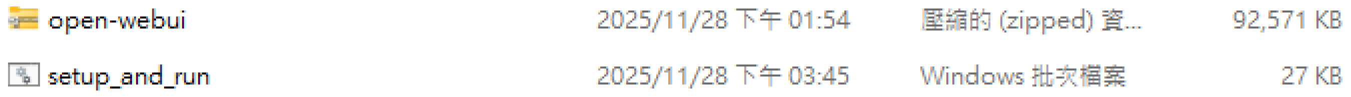
Openwebui User Manual

📦 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.

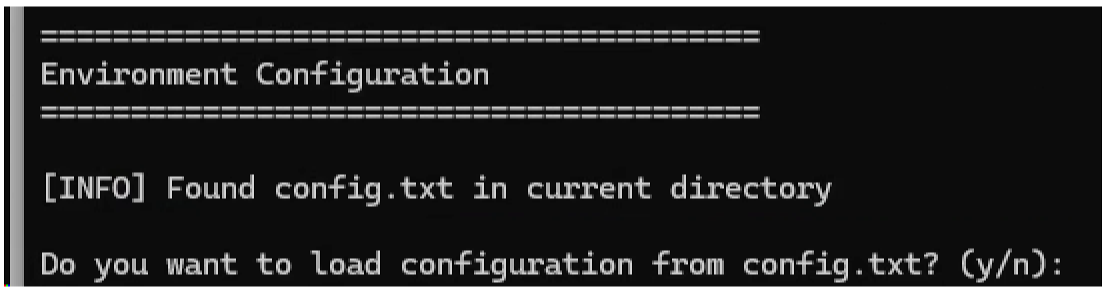


open-webui	2025/11/28 下午 01:54	壓縮的 (zipped) 資...	92,571 KB
setup_and_run	2025/11/28 下午 03:45	Windows 批次檔案	27 KB

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.

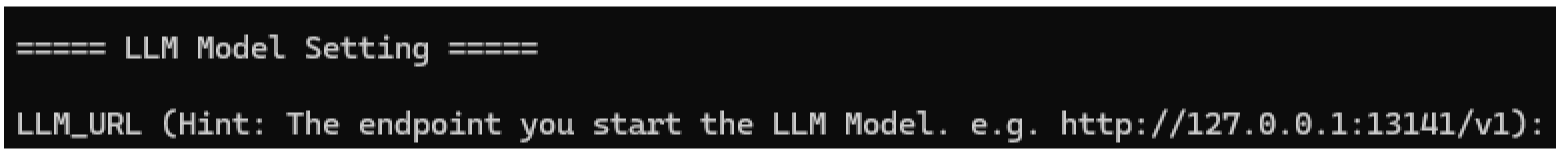
⚙️ Part 2: Setting Parameters for LLM, Embedding Model, Knowledge Management, and Open-WebUI

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.

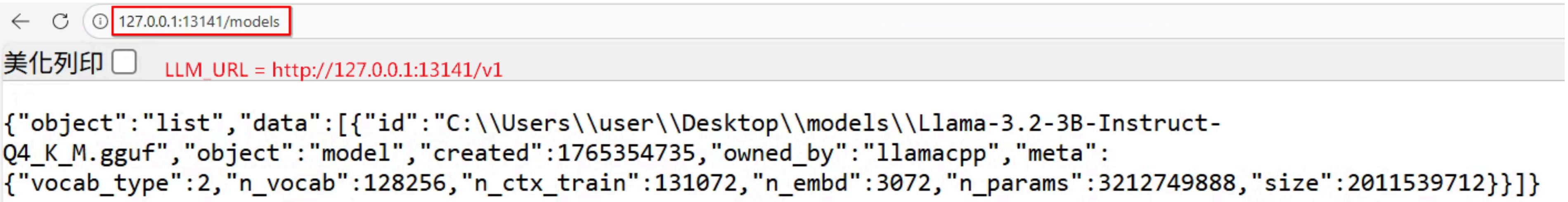


```
========================================
Environment Configuration
========================================

[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):
```

