University of Chinese Academy of Sciences

香港中文大學
The Chinese University of Hong Kong

ICML
International Conference On Machine Learning

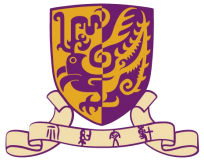
MIND: Multi-rationale INtegrated Discriminative Reasoning Framework for Multi-modal Large Language Models

arXiv:2512.05530

Github: https://github.com/YuChuang1205/MIND

Reporter: Yu Chuang

Data: 2026.06.01



香港中文大學
The Chinese University of Hong Kong

ICML
International Conference
On Machine Learning

第一章节

PART ONE

Background

Background

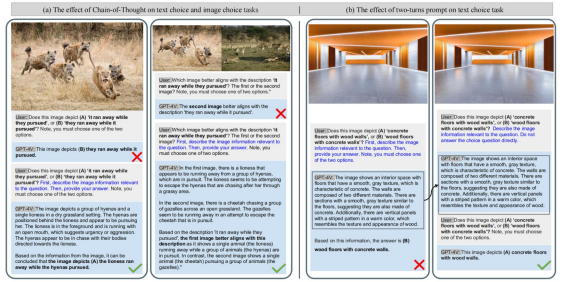


Figure 2: Examples of results with different prompt configurations. Text in blue highlights differences in the prompts. All figures shown here are from Winoground (Thrush et al., 2022)

Desp-CoT

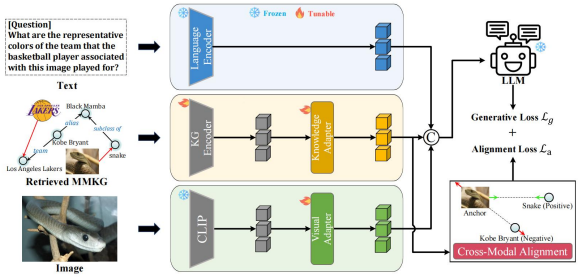


Figure 2: The overview of our MR-MKG approach. Text, multimodal knowledge graph, and image are independently embedded and then concatenated to form prompt embedding tokens. A cross-modal alignment module is designed to enhance the image-text alignment through a matching task within MMKGs.

MR-MKG

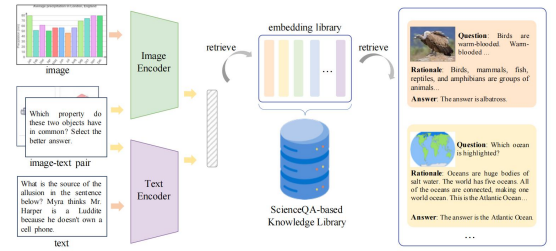


Figure 2: The overall architecture and the retrieval mechanism of the bi-modality retrieval module.

RMR

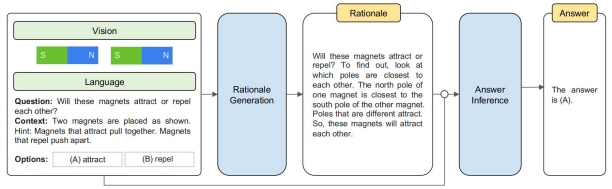


Figure 4: Overview of our Multimodal-CoT framework. Multimodal-CoT consists of two stages: (i) rationale generation and (ii) answer inference. Both stages share the same model structure but differ in the input and output. In the first stage, we feed the model with language and vision inputs to generate rationales. In the second stage, we append the original language input with the rationale generated from the first stage. Then, we feed the updated language input with the original vision input to the model to infer the answer.

Multimodal-CoT

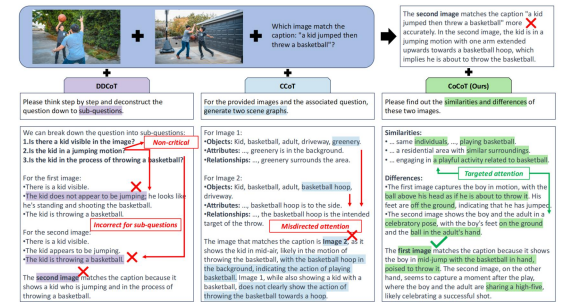
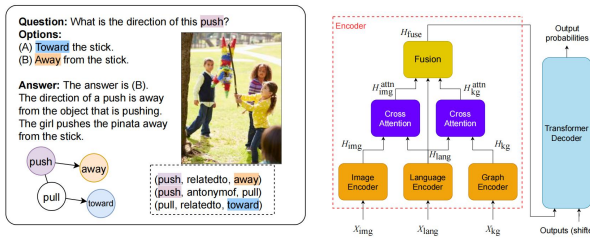


Fig. 1. Comparison between different multimodal prompting strategies. The unique components in each prompting strategy's corresponding response are highlighted in varied colors. Note that GPT-4V is used in this example.

COCoT

. . . . . .

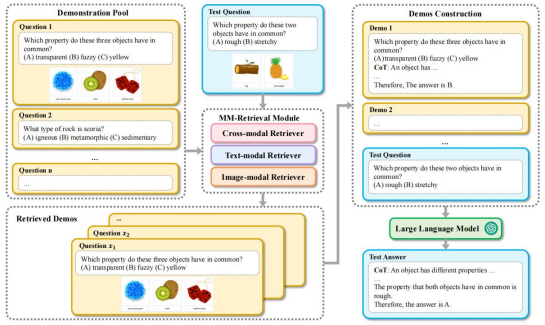
Prompt-based



KAM-CoT

. . . . . .

KG-based



CoT-MM-Retrieval

. . . . . .

RAG-based

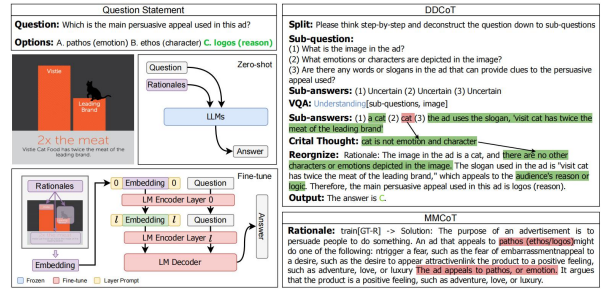


Figure 5: An overview of our DDCoT and its utilization to improve the multimodal reasoning of LMs. Note that although errors encounter in the second sub-problem during visual recognition, the language model rectifies this error in the joint reasoning step with critical thought.

DDCoT

. . . . . .

