

# Krishna vamsi Dhulipalla

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

## Summary

ML Engineer with 3+ years of experience delivering AI systems that drive measurable impact, from real-time data pipelines to LLM-powered agents. Skilled in Python, SQL, and FastAPI with expertise in semantic search and modular agent workflows (LangChain, LangGraph, MCP). Proven track record of reducing processing latency, improving query performance, and deploying cloud-native solutions (Docker, AWS, Kubernetes) that support both real-time and batch applications.

## Skills

Programming:	Python, R, SQL, JavaScript, TypeScript, Node.js, MongoDB, PostgreSQL, MySQL, FastAPI
ML & AI Frameworks:	PyTorch, TensorFlow, Scikit-learn, Hugging Face Transformers, LangChain, LangGraph, AutoGen, CrewAI, OpenCV
Data Engineering&Cloud:	Apache Spark, Kafka, Airflow, dbt, ETL Pipelines, Data Warehousing, AWS (S3, Glue, Redshift, DynamoDB, ECS, Lambda, SageMaker, CloudWatch, IAM), GCP (GCS, Dataproc, BigQuery, Dataflow, Composer), Snowflake
MLOps & Infrastructure:	Docker, Kubernetes, MLflow, CI/CD, Weights & Biases, Git/GitHub, LangSmith, LangFlow, Linux
Core Expertise:	LLM Fine-Tuning, Retrieval-Augmented Generation (RAG), Semantic Search, Agents, FastMCP, MCP, Self-Supervised Learning, Synthetic Data Generation, Real-time Pipelines, Cross-Domain Adaptation, Hyperparameter Optimization, A/B Testing, Model Evaluation, EDA

## Experience

Cloud Systems LLC	Jul. 2025 - Present
ML Engineer	Remote
<ul style="list-style-type: none"><li>Built and deployed lightweight LLM-powered agents for data retrieval and workflow automation, reducing manual query handling by 40% and enabling early-stage production adoption</li><li>Optimized SQL queries and ETL jobs for structured/unstructured data, improving query runtime by ~25% and cutting daily pipeline costs</li><li>Designed hybrid retrieval pipelines (FAISS + BM25 + cross-encoder reranking) that increased knowledge grounding accuracy by 20%, improving internal decision-making tools</li></ul>	

Virginia Tech	Sep. 2024 - Jul. 2025
ML Research Engineer	Blacksburg, VA
<ul style="list-style-type: none"><li>Achieved 94%+ accuracy in genomic sequence classification by fine-tuning LLMs with LoRA and soft prompting, accelerating research iterations by 30%</li><li>Automated preprocessing of 1M+ sequences using BioPython and Airflow, reducing data preparation time by 40% and freeing researchers to focus on modeling</li><li>Developed LangChain-powered semantic search over genomics literature, cutting information retrieval time from hours to minutes</li><li>Deployed fine-tuned LLMs via Docker + MLflow, enabling reproducible experimentation and cutting deployment friction for researchers by 25%</li></ul>	

Virginia Tech	Jun. 2023 - May. 2024
Research Assistant	Blacksburg, VA
<ul style="list-style-type: none"><li>Improved research data availability by 50% by orchestrating genomic ETL pipelines in Airflow, ensuring cleaner datasets for downstream modeling</li><li>Reduced manual intervention by 40% through automated retraining and evaluation pipelines with CI/CD scripts</li><li>Benchmarked and optimized workflows across institutional compute clusters, cutting runtime and resource consumption by 15%</li></ul>	

UJR Technologies Pvt Ltd	Jul. 2021 - Dec. 2022
Data Engineer	Hyderabad, India
<ul style="list-style-type: none"><li>Migrated batch ETL to real-time streaming using Kafka and Spark, lowering processing latency by 30% and enabling faster reporting for clients</li><li>Deployed microservices with Docker on AWS ECS, shortening release cycles by 25% and improving system reliability</li><li>Enhanced Snowflake query performance by 40% via schema redesign and materialized views, leading to faster client reporting</li><li>Maintained 99.9% system uptime through proactive infrastructure monitoring with AWS CloudWatch, reducing incident escalations</li></ul>	

## Projects

### PulseMap Agent: Community & Hazard Intelligence Map

- Built an agentic system combining official feeds (USGS, NWS, EONET, FIRMS) with community reports on a live map, enabling real-time disaster and safety intelligence
- Designed LangGraph workflows (reporting, classification, verification) and a React + FastAPI stack with GeoJSON overlays, reducing duplicate entries and improving engagement by 35%
- Solved concurrency and overlay issues with request guards and optimistic UI, delivering a smooth, reliable mapping experience
- Enabled faster situational awareness for community safety, disaster response, and local event intelligence by unifying official and user-generated data into a single, reliable map

**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
- Reduced computational overhead by 70% and accelerated inference by 30% through distributed training 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
- Integrated on-device RetinaFace ML models for real-time attention tracking, reducing call dropouts and increasing engagement by 25%
- Deployed a Transformer CNN-RNN speech-to-text pipeline with summarization, enabling automated meeting notes and enhancing productivity
- Maintained 99.9% uptime with CI/CD pipelines and modular microservices architecture

**Education**

**Virginia Tech**

*M.S. in Computer Science CGPA - 3.95/4*

Jan. 2023 - Dec. 2024

*Blacksburg, Virginia*

**Anna University Chennai**

*Bachelor's in computer science CGPA - 8.24/10*

Jul. 2018 - May. 2022

*Chennai, India*

**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**

- ▣ Building RAG Agents with LLMs — NVIDIA (2025)
- ▣ Deploying RAG Pipelines for Production at Scale — NVIDIA (2025)
- ▣ Delivering Data-Driven Decisions with AWS — AWS Training (2024)
- ▣ End-to-End Data Engineering with Snowflake — Snowflake (2024)
- ▣ Google Cloud Data Engineering Foundations — Google Cloud (2024)
- ▣ AWS Cloud Foundations — AICTE-EduSkills (2022)
- ▣ Machine Learning Specialization — Coursera (2021)