Yogesh Murala

muralayogesh@gmail.com — +91 9014802929
linkedin.com/in/yogeshmurala — github.com/Yogesh-001

Professional Summary

AI Engineer with hands-on experience in Generative AI, LLM fine-tuning, MLOps, and Python backend development. Skilled in designing scalable AI systems, backend automation, and deploying ML pipelines with CI/CD and containerization.

Technical Skills

- **Programming:** Python, C/C++, CUDA
- **AI/ML Frameworks:** PyTorch, TensorFlow, LangChain, LangGraph, OpenCV, Scikit-learn
- **DevOps Tools:** Azure DevOps, Jenkins, Docker, CI/CD Pipelines, MLFlow (basic), Databricks (learning)
- **Concepts:** Generative AI, RAG, LLMs, MLOps, Autonomous Systems, SQL, NoSQL, ONNX
- **Certifications:** Microsoft Certified: Azure Fundamentals

Professional Experience

Bosch Global Software Technologies (BGSW) **Bangalore, India**
Senior Engineer – AI Automation, Aug 2023 – Present

- Designed and deployed scalable AI/ML systems (LLMs, RAG, traditional ML) for ADAS and automation.
- Designed and implemented a modular Python backend to parse complex YAML configs into JSON schema required by internal systems, enabling automated visualizations via Yaavis.
- Post-trained LLaMA 3 chatbots using Supervised Fine-Tuning (SFT) on internal enterprise QA pairs, reducing document query resolution time by 20%
- Led GenAI initiatives using LangChain and LLaMA 3: **JIRA AI Assistant** – 30% reduction in triaging effort; **Autonomous Agent for CI/CD** – 25% improvement in pipeline recovery time with Jenkins integration.
- Designed containerized MLOps pipelines with Docker and MLflow; integrated Jenkins + Azure DevOps for CI/CD automation across experimentation and deployment stages.
- Benchmarked and optimized ONNX models using Apache TVM (auto-tuning, operator fusion, and target-specific codegen); achieved near-parity inference performance with TensorRT on non-CUDA machines.
- Collaborated across cross-functional teams to build ML lifecycle tools integrating SQL and NoSQL databases.
- Contributed to internal knowledge sharing by building training scripts, deployment, and pipeline orchestration.

Bosch Global Software Technologies (BGSW) **Bangalore, India**
Project Trainee – Computer Vision & Deep Learning, Sep 2022 – May 2023

- Developed a YOLO-based real-time object detection pipeline with TensorRT optimizations for low-latency inference on industrial datasets.
- Built a custom multi-head, multi-modal neural network with object-specific (localization, regression) heads for spatial and semantic object localization in traffic scenes.
- Improved detection accuracy (+12% mAP) by researching and applying custom anchor box designs and optimized NMS techniques.
- Converted the model to ONNX and deployed it with TensorRT using CUDA mixed-precision for real-time applications.

Projects

- **AI-Powered Debugging Assistant:** Fine-tuned LLaMA 3.2 on GitHub issues and bug descriptions to build a self-hosted assistant; achieved 80% classification accuracy on 10K+ samples.
- **Medical Diagnosis RAG Assistant:** Built a LangChain-based RAG system for symptom-based diagnosis using clinical case studies and literature.
- **Language Translation Transformer:** Developed a Transformer from scratch for English–Telugu translation using attention and positional encoding; achieved 60% accuracy on limited compute.

Education

KL University *2019 – 2023*
B.Tech in Computer Science Engineering (Specialization in Artificial Intelligence) — CGPA: 9.16 / 10