## **Project Overview: Building a Discord RAG Bot**

Welcome to your first AI project!

This assignment challenges you to build a **Discord Retrieval-Augmented Generation (RAG) Bot**. This bot will be capable of answering questions based on a specific knowledge base, going beyond what a standard Large Language Model (LLM) might know by "retrieving" relevant information before "generating" a response.

**What is a RAG Bot?** Imagine you want an AI to answer questions about this AI Internship program's internal documentation, or a training material. A regular LLM might hallucinate or not have the most up-to-date information. A RAG bot solves this by:

1. **Retrieval:** When a user asks a question, the bot first searches a dedicated knowledge base (e.g., a collection of documents) to find the most relevant pieces of information.
2. **Augmentation:** It then takes these retrieved pieces of information and "augments" the user's original query.
3. **Generation:** Finally, it sends this augmented query (original question + retrieved context) to an LLM, which uses the provided context to generate a more accurate and grounded answer.

**Why Discord?** Discord provides a familiar, real-world platform for deploying and interacting with your bot, allowing for easy testing and user feedback.

**Various Role Involvement:** Each role (Data Scientist, Frontend Engineer, Backend Engineer) has distinct responsibilities. See the tabs on the left side for role specific assignments.

**Key Concepts to Understand:**

* **Chunking:** Breaking down large documents into smaller, manageable pieces (chunks).
* **Embeddings:** Converting text chunks into numerical vectors, representing their semantic meaning.
* **Vector Database:** A specialized database optimized for storing and searching these vector embeddings (e.g., MongoDB Atlas Vector Search).
* **Vector Search:** Finding the most similar vectors (and thus, text chunks) to a given query vector.
* **LLMs:** Large Language Models, the "brain" that generates human-like text.
* **API Design:** How different parts of your application communicate with each other.
* **Deployment:** Getting your application running in a live environment.
* **Logging & Observability:** Monitoring your application's health and performance.
* **User Experience (UX) in AI:** Designing intuitive and effective ways for users to interact with AI.

Note: To assist with the development process, you are free to use any AI coding, feel free to to use Azure AI Foundry and the free DeepSeek model.

1. YT Deploy DeepSeek R1 using Azure AI Foundry and Build a Web Chatbot | No Charges for API Use =<https://www.youtube.com/watch?v=pj2knBX4S1w>
2. MSFT =<https://azure.microsoft.com/en-us/pricing/purchase-options/azure-account> [Available only to new Azure customers = $200 credit]
3. MSFT Billing for Free Trial Account =<https://learn.microsoft.com/en-us/answers/questions/2283380/billing-for-free-trial-account>
4. MSFT - DeepSeek R1 is now available on Azure AI Foundry and GitHub =<https://azure.microsoft.com/en-us/blog/deepseek-r1-is-now-available-on-azure-ai-foundry-and-github/>

Go through all of the links above so you have a good idea

Good luck, and remember to have fun learning!

### Helpful Resources:

**Here are some helpful resources that could be used in your Discord Chatbot journey.**

* Youtube:  [Code a Discord Bot with JavaScript](https://www.youtube.com/watch?v=7rU_KyudGBY)
* Youtube: [Code a Discord Bot with Python](https://www.youtube.com/watch?v=SPTfmiYiuok)
* Youtube: [Learn Discord.js V14](https://www.youtube.com/watch?v=Dgy3EJ3HtMw)
* [Workshop created by **Mentor Apporva**](https://github.com/mongodb-developer/genai-devday-notebooks/blob/main/notebooks/ai-rag-lab.ipynb) that walks through the end-to-end process of chunking, embedding, ingestion, implementing vector search, RAG, and adding memory from scratch, mostly without any frameworks, except LangChain for chunking
* Youtube: [Deploying a discord bot using docker](https://www.youtube.com/watch?v=z58g7_dHeMA)
* [Build a RAG System Using Claude 3 Opus And MongoDB](https://www.mongodb.com/developer/products/atlas/rag_with_claude_opus_mongodb/)

### Source Documents the Discord Chatbot should Reference:

1. [AI Bootcamp Journey & Learning Path](https://docs.google.com/document/d/18O8Xpbognhi3GYJeHDYbBRHGAl5-a4DL/edit?usp=sharing&ouid=108974253833672780221&rtpof=true&sd=true) - Also shows high-level schedule
2. [Training For AI Engineer Interns](https://docs.google.com/document/d/1NfmEDyxrJ7Tz7Wq4lAJHx1fQ4bPpM5v07dPTM4pjOsM/edit?usp=sharing)  - Many video links in this document
3. [Intern FAQ - AI Bootcamp](https://docs.google.com/document/d/1RDsjDgszRw5yvjca24aqhmQu2UCrdWdVB3Bb_jd6V6Q/edit?usp=sharing) =Frequently asked questions

### Office Hours *(for weeks 1-3)*:

* Please look for announcements in Discord for exact dates/times of Office Hours.
* Use Office Hours to help with roadblocks rather than as a gatekeeper holding you back. This way you start coding and make progress while not waiting for scheduled weekly Office hour sessions.