Towards the Adversarial Robustness of Vision-Language Model with Chain-of-Thought Reasoning

Zefeng Wang*
Technical University of Munich
zefeng.wang@tum.de

Zhen Han*
LMU Munich
hanzhen02111@163.com

Shuo Chen
LMU Munich
chenshuo.cs@outlook.com

Volker Tresp
LMU Munich
tresp@dbs.ifi.lmu.de

Jindong Gu†
University of Oxford
jindong.gu@outlook.com

Abstract

Recent research has indicated that the performance of vision language models on intricate reasoning tasks is significantly enhanced by applying chain-of-thought (CoT) reasoning. However, few studies investigate the impact of the reasoning chain on the model's robustness. In this work, we develop three attack methods to investigate the robustness of vision-language models with chain-of-thought reasoning. In particular, we propose a novel attack method using the stopping reasoning strategy to study the robustness of vision-language models with CoT. Experiments on SicenceQA dataset show that 1) The rationale generated by vision-language models with CoT is more susceptible to attacks compared to the answer part. 2) Vision-language models with CoT are most vulnerable to the stop reasoning attack compared to other attacks.

1. Introduction

Recent studies [16, 4] showed that large language models' (LLMs) performance on complex reasoning tasks gets significantly boosted by asking the models to generate a rationale before producing the answer. A rationale is a series of intermediate reasoning steps that decompose a complex task into several simple subtasks. The rationale and its corresponding answer together are referred to as a reasoning chain, and the technique of asking LLMs to generate a reasoning chain is called chain-of-thought (CoT) prompting [16]. Zhang *et al.* [19] extended this technique into vision-language models, called multimodal CoT, and demonstrated its effectiveness on complex reasoning tasks with visual features. The multimodal CoT takes both textual and vision information of a question as input and generates a textual rationale as the intermediate output, then takes the intermediate output and the question again as input to produce the answer.

However, the robustness of vision-language models with chain-of-thought reasoning remains unclear. In this paper, we apply three attack methods to investigate the robustness of vision-language models with CoT. First, we compare the model's performance under two attacking scenarios, i.e., attacking the model based on 1) the answers and 2) the intermediate rationales. These two scenarios use a shared attack pipeline, which leverages the projected gradient descent (PGD) method[1] [12]. They differ from the designed objective functions. For attacking the models relying on their answers, we construct an objective function between the model's answers and ground truth labels. By attacking the models based on rationales, we build an objective function using KL divergence between the rationale generated from the original image and the rationale generated from the perturbed images. This KL divergence will be used to perturb the image for the next attack iteration.

Moreover, we study the robustness of vision-language models with CoT to a novel attack method called *stop reasoning*, where the input images are perturbed to induce the model to generate $[EOS]$ token earlier, where $[EOS]$ token is used to signify the end of the output sequence. The idea here is to let CoT models stop their reasoning process, and thus, predict wrong answers. Specifically, we build a cross-entropy loss function between the rationale and the $[EOS]$ token. By reducing the loss, the tendency to generate the $[EOS]$ token is increased. And with an early occurrence of the $[EOS]$, the model will stop the reasoning process there.

We run experiments on the ScienceQA [11] dataset, a visual question-answering dataset on scientific domains.

*Equal Contribution.
†Corresponding author

[1] PGD perturbs the input image based on the gradient of the objective function in an iterative manner. In this work, PGD with l_{inf} norm is used.

The results demonstrate that **1)** The rationale generated by vision-language models with CoT is susceptible to attacks. We found that attacking the model's rationales causes a significantly larger performance drop compared to only attacking the answer part. **2)** In comparison to answer attack and rationale attack, the multi-modal CoT models are most vulnerable to our proposed *stop reasoning* attack åmethod.