Two-phase SFT-based

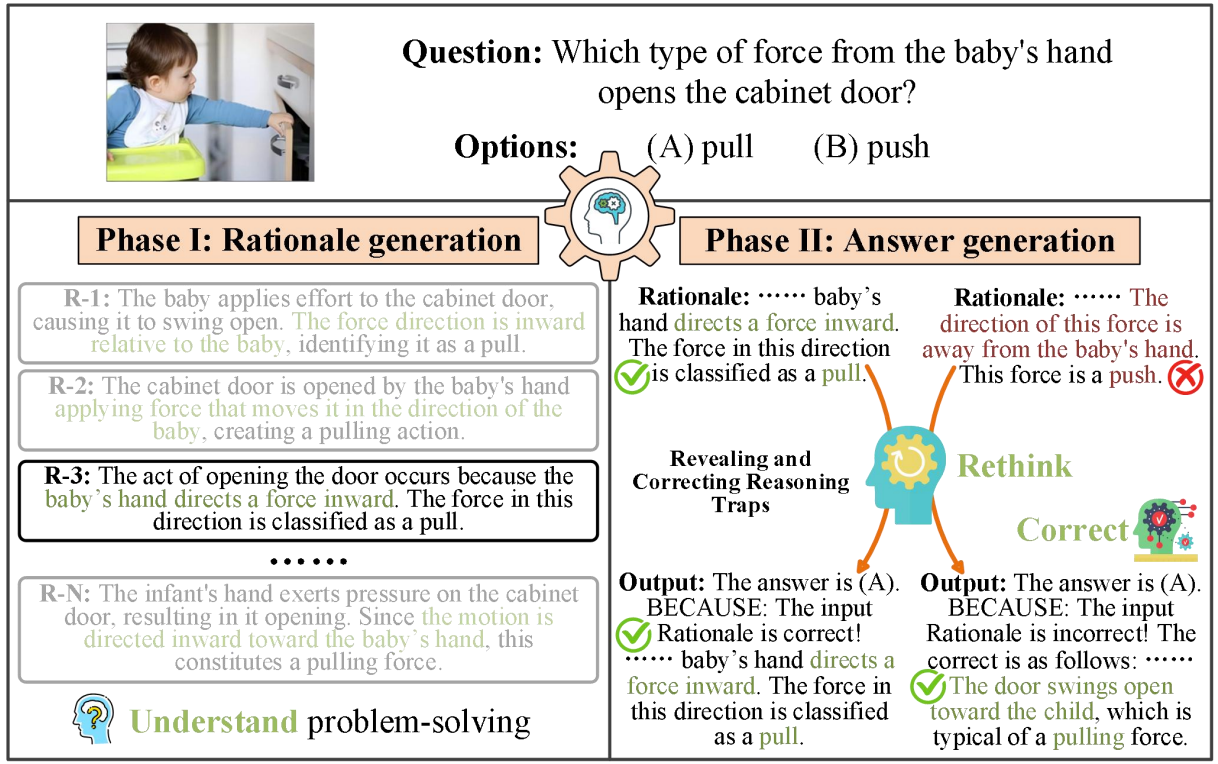
Background

Limitations of existing MCoT methods:

1. limited multi-rationale semantic modeling
2. Insufficient logical robustness
3. Vulnerability to misleading interpretations
4. Insufficient alignment of human cognition.

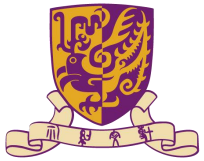
Multi-rationale INtegrated Discriminative (MIND)

"Understand → Rethink → Correct"



Unlike traditional single rationale supervision, our MIND reasoning framework **incorporates diverse positive rationales to model the diversity of human reasoning**, while simultaneously **leveraging challenging negative rationales to reveal potential reasoning pitfalls**.

中国科学院大学
University of Chinese Academy of Sciences



香港中文大學
The Chinese University of Hong Kong

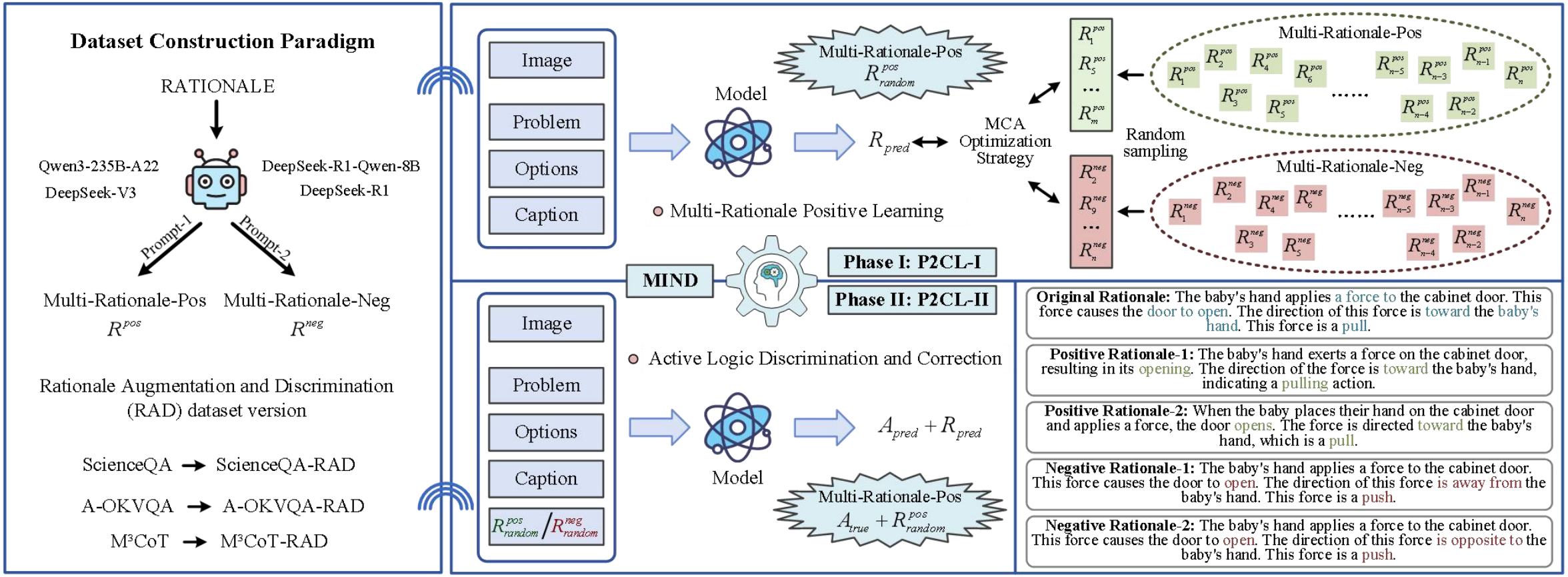
ICML
International Conference
On Machine Learning