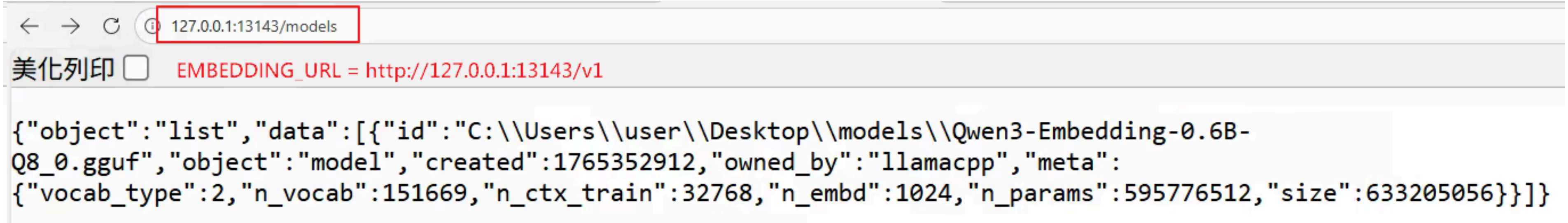


127.0.0.1:13141/models

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

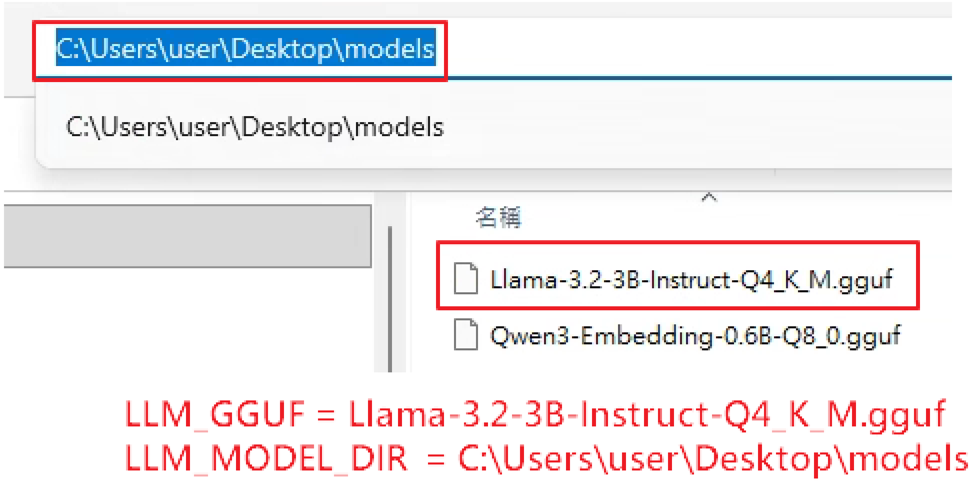


127.0.0.1:13143/models

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 Knowledge Management:



