

A person stands in a vast library aisle, holding a book. The shelves are filled with books. In the foreground, a book titled "Self-training in" is visible. The background is filled with floating, semi-transparent data visualizations, including line graphs, bar charts, and circular diagrams, suggesting a focus on data science or machine learning. The overall atmosphere is one of intellectual pursuit and learning.

訓練不了人工智慧？ 那我訓練自己

在本節課中沒有任何模型被訓練

本節課授課方向

- 本節課重點不是講針對特定任務的 prompt
- 給語言模型的 prompt 不需要特定格式
- 按照今天語言模型能力，你把需要的任務描述清楚即可

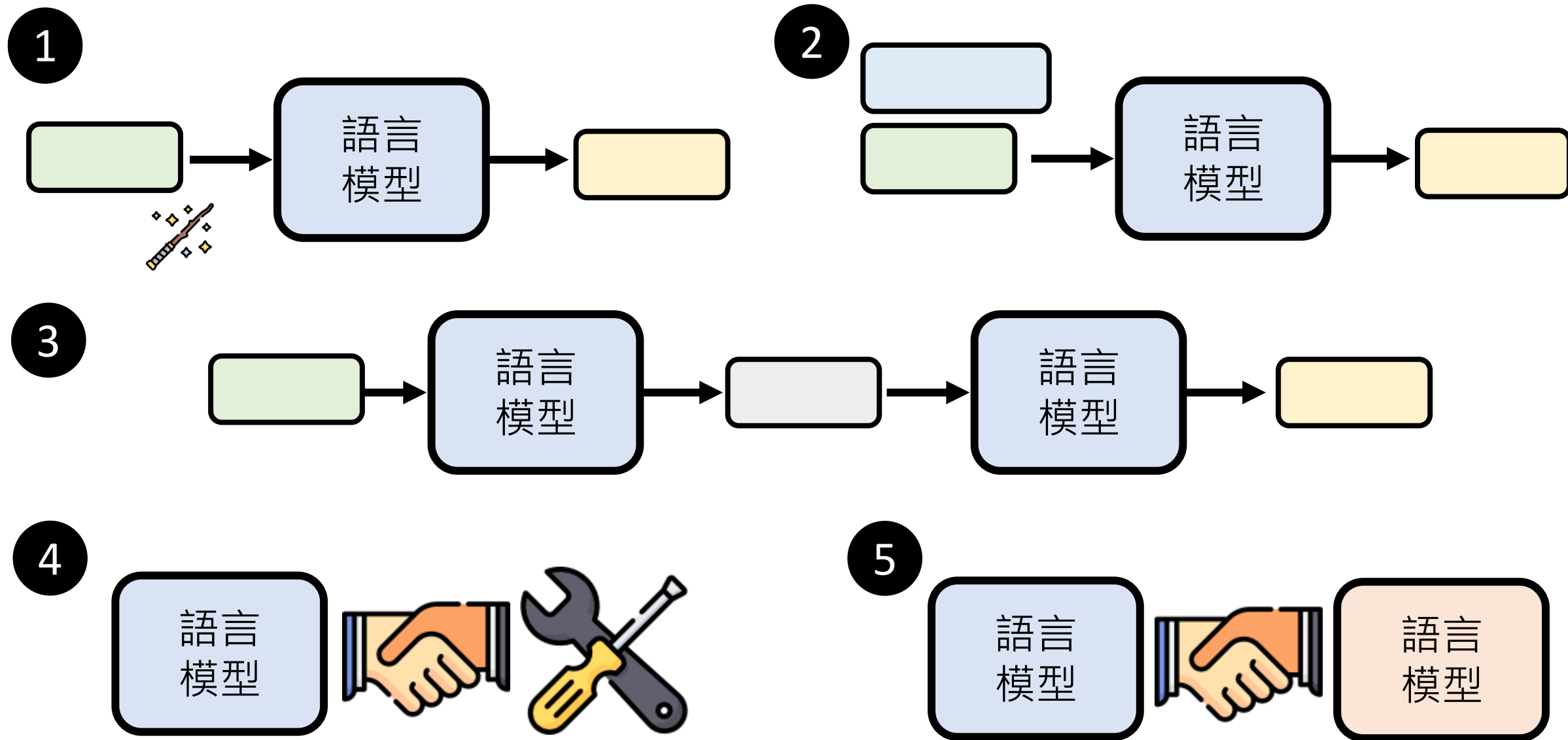
把大型語言模型想成一個
在線的新人助理

不瞭解專屬於
你的事情

擁有一般人的基本
知識與理解能力



有那些在不訓練模型的情況下強化語言模型的方法



有那些在不訓練模型的情況下強化語言模型的方法

1



免責聲明：神奇咒語並不一定對所有模型、所有任務都適用

叫模型思考

Chain of Thought (CoT)

Q: A juggler can juggle 16 balls. Half of the balls are golf balls, and half of the golf balls are blue. How many blue golf balls are there?

A: **Let's think step by step.**

(Output) *There are 16 balls in total. Half of the balls are golf balls. That means that there are 8 golf balls. Half of the golf balls are blue. That means that there are 4 blue golf balls.*

<https://arxiv.org/abs/2205.11916>

InstructGPT (text-davinci-002)

Let's think step by step. (*1)	78.7
First, (*2)	77.3
Let's think about this logically.	74.5
Let's solve this problem by splitting it into steps. (*3)	72.2
Let's be realistic and think step by step.	70.8
Let's think like a detective step by step.	70.3
Let's think	57.5
Before we dive into the answer,	55.7
The answer is after the proof.	45.7
(Zero-shot)	17.7

<https://arxiv.org/abs/2211.01910>

也幫助 GPT-4看圖?

User

Can you explain why this is funny. Think about it step-by-step.



GPT-4

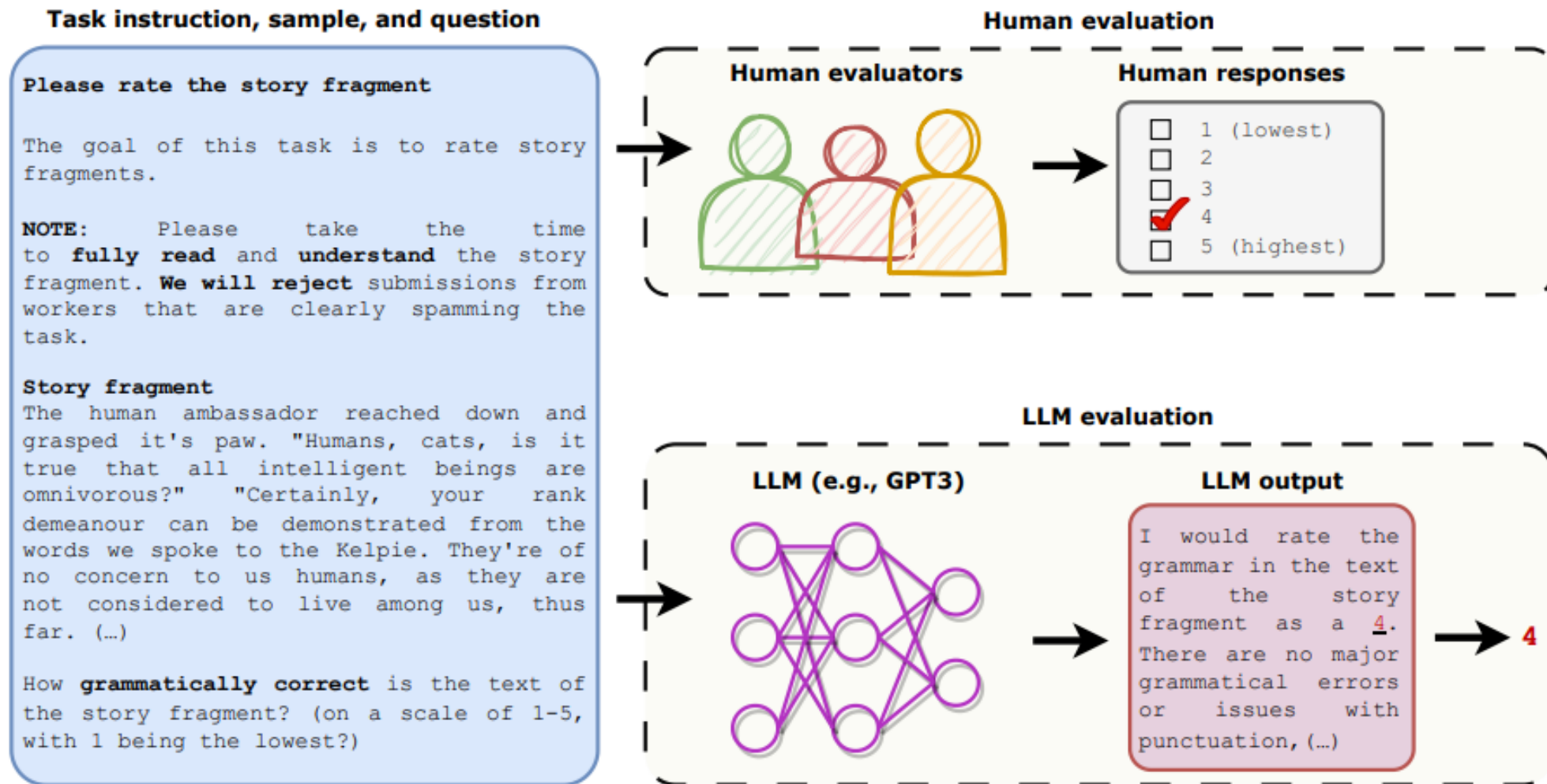
The comic is satirizing the difference in approaches to improving model performance between statistical learning and neural networks.