第二章节

PART TWO

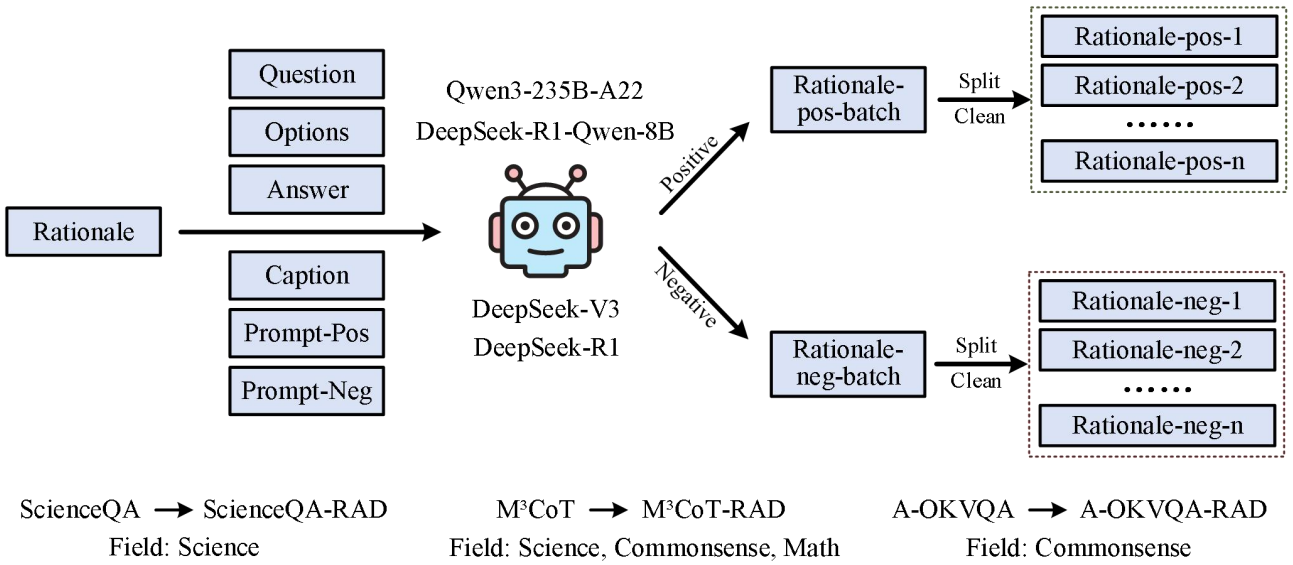
Methods

Method--Overview



We propose a **M**ulti-rationale **IN**tegrated **D**iscriminative (**MIND**) reasoning framework, which is designed to endow MLLMs with human-like cognitive abilities of "**Understand → Rethink → Correct**", and achieves **a paradigm evolution** from passive **imitation-based reasoning** to **active discriminative reasoning**.

Method--RAD Paradigm



Original Rationale: The baby's hand applies a force to the cabinet door. This force causes the door to open. The direction of this force is toward the baby's hand. This force is a pull.

Positive Rationale-1: The baby's hand exerts a force on the cabinet door, resulting in its opening. The direction of the force is toward the baby's hand, indicating a pulling action.

Positive Rationale-2: When the baby places their hand on the cabinet door and applies a force, the door opens. The force is directed toward the baby's hand, which is a pull.

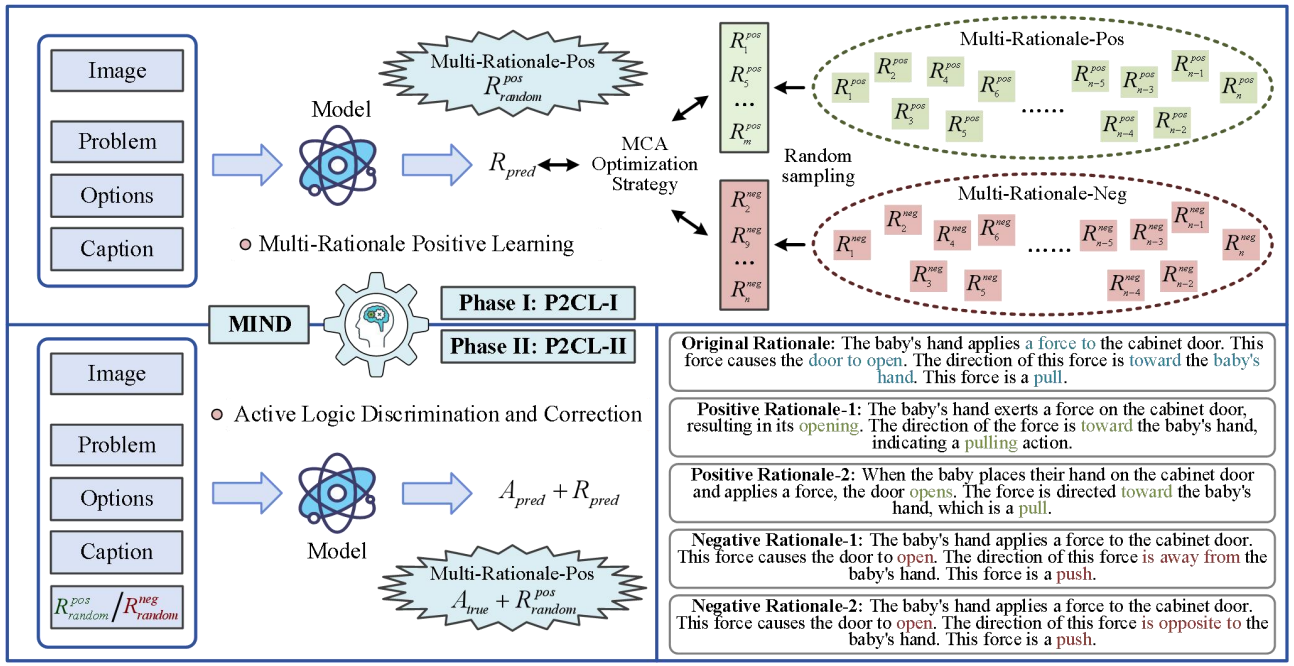
Negative Rationale-1: The baby's hand applies a force to the cabinet door. This force causes the door to open. The direction of this force is away from the baby's hand. This force is a push.

