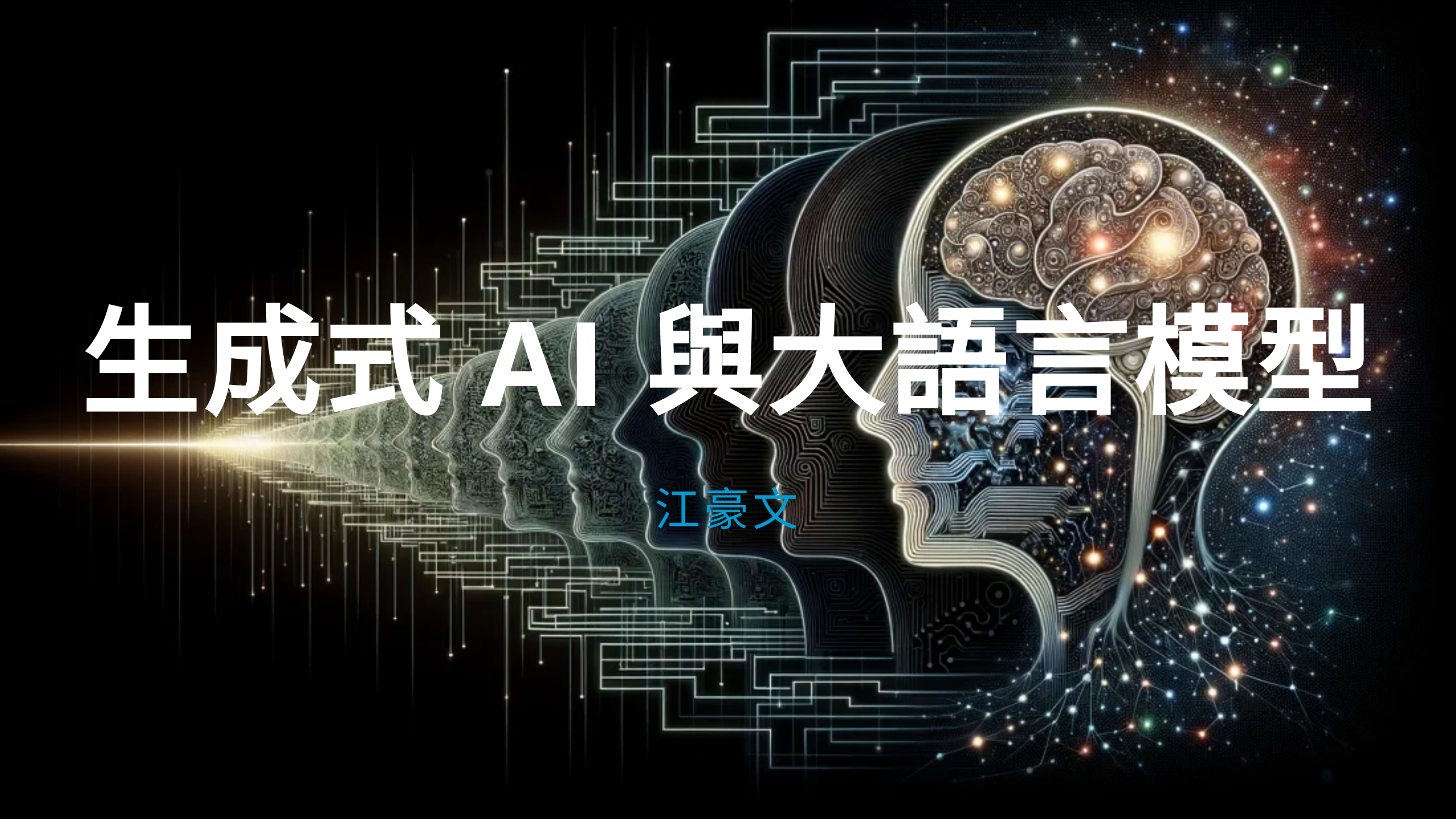
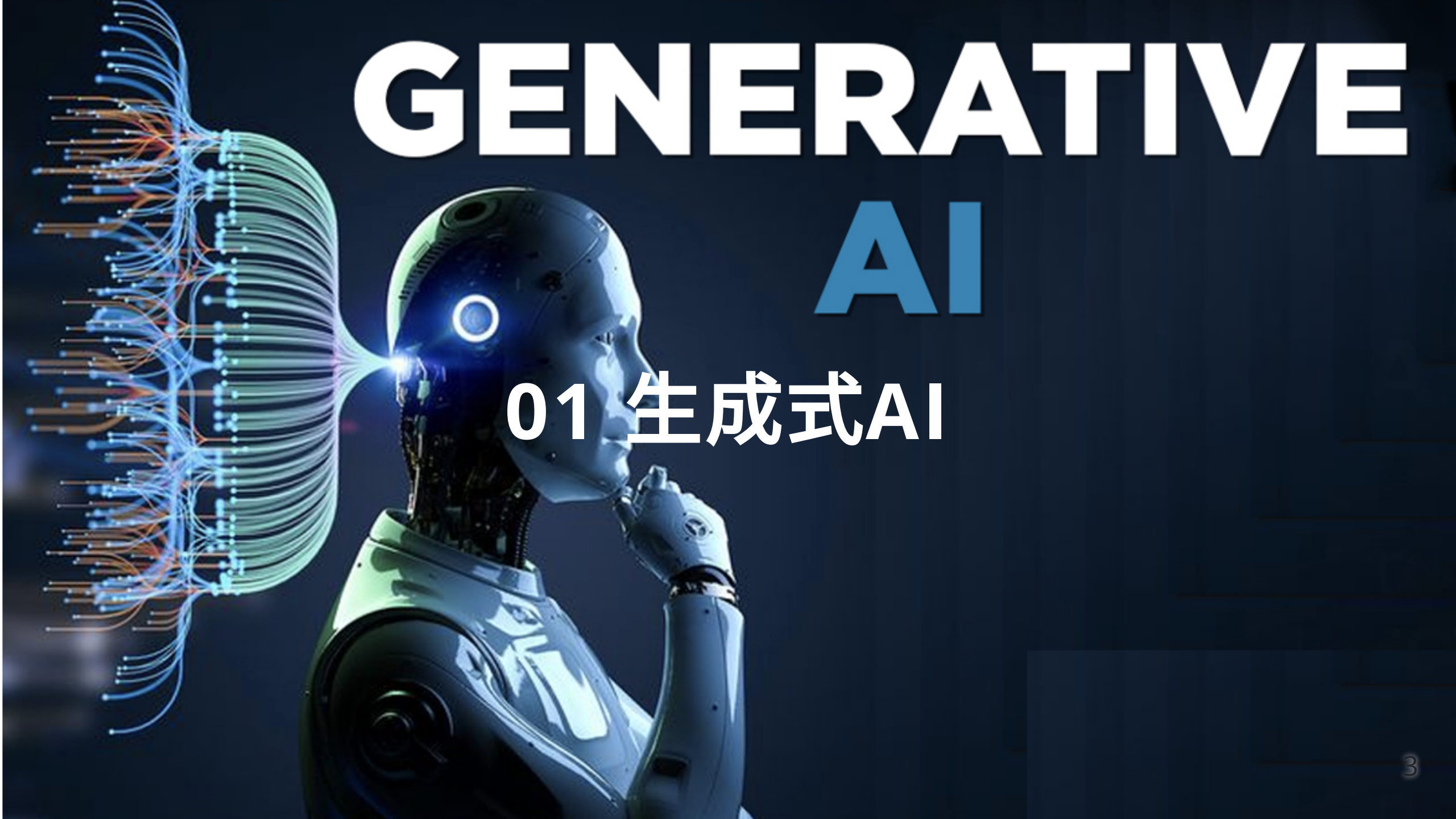


