



Configuring Open WebUI for AI Interaction

Configuring Open WebUI for AI Interaction

WHAT?

Open WebUI is a Web-based application that allows you to interact with AI models.

WHY?

To configure Open WebUI according to both user and administrator requirements and preferences.

EFFORT

To understand how to configure selected parts of Open WebUI takes less than 1 hour.

REQUIREMENTS

The procedures in this article assume that the following requirements are already fulfilled:

- You have successfully deployed the SUSE AI Factory.
- You can access the Open WebUI interface in your Web browser.
- Your Open WebUI user account has been created, and you have received the login credentials.
- You have navigated your Web browser to the URL of your instance of Open WebUI and logged in with your credentials.
- Selected tasks require the Open WebUI administrator privileges.

Publication Date: 2026-08-05

<https://documentation.suse.com> 

Contents

1	Introduction	1
1.1	What are the key benefits of using Open WebUI?	1
1.2	How does an AI chatbot process user queries?	1
1.3	Known Issues	2
	User password reset	2
	• Audio processing stops after several minutes	3
1.4	For more information	3
2	Setting base AI models	4
2.1	Downloading AI models from Ollama	4
2.2	Downloading AI models from Hugging Face	6
2.3	Using AI models from OpenAI-compatible providers	9
3	Setting the default system prompt	12
3.1	Setting a system prompt for the current chat	12
3.2	System prompt for the current user	12
3.3	System prompt for an AI model	13
	For more information	14
4	Setting the default AI model	15
4.1	AI model for the current chat	15
4.2	Default system AI model	16
5	Managing custom AI models	17
5.1	Creating custom AI models	17

5.2	Modifying existing AI models	18
	Enabling the default MCP services	19
5.3	Cloning existing AI models	20
5.4	Importing exported models	20
5.5	Exporting existing AI models	21
5.6	Exporting all existing AI models	21
5.7	Deleting existing AI models	21
6	Managing user accounts	23
6.1	User roles in Open WebUI	23
	The 'pending' role	23 • The 'user' role 23 • The 'admin' role 24
6.2	Creating user accounts	26
	Creating a single user account	26 • Creating multiple user accounts from a CSV list 27
6.3	Assigning roles to existing users	27
6.4	Modifying existing user accounts	28
6.5	Revoking access to Open WebUI	29
7	Managing a knowledge base	30
7.1	Creating a knowledge base	30
7.2	Deleting a knowledge base	31
7.3	Adding content to a knowledge base	32
7.4	Removing content from a knowledge base	34
8	Managing system prompts	36
8.1	Creating a system prompt	36
8.2	Modifying existing system prompts	37
8.3	Cloning existing system prompts	37

- 8.4 Importing exported system prompts 38
- 8.5 Exporting existing system prompts 38
- 8.6 Exporting all existing system prompts 39
- 8.7 Deleting existing system prompts 39

9 Authenticating via Okta 40

- 9.1 Configuring Okta 40
- 9.2 Gathering data for Open WebUI configuration 42
- 9.3 Configuring Open WebUI for Okta integration 43

A Glossary 46

B Copyright 51

C GNU Free Documentation License 52

- C1 0. PREAMBLE 52
- C2 1. APPLICABILITY AND DEFINITIONS 52
- C3 2. VERBATIM COPYING 54
- C4 3. COPYING IN QUANTITY 54
- C5 4. MODIFICATIONS 55
- C6 5. COMBINING DOCUMENTS 57
- C7 6. COLLECTIONS OF DOCUMENTS 57
- C8 7. AGGREGATION WITH INDEPENDENT WORKS 58
- C9 8. TRANSLATION 58
- C10 9. TERMINATION 58
- C11 1. FUTURE REVISIONS OF THIS LICENSE 59
- C12 ADDENDUM: How to use this License for your documents 59

1 Introduction

Open WebUI is a Web-based interface that allows you to interact with AI models, such as *large language models* (LLM). It simplifies working with AI by providing a graphical user interface (GUI) instead of relying on command-line tools. The goal of Open WebUI is to make AI accessible to a broader audience, including those without technical expertise in AI development.

1.1 What are the key benefits of using Open WebUI?

Interaction with AI models

You can load a preferred AI model and interact with it using prompts in your natural language.

Custom data handling for RAG workflows

Open WebUI allows you to upload documents or specify URLs of pages that the AI can parse and reference during interaction, which is particularly useful for RAG workflows.

User-friendly interface

By providing a graphical user interface, Open WebUI makes AI tools more accessible to non-technical users. It helps them use the power of advanced AI models with no need for technical knowledge.

Data privacy

You can deploy Open WebUI locally or in a private environment. Such deployment is required for users concerned about privacy and data security.

1.2 How does an AI chatbot process user queries?

When you enter a user prompt, several processes happen in the background to generate the response.

1. **Input processing.** The AI first processes the text of your prompt to understand its meaning. This involves identifying the subject, intent and any details or context provided. This process is called **Natural Language Understanding** (NLU).
2. **Contextual analysis.** If you are interacting with AI in a session where you have already asked previous questions, the AI considers the context of the conversation. This results in more relevant and coherent answers.

3. **Knowledge retrieval.** The AI retrieves information from its pre-trained knowledge base. This database includes facts, data and concepts that the AI has been trained on. AI models can also utilize retrieval-augmented generation (RAG) systems to get contextual information from the data provided by the organization. If the AI has access to real-time data, it can search for the latest information online to provide an up-to-date response.
4. **Response generation.** Using natural language generation (NLG) techniques, the AI constructs a coherent and grammatically correct response based on the information it retrieved.
5. **Output.** The AI delivers the response in a requested form, usually in a human-readable format. This can be a detailed explanation, a step-by-step guide, or a customized image, depending on the complexity of your query.
6. **Feedback Loop (Optional).** In specific AI systems, your feedback or follow-up questions can help refine future responses, allowing the AI to improve its answers over time.

1.3 Known Issues

This section lists the known issues with the Open WebUI and their solutions.

1.3.1 User password reset

An Open WebUI user can only reset their own password if they are logged in to Open WebUI and have an active Web browser session. However, if the user has lost their password and cannot log in to Open WebUI, the password must be reset by an SUSE Rancher Prime: RKE2 administrator using the `kubectl` command.

1. Generate a new password hash using the `htpasswd` command. `htpasswd` is part of the `apache2-utils` package.

```
> htpasswd -bnBC 10 "NEW_PASSWORD" | tr -d ':\n'
```

2. Identify the Pod where the Open WebUI application is running.

```
> kubectl get pod -n SUSE_AI_NAMESPACE | grep open-webui | grep -v pipelines
```

3. Update the password.

```
> kubectl exec open-webui-0-n SUSE_AI_NAMESPACE -- \
```

```
python3.11 -c 'import sqlite3; con = sqlite3.connect("/usr/lib/python3.11/site-packages/data/webui.db"); \ncon.cursor().execute("UPDATE auth SET password=\"PASSWORD_HASH\" WHERE email=\"admin@example.org\")'; con.commit()'
```

1.3.2 Audio processing stops after several minutes

When the voice recorder is used as described in the section [Chat window](https://documentation.suse.com/suse-ai/1.0/html/openwebui-using/index.html#openwebui-chat-window-usage) (<https://documentation.suse.com/suse-ai/1.0/html/openwebui-using/index.html#openwebui-chat-window-usage>), Open WebUI tries to process the audio, but it times out without any warning and no text is generated. The API endpoint `/audio/api/v1/transcriptions`, which receives the audio to be processed, results in a 504 Gateway Timeout.

1.4 For more information

- The upstream Open WebUI documentation is hosted at <https://docs.openwebui.com/>.

2 Setting base AI models

Open WebUI needs at least one base AI model configured so that you can interact with it or create custom AI models as described in [Chapter 5, Managing custom AI models](#). You can download models from repositories such as Ollama or Hugging Face, or use models via OpenAI-compatible APIs.

2.1 Downloading AI models from Ollama

Ollama (<https://ollama.com/> ) is an online repository that hosts open source AI models. This procedure describes how to download and use these models from the Open WebUI interface.

REQUIREMENTS

- You must have Open WebUI administrator privileges to access configuration screens or settings mentioned in this section.
 1. In the bottom left of the Open WebUI window, click your avatar icon to open the user menu and select *Admin Panel*.
 2. Click the *Settings* tab and select *Models* from the left menu.
 3. In the top-right corner of the screen, click the small "download" icon to open the *Manage Models* screen.

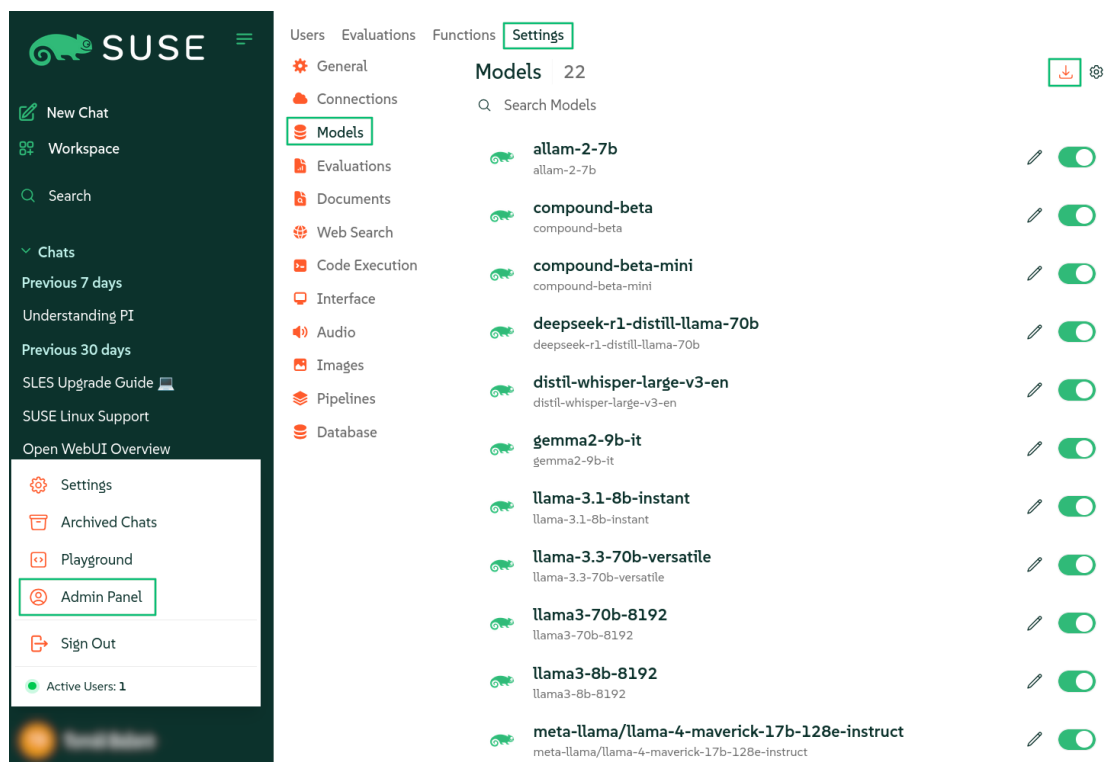


FIGURE 2.1: DOWNLOADING OLLAMA AI MODELS

4. On the [Ollama library page \(https://ollama.com/library\)](https://ollama.com/library), identify the model tag of an AI model that you want to download, for example, [gemma3:4b](#).
5. Paste the model tag in the *Pull a model from Ollama.com* input field and confirm by clicking the small download button on the right.

