DataDialect: Bridging Databases and Users

PROJECT SYNOPSIS

OF MAJOR PROJECT

BACHELOR OF TECHNOLOGY

COMPUTER SCIENCE & ENGINEERING

SUBMITTED BY:

JASPREET SINGH (2004600)

KESHAV KUMAR ARRI (2004610)

UNDER THE GUIDANCE OF MANPREET KAUR MAND

Department of Computer Science and Engineering

GURU NANAK DEV ENGINEERING COLLEGE,

LUDHIANA

TABLE OF CONTENTS

INTRODUCTION

The project aims to build an End-to-End Language Model (LLM), leveraging the power of advanced technologies to create a robust and efficient question-answer system. The system will be capable of interacting with a MySQL database to generate accurate and relevant responses, thereby enhancing the user experience, and providing valuable insights.

The project will utilize the Google Palm LLM model, a state-of-the-art language model known for its superior performance in understanding and generating human-like text. This model will be integrated with Hugging Face embeddings, a popular choice for natural language processing tasks due to their ability to capture semantic meanings of words. The project will be built on the Langchain framework, which provides a seamless interface for integrating various components of the system.

The project will also incorporate Chroma DB as the vector database, which is known for its high-speed search and scalability. The MySQL database will serve as the backbone for storing and retrieving data. The interface will be designed to be user-friendly and intuitive, ensuring ease of use for end-users. This project, with its innovative approach and use of cutting-edge technologies, promises to revolutionize the way we interact with databases in the retail domain.

.

RATIONALE

- **Need for Efficient Database Interaction:** In the retail domain, vast amounts of data are generated daily. Efficient interaction with this data is crucial for making informed decisions and providing excellent customer service. The proposed system will allow users to interact with the MySQL database using natural language, making data retrieval more intuitive and less time-consuming.
- **Utilization of Advanced Technologies:** The project leverages cutting-edge technologies like the Google Palm LLM model and Hugging Face embeddings. These technologies have proven their effectiveness in various natural language processing tasks. By integrating these technologies, the project aims to build a system that understands and generates human-like text, enhancing the overall user experience.
- **Scalability and Speed with Chroma DB:** Chroma DB, known for its high-speed search and scalability, will be used as the vector database. As the retail domain often deals with large volumes of data, having a scalable and fast database is essential. Chroma DB meets these requirements, making it an excellent choice for this project.
- **User-Friendly Interface:** A user-friendly interface is key to ensuring that the system can be used by individuals with varying levels of technical expertise. The project aims to design an intuitive interface that allows users to easily interact with the system and retrieve the information they need.

OBJECTIVES

1. Integrate Google Palm LLM and Hugging face embeddings for precise natural language understanding in question-and-answer system.
2. Utilize Langchain's SQL Database chain to establish secure communication with MySQL, enhancing data retrieval efficiency.
3. Utilize Chroma DB vector database for optimized storage and retrieval of embeddings, ensuring scalability.
4. Apply few-shot learning techniques to enhance the system's adaptability and responsiveness in generating accurate answers from the database.

FEASIBILTY STUDY

- Technical feasibility: The project is technically feasible as it leverages established technologies like the Google Palm LLM model, Hugging Face embeddings, and MySQL. These technologies have been widely used and tested, ensuring the project's technical viability.
- Legal feasibility: The project adheres to all relevant data protection and privacy laws. It uses open-source technologies and respects all licensing agreements, ensuring legal compliance.
- Economic feasibility: The project is economically feasible. The use of open-source technologies helps to keep costs low, while the potential benefits of a more efficient database interaction system in the retail domain provide a strong economic justification.

METHODOLOGY USED

- **Dataset Collection and Loading in Database:** The first step involves collecting the necessary data from the retail domain. Once collected, the data is loaded into the MySQL database for storage and retrieval.
- **Embedding:** We will create a dataset for out of box queries, Hugging Face is used to generate embeddings for the data. These embeddings capture the complex understanding of queries and are crucial for improving the accuracy of the answers.
- **Storing in Vector Database:** The generated embeddings are then stored in Chroma DB, the vector database. Chroma DB is known for its high-speed search and scalability, making it an excellent choice for handling large volumes of data.
- **Prompting to LLM:** The stored embeddings are used as prompts to the Google Palm LLM model. The model uses these prompts to generate human-like text, which forms the basis of the question-answer system.
- **Interface:** Finally, an intuitive and user-friendly interface is developed. This interface allows users to interact with the system, ask questions in natural language, and receive accurate and relevant responses from the MySQL database.

FACILITIES REQUIRED

- **Hardware Requirements:**

  1. **High-Performance Computer:** A high-performance computer with a fast processor and high RAM capacity is required to handle the computational demands of training the Google Palm LLM model and processing large volumes of data.
  2. **Storage Resources:** Adequate storage space is necessary to store the retail domain data, the trained model, and the generated embeddings.

- **Software Requirements:**

  1. **Langchain Framework:** The Langchain framework is necessary for integrating the different components of the system and facilitating their interaction.
  2. **Google Palm LLM and Hugging Face Libraries:** These libraries are required for training the language model and generating embeddings, respectively.
  3. **MySQL and Chroma DB:** These database systems are needed for storing the retail domain data and the generated embeddings, respectively.

Role of Team Members

Module 1: Data Collection and Loading

Team member 1: Jaspreet Singh

1. **Data Collection and Loading in Database:** Member 1 will be responsible for gathering the necessary data from the retail domain and loading it into the MySQL database.
2. **Embedding Generation:** After the data is loaded into the database, Member 1 will use Hugging Face to generate embeddings for the data.

Module 2: System Integration and Interface Development

Team member 2: Keshav Kumar Arri

1. **Storing in Vector Database:** Once the embeddings are generated, Member 2 will store them in Chroma DB.
2. **System Integration:** Member 2 will then integrate the trained model, the embeddings, and the databases into the Langchain framework.
3. **Interface Development:** Finally, will develop the user-friendly interface that allows users to interact with the system.

REFERENCES

1. A. Chowdhery et al., “PaLM: Scaling Language Modeling with Pathways,” arXiv preprint arXiv:2204.02311, 2022.
2. S. Doddapaneni et al., “User Embedding Model for Personalized Language Prompting,” arXiv preprint arXiv:2401.04858, 2024.
3. O. Topsakal and T. C. Akinci, “Creating Large Language Model Applications Utilizing LangChain: A Primer on Developing LLM Apps Fast,” in Proceedings of the 5th International Conference on Applied Engineering and Natural Sciences, Konya, Turkey, 2023.
4. “Learn How to Use Chroma DB: A Step-by-Step Guide,” DataCamp, Aug. 2023.
5. “Optimizing MySQL database system on information systems research, publications and community service,” in 2016 3rd International Conference on Information Technology, Computer, and Electrical Engineering (ICITACEE), 2016.