

Krishna Vamsi Dhulipalla

+1 540-558-3528 | krishnadhulipalla13@gmail.com | [GitHub](#) | [LinkedIn](#) | [Portfolio](#) | [Personal Chatbot](#)

Summary

AI Engineer with experience building production multi-agent AI platforms, LLM-powered applications, and full-stack services for enterprise workflows. Skilled in Python, LangGraph, FastAPI, React, AWS, RAG, text-to-SQL, MCP tool integrations, and DB connectivity. Experienced in improving workflow automation, reducing LLM inference costs, evaluating retrieval and generation quality using RAGAS, and building production APIs with validation, security controls, and automated regression testing.

Experience

AI Engineer | Tabner Inc. | Remote (US)

Jul 2025 - Present

- Automated 70%+ of recurring analytics requests by developing a production multi-agent AI platform using Python, LangGraph, Claude, and AWS Bedrock to support stakeholder workflows across DB, including text-to-SQL, database operations, document retrieval, and specialized MCP tool integrations
- Reduced LLM inference costs by 30% by optimizing model selection and routing, implementing semantic response caching, and controlling token usage across production AI workflows
- Reduced failed agent executions by 40% by implementing Pydantic schema validation, SQL guardrails, retry policies, and structured tool execution across multi-agent workflows.
- Built and scaled reusable FastAPI services and production REST APIs to integrate agent workflows and other enterprise data sources, standardize structured data access, and enforce PII masking before sending data to external LLM APIs
- Improved stakeholder access to enterprise knowledge by building RAG pipelines over internal documentation and integrating MCP tools for specialized operations, with retrieval and response quality evaluated using RAGAS
- Contributed to the React frontend by integrating multi-agent capabilities with backend APIs and developing stakeholder-facing workflows for database queries, document retrieval, and enterprise support
- Reduced production regressions during prompt and model updates by 30% by establishing offline evaluation datasets, automated validation, and regression testing pipelines for LLM workflows

Machine Learning Engineer | Virginia Tech | Blacksburg, VA

Aug 2024 - Jul 2025

- Increased distributed training throughput 30% on HPC GPU clusters through optimized batching, checkpointing, and GPU utilization monitoring for large scale genomics foundation model training
- Cut per experiment runtime from 12 hours to 8 hours by standardizing fine tuning and evaluation workflows, improving reproducibility across repeated runs
- Established machine learning operations tooling including Docker, MLflow tracking, and SageMaker deployment that reduced deployment failures and improved evaluation cycles

Graduate Research Assistant | Virginia Tech | Blacksburg, VA

Jun 2023 - May 2024

- Fine-tuned LLaMA and SciBERT models for cross-domain named entity recognition, improving entity extraction consistency across heterogeneous scientific datasets
- Reduced manual experiment setup 40% by automating retraining and evaluation workflows on institutional HPC clusters

Software Engineer | UJR Technologies Pvt Ltd | Hyderabad, India

Jul 2021 - Dec 2022

- Built end-to-end ETL pipelines in Python and Azure Data Factory to ingest files and API data from SFTP, Blob Storage, and relational sources into Azure SQL, reducing manual data handling by 60%
- Designed database schemas, SQL transformations, and validation rules for reporting datasets, improving downstream data accuracy by 25% and reducing rework from data issues
- Developed reusable Python utilities for transformation, file parsing, and exception handling, reducing time for new data feeds
- Reduced pipeline issue triage time 40% by adding monitoring and alerts for failed jobs and delayed loads, helping the team meet service level agreements

Software Engineer Intern | Cognizant | Chennai, India

Dec 2019 - May 2020

- Architected a Python-based backend event management system, implementing REST APIs and automated scheduling workflows to streamline tournament operations and data processing

SKILLS

Programming & Backend: Python, SQL, TypeScript, Go, FastAPI, Node.js, REST APIs, GraphQL, NoSQL, MongoDB, ReactJS, R

LLM & Agent Systems: LangGraph, LangChain, Multi-Agent Orchestration, Retrieval-Augmented Generation (RAG), Text-to-SQL, Hybrid Retrieval, Reranking, Vector Databases, Model Context Protocol (MCP), Tool Calling, Prompt Engineering, Guardrails, LLM Evaluation, RAGAS, LangSmith, Claude, OpenAI API, AWS Bedrock, Azure AI Foundry, Gemini, Langflow

AI-Assisted Development Tools: Claude Code, OpenAI Codex, Google Antigravity, Code Generation, Codebase Navigation, Refactoring, Debugging, Automated Test Generation, Agentic Development Workflows

Machine Learning: PyTorch, Hugging Face Transformers, LoRA/PEFT Fine-Tuning, Scikit-learn, XGBoost, Supervised Learning

Cloud & MLOps: AWS (SageMaker, S3, Step Functions, DynamoDB, CloudWatch), Google Cloud Platform (GCP), Azure, Azure Data Factory, Azure SQL, Docker, Kubernetes, CI/CD Pipelines, MLflow, Git, Linux

Data Engineering & Observability: PostgreSQL, Snowflake, Redis, Apache Spark, Airflow, Tableau, Prometheus, Grafana

Projects

CoreLink AI, A2A Finance Agent Runtime

- Built a finance agent runtime on A2A and MCP with deterministic compute tools, risk controls, and policy compliance guards to stabilize tool routed reasoning under constrained tasks
- Developed short-term task memory and persistent strategy journaling into a finance agent runtime, enabling completed task outcomes, validator verdicts, and compute results to guide later task execution
- Reached 87% on the OfficeQA finance benchmark using MCP document retrieval with hybrid search and reranking, validated against an evaluation set scored by LLM as judge with a regression suite to catch drift

K8s Agent Sandbox (Kubernetes Operator and Dev Tooling)

- Built a Kubernetes Operator in Go to provision isolated execution environments for LLM agents with role-based access control, resource quotas, and lifecycle management
- Implemented pre-flight validation, sandboxing, and structured execution monitoring to improve agent reliability to 98%
- Added OpenTelemetry-based tracing for per-agent latency and failure analysis across isolated execution environments

Education

Virginia Tech | Master of Science, Computer Science2024

Vel Tech University | Bachelor of Technology, Computer Science and Engineering

Publications

- Leveraging Machine Learning for Predicting Circadian Transcription in mRNAs and lncRNAs, IEEE BIBM 2024
- Harnessing DNA Foundation Models for Cross-Species TFBS Prediction in Plant Genomes, MLCB in CompBio 2025
- Multi-omics atlas of the plant nuclear envelope, Science Advances 2025, University of California, Berkeley

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 – Cloud Foundations (2024)
- Google Cloud – Data Engineering Foundations (2024)