ICCV
OCT 19-23, 2025
HONOLULU
HAWAII

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

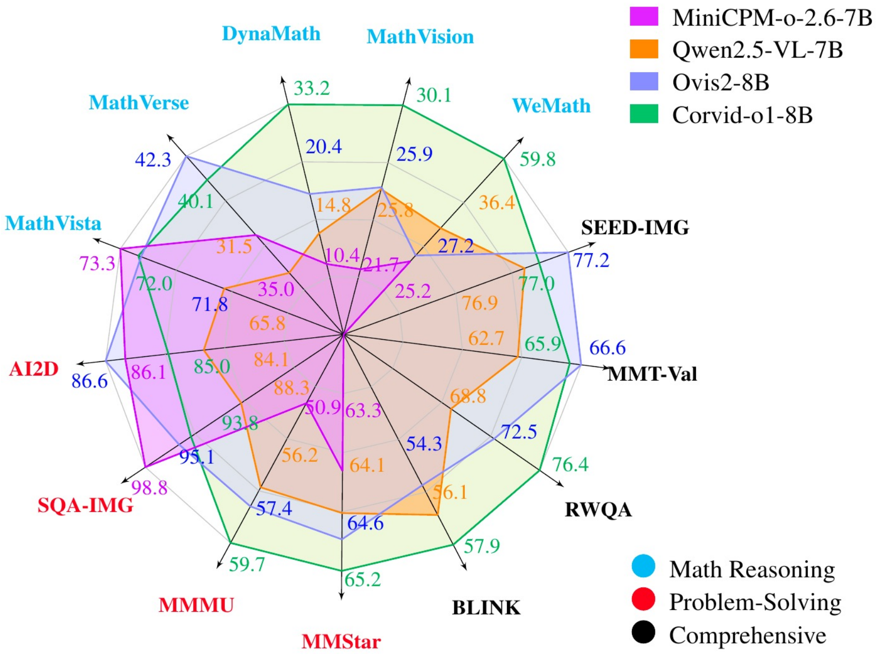
https://mm-vl.github.io/corvid

Jingjing Jiang[1,2] Chao Ma[1*] Xurui Song[2] Hanwang Zhang[2] Jun Luo[2]
[1] Shanghai Jiao Tong University [2] Nanyang Technological University

Background & Motivation

Status of MLLMs

◉ **Pioneering MLLMs still exhibit suboptimal performance on complex tasks requiring deep thinking and extrapolation for effective problem-solving → CoT Reasoning**



