□□ Llama.cpp

- 1. □□□□

1. 基本操作

1. 至AI智慧整合點選 Llama.cpp
裡的設定，點選啟用，待狀態顯示為綠色號誌，即啟用成功
llama0.png

2. 至儀表板點選紅框處即可進入頁面
llama01.png

3. 進入 llama-ui 後，會看到主要對話畫面，可直接進入新對話
左側為功能選單，包含 New chat、Search、 MCP Servers、Settings 與 Recent conversations。
中間為對話輸入區，可直接輸入訊息，或透過左下方加號上傳檔案。
下方會顯示目前使用的模型名稱與模型資訊，方便使用者確認目前 AI
模型狀態。
llama.png

4. 點選左側 Search，可搜尋歷史對話紀錄。
使用者只要在搜尋框輸入關鍵字，系統即會顯示相關對話結果。
此功能可協助快速查找過去產生的內容、提示詞、文件整理紀錄或重要對話。
llama2.png

5. MCP Servers 用於新增與管理外部工具伺服器。

若畫面顯示尚未設定 MCP Servers，代表目前尚未啟用外部工具功能。

使用者可點選右上角 Add New Server 新增 MCP Server，讓 AI
能夠串接更多工具或服務，例如檔案讀取、資料查詢、自動化流程或其他系統
功能。

llama3.png

6. 點選左側 Settings，可進入系統設定頁面。

此處可設定系統主題、 API Key、System Message
、長文字處理方式，以及是否按 Enter 送出訊息。 System Message 可用來設定 AI
的角色與回答規則，例如客服助理、文件整理員、專業顧問或企業內部助理，
設定完成後，需點選 Save settings 儲存設定。
llama4.png

7. Display 用於調整對話畫面的顯示方式。

此處可設定是否顯示生成統計、思考過程、工具呼叫過程、模型量化資訊、 Markdown 顯示效果與側邊欄顯示方式。

若是一般使用者，建議保留預設設定即可。

若是開發或測試模型，可開啟工具呼叫過程、完整模型名稱與模型量化資訊，方便觀察模型運作狀態。

llama5.png

8. Sampling 用於控制模型回答的隨機性與創意程度。

Temperature 數值越高，回答越有創意；數值越低，回答越穩定。

Top K、Top P、 Min P 則會影響模型選字範圍與輸出穩定性。
若用於客服、文件整理或正式回答，建議將 Temperature 設低一點。
若用於文案、腳本或創意發想，可將 Temperature 適度提高。
llama6.png

9. Penalties 用於降低模型重複輸出的情況。
此處可設定 Repeat last N、Repeat penalty、Presence penalty、 Frequency penalty 與 DRY 相關參數。
當模型出現一直重複句子、重複段落或回答過於單調時，可適度調整重複懲罰
參數。
一般情況下建議維持預設值，避免調整過高造成回答不自然。
llama7.png

10. Agentic 用於設定 AI 執行工具或代理流程時的最大回合數。
Agentic turns 代表模型最多可執行幾輪工具流程。
Max lines per tool preview 則是工具結果預覽最多顯示幾行。
此功能可避免模型在工具執行過程中無限循環，適合用於 MCP
工具、自動化任務與多步驟代理流程。
llama8.png

11. Developer 為進階設定頁面，主要提供開發者與系統測試人員使用。
此處可設定 KV cache、reasoning 解析、 thinking 功能、 raw output toggle 、JavaScript sandbox tool 與 Custom JSON。

若非開發用途，建議不要任意修改此頁設定，以免影響模型回覆結果或系統運作。

llama9.png

12. MCP設定頁面可調整MCP工具呼叫逾時時間。

Request timeout seconds代表每次MCP工具呼叫最多等待的秒數。

若工具需要較長時間處理資料，可將逾時秒數調高。

若只是一般工具呼叫，維持預設值即可。

llama10.png

13. Tools頁面會顯示目前可用工具。

若畫面顯示No tools available，代表目前尚未啟用任何工具。

當MCP Server或其他工具模組設定完成後，相關工具會顯示在此頁面，供模型進行呼叫使用。

llama11.png

14. Import / Export用於管理對話紀錄與系統設定。

Export conversations可匯出所有對話紀錄，方便備份或轉移資料。

Import conversations可匯入先前匯出的對話JSON檔案。

Delete all conversations可刪除所有對話紀錄，此操作無法復原，建議刪除前先完成備份。

Settings區域則可匯出或匯入系統設定，方便快速還原使用環境。

llama12.png