MediaTek Research M7 On-premises

M7 是由聯發創新基地 (MediaTek Research) 開發的大型語言模型 (LLMs)。除了具備英文能力之外，M7 還專注於繁體中文能力的開發。它利用先進的機器學習技術達到卓越的回答，在繁體中文 TC-Eval 評測下，於商業常用場域 (包含：文本理解、文本摘要和分類) 達到與 GPT3.5 (turbo-1106) 同等水準。為了因應商業需求，M7 研發過程已經考量並調配至其具有低部署成本、低營運成本的特性。再者為了回應顧客對於資料安全的疑慮，我們將允許顧客在本地端機器使用 M7，稱之為 M7 On-premises。此外我們也提供客製化的服務—進階微調服務，以優異的技術滿足開發者的專業需求。我們相信 M7 將成為您商業應用的好夥伴，歡迎與我們洽談合作。

語言能力

在商業常用場域上 M7 與 GPT3.5 同等級，以下為 TC-Eval 的評測結果：

類別	項目	M7	GPT3.5
文本理解	DRCD (EM)	77.0%	78.4%
	FGC (EM)	38.0%	36.0%
文本摘要	XSum-TC (Rouge2)	0.0389	0.0316
分類	IMDB-TC (ACC)	91.6%	94.1%

推論效率

M7 採用專為繁體中文應用場景量身打造的演算法，在同樣的運算資源下，M7 的推論速度能達到開源模型 LLaMA 2 的兩倍以上。在對速度和反應靈敏度有極高要求的情境下，M7 無疑是您打造即時應用程式的最佳選擇。

本地部署

M7 On-premises 的本地部署特性，能夠滿足那些對數據隱私和安全有嚴格要求的組織。透過本地部署，這些組織能確保所有的數據都儲存在自己的機器中，並且擁有完全的控制權。

硬體需求

為平衡模型表現與推論速度，推薦使用一張有 48GB VRAM 的 GPU（如 L40），或是二張有 24GB VRAM 的 GPU（如 RTX 4090）。

成功案例

在與多家企業的可行性評估中，我們驗證了 M7 將可運用在包括：基於內部資料的問答機器人、資訊抽取、文件摘要等場景。

免費試用

我們提供介面供開發者做簡單測試，申請方式請詳見：github.com/mtkresearch/MR-Models 如果需要 API 和 On-premises，請進一步與我們聯繫。

進階微調服務

為了回應實務上的客製化需求，我們持續研發進階微調技術，目前開放以下進階微調服務：

- 學習專業知識：在這個服務中，M7 會被微調來學習特定領域的知識以解決開發者的專業需求，在這個微調後的專業模型您將可以使用 Prompt 來完成多任務業務需求。
- 回答偏好調整：在這個服務中，我們會使用強化學習來微調 M7 以符合開發者對模型回答的偏好，例如：安全性、拒答、回答風格，為此我們也準備一些獎勵模型供使用，當然您也可以提供資料來完成特殊的客製化需求。

合作夥伴

- 基礎設施提供商： 台智雲
- 資料提供商：意藍資訊
- 應用工具：聯發科達哥團隊