Sowmith Kuppa

+1 (757) 798-9998 | soumith.odu@gmail.com | linkedin.com/in/soumith29 | github.com/s0umith29

Education

Old Dominion University Norfolk, VA

M.Sc. in Computer Science, GPA: 3.84/4.0 *Aug 2023 – May 2025*

- Relevant Coursework: MLOps for Generative AI, Natural Language Processing, Intro to Data Science and Analytics, Cybersecurity Fundamentals

Experience

AI Research Intern Jan 2025 – Present

Old Dominion University *Norfolk, Virginia*

- Designed and implemented machine learning models using PyTorch to support research on reliable AI systems, focusing on model robustness and performance evaluation.
- Developed and integrated Retrieval-Augmented Generation (RAG) pipelines to enhance dataset quality and context relevance, improving model accuracy and research insights by 20%.
- Conducted A/B testing and controlled experiments to evaluate new model features, statistically measuring their impact on accuracy and user experience.
- Created data mining and drift detection algorithms to analyze large datasets, generating actionable visual reports to monitor and explain model behavior over time.
- Performed iterative model refinement and testing, leading to measurable improvements in performance and contributing to findings on reliable and scalable AI techniques.

Graduate Research Assistant Aug 2023 – May 2025

Old Dominion University *Norfolk, Virginia*

- Deployed infrastructure with Ansible and Terraform reducing the provisioning time by 40% in various production grade systems.
- Automated CI/CD pipelines using Jenkins and Terraform, shortening deployment cycles by 30% and aligning with customer-ready ITSM practices.
- Engineered monitoring solutions with Prometheus and Grafana (akin to Datadog), maintaining 99.9% uptime by resolving bottlenecks in real-time.
- Collaborated with cross-functional teams, including ITS, to resolve over 100 hardware and software issues in Windows, Linux and MacOS systems, enhancing system reliability for Computer Science Department users.
- Conducted routine system checks, patching, and updates to ensure compliance and system health.

Projects

MedSim Jan 2025 – Present

- Re-architected the platform into a decoupled microservices architecture, utilizing React.js for a responsive client interface and FastAPI for high-concurrency asynchronous backend handling.
- Engineered a Hybrid RAG pipeline on Google Cloud Platform utilizing Vertex AI Vector Search and Gemini 2.5 Pro to retrieve precise diagnostic context from a dataset of 40,000+ PMC case reports.
- Developed an automated OSCE-style evaluation engine that analyzes student-patient chat logs in real-time, acting as a senior faculty member to generate detailed clinical performance reports.
- Deployed the full-stack application as containerized services on Google Cloud Run, implementing strict IAM roles for security and optimizing Docker builds to minimize cold-start latency.

Technical Skills

- **Machine Learning & AI**: Scikit-learn, PyTorch, TensorFlow, Agentic AI, Retrieval-Augmented Generation (RAG), Model Deployment, Hyperparameter Tuning, ML Pipeline Automation, Apache Spark
- **Data Science & Analytics**: Python (Pandas, NumPy), Data Preprocessing, Statistical Analysis, Data Visualization
- **MLOps & Automation**: Ansible, Terraform, Jenkins, ArgoCD, Helm, GitLab CI/CD, canary/blue-green deployments
- **Cloud & Containerization**: GCP (Cloud Run, Cloud Storage, Vertex AI, Vector Search, GKE), Docker, Kubernetes
- **Databases**: PostgreSQL, MySQL, Google Cloud Storage
- **Programming Languages**: Python, Go, SQL, Bash, PowerShell, C++, Java, HTML, CSS, JavaScript
- **Observability & Monitoring**: Prometheus & Grafana (ML Monitoring), Datadog (like workflows)
- **Collaboration & Workflow Tools**: JIRA, ServiceNow, SharePoint

Certifications

CKA: Certified Kubernetes Administrator April 2025

The Linux Foundation *Cluster Management, Workloads, Troubleshooting*

- Hands-on expertise in installing, configuring, managing, and troubleshooting production-grade Kubernetes clusters mastering cluster architecture, workloads, networking, storage, and efficient use of kubectl in real-world scenarios.