AI 巨頭之爭：模型性能評估與較量

近年來，AI 技術的發展可謂日新月異，各大科技公司如 OpenAI、Google 和 Anthropic 不斷推陳出新，推出更多高效能的機器學習模型。這些模型在多項性能基準測試中表現卓越，不斷地推高 AI 技術的標準。在人工智慧（AI）領域，這三家公司展開了激烈的競爭，各自努力成為開發先進 AI 技術的領導者。今天，我們將深入探討這三大巨頭所開發的模型核心特點！

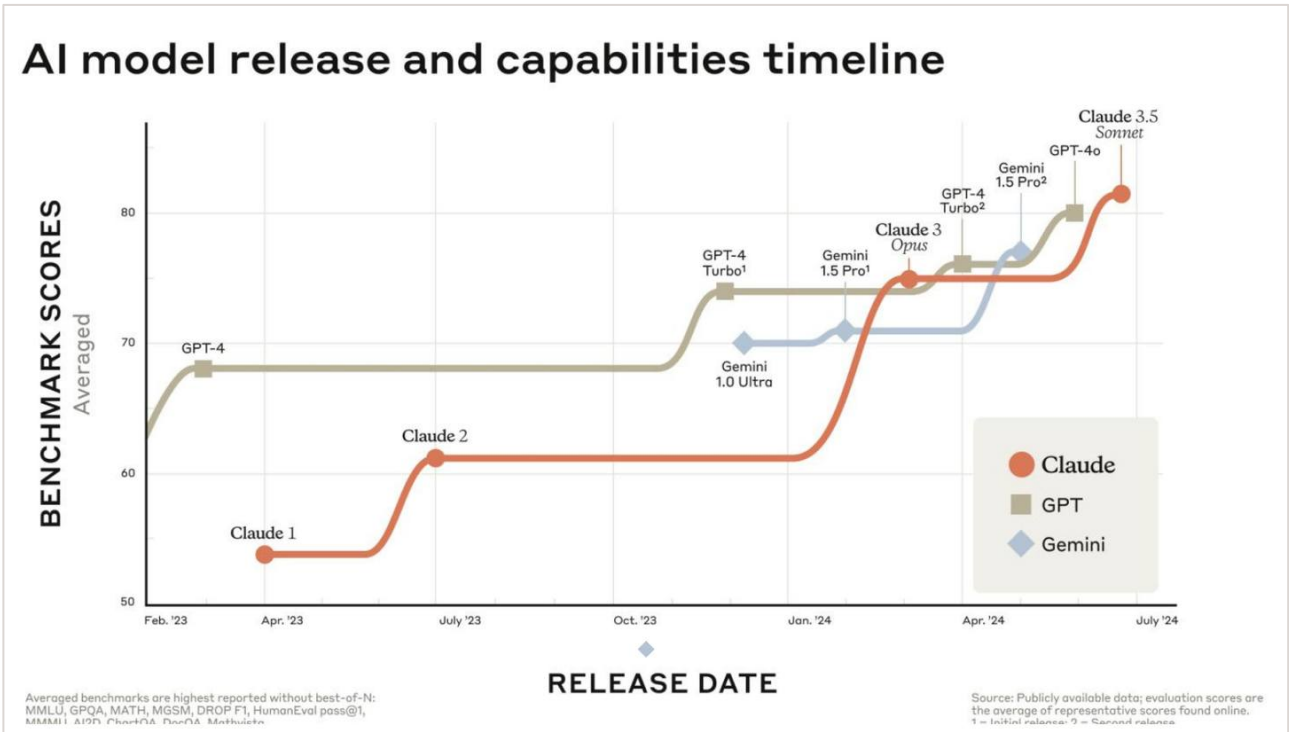
模型發展軌跡與性能評估

OpenAI、Google 和 Anthropic 在各自的模型在性能和應用場景上有所不同，OpenAI 以其 ChatGPT 等開創性模型而聞名，將自己定位為生成人工智慧領域的關鍵參與者，該公司已獲得大量投資，尤其是來自微軟的投資，OpenAI 的 GPT-4o 習得多模態能力和高效的推理能力激發優勢

Gemini 系列展現了 Google 在大型語言模型領域的優勢，透過多型能力、長上下文視窗和針對效能最佳化，為開發者提供了強大的 AI 工具，Google 已將人工智慧功能整合到其大量產品和服務中。該公司還對人工智慧新創公司進行了戰略投資，包括最近向 Anthropic 承諾投資 20 億美元，在 AI 創新公司中進行了戰略性投資。

