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 LLM-powered internal agents, multi-agent automation systems, and end-to-end ML workflows. Skilled in Python, model training and evaluation, distributed experimentation, and cloud-native engineering. Experienced in designing RAG pipelines, integrating ML models into production services, improving reliability through observability, and deploying scalable AI systems used by research and business teams.

Experience

Software Engineer – AI Platform Jul. 2025 - Present
Cloud Systems LLC *Remote*

- Reduced manual query-handling by 40% by deploying LLM-powered internal agents that autonomously retrieved SQL and API data, improving turnaround time for financial reporting teams
- Cut daily ETL runtime by 25% by rebuilding SQL and model-evaluation jobs with batched I/O, dependency pruning, and incremental refresh logic
- Maintained 99.9% uptime for agent services by containerizing workloads on Kubernetes with rolling updates, resource tuning, and HPA-based autoscaling
- Improved model integration pace by 20% by defining API schemas early, adding CI-based schema validation, and reducing back-and-forth across teams

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 repeatable training/evaluation pipelines
- Led development of an AI search interface using LangChain and semantic retrieval, enabling automated literature triage and removing manual steps for research groups
- Shortened multi-GPU experiment cycles from days to hours by orchestrating 100+ training runs with scheduling, checkpoint reuse, and efficient job dispatching on HPC clusters
- Reduced research setup time by 25% by containerizing fine-tuned models and tracking versions with MLflow for reproducible deployment

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

- Delivered full-stack features across backend APIs and UI flows, improving team delivery speed and supporting development across 3 core product modules
- Increased stability of ML-driven product features by adding model-serving endpoints and validation logic, reducing prediction-related failures by 30%
- Reduced cross-service integration issues by 40% by aligning data and ML teams on API contracts and standardizing request/response formats
- Improved deployment reliability by implementing automated tests, rollback paths, and environment-based configurations, lowering release failures by 20%

Skills

Programming:	Python, R, SQL, JavaScript, TypeScript, Node.js, MongoDB, FastAPI, Go
ML & AI Frameworks:	PyTorch, TensorFlow, Scikit-learn, Hugging Face Transformers, NLP, Computer Vision, Distributed Training (DDP), Sequence Models, HPC, Checkpointing, Optimization
Multi-Agent & LLM:	LangChain, LangGraph, AutoGen, MCP, LLM Fine-Tuning, RAG, OpenAI SDK, Prompt Engineering
Data Engineering & Cloud:	Apache 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, Grafana
Other:	LangSmith, Evals, Experiment Tracking, A/B Testing, Model Evaluation, Data Visualization, Streamlit, Semantic Search

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

Autonomous Multi-Agent Web UI Automation System

- Built an autonomous web-UI agent system that plans from screenshots, ranks DOM elements, and executes verified actions, reaching over 90% reliability across 10+ UI workflows
- Developed a robust action-executor with fast-retry, timeout bounding, and UI-hash change detection, cutting failure loops by 85% and making automation stable on real SaaS apps
- Integrated two-stage verification using DOM and vision signals, improving final action correctness to above 95% without increasing model cost
- Built detailed LangSmith traces for observability, enabling step-level debugging and cutting development time for new workflows by 35%

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)