Nagraj Desai

Kolhapur, Maharashtra, India
Email: nagrajdesaee@gmail.com — Phone: +91-7387936696 — LinkedIn — GitHub

SUMMARY

AI Engineer with hands-on experience in LLM, Deep Learning and GPU-accelerated systems, specializing in designing and deploying end-to-end ML pipelines. Skilled in Python, Docker, Kubernetes, vLLM, and PyTorch with expertise in building RAG systems and optimizing scalable inference and model serving for production environments.

EXPERIENCE

AI Engineer April 2025 – Present
Micropoint Computers Private Limited

- Building an on-premise **Model-as-a-Service** platform leveraging **Docker**, **Kubernetes**, and **vLLM**, centralizing access via a unified **Open Web UI**.
- Orchestrated AI workflows with **Docker** and **vLLM** to enable **Distributed Inference**, significantly optimizing GPU utilization and minimizing latency.
- Engineered scalable **Retrieval-Augmented Generation (RAG)** systems within **Kubernetes pods**, delivering context-aware insights from unstructured documents.
- Integrated **NVIDIA NIM GPU-accelerated microservices** and various **LLM models** into production, ensuring high scalability and seamless inference pipelines.

Data Science Intern Jul 2024 – Jan 2025
Taabi Mobility Ltd.

- Developed Vehicle Health predictive algorithm and robust API using **Python** and **GraphQL** for real-time vehicle health monitoring.
- Improved the Event Prediction Algorithm, achieving an accuracy uplift from 99.5% to **99.62%**.
- Enhanced the data extraction workflow, slashing processing time by **78%** (from 3 hours to 40 mins).
- Conducted rigorous daily analysis of vehicle performance metrics, ensuring data integrity and consistency across datasets.

SKILLS

Core Technical: Python; SQL; Machine Learning; Deep Learning; Transformers; Large Language Models (LLMs); Retrieval-Augmented Generation (RAG); PyTorch; vLLM; Scikit-learn; LangChain; Hugging Face; Model Optimization
MLOps & Deployment: Docker; Kubernetes; Triton Inference Server; NVIDIA NIM; Model Serving
Data & Analytics: Exploratory Data Analysis (EDA); Statistics; MySQL; MS Excel
Development & Frameworks: FastAPI; Streamlit; Linux; Version Control (Git)

CERTIFICATIONS

NVIDIA Technical Curriculum (DGX, AI, Compute, Generative AI, 2025)

EDUCATION

Post Graduation Diploma in Data Science 2023 – 2024
IIIT Bangalore
CGPA: 3.97/4

B. Tech 2019 – 2023
Shivaji University, Kolhapur
CGPA: 8.94/10

PROJECTS

Reasoning-Enhanced LLM Fine-Tuning — *Stack: Python, Unsloth, PyTorch, QLoRA*
Fine-tuned **Llama-3.2-3B** using **Unsloth** and **QLoRA** to integrate Chain-of-Thought (CoT) capabilities from the **R1-Distill-SFT** dataset. Optimized training on a T4 GPU via **4-bit quantization**, successfully enabling the model to generate structured steps for complex tasks.

Visual-Semantic Captioning Engine — *Stack: Python, PyTorch, CNN, GRU, Attention*
Built a image-to-text **Encoder-Decoder** pipeline combining CNNs and GRUs with **Attention Mechanism** to generate descriptive image captions. Optimized the model for semantic accuracy, attaining a competitive BLEU score of 0.6 on the Flickr8K dataset and validating superior visual-textual alignment.

RAG-Based Knowledge Retrieval System — *Stack: LLM, VectorDB, Streamlit, Python, LangChain, Docker*
Built a context-aware **Retrieval-Augmented Generation (RAG)** pipeline to extract insights from unstructured data. Leveraged **LangChain** workflows for optimized vector retrieval and fully **containerized** the application with **Docker**, ensuring scalable and reproducible deployment.