Rohan Shinde

+1-213-234-8020 | rohanshi@usc.edu | linkedin.com/in/rohanshinde24 | github.com/rohanshinde24 | Portfolio | Los Angeles, CA

Education

University of Southern California — Los Angeles, CA
Master of Science in Computer Science | GPA: 3.76 — Jan 2025 – Dec 2026 (Expected)
Coursework: Analysis of Algorithms, Database Systems, Applied NLP, ML for Data Science, Information Retrieval

Dr.Vishwanath Karad MIT World Peace University — Pune, MH
Bachelor of Technology in Computer Science and Engineering | GPA: 8.93 — 2020 – 2024

Technical Skills

Languages: JavaScript, TypeScript, Java, Python, SQL, C++, HTML, CSS
Web: Express.js, React.js, Spring Boot, Node.js, FastAPI, Flask, REST, Testing (JUnit, Mockito, pytest)
Cloud & DevOps: Docker, Kubernetes, Git, GitHub Actions, AWS(EC2, S3, RDS), Azure, Jenkins, Linux
Data & ML: PyTorch, TensorFlow, scikit-learn, Spark, NumPy, Pandas, HuggingFace, FAISS, Pinecone, RAG
Databases: PostgreSQL, MySQL, MongoDB, Redis
Concepts: Data Structures & Algorithms, Object-Oriented Design, Multi-threaded Programming, Distributed Systems, Networking Protocols, System Performance Optimization, Complex Software Debugging, CI/CD, Agile, Scrum

Experience

Software Engineer (Student Worker) — May 2025 – Present
University of Southern California – Advancement Services — Los Angeles, CA

- Developed full-stack search solution using **React/TypeScript** frontend with SharePoint **REST API** backend, implementing NLP-powered intent detection and similarity scoring algorithms for real-time suggestions
- Engineered automated data pipeline using Python with **multi-API integration** (Tableau + SharePoint), implementing **OAuth** authentication and distributed caching for 440+ users across 19 administrative groups
- Built custom monitoring integration between IBM Cognos Analytics and Azure Monitor using **REST APIs** and Python middleware, reducing monitoring overhead by **40%**
- Refactored legacy **PostgreSQL** pipelines using set-based operations and strategic indexing, reducing runtime by **80%**

Software Engineer Intern (Full Stack & AI) — Jul 2024 – Oct 2024
DMI Finance — Delhi, India

- Optimized a **low-latency C++ backend inference** service by removing hot-path bottlenecks and improving concurrency, delivering **2x throughput** on the same hardware and cutting **response time from 500 ms to 250 ms** for 95% of requests
- Re-architected BERT-based NLP pipeline with automated training/eval workflows and real-time metrics, boosting classification accuracy to **98%** on multi-class intent labels (**+17%** over baseline)
- Designed a **scalable** LLM powered hybrid streaming and batch pipeline using vLLM and Groq for synthetic data and labels, enabling **concurrent training** of domain-specific models across 5+ financial use cases
- Fine-tuned LLaMA 3.1 with QLoRA, improving precision from **0.26** to **0.75** across **35** classes (**+190%** relative; **2.9**×)

AI Engineer Intern — Dec 2023 – May 2024
ResoluteAI Software — Bangalore, India

- Architected a scalable application by delivering **FastAPI microservices** with Google OAuth 2.0, JWT, and end-to-end HTTPS behind Nginx, added validation, retries, and structured logging
- Engineered containerized, scalable embedding services using Pinecone and FAISS for secure vector search, optimizing index configurations to achieve **sub-100ms** latency and a **20%** lift in semantic recall

Projects

QueryLens (GitHub) | *Java, Spring Boot, PostgreSQL, REST API, Docker, TDD, CI/CD*

- Developed a full-stack PostgreSQL performance tool using **Java & Spring Boot**, with a **REST API** for deep query analysis, intelligent pattern detection, and automated rewriting for more efficient execution plans
- Architected a modular optimization engine using the **Strategy design pattern** to rewrite SQL anti-patterns like non-SARGable predicates, reducing query latency over **80%** on average by converting full table scans into efficient index seeks
- Engineered an automated **CI/CD** pipeline with GitHub Actions and a comprehensive **JUnit + Mockito** test suite (**TDD**), achieving **90% code coverage** on core optimizer logic

RetentionPulse (Demo) | *FastAPI, Python, TypeScript, React, Docker, CI/CD*

- Engineered an end-to-end, low-latency FastAPI microservices system using **asynchronous I/O, connection pooling, and request batching** and deploying it via a **CI/CD** pipeline with automated contract testing (**pytest**)
- Owned the full deployment and infrastructure for 4 containerized services, implementing a zero-downtime continuous deployment workflow and managing inter-service communication through a central **API gateway**

FinTrackr | *React, Node.js, PostgreSQL, Docker, GitHub Actions*

- Designed and built a full-stack finance tracker, implementing a secure **Node.js REST API** and a normalized PostgreSQL schema, while achieving **90% test coverage** to ensure backend reliability and data integrity
- Implemented responsive **React** frontend with semantic HTML/CSS and accessibility-first design, ensuring WCAG compliance; containerized with **Docker** and deployed to **Azure App Services** via GitHub Actions, reducing deployment time by **70%**