

Mettre à jour les modèles dans Librechat

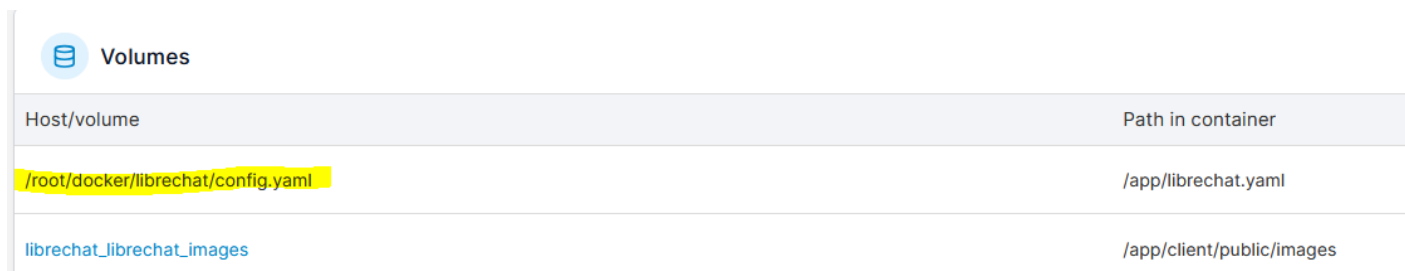
La configuration de Librechat se fait à deux niveaux :

- Dans le fichier de configuration de Librechat
- Dans la stack (docker-compose) de Librechat

Fichier de configuration de Librechat

1. Identifier la localisation du fichier

Aller sur le container Librechat puis sur la partie Volumes

A screenshot of the Docker interface showing the 'Volumes' section for a container named 'librechat'. The table lists two volumes: one for the configuration file and one for images. The configuration file path is highlighted in yellow.

Host/volume	Path in container
<code>/root/docker/librechat/config.yaml</code>	<code>/app/librechat.yaml</code>
<code>librechat_librechat_images</code>	<code>/app/client/public/images</code>

Le fichier peut être édité

- soit dans le container (via `/app/librechat/yaml`)
- soit en SSH sur le serveur (via `/root/docker/librechat/config.yaml` dans le cas présent)

2. Editer le fichier suivant la méthode choisie

Ci-dessous un exemple de configuration pour les modèles Ollama :

```
- name: "Ollama"
```

```
apiKey: "ollama"
# use 'host.docker.internal' instead of localhost if running LibreChat in a docker
container
baseURL: "http://sp26.vincennes.fr:11434/v1/"
models:
  default: [
    "llama3.2-vision:11b", "llama3.1:8b"
  ]
# fetching list of models is supported but the `name` field must start
# with `ollama` (case-insensitive), as it does in this example.
fetch: false
titleConvo: true
titleModel: "current_model"
summarize: false
summaryModel: "current_model"
forcePrompt: false
modelDisplayLabel: "Ollama"
```

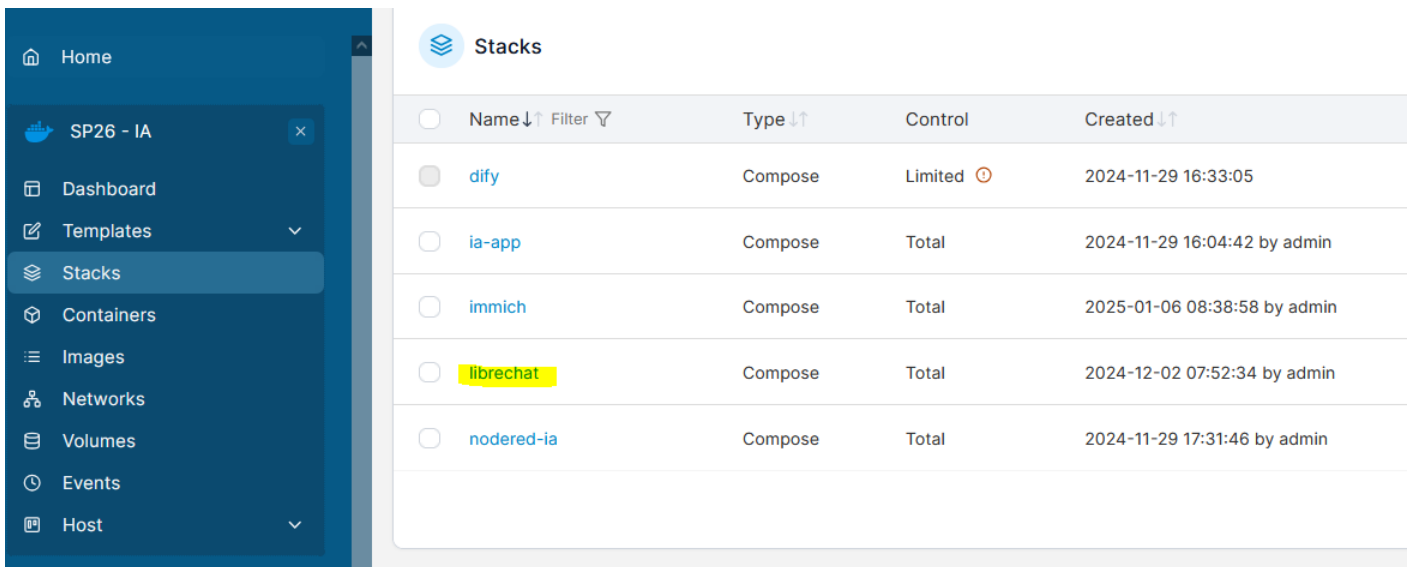
La documentation de Librechat précise la signification des différents champs :

https://www.librechat.ai/docs/configuration/librechat_yaml/object_structure/custom_endpoint

Le numéro de version du fichier config doit être changé

Fichier docker-compose de Librechat

1. Dans stacks, sélectionner Librechat



2. Cliquer sur Editor

3. Modifier la variable d'environnement ENDPOINTS

```
33     - DEBUG_LOGGING=true
34     - DEBUG_CONSOLE=true
35     - ENDPOINTS=ollama
36
37     - PROXY=
```

Elle doit correspondre au champ "name" de config.yaml

4. Embeddings

Dans le cas où le modèle d'embedding du RAG doit être modifié, les options se trouvent sur la partie "rag_api".

```
165     rag_api:
166         container_name: rag_api
167         image: ghcr.io/danny-avila/librechat-rag-api-dev:latest
168         environment:
169             - DB_HOST=vectordb
170             - RAG_PORT=${RAG_PORT:-8000}
171             - EMBEDDINGS_PROVIDER=ollama
172             - OLLAMA_BASE_URL=http://sp26.vincennes.fr:11434/
173             - EMBEDDINGS_MODEL=bge-m3
174         restart: always
175         depends_on:
176             - vectordb
```

Les explications pour les différents champs sur dans la documentation Librechat :

https://www.librechat.ai/docs/configuration/rag_api

5. Cliquer sur Deploy pour mettre à jour Librechat

Tests

Vérifier le bon fonctionnement.

Les logs du container Librechat indiqueront les problèmes éventuels.

Révision #5

Créé 20 February 2025 12:09:39 par Testelin Didier

Mis à jour 9 April 2025 09:39:40 par LE POTTIER Romain