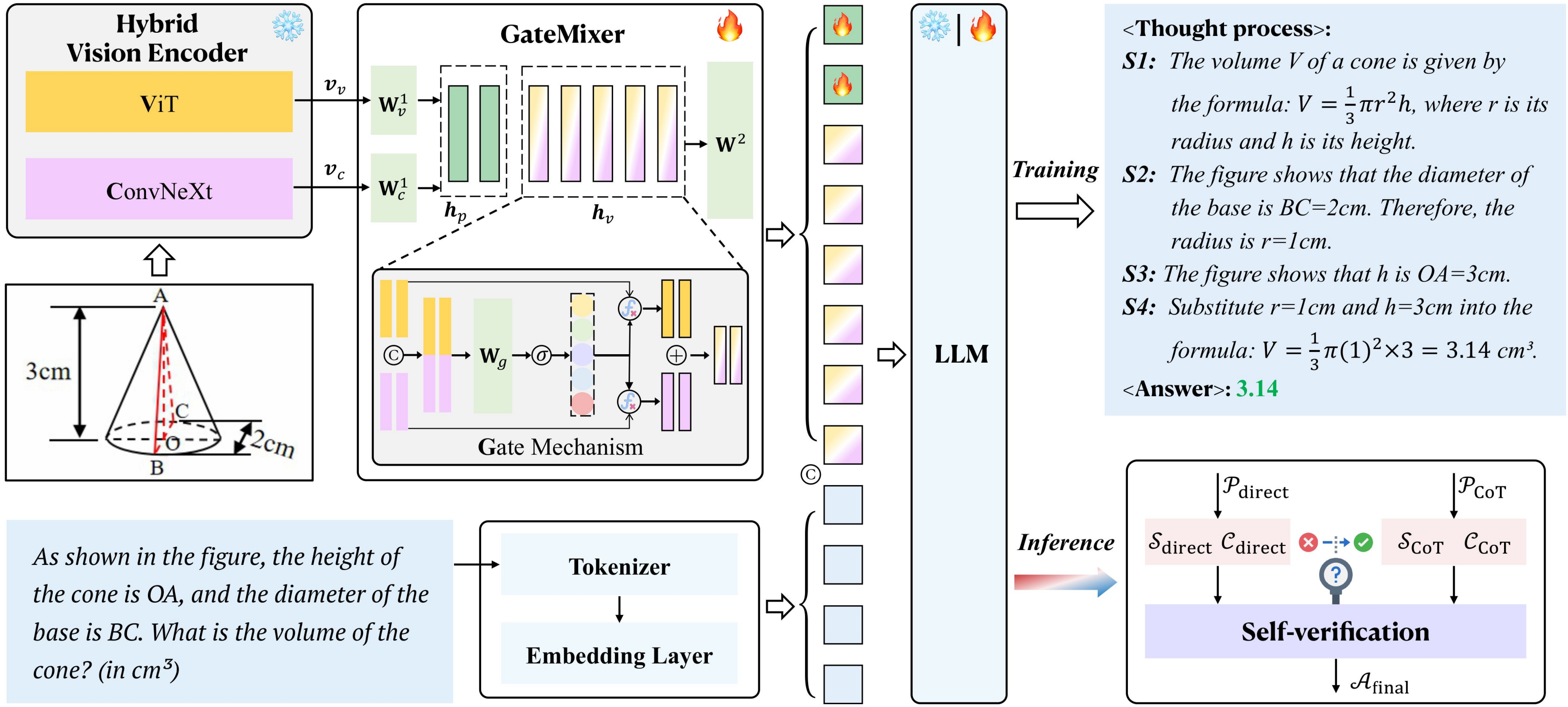
◉ **Challenges in improving MLLMs towards CoT reasoning**

- ✓ **Data:** there remains a significant shortage of high-quality multimodal CoT data
- ✓ **Model architecture:** MLLMs frequently *reason using flawed visual evidence* due to insufficient representation and misalignment
- ✓ **Reasoning strategy**: MLLMs are prone to *over-reasoning* and *under-reasoning* during inference

Methodology: Corvid Architecture

Optimizing standard MLLM architectures to facilitate cross-modal alignment

- **Hybrid Vision Encoder** to effectively represent visual content
- **GateMixer** to facilitate cross-modal alignment



Hybrid Vision Encoder
ViT
ConvNeXt
v_v
v_c
GateMixer
W_v^1
W_c^1
h_p
h_v
W^2
©
W_g
σ
$\oplus$
Gate Mechanism
A
3cm
C
O
B
2cm
Tokenizer
Embedding Layer
LLM
Training
Inference
$\mathcal{P}_{\text{direct}}$
$\mathcal{P}_{\text{CoT}}$
$\mathcal{S}_{\text{direct}}$ $\mathcal{C}_{\text{direct}}$
$\mathcal{S}_{\text{CoT}}$ $\mathcal{C}_{\text{CoT}}$
Self-verification
$\mathcal{A}_{\text{final}}$

As shown in the figure, the height of the cone is OA, and the diameter of the base is BC. What is the volume of the cone? (in cm³)

<**Thought process**>:

S1: *The volume V of a cone is given by the formula:* $V = \frac{1}{3}\pi r^2 h$*, where r is its radius and h is its height.*

S2: *The figure shows that the diameter of the base is BC=2cm. Therefore, the radius is r=1cm.*