2. Related Work and Preliminaries

2.1. Chain-of-Thought Prompting

Chain-of-thought prompting significantly improve the performance of large language models on complex tasks by using a series of intermediate reasoning steps [16]. In general, chain-of-thought prompting can be divided into two main categories according to the prompts: Zero-Shot-CoT [7] and Few-Shot-CoT [16]. Besides labeling CoT manually, Zhang *et al*. [18] introduced Auto-CoT which automatically constructs demonstrations, Wang *et al*. [15] and Zelikman *et al*. [17] proposed models to generate a rationale for other training tasks with few human-written seed rationales. Fu *et al*. [3] showed that, with more steps, the model achieves substantially better performance on multi-step reasoning tasks. Khot *et al*. [6] and Zhou *et al*. [20] proposed models to solve complex tasks by decomposing them into simpler sub-tasks. Zhang *et al*. [19] proposed MM-CoT which solves tasks with vision features. Besides greedy thinking, Wang *et al*. [14] and Li *et al*. [10] proposed models think in several paths and predict based on them.

2.2. Multimodal Large Language Model

The vision-language interaction methods can be divided into two types: fusion encoder and dual encoder [2].In the fusion encoder category, OSCAR [9] contacted text, tags, and image features and feed them into the encoder with a single-stream architecture. ALBEF [8] adopted a cross-attention mechanism to vision and text features after two separate transformers, called dual-stream architecture. Besides, FLAVA [13] first adopted a dual encoder to obtain single-modal representations. Then the single-modal embeddings are sent to a fusion encoder to obtain cross-modal representation.

2.3. Preliminary of Multimodal Chain-of-Thought

The MM-CoT model proposed by Zhang *et al*. [19] is a two-stage model, adapting UnifiedQABase [5] as the backbone language model and applying DETR [1] as the vision feature extractor. In the first stage, it takes the text and visual information of the question as input and outputs a rationale. In the second stage, the input consists of the original visual information and a new text concatenation, which combines the original text and the generated rationale and outputs a sentence containing the choice in the form "*The answer is (choice).*".

3. Methodology

We wonder, how is the vulnerability of the multimodal models with chain-of-thought reasoning to different attacks. To answer the question, we develop three distinct attack strategies. Section 3.1 introduces the general attack pipeline, which is shared in all three attack methods. In Section 3.2, the three attack methods including an efficient novel attack method named *stop-reasoning attack* are presented.

3.1. Attack Pipeline

The objective of the attacks is to diminish the predictive accuracy of the model. It is assumed that all model details, encompassing architecture, parameters, and outputs, are known when perturbing the input images to attack the model. For a more robust and compelling comparison, all attacks employ the same algorithm, ensuring minimal divergence among them.

As previously mentioned, the PGD technique is employed in all attacks, enabling the modification of pixel values in input images through specifically designed attack strategies. To prevent zero loss, the images are initially perturbed before the first attack iteration. The shared pipeline is shown in Algorithm 1 in Appendix, and the primary distinction among the attacks lies in the distinct loss functions tailored to their specific objectives.

3.2. Attack Methods

This section presents the definitions and particulars of three attacks: the *baseline attack*, the *rationale attack*, and the *stop-reasoning attack*. The prediction of the multimodal CoT model can be divided into two parts: the rationale and the answer (as illustrated in Figure 1). Throughout the remainder of the paper, **rationale** refers to the rationale component of the prediction, while **answer** refers to the answer component.

Definition 3.1. *Baseline attack* is an attack, which attacks the model via perturbing input images with increasing the cross-entropy loss between the answer and the labels.

Definition 3.2. *Rationale attack* is an attack, which attacks the model via perturbing input images with increasing the KL divergence between rationale generated from the original images and rationale generated from the perturbed images.

Definition 3.3. *Stop-reasoning attack* is an attack, which attacks the model via perturbing input images with decreasing the cross-entropy loss between rationale and a sequence of $[EOS]$ tokens.

