLM, and (2) using an expert model to transform the visual input into textual descriptions, which are then fed to the LLM. [Yin et al., 2023].

Learnable Interfaces

A learnable interface is responsible for connecting different modalities when the parameters of the pre-trained encoders are frozen [Yin et al., 2023]. For example, BLIP-2 [Li et al., 2023] uses a set of learnable queries to extract information from a frozen image encoder, and then feeds the most useful visual information to the LLM to generate the desired text. LLavA [Liu et al., 2023] uses a simple linear layer to project the image features into word embedding space.

Expert Models

Introducing an expert model to convert the visual input into textual descriptions can also bridge different modalities. Compared to implementing a learnable interface, using an expert model does not require additional training [Yin et al., 2023]. For example, ScienceQA used an image captioning model to convert the image into a caption that provides the visual semantics to the language model. However, this method might cause information loss in the captioning process [Zhang et al., 2023].

2.2.2 Learning Paradigm of CoT in VLMs

There are broadly three ways to apply CoT on VLMs, namely fine-tuning, few-shot prompting, and zero-shot prompting [Yin et al., 2023].

As an example of the fine-tuning approach, MM-CoT [Zhang et al., 2023] is fine-tuned on the ScienceQA dataset, which is designed to boost the multimodal reasoning ability of VLMs. MM-CoT first generates the CoT rationales and then predicts the final answer based on the rationales.

For the case of few-shot CoT prompting, Chameleon [Lu et al., 2023] applied few-shot CoT prompting on the ScienceQA benchmark, reporting new state-of-the-art results in the few-shot setting.

Finally, as an example of zero-shot CoT prompting, Wu et al. [2023] tested the CoT

reasoning ability of GPT-4V in a zero-shot scenario. The authors tested two settings: (1) asking the model to generate the image description and final answer in one prompt, and (2) separating the process of generating the image description and the procedure of answer inference.

2.3 Few-shot Prompting and Zero-shot Prompting

Few-shot prompting [Brown et al., 2020] refers to the setting where a model is given a few demonstrations and conditioned on these demonstrations to get the output, without updating its parameters. Figure 2.2 shows an example of few-shot prompting. For a typical dataset that contains a context and completion (e.g., English text and a French translation), few-shot prompting works by giving K exemplars that contain context and the corresponding completions, and then one final context that expects the model to provide a completion. K is typically set in the range of [2,100], as the number of exemplars in this range can fit in the context window of typical models (e.g. 2048 tokens). The major advantage of few-shot prompting relates to reducing the labor and cost of constructing a specific dataset for fine-tuning the model on a specific task. Still, there are also several disadvantages in the use of few-shot prompting. With the development of LLMs, few-shot prompting demonstrated its effectiveness by outperforming the fine-tuning models by a large margin [OpenAI et al., 2023]. However, in some tasks, the performance of few-shot prompting is still worse than that of fine-tuning state-of-the-art models. Moreover, few-shot prompting still involves the labor and cost to construct exemplars.

Zero-shot prompting [Brown et al., 2020] is similar to few-shot prompting, except that in this case one only provides a natural language task description, and no demonstrations are given to the model. Figure 2.2 also shows an example of zero-shot prompting. The advantages of this method include: (1) providing the maximum

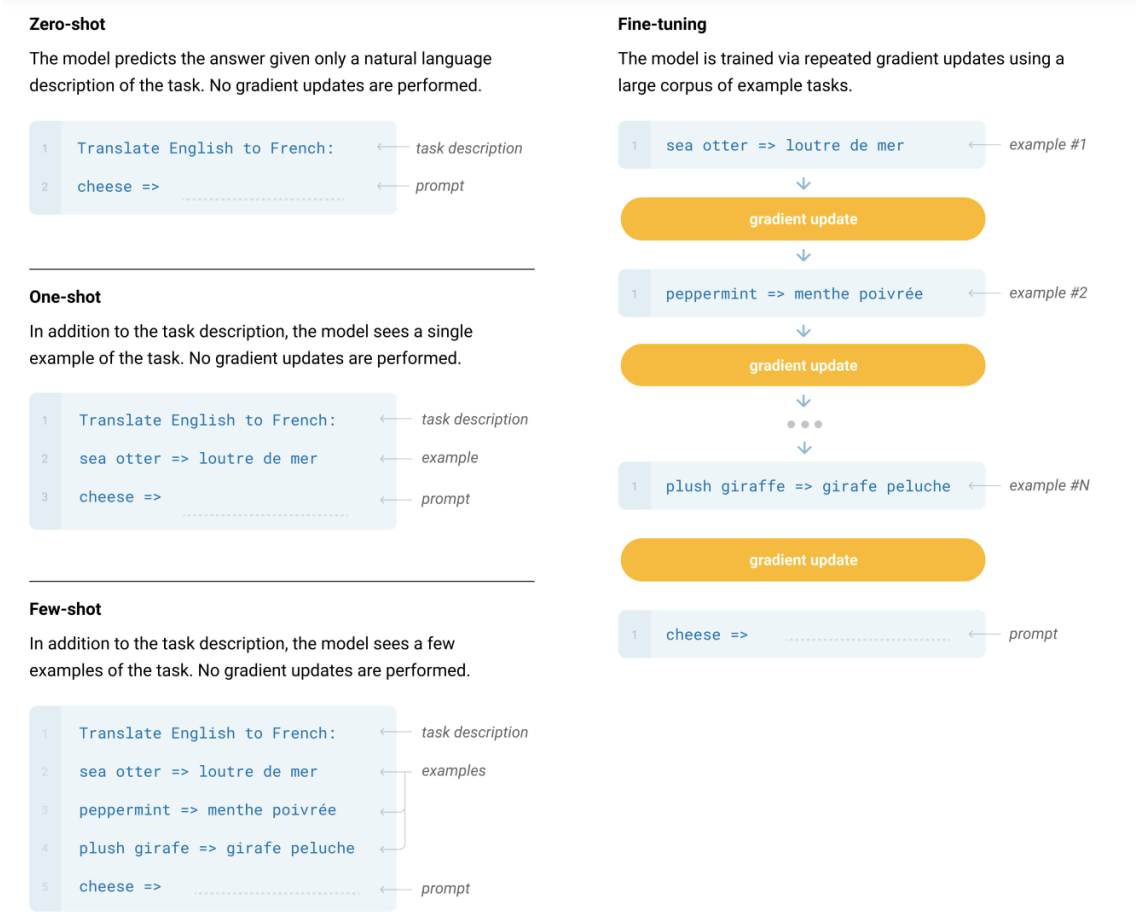


Figure 2.2: Examples of few-shot prompting, one-shot prompting, zero-shot prompting, and traditional fine-tuning. The figure is originally from Brown et al. [2020].

convenience and the potential to be robust, and (2) mimicing how humans perform tasks in some settings (for example, a human naturally knows how to perform based on the only given task instruction in the translation example in Figure 2.2.) The main disadvantage of zero-shot prompting is that it is the most challenging setting, as LLMs can only use the knowledge gained from pre-training, and for some tasks it is difficult to understand what is desired only from the task description.

2.4 Chain-of-thought Prompting

According to Wei et al. [2023], a chain of thought is a series of consistent reasoning steps. Chain-of-thought prompting (CoT) is a prompting method that either provides a chain-of-thought as a rationale before the final answer in each few-shot exemplar, or uses a trigger sentence to prompt the LLM to generate a chain-of-thought as a rationale, to help reaching the answer in reasoning tasks. Chain-of-thought prompting is purposed to tackle the challenge for LLMs to perform reasoning tasks, and has demonstrated its success by outperforming standard prompting on different reasoning tasks. There are two main chain-of-thought prompting methods, namely few-shot chain-of-thought prompting (Few-shot-CoT) and zero-shot chain-of-thought prompting (Zero-shot-CoT), these will be discussed in detail in the following subsections. Figure 2.3 shows examples of standard zero-shot prompting, standard few-shot prompting, Few-shot-CoT, and Zero-shot-CoT.

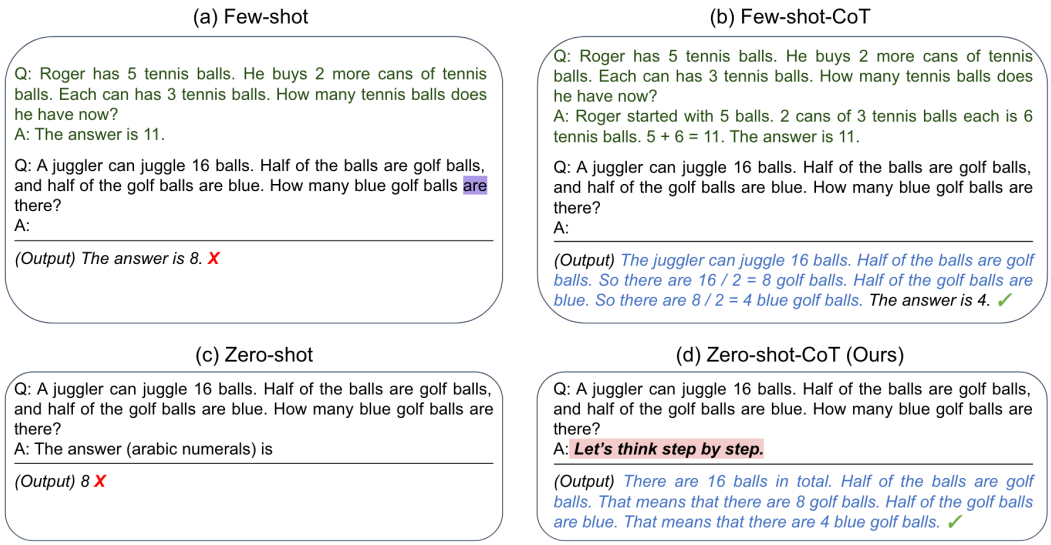


Figure 2.3: Examples of few-shot prompting, few-shot-CoT prompting, zero-shot prompting, and zero-shot-CoT prompting. Few-shot-CoT and Zero-shot-CoT instruct the model to generate consistent rationales before the answer. The figure is originally from Kojima et al. [2023].

2.4.1 Few-shot Chain-of-thought Prompting

Few-shot chain-of-thought prompting [Wei et al., 2023] provides chain-of-thought reasoning demonstrations in the few-shot in-context exemplars. Compared to fine-tuning methods and to the standard few-shot prompting method, Few-shot-CoT has the following advantages: (1) no need to create a high-quality fine-tuning dataset, which can be costly, and (2) it can improve the performance of LLMs across different reasoning tasks. Few-shot-CoT brings significant improvements to sufficiently large LLMs in symbolic reasoning tasks, commonsense reasoning tasks, and arithmetic reasoning tasks, compared to standard prompting. Few-shot-CoT can also facilitate length generalization when LLMs are evaluated on tasks that require more reasoning steps than in the few-shot exemplars they saw. The improvement of using Few-shot-CoT is consistent with the scaling law, which states that bigger LLMs can have more improvement gains compared to smaller LLMs, and smaller LLMs might not have large or positive improvements by using Few-shot-CoT.

