NARAYAN JANGID

+91 9462637251 | naaidjan.19@gmail.com | linkedin.com/in/nryn-221 | github.com/Narayan-21

EDUCATION

Indian Institute of Science Education and Research Kolkata Kolkata, WB

BS-MS (Bachelor + Master of Science) in Geological Sciences, Minor - Physics *Aug. 2017 – July 2022*

- Relevant Coursework: Linear Algebra, Probability, Programming and Data Structures, Analysis I, II

EXPERIENCE

Machine Learning Engineer March 2025 – Present

Kloudspot *Bangalore, KA*

- Optimized vLLM and SGLang frameworks with EAGLE-3 speculative decoding, including training the draft model to reach an acceptance rate of 80% of generated tokens, resulting in up to 2× higher throughput and 2.7–3.5× lower latency in production.
- Finetuned and deployed LLMs (Llama 3.1, DeepSeek, Qwen) for reasoning tasks, optimized on RTX-4090 GPUs for scalable, low-latency, high-concurrency workloads. .
- Conducted experiments and benchmarked multiple LLM inference acceleration techniques like speculative decoding, look-ahead decoding, medusa, Eagle-1/2/3 etc.
- Enhanced LISA AI with advanced Agentic AI features , graph-generation functionality and built a MCP client for seamless integration with enterprise data sources and tools.

Associate Data Scientist Nov. 2023 – Feb 2025

Celebal Technologies *Jaipur, RJ*

- Designed a hybrid VLM inference system, deploying a quantized finetuned model on edge and full-scale finetuned on cloud. Built a real-time video analytics pipeline with adaptive edge-cloud switching. Developed data prep, tagging, and retraining modules, achieving 77.9% mAP for the usecase. (YOLO11, QLoRA, PyTorch, SmolVLM)
- Modeled and optimized predictive algorithms (ARIMA & SARIMA) and machine learning algorithms (XGBoost, LightGBM, CatBoost, K Means) for predictive analytics in energy demand forecasting, achieving a 75.2% R^2 score for electricity load detection.
- Worked on an on-premise OCR solution using open-source models and fine-tuning them for specific use cases. (Microsoft Table Transformers, Tesseract OCR, Layout Parser, Detectron2, ONNX Runtime)

Data Science Intern June 2023 – Oct. 2023

Celebal Technologies *Jaipur, RJ*

- Led the design and deployment of AI-powered solutions, leveraging Azure OpenAI and document intelligence to extract, embed, and index structured/unstructured data, achieving a 15% improvement in BLEU Score and 10% boost in semantic similarity.
- Implemented enterprise-grade security with Azure AD Authentication, APIM, and versioning, ensuring robust compliance for AI and data science solutions.

Graduate Research Student May 2021 – July 2022

IISER Kolkata - Environmental Nanoscience Lab *Kolkata, WB*

- Masters Thesis Research on - Quantitative analysis of heavy metal distribution along the Hooghly estuary: Towards establishing the inter-relation between surface water, groundwater and estuarine water

TECHNICAL SKILLS

Languages: Python, C, C++, JavaScript, TypeScript
GPU programming: Cuda, cuBLAS, cuDNN, CUTLASS, Triton
Libraries: Pandas, NumPy, Matplotlib, OpenCV, PySpark, AsyncIO
Frameworks: PyTorch, TensorFlow, Langchain, Langgraph, LlamaIndex, Celery, gRPC, Flask, FastAPI
Technologies: Git, Docker, Databricks, Azure Cloud, AWS EC2, Azure Functions, CloudFlare Workers, Selenium, WebSockets, Redis, PostgreSQL, MongoDB

CERTIFICATES

Microsoft Certified: Azure Data Scientist Associate
Microsoft Certified: Azure AI Engineer Associate
Databricks Certified: Generative AI Engineer Associate

COMMUNITY ENGAGEMENT

Cohere Labs – Open Science Community: Actively collaborate with global ML researchers and contribute to open-source discussions on generative AI and ML efficiency.
vLLM Office Hours – Red Hat & UC Berkeley: Regular participant in bi-weekly sessions on LLM inference optimization and ML performance research.