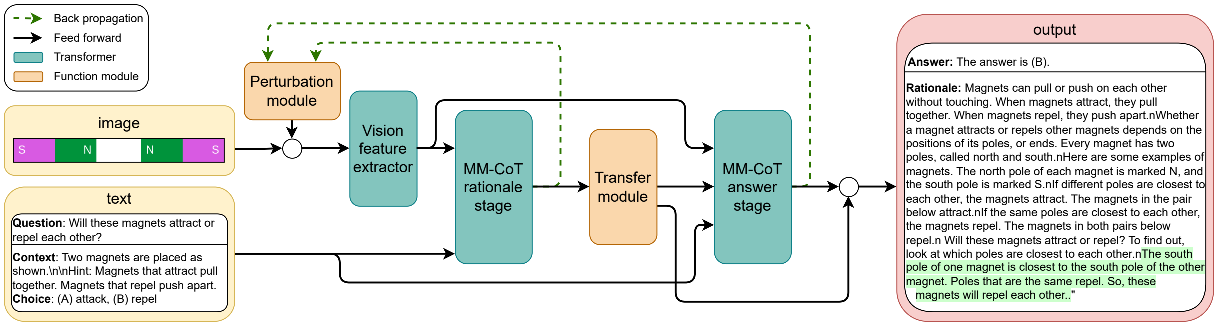


Figure 1: Joint MM-CoT model. Initially, we introduce a perturbation to the images. Subsequently, the perturbed images and original text are fed into the Joint-MM-CoT model to generate the rationale and the answer. The rationale and the answer are used to perturb the images for the next attack iteration. To ensure differentiability throughout the two stages, a transfer module is implemented, which directly incorporates the generated rationale logits in the answer stage, bypassing the decode-reencode process.

3.2.1 Baseline Attack

The first attack is the *baseline attack*. A trivial model predicts output directly from the input. The attack based on the output mirrors a general vulnerability of the multimodal CoT model. In this attack scenario, the cross-entropy loss function between the answers and the ground truth choices is used. As the loss increases, the model becomes more susceptible to selecting an alternative choice instead of the correct one. To amplify this tendency, we mitigate the interference from other tokens in the predictions by constructing a loss function that exclusively focuses on the choice, disregarding other sentence components. The loss function is shown as

$$loss_{choice} = -\log \frac{\exp(x_{answer})}{\sum_{v=1}^{V} \exp(x_v)} \tag{1}$$

where x_{answer} is the predicted logits of the labels, x_v is the logits of the token v and V is the vocabulary size used in the tokenizer.

3.2.2 Rationale Attack

To explore the influence of the CoT process on models' robustness, an attack based on the rationale is designed. This approach is to manipulate the information the rationale conveys. Since the information is implicit within the logits distribution of the generated rationale, we employ the Kullback-Leibler (KL) divergence between two generated rationales as a measure of information disparity (relative entropy). The rationale generated from the original images is established as the baseline. The relative entropy is computed between this baseline rationale and the rationale obtained from perturbed images. As the relative entropy increases, the conveyed information undergoes alteration. The loss function is shown as follows.

$$loss_{KLDiv} = \frac{1}{T} \sum_{t=1}^{T} \sum_{v=1}^{V} y_{label} \cdot (\log y_{label} - \log y_{pred}) \tag{2}$$

where T is the tokens count and V is the vocabulary size, y_{label} and y_{pred} indicate probability of corresponding logits.

3.2.3 Stop-Reasoning Attack

We introduce a novel attack method called the *stop-reasoning attack*. Multimodal CoT models leverage the information generated in intermediate stages and perform inference step by step. The *stop-reasoning attack* aims to halt this process by minimizing the information provided in the intermediate stages and to cease reasoning by reducing the length of the rationale. To accomplish this, we construct a cross-entropy loss function between the rationale logits and a sequence of $[EOS]$ tokens. As the loss decreases, the model becomes more inclined to predict the token $[EOS]$ and stops reasoning. Since the PGD attack technique increases the loss, we use the inverse cross-entropy loss as our loss.

$$loss_{length} = -\frac{1}{T} \sum_{t=1}^{T} \log \frac{\exp(x_{[EOS]})}{\sum_{v=1}^{V} \exp(x_v)} \tag{3}$$

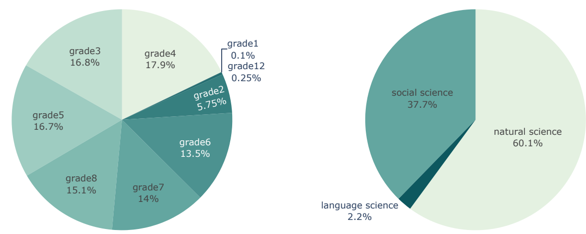
where $x_{[EOS]}$ is the logits of the $[EOS]$ token.

