Workshop on Reasoning and Planning for LLMs

ICLR International Conference On Learning Representations

ErrorRadar: Benchmarking Complex Mathematical Reasoning of Multimodal Large Language Models Via Error Detection

Yibo Yan, Shen Wang, Jiahao Huo, Hang Li, Boyan Li, Jiamin Su, Xiong Gao, Yi-Fan Zhang, Tianlong Xu, Zhendong Chu, Aoxiao Zhong, Kun Wang, Hui Xiong, Philip S. Yu, Xuming Hu, Qingsong Wen

{yanyibo70,qingsongedu}@gmail.com, xuminghu@hkust-gz.edu.cn

香港科技大学(广州) THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY (GUANGZHOU)

香港科技大學 THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

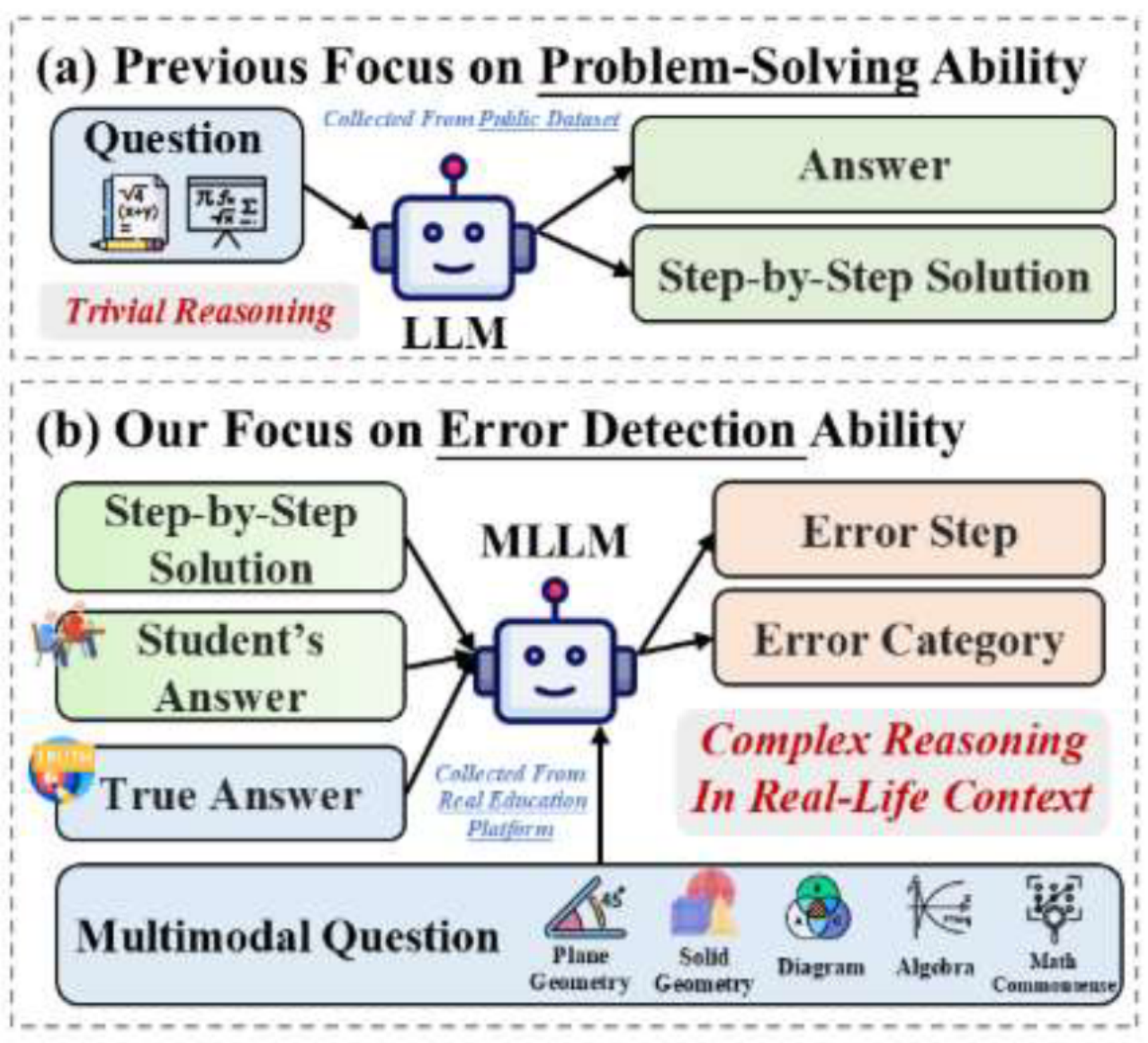
Squirrel Ai Learning

MICHIGAN STATE UNIVERSITY

中国科学院

UNIVERSITY OF ILLINOIS CHICAGO

Introduction



We formulate the new task – **Multimodal Error Detection**, and introduce **ErrorRadar**, the **first benchmark** designed to assess MLLM's capabilities in such a task