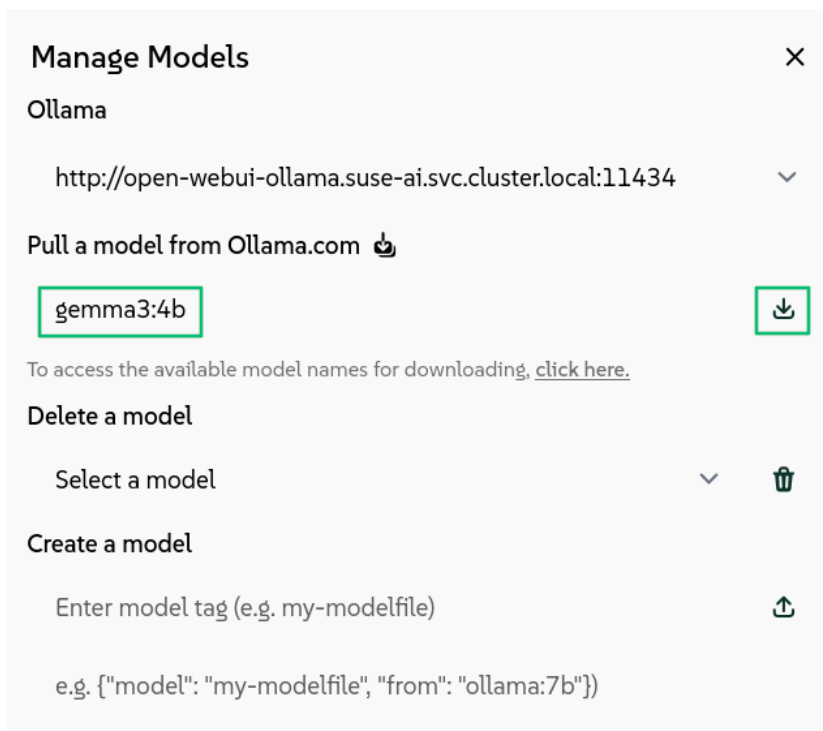


FIGURE 2.2: ENTERING MODEL TAG TO DOWNLOAD

The model download starts and, after it is finished, the model will be available as a base model for AI interaction.

2.2 Downloading AI models from Hugging Face

Hugging Face (<https://huggingface.co/models>) is an online repository for AI models. It also supports models in the GGUF binary format optimized for quick loading and saving of models. This procedure describes how to download and use these GGUF models from the Open WebUI interface.

REQUIREMENTS

- You must have Open WebUI administrator privileges to access configuration screens or settings mentioned in this section.
1. Navigate your Web browser to <https://huggingface.co/models> and in the left *Libraries* section, click *GGUF* to include only GGUF models to choose from, for example, *gemma-7b*.

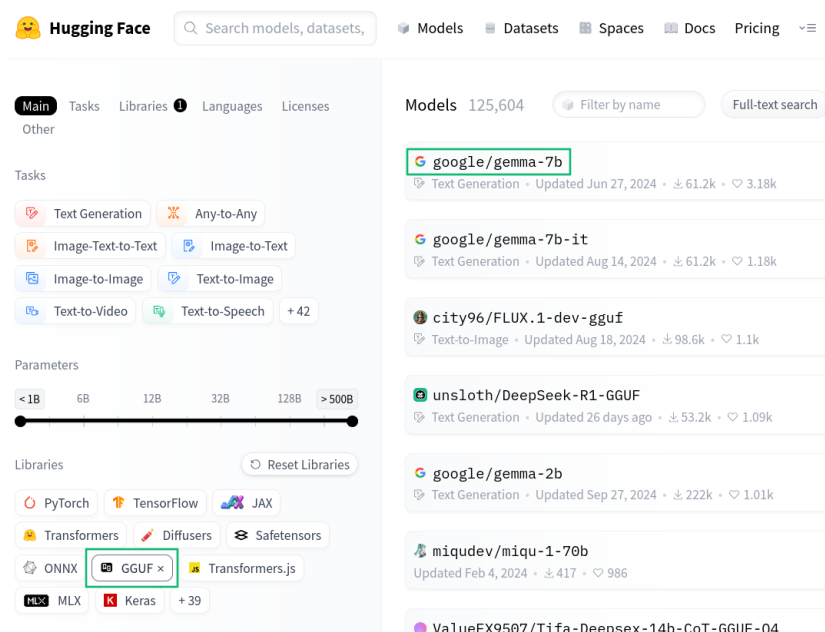


FIGURE 2.3: FINDING MODELS ON HUGGING FACE

2. Click the model name to open its model card and select *Use this model* > *Ollama*

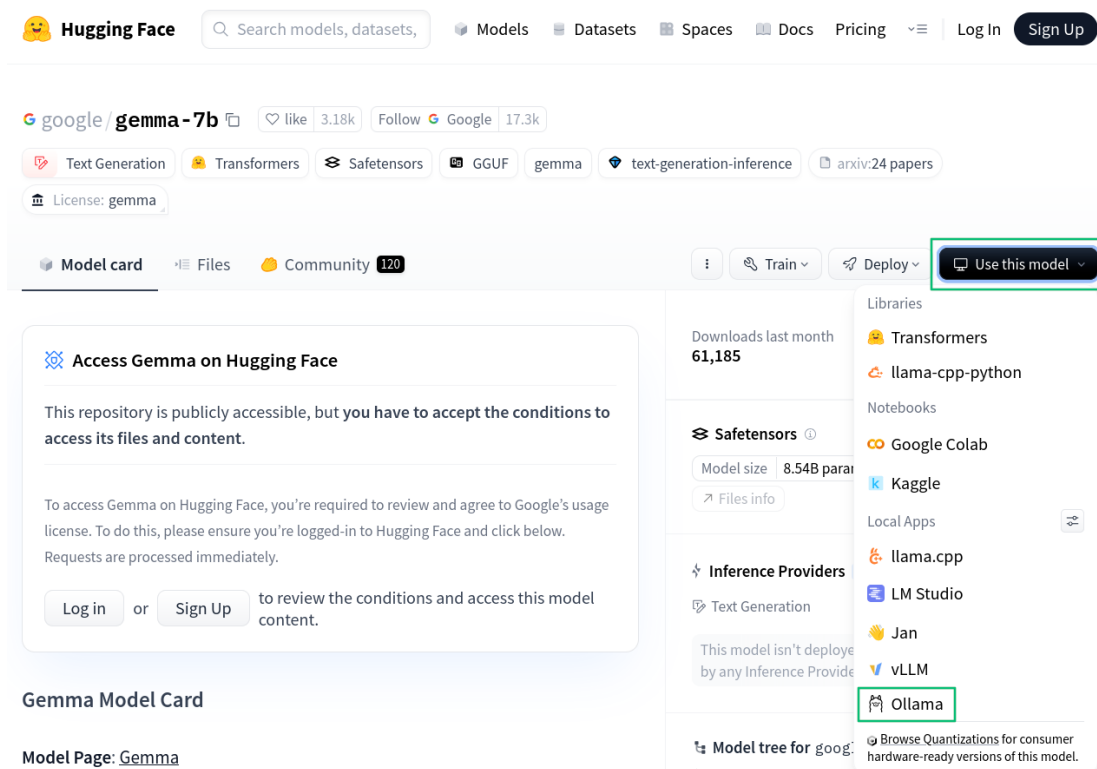


FIGURE 2.4: HUGGING FACE MODEL CARD

3. From the *How to use* pop-up, select the full path to the model, for example, [hf.co/google/gemma-7b](https://huggingface.co/google/gemma-7b).

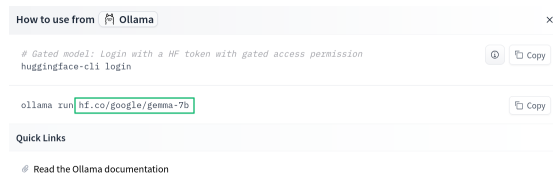


FIGURE 2.5: OLLAMA DOWNLOAD ON HUGGING FACE

4. In the bottom left of the Open WebUI window, click your avatar icon to open the user menu and select *Admin Panel*.
5. Click the *Settings* tab and select *Connections* from the left menu.
6. In the *Manage Ollama API Connections* section, click the small "download" icon on the right to open the *Manage Ollama* screen.

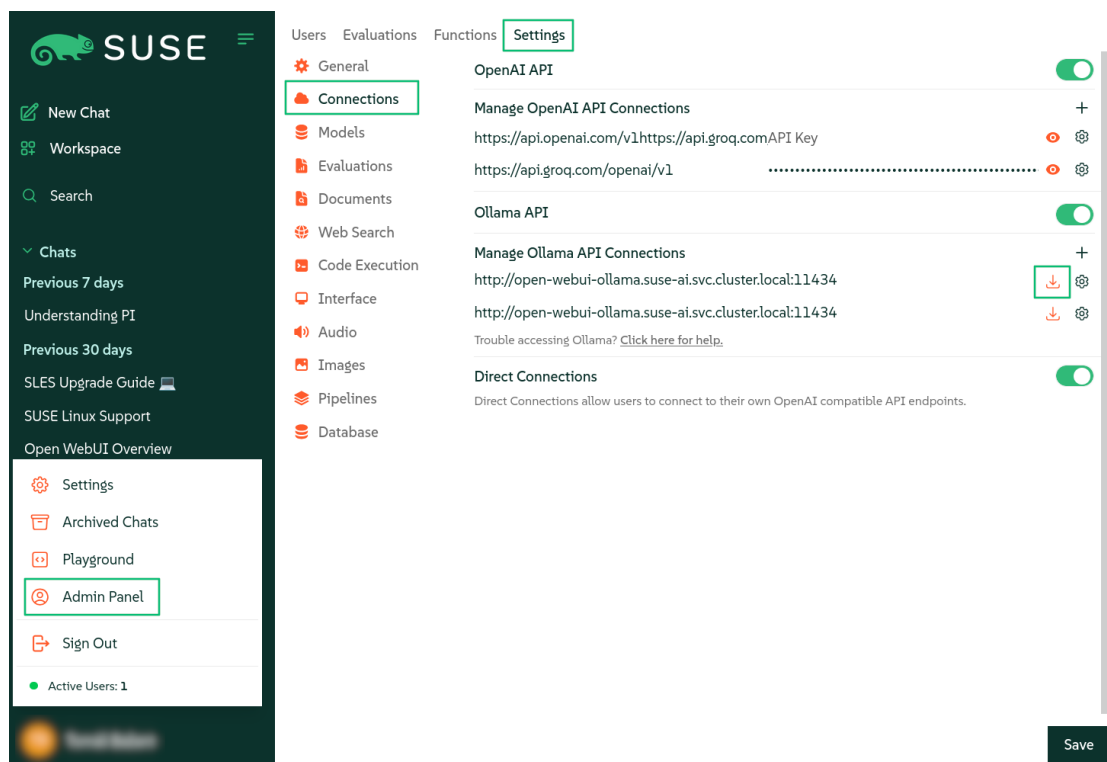


FIGURE 2.6: DOWNLOADING OLLAMA MODELS FROM HUGGING FACE

7. Paste the full path to the model tag from Hugging Face in the *Pull a model from Ollama.com* input field and confirm by clicking the small download button on the right.