4. Experiments

Dataset The ScienceQA dataset [11] is used to conduct the attacks, and it is the same dataset used to train the MM-CoT model. It is a comprehensive question-answering

	Settings	Subject			Grade				Total
		NAT	SOC	LANG	1-3	4-6	7-9	10-12	
Baseline	Original input	65.97	71.62	59.09	73.07	68.85	62.54	60.00	67.95
Attacks	Baseline attack	34.78	54.64	11.36	43.05	45.10	35.57	0.0	41.75
	Rationale attack	34.36	50.13	15.91	37.09	44.27	35.22	0.0	39.90
	Stop-reasoning attack	14.39	41.51	15.91	22.52	27.40	21.99	0.0	24.65

Table 1: The table shows the accuracy of various categories and the entire test dataset. The columns in the subgroup ”subject” display the respective accuracy values when the test set is categorized into three subjects: natural science (NAT) with 1202 samples, social science (SOL) with 754 samples, and language science (LANG) with 44 samples. The columns in the subgroup ”grade” indicate the respective accuracy values when the test set is categorized into four grade subcategories: grade 1 to 3 with 453 samples, grade 4 to 6 with 960 samples, grade 7 to 9 with 582 samples, and grade 10 to 12 with 5 samples.



(a) Grade distribution (b) Subject distribution

Figure 2: Distribution of test samples

dataset meticulously tailored for scientific domains, providing extensive answer annotations. The questions encompass various domains, spanning three subjects, 26 topics, 127 categories, and 379 skills. We utilize 2000 image-containing samples from the test split to conduct the attacks.

Attack Settings and Metric In all attacks, we employ the PGD attack algorithm with varying loss functions, while maintaining consistent hyper-parameters: maximum change of 0.0627, attack step of 0.0005, and a maximum of 200 attack iterations. The answer accuracy is utilized as the metric.

Results Table 1 presents the primary results, demonstrating that the Joint-MM-CoT model achieves an overall accuracy of 67.95% on the 2000 test samples using the original inputs. This accuracy value is considered the baseline and is compared with the accuracy values obtained under various attack scenarios. The *baseline attack* results in a 26.20% accuracy drop (from 67.95% to 41.75%). The most substantial drop occurs in questions from the language science subject, plummeting by 26.20% (from 59.09% to 11.36%). The *rationale attack* exhibits a total accuracy drop of 28.05% (from 67.95% to 39.60%), which is better than *baseline attack*. The decrease in accuracy is primarily attributed to questions in grades 1-3. The *stop-reasoning attack* demonstrates the most effective attack performance, resulting in a significant 43.3% accuracy drop (from 67.95% to 24.65%). Compared to the *baseline attack*, the *stop-reasoning attack* performs well in nearly all categories except for the language science subject.

Analysis The vulnerability of the multimodal CoT model is significantly higher when subjected to attacking the rationales than attacking the answers. Table 1 reveals that *rationale attack* causes a reduction of 1.85% in the overall prediction accuracy compared to *baseline attack* (from 41.75% to 39.90%), primarily impacting questions from lower grades. Furthermore, *stop-reasoning attack* results in a reduction of 17.1% in comparison to *baseline attack* (from 41.75% to 24.65%), which shows that the vision-language models with CoT are most vulnerable to the stop reasoning attack compared to other attacks. *Stop-reasoning attack* exhibits a substantial increase in attack performance compared to *baseline attack*, with the primary impact observed in questions from the natural science subject. We selected 20 samples and identified two common characteristics: all questions have extensive long rationales, and all rationales contain crucial information towards the end. As *stop-reasoning attack* aims to elevate the likelihood of all rationale tokens being $[EOS]$, longer rationales have a higher probability of encountering an early occurrence of $[EOS]$. Consequently, the critical information at the end of the rationale becomes lost. Without this crucial information, the advantage conferred by the CoT method is nullified, disrupting the inference pipeline.

5. Conclusion

In this study, we examine the robustness of vision-language models with chain-of-though reasoning. An efficient and novel attack method using *stop-reasoning* strategy is proposed. Our findings reveal that the model exhibits greater vulnerability when we attack the generated rationale and make the model stop reasoning.

References

[1] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020.

[2] Yifan Du, Zikang Liu, Junyi Li, and Wayne Xin Zhao. A survey of vision-language pre-trained models. *arXiv preprint arXiv:2202.10936*, 2022.