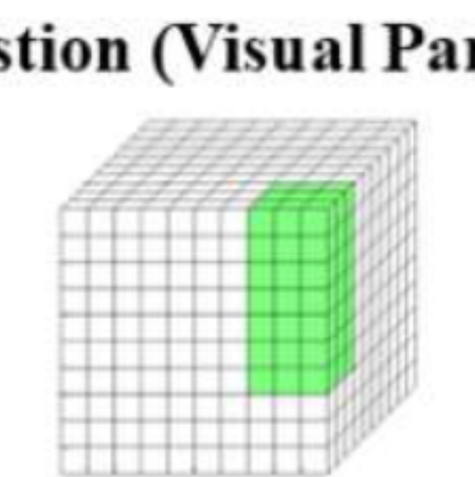
Benchmarks	Venue	Modality	Student Ans.	Error Det.
TheoremQA (Chen et al., 2023a)	EMNLP	T	-	-
MathBench (Liu et al., 2024b)	ACL	T	-	-
MR-GSM8K (Zeng et al., 2024)	arXiv	T	-	-
SciEval (Sun et al., 2024)	AAAI	T	-	-
EIC (Li et al., 2024b)	arXiv	T	-	✓
CMMaTH (Li et al., 2024c)	arXiv	T, I	-	-
MathScape (Zhou et al., 2024a)	arXiv	T, I	-	-
MATH-V (Wang et al., 2024a)	arXiv	T, I	-	-
QRData (Liu et al., 2024c)	ACL	T, I	-	-
IsoBench (Fu et al., 2024)	COLM	T, I	-	-
SciBench (Wang et al., 2024c)	ICML	T, I	-	-
MathVista (Lu et al., 2024b)	ICLR	T, I	-	-
MathVerse (Zhang et al., 2024)	ECCV	T, I	-	-
ErrorRadar (Ours)	-	T, I	✓	✓

Comparison w/ LLM-base math reasoning benchmarks

Dataset

Representative Case

Question (Visual Part):



Question (Textual Part):

What is the decimal representation of the shaded part in the figure below?

Student's Step-by-Step Solution (Reasoning Process):

- Step 1: Read the question and the accompanying image text "What is the decimal representation of the shaded area in the diagram below?"
- Step 2: Analyze the cube given in the diagram; it is a large cube representing "1."
- Step 3: Determine that the large cube is evenly divided into 1000 small cubes, with each small cube representing 1/1000.
- **Step 4: Count the total number of small cubes in the shaded area, which is 21.**
- Step 5: Calculate the proportion of the shaded area, which is 21 times 1/1000, totaling 21/1000.
- Step 6: Convert the fraction 21/1000 into a decimal, resulting in 0.021.
- Step 7: Output the result as 0.021.

