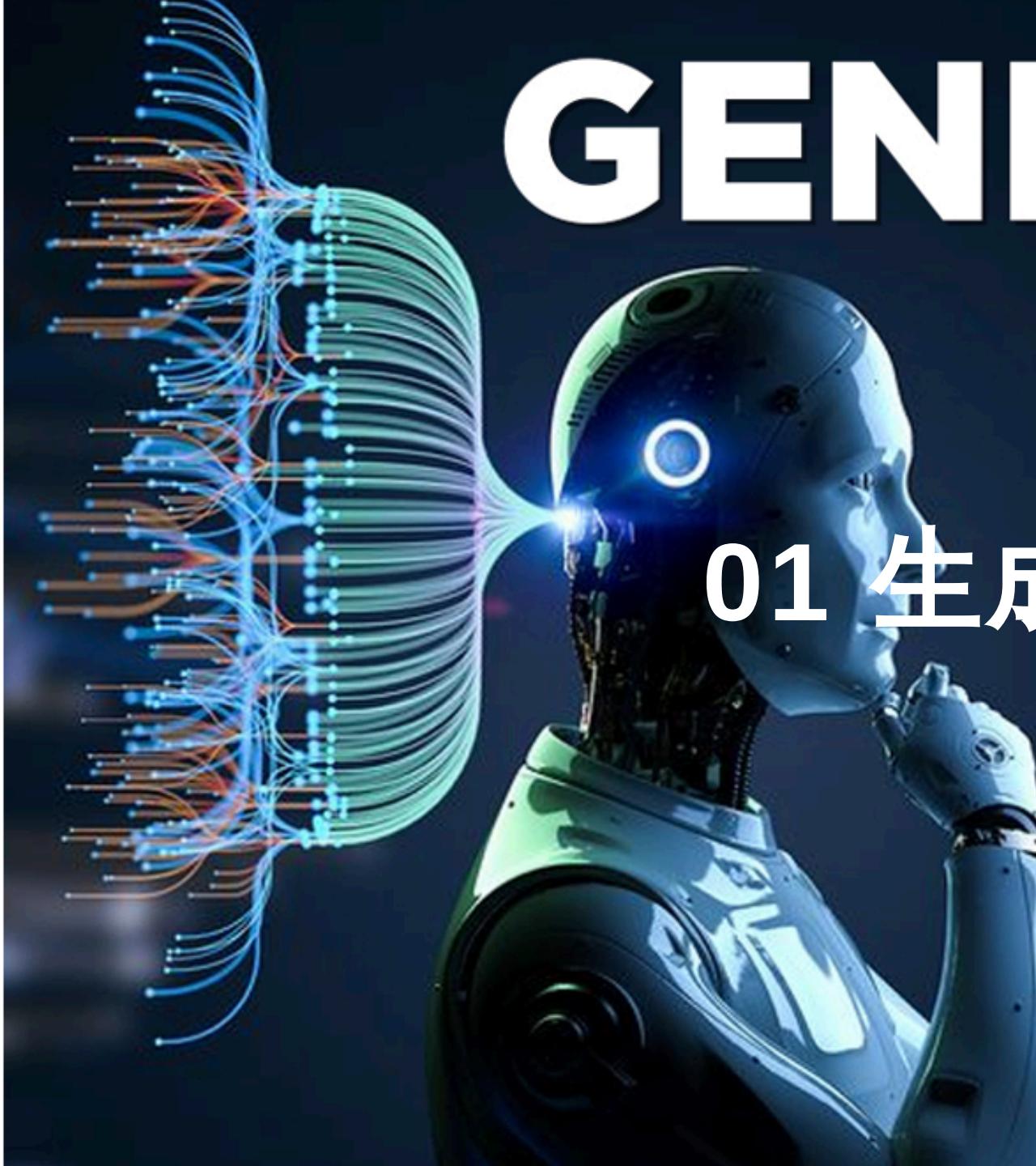


# 生成式AI與大語言模型

江豪文

# 大綱

- 生成式 AI
- 大語言模型



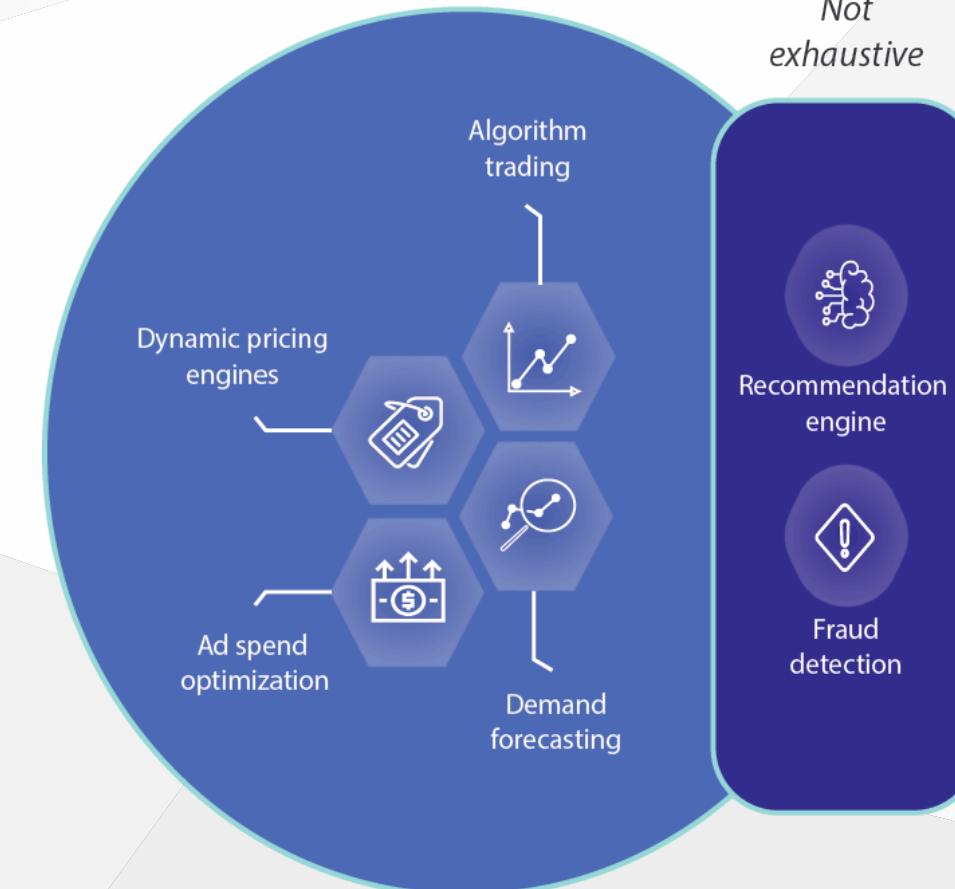
# GENERATIVE AI

01 生成式 AI

# 兩大 AI 類型

- 分辨式
- 生成式

## Discriminative uses of AI



## Generative uses of AI



# 分散式 AI: 預測數值與類別

## Regression

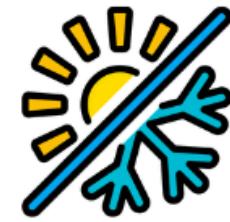


