Viresh Duvvuri

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

AI Engineer

AI Engineer specializing in LLM-powered systems and RAG pipelines, with 5+ years building production GenAI solutions from rapid prototyping to deployment. Builder-first mindset with proven track record shipping AI agent features that improved operational efficiency by 70-80% within 3 months, establishing model evaluation frameworks and prompt engineering strategies for scalable AI products on AWS. Strong Python backend developer with growing TypeScript expertise and product-focused approach to continuous improvement.

Skills

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

AI & LLM Development: LLM Integration (GPT-4, Claude, Llama), Prompt Engineering, RAG Pipelines, Vector Embeddings, FAISS, LangChain, LangGraph, Multi-Agent Systems, Model Context Protocol (MCP), Model Evaluation, Context Engineering, GenAI, Human-in-the-Loop

AI/ML Frameworks: PyTorch, TensorFlow, Scikit-learn, MLOps, Model Deployment, Feature Engineering, Responsible AI

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

Data & Databases: Vector Databases (FAISS, Pinecone), SQL, PostgreSQL, Data Pipelines, Embedding Frameworks, Semantic Search

Backend Development: FastAPI, Flask, REST APIs, Microservices, WebSocket, TypeScript Backend Development

Work Experience

Grid CoOperator — Seattle, WA

AI Engineer — *03/2025 - Present*

- Built and shipped production AI agent system from prototype to deployment, designing multi-agent architecture with LLM orchestration (GPT-4, Claude) and prompt engineering strategies that automated complex analyst workflows, reducing manual effort by 70% within 2 months through rapid iteration and continuous improvement
- Designed and deployed RAG pipeline infrastructure using LangChain on AWS with vector embeddings (FAISS) for intelligent context retrieval, established comprehensive model evaluation framework measuring agent performance, cost efficiency, and quality metrics across 50-100 daily queries with observability and monitoring dashboards
- Shipped iterative product improvements optimizing LLM prompt strategies and RAG performance based on user feedback and evaluation metrics, built monitoring systems tracking business outcomes and operational efficiency, established AI governance practices including safety guardrails and bias detection for production-ready systems

Freefly Systems — Woodinville, WA

Senior Software Engineer — *11/2023 - 10/2025*

- Built and deployed GenAI-powered diagnostic agent for automated log analysis from concept to production, integrating LLM APIs (Ollama, Llama 3.2) with RAG pipeline using vector search and semantic embeddings, implemented model evaluation frameworks and prompt engineering strategies, serving 200+ daily queries with continuous performance optimization based on user feedback
- 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 | Grid CoOperator

- Problem: Power grid analysts needed automated database querying and intelligent insights to understand complex data patterns beyond basic visualizations, requiring LLM-powered system for natural language interaction with enterprise data
- Solution: Developed multi-agent AI system using LangChain orchestration with specialized agents for SQL generation and context retrieval, implemented RAG pipeline with vector embeddings (FAISS) for intelligent querying, integrated LLM APIs (GPT-4, Claude) with prompt engineering strategies for accurate responses, established model evaluation framework tracking quality metrics and cost efficiency, deployed on AWS with FastAPI backend, observability, and logging infrastructure
- Impact: Enhanced analyst productivity by 70% through AI co-pilot that augments domain experts with automated workflows, implemented human-in-the-loop evaluation and testing pipelines for production-ready AI systems with robust error handling, serving 50-100 daily queries with 99%+ uptime through continuous iteration and performance optimization based on user feedback

Production System Optimization Tool | Freefly Systems

- Problem: Manual drone system log analysis taking hours of expert time per case, creating bottlenecks in product development and customer support, requiring AI-powered automation for diagnostic workflows
- Solution: Built full-stack application with React frontend and Python Flask backend, integrated LLM APIs (Ollama, Llama 3.2) for real-time log processing and interactive analysis, implemented RAG pipeline with semantic search for contextual diagnostics, developed prompt engineering strategies and model evaluation metrics for accurate failure detection, deployed to production with monitoring and continuous improvement based on usage patterns
- Impact: Transformed expert analysis from hours to minutes (80% reduction), deployed to production serving 200+ daily queries with significant performance improvements through rapid iteration, established evaluation framework measuring diagnostic accuracy and user satisfaction for continuous product enhancement

AI Travel Planner Agent | Personal

- Problem: Manual travel planning requiring hours of research across multiple sources with inconsistent and outdated information, needing intelligent agent for personalized itinerary generation with real-time data integration
- Solution: Built AI agent using Claude 3.5 Sonnet API with LangChain orchestration and Streamlit interface, integrated DuckDuckGo Search API for real-time information retrieval, implemented prompt engineering techniques for conversational interaction and context management, developed evaluation methodology for response quality and relevance
- Impact: Demonstrated end-to-end AI application development from prototype to functional product, learned conversational AI patterns, real-time data integration techniques, and LLM orchestration strategies through iterative development and continuous improvement, showcasing ability to rapidly build and ship LLM-powered features