

Openwebui User Manual

Part 1: Setup and Getting Started

Before starting Open-WebUI, you need to already start the LLM model and embedding model through middleware.

1. Please download `open-webui.zip` and `setup_and_run.bat` and place them in the same directory.

 `open-webui` 2025/11/28 下午 01:54 壓縮的 (zipped) 資... 92,571 KB
 `setup_and_run` 2025/11/28 下午 03:45 Windows 批次檔案 27 KB

2. Double-click `setup_and_run.bat` to launch Open-WebUI.

3. The first time you launch Open-WebUI, it will unzip `open-webui.zip`, install Python, and set up the environment. Then you can use Open-WebUI in your browser.

4. Upload files to Knowledge. We have prepared some sample questions in the `open-webui` folder. Please wait until the LLM finishes processing.



The screenshot illustrates the workflow for adding knowledge to the system:

- Top Left:** The main navigation bar includes tabs for Model, 知識 (Knowledge), Prompts, and Tools. The 知識 tab is highlighted with a red box and has a red arrow pointing to its + icon.
- Top Right:** A modal window titled "建立知識" (Create Knowledge) is open. It contains fields for "您現在的工作是什麼？" (What is your current work?) with "Test" entered, and "您正在試圖完成什麼？" (What are you trying to accomplish?) with "Test" entered. A red arrow points from the main interface's + icon to this modal's + icon.
- Middle:** The "建立知識" modal shows settings for visibility ("私有" - Private) and groups ("選擇群組" - Select Group). A red box highlights the "建立知識" button, which also has a red arrow pointing to it from the main interface's + icon.
- Bottom Left:** A sidebar for "搜尋集合" (Search Collection) includes options: "上傳檔案" (Upload File), "上傳目錄" (Upload Directory), "同步目錄" (Sync Directory), and "新增文字內容" (Add Text Content). A red box highlights the + icon, with a red arrow pointing to it from the main interface's + icon.
- Bottom Right:** A list of uploaded files from "Alpha Unlimited Inc...." with their sizes: 336.6 KB, 301.1 KB, 266.8 KB, 242.8 KB, and 116.8 KB. A red box highlights the + icon at the top of this list, with a red arrow pointing to it from the main interface's + icon.

Finished

Part 2: Running Inference + RAG

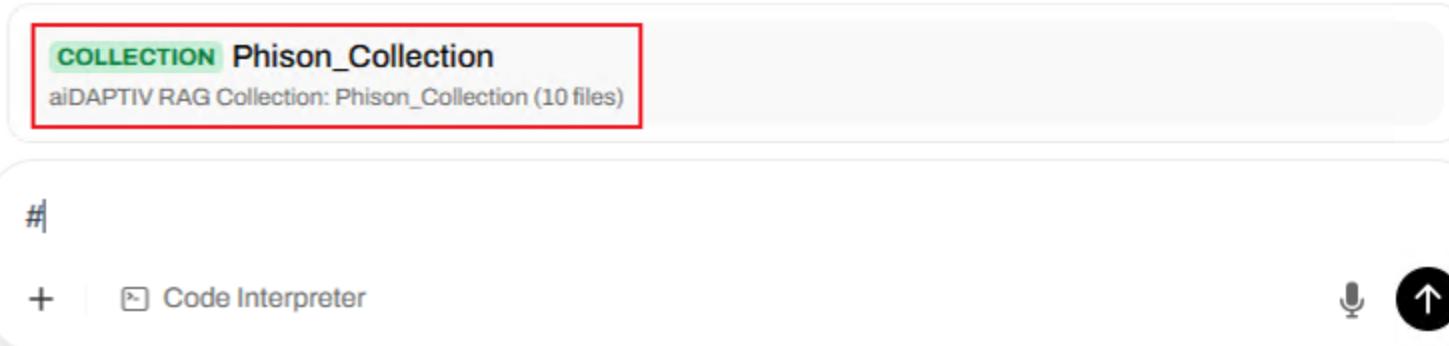
Query Methods

1. **Normal Chat:** Directly enter your question in the chat box (will not perform collection retrieval)

2. **Agent Chat:**

(1) Type <# hashtag symbol> in the chat box and click the collection you create.

(2) Enter your question after the hashtag



Verifying Execution Flow

Please confirm the following indicators:

1. **Time To First Token (TTFT) is between 2 ~ 8 seconds**

2. **RAG reference documents are displayed below the response**

A screenshot of the AI response. At the top right, there is a 'Collection' icon with the text 'Alpha Collection'. Below it, a message asks 'How many days of full-pay sick leave does the company provide annually?'. The response starts with 'C:\Users\phison\Desktop\AgentBuilderClient\Installer_0.2.1\package\inference_model\Meta-Llama-3.1-8B-...'. A red box highlights the time information: 'Time to first token: 2.32 s'. The response continues with 'The provided content does not mention the number of days of full-pay sick leave the company provides annually.' and 'Total Time: 5.21 s'. A red box highlights the reference document ID: 'a7b45d63fd154d03b670bcb96cccf596_Alpha Unlimited Inc. ...'. Below the response are several small icons for editing and sharing.