Appendix

In this section, we present additional implementation details, experiment results and discussions. The content structure is outlined as follows:

A Additional Results

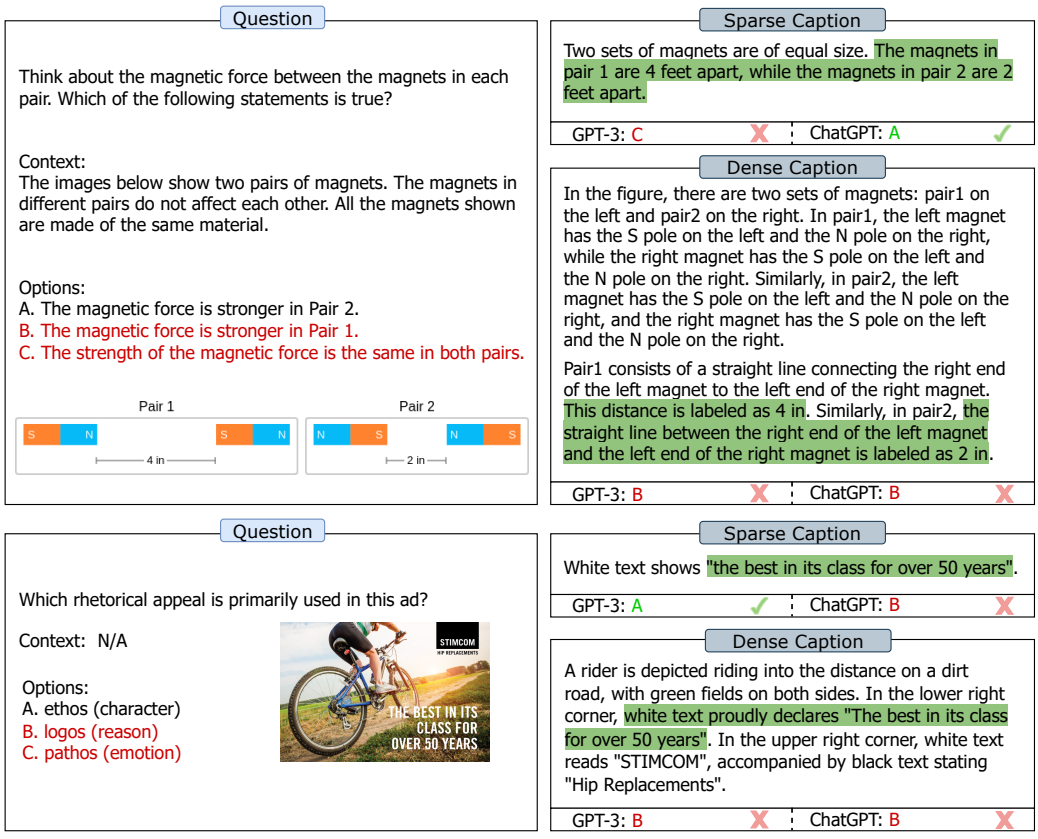


Figure 1: Cases that GPT-3 and ChatGPT face the difficulty in understanding dense image information.

A.1 Insights for Different GPT Models

In the submitted paper we have presented several findings concerning the LLMs. In this section, we aim to provide additional illustrative instances for GPT-3 [2] and the recent and potent ChatGPT [4].

Difficulty in understanding dense image information. Figure 1 presents additional instances of LLMs failing to comprehend dense image information. For sparse captions like the example in Figure 2 of the submitted paper, we observe that both GPT-3 and ChatGPT may struggle to comprehend image information in captions. Additionally, when prompted with more detailed captions, their difficulties in understanding become more pronounced. This highlights the challenges that even the most versatile and powerful language models currently available face in comprehending dense image information.

