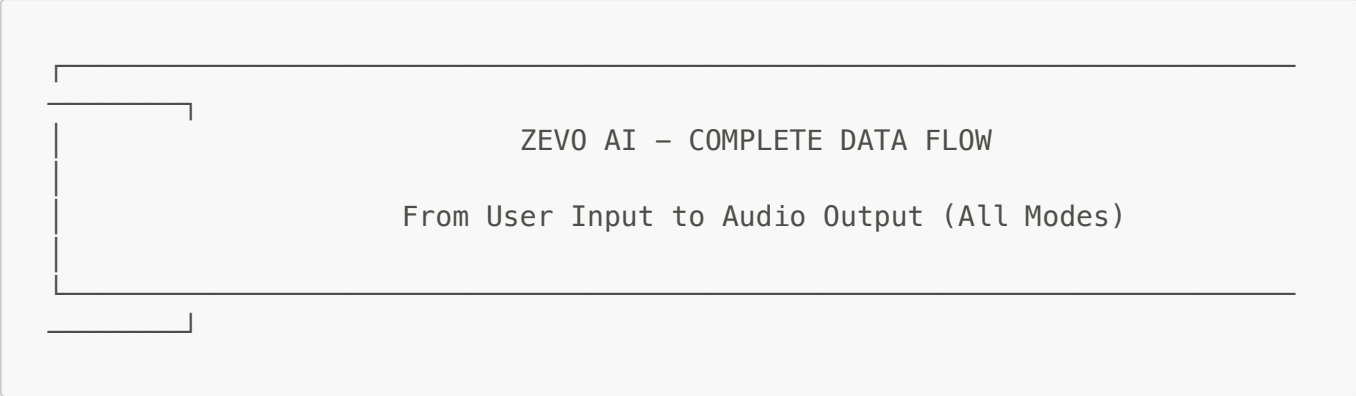
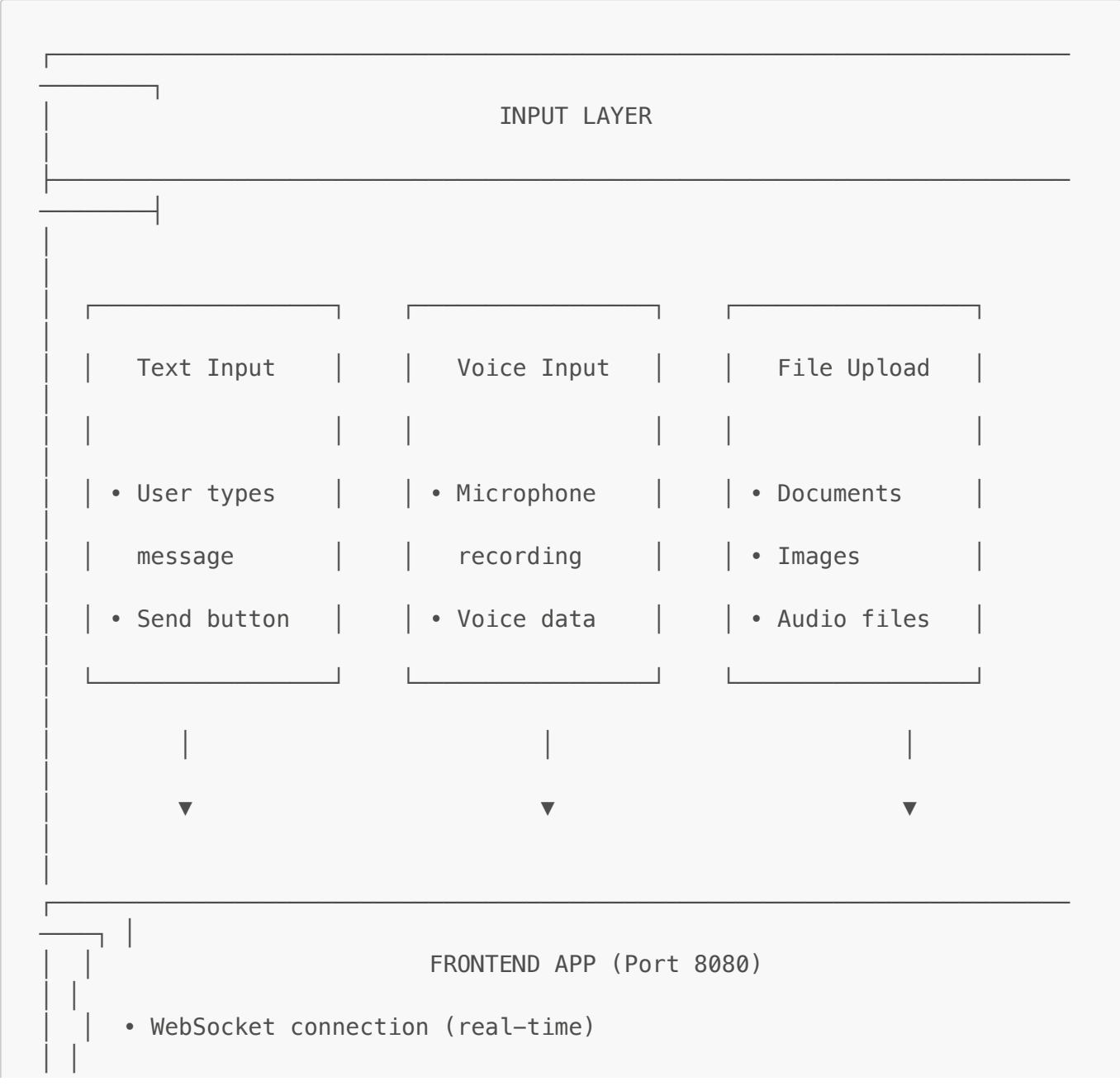
Zevo AI - Simple Data Flow Block Diagram

