# AI / LLM / Chatbot & Automation Interview Preparation

**Brahim Ouhammou**

## 1. Self-Introduction (30 seconds)

My name is **Brahim Ouhammou**. I am a Master's student in **Artificial Intelligence and Machine Learning**, and I am also studying at **1337 Coding School**. I have experience in **Python, C/C++, Docker**, and I have worked on machine learning and deep learning projects, particularly CNN-based models and AI-driven systems. I am highly motivated to work on **LLMs, chatbots, automation, and RAG systems**, and I am eager to learn and contribute to real-world AI solutions.

## 2. What is a Large Language Model (LLM)?

A Large Language Model (LLM) is a neural network trained on massive amounts of text data to understand and generate human-like language. It works by predicting the next token based on context and can perform tasks such as:

- Question answering
- Text summarization
- Translation
- Conversational chatbots

Examples include OpenAI GPT models, LLaMA, and Mistral.

## 3. How Does an AI Chatbot Work?

A typical AI chatbot pipeline:

1. User sends a message
2. Text is tokenized and converted into embeddings
3. The LLM processes the context
4. Optional: retrieve external data (documents, APIs, databases)
5. The chatbot generates and returns a response

Modern chatbots often use **Retrieval-Augmented Generation (RAG)** to improve accuracy and reduce hallucinations.

## 4. Retrieval-Augmented Generation (RAG)

RAG is a technique that combines a language model with an external knowledge base. Instead of relying only on training data, the system retrieves relevant documents and injects them into the prompt before generation.

**RAG Pipeline:**

1. Documents are converted into embeddings
2. Stored in a vector database
3. User query is embedded
4. Similarity search retrieves relevant documents
5. Context + prompt is sent to the LLM

**Benefits:**

- Reduces hallucinations
- Enables domain-specific knowledge
- Keeps information up to date

## 5. Embeddings and Vector Databases

Embeddings are numerical vector representations of text that capture semantic meaning. Texts with similar meanings have similar embeddings.

Common vector databases:

- FAISS
- Pinecone
- Chroma
- Weaviate

## 6. Automation with Chatbots

AI chatbots can automate repetitive tasks such as:

- Customer support
- Data retrieval
- Report generation
- API orchestration

They can integrate with APIs, databases, and no-code/low-code platforms to trigger intelligent workflows.

## 7. Tools and Frameworks

Relevant tools and technologies:

- Python
- LangChain
- LlamaIndex
- REST APIs

These frameworks help build structured, scalable chatbot and RAG pipelines.

## 8. Prompt Engineering

Prompt engineering is the practice of designing effective instructions to guide an LLM's behavior and output quality. It includes:

- Role prompting
- Few-shot examples
- Output constraints
- Structured formatting

## 9. Research and Learning Mindset

I stay updated by:

- Reading documentation and technical blogs
- Exploring open-source GitHub projects
- Building small experimental prototypes

## 10. Why This Internship?

This internship aligns perfectly with my interest in applied AI. I want to work on real-world systems involving LLMs, chatbots, and automation, while learning from experienced engineers and contributing with my technical skills and strong motivation.

# Préparation à l'Entretien
# AI / LLM / Chatbot & Automatisation

**Brahim Ouhammou**

## 1. Présentation personnelle (30 secondes)

Je m'appelle **Brahim Ouhammou**. Je suis étudiant en **Master Intelligence Artificielle et Machine Learning**, et également étudiant à **l'école 1337**. J'ai une solide base en **programmation (Python, C/C++)** et une expérience avec **Docker**. J'ai travaillé sur plusieurs projets en **machine learning et deep learning**, notamment des modèles basés sur les CNN et des systèmes IA appliqués. Actuellement, je m'intéresse particulièrement aux **LLM, chatbots intelligents, systèmes d'automatisation et pipelines RAG**, et je suis très motivé pour apprendre et contribuer à des projets concrets.

## 2. Qu'est-ce qu'un Large Language Model (LLM) ?

Un **Large Language Model (LLM)** est un modèle de réseau de neurones entraîné sur de très grandes quantités de données textuelles afin de comprendre et générer du langage naturel. Il fonctionne en prédisant le prochain token à partir du contexte et peut réaliser plusieurs tâches :

- Réponse à des questions
- Résumé de texte
- Traduction
- Chat conversationnel

Des exemples de LLM incluent les modèles d'OpenAI (GPT), LLaMA et Mistral.

## 3. Fonctionnement d'un Chatbot IA

Le fonctionnement général d'un chatbot IA se déroule en plusieurs étapes :

1. L'utilisateur envoie une requête
2. Le texte est transformé en embeddings
3. Le LLM analyse le contexte
4. Optionnellement, le système récupère des informations externes (documents, APIs, bases de données)

5. Le chatbot génère et retourne une réponse

Les chatbots modernes utilisent souvent la **Retrieval-Augmented Generation (RAG)** afin d'améliorer la précision et de réduire les hallucinations.

## 4. Retrieval-Augmented Generation (RAG)

La **RAG** est une approche qui combine un LLM avec une base de connaissances externe. Au lieu de se baser uniquement sur les données d'entraînement, le système récupère des documents pertinents et les injecte dans le prompt avant la génération de la réponse.

**Pipeline RAG :**

1. Conversion des documents en embeddings
2. Stockage dans une base de données vectorielle
3. Transformation de la requête utilisateur en embedding
4. Recherche de similarité
5. Envoi du contexte et du prompt au LLM

**Avantages :**

- Réduction des hallucinations
- Accès à des connaissances spécifiques au domaine
- Mise à jour facile des informations

## 5. Embeddings et Bases de Données Vectorielles

Les **embeddings** sont des représentations numériques du texte qui capturent le sens sémantique. Deux textes ayant un sens proche auront des vecteurs similaires.

Exemples de bases de données vectorielles :

- FAISS
- Pinecone
- Chroma
- Weaviate

## 6. Automatisation avec les Chatbots

Les chatbots IA permettent d'automatiser de nombreuses tâches répétitives :

- Support client

- Recherche d'informations
- Génération de rapports
- Orchestration d'APIs

Ils peuvent être intégrés à des APIs, bases de données et plateformes no-code / low-code pour déclencher des workflows intelligents.

## 7. Outils et Frameworks

Outils et technologies utilisés dans ce domaine :

- Python
- LangChain
- LlamaIndex
- APIs REST

Ces frameworks facilitent la création de chatbots avancés et de pipelines RAG.

## 8. Prompt Engineering

Le **prompt engineering** consiste à concevoir des instructions efficaces afin de guider le comportement et la qualité des réponses d'un LLM. Il inclut notamment :

- Role prompting
- Few-shot prompting
- Contraintes de format
- Structuration des réponses

## 9. Esprit de Recherche et Veille Technologique

Je maintiens une veille technologique active en :

- Lisant la documentation et des articles techniques
- Explorant des projets open-source sur GitHub
- Réalisant des prototypes expérimentaux

## 10. Pourquoi ce Stage ?

Ce stage correspond parfaitement à mon intérêt pour l'IA appliquée. Je souhaite travailler sur des systèmes réels basés sur les LLM, les chatbots et l'automatisation, tout en apprenant auprès d'ingénieurs expérimentés et en apportant ma motivation, ma rigueur et mes compétences techniques.