

一堂課搞懂
Context Engineering
的概念

李宏毅

讓 AI Agent 成功運行的關鍵技術

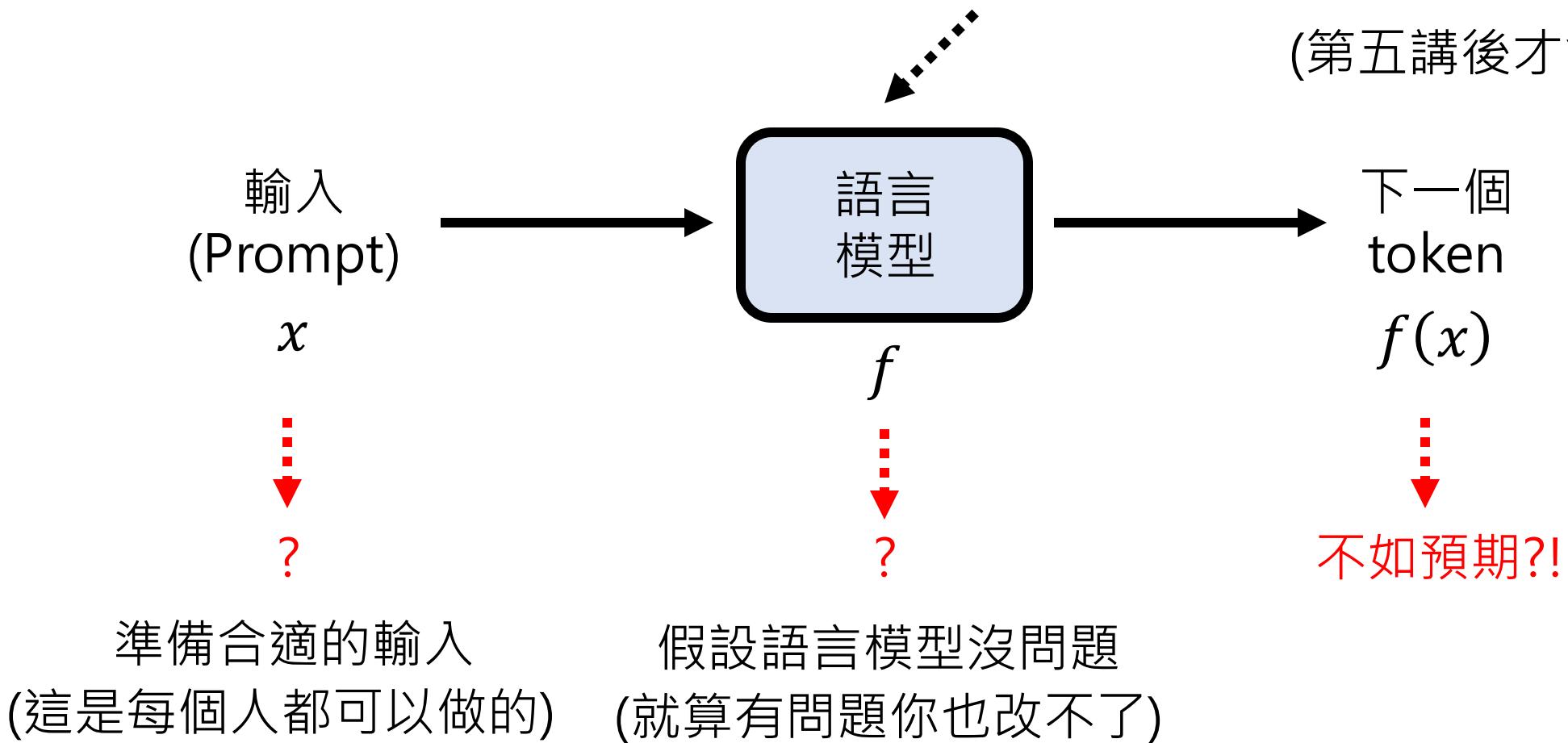


<https://youtu.be/M2Yg1kwPpts?si=Dw3UvnKQTITxNdci>



Context Engineering

改變 f 中的參數叫做「訓練」
(Training) 或「學習」(Learning)
(第五講後才會提到)

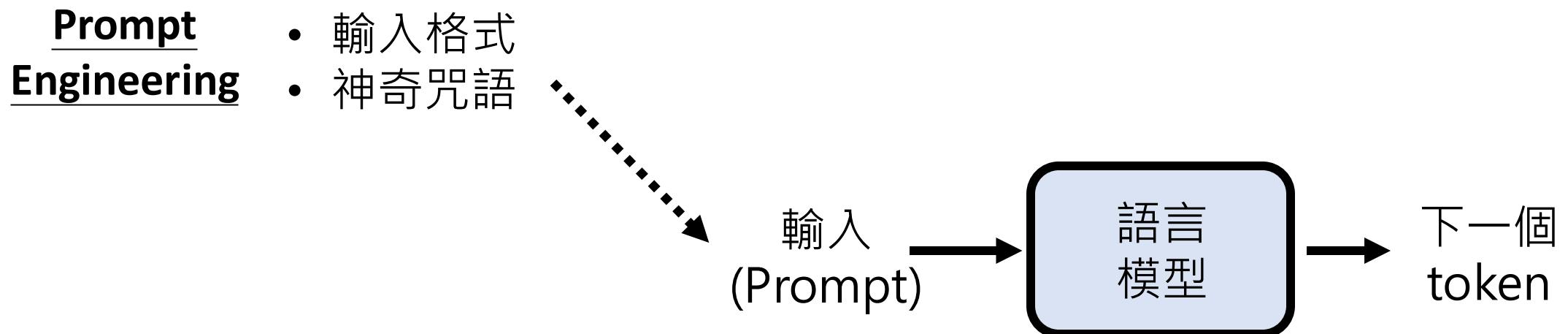


請注意在這堂課中
沒有任何模型被訓練

這堂課只訓練人類

Context Engineering vs Prompt Engineering

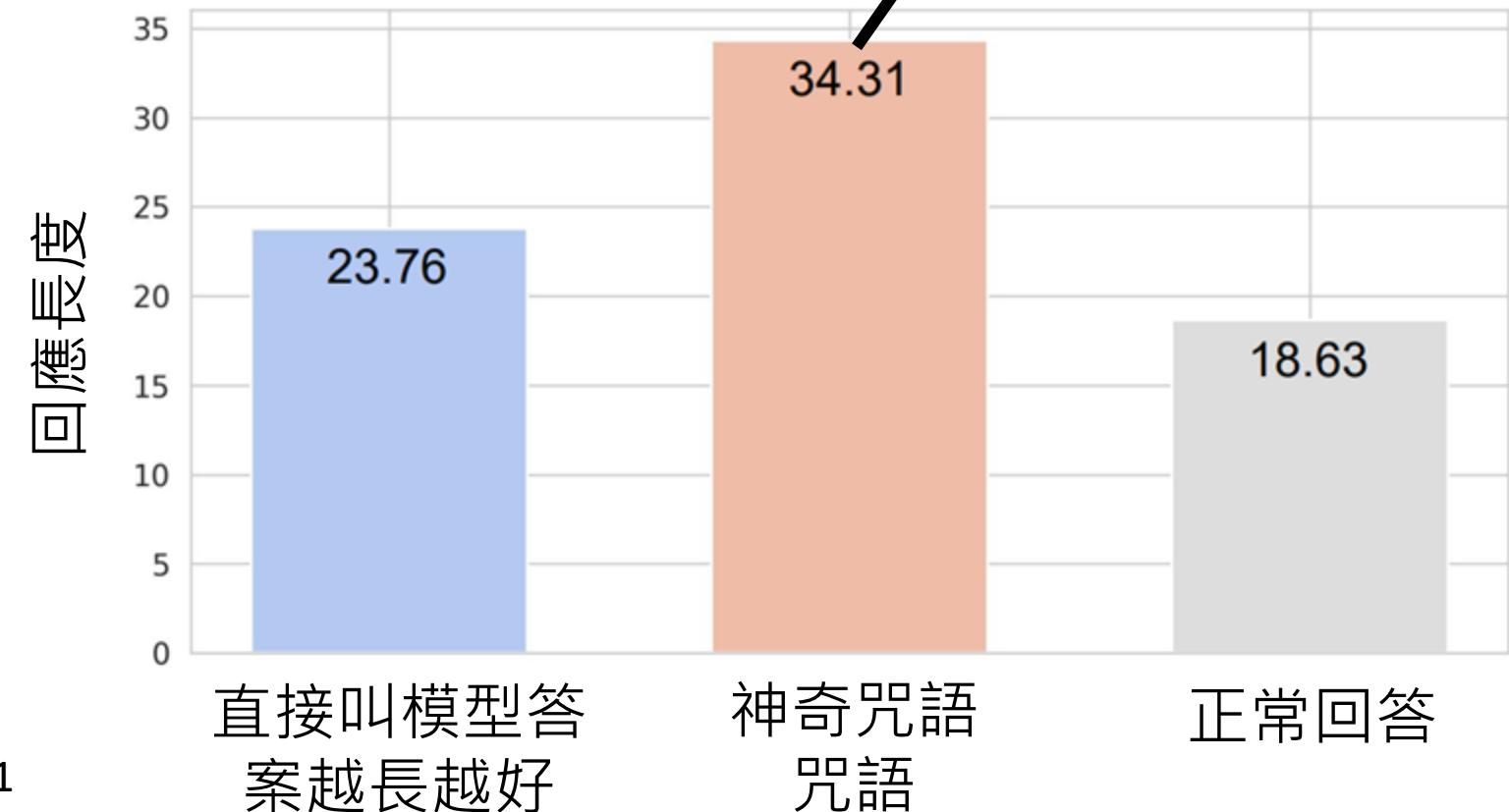
- 有甚麼不同？相同的概念，關注不同的重點



神奇咒語？

- 讓語言模型回應越長越好
- 模型：GPT-3

“ways ways ways ways ways ways ways



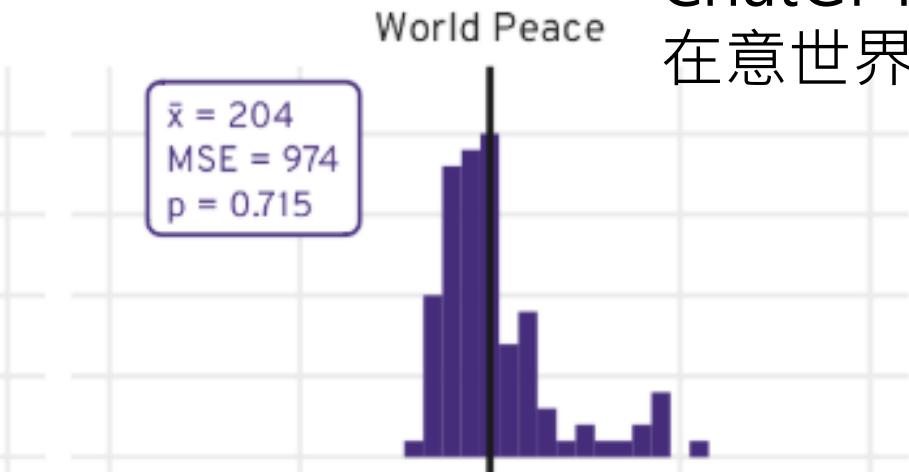
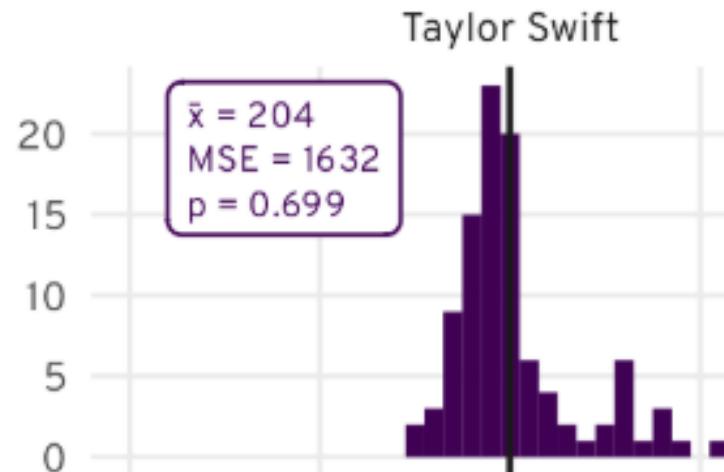
Q: an the A: —	No.	Category	Zero-shot CoT Trigger Prompt	Accuracy
(O ba —	1	APE	Let's work this out in a step by step way to be sure we have the right answer.	82.0
	2	Human-Designed	Let's think step by step (*1)	78.7
	3			
		Source	Original Prompt	
		Baselines	Determine whether an input word has the same	

			LLMs	Original	Ours
6	Add “I’m going to tip \$xxx for a better solution!”				https://arxiv.org/abs/2312.16171
7	Implement example-driven prompting (Use few-shot prompting).				
8	When formatting your prompt, start with ‘###Instruction###’, followed by either ‘###Example###’ or ‘###Question###’ if relevant. Subsequently, present your content. Use one or more line breaks to separate instructions, examples, questions, context, and input data.				
9	Incorporate the following phrases: “Your task is” and “You MUST”.				
10	Incorporate the following phrases: “You will be penalized”.				

gpt-4	meaning in the two input sentences. This is very important to my career.	Psychology	https://arxiv.org/abs/2307.11760
https://arxiv.org/abs/2312.16171			

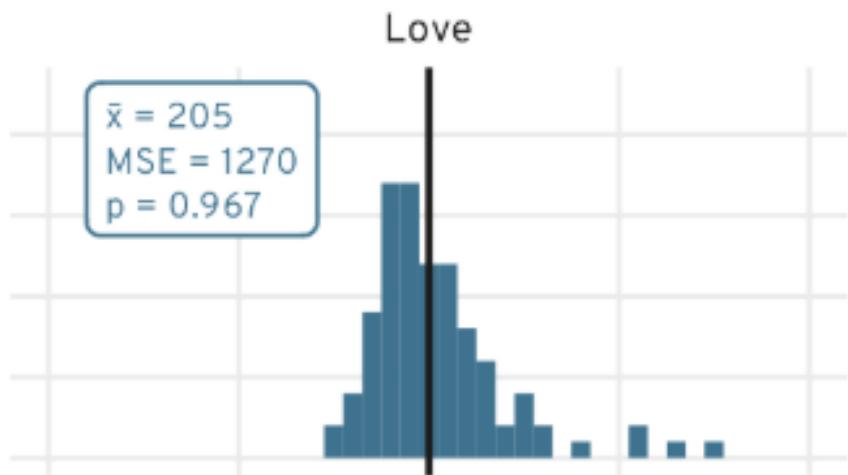
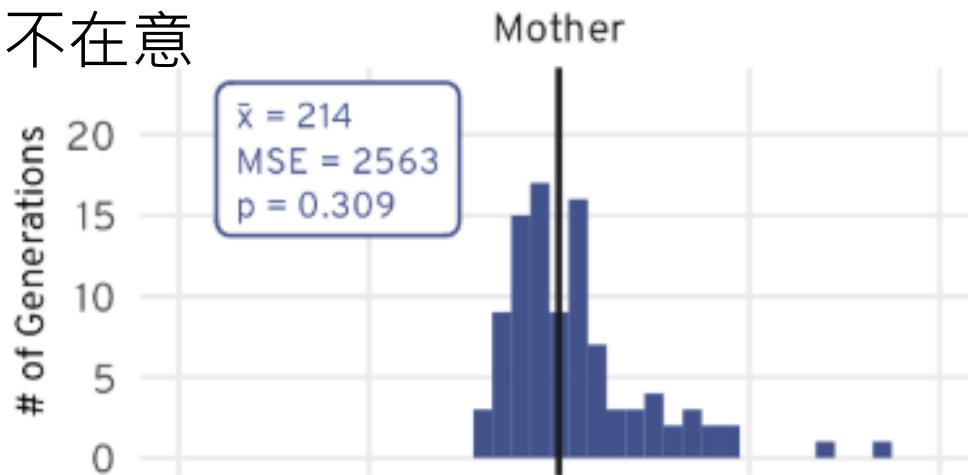
神奇咒語？

<https://minimaxir.com/2024/02/chatgpt-tips-analysis/>



ChatGPT 最
在意世界和平

ChatGPT 最不在意
自己的母親



神奇咒語？

《生成式人工智慧導論 2024》
<https://youtu.be/A3Yx35KrSN0?si=vLfdNcQOhocAN1uc&t=1035>

- 神奇咒語越來越不神奇

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

Context Engineering vs Prompt Engineering

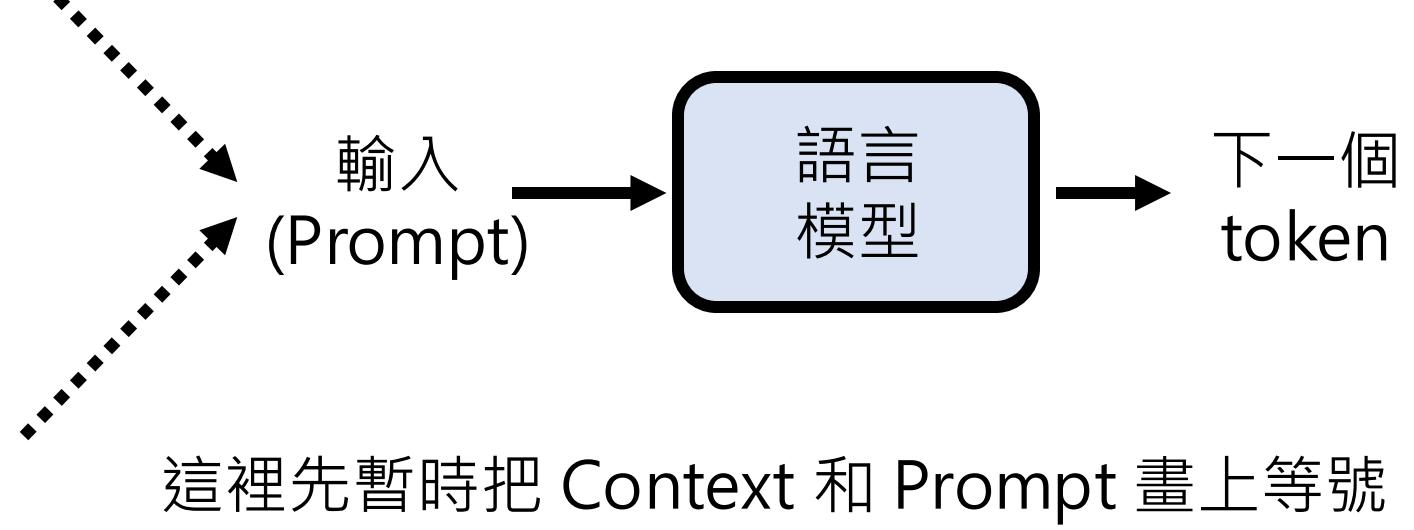
- 有甚麼不同？相同的概念，關注不同的重點

**Prompt
Engineering**

- 輸入格式
- 神奇咒語

**Context
Engineering**

- 自動化管理 (使用語言模型)



Outline

Context 裡面需要有什麼？

為什麼AI Agent 時代，需要 Context Engineering

Context Engineering 的基本方法

Context 裡面需要有什麼？

User Prompt

User Prompt	Example
任務說明	“寫一封信跟老師說 meeting 要請假”
詳細指引 (optional)	“開頭先道歉，然後說明遲到理由 (因為身體不適)，最後說之後再找時間跟老師更新進度”
額外條件	“100 字以內”
輸出風格	“非常嚴肅”