Anthropic 的 Claude 系列在 AI 界同樣引起廣泛關注，尤其是最新的 Claude 3.5 Sonnet，在眾多基準測試中表現出色，超越了 OpenAI 的部分模型。其「Artifacts」功能為用戶提供了與 AI 生成內容進行動態交互的能力，支持即時協作和內容的快速修改，從而優化用戶體驗。

在評估這些 AI 模型時，除了基準測試的分數外，更重要的是其在實際應用中的表現，如語言理解和多模態互動等關鍵領域的實用性。隨著 AI 技術的進步，這些性能評估將幫助業界和用戶更好地理解各個模型的優勢和適用場景。AI 技術的不斷推陳出新，不僅開拓了應用領域，也提升了 AI 的普及化，使得更多的人能夠利用這些強大的工具來創造價值和創新。



AI巨頭之爭：模型性能評估與較量

• Claude 模型（Anthropic）

Claude 1：於2023年3月發布

Claude 2：於2023年7月推出，增強了推理和會話能力

Claude 3（Opus）：於2023年11月發布，具備增強的視覺能力和20萬token的上下文窗口，允許更好地處理複雜任務和文件分析

■ GPT 模型（OpenAI）

GPT-3.5：於2022年3月發布，為會話型AI設定了新標準。

GPT-4：適用於更廣泛的 NLP 任務，有更好的推理能力、上下文感知能力、多語言處理能力等

GPT-4 Turbo：於2023年11月發布，擁有一個包含 12.8 萬個參數的上下文窗口

GPT-4o：一種多模態模型，於2024年5月推出，比 GPT-4 的回應速度快一倍，可用語音直接溝通

◆ Gemini 模型（Google）：

Gemini 1：最初於2023年12月推出，專注於多模態能力。

Gemini 1.5 Pro 專業版 ：於2024年5月發布，擁有 200萬 token的上下文窗口，並在翻譯、編碼和推理任務中性能提升

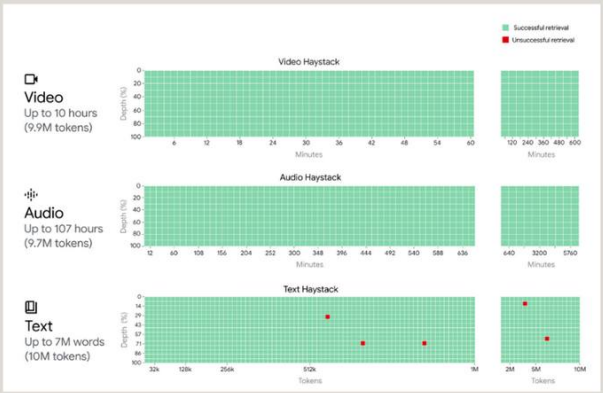
Gemini 1.5 Flash：專為快速且經濟高效的大規模服務而設計，非常適合即時回應的場景

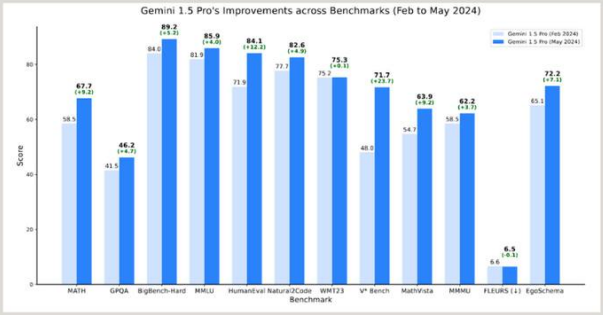
選擇合適的人工智慧模型

人工智慧模型	上下文視窗	主要特點
Claude Sonnet	200K tokens	強大的推理、文件視覺問答以及改進的各種視覺格式的處理
GPT-4o	128k tokens	多模式功能比 GPT-4 更快，改進了多語言支援
Gemini 1.5 Pro 專業版	200k tokens	複雜的推理任務，例如程式碼和文字生成 文字編輯、問題解決、資料擷取與生成

