Viresh Duvvuri

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

AI Engineer with 5+ years building scalable AI-powered tools for business users, specializing in translating early-stage ideas into production applications through multi-agent systems, RAG pipelines, and LLM integration. Proven track record collaborating with non-technical stakeholders to develop workflow automation solutions that improved operational efficiency by 70-80% within 3 months. Strong expertise in rapid prototyping, cloud deployment (AWS), and communicating complex technical concepts to drive business impact through AI innovation.

Skills

AI Agent Development: Multi-Agent Systems, LangChain, LangGraph, RAG, Agentic AI, Prompt Engineering, Model Context Protocol (MCP), Function Calling, GenAI, Human-in-the-Loop

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

AI/ML Frameworks: PyTorch, TensorFlow, Scikit-learn, OpenAI APIs (GPT-4), Anthropic Claude, Ollama, Llama, Hugging Face, NLP, Transformers, LLM Fine-Tuning, Model Evaluation, MLOps, Model Deployment

Cloud & Infrastructure: AWS, Azure, Docker, Kubernetes, CI/CD Pipelines, REST APIs, Web Services, Monitoring, Performance Tuning, Scalability, Observability

Data & Databases: Vector Databases (FAISS, Pinecone), SQL, PostgreSQL, Data Pipelines, Data Processing, Data Warehouses, Real-Time Data Integration

Software Engineering: Git, Software Testing, API Development, Microservices, Agile Methodologies, Documentation, Technical Communication

Work Experience

Grid CoOperator — Seattle, WA

AI Engineer — *03/2025 - Present*

- Partnered with business analysts and stakeholders to translate workflow automation ideas into production-ready multi-agent AI system, collaborating with non-technical users to gather requirements and iterate on AI-powered tools that automated complex data analysis tasks, reducing manual effort by 70% within 2 months through rapid prototyping and user feedback integration
- Built and deployed scalable cloud-based infrastructure on AWS with REST APIs (FastAPI), data pipelines connecting SQL databases and web services, implemented multi-agent system using LangChain orchestration with RAG techniques (FAISS vector databases) for contextually accurate responses, serving 50-100 daily queries with comprehensive monitoring and logging
- Evaluated and prioritized AI features based on technical feasibility and business impact, documented system architecture and best practices for maintainability, communicated technical concepts clearly to non-technical stakeholders through demos and documentation, contributed to culture of experimentation by researching emerging AI technologies (MCP, function calling) and proposing high-impact applications

Freefly Systems — Woodinville, WA

Senior Software Engineer — *11/2021 - 10/2025*

- Developed and deployed GenAI-powered diagnostic tool from early-stage concept to production application serving 200+ daily queries, partnering with engineering teams to translate technical requirements into AI solution integrating foundation model APIs (Ollama, Llama 3.2) with RAG semantic search, enabling non-technical users to analyze complex system logs through natural language interface
- 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 workflow solution for database querying and intelligent insights, requiring AI system that non-technical business users could interact with naturally to understand complex data patterns and make data-driven decisions
- Solution: Developed multi-agent AI system using LangChain orchestration with specialized agents for SQL generation and context retrieval, built RAG pipeline with vector databases (FAISS) for intelligent information retrieval, integrated OpenAI APIs (GPT-4, Claude) with advanced prompt engineering, deployed scalable cloud infrastructure on AWS with FastAPI REST APIs and data pipelines, established testing framework and documentation for maintainability
- Impact: Translated early-stage idea into production tool within 2 months through rapid prototyping and stakeholder collaboration, enhanced business user productivity by 70% through workflow automation, served 50-100 daily queries at 99%+ uptime with comprehensive monitoring, validated impact through user feedback and data-driven evaluation showing 40% improvement in data accuracy

Production System Optimization Tool | Freefly Systems

- Problem: Manual system analysis requiring hours of expert time per case, creating bottlenecks in product development and customer support, needing AI-powered solution to enable non-technical users to perform complex diagnostics independently
- Solution: Built full-stack application with React frontend and Python Flask backend, integrated foundation model APIs (Ollama, Llama 3.2) for real-time processing, implemented RAG pipeline with semantic search and vector retrieval for contextual diagnostics, developed REST APIs and data pipelines, deployed to production with monitoring and continuous optimization based on user feedback
- Impact: Transformed early-stage prototype into production application serving 200+ daily queries, reduced analysis time from hours to minutes (80% reduction), enabled self-service for technical teams through natural language interface, demonstrated clear business impact through user satisfaction metrics and efficiency gains

AI Travel Planner Agent | Personal

- Problem: Manual travel planning requiring hours of research across multiple sources with inconsistent information, needing intelligent agent that could integrate real-time data and provide personalized recommendations through conversational interface
- Solution: Built AI agent using Claude 3.5 Sonnet API with LangChain orchestration, implemented function calling for API integration with DuckDuckGo Search enabling real-time information retrieval, developed prompt engineering techniques for natural conversation, created Streamlit interface for user-centric experience, applied evaluation methodology for quality
- Impact: Demonstrated end-to-end AI application development from concept to functional prototype, showcased rapid experimentation and iteration with emerging AI technologies, validated conversational AI patterns and tool integration approaches, illustrated adaptability and growth mindset through hands-on learning