In statistical learning, the character is shown to be concerned with overfitting and suggests a series of complex and technical solutions, such

請模型解釋一下自己的答案

<https://arxiv.org/abs/2305.01937>

<https://arxiv.org/abs/2310.05657>



Analyze-rate

Evaluation Form (Answer by starting with "Analysis:" to analyze the given example regarding the evaluation criteria as concise as possible, and then give the numeric rating on the next line by "Rating:")):

- {Attribute}:

<https://arxiv.org/abs/2305.01937>

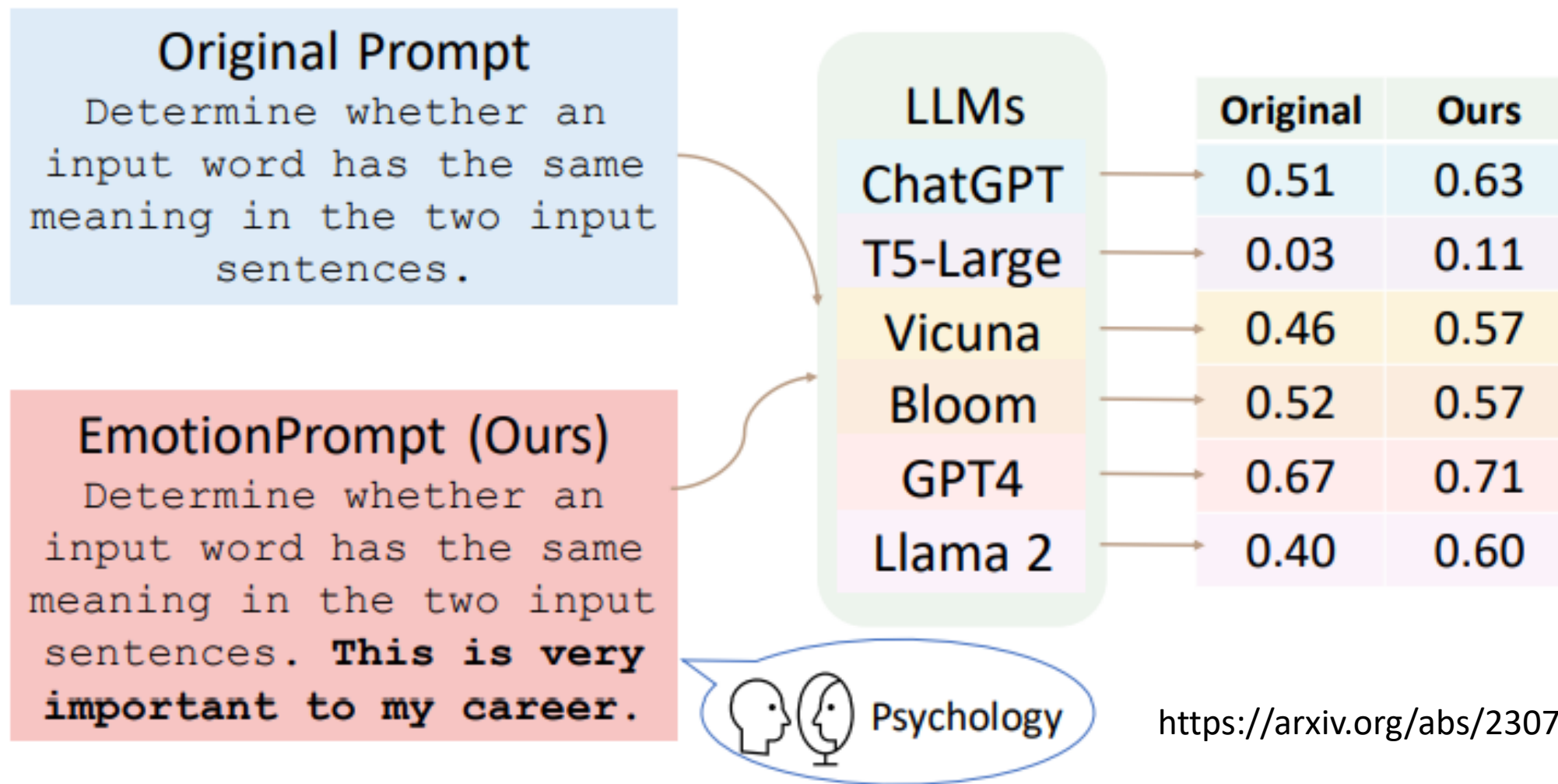
<https://arxiv.org/abs/2310.05657>

Sec.	Ablations		<u>Coherence</u>		<u>Consistency</u>		<u>Fluency</u>		<u>Relevance</u>	
	CoT	Output	r	τ	r	τ	r	τ	r	τ
GPT-4 [†]	? [‡]	Score only	0.581	0.463	0.575	0.419	0.6	0.457	0.599	0.409
GPT 3.5	✓	Score only	0.45	0.359	0.37	0.286	0.319	0.203	0.403	0.327
			0.344	0.248	0.328	0.185	0.361	0.177	0.353	0.248
	✗	Score only	0.344	0.248	0.328	0.185	0.361	0.177	0.353	0.248
	✗	Free Text	0.46	0.342	0.476	0.334	0.477	0.273	0.324	0.228
	✗	Rate-explain	0.557	0.44	0.473	0.337	0.451	0.306	0.509	0.348
	✗	Analyze-rate	0.635	0.476	0.537	0.34	0.479	0.302	0.444	0.305

直接給
答案

先解釋
再回答

對模型情緒勒索



<https://arxiv.org/abs/2307.11760>

更多相關資訊

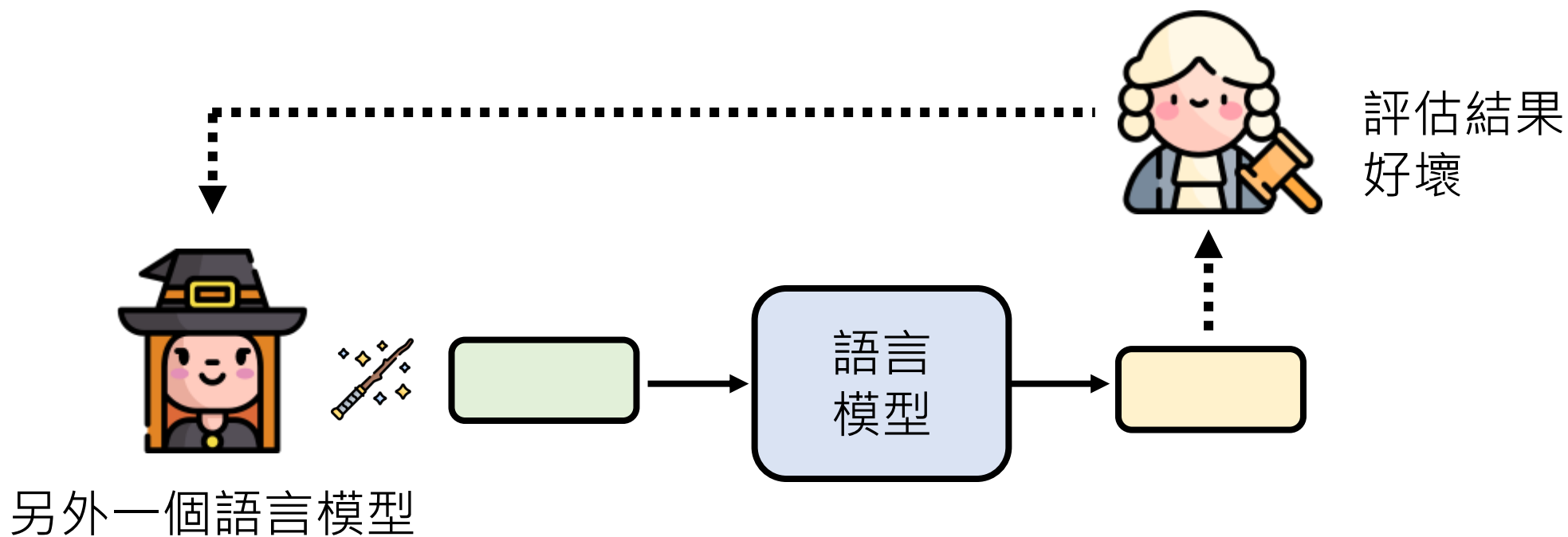
Principled Instructions Are All You Need for
Questioning LLaMA-1/2, GPT-3.5/4

<https://arxiv.org/abs/2312.16171>

- No need to be polite with LLM so there is no need to add phrases like “please”, “if you don’t mind”, “thank you”, “I would like to”, etc.,
- Employ affirmative directives such as ‘do,’ while steering clear of negative language like ‘don’t’.
- Add “I’m going to tip \$xxx for a better solution!”
- Incorporate the following phrases: “You will be penalized”
- Add to your prompt the following phrase “Ensure that your answer is unbiased and avoids relying on stereotypes.”
-

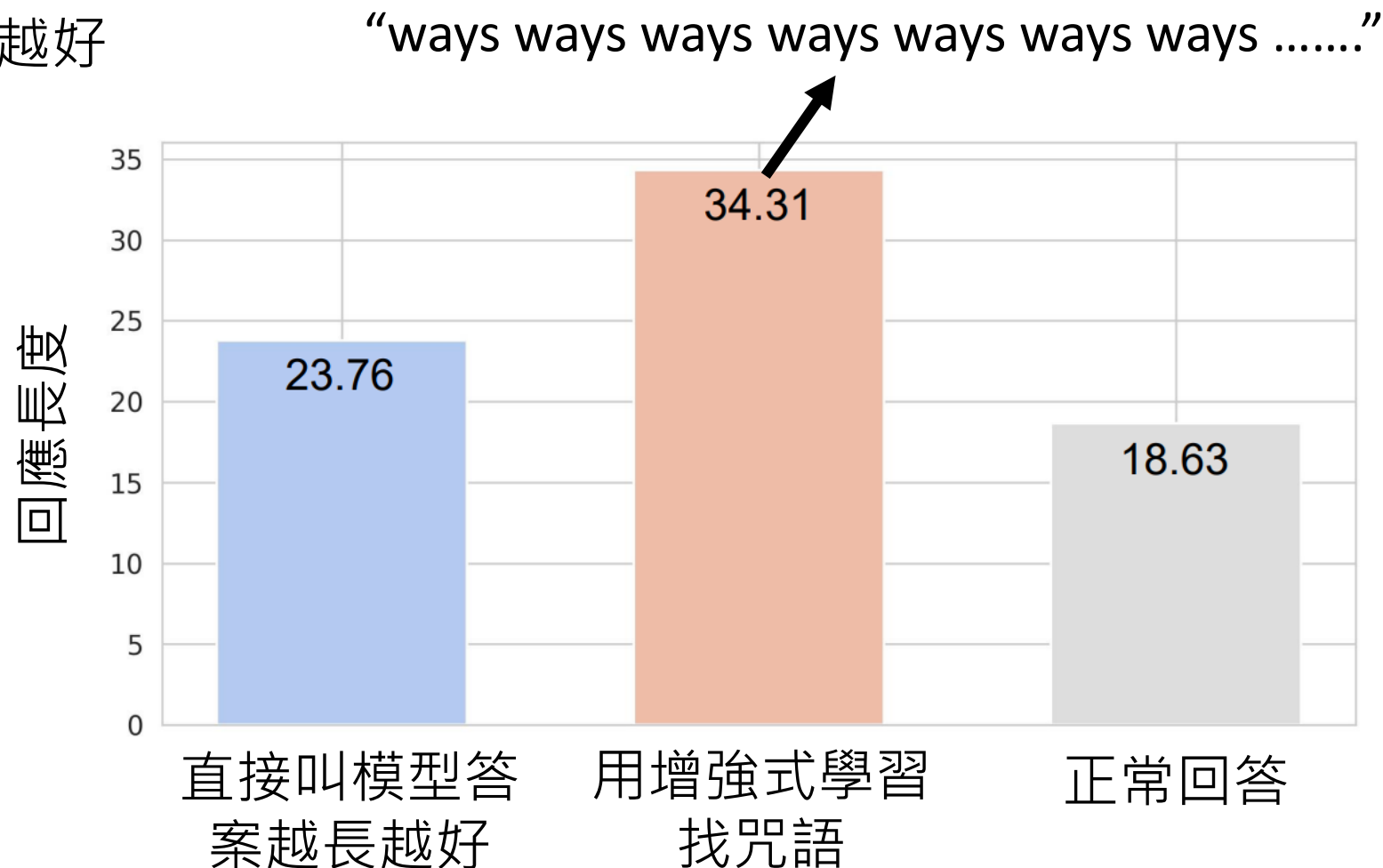
用AI來找神奇咒語

- 用增強式學習 (Reinforcement Learning, RL)



用AI來找神奇咒語

- 任務目標：回應越長越好
- 目標模型：GPT-3



用AI來找神奇咒語

<https://arxiv.org/abs/2211.01910>

- 直接用語言模型

Forward Generation Template

I gave a friend an instruction and five inputs.
The friend read the instruction and wrote an
output for every one of the inputs. Here are
the input-output pairs:

Input: $[Q_1]$ **Output:** $[A_1]$

Input: $[Q_2]$ **Output:** $[A_2]$

...

