Viresh Duvvuri

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

AI Engineer

AI Engineer specializing in multi-agent systems and GenAI solutions, with 5+ years developing production AI applications from prototype to production-grade deployment. Led cross-functional teams in delivering GenAI solutions that improved business efficiency by 50-80% within 3 months, establishing model governance, evaluation frameworks, and MLOps pipelines for scalable AI solutions on AWS/Azure.

SKILLS

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

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

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

Data & Analytics: Data Integration, Data Processing, Data Pipelines, Data Science, SQL, Enterprise Integrations, Enterprise Systems, Knowledge Graph, Analytics, Operational Efficiency

---

WORK EXPERIENCE

Grid CoOperator | AI Engineer • Full-time

03/2025 - Present | Seattle, WA

- Led end-to-end development of GenAI solution from initial prototype to production deployment, collaborating cross-functionally with business stakeholders to identify opportunities and translate operational requirements into multi-agent systems using LangChain orchestration, delivering measurable business value with 70% reduction in analyst workflows within 2 months
- Architected AI orchestration system where specialized agents communicate and coordinate for complex analytics tasks, deployed on AWS with observability and cost monitoring, established rigorous model governance including evaluation pipelines tracking quality metrics, latency, and performance to achieve reliable enterprise performance within 6 weeks across 50-100 daily queries
- Delivered analytics and insights to measure solution value, implementing comprehensive monitoring dashboards and performance tracking that demonstrated 60% acceleration in deliverables within first quarter while maintaining system reliability and bias-free operations through continuous testing and validation

Freefly Systems | Senior Software Engineer • Full-time

11/2021 - 10/2025 | Woodinville, WA

- Built GenAI-powered co-pilot for automated log analysis from concept to production, integrating foundation model APIs (Ollama, Llama 3.2) and deploying to cloud infrastructure with model evaluation metrics and monitoring, reducing manual workflows by 80% within 3 months and delivering measurable ROI through analytics tracking
- Led cross-functional collaboration translating business requirements into technical solutions, implementing software design principles, testing frameworks, and model governance practices across engineering divisions to ensure accurate and bias-free AI applications
- Enhanced distributed systems with microservices architecture and CI/CD pipelines, improving deployment efficiency by 60% over 6 months through rigorous testing and continuous improvement methodologies

Lumenier | Software Engineer - Embedded Systems • Full-time

07/2020 - 10/2021 | Sarasota, FL

- Implemented custom software using C++ and optimized data structures for specialized applications, enabling autonomous capabilities within 8 weeks through algorithm optimization and efficient data processing pipelines
- Enhanced system performance through data ingestion pipelines and algorithm improvements, improving operational efficiency by 45% across environments while establishing testing frameworks to ensure solution accuracy

York Exponential | Software Engineer - R&D • Full-time

08/2018 - 05/2020 | York, PA

- Developed Human Machine Interface using Python and ROS2, reducing operator programming complexity by 50% within 4 months through innovative AI/ML integration and workflow automation
- Built autonomous robot prototype integrating computer vision and machine learning pipelines from requirements gathering to production deployment, demonstrating end-to-end technical leadership over AI solutions

---

EDUCATION

Master of Science in Computer Science Washington State University | Pullman, WA, USA • 01/2015 - 01/2017

Bachelor Of Technology in Information Technology GITAM University | Visakhapatnam, India • 01/2011 - 01/2015

PROJECTS

GridCOP: Smart Grid Analytics Agent | Grid CoOperator

- **Business Challenge:** Power grid analysts needed automated database querying and intelligent insights to understand complex data patterns, requiring a GenAI solution that could be rapidly prototyped, validated for business value, and transformed into production-grade capability
- **Solution & Technical Leadership:** Led development of 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, integrated multiple foundation models (Claude, GPT-4), established comprehensive model governance including evaluation frameworks monitoring quality and cost metrics, deployed on AWS with auto-scaling, observability, and logging
- **Business Impact:** Enhanced analyst productivity by 70% through AI co-pilot that augments domain experts with automated workflows. Delivered analytics measuring solution value including latency tracking, accuracy metrics, and cost efficiency. Implemented human-in-the-loop (HIL) evaluation and rigorous testing pipelines ensuring production-ready AI systems remained accurate and bias-free through continuous monitoring

Production System Optimization Tool | Freefly Systems

- **Business Challenge:** Manual system analysis creating bottlenecks requiring hours of expert time, impacting product development velocity and customer support resolution timelines
- **Solution & Technical Leadership:** Built full-stack GenAI application integrating foundation model APIs (Ollama, Llama 3.2) for real-time log processing. Led model selection process, implemented prompt engineering strategies, and established model evaluation frameworks. Deployed to production cloud infrastructure with comprehensive monitoring and analytics tracking
- **Business Impact:** Transformed expert analysis from hours to minutes (80% reduction), deployed to production serving 200+ daily queries. Delivered measurable value through analytics dashboards tracking performance improvements, cost efficiency, and user satisfaction metrics. Established governance practices including testing, monitoring, and documentation ensuring solution accuracy

AI Travel Planner Agent | Personal

- Built AI agent using Claude 3.5 Sonnet API, LangChain orchestration, and DuckDuckGo Search API integration, demonstrating rapid prototyping capabilities and GenAI solution development including model selection, prompt engineering, and iterative improvement from concept to production-grade applications