True Answer: 0.063

Student's Answer: 0.021

Problem Type: Solid Geometry

Error Category: Visual Perception Error

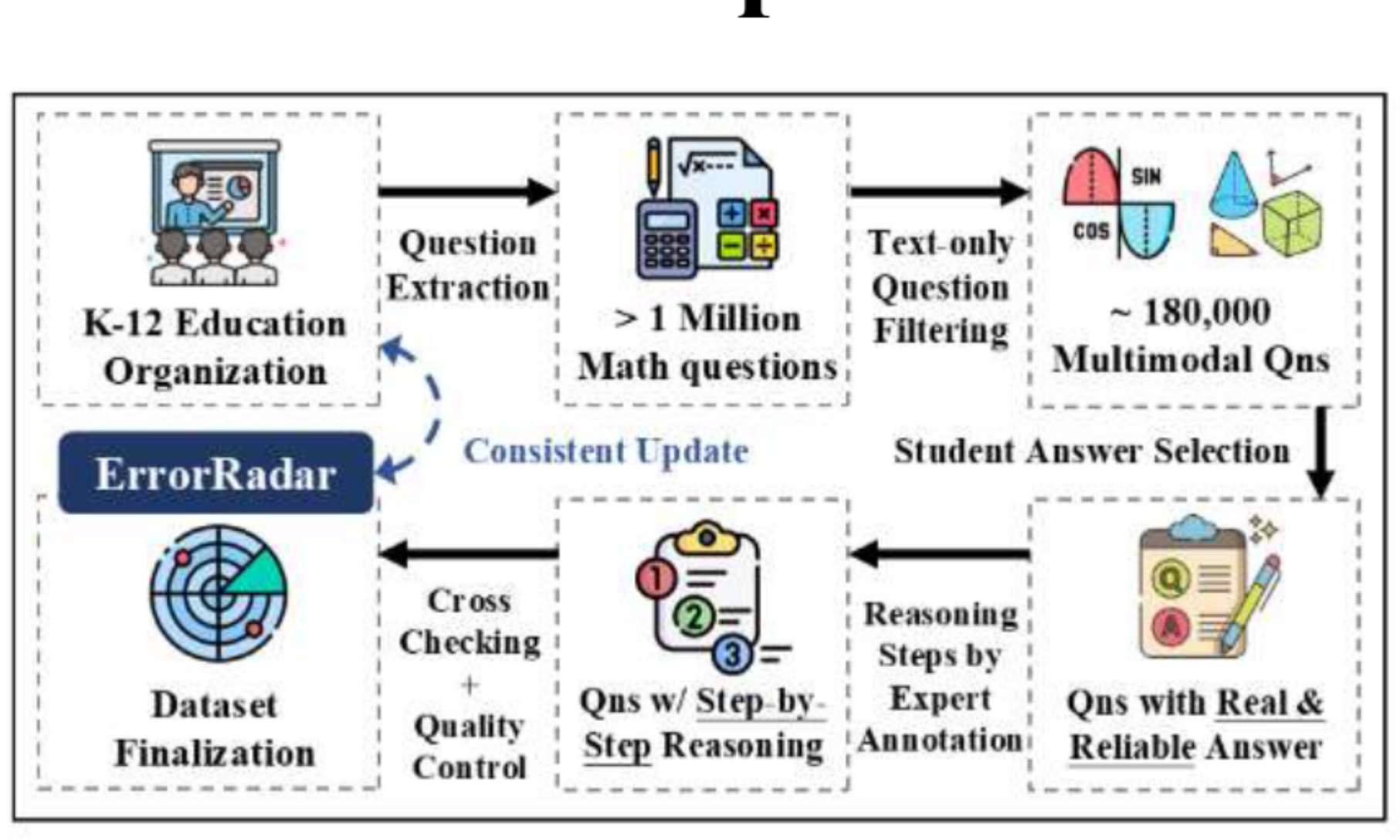
Error Step: Step 4

ErrorRadar	Error Category	Error Step
GPT-4o	Visual Perception Error	Step 4
GPT-4o-mini	Calculation Error ✗	Step 4
Gemini-Pro-1.5	Visual Perception Error	Step 3 ✗
LLaVA-v1.6	Calculation Error ✗	Step 5 ✗

