Applied Data Science Department

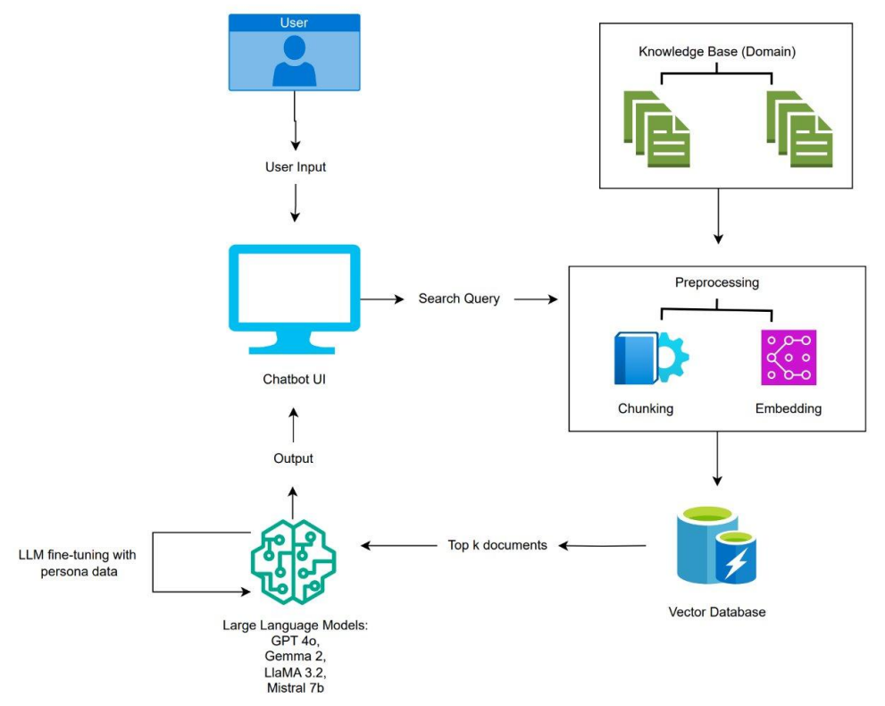
SageSoulRAG: RAG-based Chatbot with Personality

Project Advisor: Simon Shim

Rathnakar, AradhyaAlva
Basavaraju, BhavanKumar
Panda, Mahamaya
ReddyChappidi, ReddySaketh
KadariMallikarjuna, ShashiKumar

