

客製化部署：實現 LLM 於邊緣裝置的挑戰與機遇

客製化邊緣裝置部署

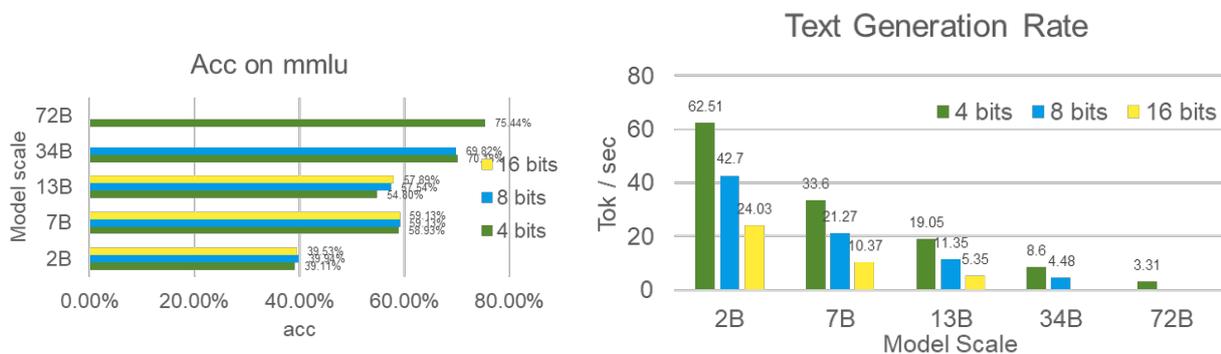
在當今數據驅動的時代，大型語言模型（LLM）的邊緣落地實現成為了企業追求智慧化服務的一大關鍵。為了在具有不同能力的邊緣裝置上部署適合的大型語言模型，我們進行了一系列的實驗。這些實驗包括評估各種大型語言模型在邊緣裝置上的可行性、性能以及模型能力。通過這些測試，我們旨在找出最適合在邊緣計算環境中運行的語言模型，實現智慧高效邊緣計算解決方案。

客製化邊緣裝置部署的實踐

我們的實驗聚焦於以下幾個核心應用場景，以展示大型語言模型在邊緣裝置上的多樣化應用潛力：

1. 智慧企業文本管理中心：透過 LLM 提升企業內部文檔的管理效率和檢索精度，從而加快信息流通和決策效率。
2. FAQ 智慧客服：利用 LLM 提供準確快速的客戶問題解答，提升客戶服務體驗。
3. 智慧工具代理：結合 LLM 與自動化工具，為企業內部的開發、運維等提供智能化支持。
4. 演講統整智慧助手：運用 LLM 進行會議內容的即時整理和摘要，提高會議效率。

不同大型語言模型在邊緣裝置的表現



大型語言模型在邊緣裝置的不同應用場景

1. 智慧企業文本管理中心



The screenshot displays the myLLM ChatRAG plus interface. On the left, a chat window shows a user query: "如何降低降腸癌死亡率?". Below the query, a series of status messages indicate the completion of various processing steps: "Search complete!", "Rerank complete!", "Filter complete!", "Compress complete!", and "Retrieving complete!". A summary paragraph follows, explaining that the reduction of colorectal cancer mortality can be achieved through regular screening, and that the current method in Taiwan is the fecal immunochemical test (FIT), which has high sensitivity for detecting asymptomatic high-risk colorectal cancer.

On the right, a PDF document is displayed, titled "壹、大腸癌篩檢效益". The document discusses the benefits of fecal immunochemical testing (FIT) for colorectal cancer screening. It compares FIT to traditional guaiac fecal occult blood tests (gFOBT) and highlights its superior sensitivity, especially for early-stage and advanced adenomas. The text also mentions that FIT can reduce colorectal cancer mortality by 22% and that FIT-based screening can reduce mortality by 50.4% compared to gFOBT.