What will be the temperature tomorrow?

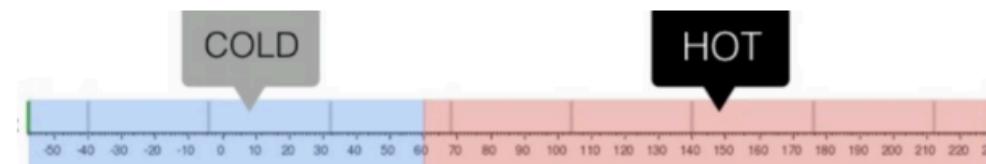


Fahrenheit

## Classification



Will it be hot or cold tomorrow?



Fahrenheit

# 生成式 AI 範例

“

Bill Gates:

**ChatGPT's history  
is as significant as  
the birth of the PC  
or the Internet**

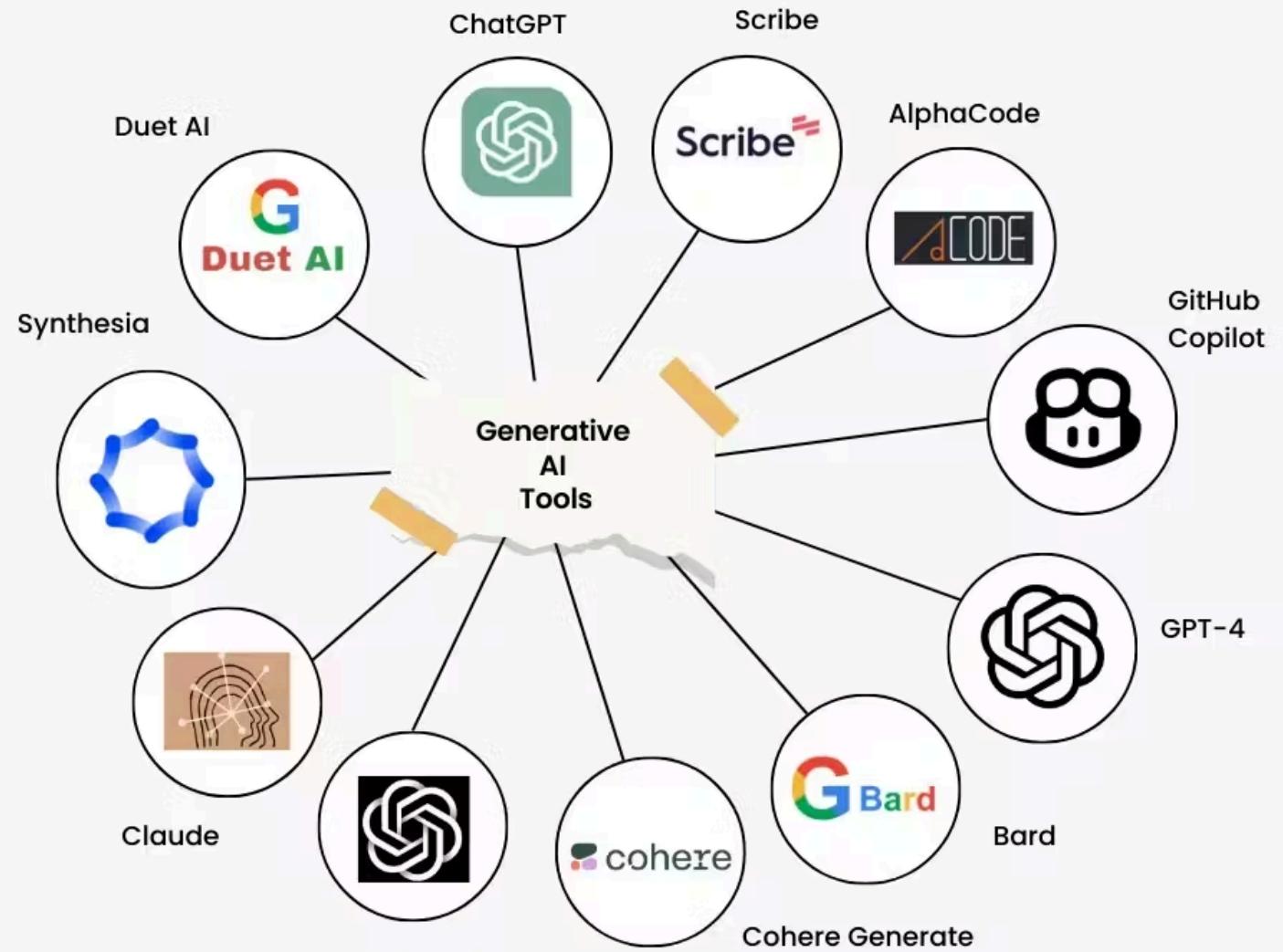
~TechGoing

”

