Vinuta Patil

469-558-8950 | vinuta.patil@sjsu.edu | LinkedIn | GitHub

Education

Master of Science in Computer Software Engineering — San Jose, CA
San Jose State University — *August 2024 – May 2026*

Bachelor of Technology in Computer Science and Engineering — Hyderabad, TG
Sreyas Institute of Engineering and Technology — *Aug 2019 – July 2023*

Technical Skills

Languages: Python, JavaScript, Java
Backend Technologies: Flask, FastAPI, Node.js, Express.js, LangChain
Frontend Technologies: React.js, TypeScript, HTML, Tailwind CSS, Streamlit
Databases: MongoDB, ChromaDB, FAISS, SQL, Redis, Amazon RDS, MySQL
Cloud: Amazon S3, AWS Lambda, AWS EC2
DevOps & CI/CD: AWS, Nginx, Docker, Kubernetes, RabbitMQ, Kafka, GitHub Actions

Experience

Software Engineer Intern — May 2025 – August 2025
Mimirchat — *San Jose,CA*

- Built and optimized a Retrieval-Augmented Generation (RAG) pipeline using AWS Lambda, Amazon Textract, and Bedrock LLMs for page-wise document ingestion and semantic search, improving retrieval accuracy by 30%.
- Engineered scalable embedding storage and retrieval workflows with S3 Vectors, OpenSearch, and FAISS, resolving bottlenecks and cutting query latency by 20% compared to Elasticsearch.
- Benchmarked and evaluated multimodal LLMs (Claude, LLaMA, Titan, DeepSeek, Nova) for accuracy, latency (P95/P99), and throughput, building a dashboard to optimize accuracy–cost trade-offs and refining prompts to reduce hallucinations by 30%.

Software Engineer — May 2023 – May 2024
Win Information Technology — *Hyderabad, India*

- Developed and deployed real-time microservices with Flask, Kafka, and RabbitMQ, enabling concurrent query handling and reducing latency by 40%.
- Containerized and deployed systems on AWS using Docker and Kubernetes, improving deployment speed by 35% and scalability across environments.
- Built a multimodal RAG chatbot with LangChain, FAISS, and Cohere, integrating OCR for image queries and achieving 85%+ intent-matching accuracy.

Undergraduate Research Assistant — August 2022 – April 2023
Sreyas Institute of Engineering and Technology — *Hyderabad, India*

- Led design and containerization of a scalable microservices-Based system for real-time data processing using Spring Boot, Kafka, and Docker, handling large data volumes and enabling efficient deployment.
- Managed automated deployment and scaling of containers with Kubernetes, increasing system reliability and resource optimization by 25%, and reducing operational overhead.

Projects

AI Customer Support Agent | *Python , RabbitMQ, Tesseract OCR, LangChain* — January 2025 – May 2025

- Built a multimodal RAG chatbot using Flask, LangChain, FAISS, and Cohere to handle text, speech, and image inputs, added multilingual support and dynamic FAQ uploads, improving query coverage by 40%.
- Integrated RabbitMQ for ticket escalation and designed a real-time query analytics dashboard; achieved 85%+ accuracy in answering user queries during testing.

Adalat Tax-Copilot | *Node.js, Next.js, Tailwind CSS, FAISS* — November 2024 – December 2024

- Built an AI-powered co-pilot that parsed Indian court judgments (PDFs) to extract legal metadata, identify precedent cases, and summarize tax litigation patterns across 25+ real-world documents.
- Generated structured outputs including case summaries and appeal recommendations, aligning with expert decisions in 85% of test cases, improving legal decision consistency by 40%.

Achievements

- **Runner-up** at *LLM x Law Hackathon* held at Stanford Law School – Recognized for innovative legal tech application using large language models.
- Completed **Udemy Certification** in *Introduction to Cloud Computing.*
- Awarded **Elite+ Silver Certificate** for *Internet of Things (IoT)* – NPTEL.
- Earned **Elite Certificate** for *Cloud Computing* – NPTEL.