LLMs is Blind: LLMs can answer the Visual-Question-Answering without seeing the image.

Yeaeun Kwon
University of Arizona
the College of Information Science
yeaeunkwon@arizona.edu

Abstract

Vision for humans is another significant channel for processing information. However, human cognitive ability can visualize an image based solely on descriptive text, using prior experiences and accumulated knowledge. This project investigates the ability of a text-only LLM in a VQA task when provided the image descriptions, instead of the original image, with QA pairs, compared to that of a multi-modal LLM. The standard and Chain-of-Thought(CoT) prompts are given to Gpt-3.5-turbo, and the image descriptions are obtained from LLaVa. The results illustrate that CoT strategy can enhance and surpass the multi-modal LLM in a VQA task even though the images are not directly provided to the text-only LLM. The experiments were conducted on the M3CoT dataset, addressing multiple domains such as mathematics, science, and commonsense.

1 Introduction

The advances in Large language models have led to development of instruction-tuned models like FLAN-T5(Chung et al., 2022) and Vicuna (Chiang et al., 2023), moving beyond previous fine-tuning language models like BERT (Devlin et al., 2019) and RoBERTa (Zhuang et al., 2021). These instruction-tuned models have achieved performance comparable to or sometimes surpassing humans in various tasks. In particular, Chain of Thought (Wei et al., 2022) has demonstrated significant improvements in the reasoning abilities of LLMs. While the ultimate goal of AI systems may have the comparable cognitive abilities with human, humans possess an additional crucial channel to processing information — vision. Human reasoning and understanding are greatly enhanced not only by text but also when visual information is provided. Images can convey concepts that are difficult to describe in words, providing intuitive representations and aiding in the comprehension of complex ideas. As a result, there has been a growing interest and importance in research on Vision models, where both text and images are used to improve the performance in various tasks (Alayrac et al., 2022; Radford et al., 2021; Liu et al., 2023, 2024).

Another distinctive human capability is imagination. Humans can visualize an image based solely on descriptive text, using prior experiences and accumulated knowledge to construct a representation that may not be identical to the actual image but captures its essence. This raises a key question for NLP research: Can LLMs, without direct visual embeddings, interpret textual descriptions of images and leverage this information to generate answers related to those images?

This project explores this question using the M3CoT dataset (Chen et al., 2024), a multi-modal and multi-domain dataset. It involves generating image descriptions from the M3CoT dataset and assessing the performance of LLMs in a QA task when given these descriptions instead of direct visual input. Ultimately, this project seeks to determine whether a text-only LLM, when provided with image descriptions and enhanced through CoT prompting and fine-tuning, can achieve performance that is comparable to or even better than multi-modal LLM. The overall framework is detailed in Figure 1.

2 Related Works

Chain of Thought(CoT)

With the advent of the instruction-tuned models such as ChatGPT, FLAN-T5, and FLAN-PaLM (Chung et al., 2022), researchers have increasingly explored methods to enhance performance through zero-shot and few-shot prompting. Chain-of-Thought(CoT) is a strategy guiding models through a step-by-step reasoning process, where the model follows a logical reasoning path to an an-

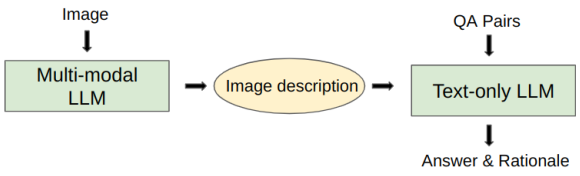

Figure 1: The overall framework of the Visual-Question-Answering task on the text-only LLM in this project.

swer, achieving significant boosts in performance across various tasks. Wei et al. (2022) proposed Chain of thought (CoT) as an application of few-shot prompting by modifying the answers in few-shot examples. Subsequently, zero-shot CoT was introduced as a method to improve model performance by simply appending the phrase "Let's think step by step" to prompts (Kojima et al., 2022). In this project, I experiment zero-shot and few-shot CoT approaches to assess the model's reasoning ability in a Visual-Question-Answering(VQA) task.

