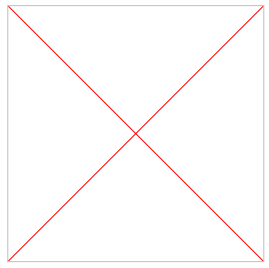
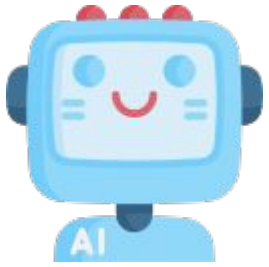
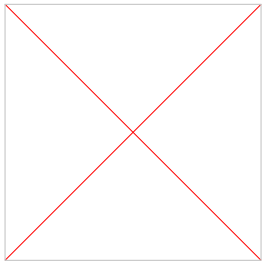
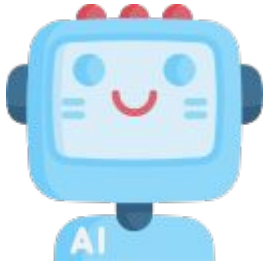
Self-Correction

Self-correction

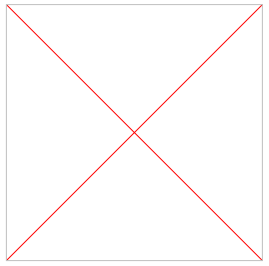


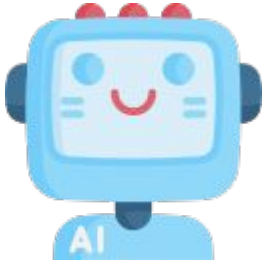


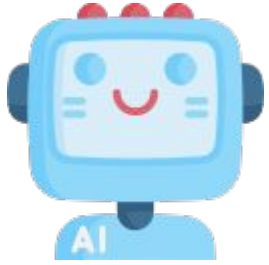


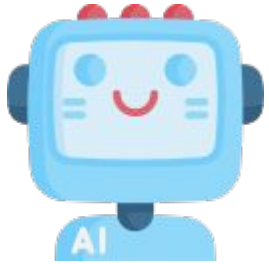


input | output | error detection | corrected output









input | output | error detection | corrected output

相關課程錄影

【生成式AI】ChatGPT 可以自我反省!

https://youtu.be/m7dUFlX-yQI?si=RPtyAmh3-2ICrgmW

(2023)

相關課程錄影

【生成式AI時代下的機器學習(2025)】第七講：DeepSeek-R1 這類大型語言模型是如何進行「深度思考」（Reasoning）的？

https://youtu.be/bJFtcwLSNxI?si=EZyt2nPudfrq1LOv

Outline

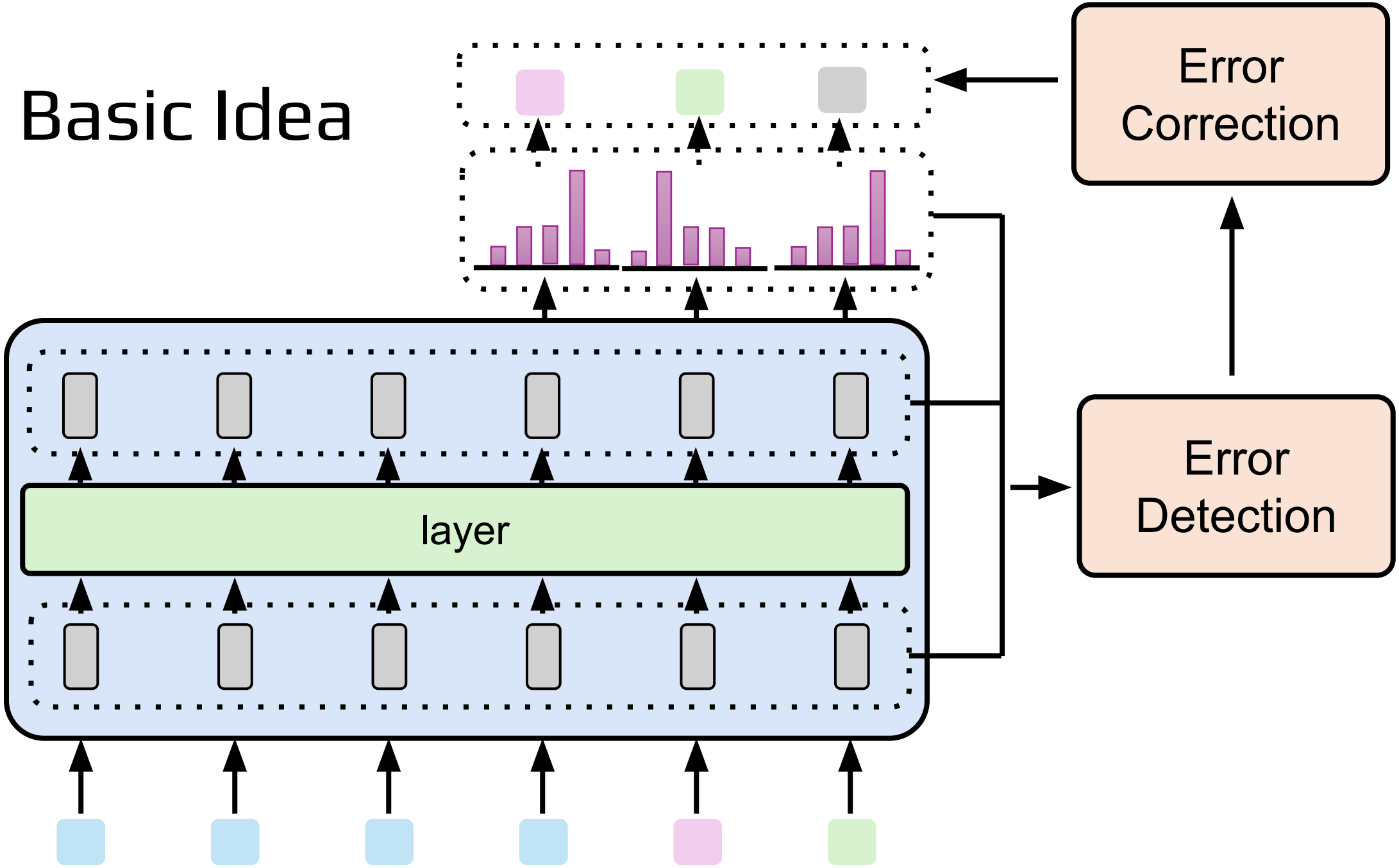
修改 Inference 過程

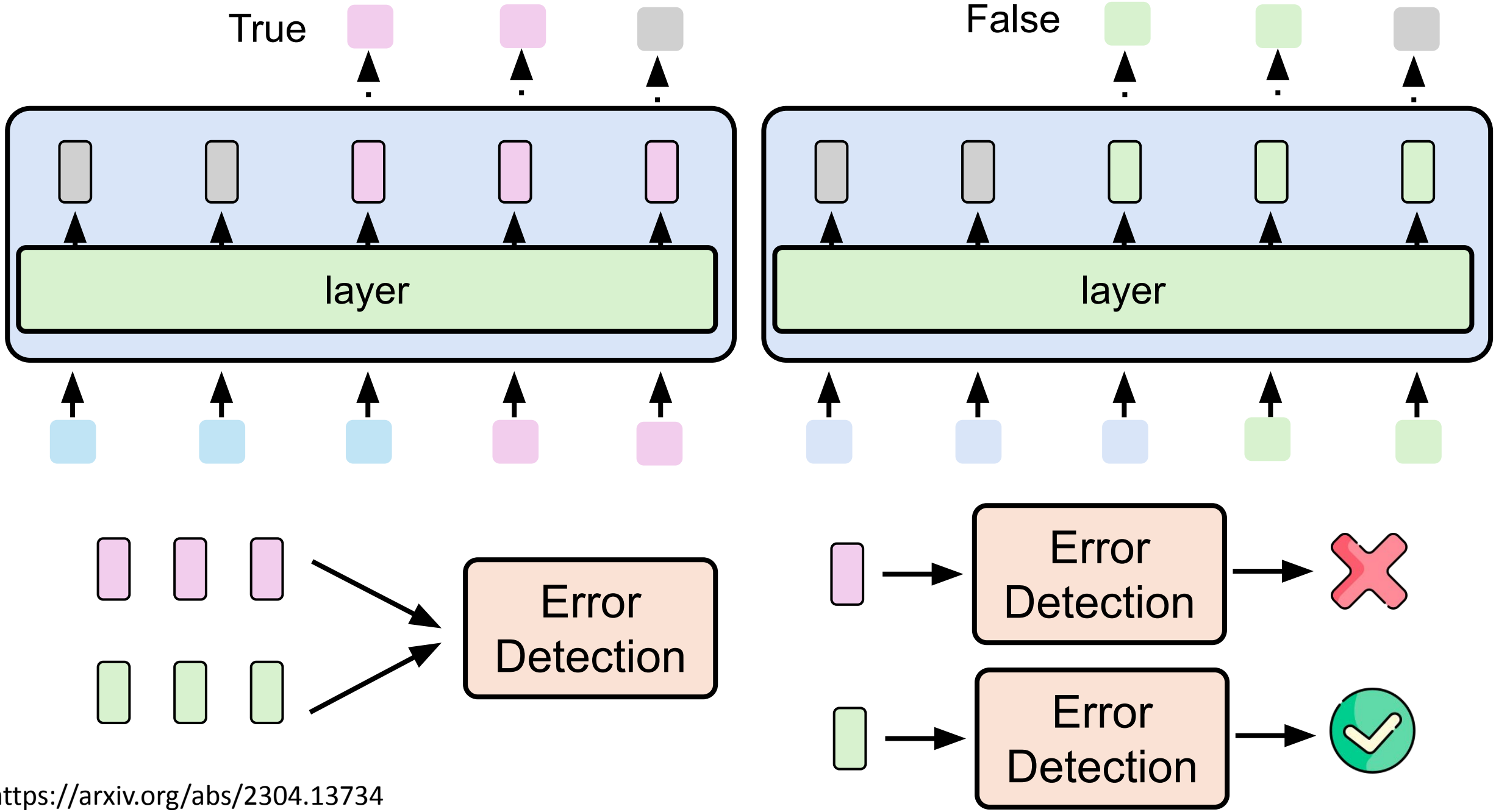
修改 Harness (Workflow)

修改 Model Parameters (Reasoning)