2.4.2 Zero-shot Chain-of-thought Prompting

Kojima et al. [2023] proposed zero-shot chain-of-thought prompting, which involves prompting an LLM with a trigger sentence such as "let's think step by step," instead of providing chain-of-thought rationales in the few-shot exemplars, to make the LLM generate the chain-of-thought rationales. Zero-shot-CoT outperforms zero-shot standard prompting on four of six arithmetic reasoning tasks (MultiArith [Roy and Roth, 2015], GSM8K [Cobbe et al., 2021], AQUA [Goswami et al., 2024], and SVAMP [Patel et al., 2021]), symbolic reasoning tasks (i.e. Last Letter and Coin Flip), and other reasoning tasks (Date Understanding and Shuffled Objects). Zero-shot-CoT can largely improve the performance of LLMs on Last Letter, GSM8K, and MultiArith compared to standard zero-shot prompting. Compared to Few-shot-CoT, Zero-shot-CoT can be applied to LLMs directly and one does not need to define the few-shot

exemplars based on the specific task. Besides, Zero-shot-CoT can be used as a simple way to probe the internal reasoning ability of LLMs gained by the per-training process.

2.4.3 Diverse Reasoning

Increasing the diversity of CoT reasoning paths can improve the performance of LLMs on reasoning tasks. Self-consistency is a simple yet effective strategy to increase the performance of applying CoT on LLMs by increasing the diversity of generated reasoning paths. As described by Wang et al. [2023b], the self-consistency method makes the model generate a set of different reasoning paths by replacing greedy decoding with a temperature sampling method during the process of generating the CoT rationale. After that, the model uses the different CoT paths to generate predictions, and selects the most consistent answer as the final prediction. Compared to using greedy decoding to generate only one reasoning path in LLMs, self-consistency can significantly improve the performance of LLMs on arithmetic reasoning tasks, and can improve the performance on commonsense reasoning tasks by 2-5%.

Fine-tune-CoT [Ho et al., 2023] uses a similar idea to self-consistency, first generating CoT rationales from very large teacher models and then using these CoT rationales to fine-tune a small student model to enable small LMs to perform complex reasoning tasks. Fine-tune-CoT generates diverse CoT reasoning paths to maximize the teaching effects on the small LMs. The experimental results show that Fine-tune-CoT with diverse reasoning can significantly outperform the Few-shot-CoT baseline on the MultiArith and SVAMP benchmarks. Also, smaller models applied with Fine-tune-CoT with diverse reasoning can outperform larger models with Few-shot-CoT, which demonstrates the effectiveness of increasing the diversity of reasoning paths to improve the reasoning abilities of small LMs.

Chapter 3

Related Work

The studies that explore similar research questions to ours

This section introduces previous research focused on explaining CoT, and on analyzing M-CoT on multimodal reasoning tasks. We only introduce studies that are closely related to our research project.

3.1 Explaining Chain-of-Thought Reasoning

Chain-of-Thought (CoT) reasoning has clearly demonstrated its success in LLMs [Wei et al., 2023]. However, the reasons to why reasoning step by step to produce an explanation can improve the performance of LLMs on reasoning tasks still remains a mystery.

Wang et al. [2023a] conducted an empirical study to know which parts are the key aspects of CoT reasoning. They ablated different aspects of CoT to see how the changes affect the model's performance. First, they speculated that the validity of the reasoning chain is the most important part of CoT. If the reason provided in

Arithmetic Reasoning	Multi-hop QA
Q: Leah had 32 chocolates and her sister had 42. If they ate 35, how many pieces do they have left in total?	Q: Who is the grandchild of Dambar Shah?
A: Originally, Leah had 32 chocolates and her sister had 42. So in total they had 32 + 42 = 74. After eating 35, they had 74 - 35 = 39 pieces left in total. The answer is 39.	A: Dambar Shah (? - 1645) was the father of Krishna Shah. Rudra Shah was the child of Krishna Shah (? - 1661). So the final answer (the name of the grandchild) is: Rudra Shah.

Figure 3.1: Examples of language templates and bridging objects for CoT rationales in arithmetic reasoning and multi-hop QA tasks. This figure was originally presented by Wang et al. [2023a].

the few-shot demonstrations is not logically valid, the model is not taught to reason correctly, which might cause the performance to decrease. They constructed invalid reasoning chains by changing the bridging objects and language templates of the valid reasoning chains. The bridging objects are the necessary objects that need to be traversed to reach the accurate answer. For instance, in an arithmetic reasoning task, the bridging objects are the numbers and equations; while for factual QA, the bridging objects are the subject and object entities. The language templates are the complementary parts of the bridging objects, which are the textual hints or the predicates/relations that need to be traversed to reach the final answer. Figure 3.1 shows examples of the language templates and bridging objects.

The experimental results show that the performance of Instruct- GPT [Ouyang et al., 2022] using invalid CoT rationales can achieve almost 90% of the performance of using valid CoT rationales on GSM8K [Cobbe et al., 2021] and Bamboogle [**?**] evaluated both intrinsically and extrinsically. The qualitative analysis also shows that the generated rationales are not distinguishable between using valid and invalid CoT rationales. Therefore, the authors concluded that the validity of the CoT rationale is not the key to the effectiveness of few-shot CoT prompting.

They also analyzed the constructed invalid CoT rationales and observed that they still

leverage information from the query, as the CoT rationales start with the bridging objects mentioned in the question, and the language templates are related to the question. Therefore, the authors speculated that relevance might be the key aspect of CoT rationales. However, noting that each step in the invalid CoT rationales follows the previous step, they also hypothesized that coherence might be the key aspect of CoT rationales. The authors conducted an ablation study on four combinations of relevance, coherence, bridging objects, and language templates, to test these two hypotheses. To ablate relevance, they changed the bridging objects or the language template of the CoT rationale by random substitutions. To ablate coherence, they shuffled the bridging objects or language templates of the invalid CoT rationales, permuting their orders. A setting of no relevance was also studied, in which the provided CoT rationales are irrelevant to the question, and similarly also a setting of no coherence. Figure A.1 in the Appendix shows examples of the ablation settings.

The experimental results revealed that relevance and coherence are the keys to CoT rationales, as the performance decreased in all ablation settings compared to using valid CoT rationales. The performance of the no relevance is the lowest in all ablation tests, so the authors concluded that it is crucial to keep relevance in the CoT rationales. They also observed that relevance matters more than coherence for bridging objects, while coherence matters more than relevance for language templates.

Overall, the study reveals that the knowledge about how to reason properly that LLMs learn from few-shot CoT demonstrations is limited. LLMs rely more on the knowledge gained from the pre-training process when performing complex reasoning tasks, than learning from the few-shot exemplars. The few shot CoT exemplars only formalize the output space of LLMs, which lets them generate rationales step by step and is relevant to the question. Besides, this study can also be regarded as a way to approximately qualify the prior knowledge that LLMs need in order to perform complex reasoning tasks. Limitations of the study include: (1) The experimental

design of ablation studies cannot be directly applied to reasoning tasks that are highly template-based, as the reasoning steps in these tasks are very similar within each example and across different examples. For example, for the last letter concatenation task that asks models to concatenate the last letters of a given sequence of words (e.g., "Barbara Betsy" $\rightarrow$ "ay"), each step in the CoT rationale has the form of "the last letter of X is Y", where X refers to the words in the sequence, and Y is the last letter of the corresponding word, except the last step. Language templates are the same and have no sense of order among the steps, so the ablation setting cannot be applied here. (2) The authors manually wrote the invalid CoT rationales and claimed that synthesizing these rationales automatically is challenging due to the informal nature of the tasks they experimented on.

Similar to Wang et al. [2023a], Madaan and Yazdanbakhsh [2022] leveraged the idea of altering a particular aspect of the in-context CoT exemplars, comparing the performance disparity of an LLM on the altered CoT demonstrations and the original ones, to explore whether the altered aspect matters in CoT demonstrations. The difference between the studies from Madaan and Yazdanbakhsh [2022] and Wang et al. [2023a] is that Madaan and Yazdanbakhsh [2022] used counterfactual prompting. They explored whether three major semantic components of a prompt are the key to CoT's efficacy: symbols, patterns, and text. Symbols are defined as the tokens the model needs to traverse in order to reach the final answer. Patterns are defined as either composition of symbols and operators, or a prompt structure that reinforces task understanding. Finally, text refers to the tokens that are neither symbols nor patterns in the prompt. The experimental results reveal that: (1) The exact type and value of symbols are primarily immaterial to the model performance. (2) The symbiosis relationship between patterns and text plays a vital role in CoT's success. Patterns help the model generate meaningful intermediate text, hinting how the model should form connections between different clauses in the intermediate text, and eliciting the

model to derive knowledge and reach the accurate answer. Text imbues patterns with commonsense knowledge, which assists the model to solve a task.

3.2 Analysis of Chain-of-Thought in Multimodal Reasoning Tasks

Most previous work on analyzing CoT focused on the text modality [Madaan and Yazdanbakhsh, 2022, Wang et al., 2023a, Prystawski et al., 2023]. With the development of MLLMs, more and more studies payed attention to the utilization of multimodal CoT in multimodal reasoning tasks [Zheng et al., 2023, Wu et al., 2023, Lu et al., 2022, Zhang et al., 2023].

Zheng et al. [2023] first conducted an in-depth analysis of CoT on multimodal reasoning. They observed that even when providing sufficient image information to GPT-3 by captioning the image, GPT-3 still cannot answer questions about the image correctly. When providing a CoT rationale that contains external commonsense knowledge and the image analysis based on the question, GPT-3 can generate the accurate answer. Therefore, the authors concluded incorporating CoT as the input can facilitate the reasoning abilities of GPT-3 in multimodal tasks. They also found that LLMs tend to reason in line with the input CoT rationale in the zero-shot prompting setting. Inaccurate rationales in the zero-shot prompts will lead LLMs to reach a wrong answer. The authors also explored how well LLMs can leverage the vision and language information in the prompt for multimodal reasoning tasks. They observed that if just giving a unimodal prompt to GPT-3, the performance of GPT-3 is poor. When captioning the image and integrating the generated description with the question, GPT-3 will generate rationales with hallucinations about the image content, as the generated image description by the captioning process might lack some necessary information about the image. Differently from this work, our study analyzes applying

M-CoT in VLMs, which might solve the problem of the information loss caused by the captioning process, as the vision encoder of VLMs can extract more effective visual features compared to just captioning the image. Also, VLMs are trained or instruction tuned over multimodal datasets, which can make them better to jointly understand and leverage the information of the image and text of the prompt to solve multimodal reasoning tasks.

Wu et al. [2023] use the Winoground [**?**] as a probing task to test whether M-CoT can improve the performance of VLMs. Winoground is a dataset designed to test the compositional reasoning ability of vision and language models. This task requires models to select the caption that matches the image from two given captions or vice versa. For applying CoT on GPT-4V, the authors provided a prompt that asks the model to generate the description of the image, and then lets the model make the decision based on the generated image description. The experimental results show that for the setting of asking the model to choose the caption that describes the image from the provided two captions, GPT-4V with CoT can improve the performance by 6% compared to GPT-4V without CoT. For the setting of asking the model to select the image that better aligns with the description from the provided two images, making GPT-4V generate the description before the prediction can improve the performance by 18.5%, compared to directly predicting the answer. Therefore, the authors concluded that applying CoT can largely improve the performance of classifying the image that better aligns with the description, and can largely fill the gap between classifying the caption that better aligns with the image. The authors also tested the setting of two-turn prompts, which is used to divide the recognition and the reasoning process. They first prompt GPT-4V to generate the image description, and then prompt GPT-4V with the generated description, the question, and the instruction, to ask the model to provide the analysis. The experimental results showed that the two-turn prompt can improve the performance by 4.75% compared to GPT-4V with CoT.