🔄 Complete Data Flow from Input to Output



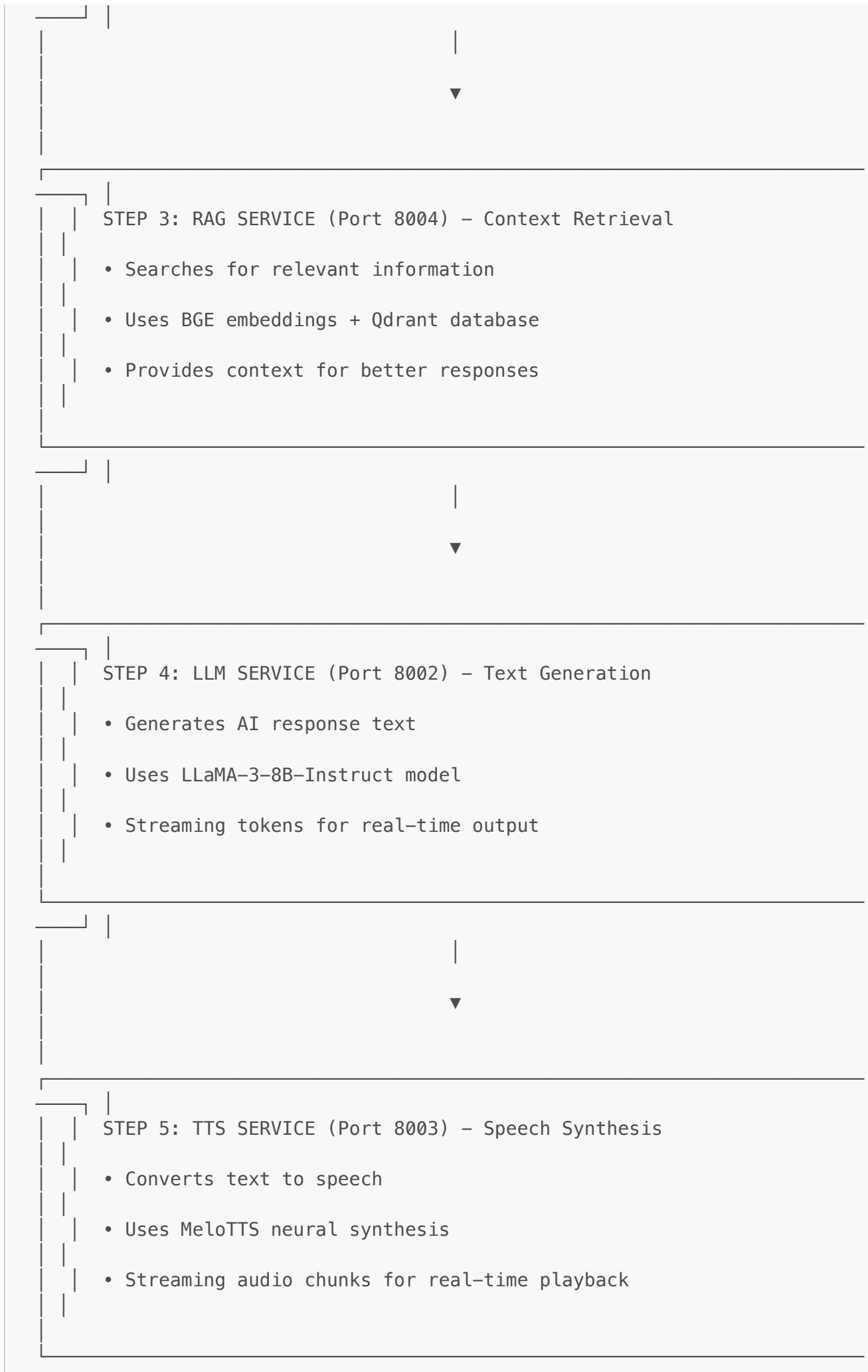
📱 Input Processing



```
│   │  • WebRTC connection (voice mode)
│   │
│   │  • HTTP API calls (file upload)
│   │
└──────────────────────────────────────────────────────────────────────────────
─────┘ │
└──────────────────────────────────────────────────────────────────────────────
──────────┘
```