Introduction

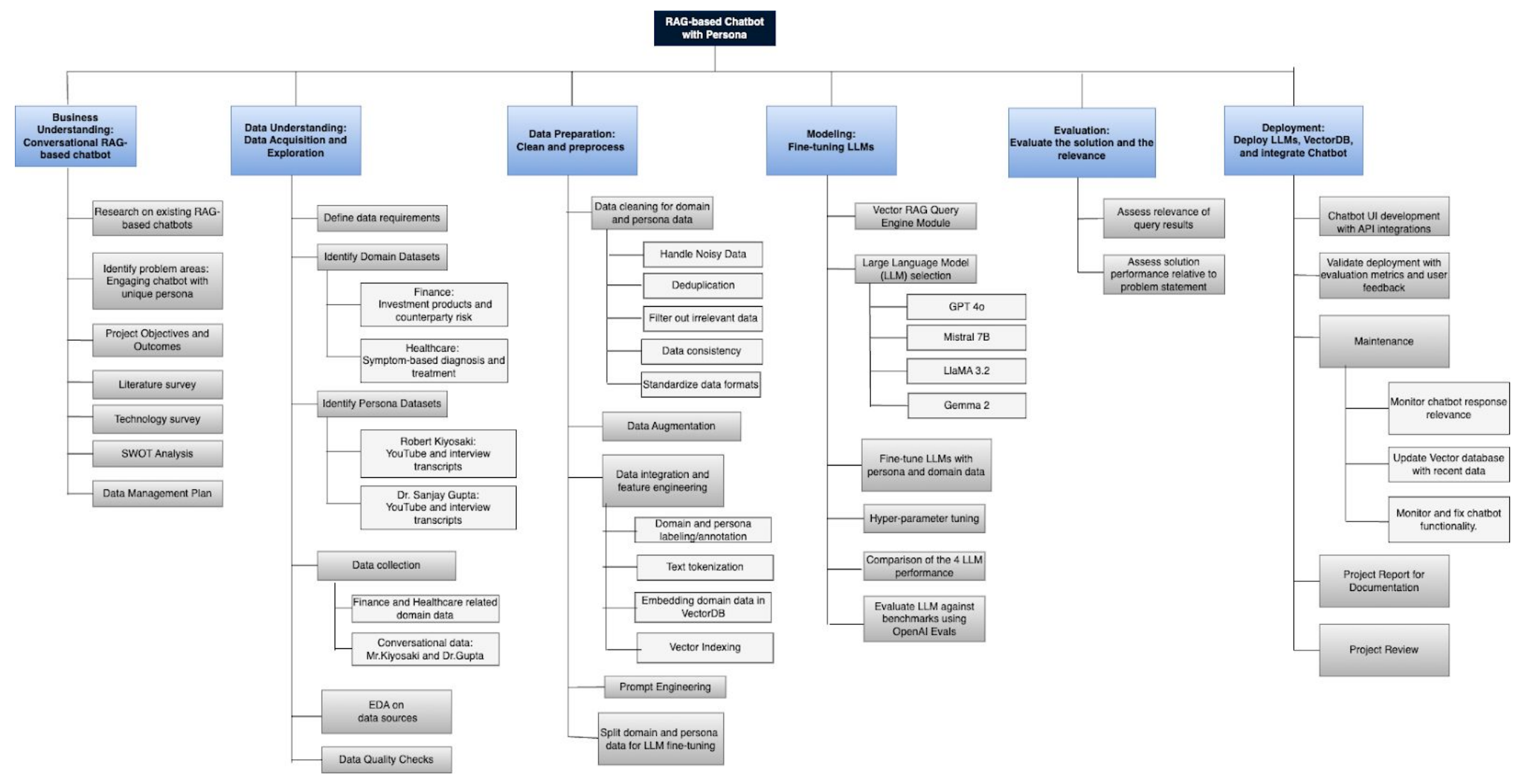
In the evolving landscape of Conversational AI, traditional chatbots often fail to deliver engaging and human-like interactions. SageSoulRAG addresses this gap by creating a Retrieval-Augmented Generation (RAG)-based chatbot that mimics the personas of finance expert Robert Kiyosaki and medical journalist Dr. Sanjay Gupta. This system leverages advanced large language models, including GPT 4o, LlaMA 3.2, Gemma 2, and Mistral 7B, integrated through APIs to provide accurate and personality-driven responses for finance and healthcare domains.



- Implements robust data preprocessing with tokenization, normalization, and embedding.
- Fine-tunes models on persona-specific datasets for expert-level responses.
- Optimizes retrieval using Pinecone Vector Database and rigorous evaluation metrics.

Methodology

The SageSoulRAG chatbot employs a structured approach combining advanced data processing, model fine-tuning, and efficient deployment. The methodology ensures accurate, engaging, and domain-specific responses through the integration of cutting-edge technologies.