S3: *The figure shows that h is OA=3cm.*

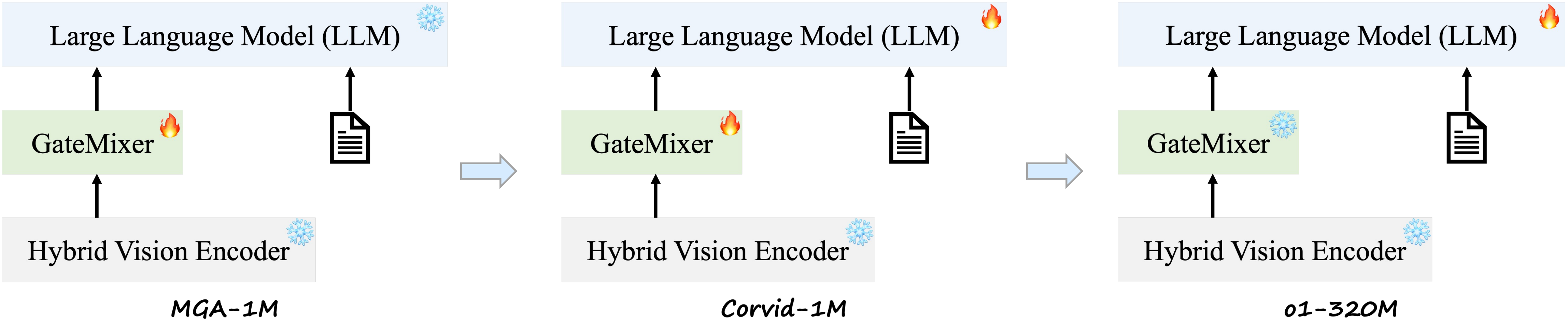
S4: *Substitute r=1cm and h=3cm into the formula:* $V = \frac{1}{3}\pi(1)^2 \times 3 = 3.14$ *cm³.*

<**Answer**>: **3.14**

Corvid incorporates a hybrid vision encoder, a GateMixer, and an LLM

Methodology: Training Procedure

- **Corvid undergoes three consecutive training stages**



S1: Multi-Grained Alignment Pre-training

Large Language Model (LLM) ❄️

GateMixer 🔥

Hybrid Vision Encoder ❄️

MGA-1M

visual-text alignment

S2: CoT-Enhanced Supervised Fine-tuning

Large Language Model (LLM) 🔥

GateMixer 🔥

Hybrid Vision Encoder ❄️

Corvid-1M

Enable Corvid with instruction-following and CoT reasoning capabilities

S3: Pure-CoT Instruction Tuning

Large Language Model (LLM) 🔥

GateMixer ❄️

Hybrid Vision Encoder ❄️

o1-320M

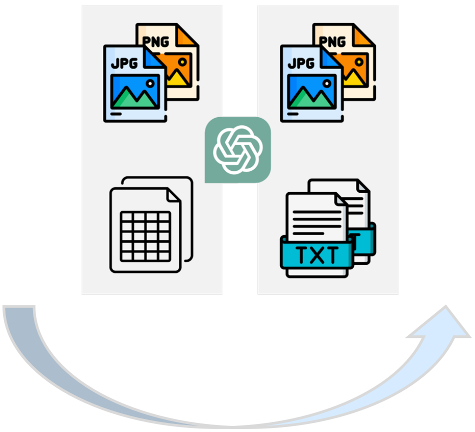
Further refining Corvid's CoT reasoning capability

Methodology: Training Data Curation

■ How to ensure data quality throughout training stages?

◉ **Creating a high-quality multimodal CoT instruction-following dataset, MCoT-Instruct-287K, refined and standardized from diverse public reasoning sources**

Reasoning Type	Raw Dataset	Size
① General visual reasoning	GPT-VQA [28]	26K
② Knowledge-intensive visual reasoning	A-OKVQA [18]	18K
③ Visual Commonsense Reasoning	VCR [26]	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



GPT-based CoT Rewriting and Data Filtering

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**

Comparison between raw and rewritten CoTs

Methodology: Training Data Curation

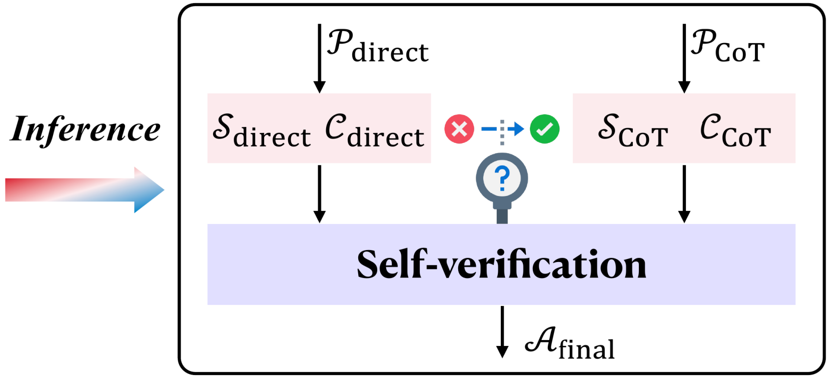
- **How to ensure data quality throughout training stages?**
  - **Collecting and curating publicly available image-text data**