🎯 Core Processing Pipeline

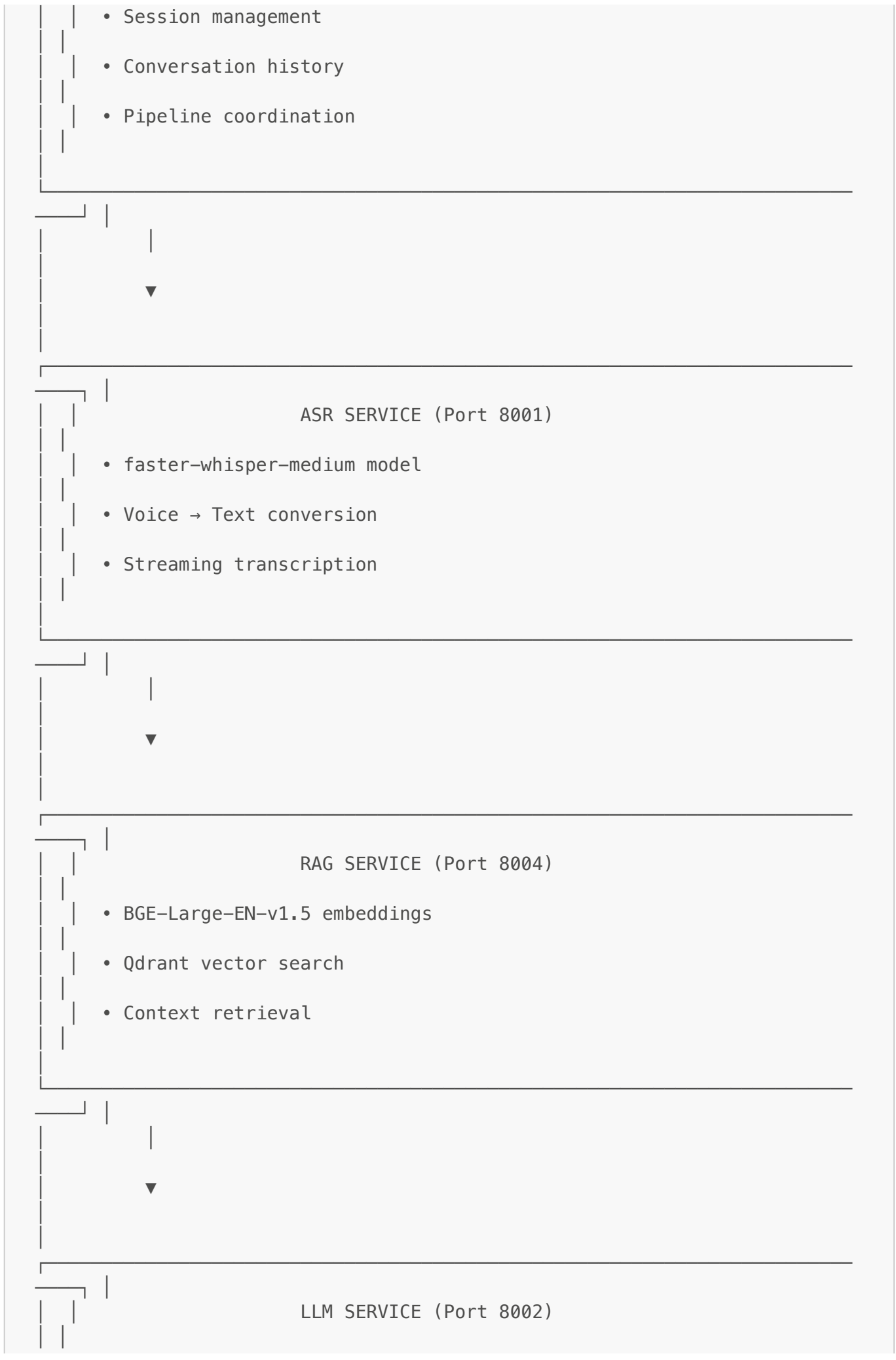
```
┌──────────────────────────────────────────────────────────────────────────────
──────────┐
│                            CORE PROCESSING PIPELINE
│
├──────────────────────────────────────────────────────────────────────────────
──────────┤
│
│
│
┌──────────────────────────────────────────────────────────────────────────────
──────────┐
│   │  STEP 1: ORCHESTRATION SERVICE (Port 8000)
│   │
│   │  • Receives input from frontend
│   │
│   │  • Manages session and conversation history
│   │
│   │  • Coordinates all downstream services
│   │
└──────────────────────────────────────────────────────────────────────────────
─────┘ │
│                                       │
│                                       ▼
│
┌──────────────────────────────────────────────────────────────────────────────
──────────┐
│   │  STEP 2: ASR SERVICE (Port 8001) - Speech Recognition
│   │
│   │  • Converts voice input to text
│   │
│   │  • Uses faster-whisper-medium model
│   │
│   │  • Streaming transcription for real-time processing
│   │
└──────────────────────────────────────────────────────────────────────────────
──────────┘
```