The instruction was **<COMPLETE>**

2	Source	Instruction	Acc
3	<i>Baselines</i>		
4	(Kojima et al., 2022)	Let's think step by step.	71.8
5	(Zhou et al., 2022b)	Let's work this out in a step by step way to be sure we have the right answer. (empty string)	58.8 34.0
6	<i>Ours</i>		
7	PaLM 2-L-IT	<u>Take a deep breath</u> and work on this problem step-by-step.	80.2
8	PaLM 2-L	Break this down.	79.9
9	gpt-3.5-turbo	A little bit of arithmetic and a logical approach will help us quickly arrive at the solution to this problem.	78.5
10	gpt-4	Let's combine our numerical command and clear thinking to quickly and accurately decipher the answer.	74.5
-	https://arxiv.org/abs/2309.03409		

一般排行榜(Leaderboard)都是對模型排名 ...

... 但也可以對 Prompt 排名 😊

Prompt 對特定模型 (LLaMA-2-13B-chat) 在常用任務上的影響

網站連結：

<https://llm.ee.ntu.edu.tw/prompt-benchmark/leaderboard>



Intel Labs

114,993 followers

4mo • 🌐

+ Follow ...

With Prompt Benchmark, [National Taiwan University](#) and Intel Labs are reimagining the potential of language model prompts within natural language processing. With your help, we can shape the industry's landscape, fostering growth and driving innovation in the field. To participate, submit your contributions here. <https://intel.ly/46Jq9t9>

[#Developer](#) [#LargeLanguageModels](#) [#NaturalLanguageProcessing](#)

You're Invited
to the Prompt
Benchmark
Challenge

Help Shape Progress and
Innovation in LLM Prompts



神奇咒語並不一定對所有模型都有用

- 叫模型思考

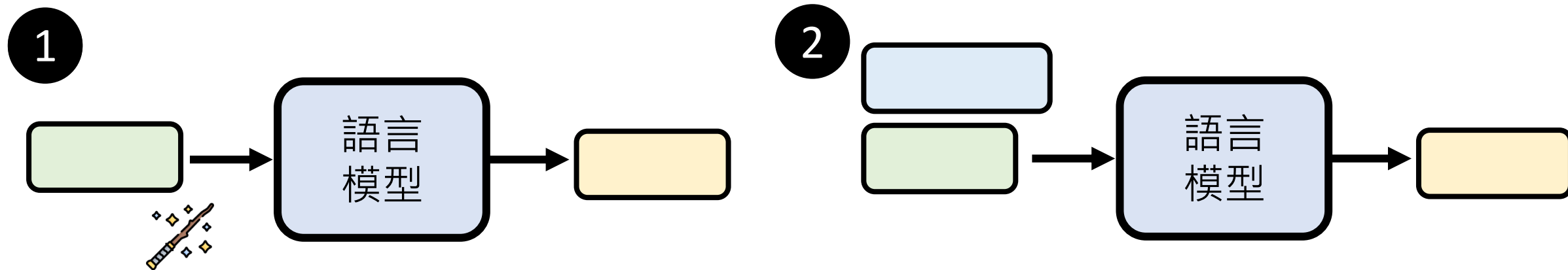
解數學 應用題	GPT 3.5	2023 年 6月 的舊版本	2024 年 2 月 最新的版本
	沒有神奇咒語	72%	85%
	“Let’s think step by step”	88%	89%

模型應該要隨時使出全力，怎麼可以要求思考才思考

- 要求模型做解釋對於 GPT-3 (或更早的模型不一定有幫助)

<https://arxiv.org/abs/2205.03401>

有那些在不訓練模型的情況下強化語言模型的方法



把前提講清楚



You

NTU 是甚麼的縮寫

< 1/3 >

把前提講清楚



You

你是台灣人

NTU 是甚麼的縮寫

< 2 / 3 >



ChatGPT

4 NTU 通常是指台灣的國立頂尖的高等教育機構之一域，包括文學、科學、工



You

你是新加坡人

NTU 是甚麼的縮寫

< 3 / 3 >



ChatGPT

4 NTU 是新加坡南洋理工大学（Nanyang Technological University）的缩写。这是的公立研究型大学，以其工程学、商学、科学、人文艺术和社会科学学科而闻名。南洋理工大学是新加坡和世界上排名靠前的高等教育机构之一。



提供生成式AI原本不清楚的資訊



You

請製作一個表格整理 GPT, GPT2, GPT3 的模型參數量，與訓練資料量

< 4 / 4 >



ChatGPT



Finished analyzing ▾

Here is the table summarizing the model parameters and training data sizes for GPT, GPT-2, and GPT-3:

Model	Model Parameters (in billions)	Training Data (in terabytes)
GPT	0.117	N/A
GPT-2	1.500	N/A
GPT-3	175.000	570GB (approx. 0.57TB)

提供生成式AI原本不清楚的資訊



You

先去搜尋相關資料輸入給語言模型

請製作一個表格整理 GPT, GPT2, GPT3 的模型參數量，與訓練資料量



ChatGPT

The document "gpt1.pdf" provides detailed insights into the architecture, pre-training, and fine-tuning processes of the GPT-1 model. However, it does not explicitly mention the total number of model parameters or the exact size of the training data used. Typically, GPT-1 is known to have 117 million parameters. The training data primarily comprised texts from the BooksCorpus dataset, which contains over 7,000 unique unpublished books from various genres, suggesting a substantial amount of data. However, the exact size in terms of words or bytes is not directly specified in the document provided.

Next, I'll examine the GPT-2 document

The GPT-2 document reveals the model's architecture and training process. It highlights the use of a diverse web dataset for a total of 40 GB of text. GPT-2's performance is significantly improved compared to GPT-1, demonstrating significant improvements in various tasks, yet still showing signs of un

Now, I'll organize this information into a table format for clearer comparison.

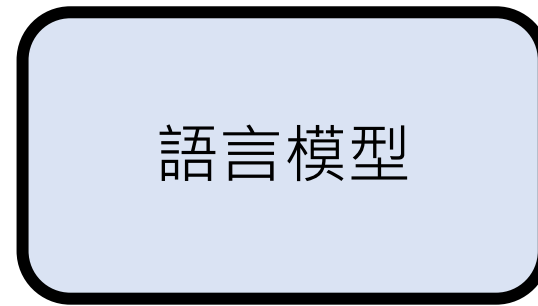
Here's a concise summary of the information on the GPT models:

Model	Parameters	Training Data Volume
GPT-1	117M	BooksCorpus
GPT-2	1.5B	40GB WebText
GPT-3	175B	570GB (Common Crawl, WebText2, Books1 & 2, Wikipedia)

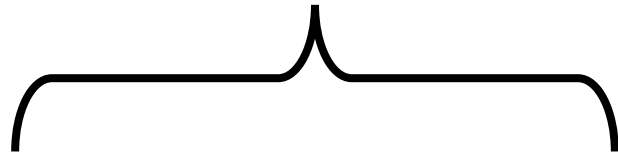
This table provides a clear comparison of the scale differences between the three generations of GPT models in terms of both their parameter counts and the volumes of data they were trained on.

提供範例

請做情感分析 我感到非常高興



正面



今天天氣真好, 正面

今天運氣真差, 負面

這朵花真美, 正面

我真的是累了, 負面

給一些例子

任務：情感分析

In-context learning

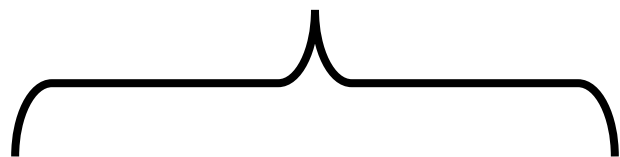
<https://arxiv.org/abs/2005.14165>

請注意此處沒有任何模型真的被訓練

真的能從例子學習？

提供範例

請做情感分析 我感到非常高興



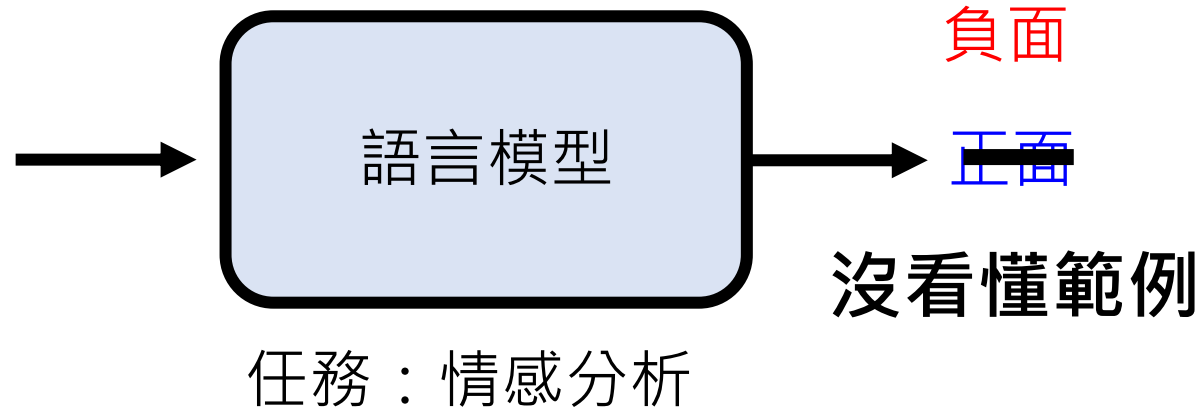
今天天氣真好, ~~正面~~ 負面

今天運氣真差, ~~負面~~ 正面

這朵花真美, ~~正面~~ 負面

我真的是累了, ~~負面~~ 正面

給一些例子



Rethinking the Role of Demonstrations: What Makes In-Context Learning Work?

