Layakishore Desireddy

+1(732)-532-9087 | Ld786@scarletmail.rutgers.edu | https://github.com/layakishorereddy11 | https://linkedin.com/in/layakishore

Education

Rutgers, the State University of New Jersey - New Brunswick New Jersey, USA
Master of Science in Data Science *August 2023 – May 2025*
Relevant Coursework - Natural Language Processing, Database Management System, Machine Learning, Regression and Time Series, Cloud and ML, Data Mining, Fundamental Algorithms, Devops

IIT Tirupati Tirupati, India
Bachelor of Technology in Electrical Engineering *August 2018 – May 2022*
Relevant Coursework - Reinforcement Learning, Advanced Deep Learning, Computer Vision, Probability and Statistics, Linear Algebra, Advanced Calculus, OOP, Data Structures and Algorithms, Operating Systems

Skills

Languages: C, C++, C#, Java, Python, HTML/CSS, SQL, Typescript, Javascript
Technologies/Frameworks: Tensorflow, Pytorch, Angular JS, React JS, NodeJs, Spring Boot, Jenkins
Databases: PostgreSQL, MySQL, Elasticsearch, MongoDB, Firestore
Technical Skills: Distributed ML Training, AI Infrastructure, API Design, Unit Testing, High-Performance Computing (HPC), CUDA Programming, NCCL, AI Infrastructure

Work Experience

Jobsforce.ai | *Machine Learning Engineer Intern, AI/ML* | *San Francisco , California* March 2025 - Present

- Led a cross-functional team to build and deploy a scalable end-to-end web-based voice-cloning platform with Flask, TTS, React, Docker, MongoDB, using AWS ECR, ECS Fargate, ALB, Auto Scaling, CloudWatch, achieving a 40% reduction in response time and supporting a 3x increase in concurrent user load
- Implemented an automated form-filling and submission platform using Python, Selenium, and Pandas, integrating GPT-4o and Gemini for personalized responses, containerized in Docker, deployed via AWS ECR and ECS Fargate

Optum | *Software Engineer* July 2022 - August 2023

- Architected and implemented end-to-end features within microfrontend/microservices architectures using Angular and Spring Boot (with Spring Data JPA and MySQL), deployed on Google Cloud Platform (GCP). Enhanced security through JWT authentication and role-based access control while delivering dynamic pages and robust RESTful APIs
- Developed and launched a .NET (C#) monitoring tool for real-time error detection and root-cause analysis, automating the reprocessing of failed records, reducing manual interventions by 75% and boosting system reliability and uptime

Ziroh Labs | *Machine Learning Engineer Intern, AI/ML* May 2021 - July 2021

- Engineered unique and sophisticated Deep Neural Network models tailored to training on encrypted data using fully homomorphic encryption, addressing the challenges standard DNN models face with FHE data
- Resolved key challenges to achieve a 99.2% accuracy rate, optimizing performance to within a 0.5% margin of standard deep neural networks trained on unencrypted data

Projects

Natural Language to Query Engine with LLMs | *Python, GPT-4, LangChain, MongoDB, MySQL* February 2025

- Built a scalable, production-grade AI system enabling non-technical users to query MySQL and MongoDB using natural language, leveraging GPT-4 and LangChain for conversational interfaces
- Implemented prompt templates, robust query validation, and fallback logic, improving translation accuracy and overall system reliability by 30%

Personalized Healthcare Advisor with Generative AI | *Python, RAG, LangChain, AWS Bedrock* January 2025

- Architected an AI-driven Health Assistance Application for real-time, personalized medical recommendations, leveraging Gemini 1.5 Pro LLM, LangChain, Retrieval Augmented Generation (RAG), and AWS Bedrock
- Integrated scalable system components including symptom diagnosis, treatment suggestions, and AI-powered preventive care, delivering business impact by enhancing healthcare decision support

NLP-Based Chess Engine for Distributed LLM Training | *PyTorch, Streamlit, Distributed ML* December 2024

- Developed a transformer-based Chess Engine inspired by GPT-3 architecture, trained on over 3 million chess games using distributed machine learning techniques and High-Performance Computing (HPC) resources
- Fine-tuned GPT-3 (124M) with LoRA and Reinforced Fine-Tuning, achieving a 10% win and 69% draw rate against Stockfish, with distributed training reducing training time by 30%, deployed using scalable AWS infrastructure

Conversational AI Chatbot | *Python, RAG, LangChain, GPT-4o, AI Infrastructure* August 2024

- Designed and deployed a scalable AI chatbot for enterprise customer support, leveraging GPT-4o, LangChain RAG orchestration, and a vector database of indexed support documents for knowledge-grounded responses
- Implemented intent classification, dynamic response generation, and retrieval-augmented prompts, cutting query resolution times by 40%, enhancing system reliability and user satisfaction