Negative Rationale-2: The baby's hand applies a force to the cabinet door. This force causes the door to open. The direction of this force is opposite to the baby's hand. This force is a push.

Positive Prompts: "{*Task-related information*}"\n You are an intelligent agent with both perception and reasoning abilities. Based on the given context, please make random content adjustments to "{*Solution*}" within a range of 10% to 50% while ensuring that the semantics remain unchanged. Please output {*Repeat number*} different solutions. Each output format is "Adjusted Solution:". Use"\n\n~~~ \n\n" to separate them. The output must be in the required format.

Negative Prompts: "{*Task-related information*}"\n You are an intelligent agent with both perception and reasoning abilities. "{Solution}" is the explanation for the above problem. Based on the given context, please make minor edits to "{*Solution*}" to reverse its meaning and ensure the correct answer cannot be logically derived, while keeping most of the original words and structure intact. Please output {*Repeat number*} different solutions. Each output format is "Negative Solution:". Use"\n\n~~~ \n\n" to separate them. The output must be in the required format.

Method--P2CL and MCA



MCA Optimization Strategy:

$$h^{+/-} = g_\phi\Big(f_\theta(I, Q, O, C, R^{+/-})\Big)$$

$$s_i^+ = \mathrm{sim}(h_{\mathrm{pred}}, h_i^+), \quad s_j^- = \mathrm{sim}(h_{\mathrm{pred}}, h_j^-)$$

$$S_{\mathrm{hard}}^+ = \text{Bottom-}k(\{s_i^+\}), \quad S_{\mathrm{hard}}^- = Top\text{-}k(\{s_i^-\})$$

$$\mathcal{L}_{\mathrm{mca}} = \mathrm{ReLU}(\bar{S}_{\mathrm{hard}}^- + m - \bar{S}_{\mathrm{hard}}^+)$$

P2CL-I: Multi-rationale positive learning

$$\mathcal{L}_{P\text{-}I} = \mathcal{L}_{\mathrm{pos}} + \alpha \cdot \mathcal{L}_{\mathrm{mca}}$$

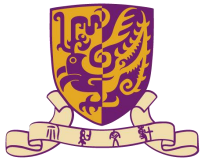
$$\mathcal{L}_{\mathrm{pos}} = -\mathbb{E}_{R^+} \sum_t \log p(\hat{R}_t = R_t^+ \mid I, Q, O, C)$$

P2CL-II: Active logic discrimination and correction

$$\mathcal{L}_{P\text{-}II} = -\mathbb{E}_{R_{\mathrm{cond}}} \sum_t \log p([\hat{A}, \hat{R}_t^+] = [A, R_t^+] \mid Q')$$

P2CL-I: learn the shared causal and logical chains expressed in different linguistic forms, avoid overfitting to a single standard rationale.

P2CL-II: guide the model in identifying and correcting erroneous reasoning, forming a self-reflective and logic repairing reasoning mechanism.



香港中文大學
The Chinese University of Hong Kong

ICML
International Conference
On Machine Learning

第三章节

PART Three

Experiments

Experimental Results

Table 1. Main results (%) on the ScienceQA dataset. Learning = Learning and training methods. Size = backbone model size. NAT = natural science, SOC = social science, LAN = language science, TXT = text context, IMG = image context, NO = no context, G1-6 = grades 1-6, G7-12 = grades 7-12. Red denotes the best result, and blue denotes the second best result.

Model	Learning	Size	NAT	SOC	LAN	TXT	IMG	NO	G1-6	G7-12	**Avg**
Random	-	-	40.28	46.13	29.25	47.45	40.08	33.66	39.35	40.67	39.83
Human	-	-	90.23	84.97	87.48	89.60	87.50	88.10	91.59	82.42	88.40
MCAN [71]	Fine-tune	95M	56.08	46.23	58.09	59.43	51.17	55.40	51.65	59.72	54.54
Top-Down [2]	Fine-tune	70M	59.50	54.33	61.82	62.90	54.88	59.79	57.27	62.16	59.02
BAN [20]	Fine-tune	112M	60.88	46.57	66.64	62.61	52.60	65.51	56.83	63.94	59.37
DFAF [10]	Fine-tune	74M	64.03	48.82	63.55	65.88	54.49	64.11	57.12	67.17	60.72
ViLT [21]	Fine-tune	113M	60.48	63.89	60.27	63.20	61.38	57.00	60.72	61.90	61.14
Patch-TRM [33]	Fine-tune	90M	65.19	46.79	65.55	66.96	55.28	64.95	58.04	67.50	61.42
VisualBERT [27]	Fine-tune	111M	59.33	69.18	61.18	62.71	62.17	58.54	62.96	59.92	61.87
GPT-3.5 [34]	Few-shot	175B	74.64	69.74	76.00	74.44	67.28	77.42	76.80	68.89	73.97
GPT-3.5 w/ coT [59]	Few-shot	175B	75.44	70.87	78.09	74.68	67.43	79.93	78.23	69.68	75.17
ChatGPT w/ coT [1]	Few-shot	175B	78.82	70.98	83.18	77.37	67.92	86.13	80.72	74.03	78.31
GPT-4 w/ coT [1]	Few-shot	-	85.48	72.44	90.27	82.65	71.49	**92.89**	86.66	79.04	83.99
Chameleon (ChatGPT) [35]	Few-shot	-	81.62	70.64	84.00	79.77	70.80	86.62	81.86	76.53	79.93
Chameleon (GPT-4) [35]	Few-shot	-	89.83	74.13	89.82	88.27	77.64	92.13	88.03	83.72	86.54
Gemini+RMR [50, 52]	Few-shot	-	91.79	94.26	89.64	91.40	89.69	91.01	92.84	89.78	91.75
LLaMA-Adapter [73]	Fine-tune	6B	84.37	88.30	84.36	83.72	80.32	86.90	85.83	84.05	85.19
LLaVA [31]	Fine-tune	13B	90.36	95.95	88.00	89.49	88.00	90.66	90.93	90.90	90.92
LaVIN [36]	Fine-tune	7B	89.25	94.94	85.24	88.51	87.46	88.08	90.16	88.07	89.41
UnifiedQA$_{base}$ [19]	Fine-tune	223M	68.16	69.18	74.91	63.78	61.38	77.84	72.98	65.00	70.12
UnifiedQA$_{base}$ w/ coT [34]	Fine-tune	223M	71.00	76.04	78.91	66.42	66.53	81.81	77.06	68.82	74.11
DDCoT$_{base}$ [76]	Fine-tune	223M	88.72	86.84	84.91	87.59	83.34	88.08	88.58	85.10	87.34
MC-CoT$_{base}$ [49]	Fine-tune	223M	91.87	84.59	**93.00**	92.28	88.30	92.75	90.64	90.64	90.64
DPMM-CoT$_{base}$ [13]	Fine-tune	306M	92.72	87.85	89.91	**92.72**	90.48	91.29	91.45	90.11	90.97
Multimodal-T-SciQ$_{base}$ [55]	Fine-tune	223M	91.52	91.45	92.45	91.94	90.33	92.26	92.11	91.10	91.75
Multimodal-CoT$_{base}$ [74]	Fine-tune	223M	84.06	92.35	82.18	82.75	82.75	84.74	85.79	84.44	85.31
MIND$_{base}$ (Ours)	Fine-tune	223M	**93.07**	**96.74**	87.09	92.42	**92.76**	89.27	**92.91**	**91.17**	**92.29**
Improvement	-	-	9.01 ⇑	4.39 ⇑	4.91 ⇑	9.67 ⇑	10.01 ⇑	4.53 ⇑	7.12 ⇑	6.73 ⇑	6.98 ⇑

