Corvid: Improving Multimodal Large Language Models Towards Chain-of-Thought Reasoning

Supplementary Material

A. Supplementary Training Details

In Tab. 10, we provide detailed hyperparameter settings and time costs for the three-stage training of Corvid. Unless otherwise specified, all training and inference are conducted using 8 NVIDIA A800 (80G) GPUs by default.

Configuration	Stage 1	Stage 2	Stage 3
Batch size	256	256	128
Peak learning rate	1e-3	2e-5	2e-6
Learning rate schedule	Cosine	Cosine	Cosine
Learning rate warm-up ratio	0.03	0.03	0.03
Weight decay	0	0	0
Epoch	1	1	3
Optimizer	AdamW	AdamW	AdamW
Float precision	bfloat16	bfloat16	bfloat16
Deepspeed configuration	zero2	zero3	zero3
Training modules	GateMixer	GateMixer, LLM	LLM
Data Size	1M	1M	320K
Training hours	$\sim$11	$\sim$28	$\sim$20

Table 10. **Training hyperparameter setting.**

B. Additional Experiment Results

B.1. Comparison with o1-Like MLLMs

In **??**, we compare our models against o1-like MLLMs on various benchmarks, including MMStar [2], MMB [12], MMVet [26], MathVista (MathV) [15], AI2D [9], and Hallusion [7], using their benchmark metrics computed with official implementations. Here, Corvid-o1†, LLaVA-o1 [23], and LlamaV-o1 [20] utilize the same baseline Llama-3.2-11B-Vision-Instruct [17]. Results in the table showcase that Corvid-o1-8B surpasses existing o1-like MLLMs on multiple benchmarks, particularly outperforming llamaV-o1 and Mulberry-o1-7B [24] on MathVista by 10.5 and 14.5 points, respectively. Additionally, Corvid-o1 achieves the best overall performance across all benchmarks. These results highlight the effectiveness of Corvid-o1, establishing it as a competitive MLLM that exceeds existing o1-like MLLMs with similar parameter sizes.

B.2. Additional Evaluation on VRC-Bench

We additionally evaluate our model on VRC-Bench [20], which is specifically designed for multimodal step-by-step reasoning tasks. The results in Tab. 11 show that Corvid-o1 achieves leading accuracy in final answers but exhibits limited performance on reasoning steps. This is because, compared to Llava-CoT and LlamaV-o1, Corvid-o1's reasoning traces do not strictly adhere to the annotated multi-step structure in VRC-Bench. It tends to generate more streamlined and simplified reasoning processes rather than following the predefined step-by-step format.

Model	Llama-3.2 Vision [17]	Mulberry [24]	LLaVA-o1 [23]	LlamaV-o1 [20]	Corvid-o1 (Ours)
Final Answer	48.40	51.90	54.09	56.49	61.90
Steps	58.37	63.86	66.21	68.93	63.93

Table 11. **Comparison with o1-like MLLMs on VRC-Bench.**

B.3. Influence of α on Self-Verification

In the proposed self-verification strategy, α is a weighting factor used to trade-off the cross-modal representation similarity $\mathcal{S}$ and the model confidence $\mathcal{C}$ for the final decision $\mathcal{SC}$. This relationship is formally expressed as:

$$\mathcal{SC} = (1-\alpha)\mathcal{S} + \alpha\mathcal{C}.$$

To analyze the influence of α, we conduct ablation studies by varying α from 0.0 to 1.0 with a step size of 0.1. Figure 5 illustrates the relationship between model performance and the weighting factor α in our self-verification strategy. As α increases from 0.0 to 0.7, accuracy rises significantly from 48.6 to a peak of 55.6, demonstrating the advantage of incorporating model confidence into the final answer selection. Beyond α = 0.7, performance gradually declines, suggesting that overweighting confidence relative to cross-modal similarity degrades effectiveness. The optimal value (α = 0.7) indicates that while both components contribute meaningfully to verification performance, a configuration that slightly prioritizes confidence yields superior results.