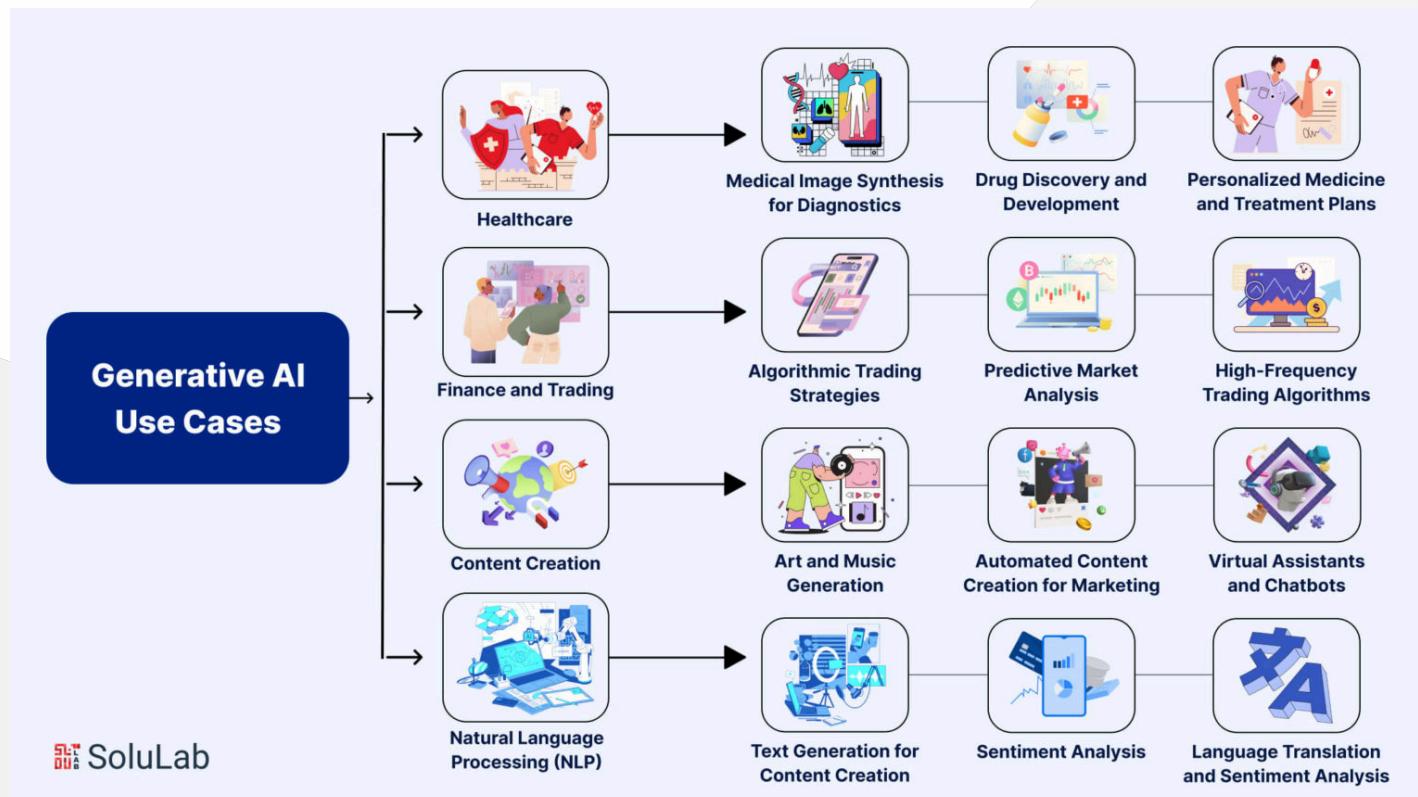


# 生成式 AI 工具

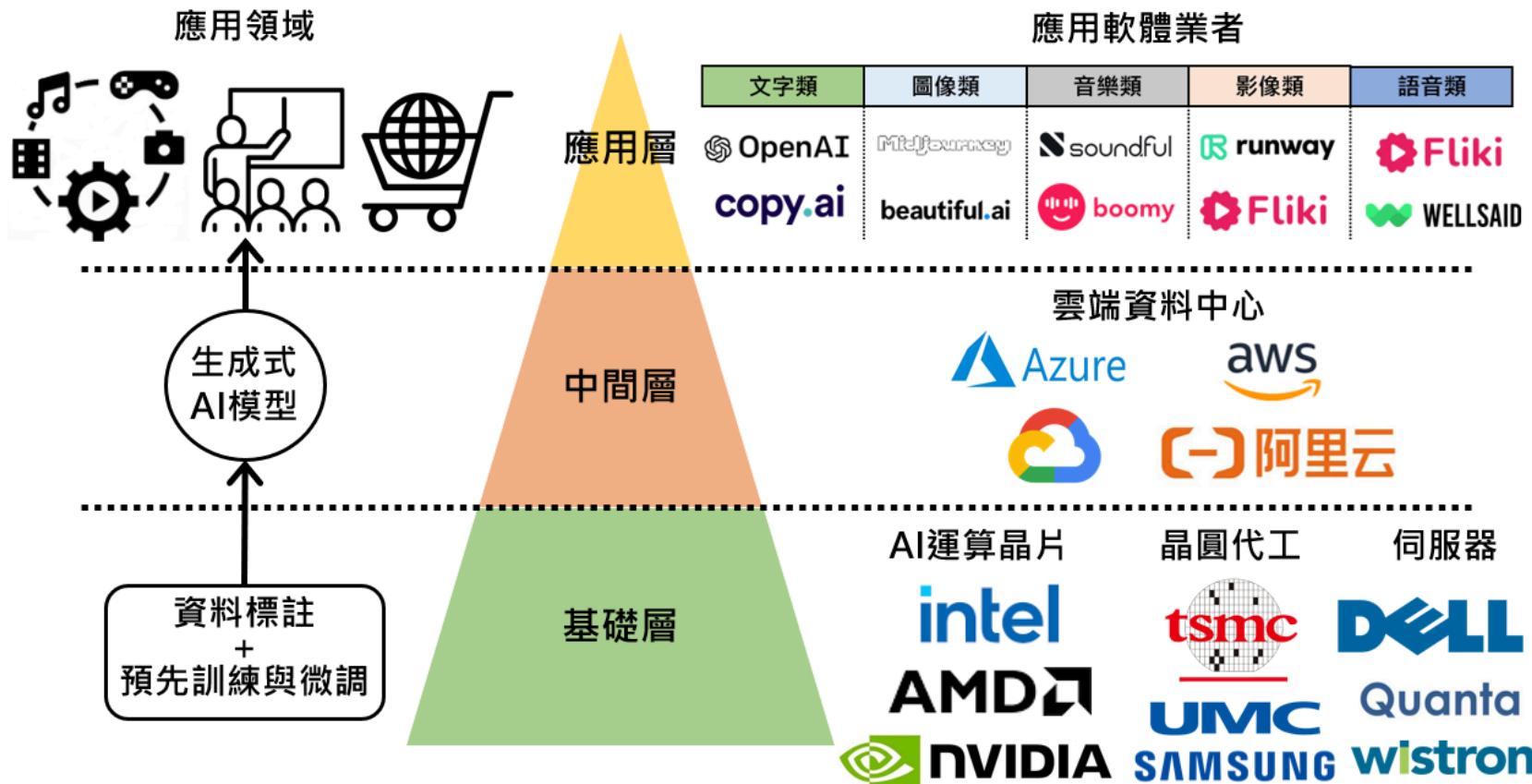
01 > 生成式 AI



# 生成式 AI 的應用情境



# 生成式 AI 生態鏈



# 生成式 AI 的市值預估

Generative AI Market Forecast  
by Revenue and Technology Spend



Source: Bloomberg 2023

Synthetia

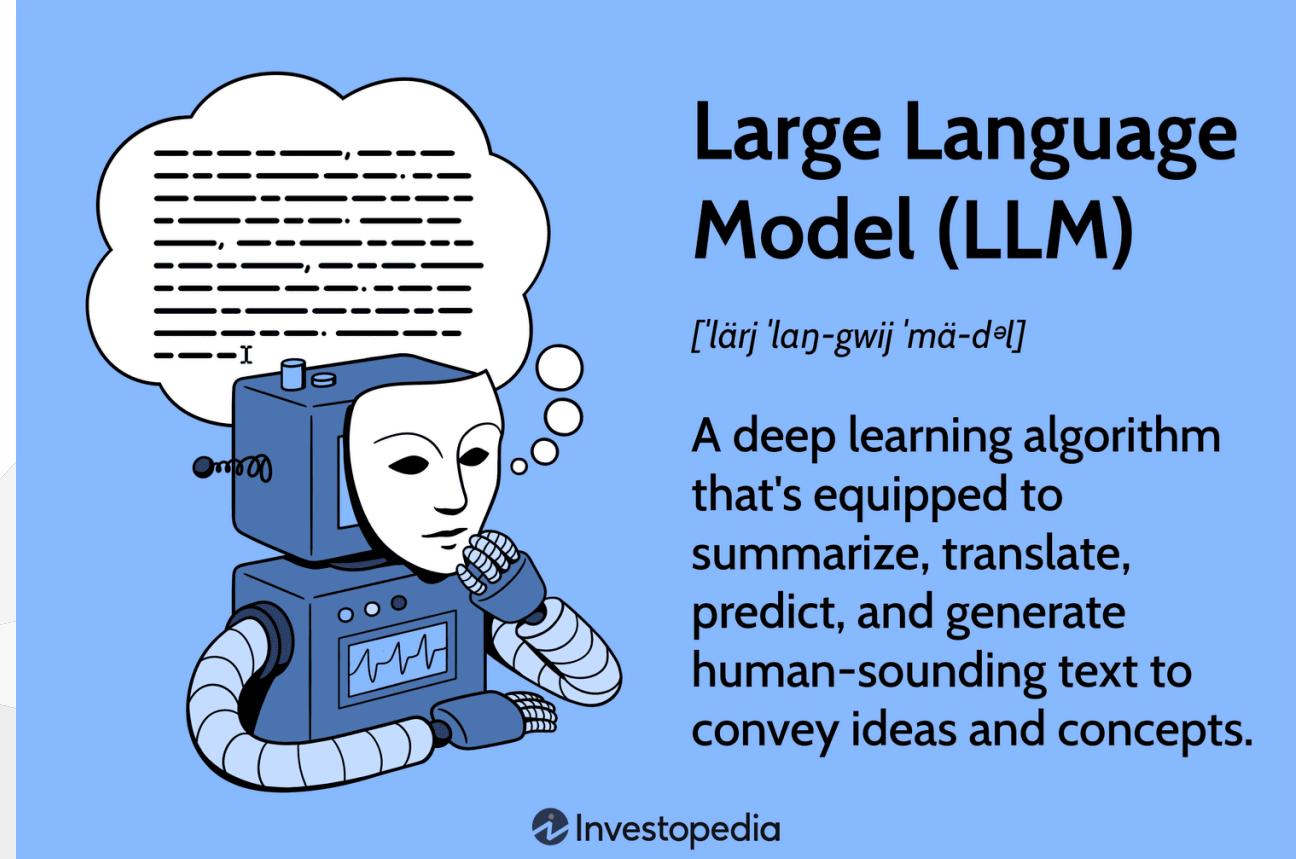
## 02 大語言模型

UNDERSTANDING THEIR IMPACT

Exploring Large  
Language Models

# 什麼是大型語言模型 (LLM)

- 一種深度學習模型
- 能從大量的文章、書籍中學習單詞和句子之間的關係，進而摘要、翻譯、生成文本



**Large Language Model (LLM)**

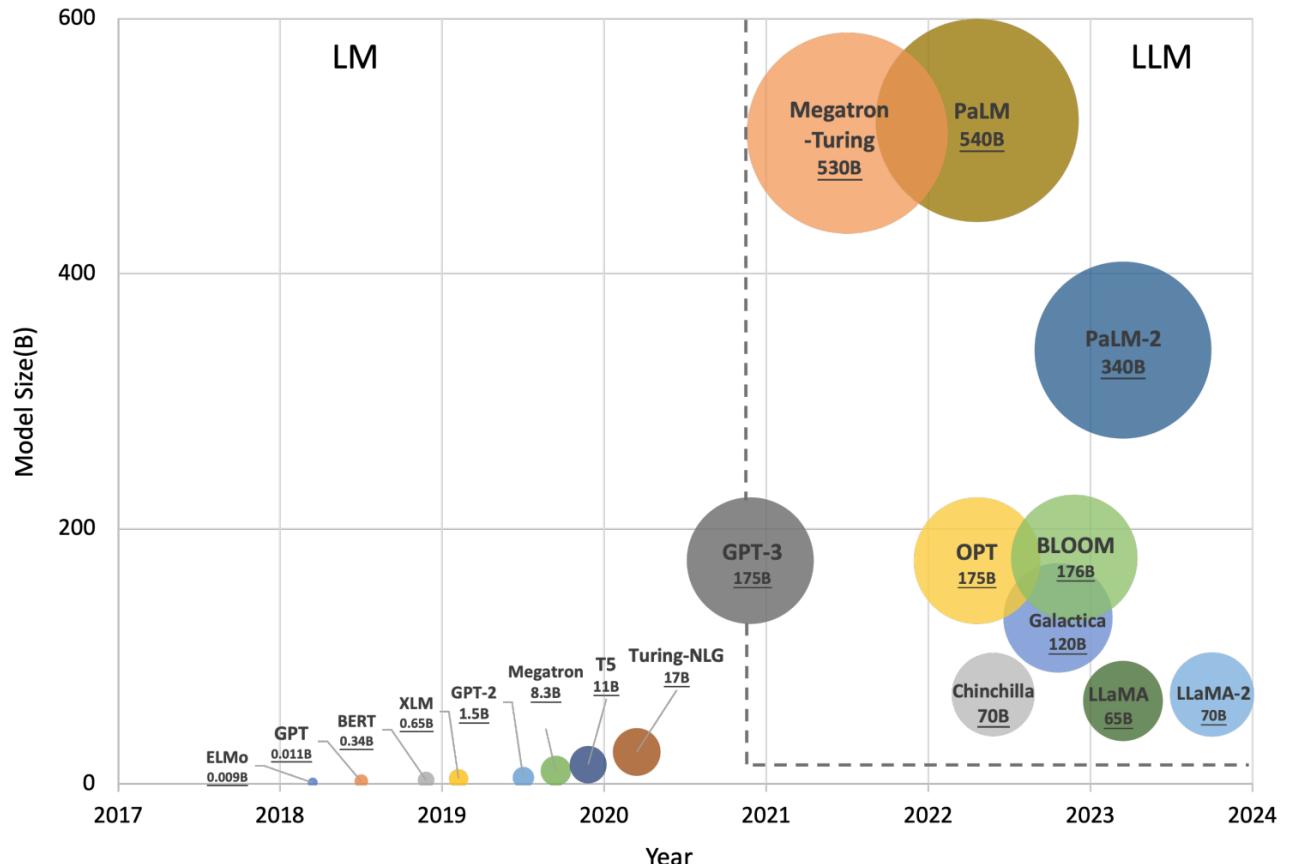
