

# Openwebui User Manual

## Part 1: Installation

Before starting Open-WebUI, make sure you have already started the LLM model and embedding model through middleware.

1. Please download [Open-webui\\_installation.zip](#)

2. Unzip Open-webui\_installation.zip , you will get open-webui.zip and setup\_and\_run.bat 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

3. Double-click setup\_and\_run.bat to launch Open-WebUI.

4. 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.

## Part 2: Setting Parameters for LLM, Embedding Model, Knowledge Management, and Open-WebUI

1. If you have already set it up, the settings will be stored in config.txt. You can directly use it. The batch file will verify that the settings still work.

```
=====
Environment Configuration
=====

[INFO] Found config.txt in current directory

Do you want to load configuration from config.txt? (y/n):
```

2. Setting the LLM model endpoint:

You can see the model information in the "models" endpoint, and provide the endpoint with version as the LLM URL.

```
===== LLM Model Setting =====

LLM_URL (Hint: The endpoint you start the LLM Model. e.g. http://127.0.0.1:13141/v1):
```

```
← → C ① 127.0.0.1:13141/models
美化列印 □ LLM_URL = http://127.0.0.1:13141/v1

{
  "object": "list",
  "data": [
    {
      "id": "C:\\\\Users\\\\user\\\\Desktop\\\\models\\\\Llama-3.2-3B-Instruct-Q4_K_M.gguf",
      "object": "model",
      "created": 1765354735,
      "owned_by": "llamacpp",
      "meta": {
        "vocab_type": 2,
        "n_vocab": 128256,
        "n_ctx_train": 131072,
        "n_embd": 3072,
        "n_params": 3212749888,
        "size": 2011539712
      }
    }
  ]
}
```

3. Setting the Embedding model endpoint for Knowledge Management:

You can see the model information in the "models" endpoint, and provide the endpoint with version as the Embedding URL.

```
===== Embedding Model Setting =====

EMBEDDING_URL (Hint: The endpoint you start the Embedding Model. e.g. http://127.0.0.1:13142/v1):
```

```
← → C ① 127.0.0.1:13143/models
美化列印 □ EMBEDDING_URL = http://127.0.0.1:13143/v1

{
  "object": "list",
  "data": [
    {
      "id": "C:\\\\Users\\\\user\\\\Desktop\\\\models\\\\Qwen3-Embedding-0.6B-Q8_0.gguf",
      "object": "model",
      "created": 1765352912,
      "owned_by": "llamacpp",
      "meta": {
        "vocab_type": 2,
        "n_vocab": 151669,
        "n_ctx_train": 32768,
        "n_embd": 1024,
        "n_params": 595776512,
        "size": 633205056
      }
    }
  ]
}
```

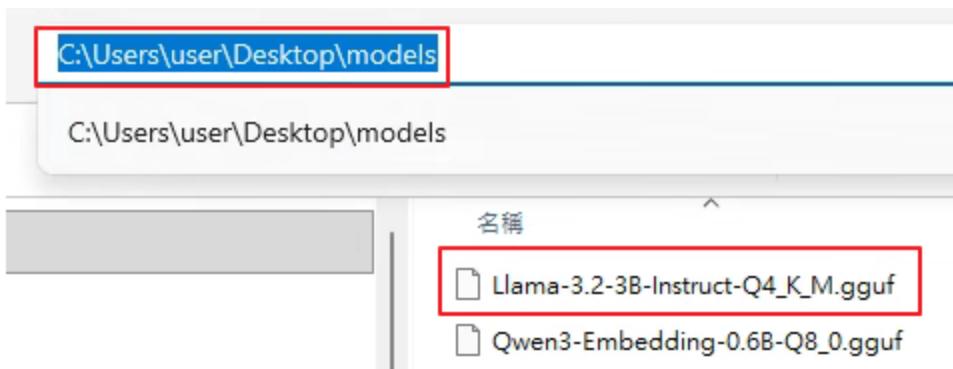
4. Setting tokenizer model for Knowledge Management:

```
===== KM Setting =====

MAX_TOKENS_PER_GROUP (Hint: The token length each group, should be smaller than context size of LLM. Advice using 80% of context size of LLM. e.g. 13000): 13000

LLM_GGUF (Hint: The file name of GGUF LLM weight file. e.g. Meta-Llama-3.1-8B-Instruct-Q4_K_M.gguf):

LLM_MODEL_DIR (Hint: The path of GGUF LLM weight file.. e.g. C:\\Program Files\\\\models): |
```