修改 Inference 過程

Basic Idea

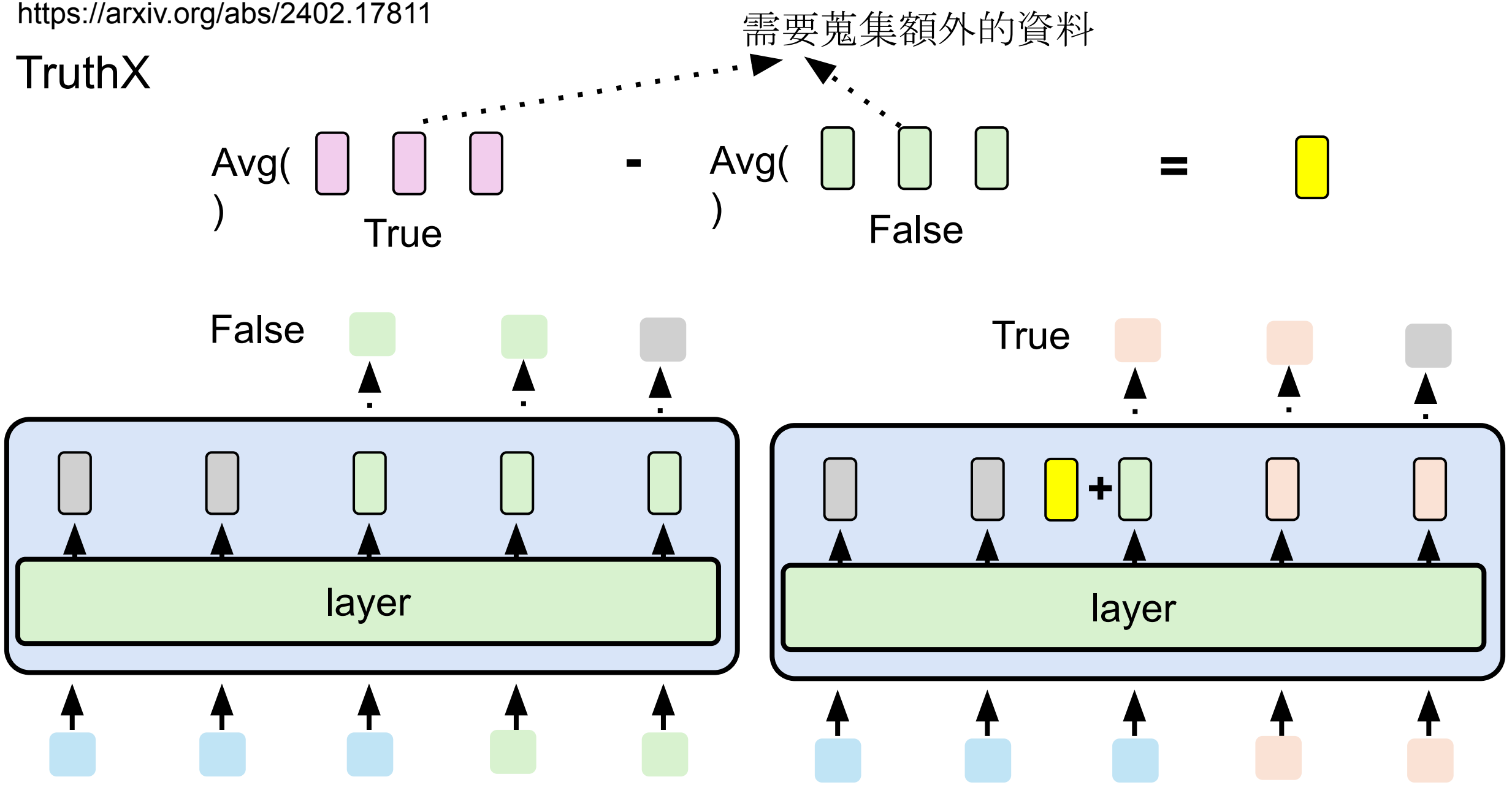




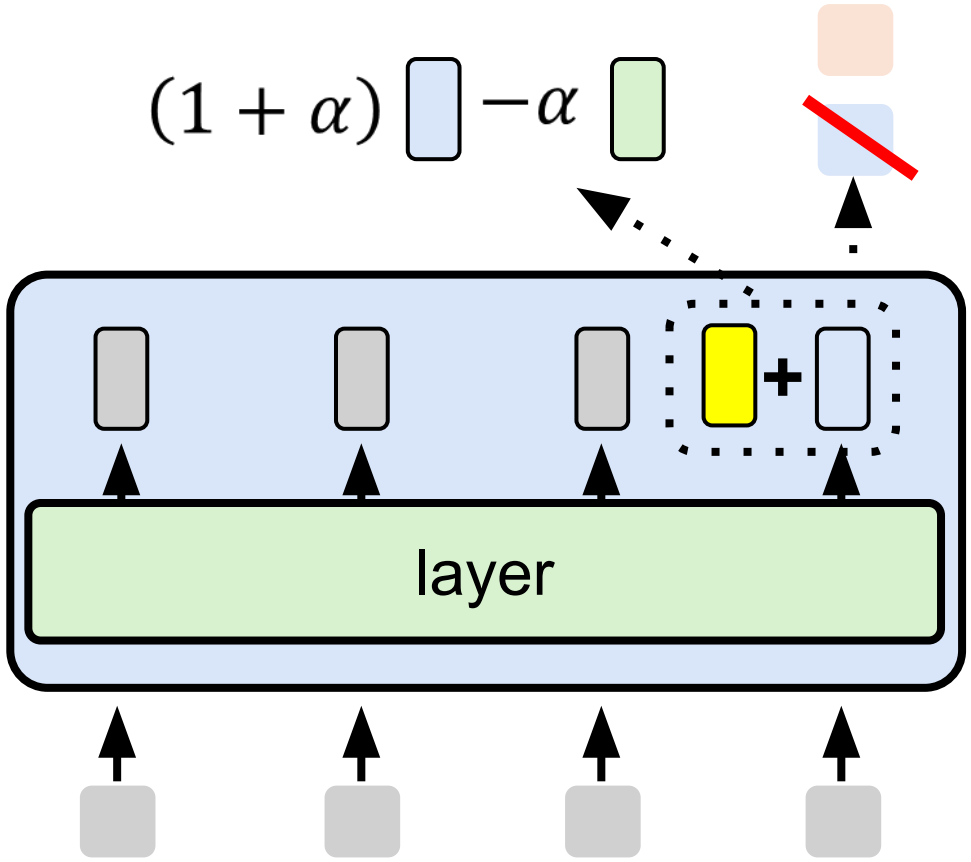
https://arxiv.org/abs/2304.13734

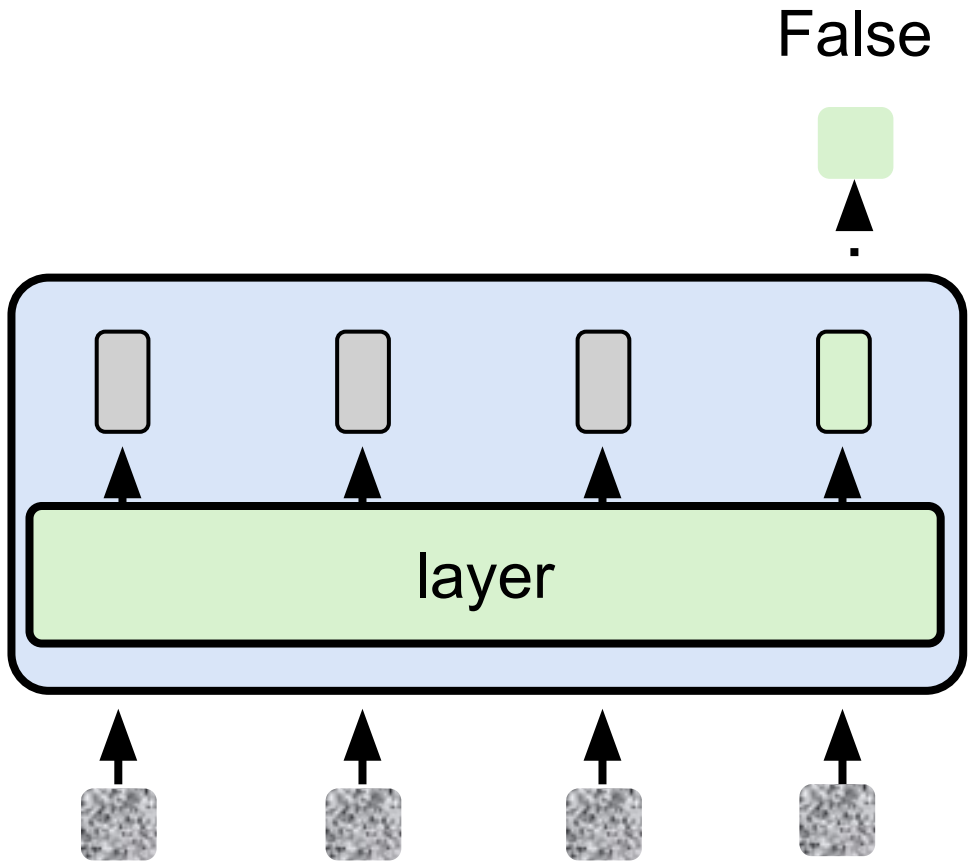
https://arxiv.org/abs/2402.17811

TruthX



Contrastive Decoding



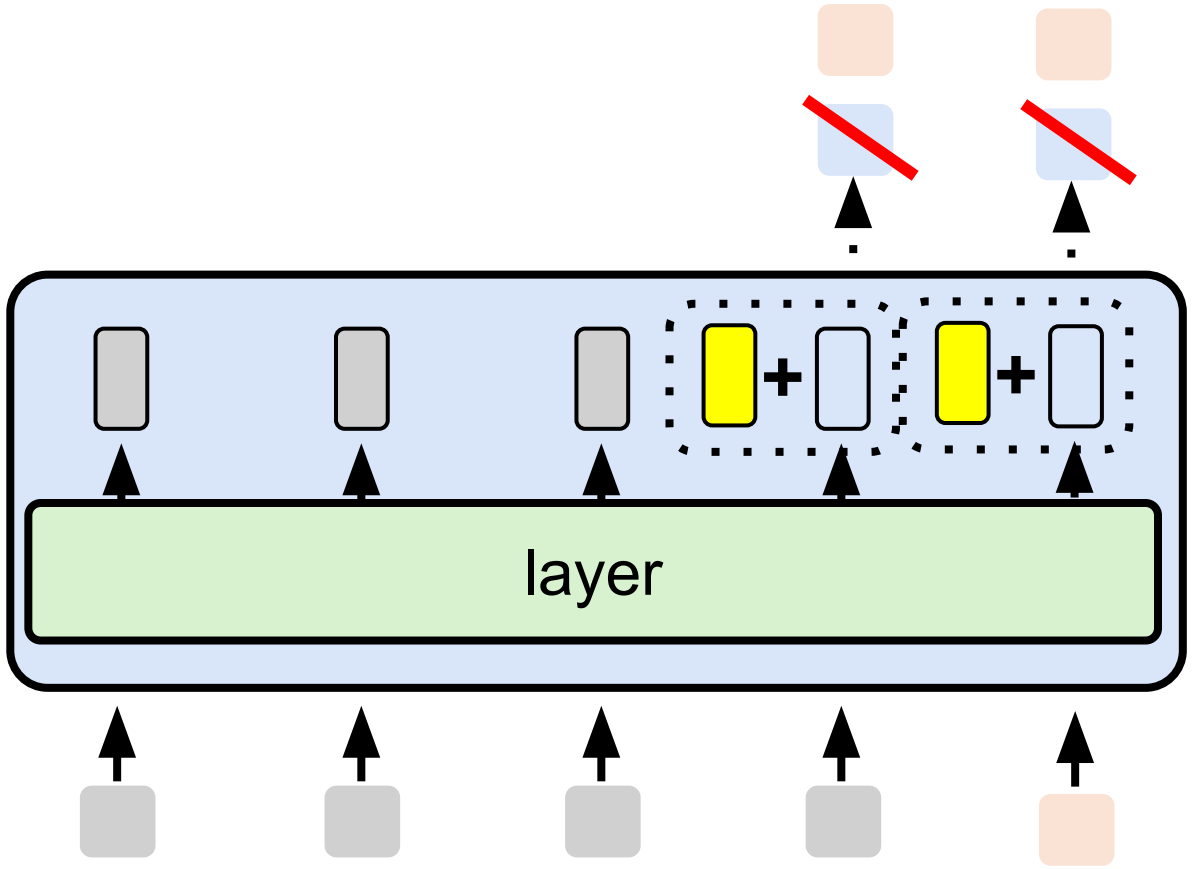


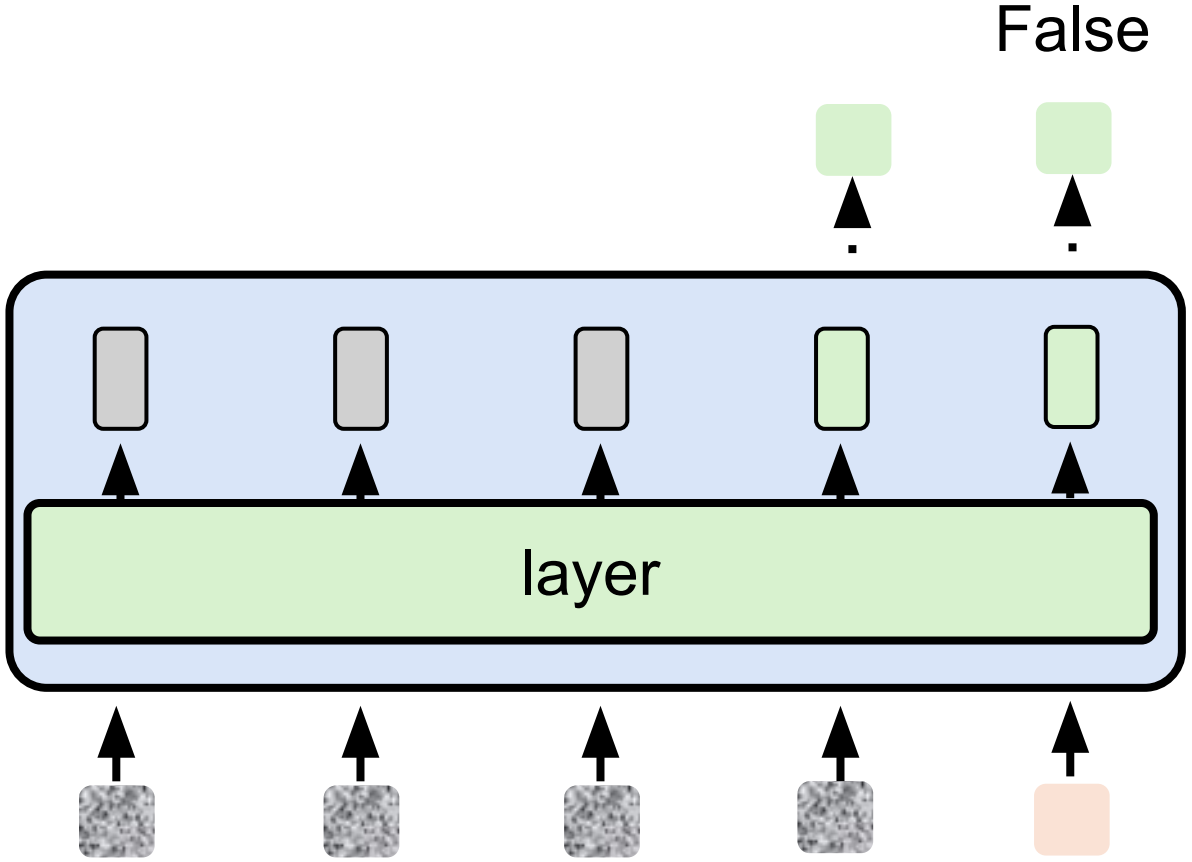
$\alpha($ - $) =$

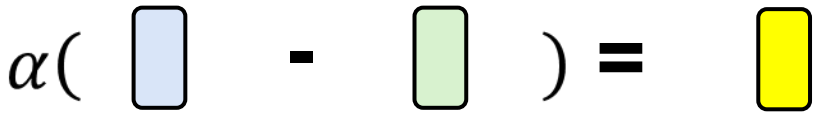
製造出可能會答錯的狀態 (?)

Contrastive Decoding

優點:沒有改變模型參數
缺點:額外運算

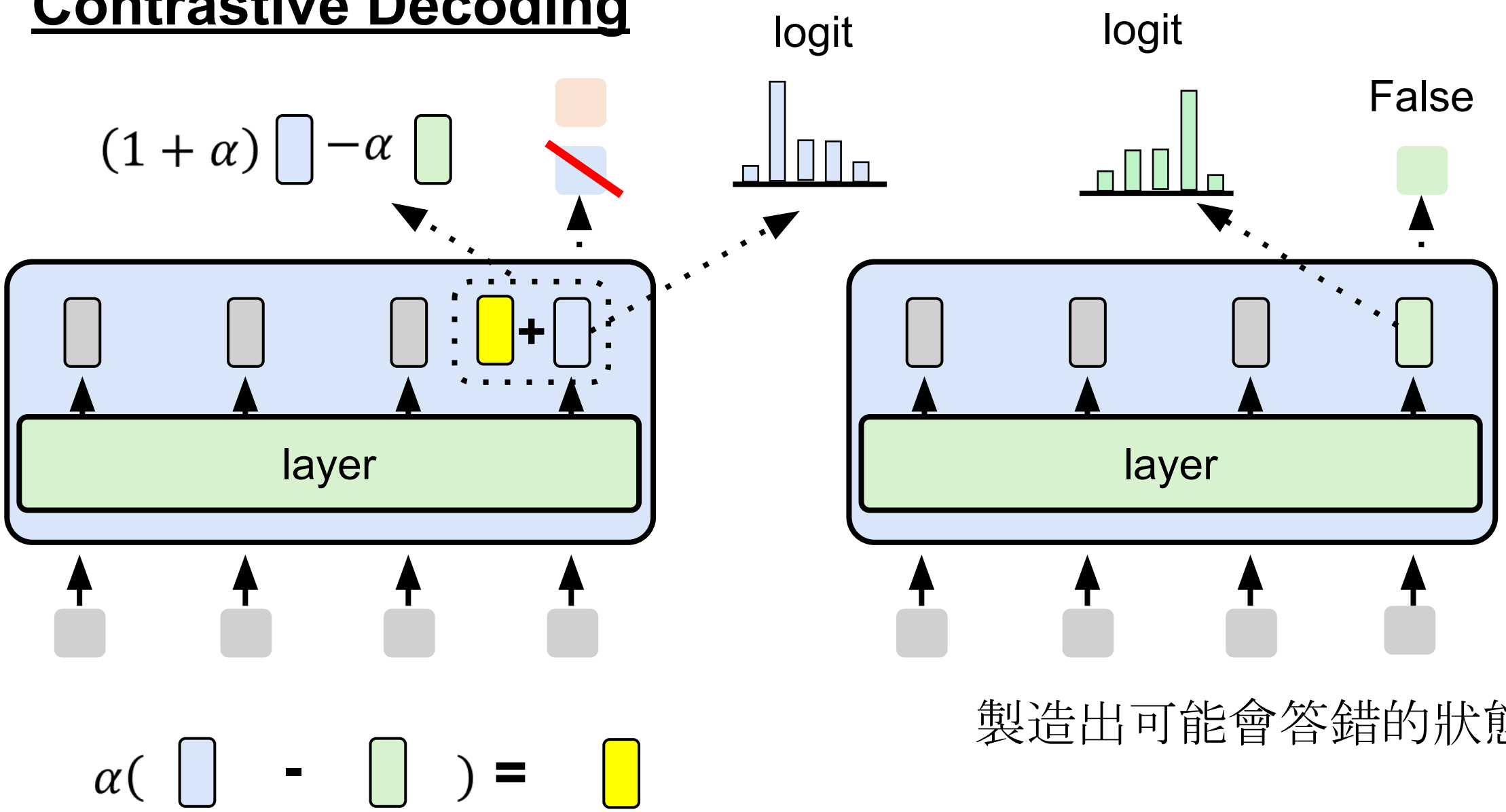






製造出可能會答錯的狀態 (?)

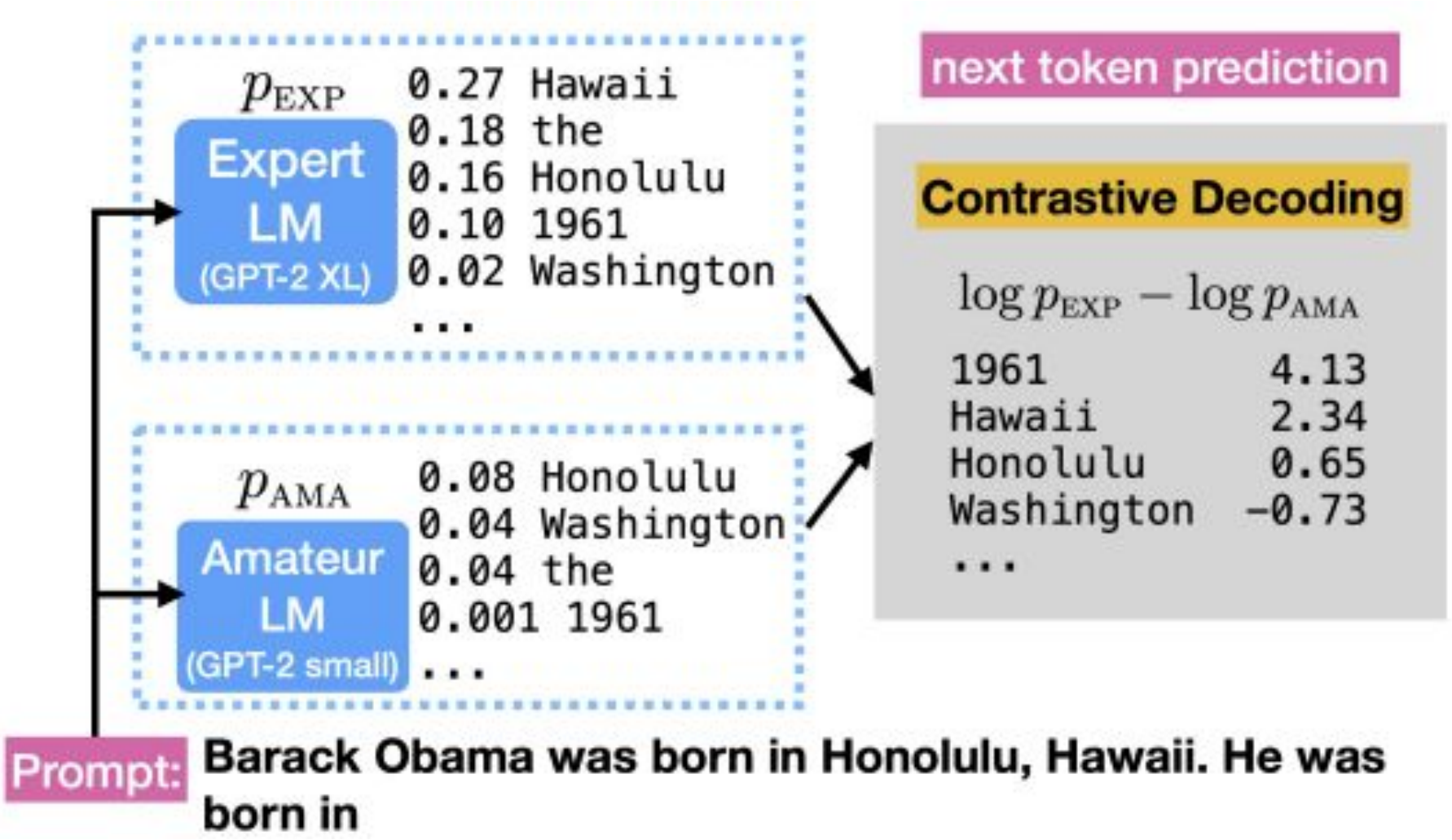
Contrastive Decoding



製造出可能會答錯的狀態 (?)

Contrastive Decoding

https://arxiv.org/abs/2210.15097

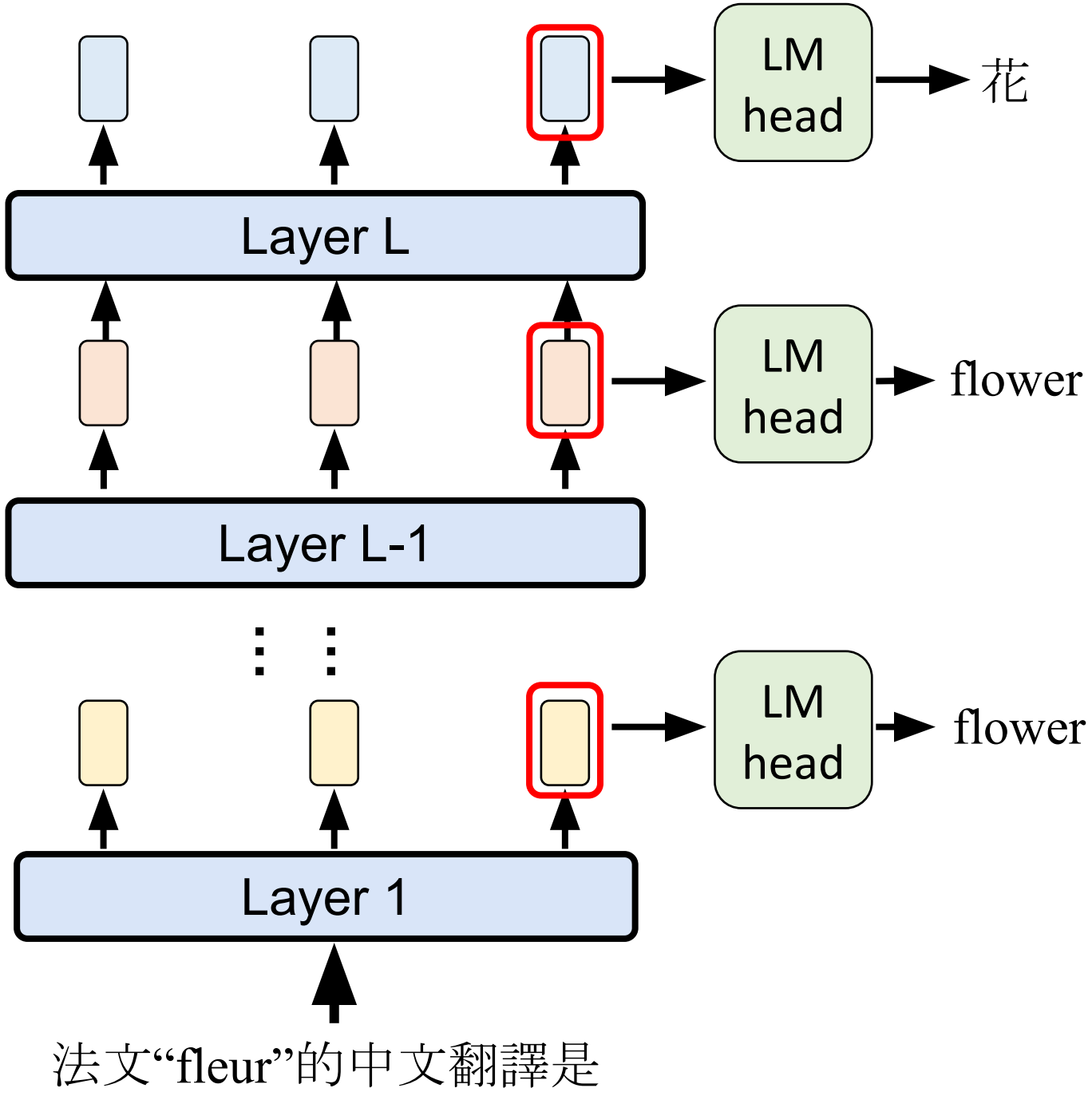


Decoding by Contrasting Layers (DoLa)

https://arxiv.org/abs/2309.03883

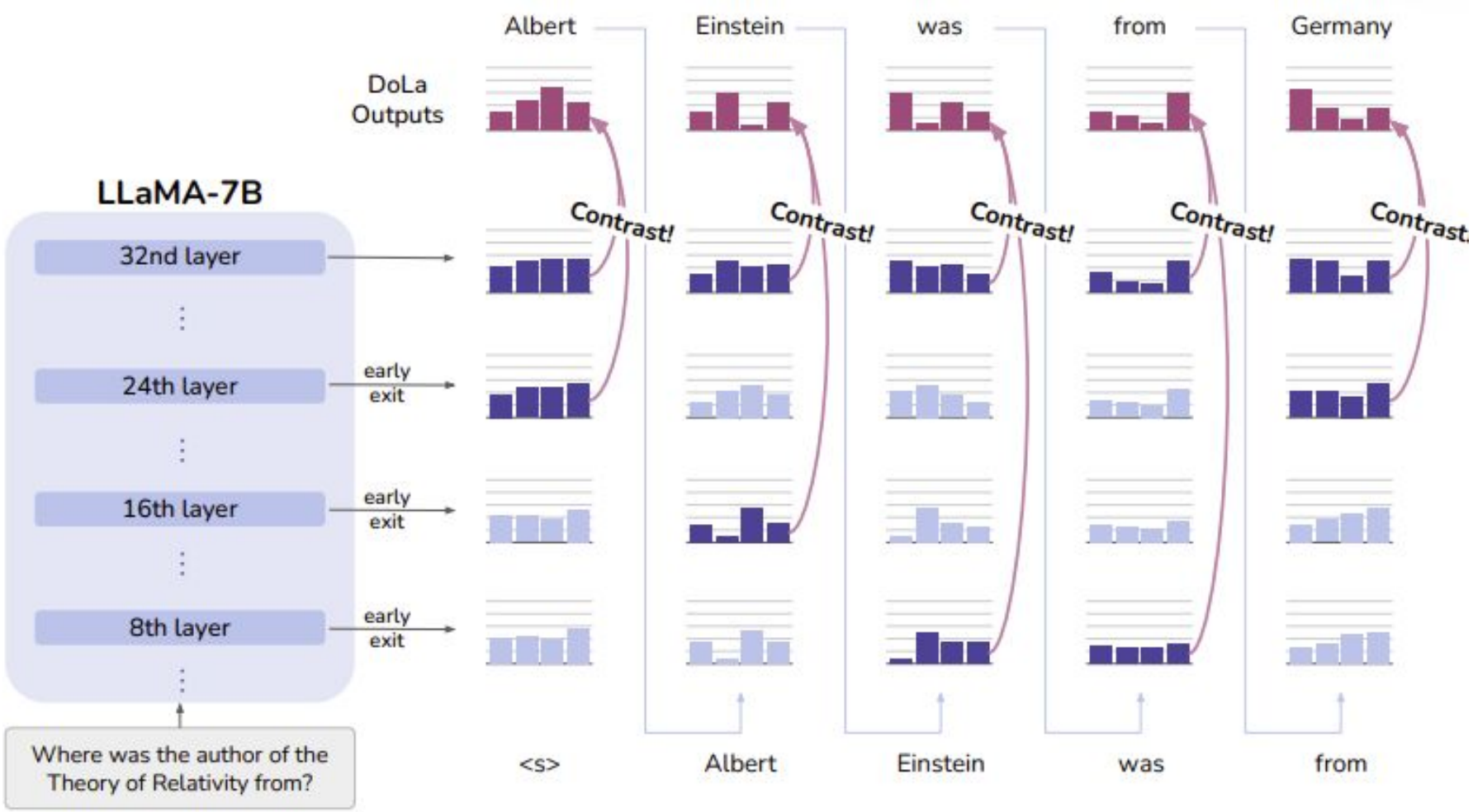
Logit lens

https://arxiv.org/abs/2001.09309
https://www.lesswrong.com/posts/AcKRB8wDpdaN6v6ru/interpreting-gpt-the-logit-lens