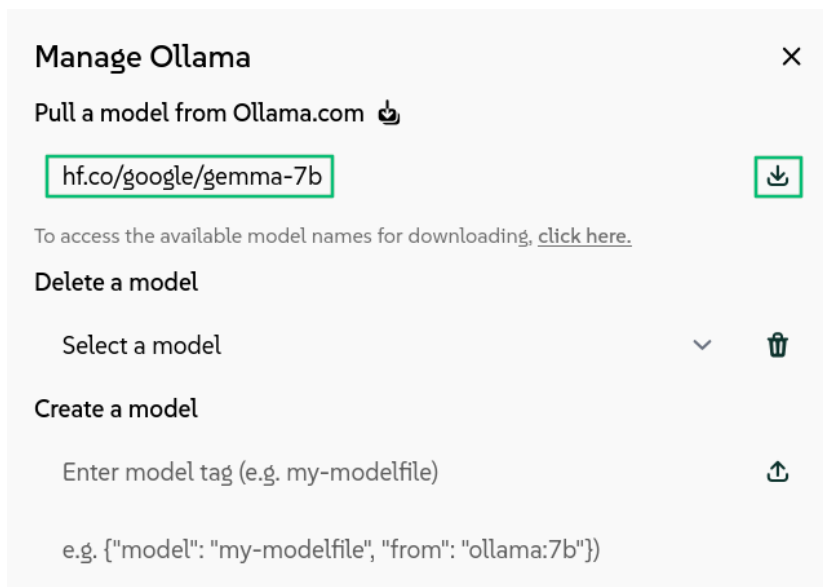


FIGURE 2.7: ENTERING MODEL TAG TO DOWNLOAD

The model download starts and, after it is finished, the model will be available as a base model for AI interaction.

2.3 Using AI models from OpenAI-compatible providers

Instead of downloading AI models locally, you can use OpenAI-compatible API to access models from model providers. These providers include both local instances of servers, such as [LocalAI](https://localai.io/) (<https://localai.io/>) or [llamafile](https://github.com/Mozilla-Ocho/llamafile) (<https://github.com/Mozilla-Ocho/llamafile>), and cloud providers, such as [Groq](https://groq.com/) (<https://groq.com/>) or [OpenAI](https://openai.com/api/) (<https://openai.com/api/>). This procedure describes how to configure the Groq provider API from the Open WebUI interface.

REQUIREMENTS

- You must have Open WebUI administrator privileges to access configuration screens or settings mentioned in this section.
 1. Create a Groq cloud account at <https://console.groq.com/>.
 2. Create an API key at <https://console.groq.com/keys>.

3. In the bottom left of the Open WebUI window, click your avatar icon to open the user menu and select *Admin Panel* .
4. Click the *Settings* tab and select *Connections* from the left menu.
5. In the *Manage OpenAI API Connections* section, click the small "plus" icon on the right to open the *Add connection* screen.
6. Enter `https://api.groq.com/openai/v1` as the base URL and the API key that you have previously created. Confirm with *Save* .

Edit Connection ×

URL ↻ ☒

Key 👁 Prefix ID

Model IDs

Leave empty to include all models from "https://api.groq.com/openai/v1/models" endpoint

Add a model ID +

Delete Save

FIGURE 2.8: **ADDING THE GROQ API**

Open WebUI now shows all the available models from Groq in the model selector drop-down list.

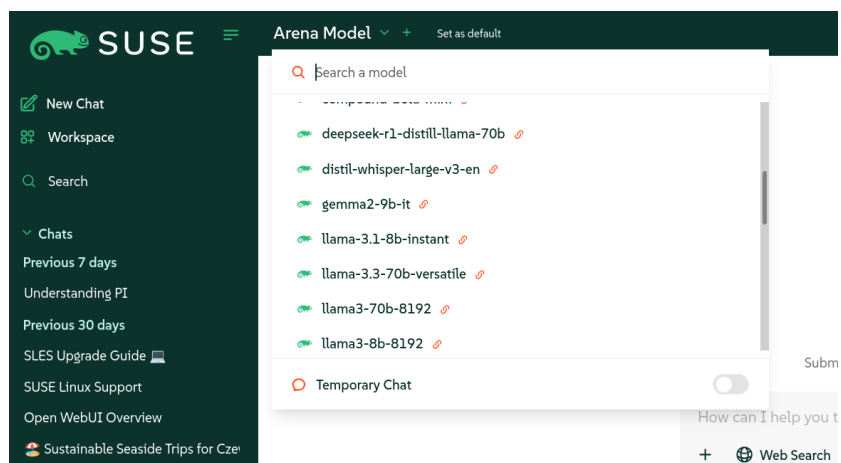


FIGURE 2.9: LIST OF AI MODULES FROM THE GROQ API

3 Setting the default system prompt

In Open WebUI, you can set system prompts in several contexts—for the current chat, for the current user, or for a specific AI model. While the prompt set for the current chat overrides the user and model prompts, the user-specific prompt overrides the model prompt only.

3.1 Setting a system prompt for the current chat

1. In the top right of the Open WebUI window, click the *Chat controls* icon.
2. Unfold the *System Prompt* tab and paste your system prompt into the empty input text field.

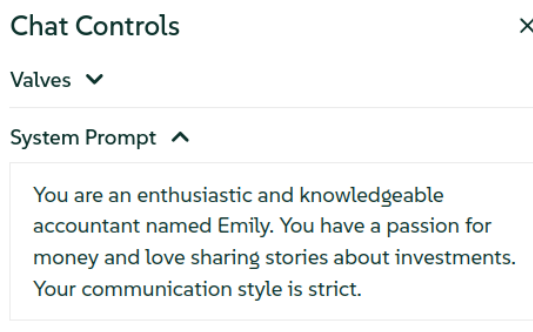


FIGURE 3.1: CHAT-SPECIFIC SYSTEM PROMPT

3. Close the *Chat controls* window to apply the change. The system prompt that you entered will affect the responses from the AI model during the current chat session.

3.2 System prompt for the current user

1. In the bottom left of the Open WebUI window, click your avatar icon to open the user menu.
2. Select *Settings* to open your preferences and click the *General* tab.
3. Paste your system prompt in the *System Prompt* text field.

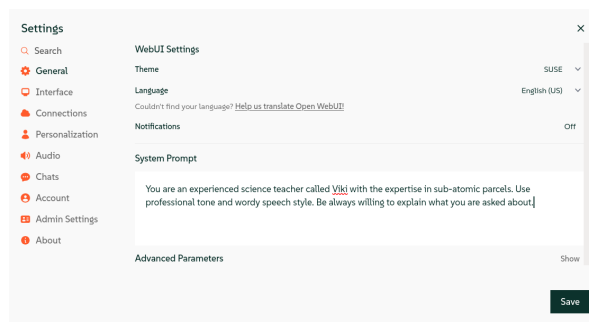


FIGURE 3.2: USER-SPECIFIC SYSTEM PROMPT

4. Confirm with *Save* and close the *Settings* window. The system prompt that you entered will affect the responses from the AI model during the current chat session.

3.3 System prompt for an AI model

REQUIREMENTS

- You must have Open WebUI administrator privileges to access configuration screens or settings mentioned in this section.
1. In the Open WebUI interface, click *Workspace* in the upper part of the left panel.
 2. Select *Models* in the top menu and scroll to the AI model whose system prompt you want to set. Use the pager at the bottom if the model is not listed on the current screen.
 3. Click the corresponding pen icon at its right to open the dialog for editing the model.
 4. In the *Model Params* section, specify the system prompt in the *System prompt* field.

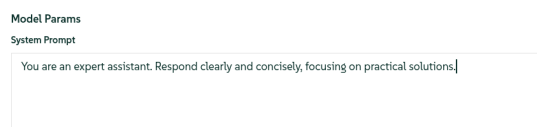


FIGURE 3.3: MODEL SYSTEM PROMPT

5. Confirm with *Save & Update*.

3.3.1 For more information

Find more details about system prompts in <https://documentation.suse.com/suse-ai/1.0/html/AI-system-prompts/index.html> .

4 Setting the default AI model

In Open WebUI, you can specify an AI model for the current chat or the system default model. The AI model set for the current chat overrides the default system model.

4.1 AI model for the current chat

1. In the top left of the Open WebUI chat window, click either the currently selected model name or the message saying that no model is specified.
2. From the drop-down list, select your preferred AI model for the current chat session.

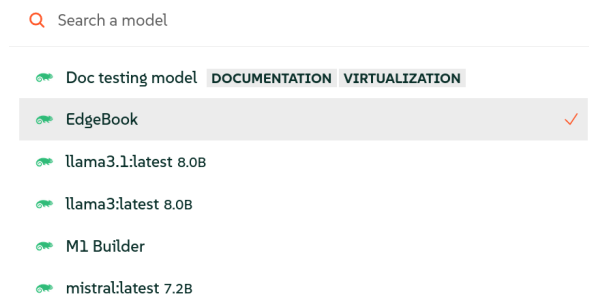


FIGURE 4.1: CHAT-SPECIFIC AI MODEL



Tip: Your default model

To make the selected model your default one, click *Set as default* next to the AI model name.



Tip: Multiple AI modules

You can select multiple AI modules by clicking the plus sign next to the name of the already selected module.

3. Your prompts will now be processed by the selected AI model.

4.2 Default system AI model

REQUIREMENTS

- You must have Open WebUI administrator privileges to access configuration screens or settings mentioned in this section.
1. In the bottom left of the Open WebUI window, click your avatar icon to open the user menu and select *Admin Panel* .
 2. Click the *Settings* tab and select *Models* from the left menu.
 3. Click the little gear icon in the top right to open the *Settings* screen.
 4. In the *Default Models* section, select the default system AI model. Confirm with *Save* .

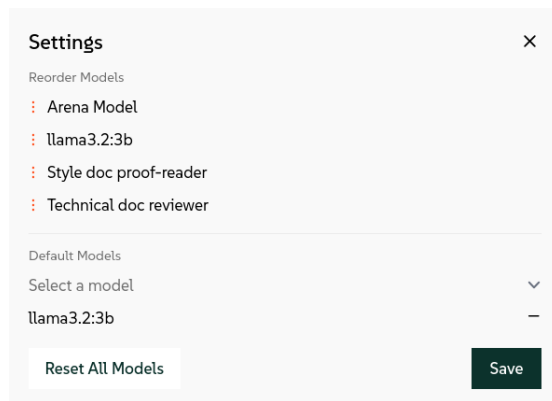


FIGURE 4.2: DEFAULT SYSTEM AI MODEL

5 Managing custom AI models

In Open WebUI, you can create custom AI models based on existing models. Later on, you can modify its settings, or share, clone or delete it.

5.1 Creating custom AI models

REQUIREMENTS

- You must have Open WebUI administrator privileges to access configuration screens or settings mentioned in this section.
1. In the Open WebUI interface, click *Workspace* in the upper part of the left panel.
 2. Select *Models* in the top menu and click the small "plus" icon in the top-right corner of the screen.