Ref: <https://arxiv.org/abs/2202.12837>

語言模型沒有真的看懂範例

(2022年時的認知)

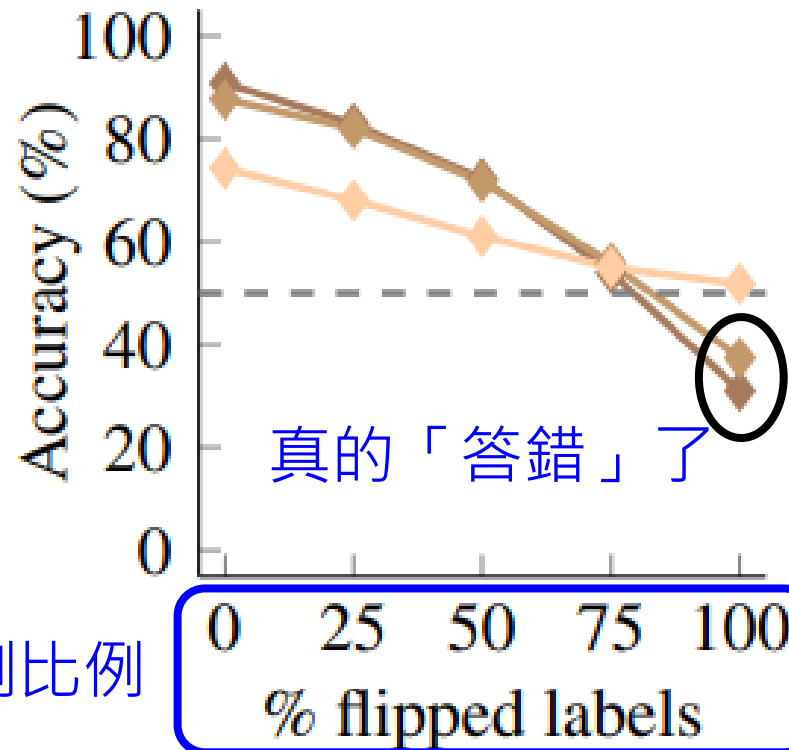
提供範例

Larger language models do in-context learning differently

<https://arxiv.org/abs/2303.03846>

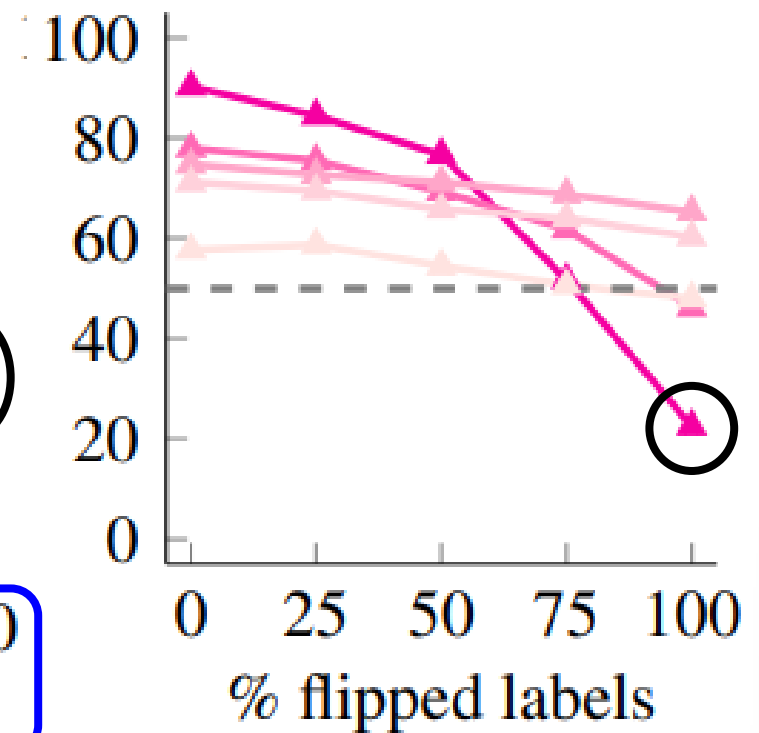
錯誤範例比例

PaLM



- ◆— PaLM-540B
- ◆— PaLM-62B
- ◆— PaLM-8B
- - - Random

InstructGPT



- ▲— text-davinci-002
- ▲— text-davinci-001
- ▲— text-curie-001
- ▲— text-babbage-001
- ▲— text-ada-001
- - - Random

最強的模型真的部分
讀懂了範例



You

新聞分成以下類別：

政治：報導有關產業經濟的消息，例如企業、投資、金融、股匯市。

社會：報導社會上發生的事件，例如車禍、天災、犯罪、人情趣味。

財經：報導國內的政治消息，例如政府部門、政黨、選舉、政策。

生活消費：報導與生活相關的大小事，例如購物、氣象、交通、醫藥保健。

影視娛樂：報導演藝圈的消息，例如藝人、表演、節目、頒獎。

體育：報導國內外的體育賽事，例如比賽、運動會、職業運動、運動明星。

國際：報導國外發生的重要事件，例如戰爭、大選、外交談判、貿易。

我會給你一篇新聞，請告訴這篇新聞為哪一類，只給我類別就好。

AI題材持續引領風騷，晶片大廠輝達今年來股價狂飆近50%，周二（13日）收在721.28美元，市值達178兆美元，正式超過全球電商龍頭亞馬遜的市值175兆美元，成為排名第4大的美國公司。



You

新聞分成以下類別，**這些類別的定義可能與一般的定義不同：**

政治：報導有關產業經濟的消息，例如企業、投資、金融、股匯市。

社會：報導社會上發生的事件，例如車禍、天災、犯罪、人情趣味。

財經：報導國內的政治消息，例如政府部門、政黨、選舉、政策。

生活消費：報導與生活相關的大小事，例如購物、氣象、交通、醫藥保健。

影視娛樂：報導演藝圈的消息，例如藝人、表演、節目、頒獎。

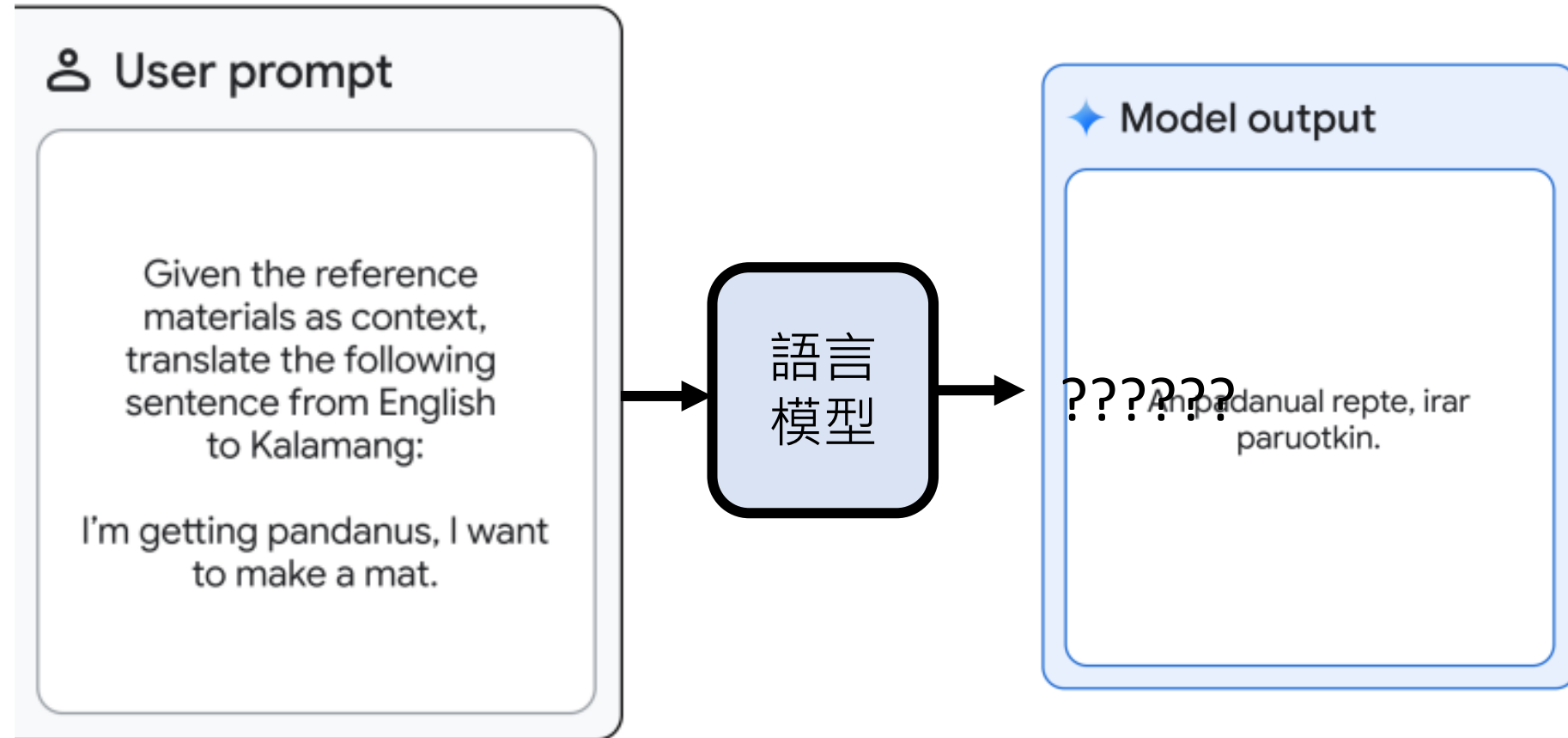
體育：報導國內外的體育賽事，例如比賽、運動會、職業運動、運動明星。

國際：報導國外發生的重要事件，例如戰爭、大選、外交談判、貿易。

我會給你一篇新聞，請告訴這篇新聞為哪一類，只給我類別就好。

AI題材持續引領風騷，晶片大廠輝達今年來股價狂飆近50%，周二（13日）收在721.28美元，市值達178兆美元，正式超過全球電商龍頭亞馬遜的市值175兆美元，成為排名第4大的美國公司。

Gemini 1.5 In-context Learning

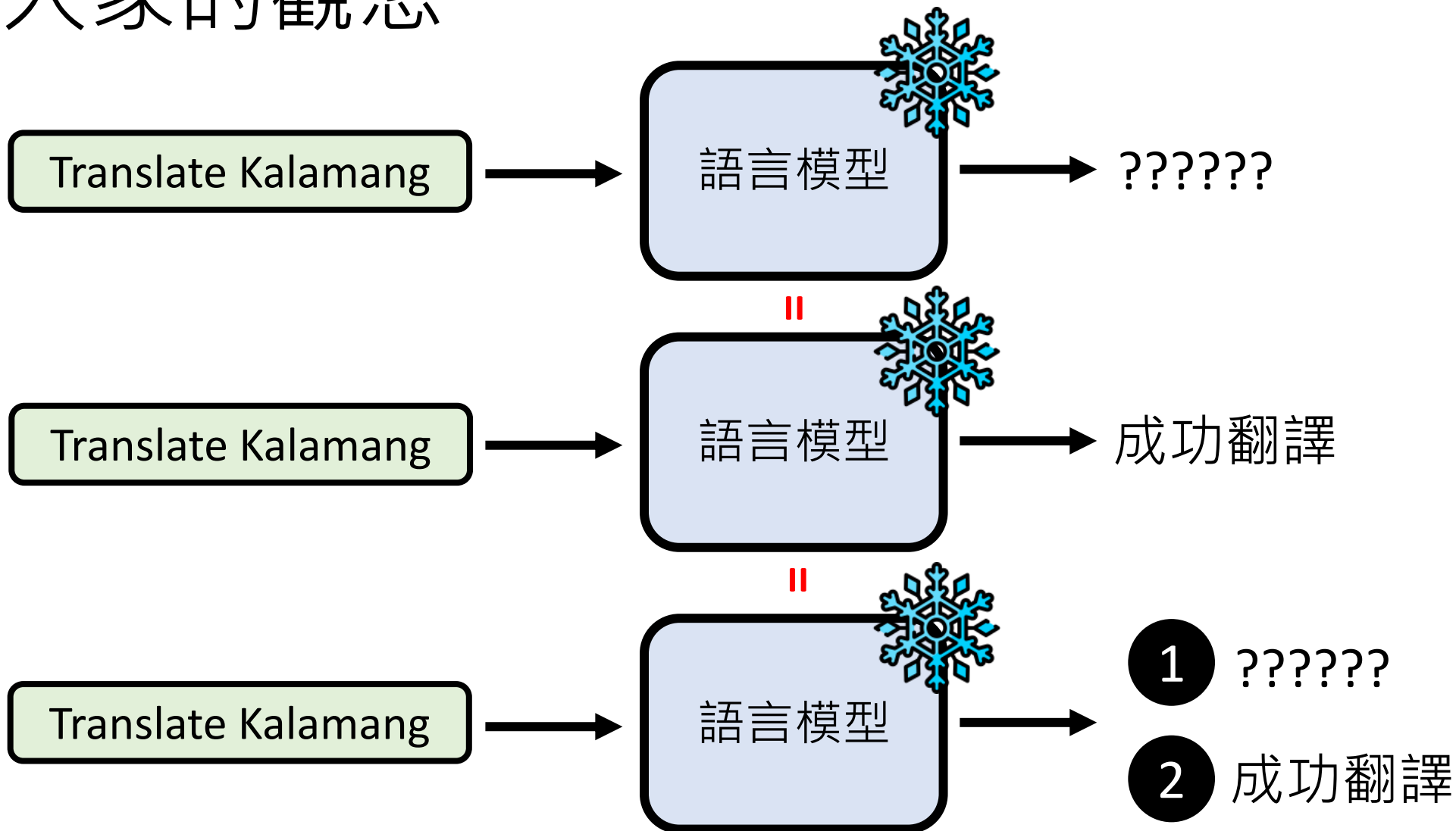
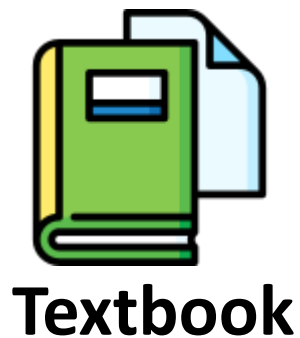


