

SHREE VARAA MANGAI V

shreevaraamangai@gmail.com • [LinkedIn](#) • [GitHub](#) • [Portfolio](#)

WORK AND RESEARCH EXPERIENCE

ML Research | University of Florida - Trustworthy Engineered Autonomy Lab

Florida, US | Sep 2025 - Present

- Generated **15,000+ videos using Wan 2.1 and HunyuanVideo** text-to-video models on NVIDIA B200 GPUs across T2VCompBench and ViBe prompts. Developed a hallucination detector using Qwen3-VL-30B, a vision-language model (VLM), to identify prompt-video mismatches by scoring how well video frames support each prompt word.
- Benchmarked the detector against 4 baselines, achieving the best performance across 6 evaluation metrics for prompt inconsistency detection on T2VCompBench with **Wan 2.1: 64.5% balanced accuracy, 0.739 AUPRC, and 0.687 AUROC**.

ML Engineering Co-op | University of Florida- IPPD | AGIS Inc.

Florida, US | Aug 2025 - Apr 2026

- Built an end-to-end pipeline to convert images and videos into 3D assets for interactive visualization in Unity, integrating SAM2 object segmentation, COLMAP camera pose estimation, and 2D Gaussian Splatting (2DGS). Compared 3DGS and 2DGS reconstruction quality, selecting 2DGS for better mesh exports; **achieved 32.72 dB PSNR and 0.0059 L1 rendering error on NVIDIA B200 GPUs**.
- Benchmarked end-to-end output against Kiri Engine on identical inputs: delivered a Unity-ready rigged asset in **8 min vs. Kiri's 15 min** for a raw PLY export, a **47% latency reduction** with richer output.

Global People Analytics Intern | Ford Motor Company

Tamil Nadu, IN | Aug 2023 - Oct 2023

- Conducted statistical analysis and forecasting on time series forecasting models such as Autoregressive Integrated Moving Average (ARIMA), Vector Autoregression (VAR), and Vector Error Correction Model (VECM) and non-time series models, such as Lasso and Ridge regression, to determine optimal female incumbency.
- Developed ETL pipelines in Python to predict salary costs using workforce datasets containing 1,000+ employee records, improving **model accuracy by 20%** through exploratory data analysis, feature engineering, and data quality improvements.

Software Engineer Intern | Spacescan Ltd.

Ontario, Canada | Sep 2022 - Jan 2023

- Built and maintained responsive, component-based React.js interfaces integrating Swagger-documented REST APIs, reducing code duplication through reusable UI components and robust state management.
- Extended the React Native mobile application with new features, collaborating on frontend architecture and software design decisions to maintain consistent UI and interaction patterns across web and mobile.

PROJECTS

CareMind: Agentic Clinical RAG Assistant | [GitHub-Link](#)

Aug - Sept 2026

- Building an agentic clinical RAG system (FastAPI, LangGraph, Supabase vector search, Redis, NVIDIA NIM, MCP tools) routing queries across document QA, report comparison, and medical education workflows, with hands-on orchestration, context engineering, and tool integration across the full agent stack.
- Developing an eval-driven RAG observability pipeline (PubMedQA-style biomedical QA, Prometheus, Grafana) to measure system quality before scaling tracking Recall@K, Precision@K, MRR, faithfulness, latency, and TTFT; current results show **recall improving from 0.598 to 0.764** and **faithfulness support rate from 0.505 to 0.677**.

Tech Stack: Python, FastAPI, LangGraph, Supabase/pgvector, Redis, NVIDIA NIM, MCP, Prometheus, Grafana, Next.js, TypeScript

Multi-Image Super Resolution for Prostate MRI (CNN, SRGAN, Diffusion Models) | [GitHub-Link](#)

Nov - Dec 2025

- Trained and benchmarked **three architectures**: a 5-block Residual Interpolation CNN, SRGAN (PatchGAN discriminator with adversarial/L1 loss), and a **Fast-DDPM diffusion model** (10-step DDIM, non-uniform β schedule) on 40,532 slice-pairs to reconstruct missing prostate MRI through-plane slices.
- SRGAN led on numerical accuracy (29.045 dB PSNR, 0.850 SSIM vs. CNN's 29.027 dB/0.843), while Fast-DDPM traded **PSNR (27.019 dB, 0.801 SSIM)** for the strongest anatomical continuity, **at 24x the inference latency** (10.5ms vs. 0.4-0.6ms).

Tech Stack: PyTorch, CNN, GAN, Diffusion Models, OpenCV, NumPy

CognitoMap: Automated Question Classification using Bloom's Taxonomy | [GitHub-Link](#)

Feb - Apr 2024

- Built a scalable ETL pipeline extracting and transforming questions from PDFs into MongoDB, achieving **94% accuracy** in question extraction using the Gemini 1.0 Pro model and classified questions into Bloom's Taxonomy levels using Sentence Transformer embeddings with Pinecone vector retrieval.
- Delivered a full-stack dashboard (Flask, React.js, Chart.js) giving educators real-time insights into the cognitive distribution of their assessments across Bloom's Taxonomy levels.

Tech Stack: Flask, React, MongoDB, Pinecone, Chart.js, TypeScript

SKILLS

AI/ML & LLM: PyTorch, TensorFlow, HuggingFace Transformers, LangChain, LangGraph, LlamaIndex, RAG, MCP, CUDA, NVIDIA NIM, Sentence Transformers,

Backend & Data: Python, FastAPI, Flask, SQL, PostgreSQL, Supabase/pgvector, Redis, MongoDB, Pinecone, REST APIs, Swagger, Firebase

Cloud & Deployment: Docker, Git, Linux, CI/CD, AWS Bedrock, Azure, SageMaker, Prometheus, Grafana, Weights & Biases, Gradio, Streamlit

Frontend & Visualization: TypeScript, JavaScript, React.js, Next.js, Chart.js, PowerBI

EDUCATION

Master of Science in Artificial Intelligence Systems | GPA: 3.90 / 4.0

Aug 2024 - May 2026

University of Florida, Gainesville, FL

Bachelor of Technology in Information Technology | GPA: 8.43 / 10

Sep 2020 - May 2024

Loyola-ICAM College of Engineering and Technology, India (affiliated under Anna University)

CERTIFICATIONS

Building with the Claude API: Anthropic (2026) · **Build & Orchestrate Agents with MS Foundry:** Microsoft (2026) · **Fundamentals of Deep Learning:** NVIDIA (2024) · **Fine-Tuning LLMs for Cybersecurity:** LinkedIn Learning (2025) · **Tesla Stock Price Prediction / Google Cloud Fundamentals:** Coursera (Dec 2023 / Aug 2022)