Visual Question Answering(VQA)

Visual Question Answering(VQA) provides a model visual information with natural language. Typically, a model is given pairs of questions and relevant images. One approach to deal with the combined information with visual and textual content involves jointly training an image encoder and text encoder as demonstrated in CLIP (Radford et al., 2021). Another approach converts an image to a text describing the image. Various image captioning techniques have been proposed such as providing single-sentence description and to more detailed paragraphs (Kim and Bansal, 2019). Recently, multi-modal large language models (LLMs) have been developed to process inputs that include text, images, and audio, such as GPT-4o (Wu et al., 2024) and LLAVA (Liu et al., 2023, 2024).

In this project, I aim to investigate text-only LLMs ability to extract important information for answering questions from image descriptions, which can be interpreted inference visual content based solely on textual descriptions. This ability is compared with that of LLaVa, an end-to-end trained large multimodal model. The image captioning is obtained from the visual instruction-tuned model, `llava-v1.6-mistral-7b-hf`.

3 Dataset

Among the various VQA datasets, The data used in this project is M3CoT (Chen et al., 2024), which is a multi-modal and multi-domain dataset. M3CoT includes content in science, mathematics, and commonsense, enabling an assessment of the understanding capabilities of LLMs across diverse domains based on image descriptions. In this project, the image, questions, answer, and the rationale from M3CoT are used to investigate the reasoning ability of LLMs when provided with an image description instead of an image. Entries without images were excluded, and image descriptions were generated using LLaVA `llava-v1.6-mistral-7b-hf` from the huggingface library. These descriptions are provided as prompts to the LLMs, alongside the corresponding questions, options, and answers.

In addition, fine-tunig the FLAN-T5 (Chung et al., 2022) model on the M3CoT dataset is conducted to compare its performance with the results from the prompting strategy. For evaluation, the MathVista Dataset (Lu et al., 2024), a consolidated Mathematical reasoning benchmark within visual contexts, is presented to the fine-tuned model. MathVista consists of three newly created datasets, IQTest, FunctionQA, and PaperQA. While this dataset doesn't contain rationales, it allows for assessing the model's ability to interpret visual information through text-based reasoning. Unlike the M3CoT dataset, where mathematical images often include textual elements, MathVista focuses more on domain-specific visual content. This comparison enables an evaluation of the fine-tuned model's understanding of images through textual descriptions. Due to the computational limitation, I utilized a subset of the M3CoT dataset, comprising 1,016 from science, 1,308 from commonsense, and 754 from mathematics domain.

4 Experiment

This project explores the performance of a text-only LLM in a VQA task using image descriptions instead of the original images. The aim is to investigate whether a text-only LLM, when utilizing Chain of Thought (CoT) prompting, can outperform multi-modal LLMs and models trained on specific context, even without additional training.

4.1 Image Description

To obtain descriptions of images from the M3CoT dataset, the prompt were provided to `llava-v1.6-mistral-7b-hf`. Given that scientific and mathematical images often contain num-

The image captures a vibrant scene at a park on a clear day. The sky, a beautiful shade of blue, is dotted with a variety of kites. There are six kites in total, each one unique in its design and color. Starting from the left, the first kite is a striking combination of pink and yellow. It's shaped like a heart, adding a touch of romance to the scene. Next to it, the second kite is a playful mix of green and red, designed to look like a fish. The third kite is a bold statement with its black and red color scheme. It's shaped like a jellyfish, floating gracefully in the sky.

Figure 2: The example of an image description from `llava-v1.6-mistral-7b-hf`.

bers and letters in images, the prompts were specifically designed to emphasize the extraction of these elements from the images. An example of the generated description is provided in Figure 2, and the detailed prompt used is available in Appendix A.

4.2 Baselines

To examine the reasoning ability of text-only LLMs with visual information, we conducted the VQA task using GPT-3.5-turbo. As baselines, zero-shot and few-shot prompts were provided to the model without Chain of Thought strategy. Each prompt consisted of an image description, a question, and multiple options, requiring the model to generate a correct answer based on the information provided. To reduce the randomness in the answers from GPT-3.5-turbo, the temperature was set to 0.01 and the maximum token was 100. Additionally, the same multi-modal LLM, LLaVa, used for image description was also experimented to generate the correct answer when provided with the corresponding question, multiple options, and the original images from M3CoT dataset instead of image descriptions. Details of the prompts used for these baseline models are provided in Appendix A.

