VIVEK KRISHNA REPALA

+1 470 636 9622 — vrepala@uab.edu — linkedin.com/in/vivek-krishna24 — github.com/vivek-krishna24

SUMMARY

Software Development Engineer with **3+ years of professional experience** building and deploying **large-scale machine learning and Generative AI systems**. Strong hands-on background in **LLM training, fine-tuning, and inference**, distributed systems, and cloud-native architectures. Experienced in **end-to-end LLM pipelines**, model optimization, and production-grade software development. Seeking to contribute to the **AWS Generative AI Innovation Center** by advancing scalable, high-performance LLM solutions.

EDUCATION

Master of Science in Data Science *Dec 2025*
University of Alabama at Birmingham GPA: 3.6

TECHNICAL SKILLS

- **Programming:** Python, SQL, Object-Oriented Design, Data Structures
- **Deep Learning:** Transformers, LLMs, Multimodal Models, RLHF
- **Distributed Training:** Data Parallelism, Model Parallelism, FSDP, DeepSpeed
- **Frameworks:** PyTorch, Hugging Face, TensorFlow
- **Generative AI:** Fine-tuning, Continued Pretraining, RAG, Prompt Engineering
- **MLOps & Systems:** Docker, CI/CD, Model Monitoring, Performance Optimization
- **AWS:** EC2, S3, Lambda, SageMaker, IAM, CloudWatch
- **Software Engineering:** SDLC, Code Reviews, Testing, Version Control (Git)

PROFESSIONAL EXPERIENCE

Machine Learning Intern – ReplyQuickAI (Remote) Oct 2025 – Present
Generative AI — LLM Systems — Cloud Deployment

- Designed and implemented **end-to-end LLM pipelines** including data preprocessing, fine-tuning, inference, and monitoring.
- Fine-tuned transformer-based language models for domain-specific tasks, improving task accuracy and response quality by **35%**.
- Built scalable **LLM inference services** using FastAPI and deployed on AWS infrastructure.
- Optimized model inference latency and throughput through batching, caching, and efficient memory utilization.
- Collaborated with cross-functional teams to translate business use cases into production-ready GenAI systems.

Machine Learning Intern – Honeywell Jan 2025 – Jun 2025
ML Systems — Predictive Analytics

- Developed and trained ML models on large-scale industrial datasets for anomaly detection and forecasting.
- Built robust data pipelines and feature engineering workflows to support scalable model training.
- Assisted in deploying and monitoring ML models in cloud-based production environments.

Data Scientist – Amadeus Dec 2021 – Dec 2023
Production ML — Software Engineering

- Built and maintained **production-grade ML systems** processing **50M+ records** across global datasets.
- Designed scalable data processing pipelines and optimized system performance for reliability and efficiency.
- Followed full SDLC practices including design reviews, testing, deployment, and ongoing maintenance.
- Communicated complex technical solutions to diverse stakeholders, including engineering and business teams.

PROJECTS

Large-Scale LLM Fine-Tuning Pipeline
PyTorch — Hugging Face — Distributed Training

- Implemented distributed LLM fine-tuning using data parallelism techniques for efficient large-batch training.
- Supported domain adaptation through continued pretraining and supervised fine-tuning.
- Evaluated model performance and optimized training stability and convergence.

Multimodal Generative AI System

Vision + Language Models

- Built a multimodal pipeline combining vision encoders and LLMs for image-text understanding tasks.
- Optimized inference workflows for scalable deployment in cloud environments.

Cloud-Optimized LLM Inference Service

AWS EC2 — Docker — FastAPI

- Developed a high-throughput LLM inference service with REST APIs for enterprise applications.
- Implemented logging, monitoring, and performance tuning to meet production SLAs.