國立臺北大學資訊工程學系114學年度專題

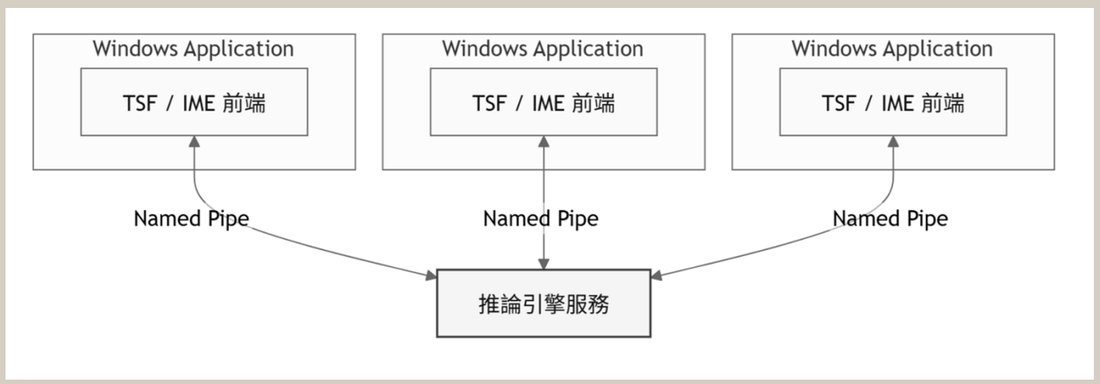
> Transformer-based 注音輸入法

成員: 劉晉嘉、陳毓霖、洪紹軒

簡介

目前多數注音輸入法採用傳統分詞引擎+n-gram頻率表的架構，缺乏對前後文語境的理解，導致時常出現預測錯誤，需要使用者手動選字的情況，這讓使用者不僅打字效率低落，還嚴重影響打字體驗，同時如今注音輸入法使用者已佔大宗，我們發現不少人遇到上述提到的困擾，並且多數願意犧牲一點速度來換取準確性，這讓我們決定使用現有 LLM 架構，以現代 LLM 工具鏈及架構，打造全新開源輸入法模型，期望能提高輸入法選字體驗。

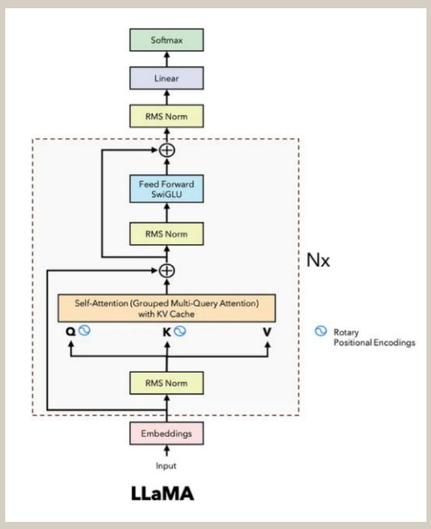
系統架構



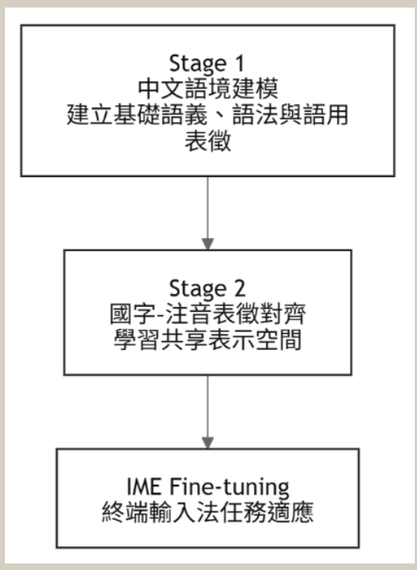
前端為基於 windows TSF COM 框架之自製前端，後端以llama.cpp進行推論，前後端以Named Pipe進行溝通

模型

模型架構



- **LLaMA for causal LM**
- Decoder-only causal language model
- layers：20
- hidden size：1024
- attention heads：16
- FFN / intermediate size：2048
- 參數量 248M



訓練流程

Stage1:
近乎全中文字 next prediction pretrain，建立基礎中文語境能力

Stage2:
訓練資料包含注音對國字與國字對注音，讓模型學習國字注音對照在內部產生共同表示空間

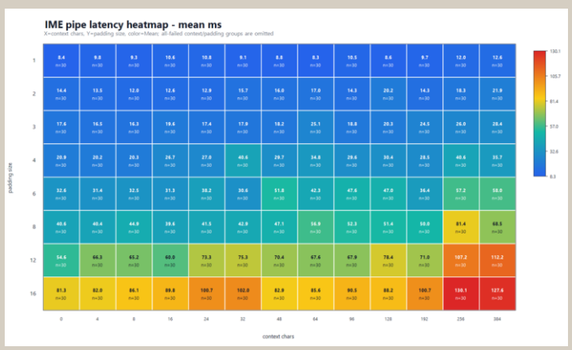
IME Fine-tuning:
結合 G2PW 標註資料與部分 Stage 2 資料進行 Multi-task Fine-tuning，使模型能根據前文與注音預測目標國字

成果

微軟注音 ✖	我們的輸入法 ○
你今天有藥丸嗎	你今天有要玩嗎
你可以幫我化成圖表嗎	你可以幫我畫成圖表嗎
你等我一夏	你等我一下
等一下憶起吃飯嗎	等一下一起吃飯嗎
剩下的資料在復健中	剩下的資料在附件中
多變亮分析	多變量分析

▲ 與微軟注音之比較

有效改善微軟注音輸入法因無法理解語境造成的錯誤



▲ 推論速度測試

大部分情況下延遲時間都在可接受範圍內

未來展望

希望未來能夠透過記錄使用者輸入紀錄，讓使用者能對模型進行微調，以達到個人化效果，同時希望能吸收目前輸入法之優點，將本模型作為 reranker，提高0-context準確率及選字穩定度，最後希望能強化表格之類等不同類型的上下文搜集，讓我們輸入法在不同場景下也能擁有更好的選字穩定度。

國立臺北大學
資訊工程學系
Department of Computer Science and Information Engineering,NTPU