INTEL UNNATI

Tanisha Keshavan

Introduction

In the dynamic landscape of artificial intelligence, Generative AI (GenAI) is at the forefront of transforming how we interact with machines. One notable advancement in this field is the development and deployment of the "phi3" model, an advanced generative language model designed to understand and generate human-like text. To maximize the performance and efficiency of such sophisticated models, leveraging optimized inference frameworks like Intel's OpenVINO is essential. OpenVINO enhances the model's ability to run efficiently on various hardware, ensuring rapid and accurate response generation. This synergy between GenAI, phi3, and OpenVINO not only enhances the model's utility but also showcases the potential of combining state-of-the-art AI technologies for practical applications in various domain

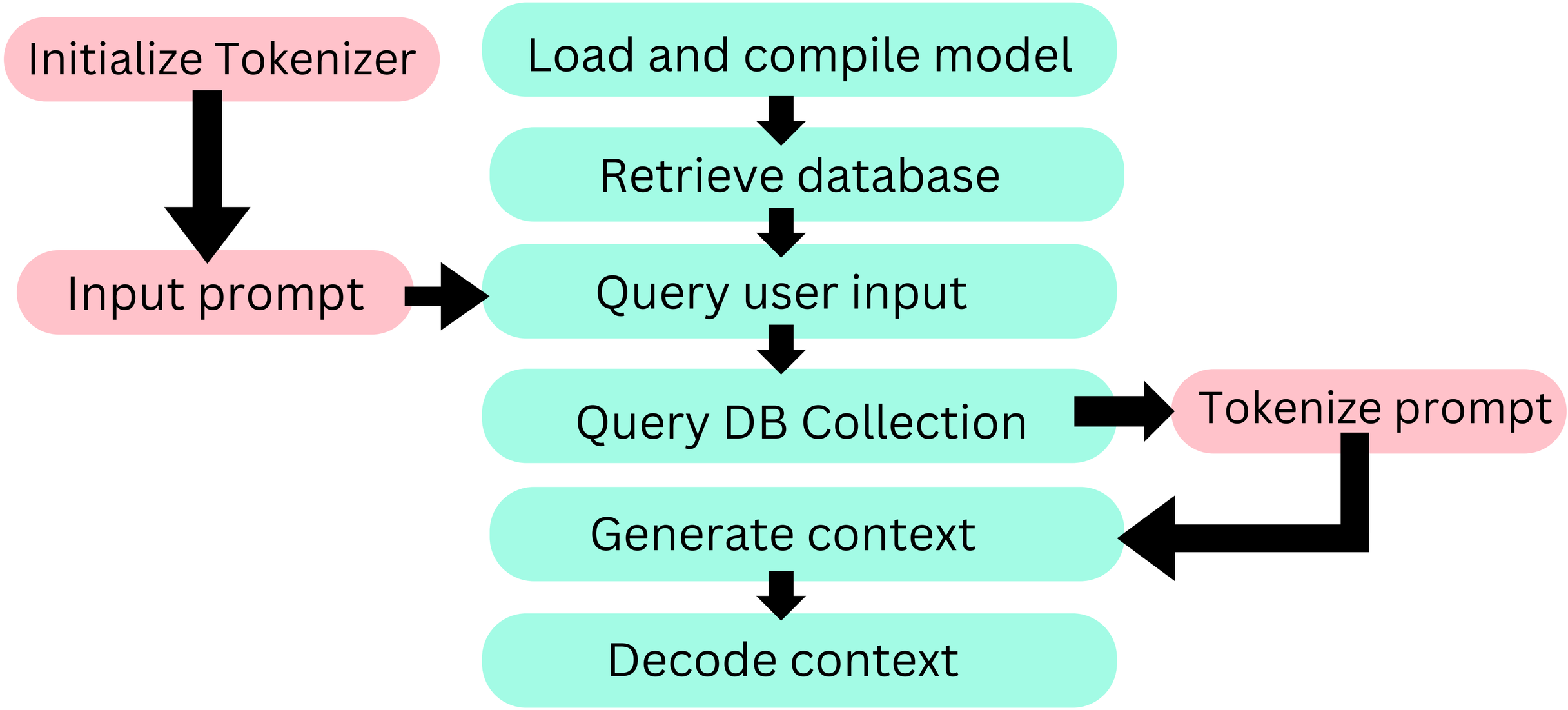
Objectives

- Developing a Gen-AI system to generate responses to input prompts using pre-trained language model.
- Creating a conversational AI system to generate answers for queries from PDFs.
- The system should retrieve relevant information from a database, generate context for the query, and use a language model to provide concise responses.

Dependencies

- Hugging Face Transformers
- NLTK (Natural Language Toolkit)
- Scikit-learn (sklearn)
- Openvino
- ChromaDB
- SentenceTransformer

Project Flow



Model Information

- The model used is **Microsoft’s Phi-3** which represents a series of advanced small language models (SLMs) developed by Microsoft. These models are designed to offer high performance and cost-effectiveness, outperforming other models of similar or larger sizes across various benchmarks.
- **OpenVINO** completes the call of generative AI by running quantitative models.It to quantized the Phi-3 model first and complete the model quantization on the command line through **optimum-cli**
- In this project, the model was downloaded and converted from a public source using the OpenVINO integration with Hugging Face Optimum and compress model weights to 4-bit or 8-bit data types using **NNCF**

Project Flow

1. **Database Interaction**: The system retrieves relevant information from ChromaDB. This step involves traditional database query techniques.
2. **Query Processing**: The user's input is processed and passed to the database querying function.
3. **Context Generation**: The results from the database are processed to generate a context. This context is crucial for providing relevant information to the language model.
4. **Prompt Preparation**: The context and the user's query are formatted into a prompt. This step involves structuring the information in a way that the language model can effectively use to generate responses.
5. **LLM Invocation**: The language model is invoked with the prepared prompt. The LLM then generates a response based on its understanding of the prompt, showcasing its generative capabilities.
6. **User Interaction**: The generated response is presented to the user, providing them with the information they requested.

Conclusion

The integration of Generative AI, specifically through the advanced phi3 model, with the performance optimization capabilities of Intel's OpenVINO toolkit, demonstrates a powerful approach to developing responsive and efficient AI systems. By leveraging ChromaDB for effective data management and embedding functionalities, this project showcases the potential of combining state-of-the-art technologies to create an intelligent system capable of handling complex queries with accuracy and speed.Overall, this project underscores the importance of combining advanced AI models, performance optimization tools, and robust data management practices to develop intelligent systems that can meet the demands of modern applications.