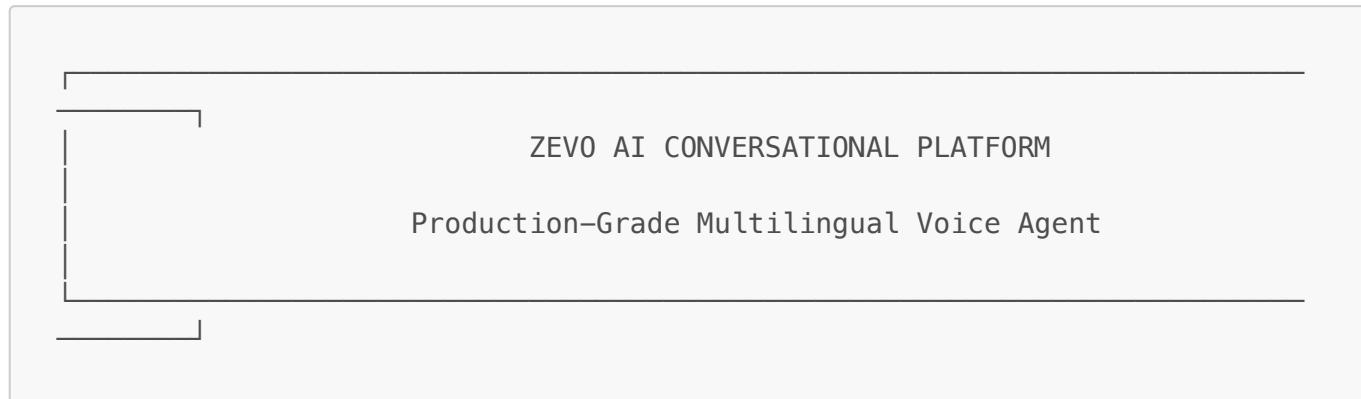


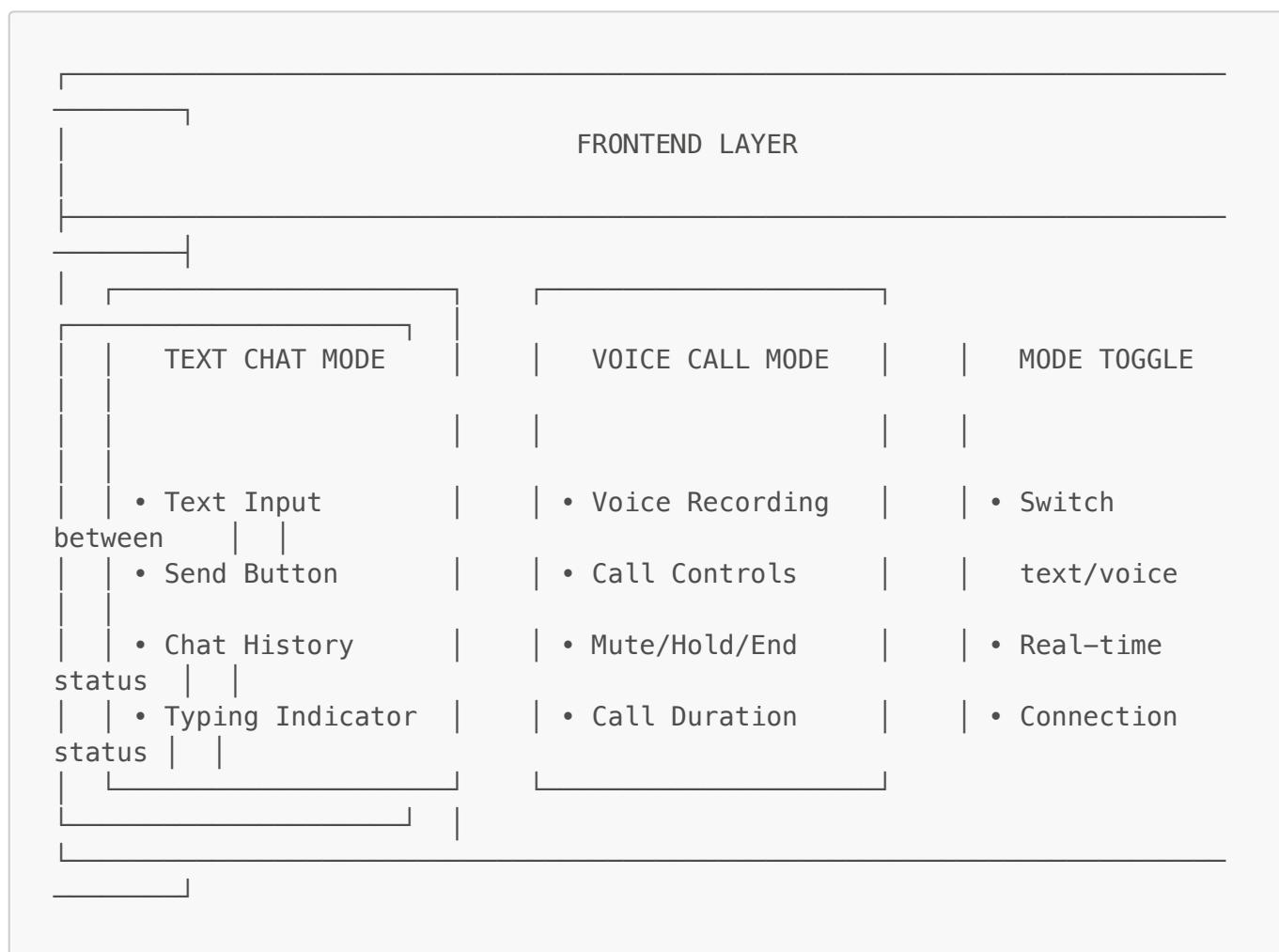
Zeko AI - End-to-End Voice & Text Processing Architecture