The authors also conducted an error analysis based on Winoground's tagging categorization. The error analysis showed that GPT-4V with the two-turn CoT prompt has difficulty in discerning abstract attributes such as size, amount, and weight. Also, it is challenging for GPT-4V to handle instances classified into the tag categories "Series, Pragmatics, Object-Centric Spatial, and Temporal" in Winoground. Similarly to this study, we also test whether M-CoT can improve the performance of vision and language tasks. The differences between the work from Wu et al. [2023] and ours are: (1) They use GPT-4V while we use InstructBLIP on the ScienceQA benchmark. Although GPT-4V might perform much better than InstructBLIP based on its pre-training knowledge, GPT-4V is not open-sourced. Thus, it is hard to interpret how GPT-4V responds to the multimodal CoT prompt, while this can be easily done in open-sourced VLMs by looking into their attention layers. (2) They use Winoground as the probing task to explore how M-CoT affects the performance of GPT-4V, noting that Winoground is designed to test the compositional reasoning ability of vision and language models. Compared to the ScienceQA benchmark used in our experiments, Winoground can only test basic vision and language reasoning abilities, such as classifying the spatial relationship of objects. ScienceQA needs external commonsense knowledge to infer the answer, which makes it more challenging than Winoground.

Chapter 4

Experiments

This chapter discusses our experimental design starting with a test that aims at assessing whether M-CoT can improve the performance of VLMs, and then presenting tests that explore why CoT can improve the performance of VLMs. For each experiment, we discuss the dataset, the backbone VLM, the method, the experimental steps, and the evaluation method in detail.

4.1 Testing Whether M-CoT Can Improve the Performance of Vision LLMs

Our first experiment explores whether zero-shot CoT prompting can improve the performance of vision LLMs on multimodal reasoning tasks. We apply zero-shot CoT prompting to InstructBLIP on the ScienceQA benchmark.

4.1.1 Dataset

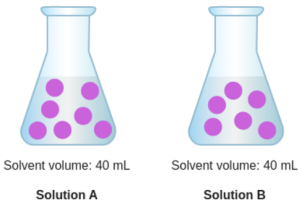
Science Question Answering (ScienceQA) [Lu et al., 2022] is a large-scale multimodal dataset that consists of multi-choice science questions with explanations. ScienceQA has 21,208 examples with lectures and explanations across three different subjects:

natural science, social science, and language science. Questions in each subject can be first categorized by topics (Biology, Earth Science, etc.), secondly categorized by categories (genes to traits, classification, etc.), and thirdly categorized into skills (use a chemical formula to describe the molecule, select object corresponding to a liquid, etc.). There are 26 topics, 127 categories, and 379 specific skills in ScienceQA. The examples in ScienceQA are collected from elementary and high school curricula. Each example has a question, multimodal contexts, multiple choices, lecture, explanation, and a correct answer. The lecture provides external commonsense knowledge, while the explanation provides a specific reason to reach the answer. Figure 4.1 shows an example of the ScienceQA dataset. The most important goal of implementing ScienceQA was to aid the development of models that can generate high-quality CoT rationales that mimic the multi-hop reasoning process when answering science questions.

The main reason for using the ScienceQA dataset in our experiments relates to

Question: Which solution has a higher concentration of pink particles?

Context: The diagram below is a model of two solutions. Each pink ball represents one particle of solute.



Choices

- neither; their concentrations are the same
- Solution B
- Solution A

Answer: Solution A

Lecture: A solution is made up of two or more substances that are completely mixed. In a solution, solute particles are mixed into a solvent. The solute cannot be separated from the solvent by a filter. For example, if you stir a spoonful of salt into a cup of water, the salt will mix into the water to make a saltwater solution. In this case, the salt is the solute. The water is the solvent. The concentration of a solute in a solution is a measure of the ratio of solute to solvent. Concentration can be described in terms of particles of solute per volume of solvent. concentration = particles of solute / volume of solvent

Solution: In Solution A and Solution B, the pink particles represent the solute. To figure out which solution has a higher concentration of pink particles, look at both the number of pink particles and the volume of the solvent in each container. Use the concentration formula to find the number of pink particles per milliliter. Solution A has more pink particles per milliliter. So, Solution A has a higher concentration of pink particles.

Figure 4.1: An example of the ScienceQA dataset.

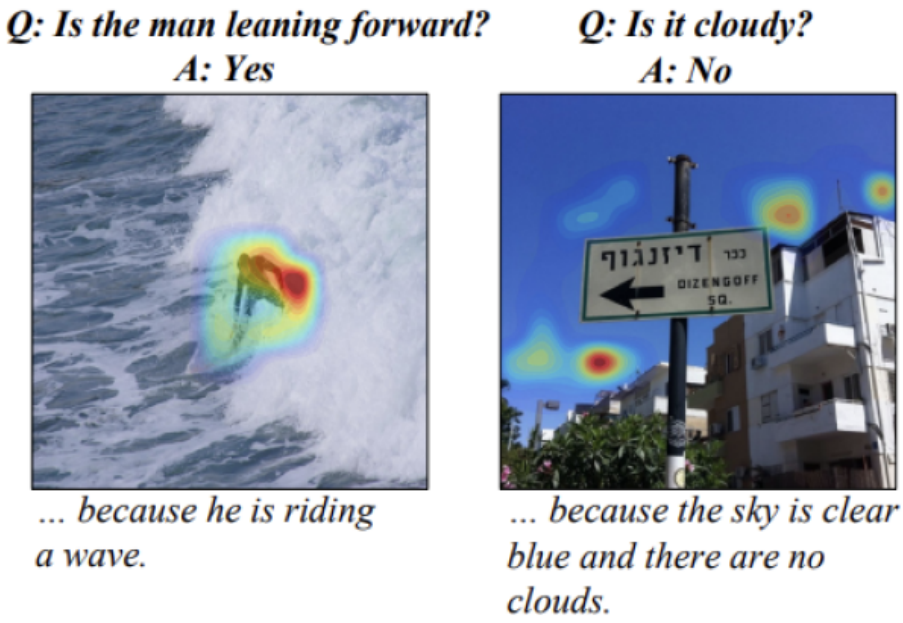


Figure 4.2: Two examples of the VQA v2 dataset. Each example contains a question, an image, an answer, and an explanation of how to reach the answer. The figure is adapted from Park et al. [2018].

the fact that this is the only open-sourced multimodal dataset designed to test the multi-hop reasoning ability of vision LLMs. In other common VQA datasets, such as VQA v2 [Goyal et al., 2017], the questions are too simple to require LLMs to use consistent CoT rationales to reach the answer. Figure 4.2 shows examples of the VQA v2 dataset.

4.1.2 Zero-shot Prompting and Zero-shot CoT Prompting

In the first experiment to test whether M-CoT can improve the performance of VLMs on the ScienceQA multimodal reasoning task, we apply zero-shot prompting to the InstructBLIP model as the baseline, and then use zero-shot CoT prompting for a comparison. We tried some prompts to make the performance gap between the zero-shot results and the zero-shot CoT results as large as possible. Among the prompts we tried, we observed that the prompt that can make the largest performance gap has the following features: (1) include hints to the model about the different components

in the prompt, referring to them explicitly and as clearly as possible, using "Context:", "Question:", "Choices:", "Solution:", and "Answer:" before the corresponding textual parts; (2) use a sentence like "Let's think step by step" after "Solution:" to let the model know that the given rationales use a multi-step reasoning strategy; (3) use a structure that can be read by human naturally. Although using "{}" to include the context, questions, and choices can have the highest zero-shot result, using "{}" can make the textual instruction unnatural and not like the text humans normally read. Therefore, we deleted "{}" to make the prompt have a structure more similar to human text, and observed that not using "{}" can make the performance gap larger compared to having these symbols. Figure 4.3 shows an example of the prompts we use in the zero-shot setting and an example of the prompts we use in the zero-shot CoT setting.

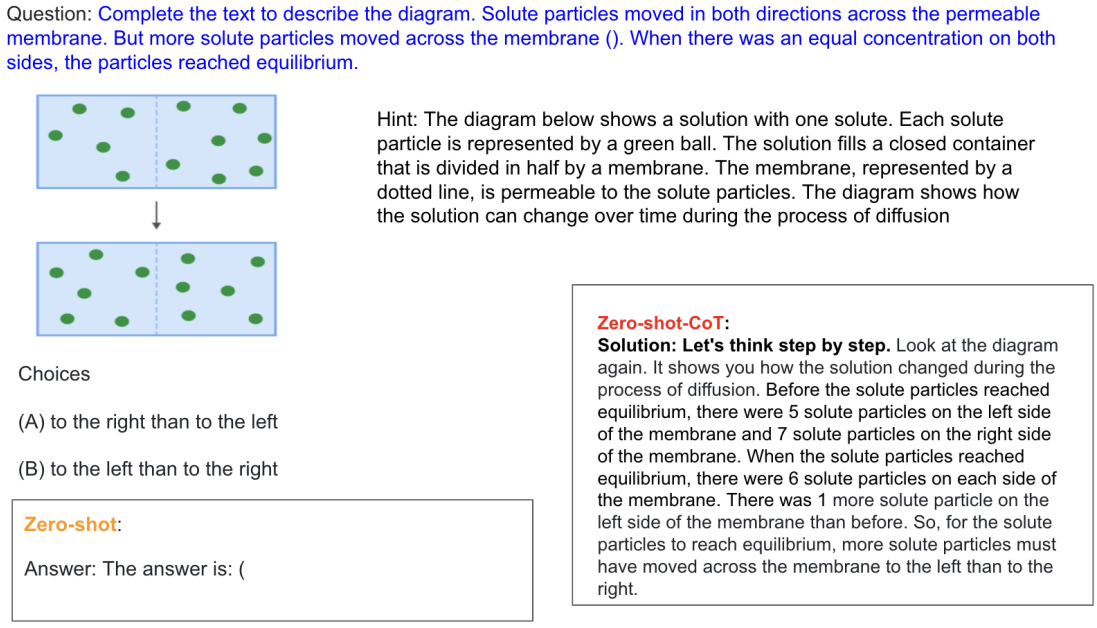


Figure 4.3: An example prompt used in the zero-shot setting, and an example prompt used in the zero-shot CoT setting.

4.1.3 Backbone Vision LLM

Introduction of InstructBLIP

InstructBLIP [Dai et al., 2023] is a general-purpose instruction-tuned vision and language model that is purposed to tackle the challenge of applying instruction tuning on vision and language tasks, and to make a unified vision and language model that can follow arbitrary instructions.