[3] Yao Fu, Hao Peng, Ashish Sabharwal, Peter Clark, and Tushar Khot. Complexity-Based Prompting for Multi-Step Reasoning, Jan. 2023. arXiv:2210.00720 [cs].

[4] Jie Huang and Kevin Chen-Chuan Chang. Towards reasoning in large language models: A survey. *arXiv preprint arXiv:2212.10403*, 2022.

[5] Daniel Khashabi, Sewon Min, Tushar Khot, Ashish Sabharwal, Oyvind Tafjord, Peter Clark, and Hannaneh Hajishirzi. Unifiedqa: Crossing format boundaries with a single qa system. *arXiv preprint arXiv:2005.00700*, 2020.

[6] Tushar Khot, Harsh Trivedi, Matthew Finlayson, Yao Fu, Kyle Richardson, Peter Clark, and Ashish Sabharwal. Decomposed Prompting: A Modular Approach for Solving Complex Tasks, Apr. 2023. arXiv:2210.02406 [cs].

[7] Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large Language Models are Zero-Shot Reasoners, Jan. 2023. arXiv:2205.11916 [cs].

[8] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34:9694–9705, 2021.

[9] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXX 16*, pages 121–137. Springer, 2020.

[10] Yifei Li, Zeqi Lin, Shizhuo Zhang, Qiang Fu, Bei Chen, Jian-Guang Lou, and Weizhu Chen. Making Large Language Models Better Reasoners with Step-Aware Verifier, May 2023. arXiv:2206.02336 [cs].

[11] 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, Oct. 2022. arXiv:2209.09513 [cs].

[12] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

[13] Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, and Douwe Kiela. Flava: A foundational language and vision alignment model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15638–15650, 2022.

[14] 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, Mar. 2023. arXiv:2203.11171 [cs].

[15] Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, and Denny Zhou. Rationale-Augmented Ensembles in Language Models, July 2022. arXiv:2207.00747 [cs].

[16] 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, Jan. 2023. arXiv:2201.11903 [cs].

[17] Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah D. Goodman. STaR: Bootstrapping Reasoning With Reasoning, May 2022. arXiv:2203.14465 [cs].

[18] Zhuosheng Zhang, Aston Zhang, Mu Li, and Alex Smola. Automatic Chain of Thought Prompting in Large Language Models, Oct. 2022. arXiv:2210.03493 [cs].

[19] Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. Multimodal Chain-of-Thought Reasoning in Language Models, Feb. 2023. arXiv:2302.00923 [cs].

[20] Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc Le, and Ed Chi. Least-to-Most Prompting Enables Complex Reasoning in Large Language Models, Apr. 2023. arXiv:2205.10625 [cs].

Appendix

Algorithm 1 Attacks with PGD technique

Require: Input text x_t, Input image x_i, Target label y, Model M, Loss function L, Perturbation budget ϵ, Step size α, Number of iterations T, PGD perturbation module F

Ensure: Model output o

```
Initialize x'_i ← x_i
Add a random initial perturbation x'_i ← x'_i + δ
for t = 1 to T do
    Model output o ← M(x'_i, x_t)
    if o is different from y then
        break
    else
        PGD with l_inf norm x'_i ← F(x'_i, y, L, ε, α)
    end if
end for
return Model output o
```

A. End-to-End Differentiable MM-CoT

We utilize the MM-CoT model as the foundational framework of our study. MM-CoT comprises two distinct stages that operate independently and lacks differentiability

Setting	Successful	Unsuccessful
Baseline attack	186.21	146.90
Rationale attack	162.18	161.97
Stop-reasoning attack	187.02	118.21
Joint attack	168.14	157.62

Table 2: Average length of baseline predictions. The table presents the average length of baseline predictions categorized into two groups. The successfully attacked questions are classified under the *successful* category, while the failed attacked questions are classified under the *unsuccessful* category. The length refers to the token count.

in between. To enable backpropagation, we merge the two stages into a unified model called Joint-MM-CoT. In the first stage, the MM-CoT model uses image and text input to generate rationale. In the second stage, the raw text input and the generated rationale are concatenated. The MM-CoT model uses the image and concatenated text to infer an answer. To achieve differentiable concatenation of the stages, we extract the logits directly from the first stage and apply a sequence of transformations to utilize them in the second stage. A transfer module that converts logits into a one-hot matrix and applies filtering is implemented. The one-hot matrix is constructed based on the position of the largest logit. And the one-hot matrix is cleaned up by removing the lines containing the special tokens. Subsequently, the filtered one-hot matrix is passed through the embedding layer. Then the embeddings from the rationale are combined with the text embeddings derived from the raw input. These combined embeddings serve as the text embeddings for the second stage. Regarding extracting visual features, the pre-trained model DETR is employed, more specifically, *detr_resnet101_dc5*. Furthermore, the inference process is transitioned from generating predictions for the entire sequence in a single run to a token-by-token generation approach.