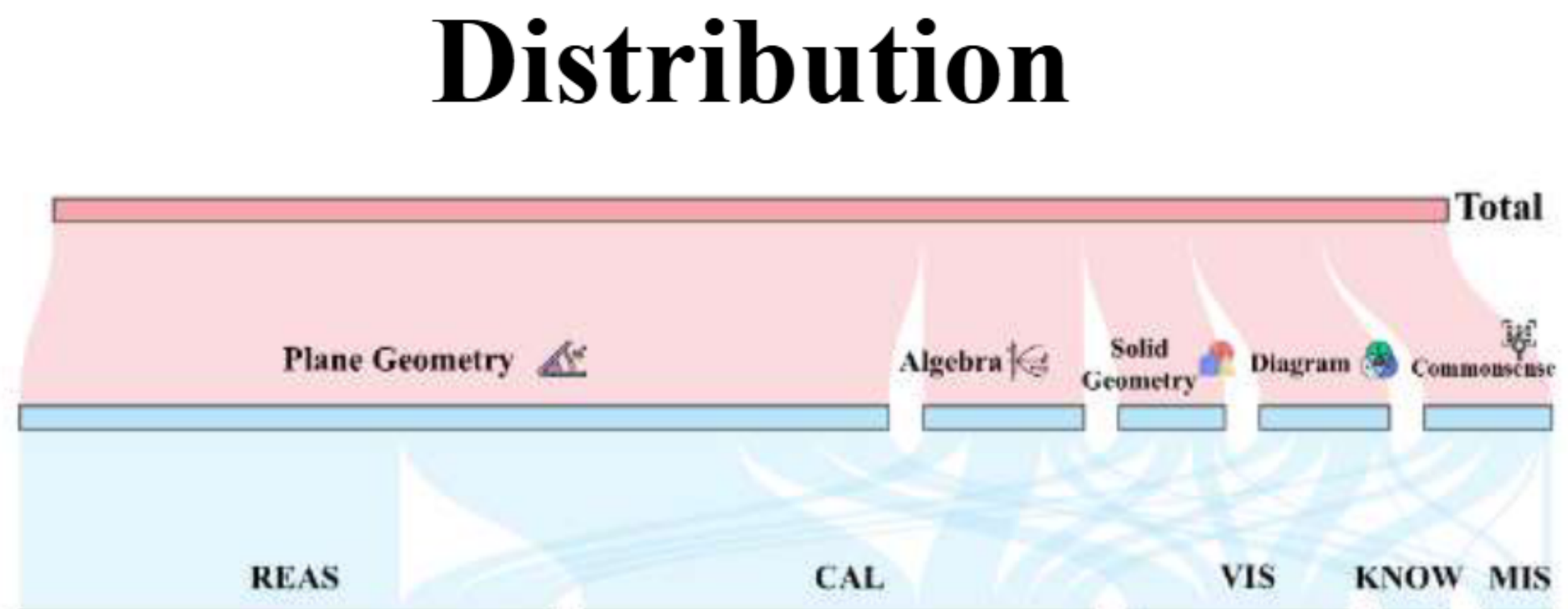
Key Statistics

Statistic	Number
Total multimodal questions	2,500
Problem Type	
- Plane Geometry	1559 (62.4%)
- Solid Geometry	191 (7.6%)
- Diagram	233 (9.3%)
- Algebra	288 (11.5%)
- Math Commonsense	229 (9.2%)
Error Category	
- Visual Perception Error	395 (15.8%)
- Calculation Error	912 (36.5%)
- Reasoning Error	951 (38.0%)
- Knowledge Error	119 (4.8%)
- Misinterpretation of the Qns	123 (4.9%)
Average Reasoning Step	7.6
Maximum Reasoning Step	20
Minimum Reasoning Step	3
Average Question Length	168
Maximum Question Length	719
Minimum Question Length	13

