Harisathwik Veerla

+91-9398822033 | harisathwikveerla@gmail.com | github.com/Harisathwik | linkedin.com/in/harisathwik-veerla

Education

M.Tech in Artificial Intelligence and Machine Learning *Pilani*
Birla Institute of Technology and Science (BITS) Pilani (Work Integrated Learning Program) *Oct 2023 - Oct 2025*

PG-Diploma in AI Robotics and Allied Technologies *Kolkata*
Centre for Development of Advanced Computing (CDAC) *Sept 2021 - March 2022*

B.Tech in Computer Science and Engineering (CSE) *Hyderabad*
JNTUH College of Engineering Sultanpur *July 2017 - Apr 2021*

Work Experience

Associate Consultant *Bangalore*
KPMG Global Services Private Limited *Sep 2025 - Present*

- Architected enterprise-grade GenAI applications for financial clients, developing retrieval-augmented generation (RAG) architectures and accelerating model inference APIs.
- Constructed end-to-end MLOps and LLMOps lifecycles on Azure and AWS, streamlining the transition from training to live deployment.
- Improved large language model efficiency via model compression (quantization) and memory caching, lowering prediction latency by 40% and cloud expenditure by 25%.
- Built resilient streaming and batch prediction workflows under strict performance guarantees, enforcing automated data integrity and formatting checks for high-throughput systems.
- Orchestrated continuous integration and delivery pipelines via GitHub Actions, accelerating deployment cycles by 30% across multiple staging environments through to live production.
- Designed comprehensive telemetry solutions to monitor system stability, detect data drift, and measure generative model quality, including response times and output factuality.

Project Engineer *Pune*
FEV India Private Ltd *May 2022 - Sep 2024*

- Spearheaded scalable machine learning workflows and large-scale data transformation jobs for ADAS systems using Python and PySpark.
- Established continuous deployment practices and infrastructure-as-code via Azure DevOps, creating consistent release templates.
- Programmed custom diagnostic dashboards in Python for continuous evaluation of system uptime, dataset integrity, and early fault detection.
- Enforced robust testing protocols for analytical data pipelines, guaranteeing consistent experimental results and traceable updates.
- Embedded stringent data formatting rules and verification gates into enterprise data streams, boosting system reliability by 35%.
- Drove cloud resource efficiency campaigns across internal tools and AI hosting platforms, cutting excessive cloud spending by over 20%.
- Served as the primary engineering liaison for ML systems, establishing rapid troubleshooting procedures and coordinating cross-team deliveries.

Skills

MLOps & LLMOps Scalable Model Serving, RAG Workflows, Vector Databases, MLflow, Inference Optimization, Telemetry & Observability.
Cloud & DevOps AWS, Azure, Docker, Kubernetes, Infrastructure as Code, CI/CD (GitHub Actions), Linux, Git.
Programming Python (PyTorch, LangChain, LangGraph, ADK, Pandas), SQL, MATLAB, Dart, Flutter.