生成式 AI 與大語言模型

江豪文

大綱

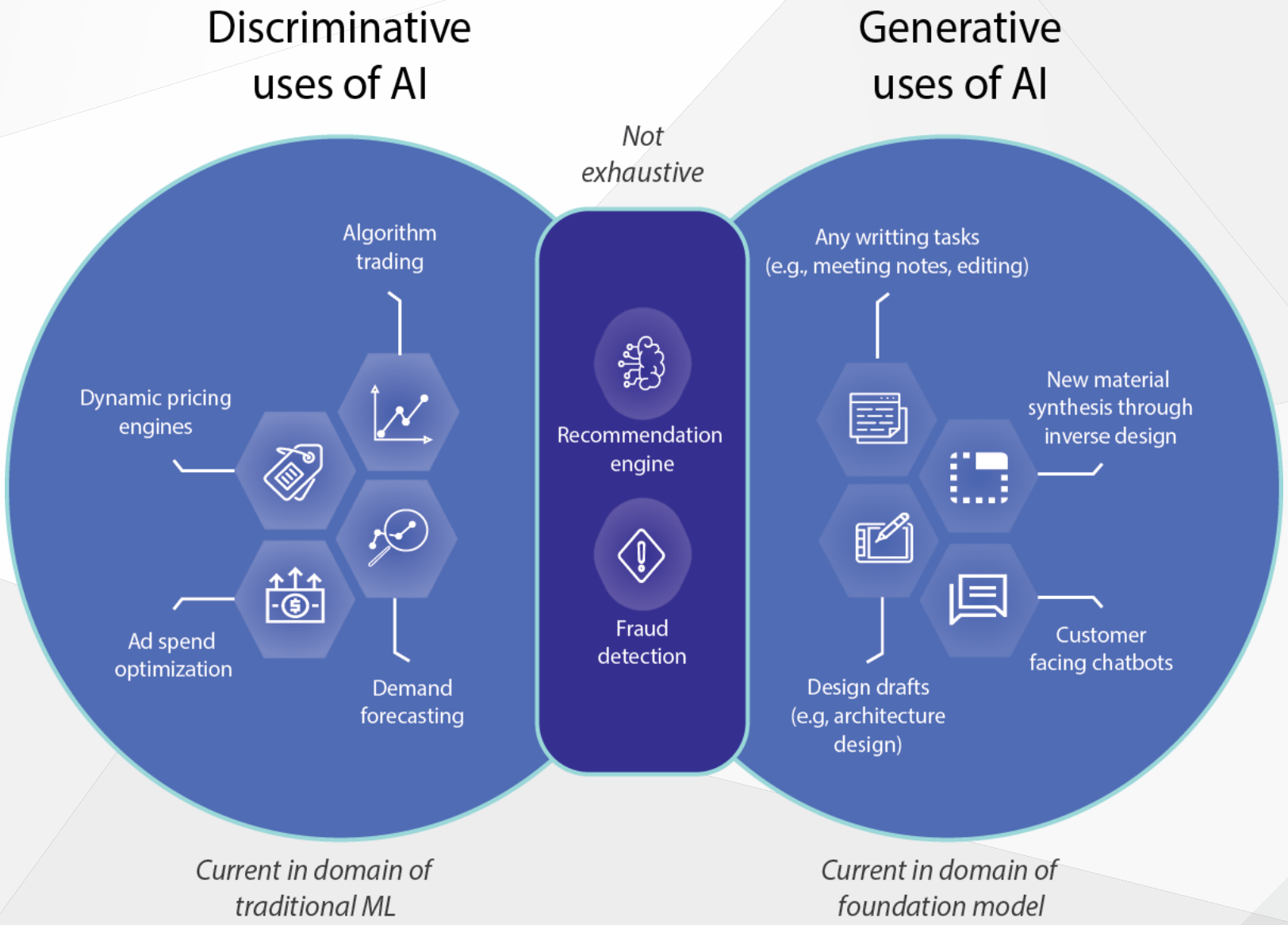
- 生成式 AI
- 大語言模型



01 生成式AI

兩大 AI 類型

- 分辨式
- 生成式



分辨式 AI: 預測數值與類別

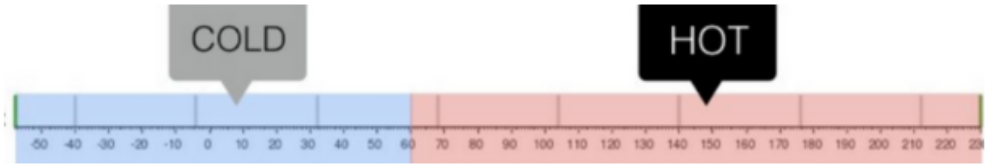
Regression

What will be the temperature tomorrow?

Fahrenheit

Classification

Will it be hot or cold tomorrow?



Fahrenheit

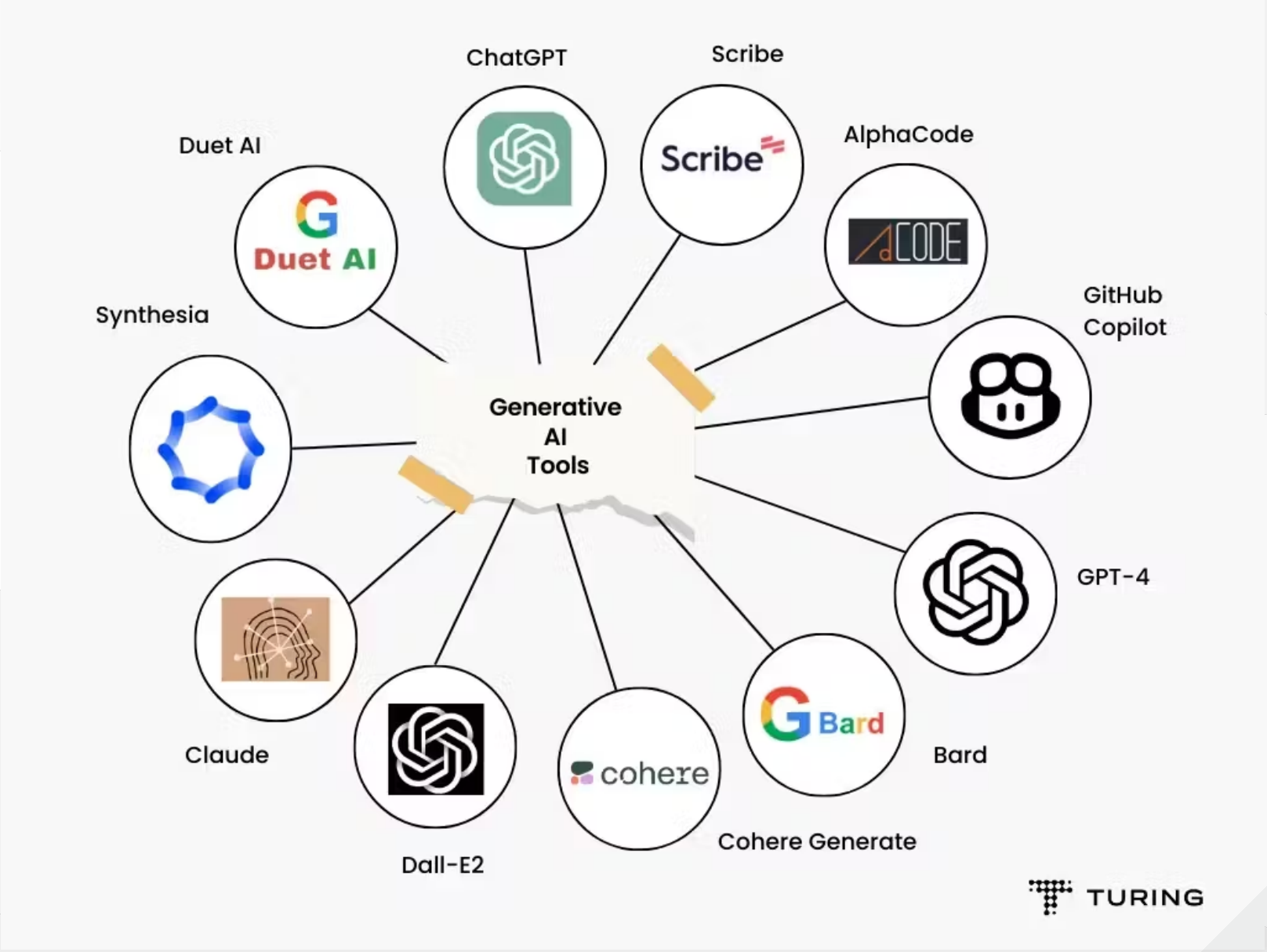
生成式 AI 範例

Bill Gates:

