Samar Kale

Worcester, MA | ks.samar.kale@gmail.com | (508) 373 6980 | linkedin.com/in/samar-kale
github.com/REZ3LIET

Experience

AI Engineer, Tata Elxsi Jul 2022 – May 2025

- Led development of an LLM-driven autonomous decision-making framework, growing the team from 2 to 4 engineers and mentoring interns/new hires; architected LangChain-based agent pipelines for complex task-planning across distributed systems.
- Built end-to-end agent architectures integrating retrieval, tool-use, self-reflection and multi-step planning; deployed pipelines that generated structured action plans executable by downstream systems.
- Created a GPT-guided workflow-generation system using few-shot prompting and custom C++ orchestration wrappers, enabling auto-generated Behavior Tree execution graphs with three-second user-to-robot task execution; system demonstrated publicly at IREX Japan 2023.
- Benchmarked 5–6 depth estimation models (MiDaS, ZoeDepth, etc.) for downstream 3D inference tasks; implemented ONNX quantization workflows to reduce model size and improve runtime for real-time pipelines.
- Designed and trained a U-Net model on a 1200 image custom dataset for footpath segmentation, achieving 70% accuracy and improving dataset consistency via error analysis and relabeling heuristics.
- Integrated ML models with production-grade systems using Docker, gRPC, and ROS2-style modular architectures; packaged perception and planning modules as containerized services for reliable deployment.

AI Intern, Tata Elxsi Aug 2021 – Jun 2022

- Developed computer vision training pipelines including dataset generation, augmentation, and validation cycles for segmentation and depth-estimation models.
- Implemented early prototypes of agent-style planning and ML inference modules integrated with simulation environments.

Projects

Resume Chat-Bot Github

- Built a Retrieval-Augmented Generation system for resume evaluation and interview simulations using Python, and LangChain.
- Implemented embedding pipelines, retrieval logic, and structured prompt flows to improve response accuracy and relevance.

Technologies

AI Systems: Large Language Models (LLM), Agentic LLM Pipelines, LangChain Tools, Prompt Engineering, Program Synthesis, Structured Planning
Machine Learning: PyTorch, Depth Estimation, Semantic Segmentation (U-Net), ONNX Quantization, Embeddings, RAG Systems
Languages: Python3, C++
Infra & Tools: Docker, gRPC, Milvus, git

Education

Worcester Polytechnic Institute, MS in Artificial Intelligence Aug 2025 – Aug 2027

Vishwakarma Institute of Technology, B.Tech in Mechanical Engineering Aug 2018 – Jun 2022

- GPA: 8.06/10.0