kuwa-taide-v.0.2.0-*.exe 2024/4/15

內建 TAIDE LX 7B 4-bit 量化版模型之 Kuwa GenAI OS v.0.2.0 安裝程式

參考：

https://kuwaai.org/

https://github.com/kuwaai/genai-os/releases

註：系統透過 llama.cpp 支援純 CPU 版，以及支援 Nvidia CUDA ver.12.1/12.2/12.3+的 GPU 版，請依您的硬體環境下載適合的版本。可在命令列執行 nvcc --version 檢查 CUDA 版本。
https://docs.nvidia.com/cuda/cuda-toolkit-release-notes/

說明：

This is a special Kuwa distribution for developers interested in the Taiwan's TAIDE model.

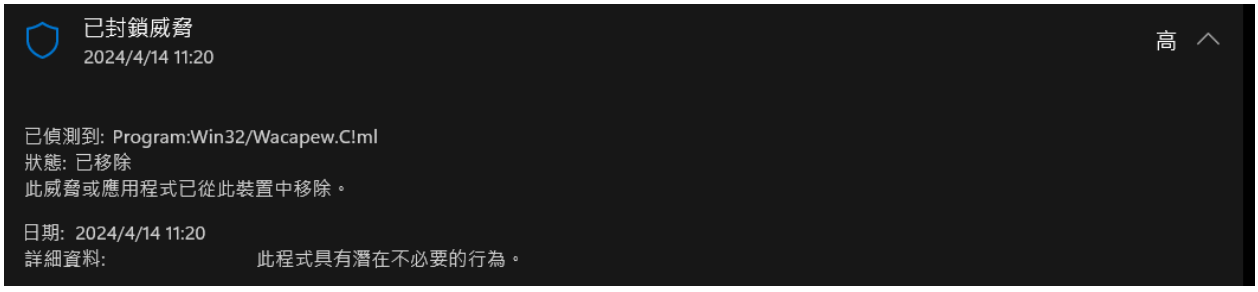
這是內建台灣 TAIDE 模型的 Windows 版 Kuwa 客製化系統。以下提醒：

1. 此客製化系統是針對已經成年並對 GenAI 有一定了解的系統開發者或使用者，提供簡易方便的安裝測試使用環境。
   * 硬體建議：至少 DRAM 8GB(建議 16GB+)的雙核心系統；擬在地端測試 RAG 者，建議 GPU 具有 VRAM 6GB+，視輸出入 Token 數而定。
   * 軟體建議：因版權限制，請先安裝微軟的 VC_redist.exe；若要使用 Nvidia 的 GPU，也請先安裝好 CUDA。載點如下，詳見 Kuwa 在 Github 上的說明。
     https://aka.ms/vs/17/release/vc_redist.x64.exe
     https://developer.nvidia.com/cuda-downloads
   * 網路需求：安裝過程中將透過網路下載所需要的第三方套件及模組，建議在具有穩定頻寬的網路環境下進行。
2. 此客製化系統已內建 TAIDE 模型，但考量算力限制，此為 TAIDE LX 7B 的量化簡化版，並取消過濾管制。因此此系統中模型的表現會與 TAIDE 原生模型有所差異，不代表國科會 TAIDE 及本系統的立場，還望諒察。
3. TAIDE 以台灣的文本翻譯、自動摘要、寫文章、寫信等任務為主。技術有其限制，AI 系統會產生幻覺，尤其可供訓練的繁體中文開放資料極為欠缺，TAIDE 的訓練資料及模型規模遠小於國外商用的模型，因此能力有其侷限。
4. Kuwa 系統及 TAIDE 模型仍還在持續開發及改善，難免會發生不穩。本系統生成的內容僅供參考，不擔保其正確性，仍需使用者再行查證；請勿將不適的對話內容公開，以免帶來不預期的困擾。
5. 此客製化系統亦支援串接 OpenAI ChatGPT 及 Google Gemini Pro，使用前請先依說明設定好全系統使用或個人使用的 API Keys。
6. 如遇到問題，歡迎到 Kuwa 的開源社群詢問。更進一步的資訊參見： https://kuwaai.org/

