Utrecht University

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

Motivation:

Chain-of-Thought (CoT) can largely improve the performance of Large Language Models (LLMs) on complex reasoning tasks.

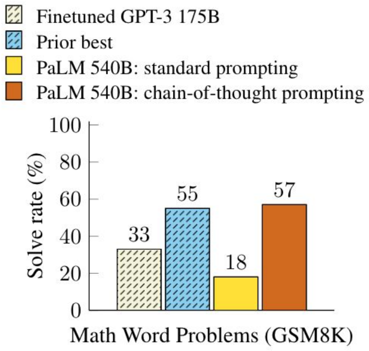


Figure 2: PaLM 540B uses chain-of-thought prompting to achieve new state-of-the-art performance on the GSM8K benchmark of math word problems. Finetuned GPT-3 and prior best are from Cobbe et al. (2021).

Visual modality is integrated with LLMs to enable them have the ability to perceive and reason about the multimodal inputs.

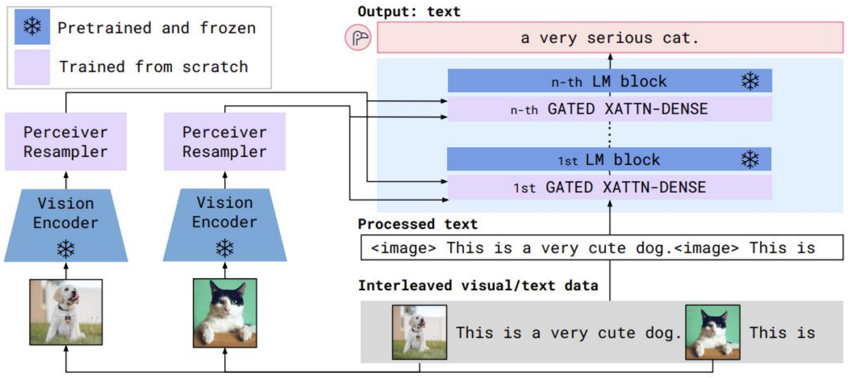


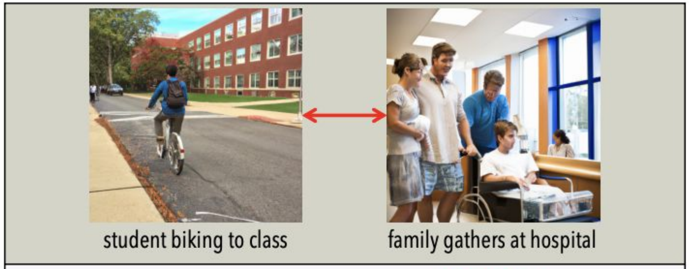
Figure 3: **Flamingo architecture overview.** Flamingo is a family of visual language models (VLMs) that take as input visual data interleaved with text and produce free-form text as output.

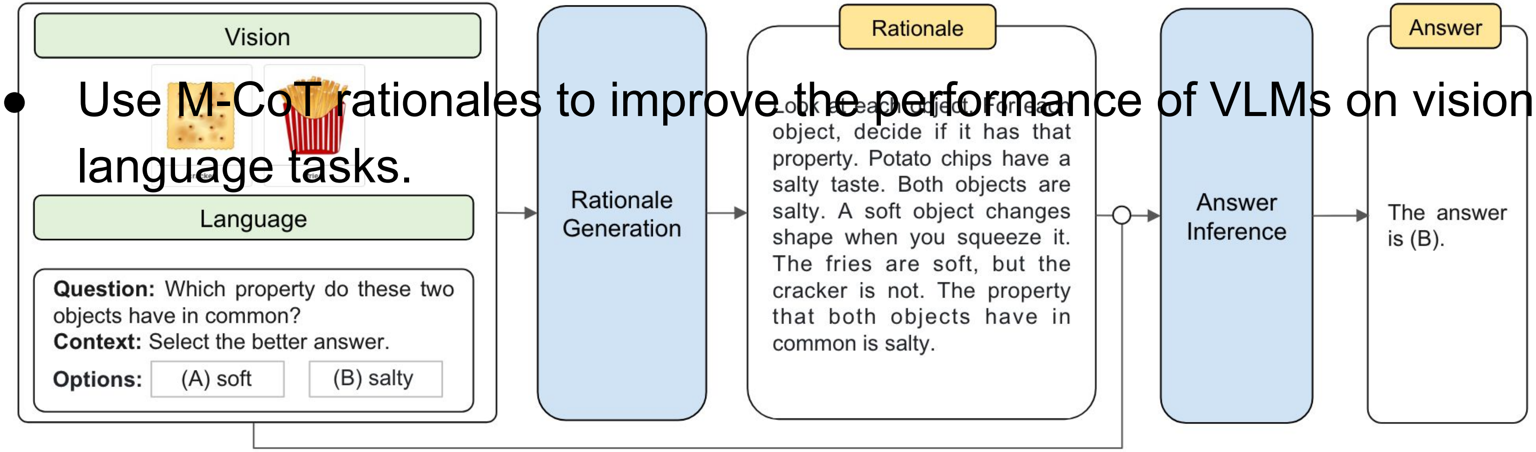
Motivation

Can we apply CoT on vision and language tasks? Can Multimodal CoT (M-CoT) have the same amazing performance improvement on Large Vision and Language Models (VLM) ?

Current studies on M-CoT focus on

- The design of the chain in the prompt.
- Use M-CoT rationales to improve the performance of VLMs on vision and language tasks.





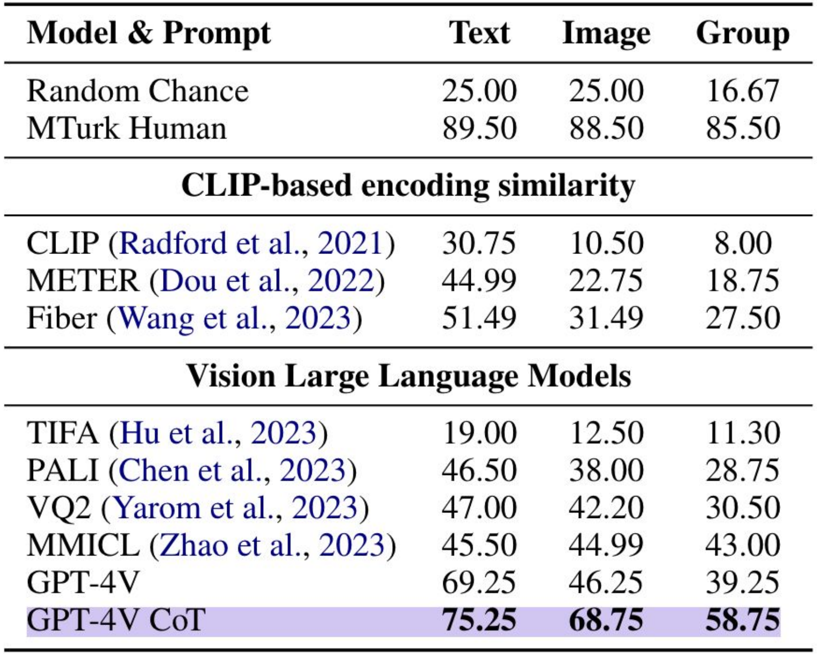
Question: Which property do these two objects have in common?
Context: Select the better answer.
Options: (A) soft (B) salty

Look at each object. For each object, decide if it has that property. Potato chips have a salty taste. Both objects are salty. A soft object changes shape when you squeeze it. The fries are soft, but the cracker is not. The property that both objects have in common is salty.

The answer is (B).

