# GenAI 的生態演變與競爭格局

從 2022 年 11 月 30 日 OpenAI 推出 Chat GPT 以來，GenAI 產業突飛猛進，百花齊放，短短不到三年的時間，全世界學習與使用 GenAI 產品的人數超過 10 億，推動 GenAI 技術快速發展成每個人、每家企業所必需的通用技術。不過 AI 產業的發展其實已有數十年，本文將重點介紹 GenAI 的生態演變，以及將不同大小的語言模型分成三種類型。了解 GenAI 的生態演變與 GenAI 競爭格局中的「大型客機、私人飛機、紙飛機」三種模型，有助於快速掌握、跟進 AI 產業的高速發展與市場需求變化。

## GenAI 的生態演變

GenAI 從早期模型到當前最先進的狀態，其發展演變的各種里程碑，在在顯示了這是一個持續追求創新、改進的生態系。

### 1. 早期 AI 模型巨人

GenAI 在早期發展階段，各家科技巨頭紛紛推出自己的大型語言模型，掀起了一場技術革新風暴。隨著時間推移，有些模型逐漸淡出舞台，有些則持續壯大，成為市場領導者。

Google 的 BERT 模型，做為 Transformer 架構的早期應用，雖然擁有 3.4 億個參數，如今看來已顯得相對小型。儘管如此，BERT 在自然語言處理領域的貢獻不可磨滅，為後續模型的發展奠定了基礎。

OpenAI 的 GPT-3，憑藉 1,750 億個參數和驚豔的文本生成能力，一度成為 GenAI 的代名詞。其衍生的應用程式 ChatGPT 更是風靡全球，截至 2025 年 10 月，每週活躍用戶超過 8 億人，引領了在對話互動和內容創作領域的應用潮流。

Jurassic-1 由 AI21 Labs 所開發，擁有 1,780 億個參數，一度是文本生成領域的佼佼者。隨著 GPT 系列模型的崛起，Jurassic-1 的影響力雖然逐漸減弱，在高品質文本生成和上下文理解方面的能力仍不容忽視。

MT-NLG 由微軟和 NVIDIA 合作開發，擁有 5,300 億個參數，曾是規模最大的單體 Transformer 語言模型。由於龐大的運算需求和潛在的風險，MT-NLG 並未公開發佈，其應用主要用於研究目的。

Google 的 PaLM，擁有高達 5,400 億個參數，是 LLM 領域的重要里程碑。PaLM 在語言理解、推理和編程等多項任務上表現出色，為 Google 的 AI 應用提供了強大的技術支援。

除了上述模型，LaMDA（1,370 億個參數）、Gopher（2,800 億個參數）和 BLOOM（1,760 億個參數，開源）等模型也在早期發展中扮演了重要角色。LaMDA 專注於對話應用，Gopher 在通用理解方面表現出色，而 BLOOM 以其多語言能力和開源特性在研究和開發社群中廣受歡迎。

綜觀全局，GenAI 的早期發展呈現百家爭鳴的局面。各家公司在模型規模、架構設計和應用領域上不斷探索，推動了技術的快速疊代和創新。如今有些模型已逐漸淡出，但它們為後續模型的開發累積了寶貴經驗， 奠定 GenAI 發展的堅實基礎。

## 2. 現代巨人：巨人級的大型語言模型

下一代 GenAI 的巨人模型更複雜，融合了先進的技術和架構，擁有更大的規模和更強的能力（見表 1-1）。

表 1-1 現代巨人

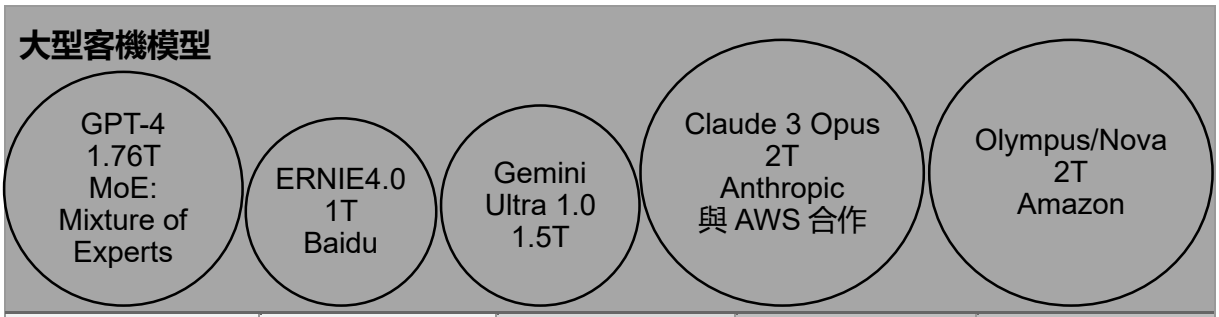
領先 AI 模型的比較分析，根據參數、訓練的 token 和架構設計和各自的發佈日期進行評估。Arch 欄位的 MoE、Dense、CoE 表示不同的模型架構。