[ˈlärj ˈlaŋ-gwij ˈmä-dəl]

A deep learning algorithm that's equipped to summarize, translate, predict, and generate human-sounding text to convey ideas and concepts.

 Investopedia

# 大型語言模型之所以大

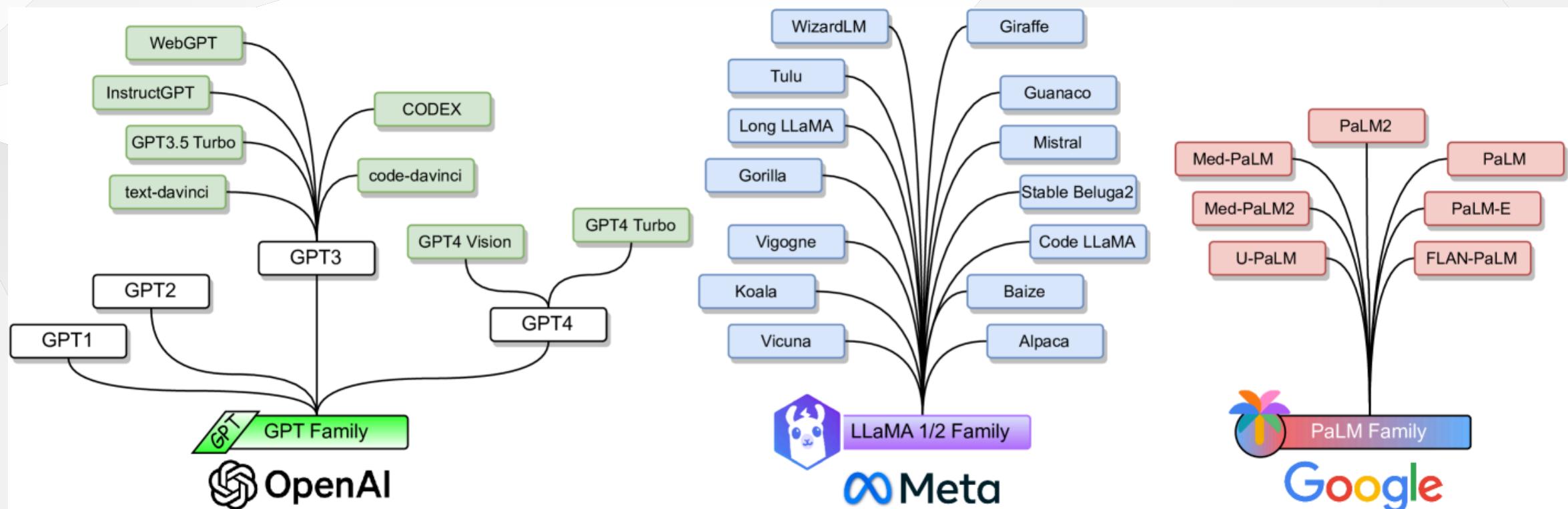
- “
- 參數量動  
輒 10 億
  - $y = 0.1x + 5$   
只有一個  
參數
- ”



# 2024 年最熱門的 LLM 對話機器人

	 ChatGPT	 Google Bard	 Claude AI	 Bing Chat	 OORT AI
<b>Release date</b>	Nov. 2022	Mar. 2023	Mar. 2023	Mar. 2023	Sept. 2023
<b>Key feature</b>	Creativity and human-like conversation	Google ecosystem assistant	Safety and ethics response	Web-based contextual response	Privacy and customization
<b>Language model</b>	GPT-3.5/Turbo	Gemini (successor to LaMDA and PaLM)	Claude	GPT-4	Enhanced vicuna
<b>Information access</b>	Offline knowledge data	Online internet data	Offline knowledge data	Online internet data	Offline knowledge data
<b>Customer data storage</b>	Centralized	Centralized	Centralized	Centralized	Decentralized
<b>Integration API</b>	Available	Not available	Available	Not available	Available