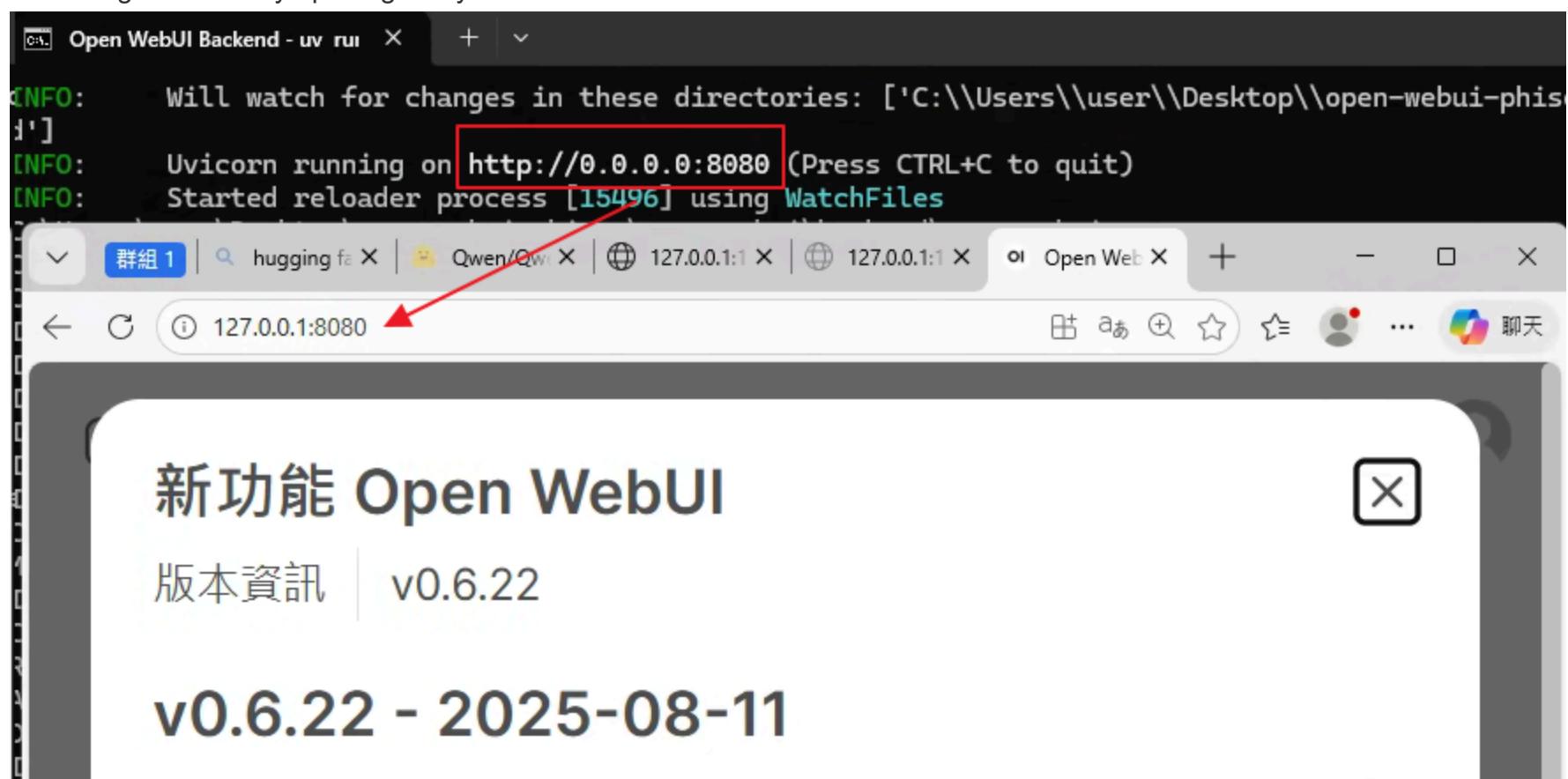
LLM\_GGUF = Llama-3.2-3B-Instruct-Q4\_K\_M.gguf  
LLM\_MODEL\_DIR = C:\Users\user\Desktop\models

