Sahai Phase 1 System Design

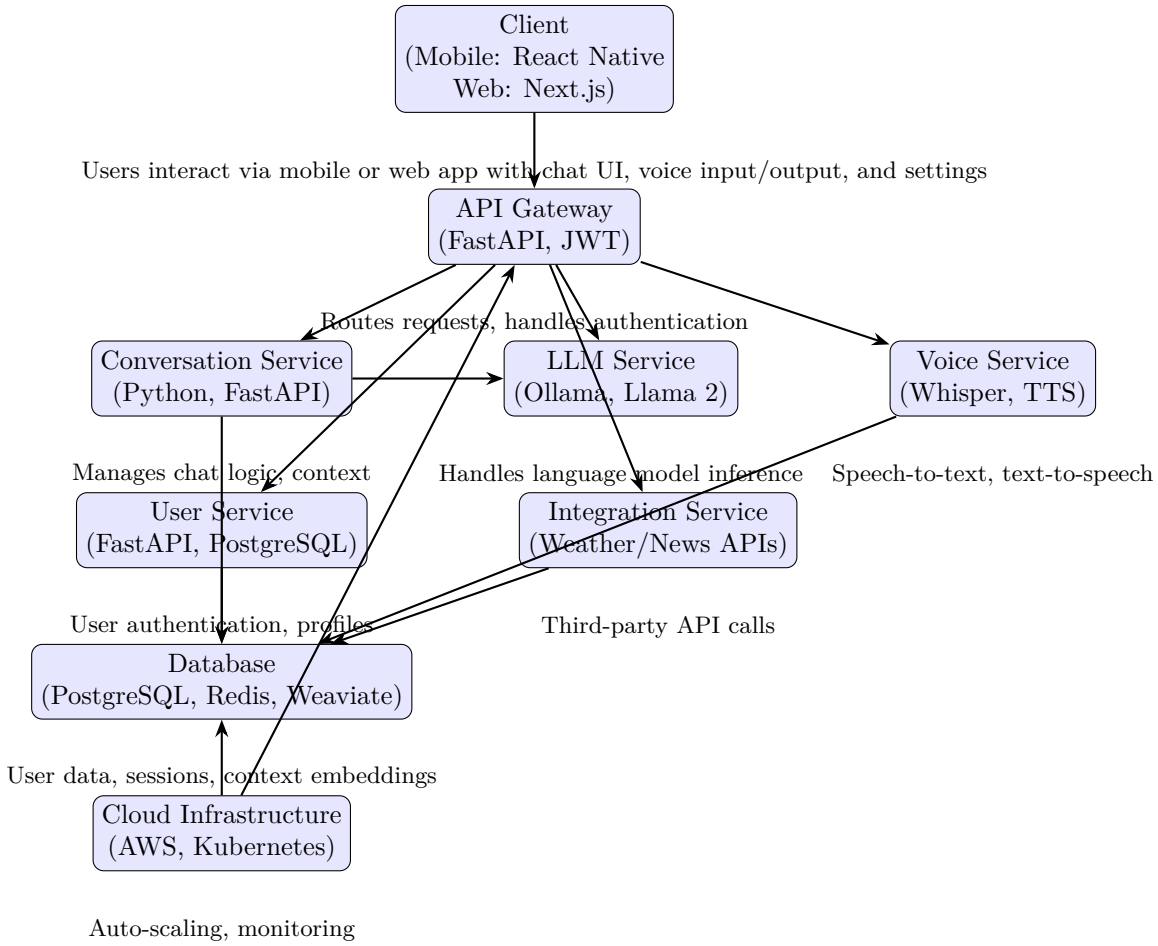

Figure 1: Sahai System Architecture

System Design Overview

The Sahai system follows a microservices architecture with an API Gateway, deployed on AWS with Kubernetes orchestration (SRS 6.1, 9.1). Key components include:

- **Client Layer**: Mobile app (React Native) and web app (Next.js) provide a WhatsApp-inspired chat UI, voice input/output, and settings for language/cultural preferences (SRS 3.1).
- **API Gateway**: Built with FastAPI, handles request routing, JWT authentication, and rate limiting (SRS 3.4, REQ-NF-005).
- **Conversation Service**: Manages chat logic, context retrieval, and LLM inference calls, storing conversation history in PostgreSQL (SRS 4.1, 6.3).
- **LLM Service**: Uses Ollama with Llama 2 7B, fine-tuned with LoRA for Indian cultural context, with fallbacks to Groq API (SRS 3.3.1, 6.4).
- **Voice Service**: Integrates Whisper for speech-to-text and a TTS engine for voice output, supporting Hindi/English (SRS 4.2).
- **User Service**: Handles user registration, profiles, and authentication, storing data in PostgreSQL (SRS 4.4).
- **Integration Service**: Connects to third-party APIs (e.g., OpenWeatherMap, NewsAPI) for weather and news (SRS 4.3).

- **Database**: PostgreSQL for user/conversation data, Redis for sessions, Weaviate for context embeddings (SRS 3.3.3, 6.3).
- **Cloud Infrastructure**: AWS with Kubernetes for auto-scaling, Prometheus/Grafana for monitoring, and CloudFlare CDN for content delivery (SRS 6.2, 9.2).

Data Flow

1. User sends text/voice input via mobile/web app.
2. API Gateway authenticates and routes requests to appropriate services.
3. Conversation Service processes input, retrieves context from Weaviate, and calls LLM Service.
4. LLM Service generates responses with cultural context, cached in Redis for efficiency.
5. Voice Service handles speech-to-text and text-to-speech conversions.
6. Integration Service fetches external data (e.g., weather, news).
7. Responses are returned to the client via WebSocket for real-time updates (SRS 3.4).

Key Considerations

- **Scalability**: Kubernetes auto-scaling supports up to 10,000 users (SRS 5.2).
- **Security**: TLS 1.3, AES-256 encryption, and JWT tokens ensure data protection (SRS 5.2).
- **Performance**: Target <3-second response time for 95% of requests (SRS 5.1).
- **Compliance**: Adheres to Indian Data Protection Act with user consent controls (SRS 5.4).