Figure 4. Overview of our Multimodal-CoT framework. Multimodal-CoT consists of two stages: (i) rationale generation and (ii) answer inference. Both stages share the same model architecture but differ in the input and output. In the first stage, we feed the model with language and vision inputs to generate rationales. In the second stage, we append the original language input with the rationale generated from the first stage. Then, we feed the updated language input with the original vision input to the model to infer the answer.

ical gaps
ing tasks;
nt method
fillings to



Model & Prompt	**Text**	**Image**	**Group**
Random Chance	25.00	25.00	16.67
MTurk Human	89.50	88.50	85.50
CLIP-based encoding similarity			
CLIP (Radford et al., 2021)	30.75	10.50	8.00
METER (Dou et al., 2022)	44.99	22.75	18.75
Fiber (Wang et al., 2023)	51.49	31.49	27.50
Vision Large Language Models			
TIFA (Hu et al., 2023)	19.00	12.50	11.30
PALI (Chen et al., 2023)	46.50	38.00	28.75
VQ2 (Yarom et al., 2023)	47.00	42.20	30.50
MMICL (Zhao et al., 2023)	45.50	44.99	43.00
GPT-4V	69.25	46.25	39.25
GPT-4V CoT	**75.25**	**68.75**	**58.75**

A systematic test on whether M-CoT can improve the performance of VLMs is still lacked.

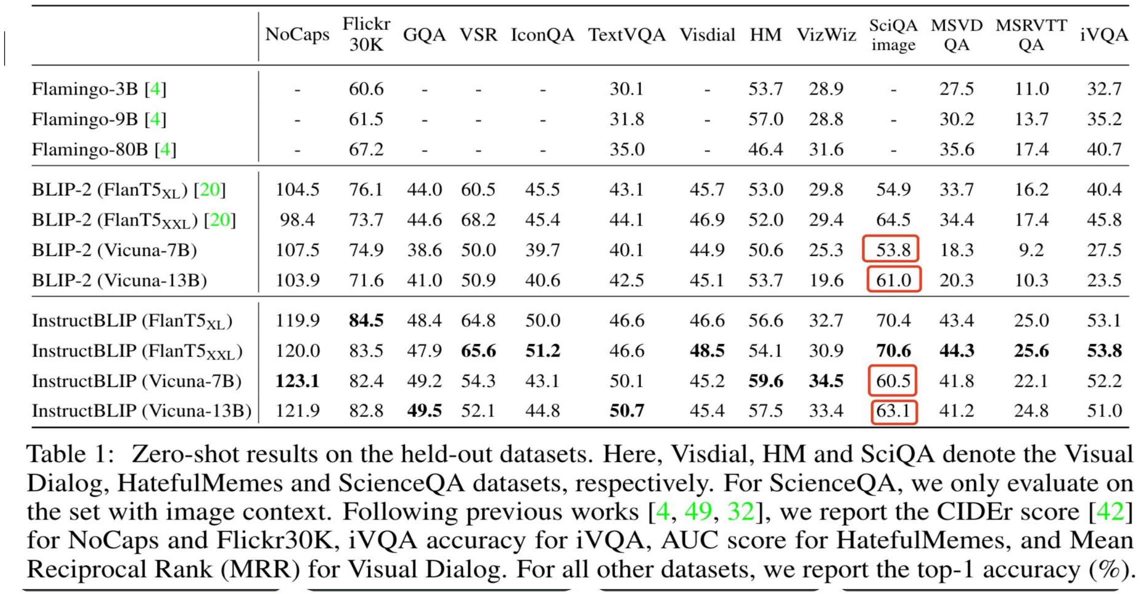
If M-CoT can have the same performance improvement on VLMs as that on LLMs, the reason of why M-CoT can have the improvement still remain mysteries.

↓

Research question: Can M-CoT improve the performance of VLMs on the complex multimodal reasoning tasks? If yes, which aspects of M-CoT are important to the performance gain of M-CoT?

Experimental design: Model, Dataset, and Method

Model: InstructBLIP（Dai et al., 2023）



	NoCaps	Flickr 30K	GQA	VSR	IconQA	TextVQA	Visdial	HM	VizWiz	SciQA image	MSVD QA	MSRVTT QA	iVQA
Flamingo-3B [4]	-	60.6	-	-	-	30.1	-	53.7	28.9	-	27.5	11.0	32.7
Flamingo-9B [4]	-	61.5	-	-	-	31.8	-	57.0	28.8	-	30.2	13.7	35.2
Flamingo-80B [4]	-	67.2	-	-	-	35.0	-	46.4	31.6	-	35.6	17.4	40.7
BLIP-2 (FlanT5$_{XL}$) [20]	104.5	76.1	44.0	60.5	45.5	43.1	45.7	53.0	29.8	54.9	33.7	16.2	40.4
BLIP-2 (FlanT5$_{XXL}$) [20]	98.4	73.7	44.6	68.2	45.4	44.1	46.9	52.0	29.4	64.5	34.4	17.4	45.8
BLIP-2 (Vicuna-7B)	107.5	74.9	38.6	50.0	39.7	40.1	44.9	50.6	25.3	53.8	18.3	9.2	27.5
BLIP-2 (Vicuna-13B)	103.9	71.6	41.0	50.9	40.6	42.5	45.1	53.7	19.6	61.0	20.3	10.3	23.5
InstructBLIP (FlanT5$_{XL}$)	119.9	**84.5**	48.4	64.8	50.0	46.6	46.6	56.6	32.7	70.4	43.4	25.0	53.1
InstructBLIP (FlanT5$_{XXL}$)	120.0	83.5	47.9	**65.6**	**51.2**	46.6	**48.5**	54.1	30.9	**70.6**	**44.3**	**25.6**	**53.8**
InstructBLIP (Vicuna-7B)	**123.1**	82.4	49.2	54.3	43.1	50.1	45.2	**59.6**	**34.5**	60.5	41.8	22.1	52.2
InstructBLIP (Vicuna-13B)	121.9	82.8	**49.5**	52.1	44.8	**50.7**	45.4	57.5	33.4	63.1	41.2	24.8	51.0

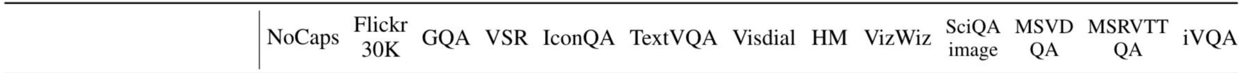
Table 1: Zero-shot results on the held-out datasets. Here, Visdial, HM and SciQA denote the Visual Dialog, HatefulMemes and ScienceQA datasets, respectively. For ScienceQA, we only evaluate on the set with image context. Following previous works [4, 49, 32], we report the CIDEr score [42] for NoCaps and Flickr30K, iVQA accuracy for iVQA, AUC score for HatefulMemes, and Mean Reciprocal Rank (MRR) for Visual Dialog. For all other datasets, we report the top-1 accuracy (%).