RAG-based Chatbot with Persona
Business Understanding: Conversational RAG-based chatbot
Research on existing RAG-based chatbots
Identify problem areas: Engaging chatbot with unique persona
Project Objectives and Outcomes
Literature survey
Technology survey
SWOT Analysis
Data Management Plan
Data Understanding: Data Acquisition and Exploration
Define data requirements
Identify Domain Datasets
Finance: Investment products and counterparty risk
Healthcare: Symptom-based diagnosis and treatment
Identify Persona Datasets
Robert Kiyosaki: YouTube and interview transcripts
Dr. Sanjay Gupta: YouTube and interview transcripts
Data collection
Finance and Healthcare related domain data
Conversational data: Mr.Kiyosaki and Dr.Gupta
EDA on data sources
Data Quality Checks
Data Preparation: Clean and preprocess
Data cleaning for domain and persona data
Handle Noisy Data
Deduplication
Filter out irrelevant data
Data consistency
Standardize data formats
Data Augmentation
Data integration and feature engineering
Domain and persona labeling/annotation
Text tokenization
Embedding domain data in VectorDB
Vector Indexing
Prompt Engineering
Split domain and persona data for LLM fine-tuning
Modeling: Fine-tuning LLMs
Vector RAG Query Engine Module
Large Language Model (LLM) selection
GPT 4o
Mistral 7B
LlaMA 3.2
Gemma 2
Fine-tune LLMs with persona and domain data
Hyper-parameter tuning
Comparison of the 4 LLM performance
Evaluate LLM against benchmarks using OpenAI Evals
Evaluation: Evaluate the solution and the relevance
Assess relevance of query results
Assess solution performance relative to problem statement
Deployment: Deploy LLMs, VectorDB, and integrate Chatbot
Chatbot UI development with API integrations
Validate deployment with evaluation metrics and user feedback
Maintenance
Monitor chatbot response relevance
Update Vector database with recent data
Monitor and fix chatbot functionality.
Project Report for Documentation
Project Review

- **Data Collection & Processing:** Domain-specific data (finance, healthcare) and persona conversational data are collected. Preprocessing techniques like cleansing, tokenization, normalization, and embedding are applied for efficient storage in the Pinecone Vector Database.
- **Model Training:** Large language models (GPT 4o, LlaMA 3.2, Gemma 2, Mistral 7B) are fine-tuned using persona-specific datasets to emulate distinct communication styles.
- **System Integration:** The chatbot is deployed on Hugging Face with a Streamlit interface, facilitating real-time, user-friendly interaction with scalable architecture.