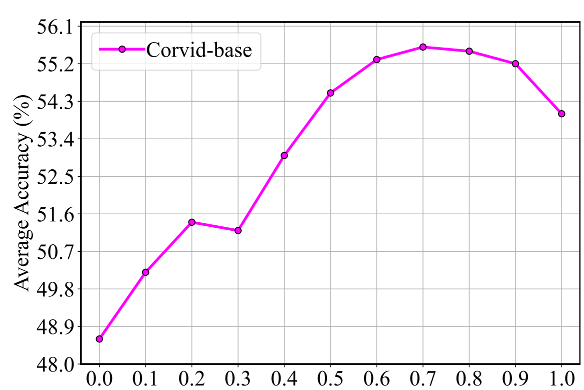
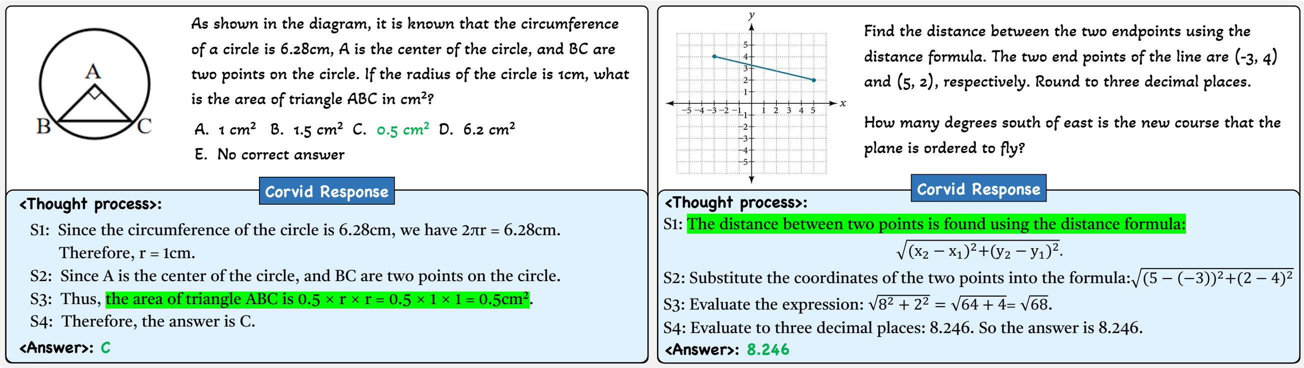
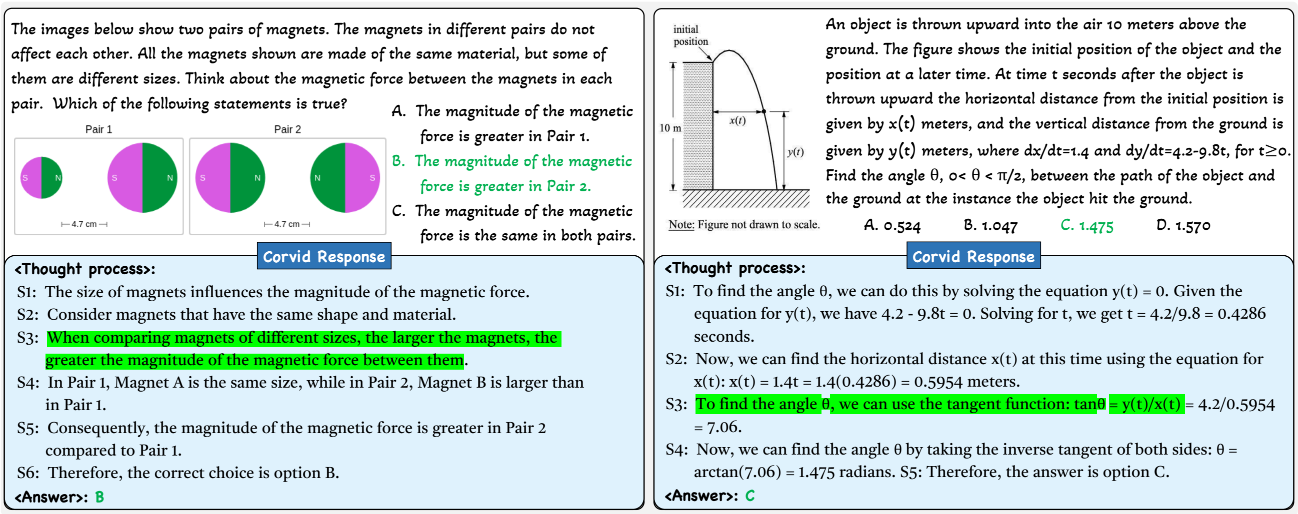