	Name	Data Type (Proportion)	Source Datasets
Stage 1	MGA-1M	Coarse-grained (32.1%)	LLaVA-Pretrain [58] (321K)
		Fine-grained (32.1%)	ALLaVA-4V [10] (195K), Docci [84] (15K), ShareGPT-4o [19] (49K), ShareGPT4V [11] (52K), VG [47] (10K)
		Chart, Math, OCR (35.8%)	ChartCap [43] (30K), MAVIS-Cap [120] (306K), TextCaps [91] (22K)
Stage 2	Corvid-1M	CoT Reasoning (20.2%)	**MCoT-Instruct** (124K), MAVIS-Instruct [120] (81K)
		Direct Reasoning (31.3%)	AI2D [44] (4K), CLEVR [40] (70K), CLEVR-Math [55] (85K), HatefulMemes [45] (8K), VSR [56] (2K), IconQA [65] (27K), Inter-GPS [64] (1K), RAVEN [117] (21K), TallyQA [1] (99K), TQA [2] (1K)
		Chart, Figure, Table (29.7%)	ChartQA [73] (18K), DVQA [41] (49K), FigureQA [42] (49K), HiTab [18] (2K), MapQA [9] (37K), PlotQA [77] (49K), SQA [37] (9K), TAT-QA [127] (2K), WikiSQL [126] (49K), WTQ [87] (38K)
		OCR, Doc (13.8%)	IAM [72] (6K), OCRVQA [79] (80K), InfoGraphicVQA [75] (2K), TextVQA [92] (22K), Visualmrc [94] (3K), DocVQA [74] (10K), ST-VQA [8] (17K)
		Language-only (5.0%)	CamelMath [50] (12K), Dolly [78] (15K), Orca-Math [80] (13K), OpenHermes-2.5 [96] (12K)
Stage 3	o1-320K	CoT Reasoning (100%)	**MCoT-Instruct** (163K), MAVIS-Instruct [120] (137K), CamelMath [50] (20K)

Methodology: Corvid Inference

Inference-Time Self-Verification

◉ **Scaling inference-time computation to mitigate the issues of over-reasoning and under-reasoning, using task difficulty determined by model confidence and cross-modal representation similarity**


Inference
$\mathcal{P}_{\text{direct}}$
$\mathcal{P}_{\text{CoT}}$
$\mathcal{S}_{\text{direct}}$ $\mathcal{C}_{\text{direct}}$
$\mathcal{S}_{\text{CoT}}$ $\mathcal{C}_{\text{CoT}}$
?
Self-verification
$\mathcal{A}_{\text{final}}$

Algorithm 1 Inference-Time Self-Verification

Input: Image: $\mathcal{I}$; Question: $\mathcal{Q}$; Task Prompts: $\{\mathcal{P}_{\text{direct}}, \mathcal{P}_{\text{CoT}}\}$;
Averaging Weight: α.

Output: Answer: $\mathcal{A}$.

```
# Computing Similarity and Confidence during inference
$\mathcal{R}_{\text{direct}}, \mathcal{S}_{\text{direct}}, \mathcal{C}_{\text{direct}} \leftarrow Corvid(\mathcal{I}, \mathcal{Q}, \mathcal{P}_{\text{direct}})$;
$\mathcal{R}_{\text{CoT}}, \mathcal{S}_{\text{CoT}}, \mathcal{C}_{\text{CoT}} \leftarrow Corvid(\mathcal{I}, \mathcal{Q}, \mathcal{P}_{\text{CoT}})$;
# Determining the final answer $\mathcal{A}$
$\mathcal{A}_{\text{direct}} \leftarrow$ Extracting an answer from $\mathcal{R}_{\text{direct}}$;
$\mathcal{A}_{\text{CoT}} \leftarrow$ Extracting an answer from $\mathcal{R}_{\text{CoT}}$;
if $\mathcal{A}_{\text{CoT}} = \mathcal{A}_{\text{direct}}$ then
    return $\mathcal{A}_{\text{CoT}}$
else
    $\mathcal{SC}_{\text{direct}} \leftarrow (1-\alpha)\mathcal{S}_{\text{direct}} + \alpha\mathcal{C}_{\text{direct}}$;
    $\mathcal{SC}_{\text{CoT}} \leftarrow (1-\alpha)\mathcal{S}_{\text{CoT}} + \alpha\mathcal{C}_{\text{CoT}}$;
    if $\mathcal{SC}_{\text{CoT}} >= \mathcal{SC}_{\text{direct}}$ then
        return $\mathcal{A}_{\text{CoT}}$
    else
        return $\mathcal{A}_{\text{direct}}$
```