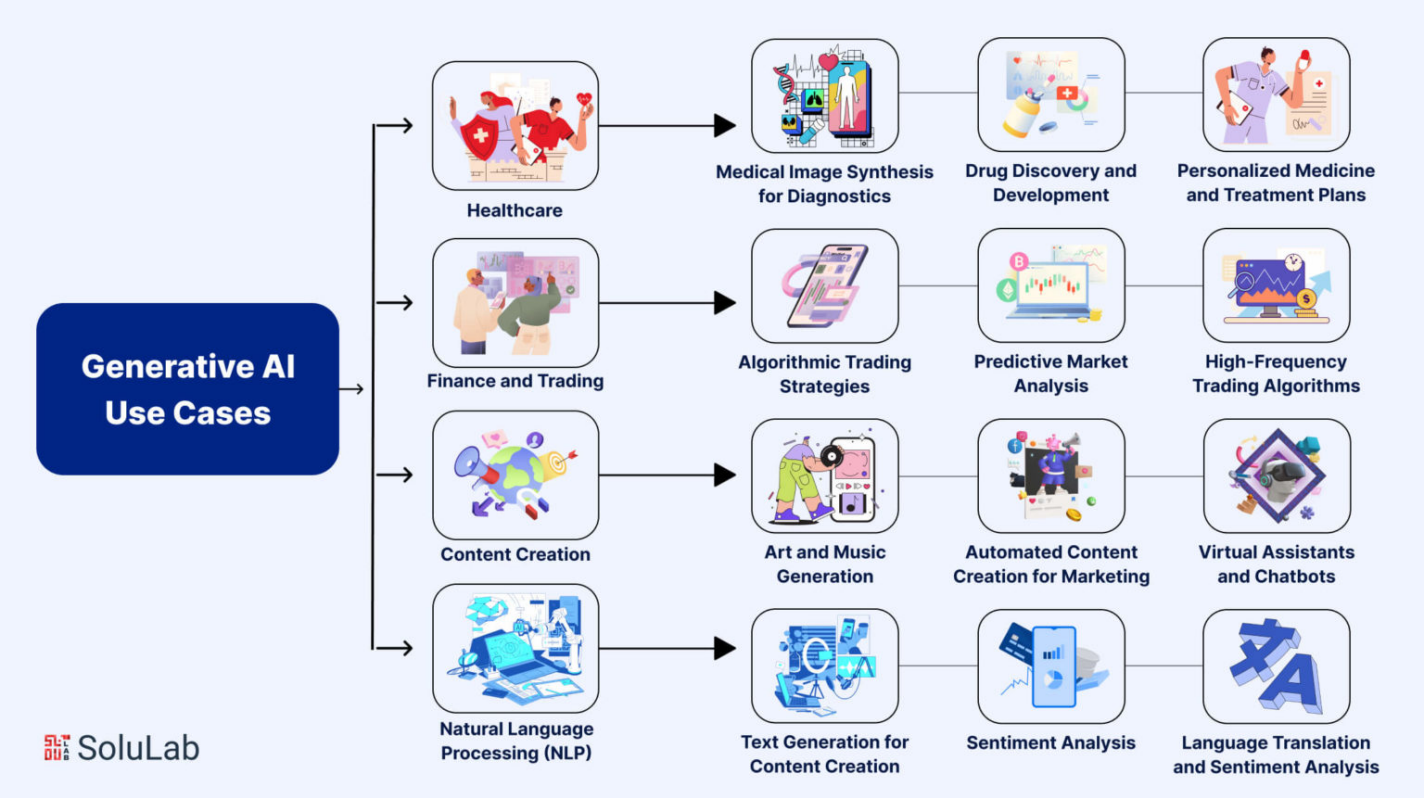
ChatGPT’s history is as significant as the birth of the PC or the Internet

~TechGoing

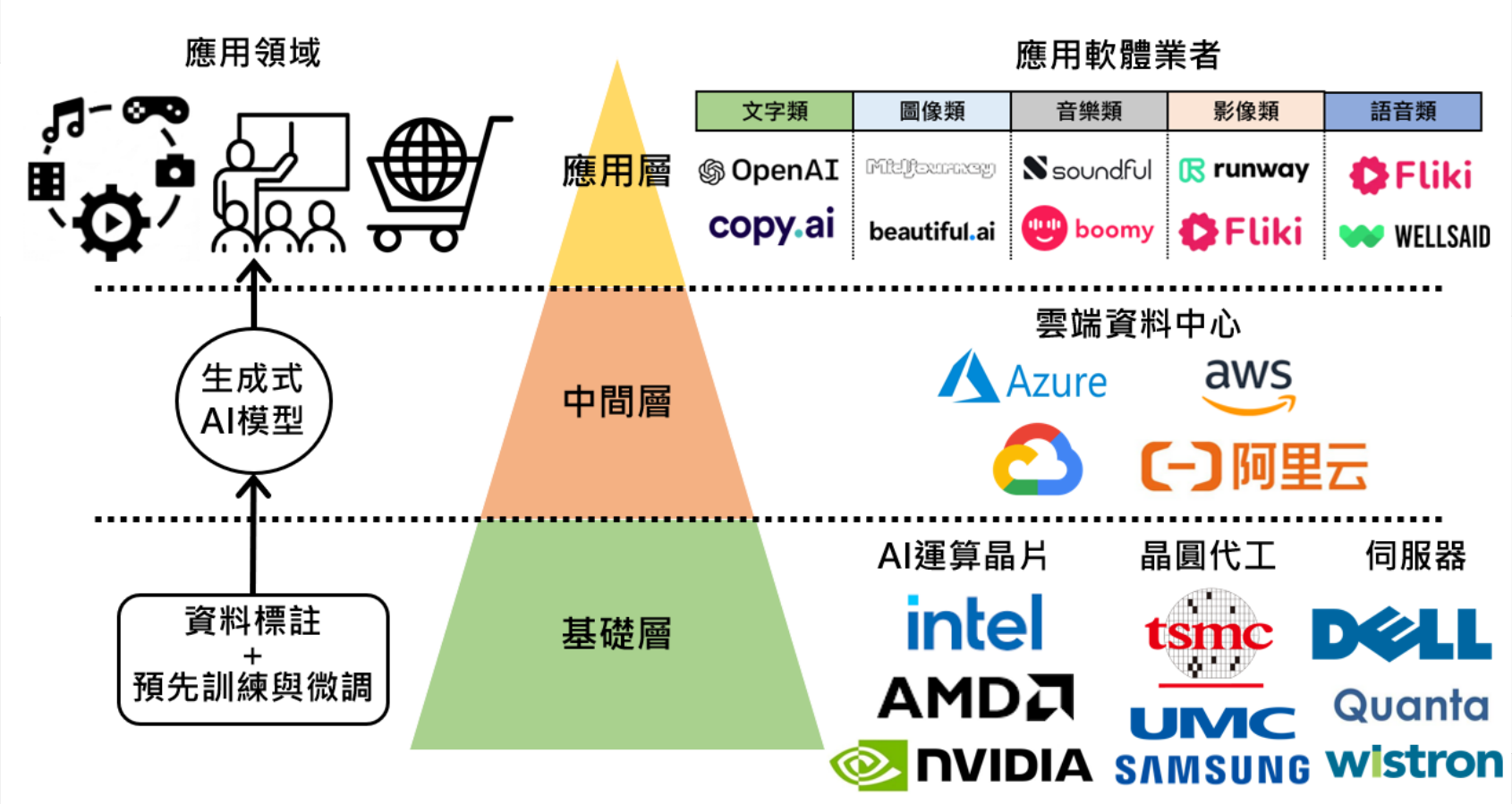
生成式 AI 工具


ChatGPT
Scribe
Duet AI
AlphaCode
Synthesia
GitHub Copilot
Generative AI Tools
GPT-4
Claude
Bard
Dall-E2
Cohere Generate
TURING