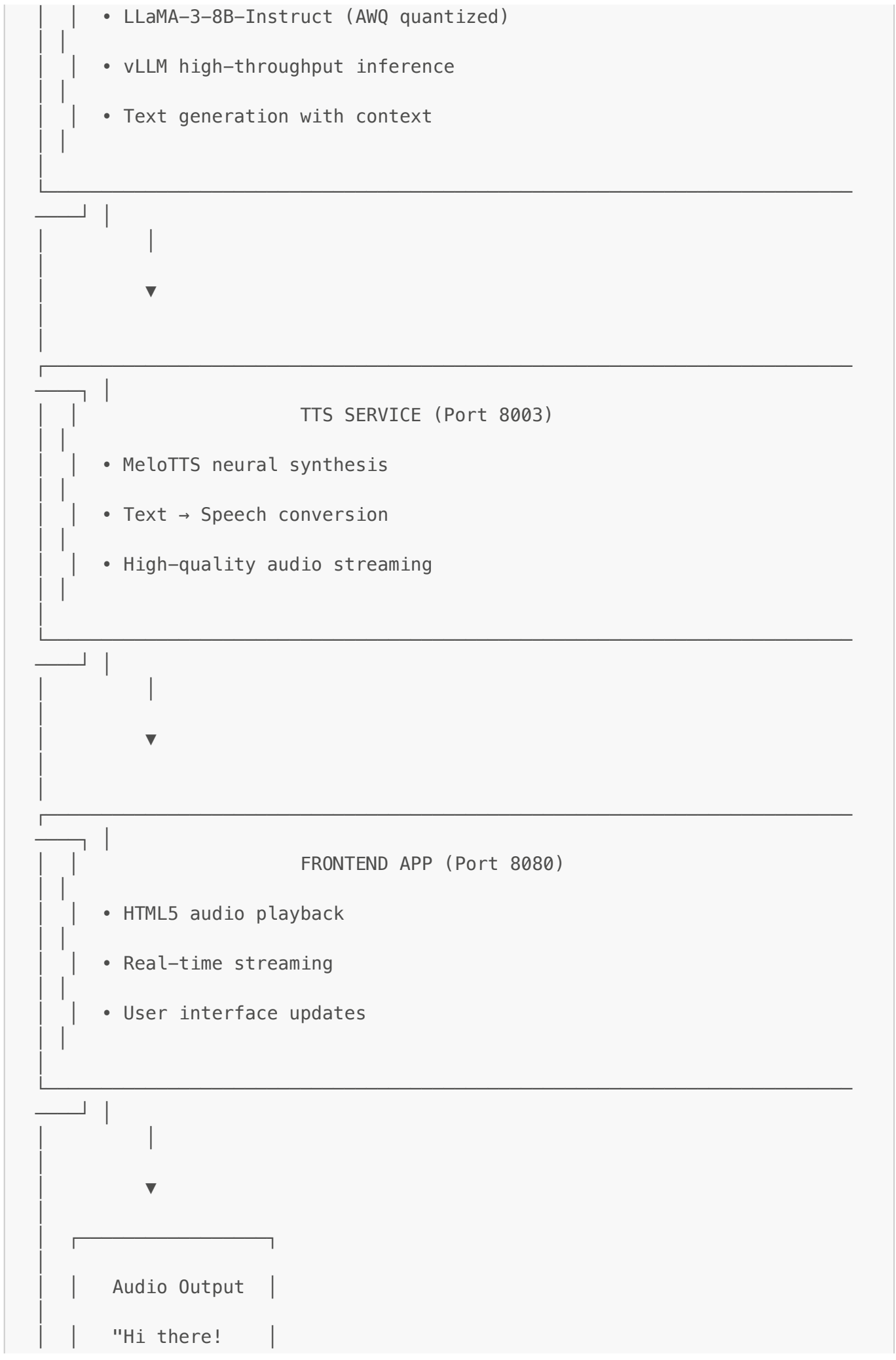

STEP 3: RAG SERVICE (Port 8004) - Context Retrieval
• Searches for relevant information
• Uses BGE embeddings + Qdrant database
• Provides context for better responses
STEP 4: LLM SERVICE (Port 8002) - Text Generation
• Generates AI response text
• Uses LLaMA-3-8B-Instruct model
• Streaming tokens for real-time output
STEP 5: TTS SERVICE (Port 8003) - Speech Synthesis
• Converts text to speech
• Uses MeloTTS neural synthesis
• Streaming audio chunks for real-time playback

```
┌──┐ │
└──────────────────────────────────────────────────────────────────────────────┘
└────────┘
```

🔄 Complete Data Flow Diagram

```
┌──────────────────────────────────────────────────────────────────────────────┐
└────────┐
                        COMPLETE DATA FLOW DIAGRAM
                                                                              │
├──────────────────────────────────────────────────────────────────────────────┘
└────────┤
│
│
│  USER INPUT
│
│  ┌──────────────────┐
│  │   "Hi there!"    │  (Text)
│  │
│  │   [Voice Audio] │  (Voice)
│  │
│  └──────────────────┘
│
│         │
│
│         ▼
│
│
┌──────────────────────────────────────────────────────────────────────────────┐
└──┐ │
│  │                     FRONTEND APP (Port 8080)
│ │
│  │  • WebSocket: wss://agent.zevo360.in/ws/chat/{session_id}
│ │
│  │  • WebRTC: Data channel for voice mode
│ │
│  │  • HTTP: File uploads and API calls
│ │
│
└──────────────────────────────────────────────────────────────────────────────┘
└──┐ │
│         │
│
│         ▼
│
│
┌──────────────────────────────────────────────────────────────────────────────┐
└──┐ │
│  │                ORCHESTRATION SERVICE (Port 8000)
│ │
```


