Viresh Duvvuri

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

Senior AI Engineer

Senior AI Engineer with 7+ years of experience architecting intelligent solutions and building production-grade systems. Background in engineering robotics systems and embedded control, now focused on designing scalable multi-agent ecosystems and optimizing AI infrastructure. Proven track record of leading cross-functional teams to deliver enterprise solutions that solve complex technical problems.

Skills

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

AI/ML Frameworks: System Design, Solution Architecture, Technical Strategy, Stakeholder Management, Compliance, Regulatory Requirements, Agentic AI, LangChain, LangGraph, Multi-Agent Systems, MCP (Model Context Protocol), RAG, Context Engineering, Prompt Engineering, Model Evaluation, MLOps, 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

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*

- Architected and deployed domain-specific agentic AI agents for smart grid analytics, collaborating cross-functionally with business stakeholders to translate operational requirements into multi-agent systems using LangChain orchestration and prompt engineering strategies that reduced analyst workflows by 70% within 2 months through rapid iteration
- Designed AI orchestration system where specialized agents communicate and coordinate for complex analytics tasks, deployed on AWS with observability and cost monitoring, established model evaluation pipelines tracking quality metrics, latency, and performance to achieve reliable enterprise performance within 6 weeks across 50-100 daily queries
- Defined technical strategy for production AI system on cloud infrastructure with CI/CD pipelines, monitoring, and performance optimization, accelerating deliverables by 60% within first quarter through rapid experimentation, iterative prompt engineering, and continuous improvement

Freefly Systems — Woodinville, WA

Senior Software Engineer — *Nov 2021 - Oct 2025*

- Built AI co-pilot for automated log analysis using React, Python Flask, and foundation model APIs (Ollama, Llama 3.2), deployed to production on cloud infrastructure with model evaluation metrics and monitoring, reducing manual workflows by 80% within 3 months through rapid iteration and prompt engineering
- Coordinated cross-functional projects translating business requirements into technical solutions, implementing software design principles and testing frameworks across engineering divisions, supporting product roadmap development and stakeholder demos
- Enhanced flight control systems with microservices architecture and CI/CD pipelines, improving deployment efficiency by 60% over 6 months
- 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

Software Engineer - Embedded Systems — *Jul 2020 - Oct 2021*

- Implemented custom software using C++ and data structures for specialized applications, enabling autonomous capabilities within 8 weeks, prototyping innovative features for product roadmap development
- Enhanced system performance through algorithms and data ingestion pipelines, improving operational efficiency by 45% across environments
- Architected testing frameworks with software design principles, reducing implementation issues by 30% within 3 months

York Exponential — York, PA

Software Engineer - R&D — *Aug 2018 - May 2020*

- Created Human Machine Interface for collaborative welding using Python, Kivy, and ROS2, reducing operator programming complexity by 50% within 4 months
- Developed autonomous robot prototype using computer vision and machine learning from requirements to working deployment

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

Enterprise AI Architecture Optimization

- 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

Autonomous Obstacle Avoidance System

- Problem: Drone platforms required robust safety mechanisms to detect and avoid obstacles in real-time during autonomous flight missions
- Solution: Trained and deployed TensorFlow/Keras object detection models on embedded edge devices, tuning control algorithms based on computer vision inputs to execute evasive maneuvers with low latency
- Impact: Delivered a critical safety feature for autonomous flight systems, enabling safe operation in complex environments and reducing collision risks by significant margins