Harish Muthyala

+1 (281) 965-2335 harishcmuthyala@gmail.com linkedin.com/in/harish-muthyala harishmuthyala.com

Skills

Credentials: AWS Solutions Architect - Associate, Accenture Trailblazer Award, AWS CodeWhisperer Article
Languages & Frameworks: Python, Java, SQL, Langchain, RAG, MLOps, Pandas, PyTorch, C/C++, JavaScript
Cloud: AWS - S3, Lambda, SageMaker, Bedrock, EC2, VPC, DynamoDB, CodePipeline, QuickSight, Azure Foundations
Developer Tools: Git, Docker, Kubernetes, Postman, VS Code, Jupyter

Experience

Senior Analyst, Generative AI Engineer Jan. 2024 – Aug. 2024
Accenture AWS Business Group (AABG) *Hyderabad, India*

- Architected **end-to-end Retrieval-Augmented Generation (RAG) pipeline** that automated credit memo generation from financial documents, reducing underwriting time by **75%** for Accenture's banking clients.
- Deployed **OpenSearch vector database** with **Titan embeddings** to enable sub-second retrieval from complex financial documents, processing **200+ page reports** into queryable insights for credit committee decision-making.
- Engineered **few-shot prompt templates** with domain-specific examples, achieving **compliance-ready** credit memo generation with accurate financial notation including bracketed negatives and industry-standard formatting.
- Built robust **ETL pipelines in Pandas** for preprocessing financial data, implementing data quality checks and chunking strategies that maintained **99%+ accuracy** in extracting structured data from diverse financial statements.

Analyst, Machine Learning Operations Engineer Aug. 2022 – Jan. 2024
Accenture AWS Business Group (AABG) *Hyderabad, India*

- Engineered an end-to-end **ML automation platform** using **CloudFormation templates** that deploys **SageMaker Autopilot pipelines** in isolated **VPC environments**, automatically generating scalable **API endpoints** from user-uploaded S3 datasets without manual intervention.
- Implemented real-time **data and model quality monitoring** via **SageMaker Model Monitor**, detecting drift by comparing live inference data against baseline statistics from ground truth datasets.
- Built **end-to-end QuickSight migration platform** with **web UI** for CloudFormation deployment and **RESTful API** for automated cross-region dashboard transfers, reducing migration time from **hours to minutes**.

Project Analyst Oct. 2024 – Present
Office of Information Technology, University of Houston *Houston, TX*

- Collaborated with **cross-functional teams** to manage and deliver **technology projects** aligned with university strategic goals, ensuring project visibility and **successful implementation**.
- Led migration from FootPrints to **TeamDynamix platform**, enabling **self-service IT request submission** for 20,000+ students and faculty while configuring **3 workflow automation systems** that replaced email-based processes with centralized dashboards—modernizing service delivery for 80+ IT staff.

Projects

LLM Inference Acceleration Research | *Quantization, Speculative Decoding*

- Conducted systematic review of **22 research papers** on LLM inference acceleration, comparing quantization and speculative decoding techniques across performance metrics (speed, memory, accuracy) to identify optimal optimization strategies.
- Identified that hybrid approaches combining both techniques achieve **2.6× speedup** and **45% memory reduction** while maintaining **99% accuracy**—outperforming individual methods by addressing complementary bottlenecks.
- Determined synergistic mechanism: quantization frees memory enabling deeper speculation, while draft-verify correction preserves accuracy despite compression—findings validated by production deployments at Google and IBM.

Education

Master of Science in Computer Science, *University of Houston* Aug. 2024 – May 2026
Machine Learning, Generative AI, Image Processing | ***GPA:*** *3.8/4.0* *Houston, TX*

Bachelor of Technology in Computer Science, *Vellore Institute of Technology* Jul. 2018 – Apr 2022
Data Structures, Software Engineering, Computer Networks, Artificial Intelligence *Vellore, India*