

OM THAKUR

New York City, NY | Open to Relocation | 929-289-9950 | ot2131@nyu.edu | Portfolio | LinkedIn | GitHub

EDUCATION

New York University, New York, NY

Sep 2024 – May 2026

Master of Science in Computer Engineering (GPA: 3.9/4.00)

Coursework: Machine Learning, Deep Learning, ML Systems Engineering & Operations, Big Data, Advanced Computer Vision.

Vishwakarma Institute of Technology, Pune, India

Aug 2019 – May 2023

Bachelor of Technology in Computer Engineering (GPA: 8.86/10)

TECHNICAL SKILLS

AI & ML: LLMs, RAG, Agentic Workflows, Generative AI, Embeddings, Vector Search, Prompt Engineering, Model Evaluation, LangChain, LangGraph, MCP, OpenAI API, Anthropic API, AWS Bedrock, Ollama, Chroma, Langfuse, NumPy, Hugging Face Transformers

Languages: Python, TypeScript, JavaScript, SQL, Java, Bash

Backend & Databases: FastAPI, Node.js/Express, REST APIs, WebSockets, SSE, SQLAlchemy, Spring Boot, PostgreSQL, SQLite, MongoDB, Redis

Frontend: React, React Native, Next.js, Tailwind, Angular

Data & Cloud: Apache Spark, Spark Structured Streaming, Kafka, ETL Pipelines, HPC, AWS (S3, EC2, Glue, Athena), Docker, Kubernetes, Helm, Terraform, GitHub Actions, CI/CD

PROFESSIONAL EXPERIENCE

NYU Langone Health, New York, NY

May 2025 – May 2026

AI Engineer Intern

- Designed an LLM evaluation framework benchmarking **10 models across 6 providers** (OpenAI, Anthropic, Bedrock-hosted Llama/Claude) on latency, cost, faithfulness, and quality across **1,950 evaluations**; surfaced silent failures in citation grounding to inform model selection.
- Built a clinician-facing RAG pipeline over **170 clinical documents** with domain-specific medical embeddings in an in-house NumPy vector store (HIPAA compliance); integrated GPT-4o with Bedrock-Llama fallback for source-cited Q&A.
- Designed the data layer of an internal lab platform with UUID-based entity resolution, ETL pipelines, and FastAPI services; consolidated 5 fragmented laboratory workflows (sample intake, sequencing, QC, analysis, reporting) into a centralized system adopted by **20+ researchers**.
- Automated sequencing and analytics pipelines, reducing per-cohort processing from 5 days to under 1 day (**80%**); processed **50+ TB** across 4 cohorts on NVIDIA A100s across 71 longitudinal timepoints.
- Engineered an ensemble risk-scoring pipeline combining logistic regression with gradient-boosted stacking and parallel HPC execution; achieved **6.5× enrichment** in the top-decile cohort across 3 prediction tasks on an **8,700-patient dataset**.

4iBiz, Pune, India

Jun 2023 – Jun 2024

Software Developer

- Scaled Fytrack, a gym/club platform handling **\$7.2M+ annual payment volume and 10M+ class bookings** across **1,000+ gyms and 20,000+ users**; revamped the Angular/TypeScript frontend, migrated Bootstrap to Angular Material, and delivered 10+ customer-requested features.
- Launched Fytrack Corporate in 4 months, designing the MongoDB data model, Node.js/Express backend, Angular frontend, and JWT-based role management for a corporate wellness platform adopted by **10 pilot organizations**.
- Built Office Assist as a multi-tenant HR platform serving 10 organizations (2,500+ employees), owning Angular frontend, Node.js APIs, MongoDB schema design with tenant isolation, and production issue resolution.
- Reduced dashboard and reporting latency from 4s to under 2s (**50%**) for **1,000+ gyms** through query optimization, server-side pagination, and database indexing; maintained sub-2s p95 page loads for 20,000+ active users.
- Led delivery of **18 production features** (corporate wellness onboarding, multi-tenant billing, reporting automation) with 12 enterprise customers; mentored 3 interns, conducted code reviews, and drove sprint planning across an 8-engineer Agile team.

4iBiz, Pune, India

May 2022 – Jun 2023

Software Developer, Intern

- Contributed production features across Fytrack, Office Assist, and client applications, owning functionality across gym management, HR, and business workflows while supporting customer requirements, issue resolution, and releases.
- Built **12 end-to-end features and 25 REST APIs** using Angular, React, Node.js, Express, and MongoDB, independently owning frontend, backend, database design, testing, and release support across multiple production applications.
- Designed Postman test suites covering **100+ API endpoints**, uncovering integration defects, validating business workflows, and supporting UAT and regression testing across customer-facing applications.

PROJECTS

DealForge: Agentic Negotiation Platform [Link]

Sep 2025 – Dec 2025

Python, FastAPI, LangGraph, Langfuse, SSE, Next.js 14, Tailwind, SQLAlchemy, SQLite, LM Studio, OpenRouter

- Built an **agentic negotiation platform** on LangGraph: **10 Qwen3-1.7B seller agents** (Firm Pricing, Aggressive Discounter, Bundler, Limited Inventory, Slow Responder, Loyalty Builder, Premium Positioner, Price Matcher, Clearance Seller, Hagglers) against a buyer; Langfuse observability and evals.
- SSE-streamed negotiations to a Next.js 14 dashboard with on-device LM Studio inference on **Snapdragon X Elite NPU** + OpenRouter fallback; integrated credit card rewards optimizer (cashback/points) and Deal Explanation Engine producing ranked deals with constraint-based reasoning.

EDGAR Lens: Self-Evaluating SEC 10-K Analysis System [Link]

Jan 2026 – May 2026

FastAPI, Next.js/TypeScript, PostgreSQL, Chroma, NumPy, Jaccard, OpenAI/Ollama-compatible LLM

- Built a **self-evaluating SEC 10-K analysis system** on FastAPI + Next.js + PostgreSQL with an agentic orchestration loop that drafts citation-backed reports, self-evaluates Jaccard token-overlap citation precision, and adjusts retrieval strategy (focused → broad) across rounds.
- Designed reproducible evaluation suite across 19 SEC 10-K filings (5 sectors, 212–1,364 chunks/filing); achieved **100% citation grounding** (Jaccard ≥ 0.20) with median 1-round convergence and zero silent failures; framework instrumented for LLM-mode ablation.