BFSI Call Center AI Assistant

Technical Documentation

Prepared By: Pavithra Veerapathiran

1. Project Overview

The BFSI Call Center AI Assistant is a lightweight, secure, and compliance-focused artificial intelligence system designed to assist banking and financial services call centers. The system provides accurate, policy-grounded, and safe responses to customer queries while minimizing operational risks and preventing financial misinformation.

2. Project Objective

• Design a lightweight AI assistant tailored for BFSI environments.

• Ensure compliance with financial regulations.

• Prevent hallucinations and unsafe financial advice.

• Provide cost-efficient architecture using local SLM.

• Deliver accurate and context-aware responses for call center queries.

3. System Architecture

3.1 Similarity Engine (Dataset-First Approach)

Matches incoming queries against a structured BFSI dataset and returns curated, pre-approved safe responses. This reduces unnecessary model generation and ensures reliable answers for common queries.

3.2 Fine-Tuned Local SLM (TinyLlama)

Acts as a fallback mechanism when similarity matching confidence is low. The model is fine-tuned on 150+ Alpaca-formatted BFSI conversational samples to ensure domain-specific accuracy and controlled generation.

3.3 RAG (Retrieval-Augmented Generation) Layer

Retrieves relevant financial policy documents indexed using FAISS and injects grounded context into responses. This significantly reduces hallucination and ensures policy-backed answers.

3.4 Guardrails & Compliance Layer

Implements safety mechanisms to block unsafe financial advice, prevent hallucinated rates or approvals, and ensure all responses align with regulatory standards.

4. Dataset Details

The dataset consists of 150+ Alpaca-formatted BFSI conversational samples covering loan eligibility, EMI calculation, interest rate queries, prepayment charges, and policy clarifications.

5. Component Summary

• Similarity Engine – Returns curated safe responses.

• SLM (TinyLlama) – Generates fallback conversational responses.

• RAG Layer – Retrieves grounded financial policy information.

• Guardrails – Ensures compliance and safety.

6. Future Deployment Note

This solution is architecturally designed to support deployment in enterprise environments. While deployment is not included in the current scope, the system can be deployed in the future using suitable frameworks and infrastructure as required.

Developed & Documented By: Pavithra Veerapathiran