4.3 Zero-shot CoT and Few-shot CoT

Zero-shot and few-shot CoT strategies were incorporated into the prompts to enhance the performance of text-only LLM with visual information compared to the standard prompting. For the zero-shot CoT, the phrase " Let's think step by step", as suggested in previous research (Kojima et al., 2022), was added at the end of the prompt. For the few-shot CoT, rationales from M3CoT dataset were included to the examples in the prompt, requiring the model to generate a rationale that supports its answer. Consistent with the baseline prompting, the temperature was set to 0.01, but the maximum token limit was increased to 512 to accommodate rationale generation. The generated rationales are evaluated by BERTScore (Zhang et al., 2020) and ROUGE (Lin, 2004) metrics to measure the similarity between the rationales from the M3CoT dataset and the generated rationales. The prompts used for the CoT strategy are also provided in Appendix A .

4.4 Fine-tuning

Fine-tuning a model on the M3CoT was also conducted to investigate whether learning specific contexts could enable a model to outperform a multi-modal LLM even when the model is trained using only image descriptions instead of original images. The model fine-tuned is FLAN-T5 (Chung et al., 2022), which is the enhanced version of T5 (Roberts et al., 2019). It has shown for its strong performance in generative tasks.

For this study, I fine-tuned the `flan-t5-small` model, which is from HuggingFace library, to generate the correct answer when given the corresponding question, the image description as context, and multiple options. Due to computational limitations, training larger models was not feasible, such as `flan-t5-base` and `flan-t5-large`. The fine-tuning process was performed on the M3CoT dataset with a batch size of 2, a learning rate of 1e-4, and the AdamW optimizer, for a total of 20 epochs. After fine-tuning, the model was evaluated on the MathVista dataset, a comprehensive benchmark for mathematical reasoning (Lu et al., 2024).

5 Results

Table 1 illustrates the accuracy of the results of Gpt-3.5-turbo when given zero-shot and few-shot prompts. Table 2 presents the BERTScores and ROUGE scores of the generated rationales in CoT prompting. Lastly, Table 3 shows the accuracy of fine-tuning the FLAN-T5 model on the M3CoT dataset and evaluation on the MathVista dataset.

1) Few-shot prompting outperforms zero-shot prompting in the VQA task with the text-only LLM.

In the VQA task with the text-only LLM, few-shot prompting yields higher scores in terms of accuracy, BERTScore, and ROUGE compared to zero-

shot prompting. Notably, the results from zero-shot prompting are significantly lower than those achieved by the multi-modal LLM, LLaVA. This results are consistent across different domains.

2) CoT enhances the text-only LLM's performance on the VQA task.

The Chain of Thought (CoT) strategy consistently improves VQA performance across all domains and the number of shots for the text-only LLM. Specifically, in the science domain, the accuracy of the 3-shot model increased from 0.621 to 0.682 when rationales were included in the prompt and were generated with the answer. Additionally, with the CoT strategy, all metrics improved as the number of shots increased.

3) Text-only LLM with CoT can surpass the multi-modal LLM.

With increasing shot numbers in the CoT approach, the 3-shot CoT model achieved the highest performance, with an accuracy of 0.559, compared to LLaVA's 0.511. However, performance varied by domain, despite the overall highest accuracy of the 3-shot CoT model, indicating a need for further error analysis.

4) Fine-tuning LLM can have the comparative result to the multi-model LLM.

The fine-tuning FLAN-T5 shows 0.511 accuracy, which is comparable to the LLaVa (0.511). However, FLAN-T5 has a lower accuracy on the MathVista, which is in a different format and concentrates on mathematical data even though M3CoT also deals with mathematical data. Given the LLaVa was provided with zero-shot prompts without any additional training, LLaVa is more likely to perform better on the unseen and random data than the fine-tuned model.

