Jack Michaels, Ally Nakamura, Jack Xiao, and Roshan Swaroop

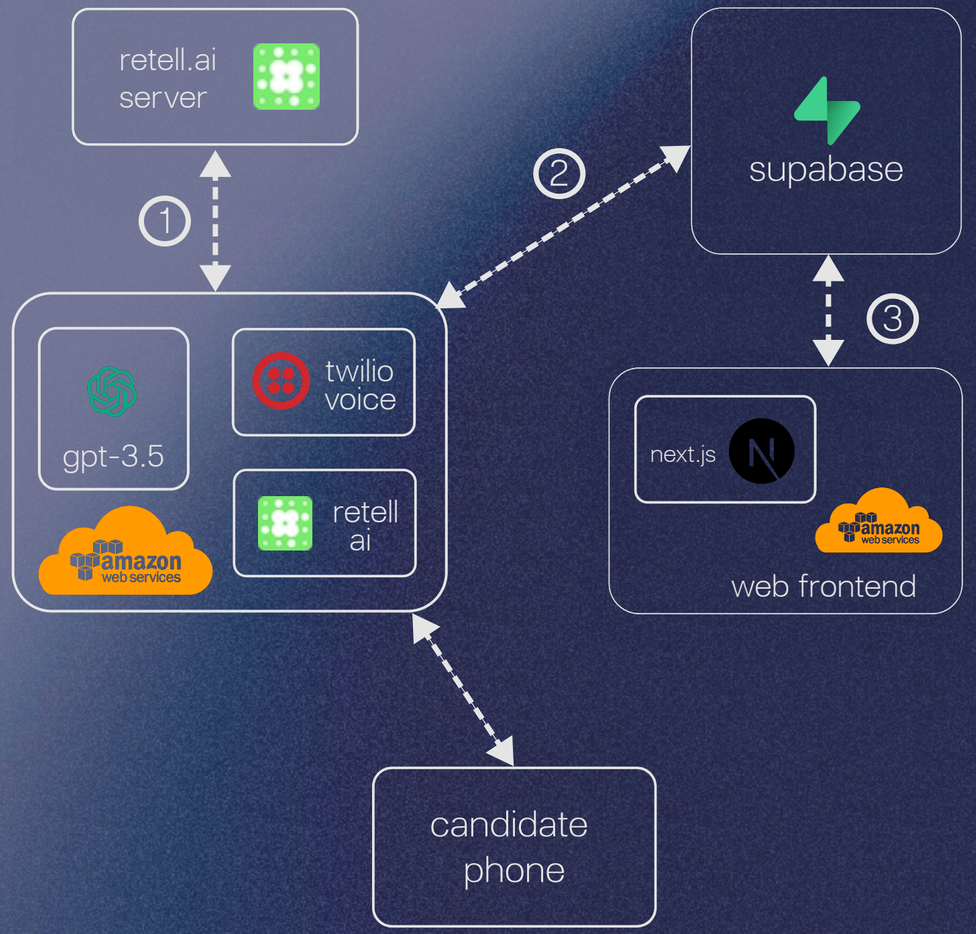
System Design

We ditched naïve integrations and rewrote our entire stack with Retell AI's purpose-built telephony product.

- Decreased latency from 2 seconds to around 800ms, best in class for telephone AI agents
- Added acknowledgement such as "uh huh" and "mhm" when candidate is speaking
- Improved prediction of candidate speech ending
- Decreased response time variability with migration from OpenAI to Azure GPT-3.5-turbo-1106, added streaming of LLM response
- TTS-agnostic between OpenAI and Eleven Labs

We also added a web interface to our product, allowing customers to provide important context for candidate outreach.

- Customers can specify contact information and the information they hope to glean from candidates
- Async calls and summarization available for customers to view



1. Our EC2 instance integrates GPT-3.5-turbo, Twilio programmable voice, and Retell AI's voice agents. We communicate with Retell's west coast-based server for lowest latency to receive our agent profiles.

2. EC2 instance logs completed calls and sends POST to Supabase with call transcript, summary, and flag that call has been completed. Supabase triggers our EC2 instance when certain operations are performed in our database, such as adding a new candidate tuple.

3. Our next.js frontend sends a POST to Supabase with new candidate context. It GETs call transcripts and summaries from Supabase for customers to view how calls have proceeded.

Prompting

catapult.ai

- Making modular prompts that can flexibly handle different inputs
- Long prompts are difficult to work with!
- Split task up into multiple smaller prompts
- Confirm Name
- Confirm Recording Consent
- Reference Checking Call