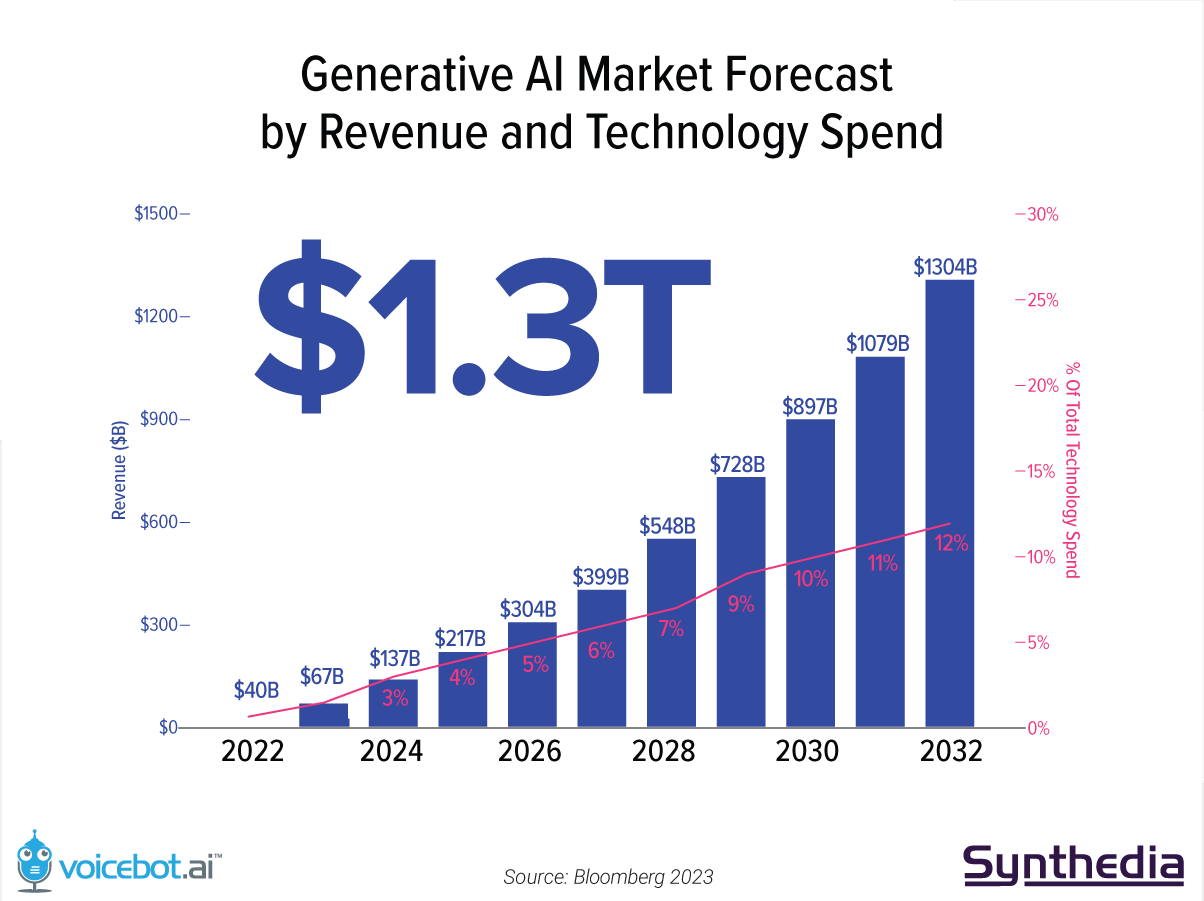
生成式 AI 的應用情境


Generative AI Use Cases
Healthcare
Medical Image Synthesis for Diagnostics
Drug Discovery and Development
Personalized Medicine and Treatment Plans
Finance and Trading
Algorithmic Trading Strategies
Predictive Market Analysis
High-Frequency Trading Algorithms
Content Creation
Art and Music Generation
Automated Content Creation for Marketing
Virtual Assistants and Chatbots
Natural Language Processing (NLP)
Text Generation for Content Creation
Sentiment Analysis
Language Translation and Sentiment Analysis
SoluLab

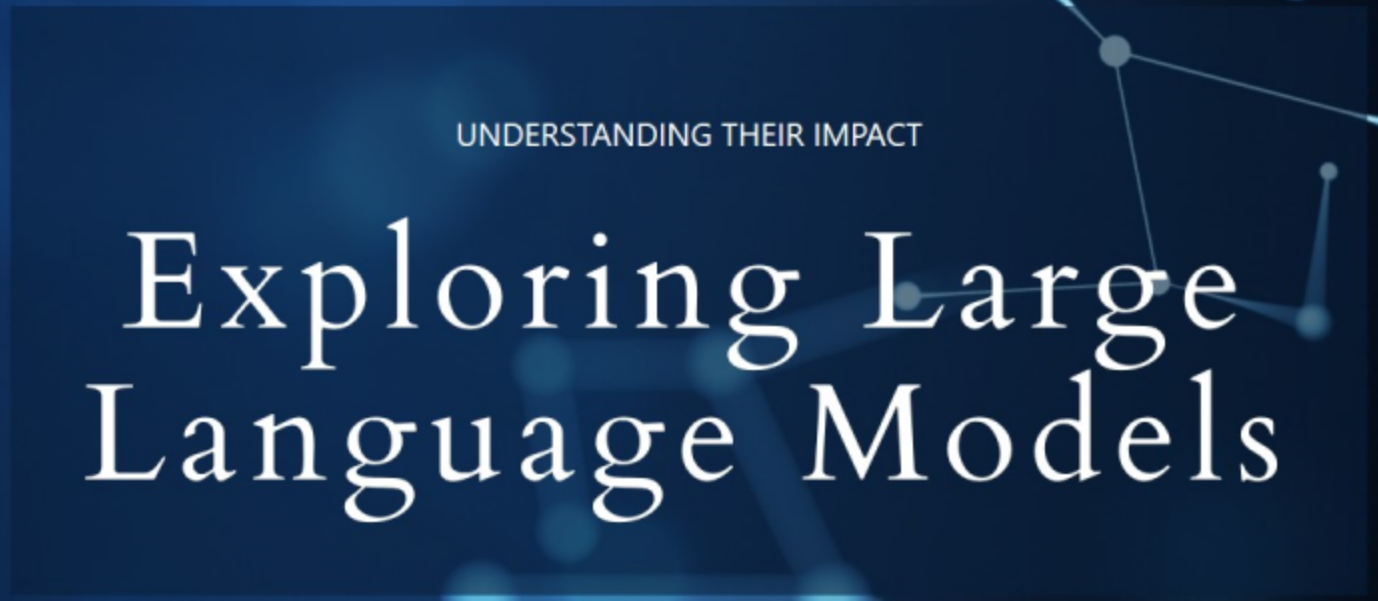
生成式 AI 生態鏈


應用領域
應用軟體業者
文字類
圖像類
音樂類
影像類
語音類
OpenAI
Midjourney
soundful
runway
Fliki
copy.ai
beautiful.ai
boomy
Fliki
WELLSAID
應用層
中間層
基礎層
生成式
AI模型
資料標註
+
預先訓練與微調
雲端資料中心
Azure
aws
阿里云
AI運算晶片
晶圓代工
伺服器
intel
tsmc
DELL
AMD
UMC
Quanta
NVIDIA
SAMSUNG
wistron

