

ARMAAN FRANCISCO SHIVPURI

AI / ML Engineer — RL & Agent Environments • Evaluation Infrastructure • Scientific ML • Applied Mathematics

armaan@shivpuri.com • (310) 906-9771 • Troy, NY • linkedin.com/in/armaan-shivpuri • github.com/ARMINSHIVES • shivpuri.com

SUMMARY

Applied-math-trained ML engineer who builds the measurement layer around models: outcome-based evaluation harnesses, held-out benchmarking pipelines, and agent tool infrastructure. Research spans surrogate modeling for computational science (GNNs/autoencoders), LLM fine-tuning for scientific reasoning, and reinforcement learning. Seeking to build RL environments where the reward is a measurement of the world.

TECHNICAL SKILLS

ML & RL: PyTorch, TensorFlow, Reinforcement Learning (REINFORCE, policy learning, reward design), Graph Neural Networks, Autoencoders, Surrogate Models, Fine-Tuning (PEFT/LoRA, TRL), Hyperparameter Optimization, scikit-learn

Agents & LLMs: Agent Workflows, Tool Calling, MCP (Model Context Protocol), Multi-Step Reasoning, Model Routing, Hugging Face Transformers, RAG, vLLM/Ollama

Evaluation & Infra: Benchmark Harness Design, Held-Out / OOD Evaluation, Reproducible Experiment Pipelines, Docker, FastAPI, REST APIs, Linux, Git, Milvus, AWS (Lambda, S3)

Scientific Computing: Numerical Methods, Differential Equations, Linear Algebra, Probability, Monte Carlo Simulation, CFD (OpenFOAM, mesh adaptation), NumPy/SciPy

Languages: Python, C/C++, Java, JavaScript/TypeScript, R, MATLAB

EXPERIENCE

AI Systems Engineer | RPI Tetherless World Constellation

Jan 2026 – Present

- Built an automated evaluation harness benchmarking ranking accuracy across embedding models and similarity metrics on held-out query sets, emitting versioned, reproducible scorecards for every corpus revision.
- Shipped an MCP tool server and FastAPI service over one shared retrieval backend, exposing the same swappable interface to autonomous agents and conventional applications.
- Architected a pluggable vector-store abstraction spanning four similarity metrics with transparent failover between a Milvus cluster and an in-process NumPy backend — zero query failures during outages.
- Designed a content-addressed caching pipeline (SHA-256 manifests) that incrementally re-embeds a 550+ item, 28-language corpus, eliminating redundant embedding-API calls.

LLM Engineer | Computational Scientific ML Lab (Pan) / FoamGPT, RPI

Mar – Sep 2025

- Fine-tuned Gemma 7B and Qwen 7B with parameter-efficient training (PEFT/LoRA) and hyperparameter optimization on OpenFOAM CFD corpora, improving scientific-reasoning benchmark performance 15%; the project produced a NeurIPS 2025 ML4PS workshop paper.
- Built cross-architecture evaluation infrastructure measuring accuracy, robustness, latency, and failure modes across competing LLMs to drive deployment decisions.
- Cut production inference latency below 800 ms in deployed voice-AI agents by optimizing streaming inference and tool-use orchestration.

Computational Science Research Assistant | CREATE-FT Capstone SDK, RPI

Sep 2024 – Mar 2025

- Developed C++ mesh-adaptation software integrating the CREATE-FT Capstone SDK with Raven CFD, applying differential equations and numerical methods to automate refinement near critical flow features in capsule-reentry simulations.

ML Research Engineer — Surrogate Modeling | DMREF Materials Research, RPI

Aug 2023 – Sep 2024

- Trained deep autoencoders and graph neural networks as surrogate models predicting material properties for aircraft-engine components under extreme conditions; GCN reached $R^2 = 0.971$ under 5-fold cross-validation.
- Established reproducible benchmarking and held-out evaluation pipelines that increased the stability and validity of experimental results across model configurations.
- Automated training, validation, and performance-analysis workflows, accelerating iteration speed and surfacing latent representations that improved downstream prediction. 2024 Summer Undergraduate Research Award.

SELECTED PROJECTS

- **Poker Equity Engine (RL + CV)** — Monte Carlo equity solver with a REINFORCE-trained reinforcement-learning agent advising fold/call/raise, fed by a fine-tuned YOLOv8 detector reading physical cards from a live camera. Python, PyTorch, OpenCV.
- **Jarvis — Local-First Multi-Agent System** — Routes requests through a local model (Ollama/vLLM) before falling back to Claude, orchestrating ~40 tool integrations across sub-agents, with CV modules using YOLOv8n + ByteTrack and ArcFace embeddings.
- **POEM Ontology Retrieval Pipeline** — Merged the embedding/retrieval pipeline (RDF/SPARQL, Dockerized Jena Fuseki + TDB2) and expanded the evaluation harness from 12 to 50 queries, fixing a retrieval-quality defect in the unscoped search path.

EDUCATION

Rensselaer Polytechnic Institute — M.S., Applied Mathematics

Expected May 2027

Rensselaer Polytechnic Institute — B.S., Computer Science & Mathematics

May 2026