您必須已經成年並了解相關的說明及風險，才繼續進行安裝。全程約需 10 分鐘或更多，視網路頻寬及電腦效能而定，請耐心等候。

提醒：

掃毒軟體如微軟的 Defender 可能誤判檔案為惡意軟體，若您是從官網直接下載，您可放心使用。

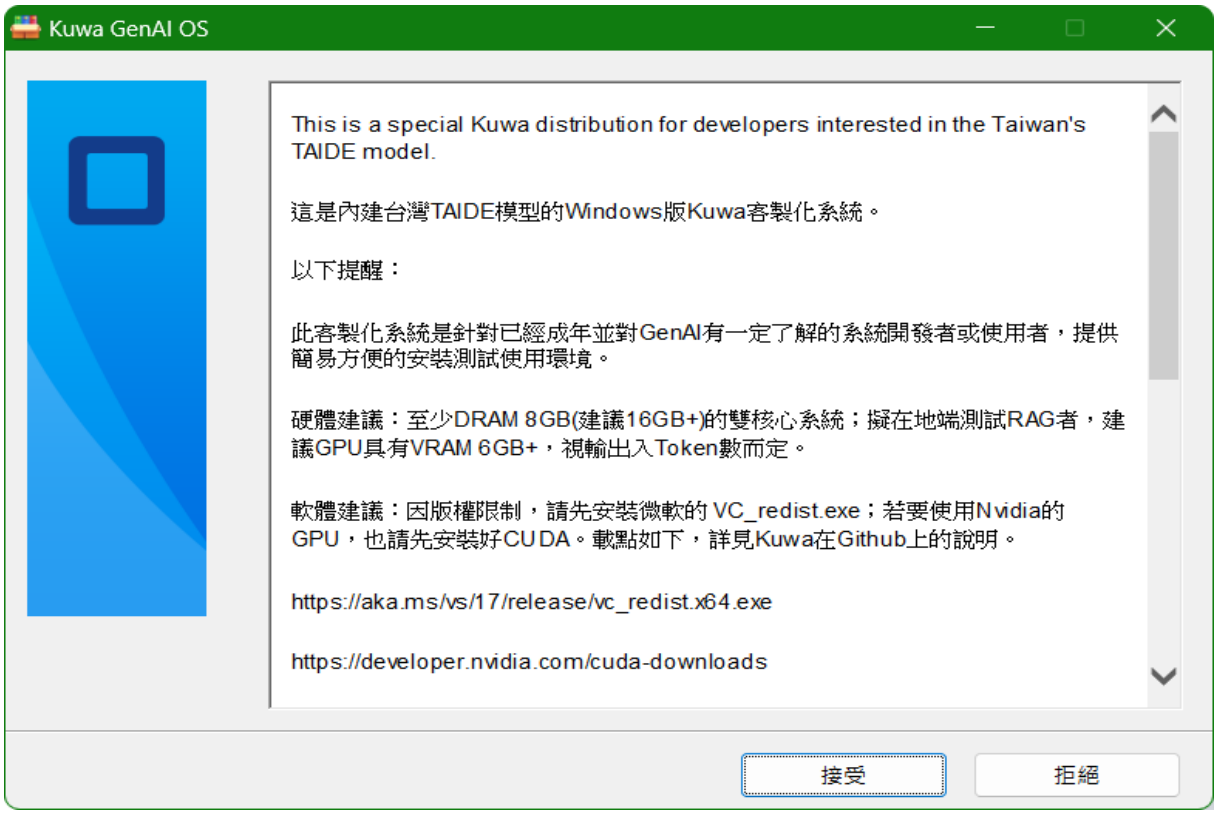


已知問題

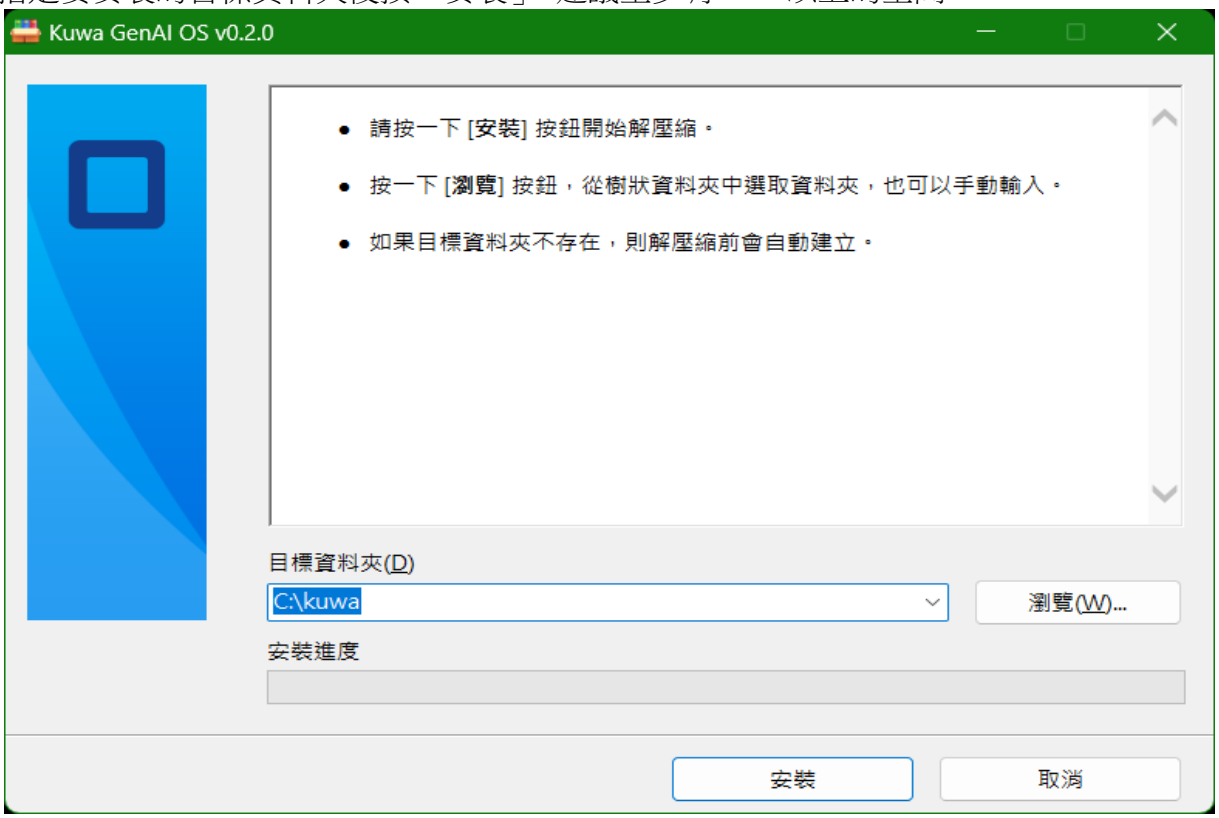
1. 目前 Windows 版的 Document QA 因為函式庫的相依性問題，可以讀取.doc 及.docx 格式的檔案，但可能無法讀取部份的.pdf 檔案。若有需要，請改用 Linux 版的 Kuwa。
2. RAG 相關應用因會產生較長的輸入，若僅使用 CPU 版串接地端模型時容易產生超時錯誤，建議串接雲端模型，或是利用 GPU 版串接地端模型再使用 RAG 應用。

