Krishna vamsi Dhulipalla

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

Summary

ML Engineer with 3+ years of experience building intelligent systems that combine AI models with scalable backend infrastructure. Skilled in Python, SQL, and FastAPI, with hands-on expertise in LLMs, semantic search, and modular agent workflows (LangChain, LangGraph, MCP). Experienced in designing end-to-end pipelines, retrieval systems, and cloud-native deployments (Docker, AWS, K8s), delivering production-ready solutions for real-time and batch data applications

Skills

Programming: Python, R, SQL, JavaScript, Typescript, Node.js, MongoDB, Fast API

Frameworks: PyTorch, TensorFlow, Scikit-Learn, Hugging Face Transformers, Lang Chain, Lang Graph, AutoGen, Crew AI

AI Systems: LLM Fine-Tuning, Retrieval-Augmented Generation (RAG), Prompt Engineering, Text & Image Generation, GANs, Agents, MCP, AutoGen

ML & Data Science: Self-Supervised Learning, Hyperparameter Optimization, A/B Testing, Synthetic Data Generation, Cross-Domain Adaptation, k-NN, Naive Bayes, SVM, Decision Trees/Random Forest, Clustering, PCA, EDA, Model Evaluation, OpenCV

Cloud & Infra: AWS (S3, Glue, Lambda, Redshift, DynamoDB, ECS, CloudWatch, IAM), GCP (GCS, Dataproc, Big Query, Dataflow, Composer), Snowflake, Docker, SageMaker, MLflow, CI/CD, Kubernetes

Data Engineering: Apache Spark, Kafka, dbt, Airflow, ETL Pipelines, Big Data Workflows, Data Warehousing

Other & Tools: Pandas, NumPy, Matplotlib, Lang Smith, Lang Flow, Weights & Biases, Git, GitHub, Shiny R, Linux

Experience

Cloud Systems LLC Jul. 2024 - Present
ML Engineer *Remote*

- Built and deployed lightweight LLM-powered agents for data retrieval and workflow automation, supporting early-stage production use
- Developed and optimized SQL queries and ETL jobs for structured/unstructured data, improving query performance by ~25%
- Designed hybrid retrieval pipelines (FAISS + BM25 + cross-encoder reranking) for accurate knowledge grounding across user profile

Virginia Tech Sep. 2024 - Jul. 2024
ML Research Engineer *Blacksburg, VA*

- Built AI pipelines using LLMs for sequence classification; applied LoRA and soft prompting to achieved 94%+ accuracy and improve iteration cycles
- Automated prepressing of 1M+ sequences using Biopython and Airflow on institutional cloud infrastructure, streamlining research workflows and reducing runtime by 40%
- Developed Lang Chain pipelines for semantic search over genomics literature, using lightweight local embeddings for retrieval
- Deployed fine-tuned LLMs using Docker and MLflow; optionally integrated AWS SageMaker and CloudWatch for experimentation

Virginia Tech Jun. 2023 - May. 2024
Research Assistant *Blacksburg, VA*

- Orchestrated genomic ETL pipelines via Airflow improving research data availability by 50%
- Automated model retraining and evaluation cycles with custom CI/CD scripts and internal tooling, reducing manual work by 40%
- Benchmark data workflows across institutional compute clusters to optimize runtime and reduce resource usage by 15%

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

- Migrated batch ETL to real-time streaming with Kafka and Spark, reducing processing latency by 30%
- Deployed microservices via Docker on AWS ECS, enabling 25% faster deployment cycles with CI/CD integration
- Optimized Snowflake performance through schema redesign and materialized views, improving query speed by 40%
- Monitored infrastructure using CloudWatch, maintaining 99.9% uptime through proactive alerting and logging

Projects

LLM-Based Android Agent for UI Task Automation

- Developed a custom LLM-powered Android agent that interprets UI layouts and user goals to generate valid actions, achieving 80%+ step accuracy on benchmark tasks
- Engineered modular prompting with few-shot examples, memory buffer summaries, and self-reflection, improving goal success rate by ~25% across multi-step episodes
- Evaluated performance over 10+ simulated mobile scenarios across apps like Settings, Chrome, and WhatsApp, logging over 200 agent actions with reasoning traces
- Diagnosed common failure cases such as UI hallucination and invalid action loops, and implemented recovery strategies like scroll/backtrack heuristics and context-aware retries

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

- Developed a proxy-tuning approach for BERT models using logit-ensembling with domain-specific expert models
- Improved average F1-score by 8% while reducing computational overhead by 70% using small domain-adapted models
- Applied gradient reversal layers for domain-invariant feature learning, boosting cross-domain accuracy by up to 15%
- Accelerated inference speed by 30% through distributed training optimization

IntelliMeet: AI-Enabled Decentralized Video Conferencing App

- Built a secure decentralized video conferencing using federated learning and encryption
- Enabled real-time attention tracking with on-device RetinaFace ML models and reduced call latency below 200ms
- Used a Transformer CNN-RNN-based speech-to-text system and NLP summarization to boost meeting engagement by 25%
- Supported rollout of new features with a stable deployment pipeline, maintaining 99.9% uptime

Education

Virginia Tech Jan. 2023 - Dec. 2024
M.S. in Computer Science CGPA - 3.95/4 *Blacksburg, Virginia*

Anna University Chennai Jul. 2018 - May. 2022
Bachelor's in computer science CGPA - 8.24/10 *Chennai, India*

Publications

L. Miao, K. V. Dhulipalla, S. Kundu et al., *"Leveraging Machine Learning for Predicting Circadian Transcription inmRNAs and lncRNAs,"* in 2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), Lisbon, Portugal, 2024, pp. 6044-6048, doi: 10.1109/BIBM62325.2024.10822684

M. Haghani, K. V. Dhulipalla, S. Li., *"Harnessing DNA Foundation Models for Cross-Species Transcription Factor Binding Sites Prediction in Plant Genomes,"* in 2025 Machine Learning in Computational Biology (co-located with NeurIPS), New York Genome Center, NYC, 2025, doi: 10.1101/2025.07.14.664780v1

Certifications

Building RAG Agents with LLMs by Nvidia

Introduction to Deploying RAG Pipelines for Production at Scale

Delivering Data-Driven Decisions with AWS (Applying Machine Learning, Data Engineering, and Generative AI) **End-to-End Real-World Data Engineering with Snowflake**

Google Cloud Data Engineering Foundations **AICTE-EduSkills Certificate in AWS Cloud**

Coursera Machine Learning (Data-Driven Insights, ML Algorithms)