The screenshot shows the configuration page for a custom AI model named 'Virtualization expert'. The page includes the following sections:

- Base Model (From):** llama3.2:3b
- Description:** virtualization linux (with a plus icon to add tags)
- Visibility:** Public (Accessible to all users)
- Model Params:**
 - System Prompt:** You are an expert on computer virtualization. Use professional and positive tone. Express your answers in short sentences. Be precise and focus on precise and secure solutions. Use examples to guide the user.
 - Advanced Params:** A table with 'Prompt suggestions' and 'Custom' settings.
- Knowledge:** To attach knowledge base here, add them to the "Knowledge" workspace first. A card for 'SUSE AI Collection' is shown with a 'Select Knowledge' button.

FIGURE 5.1: MANAGING AI MODELS

3. Click the default logo image in the top-left corner of the form to upload an avatar image for your model.
4. Enter the *Name* of your new model and change the *Model ID* if you are not happy with the provided one.

5. Specify the base AI model on which your model will be based on.
6. Enter a short description of your AI model.
7. Describe the model with custom tags so that other users can identify it easily.
8. To direct the AI model's responses, specify a *System Prompt*. Find more details about system prompts in <https://documentation.suse.com/suse-ai/1.0/html/AI-system-prompts/index.html>.



Tip

By clicking the small *Show* option below the system prompt field, you can fine-tune the model's behavior by modifying its advanced settings.

9. Specify one or more *Prompt suggestions*. They will be shown on the model's chat welcome screen.
10. To extend the model's knowledge base, select one or more documents or collections of documents from the preconfigured list.
11. In the *Tools*, *Filters*, *Actions* and *Capabilities* sections, activate the items that you want your model to use.



Tip

To view the JSON representation of your model, click the small *Show* option to the right of the *JSON Preview* label.

12. Save your changes with the *Save & Create* button.

5.2 Modifying existing AI models

REQUIREMENTS

- You must have Open WebUI administrator privileges to access configuration screens or settings mentioned in this section.

1. In the Open WebUI interface, click *Workspace* at the top of the left panel and select *Models* .
2. In the list of AI models, hover the mouse over the model that you want to modify and click the pen icon on its right.
3. On the model editing screen, you can modify the settings that are described in [Section 5.1, "Creating custom AI models"](#).
4. On the model editing screen, you can modify the settings that are described in.
5. Save your changes with the *Save & Update* button.

5.2.1 Enabling the default MCP services

To enable selected MCP services so that they are automatically enabled for a model, follow these steps:

REQUIREMENTS

- At least one MCP server must be preconfigured. Refer to [Installing mcpo \(https://documentation.suse.com/suse-ai/1.0/html/AI-deployment-intro/index.html#mcpo-installing\)](https://documentation.suse.com/suse-ai/1.0/html/AI-deployment-intro/index.html#mcpo-installing) [↗](#) to learn how to configure one.

1. In the Open WebUI interface, click *Workspace* at the top of the left panel and select *Models* .
2. In the list of AI models, hover the mouse over the model that you want to modify and click the pen icon on its right to edit its properties.
3. In the *Tools* section, select the MCP tools you want to enable by default and confirm with *Save* .

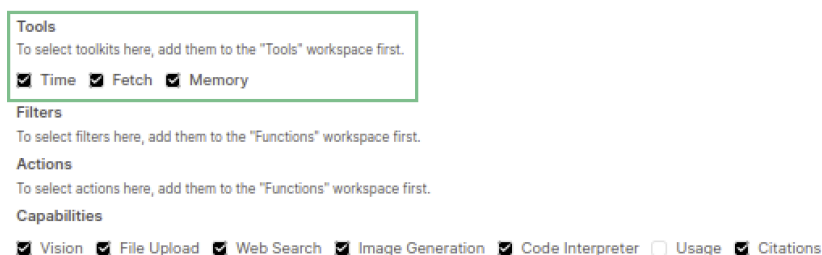


FIGURE 5.2: ENABLING THE DEFAULT MCP TOOLS

5.3 Cloning existing AI models

REQUIREMENTS

- You must have Open WebUI administrator privileges to access configuration screens or settings mentioned in this section.
1. In the Open WebUI interface, click *Workspace* at the top of the left panel and select *Models* .
 2. In the list of AI models, hover the mouse over the model that you want to clone and click the three dots on its right.
 3. From the contextual menu, select *Clone* to open the same form as described in [Section 5.1, "Creating custom AI models"](#) with all settings copied from the original model. The *Name* field has (Clone) appended.
 4. From the contextual menu, select *Clone* to open the same form as described in with all settings copied from the original model. The *Name* field has (Clone) appended.
 5. Review the settings and create the cloned AI model with *Save & Create*.

5.4 Importing exported models

REQUIREMENTS

- You must have Open WebUI administrator privileges to access configuration screens or settings mentioned in this section.
1. In the Open WebUI interface, click *Workspace* at the top of the left panel and select *Models* .
 2. Click the *Import Models* button at the bottom of the screen.
 3. In the file selector, specify the path to the JSON file with the exported models' definitions. Confirm with *Open* .

Models with identical IDs. IMPORTANT

AI models with IDs identical to existing models will not be imported.

5.5 Exporting existing AI models

REQUIREMENTS

- You must have Open WebUI administrator privileges to access configuration screens or settings mentioned in this section.
 1. In the Open WebUI interface, click *Workspace* at the top of the left panel and select *Models* .
 2. In the list of AI models, hover the mouse over the model that you want to export and click the three dots on its right.
 3. From the contextual menu, select *Export* .
 4. In the file selector, specify the path where you want to save the exported JSON file. Confirm with *Save* .

5.6 Exporting all existing AI models

REQUIREMENTS

- You must have Open WebUI administrator privileges to access configuration screens or settings mentioned in this section.
 1. In the Open WebUI interface, click *Workspace* at the top of the left panel and select *Models* .
 2. Click the menu:"Export Models"[] button at the bottom of the screen.
 3. In the file selector, specify the path where you want to save the exported JSON file. Confirm with *Save* .

5.7 Deleting existing AI models

Using the Open WebUI interface, you can delete previously created AI models.

REQUIREMENTS

- You must have Open WebUI administrator privileges to access configuration screens or settings mentioned in this section.

1. In the Open WebUI interface, click *Workspace* at the top of the left panel and select *Models* .
2. In the list of AI models, hover the mouse over the model that you want to delete and click the three dots on its right.
3. From the contextual menu, select *Delete* and then apply with *Confirm* .

6 Managing user accounts

Open WebUI provides **role-based access control** (RBAC). Each user is assigned one of the available roles. The role defines whether the user has privileges to perform administration tasks.

6.1 User roles in Open WebUI

Open WebUI offers three user roles: 'pending', 'user' and 'admin.' The following sections provide a detailed description of each role.

6.1.1 The 'pending' role

Pending users can log in but can only view the *Account Activation Pending* screen with the administrator name and e-mail. This is the default role for a newly created user account before the administrator accepts it and assigns it the "user" role.

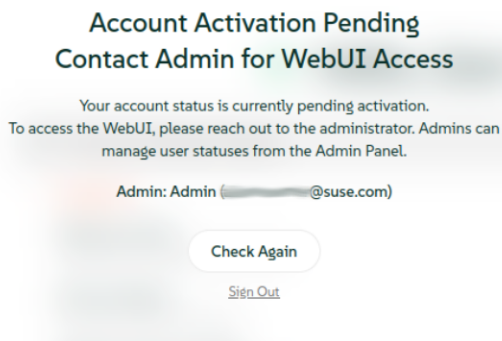


FIGURE 6.1: PENDING USER SCREEN

6.1.2 The 'user' role

Users with this role can perform the following tasks:

Chat permissions

- Create new chats with the preconfigured models.
- View own chats, including archived chats.

- Pin, rename, clone, archive or unarchive, share, delete, add tags to, and remove tags from existing chats.
- Generate an overview of a chat, view artifacts, copy content and download chats.
- Copy, edit and delete their own messages in a chat.
- Edit, copy and read aloud the responses. View response generation information. Generate images. Mark answers as good or bad. Continue responding and regenerate model messages in a chat.

Current chat controls

- Modify valves from the tools and functions available on the system.
- Define the system prompt to be used by the current chat.
- Modify advanced chat-specific generation parameters.

Current user settings

- Select theme and language. Activate or deactivate notifications. Define the default system prompt and change advanced generation parameters.
- Set the default model. Set the user interface, chat and audio related preferences.
- Personalize the interaction with the AI model by adding 'memory' —personal details about yourself—to make responses tailored to you.
- Set audio preferences.
- Import and export chats. Archive and delete all user chats.
- Change your personal details such as name, password and profile image. View JWT token and create new API key.
- View the installed Open WebUI and Ollama version.

6.1.3 The 'admin' role

Administrators have full access to Open WebUI resources and can control user management and systems settings. The first Open WebUI account has the 'admin' role. Subsequent administrators get their role from existing administrators.

Administrator's task and settings are included either on the *Admin Panel* or in the *Workspace*.

Administrators have the permissions that the 'user' role has with the following additional privileges.

Admin panel > Users

- Add new users.
- View all users and their role, name, e-mail, OAuth 2.0 ID, last active date and the user creation date.
- Edit name, e-mail and password and delete all users, except for other administrators.
- View and delete chats of all users, except for other administrators.
- Change the role of all users, except their own and the first administrator.

Admin panel > Settings

- Enable or disable new sign-ups. Set the default user role. Enable or disable viewing the administrator details on the *Account Activation Pending* screen. Enable or disable chat sharing and message rating. Set the JWT expiration and webhook URL.
- Enable or disable chat deletion, chat editing and temporary chat. Set the default model and specify allowed models.
- View, change and add OpenAI API URL and Ollama API URL.
- Update all models, pull models from [ollama.com \(https://ollama.com\)](https://ollama.com) , delete, create and upload models.
- Change documents settings, set the RAG template, and reset document storage.
- Enable or disable Web search, change the Web search engine, and set Web loader settings.
- Set the default model used for the title generation. Change the default prompt suggestions for the *New Chat* screen.
- Change audio settings.
- Enable or disable image generation and change image generation settings.
- Change pipeline settings.
- Import or export configuration files, download database, and export all chats from all users.

Workspace > Models

Manage models. Add models by pulling from ollama.com (<https://ollama.com>) or by creating custom models from base models. Change order of models or hide models from the model list.

Workspace > Knowledge

Manage collections of documents. Add documents to collections.

Workspace > Prompts

Manage system prompts. Share prompts, import and export all prompts.

Workspace > Tools

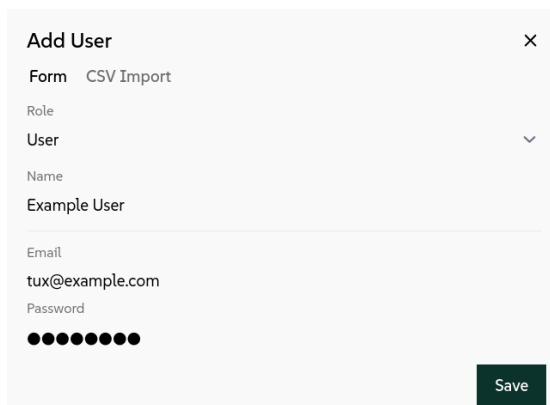
Manage tools. Import and export tools.