ScienceQA dataset

Table 3. Main results (%) on the M^{3}CoT dataset. "Random" and "Human" performance are the average accuracy by three attempts. Red denotes the best result, and blue denotes the second best result.

Model	Size	Science			Commonsense			Mathematics			**Total**
		Lang	Natural	Social	Physical	Social	Temporal	Algebra	Geometry	Theory	
Random	-	32.70	30.62	26.71	32.97	22.22	20.33	35.71	27.50	23.81	28.56
Human	-	97.83	92.62	94.31	96.28	92.41	88.71	87.23	88.75	85.71	91.61
Tool-Usage Methods											
HuggingGPT [47]	175B	17.57	20.93	10.33	8.70	14.75	9.76	11.35	22.50	9.52	14.60
VisualChatGPT [60]	>175B	30.09	36.28	7.78	43.48	29.92	33.33	21.99	21.25	28.57	25.92
IdealGPT [70]	-	31.73	31.63	26.23	56.52	50.00	26.83	20.57	30.00	38.10	32.19
Chameleon [35]	-	43.87	26.05	25.44	39.13	37.30	48.78	17.73	26.25	23.81	34.29
Zero-shot Methods											
Kosmos-2 [43]	2B	10.43	28.61	21.18	33.33	17.77	28.46	21.43	21.25	14.29	23.17
InstructBLIP [8]	13B	38.39	30.52	26.27	76.67	70.66	35.77	30.00	22.50	19.05	35.94
LLava-V1.5 [31]	13B	36.97	27.46	20.22	52.22	23.55	27.64	22.86	45.00	4.76	27.05
CogVLM [56]	17B	52.61	37.42	26.91	55.56	54.13	29.27	29.29	32.50	23.81	37.19
Gemini [52]	-	73.93	41.25	31.21	56.67	**71.49**	62.60	30.71	27.50	28.57	45.17
GPT4V [1]	-	**80.09**	54.66	43.95	**87.78**	67.77	**82.11**	42.14	43.75	42.86	56.95
Finetuning Methods											
LLaMA-Adaper [73]	7B	62.56	72.29	30.21	76.92	59.67	72.36	30.71	38.75	38.10	54.89
LLaVA-V1.5 [31]	13B	68.72	72.41	40.86	83.52	64.61	69.11	35.71	45.00	38.10	59.50
CogVLM [56]	17B	65.88	**77.52**	29.09	81.32	65.43	75.61	35.71	46.25	**47.62**	58.25
MC-CoT$_{base}$ [49]	223M	53.55	63.98	43.56	61.54	69.55	29.27	42.86	33.75	28.57	53.51
MC-CoT$_{large}$ [49]	738M	42.65	67.43	**50.56**	58.24	60.49	56.10	**57.86**	**62.50**	14.29	57.69
Multimodal-CoT$_{base}$ [74]	223M	41.71	46.49	39.90	59.34	60.91	27.64	48.57	35.00	28.57	44.85
MIND$_{base}$ (Our)	223M	72.04	63.09	43.31	82.22	65.29	44.72	49.29	61.25	33.33	57.38
Improvement	-	30.33 ⇑	16.60 ⇑	3.41 ⇑	22.88 ⇑	4.38 ⇑	⇑ 17.08	0.72 ⇑	26.25 ⇑	4.76 ⇑	12.53 ⇑
Multimodal-CoT$_{large}$ [74]	738M	45.50	50.19	43.56	63.74	64.61	33.33	40.71	61.25	28.57	48.73
MIND$_{large}$ (Our)	738M	79.62	66.41	48.57	81.11	69.42	51.22	50.71	61.25	**47.62**	**61.56**
Improvement	-	34.12 ⇑	16.22 ⇑	5.01 ⇑	17.37 ⇑	4.81 ⇑	17.89 ⇑	10.00 ⇑	0.00 ⇑	19.05 ⇑	12.83 ⇑

M^{3}CoT dataset

Table 2. Results (%) on the Multiple-Choice task of A-OKVQA dataset. Red marks the best result, and blue the second best.

Model	Learning	*Acc.*	Model	Learning	*Acc.*
CoT [59]	Few-shot	48.1	Pythia [17]	Fine-tune	49.0
Pica [68]	Few-shot	46.1	ViLBERT [32]	Fine-tune	49.1
ClipCap [41]	Few-shot	56.9	LXMERT [51]	Fine-tune	51.4
IPVR (OPT-66B) [5]	Few-shot	48.6	KRISP [38]	Fine-tune	51.9
IPVR (GPT-3) [5]	Few-shot	58.7	GPV-2 [18]	Fine-tune	60.3
Multimodal-CoT$_{base}$	Fine-tune	50.6	**MIND$_{base}$**	Fine-tune	**70.6**

A-OKVQA dataset

Compared to Multimodal-CoT, our MIND improves accuracy by **6.98% - 20.0%** under the same parameters.

Experimental Results

Table S1. Break-Down ablation experiments on ScienceQA, A-OKVQA, and M^{3}CoT datasets. *"Imp."*: Improvement.

