Sparrow AI

System Architecture

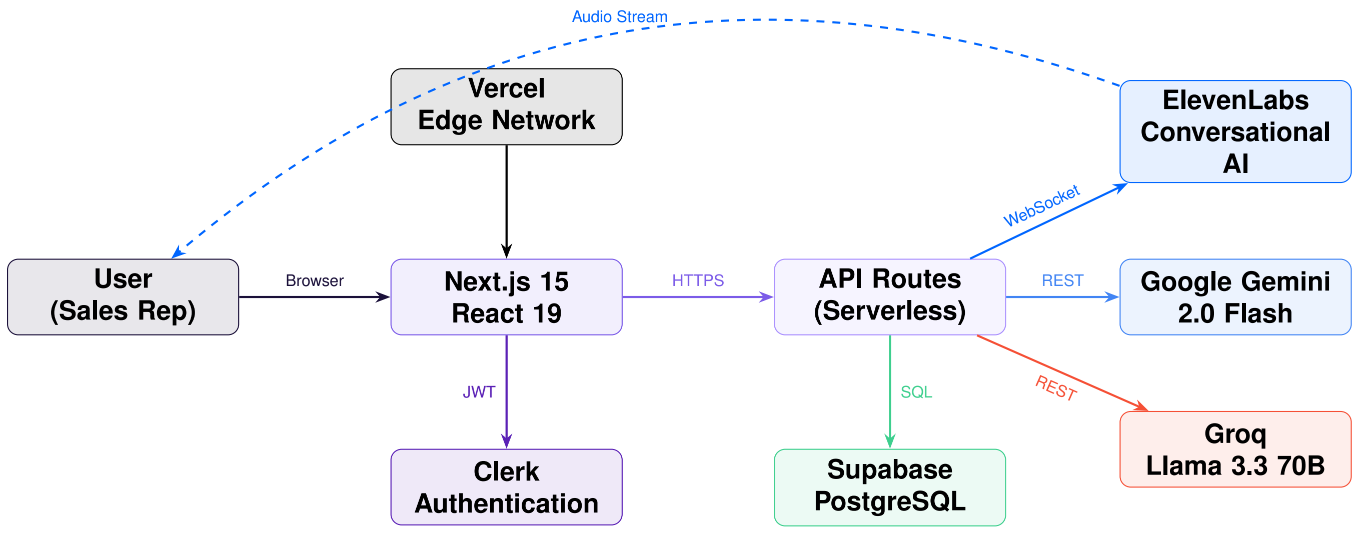
Technical Infrastructure Overview

AI Partner Catalyst Hackathon by Google Cloud

Version 1.0 | December 2025

1 Architecture Overview

Sparrow AI is built on a modern, serverless architecture that leverages best-in-class AI services for real-time voice conversations, intelligent persona generation, and performance analysis.



2 Technology Stack

sparrowPrimary!20 **Layer**	**Technology**	**Purpose**
Frontend	Next.js 15 + React 19	Server components, App Router, real-time UI
Styling	Tailwind CSS 4	Utility-first CSS, custom Sparrow theme
Voice AI	**ElevenLabs**	Real-time conversational AI with voice synthesis
AI/LLM	**Google Gemini 2.0**	Persona generation, deep analysis, feedback
Fast Inference	Groq (Llama 3.3 70B)	Real-time scoring during calls
Database	Supabase (PostgreSQL)	User data, call records, transcripts, analytics
Auth	Clerk	OAuth, user management, session handling
Hosting	Vercel	Edge deployment, serverless functions

3 Data Flow

3.1 Starting a Practice Call

1. User selects practice type (Cold Call / Discovery / Objection Gauntlet)
2. **Gemini 2.0 Flash** generates a unique AI prospect persona
3. Call record created in **Supabase**
4. **ElevenLabs Conversational AI** session initialized with persona context
5. WebSocket connection established for real-time audio

3.2 During the Call

1. User speaks → Audio streamed to **ElevenLabs**
2. ElevenLabs STT → LLM processes with persona context → TTS response
3. Audio response streamed back to user (sub-500ms latency)
4. Transcript streamed to **Supabase Realtime** for live display
5. **Groq** provides real-time coaching hints

3.3 Post-Call Analysis

1. Full transcript sent to **Gemini 2.0 Flash**
2. Detailed scoring across 5 skill dimensions
3. Timestamped feedback on key moments
4. Results stored in **Supabase** for progress tracking

4 Integration Highlights

4.1 ElevenLabs Conversational AI (Primary Partner Integration)

- **Real-time voice conversations** with AI prospects
- **Dynamic voice selection** based on persona gender/personality
- **Custom system prompts** injected per conversation
- **Sub-500ms latency** for natural conversation flow
- **Multi-account failover** for high availability

4.2 Google Cloud / Gemini (Required Integration)

- **Gemini 2.0 Flash** for intelligent persona generation
- **Deep conversation analysis** with structured JSON output
- **Skill-based scoring** with actionable feedback
- **Coach Sparrow** AI assistant powered by Gemini

5 Scalability

sparrowPrimary!20 **Component**	**Scaling Strategy**
Frontend	Vercel Edge Network - automatic global CDN
API	Serverless functions - scale to zero, infinite scale up
Voice AI	ElevenLabs managed infrastructure - enterprise SLA
Database	Supabase - managed PostgreSQL with connection pooling
AI/LLM	Google Cloud + Groq - managed, auto-scaling inference

Built for the AI Partner Catalyst Hackathon

Sparrow AI – Never wing a call again.