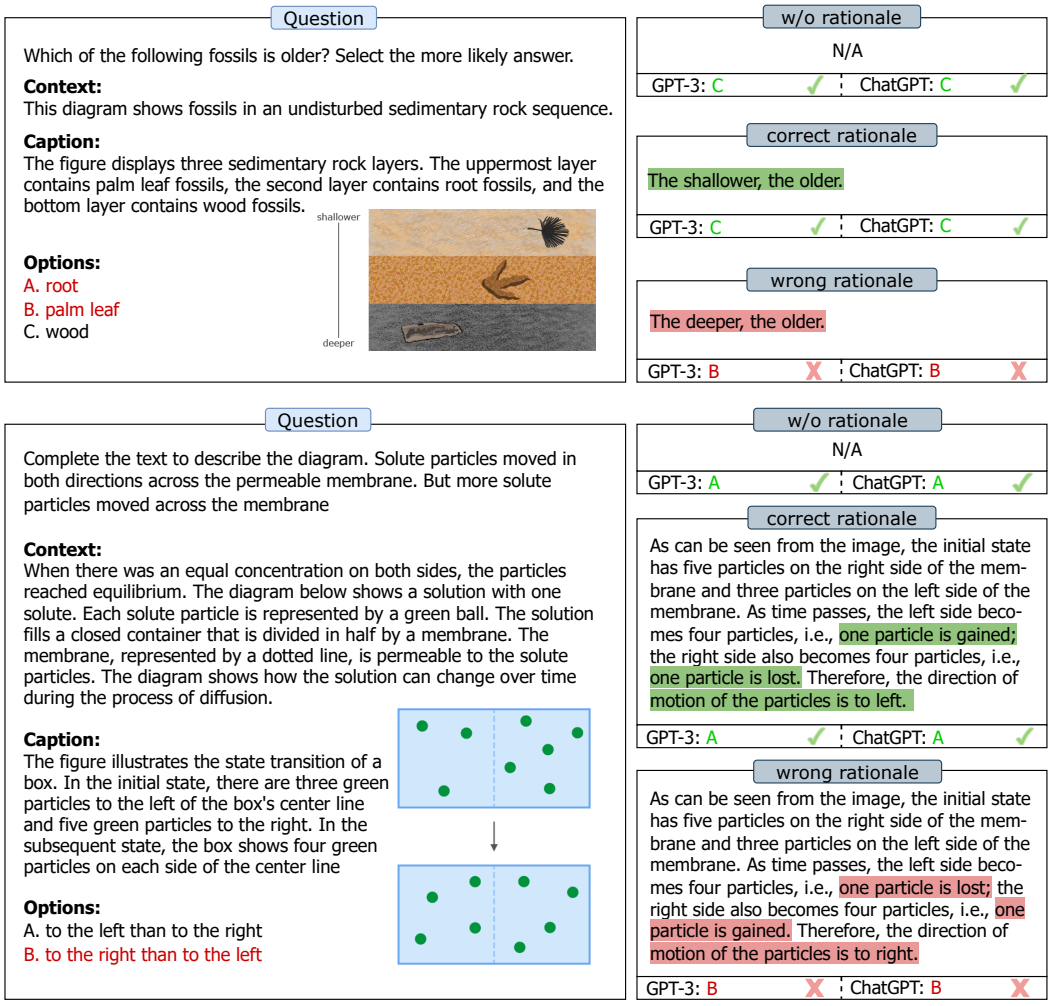


Figure 2: Cases of rationale-sensitive reasoning with GPT-3 and ChatGPT.

Rationale-sensitive reasoning. Figure 2 shows more examples where the reasoning of GPT-3 and ChatGPT is sensitive to the input rationales. The correct answers obtained without using rationales indicate that the LLMs possess commonsense knowledge to respond to the question. However, incorrect inputs can lead to misleading outcomes for both GPT-3 [2] and ChatGPT [4].

Note that the challenges involved in understanding dense image information, coupled with the difficulty of obtaining high-quality captions in practical scenarios, limit the universal applicability of reasoning without using rationales, although it may be feasible in the presented examples.

Hallucinations during generation. Figure 3 illustrates several instances where both GPT-3 and ChatGPT fall into hallucinations during rationale generation.

Our findings indicate that in cases where the provided caption information is insufficient, such as in the first example, both GPT-3 [2] and ChatGPT [4] tend to imagine image details to respond to the question. Even when we provide ample information, as demonstrated in the second and third

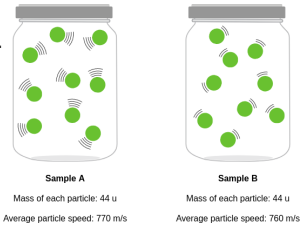
Question

Compare the average kinetic energies of the particles in each sample. Which sample has the higher temperature?

Context:
The diagrams below show two pure samples of gas in identical closed, rigid containers. Each colored ball represents one gas particle. Both samples have the same number of particles.

Caption:
The diagrams below show two pure samples A and B.

Options:
A. sample B
B. sample A
C. neither, the samples have the same temperature



GPT-3

Rationale: The particles in sample B have more kinetic energy on average than the particles in sample A. This means that sample B has a higher temperature.

ChatGPT

Rationale: The average kinetic energy of gas particles is directly proportional to the temperature of the gas. Sample B has a higher average kinetic energy because its particles are moving faster and have a wider range of velocities compared to Sample A. Therefore, Sample B has a higher temperature than Sample A.

Ours DDCoT w/o Caption

Rationale: The average kinetic energy of gas particles is directly proportional to temperature. The particles in sample A are moving at a much faster speed than the particles in sample B. Both samples have the same number of particles. Therefore, sample A has a higher temperature than sample B.

