Viresh Duvvuri

Seattle, WA | +1-509-964-5469 | vireshduvvuri@gmail.com | linkedin.com/in/viresh-duvvuri

ML Engineer with 5+ years building production AI/ML systems and data pipelines across cloud platforms and embedded systems. Currently developing multi-agent AI systems and RAG-based solutions using LangChain, PyTorch, and modern MLOps practices that improve operational efficiency by 50-80%. Recent work includes deploying LLM-powered diagnostic tools serving 200+ daily users, implementing model evaluation pipelines with deepeval and LangSmith, and fine-tuning foundation models for domain-specific applications. Experience spans full-stack ML engineering, distributed systems, DevOps infrastructure (Docker, K8s, AWS), and production model deployment with monitoring and observability.

Skills

AI/ML & Deep Learning: NLP, LLM, RAG, PyTorch, TensorFlow, Keras, LangChain, LangGraph, Multi-Agent Systems, Fine-tuning, LoRA, PEFT, Model Evaluation, deepeval, LangSmith, MLOps, Model Deployment, FAISS, Pinecone, Vector Search, Prompt Engineering, Feature Engineering

Data Engineering & Analytics: Data Pipelines, ETL, Data Processing, Data Integration, SQL, NumPy, Pandas, Scikit-learn, Data Science, Analytics, Knowledge Graph, Enterprise Systems

Programming & Development: Python, C++, JavaScript, TypeScript, FastAPI, Flask, React, API Design, OOP

Cloud & DevOps: AWS, Azure, Docker, Kubernetes, CI/CD, DevOps, Deployment, Monitoring, Performance Tuning, Scalability

Work Experience

Grid CoOperator — Seattle, WA
AI Engineer — *Mar 2025 - Present*

- Developed AI-enabled data processing system using LangChain, Python, and SQL databases from concept to deployment, reducing analyst research workflows by 70% within 2 months through intelligent query generation
- Built scalable backend service with API architecture handling 50-100 daily queries, ensuring reliable performance for real-time smart grid data analysis and operational decision support
- Implemented automated report generation pipeline accelerating stakeholder deliverables by 60% within first quarter, eliminating manual documentation processes for utility operations

Freefly Systems — Woodinville, WA
Senior Software Engineer — *Nov 2021 - Oct 2025*

- Independently designed and built AI-powered diagnostic tool using Python, fine-tuned Llama 3.2-3B on domain-specific drone telemetry using LoRA achieving 18% accuracy improvement, evaluated RAG quality with deepeval, serving production users daily
- Built automated systems to process complex technical data and identify system failures, developing knowledge base enhancements and support tools that streamlined operations
- Contributed to drone platform codebases implementing new features and optimizations for flight control systems and payload integration across multiple product lines, managed software integration projects from planning through release
- Led release management for drone platforms overseeing testing phases from alpha through production deployment, coordinating firmware updates and executing comprehensive testing protocols with cross-functional teams

Lumenier — Sarasota, FL
Drone Software Developer — *Jul 2020 - Oct 2021*

- Wrote embedded code in C++ to integrate LiDAR and optical flow sensors for obstacle avoidance and position holding with/without GPS under various lighting conditions
- Collaborated with open-source flight control software maintainers for integration, testing, and deployment of autonomous flight algorithms, prototyped innovative features like toss-to-launch for product roadmap development

York Exponential — York, PA
Software Engineer - R&D — *Aug 2018 - May 2020*

- Developed prototype software for in-house autonomous surveillance mobile robots using ROS2, SLAM, and computer vision technologies
- Built Human Machine Interface for Universal Robot welding applications using Python and Kivy framework, implemented multi-robot control systems with platform independence

Education

Washington State University — Pullman, WA
Master of Science Computer Science — *Jan 2015 - Jan 2017*

GITAM University — Visakhapatnam, India
Bachelor of Technology Information Technology — *Jan 2011 - Jan 2015*

Projects

GridCOP: Smart Grid Analytics Agent

- Problem: Power grid analysts needed automated database querying and intelligent insights to understand complex data patterns beyond basic visualizations
- Solution: Developed A2A multi-agent system using LangChain orchestration and MCP where specialized agents coordinate tasks, implemented RAG with vector search (FAISS), monitored agent decision quality using LangSmith tracing and deepeval for retrieval assessment achieving 0.85+ context precision, built data pipeline to collect and annotate 500+ user queries for continuous model refinement, deployed on AWS with observability
- Impact: Enhanced analyst productivity by 70% through AI co-pilot that augments domain experts with automated workflows, implemented human-in-the-loop (HIL) evaluation and testing pipelines for production-ready AI systems with robust error handling through rapid iteration

Production System Optimization Tool

- Problem: Manual system analysis taking hours of expert time, creating bottlenecks in product development and customer support resolution
- Solution: Built full-stack application with React frontend, Python Flask backend, integrated foundation model APIs (Ollama and Llama 3.2) for real-time log processing and interactive analysis using prompt engineering and model evaluation
- Impact: Transformed expert analysis from hours to minutes, deployed to production serving 200+ daily queries with significant performance improvements through rapid iteration and continuous optimization

AI Travel Planner Agent

- Problem: Manual travel planning requiring hours of research across multiple sources with inconsistent and outdated information
- Solution: Built AI agent using Claude 3.5 Sonnet, LangChain, Streamlit, and DuckDuckGo Search API for personalized itinerary generation using prompt engineering techniques
- Impact: Demonstrated end-to-end AI application development, learned conversational AI patterns and real-time data integration techniques through iterative development