

Créer votre IA Locale

Guide complet de A à Z

Pour les non-techniciens

Qu'est-ce qu'une IA locale ?

Une IA locale fonctionne entièrement sur votre ordinateur, sans connexion Internet.



Confidentialité totale

Vos données restent chez vous



Contrôle complet

Vous maîtrisez tout le système



Sans frais d'abonnement

Pas de coûts récurrents

De quoi avez-vous besoin ?

Matériel

- Ordinateur moderne (Windows, Mac ou Linux)
- 16 à 32 Go de RAM
- Carte graphique recommandée (NVIDIA idéalement)

Logiciels

- Python (langage de programmation)
- Outils d'IA (Ollama ou LM Studio)
- Bibliothèques spécialisées

Rassurez-vous : nous verrons tout pas à pas !

Les 5 grandes étapes

1

Définir votre besoin

Que voulez-vous faire avec votre IA ?

2

Préparer vos données

Rassembler et organiser vos documents

3

Choisir la bonne méthode

RAG ou entraînement personnalisé

4

Installer et configurer

Mettre en place votre système

5


Tester et utiliser

Votre IA est prête !

1

Définir votre besoin

Posez-vous ces questions :

 Que voulez-vous faire ?

Répondre à des questions, résumer des documents, analyser du texte...

 Quelles données avez-vous ?

Documents PDF, notes, emails, historique YouTube...

 Quelles sont vos contraintes ?

Vitesse, confidentialité, budget matériel...

Préparer vos données

Sources possibles

- Documents personnels (PDF, Word)
- Notes et transcriptions
- Historique YouTube (via Google Takeout)
- Pages web et articles
- Emails et messages

Organisation



Nettoyer le texte



Supprimer les doublons



Ajouter des métadonnées



Découper en sections

3 Deux approches principales

RAG

Recherche + Génération

- ✓ Rapide à mettre en place
- ✓ Idéal pour des documents
- ✓ Facilement modifiable
- ✓ Recommandé pour débuter

Fine-tuning

Entraînement personnalisé

- ✓ Plus de contrôle
- ✓ Style personnalisé
- ✓ Raisonnement adapté
- ⚠ Plus technique

Comment fonctionne le RAG ?



1

Indexation

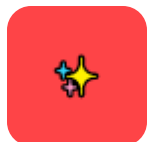
Vos documents sont convertis en vecteurs mathématiques



2

Recherche

L'IA trouve les passages pertinents pour votre question



3

Génération

L'IA formule une réponse basée sur ces passages

4

Outils à installer

Ollama

Interface simple pour faire fonctionner des modèles d'IA

Recommandé pour débuter

Outils complémentaires

Python

Langage de programmation

FAISS ou Chroma

Base de données vectorielle

Transformers

Bibliothèque d'IA

5

Mise en pratique

1

Installer Ollama

2

Télécharger un modèle (Llama 3.1)

3

Indexer vos documents

4

Créer votre système Q&R

5

Tester et affiner !

Avantages et limites

✓ Avantages

- **Confidentialité maximale** - Vos données restent locales
- **Pas de frais récurrents** - Coût unique du matériel
- **Personnalisation totale** - Adapté à vos besoins
- **Disponibilité** - Fonctionne sans Internet
- **Contrôle** - Vous maîtrisez tout

⚠ À considérer

- **Investissement matériel** - Ordinateur performant nécessaire
- **Courbe d'apprentissage** - Demande un peu de temps
- **Maintenance** - Mises à jour à gérer
- **Performance** - Dépend de votre matériel



Glossaire

IA Locale

Intelligence artificielle qui fonctionne sur votre ordinateur

RAG

Recherche + Génération. Méthode qui cherche dans vos docs puis génère une réponse

Embeddings

Représentations mathématiques du texte sous forme de vecteurs

LLM

Large Language Model - Grand modèle de langage

Fine-tuning

Entraînement personnalisé d'un modèle sur vos données

Ollama

Outil simple pour faire tourner des modèles d'IA localement

FAISS

Base de données pour stocker et rechercher des vecteurs

Quantification

Réduction de la taille d'un modèle pour économiser la mémoire

GPU

Carte graphique qui accélère les calculs de l'IA

Chunking

Découpage des documents en petits morceaux pour l'indexation

Prochaines étapes

Vous êtes prêt à commencer !

Commencez simplement avec le RAG et Ollama

1

Installez Ollama

2

Préparez vos données

3

Créez votre premier système
RAG