Question

Which continent is highlighted?

Context: N/A

Caption:
The picture shows a map of the world. In it, the area right in the middle is highlighted. It is roughly shaped as an inverted triangle. It has a sharp corner on the right and a rounded shape on the upper left. It is connected to a peninsula at the top right and a detached island at the bottom right.

Options:
A. Europe
B. Africa
C. Asia
D. South America

GPT-3

Rationale: The continent that is highlighted is Europe. This is because Europe is the only continent that is shaped as an inverted triangle and has a sharp corner on the right and a rounded shape on the upper left.

ChatGPT

Rationale: Based on the description of the highlighted area, we can identify that it is the Arabian Peninsula. The peninsula is connected to a detached island at the bottom right, which is the island of Socotra. The peninsula is also connected to a larger landmass at the top right, which is the continent of Asia. Therefore, the correct answer is Asia.

Ours DDCoT w/o Caption

Rationale: The highlighted area is a circle and has the Sahara desert as its geographical feature. The neighboring countries are not relevant to the question. By looking at the image and considering the supplementary information, we can determine that the highlighted area is in Africa.

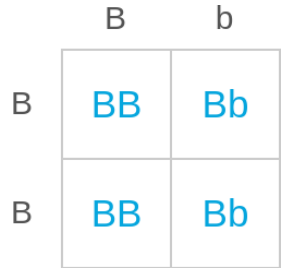
Question

What is the probability that a rat produced by this cross will have a dwarf body?

Context:
In a group of rats, some individuals have a normal-sized body and others have a dwarf body. In this group, the gene for the body size trait has two alleles. The allele for a dwarf body (b) is recessive to the allele for a normal-sized body (B). This Punnett square shows a cross between two rats.

Caption:
The image shows a table of genes. The horizontal axis is Bb, and the vertical axis is BB

Options:
A. 3/4
B. 0/4
C. 2/4
D. 1/4
E. 4/4



GPT-3

Rationale: The Punnett square only shows the possible genotypes of the offspring, not the phenotype. In this case, the phenotype is determined by the body size allele, which is controlled by the B gene. The B allele is dominant, so the phenotype of the offspring will be determined by the B allele. The B allele is present in 3 out of the 4 possible genotypes, so the probability that the offspring will have a normal-sized body is 3/4.

ChatGPT

Rationale: The Punnett square shows the possible genotypes of the offspring. To determine the probability of a rat having a dwarf body, we need to look at the genotype of the rat. If a rat has at least one copy of the recessive allele (bb), it will have a dwarf body. Looking at the Punnett square, we can see that there are four possible genotypes of the offspring: BB, Bb, Bb, and bb. Only the last genotype (bb) will result in a dwarf body. There is only one way to get a bb genotype, which is by crossing two Bb rats. Therefore, the probability of a rat produced by this cross having a dwarf body is 1/4 or 25%.

Ours DDCoT w/o Caption

Rationale: The Punnett square shows that the genotypes of the parent rats are Bb and Bb. The possible genotypes of offspring are BB, Bb, and bb. The allele for a dwarf body (b) is recessive to the allele for a normal-sized body (B). Therefore, only the bb genotype will result in a dwarf body. The probability of an offspring inheriting the recessive allele for dwarf body (bb) is 1/4 or 25%.

Figure 3: Cases of hallucinations when generating rationales with GPT-3 and ChatGPT.

examples, we observe that both language models still may fall into hallucinations that do not align with the provided information.

In contrast to generating rationales solely based on the caption and question information, our approach can alleviate the hallucinations to some extent by decomposing the questions into simple recognition tasks and emphasizing the uncertainty of image-related aspects. It is worth noting that while the former direct methods rely on manually crafted high-quality rationales, our approach utilizes BLIP-2 [3] as a visual question answering (VQA) model, serving as a visual component. However, our approach outperforms the former methods in terms of performance and interpretability.

Unfortunately, it is very challenging to entirely solve the issues of hallucinations. Although we alleviate hallucinations, we still encountered difficulties in certain cases, such as the third example, and further details regarding this limitation will be explored in section E.

	IMG	TXT	Avg
baseline(B)	72.93	85.84	79.70
B + our R	75.81	82.40	82.83
B + gt R	81.07	84.07	85.97
our model	75.16	85.25	80.45
our model + our R	83.34	91.20	87.34
our model + gt R	84.43	92.09	88.00

Table 1: Quantitative ablations on our DDCoT and visual components.