Consists of a frozen image encoder, a Q-Former, and a frozen LLM

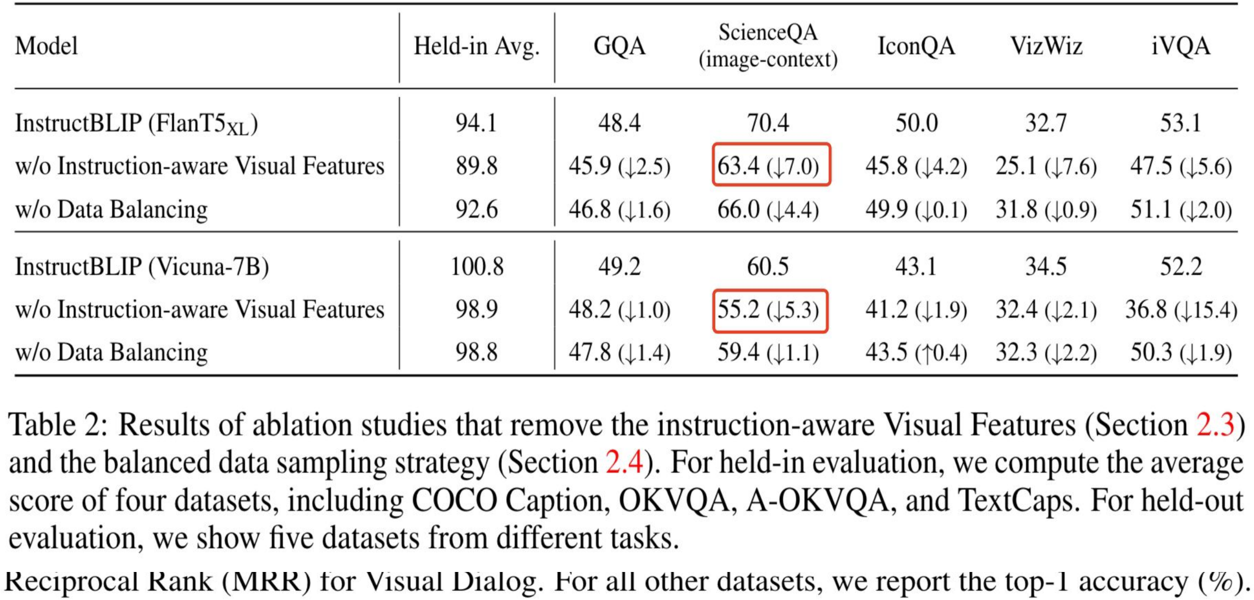
Instruct-tuned by 13 VL datasets

Zero-shot results outperform BLIP-2 across different LLMs

Reasons for using InstructBLIP Vicuna 7B/13B:



	NoCaps	Flickr 30K	GQA	VSR	IconQA	TextVQA	Visdial	HM	VizWiz	SciQA image	MSVD QA	MSRVTT QA	iVQA



Model	Held-in Avg.	GQA	ScienceQA (image-context)	IconQA	VizWiz	iVQA
InstructBLIP (FlanT5$_{XL}$)	94.1	48.4	70.4	50.0	32.7	53.1
w/o Instruction-aware Visual Features	89.8	45.9 (↓2.5)	63.4 (↓7.0)	45.8 (↓4.2)	25.1 (↓7.6)	47.5 (↓5.6)
w/o Data Balancing	92.6	46.8 (↓1.6)	66.0 (↓4.4)	49.9 (↓0.1)	31.8 (↓0.9)	51.1 (↓2.0)
InstructBLIP (Vicuna-7B)	100.8	49.2	60.5	43.1	34.5	52.2
w/o Instruction-aware Visual Features	98.9	48.2 (↓1.0)	55.2 (↓5.3)	41.2 (↓1.9)	32.4 (↓2.1)	36.8 (↓15.4)
w/o Data Balancing	98.8	47.8 (↓1.4)	59.4 (↓1.1)	43.5 (↑0.4)	32.3 (↓2.2)	50.3 (↓1.9)

Table 2: Results of ablation studies that remove the instruction-aware Visual Features (Section 2.3) and the balanced data sampling strategy (Section 2.4). For held-in evaluation, we compute the average score of four datasets, including COCO Caption, OKVQA, A-OKVQA, and TextCaps. For held-out evaluation, we show five datasets from different tasks.

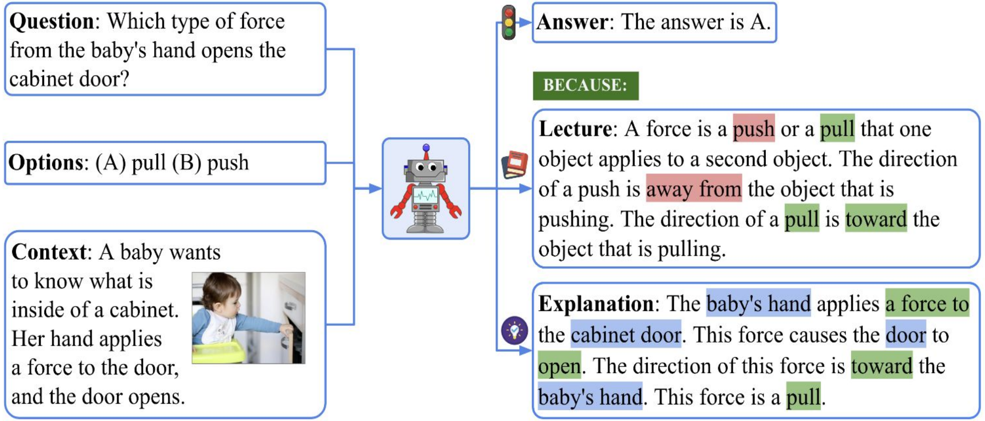
Reciprocal Rank (MRR) for Visual Dialog. For all other datasets, we report the top-1 accuracy (%).

- Ir uction-tuning
- C
- In ks.

  In tions

  w

Dataset: ScienceQA（Lu et al., 2022）



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

Options: (A) pull (B) push

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

Answer: The answer is A.

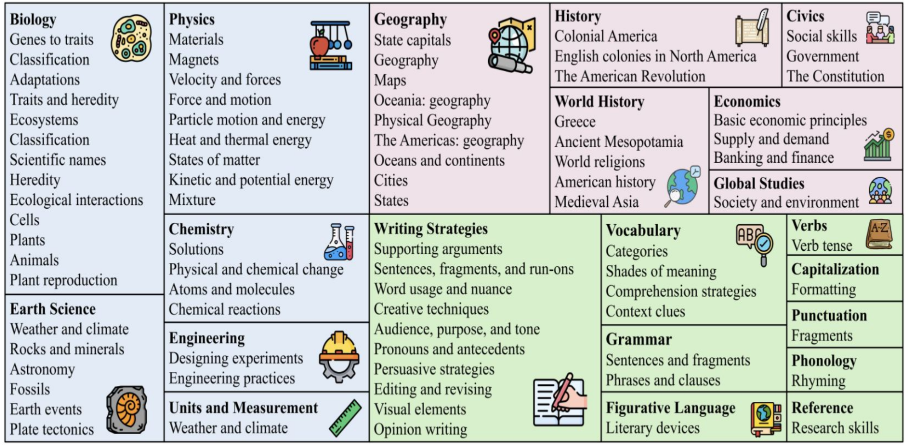
BECAUSE:

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

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

- A large scale multimodal dataset
- Multiple choice science questions with explanations
- Collected from elementary and high school curricula

ScienceQA



Biology
Genes to traits
Classification
Adaptations
Traits and heredity
Ecosystems
Classification
Scientific names
Heredity
Ecological interactions
Cells
Plants
Animals
Plant reproduction

Earth Science
Weather and climate
Rocks and minerals
Astronomy
Fossils
Earth events
Plate tectonics

Physics
Materials
Magnets
Velocity and forces
Force and motion
Particle motion and energy
Heat and thermal energy
States of matter
Kinetic and potential energy
Mixture

Chemistry
Solutions
Physical and chemical change
Atoms and molecules
Chemical reactions

Engineering
Designing experiments
Engineering practices

Units and Measurement
Weather and climate

Geography
State capitals
Geography
Maps
Oceania: geography
Physical Geography
The Americas: geography
Oceans and continents
Cities
States