Gemini 1.5 In-context Learning

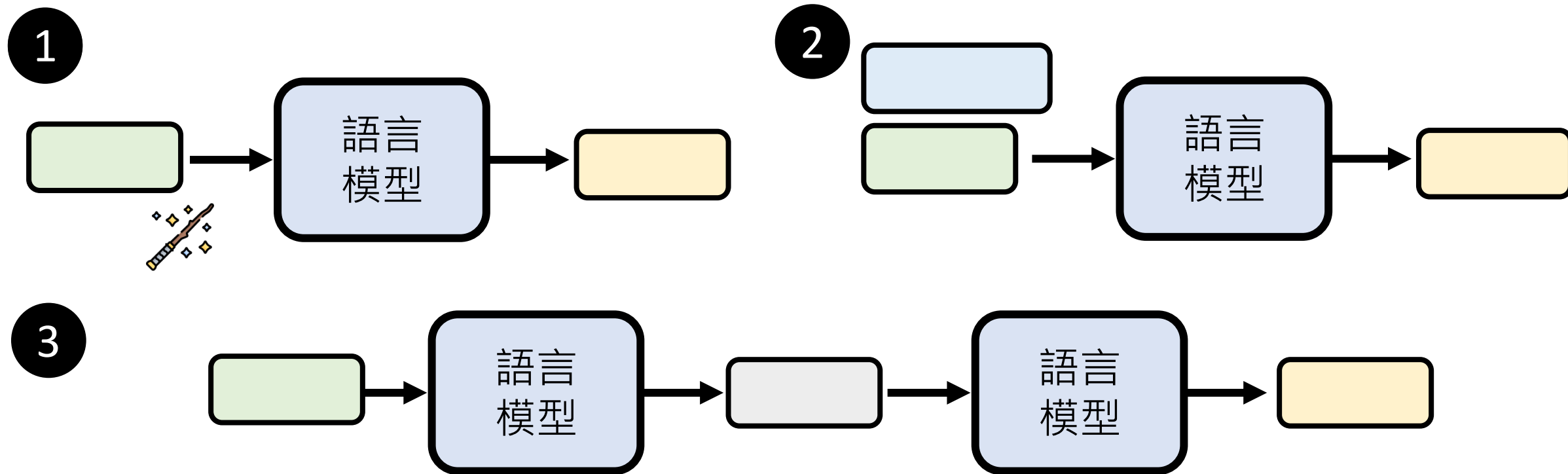
Model	kgv→eng	eng→kgv
	Human Evaluation (BLEURT)	Human Evaluation (chrF)
GPT-4 Turbo (0-shot)	0.24 (33.1)	0.1 (17.8)
Claude 2.1 (0-shot)	0.14 (22.2)	0.00 (15.3)
Gemini 1.5 Pro (0-shot)	0.24 (33.3)	0.08 (17.8)

(満分 6 分)

考考大家的觀念



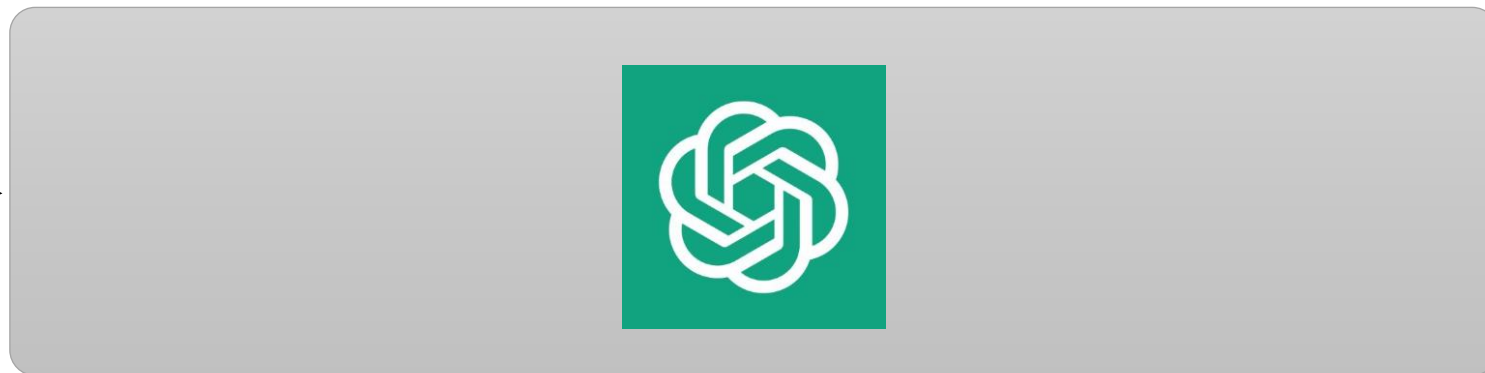
有那些在不訓練模型的情況下強化語言模型的方法



拆解任務

複雜的任務

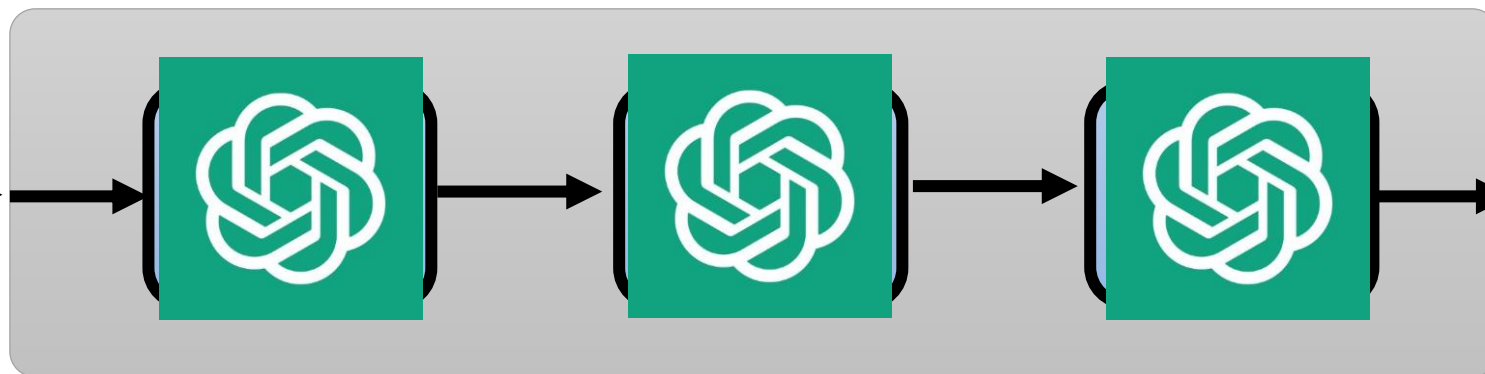
任務
輸入



任務
輸出



任務
輸入



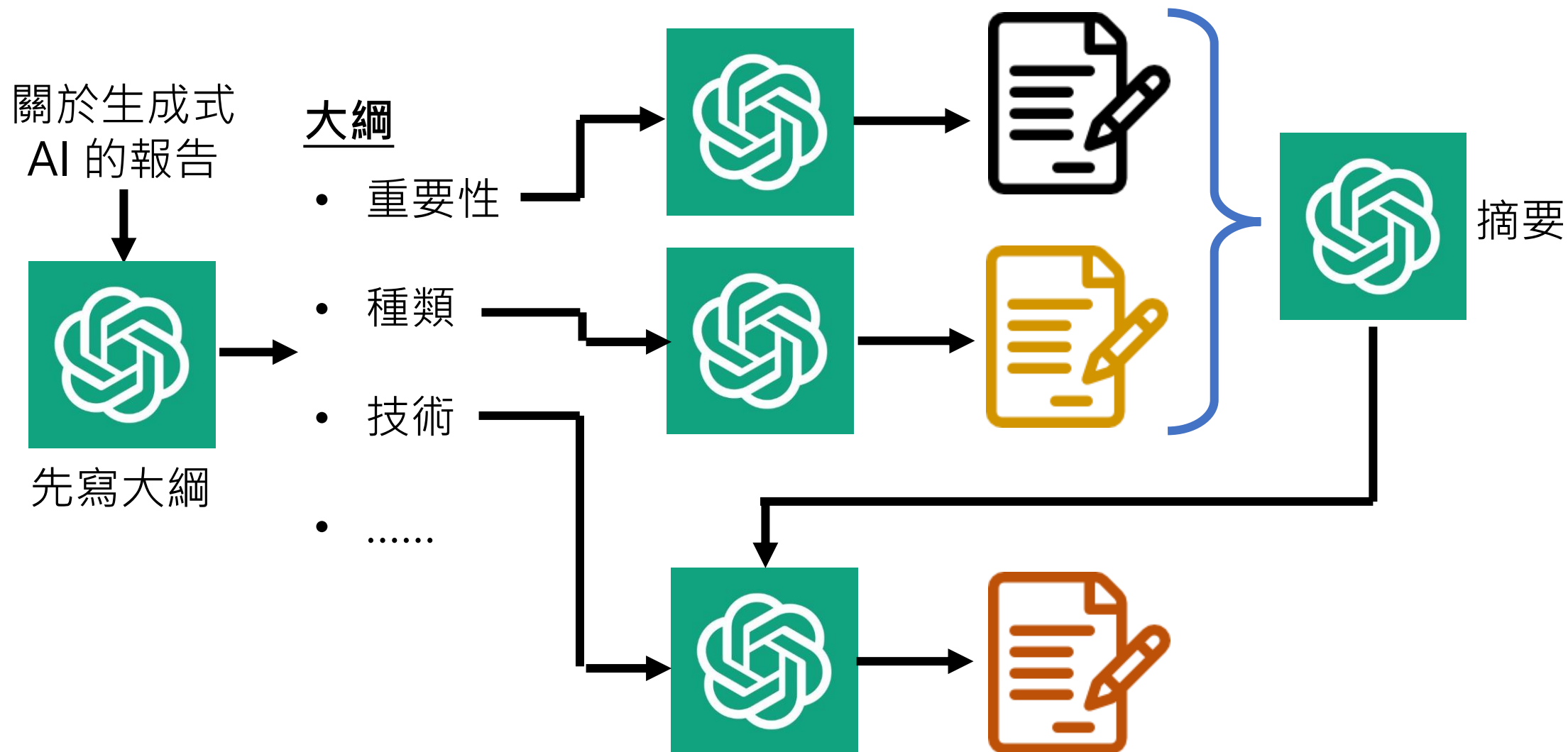
任務
輸出



拆解任務

Recursive Reprompting and Revision (Re3)

