Nikhil Bindal
San Francisco, CA – open to relocate · (857) 313-5445 · nikhil.bindal@outlook.com
LinkedIn · GitHub · Portfolio

Professional Summary

Backend engineer with **6+ years** building large-scale backend systems, real-time streaming pipelines, and LLM-powered infrastructure. Strong expertise in **Python (FastAPI, Django), Node.js, distributed systems, Postgres, AWS/GCP, WebRTC,** and **multi-agent AI architectures**. Experienced leading end-to-end engineering initiatives—from 0→1 architecture to production deployment—across fintech, media, and AI research platforms. Passionate about solving complex, ambiguous technical problems with scalable systems and intelligent automation.

Skills

Back-End: Python (FastAPI, Django, asyncio), Node.js (Express), gRPC, REST, GraphQL
Databases: PostgreSQL, MongoDB, Redis, MySQL, Qdrant, ClickHouse
Cloud & DevOps: AWS, GCP, Docker, Kubernetes, Terraform, Jenkins, CI/CD, Prometheus/Grafana
Real-Time Systems: WebRTC, WS streaming, Kafka, event-driven microservices
AI/LLM Infrastructure: Multimodal RAG, embeddings, semantic search, CrewAI, LangChain, LlamaIndex
Architecture: Microservices, event-driven systems, distributed systems, caching & rate limiting

Experience

AI Solutions Consultant — Backend & AI Infrastructure

San Francisco, CA · **Sep 2024 – Present**

- **Architected** a **real-time AI interview and voice intelligence** platform using Python (**FastAPI**, Django modules), **Postgres**, **Redis**, and **WebRTC**; achieved **<50 ms** latency for 1,000+ concurrent sessions.
- **Designed multi-agent LLM reasoning pipelines** (CrewAI + LlamaIndex) powering semantic question generation and context-aware inference.
- Built continuous streaming infrastructure for STT/TTS and model inference in production.
- Deployed the solution on GCP with **Kubernetes** and introduced monitoring (Prometheus/Grafana) and CI/CD pipelines; collaborated with clients to prioritise features and mentored junior engineers in both front- and back-end best practices.

Full-Stack Developer — Northeastern University

Boston, MA · **Feb 2023 – Dec 2023**

- **Developed a semantic search and data-visualisation platform** for biomedical research, combining a **Python FastAPI/PostgreSQL** backend with React + D3.js dashboards; enabled researchers to search **10 k+ papers** and visualise complex molecular data.
- Built an **AWS Batch pipeline** to run large-scale simulations and orchestrated results back to the front-end, cutting compute costs by **40 %**.
- Led backend design for scientific data ingestion, metadata APIs, and ML inference.
- Created accessibility features including tactile graphics and screen-reader support to broaden platform adoption.

Software Engineer — Times Internet

Noida, India · **Apr 2021 – Jul 2022**

- **Scaled the TOI+ subscription platform**, designing **NodeJS/Python** microservices and integrating Redis/Kafka; handled **8.4 M daily requests** and supported **120 k+ subscribers**, contributing to **$150 M+ ARR**.
- Led migration of **70+ city portals** from legacy stacks to **React micro-frontends**, improving page load times and boosting Lighthouse scores to **92/100** while mentoring a small team.
- Improved personalization accuracy (+9.7% CTR) through backend streaming pipeline redesign.
- Containerised services with Docker and deployed on AWS EKS, reducing infrastructure costs by **35 %** and accelerating release cycles.

Founding Software Engineer — Progcap (Fintech)

New Delhi, India · **Jan 2019 – Mar 2021**

- Built **Node.js/Express microservices** for real-time loan origination and transaction workflows; reduced latency from **8.7s** to **890ms** and processed **22k transactions/sec**.
- Designed event-driven architecture with Kafka and MongoDB; integrated credit-scoring models (XGBoost) to reduce false negatives by **19 %** and enable **$1.2B+ lending volume**.
- Partnered with business stakeholders to align technical roadmaps with regulatory requirements and mentored new hires on full-stack development and DevOps practices.

Software Engineer — Livemedia

New Delhi, India · **Aug 2017 – May 2018**

- Developed an **OCR-based document verification system** using Tesseract.js and **Python/Django**; achieved **92 %+ accuracy** and accelerated onboarding for insurance clients.
- Built a **React Native offline-first inspection app** and a **React.js/Django** claims platform, serving **50 k+ monthly inspections** and reducing claim processing times by **60 %**.
- Implemented local caching and sync logic to ensure reliability in low-connectivity environments.

Selected Projects

- **Gaussian Splatting Knowledge Graph — AI Research Infrastructure (2025):** Designed a full-stack multi-agent LLM system that ingests academic papers, extracts structured entities/relationships, and constructs a queryable knowledge graph using PostgreSQL, Hono, Drizzle ORM, and React Flow. Implemented a 3-agent pipeline (Extractor → Resolver → Validator) using GPT-4o with strict JSON schema outputs and provenance tracking across 40+ document chunks.
- **Trading Platform (2024):** End-to-end MERN + TypeScript app for live stock trading; implemented GraphQL and WebSocket feeds, reducing API calls by **60 %** and delivering responsive dashboards.
- **Conversational AI Platform (2025):** Integrated Google Gemini API and sentiment detection into a React/FastAPI app; supported **500+ concurrent users** with **40 %** lower latency, demonstrating ability to blend full-stack engineering with AI.
- **Financial Data Anomaly Detection (2024):** Built fraud-detection pipeline using Isolation Forest and One-Class SVM; achieved **95 %+ accuracy** and visualised insights via Plotly dashboards.

Education

M.Sc. in Artificial Intelligence, University of the Cumberlands — **2024 – 2025**
M.Sc. in Information Systems, Northeastern University — **2022 – 2024**
B.Tech. in Computer Science & Engineering, Kurukshetra University — **2012 – 2016**