CRÉEZ VOTRE IA LOCALE 🚀

TITRE - CRÉEZ VOTRE IA LOCALE 🚀

Un guide pratique pour maîtriser l’intelligence artificielle sur votre machine.

QU’EST-CE QU’UNE IA LOCALE ?

- **Confidentialité :** Vos données restent chez vous.
- **Autonomie :** Fonctionne sans connexion internet.
- **Maîtrise :** Contrôle total sur l’IA et son fonctionnement.

DE QUOI AVEZ-VOUS BESOIN ?

MATÉRIEL

- **Processeur (CPU) :** Intel i5/Ryzen 5 (minimum), i7/Ryzen 7 (recommandé).
- **Mémoire vive (RAM) :** 16 Go (minimum), 32 Go ou plus (recommandé).
- **Carte graphique (GPU) :** NVIDIA RTX 3060 (minimum), RTX 4070+ (recommandé) pour de meilleures performances.
- **Stockage :** SSD 500 Go (minimum), 1 To+ (recommandé).

LOGICIELS

LES 5 GRANDES ÉTAPES

1. Définir votre besoin
2. Préparer vos données
3. RAG et Fine-tuning
4. Installation complète
5. Créer votre système RAG !

ÉTAPE 1 : DÉFINIR VOTRE BESOIN

- Quel problème voulez-vous résoudre ?
- Quel type d’IA est le plus adapté ?
- Exemples : assistant personnel, résumé de documents, chatbot.

ÉTAPE 2 : PRÉPARER VOS DONNÉES

- Collecte et extraction (PDF, DOCX, TXT).
- Nettoyage et formatage (suppression HTML, déduplication).
- Exemple de script Python pour le nettoyage.

ÉTAPE 3 : RAG ET FINE-TUNING

RAG (RETRIEVAL AUGMENTED GENERATION)

L'IA "cherche" des informations pertinentes dans une base de connaissances avant de générer une réponse. Idéal pour des réponses factuelles et à jour.

```
# Pseudo-code RAG
query = "Quelle est la capitale de la France ?"
documents = vector_store.retrieve(query) # Recherche
context = combine(documents)
answer = llm.generate(query, context) # Génération
```

RAG (DIAGRAMME MERMAID)

Voici une représentation visuelle du fonctionnement du RAG.

ÉTAPE 4 : INSTALLATION COMPLÈTE

INSTALLER OLLAMA

Téléchargez et installez Ollama depuis ollama.com.

```
# Télécharger un modèle (ex: Llama 3)
ollama pull llama3

# Tester le modèle
ollama run llama3 "Bonjour, comment allez-vous ?"
```

INSTALLER PYTHON ET DÉPENDANCES

Assurez-vous d’avoir Python 3.9+ et installez les bibliothèques :

```
# Vérifier Python
python3 --version
```

VÉRIFICATION ET CHOIX DU MODÈLE

- Script Python pour vérifier l'installation.
- Tableau comparatif des modèles (Llama, Mistral, Phi-3) selon VRAM et qualité.
- Recommandations pour différents budgets matériels.

ÉTAPE 5 : CRÉER VOTRE SYSTÈME RAG !

- Code Python simplifié pour un pipeline RAG.
- Étapes : Import, Chunking, Embeddings, Vectorstore, QA.
- Commentaires en français.

EXEMPLE CONCRET : ASSISTANT DE COURS

- Cas d’usage : Étudiant avec une thèse de 350 pages.
- Résultats : Réduction du temps de recherche de 82%.
- Workflow expliqué.

PROBLÈMES COURANTS & OPTIMISATIONS

PROBLÈMES FRÉQUENTS

- **Erreur GPU :** Pilotes non à jour, VRAM insuffisante.
- **Modèle lent :** Modèle trop grand pour le matériel, pas d’accélération GPU.
- **Réponses imprécises :** Mauvaise qualité des données, chunking inadapté.
- **Ollama non trouvé :** Chemin d’accès incorrect, service non démarré.

ASTUCES D’OPTIMISATION

COMPARAISON LOCAL VS CLOUD

- **Local :** Confidentialité, coût maîtrisé, autonomie.
- **Cloud :** Scalabilité, facilité de déploiement, accès à des modèles plus grands.
- Tableau comparatif des avantages et inconvénients.

CONCLUSION : LANCEZ-VOUS !

L’IA locale est une technologie accessible et puissante qui vous offre contrôle et confidentialité. Commencez par un petit projet, expérimentez et découvrez son potentiel !

MERCI ! QUESTIONS ?

N’hésitez pas à poser vos questions. Contact :
votre.email@example.com