Design Rationale Document

Intelligent Sri Lanka IRD Tax Intelligence & Compliance Assistant

1. Introduction

   The *Intelligent Sri Lanka IRD Tax Intelligence & Compliance Assistant* was designed to provide accurate, reliable, and explainable answers to tax-related questions strictly based on official Sri Lanka Inland Revenue Department (IRD) documents. The primary motivation behind this system is to address the risk of misinformation and hallucination commonly associated with large language models when used for sensitive domains such as taxation and legal compliance.

   This design rationale explains the key architectural decisions, technology selections, and safeguards implemented to ensure correctness, transparency, and compliance with the assessment requirements.

2. Problem Definition and Design Goals

   Tax-related information is highly sensitive and must be accurate, verifiable, and sourced from authoritative documents. General-purpose AI models often generate plausible but incorrect answers when information is missing or ambiguous. Therefore, the system was designed with the following goals:

   - Ensure **answers are strictly derived from IRD-published documents**
   - Prevent **hallucination and guessing**
   - Detect **missing or ambiguous questions**
   - Provide **traceable citations**
   - Provide **traceable citations**
   - Support both **API-based** and **CLI-based** usage

   These goals guided all architectural and implementation decisions.

3. Overall System Architecture

   The system follows a **Retrieval-Augmented Generation (RAG)** architecture with strong guardrails. It is divided into two main phases: **Build-time ingestion** and **Runtime querying**.

   At build time, official IRD PDF documents are ingested, processed, and converted into a searchable vector database. At runtime, user queries are semantically matched against this database, and responses are generated only when sufficient confidence exists.

   This modular architecture ensures scalability, maintainability, and strict control over information flow.

4. Key Design Decisions

   4.1. Use of Vector Database (FAISS)

   A local FAISS vector database was chosen to store embeddings of IRD document chunks. This allows efficient semantic search while keeping all data on the local system, ensuring privacy and offline capability. FAISS was preferred over cloud-based solutions to comply with the requirement of local execution and full control over data.

   4.2. Semantic Retriever with Reranking

   A dedicated Retriever component embeds user queries using a SentenceTransformer model (all-MiniLM-L6-v2) and retrieves the most relevant document chunks. Optional reranking is applied to prioritize Public Notices over guides, aligning with IRD document authority hierarchy.
   This design ensures higher relevance and better alignment with official interpretations.

5. Guardrails and Reliability Mechanisms

A major design focus was preventing incorrect or misleading answers. Therefore, a **Guardrails module** was introduced to classify retrieval outcomes into three categories:

- **Missing**: When no sufficiently strong evidence exists in the documents
- **Ambiguous**: When multiple high-confidence but conflicting contexts are found
- **Valid:** When a clear, authoritative answer exists

If a query is missing or ambiguous, the system explicitly informs the user instead of attempting to answer. This behavior is critical for compliance-focused applications and is a core strength of the design.

6. Answer Generation Strategy

The system supports two answer generation modes:

- **Extractive fallback**: Directly returns a short excerpt from the most relevant document when LLM resources are unavailable.
- **LLM-assisted generation**: Uses a locally hosted Ollama LLM to generate concise answers strictly constrained by retrieved context.

Importantly, the LLM is never allowed to use external knowledge. This preserves factual correctness and prevents hallucination.

7. Disclaimer Enforcement

To ensure ethical and legal responsibility, a **mandatory disclaimer** is automatically appended to every response:

“This response is based solely on IRD-published documents and is not professional tax advice.”

This disclaimer is enforced at the code level and cannot be bypassed, ensuring consistent compliance with assessment and real-world expectations.

8. API and Interface Design

The system exposes functionality through a FastAPI-based REST API with clearly defined endpoints for document upload and query execution. Additionally, a command-line interface (CLI) is provided for testing and demonstration purposes.

The system exposes functionality through a FastAPI-based REST API with clearly defined endpoints for document upload and query execution. Additionally, a command-line interface (CLI) is provided for testing and demonstration purposes.

9. Assumptions and Limitations

- Only IRD-published PDFs are considered authoritative
- No external tax knowledge is used
- The system does not replace professional tax consultation
- LLM performance depends on local system resources
- Multi-language support is not included in the current version

10. Conclusion

The final system design prioritizes correctness, explainability, and safety over generative flexibility. By combining document-grounded retrieval, strict guardrails, and controlled LLM usage, the solution effectively addresses the challenges of applying AI to sensitive tax compliance scenari

This design satisfies all assessment requirements while demonstrating best practices in trustworthy AI system design.