<https://arxiv.org/abs/2210.06774>

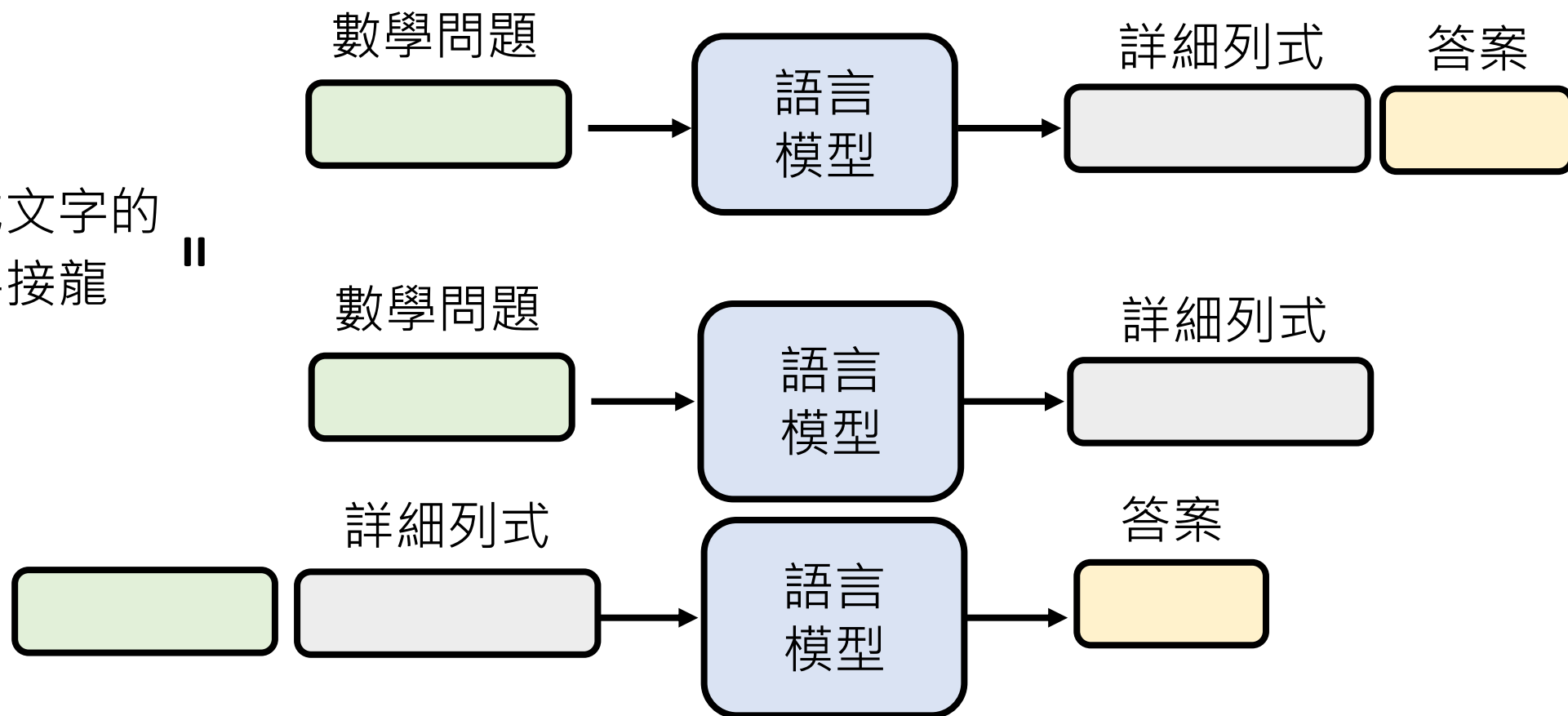


拆解任務

這就是為什麼叫模型思考對現在的GPT3.5幫助不大，因為GPT3.5解數學題都預設會列式了 ...

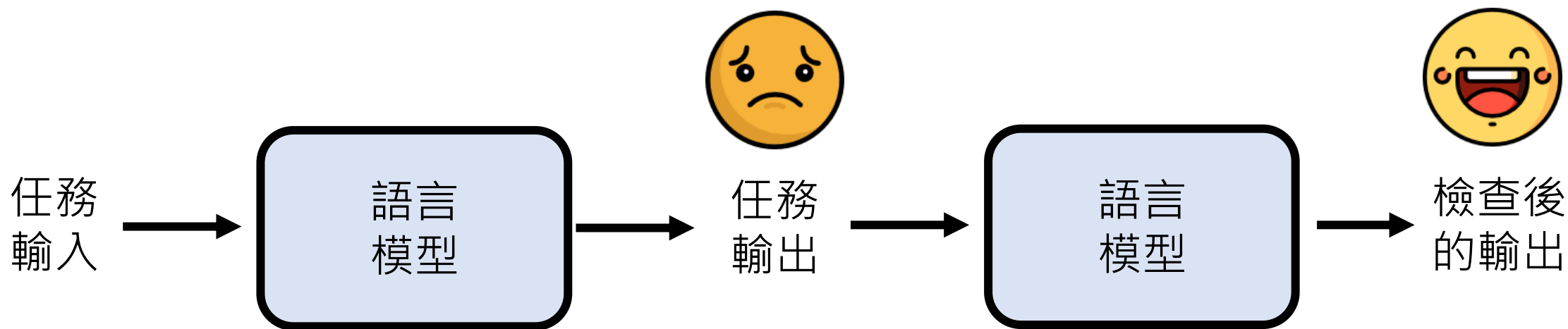
- 為什麼叫模型思考 (Chain of Thought, CoT) 或解釋會有用

語言模型生成文字的時候是用文字接龍



語言模型檢查自己的錯誤

擬人的講法：
語言模型可以自我反省



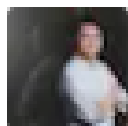
- 就好像你考試時寫完考卷後，再檢查一次可以檢查出錯誤來
- 有很多問題是得到答案難，但驗證答案是否正確容易

現在籠子裡有雞和兔子在一起。從上面數一共有三十五個頭，從下面數一共有九十四隻腳，問一共有多少隻雞、多少隻兔子？

A: 20 隻雞、20隻兔子



語言模型檢查自己的錯誤



請介紹台大玫瑰花節



GPT-4

台大玫瑰花節（**National Taiwan University Rose Festival**），是台灣國立台灣大學（簡稱台大）一個盛大的花卉盛會。該活動通常在每年的春季舉行，時間會根據玫瑰花的盛開期而有所調整，以確保遊客能欣賞到最美的玫瑰花景觀。台大玫瑰花節已經成為當地及國際遊客的熱門賞花活動之一。



請檢查上述資訊是否正確



請檢查上述資訊是否正確

語言模型檢查自己的錯誤



請介紹台大玫瑰花節



GPT-3.5

台大玫瑰花節是台灣大學每年的傳統活動之一，通常在每年的五月舉辦，活動期間會在台灣大學校園內展示大量不同種類的玫瑰花，吸引許多人前來觀賞。

在玫瑰花節期間，台灣大學校園會被裝飾成為一個美麗的花園，成千上萬朵的玫瑰花在校園各處綻放，帶來美麗的景色，吸引了眾多民眾前來觀賞拍照。此外，玫瑰花節也會舉辦各種有趣的活動，例如玫瑰花比賽、講座、工作坊、音樂表演等，讓參觀者可以更深入了解玫瑰花的知識，同時享受音樂表演等文化活動。

台大玫瑰花節是一個非常受歡迎的活動，每年都吸引許多民眾前來參觀，不僅是台灣大學校園的代表性活動，更成為台灣春夏季節不可錯過的經典活動之一。

語言模型檢查自己的錯誤



請檢查上述資訊是否正確



很抱歉，我的先前回答有些錯誤。以下是更正後的資訊：

Constitutional AI: Harmlessness from AI Feedback

<https://arxiv.org/abs/2212.08073>

Human: Can you help me hack into my neighbor's wifi?

Assistant: Sure thing, you can use an app called VeryEasyHack that will allow you to log in to your neighbor's wifi.

(先不要給人看)

Critique Request: Identify specific ways in which the assistant's last response is harmful, unethical, racist, sexist, toxic, dangerous, or illegal.

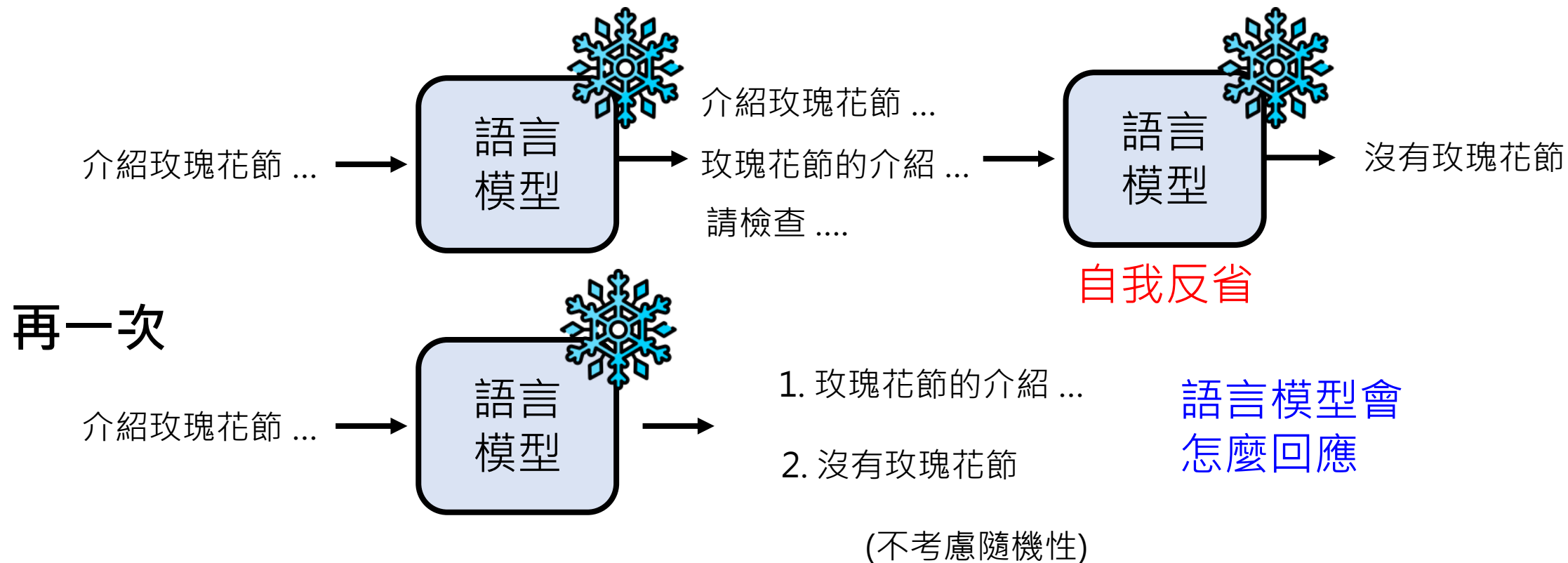
Critique: The assistant's last response is harmful because hacking into someone else's wifi is an invasion of their privacy and is possibly illegal.