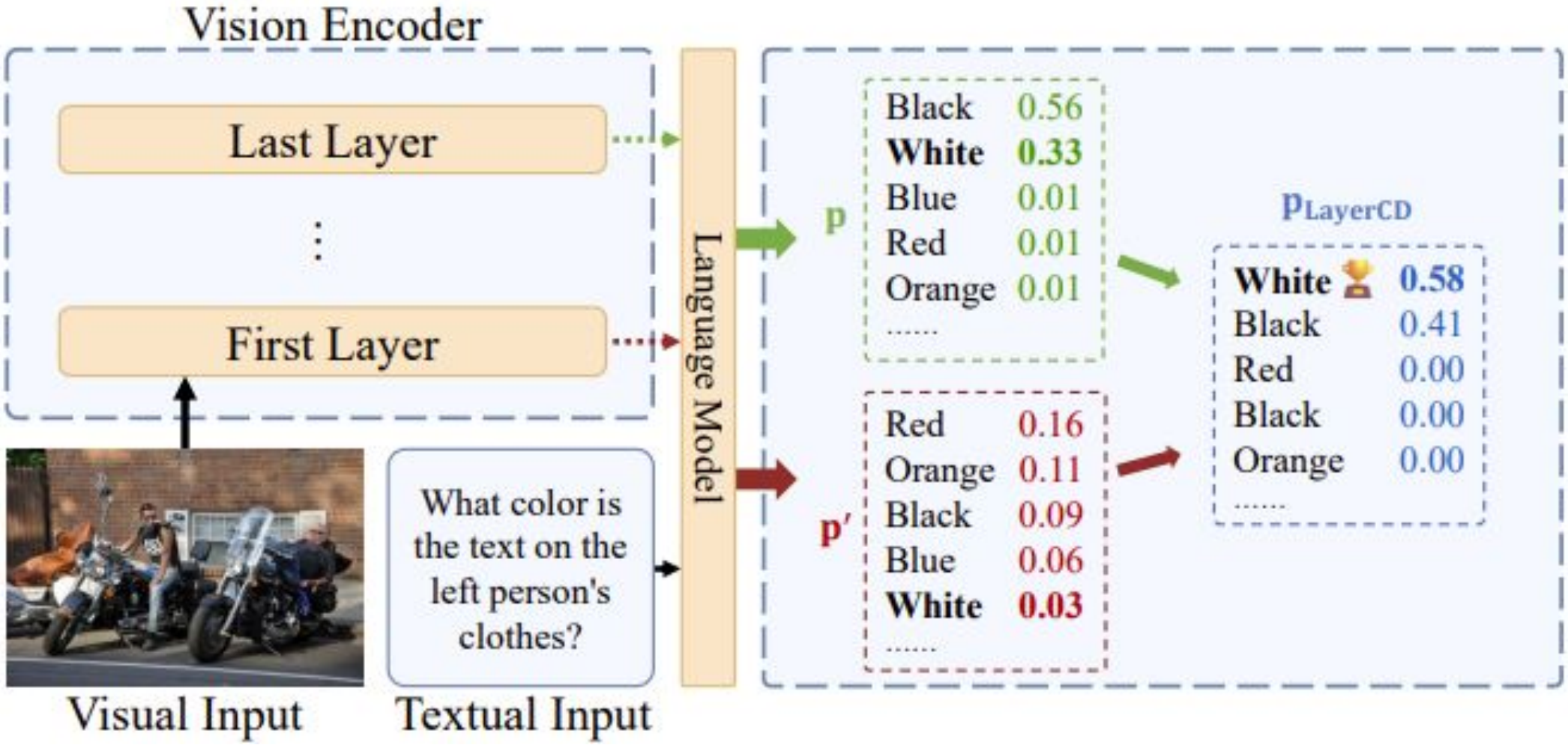
Decoding by Contrasting Layers (DoLa)

https://arxiv.org/abs/2309.03883



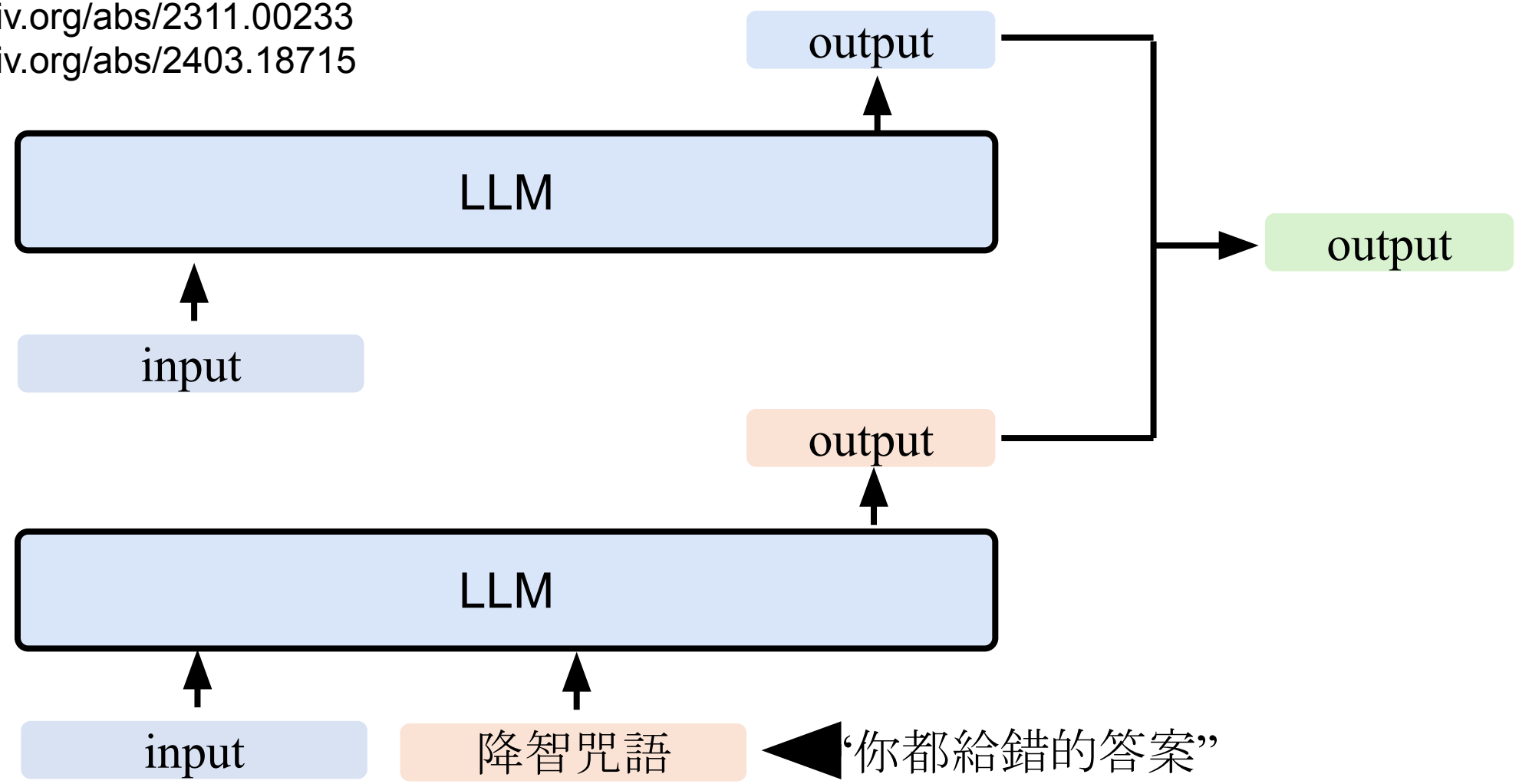
Layer Contrastive Decoding (LayerCD)

https://arxiv.org/abs/2509.25177

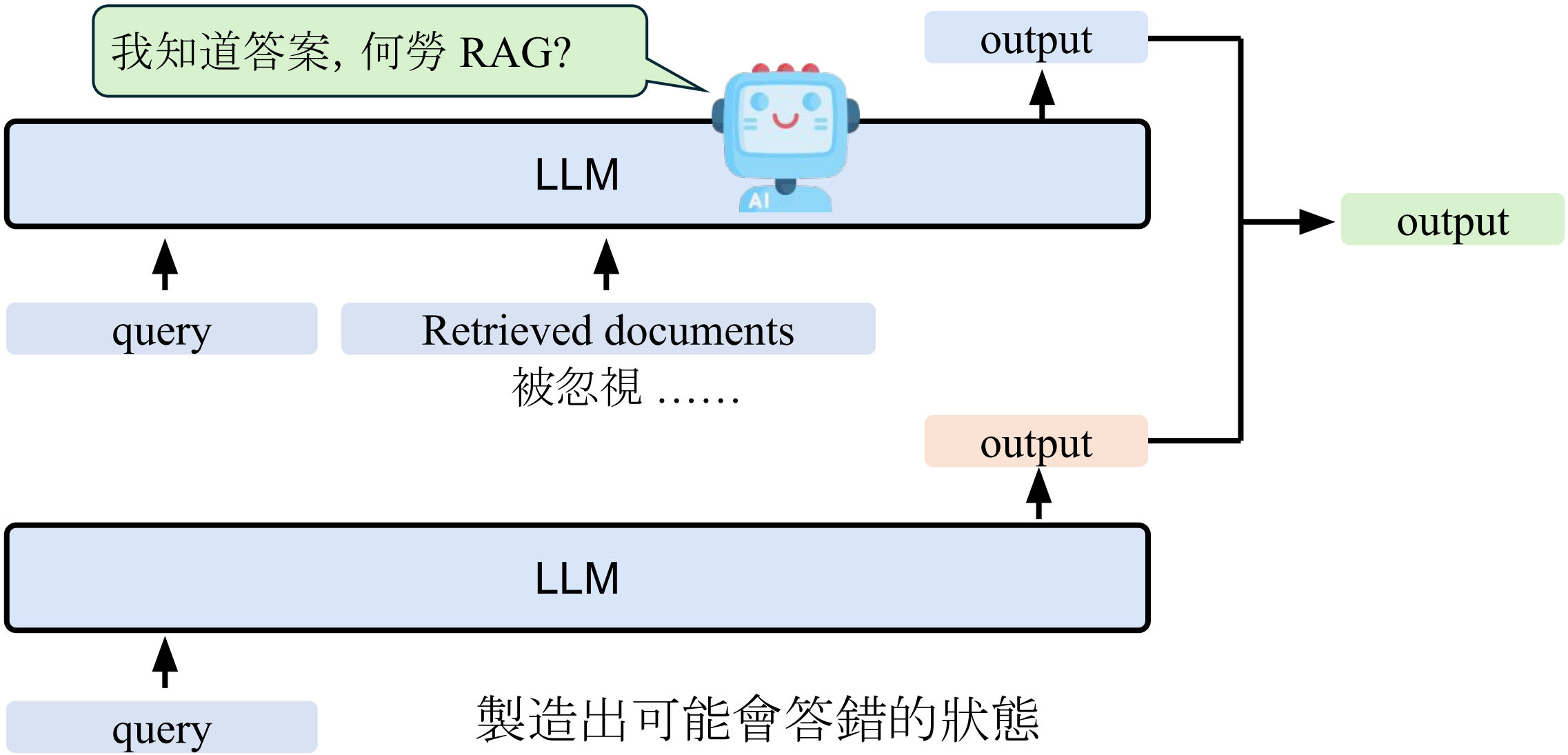


Instruction Contrastive Decoding (ICD)

https://arxiv.org/abs/2311.00233
https://arxiv.org/abs/2403.18715

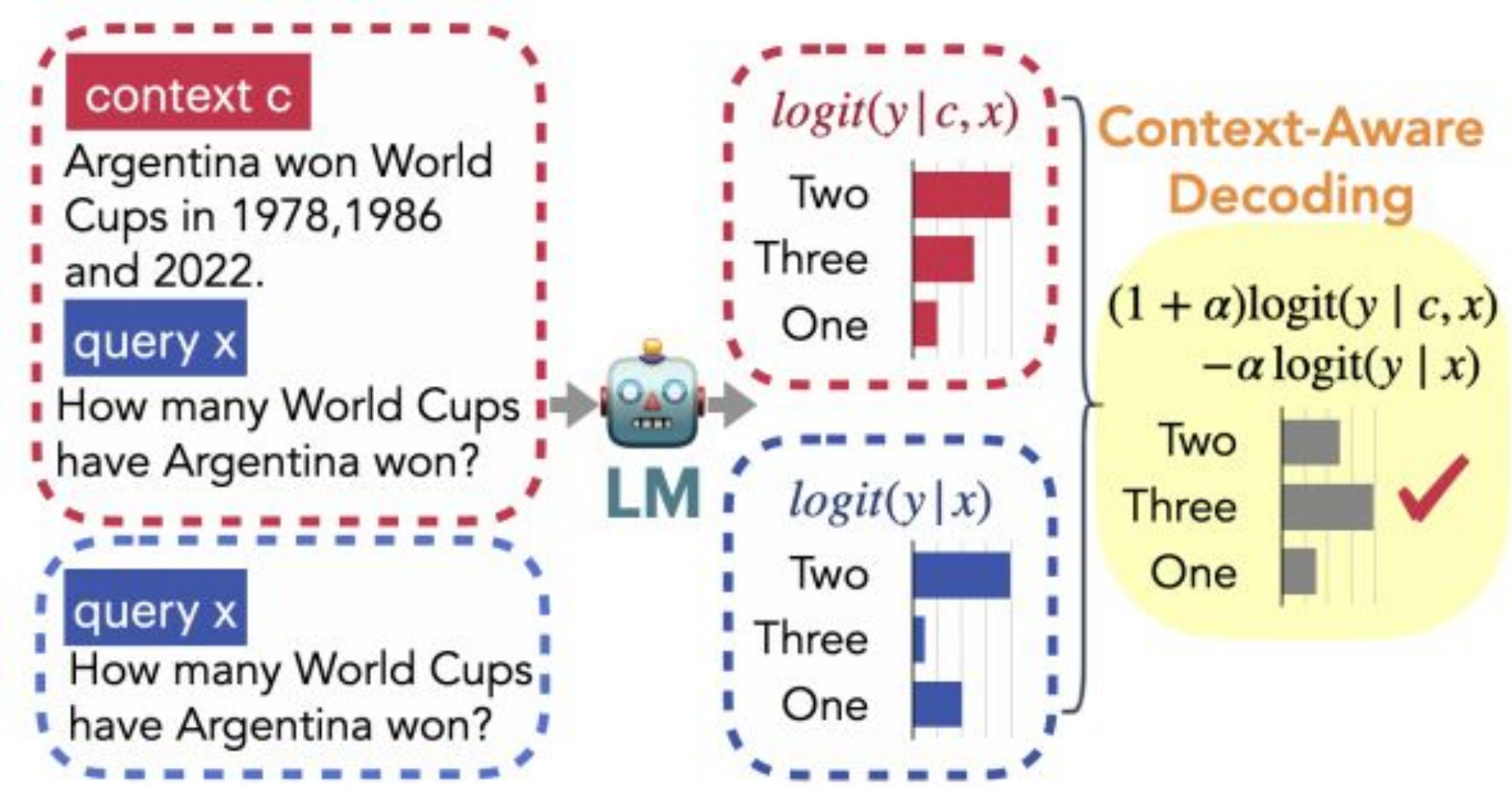


Context-aware Decoding (CAD)

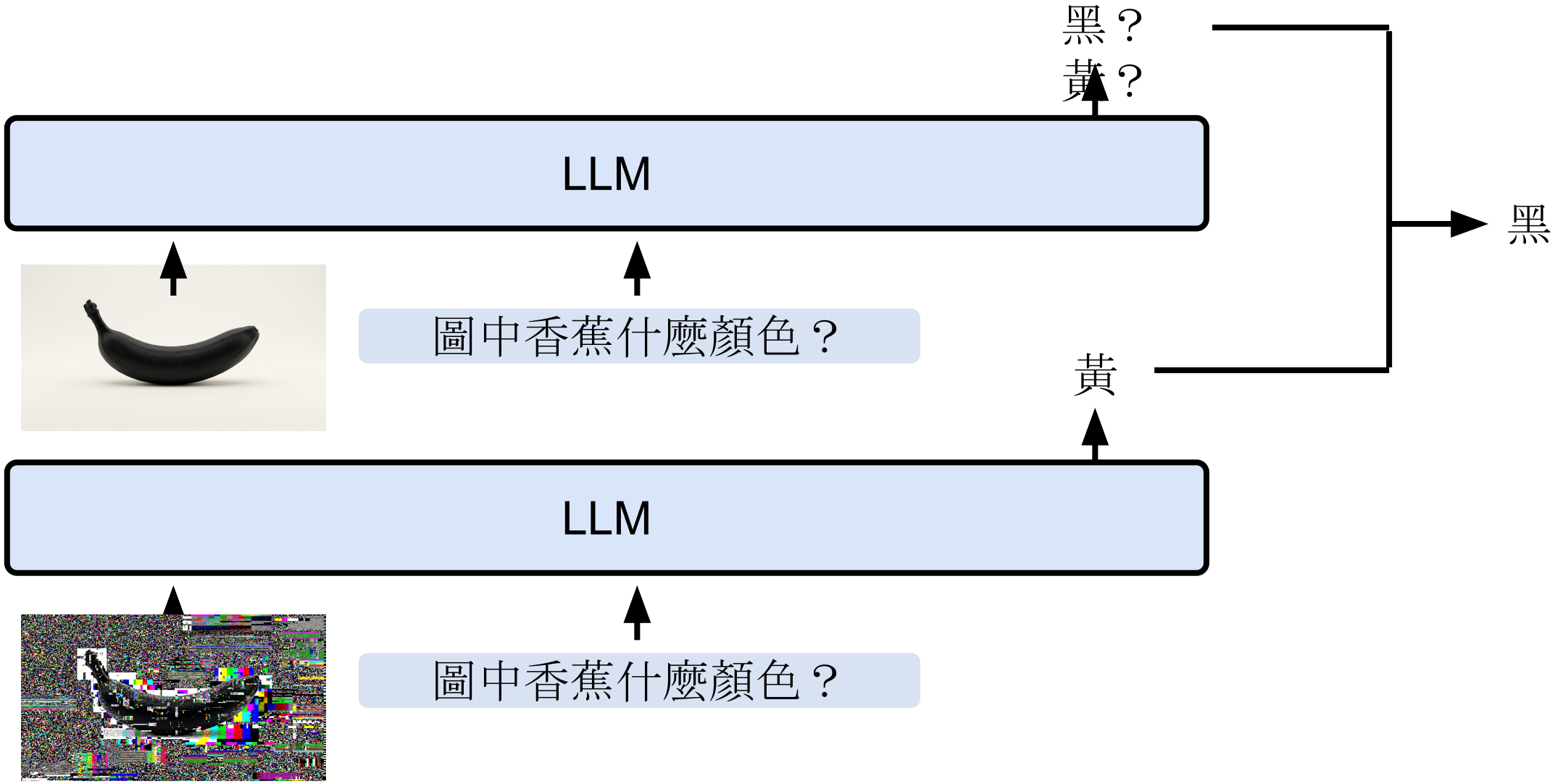


Context-aware Decoding (CAD)