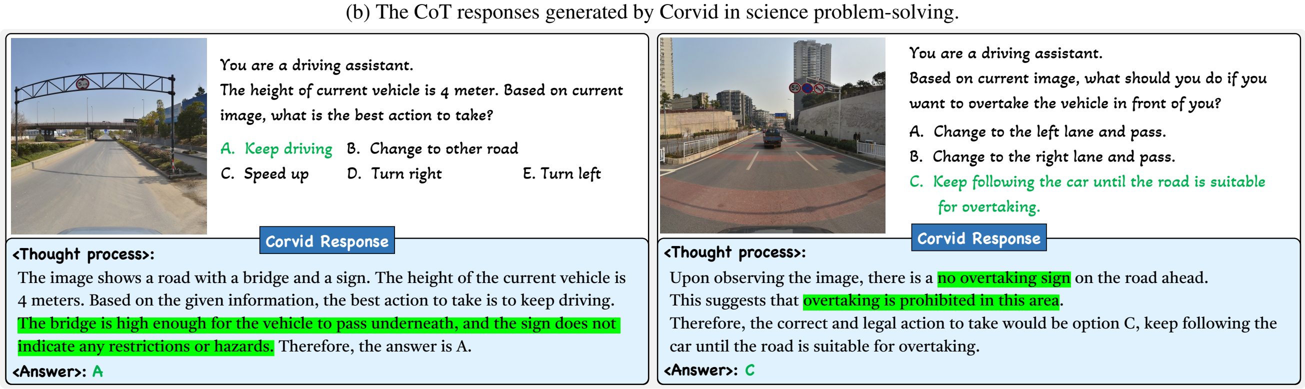
Figure 5. **Influence of α on our self-verification strategy.**



(a) The CoT responses generated by Corvid in mathematical reasoning.



(b) The CoT responses generated by Corvid in science problem-solving.

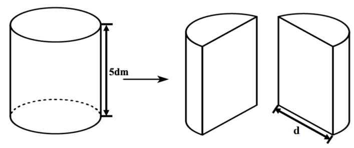


(c) The CoT responses generated by Corvid in visual comprehension.

Figure 6. **Qualitative result of Corvid** in performing mathematical reasoning, science problem-solving, and visual comprehension.

B.4. Inference Efficiency

To evaluate the efficiency and effectiveness of our inference-time scaling strategy, we compare the average inference time and accuracy of our Corvid, LLaVA-o1, and LlamaV-o1 on the MathVista benchmark comprising 1,000 test instances. LLaVA-o1 and LlamaV-o1 utilize stage-level beam search and sentence-level beam search, respectively. Following their optimal configurations, we set the beam size to 2 for LLaVA-o1 and 4 for LlamaV-o1. Tab. 12 shows that Corvid achieves significantly lower inference latency while maintaining higher accuracy. Specifically, Corvid-o1 reduces the inference time per instance to 11.4 seconds, yielding a 4.4× speedup over LLaVA-o1 and a 1.35× speedup over LlamaV-o1, while also surpassing both models in accuracy. This improvement is primarily attributed to Corvid-o1's more streamlined and simplified intermediate reasoning process during inference, as its inference time grows linearly with



Question: A cylindrical piece of wood with a height of 5 dm is cut vertically along the diameter of the base into two parts (as shown in the diagram). As a result, the surface area increases by 60 square decimeters. What is the value of d? (d = __ dm)

Options: A. 6 B. 4 C. 5 D. No correct answer

Corvid-o1

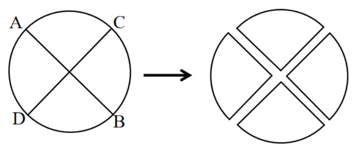
<Thought process>:

The surface area of a cylinder is given by the formula $2\pi rh$, where r is the radius and h is the height. In this case, the height of the cylinder is 5 dm. When the cylinder is cut along the diameter, the radius becomes half of the diameter, which is d/2. The surface area of the two smaller cylinders is $2\pi(d/2)h = d\pi h$. The increase in surface area is 60 square decimeters, so we can set up the...

Solving for d, we find d = 4. Therefore, the value of d is 4 dm. Therefore, the answer is option B.

<Answer>: B

Figure 7. **A failure case** where Corvid-o1 fails to reason effectively due to insufficient domain-specific knowledge.