Schemes	P2CL	MCA	ScienceQA		A-OKVQA		M^{3}CoT	
			Acc.	*Imp.*	*Acc.*	*Imp.*	*Acc.*	*Imp.*
Baseline	✗	✗	90.29	-	65.85	-	52.67	-
MIND$_{base}$ w/o P2CL	✗	✓	90.36	0.07 ⇑	68.03	2.18 ⇑	55.69	3.02 ⇑
MIND$_{base}$ w/o MCA	✓	✗	92.15	1.86 ⇑	70.13	4.28 ⇑	56.34	3.67 ⇑
MIND$_{base}$ (Ours)	✓	✓	**92.29**	**2.00 ⇑**	**70.57**	**4.72 ⇑**	**57.38**	**4.71 ⇑**

Table 5. Performance analysis of generating rationales using different large models on the ScienceQA dataset.

Large Models	*Acc.*	Large Models	*Acc.*
Baseline	90.29	DeepseekR1-Qwen8B [12]	91.49
Qwen3-235B-22A [65]	91.61	Deepseek-R1 [12]	90.73
DeepseekV3 [29]	90.87	**Final Mix**	**92.29**

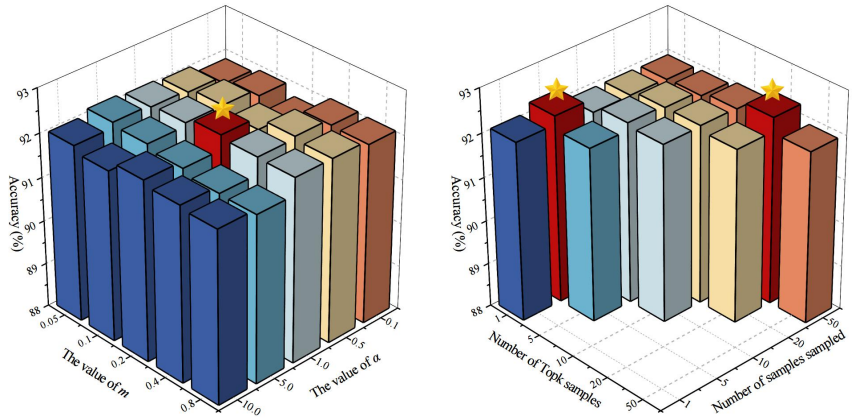
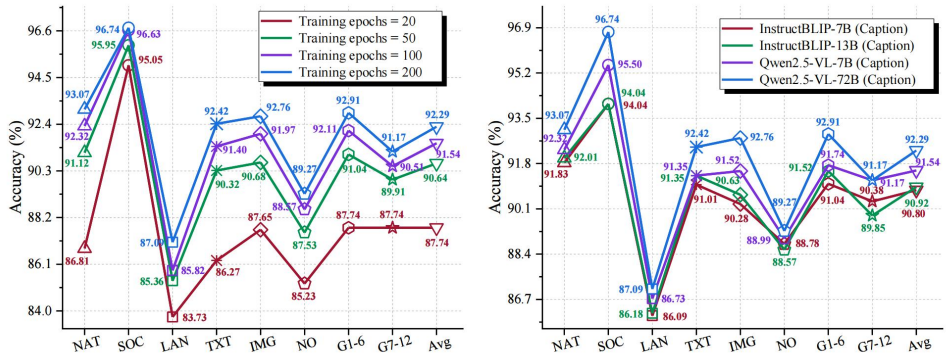


Table 6. Performance analysis of different numbers of rationales on the ScienceQA dataset. "$\times N$" denotes N-fold expansion.

Rationales	*Acc.*	Rationales	*Acc.*
Original (21K)	90.29	×100 (2.1M)	91.58 (1.29 ⇑)
×10 (210K)	91.42 (1.13 ⇑)	×500 (10.5M)	91.91 (1.62 ⇑)
×50 (1.05M)	91.56 (1.27 ⇑)	**×1000 (21M)**	**92.29 (2.00 ⇑)**

Table 7. Performance analysis of visual feature extraction methods on the ScienceQA dataset. **VFE**: Visual feature extraction

VFE Models	*Acc.*	VFE Models	*Acc.*
SAM-Base [22]	90.80	SAM-Huge [22]	91.25
DINOv2-Large [42]	91.06	DINOv2-Giant [42]	91.51
CLIP-B16 [44]	91.51	CLIP-L14-336 [44]	91.98
BLIP-Large [25]	91.49	**BLIP2-flan-t5-xxL** [26]	**92.29**



Experimental Results

Table 8. Performance comparison of MIND with different settings on the ScienceQA dataset. The left and right arrows indicate the injection operations at the input and supervision of P2CL-II, respectively. **Pos:** Positive rationales. **Neg:** Negative Rationales.

Schemes	*Acc.*	Schemes	*Acc.*
MIND w/o P2CL-I	91.63	MIND w/o P2CL-II	90.36
Pos → N/A	91.20	Pos → Pos	91.20
Neg → N/A	90.64	Neg → Pos	91.72
Pos→ N/A or Neg → N/A	91.37	Pos → Pos or Neg → N/A	91.39
Pos→ N/A or Neg → Pos	91.96	**Pos → Pos or Neg → Pos**	**92.29**

"Understand → Rethink → Correct"

1. removing the P2CL-I alone leads to a 0.66% drop.

2. removing the P2CL-II stage alone resulted in a significant performance decrease of 1.93%.

3. When the model outputs only an answer, providing a positive rationale input is more effective than a negative one.

4. Injecting a negative rationale input while supervising with a positive one yields notable gains. Compared to "Neg → N/A", "Neg → Pos" directly improves 1.08% (from 90.64% to 91.72%); compared to"Pos → N/A or Neg → N/A", "Pos→ N/A or Neg → Pos directly improves 0.59% (from 91.37 to 91.96).

Table 9. Evaluation on input perturbation task across multiple datasets (100 test samples per dataset).

