Explainable AI: Retrieval-Augmented Generation Framework for Model Interpretability

1. Objectives

- Addresses the challenge of black-box machine learning and deep learning models by providing explainability and interpretability, particularly for non-technical users in industries like healthcare, finance, and manufacturing.
- Proposes a Retrieval-Augmented Generation (RAG) based framework that combines large language models with domain-specific knowledge bases, offering clear, context-aware, and interactive explanations of model outputs.
- Aims to bridge the gap between technical complexity and practical usability while ensuring data privacy through private knowledge bases.

2. Key Contributions

- Introduces a novel RAG framework that tightly integrates retrieval of relevant data chunks with generative models to produce evidence-backed, explainable AI outputs instead of opaque predictions.
- Provides a modular approach adaptable across domains requiring accountable AI systems.
- Demonstrates through experiments that this approach improves explanation quality and enhances user trust in AI systems.
- Uses vector-based embedding techniques and private knowledge bases for efficient, scalable retrieval and explanation.

3. Framework Overview and Methodology

- **Data Chunking and Embedding**: Domain-specific data (text, images, tabular) is segmented into meaningful chunks and embedded into vector representations stored in vector databases like Pinecone.
- **Query Processing and Retrieval**: User queries are converted to vector embeddings and matched against stored embeddings using similarity metrics (cosine similarity, dot product). Results are reranked for relevance.
- **Response Generation**: Large language models (e.g., LLama-70b-8192) generate responses conditioned on retrieved data chunks, synthesizing coherent, contextually relevant outputs with clear citations.
- The system supports multi-modal data and secure, private deployment using dedicated knowledge bases.

4. Evaluation and Results

- Evaluated across use cases in healthcare (tumor classification explanation), finance (loan eligibility modeling explanation), and manufacturing (cost estimation tool explanation).
- Metrics used include BERTScore, ROUGE (n-gram overlap), Perplexity, and response time.
- Achieved good semantic similarity with ground truth references and efficient response latency suitable for real-time applications.
- Human evaluation showed improved explainability and trust due to evidence-backed, transparent responses.
- Addressed challenges like hallucination by enriching knowledge bases with detailed documentation and markdown chunks.

5. Practical Use Cases

- Healthcare chatbot providing interpretable model classifications from MRI images.
- Finance chatbot explaining logistic regression model choices with visual data insights.
- Manufacturing chatbot answering technical queries about cost prediction models, using XGBoost and Random Forest algorithms.

6. Relevance for Academic Compliance Project

- Provides a robust blueprint for implementing explainable RAG systems tailored for domain-specific knowledge bases.
- Demonstrates integration of multi-modal data, scalable vector retrieval, and LLM synthesis—relevant for processing diverse academic policy texts.
- Highlights methods for mitigating hallucination and enhancing response traceability, critical for compliance and regulatory use cases.
- Offers tested metrics and evaluation strategies that can be adapted to measure the accuracy and explainability of the Policy-as-Code engine.