# 三大 LLM 家族



## 開源與閉源的 LLM

# CLOSED SOURCE



OpenAI

Bloomberg  
GPT

ANTHROPIC



- 1 Better performance (today)
- 2 Easier to run out of the box
- 3 Accessible to broader audiences despite lack of technical skills



# OPEN SOURCE



Stable Diffusion



ADEPT



EleutherAI

- 1 More customizable
- 2 Cheaper to train and deploy
- 3 The community (and developer) maintains control

# 知名閉源 LLM

- GPT-4o: GPT 系列最新模型(2024-05-13 發布)
  1. **跨媒體推理能力**：理解和分析文字、影像、語音
  2. **實時語音對話**：通過自然的語音方式對話、識別使用者聲音中的情緒
  3. **視覺能力**：能夠使用文本和“視覺”進行互動，這意味著它可以查看用戶上傳的螢幕截圖、照片、文檔或圖表，並進行對話
  4. **效能提升**：運行速度大幅提升，比前身 GPT-4 快兩倍
  5. **價格優勢**：價格僅為前身 GPT-4 的一半

- 探索  
GPT-4o  
模型的能  
力
- 官網

## Explorations of capabilities

Select sample:

1  
Input

A first person  
typewritten  
entries:

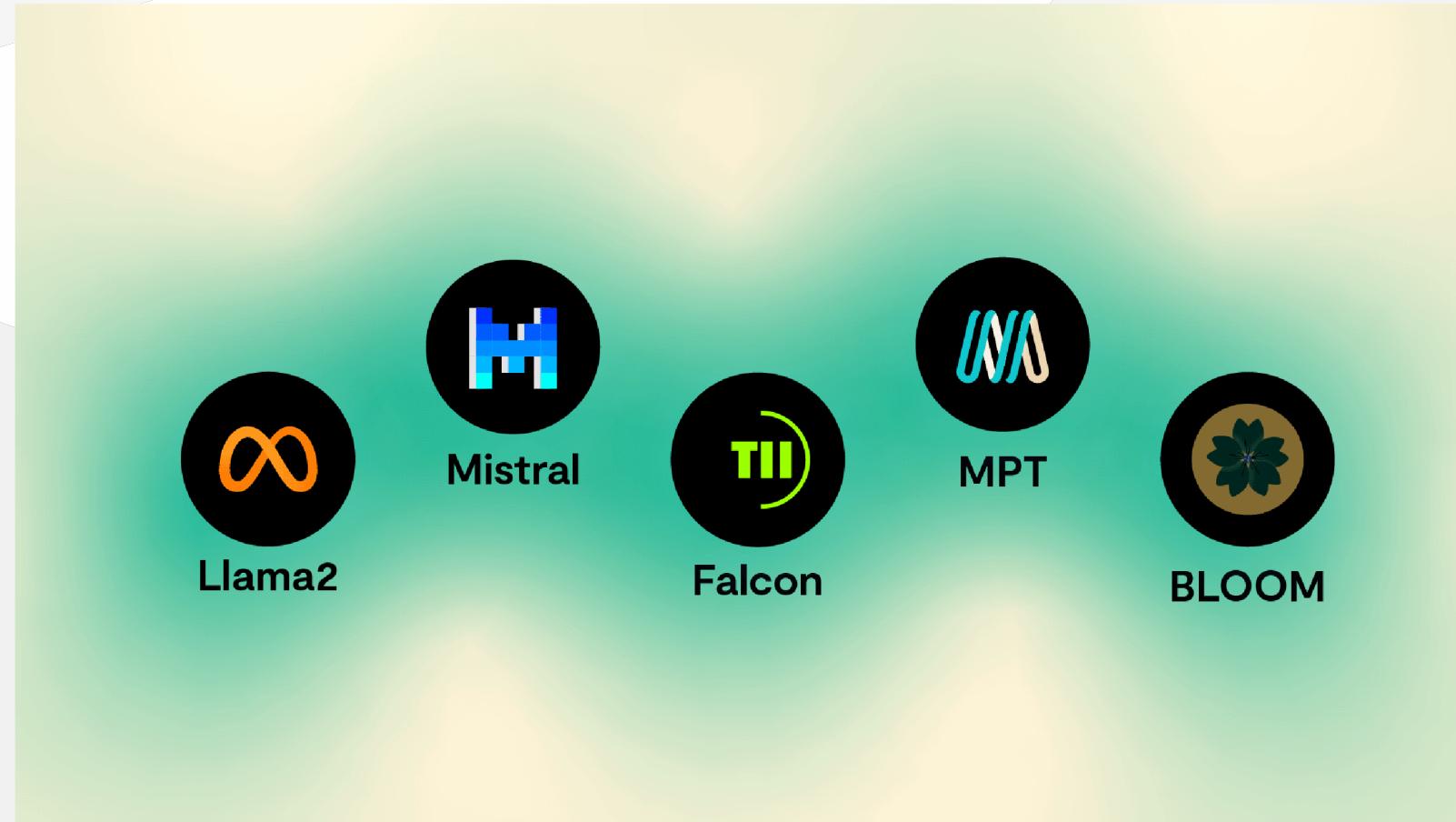
1. yo, so  
caught the  
insane, co  
makes you  
is reality'

2. sound up  
it's wild.  
now, every  
secret. ma  
am i missin

the text is large, legible and

- ✓ Visual Narratives - Robot Writer's Block
- Visual narratives - Sally the mailwoman
- Poster creation for the movie 'Detective'
- Character design - Geary the robot
- Poetic typography with iterative editing 1
- Poetic typography with iterative editing 2
- Commemorative coin design for GPT-4o
- Photo to caricature
- Text to font
- 3D object synthesis
- Brand placement - logo on coaster
- Poetic typography
- Multiline rendering - robot texting
- Meeting notes with multiple speakers
- Lecture summarization
- Variable binding - cube stacking
- Concrete poetry