2. FAQ 智慧客服

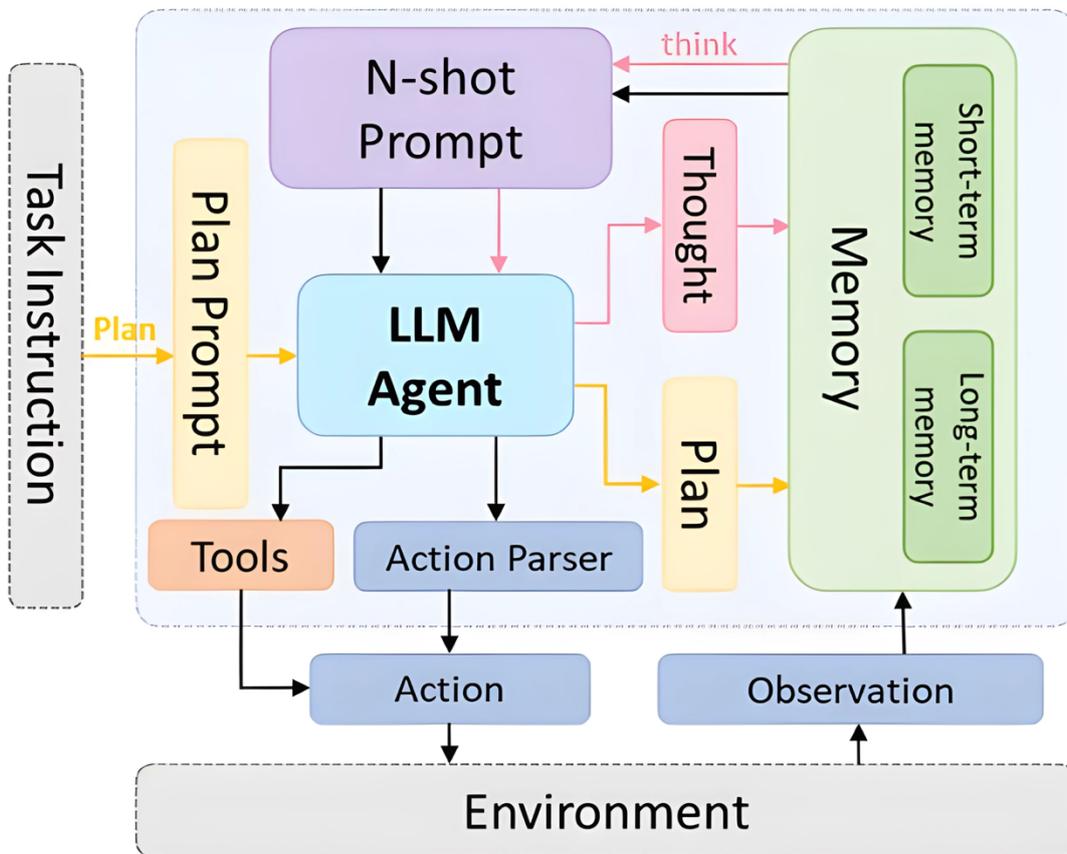


The screenshot shows the myLLM Chat (Orin) interface. A user query is displayed: "收到發票後，可否改開立為多張發票?". The system response explains that according to legal regulations and internal policies, one invoice can only correspond to one expense or one transaction. Therefore, it is not possible to split a single invoice into multiple ones. The response suggests that if the user needs multiple invoices, they should request them at the time of payment by specifying different expenses or transactions.