| 模型<br>Model | 開發者<br>Lab | 參數<br>Parameters (B) | 標記訓練<br>Tokens trained (B) | 公布時間<br>Announced | 架構<br>Arch |
|---|---|---|---|---|---|
| **Olympus/NOVA** | Amazon | 2,000 | 40,000 | Dec/2024 | |
| **GPT-5** | OpenAI | 5,250 | | TBA | |
| **GPT-6** | OpenAI | | | TBA | |
| **AuroraGPT** | Argonne National Laboratory | 1,000 | | TBA | |
| **Grok-2** | xAI | | | TBA | |
| **PanGu 5.0 Super** | Huawei | 1,000 | 20,000 | Jun/2024 | MoE |
| **Yi-XLarge** | 01-ai | 2,000 | 20,000 | May/2024 | MoE |

| | | | | | |
|---|---|---|---|---|---|
| **Yi-Large** | 01-ai | 1,000 | 15,000 | May/2024 | Dense |
| **Med-Gemini-L 1.0** | Google DeepMind | 1,500 | 30,000 | May/2024 | Dense |
| **Inflection-2.5** | Inflection AI | 1,200 | 20,000 | Mar/2024 | Dense |
| **Glaude 3 Opus** | Anthropic | 2,000 | 40,000 | Mar/2024 | Dense |
| **Samba-1** | SambaNova | 1,400 | 20,000 | Feb/2024 | CoE |
| **Gemini 1.5 Pro** | Google DeepMind | 1,500 | 30,000 | Feb/2024 | MoE |
| **Gemini Ultra 1.0** | Google DeepMind | 1,500 | 30,000 | Dec/2023 | Dense |
| **Inflection-2** | Inflection AI | 1,200 | 20,000 | Nov/2023 | Dense |
| **ERNIE 4.0** | Baidu | 1,000 | 20,000 | Oct/2023 | Dense |

銳企製表，參考資料：LifeArchitect.ai

GPT-4：OpenAI 的 GPT-4 版本，擁有驚人的 1.76 兆參數，如圖 1-11 所示。該模型採用混合專家（Mixture of Experts, MoE）方法，使模型的不同部份專門處理不同任務，其架構創造更高的準確性和上下文理解能力，能夠執行更廣泛的任務。




| 奈米（Nano） | | 超小（Extra Small） | | 小（Small） | | 中（Medium） | | 大（Large） | |
|---|---|---|---|---|---|---|---|---|---|
| Gato (Cat) | 1.0 | Gemma | 7.0 | Palmyra | 20 | Command xlarge | 52 | Yuan 2.0 | 103 |
| TinyLlama | 1.1 | Gauss | 7.0 | AlexaTM 20B | 20 | StableLM | 65 | Command-R+ | 104 |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Meta-Transformer | 2.0 | DeciLM-7B | 7.0 | Codestral | 22 | DeepSeek | 67 | InternLM | 104 |
| PanGu-Coder | 2.6 | Zephyr | 7.3 | C1.2 | 33 | Llama 3 70B | 70 | DBRX | 132 |
| Meena | 2.6 | Mistral 7B | 7.3 | Code Llama 34B | 34 | Eurus | 70 | BLOOM (tr11-176B-ml) | 176 |
| Phi-2 | 2.7 | Striped Hyena 7B | 7.7 | Yi-34B | 34 | pplx-70b-online | 70 | Jurassic-2 | 178 |
| Mamba | 2.8 | Persimmon - 8B | 8.0 | Command-R | 35 | Luminous Supreme Control | 70 | Mistral-medium | 180 |
| Apple On-Device model | 3.0 | LLaMA Pro | 8.3 | mixtral-8x7b-32kseqlen | 47 | Qwen2 | 72 | Falcon 180B | 180 |
| | | SOLAR-10.7B | 11.0 | Retro 48B | 48 | | | Grok-1 | 314 |
| | | Pythia | 12.0 | | | | | PaLM 2 | 340 |
| …Many more | | …Many more | | ...Many more | | ...Many more | | ...Many more | |
| **紙飛機模型** | | | | | | **私人飛機模型** | | | |

圖 1-1　GenAI 分大、中、小三種競爭格局

現代巨人，大小超過 1 兆個參數，顯示在這張圖片的頂部。模型規格分為奈米（Nano）、超小（Extra Small）、小（Small）、中（Medium）、大（Large），數目的單位是 Billion（十億）。大、中、小三種格局由大到小，我們可以用大型客機、私人飛機、紙飛機來比喻三種不同參數量級的模型。