Question: As shown in the diagram, there is a circular cake with a radius of 5 cm, which is divided into four parts. Compared to the original circumference of the cake, the additional edge length is ( )

Options: A. 4AB+4CD B. AB+CD C. 2AB+2CD D. No correct answer

LlamaV-o1

<Thought process>:

1. *Calculate the original circumference of the circular cake using the formula $C = 2 \times \pi \times r$, where $r = 5$ cm. This gives $C = 10 \times \pi$ cm.*
2. *The cake is divided into four equal parts, each with a central angle of 90 degrees.*
3. *The additional edges are formed by the lines AB and CD. These lines are chords of the circle.*
4. *The length of each chord (AB or CD) can be calculated using the formula for the length of a chord in a circle: $l = 2 \times r \times \sin(\theta / 2)$, where theta is the central angle.*
5. *For a central angle of 90 degrees, sin(45 degrees) = square root of 2 divided by 2. Thus, the length of each chord is $2 \times 5 \times$ (square root of $2 / 2$) = $5 \times$ square root of 2 cm.*
6. *The total additional edge length is $2 \times 5 \times$ square root of 2 = $10 \times$ square root of 2 cm.*
7. *However, the question asks for the additional edge length in terms of the original circumference, so we compare $10 \times$ square root of 2 to the original circumference $10 \times \pi$.*
8. *The correct comparison is $10 \times$ square root of 2 = $10 \times \pi \times$ (square root of $2 / \pi$), which simplifies to $10 \times \pi \times 1 = 10 \times \pi$.*
9. *Therefore, the additional edge length is equivalent to the original circumference, which matches option A.*

<Answer>: A

Corvid-o1

<Thought process>:

As shown in the diagram, connect AB and CD. Since the radius of the circle is 5 cm, the circumference of the circle is $2\pi r = 2\pi \times 5 = 10\pi$ cm. Compared to the circumference, the additional length of the edge is $2AB + 2CD = 2(AB + CD)$. Therefore, the answer is C, $2AB + 2CD$.

<Answer>: C

Figure 8. **Qualitative comparison** between LlamaV-o1 and Corvid-o1 in mathematical reasoning.

the number of generated tokens.

MLLMs	LLaVA-o1	LlamaV-o1	Corvid-o1†	Corvid-o1-8B
Time (second)	50.6	15.4	11.4	11.3
Accuracy	56.1	54.4	61.5	72.0

Table 12. **The average inference time per instance** on MathVista, evaluated using a single NVIDIA A800 (80G) GPU.

B.5. Qualitative Results

In Figure 6, we provide an intuitive understanding of Corvid's CoT reasoning capabilities. As illustrated, when performing science and math reasoning, as well as visual comprehension, Corvid-o1 consistently generates faithful and detailed thought processes before arriving at an answer, enhancing the reliability and interpretability of its answer and demonstrating exceptional CoT capabilities.

B.6. Additional Failure Case

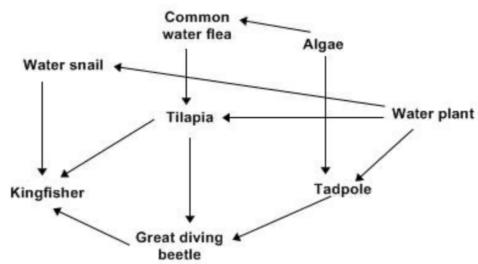
In addition to the case shown in **??**, Figure 7 presents a typical failure case in mathematical reasoning, where Corvid-o1 fails to arrive at the correct answer due to insufficient domain-specific knowledge.

B.7. Qualitative Comparison

Figures 8 to 10 visualize several qualitative comparisons between LlamaV-o1 and Corvid-o1 across tasks.

C. MCoT-Instruct

In this section, we introduce MCoT-Instruct-287K, our high-quality multimodal CoT instruction-following dataset.



Question: Which organism with be most affected if algae was eliminated?

Options:

A. Tilapia
B. Common water flea
C. Great diving beetle
D. Tadpole

LlamaV-o1

<Thought process>:

First, I will identify the organisms that directly consume algae. The arrows from algae point to common water fleas and tadpoles, indicating these organisms rely on algae for food. Next, I will trace the connections to see which organism is most dependent on algae. The common water flea is directly connected to algae, suggesting it would be most affected by the elimination of algae.

<Answer>: B

Corvid-o1

<Thought process>:

The common water flea is shown to have a direct interaction with algae, as indicated by the arrow pointing from algae to the common water flea. This suggests that the common water flea relies on algae for food and would be most affected if algae were eliminated. Therefore, the correct answer is option B.