InstructBLIP consists of a frozen image encoder, a Query Transformer (Q-Former), and a frozen LLM. The Q-Former bridges the frozen image encoder and the frozen LLM. The frozen image encoder extracts visual features. A set of learnable query embeddings is set in the Q-Former to interact with the extracted visual features in the cross-attention layer. The output of the Q-Former is linearly projected to the LLM through a fully-connected layer and is prepended before the textual input. Figure 4.4 shows the architecture of InstructBLIP.

InstructBLIP is pre-trained in two stages. In the first stage, the frozen image encoder and the Q-Former are pre-trained with image-text pairs and three objectives (image-text contrastive learning, image-grounded text generation, and image-text matching). The goal of the first stage of pre-training is to extract the image features most rele-

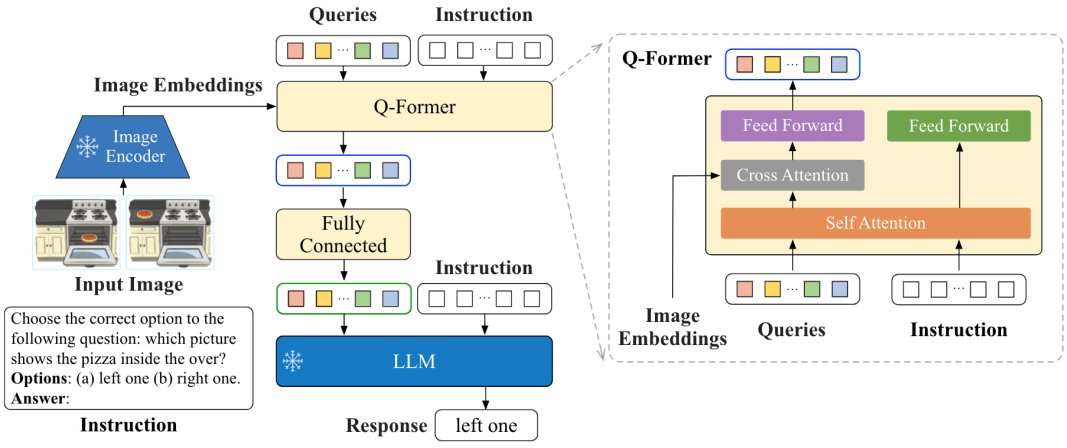


Figure 4.4: The model architecture of InstructBLIP

vant to the text. In the second pre-training stage, the Q-Former and the LLM are pre-trained together to gain the text generation ability conditioned on the soft visual prompt. InstructBLIP is instruct-tuned by 13 held-in datasets across different vision and language task categories, such as image captioning, image question answering, and image question generation. During the instruction-tuning procedure, the task instructions are used as additional input to the Q-Former, and then interact with the query embedding through the self-attention layer. This encourages the Q-Former to extract the visual features relevant to the given task, and the LLM can receive this task-related visual information.

InstructBLIP achieved state-of-art zero-shot performance on 13 held-out vision and language datasets. The zero-shot result of InstructBLIP outperforms its base architecture BLIP-2 across different LLMs, demonstrating the effectiveness of vision and language instruction tuning. InstructBLIP outperformed the previous state-of-the-art in the video question-answering task, even without video training data. This can demonstrate the zero-shot generalization capability of InstructBLIP on unseen tasks.

The reason for using InstructBLIP

Our experiments used versions of InstructBLIP in which the LLM is either Vicuna 7B or 13B [Chiang et al., 2023]. The reasons are listed as follows: (1) InstructBLIP can extract visual features which are task-related by fine-tuning the Q-Former with instruction tuning. This makes the zero-shot result of InstructBLIP outperform the zero-shot result of Flamingo-80B [Alayrac et al., 2022] on seven multimodal benchmarks [Dai et al., 2023]. (2) Some vision LLMs claim they applied M-CoT to their architectures and had the performance improvement, but these models are not open-sourced, such as Google's PaLM-E [Driess et al., 2023], Microsoft's KOSMOS-1 [Huang et al., 2023]. InstructBLIP is open-sourced and can thus be more easily adapted to perform CoT reasoning. (3) The zero-shot accuracy of InstructBLIP

with Vicuna-13B on the ScienceQA benchmark is 63.1%, and the zero-shot accuracy of InstrucBLIP with FlanT5 XXL [Chung et al., 2022] is 70.6% [Dai et al., 2023].[1] Although InstrucBLIP with FlanT5 XXL outperforms InstructBLIP with Vicuna-13B on the ScienceQA benchmark in the zero-shot setting, we still use InstructBLIP with Vicuna because this version is superior at open-ended generation tasks [Dai et al., 2023], and thus we think it can generate more accurate predictions with M-CoT rationales than FlanT5 XXL. Also, Vicuna allows using longer prompts than FlanT5.

4.1.4 Experimental Steps

We apply zero-shot prompting and zero-shot CoT prompting on InstructBLIP and compare the performance of the two settings. We sample a subset from the test examples with correct predictions after incorporating CoT rationales in the prompt. And we analyze them in order to know whether there are some similar patterns in these examples that can reveal the key to the success of M-CoT.

4.1.5 Evaluation

We experimented with two answer selection strategies. The first answer selection strategy is to use an answer trigger sentence "Answer: The answer is: (" to force the model to generate the answer label such as "A" in the first position of the generated sequences. Then, we compare the logit scores of the labels in all answer choices and select the highest one as the prediction. The second answer selection strategy is to use the same answer trigger sentence and let the model generate a text sequence. After that, we collect all generated text and use a regular expression to extract the answer label. Table 4.1 shows the zero-shot and zero-shot CoT results for the InstructBLIP 7B model with the two answer selection strategies. When applying the zero-shot

[1]These results are measured on the test set with images from the ScienceQA dataset and without generating CoT rationales.

	InstructBLIP 7B Zero-shot	InstructBLIP 7B Zero-shot CoT
Answer selection strategy 1	50.07	60.78
Answer selection strategy 2	50.07	60.19
Original answer selection strategy	60.50	– –

Table 4.1: Results to the zero-shot and zero-shot CoT strategies with InstructBLIP 7B on the ScienceQA dataset. The last row is the zero-shot accuracy by using the answer selection strategy of Dai et al. [2023] with the InstructBLIP 7B model on the ScienceQA dataset.

	InstructBLIP 7B	InstructBLIP 13B
Zero-shot	50.07	53.89
Zero-shot-CoT	60.78(+10.71)	68.77(+14.88)

Table 4.2: Results to zero-shot and zero-shot CoT prompting with InstructBLIP 7B and InstructBLIP 13B on the ScienceQA dataset.

prompting strategy on InstructBLIP 7B, the two answer selection strategies have the same performance, as the model always generates the answer label in the first position of the generated sequences. When using the zero-shot CoT strategy, the overall accuracy of the answer selection strategy 2 is slightly lower than that of the answer selection strategy 1. This is because when using the answer selection strategy 2, there are 14 examples of the test set for which the model generates a non-capital label before the answer, such as "b) phoenix", while the prompts provide capital labels in the choices. Therefore, we decided to use answer selection strategy 1 in our experiment. Dai et al. [2023] used an answer selection strategy that restricts the model's generation vocabulary to a list of answer candidates in which every answer candidate has both an answer label and answer choice. Then, they select the answer candidate with the highest log-likelihood as the answer. Different from our two answer selection strategies that only consider the probabilities of labels in the first position of the generation text, their method considers the coherence of each answer candidate in the generation, which might lead to a higher accuracy compared to our methods.

	NAT	SOC	LAN	G1-6	G7-12
7B zero-shot	49.46	50.13	65.90	55.35	37.24
7B zero-shot CoT	53.6(+4.14)	71.2(+21.07)	77.27(+11.37)	63.47(+8.12)	54.25(+17.01)
Number of examples	59.94%	37.88%	2.18%	70.85%	29.15%

Table 4.3: Zero-shot and zero-shot CoT results for InstructBLIP 7B across different subjects and difficulty levels in the ScienceQA test set. In the ScienceQA dataset, questions can be categorized into three subjects: natural science (NAT), social science (SOC), and language science (LAN). G1-6 represents questions for grades 1-6, and G7-12 represents questions for grades 7-12. Results are reported as percentages(%). The final row shows the distribution of questions across different subjects and difficulty levels in the ScienceQA test set.

4.2 Experimental Results

Table 4.2 shows the zero-shot and zero-shot CoT results of InstructBLIP 7B and InstructBLIP 13B on the ScienceQA dataset. The zero-shot results indicate that InstructBLIP 13B outperforms InstructBLIP 7B by 3.82%. After applying M-CoT, the performance gap between two models increases to 7.99%. Compared to zero-shot results, M-CoT improves the performance of InstructBLIP 7B by 10.71% and InstructBLIP 13B by 14.88%, suggesting that M-CoT can largely improve the performance of VLMs in complex multimodal reasoning tasks.

Table 4.3 shows the results of zero-shot and zero-shot CoT prompting with InstructBLIP 7B across different subjects in the SciecenQA dataset. The zero-shot results reveal that questions from the natural science subject yield the lowest accuracy at 49.46%, while questions from the language science subject achieve the highest accuracy at 65.9%. M-CoT consistently improves accuracy across all subjects, with the largest improvement observed in social science at 21.07%, and the smallest in natural science at 4.14%.

Furthermore, we analyzed the model's performance across different difficulty levels, comparing questions from grades 1-6 and those from grades 7-12. In the zero-shot scenario, InstructBLIP 7B achieves an accuracy of 55.35% for questions from grades 1-6, but only 37.24% for questions from grades 7-12, indicating the latter is more challenging for the model to answer. After applying M-CoT, the model's performance

	US-history	**Geography**	**Economics**	**Word-history**	**Civics**
7B Zero-shot	33.33	52.67	47.89	42.86	100
7B Zero-shot-CoT	55.95(+22.62)	74.67(+22)	61.97(+14.08)	42.86	100
Number of examples	10.99%	78.53%	9.29%	0.92%	0.26%

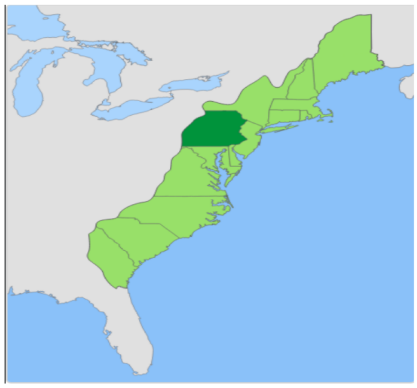
Table 4.4: Zero-shot and zero-shot CoT results for InstructBLIP 7B across different topics within the social science subject of the ScienceQA dataset. The social science subject includes five topics: us-history, geography, economics, word-history, and civics. Results are presented as percentages(%). The final row shows the distribution of topics within the social science subject in the ScienceQA test set.

has been improved significantly across all difficulty levels, with an 8.12% increase in accuracy for grades 1-6 and a 17.01% increase for grades 7-12. This demonstrates that M-CoT can effectively enhance the reasoning abilities of VLMs, enabling better performance on challenging multimodal questions that require complex reasoning.

