Sreekant Baheti

+1-213-245-9160 | sreekantbaheti13@gmail.com | linkedin.com/in/sreekantbaheti | github.com/Sreekant13

Summary

Full-stack software engineer with production Python, Java, TypeScript experience across AI startups and enterprise; agentic LLM pipelines, ML services, microservices at 1,247 req/hr, 99.7% uptime.

Education

University of Southern California — Los Angeles, USA
Master of Science in Computer Science, GPA: 3.78/4.0 — *Aug 2024 – May 2026*

- **Coursework:** Distributed Systems, Machine Learning, Database Systems, Algorithms, Software Engineering, Cloud Computing

SRM Institute of Science and Technology — Chennai, India
Bachelor of Technology in Computer Science and Engineering, GPA: 3.68/4.0 — *May 2018 – Jun 2022*

- **Coursework:** Data Structures & Algorithms, Operating Systems, Computer Networks, Object-Oriented Software Engineering, Databases, Web Technologies

Experience

Software Engineer Intern — May 2025 – Aug 2025
The Verse — *San Francisco, CA*

- Designed Python agentic LLM pipelines (GPT-4, Claude) orchestrating document search, entity extraction, summarization across 42 production workflows with streaming outputs, reducing manual ops by 43%
- Containerized inference with Docker, deployed via GitHub Actions CI/CD sustaining 1,247 req/hr peak
- Fine-tuned BART extractor from 0.74 to 0.91 F1, automated rollback cutting recovery 22 to 4 min

Software Engineer — Jul 2022 – Jul 2024
Bank of America — *Haryana, India*

- Architected enterprise Kafka messaging platform on Kubernetes processing 2.3B messages monthly across 5 environments; Ansible automation reduced deploy time from 4 hours to 11 minutes
- Built Grafana/Prometheus/Splunk stack with 47 dashboards, cutting MTTD from 3.2h to 23 min
- Engineered REST microservices with PostgreSQL, Redis caching, serving 14 downstream consumers

Software Engineer Intern — Jun 2021 – Jun 2022
HighRadius — *Bhubaneswar, India*

- Built full-stack B2B invoice platform with React dashboard, Java/Spring backend, XGBoost payment-date model achieving 92.4% accuracy across 1.2M+ historical invoices
- Designed REST APIs with MySQL schema, session auth, serving 3 downstream analytics consumers
- Automated AR aging reports, cutting manual reconciliation from 3.1h to 37 min per billing cycle

Technical Skills

Python · Java · TypeScript · JavaScript · React · Next.js · FastAPI · Node.js · Spring · PostgreSQL · MySQL · MongoDB · Redis · Docker · Kubernetes · GitHub Actions · Kafka · GCP · AWS · scikit-learn · PyTorch · LLM APIs · Grafana · Prometheus · Temporal · Spark · C++

Projects

Orq – AI Incident Response Operator | *TypeScript, Next.js, Temporal, Railway, Claude APIs* — May 2026

- Built 10-node Temporal agentic DAG in TypeScript/Next.js, cutting MTTR from 47 to 18 min
- Deployed 4 Railway microservices via GitHub Actions CI/CD with rollback hooks at 99.7% uptime
- Integrated Claude LLM APIs for root cause analysis at 91% confidence with auto-remediation

relevant-priors-api | *Python, FastAPI, scikit-learn, Render* — Apr 2026

- Deployed FastAPI ML service on Render predicting radiology exam relevance at 98.27% accuracy
- Engineered TF-IDF n-gram features with body-region extractor, lifting grouped CV accuracy to 0.94
- Applied grouped CV by case_id to prevent label leakage, 0.9402 mean accuracy on held-out cases

EcoMate-AI – Carbon Footprint Analyzer | *Python, FastAPI, LLM APIs, Docker* — Apr 2025

- Built Python LLM pipeline with Tesseract OCR, 84.7% extraction accuracy across 2,300+ test inputs
- Containerized async backend at 3.2x throughput, 47 concurrent requests under 800ms p95 latency
- Designed multimodal text/image/audio ingestion cutting parse latency from 1.4s to 0.83s