安裝步驟

1. 開啟檔案開始執行，請先詳閱說明後按「接受」。

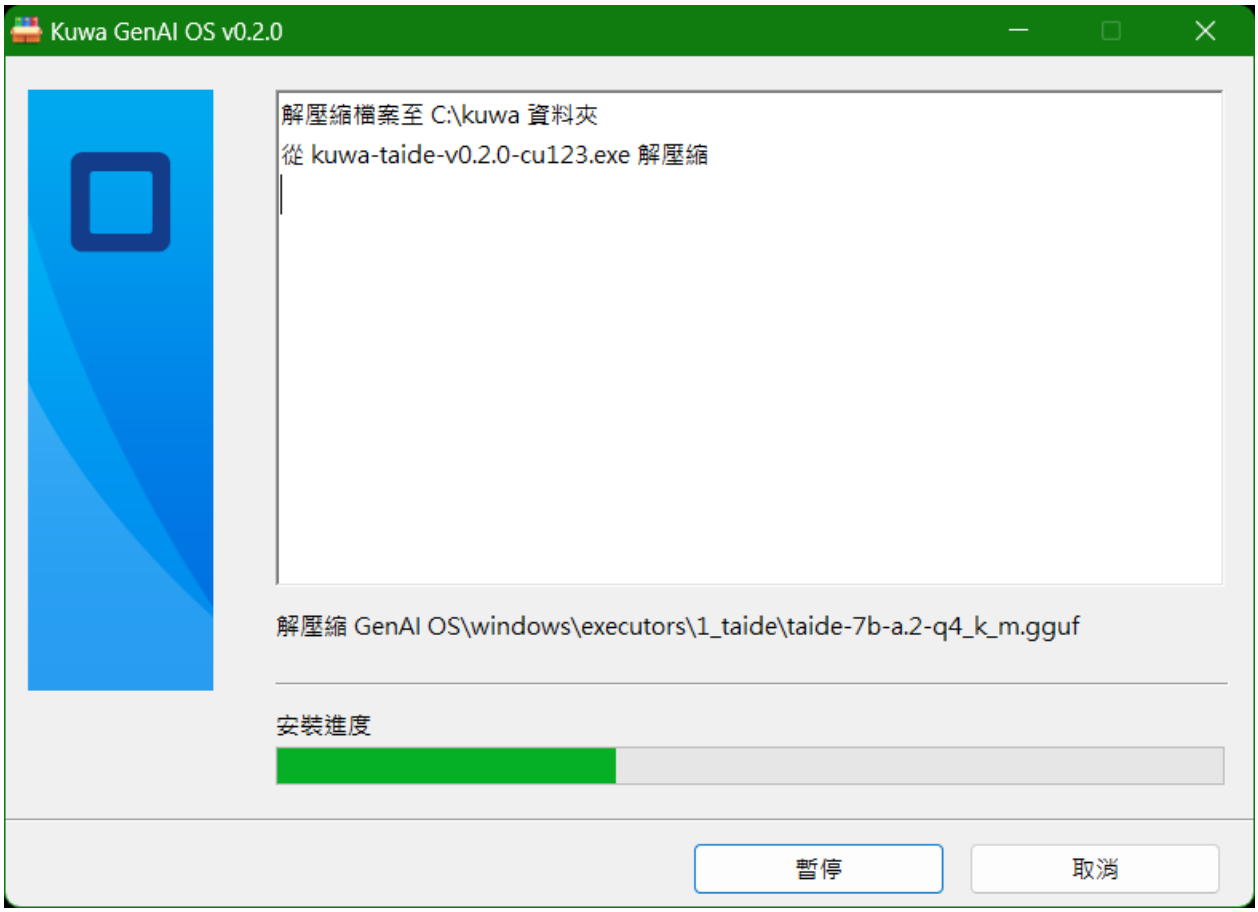


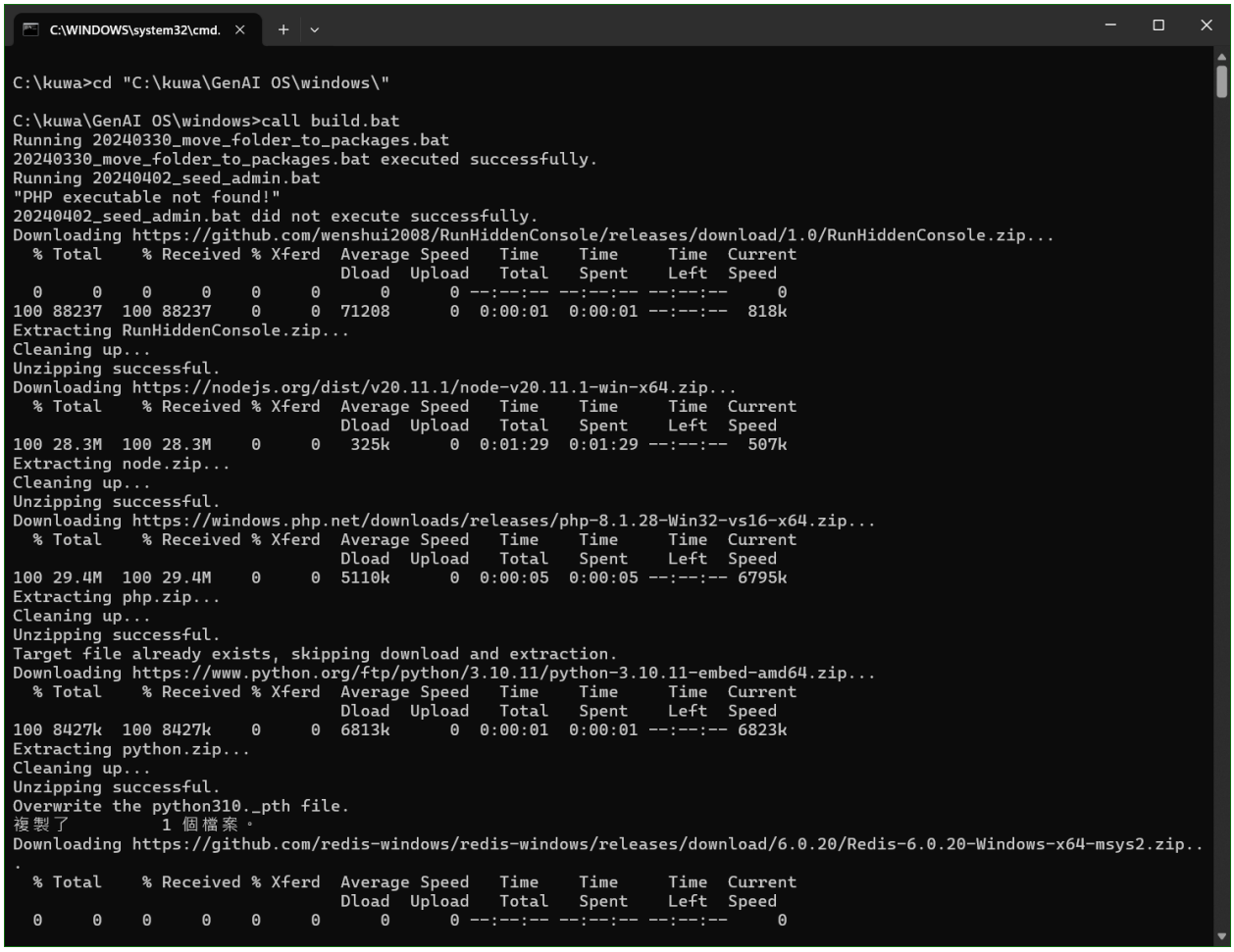
2. 指定要安裝的目標資料夾後按「安裝」。建議至少有 8GB 以上的空間。



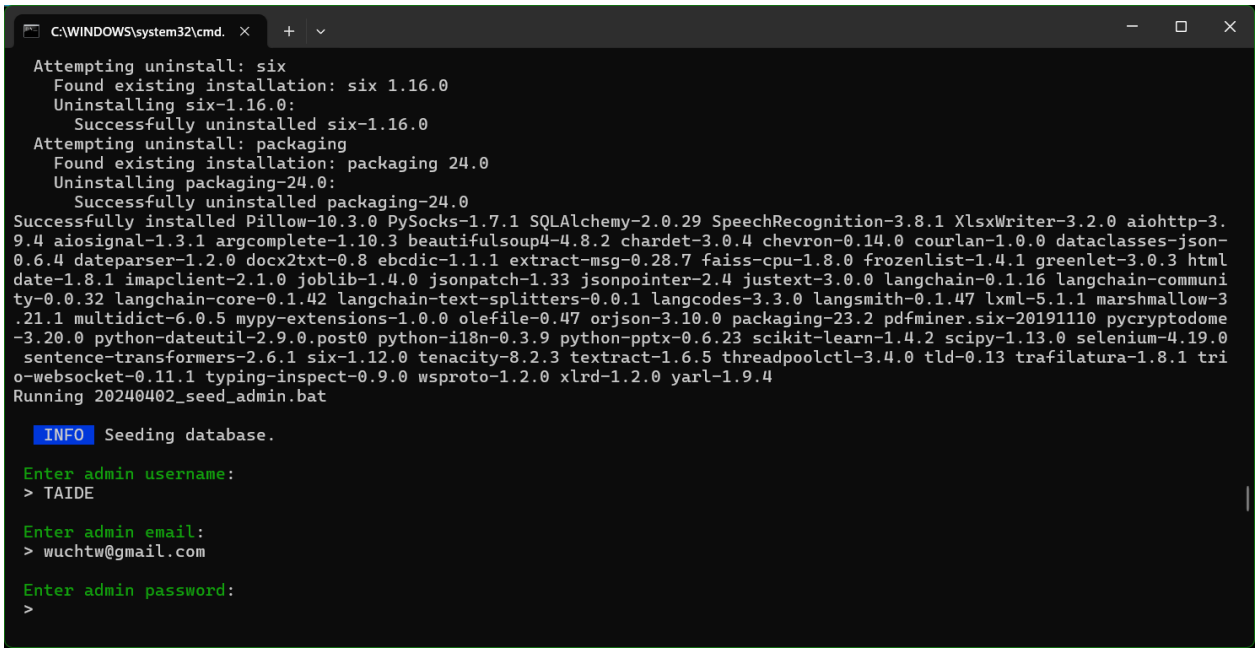
Windows 版相關套件將安裝於 kuwa\GenAI OS\windows

3. 安裝程式將開始解壓縮及初始設定環境，隨後開啟命令視窗執行自動下載相關套件及模組。
   - Windows 版相關套件將安裝於 kuwa\GenAI OS\windows 資料夾下的 packages；
     安裝程式參見 “build & start.bat” 或 build.bat
   - 全程約需 10 分鐘或更多，視網路頻寬及電腦效能而定，請耐心等候。

