Viresh Duvvuri

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

AI Engineer specializing in prompt engineering and LLM evaluation for production AI systems, with 5+ years building customer-facing AI features from rapid prototyping to deployment. Builder-first mindset with proven track record designing prompt engineering strategies and evaluation frameworks that improved efficiency by 70-80% serving hundreds of daily users. Deep experience collaborating cross-functionally with PMs, designers, and engineers to ship production AI systems on AWS/Azure with focus on quality, latency, and cost optimization.

Skills

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

LLM & Prompt Engineering: Prompt Engineering, Context Engineering, LLM Evaluation, Frontier Models (Claude, GPT-4, Ollama, Llama), Model Fine-tuning, Human-in-the-Loop (HIL)

AI Agent Development: LangChain, LangGraph, Multi-Agent Systems, RAG, Agent Orchestration, MCP (Model Context Protocol), Agentic AI

AI/ML: NLP, MLOps, PyTorch, TensorFlow, Model Deployment, Responsible AI, Feature Engineering

Cloud & Infrastructure: AWS, Azure (learning), Docker, CI/CD, API Design, Monitoring, Performance Tuning, Scalability, Observability

Work Experience

Grid CoOperator Seattle, WA
AI Engineer *03/2025 - Present*

- Designed and implemented prompt engineering strategies for multi-agent system, developed evaluation frameworks measuring response quality, correctness, and cost-efficiency, shipped production system automating analyst workflows with 70% efficiency gain within 2 months through rapid iteration and cross-functional collaboration with business stakeholders
- Built AI orchestration using LangChain where specialized agents coordinate through natural language prompting, deployed on AWS with monitoring tracking latency, token usage, and quality metrics across 50-100 daily queries, established prompt engineering best practices for production reliability and cost optimization
- Collaborated cross-functionally with PMs and domain experts to translate requirements into AI solutions, implemented human-in-the-loop evaluation with subject matter experts validating outputs, established model governance practices including bias detection and safety guardrails for production deployment

Freefly Systems Woodinville, WA
Senior Software Engineer *11/2021 - 10/2025*

- Built production LLM-powered diagnostic tool serving 200+ daily users, implemented prompt engineering and evaluation frameworks for accuracy and reliability, collaborated with engineering teams to deliver 80% productivity improvement through intelligent automation, deployed on cloud infrastructure with monitoring and performance optimization
- 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 *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 like toss-to-launch 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

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 where specialized agents coordinate through prompt engineering strategies, implemented RAG and vector search (FAISS) for intelligent querying, designed evaluation frameworks tracking quality, cost, and latency 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 customer-facing full-stack application with React frontend, Python Flask backend, integrated frontier model APIs (Ollama and Llama 3.2) for real-time log processing using prompt engineering and evaluation techniques, deployed to production serving 200+ daily users
- 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