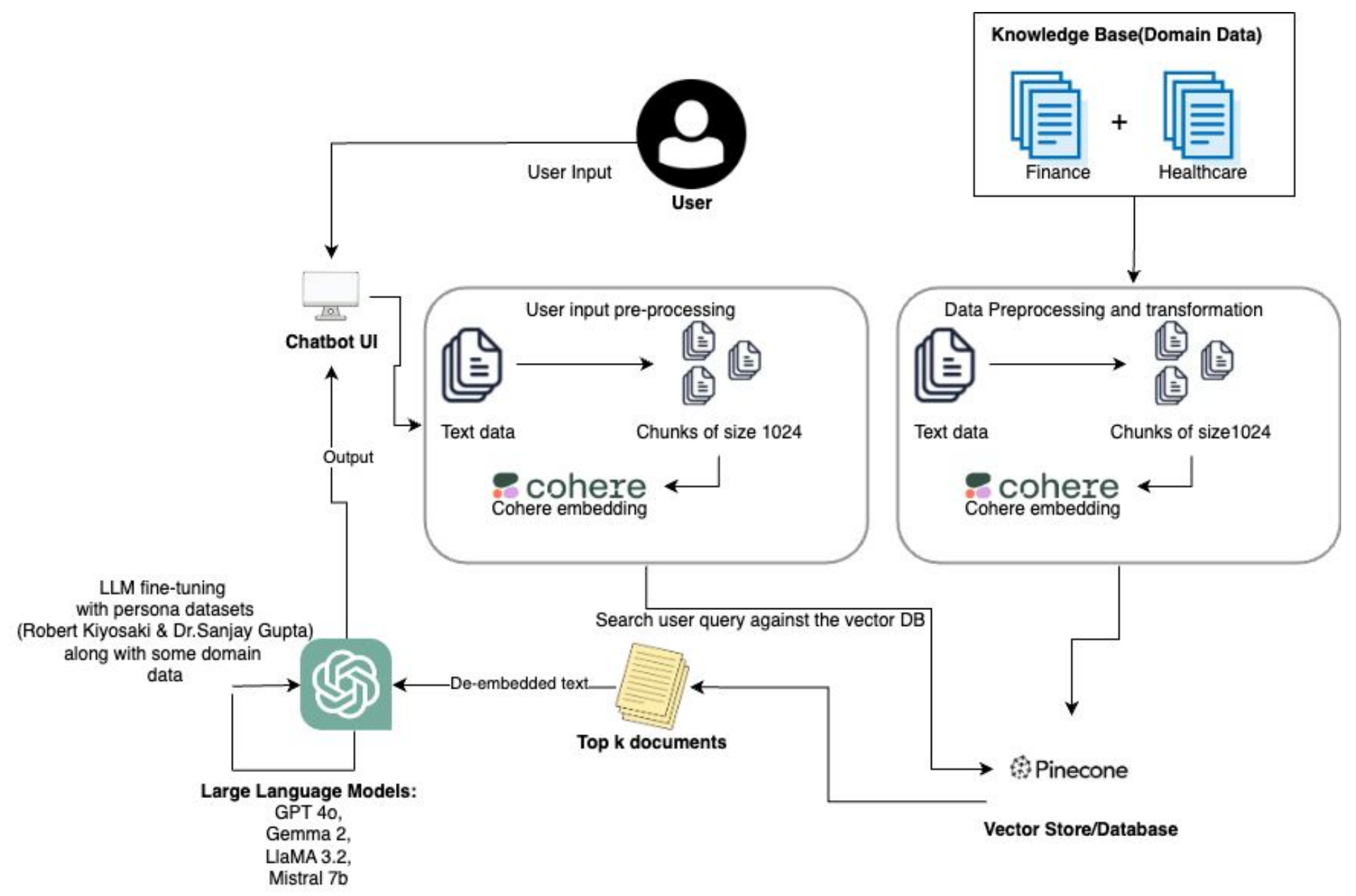
Analysis and Results

Problem Statement and Objectives

The project "SageSoulRAG: RAG-based Chatbot with Personality" addresses the limitations of traditional chatbots that lack human-like traits, often resulting in robotic and impersonal interactions. By incorporating personalities modeled after finance expert Robert Kiyosaki and medical expert Dr. Sanjay Gupta, the project aims to deliver expert-level, personalized advice in finance and healthcare. The chatbot blends natural language processing (NLP) and machine learning techniques to mimic these personas, providing users with an engaging and informative experience.