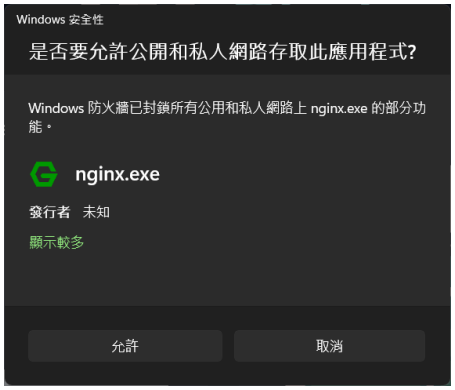




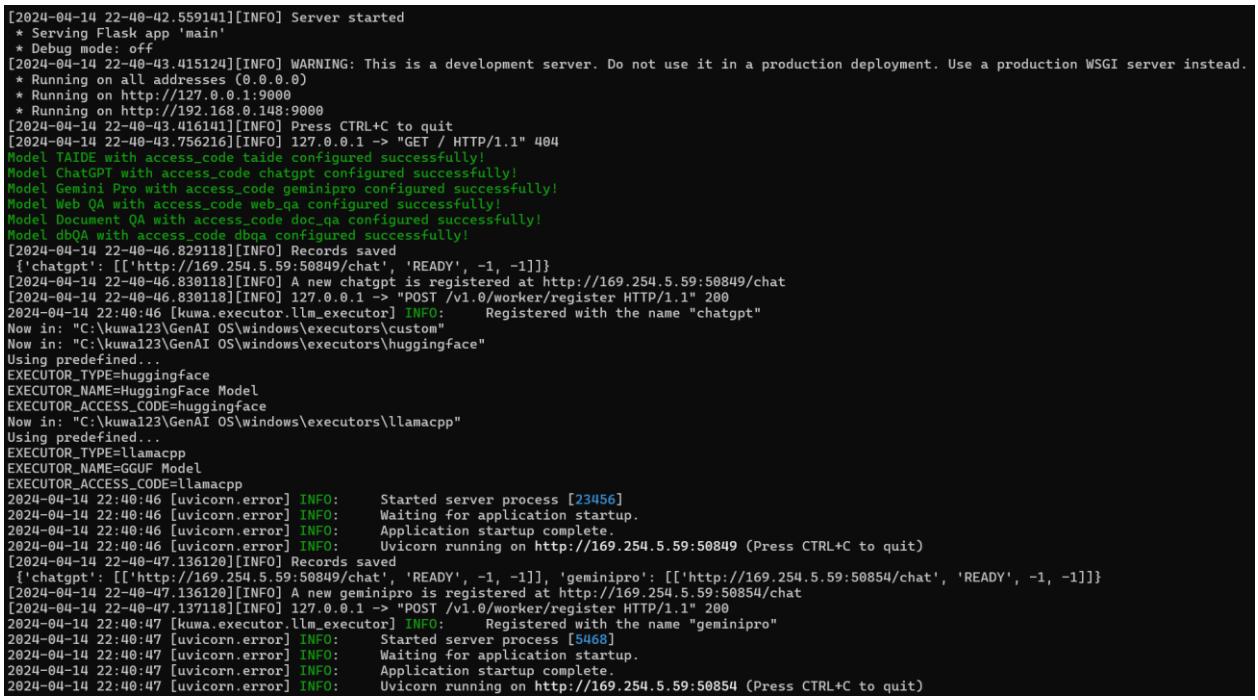
4. 下載及安裝後，會出現提示建立管理者帳號，請依序輸入 admin username (管理者的名稱，中文亦可)、admin email (管理者登入用的電子郵件帳號名稱)、admin password (管理者密碼，輸入時會隱藏)。
   - 如果錯過註冊帳號，可以稍後在視窗或執行 tool.bat 輸入 seed 指令來重新註冊管理者帳號。