## Part 3: Getting Started

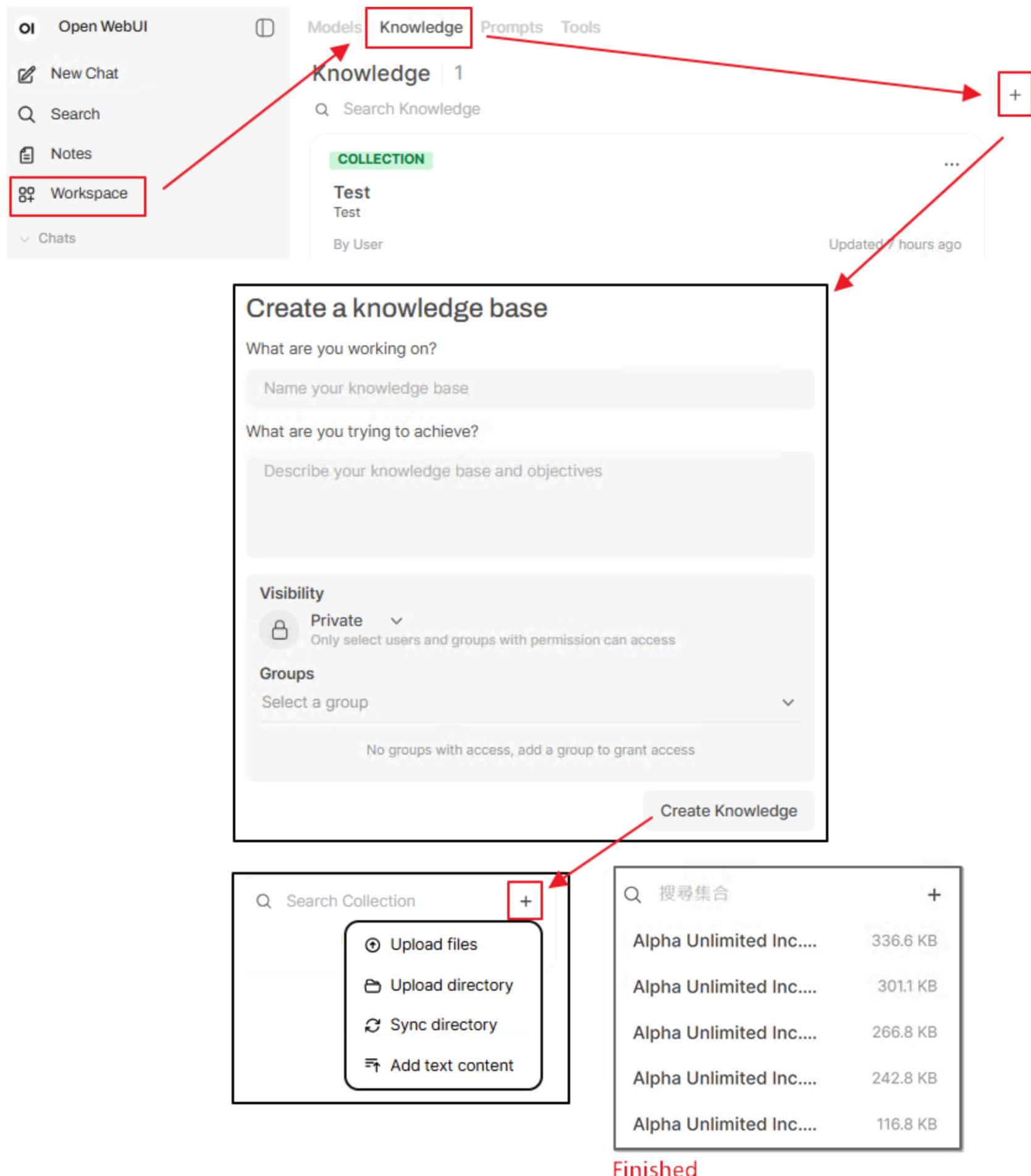
- When Open-WebUI is ready, you will see the log like the following image.

```
Open WebUI Backend - uv ruu
...
INFO:     Started server process [21956]
INFO:     Waiting for application startup.
2025-12-10 16:42:55.292 | INFO      | open_webui.utils.logger:start_logger:162 - GLOBAL_LOG_LEVEL: INFO
2025-12-10 16:42:55.293 | INFO      | open_webui.main:lifespan:530 - Installing external dependencies of functions and tools...
2025-12-10 16:42:55.316 | INFO      | open_webui.utils.plugin:install_frontmatter_requirements:241 - No requirements found in frontmatter.
```

- You can get started by opening it in your browser.



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



## Part 4: 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

COLLECTION Phison\_Collection

aiDAPTIV RAG Collection: Phison\_Collection (10 files)

#

+

Code Interpreter

Microphone icon

Upload icon

### 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



How many days of full-pay sick leave does the company provide annually?

OI C:\Users\phison\Desktop\AgentBuilderClient\Installer\_0.2.1\package\inference\_model\Meta-Llama-3.1-8B-...

● Time to first token: 2.32 s

The provided content does not mention the number of days of full-pay sick leave the company provides annually.

Total Time: 5.21 s

a7b45d63fd154d03b670bcb96cccf596\_Alpha Unlimited Inc. ...

