

HIMANSHU THAKUR

AI & Backend Engineer (Python)

Bengaluru, India · himanshuthakur8119@gmail.com · +91-7678456033

LinkedIn · GitHub · Portfolio · Medium

SUMMARY

Python backend engineer (3 yrs) shipping AI/LLM features in production fintech at Wells Fargo, with work reaching 15,000+ advisors — including an LLM+RAG "AI Teammate" chat assistant with MCP-based CRM tool-calling. Strong in Python, FastAPI, LangChain/RAG pipelines, and event-driven, cloud-native systems (AWS, Kubernetes). Built a cognitive voice AI agent with acoustic sarcasm detection and graph-based belief revision. MLH Fellow; multi-time hackathon winner.

EXPERIENCE

Senior Software Engineer, Wells Fargo

Jul 2023 – Present · Bengaluru, India

- Building **AI Teammate**, a Python LLM + RAG chat assistant that lets 15,000+ advisors retrieve information from Wells Fargo systems in natural language; connects to CRMs via MCP tool calls and skills so advisors can manage clients and log tasks and meetings directly from the chat interface.
- Shipped a GPT-4 + LangChain RAG assistant inside the advisor Client Management Application (CMA), automating routine workflows for 15,000+ financial advisors and reducing manual effort by ~20%.
- Built a **Python data-extraction service** that helps departing advisors collect and export their client data from multiple internal Wells Fargo systems into a single deliverable for a clean handoff.
- Built an NLP document-summarization pipeline (DistilBERT) integrated via REST callouts, processing 10,000+ docs/day and cutting processing time by ~30%.
- Delivered an AI-powered Next-Best-Action feature surfacing insights from client meeting notes and portfolio data, boosting advisor productivity by 25%.

Software Engineering Intern, Datahive Labs

Dec 2021 – Jun 2022 · Remote

- Built a cross-platform primary-diagnosis application (Node.js, MongoDB) connecting patients with doctors, lifting patient-doctor interaction by 40%.
- Implemented a Python collaborative-filtering recommendation engine for personalized clinical content delivery, boosting daily-active-user engagement by 35%.

PROJECTS

Revenant AI — Autonomous Outbound Engineering Agent

Multi-Agent Orchestration · GPT-4o · LLM Web Research · FastAPI · React

- Autonomous outbound engine: an LLM-driven web-research layer mines public signals (job posts, GitHub issues, stack choices) to surface prospect accounts, each with an evidence-backed fit rationale; founder-in-the-loop approves targets or refines the search via natural-language notes.
- Designed the v2 account-qualification gate: a multiplicative ROI score — $(\text{Pain} \times \text{Ownership} \times \text{Timing}) \times \text{Deal Size} \div \text{Build Cost}$ — with hard floors on each axis, so a near-zero signal on any dimension drops the account instead of being averaged away.

Cognitive Voice AI Agent

LiveKit Agents · Neo4j Graph · librosa · SenseVoice · OpenAI · Deepgram · FastAPI · WebRTC

- Built a voice agent with persistent cognitive memory (Neo4j property graph with trust-weighted edges) that remembers callers across sessions, detects sarcasm via real-time acoustic analysis (YIN pitch tracking + P90 RMS energy), and adapts voice prosody, verbosity, and greeting behavior per caller.
- Engineered four closed-loop features — Sarcasm & Truth Filter, Dynamic Prosody, Adaptive Verbosity, and Proactive Empathy — with a belief-decay policy that decays contradicted facts by 70% instead of deleting, preserving full audit history.

SYNC — AI Relationship-Manager Voice Co-pilot

FastAPI · Twilio Voice & Media Streams · WebSockets · GPT-4o · Whisper · Pipedrive

- Real-time call co-pilot: an AI voice agent bridges the RM to a client over Twilio, forks both call legs through Whisper STT, and streams live GPT-4o coaching to the RM over WebSockets — detecting verbal commitments mid-call and one-click writing follow-ups to the CRM via a pluggable adapter (Pipedrive, Salesforce, HubSpot).

Distributed Ride-Matching Backend

FastAPI · Redis GEO · DynamoDB · Pub/Sub · WebSockets

- Real-time geospatial driver discovery via Redis GEOADD/GEORADIUS (sub-200ms P95); event-driven assignment over Pub/Sub built for horizontal scale under high concurrent-ride load.

Email Orchestrator with Document Triage

FastAPI · GPT-4 · OpenAI Embeddings · MongoDB Atlas Vector Search · Kubernetes

- GPT-4 few-shot email classification; ~40% fewer LLM calls via cosine-similarity caching (>0.95) in Atlas Vector Search; scales on async FastAPI workers + K8s HPA for spiky email volume.

TECHNICAL SKILLS

Languages: Python, Go, TypeScript, JavaScript

AI & ML: GPT-4, LangChain, RAG pipelines, MCP (Model Context Protocol) & agent tool-calling, OpenAI Embeddings, Vector Search, DistilBERT, PyTorch, librosa (acoustic analysis)

Voice & Real-time: LiveKit Agents, WebRTC, Deepgram STT, OpenAI TTS, Twilio Voice & Media Streams, WebSockets

Backend & Distributed: Microservices, REST, FastAPI, gRPC, GraphQL, event-driven architecture, Pub/Sub, caching, fault tolerance

Cloud & DevOps: AWS (Lambda, EC2, S3, DynamoDB), GCP, Kubernetes, Docker, Terraform, CI/CD, Vercel, Render

Data & Storage: PostgreSQL, MongoDB, Redis, DynamoDB, MySQL, Neo4j (property graph)

EDUCATION

B.Tech, Information Technology — National Institute of Technology, Kurukshetra

2019 – 2023

CGPA: 9.01 / 10 · Continuing: Applied AI Masterclass — Arpit Bhayani (2026 cohort)

TECHNICAL WRITING

Building a Distributed Ride-Matching Backend (Uber-Style) — Medium, May 2026.

Designing a Dropbox-like File Sync System: What Actually Makes It Hard — Medium, Apr 2026.

Design Patterns for Real Backend Engineers — two-part series (Factory Method; Singleton) — Medium, Jan 2026.

ACHIEVEMENTS

- MLH Fellow 2022 — Major League Hacking Fellowship (~5% global acceptance rate).
- Top Achiever 2025 & winner, Wells Fargo Technology Hackathon 2025 (India & Philippines editions).
- Top 6 Finalist, GrowthX × Ringg AI Voice AI Buildathon 2026 — placed top 6 of 100 builders (June 2026).
- Winner, Replit × Hexaware Hackathon 2025 — production-grade AI agent built in under 24 hours.
- 1st Place, AICrowd ML Hackathon 2021 — ranked #1 of 100 across classification, regression, and CV.