補全方法比較表

方法編號	技術方式	優點	缺點
方法一	CLI 單次補全 (命令列)	- 寫法簡單、適合快速測試模型 - 支援多個句子生成 - 容易嵌入到其他流程中	- 無互動性，只能一次性生成 - 無候選選項或信心分數 - 無法持續接續或擴展語境
方法二	Gradio 即時網頁補全	- 使用者介面直觀、易上手 - 即時補全、自動更新 - 可本地或遠端部署	- 僅產生單句，互動性低 - 無選項選擇、無接句功能 - 輸出長度與控制不夠彈性
方法三	Flask API + CLI 補全迴圈	- 有互動輸入、選擇與續句能力 - 顯示 log-prob 信心分數 - 可拓展成 Web API 或 GUI 工具	- 使用門檻稍高（需理解 CLI 流程） - 沒有圖形界面（需手動建 GUI） - 結果無法直接儲存或複製分享

模型比較

模型名稱	風格傾向	簡繁體支援	品質評估	備註
原始 GPT-2 中文	混亂、口語不通	不能穩定處理繁體，易出錯	*	偏實驗用途，非實際場景推薦
ckiplab/gpt2-base-chinese	清晰分詞、風格制式	強繁體支援，分詞較穩定	**	適合繁體任務，輸出比較偏向政治
IDEA-CCNL/Wenzhong-GPT2-110M	敘事感強，有創意但語法不穩定	偏簡體，繁體輸入會轉寫或失準	***	適合模擬口語或情境式生成
uer/gpt2-chinese-cluecorpussmall	口語自然、近網民語言	偏繁體，但輸出有時候會出現簡體	***	美食評論風格明顯，實用性佳

max_length 實測分析：

長度設定	輸出風格	流暢度	結尾中斷感
max_length = 20	通常僅一兩句，偏口語	中等（需良好接句）	結尾通常突然結束或缺詞
max_length = 50	可構成段落，有情境	高（語意連貫）	結尾略突兀