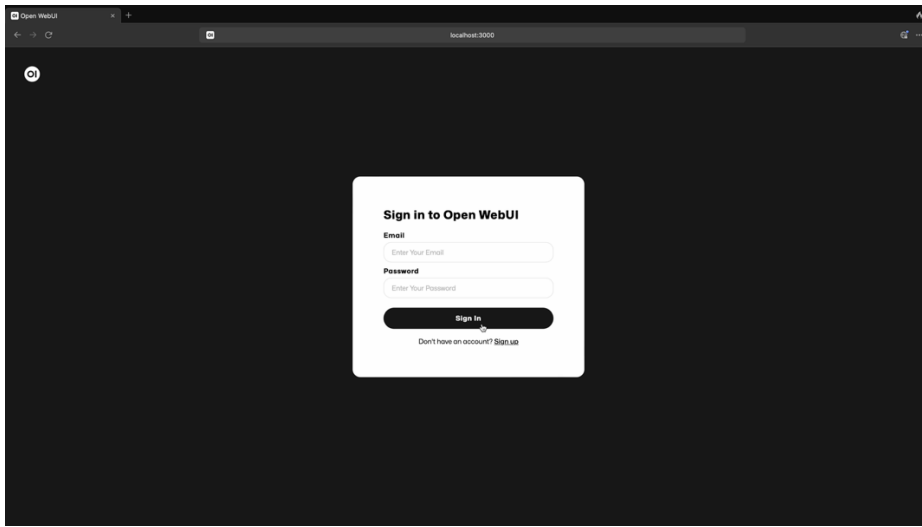


How to implement Generative AI Models



Ollama + WebUI Frameworks



The framework used is based on:

- Text Generation (Ollama + WebUI)
- Image Generation (Automatic + Comfy)
- RAG - Retrieval-Augmented Generation (RAG).

1

Models

An AI (*Artificial Intelligence*) model is a computing system designed to perform actions that normally require human intelligence. These include the capability to understand and generate natural language (NLP - *Natural Language Processing*), generating and translating text, identifying and generating images, predictive analysis and much more.

The various AI models differ from each other primarily in the following features:

- **Architecture**
 - The internal structure of the model that determines how data is processed;
- **Dimensions**
 - The number of parameters in the model, that influence its learning and accuracy capabilities;
- **Data and training techniques**
 - The type of data used to train the model and methods used to improve its performance;
- **Efficiency and scalability**
 - The model's capacity to use computational resources efficiently and to adapt to large datasets.

Ollama + WebUI Frameworks contain the following preconfigured models:

	TYPE ^[1]	USE	PARAMETERS ^[2]	WEBSITE
GEMMA	LLM	chat	9 Billion	ollama.com/library/gemma
LLAMA 3	LLM	chat	8 Billion	ollama.com/library/llama3
LLAMA 2	LLM	chat	7 Billion	ollama.com/library/llama2
MISTRAL	LLM	chat	7 Billion	ollama.com/library/mistral
PHI3	LLM	chat	3 Billion	ollama.com/library/phi3

1. LLM (Large Language Model).

2. Dimensions can vary over time.

Settings

Access the framework's general settings as follows:

- click on your profile icon shown on the bottom left;
- select **Settings** from the options;
- this will take you to **Settings** where you can work in the following areas:

General

Manage the theme and the language you want to use for the framework from here.

Connections

Manage OpenAI and Ollama APIs from here.

Models

For managing Ollama models.

Interface

Allows you to customize the WebUI interface by activating add-ons.

Personalization

Allows you to customize the iterations with LLM models.

Audio

Configure SST (Speech-to-Text) and TTS (Text-to-Speech) motors and choose the voice you want to use.

Images

Choose which model to set as the default motor for generating images.

Chats

Enable/disable chat history, import, export, archive or delete chats from this section.

Account

Manage your profile picture, change name and password from this section.

About

Current versions are listed and any available updates can be checked here.

Admin Panel - Add User

Create a new user as follows:

- click on your profile on the bottom left;
- select **Admin Panel** from the options offered;
- under **All Users** a list of all user profiles that have been created will be displayed;
- click on the + button on the top right;
- this will lead to the **Add User** option;
- choose to create a single user (**Form** tab) or upload a CSV file containing more than one user (**CSV Import** tab);
- if you have selected Form, you will be required to provide the following details:
 - o **Role** – the role you want to assign to the user (Pending, User or Admin);
 - o **Name** – the name you want to give the user;
 - o **Email** – the user's email, which will be the username that the user will need to sign in to the framework;
 - o **Password** – the password the user will use to sign in to the framework.

3

Admin Panel - Admin Settings

Access the admin settings as follows:

- click on your profile on the bottom left;
- select **Admin Panel** from the options;
- under **All Users** a list of all user profiles that have been created will be displayed;
- click on the washer-shaped icon on the top right;
- this will take you to **Admin Settings**;
- this area contains the following tabs:
 - o **General**
 - o **Users**
 - o **Database**
 - o **Banners**
 - o **Pipelines**

Prompts

A prompt is the input that guides the iteration between the user and the AI system. It is the text or the instruction given to the system.

Creating a prompt

A prompt can be created as follows:

- click on **Workspace** in the left menu;
- in the **Workspace** section select **Prompt**;
- to the right of the Search Documents area, click on **+**;
- enter **Title**, **Command** and **Prompt Content**.

Import and export a prompt

A prompt can be imported or exported as follows:

- click on **Workspace** in the left menu;
- in the **Workspace** section, select **Prompt**;
- on the right, under the Search Documents area, are two buttons:
 - o **Import Prompts**;
 - o **Export Prompts**.