B. Average Length of Baseline Predictions

Table 2 displays the average length of the baseline predictions. Under the *baseline attack*, the successfully attacked questions are longer than the unsuccessfully attacked questions (186.21 tokens vs 146.90 tokens). But the disparity is more pronounced under the *stop-reasoning attack* (187.02 tokens vs 118.21 tokens). However, in the case of the *rationale attack*, the difference in length is minimal (162.18 tokens vs 161.97 tokens).

C. Joint Attack

Since all the previously mentioned methods are solely based on a single stage, we hypothesize that the model is susceptible to attacks that exploit multiple stages. A further attack is launched, which perturbs images by gradually increasing the relative entropy, thereby impeding the model from making accurate predictions as a collective effort. In this scenario, the loss function is a combination of *baseline attack* and *rationale attack*.

$$loss_{joint} = loss_{choice} + loss_{KLDiv} \tag{4}$$

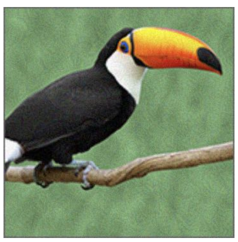
Table 3 illustrates that the attack performance of *joint attack* closely resembles that of *rationale attack* across all categories. A marginal accuracy difference of 0.6% (from 39.9% to 39.3%) exists between the two, suggesting that attacks based on multi-stages do not significantly augment the model's vulnerability.

D. Examples

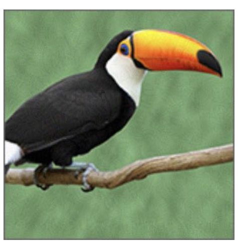
The following figures show some examples. The blocks with the orange background are the common text inputs and related answers. The blocks in the range with the red background are the inferences with the corresponding input images. The blocks in the range with the red background from top to bottom are under the settings "without attack", the "*baseline attack*", the "*rationale attack*", and the "*stop-reasoning attack*".

Question: Select the bird below.
Context: Birds have feathers, two wings, and a beak. A toco toucan is an example of a bird.
Choices: A. "great white shark", B. "white stork"
Answer: B. "white stork"

Answer: The answer is (B).
Rationale: Birds, mammals, fish, reptiles, and amphibians are groups of animals. The animals in each group have traits in common.nScientists sort animals into groups based on traits they have in common. This process is called classification. A white stork is a bird. It has feathers, two wings, and a beak.nStorks live near water. Storks eat fish, fish, reptiles, and amphibians.nA great white shark is a fish. It lives underwater. It has fins, not limbs.nGreat white sharks live near coral reefs. They have fins, not limbs.nA great white shark is a fish. It lives underwater. It has fins, not limbs.nGreat white sharks can live in shallow water..



Answer: The answer is (B).
Rationale: Birds, mammals, fish, reptiles, and amphibians are groups of animals. The animals in each group have traits in common.nScientists sort animals into groups based on traits they have in common. This process is called classification. A white stork is a bird. It has feathers, two wings, and a beak.nStorks live near water. Storks eat fish, insects, and fruit.nA great white shark is a fish. It lives underwater. It has fins, not limbs.nGreat white sharks live near water or in damp places. Storks live underwater. Storks live in damp places. They use their tails to swim underwater..



Answer: The answer is (A).
Rationale: Birds, mammals, fish, reptiles, and amphibians are groups of animals. The animals in each group have traits in common.nScientists sort animals into groups based on traits they have in common. This process is called classification. A white stork is a mammal. It has fur and feeds its young milk.nStorks do not have teeth! They swallow their food whole.nA great white shark is a fish. It lives underwater. It has fins, not limbs.nGreat white sharks live near coral reefs. They use their fins to drill into coral reefs.nA great white shark is a reptile. It has scaly, waterproof skin.nA great white shark is a reptile. It has scaly, waterproof skin..