6.2 Creating user accounts

6.2.1 Creating a single user account

REQUIREMENTS

- You must have Open WebUI administrator privileges to access configuration screens or settings mentioned in this section.
1. In the bottom left of the Open WebUI window, click your avatar icon to open the user menu and select *Admin Panel*.
 2. On the *Users* screen, click the '+' sign at the top right.



Add User X

Form CSV Import

Role

User

Name

Example User

Email

tux@example.com

Password

●●●●●●●●

Save

FIGURE 6.2: CREATING NEW USERS

3. On the *Form* screen, select the user role and enter the user name, e-mail and password.
4. Confirm with *Save* .

6.2.2 Creating multiple user accounts from a CSV list

REQUIREMENTS

- You must have Open WebUI administrator privileges to access configuration screens or settings mentioned in this section.
1. In the bottom left of the Open WebUI window, click your avatar icon to open the user menu and select *Admin Panel* .
 2. On the *Users* screen, click the '+' sign at the top right.
 3. On the *CSV Import* screen, specify the path to your CSV file with the list of user accounts to be created.



Tip

The CSV list needs to have the four fields Name,Email,Password,Role per line separated by commas, for example:

```
Example User,tux@example.com,SECRET_PASSWD,user
```

4. Confirm with *Save* .

6.3 Assigning roles to existing users

This topic explains how to assign one of the available roles to an existing user account. To assign the role when creating a new user account, refer to [Section 6.2, "Creating user accounts"](#).

REQUIREMENTS

- You must have Open WebUI administrator privileges to access configuration screens or settings mentioned in this section.

1. In the bottom left of the Open WebUI window, click your avatar icon to open the user menu and select *Admin Panel* .
2. On the *Users* screen, scroll to the user whose role you want to modify. Use the pager at the bottom if the user is not listed on the current screen.

ROLE	NAME	EMAIL	LAST ACTIVE	CREATED AT	OAUTH ID
ADMIN	Tuxmail Assistant	tuxmail.assistant@gmail.com	a few seconds ago	June 2, 2025	
USER	Example User 3	tux3@example.com	a few seconds ago	June 17, 2025	
USER	Example User 4	tux4@example.com	a few seconds ago	June 17, 2025	
USER	Example User 5	tux5@example.com	a few seconds ago	June 17, 2025	
USER	Example User 6	tux6@example.com	a few seconds ago	June 17, 2025	
USER	Example User 7	tux7@example.com	a few seconds ago	June 17, 2025	
USER	Example User 8	tux8@example.com	a few seconds ago	June 17, 2025	

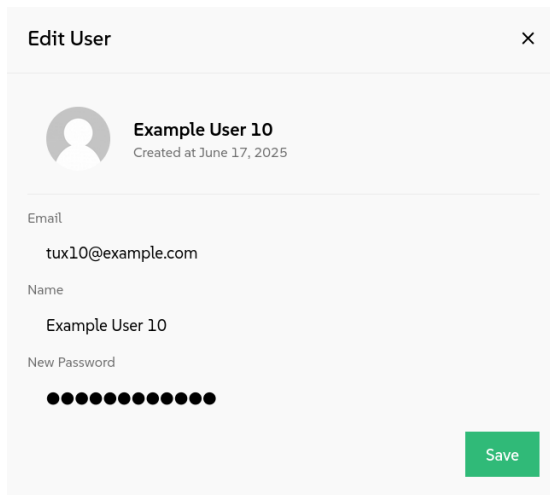
FIGURE 6.3: ASSIGNING USER ROLE

3. Click the role button to the left of the user name to cycle through "pending" → "user" → "admin" role.

6.4 Modifying existing user accounts

REQUIREMENTS

- You must have Open WebUI administrator privileges to access configuration screens or settings mentioned in this section.
1. In the bottom left of the Open WebUI window, click your avatar icon to open the user menu and select *Admin Panel* .
 2. On the *Users* screen, scroll to the user whose account you want to modify. Use the pager at the bottom if the user is not listed on the current screen.
 3. Click the corresponding pen icon at their right to open the *Edit User* dialog.



Edit User ×

 **Example User 10**
Created at June 17, 2025

Email
tux10@example.com

Name
Example User 10

New Password
●●●●●●●●●●

Save

FIGURE 6.4: MODIFYING USER ACCOUNTS

4. Modify the user's e-mail, name or password and confirm with *Save* .

6.5 Revoking access to Open WebUI

REQUIREMENTS

- You must have Open WebUI administrator privileges to access configuration screens or settings mentioned in this section.
1. In the bottom left of the Open WebUI window, click your avatar icon to open the user menu and select *Admin Panel* .
 2. On the *Users* screen, scroll to the user whose access to Open WebUI you want to revoke. Use the pager at the bottom if the user is not listed on the current screen.
 3. Choose one of the following options:
 - To revoke the access while maintaining the user account, click the user's role button to the left of their name until it becomes "pending."
 - To delete the user account from the system, click the trash icon on the right and confirm the deletion.

7 Managing a knowledge base

Knowledge base is a custom collection of resources on a specific topic. It can include documentation, tutorials, reference guides or short text snippets. Knowledge base collections extend the data on which the selected AI model is trained to make its context more specific.

7.1 Creating a knowledge base

REQUIREMENTS

- You must have Open WebUI administrator privileges to access configuration screens or settings mentioned in this section.
 1. In the Open WebUI interface, click *Workspace* in the upper part of the left panel.
 2. Select *Knowledge* in the top menu and click the "plus" sign in the right.
 3. Enter a title and a short description for the knowledge base.
 4. Specify whether the knowledge base is public or private and confirm with *Create Knowledge* .

Create a knowledge base

What are you working on?

SLES Upgrade

What are you trying to achieve?

Information about upgrades of SLES.

Visibility



Public



Accessible to all users

Create Knowledge

FIGURE 7.1: CREATING A NEW KNOWLEDGE BASE

5. Add content to the knowledge base as described in [Section 7.3, “Adding content to a knowledge base”](#).



Tip

You can access knowledge base collections from the Open WebUI chat input field by pressing the `#` character as described in [Specifying preconfigured documents, collections or external URLs from the chat input field](#) (<https://documentation.suse.com/suse-ai/1.0/html/openwebui-using/index.html#openwebui-input-field-collections>) [↗](#).

7.2 Deleting a knowledge base

REQUIREMENTS

- You must have Open WebUI administrator privileges to access configuration screens or settings mentioned in this section.

1. In the Open WebUI interface, click *Workspace* in the upper part of the left panel.
2. Select *Knowledge* in the top menu and scroll to the knowledge base that you want to delete. Use the pager at the bottom if the knowledge base is not listed on the current screen.
3. In the knowledge base tile's top right corner, click the three dots and select *Delete* . Confirm the deletion.

7.3 Adding content to a knowledge base



Important: File upload size limit

The default configuration allows file uploads up to 1 MB. To enable larger file uploads, follow these steps:

1. Update Ingress configuration by creating the `ingress.yaml` file with the following content.

```
apiVersion: networking.k8s.io/v1
kind: Ingress
metadata:
  name: open-webui
  namespace: SUSE_AI_NAMESPACE
annotations:
  nginx.ingress.kubernetes.io/proxy-body-size: 100m ❶
  nginx.ingress.kubernetes.io/proxy-connect-timeout: "360"
  nginx.ingress.kubernetes.io/proxy-read-timeout: "360"
  nginx.ingress.kubernetes.io/proxy-send-timeout: "360"
  nginx.ingress.kubernetes.io/websocket-services: "open-webui"
```

- ❶ This example sets the upload file size limit to 100 MB. Setting this value to `0` allows unlimited upload size (not recommended).

2. Apply the changes so that large file uploads are supported.

```
> kubectl apply -f ingress.yaml
```

REQUIREMENTS

- You must have Open WebUI administrator privileges to access configuration screens or settings mentioned in this section.

1. In the Open WebUI interface, click *Workspace* in the upper part of the left panel and select the *Knowledge* tab.
2. Scroll to the knowledge base that you want to add content to. Use the pager at the bottom if the knowledge base is not listed on the current screen.
3. Click the knowledge base name and in the menu:Search Collection [] files, click the "plus" sign.
4. From the drop-down list, you can select one of the following actions:

Upload files

Select one or more documents to be added to the collection.

Upload directory

Select a directory whose contents will be added to the collection one item at a time.



Warning: Wait until the processing is finished

Do not leave the collection's screen while the selected files are being uploaded and processed. Leaving the screen may disrupt the file processing.

Sync directory

The selected directory is synchronized with the content of the collection.



Warning: The collection content will be deleted

Before Open WebUI starts to synchronize the directory, all existing resources in the collection are deleted.



Warning: Wait until the processing is finished

Do not leave the collection's screen while the selected files are being uploaded and processed. Leaving the screen may disrupt the file processing.

Add text content

You can add a text snippet written in a Markdown format to the collection directly.



Warning

If you experience the error saying "Extracted content is not available for this file. Please ensure that the file is processed before proceeding," during the upload process, you need to enable installing NLTK datasets in the Open WebUI Helm chart. Refer to <https://documentation.suse.com/suse-ai/1.0/html/AI-deployment-intro/index.html#owui-helm-overrides> for details.

Confirm with *Upload*

5. The uploaded resource is processed and analyzed so that you can use it as a context for selected AI models. The processing time may vary depending on the size of the uploaded resources. After a resource is added, you can view and edit its extracted textual information by clicking its name.

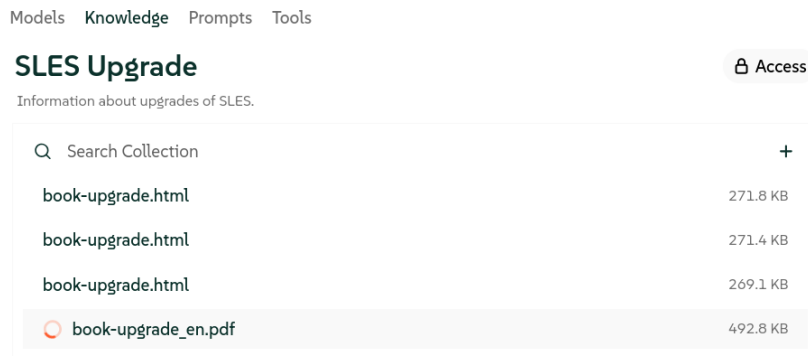


FIGURE 7.2: VIEWING AN EXTRACTED RESOURCE

7.4 Removing content from a knowledge base

REQUIREMENTS

- You must have Open WebUI administrator privileges to access configuration screens or settings mentioned in this section.

1. In the Open WebUI interface, click *Workspace* in the upper part of the left panel and select the *Knowledge* tab.
2. Scroll to the knowledge base that you want to add content to. Use the pager at the bottom if the knowledge base is not listed on the current screen.
3. Click the knowledge base name to view all the resources in the collection.
4. Hover over the resource's name and click the little cross sign at its top right. The resource will be deleted and its processed data removed from the collection.

8 Managing system prompts

System prompts are a powerful core component that shapes AI model behavior. They provide precise control over how AI models process and respond to user inputs.

Find more details about system prompts in [Guiding the AI Model with System Prompts \(https://documentation.suse.com/suse-ai/1.0/html/AI-system-prompts/index.html\)](https://documentation.suse.com/suse-ai/1.0/html/AI-system-prompts/index.html).

8.1 Creating a system prompt

REQUIREMENTS

- You must have Open WebUI administrator privileges to access configuration screens or settings mentioned in this section.