6 Error Analysis

Error analysis was conducted on the best-performing model, the 3-shot CoT. The errors were categorized into five types: *Wrong reasoning, Wrong description, Not enough description, Lack of social context, and Unclear question*. The examples are detailed in Appendix B.

1) The crucial role of the image description.

The accuracy of the image description plays a crucial role in the model's reasoning and answering abilities. Wrong description often leads to erroneous rationales and answers. Additionally, when the description offers a narrow perspective, such as ignoring details or focusing only on the main item in the image, it prevents the model from the deep reasoning. On the contrast, when the description offers the details, it can also lead to wrong answers, making the model stick to the option that includes the same word from the generated description. The issue from wrong description can be found particularly in biology problems of science domain, where both an accurate description and relevant knowledge are significant for correct reasoning.

2) Challenges in commonsense reasoning.

For commonsense reasoning tasks, even when the description and reasoning are accurate, a lack of understanding of the broader social context can result in incorrect answers. Commonsense questions

Model	COM	SCI	MAT	Total
Zero-shot	0.388	0.28	0.194	0.307
1-shot	0.577	0.64	0.27	0.523
2-shot	0.602	0.634	0.253	0.528
3-shot	0.60	0.621	0.24	0.519
Zero-shot CoT	0.52	0.325	0.156	0.371
1-shot CoT	0.625	0.679	0.273	0.557
2-shot CoT	0.616	0.673	0.273	0.551
3-shot CoT	0.617	**0.682**	0.290	**0.559**
LLaVa	**0.67**	0.45	**0.294**	0.511

Table 1: The accuracy results of prompting experiments on Gpt-3.5-turbo described in section 4. COM, SCI, and MAT indicate commonsense, science, and mathematics, respectively.

Model	B_Score	ROU-1	ROU-2	ROU-L
Zero-shot CoT	0.856	0.304	0.11	0.27
1-shot CoT	0.856	0.306	0.123	0.277
2-shot CoT	0.859	0.313	0.128	0.286
3-shot CoT	**0.862**	**0.324**	**0.134**	**0.296**

Table 2: The results of CoT on the generated rationales. B_Score and ROU indicate BERTScore and ROUGE, respectively.

Model	M3CoT	MathVista
FLAN-T5-small	0.511	0.237

Table 3: The accuracy results of fine-tuned FLAT-T5-small model on the validation set of M3CoT and evaluatoin on MathVista dataset. MathVista dataset focuses on the mathematics domain but requires more comprehensive ability to extract information from an image.

often require the knowledge beyond factual information, demanding the model to grasp underlying social or cultural implications.

3) Struggles with mathematical geometry.

Image descriptions often fail to accurately explain and capture geometrical images, leading to a lot of wrong answers in related mathematical questions. This limitation is particularly pronounced in problems requiring a precise understanding of geometric shapes and spatial relationships.

7 Discussion

This project investigates the capabilities of a text-only LLM in a VQA task across commonsense, science, and mathematical domains when provided with image descriptions instead of actual images. The experiments demonstrate that few-shot Chain of Thought (CoT) prompting can outperform or achieve results comparable to multi-modal LLMs, even without directly processing images. This suggests that the text-only LLM can effectively extract key information from descriptions to answer questions and can conceptualize an image through descriptive text.

However, this project have several limitations. First, only a subset of M3CoT dataset with a relatively small number of samples was used. This is due to the fact that the M3CoT is significantly imbalanced towards commonsense domain. Thus, I extracted a part of data from commonsense domain and the entire data from mathematics domain. The experiment on a larger dataset could yield more robust results.. Second, the experiments were limited to GPT-3.5-turbo, LLaVA, and FLAN-T5. Although there are many other text-only and multi-modal LLMs, I couldn't explore any other models when considering the time for model execution and result analysis. Addressing this limitation in future work will involve testing a broader range of models. Lastly, during the fine-tuning of FLAN-T5, there was limited exploration of hyper parameters. While some variations of learning rate, maximum length, and repetition penalty were tested, there remains considerable opportunity for further tuning to achieve optimal performance.

In the next projects and future research, I will address these types of limitations to make research more comprehensive and convincing.

References

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. 2022. Flamingo: a visual language model for few-shot learning.

