Viresh Duvvuri

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

AI Native Engineer with 7+ years building cloud-native systems and 2+ years deploying agentic solutions in production environments. Expert in designing and engineering enterprise-ready AI agents using modern frameworks (LangChain, MCP) with retrieval (RAG), orchestration, evaluation harnesses, and lifecycle observability. Proven track record collaborating with clients and stakeholders to translate business requirements into operational agentic workflows serving 200+ daily users across energy and aerospace domains, deployed on AWS with Docker containerization, CI/CD pipelines, and comprehensive monitoring for scalable, production-ready AI-native systems.

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

AI Platforms: Claude (Anthropic), OpenAI, Ollama, Llama, Foundation Models, Model APIs

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

Work Experience

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

- Partnered with energy sector stakeholders through workshops and POCs to build multi-agent analytics platform, implementing intelligent orchestration and retrieval that reduced analyst workflows by 70% in 2 months
- Engineered agent architecture with comprehensive evaluation harnesses measuring quality, latency, safety, and cost effectiveness across 50-100 daily queries, implementing lifecycle observability dashboards, automated testing pipelines, and human-in-the-loop feedback mechanisms achieving 99%+ uptime through robust error handling and continuous improvement of agent accuracy and reliability in production environment
- Deployed production AI-native system on AWS with cloud-native engineering practices including Docker containerization, microservices architecture, event-driven pipelines, CI/CD automation, and infrastructure monitoring, collaborating cross-functionally to define technical roadmap and accelerating deliverables by 60% through rapid prototyping and iterative development with stakeholder feedback integration

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

- Designed and deployed enterprise-ready full-stack AI diagnostic system serving 200+ daily users in aerospace domain, architecting React frontend with TypeScript and Python Flask REST APIs, integrating foundation model APIs (Ollama, Llama 3.2) with RAG architecture and vector search (FAISS) for context engineering, implementing evaluation framework with automated testing and quality metrics, containerized with Docker and deployed to production infrastructure with CI/CD pipelines reducing expert analysis time by 80% through operational AI-driven workflows
- 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