As shown in Table 4.3, InstructBLIP has the highest performance improvement on questions from the social science subject compared to questions from the other two subjects. The social science subject encompasses five topics: us-history, geography, economics, word history, and civics. Therefore, we decided to explore whether specific topics influence the extent of performance improvement. The accuracy of questions in us-history, geography, and economics improved by over 10% after applying M-CoT, while questions in world-history and civics only improved by less than 1%. We noticed that M-CoT improved the accuracy of us-history questions by 22.62% and the accuracy of geography questions by 22% compared to zero-shot results. After analyzing test examples in us-history and geography, we observed that most of the questions from us-history and geography are about maps. Moreover, we find that most of the M-CoT rationales designed for this kind of question directly provide the answer, which might make the model select the choice easier. Figure 4.5 and Figure 4.6 show examples of us-history and geography in the ScienceQA dataset.

We sampled 70 M-CoT rationales from a total of 400 M-CoT rationales that were shown to improve the performance of InstructBLIP 13B on the ScienceQA dataset. After analyzing them, we observed that M-CoT rationales provide information about

Question: What is the name of the colony shown?



Choices:

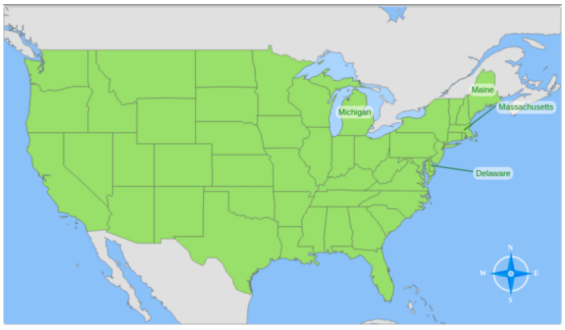
(A) Maine
(B) Pennsylvania
(C) Delaware
(D) Massachusetts

Solution: The colony is Pennsylvania.

Figure 4.5: An example of M-CoT rationale applied to us-history questions.

the image, such as the name of a place on a map, the distance between two magnets, the number of particles in solutions, the direction of arrows pointing to animals in a food web and the statistics from tables or graphs. We also find that M-CoT rationales provide commonsense knowledge to answer scientific questions, such as the relationship between the magnitude of the magnetic force and the distance, and the definitions of compounds and elementary substances. Figure 4.7 shows an M-CoT example that can improve the performance of InstructBLIP 13B on the ScienceQA dataset, and this example provides image information and commonsense knowledge for the model to reason. Combined with the fact that VLMs are weak at answering questions about maps and require commonsense knowledge to reason [Zhang et al., 2023], we hypothesize that the reason for the improvement of M-CoT might be M-CoT rationales can provide image information that the model finds difficult to discern, and the necessary commonsense knowledge for answering scientific questions.

Question: **Which of these states is farthest south?**



Choices:

(A) Maine
(B) Massachusetts
(C) Michigan
(D) Delaware

Solution: To find the answer, look at the compass rose. Look at which way the south arrow is pointing. Delaware is farthest south.

Figure 4.6: An example of M-CoT rationale applied to geography questions.

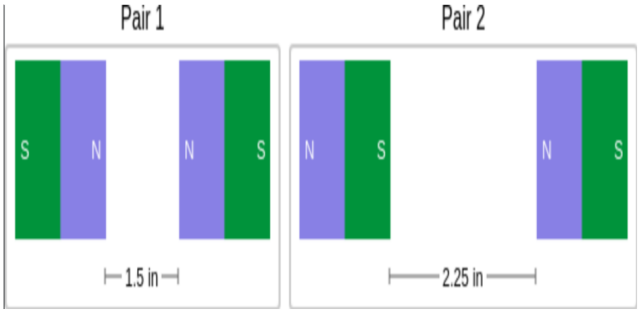
4.3 Explore Why M-CoT can Improve the Performance of VLMs

Inspired by the framework of Wang et al. [2023a] to explain what is the key part that makes CoT succeed in LLMs, we hypothesize the most important parts of M-CoT, which can lead to the improvement of performance in VLMs might be: (1) the relevance of the textual part of M-CoT rationales, (2) the validity of the reasoning chains in the textual part of M-CoT.

We adapted their experimental design, changing the components of M-CoT rationales and seeing how the performance is affected. Their work studied what is the key aspect of the CoT rationales in the few-shot setting and conducted experiments on linguistic reasoning tasks, while we explore which parts of M-CoT matter in the zero-shot setting and conduct our experiments with VLMs and on vision and language reasoning tasks. Similar to their ablation studies, we also test whether the validity and relevance affect the effectiveness.

We also conducted ablation studies to know whether our hypothesis is correct. The detailed experimental design will be discussed in the following subsections.

Question: Think about the magnetic force between the magnets in each pair. Which of the following statements is true?



Hint: The images below show two pairs of magnets. The magnets in different pairs do not affect each other. All the magnets shown are made of the same material.

Choices:

(A)The magnitude of the magnetic force is smaller in Pair 2.

(B)The magnitude of the magnetic force is the same in both pairs."

(C)The magnitude of the magnetic force is smaller in Pair 1

Solution: The magnets in Pair 2 attract. The magnets in Pair 1 repel. But whether the magnets attract or repel affects only the direction of the magnetic force. It does not affect the magnitude of the magnetic force. Distance affects the magnitude of the magnetic force. When there is a greater distance between magnets, the magnitude of the magnetic force between them is smaller. There is a greater distance between the magnets in Pair 2 than in Pair 1. So, the magnitude of the magnetic force is larger in Pair 1 than in Pair 2."

Figure 4.7: An example of M-CoT rationale that has been proven to improve the performance of InstructBLIP 13B on the ScienceQA dataset. This example provides a description of the image, highlighted in red, and commonsense knowledge, highlighted in orange. Both the image information and commonsense knowledge contribute to the model's reasoning process to infer the final answer.

4.3.1 Approaches to Conducting the Ablation Study

Prior studies [Wei et al., 2023, Madaan and Yazdanbakhsh, 2022] explored what makes CoT matter in the few-shot prompting setting, by modifying the corresponding part in the few-shot CoT exemplars and seeing how the model performance is affected by the change. However, this method cannot be directly applied in our experiments, as there is no limited number of demonstrations in the zero-shot prompting setting. We considered two ways to conduct the ablation study in the zero-shot prompting setting: (1) use a prompt to make the model generate a rationale to satisfy certain requirements. For example, use a prompt "please generate an irrelevant rationale and give the answer" to let the model generate a rationale that is irrelevant to the question; (2) modify the corresponding part we want to ablate in the rationale provided by the ScienceQA dataset as the solution to the question. Then, we provide the modified M-CoT rationale for the model. We use the second method in our experiments to

have maximal control over the parts that are to be ablated.

4.3.2 Test how the Relevance Affects the Performance of VLMs

We think the relevance of the textual part of the M-CoT rationale to the question might be the key to performance improvements when using M-CoT reasoning in VLMs. Because Zheng et al. [2023] showed LLMs tend to reason mainly based on the input rationale in the zero-shot prompting setting, we hypothesize that irrelevant M-CoT rationales could lead VLMs to the wrong prediction.

We prompted the VLM with the image, question, hints, choices, an irrelevant CoT rationale, and an answer trigger sentence "Answer: The answer is: (" to make the VLM generate the final answer. We assign the irrelevant rationale to each test example by randomly sampling a gold rationale from the same topic. Figure 4.8 shows the prompt we apply in this setting, with the modified and gold rationale for comparison.

4.3.3 Test how the Validity of Reasoning Chains Affects the Performance of VLMs

Inspired by Wang et al. [2023a], we tested whether the validity of consistent reasoning steps in the textual part of M-CoT might be the key component of M-CoT that leads to the improvement of performance in VLMs. Because logically valid reasoning chains are intuitively why CoT succeeds in LLMs, this may also be the same with VLMs. Although Wang et al. [2023a] showed that replacing valid reasoning with invalid reasoning in the few-shot exemplars can still make the LLM achieve 80% to 90% performance of the result using valid reasoning, we still speculate the validity of the reasoning will matter in the zero-shot setting and in the vision and language task. We prompted the VLM with the image, question, hints, choices, an invalid M-CoT

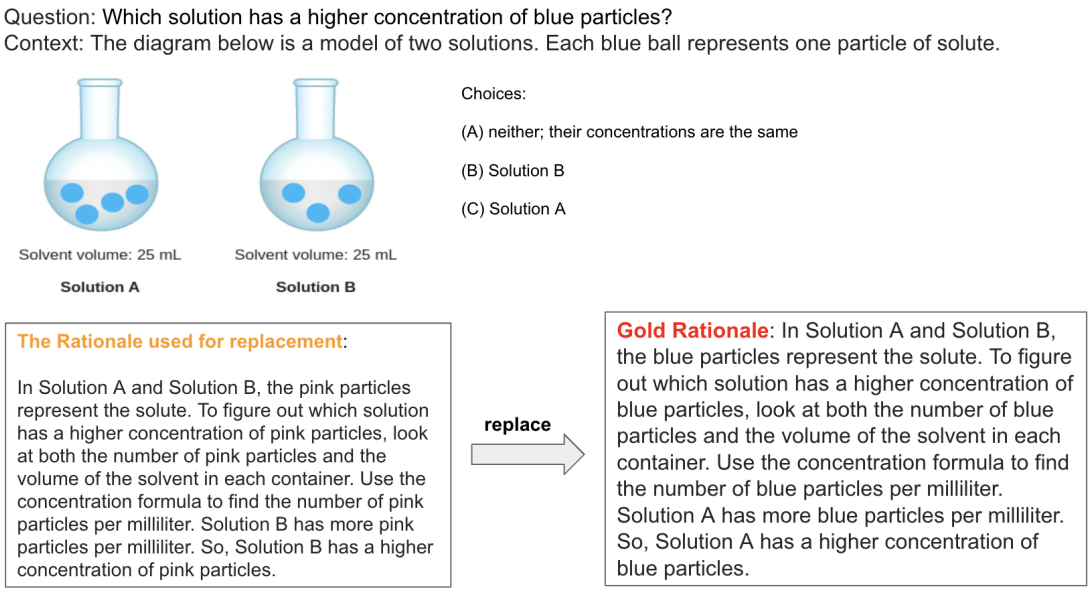


Figure 4.8: An example from the ablation study testing whether the relevance of the textual part of M-CoT rationales to the question affects the performance of VLMs. In this figure, the rationale in the lower-left corner is randomly sampled from the same topic, while the rationale in the lower-right corner is the gold rationale, which the left rationale replaces .

rationale, and the answer trigger sentence.

Constructing invalid M-CoT rationales

We sampled a subset of 30 M-CoT rationales from the M-CoT rationales that are proven to improve the performance of InstructBLIP 13B on the ScienceQA dataset. Then, we manually revised the textual part of the M-CoT rationales to make them invalid. We think there are three general ways to make M-CoT rationales invalid: (1) change the conclusion to another choice option and keep the original reasoning process; (2) revise the reasoning process and keep the original conclusion; (3) revise the reasoning procedure and also change the conclusion. We found that revising the reasoning process could make the image description in the M-CoT rationale become a confounding factor. This is because destroying the reasoning chains may need to change the part that provides some information about the image in the textual part

	InstructBLIP 13B
Zero-shot	53.89
Zero-shot CoT	68.77
No relevance	51.02

Table 4.5: The result to apply the irrelevant M-CoT rationales to InstructBLIP 13B model on the ScienceQA test set. Zero-shot and zero-shot CoT results are provided for comparison.

of the M-CoT rationale. And the image description part may be significant in visual question-answering tasks for the multimodal nature. Therefore, we decided to use the first strategy, which can avoid the confound of image description.