GOOGLE GEMINI

Google於2024年5月發布Gemini 1.5 Pro 上下文視窗可擁有200k token，不僅在文本方面有所提升，更能處理更長的文本、影片和音訊內容





◆ Gemini 1.5 的核心功能包括：

- 處理數百萬個 tokens 的超長上下文理解能力
- 增強的多模態推理能力： Gemini 1.5 在多項基準測試中展現出優於先前模型的效能，尤其是在需要跨文本、圖像、音訊和影片進行推理的任務中。
- 改進的語言理解和生成能力： Gemini 1.5 在多項語言任務中皆有顯著進步，包含機器翻譯、問答和程式碼生成等。
- 更高的效率和更小的模型尺寸： Gemini 1.5 Flash-8B 僅有 80 億個參數，卻展現出與更大模型相當的效能，證明其在資源有限的情況下仍能保持高效。
- Gemini 1.5 Pro模型在文本、視覺和音頻三大核心能力上，相比於早期版本的性能提升。

Gemini 1.5 Pro	Relative to 1.5 Pro (Feb)	Relative to 1.0 Pro	Relative to 1.0 Ultra
Long-Context Text, Video & Audio	no change	from 32k up to 10M tokens	from 32k up to 10M tokens
Core Capabilities	Win-rate: 78.1% (25/32 benchmarks)	Win-rate: 88.0% (44/50 benchmarks)	Win-rate: 77.8% (35/45 benchmarks)
Text	Win-rate: 78.6% (11/14 benchmarks)	Win-rate: 95.8% (23/24 benchmarks)	Win-rate: 84.2% (16/19 benchmarks)
Vision	Win-rate: 92.3% (12/13 benchmarks)	Win-rate: 95.2% (20/21 benchmarks)	Win-rate: 85.7% (18/21 benchmarks)
Audio*	Win-rate: 80% (4/5 benchmarks)	Win-rate: 60% (3/5 benchmarks)	Win-rate: 40% (2/5 benchmarks)

CLAUDE 3.5

Claude 3.5 Sonnet：Anthropic AI 語言模型系列中的最新模型，於2024 年6 月21 日正式推出。

● Claude 模型（Anthropic）

- Claude 3.5 Sonnet 效能提升：運行速度是 Claude 3 Opus 的兩倍，使其能夠高效執行複雜任務。它在研究生程度推理GPQA、本科程度知識 (MMLU) 和編碼能力 (HumanEval) 等領域設立了新的行業基準。
- Claude 3.5 Sonnet 以其強大的視覺處理能力而聞名，在解釋圖表、圖形和圖像方面超越了以前的模型。
- 在編碼評估中，Claude 3.5 Sonnet 解決了 64% 的問題，明顯優於 Claude 3 Opus。它可以獨立編寫、編輯和執行程式碼，使其成為軟體開發和遺留程式碼遷移的有效工具。
- Claude 3.5 Sonnet 在 Claude.ai 和 Claude iOS 應用程式上免費提供。
- 引入Artifacts變革交互方式：當使用Claude生成代碼、文本或網站設計內容時，可以透過對話旁邊的視窗快速查看、編輯和修正。

