Viresh Duvvuri

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

AI Engineer

AI Engineer with 5+ years building production AI systems serving real customers at scale. Deep expertise in LLM inference, similarity search, model evaluation frameworks, and guardrails for responsible AI deployment. Proven optimization mindset delivering 70-80% efficiency gains through performance tuning, cost optimization, and scalability design—combined with cross-functional leadership experience translating business needs into robust AI solutions on AWS.

Skills

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

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

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

Product & Collaboration: Cross-functional Leadership, Rapid Prototyping, Architectural Design, End-to-End Ownership, Responsible AI, Agile Development, Technical Communication, Mentorship

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 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
- Enhanced flight control systems with microservices architecture and CI/CD pipelines, improving deployment efficiency by 60% over 6 months

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
- 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

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

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