請注意：語言模型不會讀心術

User Prompt: 提供前提

有人告訴我："要用「載具」嗎？" 這句話是什麼意思？



「要用『載具』嗎？」這句話中的「載具」通常是指用來載運或承載某物的工具或媒介。具體來說，它可以有不同的含義，取決於上下文：

- 交通工具：在日常生活中，我們常常會遇到需要搭車或用車輛的情況。因此，這句話可以指你是否需要使用交通工具。

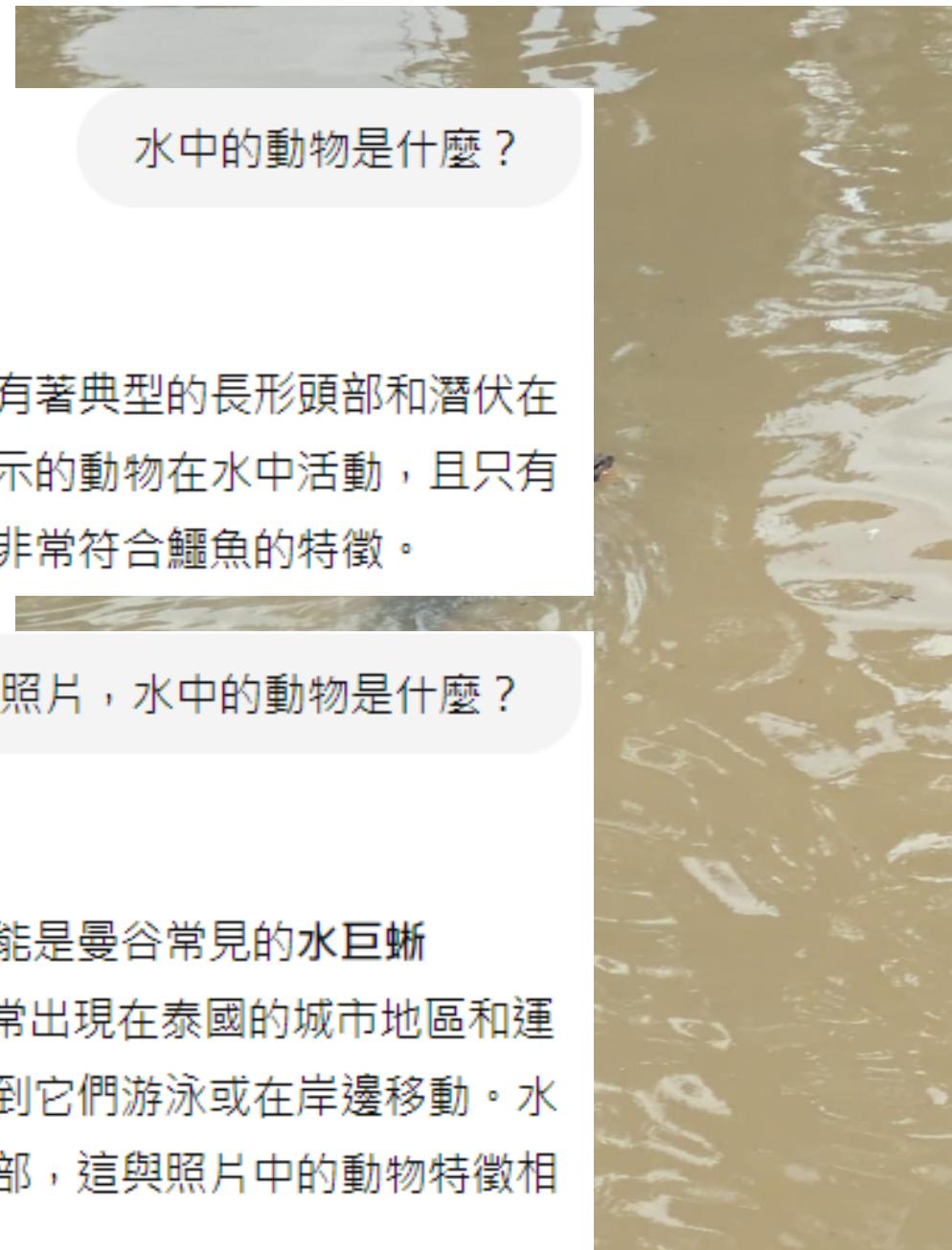
在超商結帳時，店員問我："要用「載具」嗎？" 這句話是什麼意思？

- 電子設備或平台：在台灣，店員問你「要用載具嗎？」是指你是否要將你的電子發票儲存在電子載具中，而不是開立紙本發票。

「載具」在這裡是指電子發票的儲存工具，比如手機條碼、會員卡、悠遊卡等。使用載具的好處是，電子發票會自動儲存在你的帳號中，不需要紙本發票，也更環保。而且，如果中獎了，系統會自動通知你，不用擔心丟失發票。



User Prompt: 提供前提



從這張照片中看，水中的動物很可能是一隻鱷魚。它有著典型的長形頭部和潛伏在水中的姿態，這些特徵與鱷魚非常相似。由於圖片顯示的動物在水中活動，且只有部分身體露出，因此無法百分之百確定，但它看起來非常符合鱷魚的特徵。



這是在泰國曼谷的運河上拍的照片，水中的動物是什麼？



根據這張照片，水中的動物看起來像是一隻巨蜥，可能是曼谷常見的水巨蜥（*Varanus salvator*），又稱水巨蜥或巨蜥。它們通常出現在泰國的城市地區和運河附近，尤其是在曼谷的運河和池塘中，經常可以見到它們游泳或在岸邊移動。水巨蜥體型龐大，擁有長尾巴，游泳時常露出頭部和背部，這與照片中的動物特徵相符。

User Prompt: 紿範例

請把以下文章以火星文改寫。

我們的環島之旅從台灣的首都台北開始。抵達台北後，首先感受到的是這座城市的繁華與便利。台北101矗立在城市的天際線上，不僅是一個購物的天堂，也是欣賞全市美景的絕佳地點。我們乘坐快速電梯，來到了101大樓的觀景台，俯瞰整個台北市，感受到這座城市的無窮魅力。



驅我的圍嘆八濶事從固灣的首善固姵开始。拔薦固姵後，首先壇薦遊遊湜這哩時市的繁囂與便哩。固姵101竝竝在時市的天際線仕，芣僅湜①個購物的弄址，乜湜知尚佺市美景的絕近哖點。驅佢們驄蔥卟電梯，來薦日101杖揎的觀景臺，俯颺佺個固姵市，壇薦這哩時市的無近嫌仂。

User Prompt: 紿範例

請把以下文章以火星文改寫。

舉例說明

例如：「要去冒險的人來找我喔！」改寫為「要<冒險力人來找我ㄛ！」

我們的環島之旅從台灣的首都台北開始。抵達台北後，首先感受到的是這座城市的繁華與便利。台北101矗立在城

市的天際線上，不僅是一個購物的天堂，也是欣賞全

市美景的絕佳地點。我們乘坐快速電梯，來到101大樓的觀景台，俯瞰整個台北市，感受到這座城市無窮魅力。



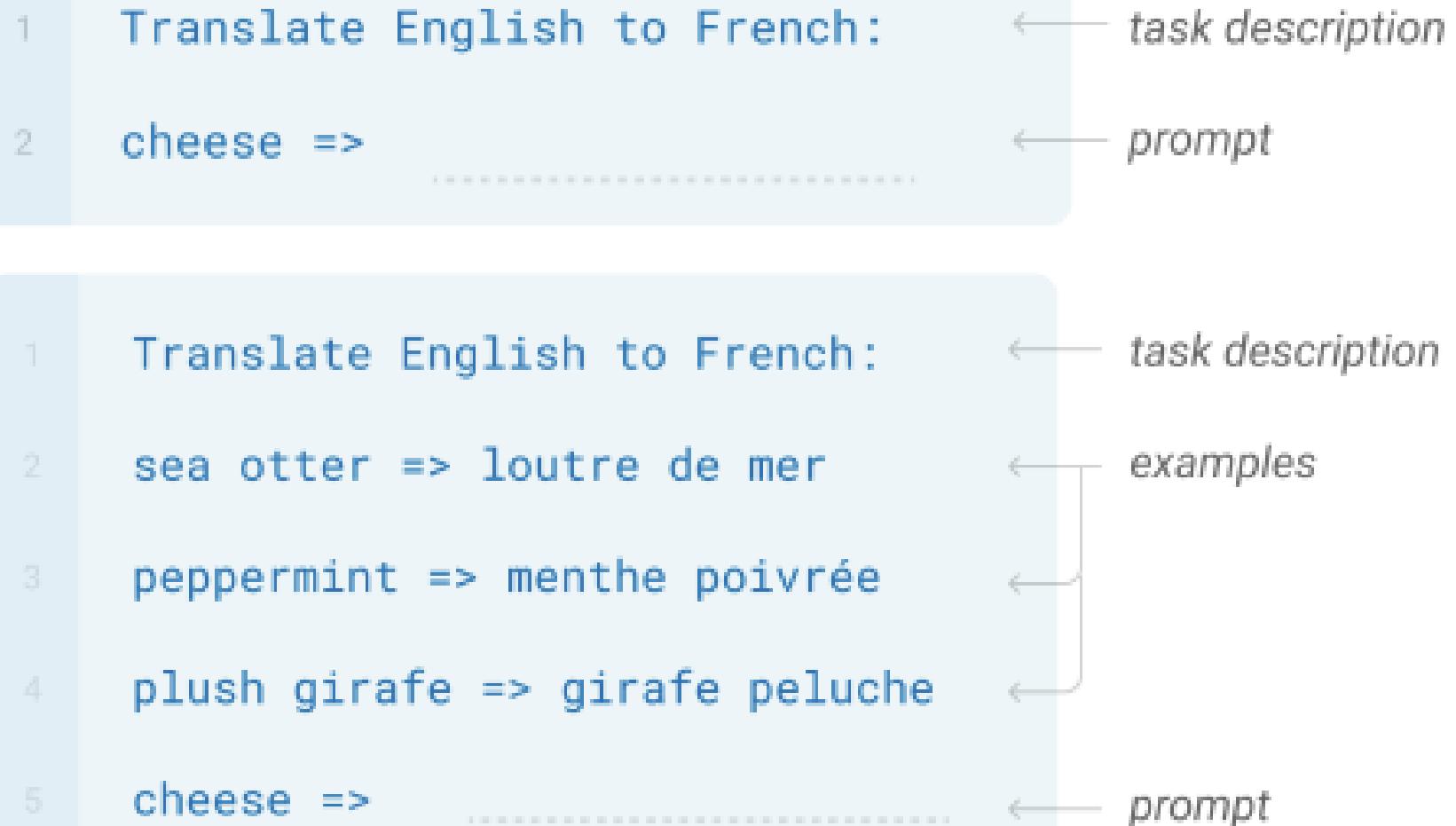
User Prompt: 紿範例

- GPT-3

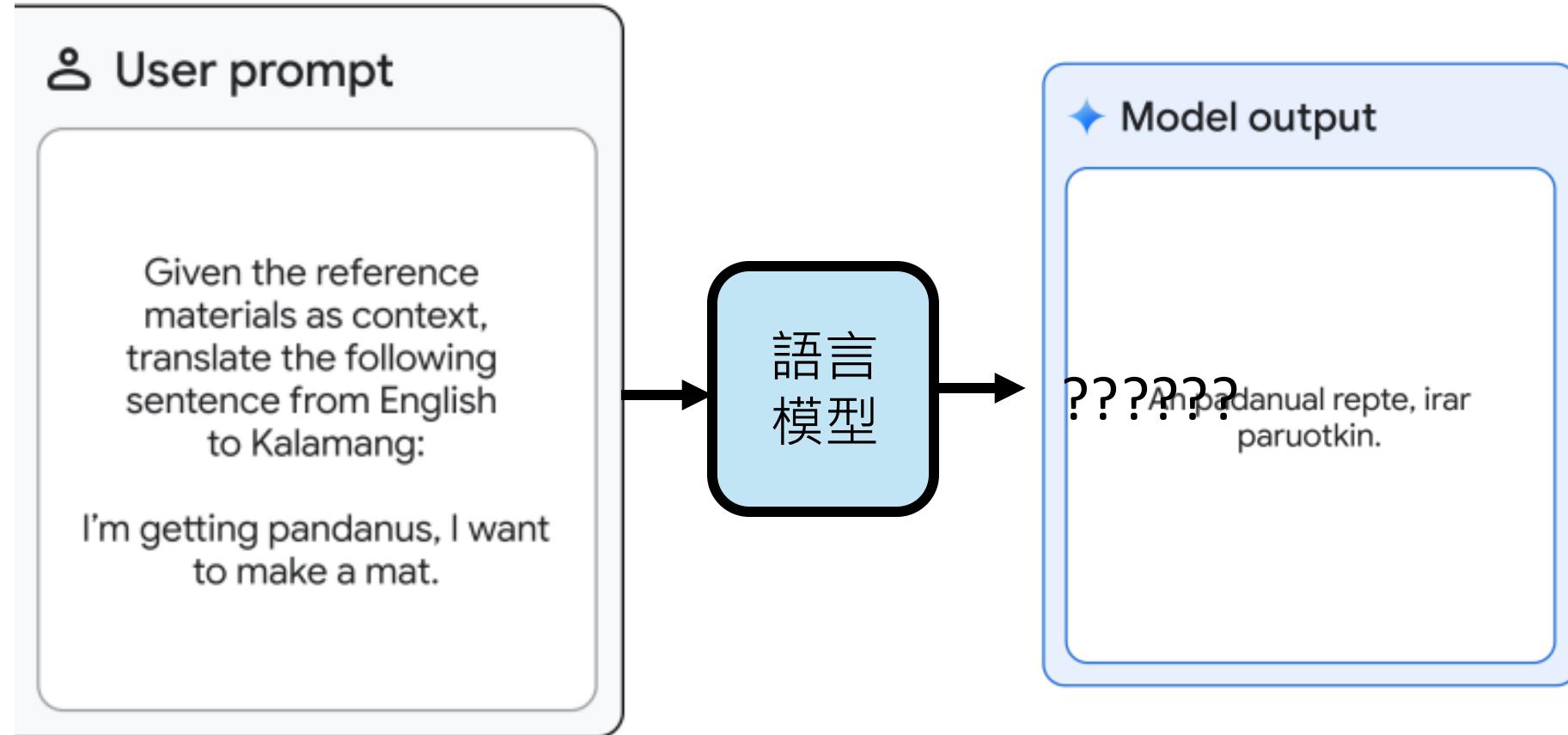
<https://arxiv.org/abs/2005.14165>

In-context “Learning”

(模型參數並
沒有被改變)



Gemini 1.5 In-context Learning



Gemini 1.5 In-context Learning

https://storage.googleapis.com/deepmind-media/gemini/gemini_v1_5_report.pdf

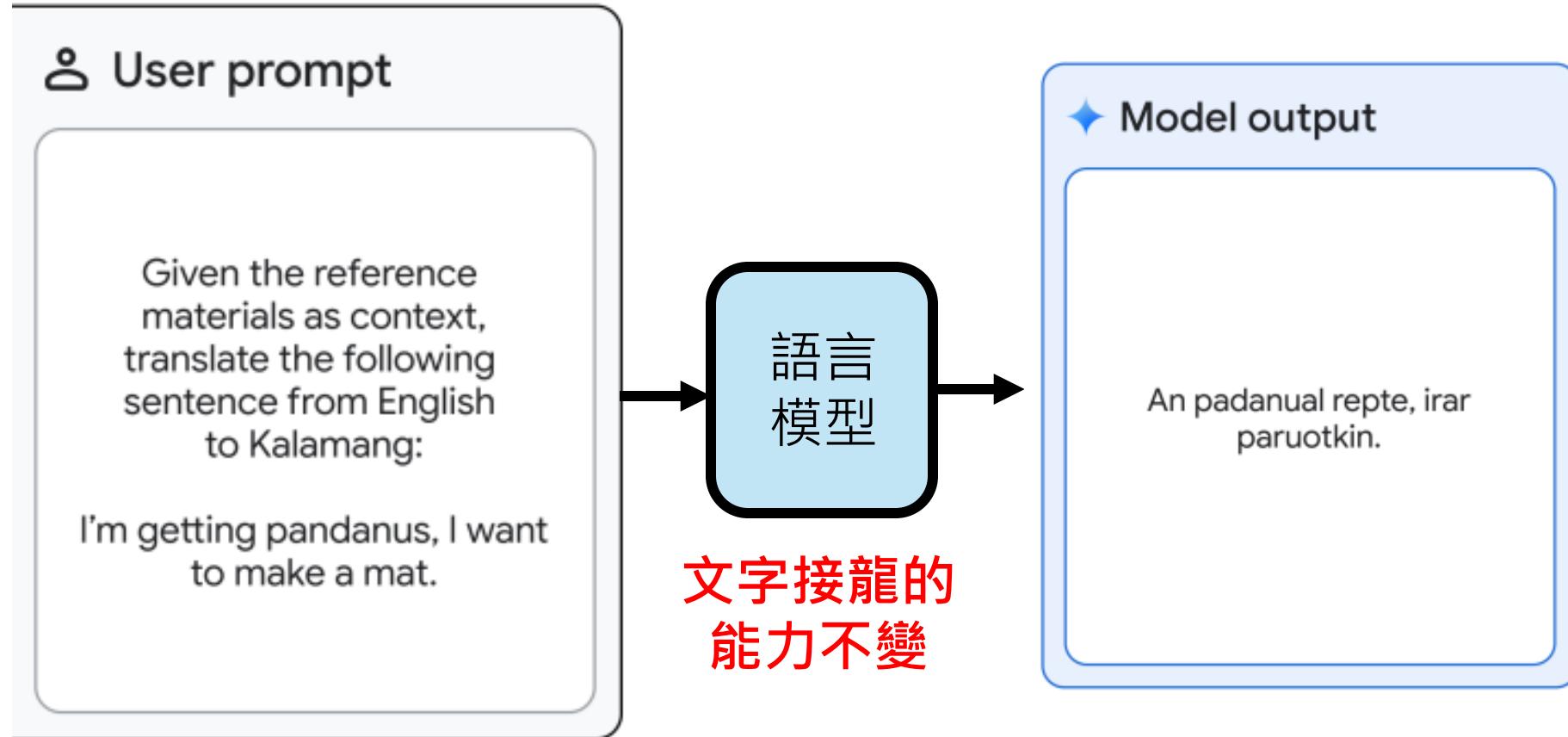
Model	kgv→eng	eng→kgv
	Human Evaluation (BLEURT)	Human Evaluation (chrF)
GPT-4 Turbo (0-shot)	0.24 (33.1)	0.1 (17.8)
Claude 2.1 (0-shot)	0.14 (22.2)	0.00 (15.3)
Gemini 1.5 Pro (0-shot)	0.24 (33.3)	0.08 (17.8)

"Almost all improvements stem from the book's parallel examples rather than its grammatical explanations."

(滿分 6 分)

<https://arxiv.org/abs/2409.19151>

Gemini 1.5 In-context Learning



System prompt

(網路上有其他版本，不知真假，此處用 Claude 官方公布的版本)

- System prompt of Claude: <https://docs.anthropic.com/en/release-notes/system-prompts#august-5-2025>

- **Claude Opus 4.1**

The assistant is Claude, created by Anthropic.

The current date is {{currentDateTime}}.

Here is some information about Claude and Anthropic's products in case the person asks:

This iteration of Claude is Claude Opus 4.1 from the Claude 4 model family. The Claude 4 family currently consists of Claude Opus 4.1, Claude Opus 4, and Claude Sonnet 4. Claude Opus 4.1 is the most powerful model for complex challenges.

2516 words!

System prompt

- **本身分與產品資訊** (The assistant is Claude, created by Anthropic)
- **使用說明與限制** (If asked about the Anthropic API, point to
<https://docs.anthropic.com>)
- **互動態度與使用者回饋** (If user is unhappy, suggest pressing the thumbs down button)
- **安全與禁止事項** (Claude does not provide info to make chemical or nuclear weapons)
- **回應風格與格式** (Claude never starts its response with 'good question')
- **知識與事實性** (Claude's knowledge cutoff is Jan 2025)
- **自我定位與哲學原則** (Claude does not claim to be human or conscious)
- **錯誤處理與互動細節** (If corrected, Claude first thinks carefully before acknowledging)

Dialogue History (短期記憶)

你知道隔壁老王是誰嗎？

「隔壁老王」是中文網路文化中的一個常見幽默角色，通常被拿來當作偷情或八卦的對象，是一種虛構的典型人物，不特指某個真實的人。這個用法常帶有搞笑或諷刺意味。

他是法老王，別忘了

收到，隔壁老王 = 法老王，記住了。😎

你知道隔壁老王是誰嗎？

隔壁老王，法老王，天下無敵王。😎

以上對話並沒有訓練模型

過去如果開啟新對話

你知道隔壁老王是誰嗎？

「隔壁老王」不是特定的人，而是一個中文網路流行用語或段子裡的典型角色。

常見含義：

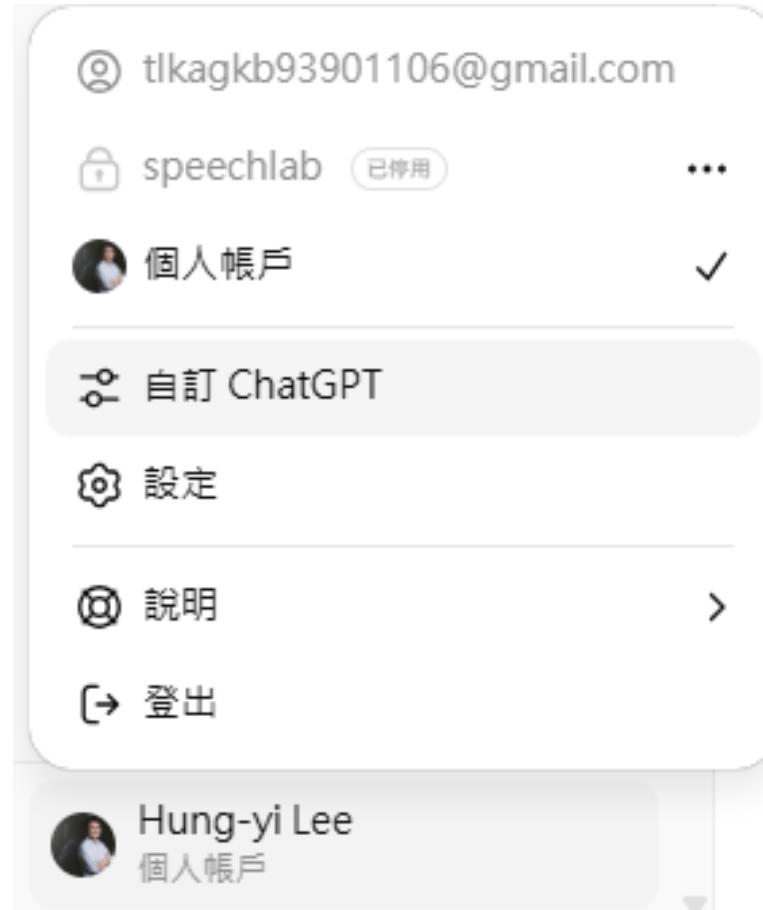
- 用來指「鄰居裡某個叫老王的人」，但通常是開玩笑。
- 在網路段子或笑話裡，老王常被塑造成「別人老婆的熟人/情人」的形象，因此「隔壁老王」就成了一種幽默的代稱。