生成式 AI 的市值預估

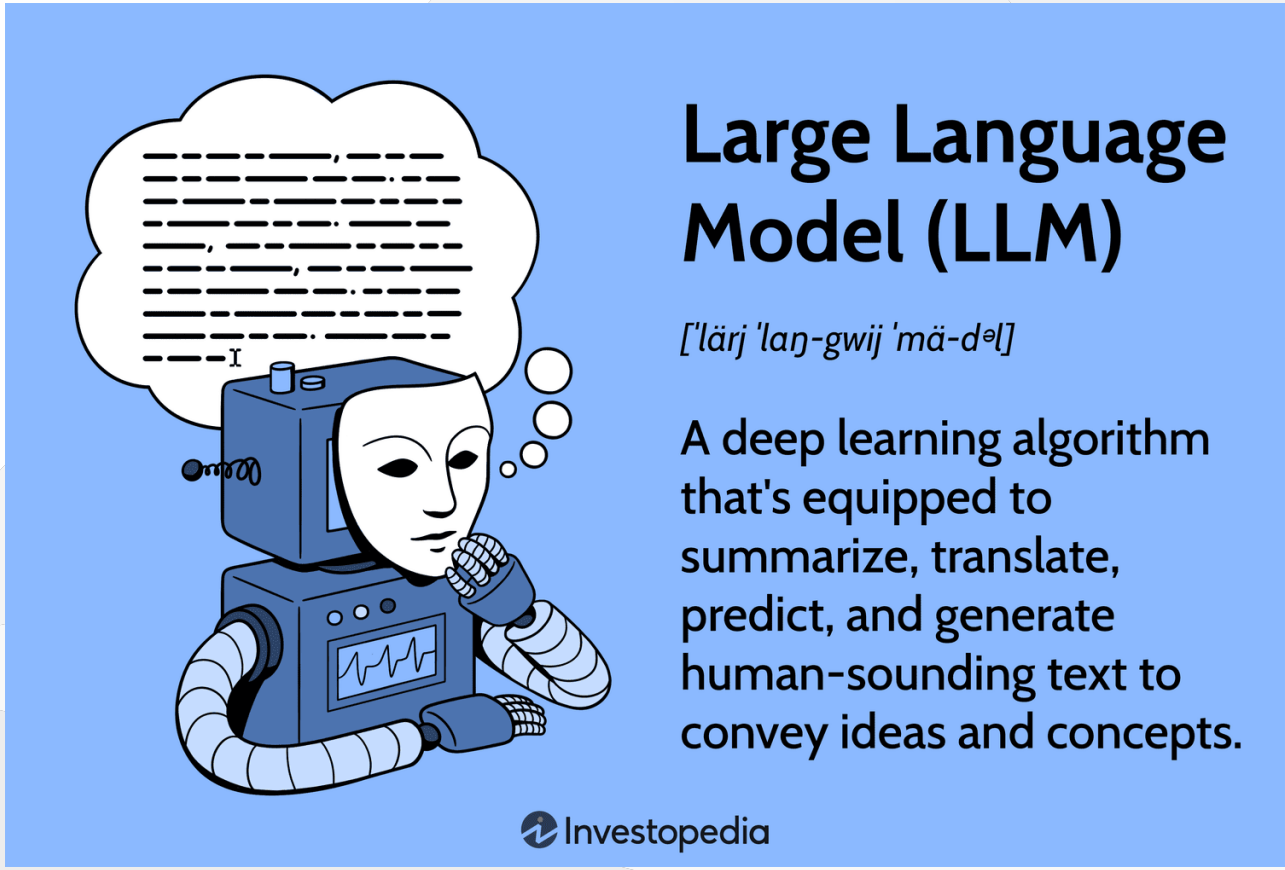

Generative AI Market Forecast
by Revenue and Technology Spend
$1.3T
Revenue ($B)
$1500
$1200
$900
$600
$300
$0
30%
25%
20%
15%
10%
5%
0%
% Of Total Technology Spend
$40B
$67B
$137B
$217B
$304B
$399B
$548B
$728B
$897B
$1079B
$1304B
3%
4%
5%
6%
7%
9%
10%
11%
12%
2022
2024
2026
2028
2030
2032
voicebot.ai
Source: Bloomberg 2023
Synthedia

02 大語言模型


UNDERSTANDING THEIR IMPACT
Exploring Large Language Models

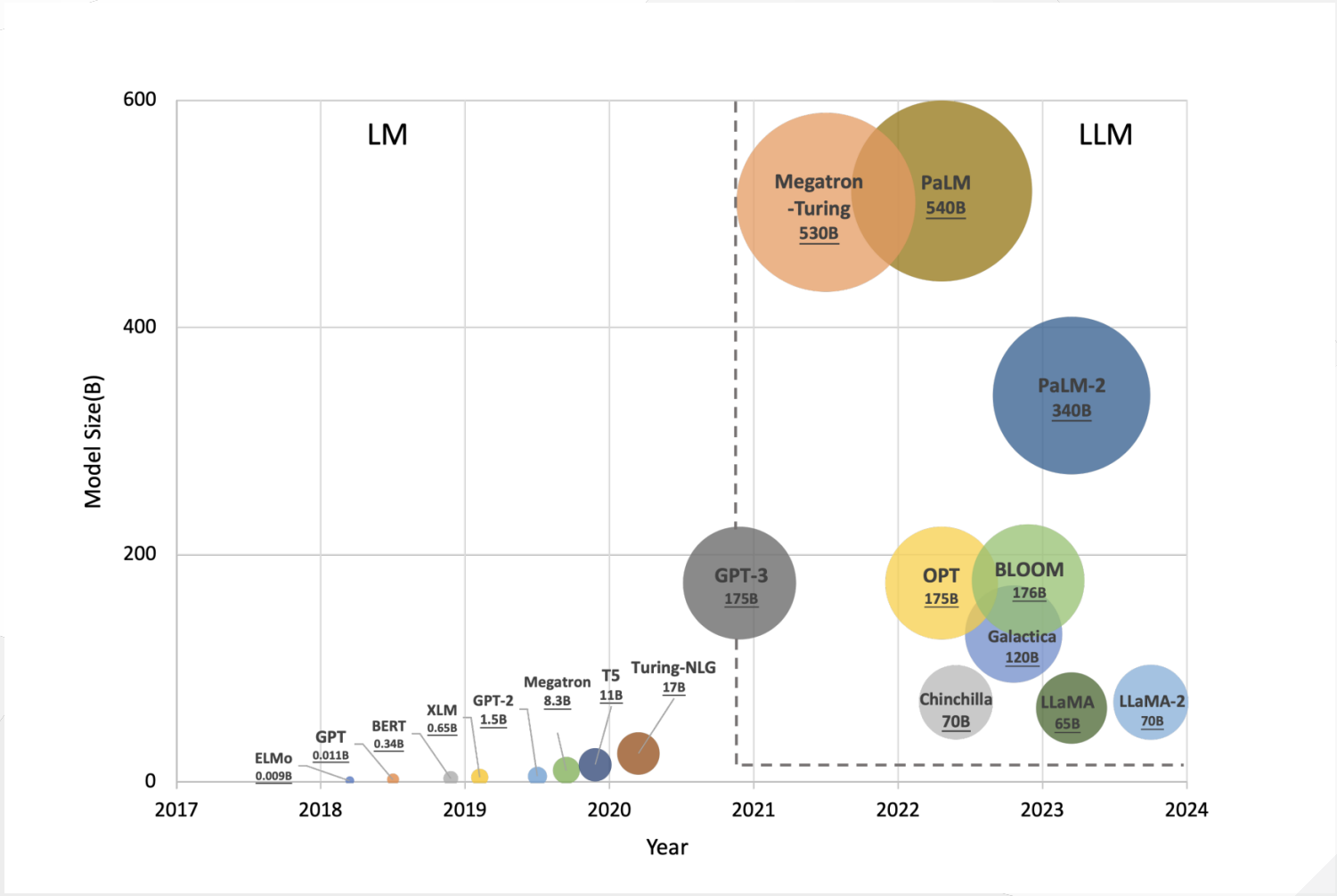
什麼是大型語言模型 (LLM)