<Answer>: B

Figure 9. **Qualitative comparison** between LlamaV-o1 and Corvid-o1 in science problem-solving.

Question: You are a driving assistant. Based on current image, what should you do if you want to overtake the vehicle in front of you?

Options:

A. Speed up to 50km/h
B. Accelerate through the vehicle gap
C. Overtaking is not allowed. Follow the vehicle in front of you slowly

LlamaV-o1

<Thought process>:

1. *Assess the weather conditions: Heavy rain and poor visibility suggest that the road is slippery and visibility is low.*
2. *Consider the road conditions: The wet road increases the risk of skidding and reduces traction.*
3. *Evaluate the traffic situation: Vehicles are closely spaced, indicating a slow-moving traffic jam.*
4. *Apply driving rules: In such conditions, overtaking is generally not advisable due to reduced visibility and increased risk of accidents.*
5. *Conclude based on safety: Given the hazardous conditions, the safest action is to follow the vehicle in front slowly.*

<Answer>: C

Corvid-o1

<Thought process>:

The image shows that the road is wet and the visibility is low. Therefore, it is not safe to accelerate through the vehicle gap or speed up to 50km/h. Instead, you should follow the vehicle in front of you slowly. Hence, the correct answer is C.

<Answer>: C

Figure 10. **Qualitative comparison** between LlamaV-o1 and Corvid-o1 in visual comprehension.

Specifically, we first describe its sources and then elaborate the process of improving the quality of raw CoTs.

C.1. Source of Raw Data

As detailed in Tab. 13, we collect data from seven manually created reasoning datasets and three AI-assisted generated reasoning datasets, totaling 292K raw instances spanning diverse reasoning types and domains, to construct MCoT-Instruct. Although all datasets provide initial rationales that serve as CoT responses, significant quality issues exist: AI-assisted generated CoTs may contain errors and duplications, while manually-created CoTs are usually brief and logically incoherent, rendering the raw data too noisy and unstandardized for effective CoT-enhancement training.

C.2. Improving the Quality of Raw CoT

To improve the quality of raw CoTs, we separately refine and standardize the aforementioned manually created and AI-

Reasoning Type	Raw Dataset	Size
① General visual reasoning	GPT-VQA [30]	26K
② Knowledge-intensive visual reasoning	A-OKVQA [19]	18K
③ Visual Commonsense Reasoning	VCR [28]	84K
④ Science Problem-Solving	M^3CoT [3]	9K
	SQA-IMG (train) [14]	8K
	ArxivQA [11]	54K
⑤ Geometric Reasoning	GeomVerse [8]	9K
	R-CoT [5]	53K
⑥ Numerical Reasoning	GeoQA [1]	7K
⑦ Mathematical reasoning	TabMWP [16]	24K

Table 13. **Raw data of MCoT-Instruct.** Here, GPT-VQA, R-CoT, and ArxivQA are the three AI-assisted generated datasets.

Given Multimodal Input

Which term matches the picture?

A. endotherms B. ectotherms

Raw CoT:
Endotherms regulate their temperature internally. Horses and other mammals are endotherms. On a hot day, horses can sweat to regulate their body temperature.

Rewritten CoT:
Upon observing the image, we can find there is a horse.
Among the given options, endotherms are organisms that have the inherent ability to regulate their body temperature internally.
Indeed, horses and other mammals are examples of endotherms.
Therefore, the term that matches the picture is option A, endotherms.
<Answer>: A

Figure 11. **Comparison between raw and rewritten CoTs.**

generated reasoning datasets with GPT assistance through the following two steps:

- **CoT Rewriting.** As shown in Figure 12, we design a specialized prompt to instruct GPT-4o to refine these raw CoTs from manually-created datasets, enhancing their diversity and logical consistency. As demonstrated in Figure 11, the rewritten CoTs remain faithful and consistent with the given context while becoming more detailed, logically coherent, and standardized.
- **Quality Verification and Data Filtering.** To guarantee the quality of all rewritten CoTs and those from AI-assisted generated datasets, we employ GPT to evaluate free-text CoTs across three dimensions: *faithfulness*, *relevance*, and *completeness*. Inspired by the success of LLMs in automatic evaluation [4, 13], we design a base prompt, as shown in Figure 13, to instruct GPT-4o to assign an overall score (0 - 1) to each rewritten CoT and its corresponding raw CoT. The CoT with the higher score is selected as the high-quality CoT. After that, we filter out instances with an overall score below 0.6.