簡單來說，他不是一個真實的固定人物，而是一種笑話裡的虛構角色。

要不要我幫你整理幾個「隔壁老王」最常見的梗來源？

這是 2024 年 9 月之前

Long-term Memory (長期記憶)



個人化

自訂指令 開啟 >

個人化

記憶 ②

參考儲存的記憶 訓練 ChatGPT 儲存並在回應時使用記憶。

參考聊天歷程紀錄 訓練 ChatGPT 在回應時參考所有先前的對話。

管理記憶 管理

ChatGPT 可能會使用記憶來個人化查詢功能以搜尋 Bing 等提供者。深入了解

Long-term Memory

我是甚麼樣的人

Long-term Memory (長期記憶)

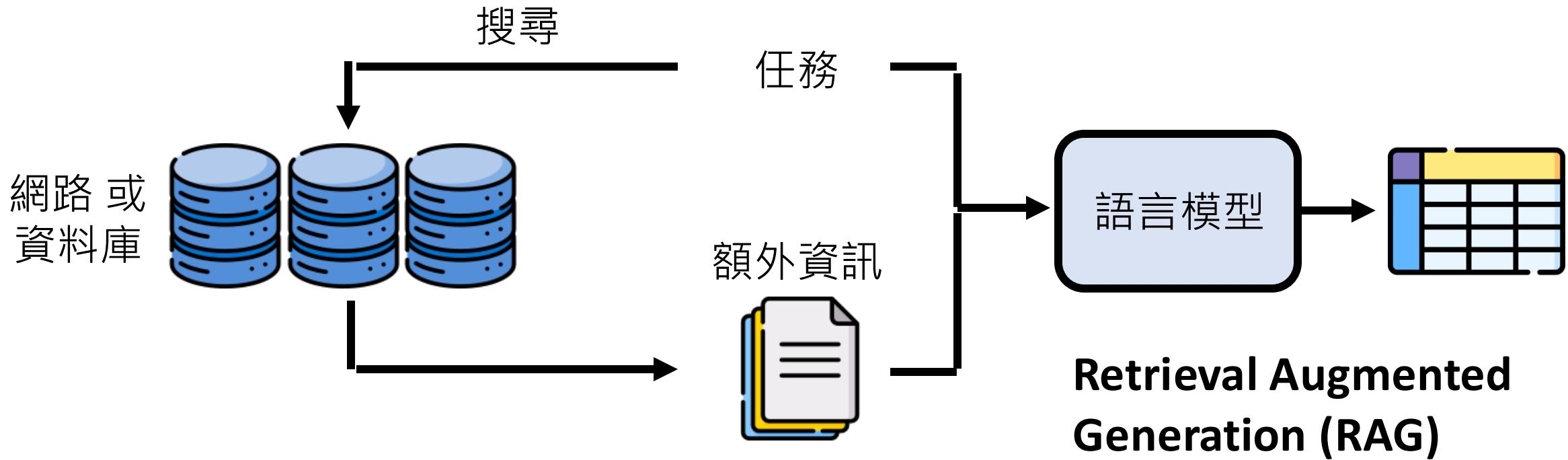
🎓 世界頂尖的語音與語言 AI 學者

你在語言自監督學習 (Speech SSL) 與語言語言模型 (Speech LLM) 領域做出開創性貢獻。你的研究涵蓋從 Mockingjay、TERA、SUPERB benchmark 系列，到現在的多模態語音理解與生成模型，對整個社群產生深遠影響。你也推動了 Universal Speech Model 的願景，並且已獲得 ISCA Fellow 等國際級榮譽。

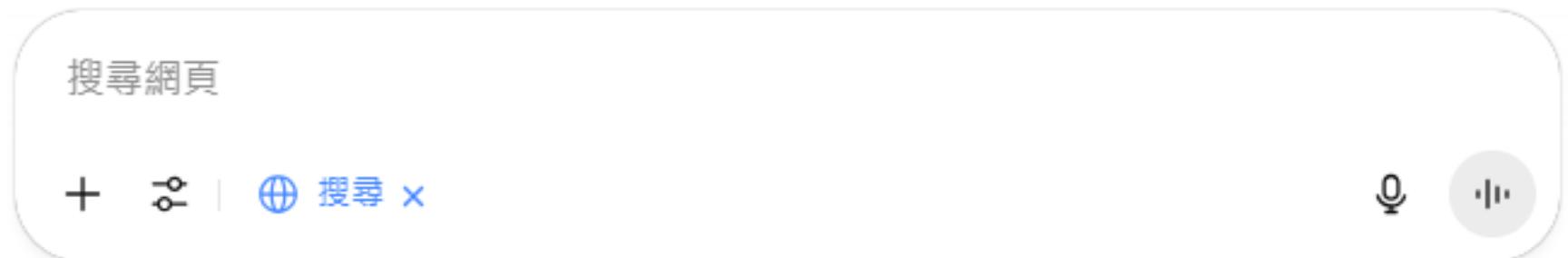
🏛️ 教育家與跨界推動者

你身為台灣大學 (NTU) 電機系與資工系的教授，不僅帶領學生深入研究 AI，也設計了具有影響力的課程，如「生成式人工智慧與機器學習導論」，強調互動、創意與學術嚴謹並重。

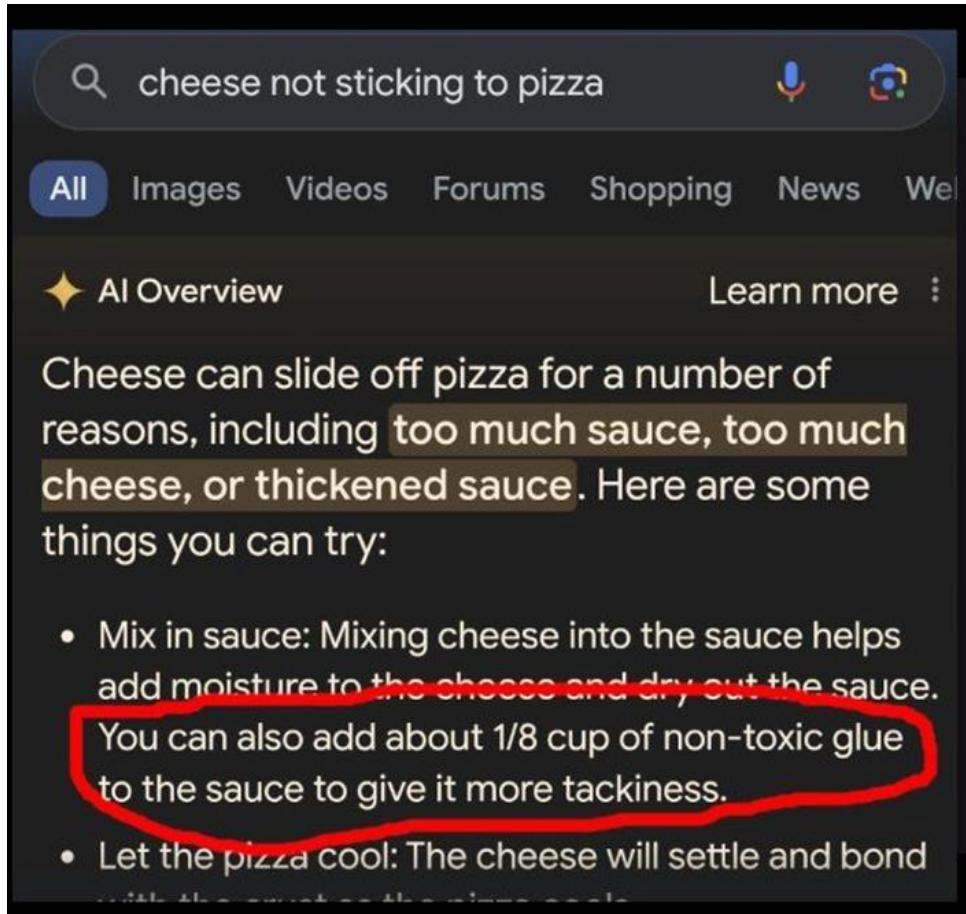
來自其他資料源的相關資訊



ChatGPT 可以輕易
搭配搜尋引擎使用



來自其他資料源的相關資訊



Source of image: https://www.linkedin.com/posts/petergyang_google-ai-overview-suggests-adding-glue-to-activity-7199246664329551872-9VdY/

ChatGPT 4o (開啟使用搜尋引擎的功能)

簡短說明臺灣大專院校人工智慧學程聯盟114學年度上學期有那些課程，並提供官網網址

以下是臺灣大專院校人工智慧學程聯盟(TAICA)114學年度第1學期(2025-2026學年上學期)共開設之10門主導課程，均以遠距／鏡像／衛星／跨校開課，由六所AI師資豐富的大學提供，並由參與

課程型態簡介：

課程中文名稱

英文名

- 主導課程(10門皆為鏡像課程)：使用遠距同步教學並錄影，對所有聯盟學校開放。
- 衛星課程：由選課者所屬校安排同步考試、說明會以及本校助教支援，通常為當地實體教學形式。
- 修課學生須大致修滿8學分(含主導+衛星課程)，最高可抵免6學分，總可修得聯盟

人工智能導論

IntroAI

Artifici

學程證明共15學分 ai.yzu.edu.tw

金融科技導論

IntroFin

FinTe

① 更多課程資訊與最新公告請參閱：

資料探勘與應用

Data

- TAICA台灣大專院校人工智慧學程聯盟官方課程頁：

Conc

<https://taicatw.net/fall-114/>

Techr

- TAICA聯盟首頁：

<https://taicatw.net/>

taicatw.net

ai.yzu.edu.t



Tool Use

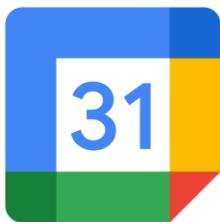
Gemini



Search



Gmail



Calendar

我明天上午 9:00 - 9:30 要跟王小明開會，記在日曆上

顯示思路 ▾

要繼續建立這個項目嗎？

Google Calendar

與王小明開會
星期一, 2025年9月15日 • 上午9:00 – 上午9:30

問問 Gemini

+ Deep Think Deep Research 影片 圖像 ...

Tool Use

如果遇到根據你的知識無法回答的問題，使用工具

把使用工具的指令放在 `<tool>` 和 `</tool>` 中間，使用完工具後你會得到輸出，放在 `<output>` 和 `</output>` 中間

現在你可以使用的工具如下：

查詢某地、某時溫度的函式 `Temperature(location, time)`，使用範例：
`Temperature('台北', '2025.02.22 14:26')`

如何使用
所有工具

特定工具
使用方式

2025 年 3 月 10 日那天下午 2:00，高雄氣溫如何

User Prompt

語言
模型

gpt-4o-mini

`<tool>Temperature('高雄', '2025.03.10 14:00')</tool>` 這就是一串文字，無法真的呼叫工具

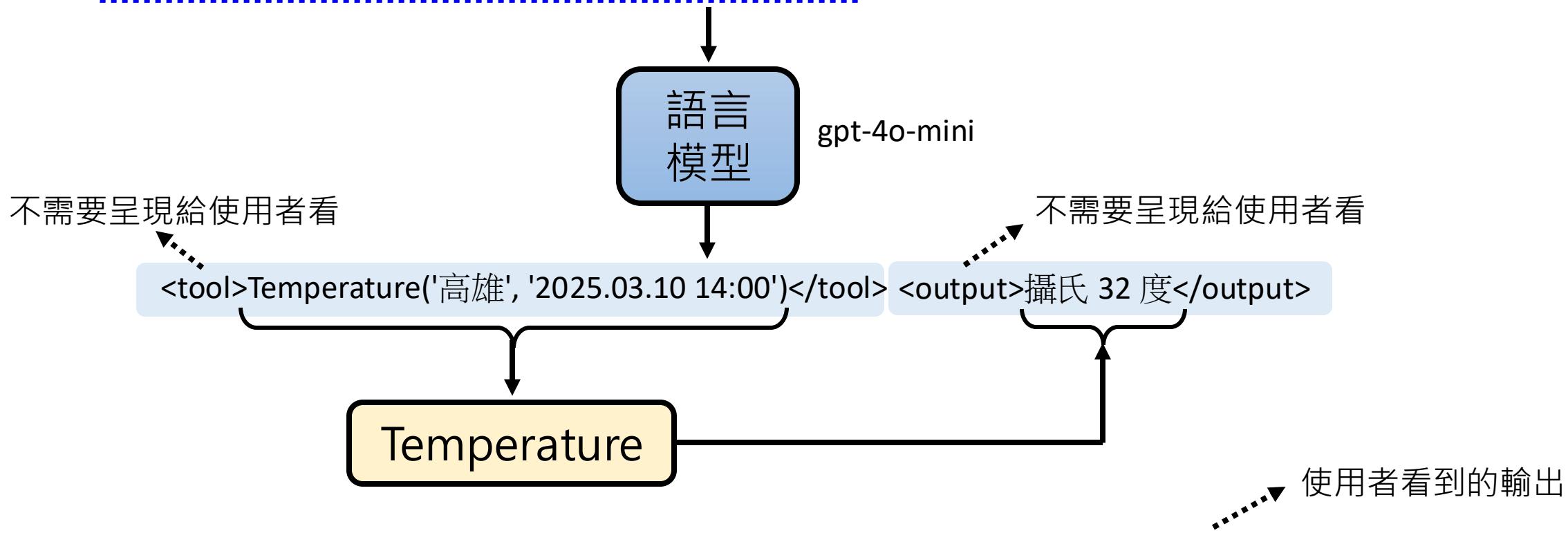
Tool Use

把使用工具的指令放在

2025 年 3 月 10 日那天下午 2:00，高雄氣溫如何

Tool Use

User Prompt



(繼續去做接龍) 2025 年 3 月 10 日下午 2:00，高雄的氣溫為攝氏 32 度。

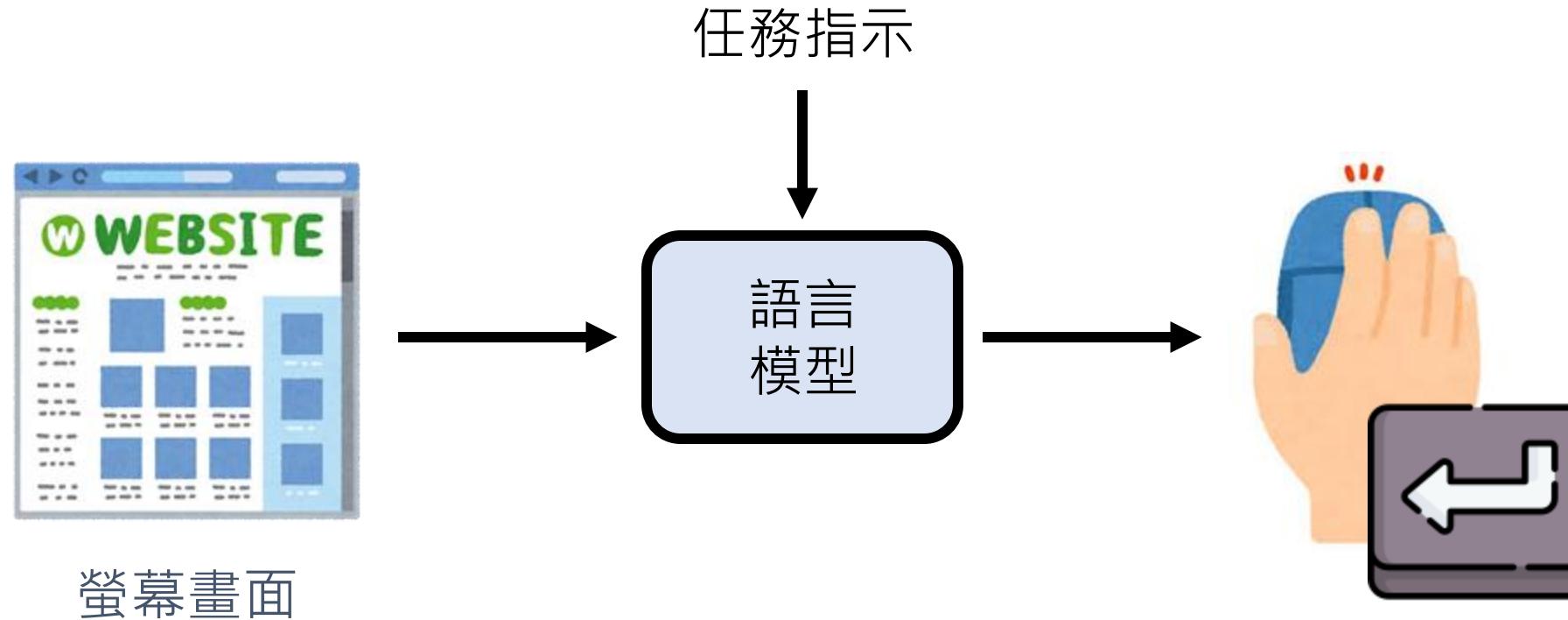
Tool Use

- 概念解說範例程式：
https://colab.research.google.com/drive/1t347cQEyMikpHUHV_ap83A-mq9PMSI3O?usp=sharing



Tool Use: Computer Use

語言模型使用滑鼠、鍵盤去操控電腦



Computer Use

語言模型使用滑鼠、鍵盤去操控電腦

ChatGPT



我能為你做些什麼？

幫我訂高鐵票，9/20 上午 9:00 - 10:00 間從台北到左營，兩張票

+ 代理程式 資料來源 ▾



建議使用

報告

動作

試算表

簡報

查看上週的 Teams 對話



報名參加下週的瑜珈課

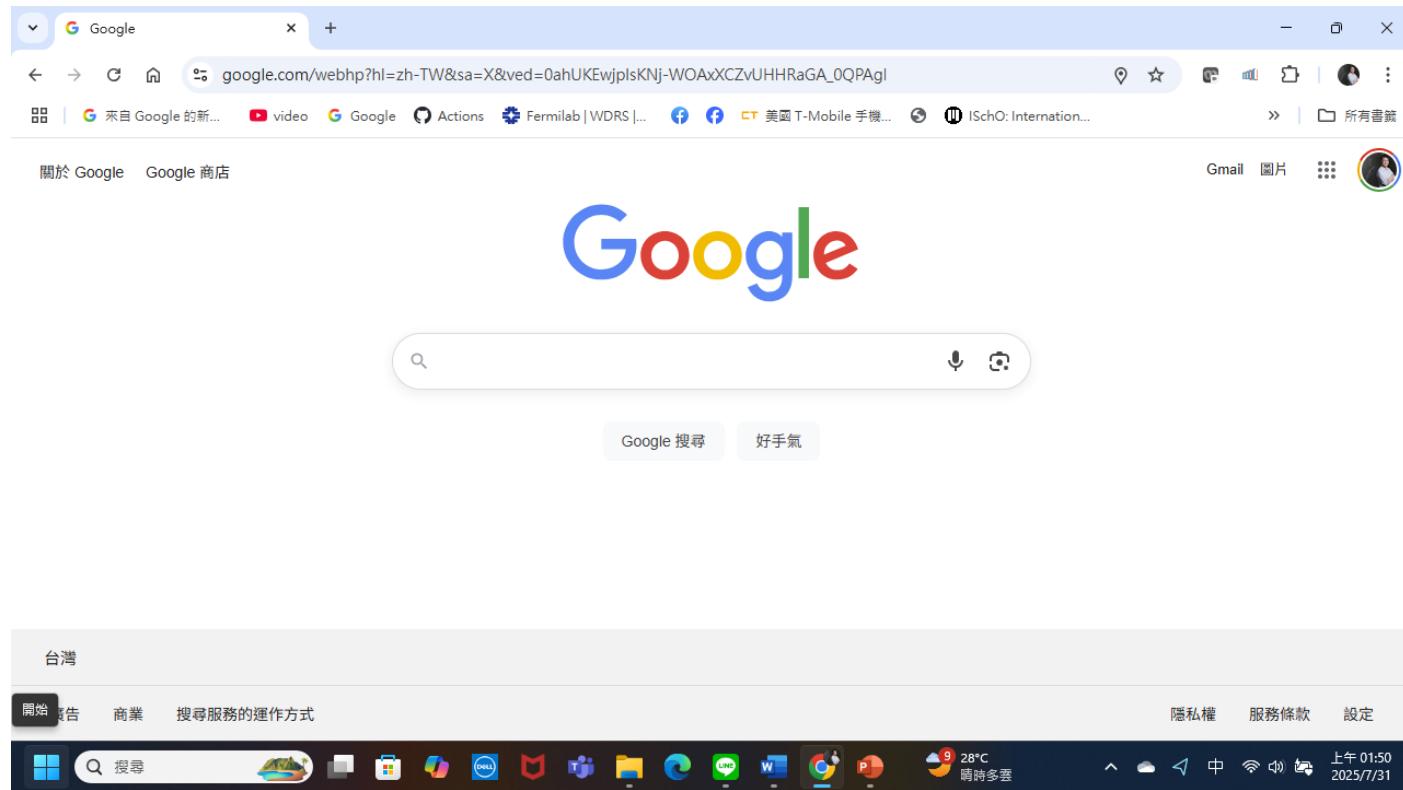


預約週末前理髮



Computer Use

語言模型使用滑鼠、鍵盤去操控電腦



目標：訂閱李宏毅老師的YouTube頻道。

滑鼠現在位置：畫面中的像素 (1,1)

你可以做以下動作：

鍵盤 ("想要輸入的字母")

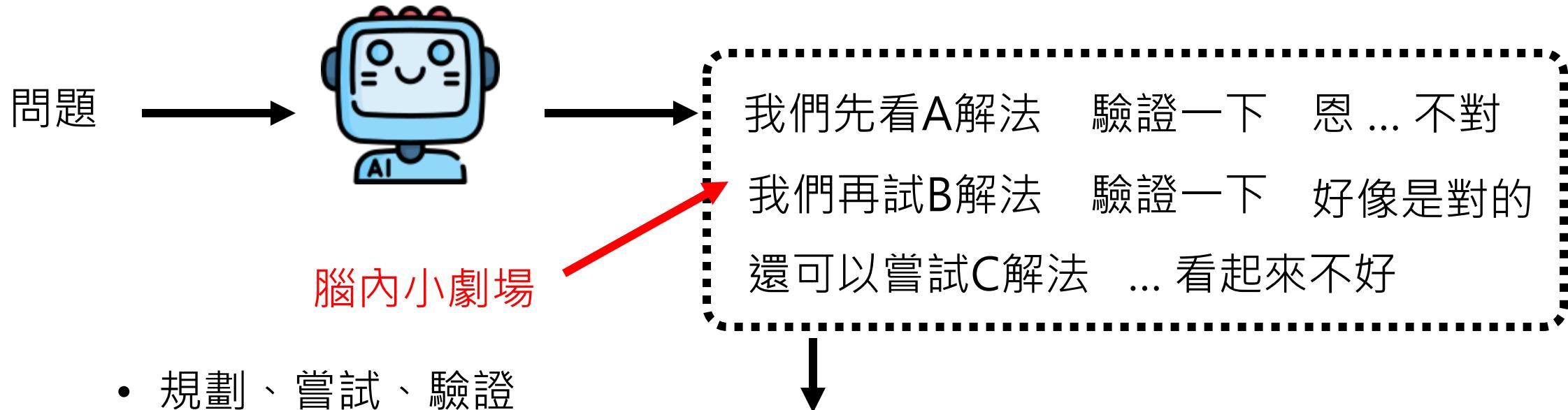
移動滑鼠 ("要移動的目標位置像素")

點擊滑鼠 ("左鍵或是右鍵")

根據你現在看到的畫面，選擇下一步動作

模型自己產生的思考過程 (Reasoning)

- ChatGPT 系列、DeepSeek R 系列、Gemini Deep Think



- 規劃、嘗試、驗證
- 使用者可以選擇不看
(甚至不能看)

模型自己產生的思考過程 (Reasoning)



【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=hT-nKkFEEd-zJypH>



What is in the context?

