Viresh Duvvuri

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

Senior AI Engineer with 5+ years building AI platforms, infrastructure, and tooling that empower data scientists, engineers, and product teams to deliver AI-first capabilities at scale. Proven expertise designing and scaling internal AI tooling using modern frameworks (MCP, LangGraph, PydanticAI), integrating LLM gateways, and building observability systems for production ML workflows. Strong track record developing scalable cloud infrastructure (AWS) for large-scale data processing and model serving, implementing end-to-end data science workflows from ingestion to monitoring, and collaborating cross-functionally to accelerate AI innovation across organizations. Deep Python expertise with AI-first mindset and passion for building platforms that enable teams to do their best work.

Skills

AI Platform & Infrastructure: MCP Servers/Clients, LangGraph, LangChain, PydanticAI, LiteLLM, Multi-Agent Frameworks, LLM Gateways, Structured Output Systems, Agent Orchestration, RAG Pipelines, Function Calling, AI Tooling Development

ML Observability & Monitoring: MLflow Patterns, Model Evaluation Frameworks, Performance Monitoring, Accuracy Tracking, Logging & Traceability, Observability Dashboards, Golden Set Testing, Synthetic Data Generation, Human-in-the-Loop Evaluation

Programming & Development: Python, FastAPI, Flask, TypeScript, JavaScript, SQL, C++, React, NumPy, Pandas, System Architecture, Internal Tooling, API Design, Library Development, Software Engineering Best Practices

Cloud & Infrastructure: AWS, Azure, Docker, Kubernetes, CI/CD Pipelines, Scalable Infrastructure, REST APIs, Containerized Development, Production Deployment, Real-Time Data Pipelines, Model Serving, Performance Optimization

Data Science Workflow: Data Ingestion, Feature Engineering, Data Processing, Experimentation, Model Deployment, Monitoring, ETL Pipelines, Vector Databases (FAISS, Pinecone), SQL, PostgreSQL, Real-Time Data Integration

Work Experience

Grid CoOperator — Seattle, WA

AI Engineer — *03/2025 - Present*

- Built and scaled AI platform infrastructure empowering analysts and stakeholders with AI-powered tools, architected multi-agent system using LangGraph orchestration with Model Context Protocol (MCP) integration, developed internal tooling and libraries enabling teams to leverage AI for complex data analysis workflows, reduced manual effort by 70% within 2 months through scalable AI platform serving 50-100 daily queries with 99%+ uptime
- Designed and implemented AI infrastructure components including LLM gateway integrations (GPT-4, Claude APIs), structured output systems with validation (PydanticAI patterns), RAG pipelines with vector databases (FAISS) for semantic retrieval, and real-time data processing infrastructure connecting SQL databases and web services, built comprehensive observability and monitoring systems tracking AI performance, accuracy, and traceability across production workloads with dashboards and logging
- Collaborated with data scientists, engineers, and product teams to build AI capabilities supporting real-world use cases, developed scalable cloud infrastructure on AWS for large-scale data processing and model serving using FastAPI, Docker, and monitoring tools, contributed to AI-first culture by researching emerging technologies (MCP, function calling, agent frameworks) and building internal platforms that accelerated AI innovation across the organization

Freefly Systems — Woodinville, WA

Senior Software Engineer — *11/2021 - 10/2025*

- Built AI-powered diagnostic platform infrastructure integrating foundation model APIs (Ollama, Llama 3.2) with internal tooling for engineering teams, implemented observability systems for ML workflows including performance monitoring and evaluation frameworks for accuracy and reliability, deployed scalable infrastructure serving 200+ daily queries with containerized deployment (Docker) and comprehensive monitoring, enabled data scientists and engineers with AI capabilities for autonomous technical analysis
- Contributed to enterprise-scale platform codebases implementing features and system optimizations, managed software integration projects from planning through production release in high-velocity environment, developed automated systems for complex data processing
- Led release management for mission-critical platforms coordinating testing phases from alpha through production deployment, troubleshooting integration issues, executing comprehensive testing protocols with cross-functional teams
- Built automated systems to process complex technical data and identify system failures, developing support tools that streamlined operations in data-intensive environment

Lumenier — Sarasota, FL

Drone Software Developer — *07/2020 - 10/2021*

- 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 for product roadmap development

York Exponential York, PA
Software Engineer - R&D *08/2018 - 05/2020*

- 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 *01/2015 - 01/2017*

GITAM University Visakhapatnam, India
Bachelor of Technology Information Technology *01/2011 - 01/2015*

Projects

AI Platform Infrastructure with MCP & LangGraph | Grid CoOperator

- Problem: Organization needed scalable AI platform infrastructure to empower data scientists, analysts, and engineers with AI-first capabilities, requiring internal tooling and libraries that integrate modern AI frameworks, LLM gateways, and observability systems to accelerate innovation and enable teams to build AI-powered applications efficiently
- Solution: Designed and built AI platform using LangGraph for multi-agent orchestration, integrated Model Context Protocol (MCP) servers/clients for tool connectivity, developed structured output systems using PydanticAI patterns for validation, implemented LLM gateway integrations (GPT-4, Claude) with LiteLLM patterns, built RAG pipeline infrastructure with FAISS vector databases, created real-time data processing infrastructure using FastAPI and SQL databases, deployed on AWS with Docker containerization, established comprehensive observability systems with MLflow-style monitoring, performance tracking, accuracy evaluation, and traceability logging
- Impact: Successfully scaled AI platform serving 50-100 daily queries with 99%+ uptime, empowered teams with 70% efficiency improvement through internal AI tooling, built observability framework validating 40% improvement in AI accuracy through evaluation systems, demonstrated AI-first mindset by researching and integrating emerging technologies (MCP, function calling) that accelerated AI innovation across organization, enabled data scientists and engineers to leverage scalable infrastructure for AI experimentation and deployment

ML Infrastructure & Foundation Model Integration | Freefly Systems

- Problem: Engineering teams needed AI platform infrastructure supporting autonomous diagnostic capabilities, requiring foundation model integration, observability for ML workflows, and scalable infrastructure enabling data scientists to deploy AI-powered tools for real-world technical analysis use cases
- Solution: Built AI platform infrastructure integrating foundation model APIs (Ollama, Llama 3.2) with internal tooling, implemented RAG pipeline with semantic search and vector retrieval, developed observability systems for ML workflows including performance monitoring and evaluation frameworks, created full-stack application (React frontend, Python Flask backend) with REST APIs, deployed containerized infrastructure (Docker) with monitoring and logging
- Impact: Delivered production AI platform serving 200+ daily queries, reduced analysis time by 80% through ML-powered infrastructure, implemented observability systems enabling continuous performance optimization, empowered engineering teams with self-service AI capabilities through scalable platform infrastructure

AI Agent Framework Prototyping | Personal

- Problem: Needed rapid experimentation with emerging AI frameworks and agent orchestration patterns to validate technical approaches for building intelligent AI systems, requiring hands-on exploration of LangChain, function calling, API integration, and conversational AI tooling
- Solution: Built AI agent using Claude API with LangChain orchestration framework, implemented function calling for tool integration with DuckDuckGo Search API, developed structured output patterns, applied prompt engineering techniques, created Streamlit-based interface for experimentation
- Impact: Demonstrated AI-first mindset through hands-on experimentation with emerging technologies, validated agent framework patterns and LLM integration approaches, showcased curiosity and passion for learning new AI tools and platforms, illustrated ability to rapidly prototype AI tooling and leverage AI in all aspects of work