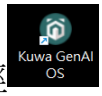


5. 輸入完成後，會詢問防火牆是否允許 Python 及 nginx.exe，請按「允許」，系統將開始啟動。

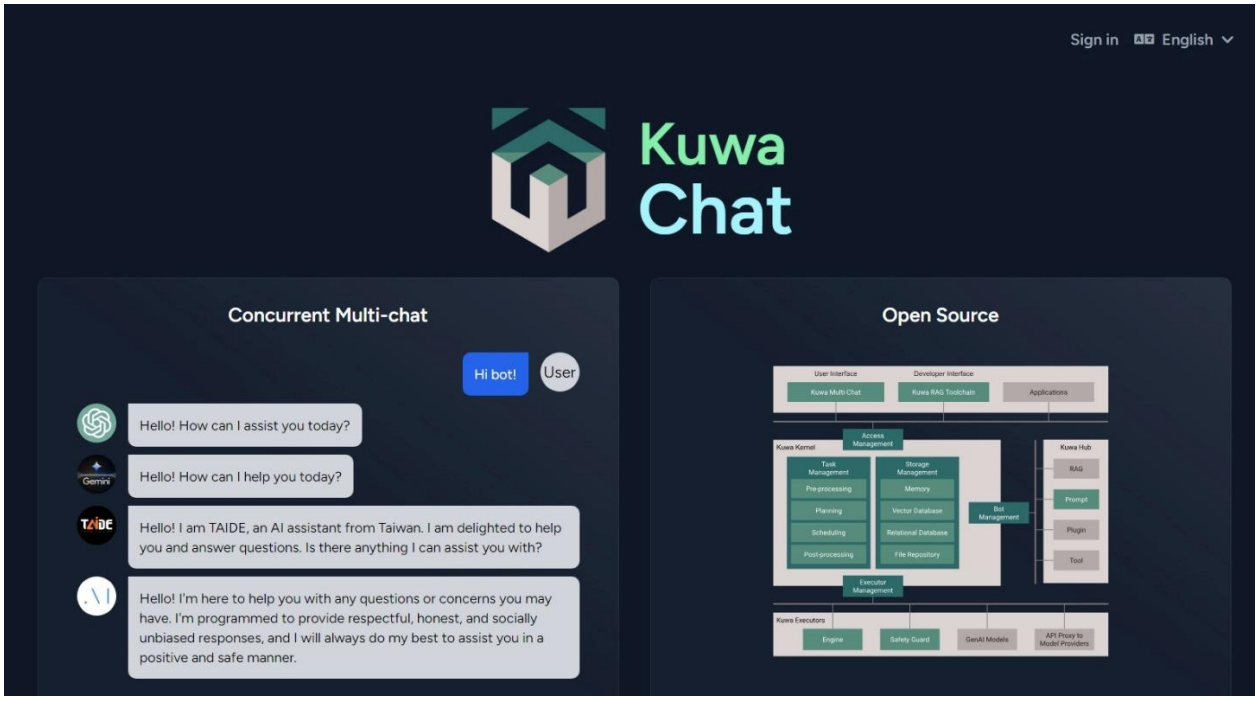


   - 畫面中間會出現各個 Model 的設定狀態，以及「llama_model_loader:」等相關訊息。



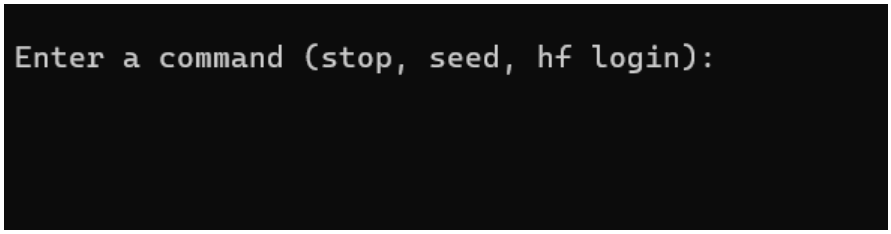
- 成功安裝設定完成後，會自動開啟預設瀏覽器並連到 Kuwa 系統 http://127.0.0.1/login
- 桌面將自動建立捷徑 