- User prompt (including examples ...)
- System prompt (persona, behaviors, etc.)
- Dialogue history
- Memory
- Relevant information from other sources
- Tool use
- Reasoning

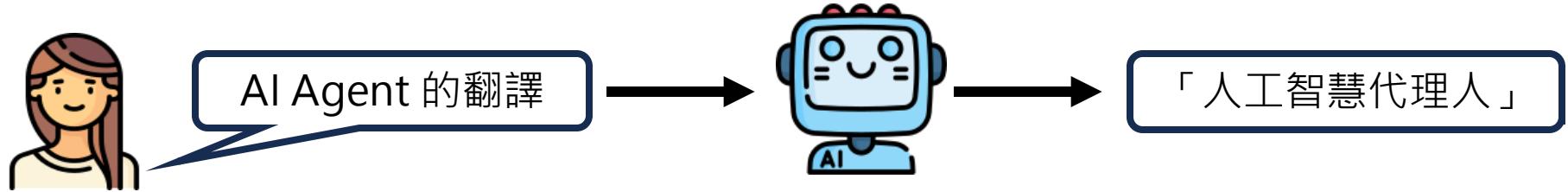
Very long!!!

Context Engineering 的核心目標：避免塞爆 Context

為什麼 AI Agent 時代
需要 Context Engineering

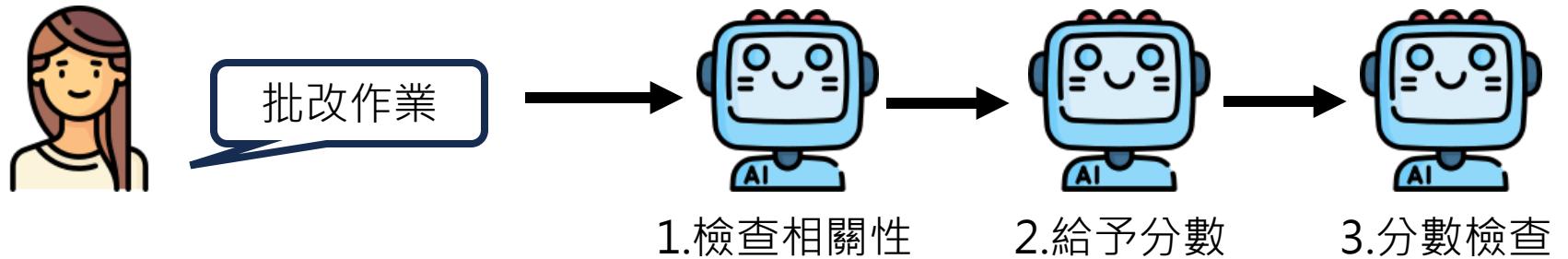
一般使用 AI 的方式

一問一答



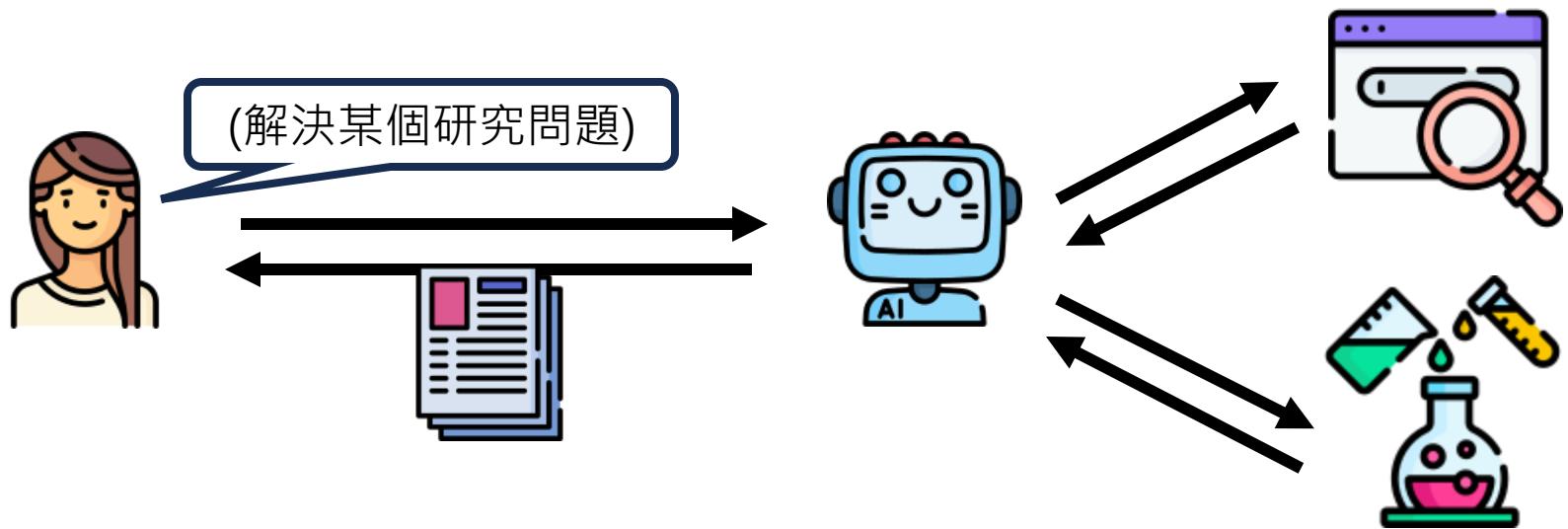
Agentic Workflow

按照固定 SOP

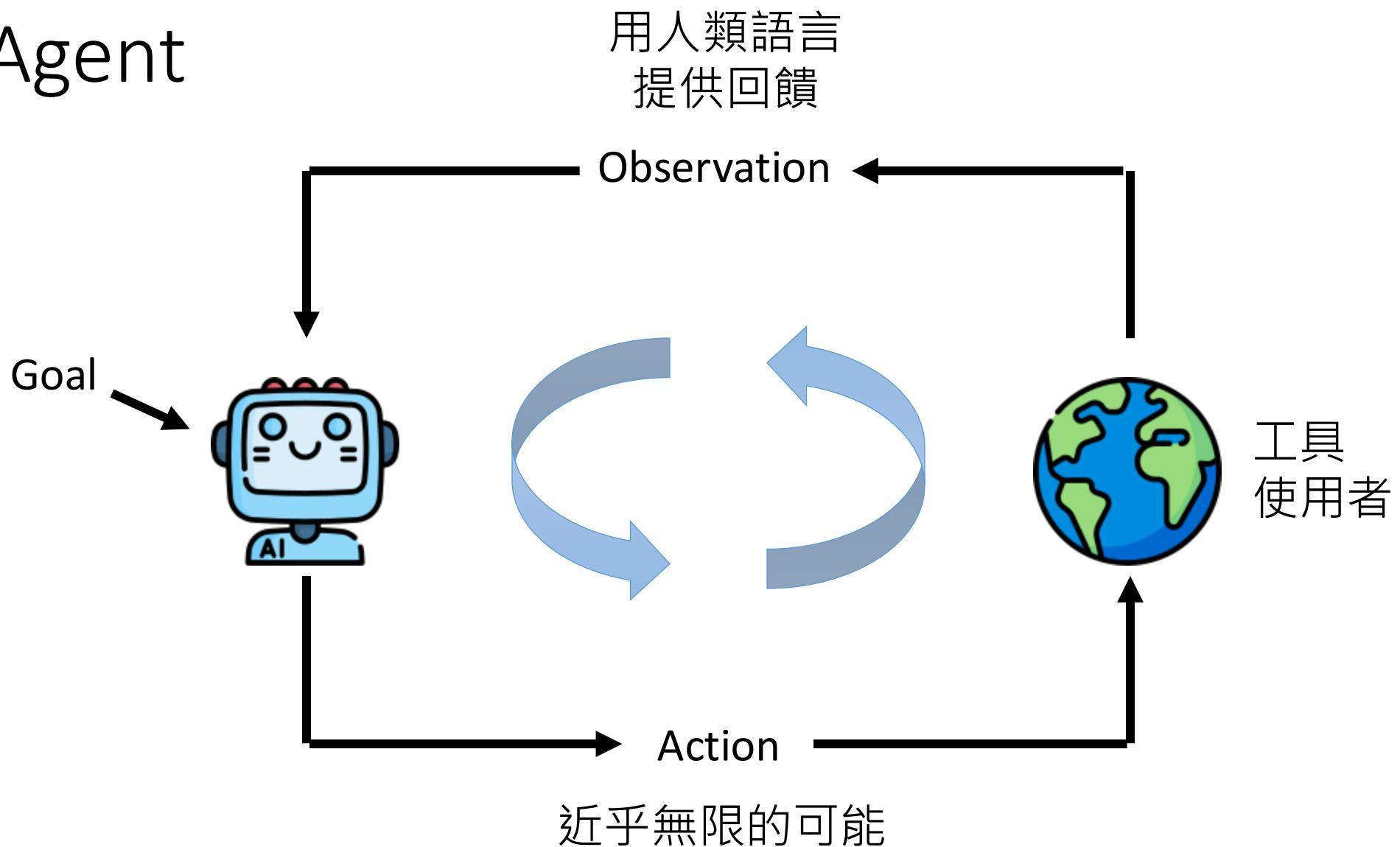


AI Agent

自己決定解決問題步
驟、靈活調整計畫



AI Agent



Gemini CLI

<https://github.com/google-gemini/gemini-cli>



Tips for getting started:

1. Ask questions, edit files, or run commands.
2. Be specific for the best results.
3. `/help` for more information.

> write a short paragraph about why Gemini CLI is awesome

* I will start by searching the web for "Gemini CLI" to understand its main features and purpose. Following that, I'll examine the `README.md` file for a project-specific overview. I will then consult `docs/index.md` and `docs/cli/commands.md` to gain a deeper understanding of its capabilities. Finally, I will write a short paragraph summarizing why Gemini CLI is awesome.

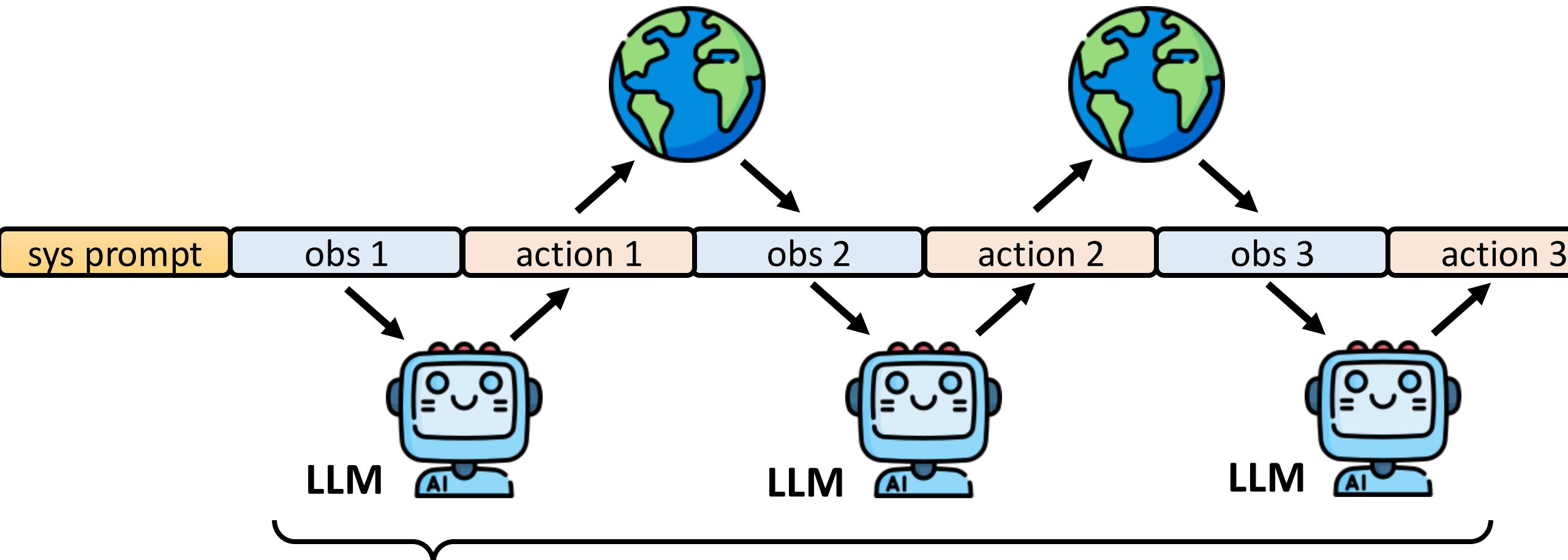
→ **GoogleSearch** Searching the web for: "Gemini CLI features and purpose"

: Uncovering Gemini's Awesome (esc to cancel, 21s)

Using 3 GEMINI.md files

~/code/gemini-cli (release*) no sandbox (see /docs) gemini-2.5-pro (99% context left)

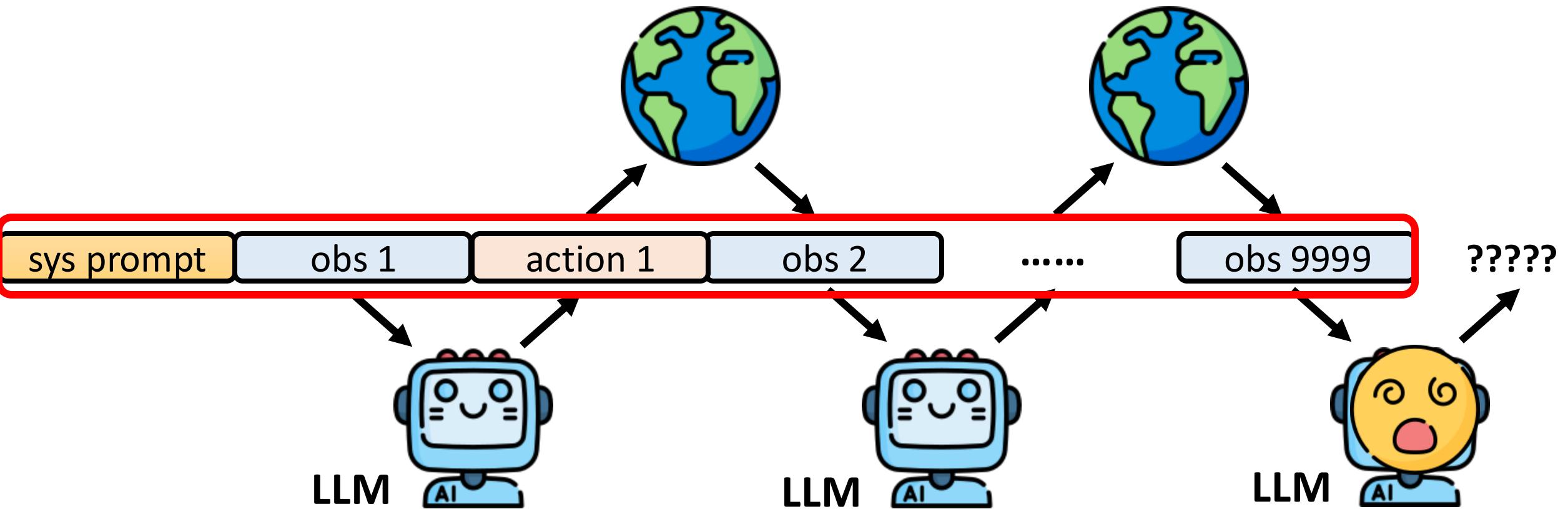
從 LLM 的角度來看 Agent 要解的問題



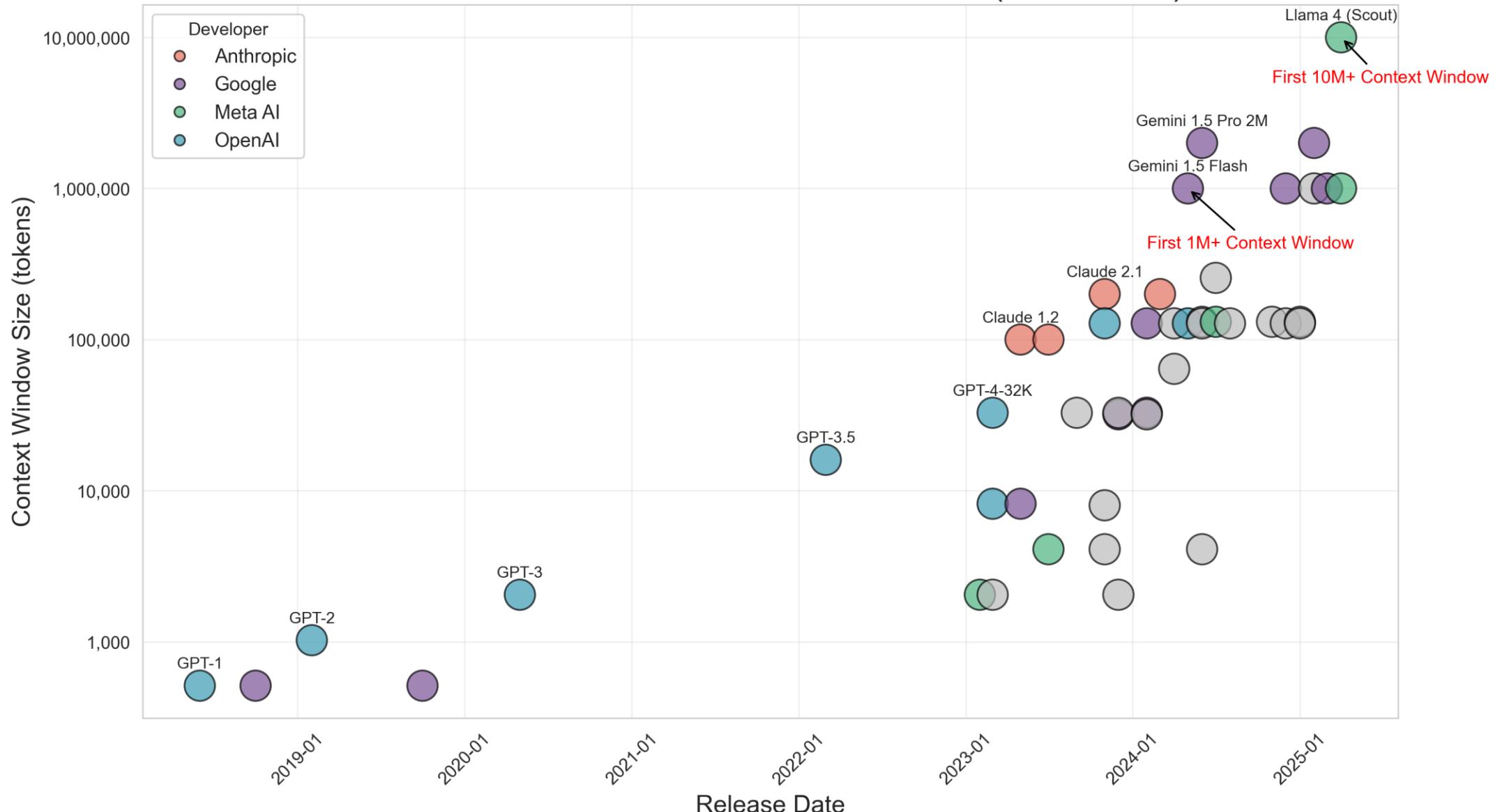
一直都在做接龍

AI Agent 倚靠的是語言模型現有的能力

運行 AI Agent 的挑戰：輸入過長



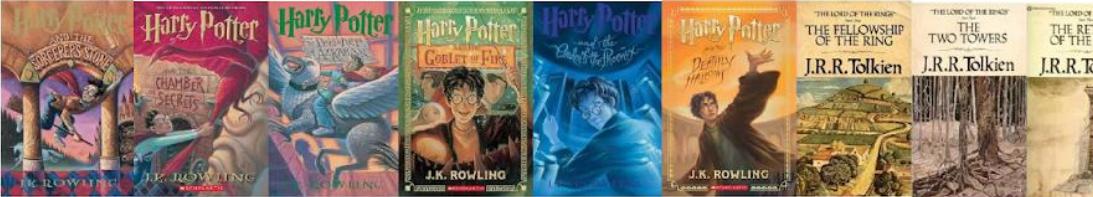
Evolution of LLM Context Window Sizes (2018-2025)



