Viresh Duvvuri

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

Senior Software/AI Engineer with 7+ years building agentic automation systems, agent reliability platforms, and full-stack AI applications. Expert in designing autonomous agents using modern AI frameworks (LangChain, LangGraph, MCP) integrated with production systems serving 200+ daily users. Proven track record deploying GenAI products with comprehensive evaluation pipelines, addressing LLM challenges like hallucination and consistency through RAG and vector search, and leading technical strategy for scalable, reliable AI-driven platforms on AWS with end-to-end ownership from backend microservices to polished React UIs.

Skills

Programming: Python, TypeScript, JavaScript, C++, SQL, FastAPI, Flask, React, Node.js, NumPy, Pandas, OOP

AI/ML Frameworks: Agentic AI, LangChain, LangGraph, Multi-Agent Systems, MCP (Model Context Protocol), RAG, Context Engineering, Prompt Engineering, Model Evaluation, MLOps, LLMOps, GenAI, FAISS, Pinecone, PyTorch, TensorFlow, Scikit-learn, Feature Engineering, Human-in-the-Loop (HIL), Model Deployment, Responsible AI, Vector Search

Cloud & Infrastructure: AWS, Azure, API Design, Deployment, DevOps, Docker, Kubernetes, Monitoring, Performance Tuning, Scalability, Workflows, Microservices, CI/CD, Event-Driven Architecture

Data & Analytics: Data Integration, Data Processing, Data Science, Enterprise Integrations, Enterprise Systems, Knowledge Graph, Operational Efficiency

Work Experience

Grid CoOperator — Seattle, WA

AI Engineer — *Mar 2025 - Present*

- Built agentic automation platform with autonomous agents that handle complex analytics tasks independently, architecting multi-agent system using LangChain and MCP frameworks where specialized agents coordinate via APIs and event-driven architecture, deployed on AWS with comprehensive evaluation pipelines and observability dashboards ensuring agent reliability and correct results, reducing analyst workflows by 70% within 2 months through continuous monitoring and quality metrics tracking
- Designed and implemented agent reliability infrastructure with evaluation frameworks measuring quality, latency, and correctness across 50-100 daily queries, achieving 99%+ uptime through robust error handling, automated recovery mechanisms, and human-in-the-loop feedback integration to continuously improve agent performance and address LLM consistency challenges
- Led technical architecture for production AI system with CI/CD pipelines, Docker containerization, and performance optimization, collaborating cross-functionally with business stakeholders to define technical roadmap and translate requirements into scalable solutions, accelerating deliverables by 60% through agile practices and implementing best practices for LLMOps including prompt versioning, model evaluation, and observability

Freefly Systems — Woodinville, WA

Senior Software Engineer — *Nov 2021 - Oct 2025*

- Built end-to-end full-stack AI diagnostic system serving 200+ daily users, architecting React frontend with TypeScript and Python Flask REST APIs, integrated foundation model APIs (Ollama, Llama 3.2) with RAG architecture and vector search (FAISS) to address LLM hallucination and consistency challenges, implemented evaluation framework with automated testing and quality metrics, containerized with Docker and deployed to production infrastructure reducing expert analysis time by 80% through reliable AI-driven insights
- 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 through prompt engineering strategies, implemented RAG and vector search (FAISS) for intelligent querying, implemented model evaluation frameworks to monitor quality and cost metrics, deployed on AWS with observability and logging
- 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