Documents

Upload a document

Upload a document as follows:

- click on **Workspace** in the left menu;
- in the **Workspace** section, select **Documents**;
- to the right of the Search Documents area, click on **+**;
- the **Add Docs** level will open where you can select a document to upload by pressing on **Click here to select documents**;
- when you upload a document, you can assign one or more tags to it by clicking **+ Add Tags**.

Tags

Tags are used to categorize uploaded documents and to allow them to be searched for when requested via chat.

Document Settings

Set up the documents that have been uploaded as follows:

- click on **Workspace** in the left menu;
- in the **Workspace** section, select **Documents**;
- click on **Document Settings** on the top right;
- in **Document Settings**, you can work on the following areas:

General

- **General settings**

- **Scan for documents from /data/docs**
Scans documents found in the /data/docs directory.
- **Embedding Modal Engine**
You can change the embedding motor, the model that creates vector representations of chunks and then saves them in a vector database.
- **Hybrid Search**
- **Embedding Modal**
- **Reset Vector Storage**

You can delete all documents from the /data/docs space.

Chunk Params

- **Chunk Params**

- **Chunk Size**
Allows you to set a maximum number of tokens for each single chunk.
- **Chunk Overlap**
Shows by how many tokens the chunks must overlap.
- **PDF Extract Images (OCR)**

Query Params

- **Query Params**

Top k shows the number of the most relevant documents from those analysed.

- **PDF Extract Images (OCR)**

This is the prompt, which you can change to behave in different ways according to use.

Web Params

- **Web Loader Settings**
- **Bypass SSL verification Websites**
- **Youtube Loader Settings**
- **Language**

Chunk

A chunk represents a block of data or information which is treated as if it is a single unit, that can be managed separately. Dividing data or information into chunks makes it easier to process, archive and search.

If the data being processed was a book, chunks could be considered to be like chapters, pages, paragraphs, sentences, etc.

In the case of NLP (Natural Language Processing), chunks are counted in tokens that, according to the AI model being used, correspond to words, characters or subwords.

Chunk Size

Represents the number of tokens, of a text document (DOC, PDF, TXT, etc.), counted to create a vector representation, that will subsequently be saved on the database by the process of embedding.

In NLP, a suitable Chunk Size could be between 1,000 and 10,000 tokens.

In other ML (Machine Learning) contexts, it could vary significantly.

Chunk Overlap

Indicates by how many tokens the chunks must overlap to guarantee contextual continuity and coherence between data.

In NLP, a suitable Chunk Overlap could be between 5% and 20% of the Chunk Size.

In other ML contexts, this could vary significantly.

There are no standard settings for Chunk Size and Chunk Overlap values, but these change according to datasets, AI models and hardware resources. The values quoted here are purely intended as examples and may not be appropriate for specific case studies.

Glossary

AI (Artificial Intelligence)

Artificial intelligence (AI) is an advanced form of technology that allows computational systems to replicate the processes of human intelligence. By implementing sophisticated and complex algorithms, AI can perform activities that would traditionally require human intervention, such as learning, reasoning, problem solving and understanding natural language.

6

Chunk

A chunk represents a block of data or information which is treated as if it is a single unit, that can be managed separately. Dividing data or information into chunks makes it easier to process, archive and search.

CUDA (Compute Unified Device Architecture)

A parallel processing hardware architecture developed by NVIDIA. It allows developers to use GPU (Graphics Processing Units) calculating capabilities to perform intensive computing tasks, significantly speeding up performance compared to using traditional CPUs (Central Processing Units).

Deep learning

Deep learning is the research field on automated learning and artificial intelligence that is based on different levels of representation, corresponding to hierarchies of features of factors or concepts, where high level concepts are defined on the basis of lower ones.

Embedding

A technique used in the field of automated learning and natural language processing (NLP) for representing complex data, such as words, sentences or objects in a vector space.

Fine-tuning

Fine-tuning represents a machine learning technique that leads to further training a pre-trained model on a specific new dataset for a particular task. This method is especially useful for large models such as, for example, Large Language Models (LLM), that have previously been trained on large volumes of generic data.

LLM (Large Language Model)

This AI model is designed to understand and generate natural language. It is noted for having a large number of parameters that are trained on huge collections of texts. Thanks to its size, it is capable of acquiring a deep knowledge of linguistic structures and contexts. It represents a major development in the processing of natural language, improving human-computer interaction more naturally and efficiently. Due to its complexity, it requires significant computational resources both for training and implementation.

ML (Machine Learning)

A subset of artificial intelligence (AI) that focuses on the development of algorithms and models that enable computers to learn from data, make predictions or take decisions. Unlike traditional programs that require explicit instructions to perform specific tasks, machine learning systems automatically improve their performance over time by means of data analysis. ML allows computers to hone their skills on specific tasks through experience gained from data.

NLP (Natural Language Processing)

Natural language processing is a field of artificial intelligence (AI) that focuses on the interaction between a computer and human language. The aim is to enable computers to understand, interpret and generate natural language in a meaningful and useful way.

RAG (Retrieval Augmented Generation)

The process of optimizing the output of a large-scale language model to ensure that it uses a reliable knowledge base beyond its own training data sources before generating a response.

STT (Speech-to-Text)

Technology that enables the spoken word to be converted into written text, allowing computers and other devices to understand and process spoken human language.

TTS (Text-to-Speech)

Technology that enables written text to be converted into spoken word using advanced ML techniques.