

Shreya Boyane

San Jose, CA | shreyaa.boyane@gmail.com | +1 (774) 519-9905 | linkedin.com/in/shreya-boyane | github.com/Sonica-B | shreyaboyane.dev

SUMMARY

Machine Learning Engineer with 2+ years of production DevOps experience. Algorverse Research Fellow with a Latent-CoT mechanistic interpretability paper accepted at ICML 2026 conference. Recognized in the Google Cloud x NVIDIA developer community (Golden Ticket Honorable Mention, NVIDIA GTC 2026) for inference optimization engagement across three completed learning pathways.

EDUCATION

Worcester Polytechnic Institute (WPI), *Master of Science (M.S.) in Artificial Intelligence*, GPA: 3.9/4.0

Aug 2024 - May 2026

- Achievement:** Data Science GQP Most valuable player (MVP), Honorable Mention, Google Cloud x NVIDIA Golden Ticket Contest, NVIDIA GTC 2026

PES Modern College of Engineering (PESMCOE), *Bachelor of Engineering in Information Technology*, GPA: 8.9/10

Jun 2017 - May 2022

- Achievement:** Honours in Artificial Intelligence and Machine Learning

RESEARCH EXPERIENCE

Latent-CoT Trajectories as Dynamic Systems

Feb 2026 - May 2026

AI Research Fellow | Algorverse AI Research Fellowship | ICML 2026

- Published mechanistic-interpretability research at ICML 2026 modeling latent chain-of-thought (CoT) reasoning in LLMs as discrete-time dynamical systems, applied Dynamic Mode Decomposition (DMD), PHATE, UMAP, t-SNE, and PCA to 768-dimensional latent trajectories across 8,792 GSM8K and 500 MATH samples.
- Reproduced two latent-reasoning training methods end to end on GPT-2, CODI self-distillation (3-loss objective with layer-wise knowledge distillation) and the COCONUT 4-stage latent curriculum, as LoRA fine-tunes using PyTorch, Hugging Face Transformers, and PEFT.
- Discovered architectural signatures distinguishing the two methods, DMD unstable mode fraction 41% vs 85%, arc-length sign flip on correctness, and a direction-consistency hinge at $t=3$ in COCONUT vs a near-period-2 oscillator in CODI, signatures preserved across GSM8K and MATH datasets.
- Engineered the reliability layer: multi-stage Bash + Python orchestrator with CUDA health-polling stabilizing 14+ hour training runs, HDF5 resume-from-cache (~90 min saved per recovery), a GPU-resident Kraskov KSG-1 mutual-information estimator, a safetensors-level checkpoint repair with byte-identical embedding preservation, and a verification harness passing 28/28 reproducible numeric claims.

FairPrep: Demographic Fairness for Table Union Search

Jan 2025 - May 2026

Research Assistant | Worcester Polytechnic Institute | Advisor: Dr. Roei Shraga | BSF Grant #2024022

- First adaptation of the FairEM framework (Shahbazi et al., VLDB 2023) from entity matching to annotator-level fairness in Table Union Search, a foundational primitive in modern data lakes; built 9 source modules across 5 experimental phases on a 480-row dataset.
- Designed a complete statistical methodology stack, 10 FairEM confusion-matrix metrics, permutation testing (10K shuffles), bootstrap confidence intervals (10K stratified resamples), Fisher's exact tests, Benjamini-Hochberg FDR correction, and dual-track logistic regression.
- Surfaced statistically significant bias against non-native English annotators on hard questions (25% vs 44% accuracy, Fisher $p=0.031$; triple-intersection effect persisting at $p_{BH}=0.037$ after FDR), robust across 5 of 5 orthogonal cleaning strategies.

WORK EXPERIENCE

AI Trainer Fellow (Machine Learning & Computer Vision) | Handshake AI | Remote (Part-time)

Apr 2026 - May 2026

- Selected as a domain-expert AI Trainer Fellow improving frontier Large Language Models (LLMs) and Vision-Language Models (VLMs) for leading AI labs, spanning prompt development, response evaluation, and benchmark authoring across software-engineering and visual-reasoning domains.
- Evaluated 40+ structured tasks, each a side-by-side assessment of two distinct frontier models (80+ LLM/VLM outputs graded for correctness, code quality, technical depth, and reasoning soundness), delivering rubric-based expert feedback used to improve and benchmark frontier models.
- Authored challenging visual and compositional reasoning benchmark items with verified ground-truth answers over real-world charts, diagrams, infographics, and scientific figures, benchmarking multimodal model performance on chart/diagram visual question answering (VQA).
- Developed domain-specific prompts for frontier LLMs across software engineering, machine learning, and agentic-reasoning topics, and conducted independent research to inform prompt design and evaluation-rubric construction.