銳企製圖，參考資料：LifeArchitect.ai

Gemini Pro 和 Gemini Ultra：代表 Google 的重大進展，基於前代模型 PaLM 的成功，它們改進了自然語言理解和生成能力。Gemini Ultra 達到 1.5 兆參數，其模型結合了自監督學習中的前沿技術，無需大量標註資料就能勝任

多種任務。

Claude 3：由 Anthropic 與 AWS 合作開發的 Claude 3 擁有 2 兆參數，顯示科技公司之間透過合作以推進 AI 的能力。該模型的特色是負責任 AI，設計結合了安全和道德考量，符合人類價值觀。

ERNIE 4.0：百度的 ERNIE 4.0 擁有 1 兆參數，展示中國科技巨頭對 AI 研究的重大投資，專注於知識和語言理解的整合。ERNIE 4.0 利用知識圖譜（knowledge graph）來增強其理解和生成能力，在複雜資訊檢索和生成任務中表現出色。

### 3. GenAI 開發的主要趨勢

從早期到現代大型語言模型的演變，不僅顯示了參數數量的增加，還展示了模型方法和應用的多樣化，指出 GenAI 開發的幾個主要趨勢：

•模型參數的指數成長

GenAI 的參數數量和規模呈爆炸性成長，GPT-4、Gemini 和 Claude 3 等模型在規模上超越了前輩，達到了數兆個參數。這種成長證明對 GenAI 研究和開發的投資和興趣正不斷增加。

•閉源模型的主導地位

雖然 BLOOM 和 OPT 等開源模型仍然存在，但從現代巨人的資訊整理，可發現閉源模型的主導地位，特別是那些由 Google、OpenAI 等主要科技公司發展的模型 Gemini 和 GPT-4，和 Anthropic 的 Claude 3。這些模型通常只能透過 API 訪問，限制它們的透明度和使用者的控制。

•專業模型的出現

現代語言模型獲得突破性進展之後，開始有能力轉向針對特定任務設計的專業 LLM。例如 Med-PaLM 2 和 PaLM-Coder 分別針對醫療和編碼任務量身定制，顯示 LLM 日益增長的客製化需求。

•競爭格局

各公司在模型大小、效能和能力方面爭奪領先地位。除了大型語言模型之外，還有其他模型顯示出不同的大小、參數和所屬開發團隊，反應全球在這個領域所投入的努力。

## GenAI 競爭格局中的三種模型

GenAI 語言模型的發展和運用非常多樣，難以通體適用（one size fits all)、「一言」以蔽之。然而，我們可以根據語言模型的大小，分為大、中、小三種競爭格局。若以飛機型號的市場區隔來做說明，由大到小可以用大型客機、私人飛機、紙飛機，象徵性地比喻不同語言模型的能力範圍和部署規模。

GenAI 快速發展且日益多樣化，每種模型都有其獨特的優點和缺點。有些功能強大且用途廣泛，有些則專業且高效；大型模型需要強大的 AI 伺服器或雲端基礎設施，較小的模型可以部署在邊緣設備（edge devices）上執行特定任務。

### 1. 大型客機模型

「大型客機模型」代表那些由大型公司運營的強大 GenAI 模型，如 GPT-4、Gemini Ultra、Claude 3 和 Amazon 的 Olympus 等，就像波音 787 商業飛機。這些模型透過網路應用程式（web applications）或行動應用程式（mobile apps）提供給大眾使用，也透過 API 向企業或開發者提供服務，但它們只能租用，不是自行構建或擁有。

大型語言模型擁有數千億至上兆個參數，具備處理複雜任務的能力，在自然語言處理、圖像生成、程式碼編寫各領域表現卓越，但龐大的模型規模相對帶來高昂運算成本。「大型客機模型」使企業可以利用最先進的 AI 技術，而無需承擔基礎設施的負擔。

大型語言模型的快速發展，源於每一代模型訓練所使用的數據集顯著擴大，短短幾年間，規模從一代到下一代增加了約十倍。有些專家認為這種成長速度可能無法持續，也有專家認為更大的數據集和模型將持續提高 GenAI 的知識和準確度。

GenAI 處於起步階段，仍有相當大的發展潛力，提高準確性可能需要更大的模型，或將通用模型細分為更多針對不同特定領域的模型。隨著數據集的提取和雲資源的可獲得性不斷提高，語言模型的參數數量將持續以指數級增長，這場創新革命可能在未來十年內才會達到技術的極限。

由於大型語言模型多採用雲端運算，企業在使用時，資料須透過 API 傳輸至雲端，具有潛在的資訊安全風險，公司機密可能被 LLM 供應商所掌握。因此，許多企業暫時禁止員工使用像是 ChatGPT、Gemini 等 GenAI，避免