- 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



大型語言模型之所以大

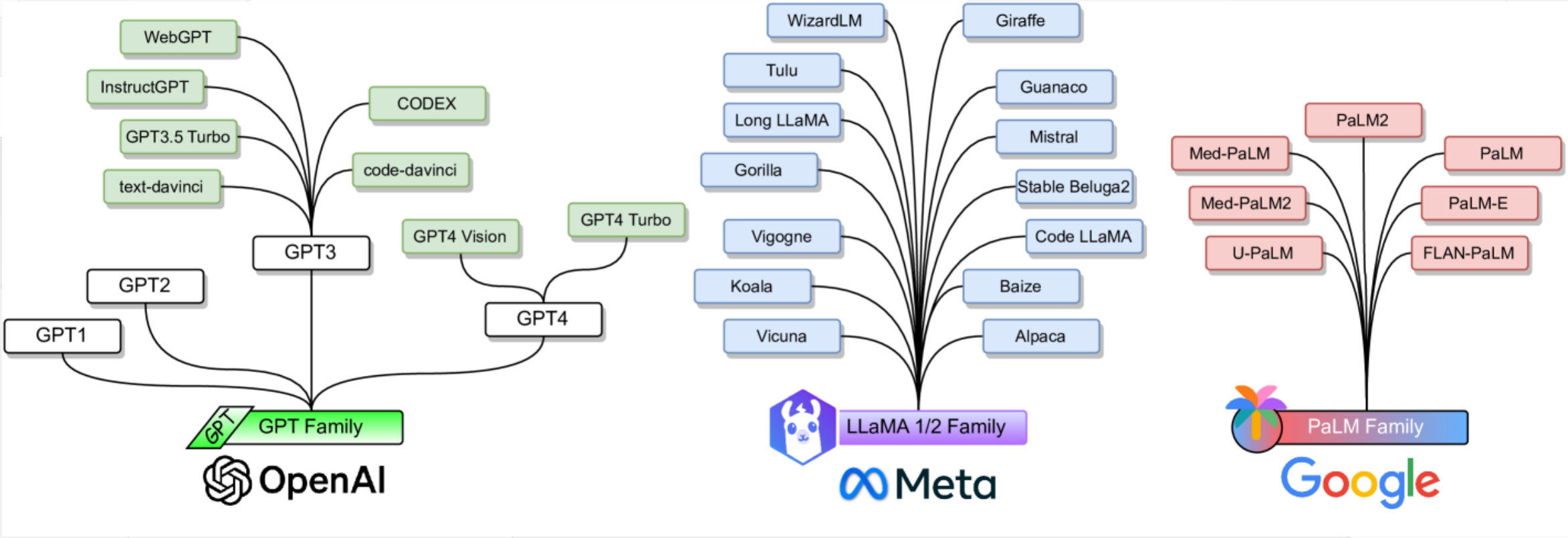
- 參數量動輒 10 億
- $y = 0.1x + 5$ 只有一個參數



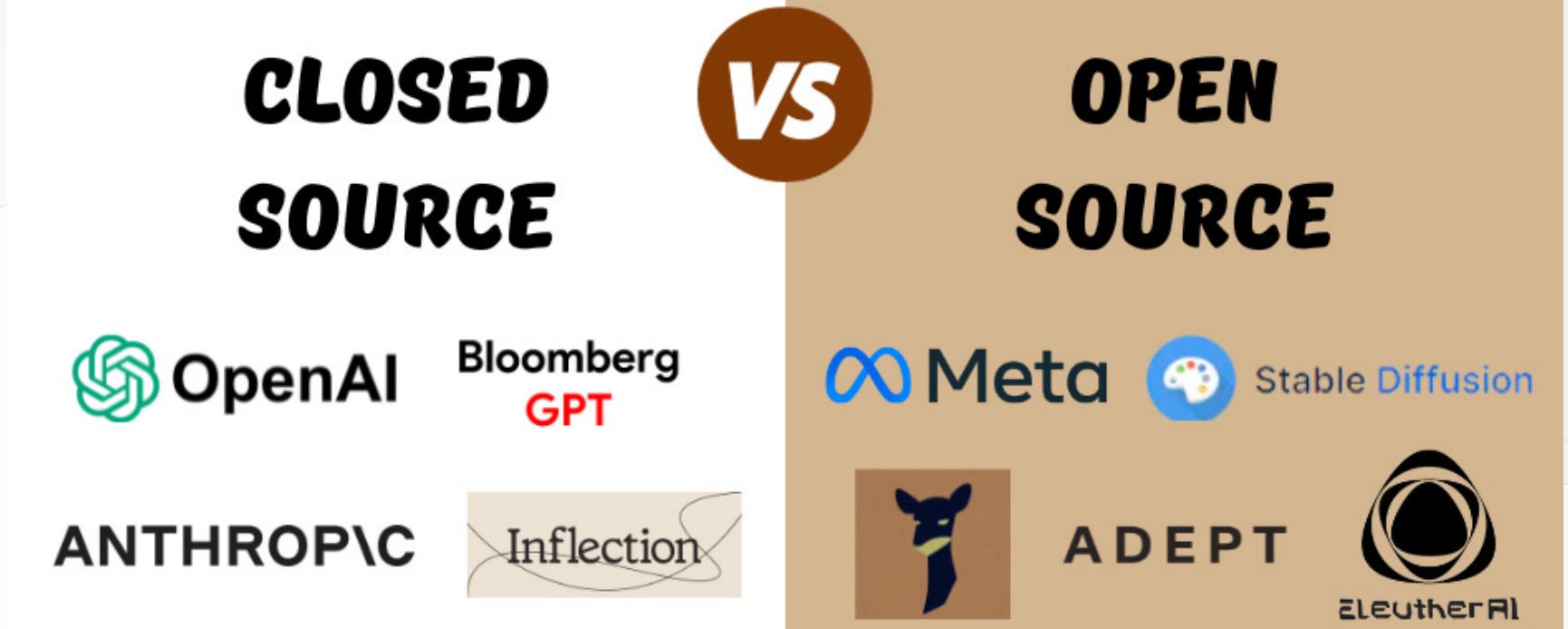
2024 年最熱門的 LLM 對話機器人

	ChatGPT	Google Bard	Claude AI	Bing Chat	OORT AI
Release date	Nov. 2022	Mar. 2023	Mar. 2023	Mar. 2023	Sept. 2023
Key feature	Creativity and human-like conversation	Google ecosystem assistant	Safety and ethics response	Web-based contextual response	Privacy and customization
Language model	GPT-3.4/GPT-4	Gemini (successor to LaMDA and PaLM)	Claude	GPT-4	Enhanced vicuna
Information access	Offline knowledge data	Online internet data	Offline knowledge data	Online internet data	Offline knowledge data
Customer data storage	Centralized	Centralized	Centralized	Centralized	Decentralized
Integration API	Available	Not available	Available	Not available	Available

三大 LLM 家族


WebGPT
InstructGPT
GPT3.5 Turbo
text-davinci
CODEX
code-davinci
GPT3
GPT4 Vision
GPT4 Turbo
GPT4
GPT2
GPT1
GPT
GPT Family
OpenAI
WizardLM
Tulu
Long LLaMA
Gorilla
Vigogne
Koala
Vicuna
Giraffe
Guanaco
Mistral
Stable Beluga2
Code LLaMA
Baize
Alpaca
LLaMA 1/2 Family
Meta
PaLM2
Med-PaLM
Med-PaLM2
U-PaLM
PaLM
PaLM-E
FLAN-PaLM
PaLM Family
Google