With these steps, we ultimately obtain 287K instance with high-quality CoT responses that are consolidated into single-turn conversation instances of MCoT-Instruct. Notably, no testing or validation instances from any evaluation benchmark were included in this process. We used only the training split of ScienceQA for data curation, and Corvid's in-domain performance was evaluated exclusively on its respective test set.

Benchmarks	Task Format	Metric	#Sample
MMStar [2]	multi-choice	Accuracy	1,500
MMMU [27]	multi-choice	Accuracy	900
SQA-IMG [14]	multi-choice	Accuracy	2,017
AI2D [9]	multi-choice	Accuracy	3,088
WeMath [18]	multi-choice	Accuracy	1,740
MathVista [15]	multi-choice&free-text	Accuracy	1,000
MathVerse [29]	multi-choice&free-text	Accuracy	3,940
MathVision [21]	multi-choice&free-text	Accuracy	3,040
DynaMath [31]	multi-choice&free-text	Accuracy	5,010
SEED-IMG [10]	multi-choice	Accuracy	14,232
MMT-Val [25]	multi-choice	Accuracy	31,325
RWQA [22]	multi-choice	Accuracy	1901
BLINK [6]	multi-choice	Accuracy	1,901
MMB [12]	multi-choice	Accuracy	6,666
MMVet [26]	free-text	GPT Score	218
Hallusion [7]	multi-choice	Accuracy	254

Table 14. **Summary of evaluation benchmarks.**

D. Benchmark Details

Tab. 14 presents all evaluation benchmarks used in this work. The task formats of MathVista, MathVerse, MathVision, and DynaMath encompass both multiple-choice question answering and free-text generation, while MMVet formats tasks as free-text generation. All other benchmarks are limited to multiple-choice question answering. Each benchmark adopts accuracy as its primary metric, except for MMVet, which utilizes a GPT-based score. Notably, SQA-IMG includes human-annotated CoTs, serving as references for assessing the quality of model's CoT responses in **??**.

References

[1] Jiaqi Chen, Jianheng Tang, Jinghui Qin, Xiaodan Liang, Lingbo Liu, Eric Xing, and Liang Lin. GeoQA: A geometric question answering benchmark towards multimodal numerical reasoning. In *ACL*, pages 513–523, 2021. 19

[2] Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, et al. Are we on the right way for evaluating large vision-language models? *arXiv:2403.20330*, 2024. 15, 19

[3] Qiguang Chen, Libo Qin, Jin Zhang, Zhi Chen, Xiao Xu, and Wanxiang Che. M^3CoT: A novel benchmark for multi-domain multi-step multi-modal chain-of-thought. *arXiv:2405.16473*, 2024. 19

[4] Cheng-Han Chiang and Hung-yi Lee. A closer look into using large language models for automatic evaluation. In *EMNLP*, pages 8928–8942, 2023. 19

[5] Linger Deng, Yuliang Liu, Bohan Li, Dongliang Luo, Liang

System message
You are an AI assistant that can do text rewritten.

Prompt
I want you to act as a Chain-of-Thought (CoT) Rewriter. Given a question with several options and its CoT response (i.e., the intermediate reasoning steps or rationales that lead to the correct answer), your objective is to rewrite the given CoT into a more standardized version.

The rewritten CoT must follow the following rules:
1) Keep the logic of reasoning-then-answering to ensure that the reasoning can be performed step by step.
2) Be faithful enough to ensure that the reasoning can accurately lead to the correct answer.
3) Be clear and concise, without factual errors or repeated content, and no key intermediate reasoning steps are omitted.
4) Do not mention or refer to the given CoT in your responses directly.

You can rewrite the given CoT using the following methods:
1. Improve existing reasoning steps or rationales to make the CoT more coherent and smooth.
2. Add more intermediate reasoning steps or rationales to make the CoT more specific and detailed.

Please make sure you have read and understood these instructions carefully.

