Comment installer un LLM open source en local

Données traitées	Outils	Public cible
Texte	OLLAMA, Docker	Tout public

Budget	Gain	Effort
€€€	★★★	

Prérequis	Connaissance des modèles de langage (LLM), Python, Docker

Contributeur

Abood MOURAD

Docteur en optimisation et recherche opérationnelle (Université Paris-Saclay), il occupe le poste de consultant senior / chef de projet chez EURODECISION. Il intervient sur une diversité de projets d'optimisation et d'aide à la décision.

PROBLÉMATIQUE

Avec l'essor des modèles de langage de grande taille (LLM), de nombreuses entreprises et utilisateurs cherchent à les utiliser pour des applications variées. Cependant, l'accès à des modèles payants peut être limité par leurs coûts élevés ou des restrictions d'utilisation. Installer un LLM open source en local permet non seulement de contourner ces obstacles, mais aussi de **garantir la confidentialité des données** et de **personnaliser le modèle selon des besoins spécifiques**. Cette fiche présente les étapes nécessaires pour installer un LLM open source sur votre propre machine.

EXPLICATION

Un LLM open source est un modèle de traitement du langage naturel accessible au public, permettant à chacun de l'utiliser et de l'améliorer. Des modèles comme **gemma-3**, **phi-3** ou **deepseek-r1** sont disponibles sur les plateformes Hugging Face et OLLAMA.

Contrairement aux modèles en ligne comme ChatGPT ou Claude, qui nécessitent une connexion Internet et l'envoi de données vers des serveurs distants, **les LLM open source peuvent être installés localement, sans transmission d'informations en ligne**.

Nous présentons ici une installation d'un LLM open source via OLLAMA.

Avantages de cette solution :

- ☑ Modèles open-source
- ☑ Petits modèles tournant en local sur une machine avec petit GPU
- ☑ Aucun envoi de données en ligne

HUB FRANCE IA

Intelligence Artificielle Générative
Fiche pratique
Catégorie : Acculturation technique

Comment installer un LLM open source en local

1. Préparer l'environnement : Installer Python/Docker

Assurez-vous d'avoir Python installé sur votre système. Ouvrez un terminal et tapez :

`python -- version` → `Python 3.12.10`

Afin de lancer vos projet python dans des containers, il faut installer docker :

- Sous Windows : https://docs.docker.com/desktop/setup/install/windows-install/
- Sous Linux : https://docs.docker.com/engine/install/

Vous pouvez vérifier l'installation de docker en tapant : `docker run hello-world`

→
```
Hello from Docker!
This message shows that your installation appears to be working correctly.
```

2. Installer un container pour lancer les modèles LLM : OLLAMA

```
docker run ollama/ollama:latest
```

Vous pouvez vérifier l'installation du container OLLAMA :

```
docker ps
```

CONTAINER ID	IMAGE	COMMAND.	CREATED
9add7add3d3fa	Ollama/ollama:latest	"/bin/ollama serve"	3 minutes

3. Une fois docker est lancé, entrer dans votre container OLLAMA en indiquant son ID qui se trouve dans étape 2 :

```
docker exec -it 9aa7aa3d3fa bash
```

4. Vous pouvez désormais choisir un modèle (phi3 par exemple) et le lancer :

```
ollama run phi3:latest
```
→ Posez votre première question au modèle !

```
>>> Pourquoi le ciel est bleu ?
La couleur du ciel pendant la journée terrestre provient principalement de l'effet Rayleigh. Les molécules d'air, comme les molécules d'oxygène et d'azote qui composent une grande partie de notre atmosphère, émettent plus efficacement des longueurs d'onde bleues par rapport aux autres couleurs du spectre visible lorsqu'elles dispersent la lumière blanche solaire.
```

Exemple d'applications

Chatbots : Créez des assistants virtuels capables de répondre aux questions des utilisateurs en langage naturel.

Analyse de sentiment : Utilisez le modèle pour évaluer les émotions dans des textes, comme des avis clients ou des publications sur les réseaux sociaux

POUR ALLER PLUS LOIN

Interface graphique : Explorez les fonctionnalités offertes par Open-WebUI, une interface graphique open source, conviviale et intuitive, qui simplifie l'utilisation et l'exécution des LLM.

Intégration dans des applications : Développez des applications en intégrant le modèle dans des chatbots, des systèmes de recommandation ou des outils d'analyse de texte.