開源與閉源的 LLM

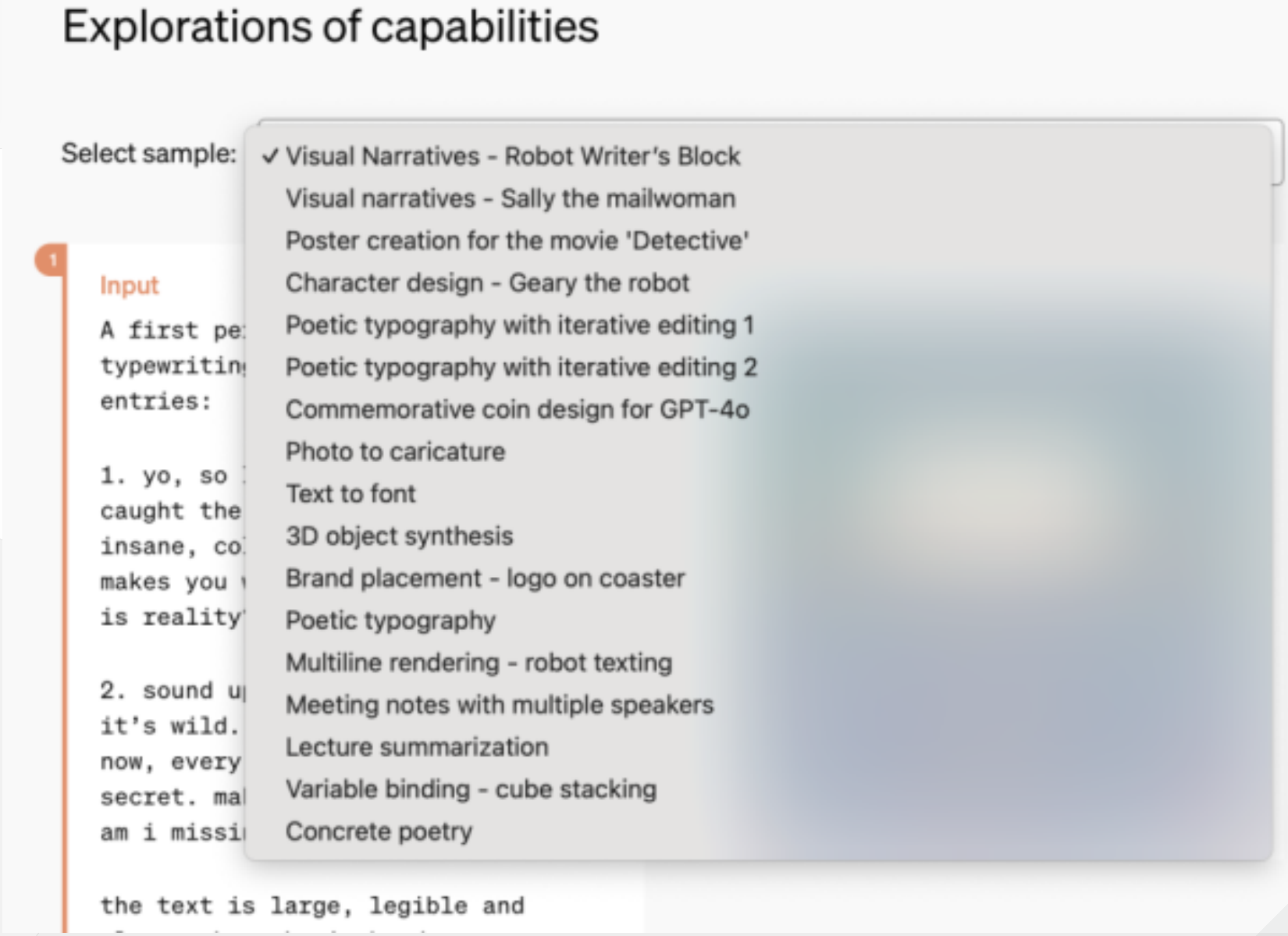

CLOSED SOURCE
VS
OPEN SOURCE
OpenAI
Bloomberg GPT
ANTHROP\C
Inflection
Meta
Stable Diffusion
ADEPT
EleutherAI
1 Better performance (today)
2 Easier to run out of the box
3 Accessible to broader audiences despite lack of technical skills
1 More customizable
2 Cheaper to train and deploy
3 The community (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



知名開源 LLM

針對繁體中文優化的開源 LLM

- 聯發科： `MR Breeze-7B`

- 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，以有 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

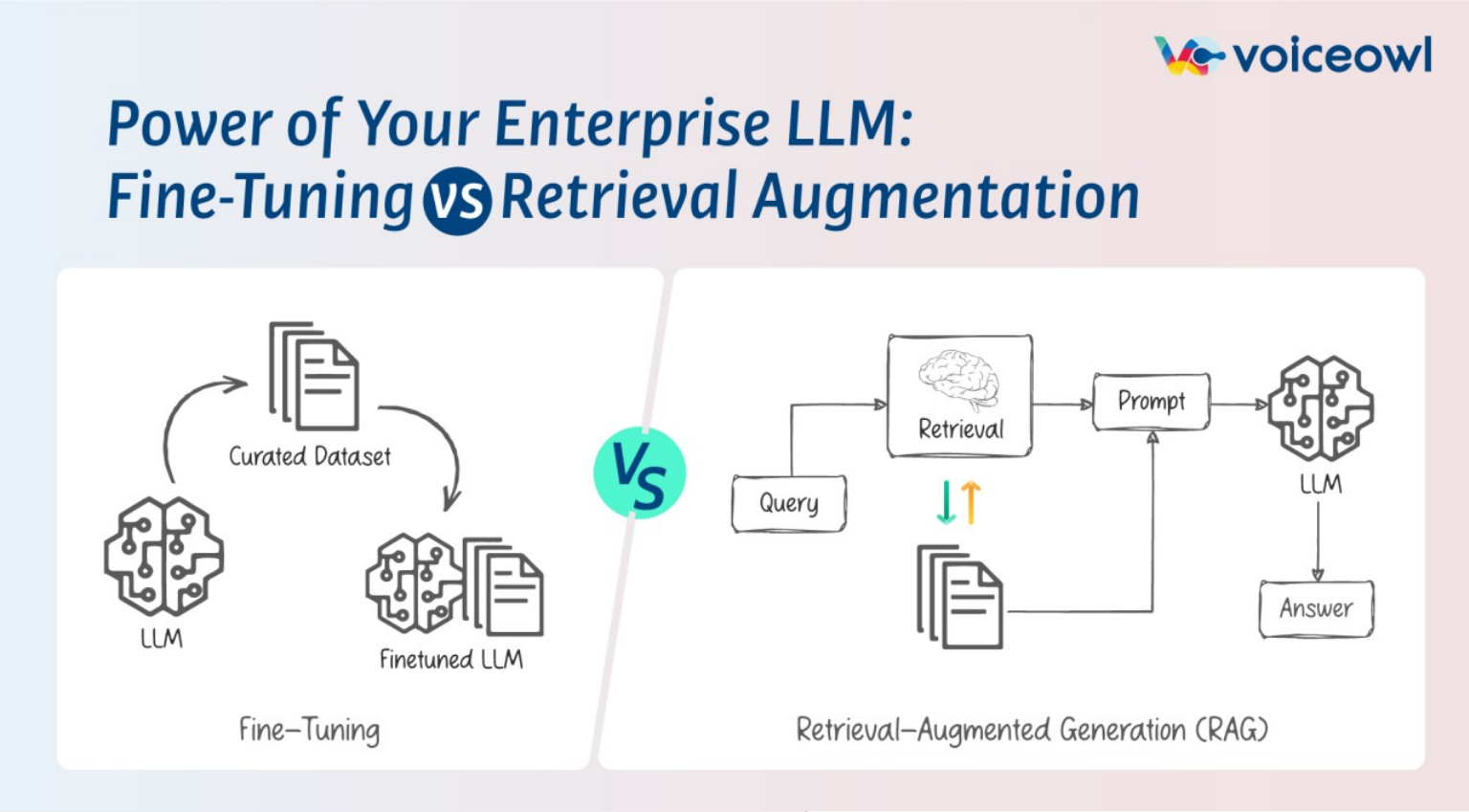
相關報導

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。
- aiDAPTIV+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。
- 叡揚資訊查詢知識庫：叡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**。