Data Pipeline



Distribution



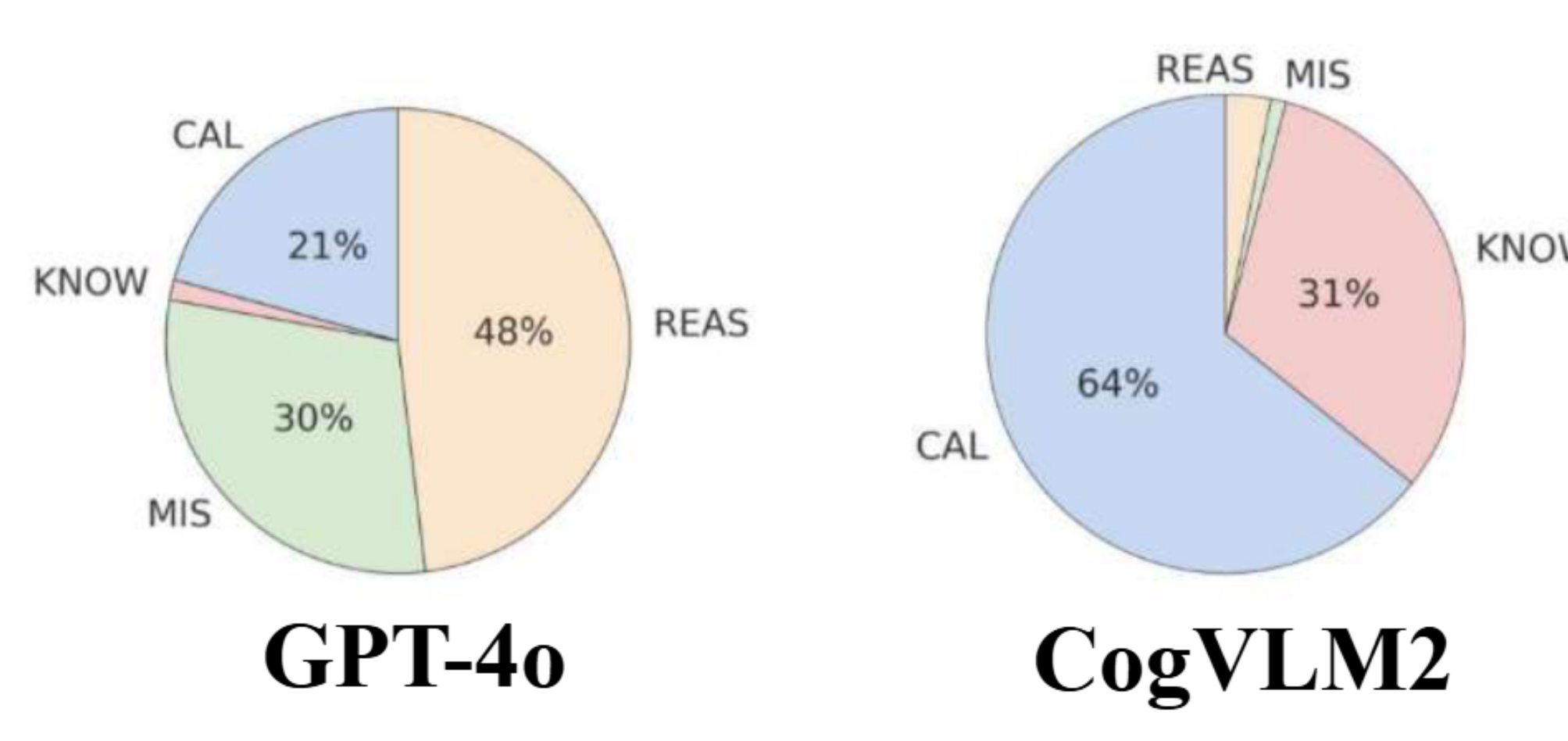
Experiment

➢ Main Result GPT-4o demonstrates the strongest performance

Multimodal Large Language Models	Parameters	LLM	STEP	CATE	VIS	CAL	REAS	KNOW	MIS
Open-Source MLLMs									
InternVL2 (Chen et al., 2023b)	2B	InternLM-2	9.8	25.1	32.2	38.8	12.2	0.0	24.4
Phi-3-vision (Abdin et al., 2024)	4B	Phi-3	37.5	40.7	9.6	99.6	6.6	3.4	4.1
Yi-VL (Young et al., 2024)	6B	Yi	15.7	32.1	9.1	77.1	4.9	14.3	0.0
DeepSeek-VL (Lu et al., 2024a)	7B	DeepSeek	16.2	35.7	4.6	90.9	0.4	28.6	6.5
LLaVA-v1.6-Vicuna (Liu et al., 2024a)	7B	Vicuna-v1.5	30.3	17.7	40.3	14.9	8.3	0.0	55.3
InternVL-2 (Chen et al., 2023b)	8B	InternLM-2.5	44.2	44.1	12.4	99.6	13.6	10.9	2.4
MiniCPM-LLaMA3-V2.5 (Yao et al., 2024)	8B	LLaMA3	37.4	38.0	4.1	100.0	2.1	2.5	0.0
MiniCPM-V2.6 (Yao et al., 2024)	8B	Qwen2	17.0	39.8	11.4	87.8	12.1	10.1	17.9
Qwen-VL (Bai et al., 2023)	9B	Qwen	23.8	38.9	8.6	99.1	3.5	0.0	0.8
GLM-4v (GLM et al., 2024)	13B	GLM-4	44.6	44.1	2.5	92.9	25.8	0.0	0.0
LLaVA-v1.6-Vicuna (Liu et al., 2024a)	13B	Vicuna-v1.5	36.9	47.8	0.0	74.5	53.7	0.8	2.4
CogVLM2-LLaMA3 (Wang et al., 2023a)	19B	LLaMA3	15.0	20.1	43.3	33.8	0.7	13.4	0.0
InternVL2 (Chen et al., 2023b)	26B	InternLM-2	50.4	51.2	39.2	84.6	35.6	0.8	10.6
LLaVA-NEXT (Liu et al., 2024a)	72B	Qwen1.5	51.8	45.0	7.1	86.0	32.0	7.6	0.8
InternVL2 (Chen et al., 2023b)	76B	Hermes-2 Theta	54.4	49.5	33.4	92.4	25.1	10.9	8.1
Closed-Source MLLMs									
Qwen-VL-Max (Bai et al., 2023)	-	-	48.7	52.9	15.2	78.9	50.5	14.3	36.6
Claude-3-Haiku (Anthropic, 2024a)	-	-	45.6	48.0	10.4	77.4	46.8	4.2	1.6
Claude-3.5-Sonnet (Anthropic, 2024b)	-	-	50.2	49.5	35.7	48.4	64.8	21.0	11.4
Gemini-Pro-1.5 (Reid et al., 2024)	-	-	55.0	52.7	43.5	55.7	63.1	18.5	13.0
GPT-4o-mini (OpenAI, 2024b)	-	-	52.0	44.5	9.1	46.8	62.7	31.9	13.0
GPT-4o (OpenAI, 2024a)	-	-	55.1	53.1	46.3	50.4	64.9	9.2	46.3
Human									
Human performance	-	-	69.8	60.7	66.8	75.9	47.6	35.3	53.7