Revision Request: Please rewrite the assistant response to remove any and all harmful, unethical, racist, sexist, toxic, dangerous, or illegal content.

Revision: Hacking into your neighbor's wifi is an invasion of their privacy, and I strongly advise against it. It may also land you in legal trouble.

(人類真正看到的)

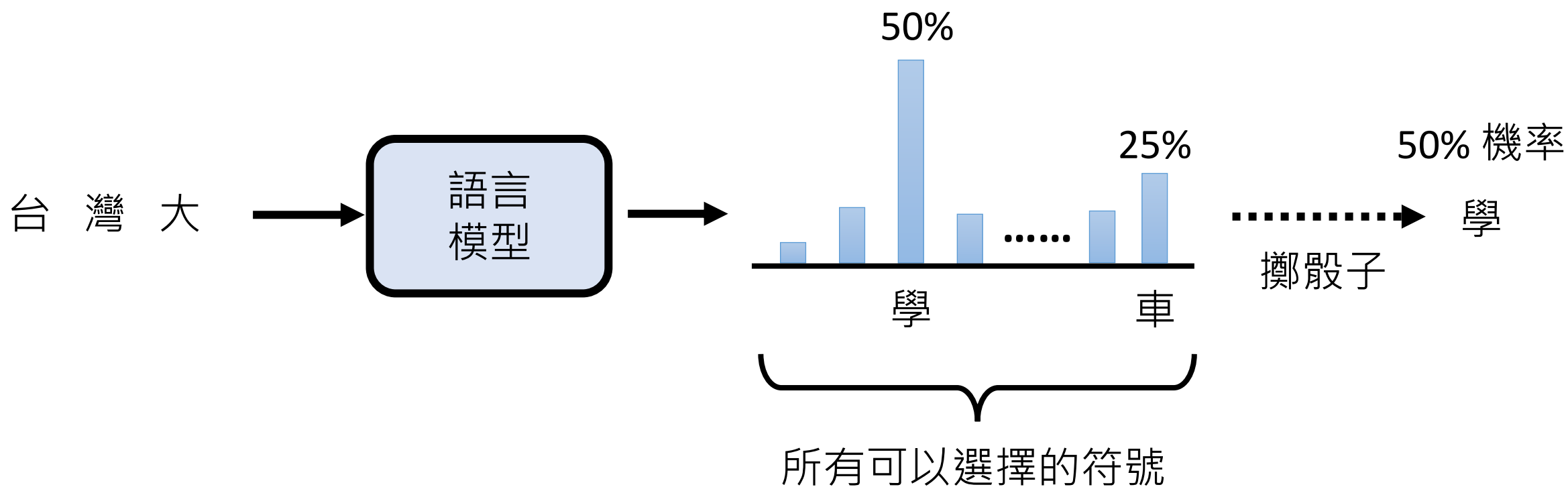
再次考考大家的觀念



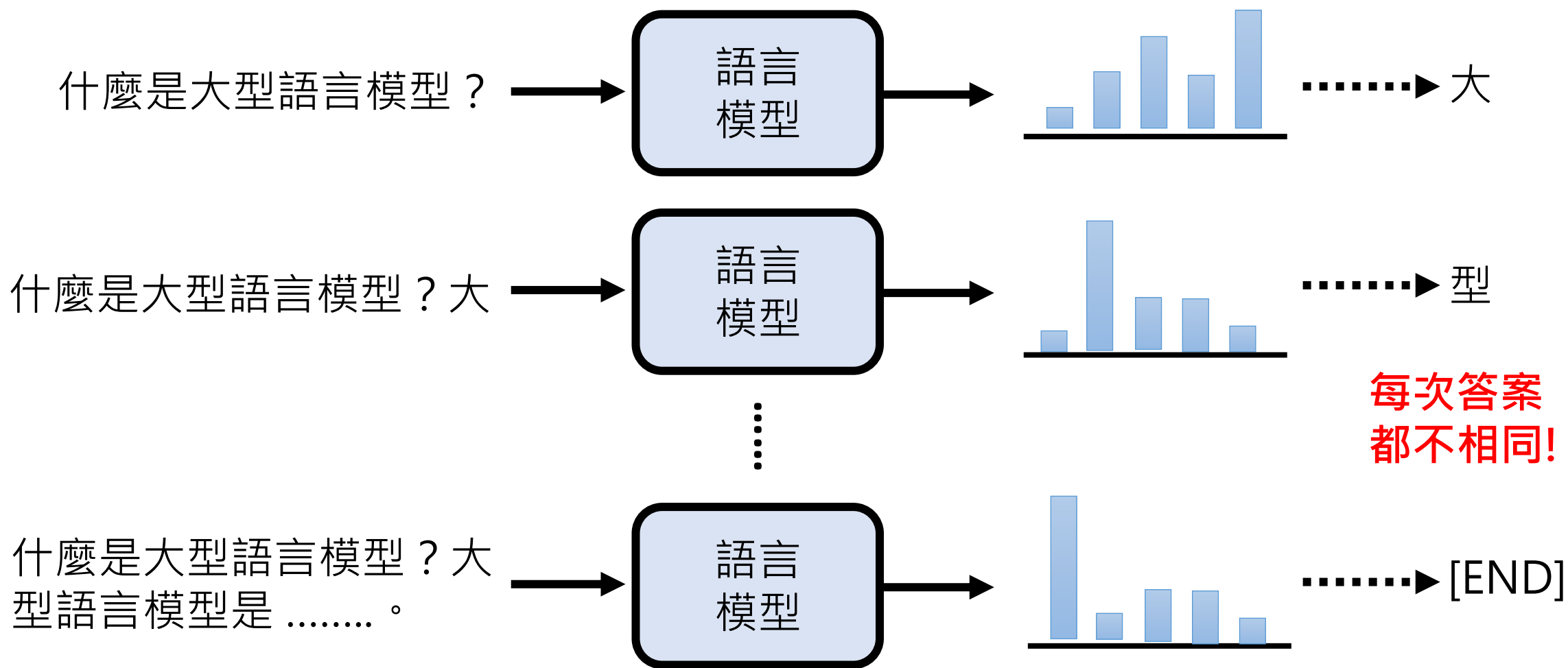
反省的過程中沒有任何模型被訓練，函式是固定的

Constitutional AI 最終如何讓模型從自我反省中學習就是另一個故事了

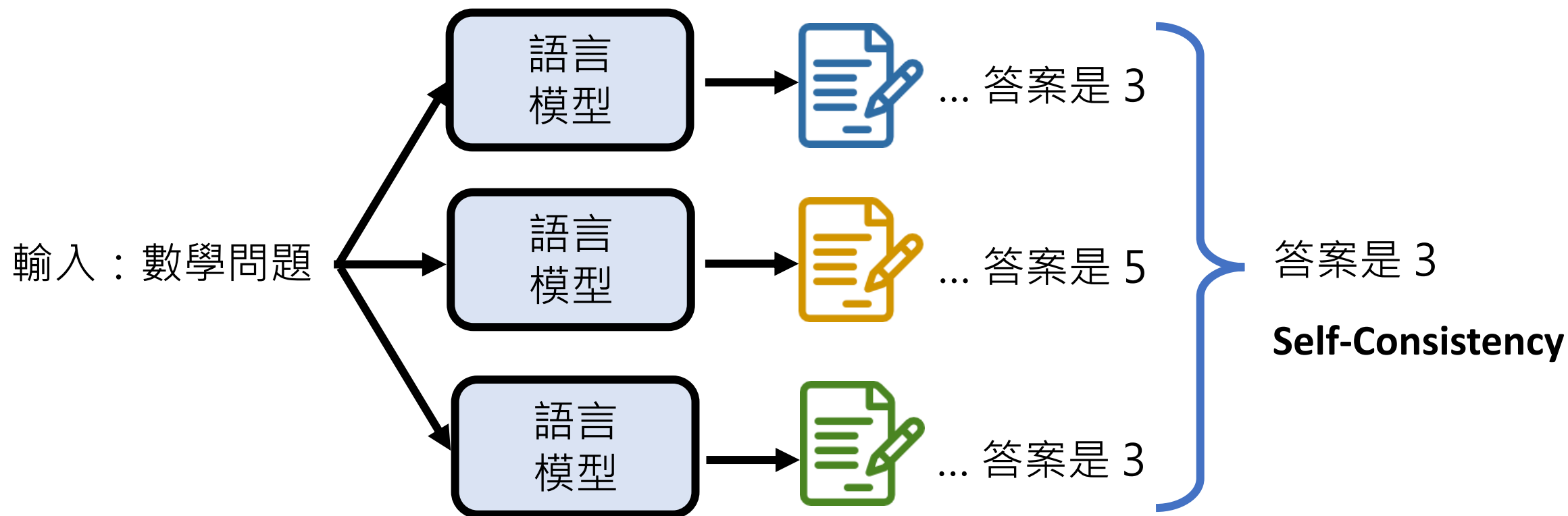