🏗 System Architecture Overview



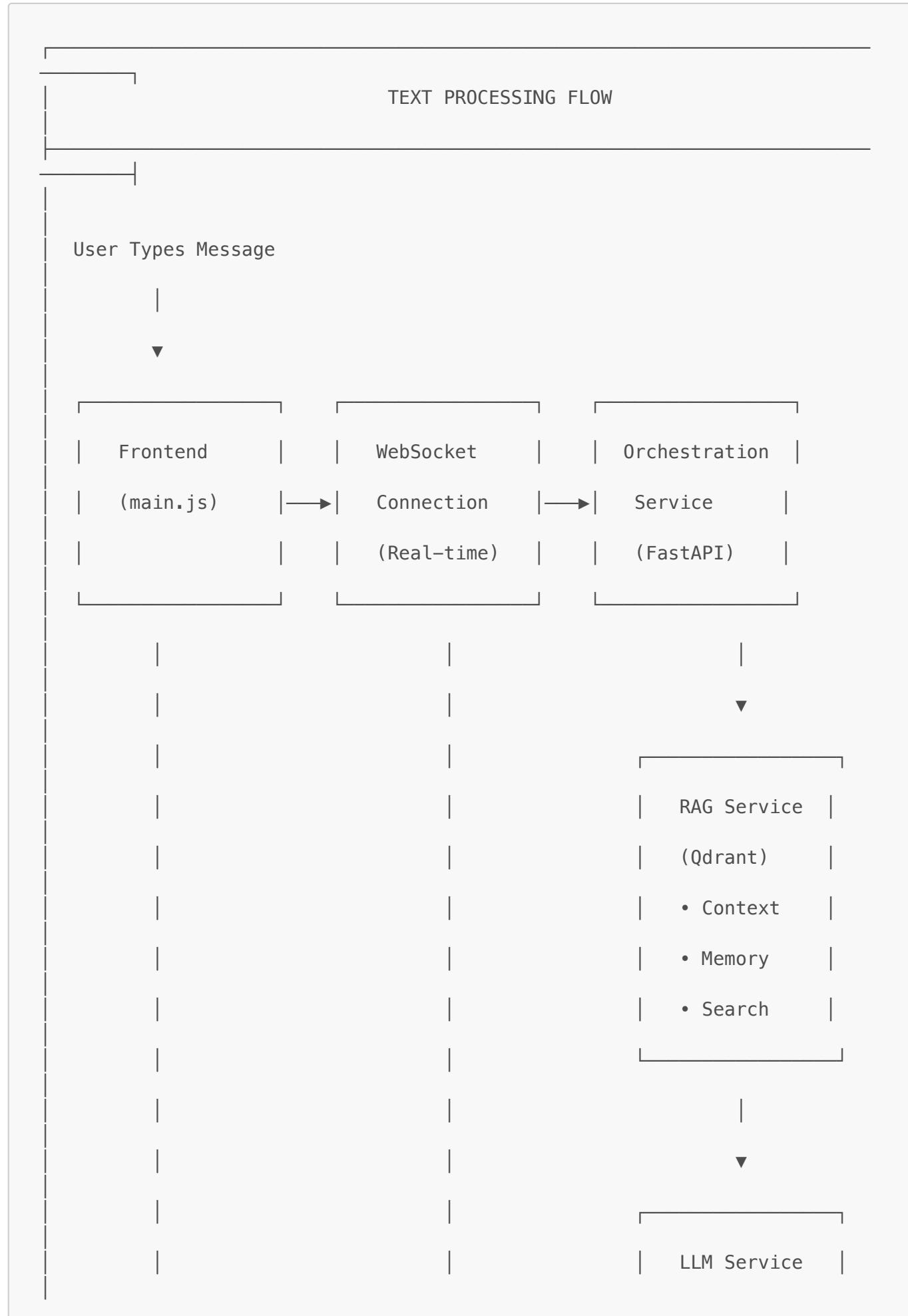
📱 Frontend Layer (User Interface)