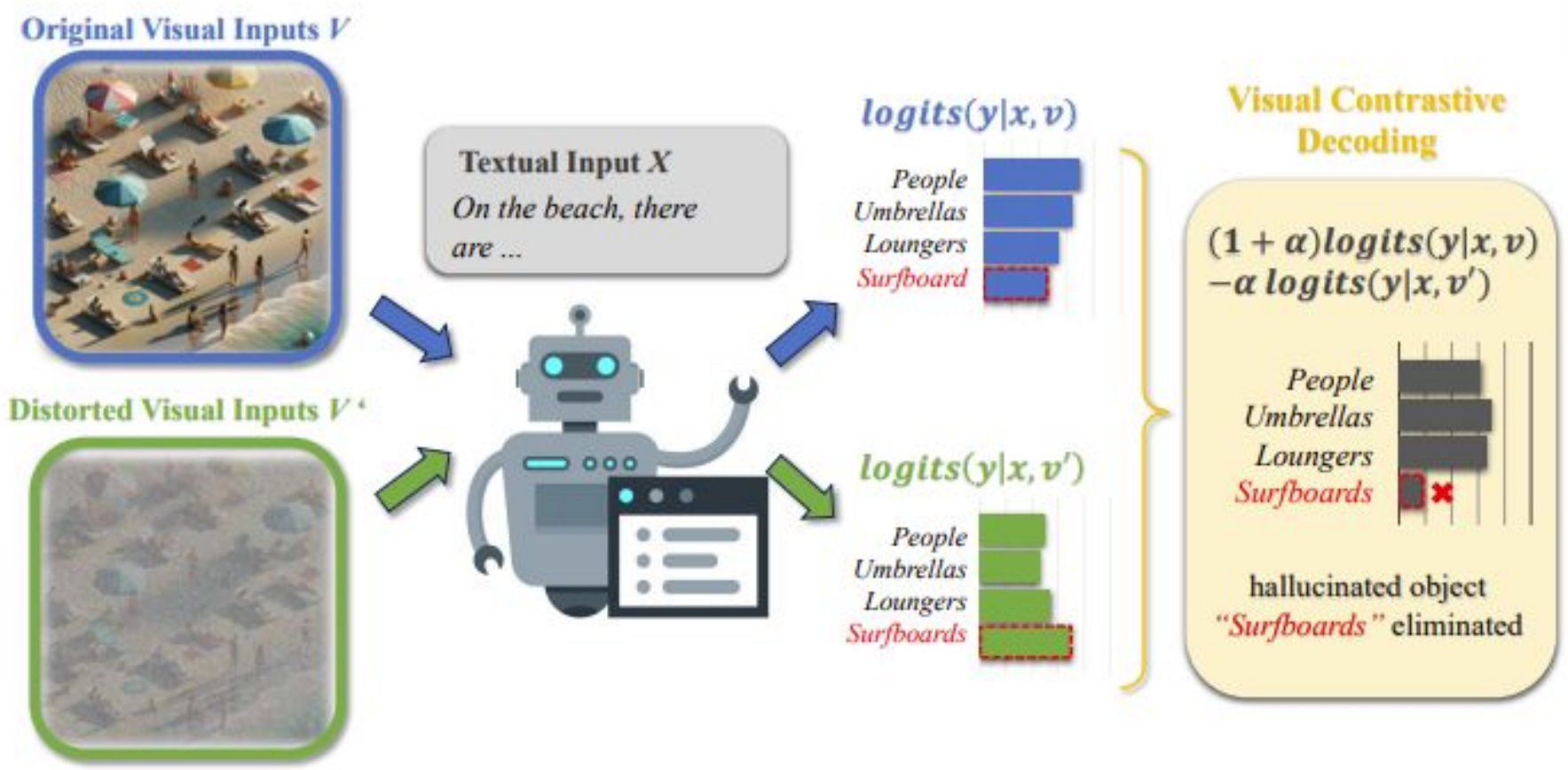
https://arxiv.org/abs/2305.14739



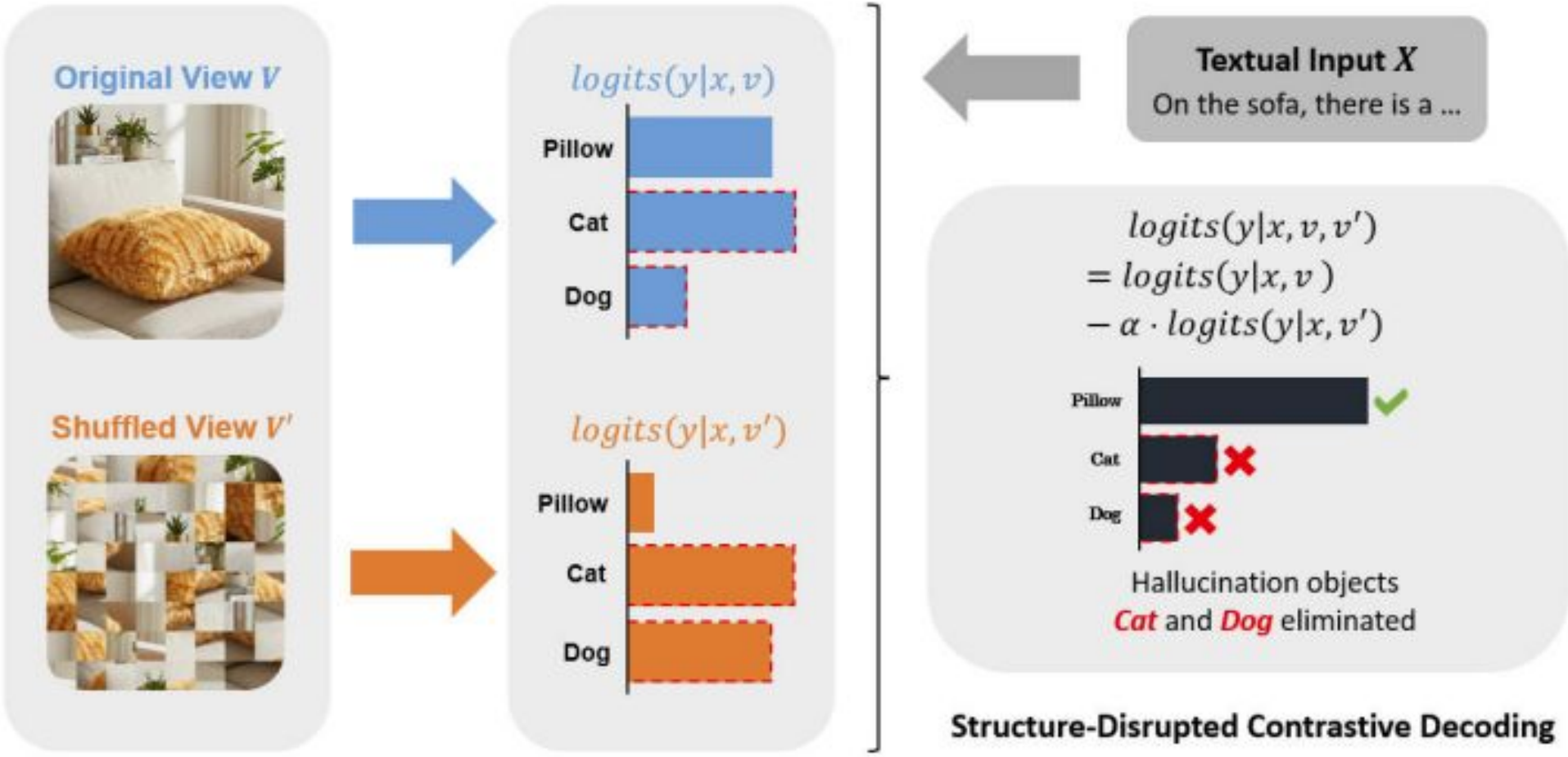
Context-aware Decoding for Image



Context-aware Decoding for Image



https://arxiv.org/pdf/2311.16922



https://arxiv.org/abs/2601.03500

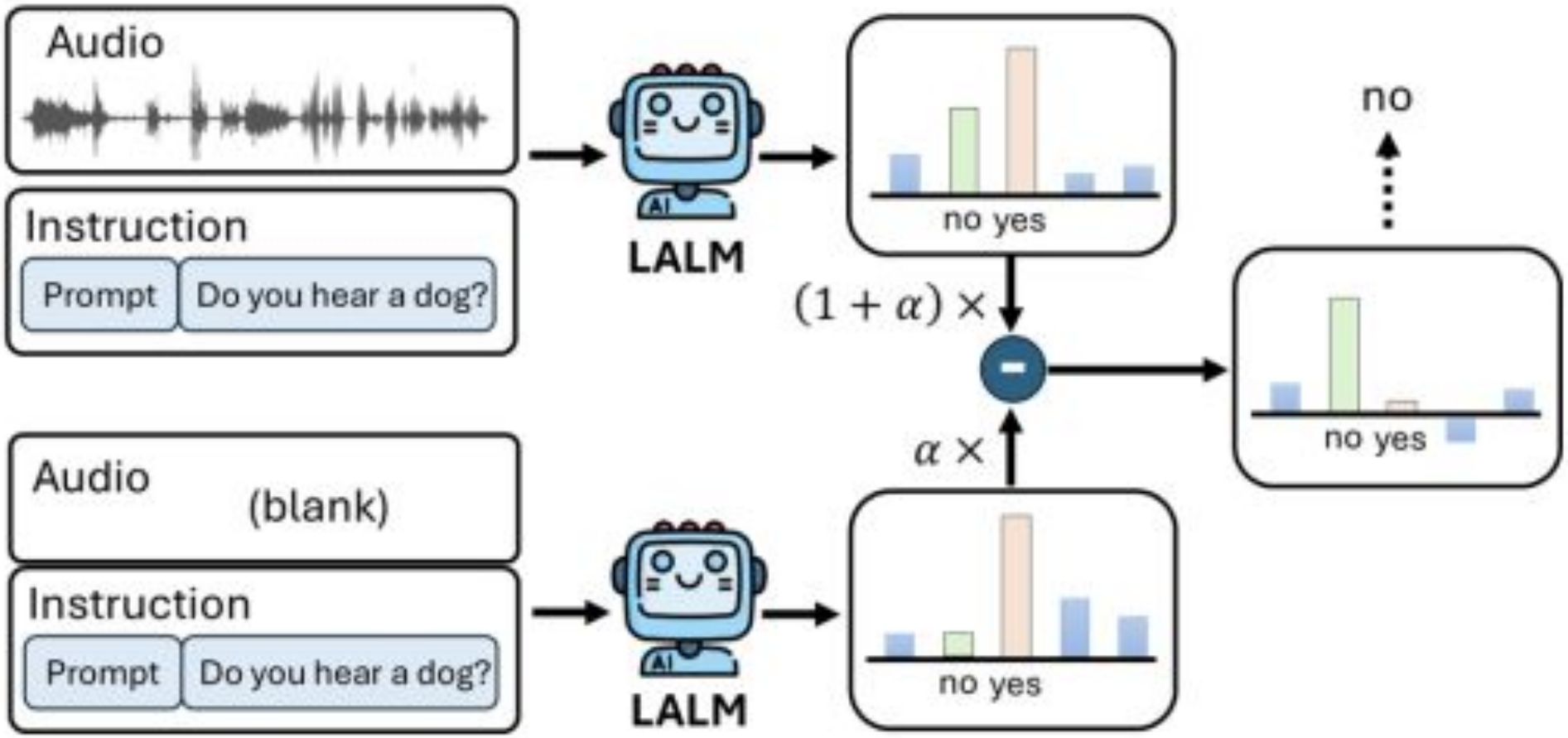
Remove the most salient visual evidence

https://arxiv.org/abs/2602.11737

Audio-Aware Decoding

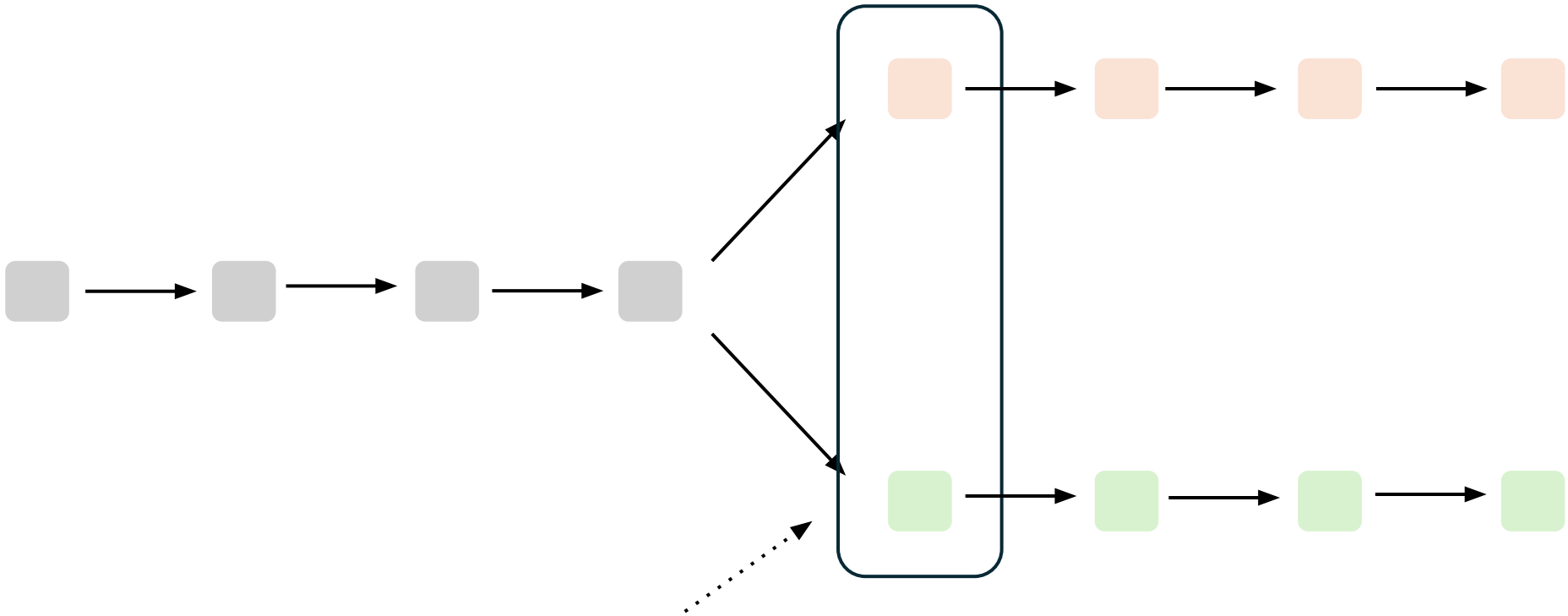
More comparison
https://arxiv.org/abs/2603.09232
https://arxiv.org/abs/2603.14636

https://arxiv.org/pdf/2506.07233

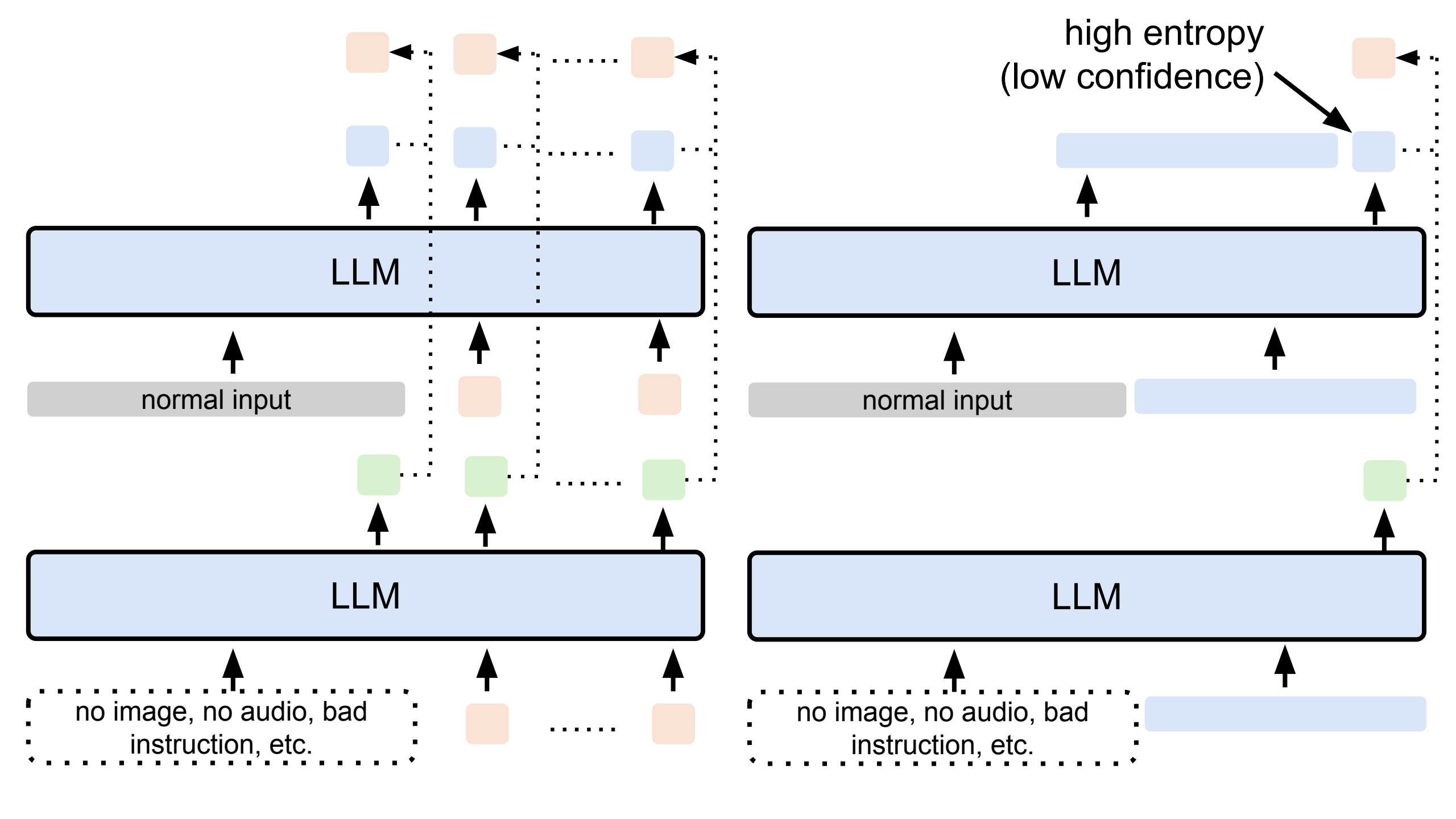


Minimal Test-Time Intervention (MTI)

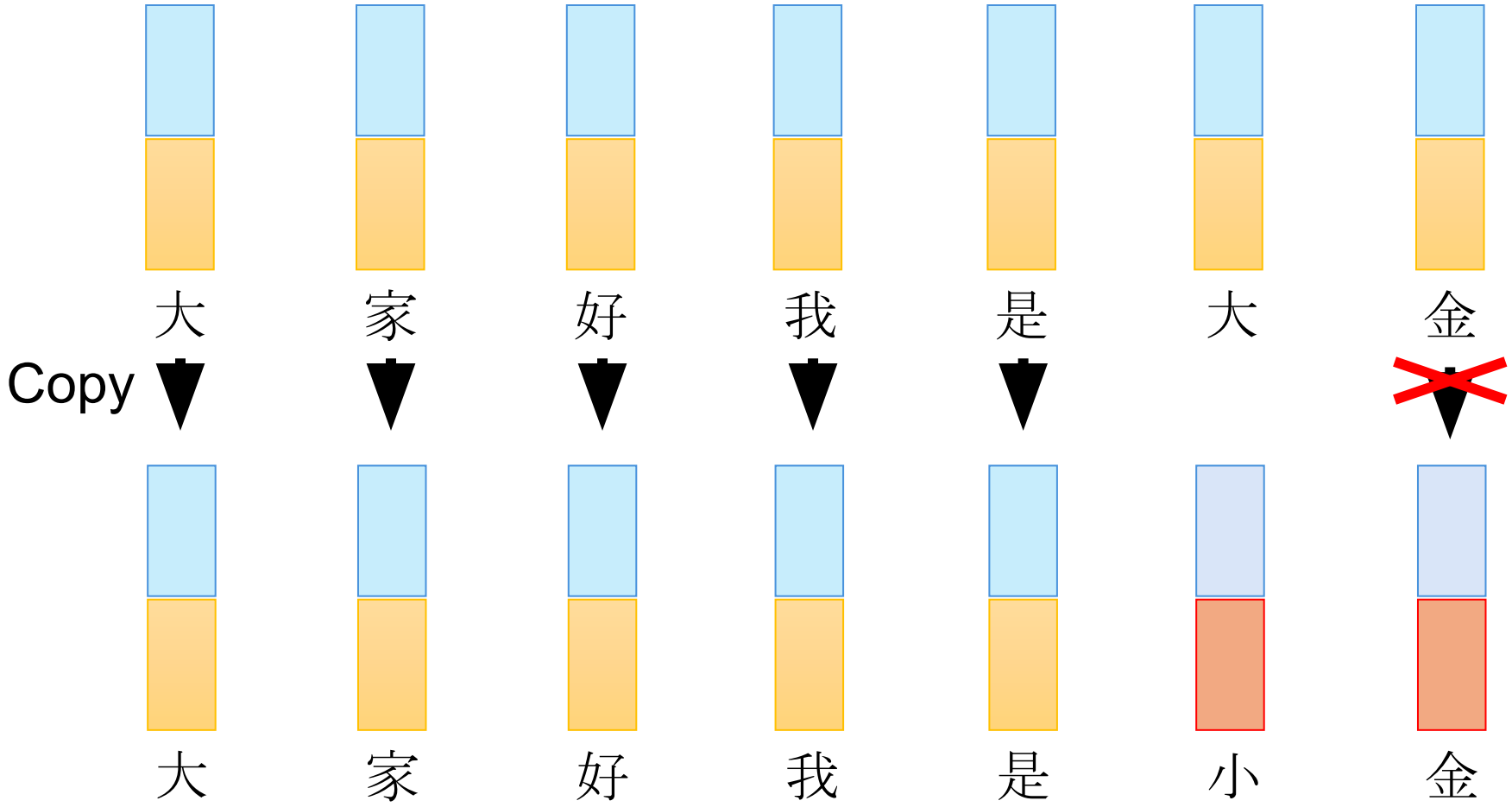
https://arxiv.org/pdf/2510.13940

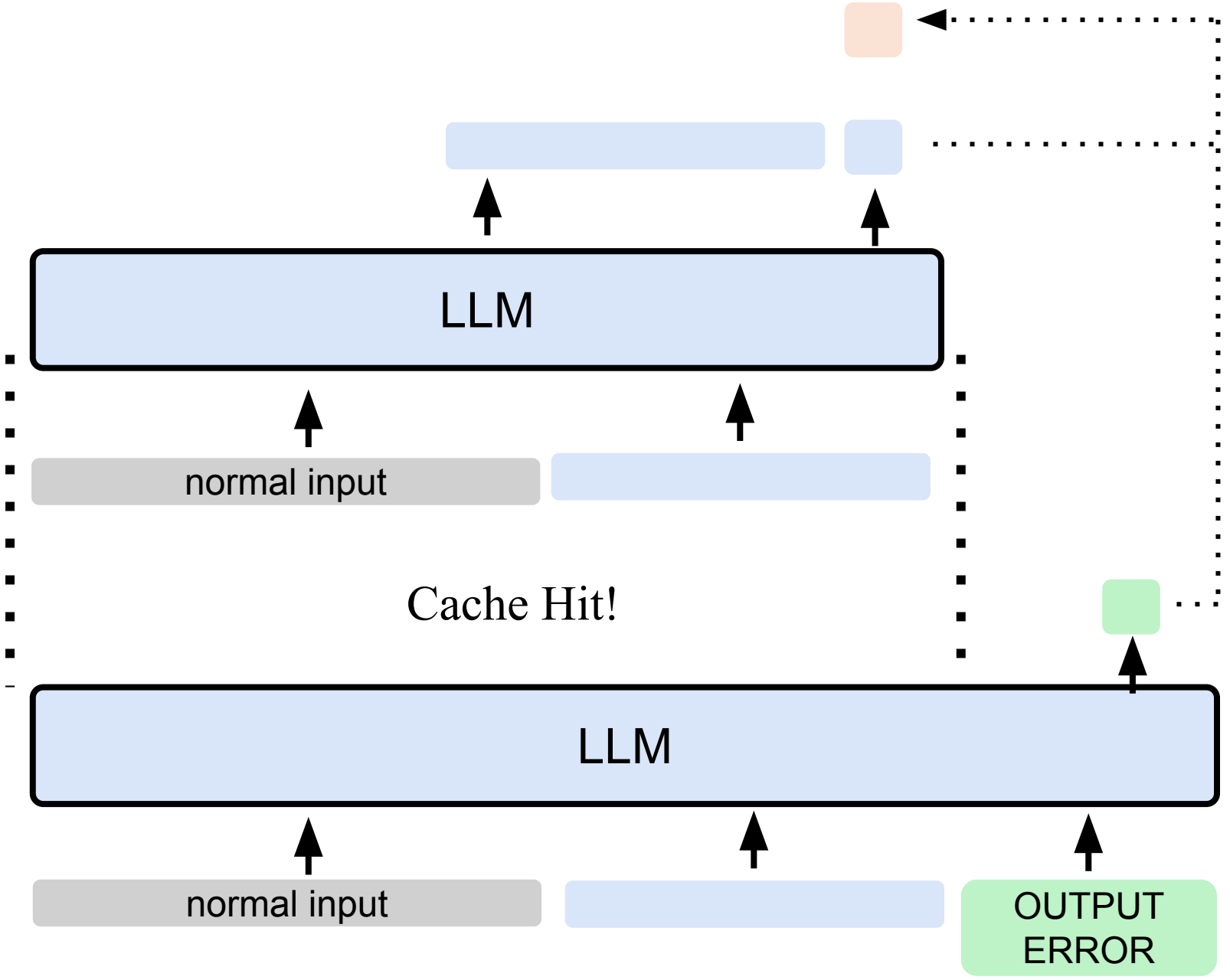


這裡才需要用 contrastive decoding 仔細考慮


high entropy
(low confidence)
LLM
normal input
LLM
no image, no audio, bad
instruction, etc.
LLM
normal input
LLM
no image, no audio, bad
instruction, etc.