History
Colonial America
English colonies in North America
The American Revolution

Civics
Social skills
Government
The Constitution

World History
Greece
Ancient Mesopotamia
World religions
American history
Medieval Asia

Economics
Basic economic principles
Supply and demand
Banking and finance

Global Studies
Society and environment

Writing Strategies
Supporting arguments
Sentences, fragments, and run-ons
Word usage and nuance
Creative techniques
Audience, purpose, and tone
Pronouns and antecedents
Persuasive strategies
Editing and revising
Visual elements
Opinion writing

Vocabulary
Categories
Shades of meaning
Comprehension strategies
Context clues

Grammar
Sentences and fragments
Phrases and clauses

Figurative Language
Literary devices

Verbs
Verb tense

Capitalization
Formatting

Punctuation
Fragments

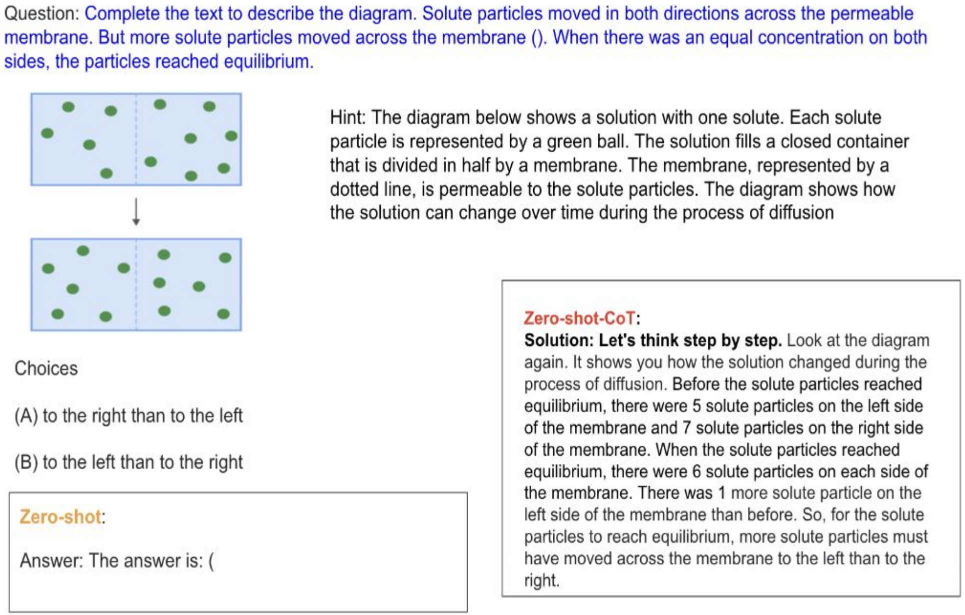
Phonology
Rhyming

Reference
Research skills

- 3 subjects: natural science, social science, language science
- 26 topics: Biology, Earth Science…
- 127 categories: genes to traits, classification…
- 379 specific skills: use a chemical formula to describe the molecule, select object corresponding to a liquid

Reason for using ScienceQA: the only open-sourced multimodal dataset designed to test the multi-hop reasoning ability of VLMs

Method: Zero-shot Prompting and Zero-shot CoT Prompting



Question: Complete the text to describe the diagram. Solute particles moved in both directions across the permeable membrane. But more solute particles moved across the membrane (). When there was an equal concentration on both sides, the particles reached equilibrium.

Hint: The diagram below shows a solution with one solute. Each solute particle is represented by a green ball. The solution fills a closed container that is divided in half by a membrane. The membrane, represented by a dotted line, is permeable to the solute particles. The diagram shows how the solution can change over time during the process of diffusion

Choices

(A) to the right than to the left

(B) to the left than to the right

Zero-shot:

Answer: The answer is: (

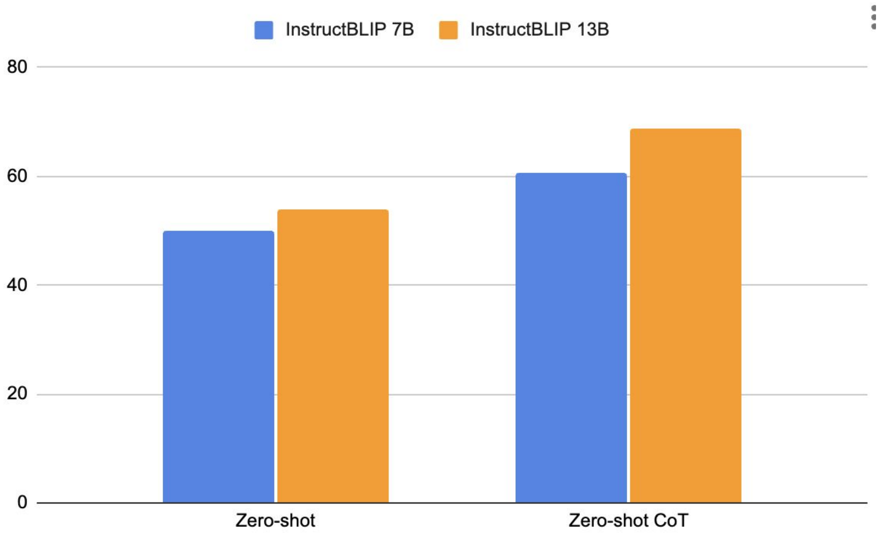
Zero-shot-CoT:

Solution: Let's think step by step. Look at the diagram again. It shows you how the solution changed during the process of diffusion. Before the solute particles reached equilibrium, there were 5 solute particles on the left side of the membrane and 7 solute particles on the right side of the membrane. When the solute particles reached equilibrium, there were 6 solute particles on each side of the membrane. There was 1 more solute particle on the left side of the membrane than before. So, for the solute particles to reach equilibrium, more solute particles must have moved across the membrane to the left than to the right.

Features of the prompt that has the largest performance gap:

- Hints about the different components
- Use a sentence like “Let’s think step by step”
- The structure can be naturally read by human

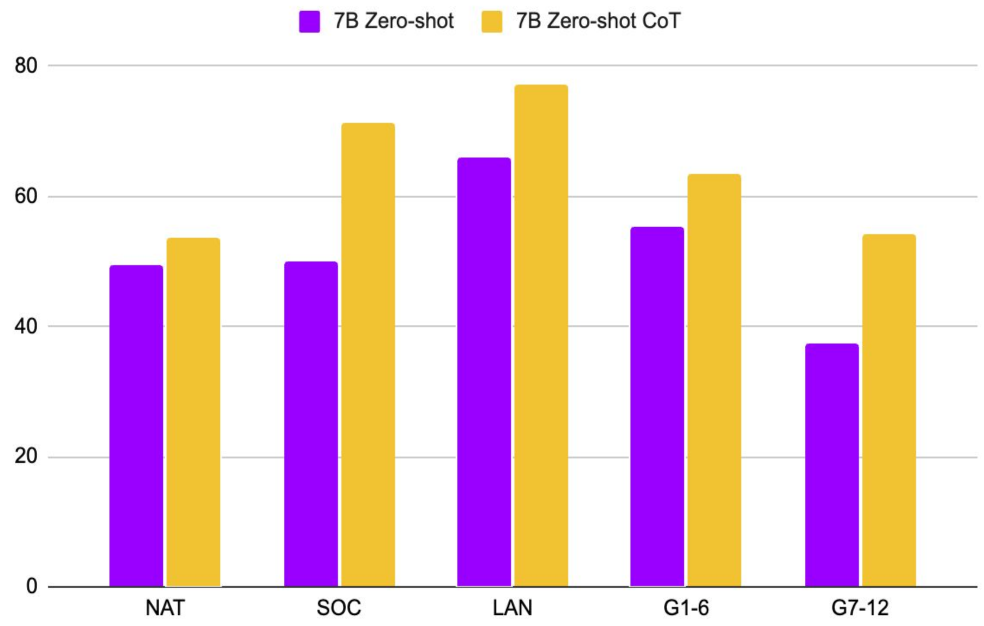
Experiment results:



Conclusions:

- M-CoT can significantly improve the performance of VLMs on complex multimodal reasoning tasks!
- The performance improvement bought by M-CoT increases with the model scale.

