Développement d'une application d'intelligence artificielle locale basée sur DeepSeek

Objet du projet :

Dans ce projet, je veux développer une application d'intelligence artificielle qui tourne entièrement en local sur l'ordinateur. L'idée est de pouvoir discuter avec un modèle de langage avancé, un peu comme avec les assistants connus sur internet, mais sans dépendre d'un serveur externe ni d'une connexion permanente. L'application devra répondre aux questions de l'utilisateur et rester utilisable même hors ligne. Pour moi, ce projet est aussi un moyen de mieux comprendre comment fonctionne un LLM et comment on peut vraiment l'intégrer dans une application complète, et pas seulement en théorie.

Contexte du projet :

Aujourd'hui, les modèles de langage de grande taille sont utilisés partout et ont beaucoup fait avancer l'IA, mais la plupart des solutions disponibles passent par le cloud. Cela pose plusieurs problèmes : les données partent sur des serveurs externes, il faut toujours une connexion internet et, dans certains cas, il y a aussi un coût d'utilisation. Depuis quelque temps, des outils comme Ollama permettent cependant de faire tourner des modèles de langage directement sur une machine locale, ce qui ouvre la porte à des assistants IA plus autonomes. C'est dans ce contexte que s'inscrit mon projet, où je cherche à créer une application capable de dialoguer avec un modèle de langage local pour proposer un assistant IA utilisable sans connexion.

Objectifs du projet :

Mon premier objectif est de mettre en place une application d'IA locale que l'utilisateur pourra lancer sur son propre PC pour poser des questions et obtenir des réponses. Je veux aussi profiter de ce projet pour me familiariser avec l'architecture complète d'un système basé sur un LLM, en incluant l'interface utilisateur, le backend, le moteur d'inférence et la partie stockage. J'ai choisi d'utiliser un modèle open-source comme DeepSeek afin de voir concrètement comment l'intégrer, le configurer et essayer d'améliorer un peu son utilisation. Enfin, je souhaite que l'application reste suffisamment modulaire pour pouvoir ajouter plus tard d'autres modèles ou de nouvelles fonctions sans tout réécrire.

Développement d'une application d'intelligence artificielle locale basée sur DeepSeek

Technologies et outils utilisés

La réalisation du projet s'appuiera sur plusieurs technologies complémentaires.

Interface utilisateur :

L'interface graphique sera développée avec Electron, permettant de créer une application desktop multiplateforme à partir de technologies web (HTML, CSS et JavaScript).

Backend API :

Le backend sera développé en Python à l'aide du framework FastAPI, qui permettra de créer une API REST pour gérer les requêtes entre l'interface utilisateur et le modèle d'intelligence artificielle.

Moteur d'inférence IA :

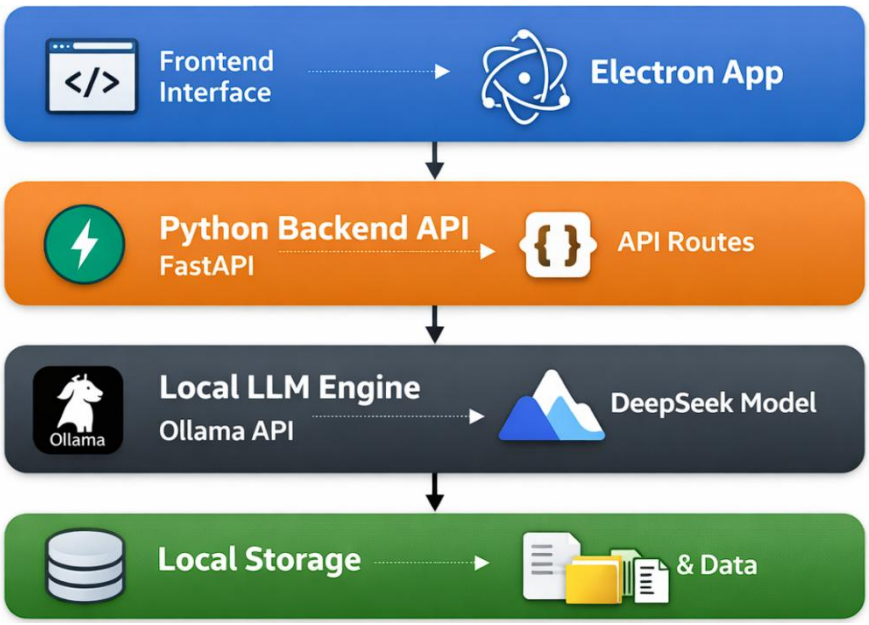
L'exécution du modèle de langage sera assurée par Ollama, qui permettra de charger et d'exécuter localement un modèle de type DeepSeek.

Modèle de langage :

Le projet utilisera un modèle LLM open-source de la famille DeepSeek, conçu pour la génération de texte et la conversation.

Stockage local :

Les données telles que l'historique des conversations pourront être stockées localement à l'aide d'une base de données légère comme SQLite.



Développement d'une application d'intelligence artificielle locale basée sur DeepSeek

Architecture générale du système

L'application est organisée en plusieurs parties qui communiquent entre elles. L'interface utilisateur, développée avec Electron, est la partie visible par l'utilisateur et lui permet d'envoyer des messages à l'IA et de lire les réponses. Derrière, le backend écrit avec FastAPI reçoit ces messages et les transmet au moteur d'inférence. Celui-ci, basé sur Ollama, exécute le modèle de langage DeepSeek et renvoie le texte généré. Les échanges importants, comme l'historique des discussions, sont enregistrés dans une base SQLite pour pouvoir garder une trace des conversations et, éventuellement, améliorer l'outil plus tard.

But du projet

À la fin du projet, je souhaite disposer d'une application d'assistant IA qui fonctionne correctement en local, sans connexion internet obligatoire, tout en respectant la confidentialité des données de l'utilisateur. J'aimerais que l'interface reste assez simple à prendre en main, même pour quelqu'un qui n'est pas développeur, et que les réponses générées soient suffisamment pertinentes pour que l'outil soit réellement utile. Ce travail me permet aussi de progresser sur plusieurs aspects que je vois en cours, comme l'intelligence artificielle, l'architecture logicielle ou encore le développement backend et frontend. C'est une occasion concrète de mettre en pratique ce que j'ai appris pendant ma formation et de mieux comprendre les limites et les contraintes des modèles de langage utilisés dans des applications réelles et comblé des problèmes de lacune en programmation. Pour finir le projet terminal seras mis avec l'intégralité du code seras sur mon portfolio : https://nabilchk.fr/ .