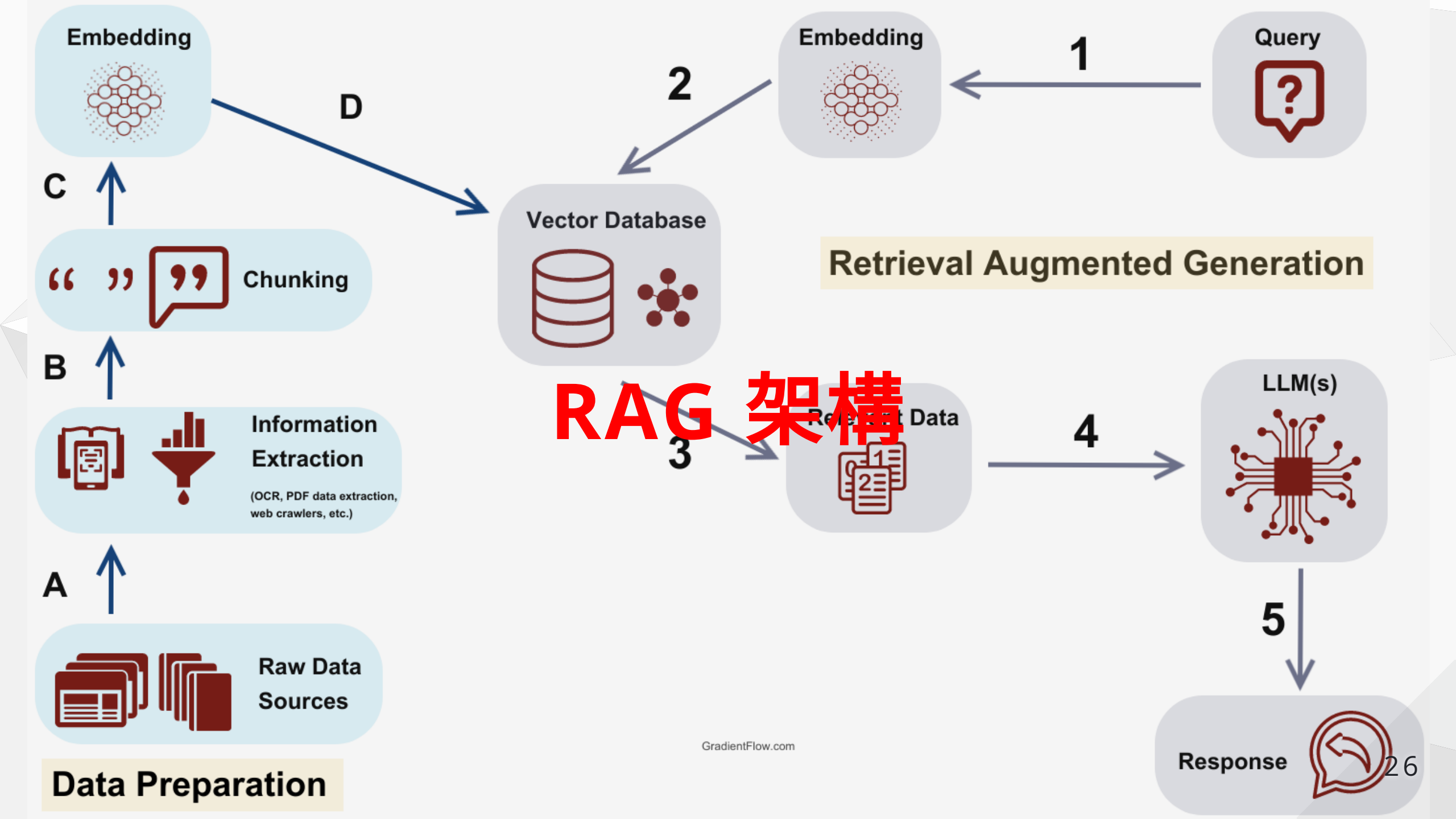
兩種企業級 LLM 應用

- 檢索增強生成 (RAG)
- 微調(Fine-tuning)



什麼是 RAG

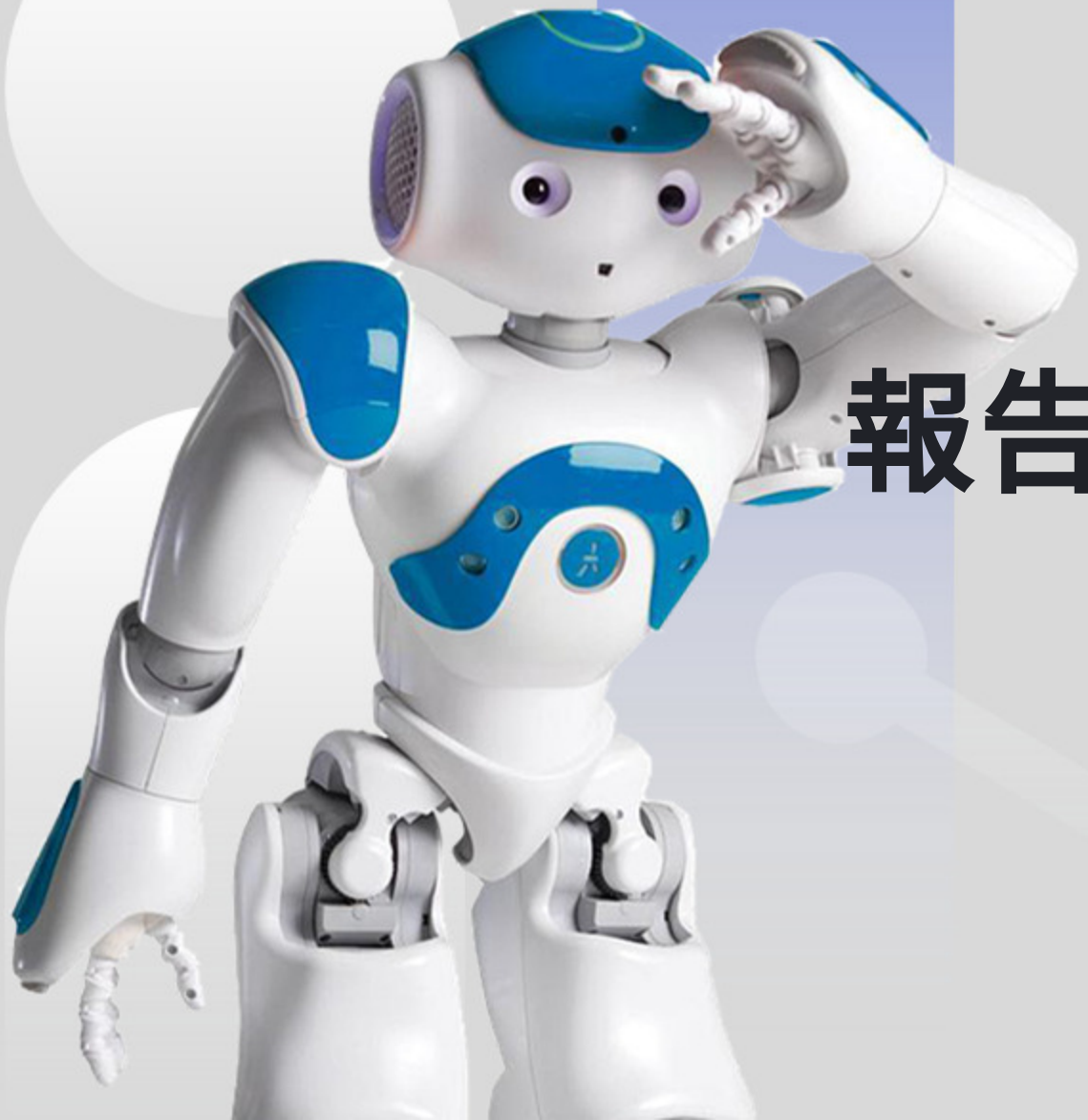
RAG（Retrieval-Augmented Generation，檢索增強生成）是一種新的技術，它允許大型語言模型（LLM）從**外部資料源**獲取並整合資訊


Embedding
D
2
Embedding
1
Query
C
Vector Database
Retrieval Augmented Generation
Chunking
B
RAG 架構
Information
Extraction
(OCR, PDF data extraction,
web crawlers, etc.)
3
Data
LLM(s)
4
A
5
Raw Data
Sources
Response
Data Preparation
GradientFlow.com

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