# 知名開源 LLM



# 針對繁體中文優化的開源 LLM

- 聯發科：MR Breeze-7B



聯發科技

- 國科會：Llama 3-TAIDE-LX-8B



# 聯發科 Breeze

- 聯發科在2024-03-07日釋出開源的模型 **MR Breeze-7B**，是以**Mistral-7B**為基礎進行訓練
- 以模型參數量小為特色，能在短的時間內提供更流暢、更精準的回應，在繁體中文與英文提供出色的表達能力
- 在處理表格數據方面的性能為 7B 中英雙語模型裡最優，能夠更加精確地解讀和生成表格內容

相關報導

# 國科會 TAIDE

- TAIDE (Trustworthy AI Dialogue Engine) 是由台灣國家科學技術委員會 (NSTC) 於 2023 年 4 月啟動的專案，旨在創建一個專為台灣設計的繁體中文生成式 AI 對話引擎
- TAIDE 團隊於 2024-04-29 度釋出 Llama 3-TAIDE-LX-8B-Chat-Alpha1 模型，這是以 **Meta Llama-3-8B** 為基礎進行訓練，是最新具臺灣文化的大型繁體中文模型
- 可商用，已有 10 餘家廠商及多個學研團隊及公部門開始導入 TAIDE 模型開發相關應用系統

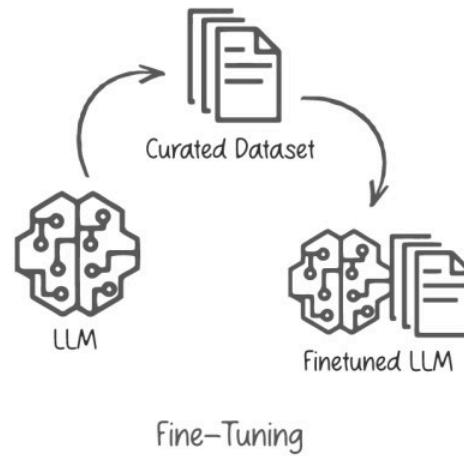
# TAIDE 模型的應用

- Kuwa GenAI OS：這是本土開發且開放源碼的新一代生成式 AI 應用平台。高雄大學運用預載的 TAIDE 模型，能輕鬆在地端使用 TAIDE 生成式 AI 服務或開發創新應用。
- aiDAPTIIV+平台：群聯獨家專利研發的平台已與 TAIDE 模型驗證整合；此外，此專案攜手十家以上的國際電腦大廠聯手，推出平民化的生成式 AI 地端運算平台，大幅加速台灣及全球的生成式 AI 應用普及。
- 叢揚資訊查詢知識庫：叢揚資訊在不同產品間整合 TAIDE 模型，讓使用者以直白語句查詢知識庫。

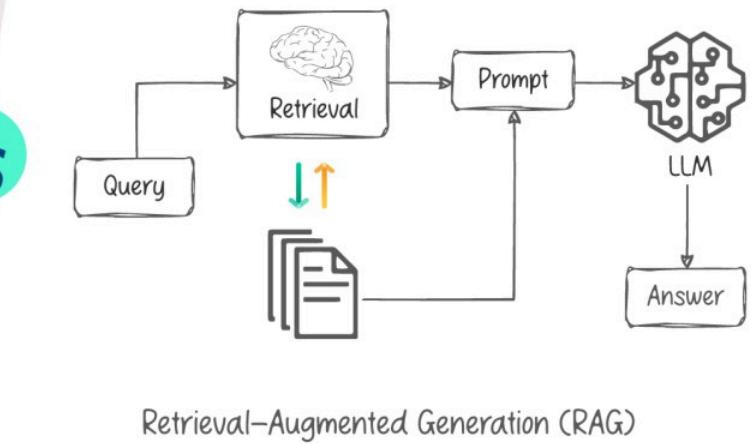
## 兩種企業級 LLM 應用

- 檢索增強生成 (RAG)
- 微調(Fine-tuning)