Experiment: Quantitative Comparison

Multimodal Reasoning Benchmarks

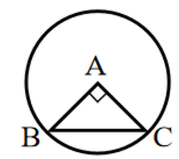
MLLMs	*Problem Solving*				*Mathematical Reasoning*				
	MMStar	MMMU	SQA-IMG	AI2D	MathVista	MathVerse	WeMath	MathVision	DynaMath
LLaVA-v1.5-7B [65]	33.1	35.7	69.2	55.5	25.5	4.3	7.0	11.4	1.4
Janus-Pro-7B [15]	46.5	41.6	83.2	68.1	42.5	15.9	9.7	14.7	4.0
Molmo-7B-D [25]	54.4	48.7	92.2	79.6	48.7	4.2	-	16.2	12.6
GLM-4v-9B [35]	54.8	46.9	96.7	71.2	52.2	15.9	11.8	15.0	8.6
MiniCPM-V-2.6-7B [122]	57.5	49.8	96.7	82.1	60.8	17.6	-	18.4	9.8
URSA-8B [81]	-	-	-	-	58.8	31.0	32.8	28.7	13.2
InternVL2.5-4B-MPO [113]	-	-	-	-	64.1	26.0	-	22.5	10.0
VITA-v1.5-7B [30]	60.2	52.6	95.8	79.2	66.2	23.4	19.4	19.5	9.6
POINTS1.5-7B [69]	61.1	53.8	95.0	81.4	66.4	26.6	24.6	22.0	14.2
Ovis2-4B [79]	61.6	49.0	94.0	85.7	69.6	38.5	16.9	21.5	18.0
InternVL2.5-8B [16]	63.2	56.2	-	84.6	64.5	22.8	23.5	17.0	9.4
MiniCPM-o-2.6-7B [122]	63.3	50.9	**98.8**	86.1	**73.3**	35.0	25.2	21.7	10.4
Qwen2.5-VL-7B [6]	64.1	56.2	88.3	84.1	65.8	31.5	36.4	25.8	14.8
Ovis2-8B [79]	64.6	57.4	95.1	**86.6**	71.8	**42.3**	27.2	25.9	20.4
▼ *base LLM: Llama3-8B-Instruct*									
VILA1.5-8B [61]	39.7	37.4	73.2	58.8	37.4	-	-	-	-
Mantis-8B [44]	41.3	41.1	75.5	60.4	32.7	-	-	-	-
Slime-8B [134]	43.5	38.8	78.0	68.5	41.8	22.9	-	-	-
LLaVA-NeXT-8B [66]	43.9	43.1	73.1	72.8	37.7	-	-	-	-
Idefics3-8B [54]	55.0	46.6	91.3	76.5	58.7	-	-	-	-
MiniCPM-V-2.5-8B [122]	51.8	45.8	89.2	78.4	54.5	-	-	-	-
Bunny-8B [39]	45.4	43.4	79.1	69.4	35.2	-	-	-	-
Ovis1.5-8B [79]	57.3	48.3	88.8	82.5	63.0	-	-	-	-
Cambrian-8B [109]	-	42.7	80.4	73.0	49.0	-	-	-	-
Eagle-X5-8B [101]	-	43.8	84.3	76.1	52.7	-	-	-	-
Corvid-base-8B	62.4	57.4	93.2	82.8	64.8	34.8	54.0	26.8	24.5
Corvid-o1-8B	**65.2**	**59.7**	93.8	85.0	72.0	40.1	**59.8**	**30.1**	**33.2**