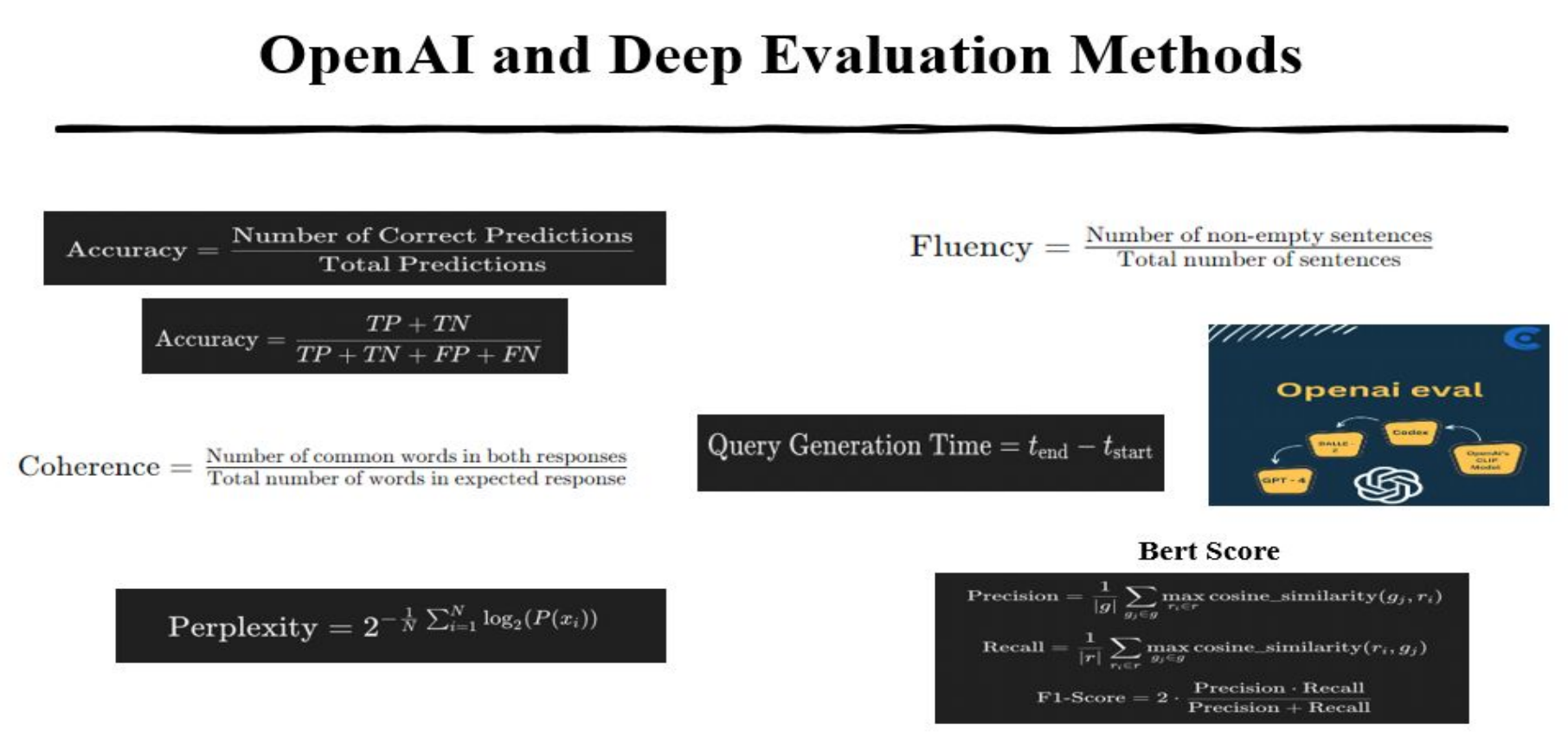
Technical Approach

The chatbot leverages a **Retrieval Augmented Generation (RAG)** framework to enhance response accuracy by integrating a robust retrieval system with generative language capabilities. A comprehensive knowledge base was created using the Pinecone Vector Database, which stores domain-specific data (finance and healthcare) alongside persona-based conversational datasets. Advanced preprocessing techniques, including tokenization, normalization, and embedding, ensure efficient data handling. Four large language models (LLMs)—GPT-4o, LlaMA 3.2, Gemma 2, and Mistral 7B—are fine-tuned to align with the specific communication styles of the personas. A system architecture diagram below illustrates the integration of LLMs, Pinecone database, and the RAG framework.



Evaluation Metrics

The chatbot's performance was rigorously assessed using OpenAI Eval and DeepEval frameworks to ensure its reliability and alignment with project goals. The metrics included **accuracy**, **fluency**, and **coherence.** Quantitative metrics such as Precision, Recall, and F1-score are employed to evaluate the classification tasks, ensuring balance between false positives and false negatives. The **BERT Score** and **Perplexity** metrics provided insights into the semantic and predictive accuracy of the responses, while human evaluators qualitatively rated fluency and coherence on a predefined scale. Additionally, **Query Generation Time** is used as a critical metric to measure the system's responsiveness, ensuring suitability for real-time applications.



OpenAI and Deep Evaluation Methods

$$\text{Accuracy} = \frac{\text{Number of Correct Predictions}}{\text{Total Predictions}}$$

$$\text{Fluency} = \frac{\text{Number of non-empty sentences}}{\text{Total number of sentences}}$$

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

$$\text{Coherence} = \frac{\text{Number of common words in both responses}}{\text{Total number of words in expected response}}$$