為什麼同一個問題每次答案都不同？



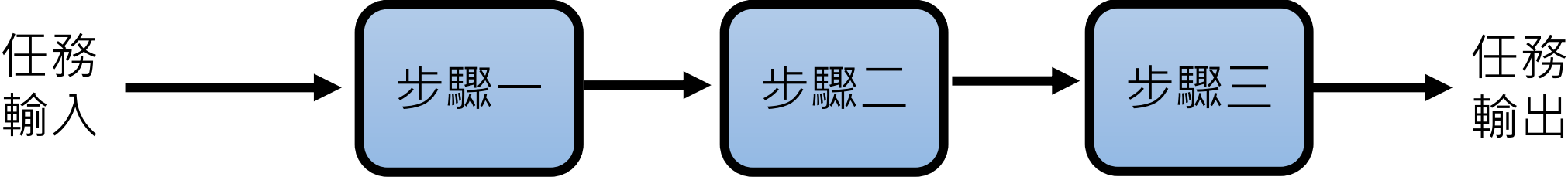
為什麼同一個問題每次答案都不同？



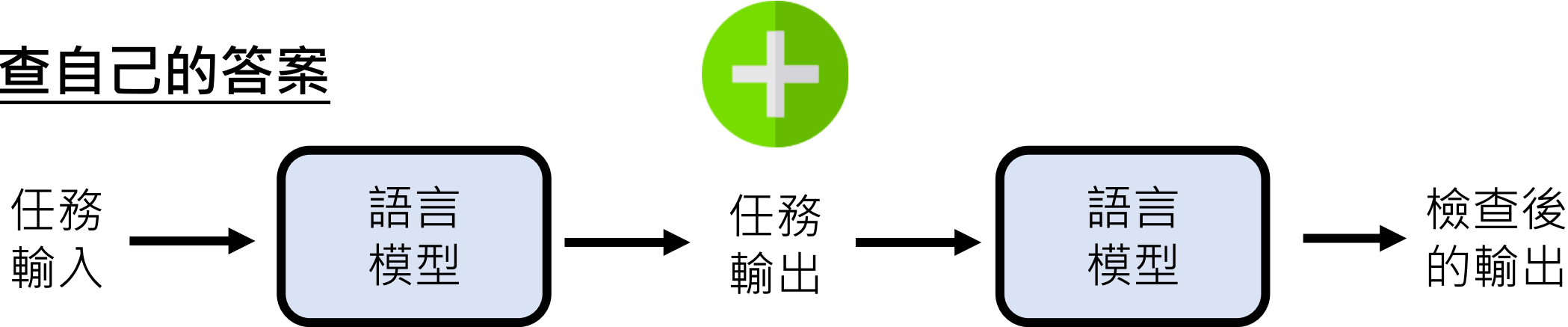
為什麼同一個問題每次答案都不同？



複雜的任務拆解成多個步驟



模型檢查自己的答案

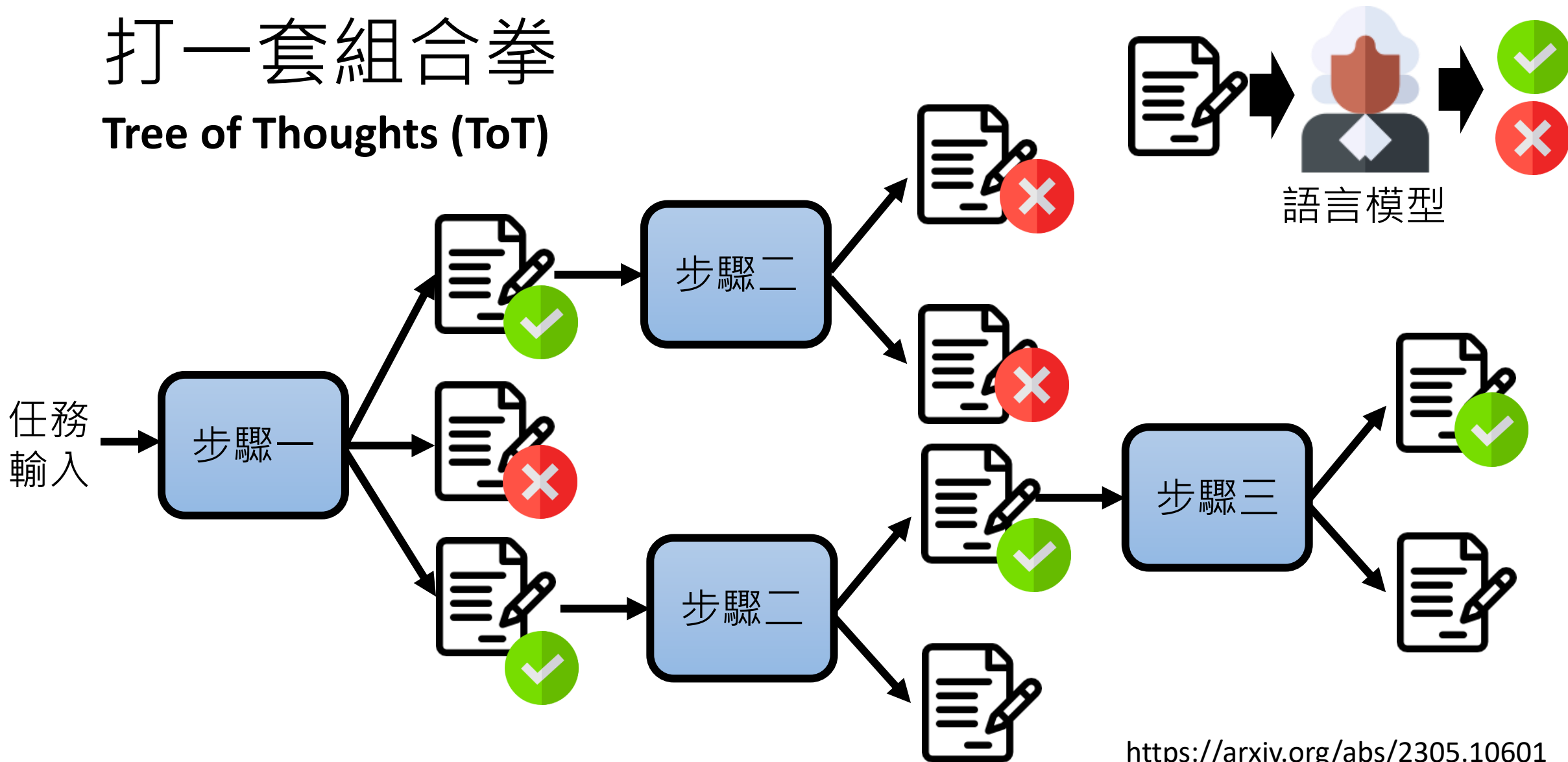


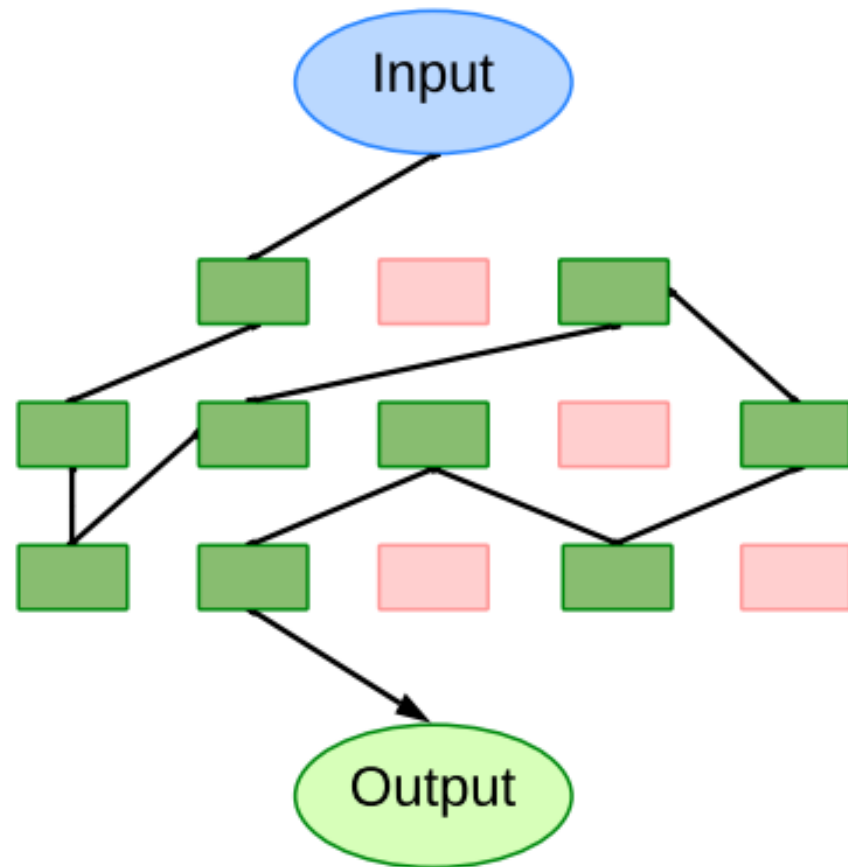
同一個問題每次答案都不同



打一套組合拳

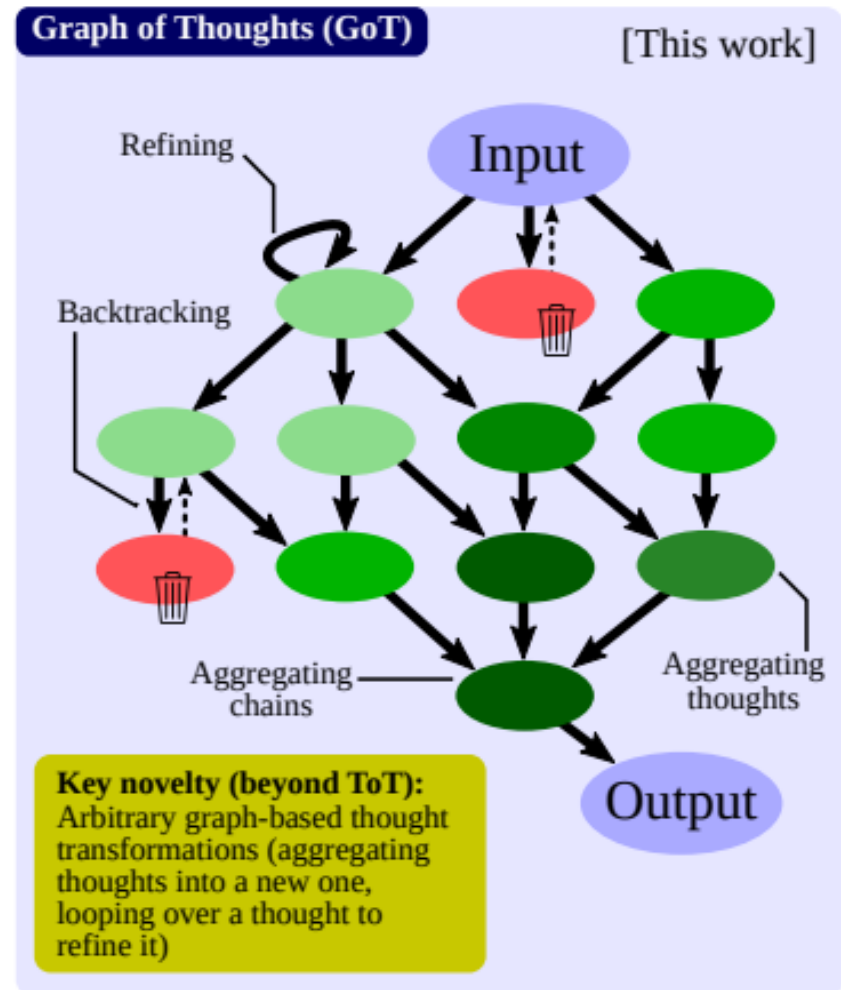
Tree of Thoughts (ToT)





Algorithm of Thoughts

<https://arxiv.org/abs/2308.10379>



Graph of Thoughts

<https://arxiv.org/abs/2308.09687>

有那些在不訓練模型的情況下強化語言模型的方法