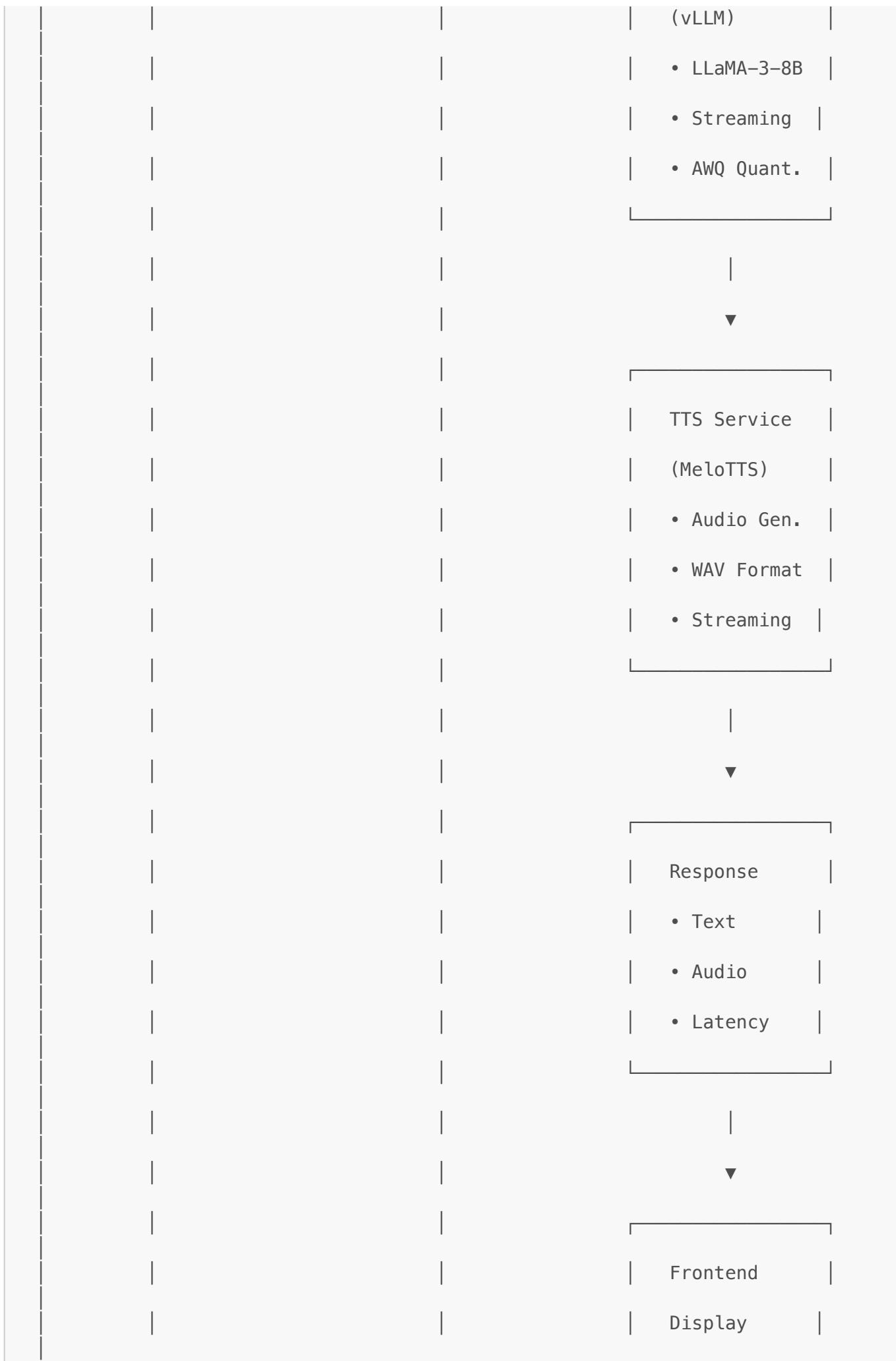
Dual-Mode Interface

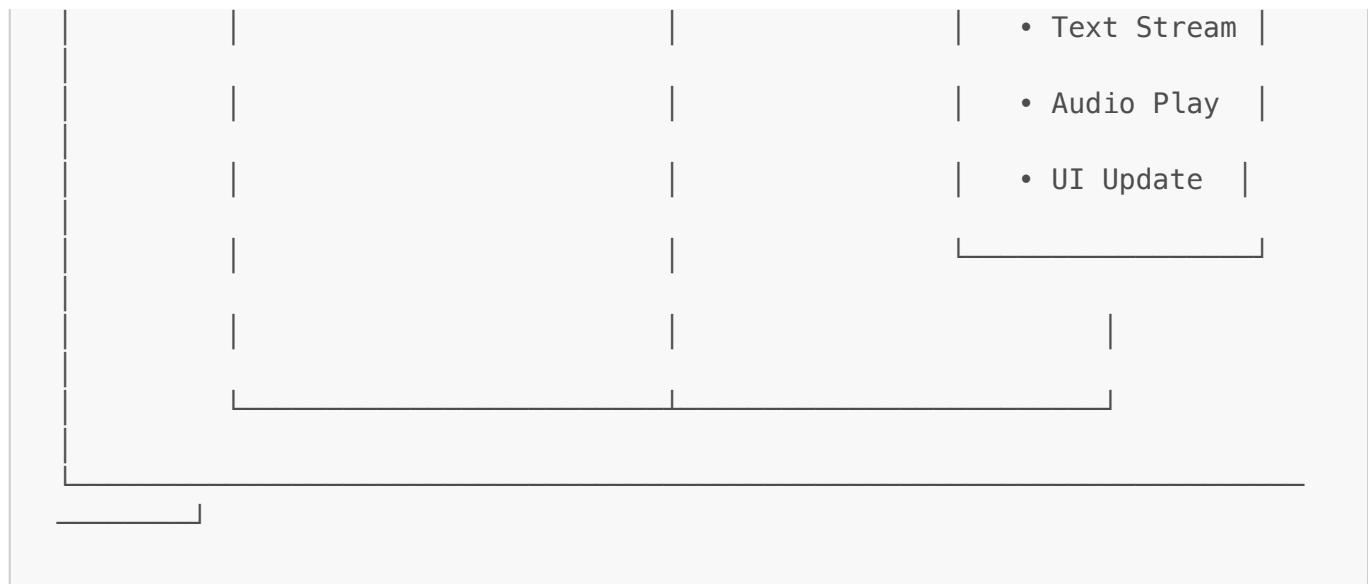


🔄 Processing Pipeline Architecture

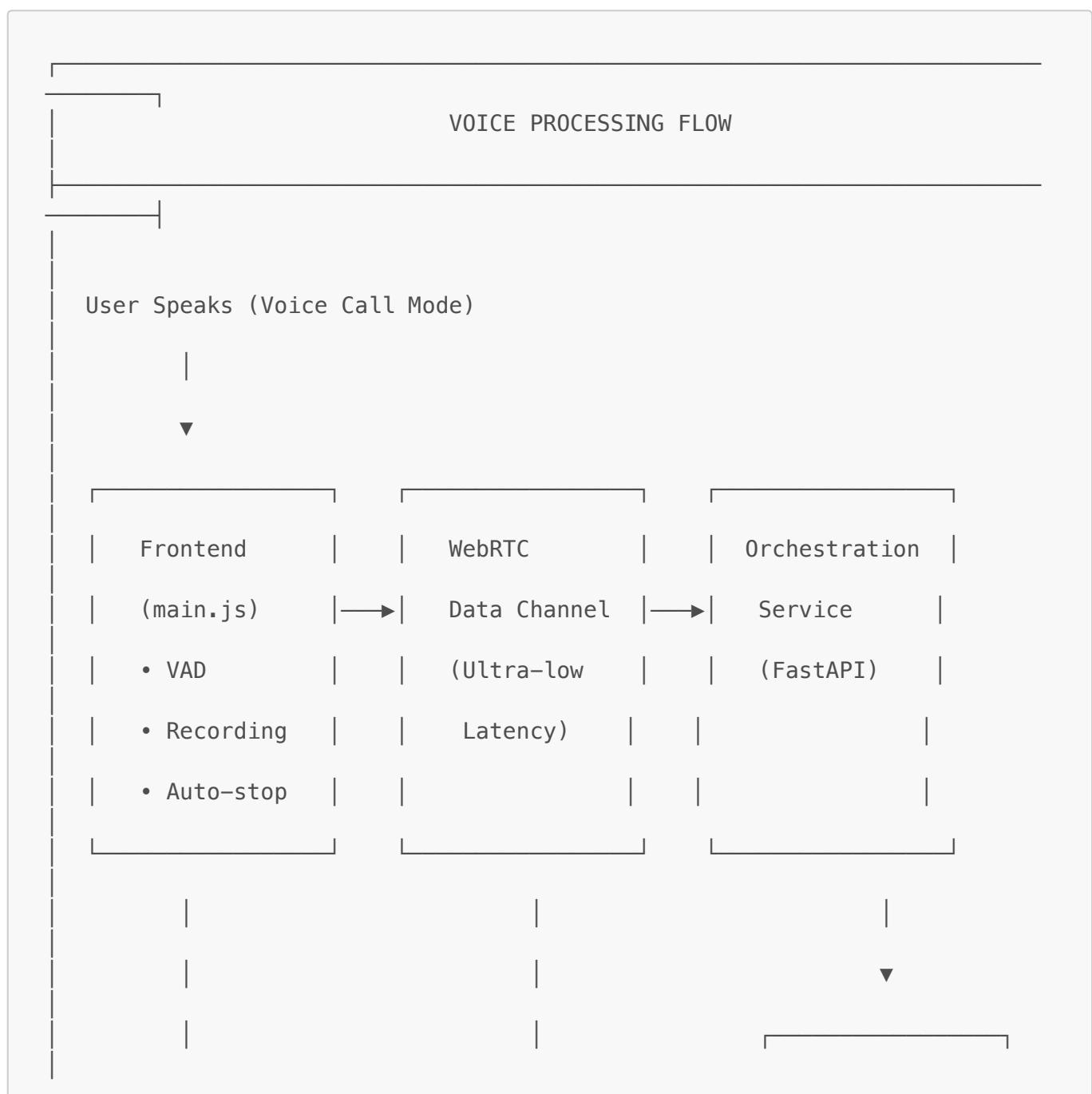
1. TEXT MODE FLOW

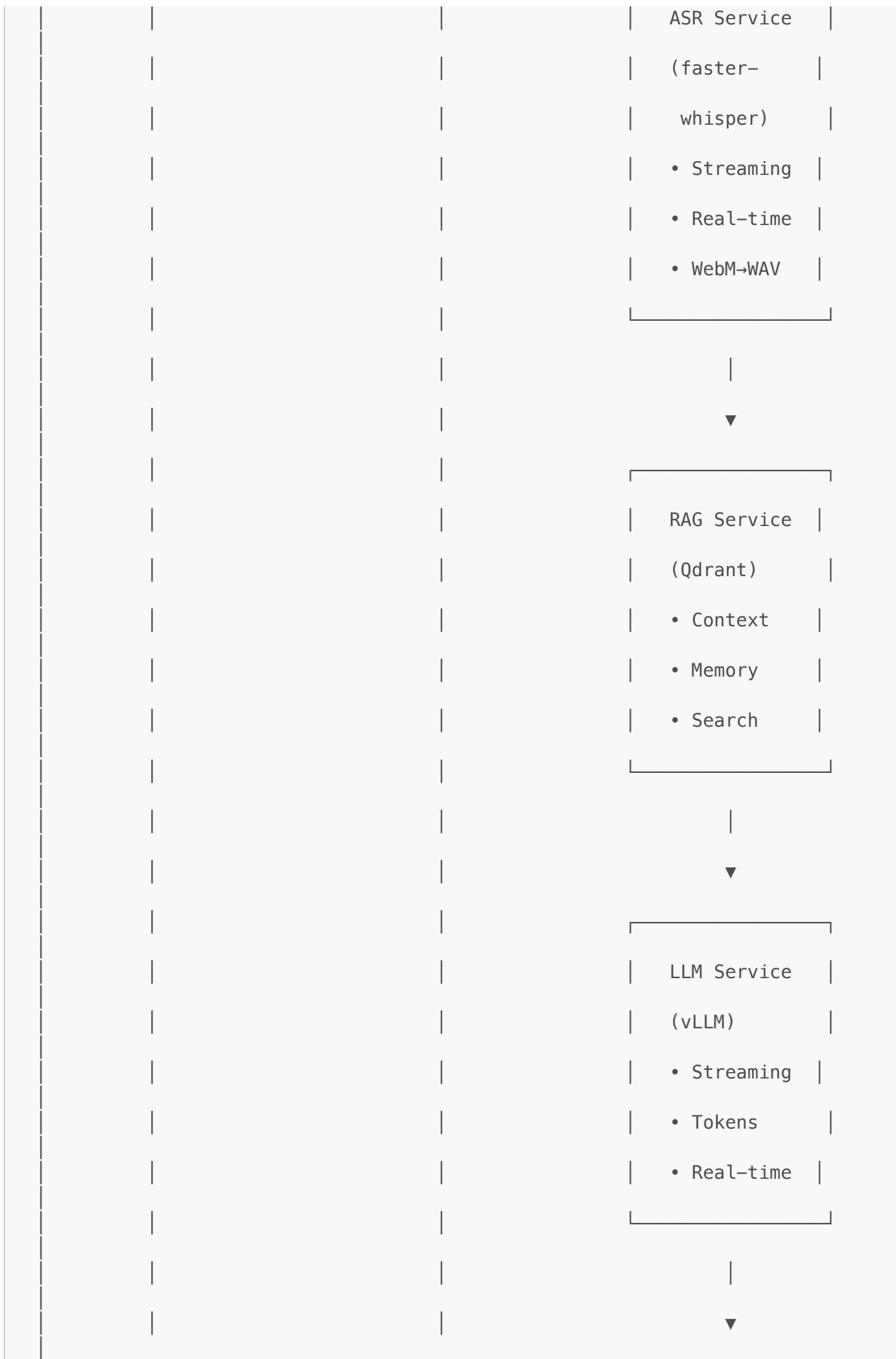


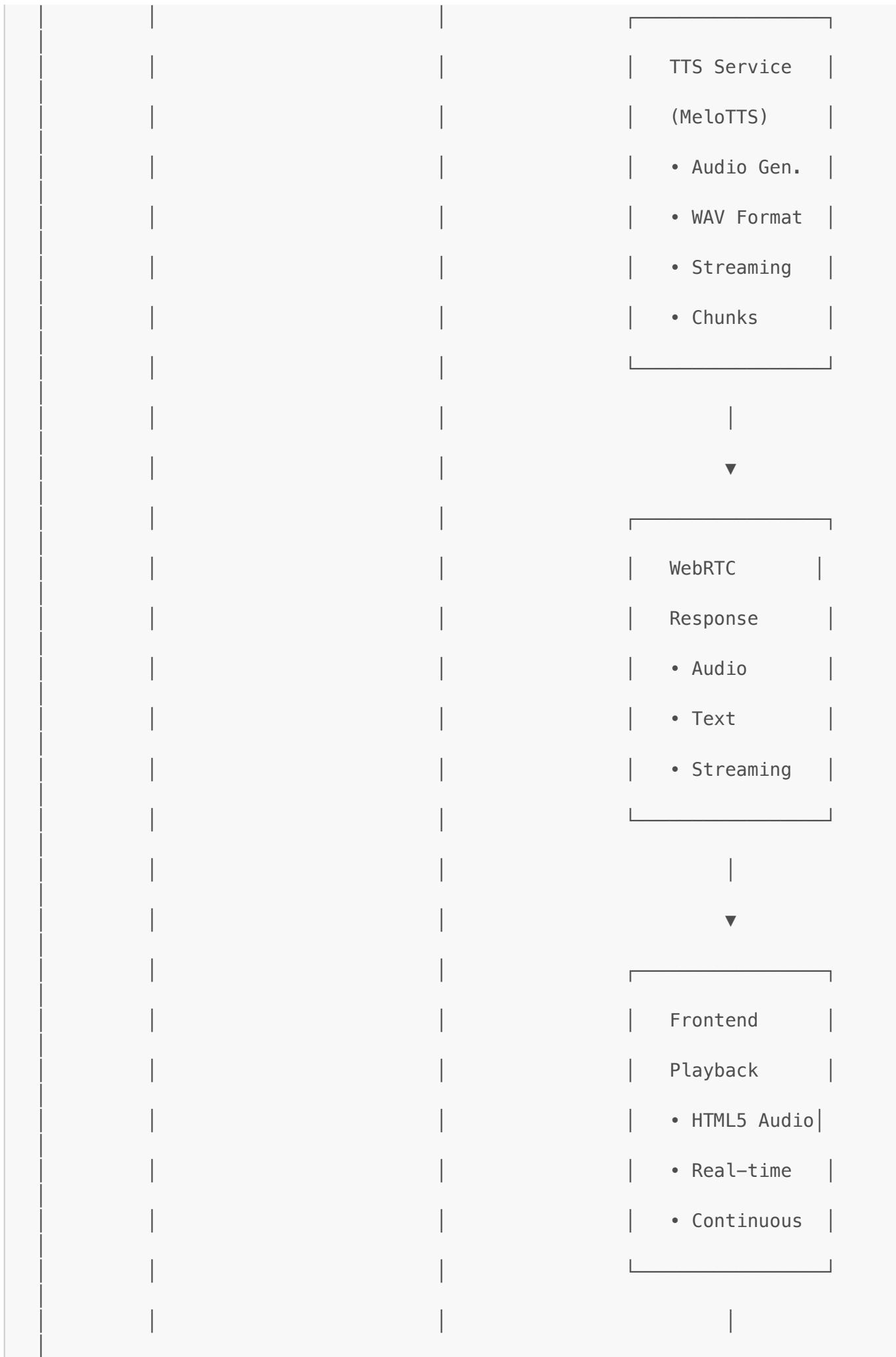




2. VOICE MODE FLOW



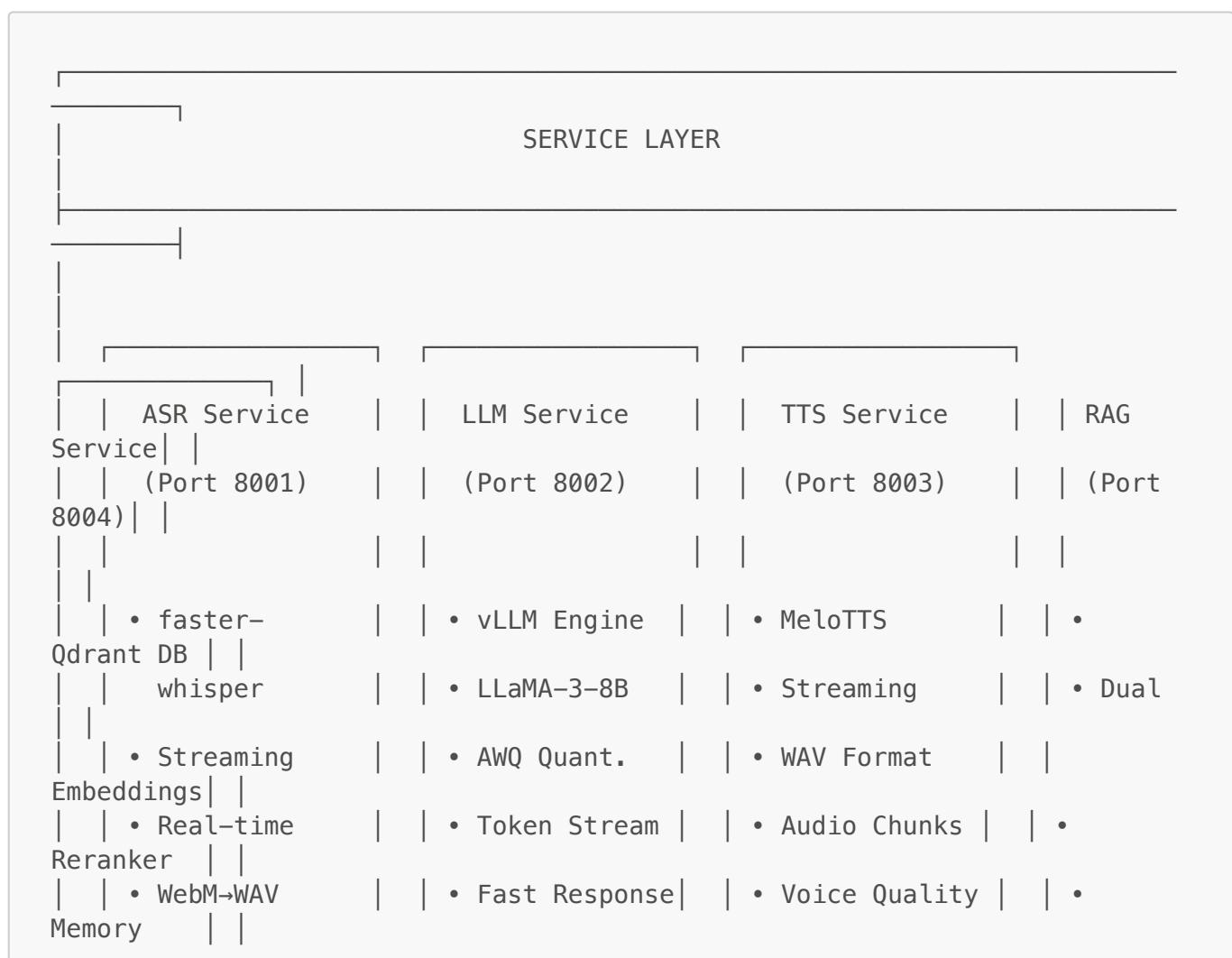


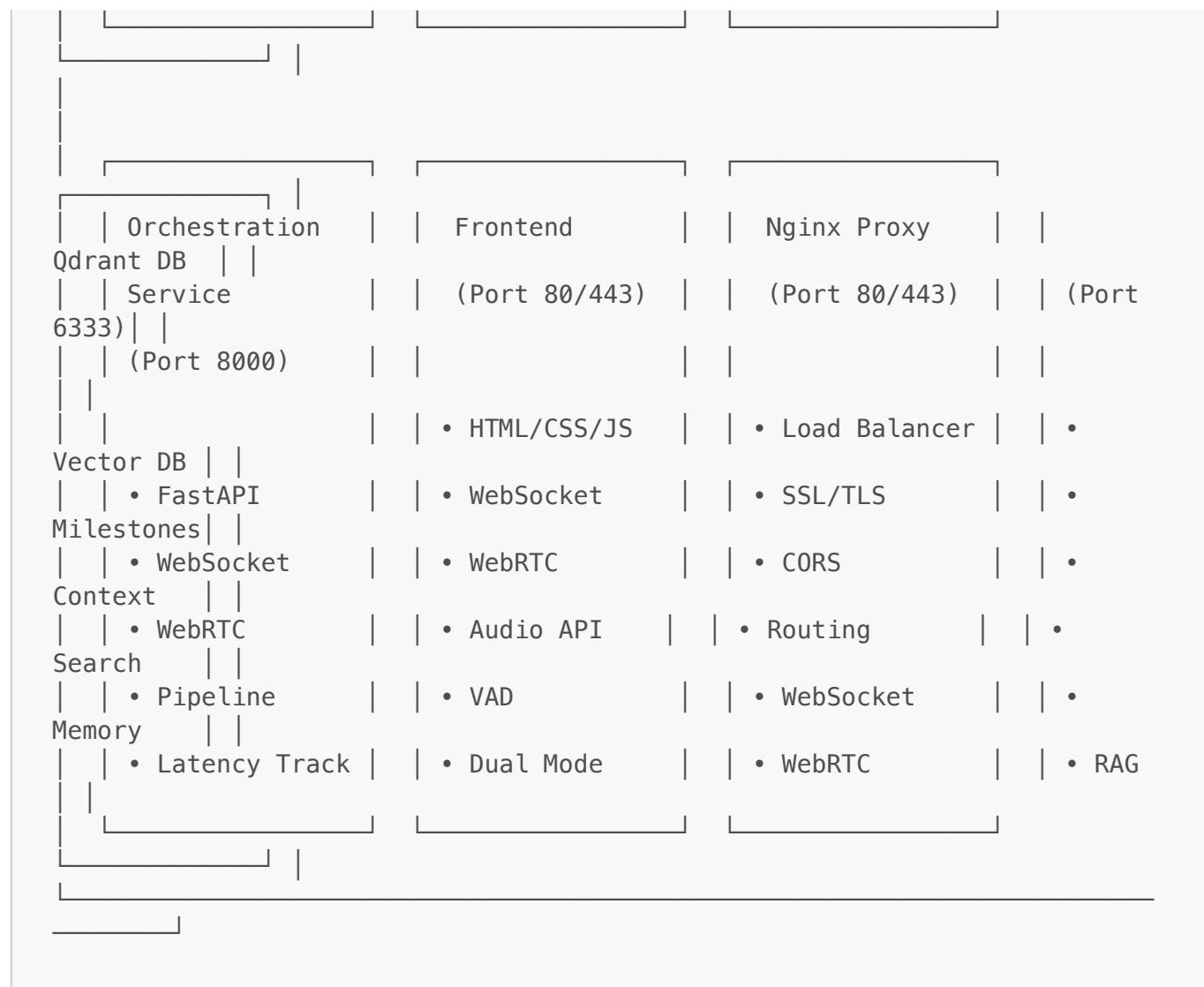




🔧 Service Layer Architecture