When selecting examples for constructing the invalid M-CoT subset, we excluded the map-related questions that contain only a single sentence of reasoning, as these cases lack a clear separation between the reasoning process and conclusion. Figure 4.9 illustrates an example of such a question. The examples included in the invalid M-CoT subset have different degrees of consistency in their reasoning processes. Some M-CoT rationales have a reasoning process that lacks some steps to perform rigorous reasoning, while some M-CoT rationales have a step-by-step reasoning process. Figure 4.11 provides an example of an M-CoT rationale with an inconsistent reasoning process, and Figure 4.10 shows an M-CoT rationale with a step-by-step reasoning procedure. We combine different strategies to make the conclusion invalid, including altering comparative adjectives to their antonyms, introducing negations to reverse the sentence meaning, swapping numerical values, and modifying key terms to direct the reasoning chain toward an incorrect option. Figure 4.12 shows an example of an invalid M-CoT rationale from the subset used in our experiment.

Question: **Which continent is highlighted?**

Choices:

(A) Australia

(B) North America

(C) South America

(D) Asia

Solution: This continent is Asia.

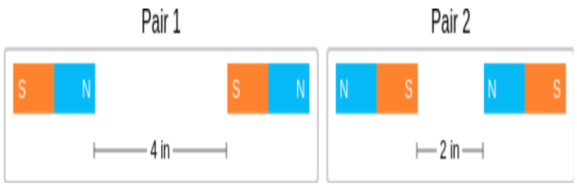
Figure 4.9: An example of the map question that provides only a sentence as the solution. These examples are excluded from the invalid M-CoT subset, as they lack a distinct separation between reasoning and conclusion.

4.4 Experiment Results

Table 4.5 shows the result of using irrelevant M-CoT rationales to test whether relevance is important to M-CoT in the multimodal reasoning task. When InstructBLIP 13B was tested with irrelevant M-CoT rationales, its performance on the ScienceQA dataset had an accuracy of 51.02%, which is 17.7% lower than the accuracy achieved using relevant M-CoT rationales. Furthermore, this result is 2.87% below the accuracy observed in the zero-shot scenario. These findings underscore the critical role of relevance in the success of M-CoT. The use of irrelevant M-CoT rationales may mislead the model toward incorrect answers, resulting in slightly lower performance compared to the performance of the zero-shot setting.

Table 4.6 shows the experiment result of applying the valid and invalid M-CoT rationales to the InstructBLIP 13B model. The use of M-CoT rationales with invalid reasoning leads to a significant performance decline of 66.7% compared to using valid M-CoT rationales. This can demonstrate that the validity of reasoning in the textual

Question: **Think about the magnetic force between the magnets in each pair. Which of the following statements is true?**



Hint: The images below show two pairs of magnets. The magnets in different pairs do not affect each other. All the magnets shown are made of the same material.

Choices:

(A) toward the boy's thumb
(B) away from the boy's thumb

Solution: Distance affects the strength of the magnetic force. When magnets are closer together, the magnetic force between them is stronger. The magnets in Pair 2 are closer together than the magnets in Pair 1. So, the magnetic force is stronger in Pair 2 than in Pair 1.

Figure 4.10: An example of M-CoT rationales with a consistent reasoning process. In the textual part of this rationale, each step logically follows the preceding one, forming a step-by-step reasoning chain.

	InstructBLIP 13B
valid M-CoT reasoning	100
invalid M-CoT reasoning	33.33

Table 4.6: The experimental result of applying invalid M-CoT rationales to the InstructBLIP 13B model. For comparison, the performance of the model using valid M-CoT ratioanles is also provided.

part of the M-CoT rationale is crucial for VLMs in answer selection. We hypothesize that VLMs rely heavily on the validity of the conclusions in the textual part of the M-CoT rationale for answer prediction, as our experiment shows that altering only the conclusion can substantially reduce accuracy.

Question: **What is the direction of this push?**

Hint: A boy plays with marbles. He pushes one of the marbles with his thumb.

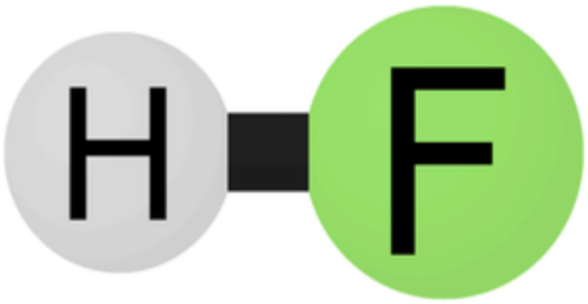
Choices:

(A) toward the boy's thumb
(B) away from the boy's thumb

Solution: The boy pushes his marble away from his thumb. The direction of the push is away from the boy's thumb.

Figure 4.11: An example of M-CoT rationales that lack certain reasoning steps necessary for a consistent reasoning procedure. A complete reasoning process would be: "The boy pushes his marble away from his thumb. The direction of a push is away from the object that is pushing. So, the direction of the push is away from the boy's thumb."

Question: **Complete the statement. Hydrogen fluoride is ().**



Hint: The model below represents a molecule of hydrogen fluoride. Hydrogen fluoride is used to make chemicals that can help keep refrigerators cool.

Choices:

(A) an elementary substance
(B) a compound

Solution: Count the number of chemical elements represented in the model. Then, decide if hydrogen fluoride is an elementary substance or a compound. In this model, each ball is labeled with H for hydrogen or F for fluorine. So, the model shows you that hydrogen fluoride is made of two chemical elements bonded together. Substances made of two or more chemical elements bonded together are compounds. So, hydrogen fluoride is **an elementary substance**.

Figure 4.12: An example of an invalid M-CoT used in our experiment. In this example, the reasoning process cannot logically lead to the final conclusion, as the conclusion has been altered to point to an incorrect option. The revised part is highlighted in orange. The correct conclusion for this example should be: "So, hydrogen fluoride is a compound."

Chapter 5

Limitations, Future Work, and Conclusions

In this chapter, we address the limitations of our experiments, and suggest potential avenues for future research. We conclude with a summary of the main contributions of our study.

5.1 Limitations

This section discusses the limitations inherent in our experiment design.

First, our experiments utilized only a single image as the visual component of the context. However, the ScienceQA test set includes examples that require multiple images to answer a question. We identified three types of such examples:

(1) Examples that lack the image as the visual context but include images as supplementary information for the textual choice. These examples are not used in our experiments due to the absence of an image context attribute. Figure 5.1 provides an illustration of this kind of question. We hypothesize that in these cases, the images serve to enrich the textual choice with visual data, potentially enabling the model to have a deeper understanding of the textual options, and thereby improving the

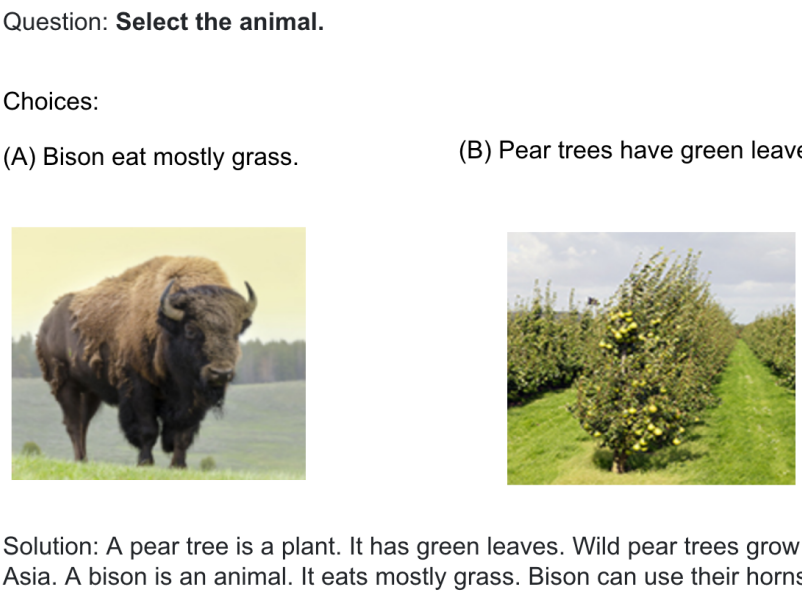


Figure 5.1: An example of a question without image context but with multiple images corresponding to the choices. These images can provide complementary visual information to the textual choices, and thus enhance the understanding of the model to the textual options.

accuracy of its final predictions.

(2) Examples that include an image as the context and other images that provide complementary visual information, which can enhance the model's understanding of the textual choice. In this case, we only feed the image context to the model, omitting the images to explain the textual choices. We hypothesize that these choice-related images may offer a similar advantage to that observed in the previous case, by enabling the model to develop a deeper comprehension of the options. Furthermore, we speculate that the model could combine the visual information about the choices and the textual information to perform better reasoning, and thus enhance the performance. Figure 5.2 illustrates an example of this type of question.

(3) Questions that require multiple images to answer. These questions are about trading. Figure 5.3 illustrates an example of this kind of question. In our experiments, we only feed the first image to the model, which makes the model lack some essential information from the second image that is necessary for answering the question

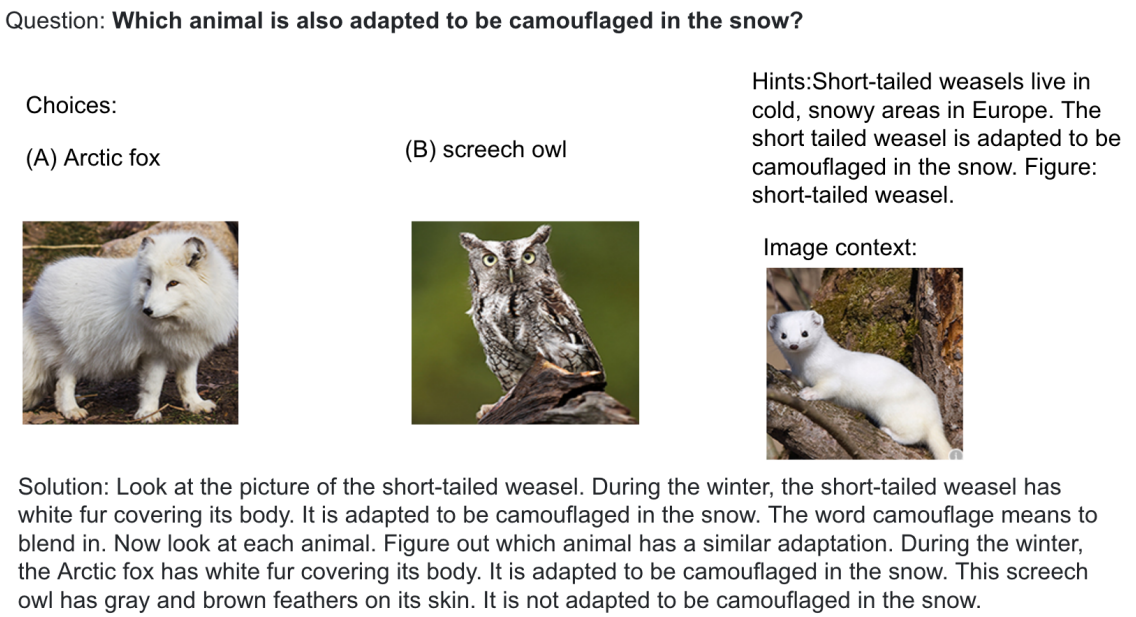


Figure 5.2: An example of a question with an image as the context and multiple images representing the choices. The model can leverage the visual information for the choices and the textual information to perform better reasoning.

accurately. This limitation could lead to reduced performance compared to feeding all images to the model. However, the trading questions constitute only 3.5% of the test set. Therefore, we do not anticipate a significant performance gap between our implementation, which only provides the model with the first image, and an approach that feeds all images to the model.

