Sashikumar Nehrumohan

California (Open to Relocate), USA | sashikumar.nehrumohan@gmail.com | 716-305-8512 | LinkedIn | GitHub

PROFESSIONAL SUMMARY

AI Engineer with 3+ years of experience building production-grade LLM systems, multi-agent orchestration pipelines, and ML infrastructure at scale, delivering measurable impact across document intelligence, fraud detection, and agentic workflows.

EDUCATION

State University of New York at Buffalo — **Buffalo, NY**
Master of Science in Computer Science Engineering, AI/ML. GPA: 3.7/4.0. — *Dec 2025*

Anna University — **Chennai, India**
Bachelor of Science in Computer Science Engineering. GPA: 8.4/10.0. — *May 2022*

EXPERIENCE

AI Engineer — **Mar 2026 – Present**
Rebecca Everlene Trust Company — *Remote*

- Architect AI Agentic volunteer management platform supporting 50+ active projects and 20+ coordinators through LangChain orchestration, FastAPI, PostgreSQL state management, and workflow automation.
- Engineer task-management framework supporting 1,500+ volunteers across 5+ channels, reducing coordination bottlenecks through Slack integrations, Pinecone retrieval, and automated workflow routing.

AI Engineer Intern — **May 2025 – Aug 2025**
EVU — *NY, USA*

- Developed LLM-powered document intelligence system using LangGraph and structured output parsing, reducing review time from 4 hours to 1 hour per application.
- Reduced hallucination risk by 20% across RAG pipeline using Ragas evaluation with regression checks and a HITL validation layer on extracted fields.
- Designed LangGraph multi-agent orchestration with agent routing and database-backed state tracking, reducing document-processing errors by 18%.
- Built underwriting analytics dashboard with RAG-powered risk visualizations, reducing decision time by 50%.

Software Engineer, AI/ML — **Jun 2022 – Jun 2024**
Cognizant Technology Solutions — *Chennai, India*

- Accelerated lending decisions from 14 to 10 days across 10M+ financial records using XGBoost models, PySpark feature engineering, and Delta Lake pipelines.
- Processed 6,000 monthly loan applications for a $184B+ U.S. financial institution through automated fraud-risk scoring using Databricks and AWS.
- Reduced manual retraining effort by 50%+ across 30 production models through automated drift detection, MLflow monitoring, and CI/CD workflows.

PROJECTS

The Digital Jurist: Autonomous Legal Research Agent — Demo — **Jan 2026**

- Built autonomous legal-research agents with multi-step reasoning and dynamic retrieval using LangGraph orchestration and confidence-gated pipelines, reducing hallucination rates by 35% across 100+ concurrent sessions.

MindTrack AI: Tuned Voice Agent for Mental Health — Demo — **Apr 2026**

- Deployed a voice-first mental health check-in agent with fine-tuned transformer models and a real-time STT/TTS pipeline, capturing mood data through multi-turn natural conversation.
- Built longitudinal mood-tracking system using time-series analysis and anomaly detection to surface deviation patterns and trigger proactive wellness alerts.

SKILLS

LLM and Agentic AI: LangGraph, LangChain, OpenAI API, Claude API, Prompt Engineering, Multi-Agent Systems, HITL Workflows
RAG and Evaluation: Qdrant, Pinecone, pgvector, Vector Databases, Hybrid Search, Semantic Search, Ragas
Machine Learning and MLOps: PyTorch, XGBoost, MLflow, Databricks, Delta Lake, PySpark
Backend and Data Systems: FastAPI, PostgreSQL, Redis, REST APIs
Cloud and DevOps: AWS, Azure AI, Docker, Kubernetes, GitHub Actions, CI/CD, Prometheus
Programming Languages: Python, JavaScript, TypeScript, SQL