A.2 Quantitative ablations on our DDCoT and visual components

We conduct additional ablation study on the extent of impact exerted by our DDCoT prompting and visual components, as shown in Table 1. DDCoT prompting and visual components cooperatively facilitate inducing visual information to language models for multimodal reasoning. We can observe that rationales generated by our DDCoT and visual components individually exhibit certain gains in terms of IMG improvement. However, when combined, they yield substantial gains.

Besides, please note that the annotated ground truth rationales within the ScienceQA dataset inherently encompass the final prediction, i.e., correct answers. To ensure a fair comparison, we manually exclude the answers from these annotations, using the remaining text as input rationales for fine-tuning. Under this configuration, our proposed rationale achieves a fine-tuning performance comparable to the annotated rationales.

N_p	1	3	5	N_r	8	16	32
Avg	86.72	**87.34**	86.02	Avg	86.63	**87.34**	86.30

Table 2: Ablation of N_p and N_r.

A.3 Hyperparameters for Fine-tuning Components

Table 2 presents the results of our ablation studies on N_p, which denotes the number of learnable prompt tokens, and N_r, which represents the number of low-rank vectors. Regarding N_p, we conducted experiments using values of 1, 3, and 5. Increasing the number of prompts introduces more learnable parameters and provide stronger guidance to align vision and language, while too many prompts may disrupt the model's comprehension of visual features and result in a decline in performance. Considering N_r, we investigated the influence of different filtering intensity in the Rational-Compressed Visual Embedding process by experimenting with values of 8, 16, and 32. The experimental results indicate that selecting 3 for N_p and 16 for N_r yields the best performance.

A.4 Additional experiments on the effectiveness of our DDCoT with existing pre-trained VLMs and multimodal reasoning models

We also validate the effectiveness of our DDCoT with existing pre-trained VLMs and multimodal reasoning models, as shown in Table 3. We observe that rationales generated by our proposed DDCoT is compatible with such pre-trained VLMs. Without correct rationales, existing pretrained

Model	NAT	SOC	LAN	TXT	IMG	NO	G1-6	G7-12	Avg
Flamingo [1]	40.28	46.13	29.25	47.45	40.08	33.66	39.35	40.67	39.83
Flamingo with our R	40.28	46.13	29.25	47.45	40.08	33.66	39.35	40.67	39.83
MiniGPT-4 [6]	40.28	46.13	29.25	47.45	40.08	33.66	39.35	40.67	39.83
MiniGPT-4 with our R	40.28	46.13	29.25	47.45	40.08	33.66	39.35	40.67	39.83

Table 3: The effectiveness of our DDCoT with Flamingo [1] and MiniGPT-4 [6]

VLMs and multimodal reasoning models have difficulty in complex reasoning tasks. Fortunately, the generalizable rationales generated by our DDCoT prompting can help existing VLMs to comprehend visual information and reason with rich knowledge, achieving significant improvement of 11.14% and 10.96% based on Flamingo [1] and Mini GPT-4 [6] as Table 3 shows.

	NoCaps				MSVD-QA
	CIDEr	SkipThoughtCS	EmbeddingAverageCS	GreedyMatchingScore	Acc
BLIP-2	76.15	49.84	89.20	77.94	34.4
Ours	46.26	84.78	92.35	79.12	39.3

Table 4: Addition experiments on NoCaps and MSVD-QA.

A.5 Additional experiments on Captioning and Video Question Answering tasks

We extend our approach to more appropriate datasets. While other existing datasets may not fully exploit the benefits of our approach, we venture into exploring the captioning task on NoCaps and the video question-answering task on MSVD-QA, as shown in Table 4.

For captioning, we prompt the LLM to solve sub-problems derived from a simple caption, aiming to optimize and enrich it using the corresponding sub-answers. The substantial knowledge within LLM enables the generation of semantically enriched captions, leading to improvements in metrics evaluating sentence semantics, i.e. 34.94%, 3.15%, and 1.18% in terms of SkipThoughtCS, EmbeddingAverageCS, and GreedyMatchingScore. Note that the CIDEr metric to evaluate ours is limiting. It is designed to measure the similarity between the tested caption and reference captions without considering the diversity and high-level semantics.

For video question answering, LLM deconstructs problems like the decomposition step on ScienceQA. We sample video frames for VQA recognition and integrate frame information for multimodal rationale and answers. Leveraging the sequence understanding in LLM and visual information returned by the VQA model, we achieve a 4.9% improvement over BLIP-2.

Note that we randomly evaluated only 1000 images from NoCaps and 1000 videos from the MSVD test dataset in a zero-shot setting.