$$\text{Query Generation Time} = t_{\text{end}} - t_{\text{start}}$$

$$\text{Perplexity} = 2^{-\frac{1}{N}\sum_{i=1}^{N}\log_2(P(x_i))}$$

Bert Score

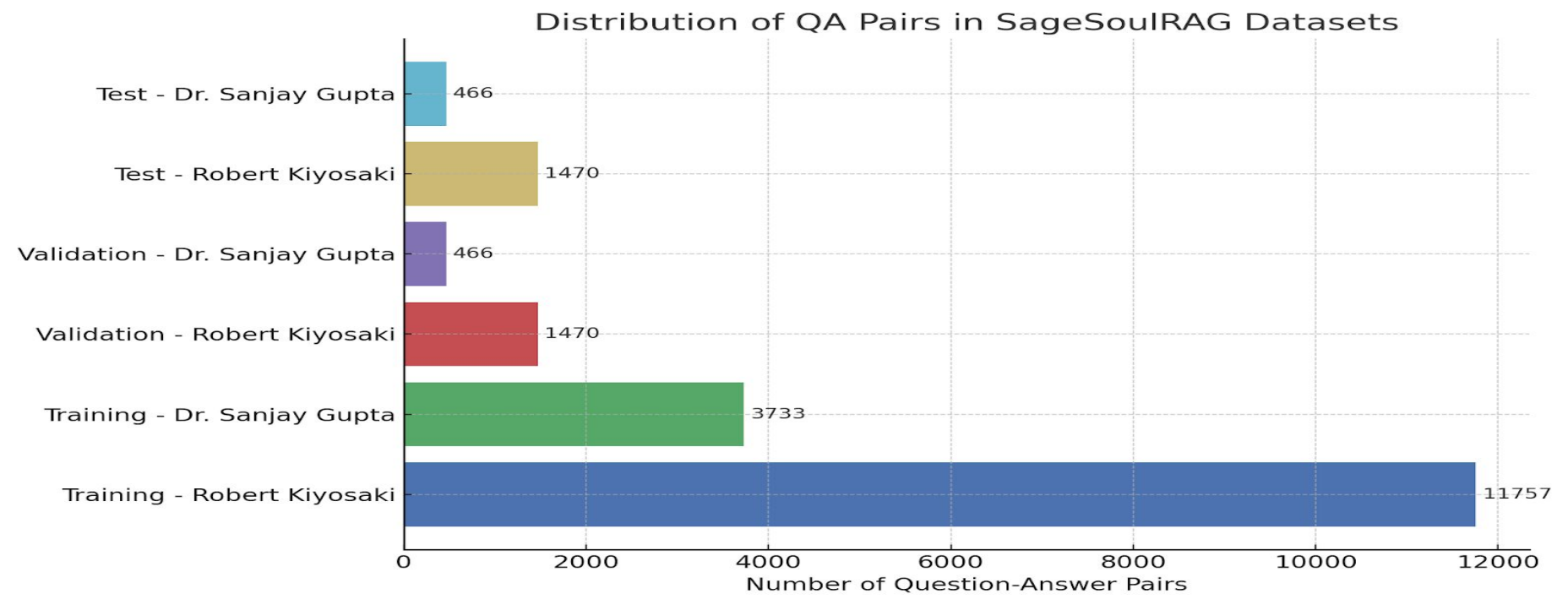
$$\text{Precision} = \frac{1}{|g|}\sum_{g_j \in g}\max_{r_i \in r} \text{cosine\_similarity}(g_j, r_i)$$

$$\text{Recall} = \frac{1}{|r|}\sum_{r_i \in r}\max_{g_j \in g} \text{cosine\_similarity}(r_i, g_j)$$

$$\text{F1-Score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

Project Data

Data Sources and Preparation: The knowledge base combines static sources such as books and articles with dynamic data streams accessed via APIs. The finance dataset contains 11,757 question-answer pairs derived from investment banking and market trends, while the healthcare dataset features 3,733 pairs related to symptom diagnosis and treatment. Persona-specific content, such as transcripts, interviews, and authored books, was curated to emulate the experts' distinct styles. The datasets underwent rigorous cleaning, augmentation, and embedding processes to prepare them for model training and retrieval. A visualization of the training, validation, and test data split is shown below.



Results

Model Performance Evaluation: The chatbot's performance was assessed using OpenAI Eval and DeepEval frameworks. Metrics such as accuracy, fluency, coherence, Precision, Recall, and F1-scores highlighted the system's ability to generate reliable, contextually relevant responses. BERT Scores and Perplexity metrics further demonstrated semantic and predictive accuracy. Human evaluators confirmed the chatbot's fluency and coherence through qualitative ratings, while query generation time metrics verified its suitability for real-time interactions.