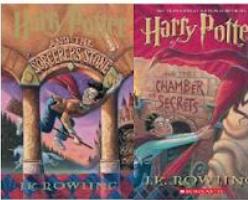
Source of image: <https://www.meibel.ai/post/understanding-the-impact-of-increasing-lm-context-windows>

Google's Gemini 1.5 can (almost) fit the entire Harry Potter + Lord of the Ring series in its 2 million context window

Gemini 1.5 2M
(June 2024)



Claude 2.1
(July 2023)



Source of image: <https://www.artfish.ai/p/long-context-langs>

GPT-4 Turbo
(March 2023)



能讀上百萬個 token，
並不代表能讀「懂」上百萬個 token

GPT-3.5 Turbo
(March 2022)

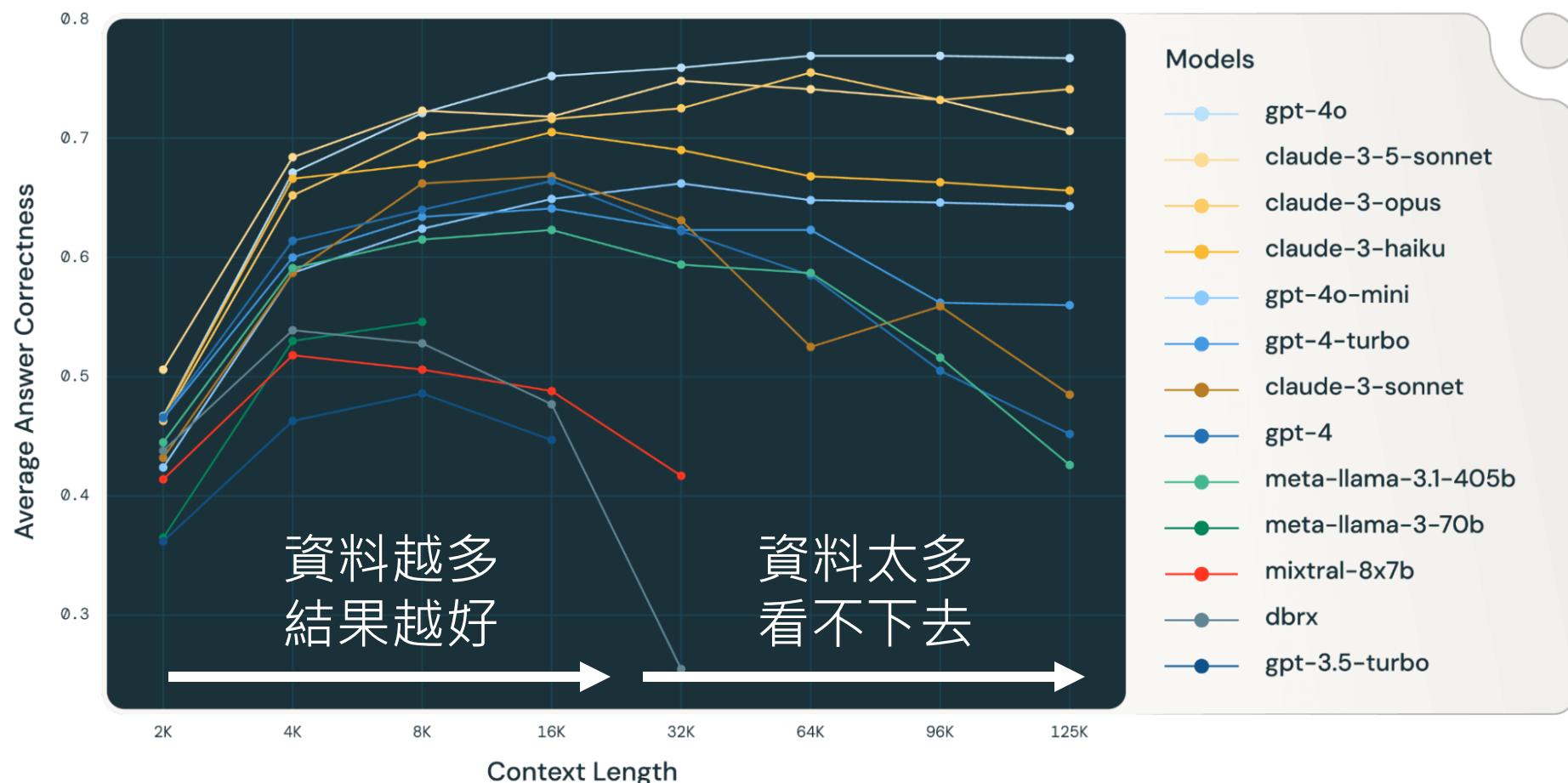


<https://www.dbreunig.com/2025/06/22/how-contexts-fail-and-how-to-fix-them.html>

輸入過長會有甚麼問題

<https://www.databricks.com/blog/long-context-rag-performance-langs>

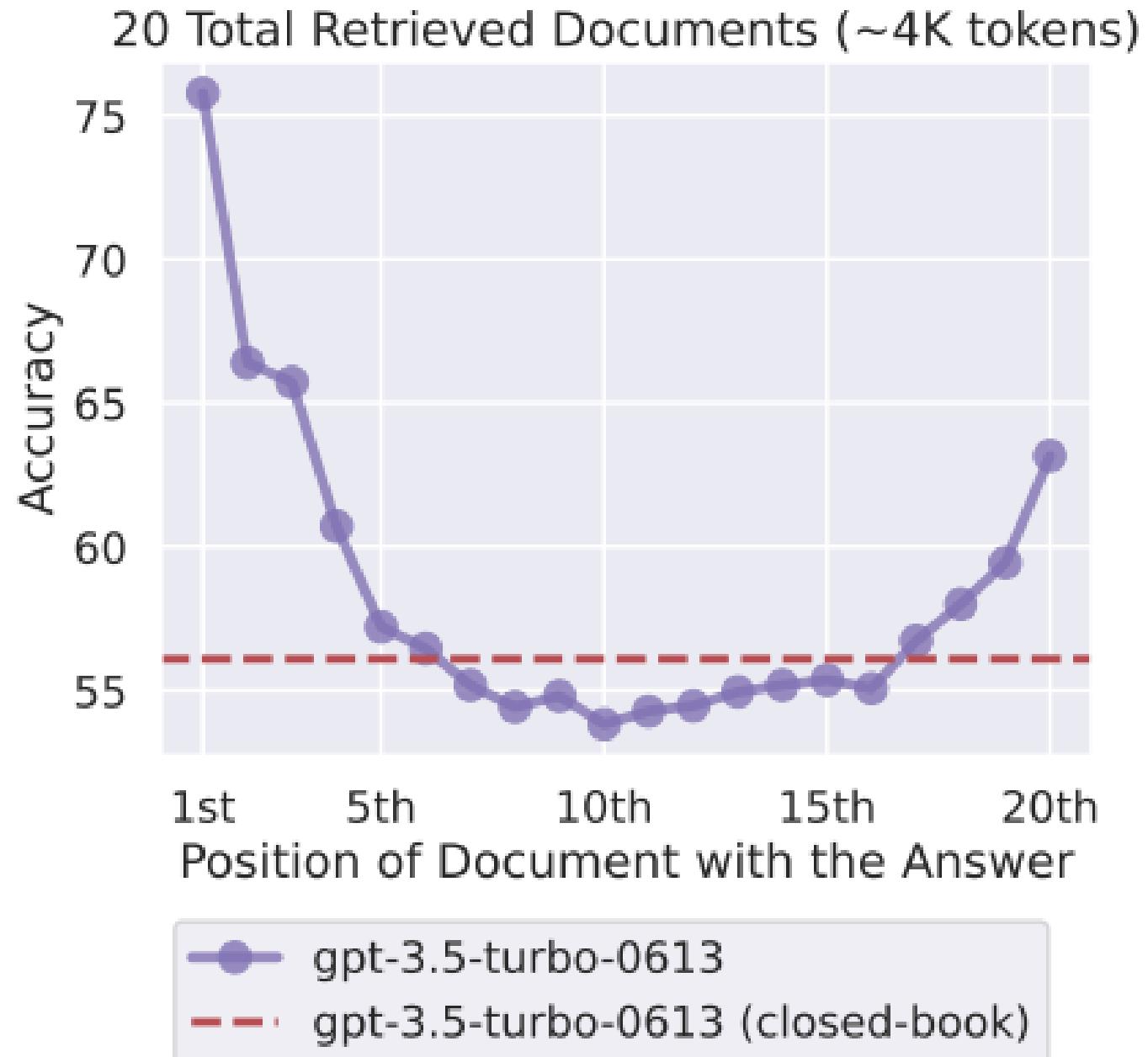
- RAG: 搜尋到的資料越多越好嗎？



Lost in the Middle

<https://arxiv.org/abs/2307.03172>

比較記得開頭跟結尾？



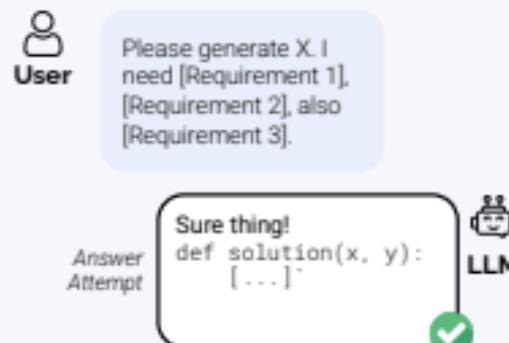
Lost in the Conversation?

<https://arxiv.org/pdf/2505.06120v1>

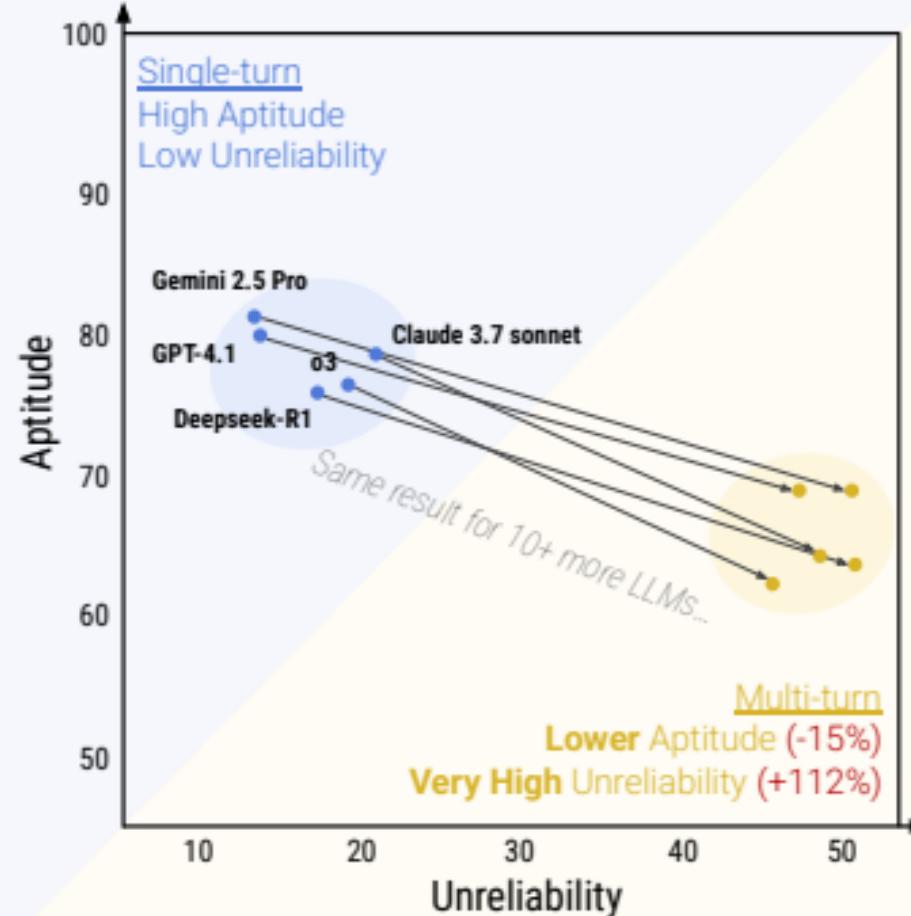
Fully-Specified

Jay is making snowballs for a snowman. He can build 2 snowmen per hour. If it takes 2 hours to melt every snowball, how long will it take to melt all the balls?

Single-Turn Fully-Specified



LLMs get Lost in Conversation



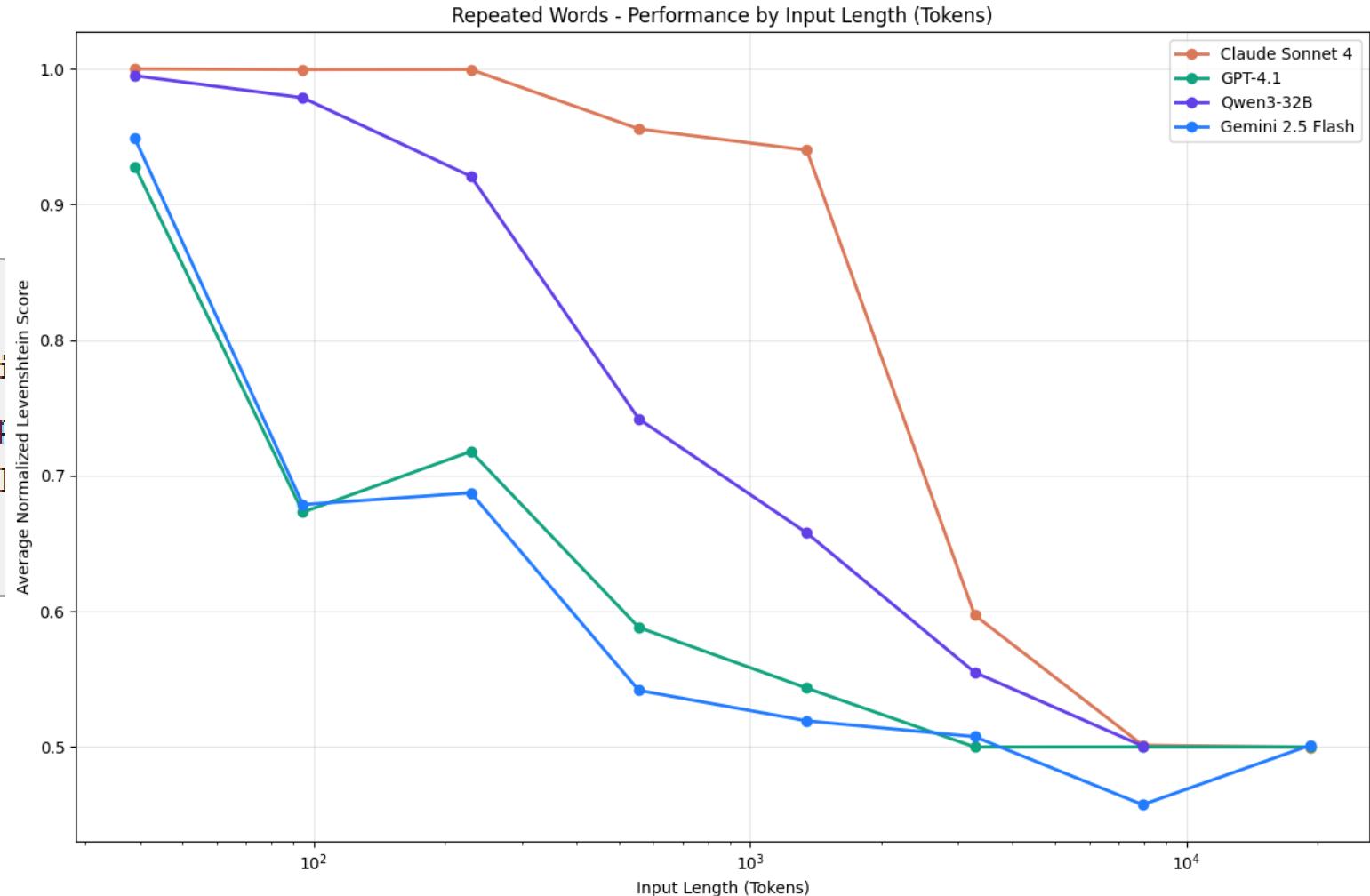
Multi-Turn Underspecified

- User: I'm trying to implement X.
- Clarification: Do you mean X'?
- User: No I want [Requirement 1].
- Premature Answer Attempt: Sure thing! def function(x): [...]
- User: Well, I also need that [Requirement 3].
- Incorrect Assumption: Oh, in that case: def function(x, y): [...]
- User: One more thing, can you include [Requirement 2]?
- Bloated Answer: Absolutely, here it is: def function(y, x): [...]