C:\Users\user\Desktop\models

C:\Users\user\Desktop\models

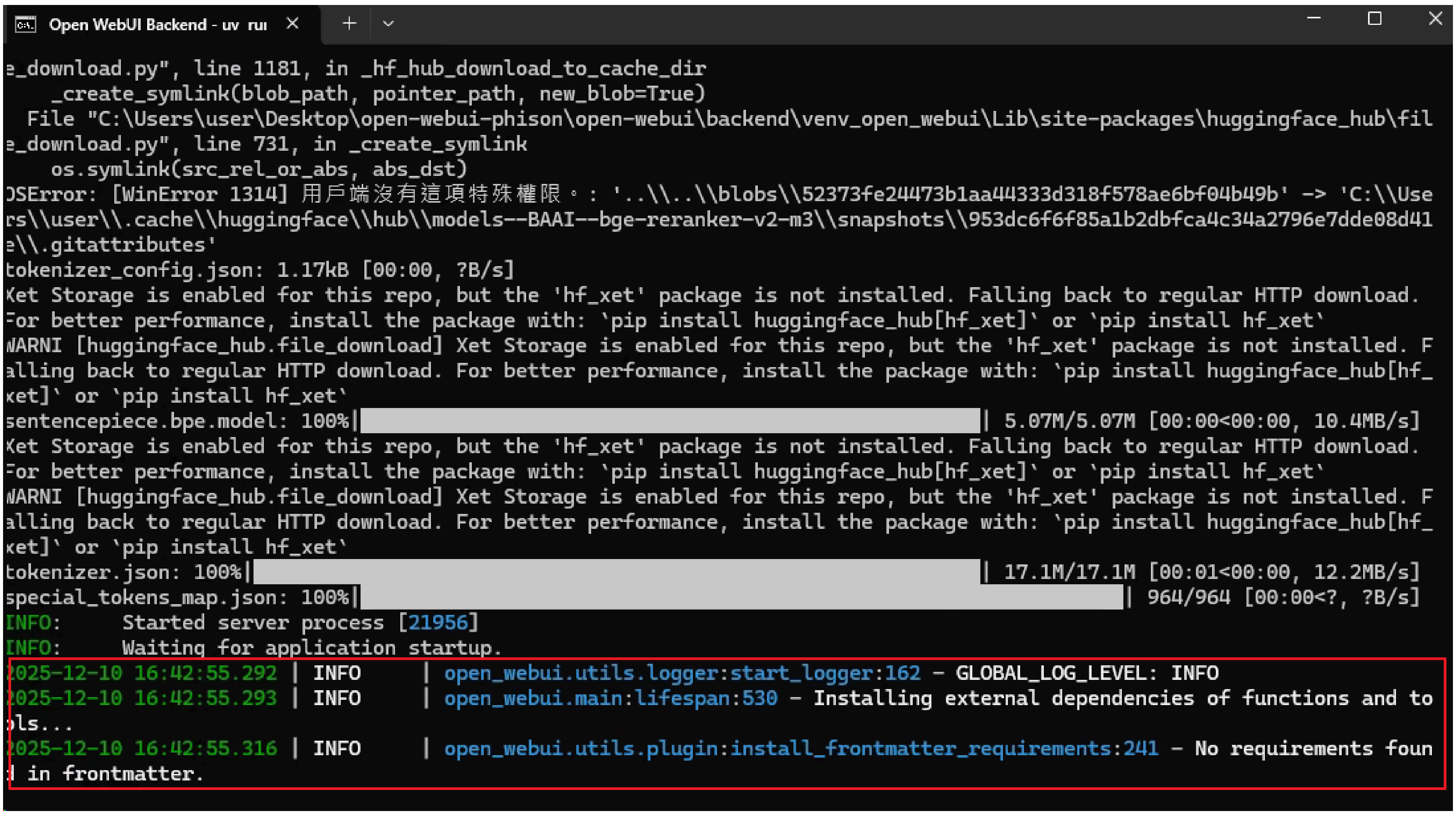
Llama-3.2-3B-Instruct-Q4_K_M.gguf

Qwen3-Embedding-0.6B-Q8_0.gguf

LLM_GGUF = Llama-3.2-3B-Instruct-Q4_K_M.gguf
LLM_MODEL_DIR = C:\Users\user\Desktop\models