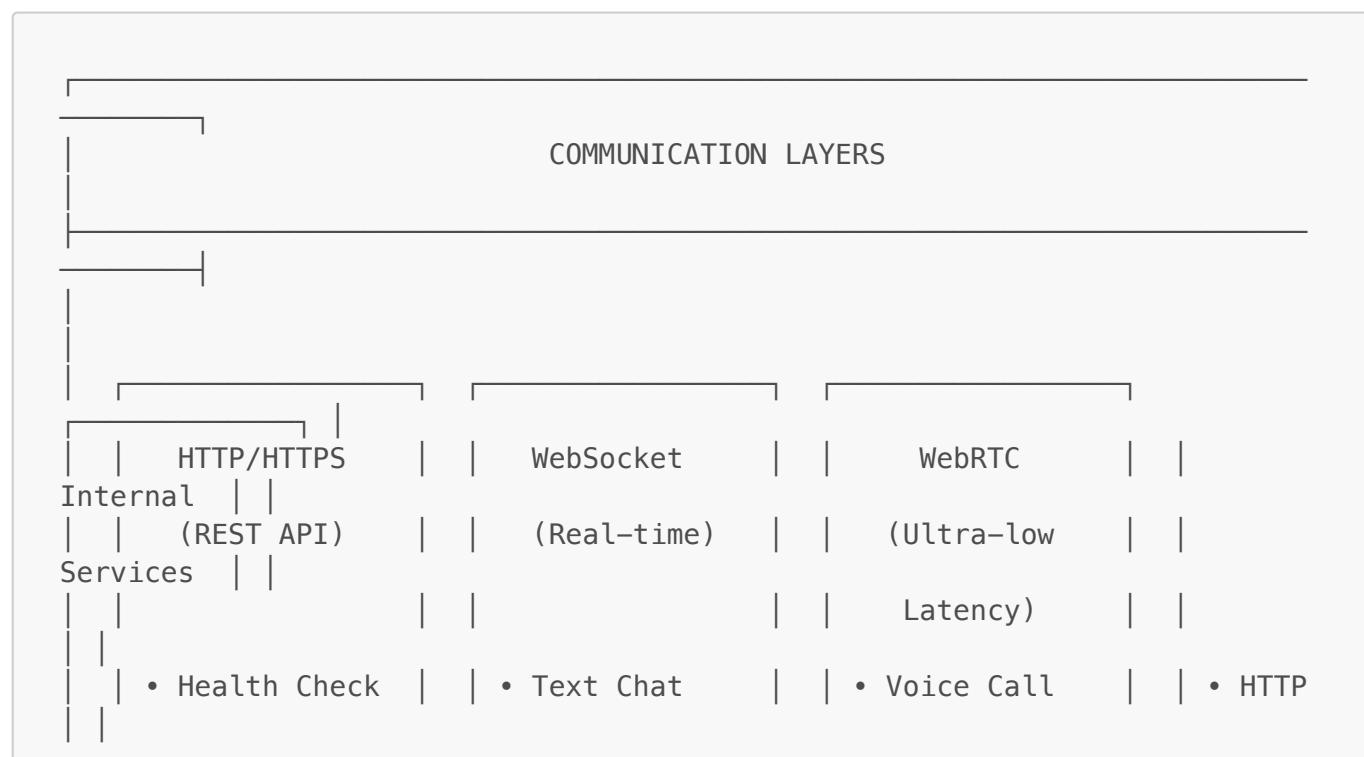
Core Services

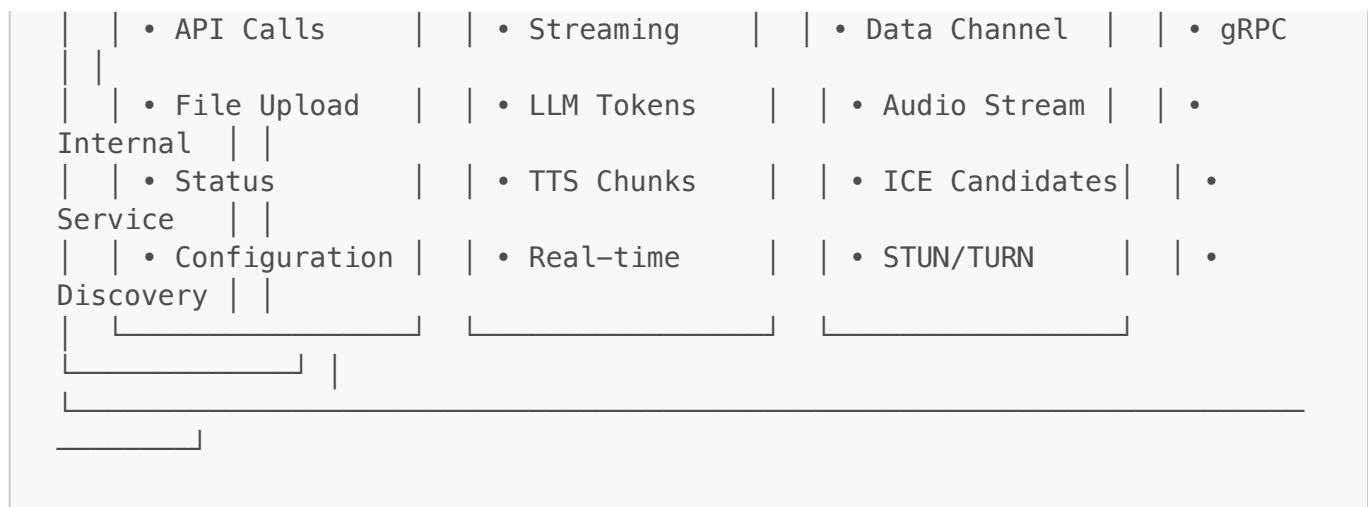




🌐 Network & Communication Flow

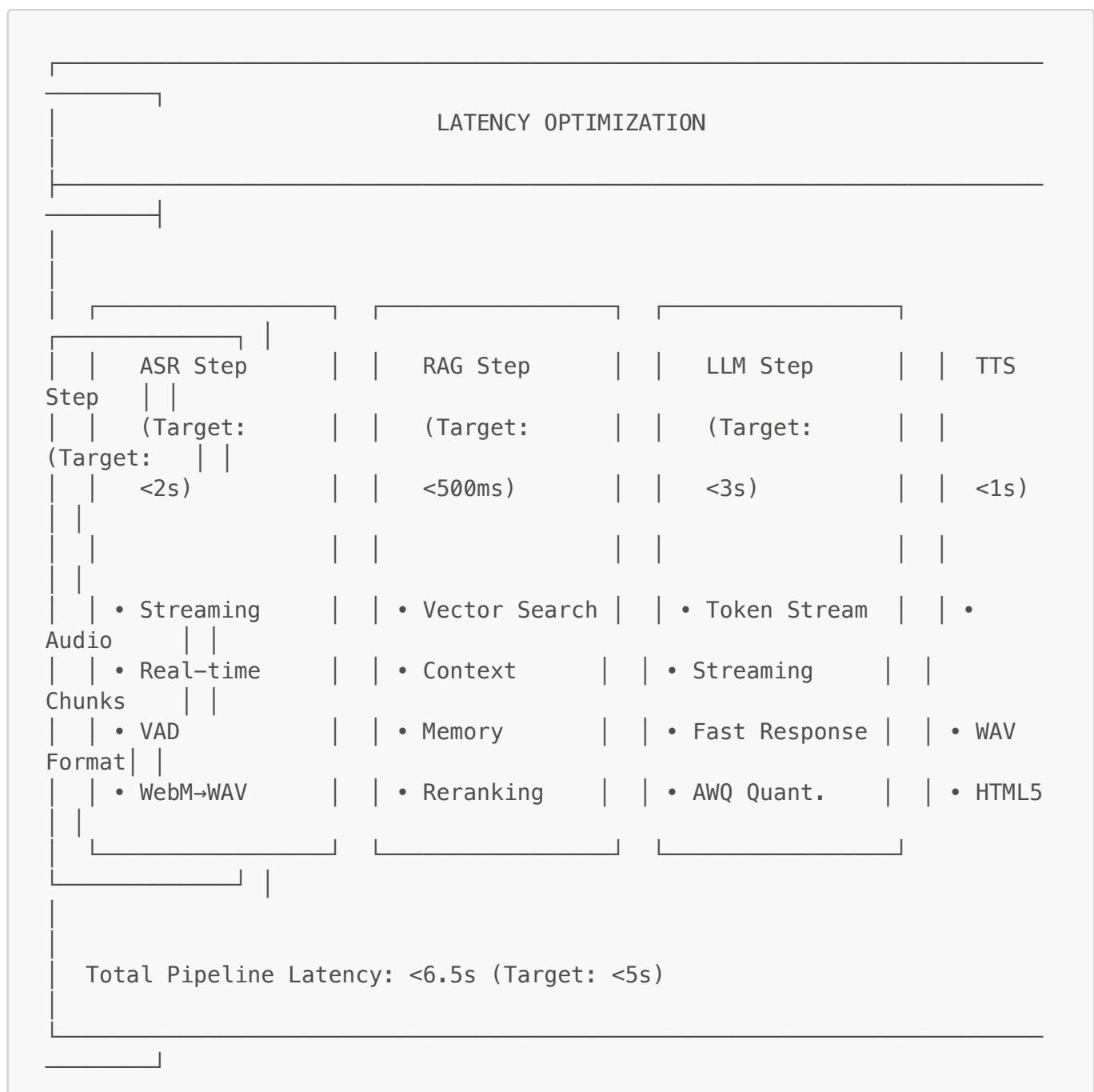
Protocol Stack





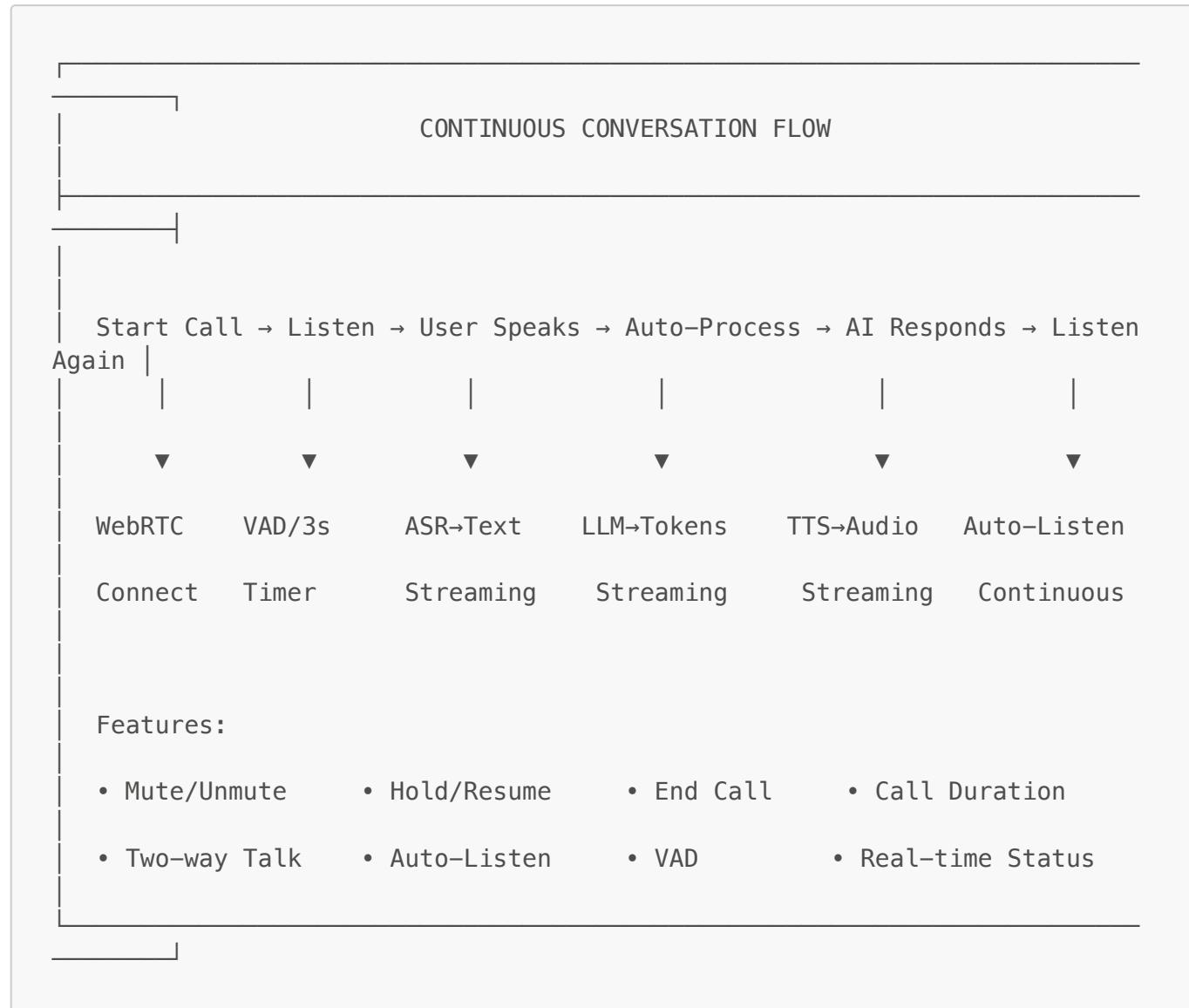
>Data Flow & Latency Optimization

Latency Tracking



⌚ Continuous Conversation Flow

Voice Call Mode (Phone-like)



🛠️ Technology Stack

Frontend Technologies

- **HTML5/CSS3:** Modern UI with Tailwind CSS