洩漏商業機密，部份企業選擇轉向開發自有的 AI 語言模型。

2. **私人飛機模型**

「私人飛機模型」就像企業可以向小型通用航空飛機製造商 Cessna 購買私人飛機，代表企業可以擁有和控制的 GenAI 模型，如 Mistral、DeepSeek、Falcon 和 Grok-1 等。這些 GenAI 模型通常由中小型企業或研究機構開發，具有更高的靈活性和客製化潛力。在特定的應用領域中，它們的表現不亞於大型模型，且更易於部署和維護。

這些模型的優勢在於，企業可以將其部署在自己的 IT 基礎設施中，確保數據的安全性和隱私性。對於處理敏感資訊的企業，例如銀行、醫院或半導體公司，基於資訊安全考量，通常不會使用雲端 GenAI 服務，而傾向使用內部托管的 GenAI 模型。

隨著 GenAI 技術的普及，企業必須現代化其 IT 基礎設施，聘請專業人員來管理和維護。然而，招募 AI 專業人才面臨諸多挑戰。根據 IDC 的研究，全球近 70% 的企業在運營 AI 系統時遇到人才短缺的問題，這一缺口將持續到 2026 年，並可能對經濟造成約 5.5 兆美元的影響。麥肯錫顧問公司指出，越來越多的中小企業開始尋求外部的 GenAI 技術服務，這一市場預計到 2029 年將超過 2,000 億美元。

3. **紙飛機模型**

「紙飛機模型」是指運行在邊緣設備上的小型模型，例如 Llama、Mistral 和 Apple OpenELM。這些模型由企業或個人管理，通常部署在本地或邊緣設備上，主要優勢在於減少對雲端和數據中心的依賴，能提供更快速的反應和更個別化的服務，並降低運營成本。

邊緣 GenAI（edge GenAI）能夠保護數據隱私，因為大部份處理都可以在本地完成，無需將敏感數據傳輸至雲端。和大型商業 GenAI 模型比起來，這些小型語言模型能夠快速回應用戶需求，降低延遲並提高數據安全性，這對於汽車自動駕駛、個人助理的應用尤其重要。此外，這些模型的設計使企業能夠根據特定需求進行調整和優化，適用於智慧手機、汽車和物聯網設備等多種場景。

根據 Tirias Research 的報告，預計到 2028 年，邊緣 GenAI 技術能夠減少 20% 的雲端處理需求，為企業節省高達 160 億美元的運營成本。這一

變化來自於將部份 AI 工作負載轉移到本地設備，減少對數據中心的傳輸與運算成本並提升運行效率。隨著邊緣 AI 的普及，GenAI 的應用將變得更加靈活，為企業和消費者提供更強大的服務。

## 結語

了解 GenAI 的生態演變與 GenAI 競爭格局中的大型客機、私人飛機、紙飛機三種模型，有助於快速掌握、跟進 AI 產業的高速發展與市場需求變化。自 OpenAI 於 2022 年 11 月 30 日推出 Chat GPT 以來，激發 GenAI 產業突飛猛進，該程式也在近三年的時間，從 GPT-3.5、GPT-4、GPT-4o、GPT-4.5 到 GPT-5 架構快速疊代。FirstPageSage 的統計數據顯示，該程式自推出以來，持續維持超過七成的大型語言模型（LLM）市占。OpenAI 2025 年營收預估達到 130 億美元，主要營收便是來自 ChatGPT 付費用戶和企業 API 授權，企業客戶佔三成，七成來自消費者，全球下載人數超過 10 億。

不過，這段期間，市場需求與產業競爭態勢也在快速變遷，OpenAI 不再壟斷全部市場，尤其是在企業級語言模型使用市場，Anthropic、Google、Meta 等大廠也都持續推出自主開發的 LLM，GenAI 產業進入百家爭鳴的爆發式發展狀態。之所以容得下大量湧入的競爭者，也是因為使用者人數快速上升，對 GenAI 產品的使用需求快速增長，應用面越來越廣，應用方式豐富多樣，帶動市場細分。

Menlo Ventures 的生成式 AI 報告指出，GenAI 的技術發展，正從追求更大模型開枝散葉，更專精於某些任務的語言模型正快速被推出，這類模型的建置和維運成本都可以降低很多，客製化程度也更高。同時，各大雲端服務商為了提供用戶更好的使用體驗，也積極整合多模型，例如微軟於 2025 年 9 月宣布將 Anthropic 的 Claude 納入其打造的 Copilot 生態，凸顯企業之間的競合關係也將愈加複雜。

## 參考資料

•ChatGPT 市佔穩居七成！OpenAI 再融資 83 億鎂，AI 資金戰持續升溫。CRYPTO CITY。2025

•當 OpenAI 不再一家獨大：LLM 戰爭進入群雄割據時代。Vocus。2025

·2025 Mid-Year LLM Market Update: Foundation Model Landscape + Economics. MENLO VENTURES (2025)