TR-LLM-PL - High Level Design

Overview

A Retrieval-Augmented Generation (RAG) system that answers questions about the Premier League using current data.

Core Components

1. Data Sources

- **Player Information CSV**: Pre-scraped dataset containing:
    - Player names
    - Current team affiliations
    - Playing positions
    - Basic player attributes
- **Scope**: Limited to current Premier League players only
- **Example queries supported**:
    - "Tell me the players of Manchester United"
    - "What team does Gyökeres play for?"
    - "What position does Robert Sánchez play?"

2. Data Processing Pipeline

- **Initial State**: CSV file is pre-prepared and ready to use
- **Loading**: Direct CSV ingestion into memory/database
- **No preprocessing required**: Data assumed to be clean and structured
- **Storage**: simple database table

3. Retrieval System

- **Main LLM Router** (gpt 5 chat): Azure OpenAI determines query classification
    - General Premier League queries (rules, history, general knowledge)
    - Player data queries requiring database lookup
- **SQL Agent** (gpt 5 mini): Activated only when player data is needed
    - Natural language to SQL query conversion
    - Direct CSV/database querying using SQL (to resemble deployment environment closer)
    - Query types: Filter by team, player name, or position
- **Validation Loop**: SQL results returned to main LLM for context and verification

4. Generation System

- **Main LLM**: OpenAI GPT 5 chat model as central orchestrator
    - **Input**: User queries of any type
    - **Decision Making**: Classify query type and route appropriately
    - **Context Integration**: Combine SQL results with general knowledge
    - **Output**: Unified natural language responses
- **Prompt Engineering**: Tailored for football terminology and query classification

5. User Interface

- **Streamlit**: Python-based web interface
- **Features**:
    - Text input for questions
    - Display formatted answers
    - Show underlying data tables when relevant

High-Level Architecture

User Query (Streamlit) → Main LLM → Query Classification

↓

General Query → Direct Answer

↓

Player Data Query → SQL Agent → CSV/Database

↓

Retrieved Data → Main LLM → Final Response

Data Flow

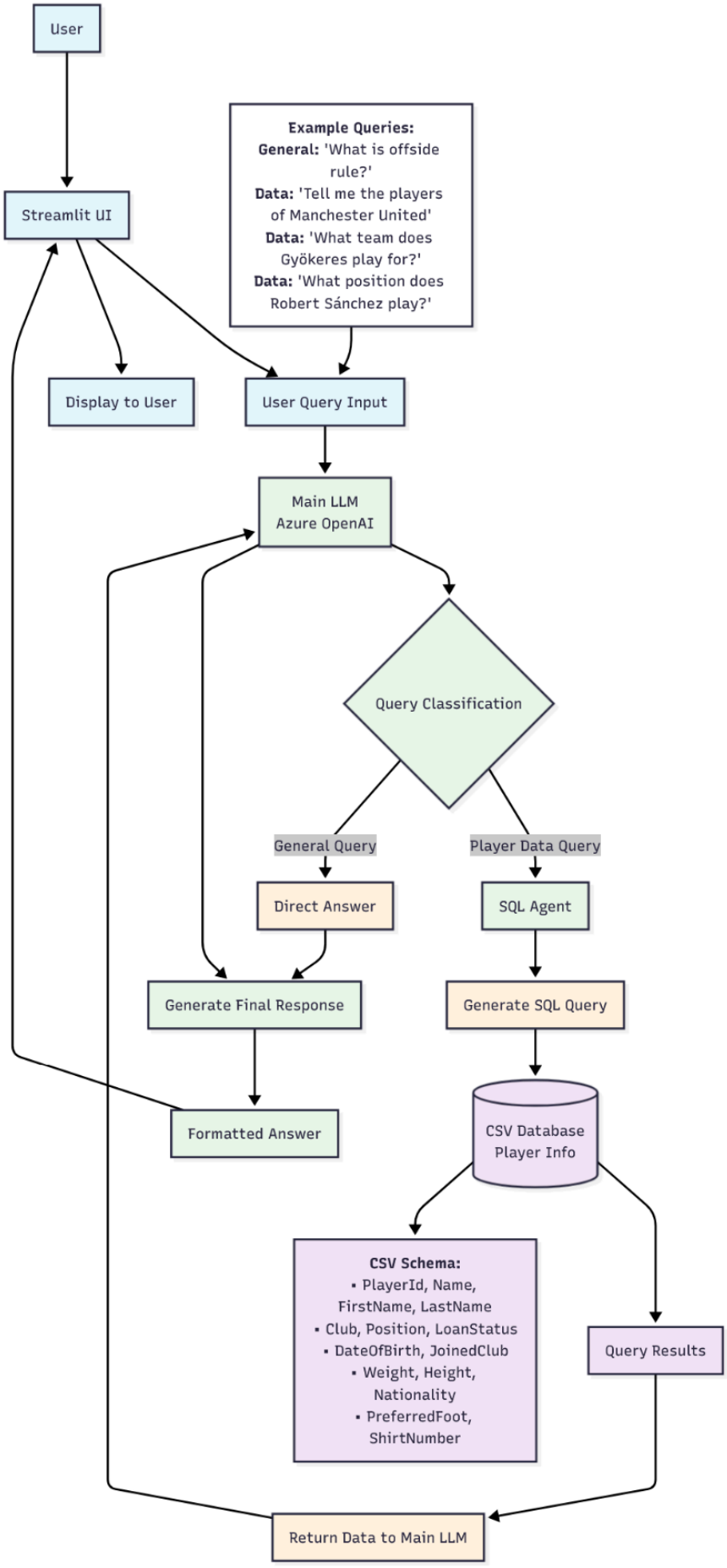
1. User asks question via Streamlit interface
2. Main LLM (Azure OpenAI) receives and classifies the query
3. **Path A - General Query**: Main LLM answers directly using its knowledge
4. **Path B - Player Data Query**:
    - Main LLM routes to SQL Agent
    - SQL Agent converts natural language to SQL and queries CSV
    - Results returned to Main LLM for validation and contextualization
5. Main LLM generates unified natural language response
6. Answer displayed in Streamlit UI

Key Requirements

- **Data**: CSV file with current Premier League player data
- **Query Handling**: Support both general Premier League questions and specific player data queries
- **Intelligence**: Main LLM classifies and routes queries appropriately
- **Response Time**: Fast routing decisions and SQL queries on small dataset
- **Accuracy**: Correct player-team-position mappings and general knowledge
- **Interface**: Streamlit app

Technology Stack

- **Main LLM**: Azure OpenAI GPT models (query classification and response generation)
- **Data Storage**: CSV file / SQLite database
- **SQL Agent**: LangChain SQL Agent or custom implementation
- **Frontend**: Streamlit
- **Backend**: Python with pandas/sqlite3


User
Streamlit UI
Example Queries:
General: 'What is offside rule?'
Data: 'Tell me the players of Manchester United'
Data: 'What team does Gyökeres play for?'
Data: 'What position does Robert Sánchez play?'
Display to User
User Query Input
Main LLM
Azure OpenAI
Query Classification
General Query
Player Data Query
Direct Answer
SQL Agent
Generate Final Response
Generate SQL Query
Formatted Answer
CSV Database
Player Info
CSV Schema:
▪ PlayerId, Name, FirstName, LastName
▪ Club, Position, LoanStatus
▪ DateOfBirth, JoinedClub
▪ Weight, Height, Nationality
▪ PreferredFoot, ShirtNumber
Query Results
Return Data to Main LLM