Viresh Duvvuri

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

AI Engineer with 5+ years building production agentic systems and RAG pipelines using LangGraph, LangChain, and foundational LLM APIs (Anthropic, OpenAI). Proven track record shipping intelligent document understanding and reasoning systems from prototyping through production deployment, with evaluation frameworks and observability—delivering 70-80% efficiency gains through rapid iteration. Strong software engineering fundamentals combined with end-to-end ownership of ML systems, cross-functional collaboration, and continuous improvement of AI workflows.

Skills

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

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

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

Product & Collaboration: Cross-functional Leadership, Rapid Prototyping, Consumer AI Products, Architectural Design, Mentorship, Agile Development, Technical Communication

Work Experience

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

- Architected production multi-agent system using LangGraph and LangChain for smart grid analytics, leading cross-functional teams to translate business requirements into specialized agents that coordinate complex tasks through AI orchestration, reducing analyst workflows by 70% within 2 months through rapid prototyping and iteration
- Built evaluation frameworks and observability systems tracking model quality metrics, latency, cost monitoring, and performance across production AI deployment on AWS, achieving reliable enterprise performance within 6 weeks serving 50-100 daily queries with robust error handling and human-in-the-loop validation
- Deployed production-quality code with CI/CD pipelines, monitoring, and performance optimization, accelerating deliverables by 60% within first quarter through rapid experimentation, iterative prompt engineering, and continuous improvement of agentic AI capabilities

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

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: Architected production multi-agent system using LangGraph and LangChain orchestration where specialized agents coordinate complex analytics tasks, implemented RAG and vector search (FAISS) for intelligent querying, built model evaluation frameworks tracking quality metrics and cost, deployed on AWS with observability and monitoring
- Impact: Enhanced analyst productivity by 70% through AI co-pilot that augments domain experts, implemented human-in-the-loop (HIL) evaluation and testing pipelines for production-ready AI systems with robust error handling, achieved enterprise performance within 6 weeks through rapid prototyping and 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 production-quality full-stack AI application with TypeScript/React frontend, Python Flask backend, integrated LLM APIs (Ollama, Llama 3.2) for real-time log processing and interactive analysis using prompt engineering, model evaluation, and observability
- Impact: Transformed expert analysis from hours to minutes, deployed to production serving 200+ daily queries with continuous monitoring and optimization, demonstrated rapid prototyping and iterative development for consumer-facing AI experiences

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