### Power of Your Enterprise LLM: Fine-Tuning vs Retrieval Augmentation



VS

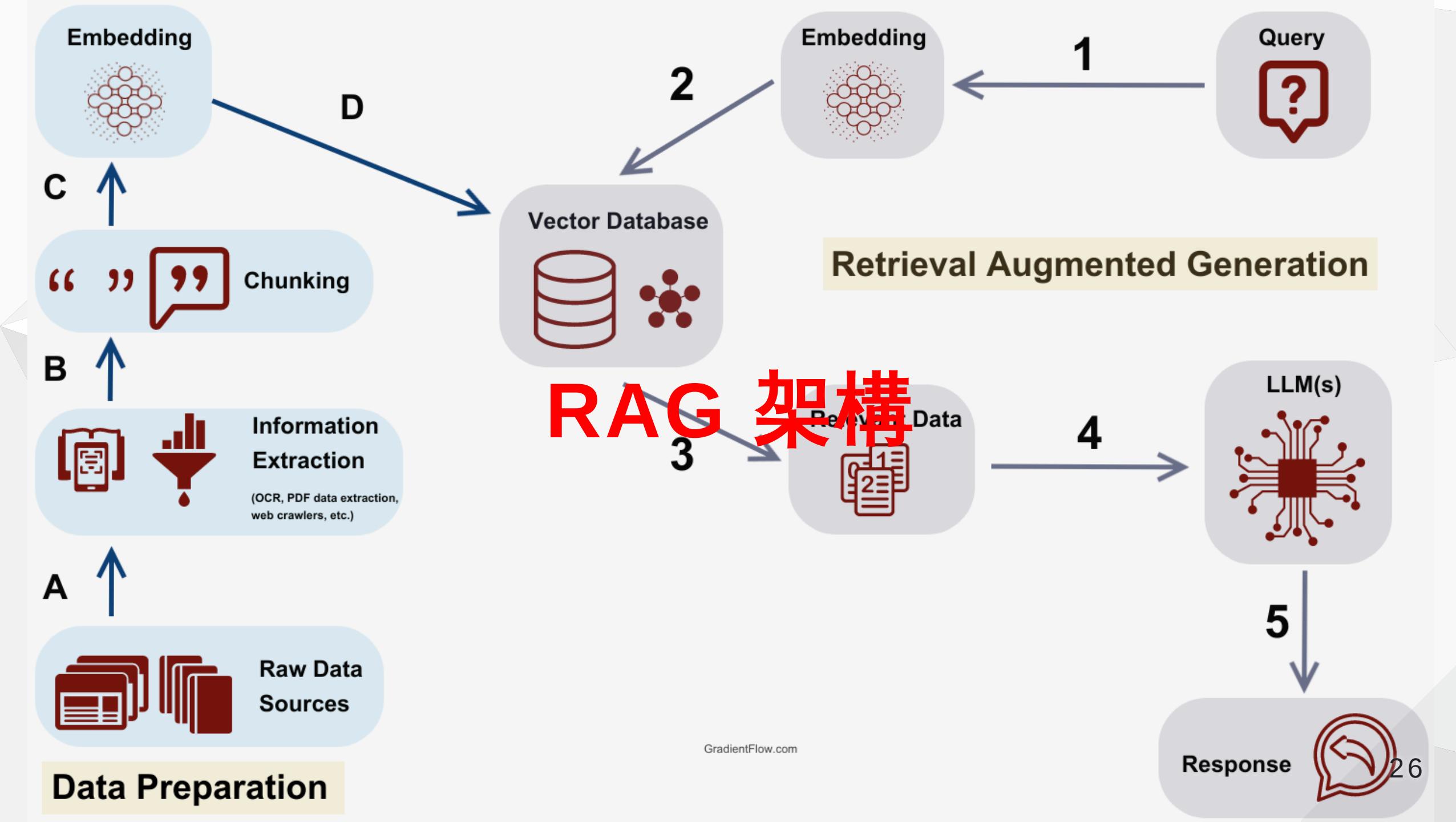


# 什麼是 RAG

RAG (Retrieval-Augmented

Generation，檢索增強生成) 是一種新的技術，它允許大型語言模型 (LLM) 從外部資料源獲取並整合資訊



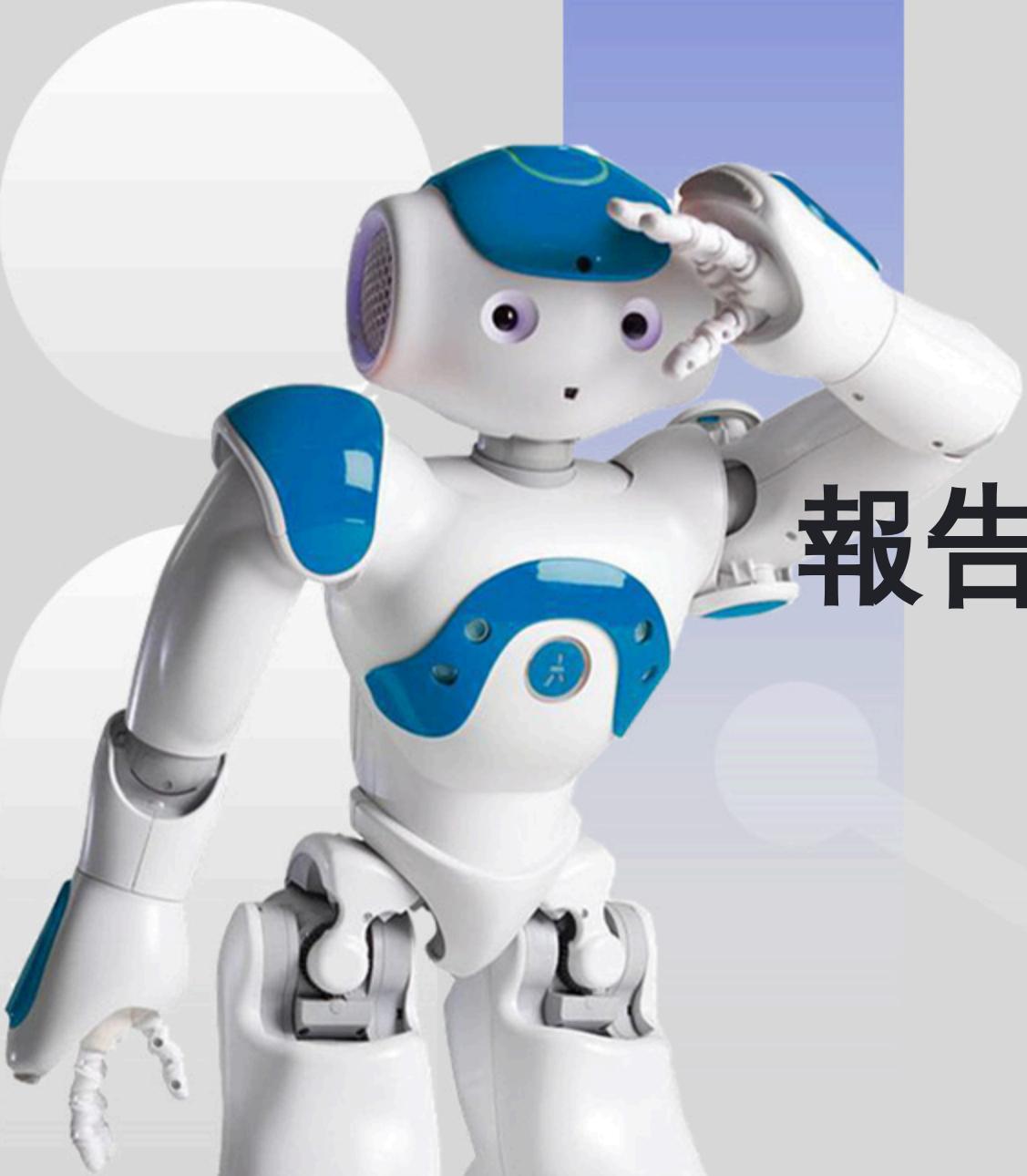


# RAG 比喻

- LLM: 法官
- RAG: 書記官

“書記官從法律書籍  
(即外部資料源) 中  
查找相關的法條，幫  
助法官做出更精準的  
判決。”





報告完畢

ARTIFICIAL INTELLIGENCE

RISE OF ROBOTS

THANKS