1. In the Open WebUI interface, click *Workspace* in the upper part of the left panel.
2. Select *Prompts* in the top menu and click the "plus" sign on the right.

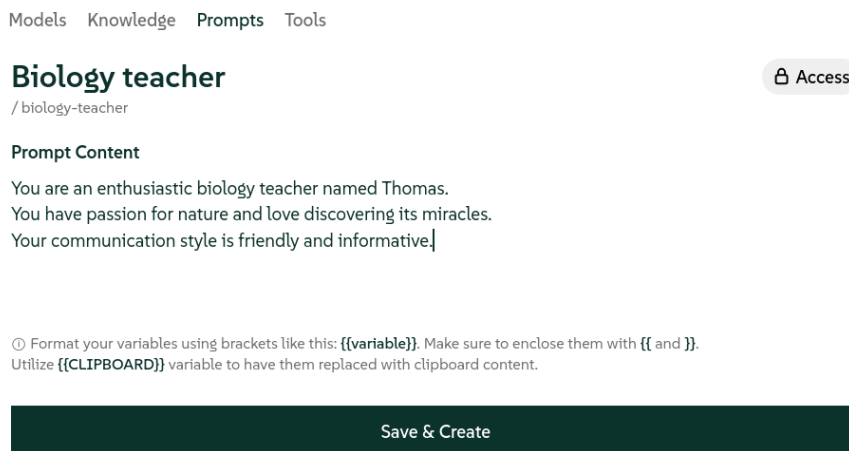


FIGURE 8.1: CREATING A NEW SYSTEM PROMPT

3. Enter the title and content of the new system prompt, and optionally, update the command that invokes the prompt from the chat input field.
4. Confirm with *Save & Create*.



Tip

You can access preconfigured system prompts from the Open WebUI chat input field by pressing the `/` character as described in [Selecting a preconfigured system prompt from the chat input field](https://documentation.suse.com/suse-ai/1.0/html/openwebui-using/index.html#openwebui-input-field-specify-system-prompt) (<https://documentation.suse.com/suse-ai/1.0/html/openwebui-using/index.html#openwebui-input-field-specify-system-prompt>) .

8.2 Modifying existing system prompts

REQUIREMENTS

- You must have Open WebUI administrator privileges to access configuration screens or settings mentioned in this section.
1. In the Open WebUI interface, click *Workspace* in the upper part of the left panel and select *Prompts* .
 2. Scroll to the prompt that you want to modify. Use the pager at the bottom if the prompt is not listed on the current screen.
 3. Click the corresponding pen icon on the right to open the form that enables you to edit the title and the description of the prompt.
 4. Confirm with *Save & Update* .

8.3 Cloning existing system prompts

REQUIREMENTS

- You must have Open WebUI administrator privileges to access configuration screens or settings mentioned in this section.
1. In the Open WebUI interface, click *Workspace* in the upper part of the left panel and select *Prompts* .
 2. Scroll to the prompt that you want to clone. Use the pager at the bottom if the prompt is not listed on the current screen.
 3. Click the three dots to the right of the prompt that you want to clone and select *Clone* .

4. Optionally update the title, the content and the activating command of the prompt. The command must be unique among all other existing system prompts.
5. Confirm with *Save & Create* .

8.4 Importing exported system prompts

REQUIREMENTS

- You must have Open WebUI administrator privileges to access configuration screens or settings mentioned in this section.
1. In the Open WebUI interface, click *Workspace* in the upper part of the left panel and select *Prompts* .
 2. Click *Import Prompts* in the bottom right of the prompt list.
 3. Navigate to the JSON file with previously exported prompts and confirm with *Open* .



Warning: Unique command

The activating command of the imported prompts must be unique among existing prompts. Prompts with an existing activating command will not be imported.

8.5 Exporting existing system prompts

REQUIREMENTS

- You must have Open WebUI administrator privileges to access configuration screens or settings mentioned in this section.
1. In the Open WebUI interface, click *Workspace* in the upper part of the left panel and select *Prompts* .
 2. Scroll to the prompt that you want to export. Use the pager at the bottom if the prompt is not listed on the current screen.
 3. Click the three dots to the right of the prompt that you want to clone and select *Export* .
 4. Navigate to the path where you want to save the JSON file with the exported prompt. Confirm with *Save* .

8.6 Exporting all existing system prompts

REQUIREMENTS

- You must have Open WebUI administrator privileges to access configuration screens or settings mentioned in this section.
1. In the Open WebUI interface, click *Workspace* in the upper part of the left panel and select *Prompts* .
 2. Click *Export Prompts* in the bottom right of the prompt list.
 3. Enter the path to the JSON file that will contain the exported prompts and confirm with *Save*.

8.7 Deleting existing system prompts

REQUIREMENTS

- You must have Open WebUI administrator privileges to access configuration screens or settings mentioned in this section.
1. In the Open WebUI interface, click *Workspace* in the upper part of the left panel and select *Prompts* .
 2. Scroll to the prompt that you want to clone. Use the pager at the bottom if the prompt is not listed on the current screen.
 3. Click the three dots to the right of the prompt that you want to clone and select *Delete* . Confirm the deletion.

9 Authenticating via Okta

Okta is an identity management service that allows you to manage users and user access to applications. This topic describes how to use Okta as the identity manager for Open WebUI.

The process of enabling Okta as the identity manager for Open WebUI consists of several steps:

1. Configuring Okta ([Section 9.1, “Configuring Okta”](#))
2. Gathering data for Open WebUI configuration ([Section 9.2, “Gathering data for Open WebUI configuration”](#))
3. Configuring Open WebUI for Okta integration ([Section 9.3, “Configuring Open WebUI for Okta integration”](#))



Tip

This example procedure shows the minimal configuration required to use Okta with Open WebUI. Customize the configuration to your specific scenario and refer to the Okta documentation at <https://help.okta.com/en-us/content/index.htm?cshid=csh-index> for more information.

9.1 Configuring Okta



Important

You must have administrator privileges to perform the following steps.

1. Log in to Okta and access the *Admin* panel.



FIGURE 9.1: OKTA ADMINISTRATOR PANEL

2. Go to *Applications > Create App Integration*.
3. In the *Sign-in method* section, select *OIDC - OpenID Connect*. In the *Application type* section, select *Web Application*. Click *Next* to proceed.

Create a new app integration

×

Sign-in method

Learn More

☒ **OIDC - OpenID Connect**
Token-based OAuth 2.0 authentication for Single Sign-On (SSO) through API endpoints. Recommended if you intend to build a custom app integration with the Okta Sign-In Widget.

☐ **SAML 2.0**
XML-based open standard for SSO. Use if the Identity Provider for your application only supports SAML.

☐ **SWA - Secure Web Authentication**
Okta-specific SSO method. Use if your application doesn't support OIDC or SAML.

☐ **API Services**
Interact with Okta APIs using the scoped OAuth 2.0 access tokens for machine-to-machine authentication.

Application type

What kind of application are you trying to integrate with Okta?

Specifying an application type customizes your experience and provides the best configuration, SDK, and sample recommendations.

☐ **Web Application**
Server-side applications where authentication and tokens are handled on the server (for example, Go, Java, ASP.Net, Node.js, PHP)

☐ **Single-Page Application**
Single-page web applications that run in the browser where the client receives tokens (for example, Javascript, Angular, React, Vue)

☐ **Native Application**
Desktop or mobile applications that run natively on a device and redirect users to a non-HTTP callback (for example, iOS, Android, React Native)

Cancel

Next

FIGURE 9.2: OKTA CONFIGURATION WINDOW

- In the *General Settings* section, set the *App integration name* to *Open WebUI* . In the *Grant type* section, toggle the *Refresh token* option. Optionally, specify an image for the application logo.
- Set *Sign-in redirect URIs* to the callback URI, for example, <http://localhost:8080/oauth/oidc/callback>. The URI follows this pattern:

```
PROTOCOL://OPENWEBUI_DOMAIN:OPTIONAL_PORT_NUMBER/oauth/oidc/callback
```

- Similarly, set *Sign-out redirect URIs* to

```
PROTOCOL://OPENWEBUI_DOMAIN:OPTIONAL_PORT_NUMBER
```

- In the *Assignments* section, toggle the *Skip group assignment for now* option. Save changes.



Tip: Open WebUI application settings

After the integration is created, you can access the Open WebUI application settings by clicking the related hyperlink. It is required for final configurations and for retrieving the necessary information for the Open WebUI configuration further on.



FIGURE 9.3: OPEN WEBUI APPLICATION ACCESS IN OKTA

8. In the *General Settings* section, click *Edit* and in the *Login* section, set the *Sign-in redirect URIs* option to the URI in this format:

```
PROTOCOL://OPENWEBUI_DOMAIN:OPTIONAL_PORT_NUMBER/oauth/oidc/callback
```

9. In the *Login initiated by* section, select *App Only* . Confirm with *Save* .

9.2 Gathering data for Open WebUI configuration

1. In the top-right menu, copy the Okta domain. In this example, it is my-domain.okta.com.

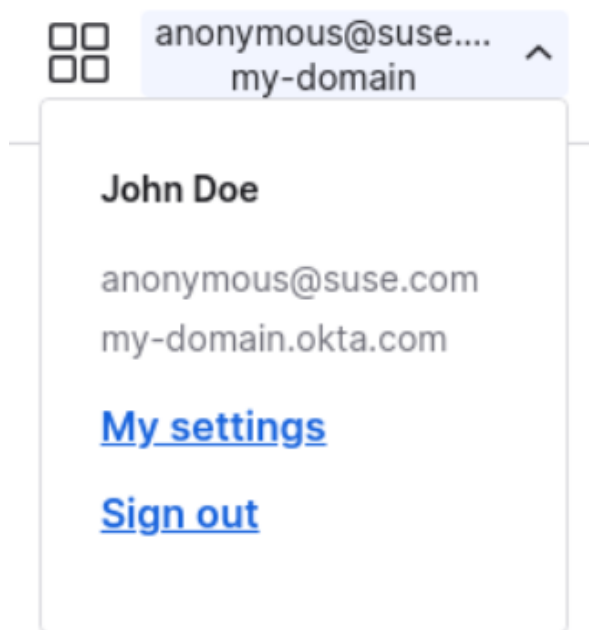


FIGURE 9.4: OKTA DOMAIN NAME

2. Go to *Applications > Open WebUI* and retrieve the *Client ID* and *Client Secret* .

FIGURE 9.5: CLIENT ID AND SECRET IN OKTA



Tip

The previously gathered values are used by Open WebUI via the following environment variables as described in [Section 9.3, “Configuring Open WebUI for Okta integration”](#).

```
OPENID_PROVIDER_URL="OKTA_DOMAIN/.well-known/openid-configuration"
OAUTH_CLIENT_ID="CLIENT_ID"
OAUTH_CLIENT_SECRET="CLIENT_SECRET"
```

9.3 Configuring Open WebUI for Okta integration

To configure Open WebUI to use Okta as an authentication provider, you must set up specific environment variables. For more information about OAuth 2.0 in Open WebUI and a complete description of each environment variable, refer to [Open WebUI SSO \(https://docs.openwebui.com/features/authentication-access/auth/sso/\)](https://docs.openwebui.com/features/authentication-access/auth/sso/) [documentation](#).

