Mallikarjuna Reddy Gayam

mallikarjunareddygayam77@gmail.com | 3143339394 | St. Louis, Missouri | LinkedIn | GitHub | Portfolio

SUMMARY

Innovative Machine Learning Engineer with a proven track record in designing and deploying predictive AI solutions, specializing in fine-tuning Large Language Models (LLMs) including GPT and leveraging frameworks like TensorFlow and PyTorch. Spearheaded the development of AI-driven applications, achieving a 30% increase in response reliability through optimized prompt engineering, and successfully integrated machine learning models into enterprise environments via secure REST APIs. Adept at implementing Retrieval-Augmented Generation (RAG) techniques and deploying scalable AI solutions on cloud platforms, while ensuring high standards of performance and ethical compliance.

EDUCATION

Master's in Information Systems | Aug 2023 - May 2025 | GPA: 3.9
Saint Louis University, St. Louis, MO

Bachelor's in Computer Science and Engineering | Jun 2018 - May 2022 | GPA: 3.7
Lakireddy Bali Reddy College of Engineering, Vijayawada, India

TECHNICAL SKILLS

Programming Languages: Python
AI & ML Frameworks: TensorFlow, PyTorch, Hugging Face Transformers
Large Language Models: GPT, Llama, Claude, Mistral, OpenAI API, Anthropic
Prompt Engineering & RAG: Prompt engineering techniques, Retrieval-Augmented Generation, embeddings
Containerization & Orchestration: Docker, Kubernetes
Cloud & Deployment: AWS, Azure, GCP, model deployment strategies
APIs & Integration: RESTful APIs, AI model integration into applications
Model Optimization: Fine-tuning techniques, reinforcement learning, model performance evaluation
Collaboration & Compliance: Cross-functional team collaboration, ethical AI practices

PROFESSIONAL EXPERIENCE

Machine Learning Engineer | Community Dreams Foundation, Remote | Aug 2025 – Present

• Designed and deployed **end-to-end predictive AI models**, leveraging TensorFlow and PyTorch to enhance business insights and decision-making.
• Fine-tuned **Generative AI models** for text generation, optimizing prompt engineering techniques to improve response quality by **25%**.
• Developed and integrated API-driven generative AI solutions into enterprise applications, ensuring model scalability and compliance with ethical AI standards.
• Implemented **Retrieval-Augmented Generation (RAG)** strategies using embeddings and vector databases, resulting in **30% improved accuracy** in information retrieval processes.
• Collaborated cross-functionally to deliver comprehensive AI solutions, significantly enhancing stakeholder satisfaction through data-driven insights.

Software Engineer | Cognizant, Hyderabad, India | Feb 2023 – Aug 2023

• Engineered and deployed **secure RESTful APIs** for AI model integration, optimizing response times and ensuring seamless application connectivity.
• Developed and maintained **scalable data pipelines** for training purposes, reducing data processing time by **40%** through advanced caching techniques.
• Built **TensorFlow-based classification models** for customer support, automating workflows and reducing manual review workloads by **35%**.
• Led CI/CD pipeline automation with GitHub Actions, shortening deployment cycles by **40%** and minimizing production errors significantly.
• Refactored existing database models in **MongoDB/PostgreSQL**, enhancing data retrieval efficiency and cutting mean query times by **50%**.

Software Engineer Intern | Cognizant, Hyderabad, India | Feb 2022 – Aug 2022

• Developed an interactive internal dashboard using **React and FastAPI**, streamlining record tracking and reducing administrative workload by **30%**.
• Integrated **NLP capabilities** using spaCy and Scikit-learn, improving search precision over semi-structured records by leveraging fuzzy matching techniques.
• Created a **secure API layer** with JWT authentication, enforcing stringent access controls for diverse user roles and enhancing data security.
• Deployed services on **AWS Lambda**, effectively scaling applications to meet peak demands and reducing infrastructure costs by **20%**.
• Participated actively in full Agile cycles, contributing to UI planning, API development, and production-grade deployments.

PROJECTS

Acco Finder – AI-Powered Housing Platform

• Designed an **AI-driven recommendation system** using a combination of cosine similarity and user behavior data, improving listing match accuracy by **30%**.
• Implemented real-time chat functionality supported by WebSockets, enhancing user engagement and reducing response times by **35%** during peak usage.

• Deployed the full-stack application on **Vercel** with a serverless architecture, allowing elastic scalability and handling up to **10,000 concurrent users** without degradation.

AI-Powered Resume & Cover Letter Generator

• Built a platform that leverages **OpenAI's NLP capabilities**, achieving a **93% success rate** in generating ATS-optimized resumes through intelligent natural language generation.
• Designed a **containerized environment using Docker**, optimizing deployment processes and ensuring consistent performance across various environments.
• Integrated automated PDF rendering via **Selenium + headless Chrome**, streamlining the resume generation process and reducing manual intervention time by **50%**.

Generative AI Model Fine-Tuning Application

• Developed an application to **fine-tune large language models (LLMs)**, utilizing Hugging Face and TensorFlow, which improved output quality for specific domain use cases by **25%**.
• Implemented **prompt engineering techniques** tailored to diverse user intents, enhancing response relevance and reliability across generated outputs.
• Utilized **retrieval-augmented generation (RAG)** strategies with vector databases, reducing model inference time by **40%** while improving content accuracy for user queries.

CERTIFICATIONS & ACHIEVEMENTS

• **TensorFlow Developer Professional Certificate** – Validated hands-on experience in developing and optimizing deep learning models using TensorFlow.
• **Presented at the International Conference on Sustainable Computing and Data Communication Systems (ICSCDS 2022)** – Delivered a presentation on "Emotion Based Music Player Using Machine Learning Techniques," showcasing application of AI in music technology.