Following is an exemplar:

Given question & options:
Which person is in charge?
A. person3 [0.1231, 0.1601, 0.4296, 0.9848] is the leader
B. person2 [0.5440, 0.1151, 0.7528, 0.9990] is in charge
C. person3 [0.1231, 0.1601, 0.4296, 0.9848] seems to be the one in charge
D. person3 [0.1231, 0.1601, 0.4296, 0.9848] has the higher rank
Given CoT:
person3 [0.1231, 0.1601, 0.4296, 0.9848] is walking ahead of the other two and speaking while person1 [0.3602, 0.0979, 0.5477, 0.9689] looks to be more of a worker.
Rewritten CoT:
The person who is in charge can be determined by observing their behavior and position. In this case, person3 [0.1231, 0.1601, 0.4296, 0.9848] is walking ahead of the other two individuals and is also the one who is speaking, indicating that they have a leadership role. On the other hand, person1 [0.3602, 0.0979, 0.5477, 0.9689] appears to be more of a worker. Based on these observations, it can be concluded that person3 seems to be the one in charge.
Hence, the correct answer is B.

Here is the example to be rewritten:
Given Question & Options:
{}
Given CoT:
{}
Rewritten CoT:

Now you can start to rewrite the given CoT.

Figure 12. **Prompt template of CoT rewriting.** Here, we use one instance from the VCR dataset [28] as the exemplar. For other source datasets, replace the given exemplar with data-specific instances.

Wu, Chengquan Zhang, Pengyuan Lyu, Ziyang Zhang, Gang Zhang, Errui Ding, et al. R-cot: Reverse chain-of-thought problem generation for geometric reasoning in large multimodal models. *arXiv:2410.17885*, 2024. 19

[6] Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A Smith, Wei-Chiu Ma, and Ranjay Krishna. Blink: Multimodal large language models can see but not perceive. *arXiv:2404.12390*, 2024. 19

[7] Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, et al. Hallusionbench: an advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *CVPR*, pages 14375–14385, 2024. 15, 19

[8] Mehran Kazemi, Hamidreza Alvari, Ankit Anand, Jialin Wu, Xi Chen, and Radu Soricut. Geomverse: A systematic evaluation of large models for geometric reasoning. *arXiv:2312.12241*, 2023. 19

[9] Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. A diagram is worth a dozen images. In *ECCV*, pages 235–251, 2016. 15, 19

[10] Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. Seed-bench: Benchmarking multimodal llms with generative comprehension. *arXiv:2307.16125*, 2023. 19

[11] Lei Li, Yuqi Wang, Runxin Xu, Peiyi Wang, Xiachong Feng, Lingpeng Kong, and Qi Liu. Multimodal arxiv: A dataset for improving scientific comprehension of large vision-language

System message

You are a helpful AI assistant that can evaluate the quality of free-text chain-of-thought (CoT) responses generated by a multimodal large language model (MLLM).

Prompt

You will be provided with the input context to the MLLM (i.e., an image description, a question, and several options for the question), along with the corresponding CoT response generated by the MLLM. Your task is to evaluate the free-text CoT responses and give a final overall score (0 - 1) based on the following three perspectives:

- ❒ **Faithfulness** (0 - 1): it refers to how accurately the CoT response reflect the actual reasoning process of the MLLM. A faithful CoT response is one that genuinely represents the factors and logic the MLLM used to arrive at its answer. For example, if the MLLM generates an answer based on certain key points in the given context, a faithful CoT response would accurately describe how it picked those points and how they led to the answer. The focus of faithfulness is on the transparency and truthfulness of the explanation.
- ❒ **Relevance** (0 - 1): it measures how the CoT response aligns with and supports the answer generated by the MLLM. A consistent CoT response should logically justify the answer, demonstrating a clear and direct connection between the CoT response and the inferred answer. That is, a consistent CoT response should not only be aligned with the answer but also provide sufficient and convincing reasons for why the answer is valid.
- ❒ **Completeness** (0 - 1): it evaluates whether the CoT response provided by the MLLM encompasses all essential information and reasoning necessary to understand the MLLM's answer reasoning process. A complete CoT response should cover all critical aspects and steps of the MLLM's reasoning without omitting key details.

Evaluation Steps:

1. Understand and analyze the provided image description, question, and options.
2. Read the MLLM's response and systematically assess the CoT response from the three perspectives of Faithfulness, Relevance, and Completeness.
3. Assign a final overall score (0 - 1) by averaging Faithfulness, Relevance, and Completeness.

Please make sure you read and understand these instructions carefully.