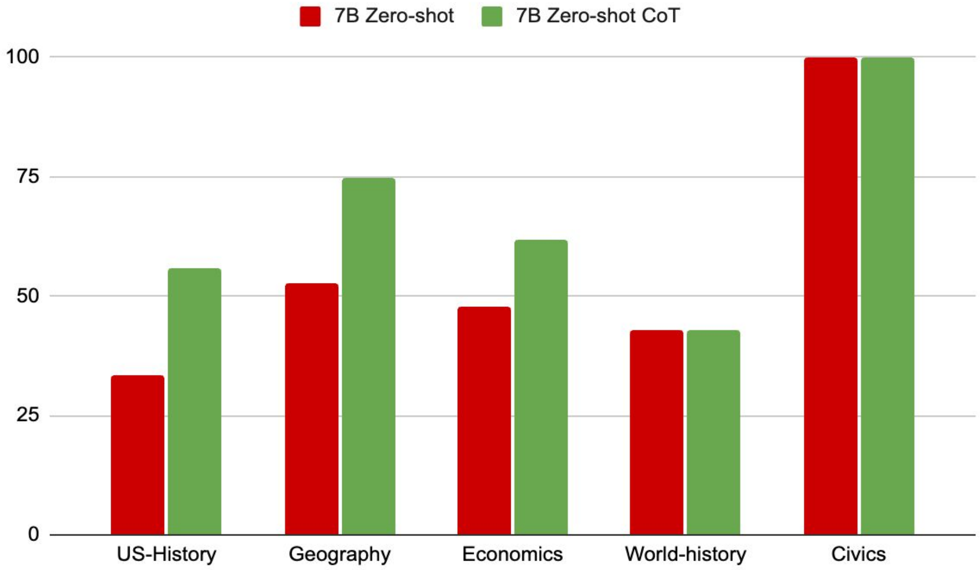
Results for different subjects and different degrees of difficulty of the question



Conclusions:

- M-CoT can improve the performance across all subjects.
- Social science questions have the highest improvement, and natural language questions have the lowest improvement.
- M-CoT can largely improve the accuracy of G7-12 questions.

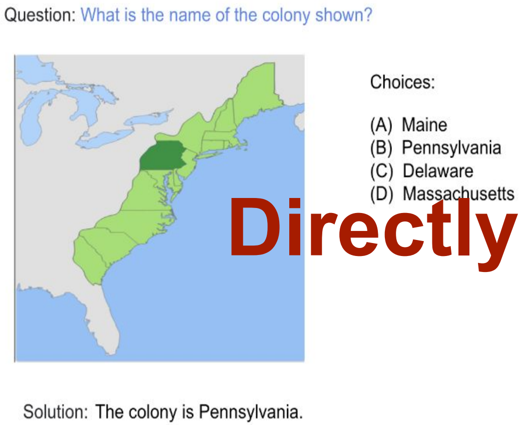
Results for five topics under the social science subject



Conclusions:

Among the five topics under the social science subject, questions in US-history and Geography have performance improvement over 20% after applying M-CoT.

Examples of US-history questions and Geography questions



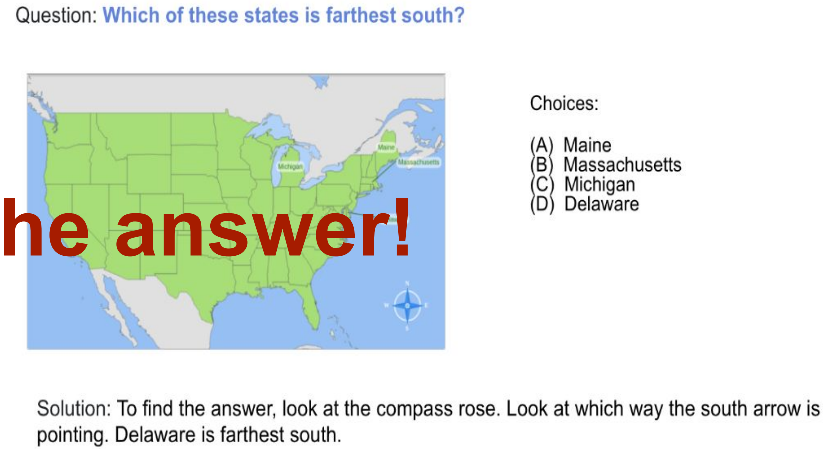
Question: What is the name of the colony shown?

Choices:

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

Solution: The colony is Pennsylvania.

Directly provide the answer!



Question: Which of these states is farthest south?

Choices:

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

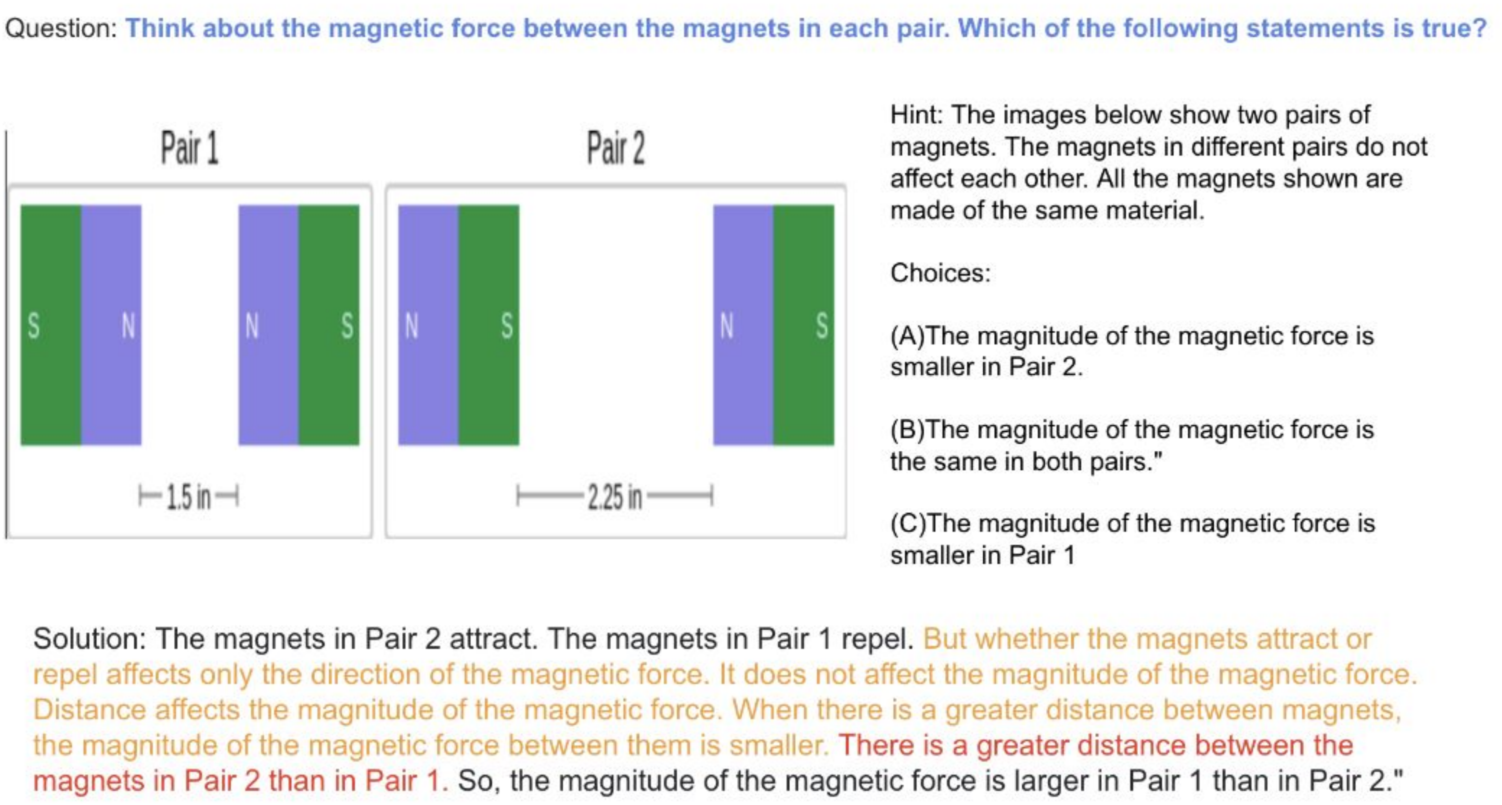
Solution: To find the answer, look at the compass rose. Look at which way the south arrow is pointing. Delaware is farthest south.

