[LinkedIn] **Karthik Nayak** [nayakkarthik02@gmail.com]

EDUCATION BS in Computer Science, University of Kentucky – GPA: 3.6 / 4.0 2015-2019

SOFT SKILLS

Independent problem solver, quick learner, strong conviction in long-term product vision, challenges conventional processes, adaptable, and clear communicator.

TECHNICAL SKILLS

Concepts	Technologies	Cloud	Programming
LLM Optimization	LLM Serving Containers	Azure ML & OpenAI	Python
AI Model Benchmarking	Monitoring & Observability	AWS SageMaker	C#
ML Infrastructure	ML Orchestration	Google Vertex AI	Java
Inference Deployments	Kubernetes	Hybrid & On-Prem ML	C++

EXPERIENCE

Goldman Sachs | Vice President – Data Science & Machine Learning Platform **Oct. 2022 – Present**

- Provisioned OpenAI models on Azure using Bicep IaC, establishing the infrastructure for Goldman Sachs' firmwide chat application to enable knowledge retrieval and support within the organization.
- Designed and implemented a custom benchmarking framework for assessing LLM throughput, leveraging various serving containers (e.g., TGI, DJL, Triton, vLLM) to drive data-backed decisions on multimillion-dollar GPU purchases.
- Created new Analytics Service to monitor and alert on inference deployments, reducing underutilized endpoints by over 20% and cutting costs charged back to customers' Business Units.
- Enabled Async Inference capabilities within the Inference Service utilizing AWS SageMaker, including comprehensive end-2-end development across control plane Lambdas, CLI, IAM roles, and probers.
- Enhanced Model Registry and Inference Service APIs, improved CLI usability, and led network and platform migrations with zero downtime contributing to team success stories.
- Facilitated onboarding by developing end-to-end ML pipeline examples making use of custom ML images, JupyterHub, and CI/ML templates, automating training jobs, model uploads, and inference deployments.

Microsoft | Software Engineer II – Azure ML Pipelines & Datasets **July. 2019 – Oct. 2022**

- Justified & developed ETL telemetry pipeline for 1P ML Experimentation platform, using Azure Event Hub and Synapse Data Warehouse increasing visibility to drive SLA improvement to 99.9% across various compute types.
- Expanded deployment capabilities to private clouds by leading development requiring provisioning 23 ML Kubernetes services with network restriction & new CI-CD workflows.
- Designed\Developed\Validated various ML services that perform job orchestration, policy & access management, data ingestion, caching, and ML Dataset management while maintaining high reliability.
- Worked closely with data scientists and other dependency service teams to migrate users from the 1P->3P platform by developing new API requirements, SDK improvements, and monitoring.

Kinemetix | Software Research Intern **Jan. 2019 – Apr. 2019**

- Improved 3D pose estimation algorithm runtime in OpenCV by reducing redundant operations, enhancing efficiency for large-scale point cloud data. Used convex decomposition to simulate complex physics collisions and render synthetic data.

IBM | Software Developer Intern **June. 2018 – Aug. 2018**

- Transitioned IBM's BTI SaaS product to a microservices architecture by developing REST APIs in Java and a containerized UI with Angular, leveraging CI/CD tools like Maven and Jenkins.

Tesla | Automation Controls Engineering Intern **May. 2017 – Aug. 2017**

- Led the drive unit team to build a custom PLC simulator for AGV interface testing, enhancing debugging capabilities and supporting safety buy-offs on Model 3 battery and drive-unit production machines with global automation engineering contractors.

Visumatic Industrial Products | Computer Engineering Intern **May. 2016 – Aug. 2016**

- Built automation machines from the ground up, which involved building, electrical wiring, developing program logic and UI for pneumatic/DC drivers, touch HMI displays, and rotating indexing systems.

Computer Science Dept, U of Kentucky | Undergrad Research Assistant **Oct. 2017 – May. 2018**

- NSF-funded research project to test the performance of Software-Defined Networks (SDN) on campus networks