• Session management
• Conversation history
• Pipeline coordination
ASR SERVICE (Port 8001)
• faster-whisper-medium model
• Voice → Text conversion
• Streaming transcription
RAG SERVICE (Port 8004)
• BGE-Large-EN-v1.5 embeddings
• Qdrant vector search
• Context retrieval
LLM SERVICE (Port 8002)


• LLaMA-3-8B-Instruct (AWQ quantized)
• vLLM high-throughput inference
• Text generation with context
TTS SERVICE (Port 8003)
• MeloTTS neural synthesis
• Text → Speech conversion
• High-quality audio streaming
FRONTEND APP (Port 8080)
• HTML5 audio playback
• Real-time streaming
• User interface updates
Audio Output
"Hi there!

```
│
│   │   It's nice to   │
│   │   chat with you"│
│   └──────────────────┘
└──────────────────────────────────────────────────────────────────────────────
──────────┘
```

🎯 Service Communication Flow

```
┌──────────────────────────────────────────────────────────────────────────────
──────────┐
│                          SERVICE COMMUNICATION FLOW
│
├──────────────────────────────────────────────────────────────────────────────
──────────┤
│
│   FRONTEND APP
│   ├── WebSocket ──────────────► ORCHESTRATION SERVICE
│   ├── WebRTC ─────────────────► ORCHESTRATION SERVICE (Voice mode)
│   └── HTTP ───────────────────► ORCHESTRATION SERVICE (File uploads)
│
│
│   ORCHESTRATION SERVICE
│   ├── HTTP ──────────────────► ASR SERVICE (Voice → Text)
│   ├── HTTP ──────────────────► RAG SERVICE (Context retrieval)
│   ├── HTTP ──────────────────► LLM SERVICE (Text generation)
│   └── HTTP ──────────────────► TTS SERVICE (Text → Speech)
│
│
│   RAG SERVICE
│   └── HTTP ──────────────────► QDRANT DB (Vector search)
│
│
│   ALL SERVICES
│   └── Health Checks ────► ORCHESTRATION SERVICE (Monitoring)
```

```
│
└──────────────────────────────────────────────────────────────────────────
─────────┘
```

📊 Data Processing Summary

```
┌──────────────────────────────────────────────────────────────────────────
─────────┐
│                              DATA PROCESSING SUMMARY
│
├──────────────────────────────────────────────────────────────────────────
─────────┤
│
│
│  INPUT TYPES:
│
│  • Text: Direct text input from user
│
│  • Voice: Audio recording from microphone
│
│  • Files: Document uploads for context
│
│
│
│  PROCESSING STEPS:
│
│  1. ASR: Voice → Text (faster-whisper-medium)
│
│  2. RAG: Context retrieval (BGE + Qdrant)
│
│  3. LLM: Text generation (LLaMA-3-8B)
│
│  4. TTS: Text → Speech (MeloTTS)
│
│
│
│  OUTPUT:
│
│  • High-quality audio response
│
│  • Real-time streaming
│
│  • Context-aware conversations
│
│
│
│  TECHNOLOGIES:
│
│  • WebSocket: Real-time communication
│
│  • WebRTC: Ultra-low latency voice
```

```
│
│  • HTTP: Service-to-service communication
│
│  • Docker: Containerized services
│
└──────────────────────────────────────────────────────────────────────────────
──────────┘
```

This simple block diagram shows the complete data flow including ASR (Automatic Speech Recognition) service, which converts voice input to text before processing through the RAG, LLM, and TTS services. The flow is clear and easy to understand! 🚀