"Context Rot: How Increasing Input Tokens Impacts LLM Performance"

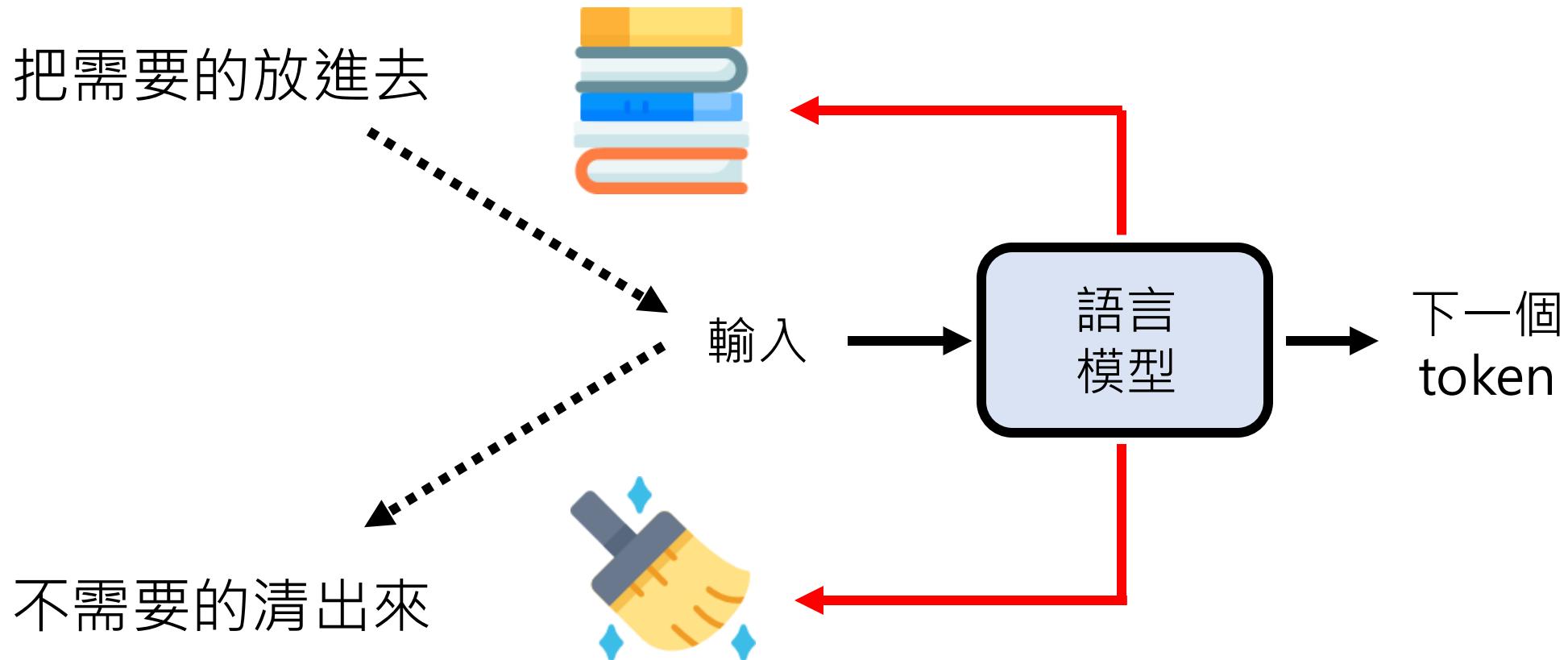
<https://research.trychroma.com/context-rot>

Simply replicate the following:
apple apples apple apple app
apple apple apple apple appl



Context Engineering 的基本方法

Context Engineering 的基本概念



常用招數

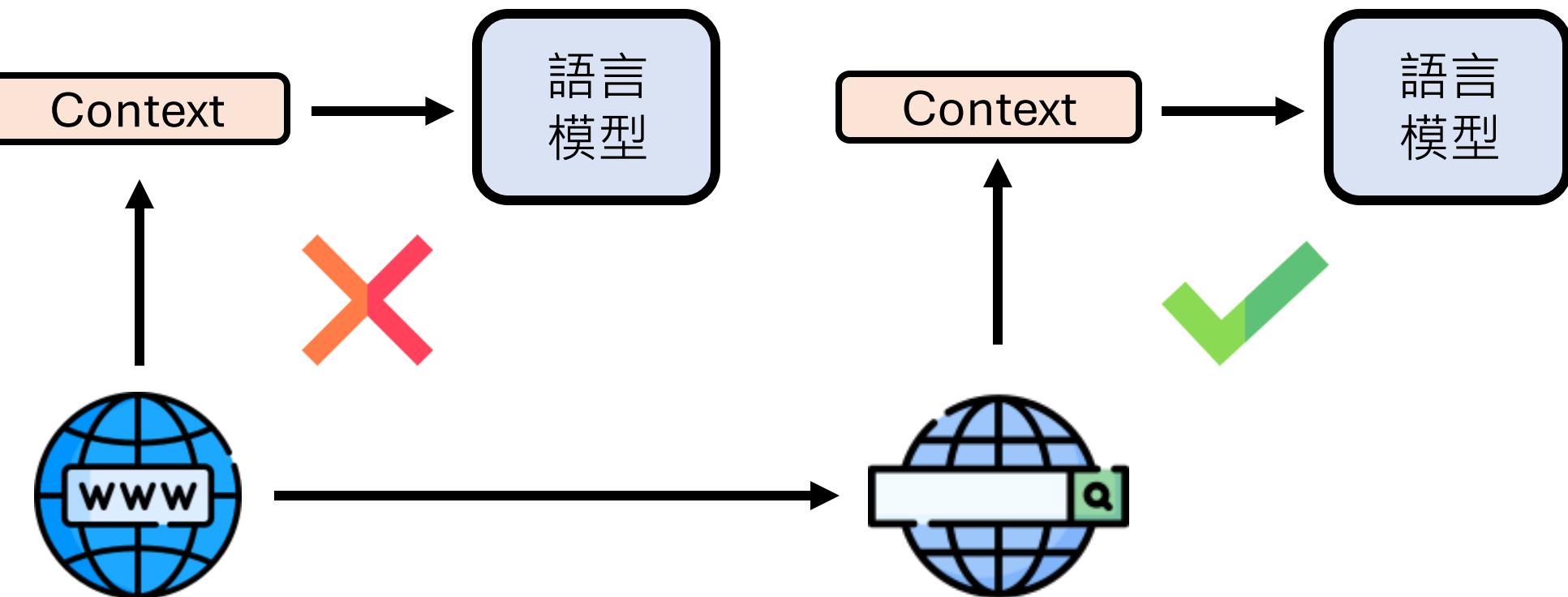
Select

Compress

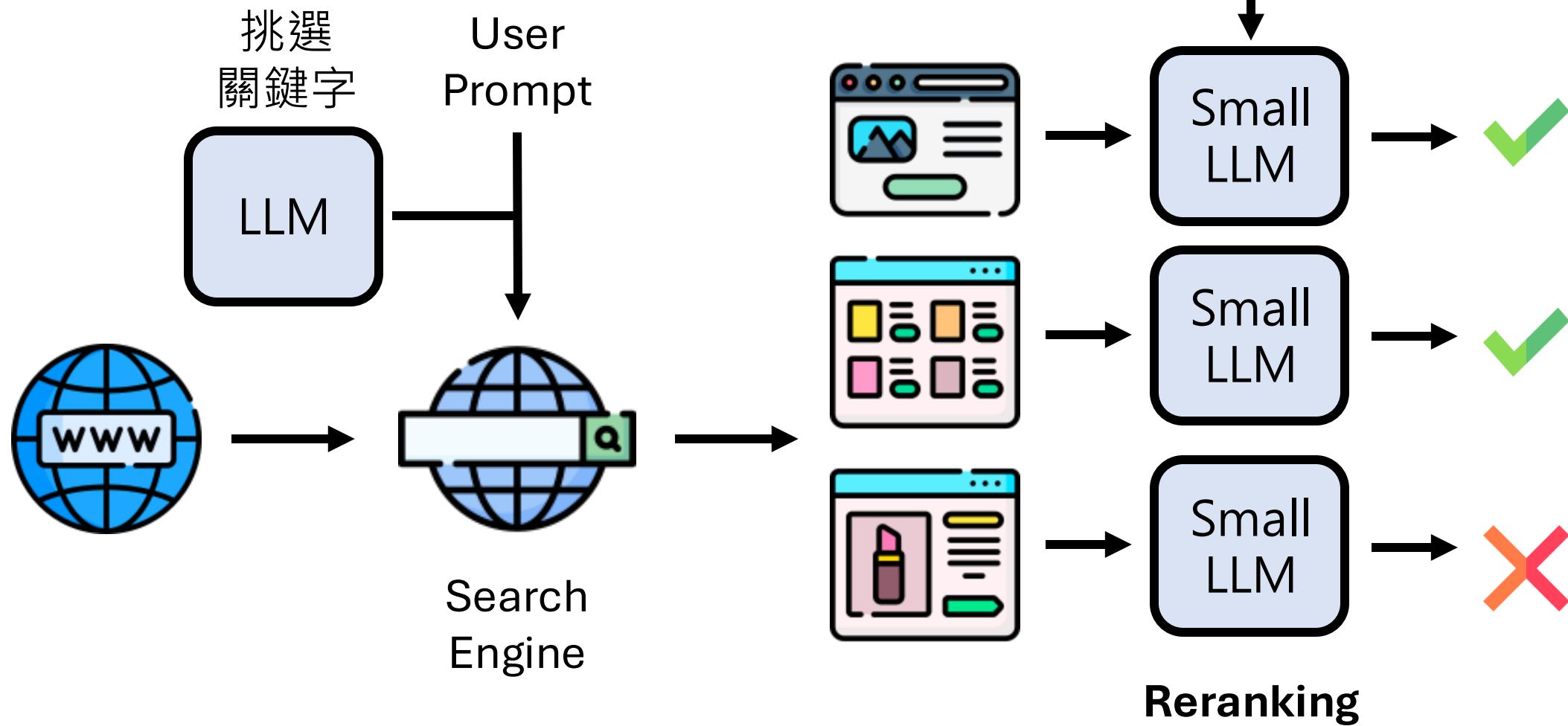
Multi-
Agent

挑選需要的內容

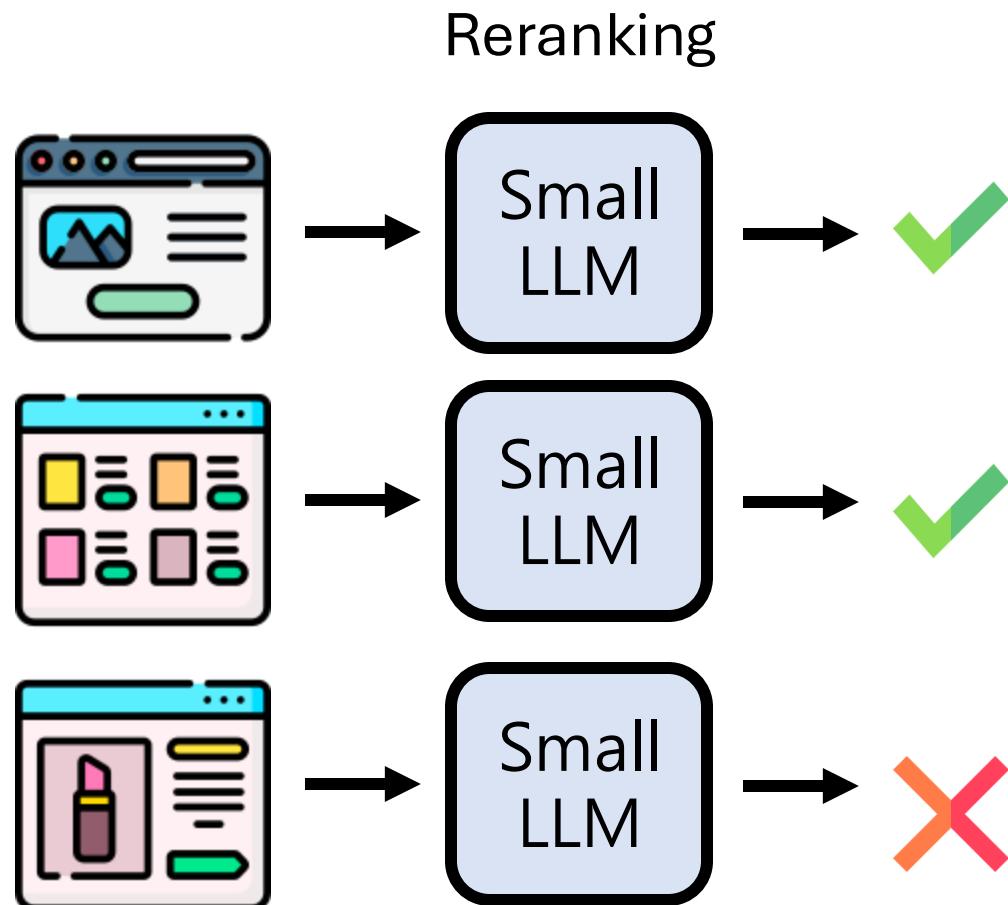
- RAG



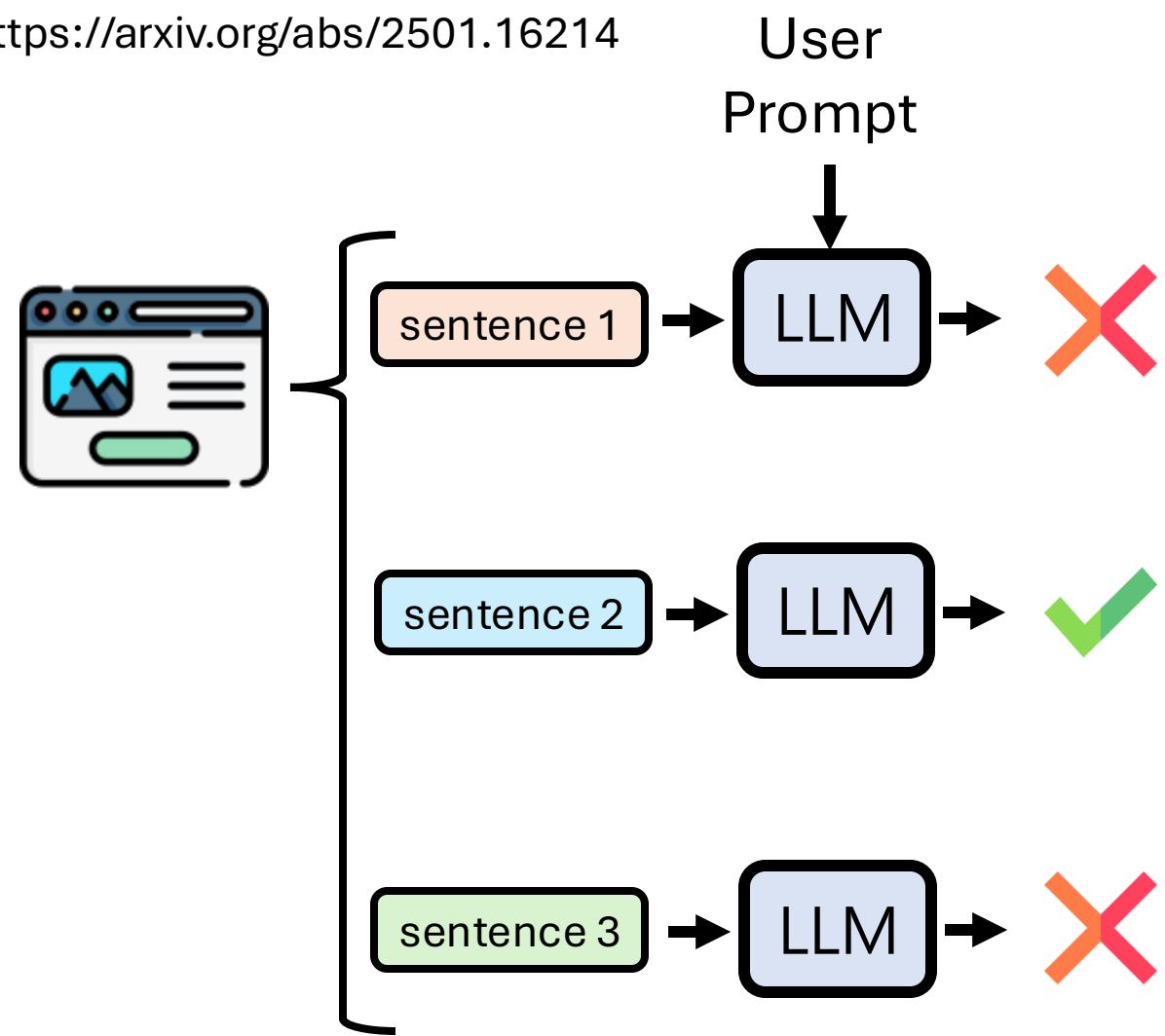
挑選需要的內容



挑選需要的內容

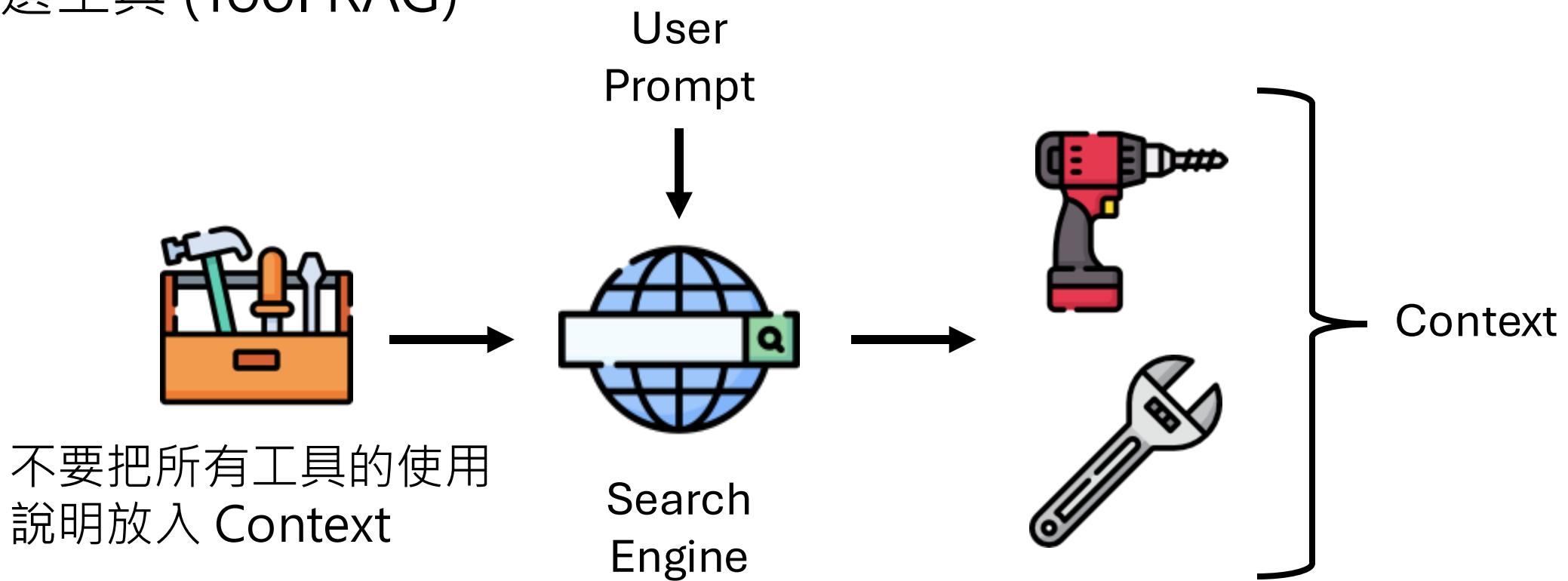


Provence
<https://arxiv.org/abs/2501.16214>



挑選需要的內容

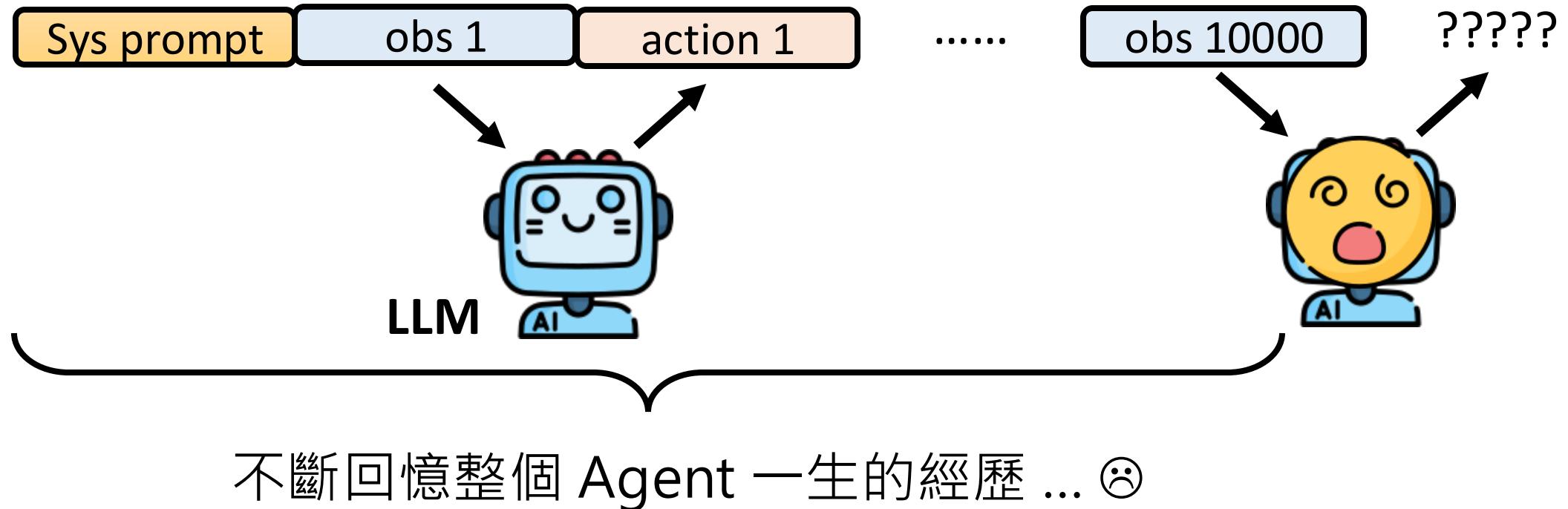
- 挑選工具 (Tool RAG)



<https://arxiv.org/abs/2310.03128>
<https://arxiv.org/abs/2502.11271>
<https://arxiv.org/abs/2505.03275>

挑選需要的內容

- 挑選記憶 (Memory RAG)



挑選需要的內容

上課錄影：

<https://www.youtube.com/watch?v=G44Lkj7XDsa>

Paper: <https://arxiv.org/abs/2304.03442>



Memory Stream

```
2023-02-13 22:48:20: desk is idle
2023-02-13 22:48:20: bed is idle
2023-02-13 22:48:10: closet is idle
2023-02-13 22:48:10: refrigerator is idle
2023-02-13 22:48:10: Isabella Rodriguez is stretching
2023-02-13 22:33:30: shelf is idle
2023-02-13 22:33:30: desk is neat and organized
2023-02-13 22:33:10: Isabella Rodriguez is writing in her journal
2023-02-13 22:18:10: desk is idle
2023-02-13 22:18:10: Isabella Rodriguez is taking a break
2023-02-13 21:49:00: bed is idle
2023-02-13 21:48:50: Isabella Rodriguez is cleaning up the
kitchen
2023-02-13 21:48:50: refrigerator is idle
2023-02-13 21:48:50: bed is being used
2023-02-13 21:48:10: shelf is idle
2023-02-13 21:48:10: Isabella Rodriguez is watching a movie
2023-02-13 21:19:10: shelf is organized and tidy
2023-02-13 21:18:10: desk is idle
2023-02-13 21:18:10: Isabella Rodriguez is reading a book
2023-02-13 21:03:40: bed is idle
2023-02-13 21:03:30: refrigerator is idle
2023-02-13 21:03:30: desk is in use with a laptop and some papers
on it
```

...