The model evaluation compares various language models—GPT-4o, LlaMA 3.2, Mistral 7B, and Gemma 2—across key performance metrics. GPT-4o leads with the highest BERT score of 0.366, indicating strong semantic alignment, and achieves perfect fluency and accuracy, though its perplexity of 392.81 suggests some challenges with word prediction. LlaMA 3.2 performs efficiently with a lower perplexity of 29.65 and a BERT score of 0.328, but its query generation time of 5.08 seconds is longer than GPT-4o's 2.90 seconds. Mistral 7B has the lowest BERT score of 0.302 but excels in coherence with a score of 0.75, though it struggles with high perplexity (927.61) and slow query generation (1955.78 seconds). Gemma 2 balances between these metrics with a BERT score of 0.337, a moderate perplexity of 42.20, and a coherence score of 0.47, though its query generation time of 6.16 seconds is the longest. The comparison table is shown below.

Model Name	BERT Score	Accuracy	Fluency	Perplexity	Coherence	Query Generation Time
GPT 4o	0.3663818836	1	1	392.8128671	0.1924882629	2.902482033
Llama 3.2	0.3285864592	1	1	29.65289497	0.1524882629	5.082540512
Mistral 7B	0.302709	1	1	927.6068975	0.75	9.852
Gemma 2	0.3378536999	1	1	42.1978569	0.47	6.155407906

Summary/Conclusions

The SageSoulRAG project successfully developed a Retrieval Augmented Generation (RAG)-based chatbot with distinct personas, Robert Kiyosaki and Dr. Sanjay Gupta. By integrating advanced language models and utilizing a robust data retrieval system, the chatbot provided domain-specific expert advice in finance and healthcare, ensuring accurate, fluent, and contextually relevant responses. The evaluation results showcased high performance across various metrics, with GPT-4o leading in accuracy and fluency, while Mistral 7B excelled in coherence.

The SageSoulRAG chatbot demonstrated significant advancements in conversational AI by mimicking expert personalities and delivering tailored responses. The use of fine-tuned LLMs and efficient data retrieval techniques helped achieve robust performance, making the chatbot a valuable tool for personalized guidance in complex domains like finance and healthcare. Future work may focus on enhancing model efficiency and expanding its application to other fields.

Key References

[1] An in-depth guide to building a custom GPT-4 chatbot on your data. Mercity AI, Pranav. (n.d.). https://www.mercity.ai/blog-post/custom-gpt-4-chatbot

[2] Anil, R., Dai, A. M., Firat, O., Johnson, M., Lepikhin, D., Passos, A. M. a. D., Shakeri, S., Taropa, E., Bailey, P., Chen, Z., Chu, E., Clark, J. H., Shafey, L. E., Huang, Y., Meier-Hellstern, K., Mishra, G., Moreira, É. R., Omernick, M., Robinson, K., . . . Wu, Y. (2023). Gemma 2 Technical Report. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2305.10403

[3] Bonnet, A. (2024, October 8). From vision to edge: Meta's Llama 3.2 explained. Encord. https://encord.com/blog/lama-3-2-explained/

[4] DeepMind\authfootnotemark1, G. T. G., Team, G., & DeepMind\authfootnotemark1, G. (n.d.). Gemma: Open models based on Gemini Research and Technology. https://arxiv.org/html/2403.08295v1

[5] DeepMind\authfootnotemark1, G. T. G., Team, G., & DeepMind\authfootnotemark1, G. (n.d.). Gemma: Open models based on Gemini Research and Technology. https://arxiv.org/html/2403.08295v1

Acknowledgements

We extend our deepest gratitude to Dr. Simon Shim and Dr. Lee C. Chang for their invaluable guidance and expert advice throughout the course of this project. Their insights and expertise in the field of Applied Data Science have been instrumental in shaping our approach and methodologies. Their willingness to impart knowledge and provide continuous support has profoundly impacted our academic journey and professional growth. We are truly grateful for their patience, mentorship, and significant contributions to our learning experience and project success.