# Sairam Bodapothula

+1 (573) 647-4099 | sairambodapothula0990@gmail.com | LinkedIn | GitHub | Portfolio

## Summary

Results-oriented AI/LLM Engineer with 5+ years of experience in developing intelligent systems, machine learning pipelines, and scalable backend architectures. Expertise in fine-tuning large language models, building RAG systems, and deploying ML models in production environments. Proven track record of improving system performance by 25% through optimization of ML inference pipelines and data processing workflows. Skilled in full-stack development with a focus on AI-powered applications, delivering innovative solutions that enhance user experiences and drive business value.

## Technical Skills

**AI/ML & LLM Technologies**: PyTorch, TensorFlow, Hugging Face Transformers, LangChain, LlamaIndex, OpenAI API, Anthropic Claude, RAG (Retrieval-Augmented Generation), Vector Databases (Pinecone, Weaviate, FAISS), Fine-tuning, Prompt Engineering
**Programming Languages**: Python, Java, C++, JavaScript, TypeScript, SQL
**Backend & Frameworks**: Spring Boot, Node.js, Express.js, FastAPI, Flask, RESTful APIs, GraphQL, Microservices
**Frontend Technologies**: React, Angular, HTML5, CSS3, Bootstrap, Material-UI
**ML Infrastructure & MLOps**: Docker, Kubernetes, AWS SageMaker, Azure ML, MLflow, Weights & Biases, Model Versioning
**Database & Data Processing**: PostgreSQL, MongoDB, MySQL, Redis, Apache Kafka, Spark, Pandas, NumPy
**Cloud & DevOps**: AWS (EC2, S3, Lambda, Bedrock), GCP, CI/CD Pipelines, Jenkins, Terraform
**Testing & Monitoring**: JUnit, Pytest, Selenium, Postman, Prometheus, Grafana
**Version Control & Collaboration**: Git, GitHub, Bitbucket, JIRA, Agile/Scrum

## Professional Experience

**AI/LLM Engineer** May 2024 – Present
*Netflix* *Remote, USA*

- Led architectural decision-making for LLM-powered recommendation systems, selecting RAG over end-to-end fine-tuning to balance scalability and cost efficiency, resulting in a 30% improvement in content discovery accuracy and an 18% increase in user engagement
- Drove model optimization strategy by evaluating LoRA vs QLoRA trade-offs, choosing QLoRA for production deployment to reduce inference latency by 35% while preserving 95% model accuracy across Netflix-scale content metadata
- Built scalable ML inference pipelines using FastAPI, Docker, and Kubernetes, handling 10M+ daily requests with 99.9% uptime and sub-200ms response times
- Developed prompt engineering framework for personalized content summarization and generation, enhancing user profile features and reducing content preview generation time by 40%
- Implemented vector search system using FAISS and Pinecone for semantic content matching, enabling real-time similar content recommendations across 200M+ users

**Software Engineer (AI/ML Focus)** May 2019 – Aug 2022
*Cognizant* *Remote, India*

- Developed NLP-powered fraud detection system using BERT and XGBoost for banking client, reducing fraudulent transactions by 40% and saving $2M+ annually
- Built conversational AI chatbot using Rasa and transformer models for customer support, handling 50K+ daily interactions with 85% resolution rate and improving customer satisfaction by 25%
- Implemented automated document processing pipeline using OCR, spaCy, and custom NLP models, reducing manual processing time by 60% and improving accuracy to 94%
- Designed end-to-end ML pipeline for credit risk assessment using ensemble methods (Random Forest, Gradient Boosting), deployed on AWS SageMaker with real-time inference capabilities

## Education

**Missouri University of Science and Technology** Missouri, USA
*Master of Science in Computer Science (Focus: Machine Learning, AI)* *Aug 2022 – May 2024*

**Sathyabama University** Chennai, India
*Bachelor of Engineering in Electronics and Communication Engineering* *May 2017 – May 2021*