Qiguang Chen, Libo Qin, Jin Zhang, Zhi Chen, Xiao Xu, and Wanxiang Che. 2024. M^3cot: A novel benchmark for multi-domain multi-step multi-modal chain-of-thought. In *Proc. of ACL*.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, march 2023. *URL https://lmsys. org/blog/2023-03-30-vicuna*, 3(5).

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. 2022. Scaling instruction-finetuned language models.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Hyounghun Kim and Mohit Bansal. 2019. Improving visual question answering by referring to generated paragraph captions. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3606–3612, Florence, Italy. Association for Computational Linguistics.

Takeshi Kojima, Shixiang (Shane) Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. In *Advances in Neural Information Processing Systems*, volume 35, pages 22199–22213. Curran Associates, Inc.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. 2024. Llava-next: Improved reasoning, ocr, and world knowledge.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. In *Advances in Neural Information Processing Systems*, volume 36, pages 34892–34916. Curran Associates, Inc.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. 2024. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts.

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. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.

Adam Roberts, Colin Raffel, Katherine Lee, Michael Matena, Noam Shazeer, Peter J. Liu, Sharan Narang, Wei Li, and Yanqi Zhou. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. Technical report, Google.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837. Curran Associates, Inc.

Yiqi Wu, Xiaodan Hu, Ziming Fu, Siling Zhou, and Jiangong Li. 2024. Gpt-4o: Visual perception performance of multimodal large language models in piglet activity understanding.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert.

Liu Zhuang, Lin Wayne, Shi Ya, and Zhao Jun. 2021. A robustly optimized BERT pre-training approach with post-training. In *Proceedings of the 20th Chinese National Conference on Computational Linguistics*, pages 1218–1227, Huhhot, China. Chinese Information Processing Society of China.

A Prompts

A.1 Image Description

A.2 Standard Prompting

```
<text prompt>
You are provided with a question and an image, which is the
context of the question. Your task is to give the correct
answer to the following question by selecting one of the
given options.
Follow the specific JSON format provided below:
{"answer": <your answer>}
Be sure to fill "<your answer>" with the one of the given
options and the correct answer.Utilize the relevant
information from the image and avoid any unrelated
explanations.
Question: <question>
Options: <multiple choices>
```

Figure 3: The prompt used for generating an image description. This is provided with an image to LLaVa.

```
You are provided with a question and an image description, which
is the context of the question.Your task is to give the correct
answer to the following question by selecting one of the given
options. Follow the specific JSON format provided below:
{"answer": <your answer>}
Be sure to fill "<your answer>" with the one of the given correct
answer options. Utilize the relevant information from the
description and avoid any unrelated explanations.
Question: <question>
Context : <image description>
Options : <multiple choices> e.g (A) 3 (B) 4 (C) 5 (D) 6
```

```
You are provided with a question and an image description,
which is the context of the question.Your task is to give the
correct answer to the following question by selecting one of
the given options. Follow the specific JSON format provided
below:
{"answer": <your answer>}
Be sure to fill "<your answer>" with the one of the given
correct answer options. Utilize the relevant information from
the description and avoid any unrelated explanations.
Question: <question_1>
Context : <image description_1>
Options : <multiple choices_1> e.g (A) 3 (B) 4 (C) 5 (D) 6
Answer :  <correct answer_1>
  .
  .
  .
Question: <question_n>
Context : <image description_n>
Options : <multiple choices_n>
Answer :  <correct answer_n>

Question: <question>
Context : <image description>
Options : <multiple choices>
```

Figure 4: The standard zero-shot and few-shot prompts to generate the correct answer given a question, context, which is the image description, and multiple options.