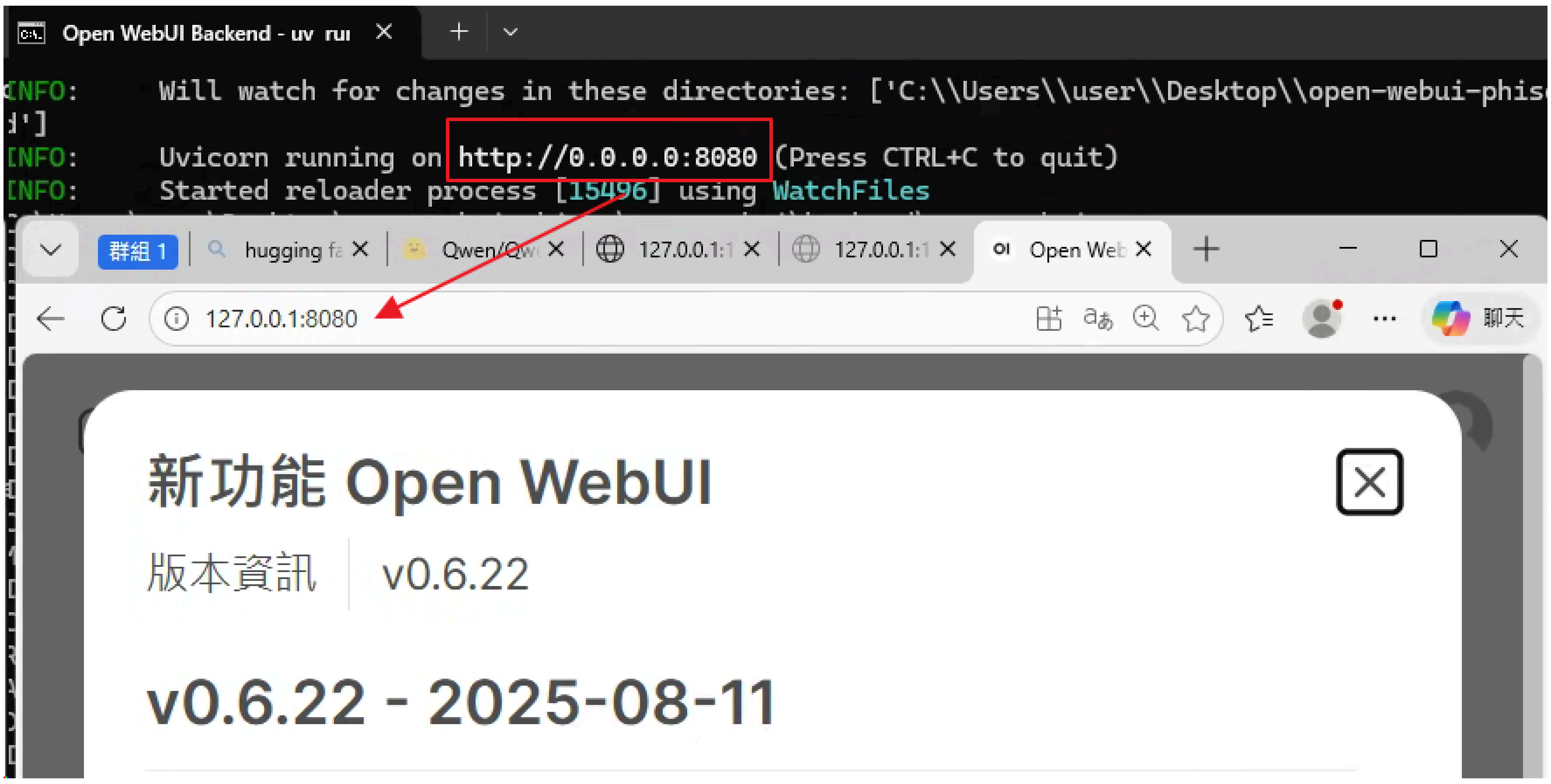
🎯 Part 3: Getting Started

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



```
e_download.py", line 1181, in _hf_hub_download_to_cache_dir
    _create_symlink(blob_path, pointer_path, new_blob=True)
  File "C:\Users\user\Desktop\open-webui-phison\open-webui\backend\venv_open_webui\Lib\site-packages\huggingface_hub\fil
e_download.py", line 731, in _create_symlink
    os.symlink(src_rel_or_abs, abs_dst)
OSError: [WinError 1314] 用戶端沒有這項特殊權限。: '..\\..\\blobs\\52373fe24473b1aa44333d318f578ae6bf04b49b' -> 'C:\\Use
rs\\user\\.cache\\huggingface\\hub\\models--BAAI--bge-reranker-v2-m3\\snapshots\\953dc6f6f85a1b2dbfca4c34a2796e7dde08d41
e\\.gitattributes'
tokenizer_config.json: 1.17kB [00:00, ?B/s]
Xet Storage is enabled for this repo, but the 'hf_xet' package is not installed. Falling back to regular HTTP download.
For better performance, install the package with: `pip install huggingface_hub[hf_xet]` or `pip install hf_xet`
WARNI [huggingface_hub.file_download] Xet Storage is enabled for this repo, but the 'hf_xet' package is not installed. F
alling back to regular HTTP download. For better performance, install the package with: `pip install huggingface_hub[hf_
xet]` or `pip install hf_xet`
sentencepiece.bpe.model: 100%|████████████████████████████████| 5.07M/5.07M [00:00<00:00, 10.4MB/s]
Xet Storage is enabled for this repo, but the 'hf_xet' package is not installed. Falling back to regular HTTP download.
For better performance, install the package with: `pip install huggingface_hub[hf_xet]` or `pip install hf_xet`
WARNI [huggingface_hub.file_download] Xet Storage is enabled for this repo, but the 'hf_xet' package is not installed. F
alling back to regular HTTP download. For better performance, install the package with: `pip install huggingface_hub[hf_
xet]` or `pip install hf_xet`
tokenizer.json: 100%|██████████████████████████████████████████| 17.1M/17.1M [00:01<00:00, 12.2MB/s]
special_tokens_map.json: 100%|██████████████████████████████████████████| 964/964 [00:00<?, ?B/s]
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 to
ols...
2025-12-10 16:42:55.316 | INFO     | open_webui.utils.plugin:install_frontmatter_requirements:241 - No requirements foun
d in frontmatter.
```