Recall: 跨對話的 KV Cache

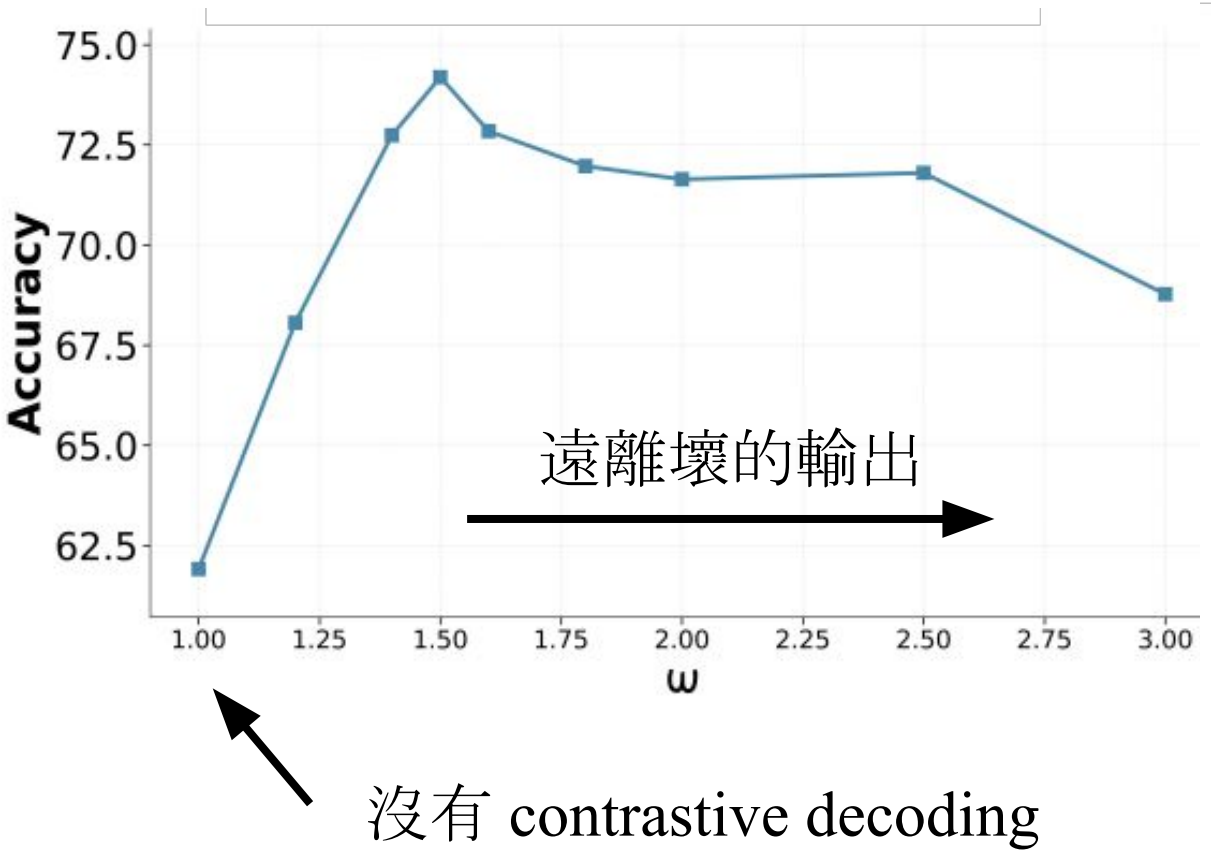


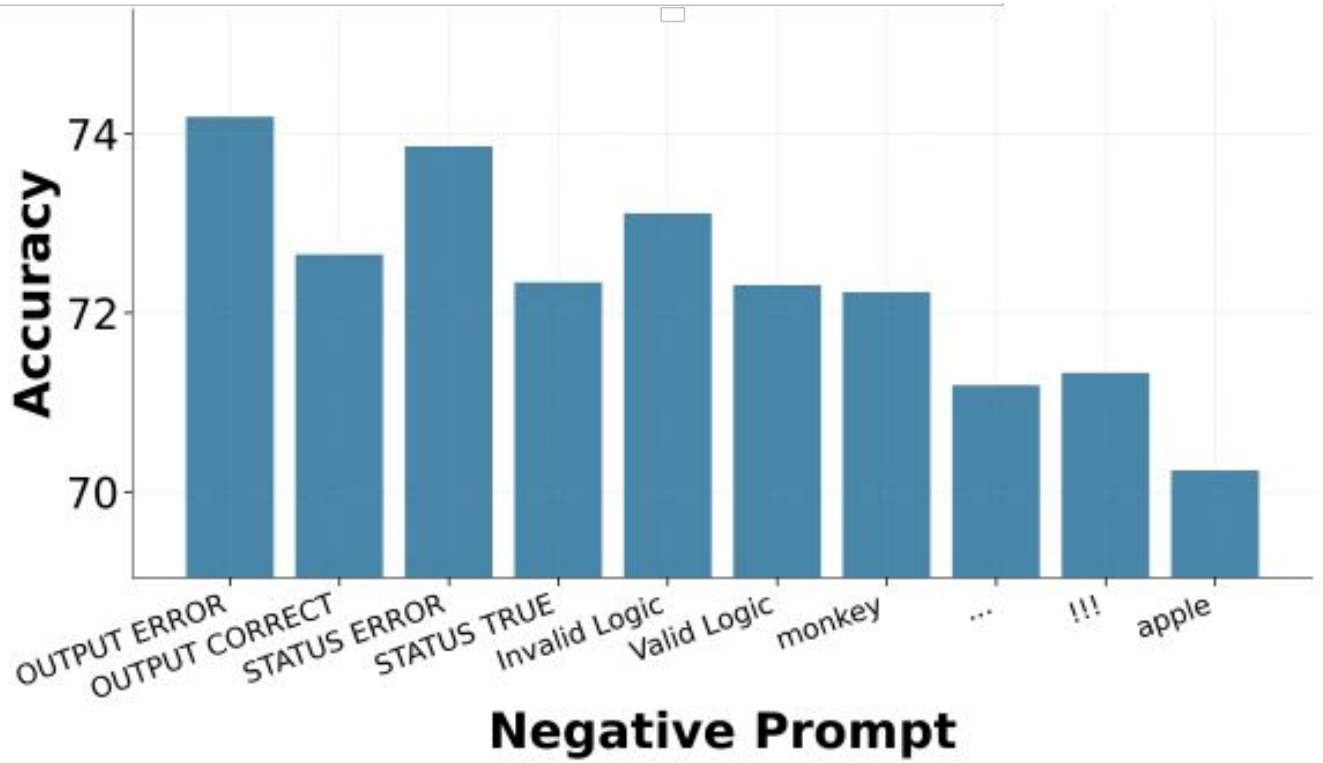


only two tokens

Minimal Test-Time Intervention (MTI)

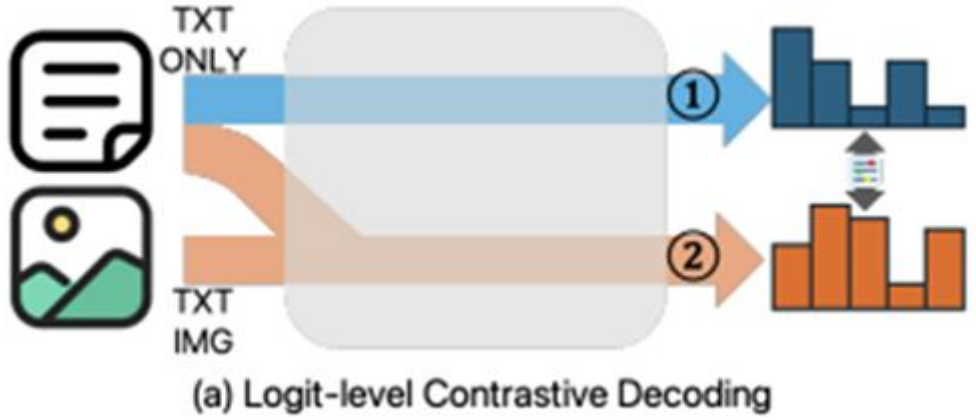
https://arxiv.org/pdf/2510.13940





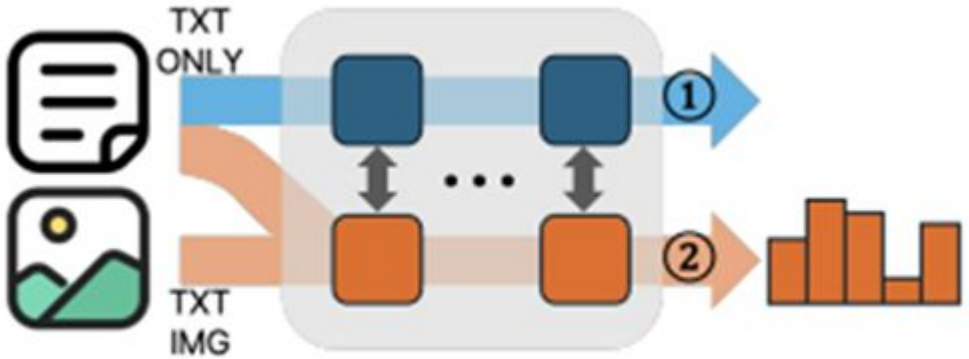
source of image: https://arxiv.org/pdf/2601.13707

可以改 logit 以外的地方



(a) Logit-level Contrastive Decoding

Most approaches



(b) Hidden States-level Latent Steering

VISTA

https://arxiv.org/pdf/2502.03628

(d) Ours (Attention-space Contrastive Guidance)

ACG

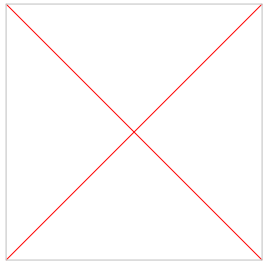
https://arxiv.org/pdf/2601.13707

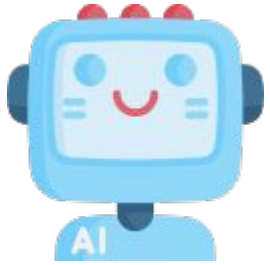
Summary

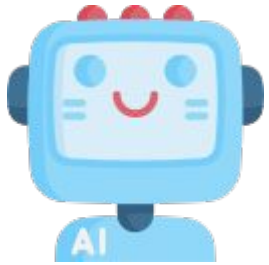
	怎麼拿到錯誤結果	改哪裡
Contrastive Decoding (CD)	小模型	output
Decoding by Contrasting Layers (DoLa)	淺層用 logit lens 生成的結果	output
Layer Contrastive Decoding (LayerCD)	用淺層的 Image encoder layer (影像)	output
Instruction Contrastive Decoding (ICD)	降智咒語	output
Context-aware Decoding (CAD)	拿掉 Context (e.g., RAG 中的 Retrieved documents)	output
Visual Contrastive Decoding (VCD)	影像加雜訊、打亂 Patch、蓋住重要部分	output
Audio-aware Decoding (AAD)	移除聲音	output
Minimal Test-Time Intervention (MTI)	降智咒語，目標: 減少算力消耗	output
Visual Information Steering with Token-logit Augmentation (VISTA)	移除影像	hidden representation
Attention-space Contrastive		

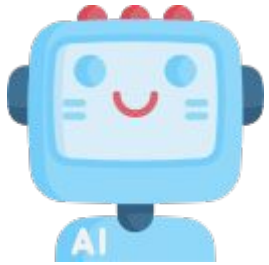
修改 Harness (Workflow)

Generation → Verification









input	output	reflection instruction	error detection?	corrected output?

“再檢查一下”

(與問題、前面的答案無關, 程式自動插入)

- 批判比生成容易 :我不用會寫小說也能判斷一部小說好不好看
- 生成的過程無法回頭, 有錯也無法修正。自動插入的 “reflection instruction”給模型 “機會”生成修正的的答案

相關課程錄影

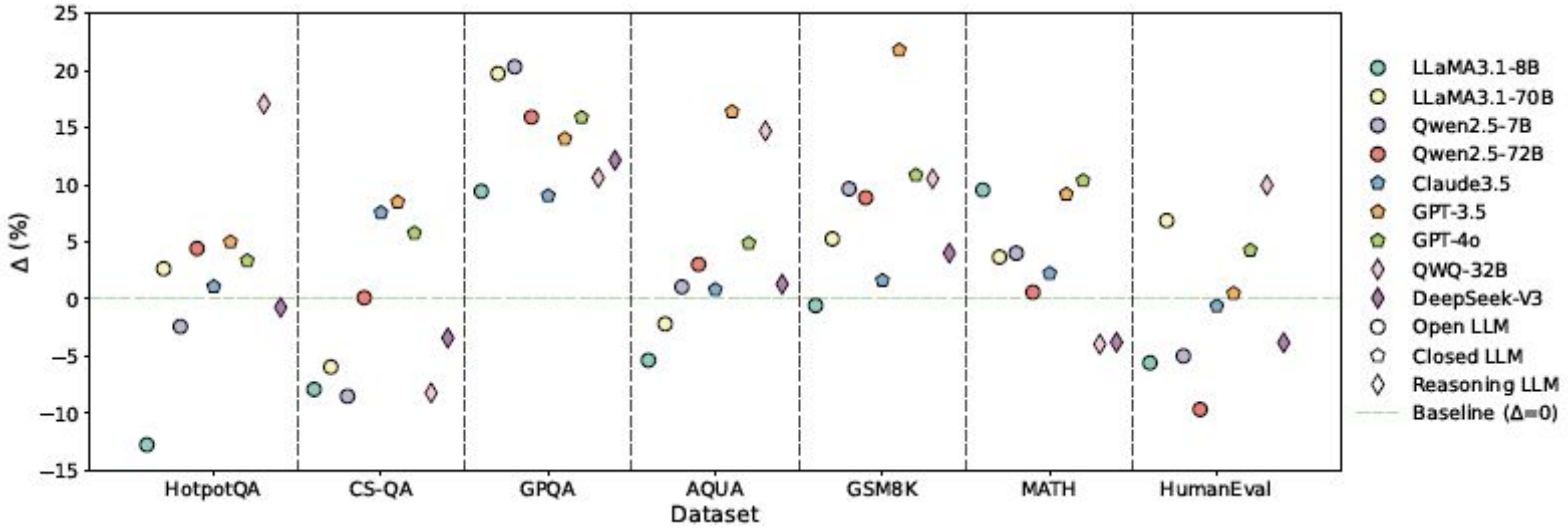
【生成式AI時代下的機器學習(2025)】第七講：DeepSeek-R1 這類大型語言模型是如何進行「深度思考」（Reasoning）的？

https://youtu.be/bJFtcwLSNxI?si=cjZTecajrM5O5fx0&t=1083

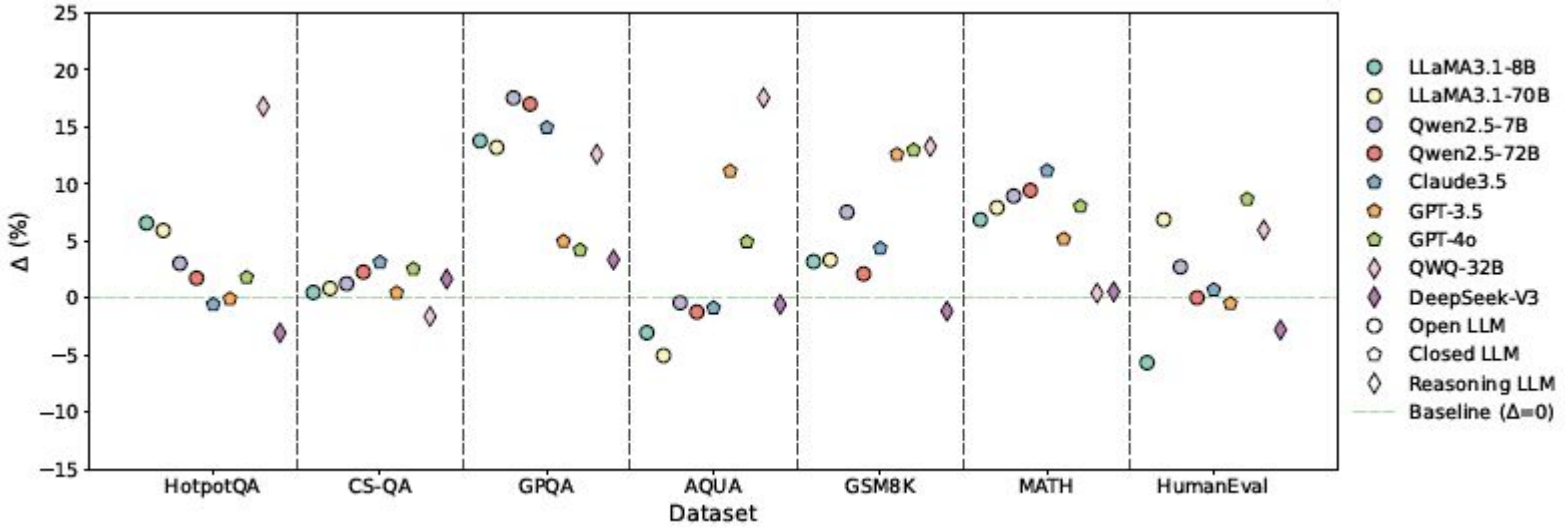
Can LLMs Correct Themselves?
A Benchmark of Self-Correction in LLMs

https://arxiv.org/pdf/2510.16062

Internal

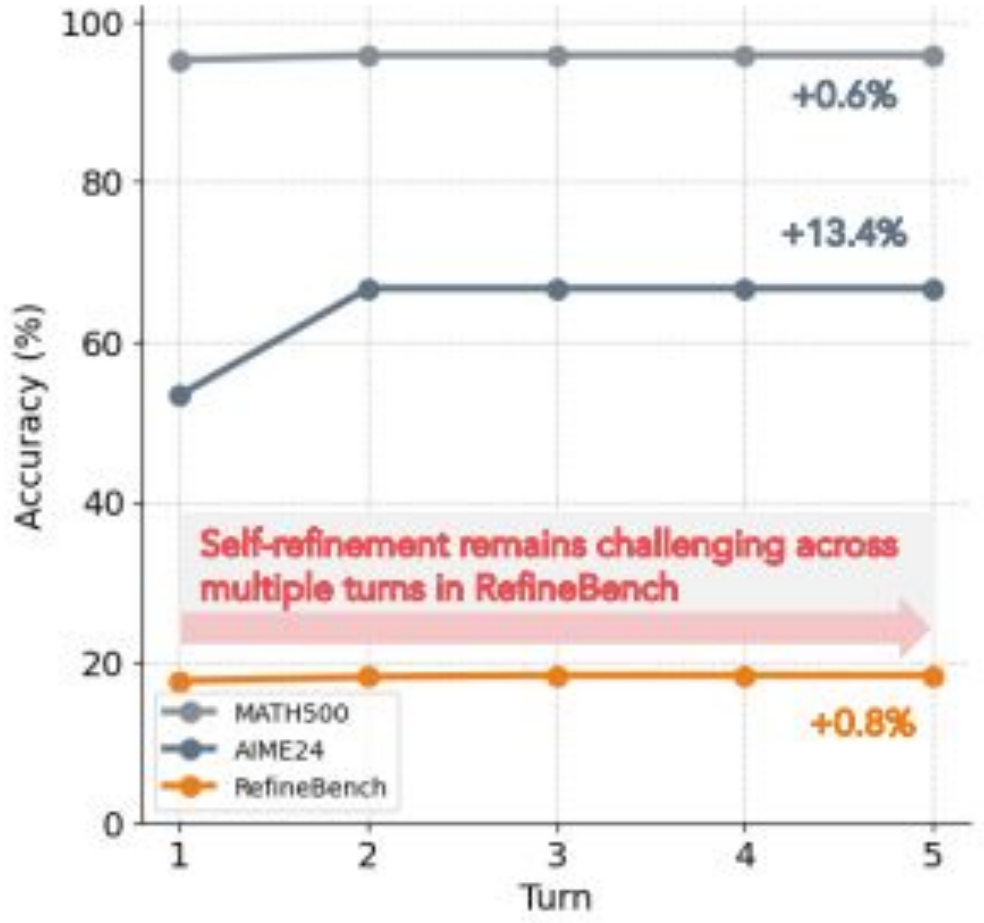


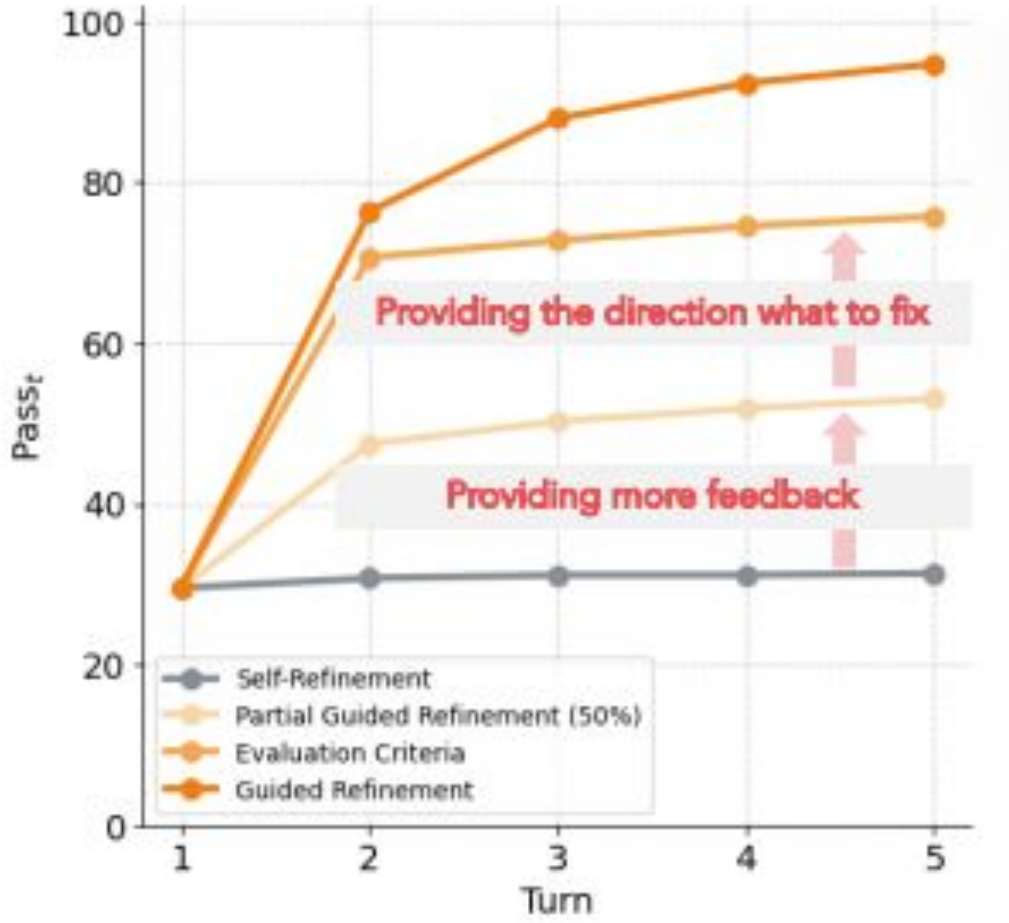
External



RefineBench: Evaluating Refinement Capability of Language Models via Checklists

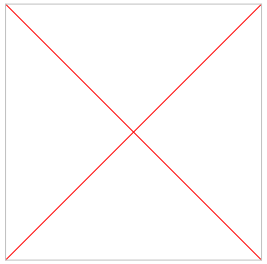
https://arxiv.org/pdf/2511.22173

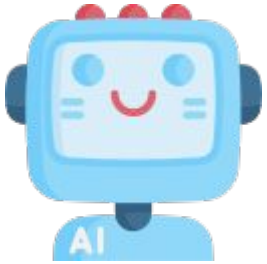


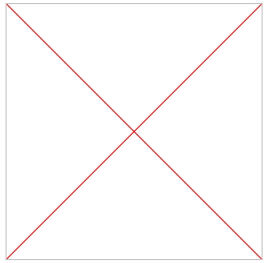


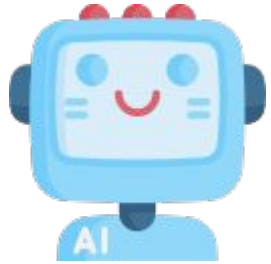
進一步分析模型修正行為

https://arxiv.org/pdf/2412.19513









input	output	reflection instruction	corrected output
	✗		✓
	✓		✓
	✗		✗
	✓		✗

進一步分析模型修正行為

https://arxiv.org/pdf/2412.19513

Confidence Level (CL)

P(✓ → ✓)

Critique Score (CS)

P(✗ → ✓)

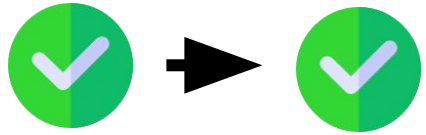
ACC1:
accuracy before correction

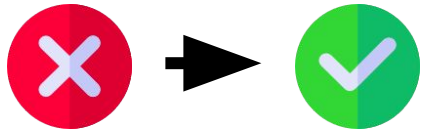
ACC2:
accuracy after correction

ACC2 =

ACC1 x **CL** + (1 – ACC1) x **CS**

Confidence Level (CL)
Critique Score (CS)

P( ) 

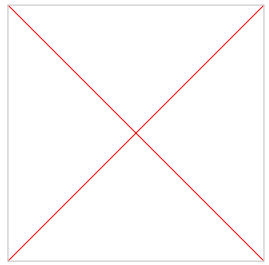
P( ) 

