# Rohan Shinde

(213) 234-8020 | rohanshi@usc.edu | linkedin.com/in/rohanshinde24 | github.com/rohanshinde24 | 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; Machine Learning for Data Science; Information Retrieval & Web Search Engines*

**Maharashtra Institute of Technology** — Pune, MH
*Bachelor of Technology in Computer Science and Engineering | GPA: 3.74* — *2020 – 2024*
*Relevant Coursework: Software Engineering; Database Management Systems; Operating Systems; Big Data Technologies*

## Experience

**Data Analyst (Student)** — May 2025 – Present
***University of Southern California – Advancement Services*** — *Los Angeles, CA*
- Refactored and **parallelized SQL pipelines** across **10M+ row** donor datasets, reducing average query latency by **90%** via index optimization and join strategy tuning; restructured legacy logic into modular views, cutting maintenance overhead
- Engineered and deployed a distributed Spark MLlib pipeline to score donor engagement likelihood across 10M+ records, automating feature selection and cross-validation; validated via A/B tests, applying **rigorous statistical methods** to measure impact and boosting campaign email open rates by **15%**
- Developed interactive Tableau dashboards proactively translating business goals into key fundraiser KPIs and donor lifecycle insights, enabling self-service across 4+ departments and cut reporting time by **30%**

**Data Science Intern** — Jul 2024 – Oct 2024
***DMI Finance*** — *Delhi, India*
- Fine-tuned LLaMA 3.1 models using **QLoRA** on synthetic financial data generated from structured transactions and prompts, applied token compression and Chain of Thought prompting to boost classification precision by **190% across 35** intent classes
- Re-architected BERT-based NLP pipeline for financial field classification, integrating automated training/eval workflows with real-time metrics; achieved **98% accuracy** on multi-class intent labels, outperforming prior baseline by **17%**
- Built **highly scalable** LLM powered data generation and labeling pipelines using **vLLM and Groq APIs**, supporting both real-time and batch modes; enabled **concurrent training** of domain-specific models across 5+ financial use cases
- Boosted inference throughput by **100%** in LlamaCpp by profiling core server logic and optimizing token streaming, cache reuse, and thread affinity for deployment in **low-memory, latency-sensitive environments**

**AI Engineer Intern** — Dec 2023 – May 2024
***ResoluteAI Software*** — *Bangalore, India*
- Architected and deployed a scalable GenAI app using LangChain and OpenAI, integrating **Whisper (audio)** and **LLM summarization (text)** into a cohesive **multimodal pipeline**; delivered secure FastAPI **microservices** with **Google OAuth2 authentication** and HTTPS for production-ready API access
- Engineered embedding pipelines with Pinecone and FAISS for scalable, secure **vector search**; optimized index configs and k-NN retrieval for **sub-50ms latency and 20% lift** in semantic recall under containerized deployment
- Reduced API expenses by **$15k annually** by optimizing token usage via prompt engineering, response truncation, and compression; **partnered cross-functionally** with MLEs to monitor latency-cost tradeoffs across real-time inference pipelines

## Projects

**AgentPilot** | *SwiftUI, GPT-4, CoreML, SiriKit* — May 2025 – Present
- Built an intelligent daily planner, **engineering robust, asynchronous API integrations** for automatic task summarization and schedule generation; conducted **beta testing with 15 users**, with **92% task relevance** and high user satisfaction
- Engineered real-time **task synchronization** using App Intents, Live Activities, and CoreData; reduced manual user input by **60%**, enhancing planning efficiency through automation and context-aware updates

**FinTrackr** | *Angular, NodeJS, PostgreSQL, Docker, GitHub Actions* — Jul 2024 – Aug 2024
- Built a full-stack finance tracker using **Angular**, **NodeJS**, and **PostgreSQL**, supporting secure login, real-time expense tracking, and dynamic savings analytics; implemented **90% test coverage** across backend APIs to ensure system reliability
- Containerized the application with **Docker** and automated deployment pipelines using **GitHub Actions**, achieving a **70% reduction in deployment time**

## Technical Skills

**Languages**: Python, Java, SQL, JavaScript, TypeScript, Swift, HTML, CSS
**ML & Rec. Systems**: PyTorch, TensorFlow, scikit-learn, Spark MLlib, HuggingFace, LangChain, FAISS, Pinecone
**Web & Mobile Frameworks**: Angular, Node.js, Express.js, React.js, Flask, FastAPI, SwiftUI
**Developer Tools & Cloud**: AWS (EC2, S3, SageMaker), Git, GitHub Actions, Jenkins, Docker, JIRA, Excel, Unix, Cursor
**Databases & BI**: PostgreSQL, MongoDB, Redis, Tableau, Cognos
**Concepts**: Statistical & Predictive Modeling, Object-oriented Design, Agile, CI/CD, Real-time Inference, NLP