Viresh Duvvuri

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

AI Product Engineer

AI Product Engineer with 5+ years building full stack LLM applications serving real users at scale. Proficient in React, TypeScript, and Python for production LLM-powered products with real-world experience deploying RAG pipelines, prompt engineering, and evaluation frameworks. Product-first mindset delivering 200+ daily active users through rapid iteration, user feedback loops, and high velocity independent execution from concept to production.

Skills

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

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

Cloud & Infrastructure: AWS, Production Deployment, CI/CD Pipelines, Monitoring, Performance Tuning, Scalability, API Design, DevOps, Docker, Kubernetes

Product & Collaboration: Product-Focused Engineering, Cross-functional Collaboration, Rapid Prototyping, User-Facing Features, High Velocity Development, Agile Development, Technical Communication, Independent Execution

Work Experience

Grid CoOperator — Seattle, WA

AI Engineer — *Mar 2025 - Present*

- Led design and deployment of 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
- Architected 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
- Deployed production AI system to 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 and deployed GenAI-powered agent for automated log analysis from concept to production, integrating foundation model APIs (Ollama, Llama 3.2) with evaluation frameworks and model governance practices, serving 200+ daily queries
- 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
- Built automated systems to process complex technical data and identify system failures, developing knowledge base enhancements and support tools that streamlined operations

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

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

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

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