

Sri Ramya Siramsetty

AI/ML Engineer

Brooklyn, NY | +1 (332) 295-7353 | sriramyaasiramsetty05@gmail.com | [GitHub](#) | [LinkedIn](#)

SUMMARY

Gen AI Engineer with 4+ years of experience in software engineering and Generative AI, specializing in Large Language Models (LLMs), Retrieval-Augmented Generation (RAG), Agentic AI, multi-agent workflows, and scalable AI application development. Hands-on experience with Python, FastAPI, LangChain, LangGraph, Model Context Protocol (MCP), vector databases, LLM fine-tuning, Docker, AWS, and CI/CD. Experienced in building secure, production-oriented AI systems for document intelligence, tool-enabled LLM workflows, information retrieval, and enterprise API integration.

SKILLS

Agentic AI & LLM Orchestration: Agentic AI, AI Agents, Multi-Agent Systems, LangGraph, Model Context Protocol (MCP), Tool Calling, Conditional Routing, Agent State/Memory, Workflow Orchestration

Generative AI & Machine Learning: LLMs, RAG, Prompt Engineering, NLP, Transformer Architecture, Embeddings, LLM Evaluation

LLM Fine-Tuning: Supervised Fine-Tuning (SFT), Parameter-Efficient Fine-Tuning (PEFT), LoRA, Instruction Tuning, Domain Adaptation

LangChain Ecosystem: LangChain, LangGraph, LangSmith, LangServe, LangChain Community, LangChain Core

Programming & Backend: Python, JavaScript, FastAPI, Flask, Django, REST APIs, Microservices, JWT, OAuth2, OOP, Data Structures & Algorithms

ML / Deep Learning: PyTorch, TensorFlow, Keras, Hugging Face Transformers, Sentence Transformers, Scikit-learn, Pandas, NumPy, RNN, LSTM, GRU

Data & Vector Databases: PostgreSQL, MongoDB, ChromaDB, FAISS, Vector Databases

Cloud & MLOps: AWS Bedrock, Lambda, API Gateway, S3, EC2, ECR, ECS/Fargate, GCP, Docker, Kubernetes, CI/CD, Model Monitoring

Tools & Methodologies: Git, JIRA, Confluence, VS Code, PyCharm, Postman, Agile, Scrum, Kanban, Sprint Planning

PROFESSIONAL EXPERIENCE

Gen AI Engineer

Morgan Stanley

July 2026 - Present | New York, NY

- Designed and developed Agentic AI workflows using LangChain and LangGraph, implementing planner, retrieval, tool-execution, and response-generation nodes for financial document intelligence and internal knowledge workflows.
- Integrated Model Context Protocol (MCP)-based tools and tool-calling patterns so LLM agents could access enterprise APIs, document retrieval services, and structured data sources through controlled interfaces.
- Engineered RAG pipelines using Python, FastAPI, LangChain, embeddings, and vector databases for structured and unstructured financial content, improving retrieval efficiency by 40%.
- Implemented conditional routing, agent state management, prompt/context optimization, error handling, and response validation to improve reliability and reduce hallucination risk in multi-step LLM workflows.
- Developed secure REST APIs with JWT/OAuth2 authentication and supported production deployment using Docker, AWS/GCP, and CI/CD pipelines, maintaining 99.9% system uptime and reducing production downtime by 25%.
- Implemented logging, monitoring, latency tracking, and debugging for RAG and agentic workflows, reducing response latency by 35% and improving application stability.

Gen AI Engineer

JPMorgan Chase

April 2025 - April 2026 | New York, NY

- Developed RAG-based features using Python, FastAPI, LangChain, and ChromaDB, improving financial document retrieval accuracy by approximately 30% for internal knowledge search.
- Integrated secure RESTful APIs with JWT/OAuth2 authentication to connect LLM-powered services with enterprise applications, ensuring reliable and secure access for internal users.
- Assisted in deploying and monitoring AI applications using Docker, AWS, and CI/CD pipelines, reducing deployment time by approximately 25% and improving release consistency.
- Improved LLM response quality through prompt refinement, embedding optimization, and performance testing, reducing response latency by approximately 30%.

Software Engineer

CodeNest Solutions

May 2021 - July 2024 | Hyderabad, India

- Developed scalable backend services and RESTful APIs using Python, Django, and Flask, reducing API response time by 30% and improving application performance for multiple client-facing web applications.
- Designed and optimized PostgreSQL and MongoDB queries while integrating third-party APIs, improving data processing efficiency by 25% and ensuring reliable communication between application components.
- Built responsive UI components using JavaScript, HTML, and CSS, enhancing user experience and reducing reported frontend issues by 20% across multiple client projects.
- Deployed and maintained applications using Docker, AWS EC2/S3, Git, and CI/CD pipelines, reducing deployment time by 25% while improving release consistency and production stability.

EDUCATION

Master of Science in Computer Science

Rivier University

September 2024 - June 2026 | Nashua, New Hampshire

Bachelor of Technology in Industrial Biotechnology

Bharath Institute of Higher Education and Research

June 2019 - May 2023 | Chennai, Tamil Nadu