In general, we hypothesize that providing all images to the model could lead to a better performance than just providing one image across the three cases discussed. However, our current implementation sufficiently addresses our research question, and the conclusions drawn remain valid within this framework. Therefore, we contend that our implementation is adequate for the scope of this project, while the inclusion of additional images could be explored in future work.

Secondly, our study does not employ the same answer selection strategy as InstructBLIP. Adopting the answer selection strategy of InstructBLIP might improve the accuracy of our reported experimental results, as it has the advantage of considering

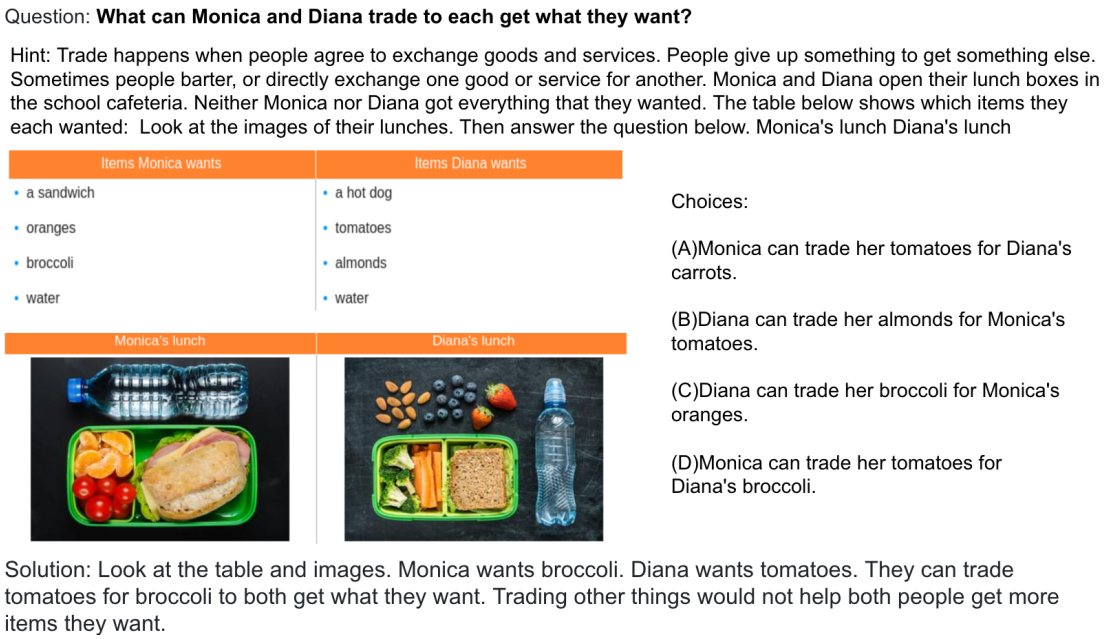


Figure 5.3: An example of a trading-related question that requires information from two images to be accurately answered.

the coherence between the label and the choice. However, our current answer selection strategy can already satisfy the requirements of exploring our research question. We suggest that implementing the answer selection strategy of InstructBLIP be considered in future research.

5.2 Future Work

In this chapter, we discuss possible research directions of our project in the future. In the experiment assessing whether the relevance of the textual part of the M-CoT rationale to the question is important to its overall success, we replaced the gold rationale with the rationale sampled from the same topic. We are interested in exploring whether substituting these with rationales from the same skill, which are more similar to the gold rationales, could yield the model performance close to the zero-shot CoT results.

In our experiment evaluating the importance of the validity of the reasoning chain to the effectiveness of M-CoT, we conducted the test on a small subset of examples. The experimental result indicated that the validity of the reasoning chain is important to M-CoT's performance improvement. However, this conclusion has yet to be confirmed across a larger sample size. Therefore, we propose to extend this investigation to encompass all test examples in the ScienceQA dataset. We plan to use automatic methods for revising conclusions, such as leveraging an LLM to modify the conclusion to make the rationale lead to an alternative option.

We intend to explore different ways to destroy the validity of the M-CoT rationale. In addition to altering only the conclusion, we could revise the textual elements that provide commonsense knowledge, excluding those that contain image information, to determine the specific contribution of the commonsense component to the reasoning ability of VLMs. Furthermore, we could revise both the conclusion and the reasoning process, as well as disrupt the order of sentences in the M-CoT rationale. If we experiment with these methods separately, we can compare the results to determine which aspects of M-CoT rationales VLMs rely on to infer the final answer. We can also combine all these validity-destroying techniques in one experiment to assess their impact on M-CoT performance. Moreover, as many M-CoT rationales in ScienceQA examples lack certain steps to form a step-by-step reasoning chain, we plan to manually introduce necessary reasoning steps or supportive sentences. This enhancement aims to make the reasoning process more complete and thereby improve the reasoning ability of the model to arrive at the correct answer.

We hypothesize that the image description is the most important component of M-CoT for two reasons: (1) M-CoT rationales that have been shown to improve the performance of VLMs include image-related information, which can address the model's limitations in accurately recognizing and interpreting visual content; and (2) in most cases, VLMs are required to answer questions based on the image-derived information.

Consequently, the image-related component in the textual part of M-CoT rationales is likely to play a key role in determining the final prediction of VLMs in visual question-answering tasks. To investigate this hypothesis, we plan to explore whether image-related textual components in the M-CoT rationale affect the performance of VLMs. Specifically, we intend to design an ablation experiment using CLIP [Radford et al., 2021] to identify and remove the sentences most closely aligned with the image. In addition to the previously discussed experiments aimed at identifying the most critical components of M-CoT for its effectiveness, we also consider conducting attention analysis across different layers of VLMs. This approach will allow us to determine which tokens receive higher attention scores than others, thereby enabling us to infer that VLMs rely on specific types of information or patterns to make their decisions. Identifying these patterns will help us understand the elements important to M-CoT's success.

Our study primarily investigates whether M-CoT can improve the performance of VLMs on complex multimodal reasoning tasks in a zero-shot setting. However, it would be interesting to explore whether M-CoT could also achieve the same significant improvement in a few-shot scenario through in-context learning. We hypothesize that VLMs that pre-trained with interleaved image and text datasets will have a better performance than those that do not include this kind of datasets in their pre-training process in the few-shot prompting setting. This hypothesis is grounded in the assumption that the use of interleaved image and text datasets during the pre-training of VLMs is suspected to be the key to the improvements when applying few-shot exemplars [Li et al., 2023]. To test this, we plan to conduct few-shot prompting experiments on VLMs such as OpenFlamingo [Awadalla et al., 2023], which are pre-trained with interleaved image and text datasets, and compare their results with those of VLMs like InstructBLIP, which do not incorporate interleaved image and text datasets in their per-training process.

5.3 Conclusions

We employ the zero-shot CoT prompting strategy to evaluate the effectiveness of M-CoT in enhancing the reasoning capabilities of VLMs in multimodal reasoning tasks. Through a detailed analysis of M-CoT rationales, we seek to identify the key factors that contribute to its success in these tasks. To this end, we design and conduct ablation experiments to determine which components of M-CoT are most influential in driving performance improvements. Our experimental results indicate that M-CoT can significantly enhance the reasoning abilities of VLMs in complex multimodal reasoning tasks. The relevance of M-CoT to the questions and the validity of its textual reasoning chain are critical factors in achieving these improvements. Our study offers new insights into the mechanisms underlying M-CoT's effectiveness and presents a framework for explaining its impact. We anticipate that our findings will inform the development of novel strategies to further enhance the reasoning capabilities of VLMs, such as the design of a more advanced M-CoT paradigm that delivers even greater performance gains.

Appendix A

A.1

Figure A.1 shows the ablation settings of the experiment conducted in Wang et al. [2023a], to explore what is the key aspect of CoT.

Prompt Setting	**Example Query (Arithmetic Reasoning)** *Leah had 32 chocolates and her sister had 42. If they ate 35, how many pieces do they have left in total?*	**Example Query (Factual QA)** *Who is the grandchild of Dambar Shah?*
STD(Standard prompting)	39	So the final answer is: Rudra Shah.
CoT (Chain-of-Thought)	Originally, Leah had 32 chocolates and her sister had 42. So in total they had 32 + 42 = 74. After eating 35, they had 74 - 35 = 39 pieces left in total. The answer is 39.	Dambar Shah (? - 1645) was the father of Krishna Shah. Rudra Shah was the child of Krishna Shah (? 1661). So the final answer (the name of the grandchild) is: Rudra Shah.
Invalid Reasoning	Originally, Leah had 32 chocolates and her sister had 42. So her sister had 42 - 32 = 10 chocolates more than Leah has. After eating 35, since 10 + 35 = 45, they had 45 - 6 = 39 pieces left in total. The answer is 39.	Dambar Shah (? - 1645) was the king of the Gorkha Kingdom. The Gorkha Kingdom was established by Prince Dravya Shah. Dravya Shah has a child named Rudra Shah. So the final answer (the name of the grandchild) is: Rudra Shah.
No coherence for bridging objects	Originally, Leah had 32 + 42 = 74 chocolates and her sister had 32. So in total they had 74 - 35 = 39. After eating 35, they had 42 pieces left in total. The answer is 39.	Krishna Shah was the father of Rudra Shah. Dambar Shah (? - 1645) was the child of Krishna Shah (? 1661). So the final answer (the name of the grandchild) is: Rudra Shah.
No relevance for bridging objects	Originally, Leah had 19 chocolates and her sister had 31. So in total they had 19 + 31 = 50. After eating 29, they had 50 - 29 = 21 pieces left in total. The answer is 21.	Metis Amando was the father of David Amando. Randall Amando was the child of David Amando. So the final answer (the name of the grandchild) is: Randall Amando.
No coherence for language templates	After eating 32, they had 42 pieces left in total. Originally, Leah had 32 + 42 = 74 chocolates and her sister had 35. So in total they had 74 - 35 = 39. The answer is 39.	Dambar Shah (? - 1645) was the child of Krishna Shah. Krishna Shah (? - 1661) was the father of Rudra Shah. So the final answer (the name of the grandchild) is: Rudra Shah.
No relevance for language templates	Patricia needs to donate 32 inches, and wants her hair to be 42 inches long after the donation. Her hair is 35 inches long currently. Her hair needs to be 32 + 42 = 74 inches long when she cuts it. So she needs to grow 74 - 35 = 39 more inches. The answer is 39.	The husband of Dambar Shah (? - 1645) is Krishna Shah. Krishna Shah (? - 1661) has a brother called Rudra Shah. So the final answer (the name of the brother-in-law) is: Rudra Shah.
No coherence	After eating 32 + 42 = 74, they had 32 pieces left in total. Originally, Leah had 74 - 35 = 39 chocolates and her sister had 35. So in total they had 42. The answer is 39.	Krishna Shah was the child of Rudra Shah. Dambar Shah (? - 1645) was the father of Krishna Shah (? 1661). So the final answer (the name of the grandchild) is: Rudra Shah.
No relevance	Patricia needs to donate 19 inches, and wants her hair to be 31 inches long after the donation. Her hair is 29 inches long currently. Her hair needs to be 19 + 31 = 50 inc long when she cuts it. So she needs to grow 50 - 29 = 21 more inches. The answer is 21.	The husband of Metis Amando is David Amando. David Amando has a brother called Randall Amando. So the final answer (the name of the brother-in-law) is: Randall Amando.

