Text

Description automatically generated with low confidence

AI-Powered Coffee Shop Chatbot

**EAI 6010 – Applications of Artificial Intelligence**

Final Project Draft

Prof Sergiy Shevchenko

Group: Husky-2

Rohit Kosamkar

Sapna Chavan

Leena Doultani

Harshal Joshi

## Abstract

This project focuses on developing an AI-powered chatbot tailored for a coffee shop's customer service operations. Leveraging Large Language Models (LLMs), Python, RunPod(Cloud Computing), Hugging Face, React Native, Pinecone, Firebase, and Retrieval-Augmented Generation (RAG), the chatbot is designed to take orders, provide menu details, filter irrelevant conversations, and offer product recommendations based on a custom recommendation engine. The implementation aims to enhance customer experience, provide scalability, and deliver personalized interactions. Challenges such as data privacy, model performance, and understanding complex queries are also addressed.

### Project Description

In the evolving landscape of artificial intelligence, chatbots have become integral to enhancing customer experiences across various industries. This project entails the development of an AI-powered chatbot specifically designed for a coffee shop's customer service operations. The chatbot is equipped to take orders, provide detailed menu information, filter out irrelevant conversations, and make intelligent product recommendations based on customer interactions.

### Technology Choice

The selection of technologies for this project includes Large Language Models (LLMs), Python, RunPod, Hugging Face, React Native, Pinecone, Firebase, and Retrieval-Augmented Generation (RAG). These choices are predicated on their industry relevance, scalability, and efficiency in deploying intelligent chatbots.

* **Hugging Face:** Offers access to pre-trained LLMs and embedding models, enhancing the chatbot's intelligence.
* **Llama 3.1-70B-Instruct:** Pre-trained LLM offers advanced natural language understanding and response generation capabilities.
* **Python:** Provides a robust ecosystem for AI development, facilitating seamless integration of machine learning and chatbot functionalities.
* **RunPod:** Enables cost-effective, scalable serverless deployment for AI models, reducing infrastructure complexity.
* **Pinecone:** Serves as a vector database that indexes and stores vector embeddings for fast retrieval and similarity search, enhancing the chatbot's ability to provide relevant responses.
* **Retrieval-Augmented Generation (RAG):** Combines the strengths of traditional information retrieval systems with generative AI, allowing the chatbot to access external knowledge bases to generate more accurate and contextually relevant responses.
* **Firebase:** Acts as a real-time database and backend service, managing coffee shop products details, authentication, and synchronization across devices.
* **React Native:** Ensures seamless, cross-platform mobile application experience for customers.

### Benefits

1. **Enhanced Customer Experience:** The chatbot provides prompt responses to customer inquiries, streamlines order processing, and offers personal recommendations, thereby improving customer satisfaction.
2. **Scalability and Cost Efficiency:** Utilizing RunPod's serverless LLM deployment allows businesses to pay only for active usage, making it an affordable solution.
3. **Personalized Recommendations:** Employing Market Basket Analysis enables the chatbot to suggest complementary products, increasing the average order value.
4. **Context-Aware Responses:** Retrieval-Augmented Generation (RAG) allows the chatbot to provide personalized and accurate responses based on specific coffee shop data.
5. **Automation and Efficiency:** The AI chatbot reduces the workload on human employees, allowing them to focus on more complex customer interactions.

### Drawbacks and Challenges

While the project offers numerous advantages, it also presents certain challenges:

* **Model Performance and Latency:** Deploying LLMs in serverless environments like RunPod may introduce latency in response generation.
* **Understanding Complex Queries:** Despite advanced NLP capabilities, the chatbot may struggle with highly ambiguous or sarcastic customer inputs.
* **Initial Deployment Costs:** Although long-term costs are optimized, the initial deployment and model fine-tuning require substantial computational resources and expertise.

### Project Demonstration

The chatbot is implemented as a full-stack AI-powered application with the following features:

1. **User Interface (React Native):** Customers interact with the chatbot through a mobile-friendly interface.
2. **Backend (Python and RunPod):** The chatbot is hosted on RunPod’s serverless LLM endpoint, providing real-time responses.
3. **Order Processing:** Customers can place orders, customize their preferences, and receive order confirmations.
4. **Recommendation Engine:** A Market Basket Analysis-based AI recommendation system suggests relevant products to customers.
5. **Data Integration:** The chatbot retrieves menu details and coffee shop information from a Pinecone vector store to ensure accurate responses.
6. **Deployment and Scaling:** RunPod and Hugging Face models are utilized to deploy and manage chatbot functionalities in a cloud-based, scalable environment.

### Conclusion

This project demonstrates the potential of AI-powered chatbots in the food and beverage industry, making customer interactions smarter, more efficient, and engaging. By leveraging the capabilities of LLMs, RunPod, and React Native, the chatbot offers an end-to-end intelligent solution for coffee shop businesses. As AI continues to advance, such technology will pave the way for more intuitive, context-aware, and personalized customer experiences in the retail and hospitality sectors.

**References**

American Psychological Association. (2019). *Publication manual of the American Psychological Association* (7th ed.). https://doi.org/10.1037/0000165-000

Purdue Online Writing Lab. (n.d.). *APA sample paper*. Purdue University. Retrieved from https://owl.purdue.edu/owl/research\_and\_citation/apa\_style/apa\_formatting\_and\_style\_guide/apa\_sample\_paper.html