(a) An US-history question

(b) An geography question

M-(he
cor

-
- ce,



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.

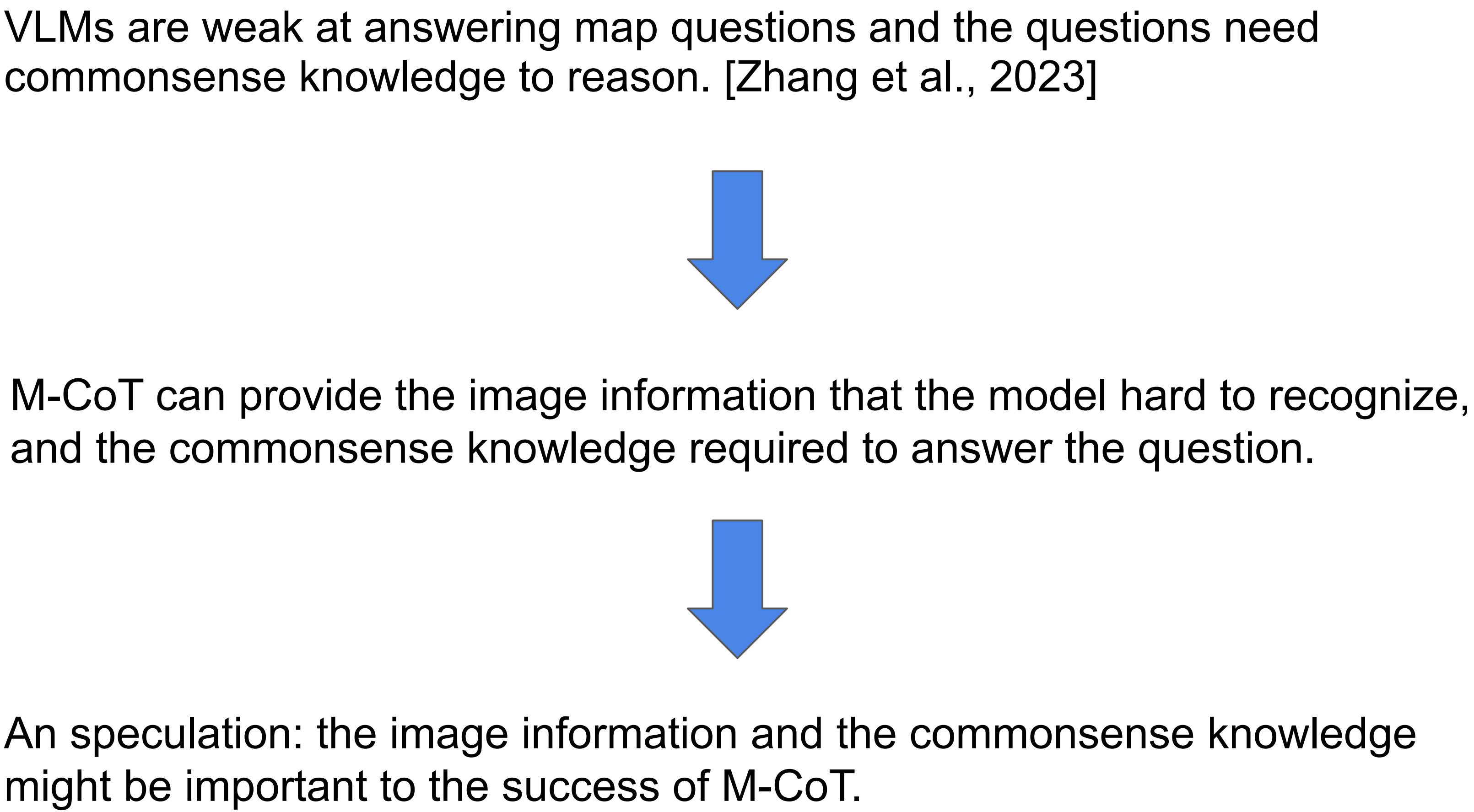
Choices:

(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."



VLMs are weak at answering map questions and the questions need commonsense knowledge to reason. [Zhang et al., 2023]

M-CoT can provide the image information that the model hard to recognize, and the commonsense knowledge required to answer the question.

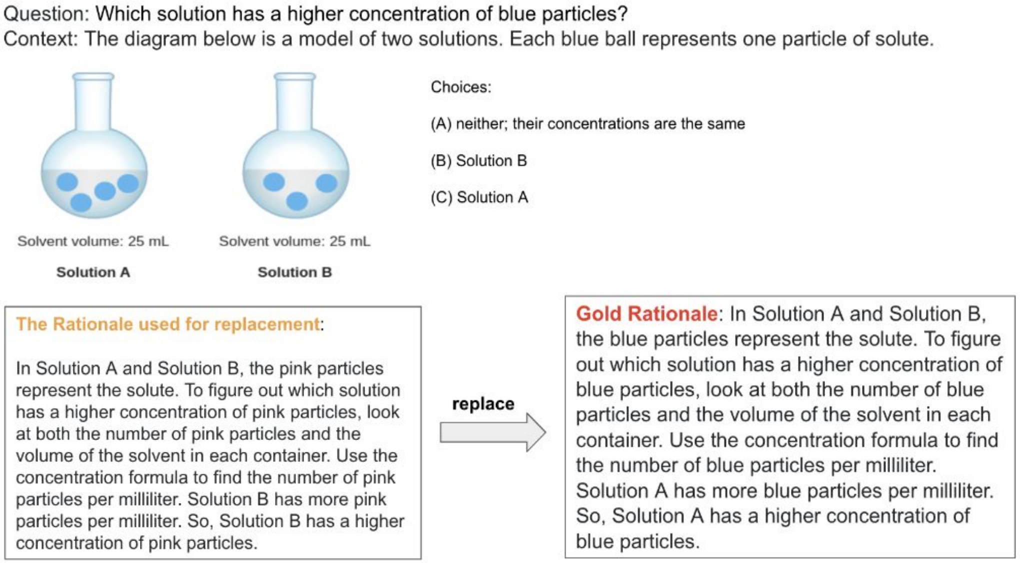
An speculation: the image information and the commonsense knowledge might be important to the success of M-CoT.

