Viresh Duvvuri

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

AI Engineer specializing in agentic AI systems and LLM orchestration, with 5+ years building production GenAI solutions through prompt engineering, RAG pipelines, and multi-agent architectures. Proven track record deploying AI systems that improved operational efficiency by 70-80% within 3 months, leveraging Model Context Protocol (MCP), function calling, and vector embeddings to automate complex workflows. Strong expertise in cross-functional collaboration, real-world model evaluation, and iterative optimization based on live interaction data at scale (200+ daily queries).

Skills

AI Agent Development: Agentic AI, LangChain, LangGraph, Multi-Agent Systems, Model Context Protocol (MCP), Agent-to-Agent (A2A), RAG, Prompt Engineering, Function Calling, API-based Reasoning, Structured Outputs, GenAI

AI/ML Frameworks: PyTorch, TensorFlow, Scikit-learn, LLM Fine-Tuning, Prompt Optimization, Few-Shot Learning, Model Evaluation, MLOps, Model Deployment, Reinforcement Learning (Learning), Feature Engineering, Responsible AI

LLM Integration: OpenAI APIs (GPT-4), Anthropic Claude, Ollama, Llama, Hugging Face, Context Engineering, Human-in-the-Loop

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

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

Data & Databases: Vector Databases (FAISS, Pinecone), Vector Search & Retrieval, Embedding Clustering, SQL, PostgreSQL, Data Pipelines, Real-Time Data Processing

Work Experience

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

- Built and shipped agentic AI system from prototype to production using multi-agent architecture with Model Context Protocol (MCP) integration, orchestrating specialized agents through LangChain for SQL generation and context retrieval, implementing function calling and API-based reasoning to automate complex analyst workflows, reducing manual effort by 70% within 2 months through iterative prompt engineering and structured reasoning
- Developed RAG pipeline with vector embeddings (FAISS) for contextually accurate real-time data retrieval, established comprehensive model evaluation frameworks tracking agent performance and cost efficiency across 50-100 daily queries, implemented observability dashboards monitoring business outcomes and proactively iterating on model behavior based on live interaction data
- Optimized agentic AI system performance through advanced prompt engineering techniques and multi-step task execution workflows, built internal tooling for automated deployment and evaluation, established AI governance practices including safety guardrails and bias detection, continuously improving model quality based on user feedback and golden set evaluations

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

- Built and deployed GenAI-powered diagnostic agent from concept to production integrating foundation model APIs (Ollama, Llama 3.2) with RAG semantic search for contextual accuracy, developed prompt engineering strategies and structured outputs to drive automated diagnostic workflows, served 200+ daily queries with observability monitoring and continuous performance optimization based on real-world interaction data
- 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, requiring agentic AI system for natural language interaction with enterprise data using multi-step reasoning and tool integration
- Solution: Developed agentic AI system with multi-agent orchestration using LangChain and Model Context Protocol (MCP), implementing specialized agents for SQL generation, web search, and context retrieval that coordinate through function calling and API-based reasoning. Built RAG pipeline with vector embeddings (FAISS) for intelligent context retrieval, integrated foundation model APIs (GPT-4, Claude) with advanced prompt engineering for structured outputs, established model evaluation framework with golden sets and synthetic data generation, deployed on AWS with FastAPI backend, observability tools, and real-time monitoring
- Impact: Enhanced analyst productivity by 70% through agentic AI that accomplishes complex tasks by invoking internal tools and APIs, demonstrated tangible results with 50-100 daily queries at 99%+ uptime, implemented human-in-the-loop evaluation and continuous improvement mechanisms based on live interaction data, validated optimizations through data-driven testing showing 40% improvement in response accuracy

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 with contextually accurate responses grounded in real-time data
- Solution: Built full-stack agentic application with React frontend and Python Flask backend, integrated foundation model APIs (Ollama, Llama 3.2) for real-time log processing, implemented RAG pipeline with semantic search and vector retrieval for contextual diagnostics, developed advanced prompt engineering strategies with few-shot learning and structured outputs for accurate failure detection, deployed to production with observability monitoring (logging, dashboards) enabling proactive iteration on model behavior
- Impact: Transformed expert analysis from hours to minutes (80% reduction), deployed to production serving 200+ daily queries, demonstrated 'show don't tell' mentality through tangible performance improvements validated by interactive testing, established evaluation framework measuring diagnostic accuracy and user satisfaction with continuous optimization based on real-world usage patterns

AI Travel Planner Agent | Personal

- Problem: Manual travel planning requiring hours of research across multiple sources with inconsistent information, needing intelligent agent with function calling capabilities for real-time data integration and multi-step task execution
- Solution: Built AI agent using Claude 3.5 Sonnet API with LangChain orchestration, implemented function calling for API-based reasoning with DuckDuckGo Search API enabling real-time information retrieval, developed advanced prompt engineering techniques for conversational interaction with structured outputs and improved contextual memory, created Streamlit interface for user-centric experience, applied evaluation methodology for response quality
- Impact: Demonstrated end-to-end AI application development from prototype to functional product, showcased ability to rapidly iterate and experiment with new techniques, learned conversational AI patterns and tool integration protocols through hands-on development, validated agentic AI design principles for multi-step reasoning tasks