Machine Learning Engineer | Industry-Sponsored Graduate Qualifying Project | Crossing x WPI | Worcester, MA

Aug 2025 - Dec 2025

- Owned client-facing delivery for an industry sponsor, ran requirements-discovery sessions with company engineers, translated ambiguous legal-workflow needs into acceptance tests, led weekly stakeholder demos through end-to-end integration and handoff.
- Built a production multi-agent LLM document-intelligence system for H-1B immigration workflows on AWS (Amazon Bedrock, S3 ingest, Lambda, event driven triggers), extracting structured fields into schema-validated canonical JSON and generating unified quality-control reports for Form I-129 evidence.
- Cut petition processing from 3-4 days to ~10 minutes at 95-99% structural accuracy and 99% missing-document detection with 85% repeatable automated checks, reducing manual verification 70-90% and cost 50-70%, freeing 2,000+ hours for higher-value legal work.

Machine Learning Student Researcher | Worcester Polytechnic Institute | Worcester, MA

Jan 2025 - Aug 2025

- Built confidence-aware clinical triage orchestration in Python, PyTorch, and scikit-learn with human-in-the-loop review, raising triage precision 20% and cutting false escalations 40%.
- Engineered a dual-channel decision flow separating high-risk escalation from routine guidance, reducing LLM hallucinations 50% and improving patient provider alignment 2.3x.
- Implemented difficulty-weighted training and GPT-style calibration scoring, errors down 30%, reliability up 35%, early critical-case detection up 60%, non-critical misclassification down 70%.

Lead DevOps Engineer | Persistent Systems | Pune, India

Jun 2022 - Jul 2024

- Accelerated software delivery 95% by building CI/CD pipelines with Azure DevOps, Docker, and Terraform with integrated automated testing across 20+ projects, earning Top Talent 2024 and Bravo individual awards.
- Standardized VM provisioning and cluster setup with Ansible and Terraform Infrastructure-as-Code (IaC), cutting manual configuration 70% and infrastructure provisioning time 40%, including AWS S3/Lambda codified in Terraform.
- Improved system uptime 25% by deploying Prometheus and Grafana observability dashboards monitoring latency, queue time, and build health in real time.
- Shortened integration cycles 70% by coordinating release tooling across frontend and platform teams, improving delivery cadence for multi-team programs.

PROJECTS

MADWE: Multi-Agent Diffusion World Engine

Mar 2025 - Aug 2025

- Built a multi-agent generative runtime with diffusion models and LoRA fine-tuning, generating 120 game assets/minute and cutting manual design time 73%.
- Optimized GPU inference via ONNX & TensorRT kernel fusion, cut p95 latency 200ms to 95ms, lifted FPS 60 to 120, reduced binary size 38%.
- Pushed efficiency further with mixed precision, INT8 quantization, CUDA graphs, 1.8x throughput, 24% less memory, FID 12.3, CLIP similarity 0.89 under 2GB VRAM.

Einstein Vision: 3D Visualization for Autonomous Vehicles

Mar 2025 - Apr 2025

- Implemented a production-style perception stack integrating Depth Anything (monocular depth), Mask R-CNN and Detic (detection/segmentation), and OpenPifPaf (pose estimation) with the Blender Python API, improving 3D placement accuracy 50% and scene realism 75% across 13 driving sequences.
- Synthesized LiDAR-like point clouds and lane contours for ground truth, streamed masks and poses at 30 FPS, and deployed detectors via TorchScript to ONNX to TensorRT FP16 with layer fusion, 1.5x multi-object throughput, 35% lower p95 latency (validated with NVIDIA Nsight).

TumorLytx Unsupervised Brain Tumor Detection and Segmentation

Oct 2024 - Dec 2024

- Built a scalable medical-imaging pipeline using CycleGAN pairing and diffusion refinement in PyTorch and TensorFlow across 65,072 MRI images, achieving more than 90% segmentation accuracy without manual labels.
- Automated labeling with OpenCV morphological operations and entropy-based filtering (+15% precision); packaged TorchScript batched mixed-precision inference for 1.6x images/sec, profiled with Nsight Compute.

SKILLS

- **Machine Learning, AI & Data Science:** PyTorch, PyTorch Lightning, TensorFlow, Keras, scikit-learn, Hugging Face Transformers, PEFT/LoRA fine-tuning, LLMs, VLMs, LangChain, LangGraph, multi-agent systems, prompt engineering, model evaluation & benchmarking, diffusion models, CycleGAN, NeRF, reinforcement learning, OpenCV, object detection & segmentation (Mask R-CNN, Detic), depth estimation, pose estimation, Structure-from-Motion (SfM), visual question answering (VQA), chart/document understanding, image processing, CNNs
- **Inference & GPU Optimization:** CUDA, TensorRT, ONNX, TorchScript, INT8 quantization, mixed precision, CUDA graphs, kernel fusion, Nsight profiling
- **MLOps, Cloud & DevOps:** AWS (Bedrock, S3, Lambda, Step Functions, EventBridge, CloudWatch), Azure DevOps, Azure ML, Docker, Terraform, Ansible, CI/CD pipelines, MLflow, Prometheus, Grafana, Infrastructure-as-Code, Git, PostgreSQL, JSON Schema
- **Programming Languages:** Python, C++, C#, Java, SQL, R, MATLAB, Scala, Bash, PowerShell, React, Angular