A.3 CoT Prompting

```
You are provided with a question and an image description, which is
the context of the question. Your task is to give the correct
answer to the questions by selecting one of the given options and
explain why the answer is correct by analyzing the given image
description and applying relevant knowledge.Ensure your explanation
stays concise and to the point, and the answer should be one of the
given options. Follow the specific JSON format provided below:
{"answer": "<your answer>",
"rationale": "<your rationale>"}
Fill "<your answer>" and "<your rationale>" with one of the given
options, the answer and the clear rationale supporting the
correctness of the answer. Utilize the relevant information from
the description and avoid any unrelated explanations.

Question: <question>
Context : <image description>
Options : <multiple choices> e.g (A) 3 (B) 4 (C) 5 (D) 6

Let's think step by step.
```

```
You are provided with a question and an image description, which
is the context of the question. Your task is to give the correct
answer to the questions by selecting one of the given options and
explain why the answer is correct by analyzing the given image
description and applying relevant knowledge.Ensure your
explanation stays concise and to the point, and the answer should
be one of the given options. Follow the specific JSON format
provided below:
{"answer": "<your answer>",
"rationale": "<your rationale>"}
Fill "<your answer>" and "<your rationale>" with one of the given
options, the answer and the clear rationale supporting the
correctness of the answer. Utilize the relevant information from
the description and avoid any unrelated explanations.
Question: <question_1>
Context : <image description_1>
Options : <multiple choices_1> e.g (A) 3 (B) 4 (C) 5 (D) 6
Answer :  <correct answer_1>
Rationale : <Rationale_1>
  .
  .
  .
Question: <question_n>
Context : <image description_n>
Options : <multiple choices_n>
Answer :  <correct answer_n>
Rationale : <Rationale_2>

Question: <question>
Context : <image description>
Options : <multiple choices>
```

Figure 5: The zero-shot and few-shot CoT prompts to generate the correct answer along with a rationale given a question, context, which is the image description, and multiple options.

B Examples of some error types

B.1 Wrong description

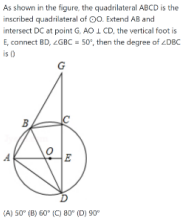

Image description: The image displays a geometric figure, which appears to be a quadrilateral ABCD with a right-angled triangle OAB inscribed within it. The quadrilateral is outlined in black, and the right-angled triangle is filled in with a solid color. The vertices of the quadrilateral are labeled with the letters A, B, C, and D, and the vertices of the inscribed triangle are labeled with the letters O and E.\n\nThe right-angled triangle OAB
Options: (A) the answer is option A in the image (B) the correct choice is option D in the image (C) All given choices given in the text are wrong (D) option C in the image is the answer
The generated rationale: The question does not provide any specific details or criteria to determine the correct answer among the given options. Since the context only describes the geometric figure without any additional information or problem statement, it is not possible to ascertain the correct option. Therefore, the correct choice is that all given choices in the text are wrong.
The generated answer: C
The correct answer:D

B.2 Lack of social context

Image description: The image depicts a serene beach scene. Dominating the center of the image is a black folding chair, which is currently unfolded. The chair is positioned on the left side of the image, facing towards the right.\n\nAbove the chair, there's a beach umbrella that adds a pop of color to the scene. The umbrella is adorned with a floral pattern in shades of pink.
Options: (A) A group of friends (B)The photographer (C) A random passerby (D) Unknown
The generated rationale: Based on the context provided, the most likely individual to set up the floral umbrella in the sand would be a random passerby. The presence of the unfolded black folding chair suggests that someone took the time to set up the beach umbrella, and it is common for beachgoers or passersby to set up such umbrellas for shade and relaxation. Therefore, option (C) a random passerby is the correct choice.
The generated answer: C
The correct answer:B

B.3 Not enough description

Image description: The image shows a vintage television set placed on a wooden surface, possibly a table or a stand. The television has a boxy shape with a curved front panel, typical of older CRT (cathode ray tube) televisions. The screen is turned off, and the reflection of the room can be seen on the glass surface in front of the television.\n\nThe television is positioned in front of a glass display case that contains various items, including what appears to be clothing and possibly.
Options: (A) A movie theater (B) A grocery store (C) An electronics store (D) A restaurant
The generated rationale: The image description indicates the presence of a vintage television set, which suggests an electronics-related environment. The presence of a glass display case with various items further supports the likelihood of it being an electronics store where electronic items are displayed for sale. Therefore, the most likely location where this picture was taken is an electronics store, making option (C) the correct answer.
The generated answer: C
The correct answer:B