《邊緣計算新紀元》- 大型語言模型邊緣落地評測

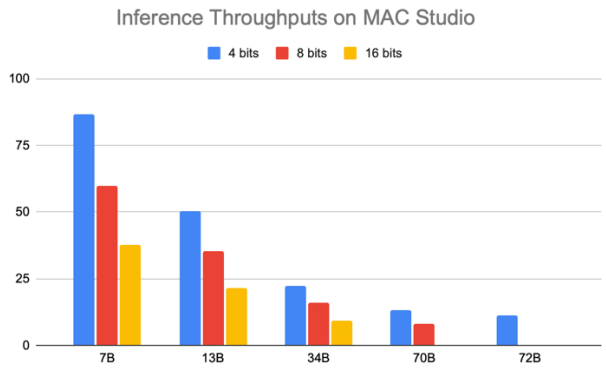
大型語言模型 邊緣落地 實現

隨著 AI 技術的快速發展，大型語言模型（LLM）的應用範圍不斷擴大，從雲端運算到邊緣裝置，其實現方式也越來越多元化。

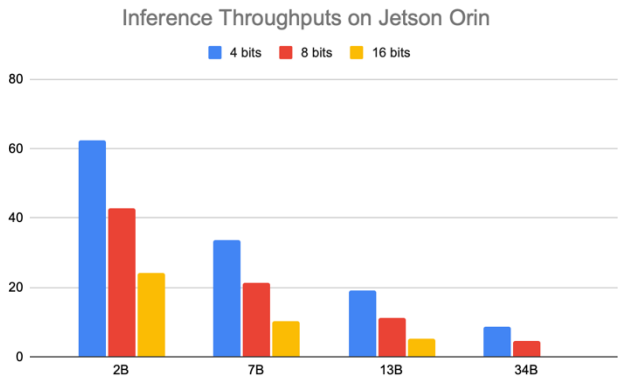
本次展示涉及一系列市面上容易獲得的高、中、低端邊緣裝置，對大型語言模型在邊緣裝置上的推論性能進行全面評測。我們測試的裝置包括：MAC Studio、NVIDIA RTX 4090、NVIDIA Jetson Orin、NVIDIA Jetson Xavier、Neuchips N3000、Raspberry Pi 4 / 5

本次測試中使用的大型語言模型包括多個版本的 LLaMA2（70B、13B、7B、1B）、Mistral-7B、Breeze-7B 以及 Smaug（34B、72B），這些模型的參數量和架構差異，使它們在不同的硬件上表現出各自獨特的性能特點。

MAC Studio

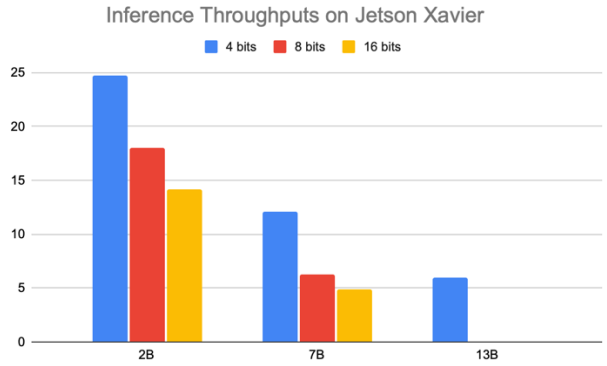


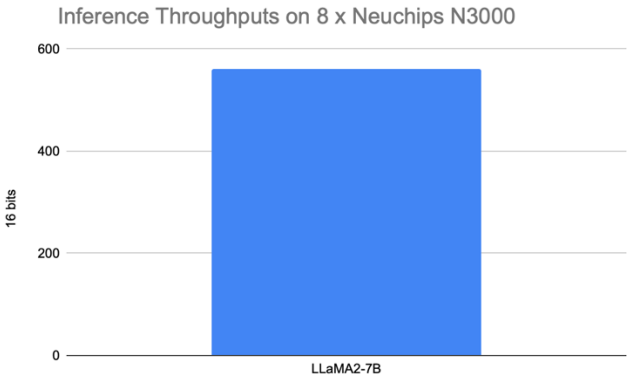
Jetson Orin



Jetson Xavier

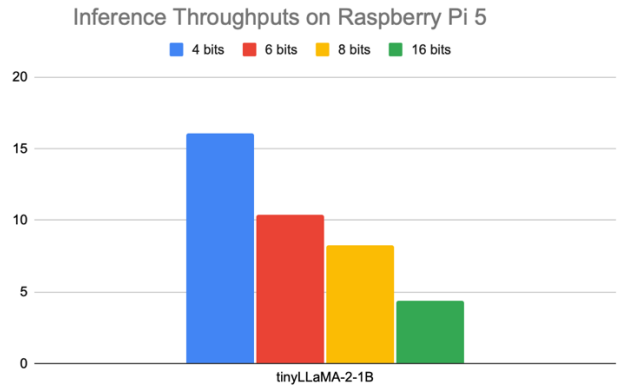
Neuchips N3000

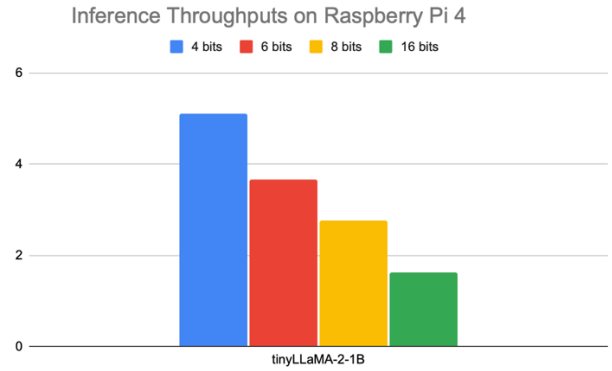




RPi 5

RPi 4





本次評測的結果將為 AI 研究者、開發者和愛好者提供寶貴的參考資訊，幫助大家更好地選擇適合自己需求的邊緣裝置。無論您是對 AI 技術應用於實際生活中感興趣，還是希望探索 LLM 在不同硬件上的運行效能，本期電子報都將為您提供豐富的資訊和洞見。