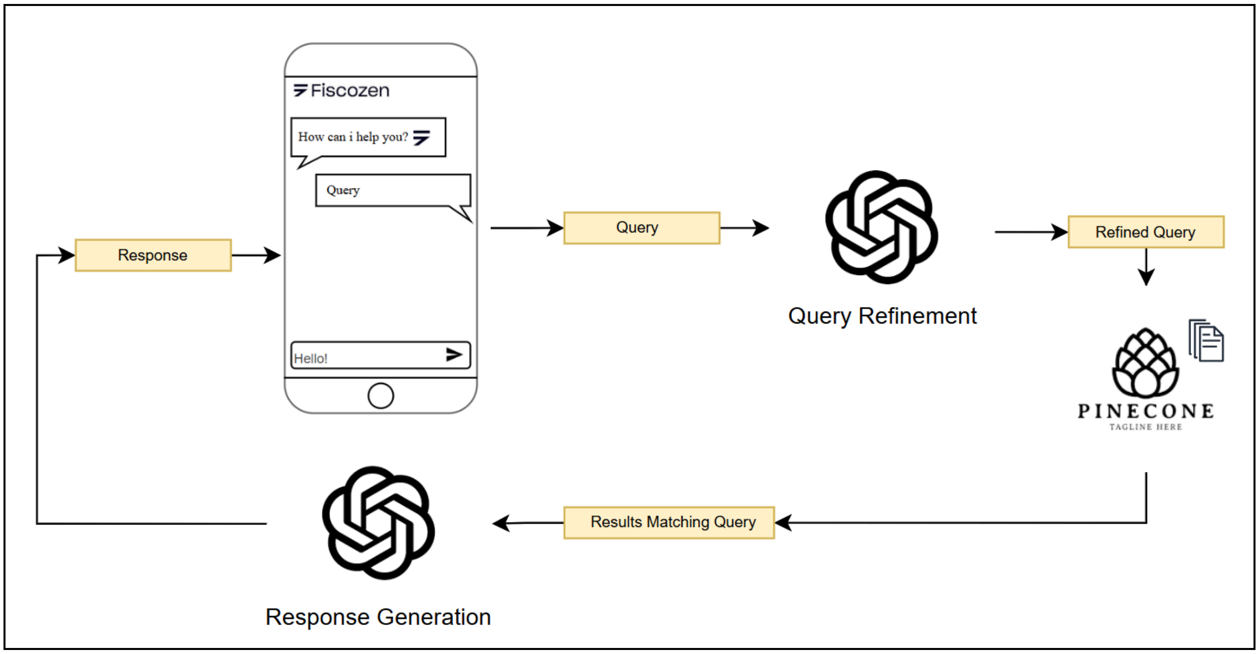
FiscoChat MVP Design Document

For the first version of FiscoChat, we will implement an AI-powered Retrieval-Augmented Generation (RAG) system to efficiently answer user questions related to tax and Partita IVA topics.



System Workflow:

1. **User Query Input**
 The user submits a natural language question through the FiscoChat interface.

2. **Query Refinement (GPT)**
 The raw query is passed to a GPT-based component that refines and optimizes the query for information retrieval. This step improves search relevance and handles variations in phrasing.

3. **Semantic Search via Pinecone**
 The refined query is submitted to **Pinecone**, which performs **vector-based semantic search** using **cosine similarity**. This allows the system to retrieve the most relevant content segments.

4. **Context Retrieval (Top-K Sections)**
 The system extracts the **top-K relevant documentation sections** from the Pinecone index. These sections form the basis of the response.

5. **Response Generation (GPT)**
   The retrieved content is passed through GPT again, which synthesizes a natural, helpful, and contextually accurate answer to the user's question.

6. **User Response Output**
   The final response is delivered via the chatbot interface for the user to read.

Data Sources for Pinecone Indexing

The searchable knowledge base will be constructed using data extracted and cleaned with **Argilla**, and indexed into Pinecone. The initial data sources include:

- **The Fiscozen website** – for company-specific tax workflows, FAQs, and service guidance.

- **Agenzia delle Entrate** – the Italian tax authority's official documentation, legal references, and procedures.