# Sai Lokesh Reddy Gayam

San Francisco, CA | (510)-255-0650 | sailokeshreddyg@gmail.com | Linkedin | Github

## Summary

Full Stack Software Engineer with 3+ years building scalable applications with React, FastAPI, and Python, specializing in AI and Machine Learning solutions. Expertise in designing and implementing RAG systems, LLM workflows, and Vector Database architectures that reduce manual workload by up to 70% and improve resolution times by 60%. Demonstrated ability to reduce API latency by 30%, accelerate deployment frequency 25% via CI/CD, and cut cloud costs through AWS optimizations.

## Education

**California State University, East Bay** Hayward, CA
*Masters in Computer Science* *Aug 2023 – May 2025*

**CMR Institute of Technology** Hyderabad, India
*Bachelors in Computer Science* *Aug 2018 – May 2022*

## Technical Skills

**Languages:** Python, Java, JavaScript, TypeScript, SQL
**Frontend:** React, Flutter, Redux, Zustand, Material UI, Tailwind CSS
**Backend:** FastAPI, Spring Boot, Firebase, REST, GraphQL
**AI & Machine Learning:** LLM, RAG, VectorDB, OpenAI API, Prompt Engineering, Machine Learning
**Databases:** Vector Databases, PostgreSQL, MongoDB, SQL, NoSQL, FireStore
**Libraries/Frameworks:** React, FastAPI, Node.js, Pytest, Jest
**Cloud Technologies:** AWS (EC2, S3, Lambda, ECS, EKS), Docker, Kubernetes, Jenkins, Firebase
**Practices:** Microservices, Distributed Systems, Agile, Scrum, SDLC, System Design, Design Patterns, API Design, ETL, Automated Testing

## Experience

**Founding Software Engineer** Aug 2025 – Present
*Subscrbe AI* *San Francisco, CA*

- Led the migration from deprecated Firebase **Dynamic Links** to native **Universal Links (iOS)** and **App Links (Android)** using a custom domain, improving deep link reliability by up to **50%**, reducing link failures, and enhancing cross-platform **user engagement and retention**.
- Engineered ultra-short, user-friendly invite links, boosting link shareability and increasing **conversion rates by over 25%**, while seamlessly supporting legacy formats and providing explicit **UI feedback** for status and errors.
- Spearheaded core **architectural decisions** and developed **mission-critical features** from inception, shaping product direction and technical standards as a founding engineer, enabling the platform to **scale efficiently** to thousands of daily active users in a high-growth startup environment.
- **Optimized** user profile synchronization by detecting changes to users' social profile photos and syncing updates in real-time with the database, ensuring **100% profile accuracy** and consistent visual presence.

**Software Engineer 2** May 2024 – July 2025
*WashMetrix* *San Francisco, CA*

- Built a customer onboarding bot using **RAG, vector DB**, and **OpenAI API**; leveraged product docs to **resolve 70%** of customer queries autonomously, cutting onboarding time by 40%.
- Developed and optimized a **React**-based analytics platform with dynamic dashboards & **real-time data visualization**, boosting user engagement by **25%**
- Engineered **scalable backend** services using **Python** & **FastAPI** and integrated them with React frontends, implementing reusable components and **microservices** that accelerated development timelines by **15%**
- Implemented **comprehensive testing** (Jest, Pytest) achieving **90% coverage**, enhancing system reliability and reducing production bug reports by **20%**
- Architected and automated **ETL pipelines** (**Python**, **SQL**) extracting data from POS systems into **cloud databases** (Redshift, PostgreSQL), improving data processing efficiency by **20%**
- Collaborated within **cross-functional teams**, ensuring timely completion of bootcamps and successful project delivery, achieving a **100% on-time** delivery rate across all assigned projects

