Krishnva Vamsi Dhulipalla

(+1) 540-558-3528 | dhulipallakrishnavamsi@gmail.com | GitHub: Krishna-dhulipalla | LinkedIn: krishnavamsidhulipalla
Portfolio | Personal Chatbot

Summary

Software and ML Engineer with 3+ years of experience building scalable AI systems, data pipelines, and production ML platforms. Skilled in Python, SQL, and cloud-native development with deep learning expertise in PyTorch. Experienced in distributed training, model deployment, and MLOps across AWS and GCP environments. Built containerized microservices and ML workflows on Kubernetes to improve system reliability, performance, and developer productivity.

Skills

Programming:	Python, R, SQL, JavaScript, TypeScript, Node.js, MongoDB, FastAPI, Go
ML & AI Frameworks:	PyTorch, TensorFlow, Scikit-learn, JAX, Hugging Face Transformers, OpenCV
Multi-Agent & LLM:	LangChain, LangGraph, AutoGen, MCP, LLM Fine-Tuning, RAG Systems, LlamaIndex, pydantic, OpenAI, Prompt Engineering, LangServe
Data Engineering & Cloud:	Apache Spark, Kafka, Airflow, ETL Pipelines, AWS (S3, Redshift, ECS, SageMaker), GCP (GCS, BigQuery, Dataflow), Snowflake, Vector Databases (FAISS, Pinecone), Redis
MLOps & Infrastructure:	Docker, Kubernetes, Jenkins, MLflow, CI/CD, Weights & Biases, Git, Linux, GitHub Actions, CloudWatch
Other:	Semantic Search, LangSmith, Synthetic Data Generation, LLMOps, Hyperparameter Optimization, A/B Testing, Model Evaluation, Pandas, Data Visualization, Streamlit

Experience

Software Engineer – AI/ML Infrastructure
Jul. 2025 - Present
Cloud Systems LLC
Remote

- Designed and deployed LLM-powered internal automation agents that streamlined data retrieval and reduced manual query handling by 40%, improving team response time across departments
- Rebuilt analytics pipelines and SQL-based ETL workflows, improving runtime efficiency by 25% and reducing daily compute costs
- Deployed containerized agent services on Kubernetes with Docker + Helm, ensuring scalable uptime for internal AI products
- Worked in Agile sprints with cross-functional teams to deliver ML features on schedule, reducing integration delays by 20%

Machine Learning Engineer
Aug. 2024 - Jul. 2025
Virginia Tech
Blacksburg, VA

- Developed genomic sequence classification, fine-tuning LLMs with LoRA and soft prompting, as part of a bioinformatics platform and accelerating research iterations by 30%
- Leveraged HPC clusters and multi-GPU training to run 100+ large-scale biosequence experiments, cutting training cycles from days to hours and enabling faster model evaluation
- Developed a LangChain-based semantic search system for genomics literature, replacing manual reviews with an on-demand AI search interface used by lab researchers
- Deployed fine-tuned LLMs via Docker + MLflow, enabling reproducible experimentation and cutting deployment friction for researchers by 25%

Data Engineer
Jul. 2021 - Dec. 2022
UJR Technologies Pvt Ltd
Hyderabad, India

- Migrated client ETL pipelines into a real-time financial reporting platform using Kafka and Spark, lowering latency by 30% and improving client decision-making
- Deployed Dockerized financial microservices to AWS EKS (Kubernetes), strengthening the company's SaaS product reliability and cutting release rollback time by 40%
- Optimized Snowflake-based analytics platform by redesigning schemas and materialized views, boosting query speed by 40% for customer dashboards
- Built proactive monitoring dashboards in AWS CloudWatch, maintaining 99.9% uptime for client-facing reporting systems
- Partnered with BI and frontend teams to deliver Power BI dashboards integrated with Snowflake, improving client decision visibility and cutting report generation time by 35%

Projects

Multi-Agent AI System: Community & Hazard Intelligence Map

- Built a multi-agent AI system combining official feeds (USGS, NWS, EONET, FIRMS) with community inputs on a live map for real-time disaster intelligence
- Designed LangGraph workflows (reporting, classification, verification) and a React + FastAPI stack with GeoJSON overlays, reducing duplicate entries and improving engagement by 35%
- Resolved concurrency and overlay issues using request guards and optimistic UI, ensuring reliable user experience
- Integrated MCP protocol for agent orchestration, aligning with emerging interoperability standards

Proxy TuNER: Advancing Cross-Domain Named Entity Recognition through Proxy Tuning

- Implemented a proxy-tuning approach for BERT using logit ensembling with domain-specific expert models, improving F1-score by 8% across diverse datasets
- Optimized distributed training to cut computational overhead by 70% and speed up inference by 30%
- Applied gradient reversal for domain-invariant feature learning, boosting cross-domain accuracy by up to 15%

IntelliMeet: AI-Enabled Decentralized Video Conferencing App

- Built a secure video conferencing platform with federated learning and encryption to protect user data
- Integrated on-device Retinaface ML models for real-time attention tracking, reducing dropouts by 25%
- Deployed a Transformer-based CNN-RNN speech-to-text + summarization pipeline for automated meeting notes
- Ensured GDPR-aligned data practices with federated learning and encrypted pipeline
- Maintained 99.9% uptime with CI/CD pipelines and modular microservices architecture

Education

Virginia Tech   Blacksburg, VA
M.S. in Computer Science   *CGPA - 3.9/4*

Vel Tech University   Chennai, India
Bachelor's in computer science   *CGPA - 8.24/10*

Publications

Predicting Circadian Transcription in mRNAs and lncRNAs — IEEE BIBM 2024
Applied ML models to genomic data; improved transcription prediction accuracy, enabling deeper insights into biological rhythms. [DOI: 10.1109/BIBM62325.2024.10822684]

DNA Foundation Models for Cross-Species TF Binding Prediction — NeurIPS ML in CompBio 2025
Leveraged DNABERT-style architectures for plant genomics; advanced cross-species binding prediction with improved generalization. [bioRxiv: 10.1101/2025.07.14.664780v1]

Certifications

- NVIDIA – Building RAG Agents with LLMs (2025)
- NVIDIA – Deploying RAG Pipelines for Production at Scale (2025)
- IBM – Unleashing the Power of AI Agents (2025)
- AWS – Delivering Data-Driven Decisions (2024)
- Snowflake – End-to-End Data Engineering (2024)
- Google Cloud – Data Engineering Foundations (2024)
- AWS – Cloud Foundations (2022)
- Coursera – Machine Learning Specialization (2021)