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



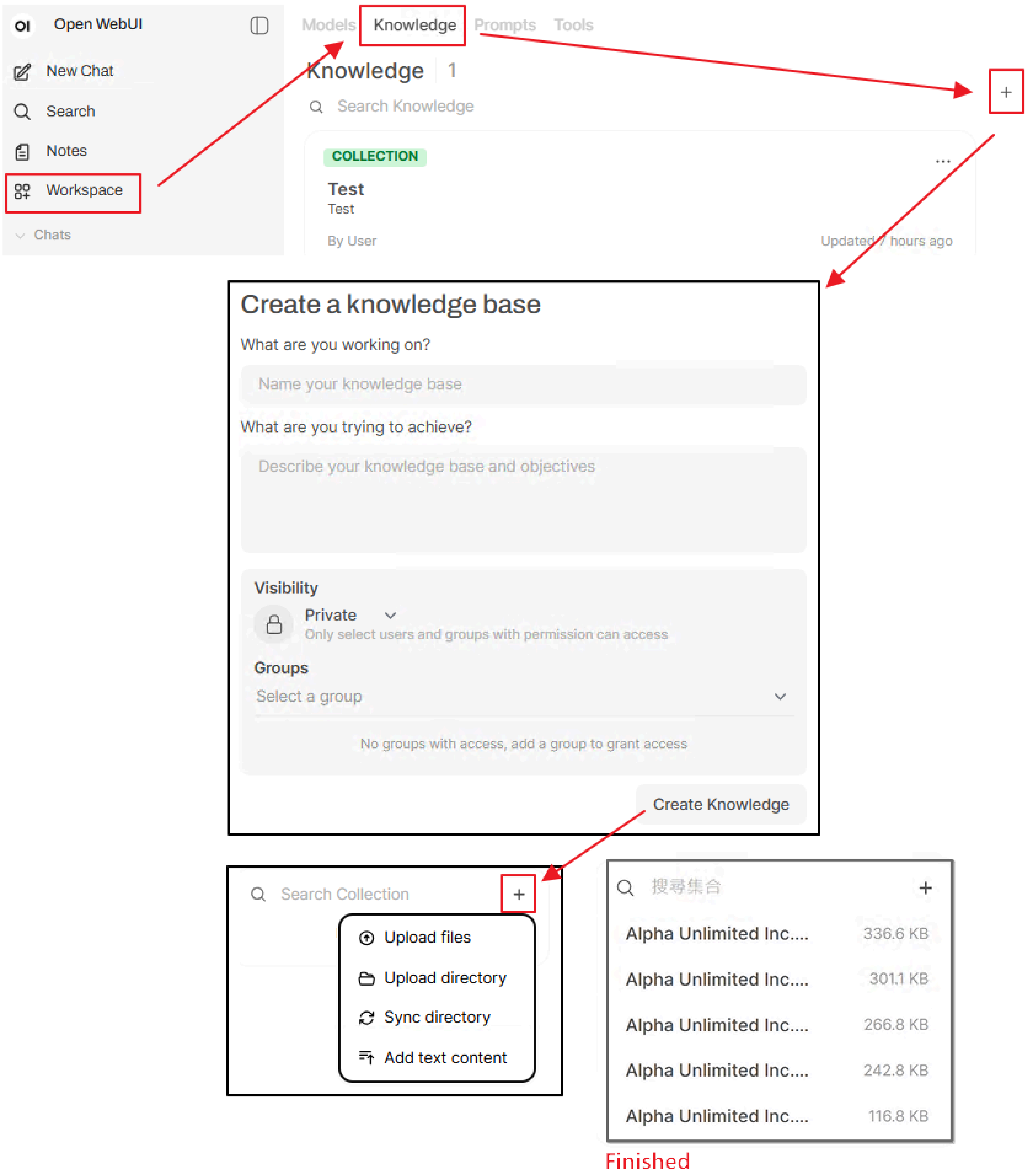
```
INFO:     Will watch for changes in these directories: ['C:\\Users\\user\\Desktop\\open-webui-phis
d']
INFO:     Uvicorn running on http://0.0.0.0:8080 (Press CTRL+C to quit)
INFO:     Started reloader process [15496] using WatchFiles
```

新功能 Open WebUI

版本資訊 | v0.6.22

v0.6.22 - 2025-08-11

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



Create a knowledge base

What are you working on?

Name your knowledge base

What are you trying to achieve?