Dataset	Methods	Normal	Original Error Type	Plausible but Wrong	Irrelevant Distractor	Incomplete Reasoning	Selective Evidence	Logical Misbinding
ScienceQA	Multimodal-CoT$_{base}$	84.0%	34.0%	43.0%	47.0%	66.0%	35.0%	41.0%
	MIND$_{base}$	**91.0%**	**90.0%**	**91.0%**	**88.0%**	**93.0%**	**89.0%**	**91.0%**
A-OKVQA	Multimodal-CoT$_{base}$	52.0%	16.0%	19.0%	16.0%	37.0%	4.0%	19.0%
	MIND$_{base}$	**68.0%**	**67.0%**	**68.0%**	**67.0%**	**67.0%**	**65.0%**	**69.0%**
M^{3}CoT	Multimodal-CoT$_{base}$	42.0%	25.0%	34.0%	14.0%	46.0%	8.0%	13.0%
	MIND$_{base}$	**55.0%**	**59.0%**	**55.0%**	**59.0%**	**60.0%**	**56.0%**	**60.0%**

Table 10. Compatibility and generalization assessment of MIND. The left and right arrows denote training and testing.

Setting	M^{3}CoT (Train set)–>M^{3}CoT (Test set)			M^{3}CoT (Train set)–>ScienceQA (Test set)		
Models	Qwen2.5-VL-7B	Qwen3-VL-8B	Qwen3.5-9B	Qwen2.5-VL-7B	Qwen3-VL-8B	Qwen3.5-9B
Origin (No SFT)	60.96 (+24.89)	65.83 (+22.31)	70.02 (+18.63)	82.74 (+6.32)	90.57 (+1.18)	90.71 (+1.39)
Baseline (MIND w/o MCA, P2CL)	77.31 (+8.54)	77.74 (+10.40)	75.97 (+12.68)	83.80 (+5.26)	88.16 (+3.59)	85.00 (+7.10)
MIND (Final)	**85.85**	**88.14**	**88.65**	**89.06**	**91.75**	**92.10**

Table 11. Rationale quality assessment on multiple datasets. Total score: 10. The scores are "ScienceQA / A-OKVQA / M^{3}CoT".

Metric	Overall	Correctness	Relevance	Coherence	Solving Process
Multimodal-CoT$_{base}$	8.12 / 6.85 / 3.63	7.49 / 5.49 / 2.22	8.64 / 8.03 / 4.69	8.35 / 7.88 / 4.12	8.00 / 6.00 / 3.47
MIND$_{base}$ (P2CL-I)	8.40 / 7.47 / 5.34	8.01 / 6.48 / **3.94**	8.73 / 8.30 / 6.40	8.55 / 7.99 / 5.85	8.32 / 7.09 / 5.18
MIND$_{base}$ (P2CL-II)	**8.53 / 7.94 / 5.54**	**8.08 / 6.71 /** 3.84	**8.88 / 8.98 / 6.81**	**8.71 / 8.60 / 6.17**	**8.44 / 7.46 / 5.34**

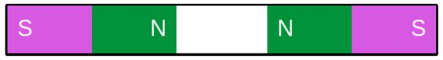
Experimental Results

Question: Will these magnets attract or repel each other?

Options:

A. attract B. repel

ScienceQA #843



Gold Rationale: Will these magnets attract or repel? To find out, look at which poles are closest to each other. The north pole of one magnet is closest to the north pole of the other magnet. Poles that are the same repel. So, these magnets will repel each other.

True Answer: The answer is (B).

Multimodal-CoT

Rationale: Will these magnets attract or repel? To find out, look at which poles are closest to each other. The south pole of one magnet is closest to the south pole of the other magnet. Poles that are different attract. So, these magnets will attract each other.

Answer: The answer is (A).

MIND

Rationale: To decide if the magnets attract or repel, check the poles closest to each other. The north pole of one magnet is near the north pole of the other. Like poles repel, so these magnets will repel.

Answer: The answer is (B). BECAUSE: The input Rationale is correct! To decide if the magnets attract or repel, check the poles closest to each other. The north pole of one magnet is near the north pole of the other. Like poles repel, so these magnets will repel.

Problem: They are likely having pizza at what kind of event?

Options:

A. family B. gaming C. academic D. social

A-OKVQA #17111

Gold Rationale: The chairs and the relative ages of the people in the image is consistent with answer a. The people are mostly about the same age. they are sitting in desks. The men look like students.

True Answer: The answer is (C).

Multimodal-CoT

Rationale: The people are all dressed up and eating pizza. they are probably having a social pizza party. The people are dressed up and eating pizza. They are all dressed up and eating pizza.

Answer: The answer is (D).

MIND

Rationale: The group is eating pizza in an office setting. The pizza is a common sight at social events, but the context suggests a more casual, family-oriented gathering.

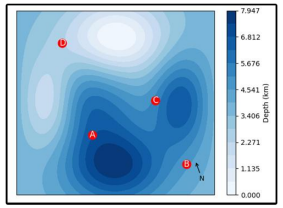
Answer: The answer is (C). BECAUSE: The input Rationale is incorrect! The correct one is as follows: The group is eating pizza in an office. Pizza is a staple at academic gatherings. They are in an office.

Question: Do you know the depth of the point farthest west?

Options:

A. 4.7 km B. 6.8 km C. 2.4 km D. 5.3 km

M³CoT #11139



Gold Rationale:
Step 1: To start, pinpoint the directional marker on the map. The directional marker indicates the cardinal directions (North) and shows the orientation of the map.
Step 2: Identify the westernmost point. The western extremity is represented by A.
Step 3: Once you've located the westernmost point, seek out the contour lines that intersect with it or pass nearby. Map contour lines connect points with matching depths, demonstrating variations in height.
Step 4: Ascertain the Depth.
- Identify the contour line that intersects or is nearest to the westernmost point.
- Read the corresponding color, and compare it with the color in the legend.
- We can roughly gauge that the depth stands at approximately 6.8 kilometers.

Step 5: Therefore, choice "(B) 6.8 km" is the answer.

True Answer: The answer is (B).

Multimodal-CoT

Rationale:
Step 1: First, locate the directional marker on the map. This marker signifies the cardinal directions (North) and indicates the map's orientation.
Step 2: Find the point farthest to the west. We can find that C the westernmost point.
Step 3: Once you've located the westernmost point, seek out the contour lines that intersect with it or pass nearby. Contour lines are lines on the map that connect points of equal depth. They represent changes in depth.
Step 4: Establishing the Depth.
- Locate the contour line that intersects or is nearest to the westernmost point.
- Analyze the associated color and match it with the legend.
- We can make an approximate estimate that the depth is around 6.8 kilometers.

Step 5: Hence, choice \"(B) 6.8 km\" is the answer.

Answer: The answer is (D).