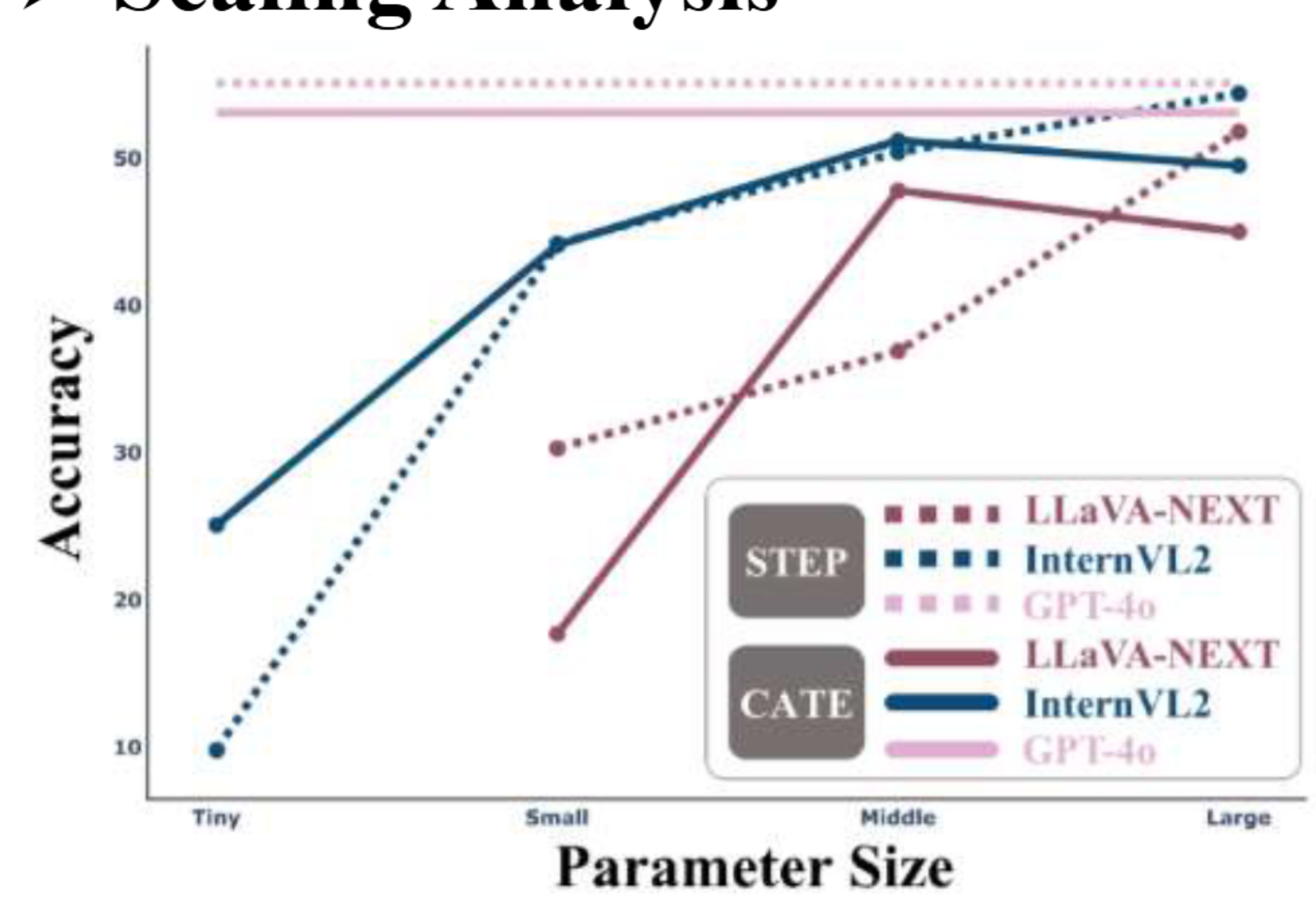
- **CAL is the easiest category, while KNOW is the most difficult**
- **MLLMs still have a gap to reach human-level error detection**