Describe your knowledge base and objectives

Visibility

Private

Only select users and groups with permission can access

Groups

Select a group

No groups with access, add a group to grant access

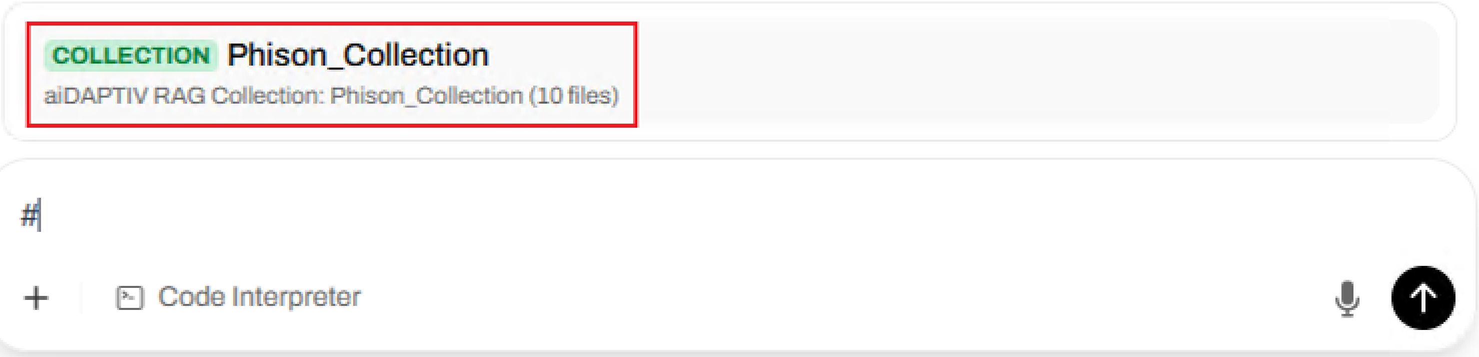
Alpha Unlimited Inc....	336.6 KB
Alpha Unlimited Inc....	301.1 KB
Alpha Unlimited Inc....	266.8 KB
Alpha Unlimited Inc....	242.8 KB
Alpha Unlimited Inc....	116.8 KB

Finished

🤖 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

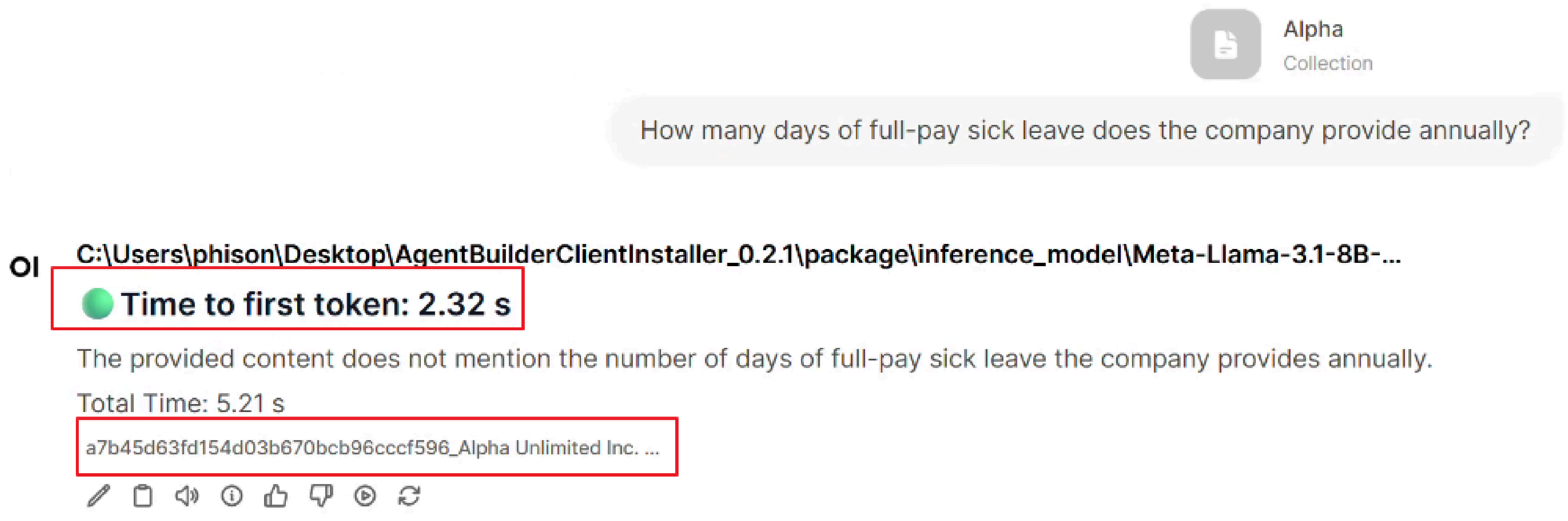

COLLECTION Phison_Collection
aiDAPTIV RAG Collection: Phison_Collection (10 files)

#

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?

C:\Users\phison\Desktop\AgentBuilderClientInstaller_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. ...