Why can M-CoT Improve the Performance of VLMs?

Hypothesis 1: The relevance in the textual part of M-CoT to the question is the important to the success of M-CoT.

Ablation experiment:

Use textual rationale sampled from the same topic to replace the original textual rationale.



Question: Which solution has a higher concentration of blue particles?
Context: The diagram below is a model of two solutions. Each blue ball represents one particle of solute.

Choices:

(A) neither; their concentrations are the same

(B) Solution B

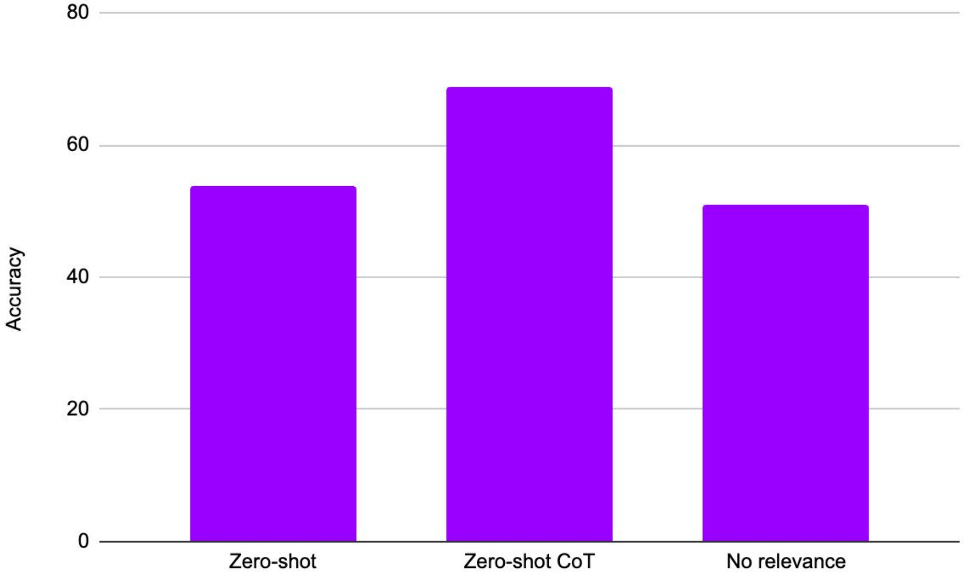
(C) Solution A

The Rationale used for replacement:

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 B has more pink particles per milliliter. So, Solution B has a higher concentration of pink particles.

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

Results to the relevance ablation experiment

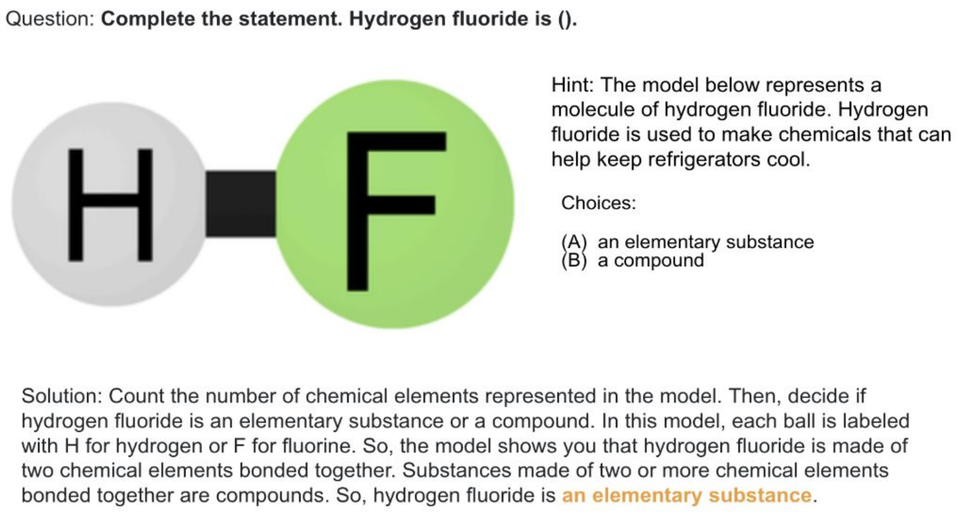


Conclusions:

- The relevance of textual part of M-CoT to the question is important to the improvement of M-CoT.
- Irrelevant M-CoT might mislead the model to make wrong decisions, which leads the irrelevant result lower than the zero-shot result.

Hypothesis 2: the validity of the reasoning chain affects the performance of M-CoT

Ablation experiment: destroy the validity by making the conclusion point to another option on a small M-CoT subset



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

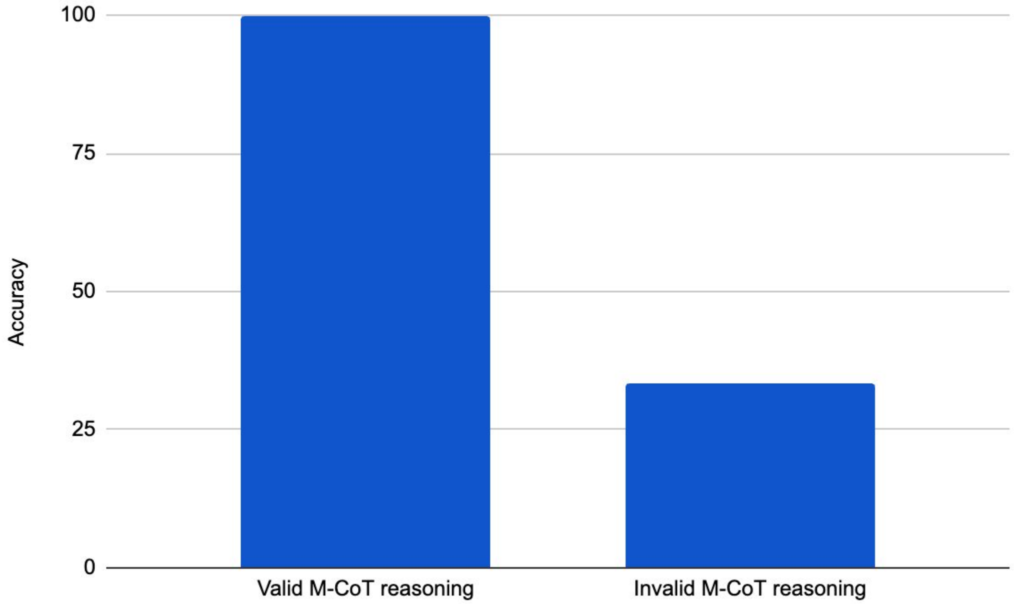
Hint: The model below represents a molecule of hydrogen fluoride. Hydrogen fluoride is used to make chemicals that can help keep refrigerators cool.

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**.

Results to the validity ablation experiment



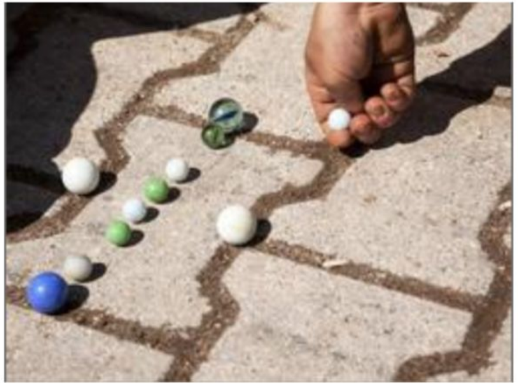
Conclusions:

- The validity of reasoning chain is important to the improvement of M-CoT.
- VLMs might reply on the conclusion part to answer the question.