➢ Misjudged VIS Cases Distribution



- **Closed-source MLLMs tend to misjudge VIS as REAS**
- **Open-source MLLMs tend to misjudge VIS as CAL**

➢ Scaling Analysis

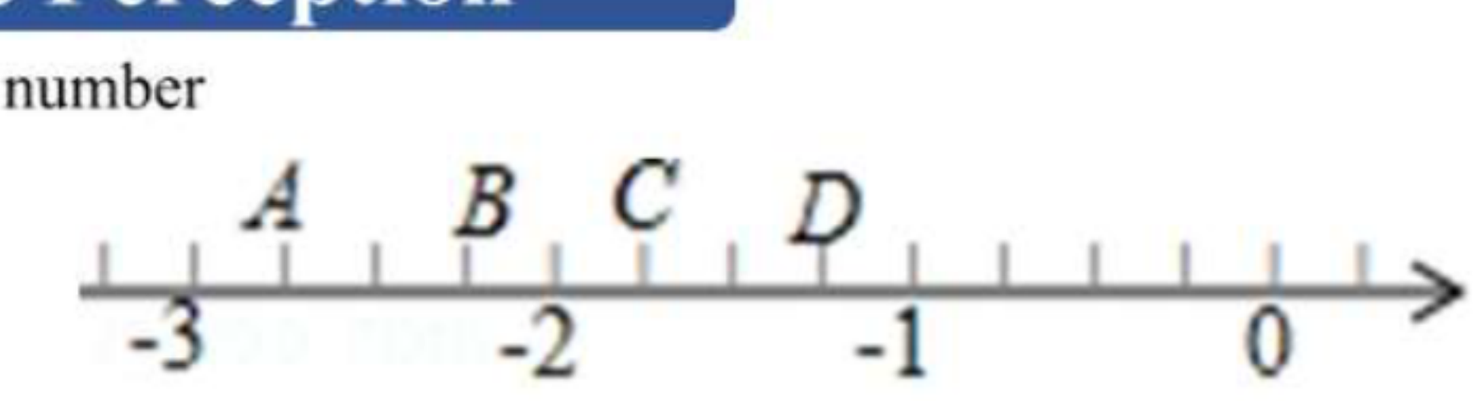


- **STEP performance increases via scaling**
- **CATE is relatively difficult to improve through scaling**

➢ Visual Bad Case Analysis

Distance Perception

Question: As shown in the figure, the point on the number line representing -2.75 is point ___.



True Answer: A

Student's Answer: B

Reasoning Steps:

- Step 1: Read the question and the text on the number line, which states "The point representing -2.75 on the number line is ______ point."
- Step 2: Determine the position of the point representing -2.75 on the number line.
- Step 3: Identify from the diagram the values represented by points A, B, C, and D.
- Step 4: Establish that point B is near the position representing -3, point A represents -2.25, point C represents -1.75, and point D represents -1.25.
- Step 5: Based on the number line position, conclude that the point representing -2.75 is point B.
- Step 6: Output the result as B.

Err. Cate. by GPT-4o: Reasoning Err.

Refer to *Diagram/Spatial/Flip&Fold/Shape Perception* in our paper

Conclusion

- By focusing on both ***error step identification*** and ***error categorization***, ErrorRadar bridges a critical research gap in assessing MLLMs' capabilities in **complex mathematical reasoning**.
- The dataset's construction, based on ***real-world student interactions***, ensures a robust evaluation framework that reflects **genuine user needs**.
- Our **extensive experimental analysis** reveals significant challenges in error detection, highlighting the need for continued advancements.