Comprehensive Benchmarks

MLLMs	SEED-IMG	MMT-Val	RWQA	BLINK
Monkey-Chat-7B [59]	68.9	53.3	52.4	47.1
DeepSeek-VL-7B [72]	70.1	53.5	54.2	40.9
Molmo-7B-D [25]	74.1	56.8	68.2	46.1
VITA-v1.5-7B [30]	74.1	59.5	66.9	45.0
POINTS1.5-7B [69]	75.1	61.8	67.5	44.0
MiniCPM-V-2.6-7B [122]	74.0	60.8	65.0	55.2
Qwen2.5-VL-7B [6]	76.9	62.7	68.8	56.1
InternVL2.5-8B-MPO [113]	76.8	62.5	68.8	56.6
Ovis2-8B [79]	**77.2**	**66.6**	72.5	54.3
▼ *base LLM: Llama3-8B-Instruct*				
VILA1.5-8B [61]	65.0	48.7	43.4	39.5
Slime-8B [134]	69.8	50.2	58.0	38.8
Mantis-8B [44]	71.2	54.3	59.5	50.1
LLaVA-NeXT-8B [66]	72.5	53.1	58.4	43.5
Bunny-8B [39]	73.5	54.8	60.4	41.6
Eagle-X5-7B [101]	73.6	52.6	63.8	22.4
Ovis1.5-8B [79]	75.4	60.7	64.2	39.8
Idefics3-8B [54]	73.8	58.4	62.6	50.3
Corvid-base-8B	76.4	62.5	**77.0**	55.7
Corvid-o1-8B	77.0	65.9	76.4	**57.9**

Experiment: Qualitative Results

✏️ Mathematical Reasoning



As shown in the diagram, it is known that the circumference of a circle is 6.28cm, A is the center of the circle, and BC are two points on the circle. If the radius of the circle is 1cm, what is the area of triangle ABC in cm²?

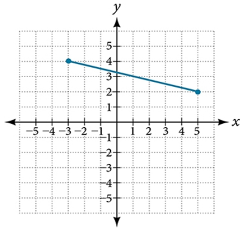
A. 1 cm² B. 1.5 cm² C. 0.5 cm² D. 6.2 cm²
E. No correct answer

Corvid Response

<Thought process>:

S1: Since the circumference of the circle is 6.28cm, we have 2πr = 6.28cm. Therefore, r = 1cm.
S2: Since A is the center of the circle, and BC are two points on the circle.
S3: Thus, the area of triangle ABC is 0.5 × r × r = 0.5 × 1 × 1 = 0.5cm².
S4: Therefore, the answer is C.

<Answer>: C



Find the distance between the two endpoints using the distance formula. The two end points of the line are (-3, 4) and (5, 2), respectively. Round to three decimal places.

How many degrees south of east is the new course that the plane is ordered to fly?

Corvid Response

<Thought process>:

S1: The distance between two points is found using the distance formula:

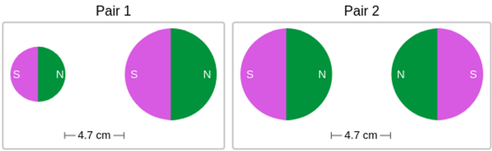
$$\sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}.$$

S2: Substitute the coordinates of the two points into the formula: $\sqrt{(5 - (-3))^2 + (2 - 4)^2}$
S3: Evaluate the expression: $\sqrt{8^2 + 2^2} = \sqrt{64 + 4} = \sqrt{68}$.
S4: Evaluate to three decimal places: 8.246. So the answer is 8.246.

<Answer>: 8.246

✏️ Science Problem Solving

The images below show two pairs of magnets. The magnets in different pairs do not affect each other. All the magnets shown are made of the same material, but some of them are different sizes. Think about the magnetic force between the magnets in each pair. Which of the following statements is true?



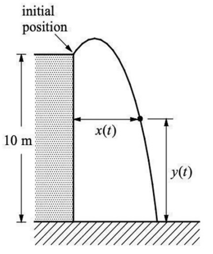
A. The magnitude of the magnetic force is greater in Pair 1.
B. The magnitude of the magnetic force is greater in Pair 2.
C. The magnitude of the magnetic force is the same in both pairs.

Corvid Response

<Thought process>:

S1: The size of magnets influences the magnitude of the magnetic force.
S2: Consider magnets that have the same shape and material.
S3: When comparing magnets of different sizes, the larger the magnets, the greater the magnitude of the magnetic force between them.
S4: In Pair 1, Magnet A is the same size, while in Pair 2, Magnet B is larger than in Pair 1.
S5: Consequently, the magnitude of the magnetic force is greater in Pair 2 compared to Pair 1.
S6: Therefore, the correct choice is option B.

<Answer>: B



Note: Figure not drawn to scale.

