高性能、低成本的突破
OpenAI 推出 GPT-4o mini 小型語言模型

OpenAI 又帶來了一項令人振奮的消息——全新的 GPT-4o mini 小型語言模型已經推出，直接取代 GPT-3.5 Turbo ，而且現在用戶已經可以直接使用拉！不僅性能卓越，成本極低，無疑將開啟 AI 應用的新篇章。

GPT-4o mini 的出現是 AI 技術中的一次重大創新。在多項基準測試中，這款模型展現了驚人的實力，尤其在大規模多任務語言理解（MMLU）測試中取得了 82% 的高分，甚至在 LMSYS 聊天排行榜上超越了 GPT-4。 在解決數學和編程問題方面，GPT-4o mini 也以 87% 的卓越表現領先於 Gemini Flash 和 Claude Haiku 等競爭對手。

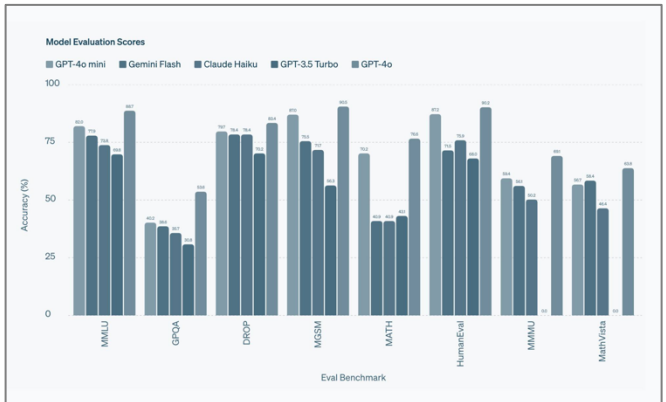
這款小型模型的定價真的很甜，GPT-4o mini 的商用價格以每百萬輸入 Token 僅需 0.15 美元，每百萬輸出 Token 則為 0.6 美元，比 GPT-3.5 Turbo 的成本降低了超過 60%，比 GPT-4o 更是降低了 96%，讓更多開發者和企業能夠輕鬆負擔並部署先進的 AI 應用啦！

不僅如此，GPT-4o mini 支持多模態輸入輸出，未來將擴展到影像、視訊和音訊處理，大大拓寬了其應用範圍。該模型擁有 128K tokens 的上下文窗口和 16K 的最大輸出長度，並支持串行和並行調用，展現了極致的靈活性和可擴展性。 由於採用了 GPT-4o 的改進版 tokenizer，在處理非英文資料時更加高效。

GPT-4o mini 透過 Assistants API、Chat Completions API 和 Batch API 提供服務，雖然目前 GPT-4o mini 的 API 僅支持文本和視覺，但之後會在釋出「語音」的測試版，未來也會支持更廣泛的運用！

最後想問大家覺得，「模型的大小重要嗎」？

在深度學習和人工智慧領域，模型的大小確實是一个重要的考量因素，但它並非唯一的決定性因素，較大的模型，通常參數數量和計算複雜度也會更大！而 GPT-4o mini 透過小型化設計，在性能和成本效率方面都取得了很好的平衡，OpenAI 承諾將持續提升模型性能並降低成本，為每個人帶來便利。讓我們拭目以待。



OpenAI 的 GPT-4o mini 模型

在多個關鍵基準測試中，GPT-4o mini
在涉及文字和視覺的推理、數學推理和編碼能力，以及多模態推理上均表現出色，
相較於其他小型模型如 Gemini Flash 和 Claude Haiku 有較高的得分

參考連結：
https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/?fbclid=IwY2xjawEo79VleHRuA2FlbQIxMAABHXsfDK80ncKQvp1M-hxD2a8vthJIEGRTx1G2sEr-usRW77_rPD_-B2Oqog_aem_Re-6P0WUvsRHjJjOwgX2uA

https://artificialanalysis.ai/models/gpt-4o-mini?fbclid=IwY2xjawEo7-ZleHRuA2FlbQIxMAABHXWhYxwQ-9POuBTvDjjsxEMdLY0nQ0MGAy1fmm6jmdkhN6Rtg8Kryfg3SQ_aem_xoo3pCD6W601lhGRAvhy_Q