Answer: The answer is (A).
Rationale: a stork is a fish. It has scaly, waterproof skin.nA stag ape is a stag. The stag ape is a stag. The ape shrew is a stag..

Figure 3: Example 1. Question id: 6144. Grade: grade 3. Subject: natural science. The blocks from top to bottom are common text input, image input and inference without attack, image and inference with *baseline attack*, image and inference with *rationale attack*, and image and inference with *stop-reasoning attack*.

Question: A baby blue whale is as big as ().
Context: Read the first part of the passage about blue whales.\nA blue whale is big. It is the biggest animal in the world.\nWhen a blue whale is a new baby, it is as big as a car. When it is older, it is even bigger. It is as long as two school buses put together.\nComplete the sentence.
Choices: A. 'an adult whale', B. 'a car', C. 'two buses'
Answer: B. 'a car'

Answer: The answer is (B).
Rationale: The passage tells you that a blue whale is the biggest animal in the world..

Answer: The answer is (A).
Rationale: The passage says that an adult whale is as big as a car..

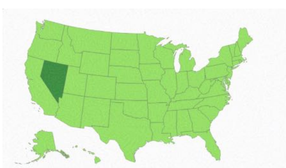
Answer: The answer is (A).
Rationale: The passage says that a blue whale is big. It is the biggest animal in the world.nWhen a blue whale is a new baby, it is as big as a car. When it is older, it is even bigger. It is as long as two school buses put together.nThe passage says that an adult whale is as big as a car. When a blue whale is a new baby, it is as big as a car. When it is older, it is even bigger. It is as long as two school buses put together..

Answer: The answer is (B).
Rationale:The passage says that a blue whale is the biggest animal in the world..

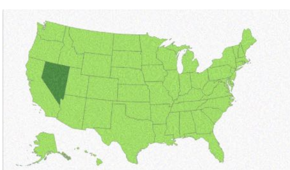
Figure 4: Example 2. Question id: 18387. Grade: grade 1. Subject: language science. The blocks from top to bottom are common text input, image input and inference without attack, image and inference with *baseline attack*, image and inference with *rationale attack*, and image and inference with *stop-reasoning attack*.

Question: What is the capital of Nevada?
Context:
Choices: A.'Carson City', B. 'Reno', C. 'Salem', D. 'Las Vegas'
Answer: A.'Carson City'

Answer: The answer is (A).
Rationale: Carson City is the capital of Nevada..



Answer: The answer is (A).
Rationale: Carson City is the capital of Nevada..



Answer: The answer is (A).
Rationale: Carson City is the capital of Nevada..

Answer: The answer is (D).
Rationale:

Figure 5: Example 3. Question id: 14219. Grade: grade 4. Subject: social science. The blocks from top to bottom are common text input, image input and inference without attack, image and inference with *baseline attack*, image and inference with *rationale attack*, and image and inference with *stop-reasoning attack*.

	Settings	Subject			Grade				Total
		NAT	SOC	LANG	1-3	4-6	7-9	10-12	
Baseline	Original input	65.97	71.62	59.09	73.07	68.85	62.54	60.00	67.95
Attacks	Baseline attack	34.78	54.64	11.36	43.05	45.10	35.57	0.0	41.75
	Rationale attack	34.36	50.13	15.91	37.09	44.27	35.22	0.0	39.90
	Stop-reasoning attack	14.39	41.51	15.91	22.52	27.40	21.99	0.0	24.65
Ablation study	Joint attack	33.86	49.34	15.91	36.64	43.02	35.57	0.0	39.30

Table 3: Results with joint attack. The table presents accuracy values in percentages for various categories and the entire test dataset. The columns in the subgroup "subject" display the respective accuracy values when the test set is categorized into three subjects: natural science (NAT) with 1202 samples, social science (SOL) with 754 samples, and language science (LANG) with 44 samples. The columns in the subgroup "grade" indicate the respective accuracy values when the test set is categorized into four grade subcategories: grade 1 to 3 with 453 samples, grade 4 to 6 with 960 samples, grade 7 to 9 with 582 samples, and grade 10 to 12 with 5 samples.