Figure A.1: Examples of the ablation settings to explore what is the key part of CoT in Wang et al. [2023a]. This figure is from Wang et al. [2023a].

Bibliography

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andrew Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karen Simonyan. Flamingo: a visual language model for few-shot learning, 2022. URL `https://arxiv.org/abs/2204.14198`.

Anas Awadalla, Irena Gao, Josh Gardner, Jack Hessel, Yusuf Hanafy, Wanrong Zhu, Kalyani Marathe, Yonatan Bitton, Samir Gadre, Shiori Sagawa, Jenia Jitsev, Simon Kornblith, Pang Wei Koh, Gabriel Ilharco, Mitchell Wortsman, and Ludwig Schmidt. Openflamingo: An open-source framework for training large autoregressive vision-language models, 2023. URL `https://arxiv.org/abs/2308.01390`.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners, 2020. URL `https://arxiv.org/abs/2005.14165`.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, March 2023. URL https://lmsys.org/blog/2023-03-30-vicuna/.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. Palm: Scaling language modeling with pathways, 2022. URL https://arxiv.org/abs/2204.02311.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. Scaling instruction-finetuned language models, 2022.

URL https://arxiv.org/abs/2210.11416.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems, 2021. URL https://doi.org/10.48550/arXiv.2110.14168.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning, 2023. URL https://arxiv.org/abs/2305.06500.

Danny Driess, Fei Xia, Mehdi S. M. Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, Wenlong Huang, Yevgen Chebotar, Pierre Sermanet, Daniel Duckworth, Sergey Levine, Vincent Vanhoucke, Karol Hausman, Marc Toussaint, Klaus Greff, Andy Zeng, Igor Mordatch, and Pete Florence. Palm-e: An embodied multimodal language model, 2023. URL https://arxiv.org/abs/2303.03378.

Mononito Goswami, Vedant Sanil, Arjun Choudhry, Arvind Srinivasan, Chalisa Udompanyawit, and Artur Dubrawski. Aqua: A benchmarking tool for label quality assessment, 2024. URL https://arxiv.org/abs/2306.09467.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering, 2017. URL https://arxiv.org/abs/1612.00837.

Vaishnavi Himakunthala, Andy Ouyang, Daniel Rose, Ryan He, Alex Mei, Yujie Lu, Chinmay Sonar, Michael Saxon, and William Yang Wang. Let's think frame by frame with vip: A video infilling and prediction dataset for evaluating video chain-of-thought, 2023. URL https://arxiv.org/abs/2305.13903.

Namgyu Ho, Laura Schmid, and Se-Young Yun. Large language models are reasoning teachers, 2023. URL https://arxiv.org/abs/2212.10071.

Shaohan Huang, Li Dong, Wenhui Wang, Yaru Hao, Saksham Singhal, Shuming Ma, Tengchao Lv, Lei Cui, Owais Khan Mohammed, Barun Patra, Qiang Liu, Kriti Aggarwal, Zewen Chi, Johan Bjorck, Vishrav Chaudhary, Subhojit Som, Xia Song, and Furu Wei. Language is not all you need: Aligning perception with language models, 2023. URL https://arxiv.org/abs/2302.14045.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners, 2023. URL https://doi.org/10.48550/arXiv.2205.11916.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models, 2023. URL https://arxiv.org/abs/2301.12597.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning, 2023. URL https://arxiv.org/abs/2304.08485.

Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *The 36th Conference on Neural Information Processing Systems (NeurIPS)*, 2022. URL https://arxiv.org/abs/2209.09513.

Pan Lu, Baolin Peng, Hao Cheng, Michel Galley, Kai-Wei Chang, Ying Nian Wu, Song-Chun Zhu, and Jianfeng Gao. Chameleon: Plug-and-play compositional reasoning with large language models, 2023. URL https://arxiv.org/abs/2304.09842.

Aman Madaan and Amir Yazdanbakhsh. Text and patterns: For effective chain of

thought, it takes two to tango, 2022. URL https://doi.org/10.48550/arXiv.2209.07686.

OpenAI, :, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mo Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel

Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan

Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. Gpt-4 technical report, 2023. URL https://arxiv.org/abs/2303.08774.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback, 2022. URL https://arxiv.org/abs/2203.02155.

Dong Huk Park, Lisa Anne Hendricks, Zeynep Akata, Anna Rohrbach, Bernt Schiele, Trevor Darrell, and Marcus Rohrbach. Multimodal explanations: Justifying decisions and pointing to the evidence, 2018. URL https://arxiv.org/abs/1802.08129.

Arkil Patel, Satwik Bhattamishra, and Navin Goyal. Are nlp models really able to solve simple math word problems?, 2021. URL https://arxiv.org/abs/2103.07191.

Ben Prystawski, Michael Y. Li, and Noah D. Goodman. Why think step by step? reasoning emerges from the locality of experience, 2023. URL https://doi.org/10.48550/arXiv.2304.03843.

Alec Radford and Karthik Narasimhan. Improving language understanding by generative pre-training. 2018. URL https://api.semanticscholar.org/CorpusID:49313245.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019. URL https://api.semanticscholar.org/CorpusID:160025533.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen

Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021. URL https://arxiv.org/abs/2103.00020.

Daniel Rose, Vaishnavi Himakunthala, Andy Ouyang, Ryan He, Alex Mei, Yujie Lu, Michael Saxon, Chinmay Sonar, Diba Mirza, and William Yang Wang. Visual chain of thought: Bridging logical gaps with multimodal infillings, 2024. URL https://arxiv.org/abs/2305.02317.

Subhro Roy and Dan Roth. Solving general arithmetic word problems. In Lluís Màrquez, Chris Callison-Burch, and Jian Su, editors, *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1743–1752, Lisbon, Portugal, September 2015. Association for Computational Linguistics. doi: 10.18653/v1/D15-1202. URL https://aclanthology.org/D15-1202.

Victor Sanh, Albert Webson, Colin Raffel, Stephen H. Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Teven Le Scao, Arun Raja, Manan Dey, M Saiful Bari, Canwen Xu, Urmish Thakker, Shanya Sharma Sharma, Eliza Szczechla, Taewoon Kim, Gunjan Chhablani, Nihal Nayak, Debajyoti Datta, Jonathan Chang, Mike Tian-Jian Jiang, Han Wang, Matteo Manica, Sheng Shen, Zheng Xin Yong, Harshit Pandey, Rachel Bawden, Thomas Wang, Trishala Neeraj, Jos Rozen, Abheesht Sharma, Andrea Santilli, Thibault Fevry, Jason Alan Fries, Ryan Teehan, Tali Bers, Stella Biderman, Leo Gao, Thomas Wolf, and Alexander M. Rush. Multitask prompted training enables zero-shot task generalization, 2022. URL https://arxiv.org/abs/2110.08207.

Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, YaGuang Li, Hongrae Lee, Huaixiu Steven Zheng, Amin Ghafouri, Marcelo Menegali, Yanping Huang, Maxim Krikun, Dmitry Lepikhin, James Qin, Dehao Chen, Yuanzhong Xu, Zhifeng Chen, Adam Roberts, Maarten Bosma, Vincent Zhao, Yanqi Zhou, Chung-Ching

Chang, Igor Krivokon, Will Rusch, Marc Pickett, Pranesh Srinivasan, Laichee Man, Kathleen Meier-Hellstern, Meredith Ringel Morris, Tulsee Doshi, Renelito Delos Santos, Toju Duke, Johnny Soraker, Ben Zevenbergen, Vinodkumar Prabhakaran, Mark Diaz, Ben Hutchinson, Kristen Olson, Alejandra Molina, Erin Hoffman-John, Josh Lee, Lora Aroyo, Ravi Rajakumar, Alena Butryna, Matthew Lamm, Viktoriya Kuzmina, Joe Fenton, Aaron Cohen, Rachel Bernstein, Ray Kurzweil, Blaise Aguera-Arcas, Claire Cui, Marian Croak, Ed Chi, and Quoc Le. Lamda: Language models for dialog applications, 2022. URL https://arxiv.org/abs/2201.08239.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models, 2023. URL https://arxiv.org/abs/2302.13971.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2023. URL https://arxiv.org/abs/1706.03762.

Boshi Wang, Sewon Min, Xiang Deng, Jiaming Shen, You Wu, Luke Zettlemoyer, and Huan Sun. Towards understanding chain-of-thought prompting: An empirical study of what matters, 2023a. URL https://doi.org/10.48550/arXiv.2212.10001.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models, 2023b. URL https://arxiv.org/abs/2203.11171.

Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. Finetuned language models are zero-shot learners, 2022a. URL https://arxiv.org/abs/2109.01652.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. Emergent abilities of large language models, 2022b. URL https://arxiv.org/abs/2206.07682.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models, 2023. URL https://doi.org/10.48550/arXiv.2201.11903.

Yifan Wu, Pengchuan Zhang, Wenhan Xiong, Barlas Oguz, James C. Gee, and Yixin Nie. The role of chain-of-thought in complex vision-language reasoning task, 2023. URL https://doi.org/10.48550/arXiv.2311.09193.

Zhengyuan Yang, Linjie Li, Kevin Lin, Jianfeng Wang, Chung-Ching Lin, Zicheng Liu, and Lijuan Wang. The dawn of lmms: Preliminary explorations with gpt-4v(ision), 2023. URL https://arxiv.org/abs/2309.17421.

Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. A survey on multimodal large language models, 2023. URL https://arxiv.org/abs/2306.13549.

Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. Multimodal chain-of-thought reasoning in language models, 2023. URL https://doi.org/10.48550/arXiv.2302.00923.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang,

Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. A survey of large language models, 2023. URL https://arxiv.org/abs/2303.18223.

Ge Zheng, Bin Yang, Jiajin Tang, Hong-Yu Zhou, and Sibei Yang. Ddcot: Duty-distinct chain-of-thought prompting for multimodal reasoning in language models, 2023. URL https://doi.org/10.48550/arXiv.2310.16436.