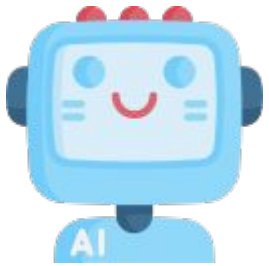
Models	GSM8k				MMLU				BoolQ			
	Acc_1	Acc_2	CL	CS	Acc_1	Acc_2	CL	CS	Acc_1	Acc_2	CL	CS
Llama3-8B-Instruct	71.0	78.1	91.7	**44.9**	62.2	64.0	94.9	13.1	62.3	64.8	86.0	29.8
Deepseek-7B-Chat	61.2	60.9	95.9	5.6	47.8	47.9	**98.7**	1.3	57.8	57.6	**98.8**	1.2
Mistral-7B-Instruct	50.1	51.1	90.9	11.0	59.2	59.2	98.4	2.3	61.4	62.5	98.5	5.4
Qwen2.5-7B-Chat	91.9	92.4	**99.4**	14.5	71.0	71.5	93.3	18.0	58.8	60.9	**93.9**	13.8
GLM4-9B-Chat	64.9	63.7	87.9	19.0	63.5	64.6	83.3	**32.1**	61.1	64.8	77.1	**45.5**
Llama3-70B-Instruct	90.7	92.7	97.3	**48.1**	78.2	79.5	97.2	**16.2**	76.3	76.4	84.7	49.3
Deepseek-67B-Chat	82.4	82.3	99.1	3.7	65.3	66.3	94.8	12.9	69.8	69.8	89.9	23.4
Qwen2.5-72B-Chat	95.7	95.9	**99.9**	7.5	82.6	83.4	**98.2**	13.5	65.5	75.9	93.9	41.5
Qwen-Max	96.1	96.4	**99.9**	11.5	83.8	85.0	**99.2**	11.6	71.3	73.6	98.2	12.5
GPT-3.5 Turbo	81.3	84.0	95.6	**33.8**	65.3	65.6	89.6	20.5	68.5	70.3	75.7	**58.8**
GPT-4 Turbo	93.6	92.1	96.8	23.9	84.3	82.3	88.4	**49.6**	80.5	78.6	87.8	40.6

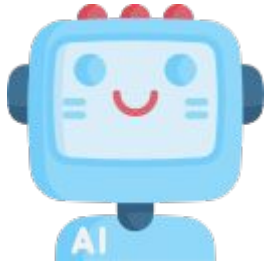
https://arxiv.org/pdf/2412.19513

Reflection Instruction 的用詞很重要

https://arxiv.org/pdf/2412.19513







input | output | reflection instruction | corrected output

(與問題、前面的答案無關，程式自動插入)

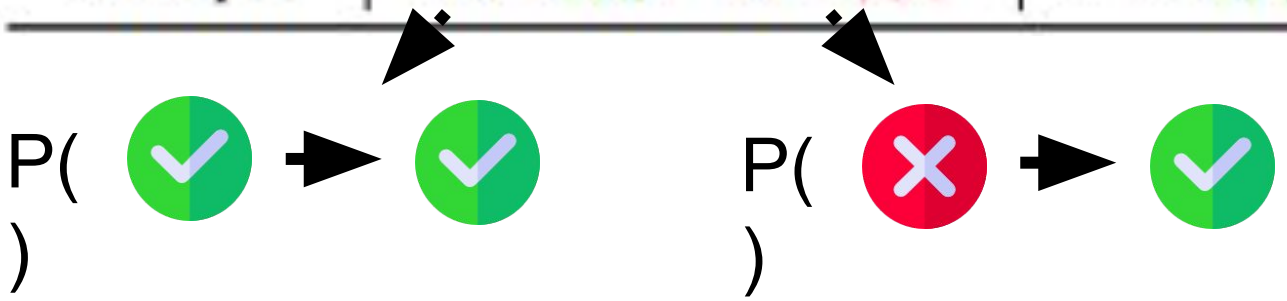
這句話怎麼說可能會大幅影響模型的行為

Reflection Instruction 的用詞很重要

https://arxiv.org/pdf/2412.19513

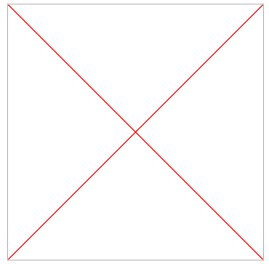
Reask: 再做一次
Confidence: 你應該是對的。給我最終答案
Critique: 你確定嗎? 再好好想想

Prompt	GSM8k		MMLU		BoolQ	
	CL	*CS*	*CL*	*CS*	*CL*	*CS*
Reask	$91.7_{0.0}$	$44.9_{0.0}$	$94.9_{0.0}$	$13.1_{0.0}$	$86.0_{0.0}$	$29.8_{0.0}$
Confidence	$93.5_{+1.8}$	$32.9_{-12.0}$	$99.0_{+4.1}$	$2.0_{-11.1}$	$96.1_{+10.1}$	$8.9_{-20.9}$
Critique	$77.7_{-14.0}$	$47.9_{+3.0}$	$71.1_{-23.8}$	$26.0_{+22.9}$	$54.6_{-31.4}$	$62.3_{+32.5}$

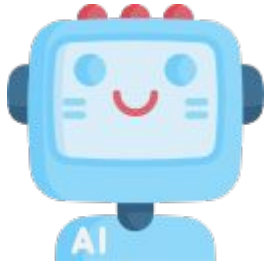


Verification 真的划算嗎？

https://arxiv.org/abs/2504.01005

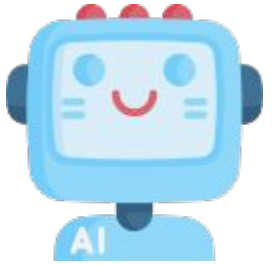


input



output

reflection instruction

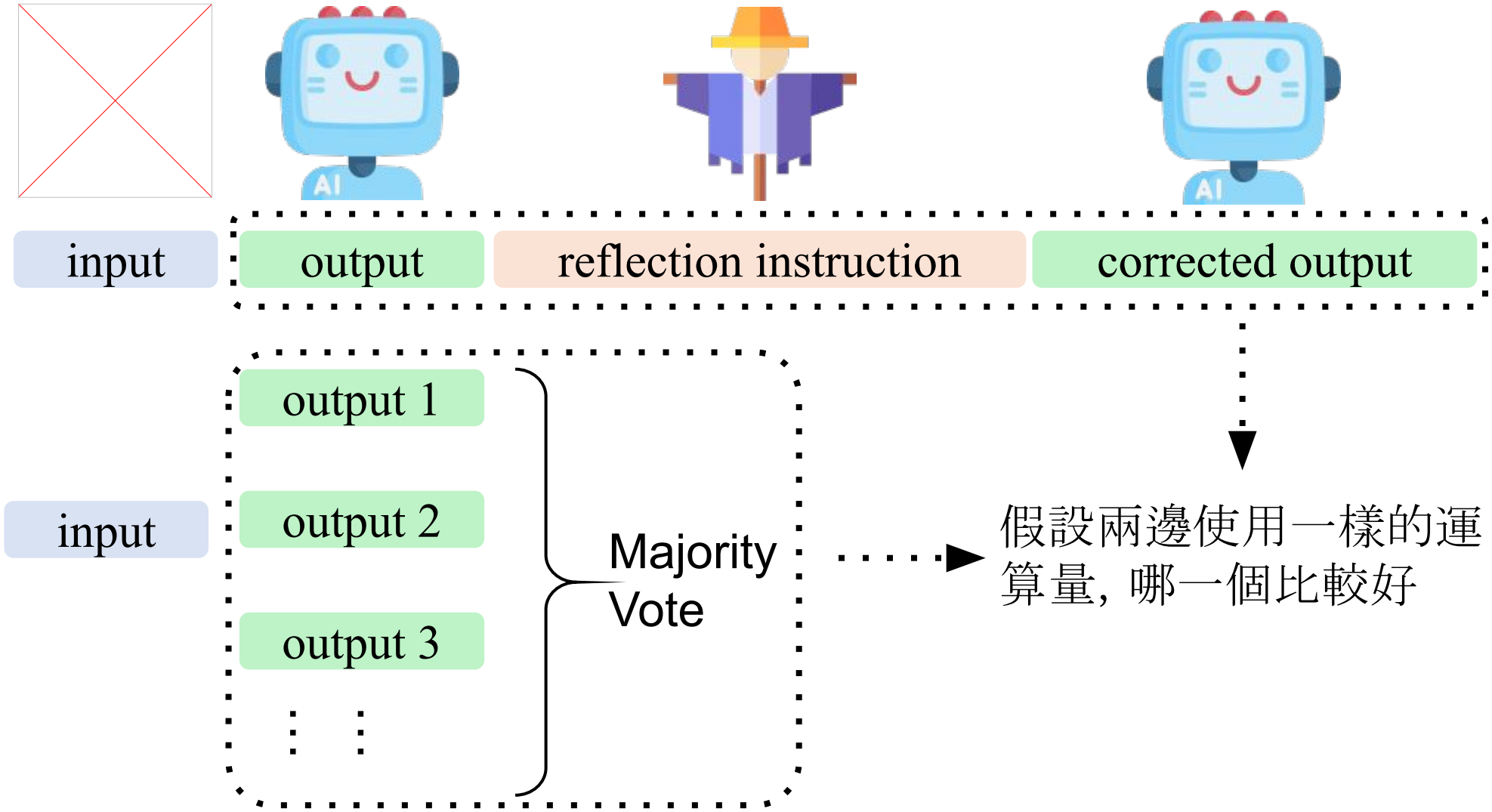


corrected output

需要投資額外的運算，這個投資划算嗎？

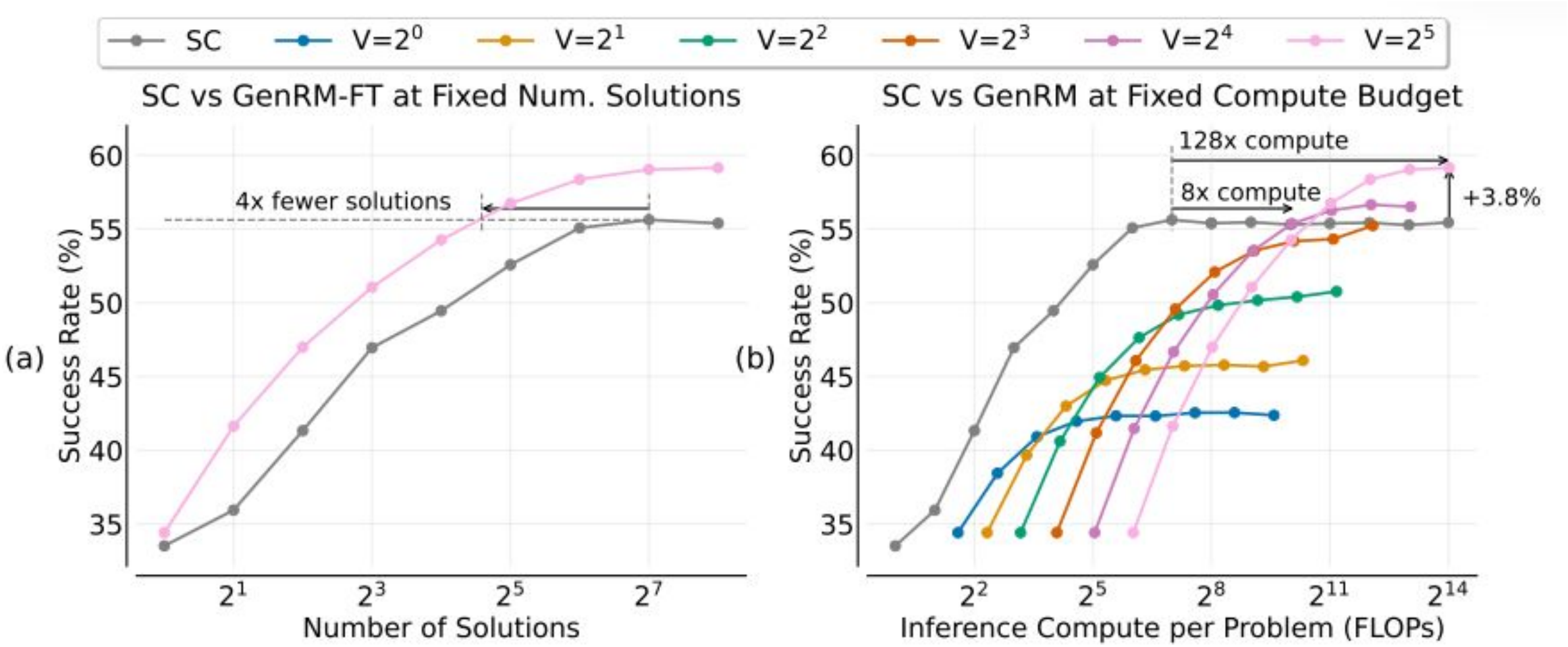
Verification 真的划算嗎？

https://arxiv.org/abs/2504.01005



Verification 真的划算嗎？

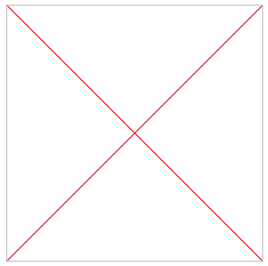
https://arxiv.org/abs/2504.01005

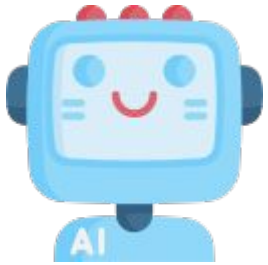


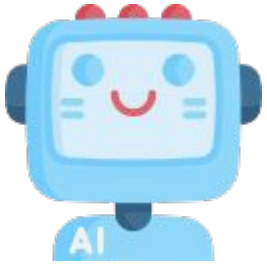
修改 Model Parameters (Reasoning)

Workflow → Reasoning

Workflow

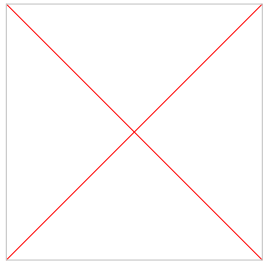


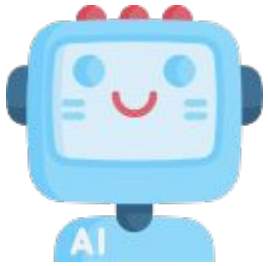


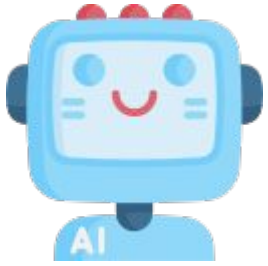


input | output | reflection instruction | corrected output

Reasoning







input | output | corrected output

直接教模型自我修正

- 正確知識 ≠ 自我修正

  https://arxiv.org/pdf/2505.16170

相關課程錄影

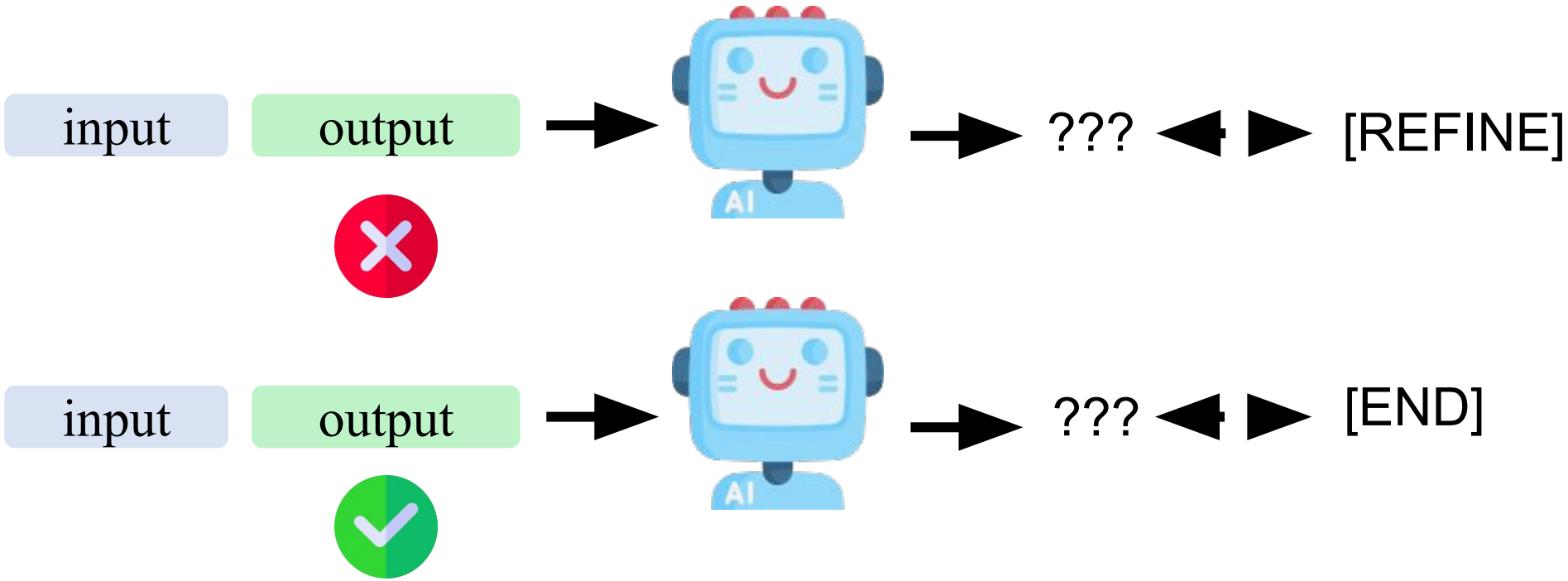
【生成式AI時代下的機器學習(2025)】第七講：DeepSeek-R1 這類大型語言模型是如何進行「深度思考」（Reasoning）的？