- **JavaScript:** ES6+ with WebSocket & WebRTC APIs
- **Web Audio API:** Advanced audio processing
- **WebRTC:** Ultra-low latency communication
- **VAD:** Voice Activity Detection

Backend Technologies

- **FastAPI:** High-performance Python web framework
- **vLLM:** High-throughput LLM inference
- **MeloTTS:** High-quality text-to-speech

- **faster-whisper**: Optimized ASR
- **Qdrant**: Vector database for RAG
- **WebSocket**: Real-time communication
- **Docker**: Containerized services

Infrastructure

- **Nginx**: Reverse proxy & load balancer
- **Docker Compose**: Service orchestration
- **SSL/TLS**: Secure communication
- **WebRTC**: P2P communication
- **STUN/TURN**: NAT traversal

⌚ Key Features

Text Mode

- Real-time text chat
- Streaming LLM responses
- Audio playback
- Conversation history
- Latency monitoring

Voice Mode

- Continuous voice calls
- Phone-like conversation
- Mute/Hold/End controls
- Auto-listening
- Two-way communication
- Call duration tracking

Performance

- <6.5s total latency
- Real-time streaming
- Ultra-low latency WebRTC
- Optimized audio processing
- Memory management
- Error recovery

🔧 Configuration

Environment Variables

```
# Service Ports
ASR_SERVICE_URL=http://asr-service:8001
LLM_SERVICE_URL=http://llm-service:8002
TTS_SERVICE_URL=http://tts-service:8003
```

```
RAG_SERVICE_URL=http://rag-service:8004

# WebRTC Configuration
WEBRTC_ENABLED=true
ICE_SERVERS=[{"urls": "stun:stun.l.google.com:19302"}]

# Audio Configuration
SAMPLE_RATE=16000
CHUNK_DURATION_MS=30
USE_OPUS=false
BITRATE=16
```

This architecture provides a complete, production-ready conversational AI platform with both text and voice capabilities, optimized for real-time performance and user experience.