Krishna 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.

Experience

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

- Cut manual query handling by 40% by deploying LLM-powered internal agents that auto-retrieved data from SQL and internal APIs, improving response time for financial reporting teams
- Reduced daily ETL runtime by 25% by rebuilding SQL and model-evaluation workflows with batched I/O, pruned steps, and incremental loads, keeping agent data up to date
- Ensured service reliability and uptime (>99.9%) by containerizing agent services on Kubernetes with rolling updates and horizontal pod autoscaling
- Shortened model integration delays by 20% by defining API contracts early and adding CI schema validation to speed up feature delivery

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

- Improved genomic sequence classification throughput by 30% by fine-tuning LLMs with LoRA and soft prompting and packaging training/eval as repeatable runs
- Shortened multi-GPU training cycles from days to hours by orchestrating 100+ biosequence experiments with scheduling and checkpoint reuse on HPC clusters
- Led development of an AI search interface with LangChain and semantic retrieval, replacing manual literature review for research teams
- Reduced setup time for researchers by 25% by containerizing fine-tuned models with MLflow for reproducible deployment

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

- Lowered analytics latency by 30% by moving ETL to a Kafka+Spark streaming pipeline for real-time financial reporting
- Reduced release rollbacks by 40% by shipping Dockerized microservices to AWS EKS with health checks and staged rollouts
- Boosted dashboard query speed by 40% by redesigning Snowflake schemas and materialized views for common reporting paths
- Maintained 99.9% uptime by adding CloudWatch metrics and alerts for ingestion lag, error rates, and throughput
- Cut client report generation time by 35% by integrating Power BI with Snowflake models co-designed with the BI and frontend teams

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, OpenAI SDK, Prompt Engineering
Data Engineering & Cloud:	Apache Spark, Kafka, Airflow, 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, Grafana
Other:	LangSmith, Evals, Experiment Tracking, A/B Testing, Model Evaluation, Data Visualization, Streamlit, Semantic Search, PowerBI

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*

Projects

Multi-Agent AI System: Community & Hazard Intelligence Map

- Unified official feeds and user reports on a live map powered by LangGraph agent, lifting engagement 35%
- Built reporting, classification, and verification workflows to surface high-confidence alerts faster
- Fixed concurrency and map-overlay bugs with request guards and optimistic updates, improving update consistency
- 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
- Cut training cost 70% and sped up inference 30% through distributed runs and model-path optimizations
- 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
- Reduced meeting dropouts 25% by running on-device attention detection and handling network hiccups gracefully
- Delivered automatic notes with speech-to-text and summarization, removing manual minutes for every call
- Kept user data on device and encrypted traffic end-to-end; maintained 99.9% uptime with staged releases

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)
- AWS – Delivering Data-Driven Decisions (2024)
- Snowflake – End-to-End Data Engineering (2024)
- Google Cloud – Data Engineering Foundations (2024)
- AWS – Cloud Foundations (2022)