B Human Evaluation

This section introduce the details of our human evaluation. Figure 4(b) shows an example of our question. Each sample comprises the question, context, options, and image. Evaluators are asked to rate the rationales generated by GPT [2], MM-COT [5], and our method in five aspects: relevance (related to the question), correctness (accuracy of reasoning and answer), completeness (logical reasoning's comprehensiveness), coherence (consistency of reason), and explainability (interpretability of reasoning and answer). The rating scale ranges from 0 to 5. Additionally, the questions and rationales are organized into 12 groups, with each group assigned to three evaluators. Finally, we average the scores for each aspect of each rationale, resulting in overall scores and ratios relative to the maximum score.

C Detailed Prompts

This section introduce more details about our zero-shot DDCoT prompting. We employ ChatGPT[4] as the most important component of our rationale generator. Specifically, it is utilized in breaking

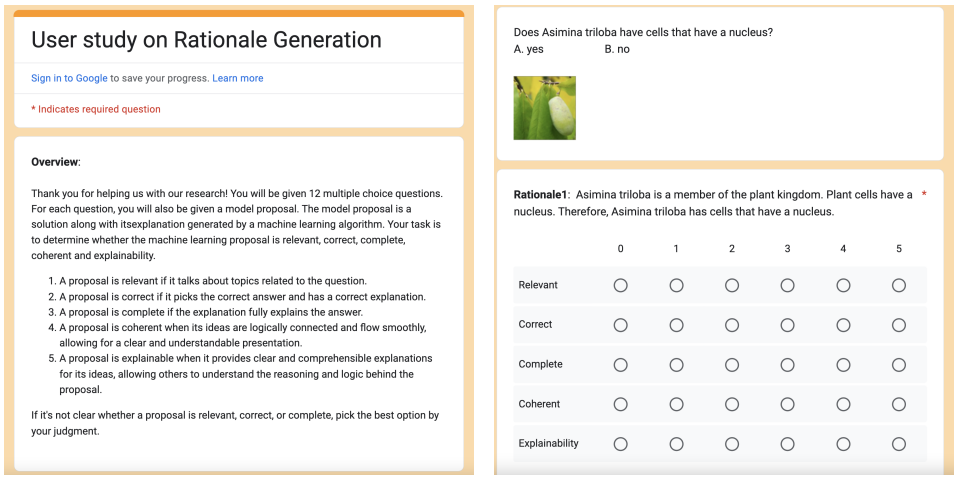


Figure 4: Interface of human evaluation. Figure(a) shows the instructions, figure(b) shows one example of our question.

duties of reasoning and recognition step and joint reasoning step. Figure 5 shows complete prompts for zero-shot DDCoT prompting.

D Case Studies

To better understand the effectiveness of our proposed method in generating rationales, we randomly selected several cases from the test set along with the process of rationale generation. Figure 6 showcases several map-related questions, demonstrating how our method integrates simple visual features (such as the shape of highlighted areas) with common knowledge to obtain correct reasoning and answers. In the examples presented in Figure 7, our method successfully identifies within the images, acquiring relevant knowledge. Figure 8 illustrates four more complex questions, where our method leverages information obtained from the images to perform intricate reasoning. However, when it comes to the complex interaction between images and textual context, our method still fall into hallucinations, leading to erroneous reasoning and answers.

E Limitations

While we have succeeded in mitigating a portion of the hallucination problem arising from multimodal inputs, it is important to acknowledge that this problem is not entirely resolved. As illustrated in Figure 8, our model is also susceptible to the risk of hallucinations. Investigating methods to suppress hallucinations is a potential topic for further research and exploration. In addition, we did not use extra image-text pairs to pre-train the alignment between vision and language modalities. Such pre-trianing is expected to further improve the alignment for joint reasoning.

As our approach involves zero-shot prompting of the LLM to generate rationales, there exists a potential risk of inheriting social biases from the LLM. These biases, which encompass cultural, ethical, and various other dimensions, might be reflected in the generated rationales, potentially leading to adverse effects on users. To mitigate this issue in the future, potential solutions could involve designing constraints at each prompting stage or utilizing more advanced LLMs trained on unbiased resources.

References

[1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35:23716–23736, 2022.

DDCoT

Breaking Reasoning Down to Recognition Steps with Negative-Space Prompting

System: You are a helpful, highly intelligent guided assistant. You will do your best to guide humans in choosing the right answer to the question. Note that insufficient information to answer questions is common, because you do not have any information about the picture. The final answer should be one of the options.