1. Add the following environment variables to your Helm charts `values.yaml` file in the `extraEnvVars` section.

```
ENABLE_OAUTH_SIGNUP=true ❶
OAUTH_MERGE_ACCOUNTS_BY_EMAIL=true ❷
OAUTH_CLIENT_ID=CLIENT_ID ❸
```

```
OAuth_CLIENT_SECRET=CLIENT_SECRET
OPENID_PROVIDER_URL=OKTA_DOMAIN/.well-known/openid-configuration
OAuth_PROVIDER_NAME=Okta
OAuth_SCOPES=openid email profile ④
```

- ① Allow account creation when logging in with OAuth 2.0.
- ② Allow logging into an account that matches the e-mail address provided by the OAuth 2.0 provider (*Optional*).
- ③ Variable set for the authentication provider gathered from Okta.
- ④ Specify scopes to request (optional, defaults to openid email profile).

The code block should look as follows:

```
extraEnvVars:
  - name: ENABLE_OAuth_SIGNUP
    value: "True"
  - name: OAuth_MERGE_ACCOUNTS_BY_EMAIL
    value: "True"
  - name: OAuth_CLIENT_ID
    value: "CLIENT_ID"
  - name: OAuth_CLIENT_SECRET
    value: "CLIENT_SECRET"
  - name: OPENID_PROVIDER_URL
    value: "OKTA_DOMAIN/.well-known/openid-configuration"
  - name: OAuth_PROVIDER_NAME
    value: "Okta"
  - name: OAuth_SCOPES
    value: "openid email profile"
```

2. Restart Open WebUI to apply the changes.
3. If the configuration went well, you should see the *Continue with Okta* button as an option on both sign-in and sign-up pages. After you click the button, you will be redirected to the Okta login page, or automatically signed in if you are already logged in to Okta.

Sign in to SUSE AI (Open WebUI)

Email

Enter Your Email

Password

Enter Your Password

Sign in

Don't have an account? [Sign up](#)

or



Continue with Okta

A Glossary

Air gap

A security measure where a computer network is physically isolated from unsecured networks, including the public Internet.

AI, artificial intelligence

Refers to the simulation of human intelligence in machines that are designed to learn and solve problems like humans. Enables computers to understand language, make decisions and improve from experience.

Batch size

The number of samples processed simultaneously during model inference, affecting processing speed and resource utilization.

BYOC, bring your own certificate

A practice allowing users to provide their own SSL/TLS certificates for securing communications instead of using default or auto-generated ones.

CA, certification authority

An entity that issues digital certificates to verify the identity of certificate holders and ensure secure communications.

Chat template

A structured format for organizing conversations between users and AI models, defining how system prompts, user inputs, and AI responses are formatted and processed.

Context window

The maximum amount of text (tokens) that an AI model can process at once, including both the input prompt and generated response.

Chain-of-thought (CoT) prompting

A prompting technique that guides AI models to break down complex problems into step-by-step reasoning processes, improving response accuracy and transparency.

CRD, custom resource definitions

Extensions of the Kubernetes API that allow users to define custom resources and their controllers in a Kubernetes cluster.

CUDA, Compute Unified Device Architecture

NVIDIA's parallel computing platform and programming model used to accelerate AI workloads on GPU hardware.

Data leakage

The unintended exposure of sensitive information through AI model responses, potentially compromising data security and privacy.

Embeddings

Numerical representations of data (text, images, etc.) in a high-dimensional space that capture semantic relationships and enable AI models to process information effectively.

Fine-tuning

The process of further training a pre-trained AI model on specific data to adapt it for particular tasks or domains, improving its performance for targeted applications.

GenAI, generative AI

A type of artificial intelligence that can create new content such as text, images or music.

GPU, graphics processing unit

Specialized hardware designed for parallel processing. In AI applications, GPUs accelerate model training and inference tasks.

Hallucination

An AI behavior where the model generates false or unsupported information that appears plausible but has no basis in provided context or real facts.

Helm

A package manager for Kubernetes that helps install and manage applications. Helm uses charts to define, install and upgrade complex Kubernetes applications.

Helm chart

A package format for Kubernetes applications that contains all resource definitions needed to deploy and configure application workloads.

IaC, infrastructure as code

The practice of managing and provisioning infrastructure through machine-readable definition files rather than manual processes.

Inference

The process of using a trained AI model to make predictions or generate outputs based on new input data.

Kubernetes pods

The smallest deployable units in Kubernetes that can host one or more containers, sharing networking and storage resources.

LLM, large language model

An advanced AI model trained on amounts of text data to understand and generate human-like text. Can perform tasks like translation, summarization and answering questions.

Model weights

The learned parameters of an AI model that determine how it processes inputs and generates outputs. These weights are adjusted during training to optimize model performance.

NLG, natural language generation

A process of automatically generating human-like text from structured data or other forms of input. Designed to convert raw data into coherent and meaningful language easily understood by humans.

NLU, natural language understanding

A process AI uses to analyze and understand the meaning of the input query.

NVIDIA GPU driver

Software that enables communication between the operating system and NVIDIA graphics hardware, essential for GPU-accelerated AI workloads.

NVIDIA GPU Operator

A Kubernetes operator that automates the management of NVIDIA GPUs in container environments, handling driver deployment, runtime configuration, and monitoring.

Ollama

An open source framework for running and serving AI models locally. Ollama simplifies the process of downloading, running and managing large language models.

OpenGL

A cross-platform API for rendering 2D and 3D graphics, commonly used in visualization applications and GPU-accelerated computing.

Prompt Engineering

The practice of crafting effective input queries to AI models to obtain desired and accurate outputs. Good prompt engineering helps prevent hallucinations and improves response quality.

Prompt injection

A security vulnerability where malicious inputs attempt to override or bypass an AI model's system prompt or safety constraints.

Quantization

A technique to reduce AI model size and computational requirements by converting model parameters to lower precision formats while maintaining acceptable performance.

RAG, retrieval-augmented generation

A technique that enhances AI responses by retrieving relevant information from a knowledge base before generating answers, improving accuracy and reducing hallucinations.

RBAC, role-based access control

A security model that restricts system access based on roles assigned to users, managing permissions and authorization in Kubernetes clusters.

Semantic search

A search method using AI to understand the meaning and context of queries rather than just matching keywords, enabling more relevant results.

System prompt

Initial instructions given to an AI model that define its behavior, role and response parameters. System prompts help maintain consistent and appropriate AI responses.

Temperature

A parameter controlling the randomness in AI model outputs. Lower values produce more focused and deterministic responses, while higher values increase creativity and variability.

Token

The basic unit of text processing in AI models, representing parts of words, characters or symbols. Models process text by breaking it into tokens for analysis and generation.

Top-K

A parameter that limits token selection during text generation to the K most likely next tokens, helping control output quality and relevance.

Top-P

Also known as nucleus sampling, a parameter that selects from the smallest set of tokens whose cumulative probability exceeds P, providing dynamic control over text generation diversity.

Vector database

A specialized database designed to store and efficiently query high-dimensional vectors that represent data in AI applications, enabling similarity searches and semantic operations.

Vector store

A specialized storage system optimized for managing and querying vector embeddings, essential for semantic search and RAG implementations in AI applications.

B Copyright

Copyright © 2026–2026-08-05 SUSE LLC and contributors. All rights reserved.

Permission is granted to copy, distribute and/or modify this document under the terms of the GNU Free Documentation License, Version 1.2 or (at your option) version 1.3; with the Invariant section being this copyright notice and license. A copy of the license version 1.2 is included in the section entitled 'GNU Free Documentation License'.

For SUSE trademarks, see <https://www.suse.com/company/legal/> . All third-party trademarks are the property of their respective owners. Trademark symbols (®, ™ etc.) denote trademarks of SUSE and its affiliates. Asterisks (*) denote third-party trademarks.

All information found in this book has been compiled with utmost attention to detail. However, this does not guarantee complete accuracy. Neither SUSE LLC, its affiliates, the authors nor the translators shall be held liable for possible errors or the consequences thereof.

C GNU Free Documentation License

Copyright © 2000, 2001, 2002 Free Software Foundation, Inc. 51 Franklin St, Fifth Floor, Boston, MA 02110-1301 USA. Everyone is permitted to copy and distribute verbatim copies of this license document, but changing it is not allowed.

C1 0. PREAMBLE

The purpose of this License is to make a manual, textbook, or other functional and useful document "free" in the sense of freedom: to assure everyone the effective freedom to copy and redistribute it, with or without modifying it, either commercially or non-commercially. Secondly, this License preserves for the author and publisher a way to get credit for their work, while not being considered responsible for modifications made by others.

This License is a kind of "copyleft", which means that derivative works of the document must themselves be free in the same sense. It complements the GNU General Public License, which is a copyleft license designed for free software.

We have designed this License to use it for manuals for free software, because free software needs free documentation: a free program should come with manuals providing the same freedoms that the software does. But this License is not limited to software manuals; it can be used for any textual work, regardless of subject matter or whether it is published as a printed book. We recommend this License principally for works whose purpose is instruction or reference.

C2 1. APPLICABILITY AND DEFINITIONS

This License applies to any manual or other work, in any medium, that contains a notice placed by the copyright holder saying it can be distributed under the terms of this License. Such a notice grants a world-wide, royalty-free license, unlimited in duration, to use that work under the conditions stated herein. The "Document", below, refers to any such manual or work. Any member of the public is a licensee, and is addressed as "you". You accept the license if you copy, modify or distribute the work in a way requiring permission under copyright law.

A "Modified Version" of the Document means any work containing the Document or a portion of it, either copied verbatim, or with modifications and/or translated into another language.

A "Secondary Section" is a named appendix or a front-matter section of the Document that deals exclusively with the relationship of the publishers or authors of the Document to the Document's overall subject (or to related matters) and contains nothing that could fall directly within that overall subject. (Thus, if the Document is in part a textbook of mathematics, a Secondary Section may not explain any mathematics.) The relationship could be a matter of historical connection with the subject or with related matters, or of legal, commercial, philosophical, ethical or political position regarding them.

The "Invariant Sections" are certain Secondary Sections whose titles are designated, as being those of Invariant Sections, in the notice that says that the Document is released under this License. If a section does not fit the above definition of Secondary then it is not allowed to be designated as Invariant. The Document may contain zero Invariant Sections. If the Document does not identify any Invariant Sections then there are none.

The "Cover Texts" are certain short passages of text that are listed, as Front-Cover Texts or Back-Cover Texts, in the notice that says that the Document is released under this License. A Front-Cover Text may be at most 5 words, and a Back-Cover Text may be at most 25 words.

A "Transparent" copy of the Document means a machine-readable copy, represented in a format whose specification is available to the general public, that is suitable for revising the document straightforwardly with generic text editors or (for images composed of pixels) generic paint programs or (for drawings) some widely available drawing editor, and that is suitable for input to text formatters or for automatic translation to a variety of formats suitable for input to text formatters. A copy made in an otherwise Transparent file format whose markup, or absence of markup, has been arranged to thwart or discourage subsequent modification by readers is not Transparent. An image format is not Transparent if used for any substantial amount of text. A copy that is not "Transparent" is called "Opaque".