https://youtu.be/bJFtcwLSNxI?si=ocBy0Y5EuOmLrY3T&t=2437

直接教模型自我修正

ReVISE
https://arxiv.org/pdf/2502.14565

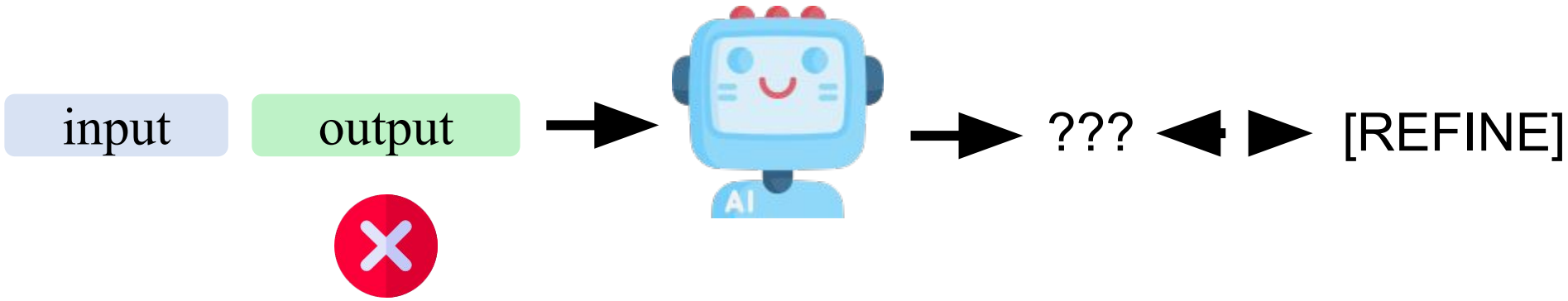
Stage 1: 先教錯誤偵測



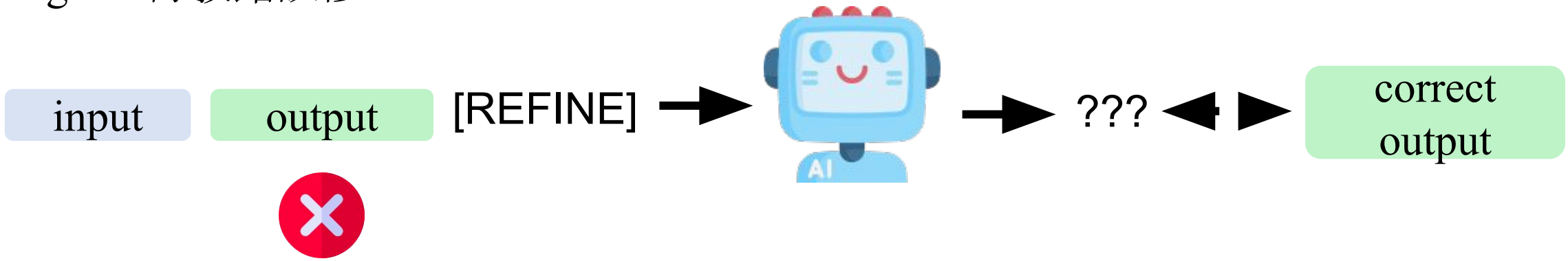
直接教模型自我修正

ReVISE
https://arxiv.org/pdf/2502.14565

Stage 1: 先教錯誤偵測

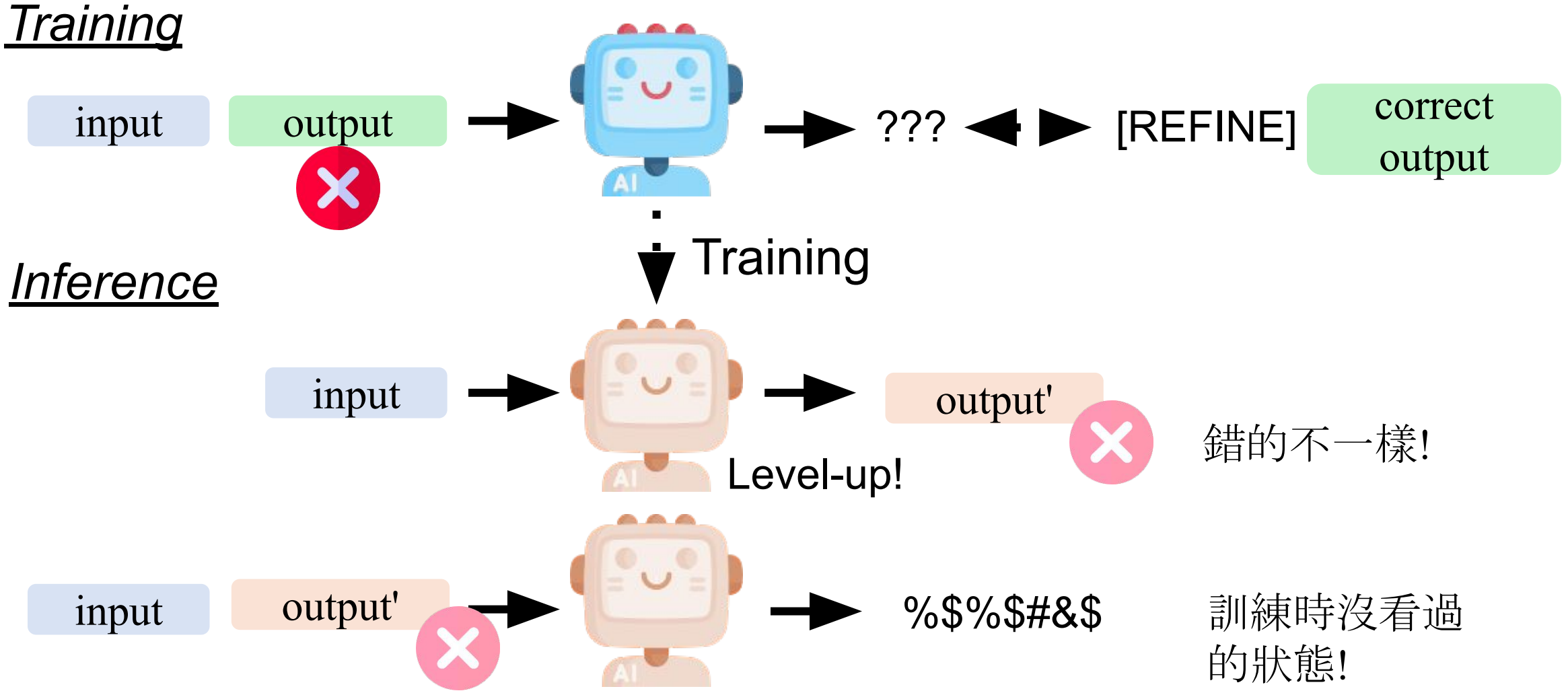


Stage 2: 再教錯誤修正



直接教模型自我修正

https://arxiv.org/abs/2409.12917

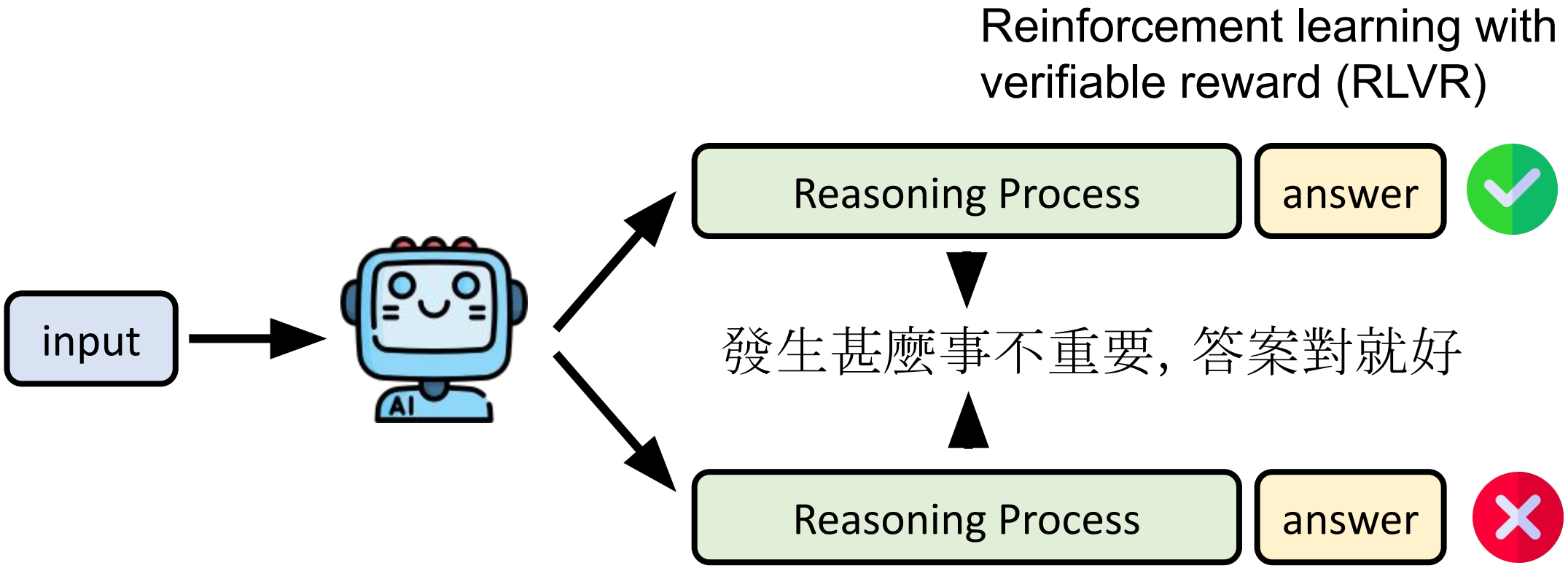


相關課程錄影

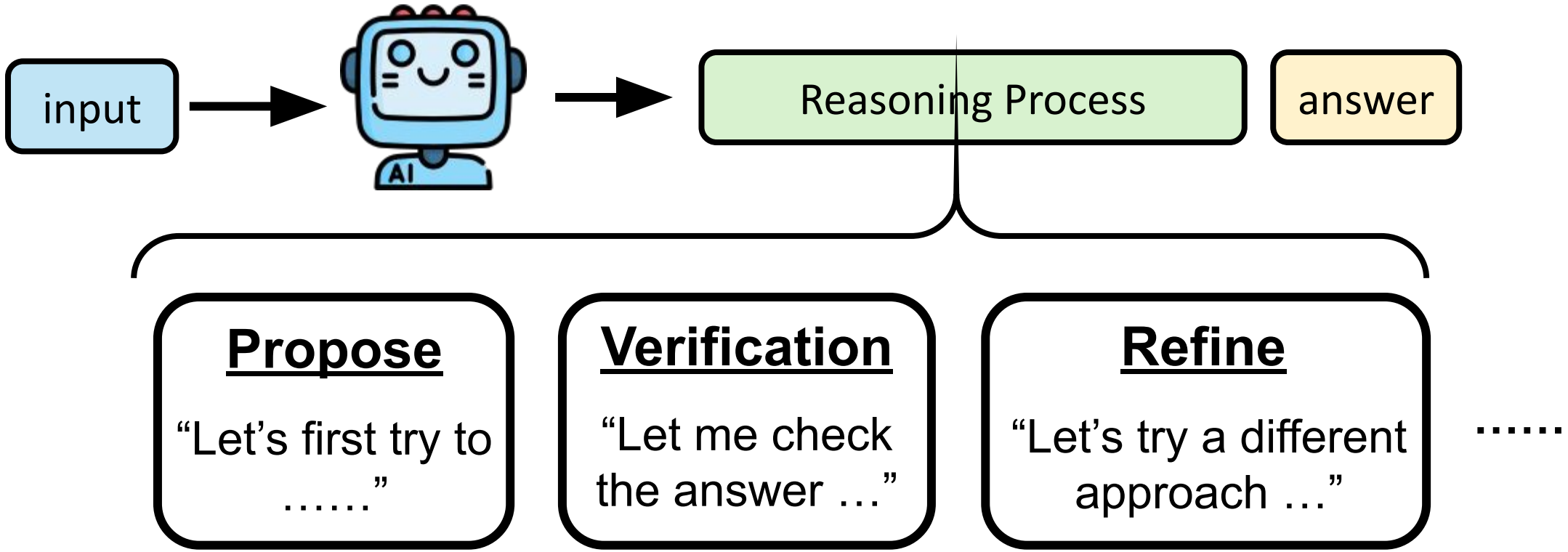
【生成式AI時代下的機器學習(2025)】第七講：DeepSeek-R1 這類大型語言模型是如何進行「深度思考」（Reasoning）的？

https://youtu.be/bJFtcwLSNxI?si=uBsl_Va4xArE6F2J&t=3327

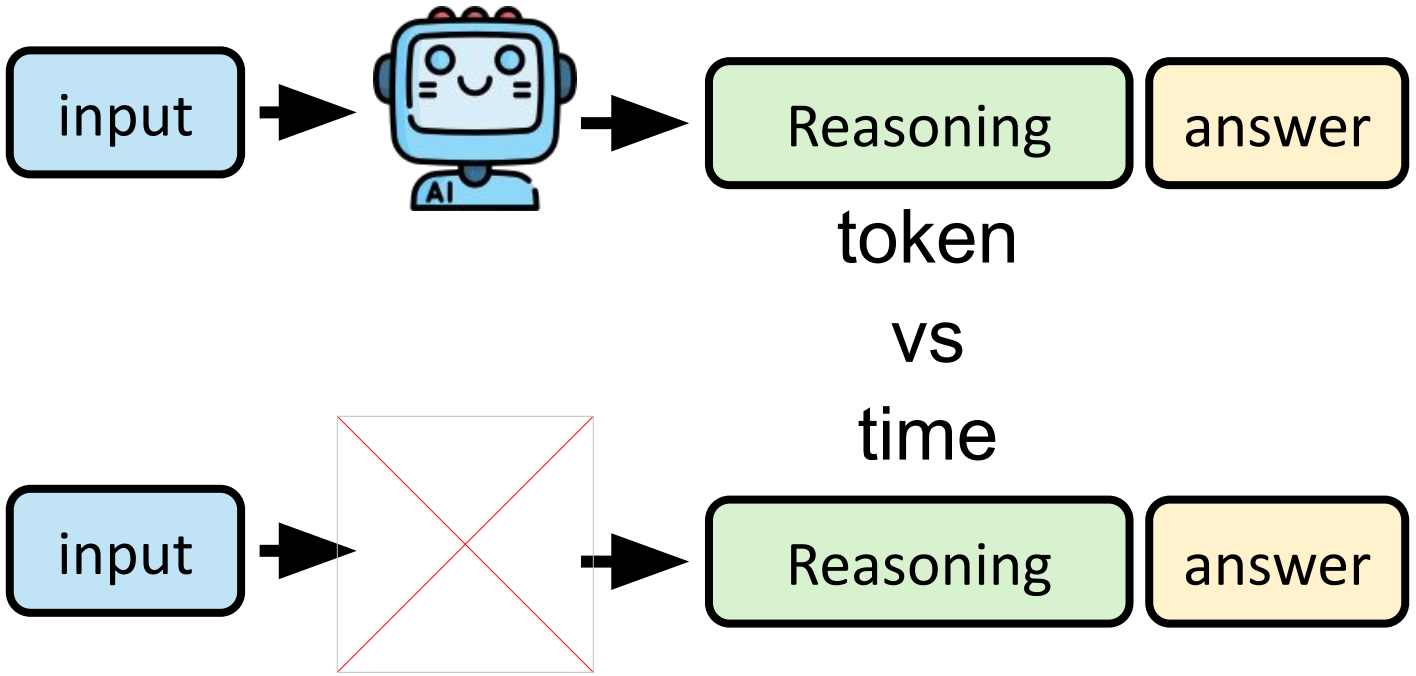
Reinforcement Learning

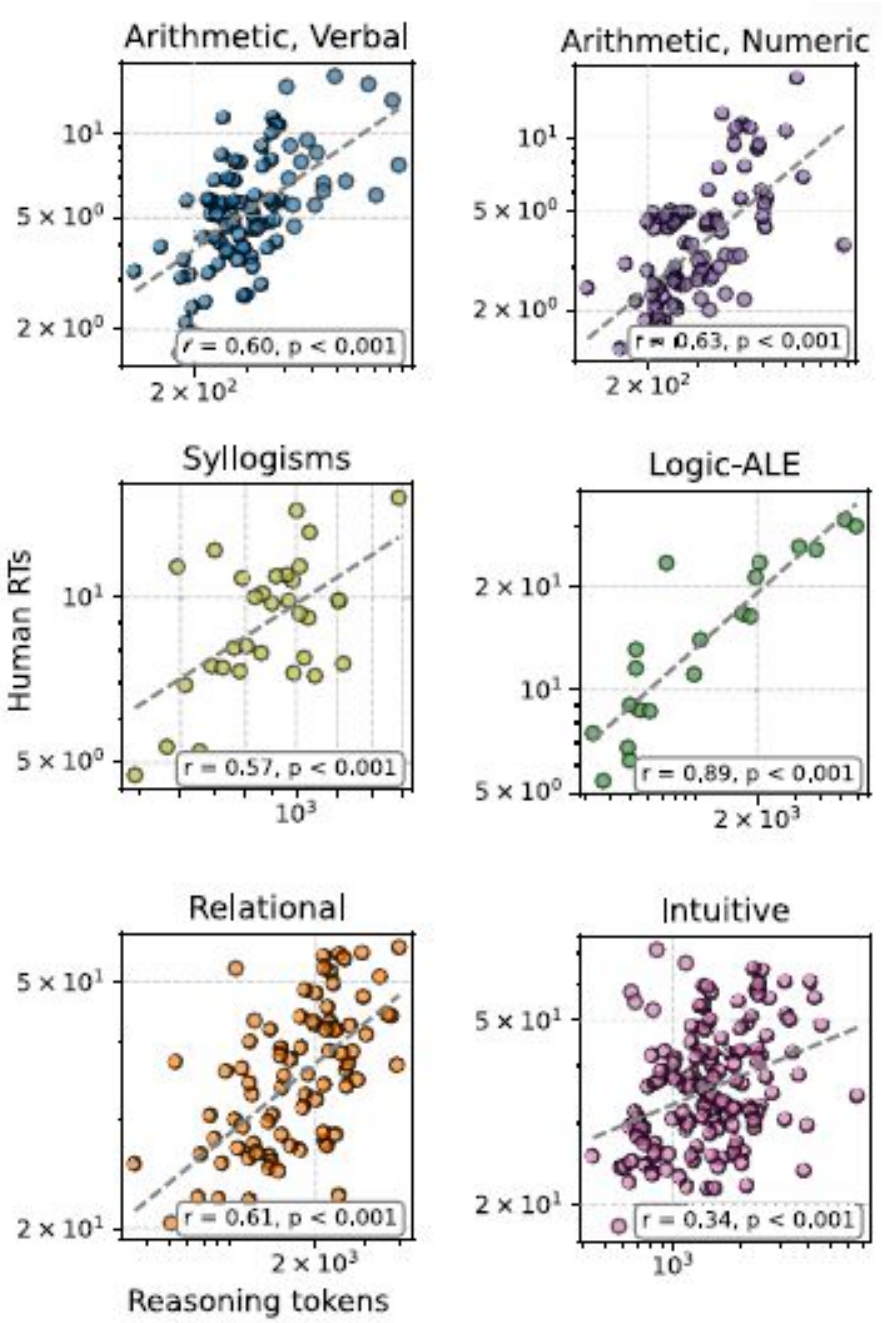


Reinforcement Learning



為什麼不一開始就做對？

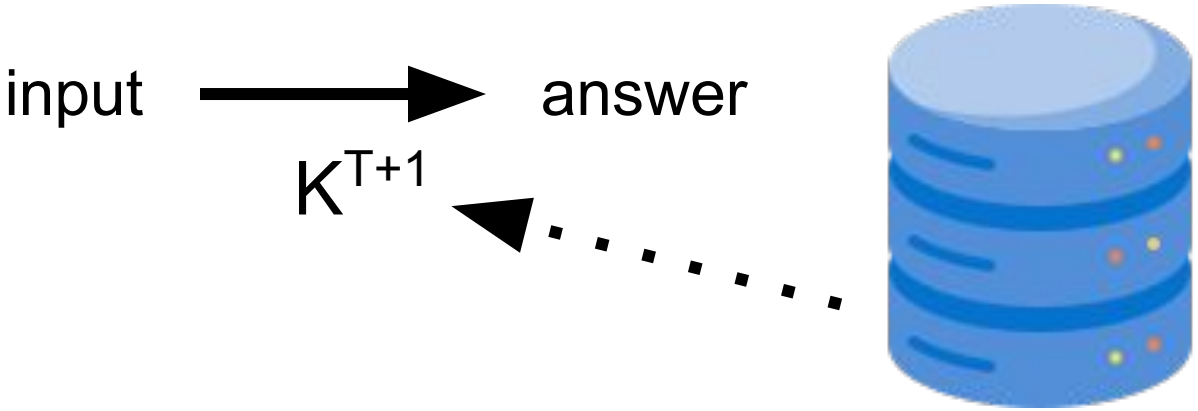


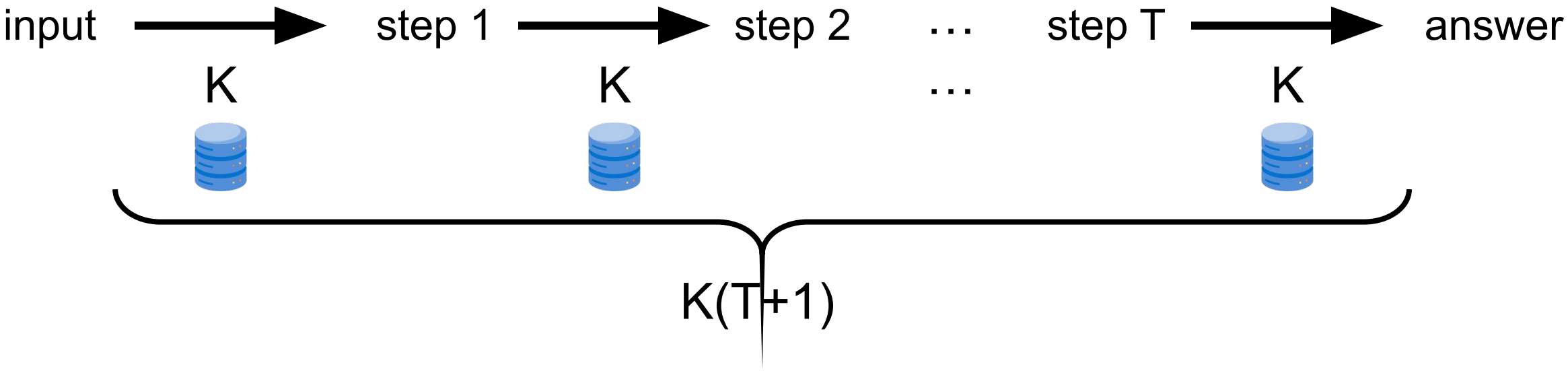


https://news.mit.edu/2025/cost-of-thinking-1119

為什麼不一開始就做對？

https://arxiv.org/abs/2502.04667
https://arxiv.org/abs/2502.08991
https://arxiv.org/abs/2502.18273