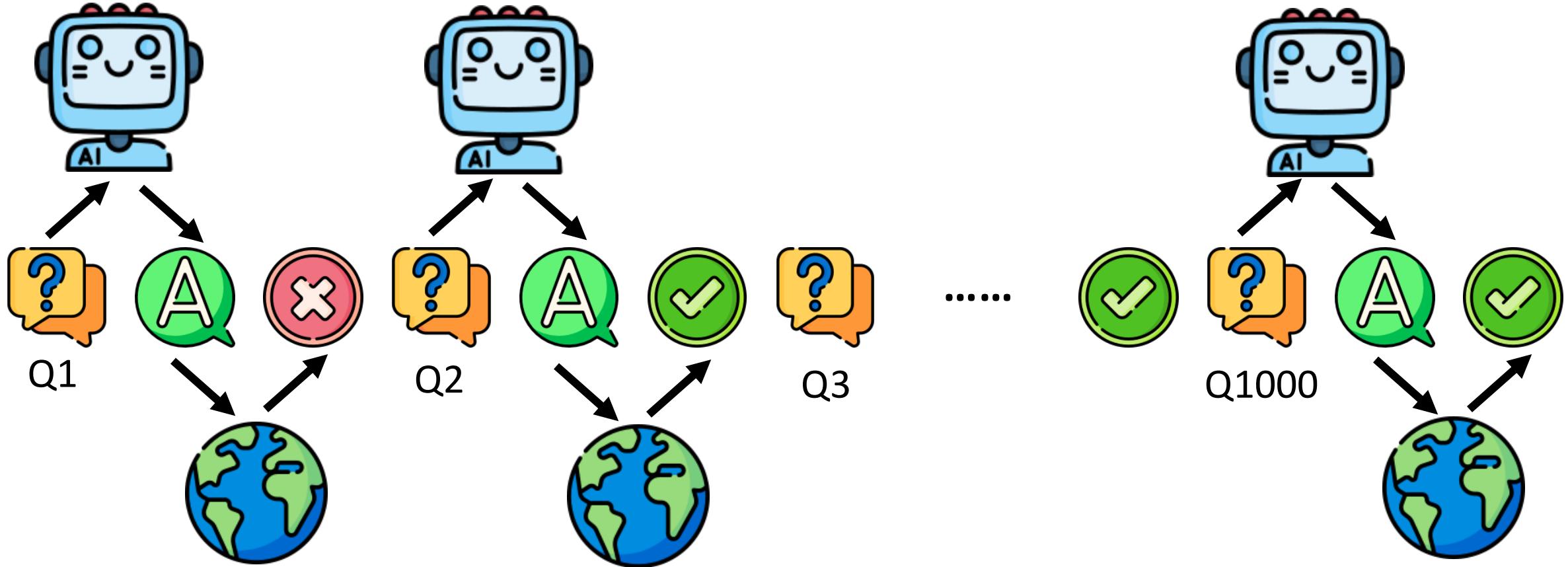
不進 Context，另外儲存



StreamBench

<https://arxiv.org/abs/2406.08747>

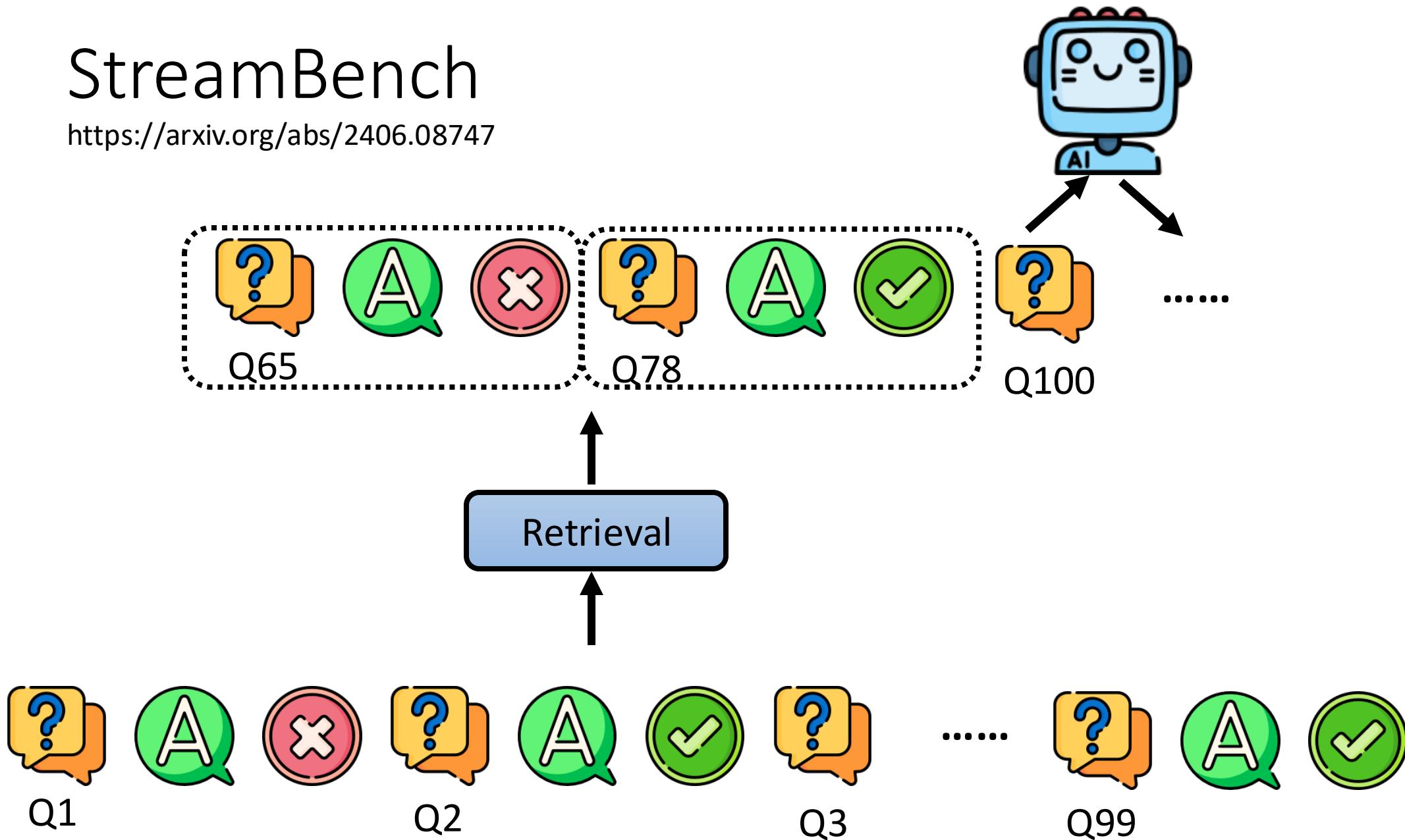
<https://stream-bench.github.io/>
(done by Appier Researchers)