Examples of suitable formats for Transparent copies include plain ASCII without markup, Texinfo input format, LaTeX input format, SGML or XML using a publicly available DTD, and standard-conforming simple HTML, PostScript or PDF designed for human modification. Examples of transparent image formats include PNG, XCF and JPG. Opaque formats include proprietary formats that can be read and edited only by proprietary word processors, SGML or XML for which the DTD and/or processing tools are not generally available, and the machine-generated HTML, PostScript or PDF produced by some word processors for output purposes only.

The "Title Page" means, for a printed book, the title page itself, plus such following pages as are needed to hold, legibly, the material this License requires to appear in the title page. For works in formats which do not have any title page as such, "Title Page" means the text near the most prominent appearance of the work's title, preceding the beginning of the body of the text.

A section "Entitled XYZ" means a named subunit of the Document whose title either is precisely XYZ or contains XYZ in parentheses following text that translates XYZ in another language. (Here XYZ stands for a specific section name mentioned below, such as "Acknowledgements", "Dedications", "Endorsements", or "History".) To "Preserve the Title" of such a section when you modify the Document means that it remains a section "Entitled XYZ" according to this definition.

The Document may include Warranty Disclaimers next to the notice which states that this License applies to the Document. These Warranty Disclaimers are considered to be included by reference in this License, but only as regards disclaiming warranties: any other implication that these Warranty Disclaimers may have is void and has no effect on the meaning of this License.

C3 2. VERBATIM COPYING

You may copy and distribute the Document in any medium, either commercially or non-commercially, provided that this License, the copyright notices, and the license notice saying this License applies to the Document are reproduced in all copies, and that you add no other conditions whatsoever to those of this License. You may not use technical measures to obstruct or control the reading or further copying of the copies you make or distribute. However, you may accept compensation in exchange for copies. If you distribute a large enough number of copies you must also follow the conditions in section 3.

You may also lend copies, under the same conditions stated above, and you may publicly display copies.

C4 3. COPYING IN QUANTITY

If you publish printed copies (or copies in media that commonly have printed covers) of the Document, numbering more than 100, and the Document's license notice requires Cover Texts, you must enclose the copies in covers that carry, clearly and legibly, all these Cover Texts: Front-Cover Texts on the front cover, and Back-Cover Texts on the back cover. Both covers must also clearly and legibly identify you as the publisher of these copies. The front cover must present the full title with all words of the title equally prominent and visible. You may add other material on the covers in addition. Copying with changes limited to the covers, as long as they preserve the title of the Document and satisfy these conditions, can be treated as verbatim copying in other respects.

If the required texts for either cover are too voluminous to fit legibly, you should put the first ones listed (as many as fit reasonably) on the actual cover, and continue the rest onto adjacent pages.

If you publish or distribute Opaque copies of the Document numbering more than 100, you must either include a machine-readable Transparent copy along with each Opaque copy, or state in or with each Opaque copy a computer-network location from which the general network-using public has access to download using public-standard network protocols a complete Transparent copy of the Document, free of added material. If you use the latter option, you must take reasonably prudent steps, when you begin distribution of Opaque copies in quantity, to ensure that this Transparent copy will remain thus accessible at the stated location until at least one year after the last time you distribute an Opaque copy (directly or through your agents or retailers) of that edition to the public.

It is requested, but not required, that you contact the authors of the Document well before redistributing any large number of copies, to give them a chance to provide you with an updated version of the Document.

C5 4. MODIFICATIONS

You may copy and distribute a Modified Version of the Document under the conditions of sections 2 and 3 above, provided that you release the Modified Version under precisely this License, with the Modified Version filling the role of the Document, thus licensing distribution and modification of the Modified Version to whoever possesses a copy of it. In addition, you must do these things in the Modified Version:

1. Use in the Title Page (and on the covers, if any) a title distinct from that of the Document, and from those of previous versions (which should, if there were any, be listed in the History section of the Document). You may use the same title as a previous version if the original publisher of that version gives permission.
2. List on the Title Page, as authors, one or more persons or entities responsible for authorship of the modifications in the Modified Version, together with at least five of the principal authors of the Document (all of its principal authors, if it has fewer than five), unless they release you from this requirement.
3. State on the Title page the name of the publisher of the Modified Version, as the publisher.
4. Preserve all the copyright notices of the Document.
5. Add an appropriate copyright notice for your modifications adjacent to the other copyright notices.
6. Include, immediately after the copyright notices, a license notice giving the public permission to use the Modified Version under the terms of this License, in the form shown in the Addendum below.
7. Preserve in that license notice the full lists of Invariant Sections and required Cover Texts given in the Document's license notice.

8. Include an unaltered copy of this License.
9. Preserve the section Entitled "History", Preserve its Title, and add to it an item stating at least the title, year, new authors, and publisher of the Modified Version as given on the Title Page. If there is no section Entitled "History" in the Document, create one stating the title, year, authors, and publisher of the Document as given on its Title Page, then add an item describing the Modified Version as stated in the previous sentence.
10. Preserve the network location, if any, given in the Document for public access to a Transparent copy of the Document, and likewise the network locations given in the Document for previous versions it was based on. These may be placed in the "History" section. You may omit a network location for a work that was published at least four years before the Document itself, or if the original publisher of the version it refers to gives permission.
11. For any section Entitled "Acknowledgements" or "Dedications", Preserve the Title of the section, and preserve in the section all the substance and tone of each of the contributor acknowledgements and/or dedications given therein.
12. Preserve all the Invariant Sections of the Document, unaltered in their text and in their titles. Section numbers or the equivalent are not considered part of the section titles.
13. Delete any section Entitled "Endorsements". Such a section may not be included in the Modified Version.
14. Do not retile any existing section to be Entitled "Endorsements" or to conflict in title with any Invariant Section.
15. Preserve any Warranty Disclaimers.

If the Modified Version includes new front-matter sections or appendices that qualify as Secondary Sections and contain no material copied from the Document, you may at your option designate some or all of these sections as invariant. To do this, add their titles to the list of Invariant Sections in the Modified Version's license notice. These titles must be distinct from any other section titles.

You may add a section Entitled "Endorsements", provided it contains nothing but endorsements of your Modified Version by various parties—for example, statements of peer review or that the text has been approved by an organization as the authoritative definition of a standard.

You may add a passage of up to five words as a Front-Cover Text, and a passage of up to 25 words as a Back-Cover Text, to the end of the list of Cover Texts in the Modified Version. Only one passage of Front-Cover Text and one of Back-Cover Text may be added by (or through arrangements made by) any

one entity. If the Document already includes a cover text for the same cover, previously added by you or by arrangement made by the same entity you are acting on behalf of, you may not add another; but you may replace the old one, on explicit permission from the previous publisher that added the old one.

The author(s) and publisher(s) of the Document do not by this License give permission to use their names for publicity for or to assert or imply endorsement of any Modified Version.

C6 5. COMBINING DOCUMENTS

You may combine the Document with other documents released under this License, under the terms defined in section 4 above for modified versions, provided that you include in the combination all of the Invariant Sections of all of the original documents, unmodified, and list them all as Invariant Sections of your combined work in its license notice, and that you preserve all their Warranty Disclaimers.

The combined work need only contain one copy of this License, and multiple identical Invariant Sections may be replaced with a single copy. If there are multiple Invariant Sections with the same name but different contents, make the title of each such section unique by adding at the end of it, in parentheses, the name of the original author or publisher of that section if known, or else a unique number. Make the same adjustment to the section titles in the list of Invariant Sections in the license notice of the combined work.

In the combination, you must combine any sections Entitled "History" in the various original documents, forming one section Entitled "History"; likewise combine any sections Entitled "Acknowledgements", and any sections Entitled "Dedications". You must delete all sections Entitled "Endorsements".

C7 6. COLLECTIONS OF DOCUMENTS

You may make a collection consisting of the Document and other documents released under this License, and replace the individual copies of this License in the various documents with a single copy that is included in the collection, provided that you follow the rules of this License for verbatim copying of each of the documents in all other respects.

You may extract a single document from such a collection, and distribute it individually under this License, provided you insert a copy of this License into the extracted document, and follow this License in all other respects regarding verbatim copying of that document.

C8 7. AGGREGATION WITH INDEPENDENT WORKS

A compilation of the Document or its derivatives with other separate and independent documents or works, in or on a volume of a storage or distribution medium, is called an "aggregate" if the copyright resulting from the compilation is not used to limit the legal rights of the compilation's users beyond what the individual works permit. When the Document is included in an aggregate, this License does not apply to the other works in the aggregate which are not themselves derivative works of the Document.

If the Cover Text requirement of section 3 is applicable to these copies of the Document, then if the Document is less than one half of the entire aggregate, the Document's Cover Texts may be placed on covers that bracket the Document within the aggregate, or the electronic equivalent of covers if the Document is in electronic form. Otherwise they must appear on printed covers that bracket the whole aggregate.

C9 8. TRANSLATION

Translation is considered a kind of modification, so you may distribute translations of the Document under the terms of section 4. Replacing Invariant Sections with translations requires special permission from their copyright holders, but you may include translations of some or all Invariant Sections in addition to the original versions of these Invariant Sections. You may include a translation of this License, and all the license notices in the Document, and any Warranty Disclaimers, provided that you also include the original English version of this License and the original versions of those notices and disclaimers. In case of a disagreement between the translation and the original version of this License or a notice or disclaimer, the original version will prevail.

If a section in the Document is Entitled "Acknowledgements", "Dedications", or "History", the requirement (section 4) to Preserve its Title (section 1) will typically require changing the actual title.

C10 9. TERMINATION

You may not copy, modify, sublicense, or distribute the Document except as expressly provided for under this License. Any other attempt to copy, modify, sublicense or distribute the Document is void, and will automatically terminate your rights under this License. However, parties who have received copies, or rights, from you under this License will not have their licenses terminated so long as such parties remain in full compliance.

C11 1. FUTURE REVISIONS OF THIS LICENSE

The Free Software Foundation may publish new, revised versions of the GNU Free Documentation License from time to time. Such new versions will be similar in spirit to the present version, but may differ in detail to address new problems or concerns. See <https://www.gnu.org/copyleft/> .

Each version of the License is given a distinguishing version number. If the Document specifies that a particular numbered version of this License "or any later version" applies to it, you have the option of following the terms and conditions either of that specified version or of any later version that has been published (not as a draft) by the Free Software Foundation. If the Document does not specify a version number of this License, you may choose any version ever published (not as a draft) by the Free Software Foundation.

C12 ADDENDUM: How to use this License for your documents

```
Copyright (c) YEAR YOUR NAME.  
Permission is granted to copy, distribute and/or modify this document  
under the terms of the GNU Free Documentation License, Version 1.2  
or any later version published by the Free Software Foundation;  
with no Invariant Sections, no Front-Cover Texts, and no Back-Cover Texts.  
A copy of the license is included in the section entitled "GNU  
Free Documentation License".
```

If you have Invariant Sections, Front-Cover Texts and Back-Cover Texts, replace the "with...Texts." line with this:

```
with the Invariant Sections being LIST THEIR TITLES, with the  
Front-Cover Texts being LIST, and with the Back-Cover Texts being LIST.
```

If you have Invariant Sections without Cover Texts, or some other combination of the three, merge those two alternatives to suit the situation.

If your document contains nontrivial examples of program code, we recommend releasing these examples in parallel under your choice of free software license, such as the GNU General Public License, to permit their use in free software.