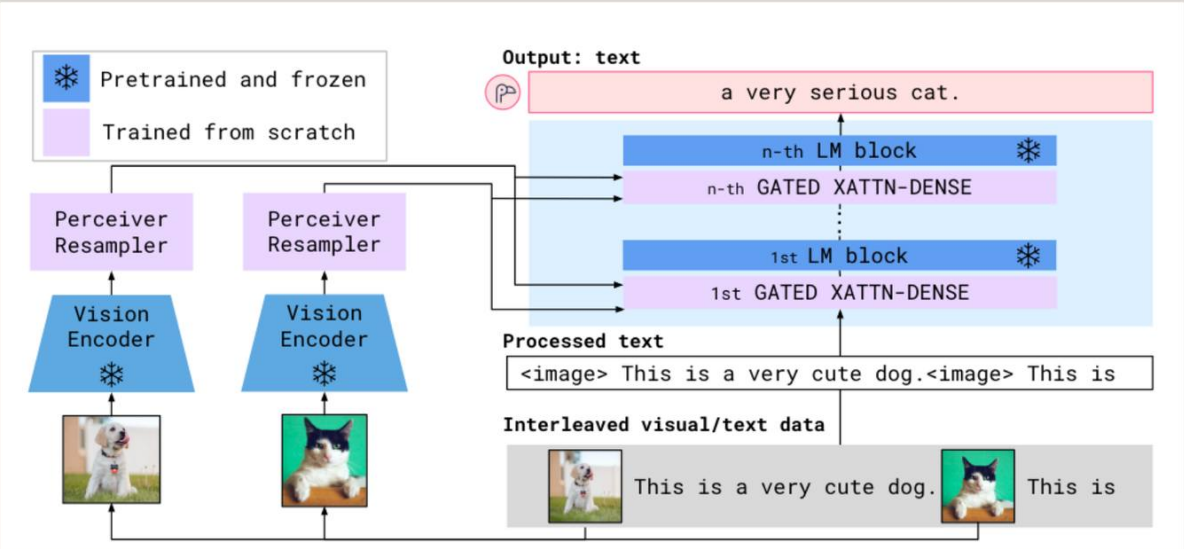
	Claude 3.5 Sonnet	Claude 3 Opus	GPT-4o	Gemini 1.5 Pro	Llama-400b (early snapshot)
Graduate level reasoning *GPQA, Diamond*	59.4%* 0-shot CoT	50.4% 0-shot CoT	53.6% 0-shot CoT	–	–
Undergraduate level knowledge *MMLU*	88.7%** 5-shot 88.3% 0-shot CoT	86.8% 5-shot 85.7% 0-shot CoT	– 88.7% 0-shot CoT	85.9% 5-shot –	86.1% 5-shot –
Code *HumanEval*	92.0% 0-shot	84.9% 0-shot	90.2% 0-shot	84.1% 0-shot	84.1% 0-shot
Multilingual math *MGSM*	91.6% 0-shot CoT	90.7% 0-shot CoT	90.5% 0-shot CoT	87.5% 8-shot	–
Reasoning over text *DROP, F1 score*	87.1 3-shot	83.1 3-shot	83.4 3-shot	74.9 Variable shots	83.5 3-shot Pre-trained model
Mixed evaluations *BIG-Bench-Hard*	93.1% 3-shot CoT	86.8% 3-shot CoT	–	89.2% 3-shot CoT	85.3% 3-shot CoT Pre-trained model
Math problem-solving *MATH*	71.1% 0-shot CoT	60.1% 0-shot CoT	76.6% 0-shot CoT	67.7% 4-shot	57.8% 4-shot CoT
Grade school math *GSM8K*	96.4% 0-shot CoT	95.0% 0-shot CoT	–	90.8% 11-shot	94.1% 8-shot CoT

* Claude 3.5 Sonnet scores 67.2% on 5-shot CoT GPQA with maj@32
** Claude 3.5 Sonnet scores 90.4% on MMLU with 5-shot CoT prompting

端對端多模態技術

Flamingo 是一系列視覺語言模型 (VLM)，能夠透過少樣本學習快速適應新任務。 它們結合了預先訓練好的視覺模型和語言模型，可以處理交錯的視覺和文本數據序列，並無縫地接收圖像或視訊作為輸入。

- 視覺編碼器: 將圖像或影片轉換為視覺特徵。
- 感知器重採樣器: 接收來自視覺編碼器的視覺特徵，並輸出固定數量的視覺標記。
- 凍結的語言模型: 使用預先訓練好的大型語言模型 (LLM)，並在其中插入新的交叉注意力層來整合視覺訊息。
- 門控交叉注意力-密集層 (GATED XATTN-DENSE): 讓語言模型在預先訓練好的 LLM 層之間加入視覺資訊，並透過 tanh 門控機制來穩定訓練。



Flamingo 架構概述 | Flamingo 是一系列視覺語言模型 (VLM)
它將與文字交錯的視覺資料作為輸入，並產生自由格式的文字作為輸出

參考資料：

https://www.kapler.cz/wp-content/uploads/gemini_v1_5_report.pdf?fbclid=IwY2xjawEo7W5leHRuA2FlbQIxMAABHTostEMafCd32Cg8mlEM2JQTev80s0VJ6BFAoE6uqgIVHEsAbJ_O-lYr8g_aem_K8ihAVu5zKyzS-3x9rhr6Q

https://arxiv.org/abs/2204.14198?fbclid=IwY2xjawEo7aVleHRuA2FlbQIxMAABHSytdKBE-vyM1KncJyFtlw6UpVLAc71-PhcH5u9GYKtrcWy7J6dzyzmj7A_aem_EKtWTrYSCEYsEdnnEHpx0g