Parity Check

0 0 1 0 0 0 → 1

0 0 1 0 0 1 → 0

1 1 0 1 1 1 → 1

⋮ ⋮ ⋮ ⋮

2^6 = 64 samples to teach

0 0 1 0 0 0

→ 0 1 1 1 1

1 1 0 1 1 1

→ 0 0 1 0 1

4 x 5 samples to teach

Each step

0 0 → 0

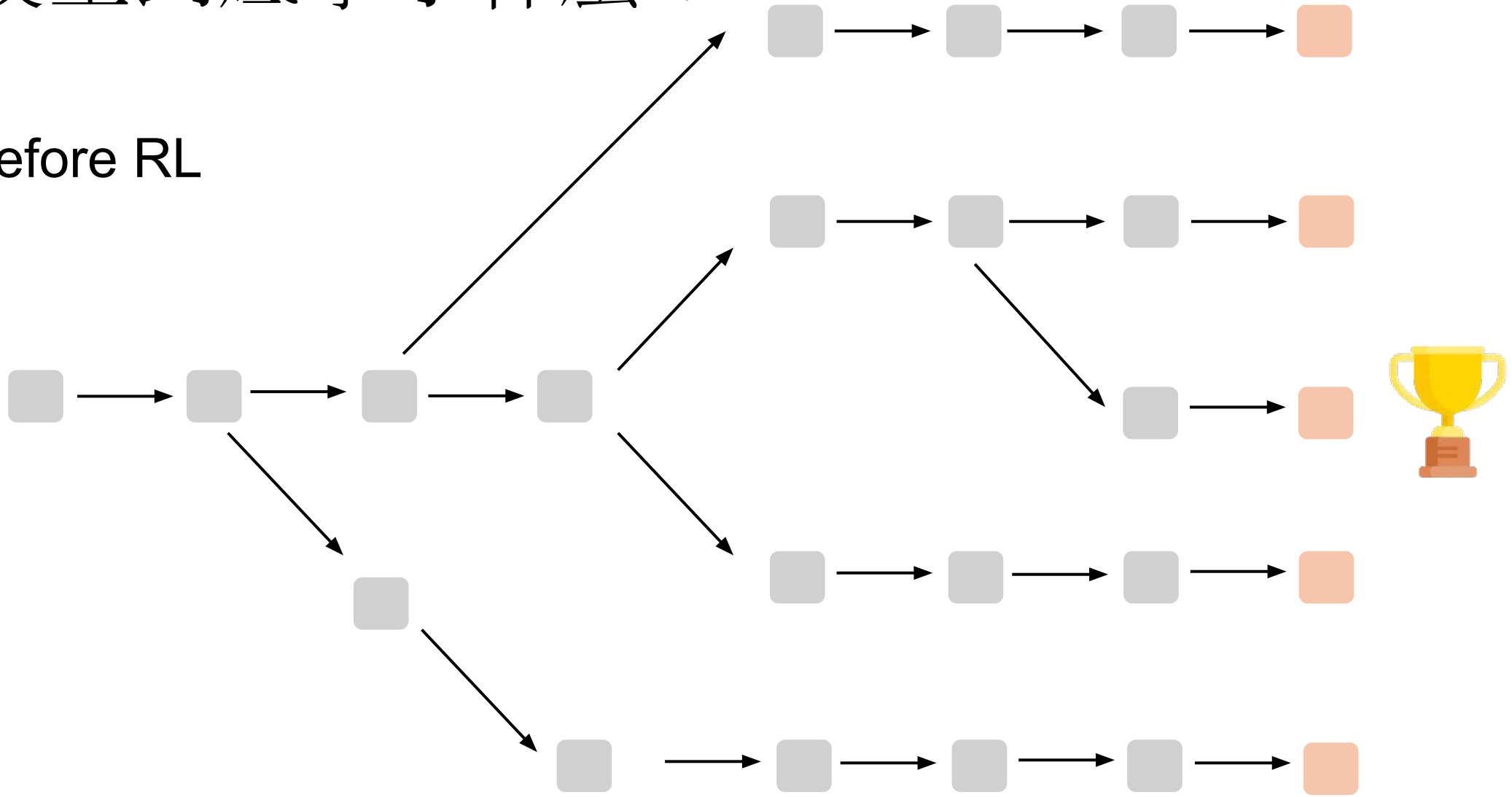
0 1 → 1

1 0 → 1

1 1 → 0

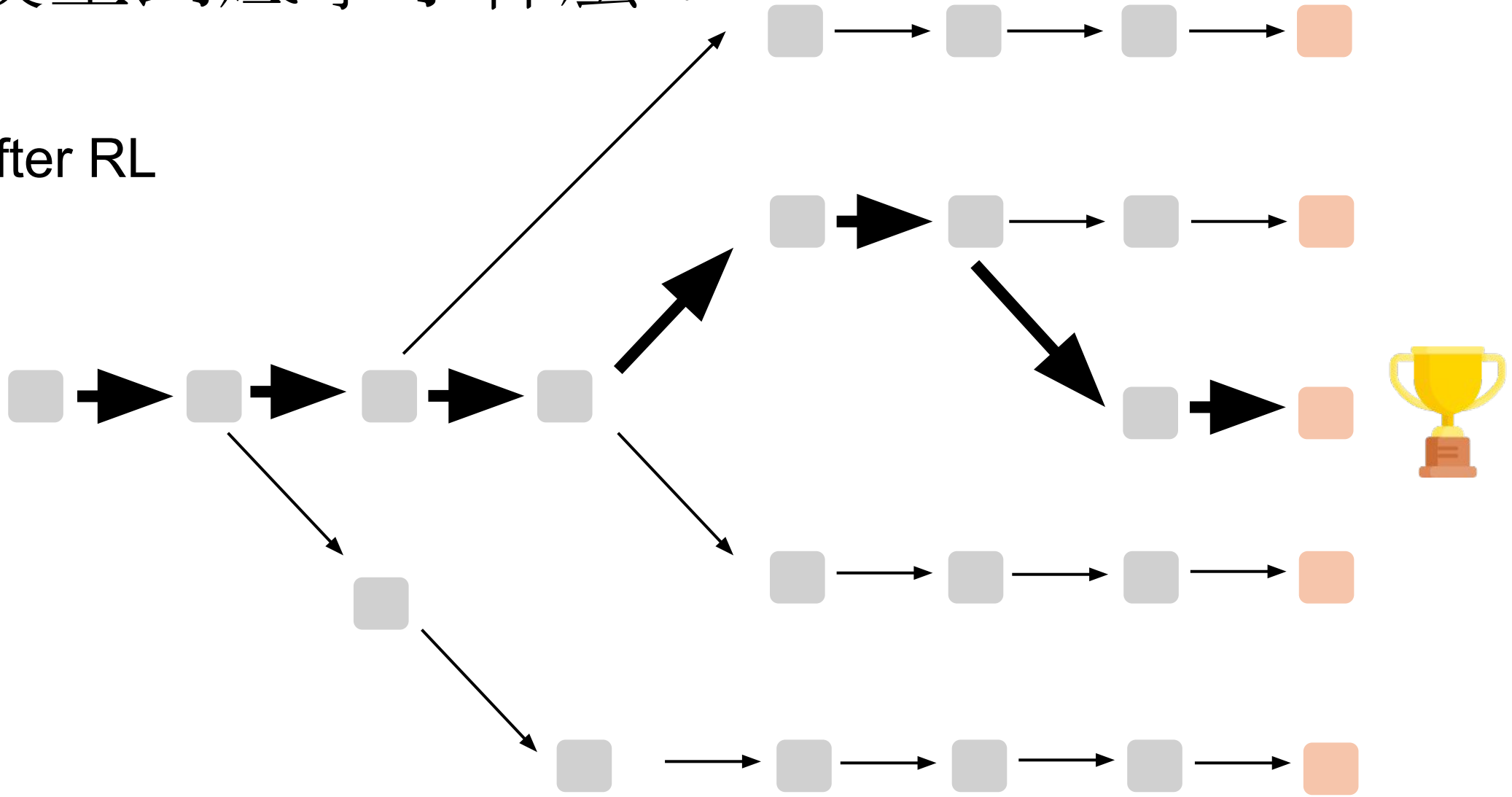
模型到底學了什麼？

Before RL

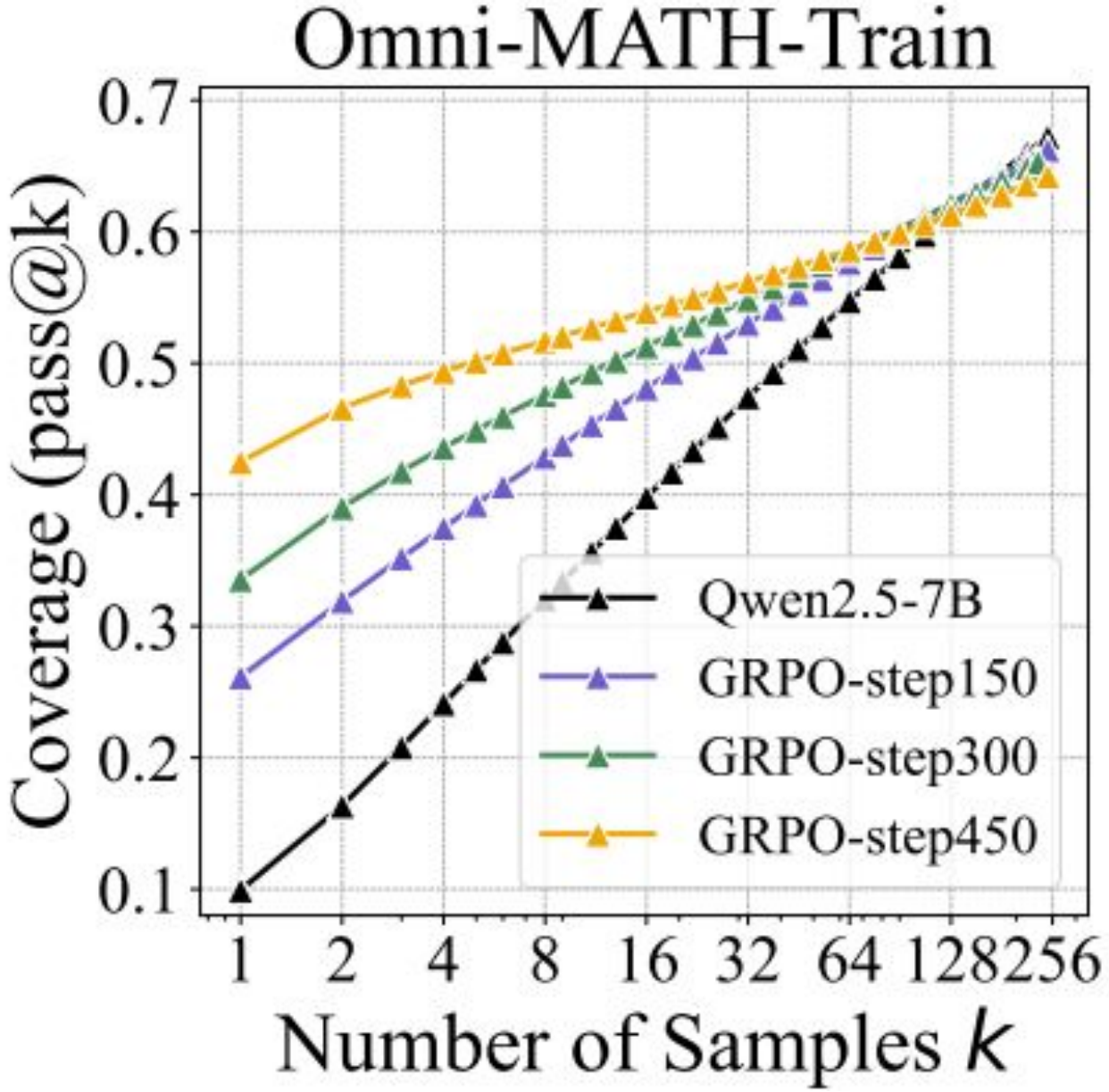


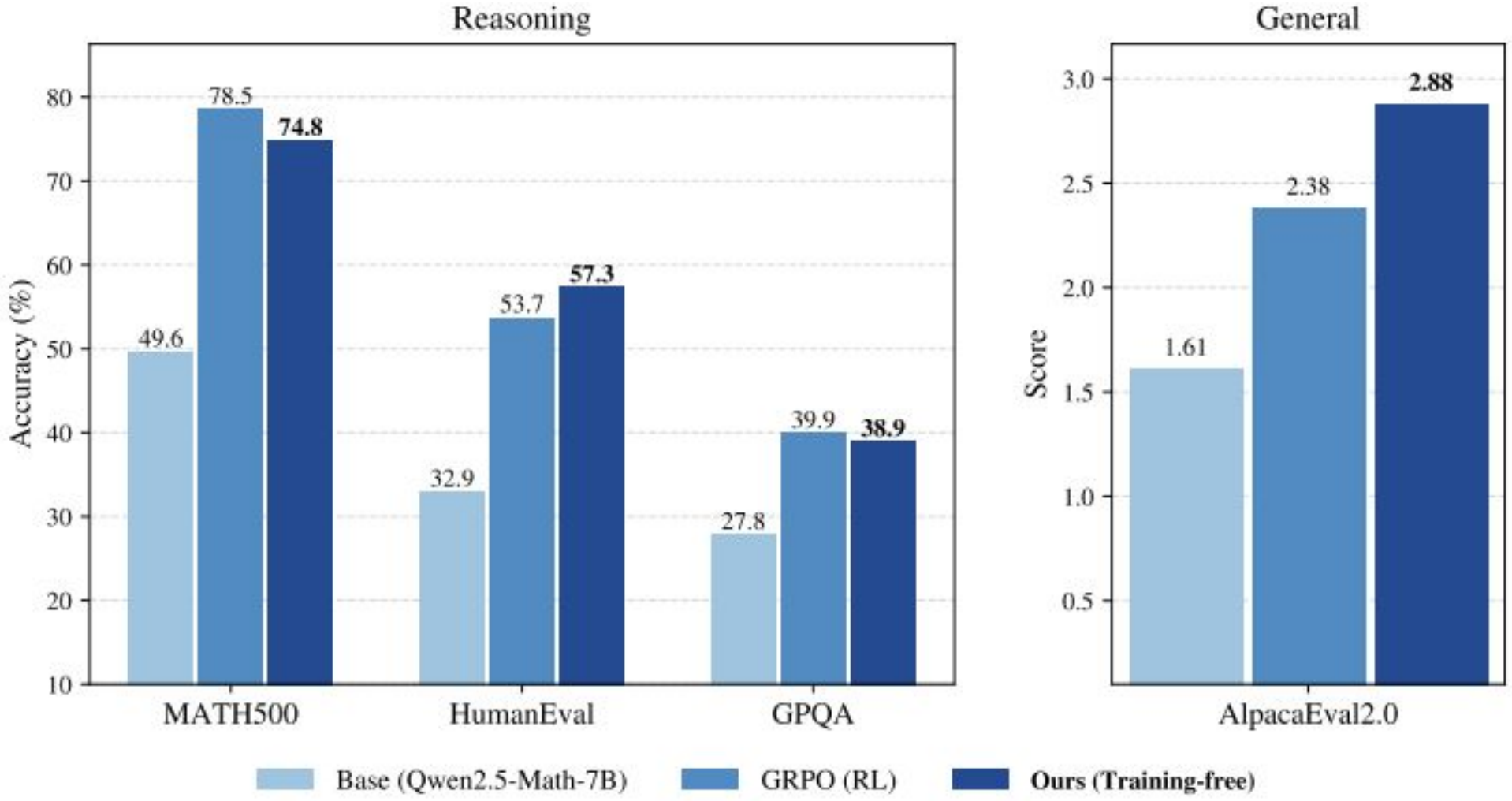
模型到底學了什麼？

After RL



Source of image:
https://arxiv.org/abs/2504.13837



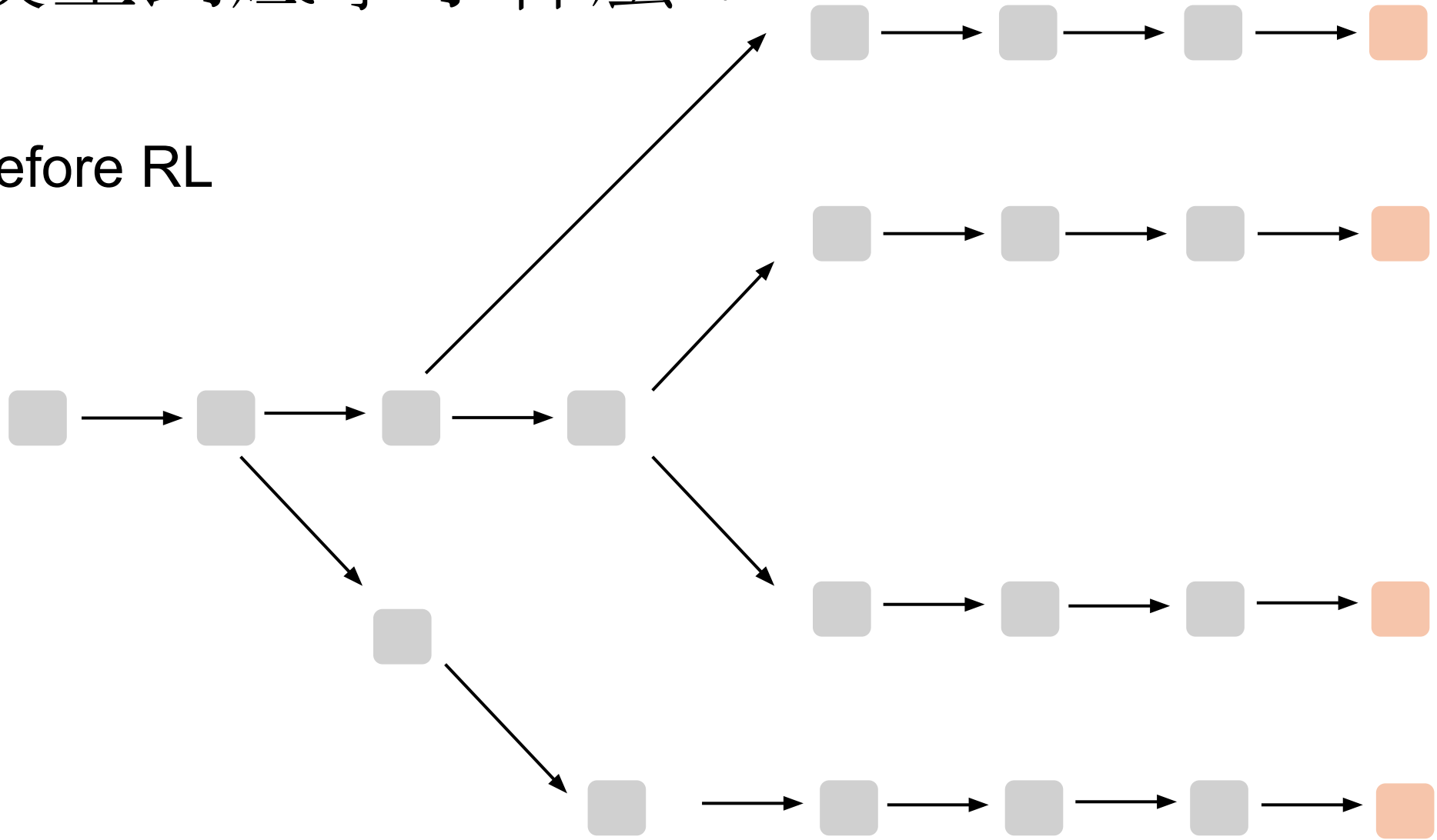


不需要訓練模型，只需要更好的 sampling 方法

https://arxiv.org/abs/2510.14901

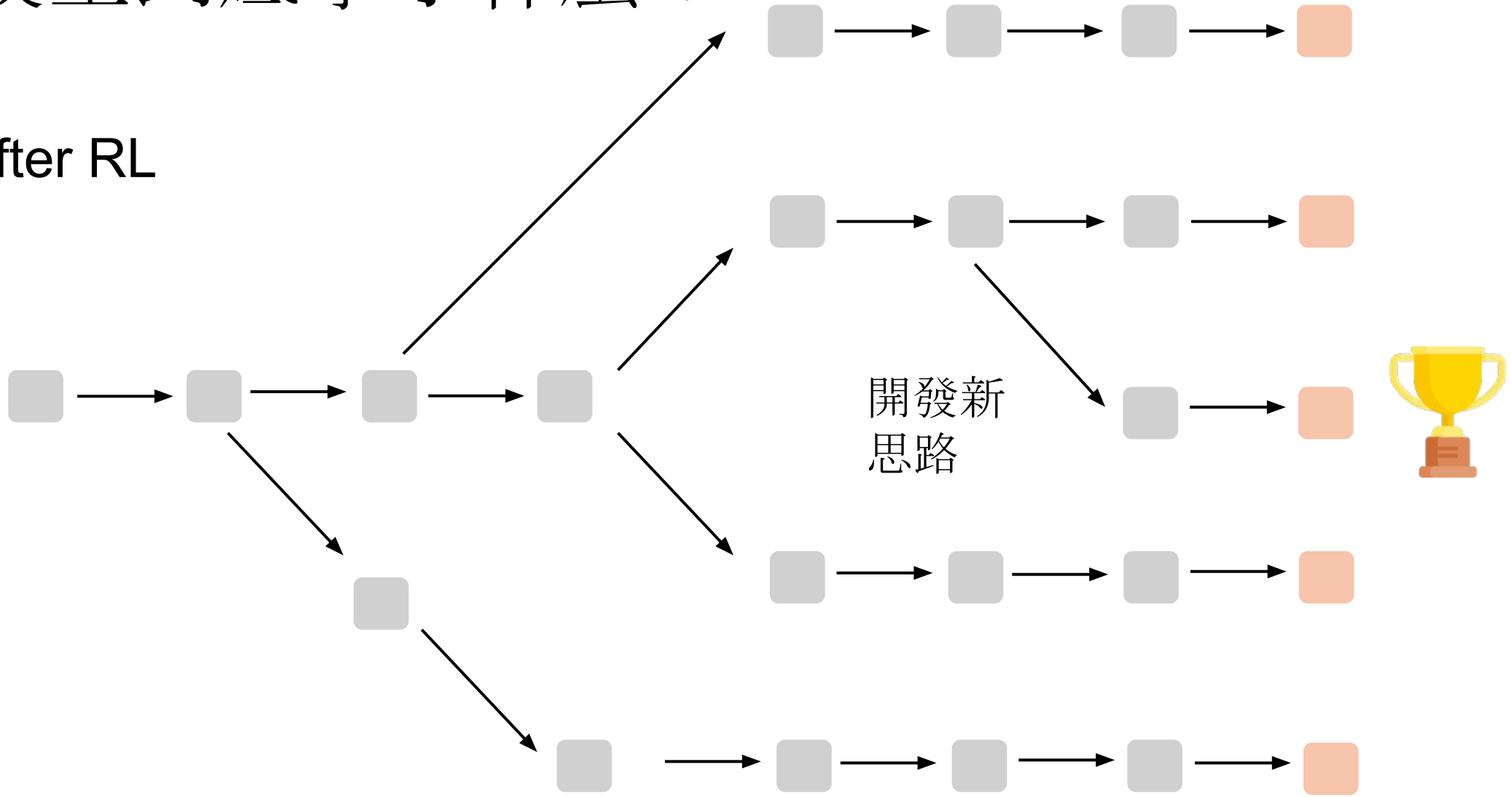
模型到底學了什麼？

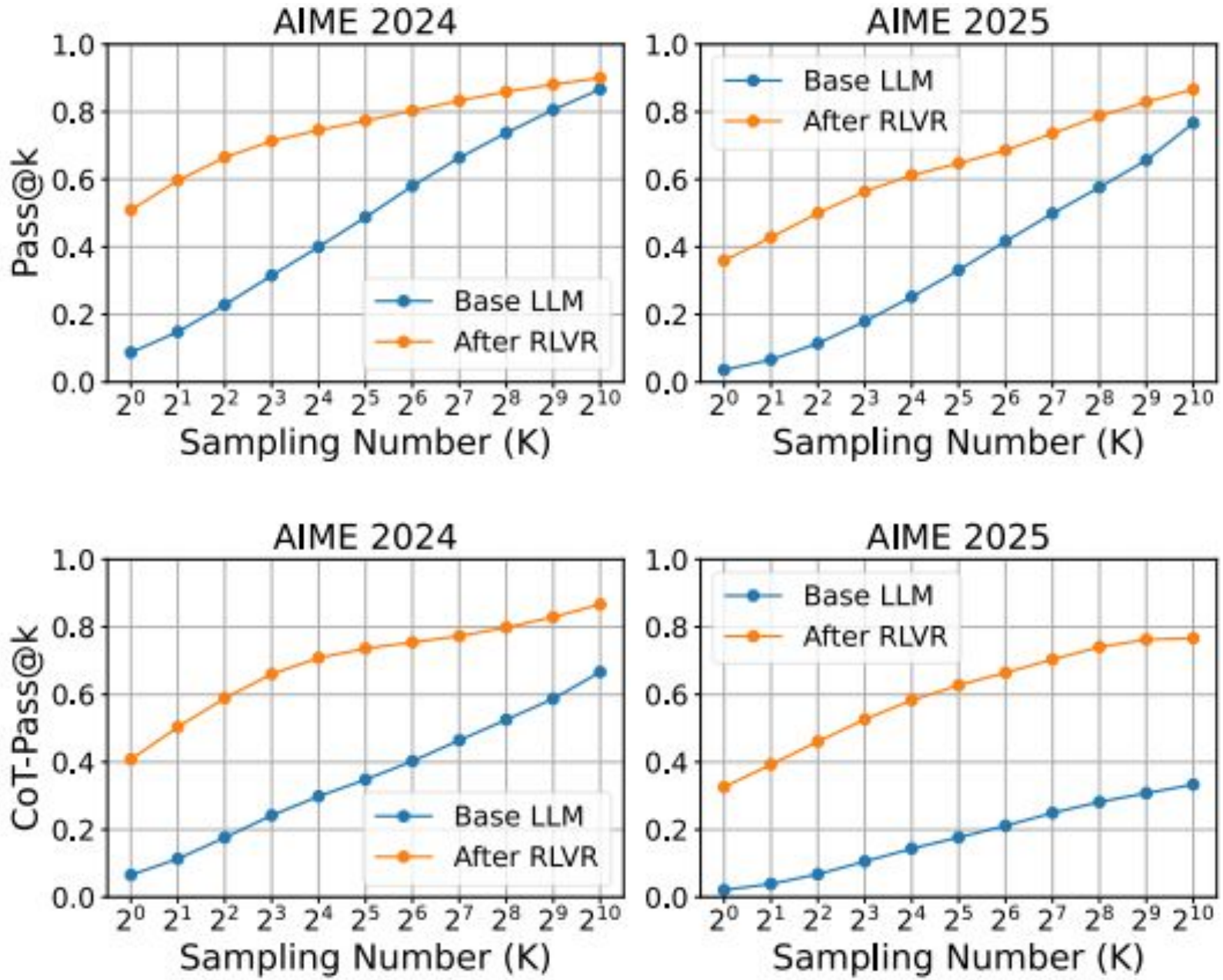
Before RL



模型到底學了什麼？

After RL





https://arxiv.org/pdf/2506.14245

Image from ChatGPT

LLM before RL Reasoning capabilities

The Debate on RLVR Reasoning Capability Boundary: Shrinkage, Expansion, or Both? A Two-Stage Dynamic View

https://arxiv.org/abs/2510.04028

Concluding Remarks

修改 Inference 過程

修改 Harness (Workflow)

修改 Model Parameters (Reasoning)