# Déploiement du modèle Phi-3 Mini sur Debian 13

Ce document présente, étape par étape, la procédure complète pour installer et exécuter localement le modèle Phi-3 Mini — un modèle de langage (LLM) développé par Microsoft — sur une distribution Debian 13. L’objectif est de déployer une version légère, fluide et performante d’une intelligence artificielle conversationnelle de type ChatGPT, mais entièrement locale et gratuite.

## 1️⃣ Préparation du système

Avant toute installation, il est essentiel de s’assurer que le système est à jour afin de garantir la stabilité et la compatibilité des dépendances.

* Commande exécutée :
* sudo apt update && sudo apt upgrade -y

## 2️⃣ Installation de l’outil Ollama

Ollama est la plateforme permettant d’exécuter des modèles d’intelligence artificielle localement, sans dépendance cloud. Elle offre une interface simple et compatible avec la majorité des modèles open source.

* Commandes exécutées :
* curl -fsSL https://ollama.ai/install.sh | sh
* ollama --version

## 3️⃣ Vérification du service Ollama

Une fois installé, Ollama crée automatiquement un service systemd pour gérer l’exécution des modèles. Les commandes suivantes permettent de vérifier et de s’assurer que le service tourne correctement :

* systemctl status ollama
* systemctl start ollama
* systemctl enable ollama
* curl http://127.0.0.1:11434/api/version

## 4️⃣ Téléchargement du modèle Phi-3 Mini

Le modèle choisi pour cette mission est \*\*Phi-3 Mini\*\*, développé par Microsoft. Il s’agit d’un modèle de 3,8 milliards de paramètres, capable de générer du texte de haute qualité tout en restant extrêmement léger et rapide. Il constitue une alternative idéale à ChatGPT pour une exécution locale.

* Commande exécutée :
* ollama pull phi3:mini

## 5️⃣ Démarrage du modèle en local

Une fois le modèle téléchargé, il peut être lancé immédiatement depuis le terminal. Le modèle se charge alors en mémoire et attend vos instructions.

* Commande exécutée :
* ollama run phi3:mini

À partir de ce moment, vous pouvez converser directement avec votre IA locale, sans connexion Internet. L’assistant est capable de répondre à vos questions, générer du texte et simuler des dialogues comme ChatGPT.

## 6️⃣ (Optionnel) Interface graphique type ChatGPT

Pour une expérience visuelle similaire à ChatGPT, une interface web peut être déployée via Docker. Cette interface permet d’échanger avec le modèle local dans un environnement moderne et intuitif.

* Commande exécutée :
* docker run -d -p 3000:8080 -e OLLAMA\_API\_BASE\_URL=http://127.0.0.1:11434 --name openwebui ghcr.io/open-webui/open-webui:main
* Accès via navigateur : http://localhost:3000

## 7️⃣ Vérification de la présence du modèle

* Commande exécutée :
* ollama list

Cette commande affiche la liste des modèles disponibles localement. Une ligne similaire à celle-ci devrait apparaître :  
  
phi3:mini 3.8B parameters

## 🌟 Prompts utilisés durant la mission

/ »

• « C’est possible d’avoir une version gratuite de ChatGPT à installer en local sur Debian 13 ? »

• « Je souhaite utiliser le modèle Phi-3 Mini, peux-tu me guider pas à pas dans son installation ? »

• « Peux-tu m’expliquer si ChatGPT-5 peut être déployé en local et quelles sont les alternatives légales ? »

• « Crée-moi un document Word récapitulant les commandes et prompts utilisés pour installer mon LLM local. »

## ✅ Conclusion

Le modèle \*\*Phi-3 Mini\*\* a été installé et configuré avec succès sur Debian 13 grâce à l’outil Ollama. Ce modèle incarne la parfaite alliance entre performance, légèreté et simplicité d’exécution. Il permet de disposer d’un assistant conversationnel fonctionnant entièrement en local, sans dépendance extérieure, illustrant la puissance des modèles open source modernes dans le domaine de l’intelligence artificielle.