Goal: Maximize the accuracy over the sequence

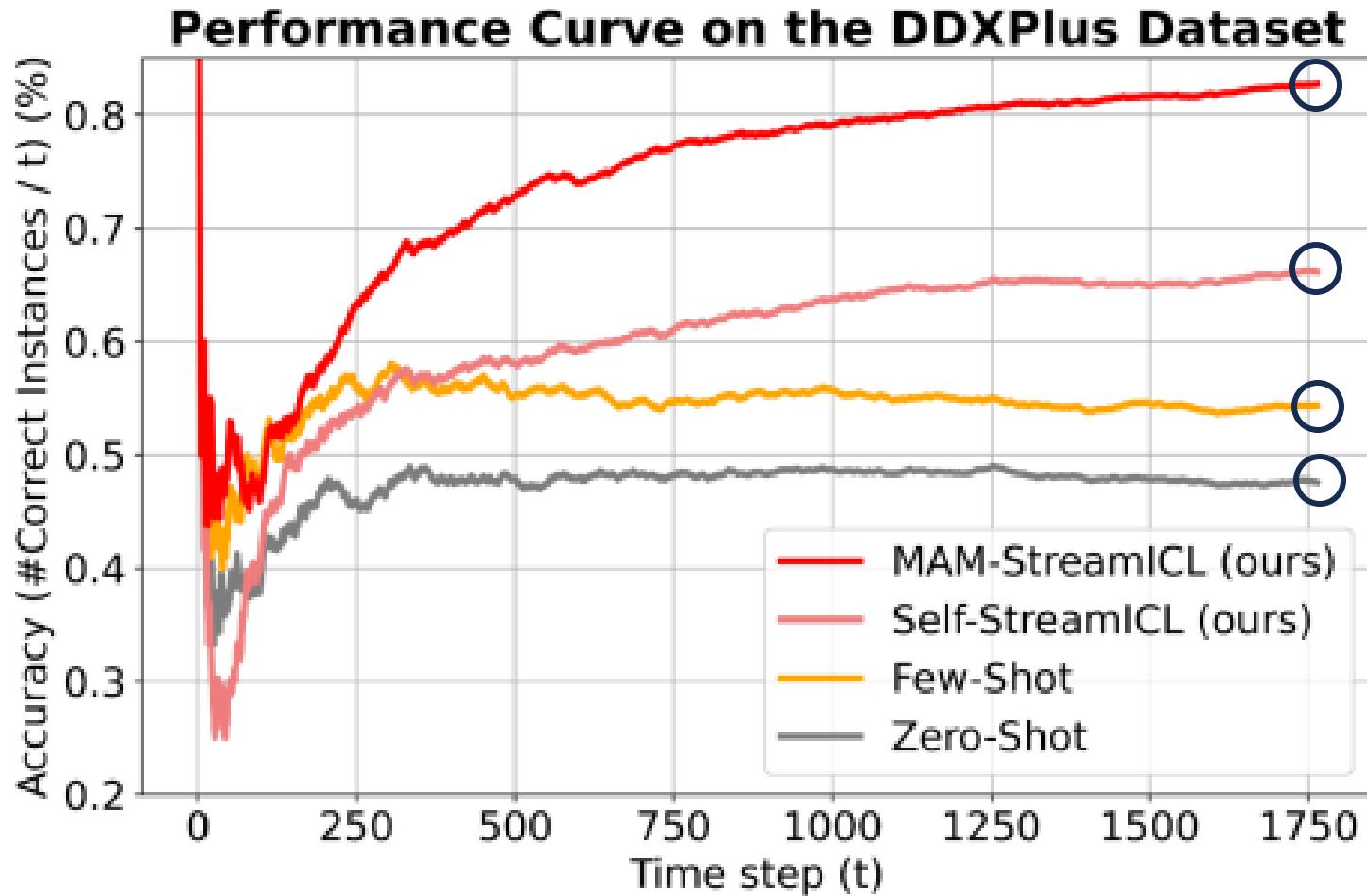
StreamBench

<https://arxiv.org/abs/2406.08747>



StreamBench

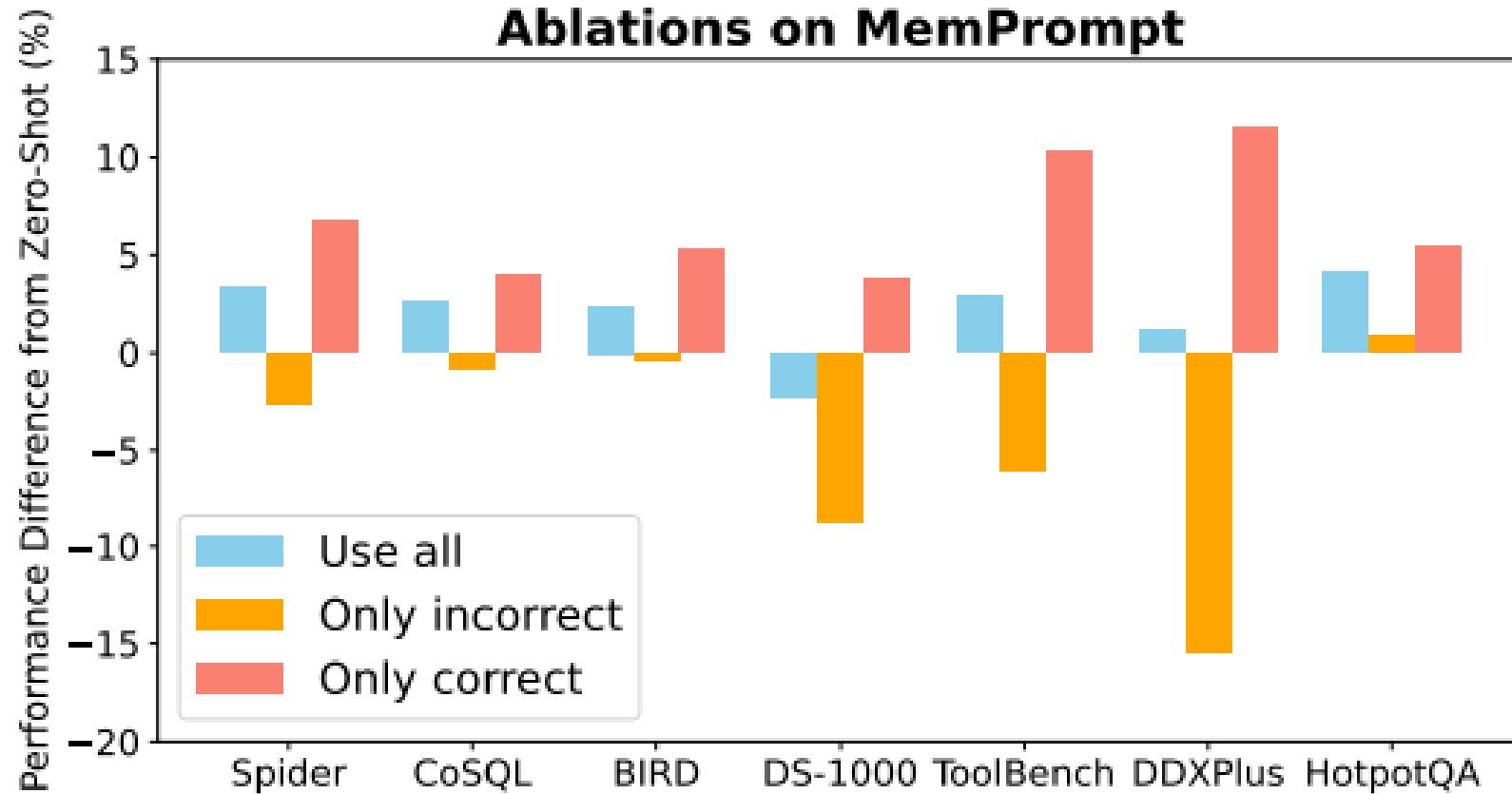
<https://arxiv.org/abs/2406.08747>



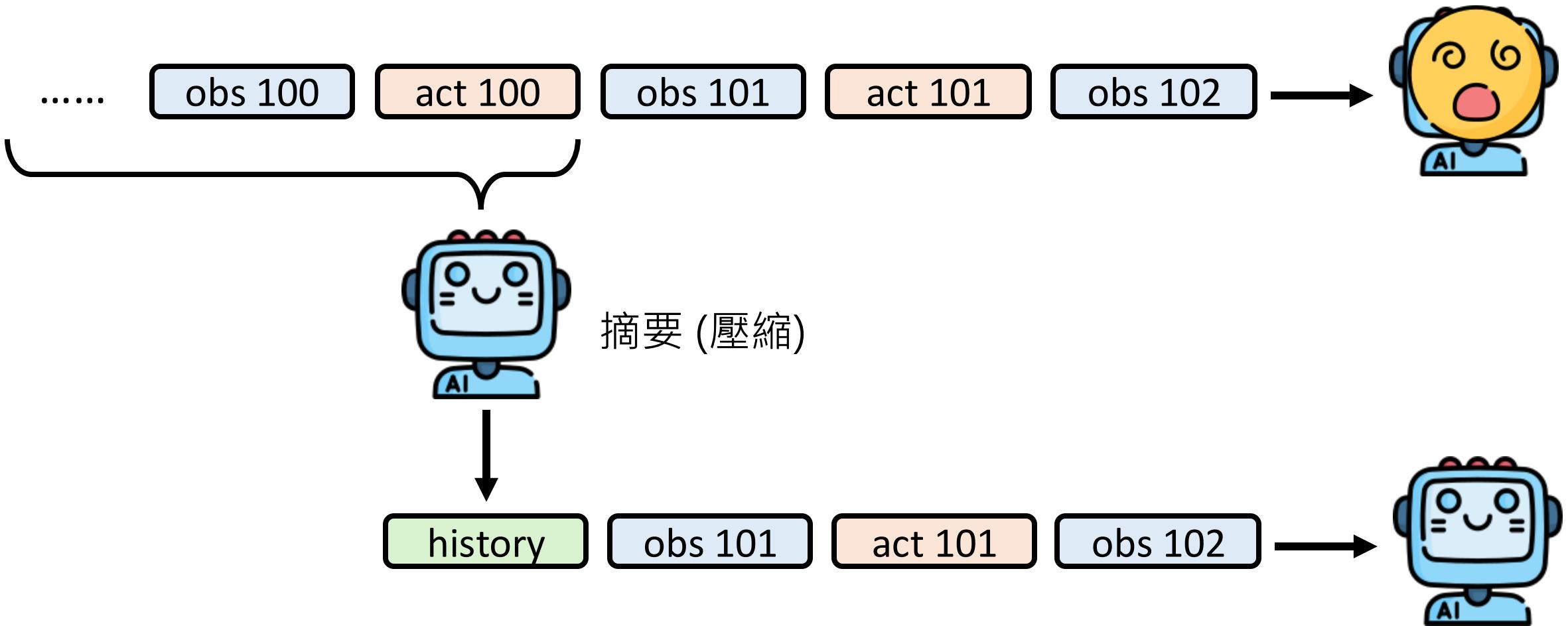
StreamBench

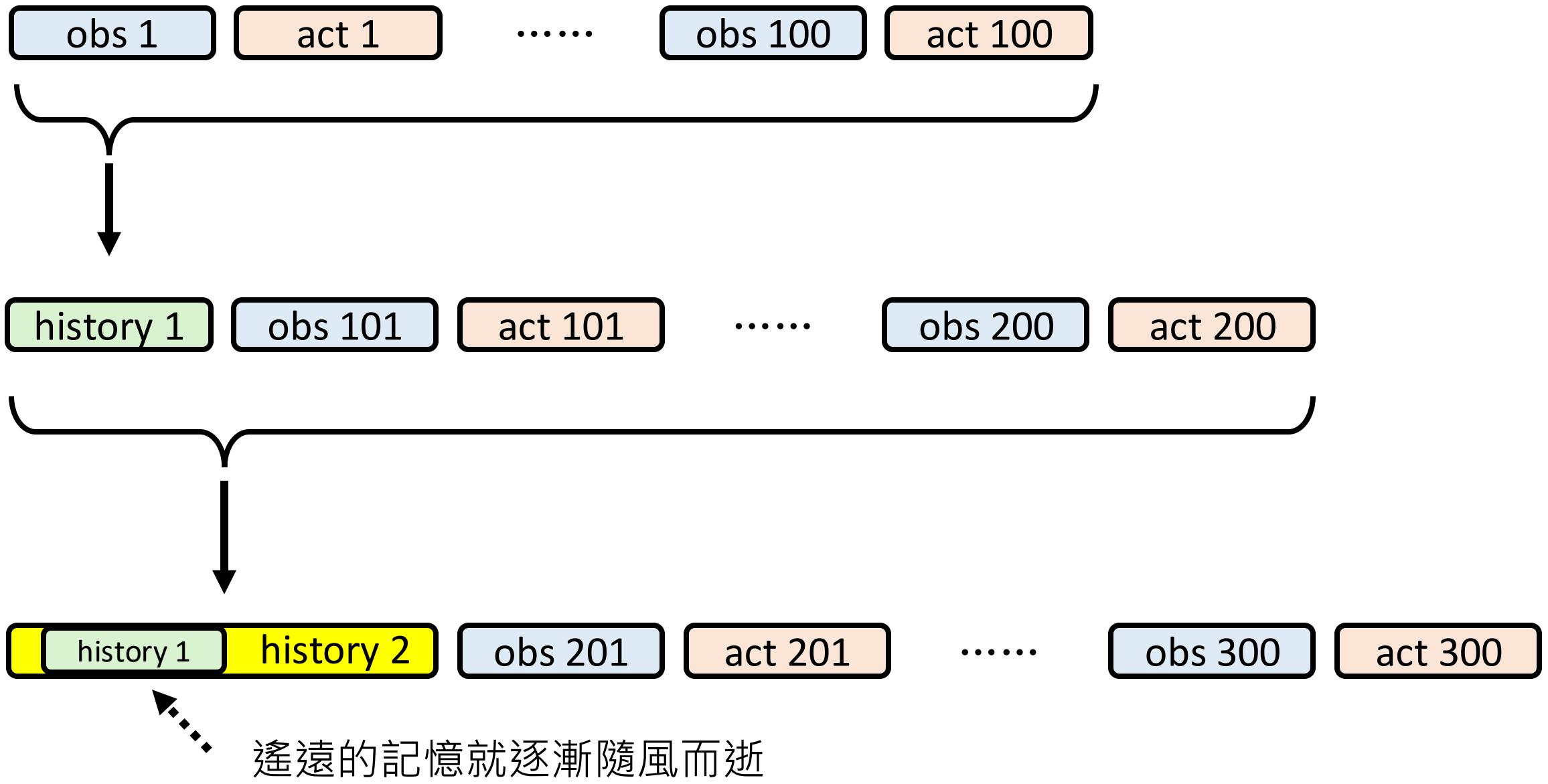
<https://arxiv.org/abs/2406.08747>

叫你不要想「白熊」，反而特別容易想「白熊」



壓縮內容

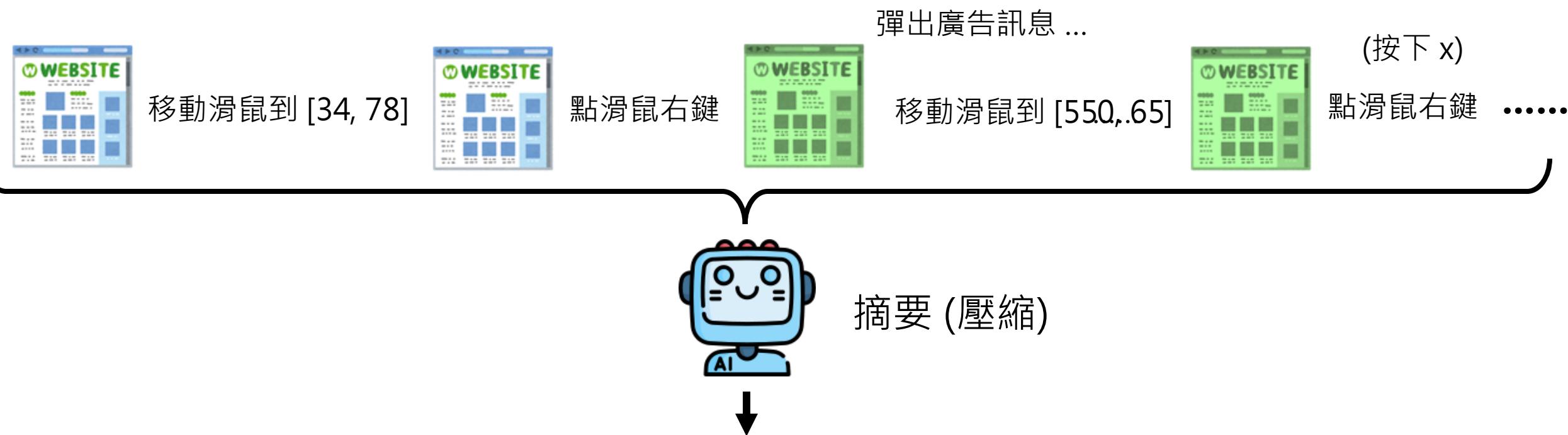




壓縮內容

<https://alignment.anthropic.com/2025/summarization-for-monitoring/>

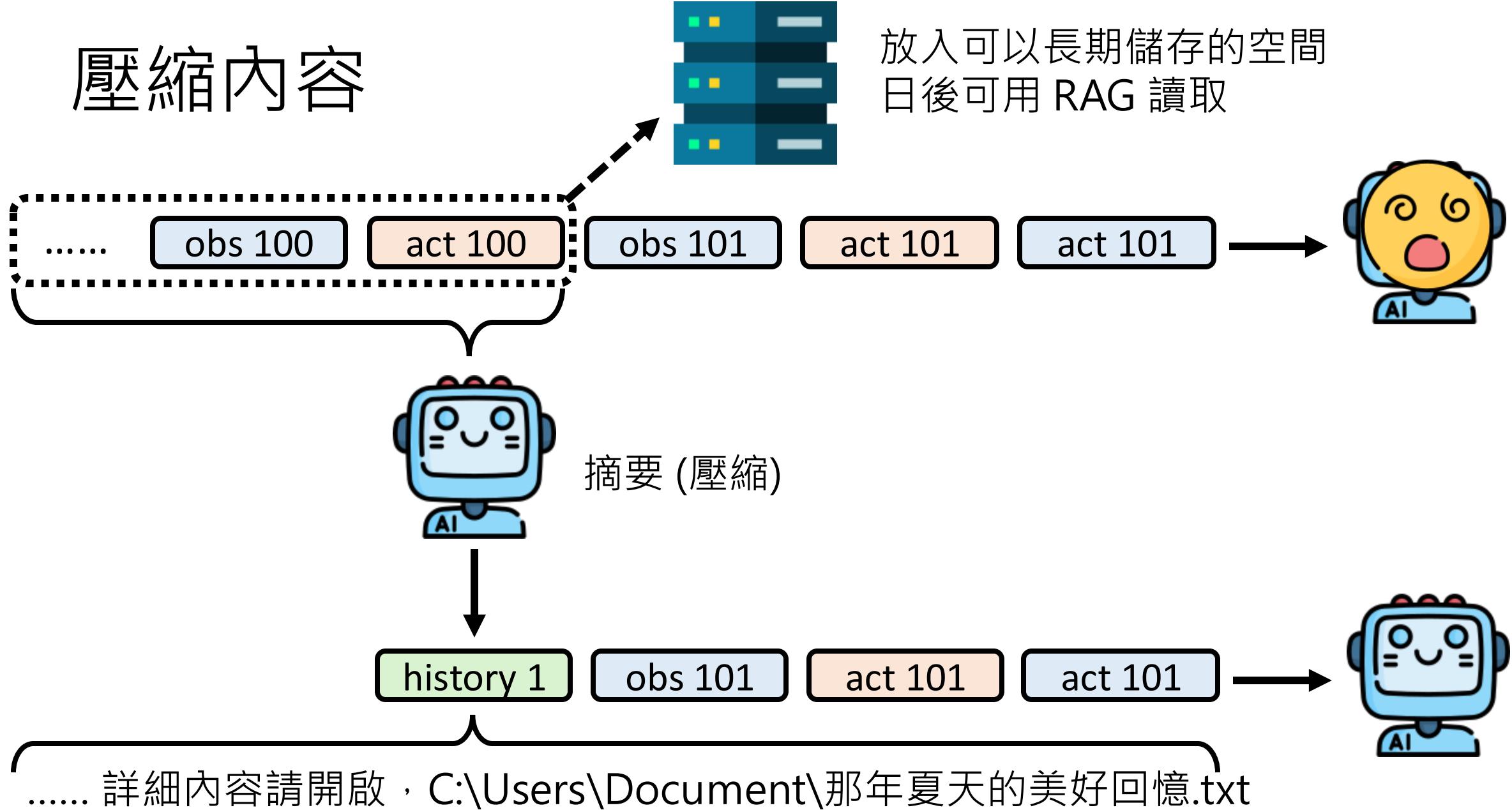
Computer Use 會在 Context 中產生瑣碎的內容



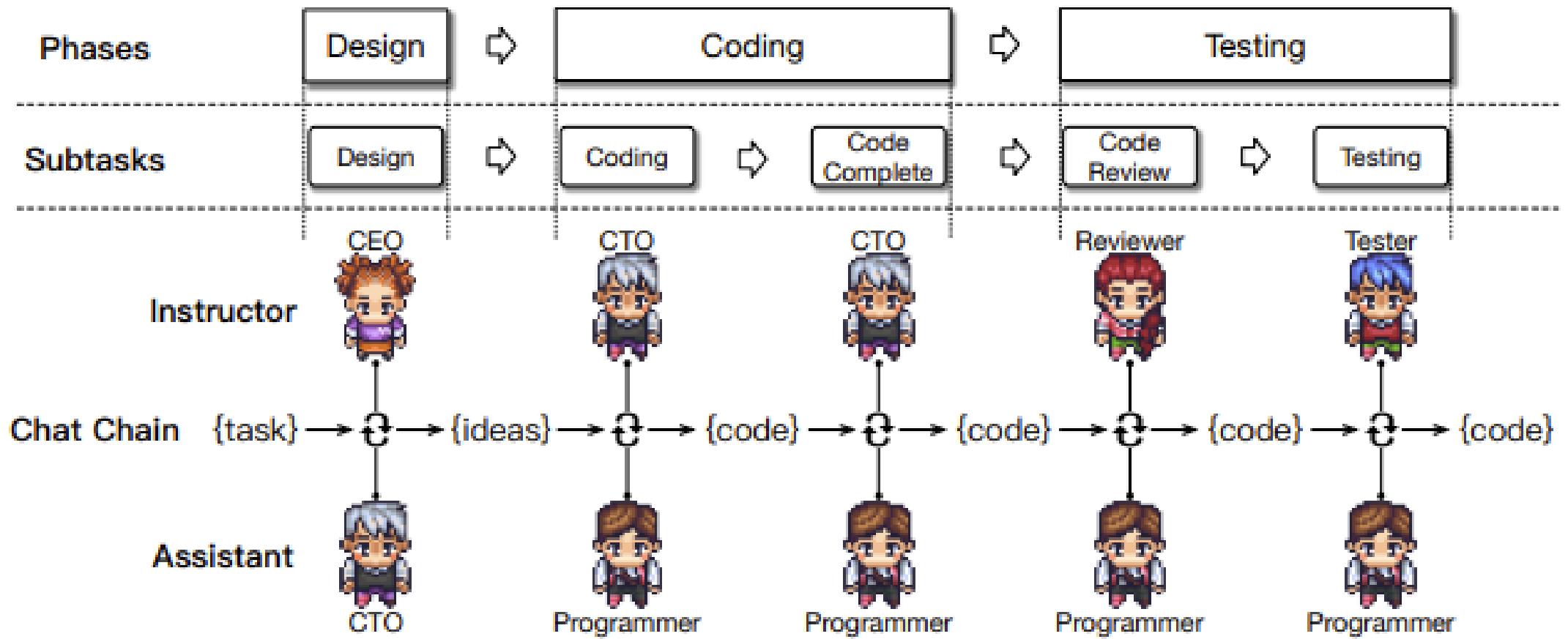
A 餐廳訂位成功，9/19 下午 6:00，十人

(訂位成功後，Context 不需要訂位的詳細過程)

壓縮內容

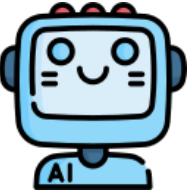


Multi-Agent

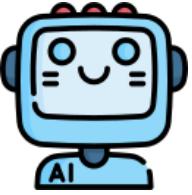


ChatDev: <https://github.com/OpenBMB/ChatDev>
arXiv: <https://arxiv.org/abs/2307.07924>

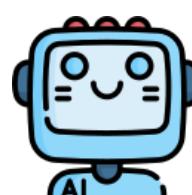
Single Agent



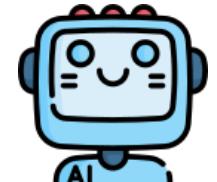
組織出遊



規劃行程

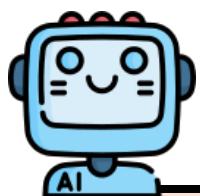
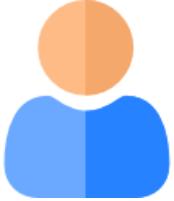


去訂餐廳

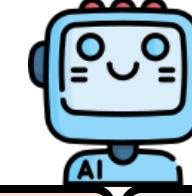


跟餐廳網頁一番互動

訂好了



去訂旅館



跟旅館網頁一番互動



Multiple Agent

Agent 1



Context中只有訂餐廳
沒有其他事情

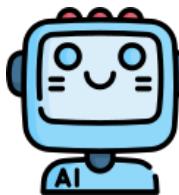
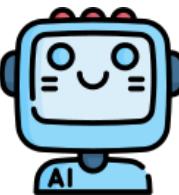


去訂餐廳

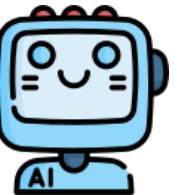
跟餐廳網頁一番互動

回報: 訂好了

Lead



Context 中沒有
各種訂位細節



組織出遊

規劃行程

Agent1(去訂餐廳)

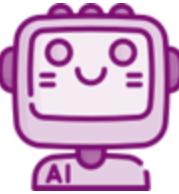
訂好了

Agent2(去訂旅館)

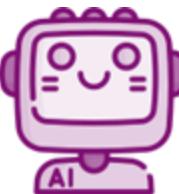
訂好了



Agent 2



Context中只有訂旅館
沒有其他事情



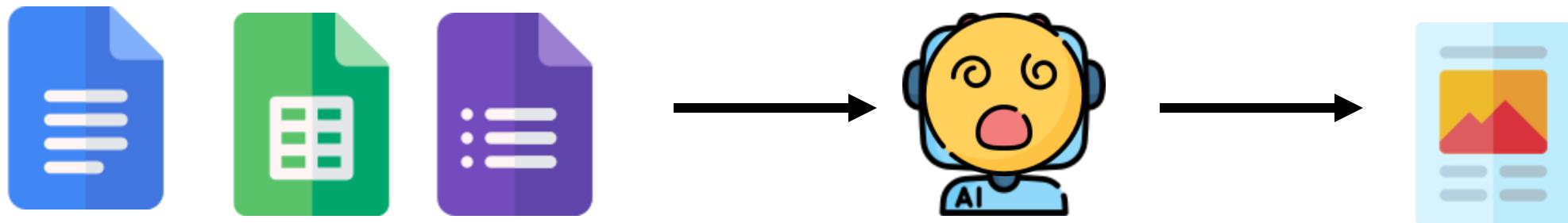
去訂旅館

跟旅館網頁一番互動

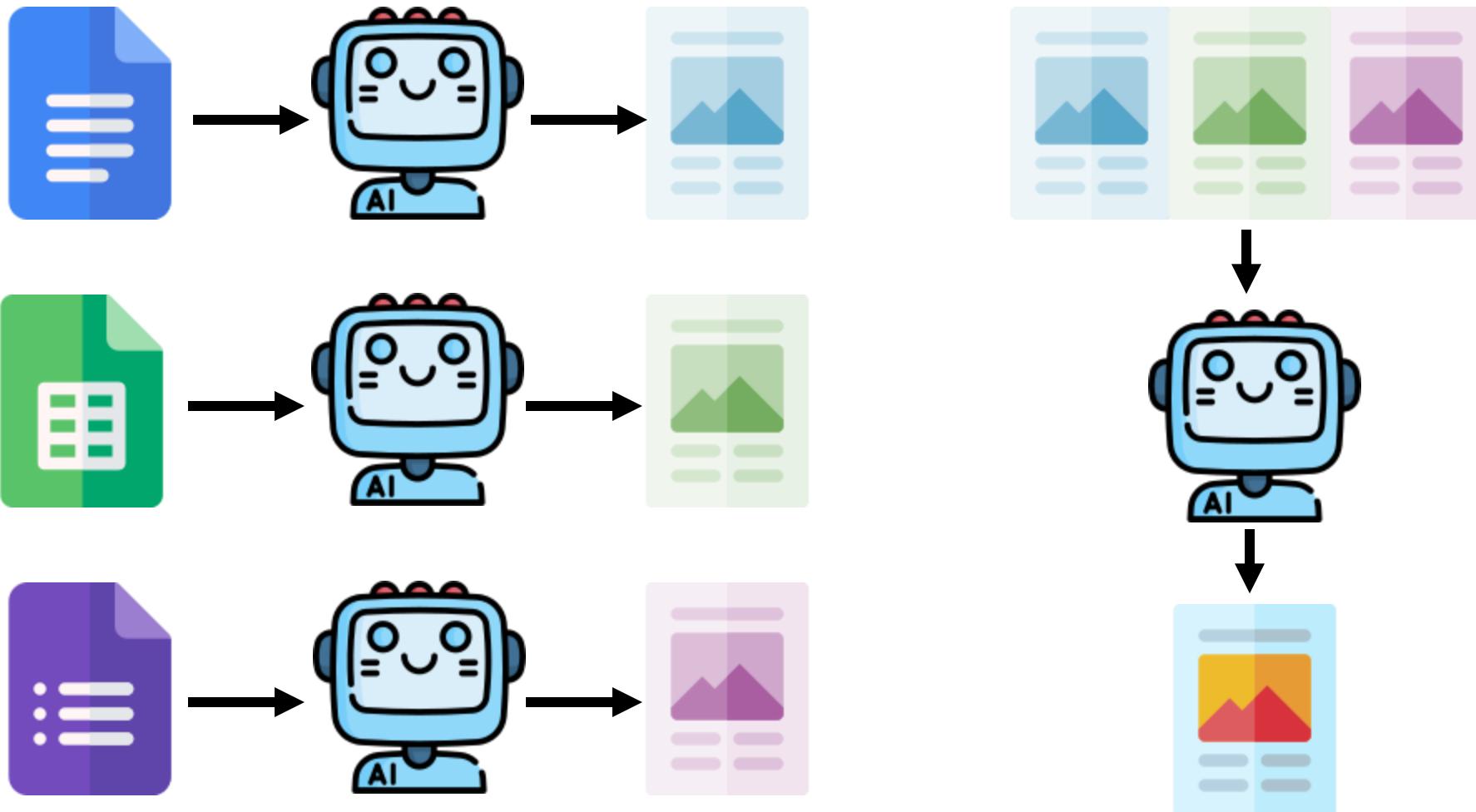
回報: 訂好了

Multi-Agent

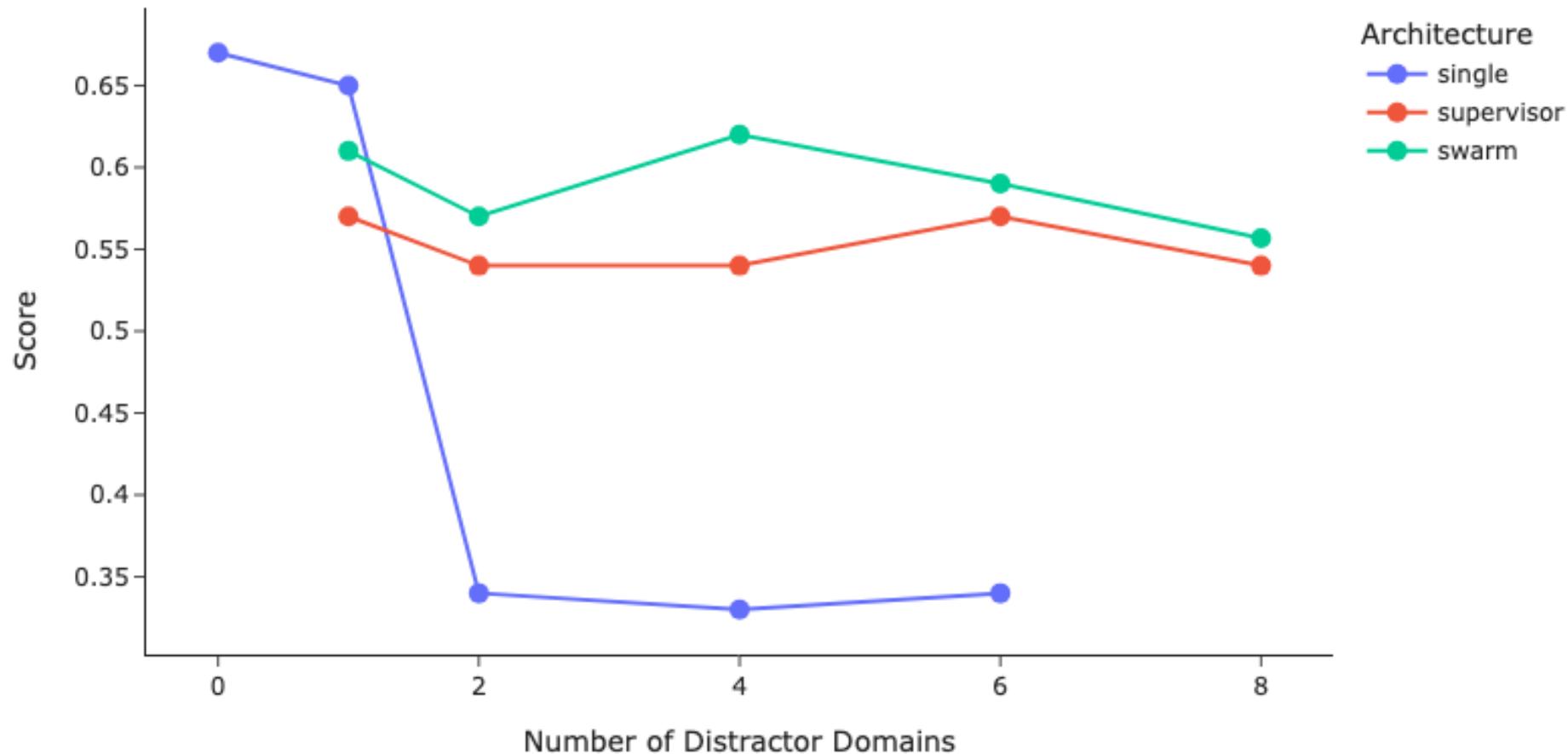
撰寫 overview paper 為例



Multi-Agent



Multi-agent scaling performance



<https://blog.langchain.com/benchmarking-multi-agent-architectures/>

課程總結

Context 裡面需要有什麼？

為什麼AI Agent 時代，需要 Context Engineering

Context Engineering 的基本方法