User: Given the context, questions and options, please think step-by-step about the preliminary knowledge to answer the question, deconstruct the question as completely as possible down to necessary sub-questions based on context, questions and options. Then with the aim of helping humans answer the original question, try to answer the sub-questions. The expected answering form is as follows:
Sub-questions:
1. <sub-question 1>
2. <sub-question 2>
...
Sub-answers:
1. <sub-answer 1> or 'Uncertain'
2. <sub-answer 2> or 'Uncertain'
...
Answer: <One of the options> or 'Uncertain'

For a question, assume that you do not have any information about the picture, but try to answer the sub-questions and prioritize whether your general knowledge can answer it, and then consider whether the context can help. If sub-questions can be answered, then answer in as short a sentence as possible. If sub-questions cannot be determined without information in images, please formulate corresponding sub-answer into "Uncertain".
Only use "Uncertain" as an answer if it appears in the sub-answers. All answers are expected as concise as possible.
Here is an attempt:
Context: {context}
Has An Image: {has_image}
Question: {question}
Options: {option}"

Visual Recognition to Obtain Visual Complements

~ Aribitrary Visual Question Answering (VQA) Model ~

Integrate to Joint Reasoning

System: You are a helpful, highly intelligent teacher. You will not only do your best to guide humans to the correct answer, but you will also give the rationales as a reference.

User: Given the context, questions, options, supplementary information, think step by step and answer the questions. Please note that we need not only the answer, but more importantly the rationales of getting the answer. The expected answering form is as follows:
Rationale: <rationale>
Answers: <one of the options>

Please note that the supplementary information given may not always be valid. Please select valid information to form the rationale and choose the relatively correct option as your answer.
Here is an attempt:
Context: {context}
Has An Image: {has_image}
Question: {question}
Options: {option}
Supplementary information: {supplementary_information}

Figure 5: The complete prompts of our zero-shot DDCoT prompting. The underlined parts are introduced in Section 3.2 of the submitted paper. The yellow background represents designs for critical thinking, and the blue background represents designs for differentiate duties of reasoning and recognition. "{}" denotes the corresponding inputs, and "{supplymentary_information}" represents the sub-questions and sub-answers obtained before.

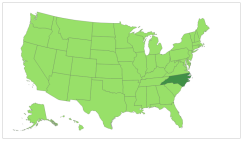
Question

What is the capital of North Carolina?

Context: Select the better answer.

Options:

A. Saint Paul B. Fargo C. Raleigh D. Charlotte



Ours Outputs

Sub-question:
(1) What is the location of North Carolina?
(2) What is the capital of North Carolina?

Sub-answers: (1) North Carolina is a state located in the southeastern region of the United States. (2) The capital of North Carolina is Raleigh.

VQA: Understanding[sub-questions, image]

Sub-answers': (1) North Carolina is a state located in the southeastern region of the United States. (2) The capital of North Carolina is Raleigh.

Rationale: The question is asking for the capital of North Carolina. The supplementary information confirms that North Carolina is a state located in the southeastern region of the United States and the capital of North Carolina is Raleigh.

GPT3(COT) Outputs

BECAUSE: Raleigh is the capital of North Carolina.

Question

Which property matches this object?

Context: Select the better answer.

Options:

A. North America B. South America
C. Antarctica D. Australia

Ours Outputs

Sub-question:
(1) What is the shape of the highlighted area?
(2) What are the geographical features of the highlighted area?
(3) What are the surrounding bodies of water of the highlighted area?

Sub-answers: (1) Uncertain (2) Uncertain (3) Uncertain

VQA: Understanding[sub-questions, image]

Sub-answers': (1) a circle (2)a mountain range (3) atlantic ocean, pacific ocean, and arctic ocean

Rationale: The highlighted area is a circle with a mountain range, and it is surrounded by the Atlantic Ocean, Pacific Ocean, and Arctic Ocean. Based on this information, we can eliminate Antarctica and Australia as they are not surrounded by these three oceans. The mountain range in the highlighted area is likely the Rocky Mountains, which are located in North America.

GPT3(COT) Outputs

BECAUSE: Antarctica is the world's fifth largest continent. It is also the coldest, windiest, and driest continent.

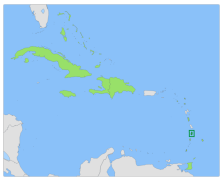
Question

Which country is highlighted?

Context: Select the better answer.

Options:

A. Haiti B. Saint Lucia
C. Saint Vincent and the Grenadines



Ours Outputs

Sub-question:
(1) What is the image showing?
(2) What are the distinguishing features of the highlighted country?
(3) What are the geographical locations of the countries in the options?

Sub-answers: (1) Uncertain (2) Uncertain (3) Uncertain

VQA: Understanding[sub-questions, image]

Sub-answers': (1) the location of the caribbean (2) the southeastern coast is a peninsula (3) the caribbean islands are located in the western hemisphere

Rationale: The image is showing the location of the Caribbean. The supplementary information tells us that the highlighted country has a southeastern coast that is a peninsula. Based on the options, only Haiti and Saint Lucia have a southeastern coast that is a peninsula. However, Haiti does not have a peninsula that is as pronounced as the one in the image.