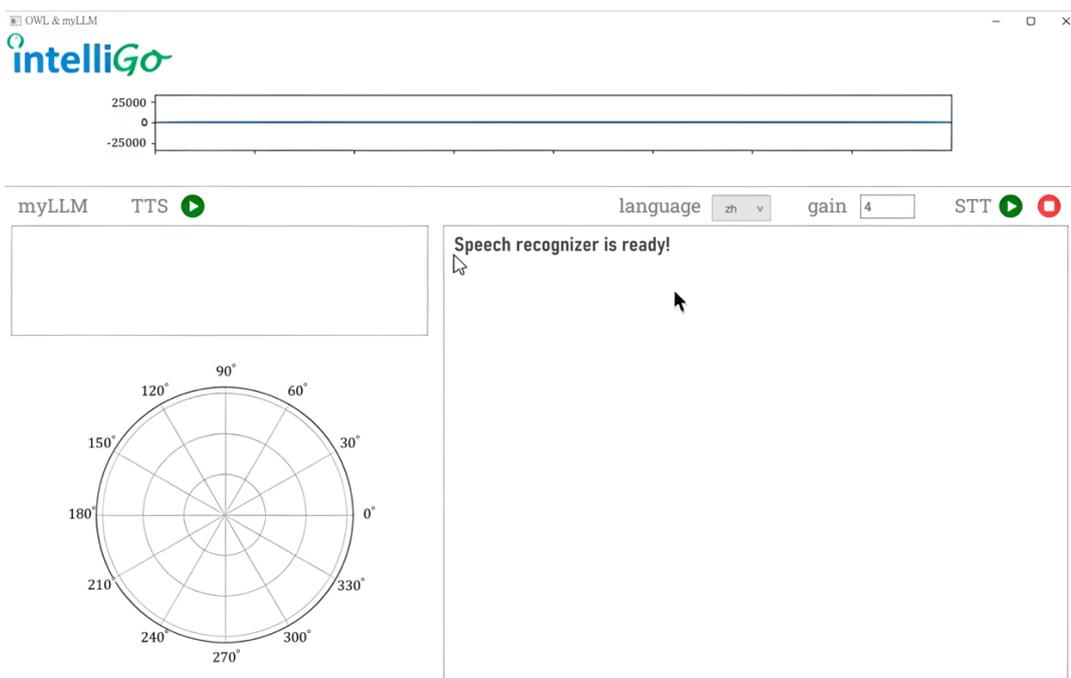
Below the main response, there are two additional FAQ items:

- 收到發票後，可否改開立為多張發票?**
根據法律規定和我們的內部政策，一個發票只能對應一個費用或一個交易。因此，收到發票後無法將它改成多張發票。建議您在下一次支付時，請提出不同的費用或交易，以獲取不同的發票。如果您有其他問題或需要更多幫助，請隨時告訴我。
- 如果覺得自己的帳號遭到駭客入侵，或遭到別人盜用了，該怎麼辦?**
首先，您需要確保您的電子信箱和其他相關帳戶（例如Facebook、Gmail等）的安全性。以下是一些可以做的步驟：
 - 更改被入侵的帳戶密碼：請使用強密碼（至少包含英文、數字和特殊字元），並避免使用易記或常用密碼。
 - 檢查您的電子信箱和其他帳戶是否被更改：確保您的電子信箱收到的任何新密碼重設請求或其他不常見的通知是來自您本人。
 - 檢查您的電子信箱和其他帳戶是否被添加了任何新的應用程式或設備：這可能表明您的帳戶已被盜用。
 - 檢查您的電子信箱和其他帳戶是否被發送任何不常見的電子郵件：這可能包括潛在的惡意連結或惡意程式。
 - 檢查您的電子信箱和其他帳戶是否被黑客留下的任何惡意程式或惡意連結：請避免點擊任何不確定的連結，並使用可靠的安全軟件進行掃描。
 - 通知相關平臺或服務提供商：請通知您使用的平臺或服務提供商（例如Facebook、Gmail等），告訴他們您的帳戶可能被盜用，並請求他們進行相應的調查和安全措施。
 - 更改其他相關帳戶密碼：如果您認為您的其他帳戶可能被黑客所竊取，請立即更改它們的密碼，並使用強密碼。
 - 保持警覺：請保持警覺，並定期檢查您的電子信箱和其他帳戶，以確保它們沒有被黑客所進行任何惡意操作。

3. 智慧工具代理



4. 講統整智慧助手



本次 Demo 使用機型



通過這一系列的實驗和探索，我們不僅擴展了大型語言模型在邊緣裝置上的應用邊界，更為企業提供了一套靈活、高效的智慧化邊緣計算解決方案。這不僅標誌著企業數字化轉型的一大步，也為未來邊緣計算的發展開闢了新的道路。隨著技術的不斷進步和創新，我們相信，大型語言模型在邊緣裝置的應用將會更加廣泛和深入。