**Software Engineer 1** Nov 2021 – Aug 2023
*Zemoso Technologies Pvt. Ltd* *Hyderabad, India*

- Engineered **scalable React components** and optimized **GraphQL** schemas with **RESTful APIs**, achieving a **30% reduction** in response time and boosting performance
- Integrated React **state management** (Redux/Zustand) to streamline complex UI workflows, reducing bug occurrences by 15% and improving overall maintainability.
- Integrated **Amazon S3** and **QuickBooks APIs** into a modular **event-driven** architecture, enabling seamless data flow that enhanced product functionality and increased **user satisfaction by 20%**
- Deployed and managed application components on **AWS (EC2, S3)** using **Docker** & **Kubernetes** (EKS), implementing **auto-scaling** to handle 30% traffic spikes and reducing **cloud costs by 15%**

## Projects

**AI-Powered Customer Feedback Intelligence Platform** | *RAG, VectorDB, FastAPI, ML* Oct 2025 - Nov 2025

- Engineered **RAG-powered conversational AI** using **LangChain agents** and **Chroma vector database**, enabling natural language queries about customer feedback patterns with semantic search
- Architected scalable **FastAPI microservices** with JWT authentication and rate limiting, supporting **ML inference** for batch processing and real-time sentiment analysis using **Hugging Face models**
- Built automated data pipelines with RQ workers and Redis queuing, processing customer feedback through ML models and storing enriched data in PostgreSQL with vector embeddings for enhanced analytics
- Developed real-time analytics dashboard with React and Zustand, visualizing sentiment trends and topic clustering results from ML models, reducing manual analysis time by **50%** for customer insights teams

**AI-Powered LaTeX Resume Tailoring Engine** | *RAG, VectorDB, LLM, FastAPI* Sep 2025 - Oct 2025

- Engineered **RAG-powered** resume optimization system using **vector database** technology and advanced **prompt engineering**, reducing resume tailoring time from hours to under **2 minutes** while maintaining ATS-compatible LaTeX formatting
- Implemented multi-model **LLM orchestration** with customizable model selection palette, enabling dynamic adjustment of output length, tone, and style while ensuring keyword optimization for specific job descriptions
- Architected real-time **FastAPI backend** with live LaTeX compilation pipeline, enabling simultaneous code editing and PDF preview while processing NLP transformations through optimized **retrieval-augmented** generation workflows
- Developed intelligent content extraction and semantic analysis modules using **vector embeddings**, allowing context-aware resume customization and automated skill-to-requirement matching.

**AI-Powered Chatbot Solutions Using Gemini API** | *AI, LLM, AWS* Nov 2024 - Dec 2024

- Engineered RAG-powered **AI chatbot** using **Gemini API** and FastAPI with advanced **prompt engineering**, reducing repetitive queries by **40%** and resolution time by **60%** through optimized LLM workflows
- Implemented **vector database** architecture for semantic search capabilities, enabling context-aware responses and improving answer accuracy by **35%** compared to traditional chatbot systems
- Deployed **containerized GenAI** solutions on AWS ECS with OAuth 2.0 security, Redis caching, and CI/CD pipelines, enabling real-time AI support for **500+ users** while reducing manual workload by **70%**
- Integrated comprehensive monitoring and evaluation framework for **LLM performance tracking**, implementing A/B testing of different prompt strategies to optimize response quality and user satisfaction

## Certifications

**Oracle Cloud Infrastructure 2025 Certified Generative AI Professional** | *Oracle* Sep 2025 - Sep 2027

- Credential URL: Oracle GenAI Certification
- Validates advanced expertise in Large Language Models (LLMs), OCI Generative AI Service, RAG, Semantic Search, Vector Databases, and LangChain for building and deploying production LLM applications

## Publications

**Advancing Nursing Education Through Virtual Reality Training** | *VR, Unity* Oct 2024 - Feb 2025

- Publication link: Springer Nature
- Developed a Virtual Reality (VR) training application to help nursing students identify unsafe clinical practices in simulated hospital environments. Published research demonstrating improved engagement and hazard recognition through immersive, risk-free VR learning experiences.