GPT3(COT) Outputs

BECAUSE: The country highlighted is Saint Vincent and the Grenadines.

Question

Which of these continents does the prime meridian intersect?

Context: Select the better answer.

Options:

A. South America B. Asia C. Antarctica

Ours Outputs

Sub-question:
(1) What is the prime meridian?
(2) Which continents does the prime meridian pass through?
(3) Is the prime meridian shown in the image?

Sub-answers: (1) The prime meridian is the line of longitude that is designated as 0° longitude. It is the line that divides the Earth into the Eastern and Western Hemispheres. (2)The prime meridian passes through Europe, Africa, and Antarctica. (3) Uncertain

VQA: Understanding[sub-questions, image]

Sub-answers': (1) The prime meridian is the line of longitude that is designated as 0° longitude. It is the line that divides the Earth into the Eastern and Western Hemispheres. (2)The prime meridian passes through Europe, Africa, and Antarctica. (3) no

Rationale: The prime meridian is the line of longitude that is designated as 0° longitude. It is the line that divides the Earth into the Eastern and Western Hemispheres. The supplementary information states that the prime meridian passes through Europe, Africa, and Antarctica. The image does not show the prime meridian.

GPT3(COT) Outputs

BECAUSE: The prime meridian intersects with Asia.

Figure 6: Cases that need to understand the map.

Question

Select the mammal below.

Context: Mammals have hair or fur and feed their young milk. A gorilla is an example of a mammal.

Options:

A. poison dart frog B. rabbit

Ours Outputs

Sub-question:
(1) Does the image show a mammal?
(2) Is the animal in the image a poison dart frog or a rabbit?

Sub-answers: (1) Uncertain (2) Uncertain

VQA: Understanding[sub-questions, image]

Sub-answers': (1) Yes (2)poison dart frog

Rationale: The given context states that mammals have hair or fur and feed their young milk. The supplementary information confirms that the animal in the image is a mammal. Therefore, we need to select the option that represents a mammal. The option 'poison dart frog' is not a mammal as it does not have hair or fur and does not feed its young milk. On the other hand, the option 'rabbit' is a mammal as it has fur and feeds its young milk.

GPT3(COT) Outputs

BECAUSE:A poison dart frog is not a mammal.

Question

Select the fish below.

Context: Fish live underwater. They have fins, not limbs. A tiger shark is an example of a fish.

Options:

A. salmon B. woodpecker

Ours Outputs

Sub-question:
(1) Is salmon a fish?
(2) Is woodpecker a fish?

Sub-answers: (1) Yes (2) Uncertain

VQA: Understanding[sub-questions, image]

Sub-answers': (1) Yes (2)no, it is a bird

Rationale: Fish live underwater and have fins, not limbs. The given image shows a fish, which means we need to select an option that is also a fish. The supplementary information confirms that salmon is a fish, while woodpecker is a bird.

GPT3(COT) Outputs

BECAUSE:Salmon are fish. Woodpeckers are not fish.

Question

Which material is this teapot made of?

Context: Select the better answer.

Options:

A. porcelain
B. rock

Ours Outputs

Sub-question:
(1) What does the teapot look like?
(2) What are the characteristics of porcelain and rock?

Sub-answers: (1) Uncertain (2) Porcelain is a type of ceramic material that is often white and translucent. It is made by heating materials, generally including clay in the form of kaolin, in a kiln to temperatures between 1,200 and 1,400 °C. Rock is a naturally occurring solid substance composed of minerals.

VQA: Understanding[sub-questions, image]

Sub-answers': (1) a pink and green teapot with pink roses (2) Porcelain is a type of ceramic material that is often white and translucent. It is made by heating materials, generally including clay in the form of kaolin, in a kiln to temperatures between 1,200 and 1,400 °C. Rock is a naturally occurring solid substance composed of minerals.

Rationale: The question asks about the material of the teapot, and the options are porcelain and rock. Porcelain is a type of ceramic material that is often white and translucent, while rock is a naturally occurring solid substance composed of minerals. The supplementary information describes porcelain as being made by heating materials, generally including clay in the form of kaolin, in a kiln to temperatures between 1,200 and 1,400 °C. The image shows a pink and green teapot with pink roses, which is consistent with the appearance of porcelain teapots. Therefore, the material of this teapot is most likely porcelain.

GPT3(COT) Outputs

BECAUSE:Teapots can be made from a variety of materials, including porcelain, metal, glass, and clay.

Question

Which property matches this object?

Context: Select the better answer.

Options: A. transparent.
B. bouncy

glass bottle

Ours Outputs

Sub-question:
(1) What is the color of the object?
(2) What is the shape of the object?

