■ Roadmap IA pour Développeurs Web

1. Prérequis

- Langages : Python (pour l'IA) + ton stack habituel (JS/TS, Vue/Nuxt).
- Bases des APIs & REST.
- Git & Docker pour le versionnement et le déploiement.
- Bases en mathématiques (optionnel) : pas nécessaire si tu utilises des modèles pré-entraînés.

2. Découverte de l'IA moderne

- Comprendre la différence entre IA, ML, DL, LLMs.
- Utiliser des APIs pré-entraînées : OpenAI, Hugging Face, Mistral, Anthropic.
- Créer de petits projets : chatbot, résumé de texte, génération d'images.

3. RAG (Retrieval-Augmented Generation)

- Comprendre les embeddings (texte → vecteurs).
- Bases de données vectorielles : Pinecone, Weaviate, Milvus, ou pgvector.
- Pipeline : Indexer → Stocker en vecteurs → Rechercher → Fournir contexte → Appeler le LLM.

4. Agents IA

- Utiliser des frameworks : LangChain, LlamaIndex.
- Connecter les agents à des APIs, bases de données et outils.
- Gérer la mémoire, les outils et le raisonnement multi-étapes.

5. Fine-tuning & Modèles personnalisés

- Fine-tuning avec LoRA/PEFT pour spécialiser un modèle.
- Instruction tuning pour adapter les réponses.
- Utiliser des modèles open-source (LLaMA 3, Mistral, Falcon).

6. Déploiement & Optimisation

- Servir les modèles avec Hugging Face Hub ou via une API custom.
- Utiliser FastAPI/Express + Docker pour la prod.
- Optimiser avec quantization (4bit/8bit).
- Déployer sur VPS, Cloud ou services managés comme Modal/RunPod.

7. Aller plus loin

- Vision + LLM (analyse d'images).
- Voix + LLM (ASR + TTS).
- Sécurité, éthique et gestion des biais.

■ Plan d’apprentissage (3–6 mois)

- Mois 1-2 : Découverte APIs + construire un chatbot ou résumeur.
- Mois 3 : Implémenter un RAG avec une base vectorielle.
- Mois 4 : Créer un agent avec LangChain.
- Mois 5-6 : Fine-tuning + déploiement avec FastAPI/Docker.

■ Références & Ressources

Bases & Prérequis

• Python Basics
• Docker Guide

Découverte de l’IA moderne

• Documentation OpenAI
• API Hugging Face

RAG & Embeddings

• Pinecone Docs
• Weaviate Docs
• pgvector (Postgres)

Agents IA

• LangChain Docs
• LlamaIndex Docs

Fine-tuning & Modèles

• LoRA / PEFT
• Hugging Face Models Hub

Déploiement & Optimisation

• FastAPI Docs
• RunPod (GPU Cloud)

Extensions (Vision, Voix, Multimodalité)

• OpenAI Whisper
• Hugging Face Speech Models