，後續也可以雙擊此圖案來開啟 Kuwa 系統

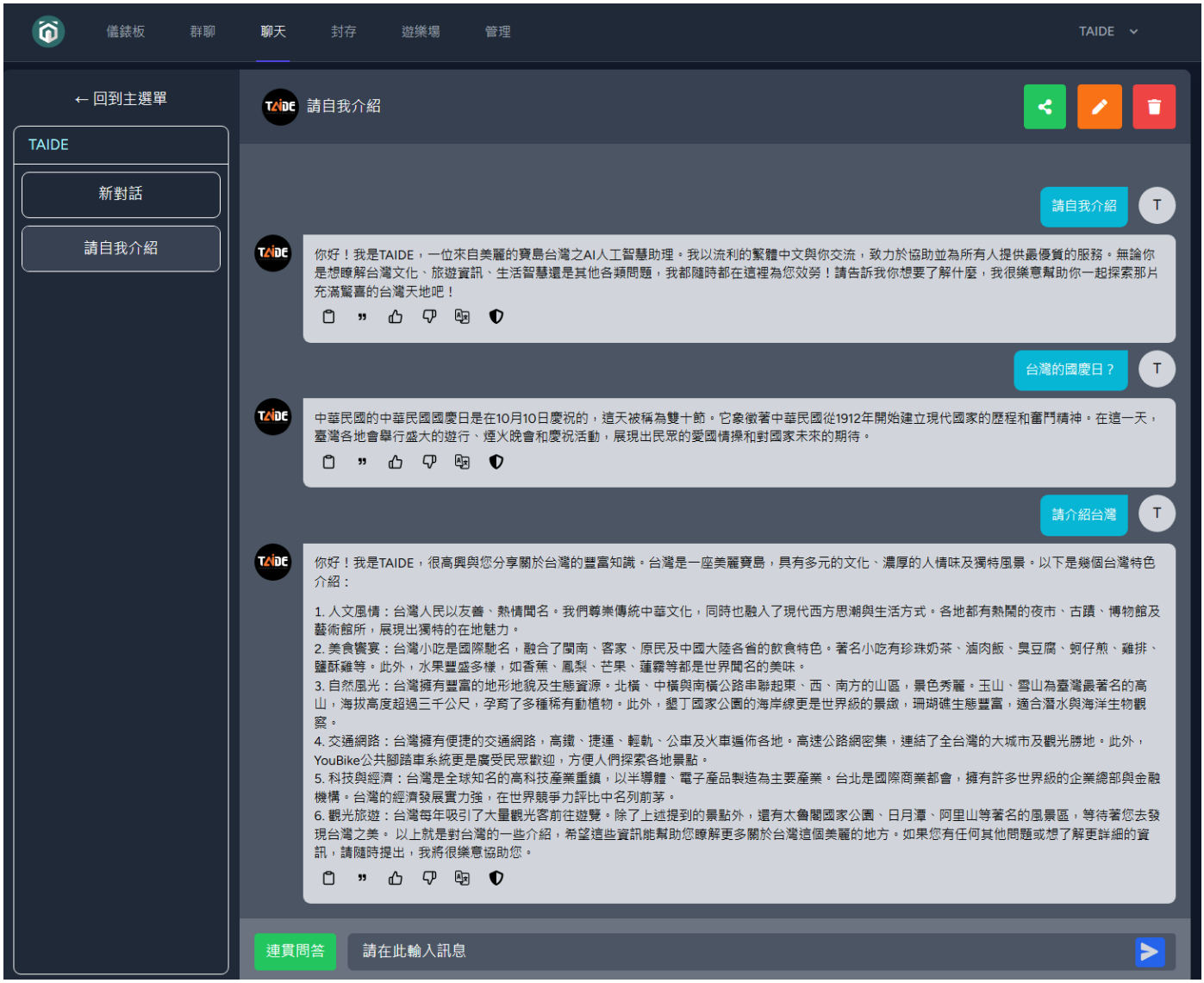


6. 按右上方「English」可改用「中文(台灣)」顯示。按「登入(或 Login)」，會出現登入畫面，請輸入安裝過程中設定的管理者 Email 及密碼，即可登入。

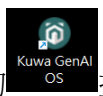
因為尚未完成 Email 設定，網頁上的「忘記密碼」尚未啟用。
若忘記密碼，可以在視窗按下 Enter，再輸入 seed 指令來設定新的、不重複的管理者帳號。



7. 此 Kuwa 客製化版本支援 TAIDE LX 7B 量化簡化版模型及相關 RAG，也可以從右上方點選「設定」去指定您的 ChatGPT 及 Gemini Pro 的 API Keys。您可以按「聊天」選取模型後開始使用。



8. 在命令列視窗按 Enter 後，可以輸入 stop 指令來結束 Kuwa 系統。

   後續可以再按桌面的 

捷徑來重啟 Kuwa，或直接執行 kuwa\GenAI OS\windows\start.bat 亦可。

Enjoy! 歡迎加入 Kuwa GenAI OS 的開源專案 https://kuwaai.org/
TAIDE 計畫詳見： https://taide.tw/