KEDAR GAIKWAD

kedar.gaikwad@asu.edu ◇ www.linkedin.com/in/kedardgaikwad ◇ (602) 662-6935 ◇ Phoenix, AZ

EDUCATION

MS, Robotics Autonomous Systems (Artificial Intelligence) | Arizona State University — Aug 2023 - May 2025
Courses: Frontiers in GenAI, Operational Deep Learning, ML accelerator design

BE, Computer Engineering | University of Mumbai — Aug 2015 - May 2019
Courses: Machine Learning, Artificial Intelligence

SKILLS

Languages & Technologies	Python, C++, Docker, Git, Jira, GCP, AWS
ML Frameworks	PyTorch, TensorFlow, Scikit-learn, ONNX, TensorRT, OpenVino
LLM Frameworks	OpenAI, LangGraph, Langchain, LlamaIndex, Ollama, LangSmith, CrewAI
Embedded Systems	Nvidia Jetson, Ambarella CV22, Raspberry Pi

EXPERIENCE

Al/ML Software Engineering Intern, Stealth Startup, ASU - Phoenix, AZ — February 2025 - Present

- As the first AI/ML developer, led the creation of core application functionalities.
- Developed a robust web scraping pipeline for data extraction, utilizing LLM APIs for metadata enhancement and automating database uploads. Deployed and orchestrated the scraper using Apache Airflow for continuous monitoring and scheduling.
- Designed and implemented AI-driven document enhancement endpoints, deploying them via GCP Cloud Run. Integrated with GitHub for streamlined CI/CD, ensuring rapid and reliable deployments.
- Created an audio podcast generator leveraging Gemini 1.5 Pro within a Retrieval-Augmented Generation framework. Integrated Google Text-to-Speech Neural and Studio voice models, employing Speech Synthesis Markup Language (SSML) for natural and expressive voice output.

Research Assistant, exsight.ai, ASU - Phoenix, AZ — October 2023 - Present

- Integrated neuro-symbolic approaches with Object Detection models to create Explainable AI (**XAI**) solutions, enhancing interpretability and increased object detection recall by 30% in military geospatial imaging applications
- Secured an STTR Phase 1 Air Force/Space Force contract, gaining recognition as featured by W. P. Carey News.
- Engineered robust **stress testing framework** utilizing **Meta SAM 2** for precise segmentation, enabling targeted adversarial patch and camouflage attacks that identified and addressed key vulnerabilities in mission-critical AI systems
- Optimized **XAI models**for **Nvidia Jetson** edge deployment through quantization, pruning, and layer fusion, maintaining accuracy while enabling real-time inference capabilities

AI Research Intern, RagaAI - Fremont, CA — Jun 2024 - Aug 2024

- Built an observability tool, RagaAI Catalyst to provide trace recording inside RAG applications with one-click deployable solution allowing fine-tuning and evaluation for LLM applications
- Collaborated on creation of Raga LLM Hub, employing metrics to evaluate LLMs, and established critical guardrails for LLMs and RAGapplications, culminating in a robust open-source framework enriched with over 100 comprehensive tests
- Benchmarked and optimized custom RAG pipelines for prompt response quality across **Llama**, **Gemma**, and **Mistral** models, significantly reducing **token costs** while enabling engineering teams to identify the most cost-effective solutions for deployment.

Senior Data Scientist, RagaAI - *Bangalore, IN* — January 2022 - August 2023

- Led implementation of custom **autoencoder** network for drift tracking and outlier detection in **ADAS**, achieving 95% test accuracy which was featured at **2023 CES** in Las Vegas
- Collaborated on creation of RagaAI Platform for computer vision **drift detection** using CNNs and **anomaly detection**, directly contributing to securing $4.7 million in seed funding
- Performed research for Out-of-Distribution (**OOD**) detection and AI stress testing in medical imaging, retail checkout, ADAS, and market research. This helped the company scale and reach out to 8 organizations.
- Designed and deployed an API pipeline with dashboard for interactive visualization and clustering of DNN embeddings using techniques like **t-SNE**, **UMAP** and **PCA**, enabling real-time analysis and interpretation of high-dimensional data.
- Implemented Maximum Mean Discrepancy **(MMD)** and Kolmogorov-Smirnov tests for **drift detection** in image datasets, reducing undetected data drift and enhancing model stability.
- Leveraged **AE**, **VAE**, Variational Auto-Encoding Gaussian Mixture Model **(VAEGMM)** algorithms to identify outliers in high-dimensional datasets, improving anomaly detection accuracy by **40%**

Deep Learning Engineer, Biocliq Technologies - *Bangalore, IN* October 2021 - January 2022

- Implemented model for 3D-segmentation on Brain CT-Scans to detect cancerous tumors based on **UNET-R** architecture improving the DICE score over existing models by 10%.
- Extracted inferences and results of CT scan reports of pneumonia patients from PDF files using camelot to create a dataset for pneumonia categorization and detection during COVID.