An object is thrown upward into the air 10 meters above the ground. The figure shows the initial position of the object and the position at a later time. At time t seconds after the object is thrown upward the horizontal distance from the initial position is given by x(t) meters, and the vertical distance from the ground is given by y(t) meters, where dx/dt=1.4 and dy/dt=4.2-9.8t, for t≥0. Find the angle θ, 0< θ < π/2, between the path of the object and the ground at the instance the object hit the ground.

A. 0.524 B. 1.047 C. 1.475 D. 1.570

Corvid Response

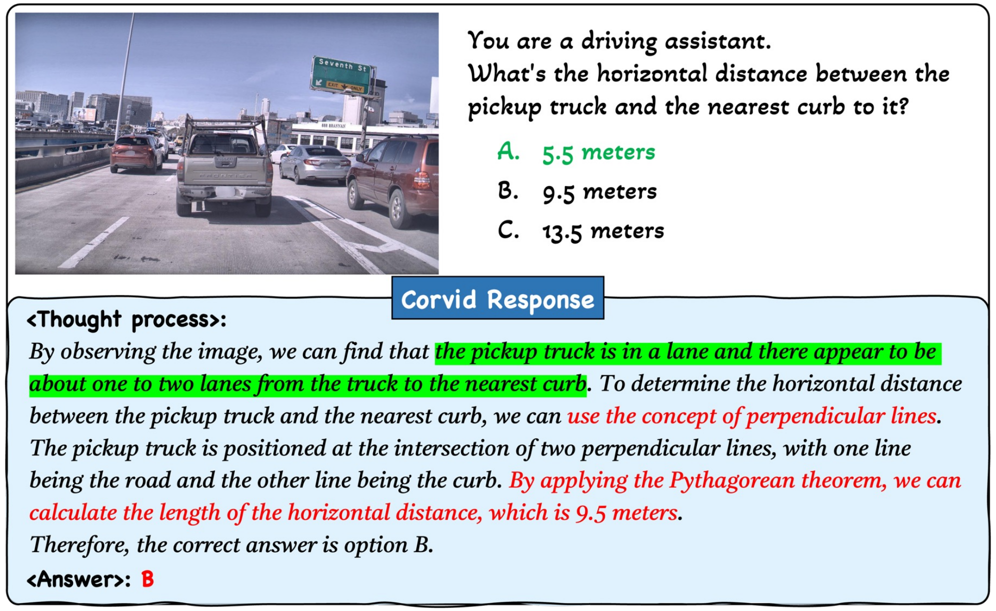
<Thought process>:

S1: To find the angle θ, we can do this by solving the equation y(t) = 0. Given the equation for y(t), we have 4.2 - 9.8t = 0. Solving for t, we get t = 4.2/9.8 = 0.4286 seconds.
S2: Now, we can find the horizontal distance x(t) at this time using the equation for x(t): x(t) = 1.4t = 1.4(0.4286) = 0.5954 meters.
S3: To find the angle θ, we can use the tangent function: tanθ = y(t)/x(t) = 4.2/0.5954 = 7.06.
S4: Now, we can find the angle θ by taking the inverse tangent of both sides: θ = arctan(7.06) = 1.475 radians. S5: Therefore, the answer is option C.

<Answer>: C

Conclusion

Failure Case Study



You are a driving assistant.
What's the horizontal distance between the pickup truck and the nearest curb to it?

A. 5.5 meters
B. 9.5 meters
C. 13.5 meters

Corvid Response

<Thought process>:
By observing the image, we can find that the pickup truck is in a lane and there appear to be about one to two lanes from the truck to the nearest curb. To determine the horizontal distance between the pickup truck and the nearest curb, we can use the concept of perpendicular lines. The pickup truck is positioned at the intersection of two perpendicular lines, with one line being the road and the other line being the curb. By applying the Pythagorean theorem, we can calculate the length of the horizontal distance, which is 9.5 meters. Therefore, the correct answer is option B.
<Answer>: **B**

◈ Corvid accurately locates the positions of the pickup truck and the nearest curb, it fails to reason correctly due to a lack of world commonsense, i.e., typical lane widths on urban roads or highways range from about 3.5 to 4.5 meters

Conclusion

① Corvid: an MLLM with advanced CoT reasoning capabilities

② Inference-time self-verification strategy

③ MCoT-Instruct-287K

Thanks & QA