Main limitation:

We only use one image for a question, which is the image context.

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



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

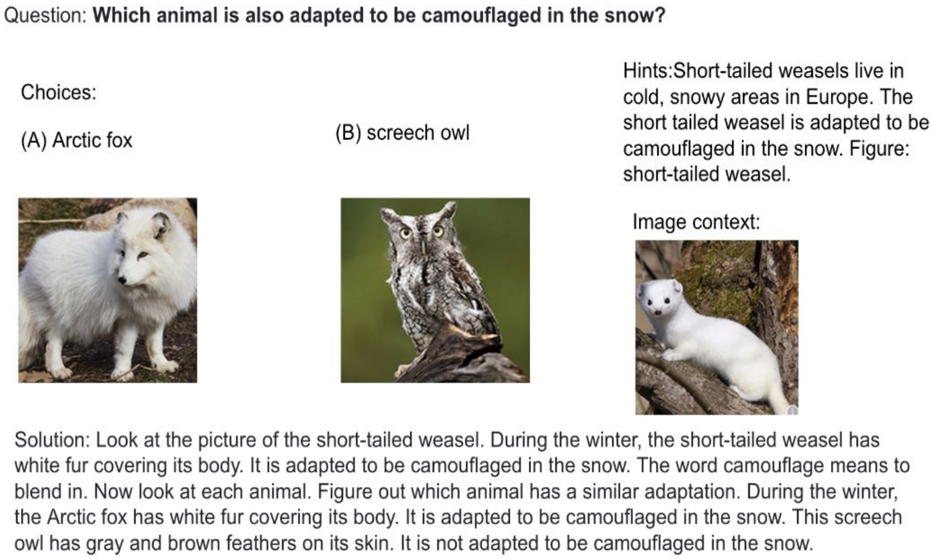
Choices:

(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.

However, there are some questions have more than one image.

The questions that have images for choices:



Question: **Which animal is also adapted to be camouflaged in the snow?**

Choices:

(A) Arctic fox

(B) screech owl

Hints:Short-tailed weasels live in cold, snowy areas in Europe. The short tailed weasel is adapted to be camouflaged in the snow. Figure: short-tailed weasel.

Image context:

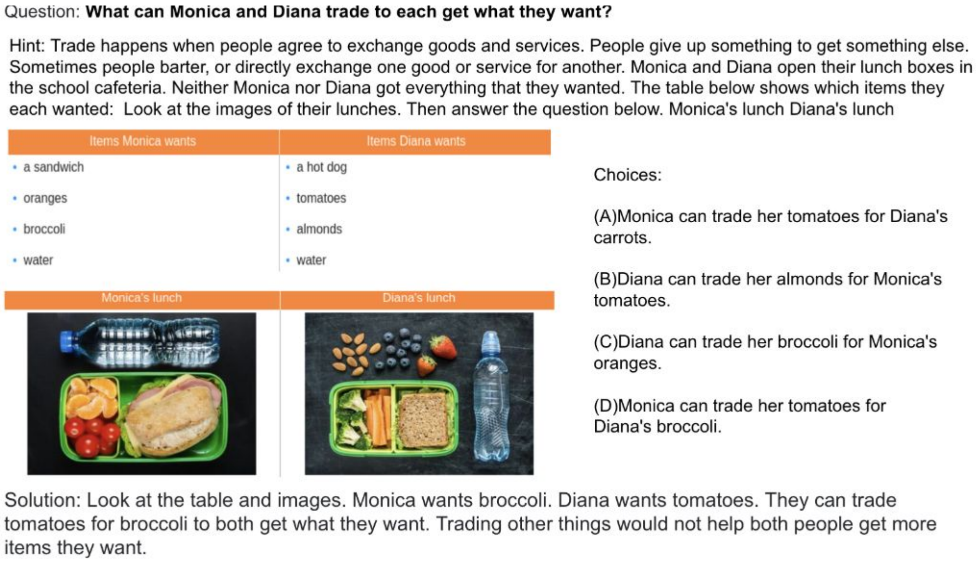
Solution: Look at the picture of the short-tailed weasel. During the winter, the short-tailed weasel has white fur covering its body. It is adapted to be camouflaged in the snow. The word camouflage means to blend in. Now look at each animal. Figure out which animal has a similar adaptation. During the winter, the Arctic fox has white fur covering its body. It is adapted to be camouflaged in the snow. This screech owl has gray and brown feathers on its skin. It is not adapted to be camouflaged in the snow.

The use of the images for choices: add complementary visual information.

The model might reason better with these images for choices.

But we neglect the images for choices for simplicity.

The trading questions that need two images to answer:



Question: **What can Monica and Diana trade to each get what they want?**

Hint: Trade happens when people agree to exchange goods and services. People give up something to get something else. Sometimes people barter, or directly exchange one good or service for another. Monica and Diana open their lunch boxes in the school cafeteria. Neither Monica nor Diana got everything that they wanted. The table below shows which items they each wanted: Look at the images of their lunches. Then answer the question below. Monica's lunch Diana's lunch

Items Monica wants	Items Diana wants
• a sandwich	• a hot dog
• oranges	• tomatoes
• broccoli	• almonds
• water	• water

Monica's lunch	Diana's lunch

Choices:

(A)Monica can trade her tomatoes for Diana's carrots.

(B)Diana can trade her almonds for Monica's tomatoes.

(C)Diana can trade her broccoli for Monica's oranges.

(D)Monica can trade her tomatoes for Diana's broccoli.

Solution: Look at the table and images. Monica wants broccoli. Diana wants tomatoes. They can trade tomatoes for broccoli to both get what they want. Trading other things would not help both people get more items they want.

Only occupies 3.5% of the test set.

We only feed the first image.

Providing all images to the model might lead to better performance.

Our conclusions are still valid under our implementation.

Our implementation can already explore our research questions.

↓

Only feeding the image context is enough for our experiment.

Providing all images to the model should be considered in the future.

Ideas for future work

- The improvement for the relevance experiment: use the rationales that are more similar to the gold rationale to replace

- Improvements for the validity experiment: test on all test set, use an LLM to revise the conclusion; apply different methods to destroy the validity (arbitrarily disrupting the order of sentences, revise the reasoning process but keep the conclusion, change both of the reasoning process and the conclusion, combine all the ways…); manually add more reasoning steps to make the reasoning process more rigorous

Ideas for future work

- The improvement for the relevance experiment: use the rationales that are more similar to the gold rationale to replace

- Improvements for the validity experiment: test on all test set, use an LLM to revise the conclusion; apply different methods to destroy the validity (arbitrarily disrupting the order of sentences, revise the reasoning process but keep the conclusion, change both of the reasoning process and the conclusion, combine all the ways…); manually add more reasoning steps to make the reasoning process more rigorous

Ideas for future work

- Explore whether image description is important for the success of M-CoT
- Conduct the attention analysis on different layers of VLMs
- Use few-shot exemplars to test whether M-CoT can improve the performance of VLMs via in-context learning

Key takeaways:

- M-CoT can largely improve the performance of VLMs in complex multimodal reasoning tasks, and the improvement is proportional to the model scale.
- M-CoT has proven its effectiveness in improving the accuracy across different subjects and difficulty levels of science questions in the ScienceQA dataset.
- The image information and the commonsense knowledge in the M-CoT prompt might contribute to boosting the reasoning ability of the model to answer challenging questions
- The relevance of the textual part of M-CoT and the validity of the reasoning chain of M-CoT are important to the improvement of M-CoT.
- VLMs might rely on the conclusion of the textual part of the M-CoT rationale to answer the question.

Utrecht University

Sharing science,
shaping tomorrow