Deep Learning Engineer, Uncanny Vision - *Bangalore, IN* June 2019 - July 2021

- Improved verification environment for system-level and intra-module testing of a custom DL framework for an **edge-AI** FPGA device, resulting in better memory utilization and tenfold faster output generation
- Streamlined FPGA device performance by adapting seven major neural network architectures from eight DL frameworks to a custom framework, significantly reducing model sizes by 50% and enhancing operational speed.
- Conducted ongoing research in computer vision and consulted clients on further developing the custom DL framework.
- Streamlined and deployed an annotation tool that **automated annotation** processes, resulting in a 40% increase in throughput and equipping the annotation team with essential tool proficiency
- Engineered cutting-edge **face mask recognition** model with the industry-specific EfficientNet series, customized for lower resolution images to combat COVID challenges, achieving an exceptional 97% accuracy rate
- Trained boom barrier monitoring model with 95% accuracy for smooth operations in automated parking checkout.

PROJECTS

Job Application Enhancement Tool (GitHub - kedardg/job-applications)

- Engineered an AI-powered tool to refine resumes and cover letters by analyzing job descriptions and extracting key qualifications. Integrated multiple LLM APIs (OpenAI, Claude, Google) to generate personalized job application insights.

Deep Learning Model Optimization for Concurrent Data Processing (ASU)

- Developed an advanced deep learning optimization technique that achieved an 80% reduction in model size without compromising performance. Led a team to adapt deep learning architectures for high-throughput applications, enhancing neural network efficiency.

Lightning NeRF Extension with Semantic Information (ASU)

- Re-engineered and enhanced the state-of-the-art Lightning NeRF framework for autonomous driving applications by incorporating semantic information, enabling the model to comprehend and interpret scene components semantically.
- Achieved a 10% improvement in Peak Signal-to-Noise Ratio (PSNR), demonstrating enhanced accuracy and fidelity in scene reconstruction.

AI Stress Testing Framework for Computer Vision (RagaAI)

- Developed an AI stress testing framework for computer vision, employing synthetic data from advanced generative models to simulate complex edge cases for thorough pre-deployment evaluation
- Identified five unique failure scenarios, offering insights for model enhancement and data-driven fine-tuning

ADAS Outlier Detection Project (RagaAI)

- Orchestrated the development of an advanced DNN for ADAS, achieving a 95% accuracy in tracking **data drift** and identifying outliers, with successful deployment on the Ambarella CV22 platform
- Implemented a strategic approach to improve model performance by selectively transmitting key outlier data, streamlining the fine-tuning process as opposed to annotating the entirety of collected data

GPT-based Language Model for Custom Script Generation

- Developed a text generation model using Transformer architecture, multi-head self-attention, and feed-forward networks, inspired by "Attention is all you need" research paper, intending to understand the core principles and implementation of LLMs
- Utilized trained model parameters to generate diverse scripts from varied seed texts, showcasing adaptability and enhancing storytelling depth

Retrieval-Augmented Generation Pipeline (RagaAI)

- Developed a custom-built Retrieval-Augmented Generation (RAG) pipeline, integrating VectorDB for optimized data storage and retrieval with the GPT-4 API for advanced natural language processing capabilities

DL Model Optimization for FPGA (Uncanny Vision)

- Streamlined FPGA device performance by adapting key neural network architectures (ResNet, YOLO, SR-GAN, VGG, Mask-RCNN, MobileNet, SqueezeNet) from six deep learning frameworks (TensorFlow, PyTorch, MXNet, Caffe, Chainer, PaddlePaddle) to the custom framework for an FPGA device

- Halved model sizes and accelerated operations using layer fusion and sparsity optimizations for model compiler

LLM Testing Platform with Advanced Metrics (RagaAI)

- Developed an LLM testing platform incorporating metrics like context precision, answer relevancy, similarity, and faithfulness. Implemented a Retrieval-Augmented Generation (RAG) model using VectorDB and GPT-4 API, establishing the foundation for Raga's LLM testing framework. This platform enables comprehensive evaluation of language models, providing a robust tool for enhancing model accuracy and reliability through targeted metrics.

CERTIFICATIONS

- Deep Learning Specialization, a five-course specialization in Coursera
- Ranked 4th in final exam of High-Performance Computing course conducted by Indian Institute of Technology, Goa
- Robotics: Perception, a course offered by the University of Pennsylvania in Coursera