MIND

Rationale:
Step 1: Start by locating the directional marker on the map. This marker shows the cardinal directions (North) and sets the map's orientation.
Step 2: Determine the westernmost point. Point A is the furthest west.
Step 3: After identifying the westernmost point, examine the contour lines that pass through or near it. These lines connect points of equal depth and show depth variations.
Step 4: Estimating the Depth.
- Find the contour line closest to or intersecting the westernmost point.
- Compare the color to the legend.
- The depth is roughly 6.8 km.

Step 5: Hence, the correct choice is \"(B) 6.8 km.\"

Answer: The answer is (B). BECAUSE: The input Rationale is correct!
Step 1: Start by locating the directional marker on the map. This marker shows the cardinal directions (North) and sets the map's orientation.
Step 2: Determine the westernmost point. Point A is the furthest west.
Step 3: After identifying the westernmost point, examine the contour lines that pass through or near it. These lines connect points of equal depth and show depth variations.
Step 4: Estimating the Depth.
- Find the contour line closest to or intersecting the westernmost point.
- Compare the color to the legend.
- The depth is roughly 6.8 km.

Step 5: Hence, the correct choice is \"(B) 6.8 km.\"

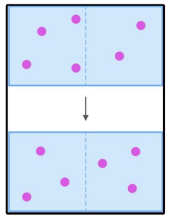
Experimental Results

Question: Complete the text to describe the diagram. Solute particles moved in both directions across the permeable membrane. But more solute particles moved across the membrane (). When there was an equal concentration on both sides, the particles reached equilibrium.

Options:

A. to the left than to the right | B. to the right than to the left

ScienceQA #13736



Gold Rationale: Look at the diagram again. It shows you how the solution changed during the process of diffusion. Before the solute particles reached equilibrium, there were 4 solute particles on the left side of the membrane and 2 solute particles on the right side of the membrane. When the solute particles reached equilibrium, there were 3 solute particles on each side of the membrane. There was 1 more solute particle on the right side of the membrane than before. So, for the solute particles to reach equilibrium, more solute particles must have moved across the membrane to the right than to the left.

True Answer: The answer is (B).

Phase I: Rationale generation

Rationale: The diagram demonstrates diffusion. Initially, the left side had 4 particles, and the right side had 2. At equilibrium, both sides had 4 particles. The right side's gain of 1 particle indicates more movement to the right than to the left.

Phase II: Answer generation

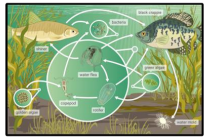
Answer: The answer is (B). BECAUSE: The input Rationale is incorrect! The correct one is as follows: The diagram illustrates diffusion. Initially, 4 particles were on the left and 2 on the right. At equilibrium, both sides had 3 particles, indicating more particles moved to the right.

Question: Which of the following organisms is the decomposer in this food web?

Options:

A. green algae | B. water mold | C. golden algae

ScienceQA #1392



Gold Rationale: Decomposers help break down dead organisms into simpler matter, such as nutrients. These nutrients can then help plants and other organisms grow. In a food web, there is an arrow pointing from another organism to a decomposer. There are no arrows pointing from a decomposer to another organism. The water mold does not have arrows pointing from it to other organisms. So, the water mold is a decomposer. The bacteria do not have arrows pointing from them to other organisms. So, the bacteria are decomposers. The green algae has arrows pointing from it. So, the green algae is not a decomposer. The golden algae has an arrow pointing from it. So, the golden algae is not a decomposer.

True Answer: The answer is (B).

Phase I: Rationale generation

Rationale: Decomposers break down dead organisms into simpler nutrients, supporting plants and other organisms. In food webs, arrows point to decomposers, not from them. Green algae has arrows pointing to it, so it's not a decomposer. Water mold has arrows pointing to it, so it's not a decomposer. Golden algae has arrows pointing to it, so it's not a decomposer.

Phase II: Answer generation

Answer: The answer is (B). BECAUSE: The input Rationale is incorrect! The correct one is as follows: Decomposers break down dead organisms into simpler nutrients, supporting plants and other organisms. In food webs, arrows point to decomposers, but not from them. Green algae has outgoing arrows, so it's not a decomposer. Water mold has no outgoing arrows, so it's a decomposer. Golden algae has outgoing arrows, so it's not a decomposer. Water mold has no outgoing arrows, so it's a decomposer.

Question: What kind of cake is in the bottom row?

Options:

A. red velvet | B. cherry | C. lemon meringue | D. apple

A-OKVQA #17623

Gold Rationale: This kind of cake is known for its texture and color, giving it its name. The cake is red. The coloring of this cake will give it away upon seeing it. it is usually topped with a cream cheese icing.

True Answer: The answer is (A).

Phase I: Rationale generation

Rationale: The cake in the bottom row is a cherry. Its spherical shape and reddish color are clear indicators.

Phase II: Answer generation

Answer: The answer is (A). BECAUSE: The input Rationale is incorrect! The correct one is as follows: The bottom row's cake is red velvet. Its petite size and reddish hue are unmistakable.

Question: What kind of place is this most likely to be?

Options:

A. A motorcycle gang meetup spot
B. A shopping mall
C. A park
D. A restaurant

M³CoT #9153

Gold Rationale: The motorcycles lined up on the street together suggest that this place is a gathering spot for a motorcycle gang.There are many people here. The booths lined up in the background suggest that there are food stalls. Many green trees in the background suggest that this place icould be a park. Based on the inferences, it is most likely that this is a park (C) rather than a motorcycle gang meetup spot, a shopping mall, or a restaurant.

True Answer: The answer is (C).

Phase I: Rationale generation

Rationale: The motorcycles parked in a row on the pavement, indicating a motorcycle-centric setting, and the green stalls positioned nearby suggest a food-serving area. The gathered crowd indicates a social gathering, typical of a motorcycle gang meetup spot. The greenery indicates a green space, which is often used for public events. Therefore, the correct answer is A) A motorcycle gang meetup spot.

Phase II: Answer generation

Answer: The answer is (C). BECAUSE: The input Rationale is incorrect! The correct one is as follows: The presence of motorcycles parked in a row on the pavement indicates a public area, and the green stalls positioned nearby suggest a food-serving area. The crowd gathered indicates a social gathering, typical of a motorcycle gang meetup spot. The greenery indicates a green space, often used for public events. Therefore, the correct answer is C) A park.

University of Chinese Academy of Sciences

香港中文大學
The Chinese University of Hong Kong

ICML
International Conference On Machine Learning

Thanks !

Code

arXiv:2512.05530

Github: https://github.com/YuChuang1205/MIND

Paper