The sample to be scored:
Image Description:
{}
Question & Options:
{}
CoT Response:
{}

Evaluation Form:
Answer by starting with "Scoring:" and then give the explanation of the score by "Explanation:"
- Overall:

Figure 13. **Prompt template for CoT quality evaluation.**

models. *arXiv:2403.00231*, 2024. 19

[12] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. MMBench: Is your multi-modal model an all-around player? *arXiv:2307.06281*, 2023. 15, 19

[13] Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. G-eval: NLG evaluation using gpt-4 with better human alignment. In *EMNLP*, pages 2511–2522, 2023. 19

[14] Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *NeurIPS*, pages 2507–2521, 2022. 19

[15] Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. *arXiv:2310.02255*, 2023. 15, 19

[16] Pan Lu, Liang Qiu, Kai-Wei Chang, Ying Nian Wu, Song-Chun Zhu, Tanmay Rajpurohit, Peter Clark, and Ashwin Kalyan. Dynamic prompt learning via policy gradient for semi-structured mathematical reasoning. In *ICLR*, 2023. 19

[17] Meta AI. Llama 3.2: Revolutionizing edge ai and vision with open, customizable models. https://ai.meta.com/blog/llama-3-2-connect-2024-vision-edge-mobile-devices/, 2024. 15

[18] Runqi Qiao, Qiuna Tan, Guanting Dong, Minhui Wu, Chong Sun, Xiaoshuai Song, Zhuoma GongQue, Shanglin Lei, Zhe Wei, Miaoxuan Zhang, et al. We-math: Does your large multimodal model achieve human-like mathematical reasoning? *arXiv:2407.01284*, 2024. 19

[19] Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. A-OKVQA: A benchmark for visual question answering using world knowledge. In *ECCV*, pages 146–162, 2022. 19

[20] Omkar Thawakar, Dinura Dissanayake, Ketan More, Ritesh Thawkar, Ahmed Heakl, Noor Ahsan, Yuhao Li, Mo-

hammed Zumri, Jean Lahoud, Rao Muhammad Anwer, et al. Llamav-o1: Rethinking step-by-step visual reasoning in llms. *arXiv:2501.06186*, 2025. 15

[21] Ke Wang, Junting Pan, Weikang Shi, Zimu Lu, Mingjie Zhan, and Hongsheng Li. Measuring multimodal mathematical reasoning with math-vision dataset. *arXiv:2402.14804*, 2024. 19

[22] X.AI. Grok-1.5 vision preview. `https://x.ai/blog/grok-1.5v`, 2024. 19

[23] Guowei Xu, Peng Jin, Li Hao, Yibing Song, Lichao Sun, and Li Yuan. Llava-o1: Let vision language models reason step-by-step. *arXiv:2411.10440*, 2024. 15

[24] Huanjin Yao, Jiaxing Huang, Wenhao Wu, Jingyi Zhang, Yibo Wang, Shunyu Liu, Yingjie Wang, Yuxin Song, Haocheng Feng, Li Shen, et al. Mulberry: Empowering mllm with o1-like reasoning and reflection via collective monte carlo tree search. *arXiv:2412.18319*, 2024. 15

[25] Kaining Ying, Fanqing Meng, Jin Wang, Zhiqian Li, Han Lin, Yue Yang, Hao Zhang, Wenbo Zhang, Yuqi Lin, Shuo Liu, et al. Mmt-bench: A comprehensive multimodal benchmark for evaluating large vision-language models towards multitask agi. *arXiv:2404.16006*, 2024. 19

[26] Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. Mm-vet: Evaluating large multimodal models for integrated capabilities. *arXiv:2308.02490*, 2023. 15, 19

[27] Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *CVPR*, pages 9556–9567, 2024. 19

[28] Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. From recognition to cognition: Visual commonsense reasoning. In *CVPR*, pages 6720–6731, 2019. 19, 20

[29] Renrui Zhang, Dongzhi Jiang, Yichi Zhang, Haokun Lin, Ziyu Guo, Pengshuo Qiu, Aojun Zhou, Pan Lu, Kai-Wei Chang, Peng Gao, et al. Mathverse: Does your multi-modal llm truly see the diagrams in visual math problems? *arXiv:2403.14624*, 2024. 19

[30] Zhiyuan Zhao, Linke Ouyang, Bin Wang, Siyuan Huang, Pan Zhang, Xiaoyi Dong, Jiaqi Wang, and Conghui He. Mllm-dataengine: An iterative refinement approach for mllm. *arXiv:2308.13566*, 2023. 19

[31] Chengke Zou, Xingang Guo, Rui Yang, Junyu Zhang, Bin Hu, and Huan Zhang. Dynamath: A dynamic visual benchmark for evaluating mathematical reasoning robustness of vision language models. *arXiv:2411.00836*, 2024. 19