Rafiki AI System Design

Introduction

This document outlines the system design for an SMS-based AI interaction platform. Users can send queries via SMS and receive responses from a Large Language Model (LLM). The system integrates a Database for storing user data and provides context to both Standard and Real-time LLMs for personalized responses. It supports general queries and those requiring real-time data, ensuring accessibility and efficiency.

System Components

Component	Component
User	Sends SMS queries to a designated phone number using a mobile device.
SMS Gateway API	Manages SMS communication, interfacing between the User and Server.
Server	Processes queries, retrieves context, routes to LLMs, and manages responses.
Database	Stores user data, query history, and responses for context and analytics.
Standard LLM	Handles general queries using pre-trained knowledge and Database context.
Real-time LLM	Processes queries needing real-time data, using internal sources and context.

System Flow

1. User Interaction

- User sends an SMS query to the SMS Gateway API.
- The API forwards the query to the Server.

2. Query Logging and Context Retrieval

- Server logs the query in the Database
- Server retrieves user context from the Database.

3. Query Routing

- Server analyzes the query to determine if real-time data is needed.
- **General Query**: Routed to Standard LLM.
- **Real-time Query**: Routed to Real-time LLM.

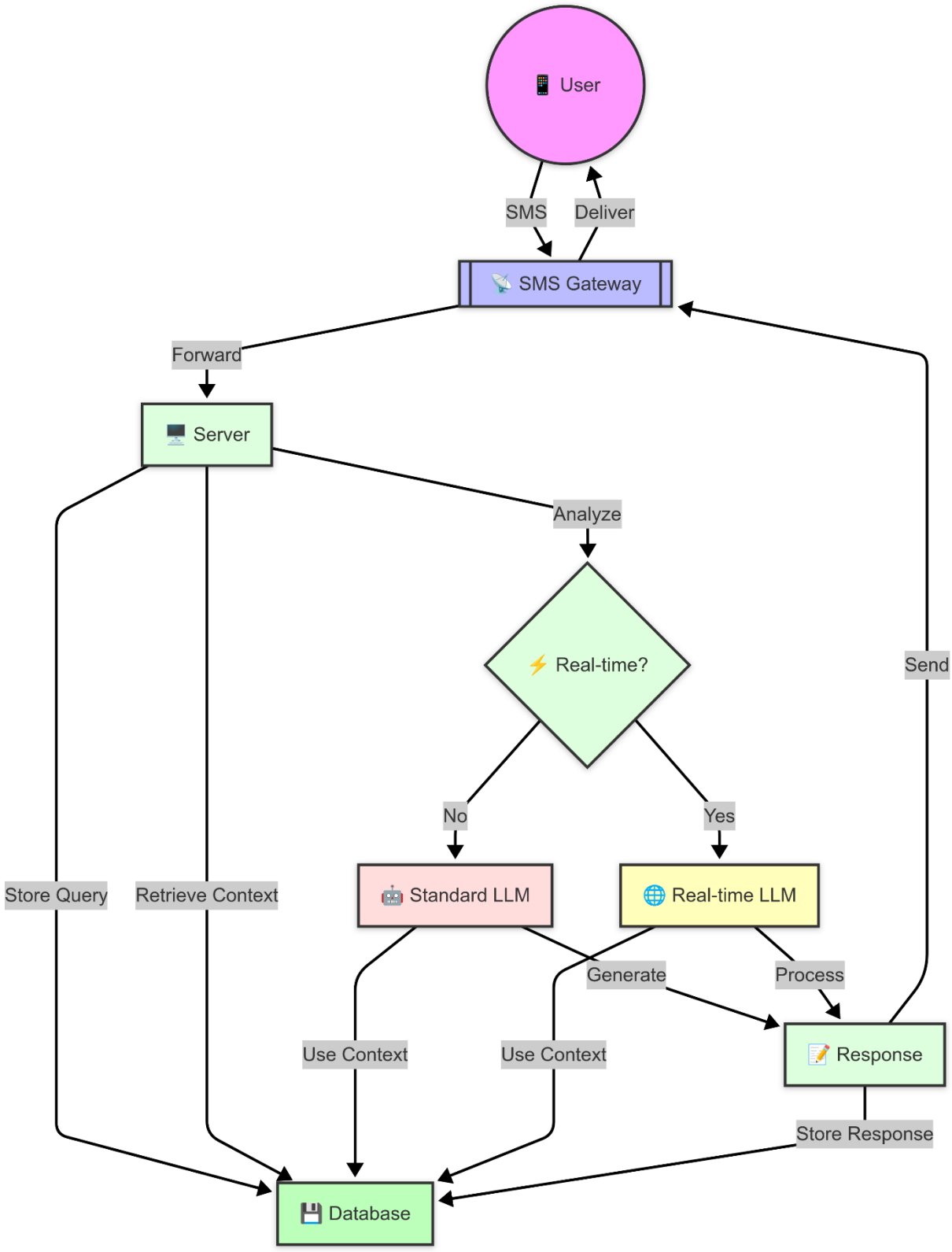
4. LLM Processing

- **Standard LLM**: Generates a response using pre-trained knowledge and context.
- **Real-time LLM**: Generates a response with real-time data and context.

5. Response Handling

- Server stores the response in the Database.
- Server sends the response back to the User via the SMS Gateway API.

Flowchart Diagram



Implementation Details

SMS Gateway API

- **Technology**: Custom API Gateway.
- **Role**: Handles SMS sending and receiving.

Server

- **Technology**: TypeScript (NodeJS).
- **Role**: Query analysis, routing, and response management.

Database

- **Technology**: PostgreSQL.

Standard LLM

- **Technology**: Powerful LLMs.
- **Role**: General query processing.

Real-time LLM

- **Technology**: Real-time capable LLM
- **Role**: Real-time query processing.

Security

- Data encryption for sensitive information.
- Access controls on Database and APIs.
- Compliance with GDPR and privacy regulations.

Benefits

- **Personalization**: Context-aware responses.
- **Real-time Access**: Up-to-date information when needed.
- **Accessibility**: SMS-based, no smartphone required.

Conclusion

Rafiki AI design provides a robust, secure, and accessible SMS-based AI interaction system, leveraging Standard and Real-time LLMs with a Database for enhanced user experience.