

RAVINDER PANDEY

AI Backend Engineer | Real-Time Voice, Node.js, TypeScript

Gurugram, Haryana, India | +91-9729041423 | pandeyravinder078@gmail.com
linkedin.com/in/ravinderpandey | github.com/ravin972 | Available immediately

SUMMARY

Backend engineer with 3+ years building production systems, focused on the infrastructure that makes real-time AI work under load. Owned the live call path of a multi-tenant AI voice platform in Node.js and TypeScript, covering Twilio Media Streams, streaming speech-to-text, LLM turn handling and text-to-speech playback, cutting first-turn time-to-first-byte from 666 ms to 100 ms on a runtime load-tested to approximately 100 concurrent calls. Builds deterministic control logic around language models rather than delegating decisions to them.

SKILLS

Real-time: WebSockets, Twilio Media Streams, streaming speech-to-text, text-to-speech, bidirectional audio pipelines

AI integration: LLM API integration with OpenAI, Anthropic and Gemini, deterministic control logic around models, provider fallback, streaming voice pipelines

Languages: TypeScript, JavaScript, Python, SQL

Backend: Node.js, Express, REST API design, Prisma, FastAPI, Django

Distributed and async: BullMQ, Celery, worker queues, retries, idempotency, dead-letter queues, distributed locks, Redis Lua scripting, schedulers

Data: PostgreSQL, MySQL, Redis, schema and index design, query tuning, connection-pool tuning

Infrastructure: Docker, Docker Compose, GitHub Actions, CI/CD, AWS EC2

Security: JWT, RBAC, OAuth2, HMAC webhook verification, API-key hashing, multi-tenant isolation

WORK EXPERIENCE

Full Stack Engineer, Backend and AI Systems

MatchBest Software Pvt. Ltd. | Gurugram, Haryana | Feb 2026 to Sep 2026

- Owned the live call path of a multi-tenant AI voice platform using Twilio Media Streams over WebSocket, streaming speech-to-text, LLM turn handling, sentence-queue text-to-speech and playback, including barge-in cancellation and sentence-queue state management.
- Cut first-turn time-to-first-byte from 666 ms to 100 ms, and cumulative text-to-speech time across a 12-turn conversation from 10,075 ms to 7,111 ms, by restructuring synthesis and playback sequencing on the Cartesia HTTP-to-WsSocket path.
- Restored voice reliability under provider failure by root-causing a speech-to-text never-connects outage, implementing Fish to Deepgram fallback and making text-to-speech completion deterministic to eliminate hung turns.
- Moved dialling, extraction, enrichment and webhook delivery off the request path onto 11 BullMQ workers across 12 queues including an integration-events dead-letter queue, with exponential retry and idempotency keys on CRM webhook delivery.
- Reduced Redis concurrency-lock round trips from 33 to 22 per admitted call by returning occupancy count from a Lua script, preserving admission correctness.
- Removed a PostgreSQL capacity ceiling by raising the connection pool from 10 to 25 with a bounded acquire timeout and revalidating capacity at roughly 40 to 60 concurrent calls, on a runtime load-tested to approximately 100 concurrent calls.
- Implemented tenant and sub-user RBAC with org-scoped data access, API-key hashing, call recording with disclosure handling, worker concurrency controls and deterministic cleanup of Redis-backed execution slots.
- Root-caused an 83-minute capacity-blocked job deferral to a time-unit mismatch in a default retry value, cutting deferral to 10 seconds.
- Built the OAuth and synchronisation backend that replaced a third-party aggregator with first-party integrations across Facebook, Instagram, YouTube and TikTok, including 60-day self-refreshing tokens driven by a scheduler and a migration from the deprecated Instagram Basic Display API to Instagram Login.
- Resolved a staging login outage caused by migration-history drift and a missing refresh_tokens.last_jti column, then shipped fail-closed migration handling, database identity checks, PR drift gates and post-deployment login smoke tests.

Backend Developer, Python/Django and Node.js

Venkara Infotech Pvt Ltd | Gurugram, Haryana | Mar 2023 to Nov 2025

- Moved from frontend-led delivery in the first year to backend ownership by the third, finishing at roughly 70 percent backend work across Python and Node.js services with React and Next.js front ends.
- Built a customer support and ticketing system end to end, covering data models, REST APIs, ticket creation and state transitions, assignment routing and notifications.
- Replaced manual SSH deployments with an automated GitHub Actions pipeline that builds and ships Docker Compose services to EC2, with rollback on failed deployment.
- Built and deployed the API layer for an internal CMS used by admins to manage client accounts and records, with role-based access control.

SELECTED INDEPENDENT PROJECTS

Voice Qualification Engine — Independent Project

github.com/ravin972/voice-qualification-bot | FastAPI, asyncio, Twilio Media Streams, Deepgram, OpenAI, React, Vite | Deployed

- Built a config-driven voice qualification engine on FastAPI and asyncio over Twilio Media Streams, with Deepgram streaming speech-to-text and OpenAI JSON-mode classification, where each bot is defined as YAML configuration rather than code.
- Kept eligibility decisions deterministic by restricting the model to normalising speech into YES, NO, REPEAT or UNCLEAR, and implementing qualification logic as a table-driven finite state machine in Python.
- Implemented a four-provider text-to-speech fallback chain, exercised by a real provider billing failure, with per-turn speech-to-text, model and text-to-speech latency logged in structured JSON and tagged by call identifier.

Voxly OSS — Founder and Builder

github.com/voxly-oss/voxly-oss | FastAPI, SQLAlchemy 2.0, Alembic, Pydantic v2, PostgreSQL, Redis, Celery, Next.js, React, Docker | MIT

- Built an open-source agency operations platform that ingests GitHub webhooks and delivers client updates over WhatsApp, Telegram and Slack, on FastAPI with SQLAlchemy 2.0, Alembic and Pydantic v2 over PostgreSQL, Redis and Celery, with a Next.js and React dashboard.
- Implemented an agentic reason-and-act workflow with bring-your-own-key support across OpenAI, Anthropic and Gemini, including a multimodal path that turns a screenshot bug report into a GitHub issue.
- Built one-command project scaffolding via create-voxly, and GitHub Actions workflows covering continuous integration, container scanning, secret scanning and releases.

EDUCATION

Master of Computer Applications (MCA), Gurugram University, Haryana | Pursuing

Bachelor of Computer Applications (BCA), Gurugram University, Haryana | 2022