Sub-answers: (1) Uncertain (2) Uncertain

VQA: Understanding[sub-questions, image]

Sub-answers': (1) white (2) a glass bottle

Rationale: The object is a glass bottle and the color is white. Based on this information, the property that matches the object is "transparent" as glass is a transparent material and allows light to pass through it.

GPT3(COT) Outputs

BECAUSE: The definition of transparent is "able to be seen through with clarity." The definition of bouncy is "able to rebound or spring back into shape after being compressed."

Figure 7: Cases that need to identify objects and acquire relevant knowledge.

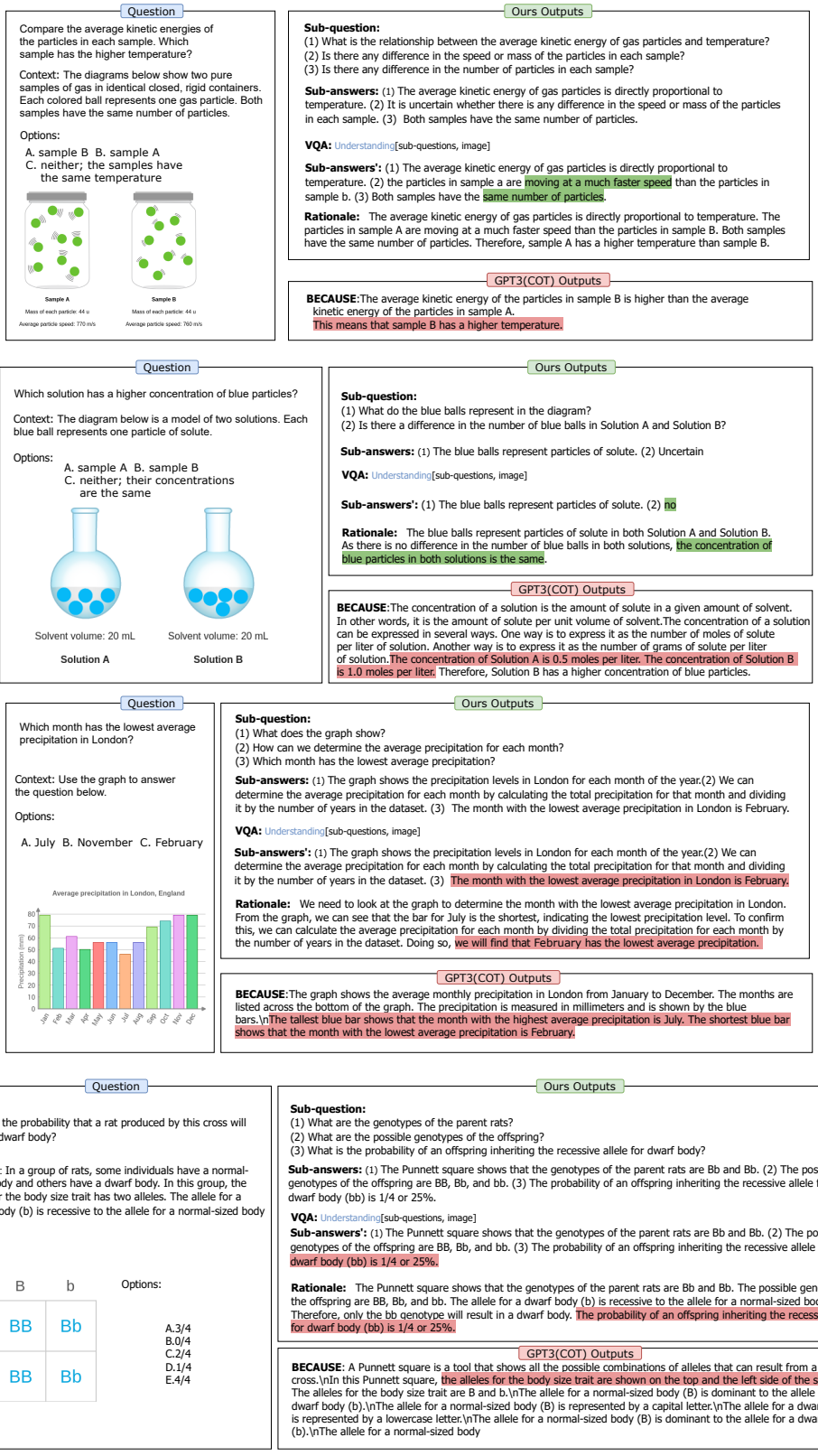


Figure 8: Complex and Challenging cases.

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

[3] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023.

[4] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744, 2022.

[5] Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. Multimodal chain-of-thought reasoning in language models. *arXiv preprint arXiv:2302.00923*, 2023.

[6] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023.