Aadi Sharma

+91 9312321893 | aadisharma2808@gmail.com | aadisharma.xyz | linkedin.com/in/aadisharma11 | github.com/AadiSharma1

Summary

Machine Learning & Systems Engineer (**B.Tech AI/ML, '28**) specializing in **distributed LLM post-training** (**RLVR**, **DeepSpeed ZeRO-2**), **privacy-preserving ML**, medical imaging, and **edge-AI systems**. Author across **4 research manuscripts** spanning LLM safety alignment, decentralized edge swarms, neuroimaging, and encrypted deep learning, experienced in scaling **multi-GPU training clusters** and shipping production full-stack software.

Education

SRM Institute of Science and Technology **Chennai, India**
B.Tech in Computer Science and Engineering (AI/ML Specialization) | ***Full Merit Scholarship*** *Aug 2024 – May 2028*

- **Relevant Coursework:** Data Structures & Algorithms, Object-Oriented Programming, Operating Systems, DBMS, Computer Networks, Linear Algebra, Probability & Statistics.

Technical Skills

Languages: Python, C++, Java, SQL, TypeScript, JavaScript, C
ML & Deep Learning: PyTorch, DeepSpeed (ZeRO-2), RLHF/RLVR, vLLM, Transformers, Hugging Face, scikit-learn, OpenCV, TenSEAL (CKKS), GNNs
Systems, Backend & Infra: CUDA, Distributed Data Parallel (DDP), FastAPI, Flask, Next.js, React, REST APIs, Docker, PostgreSQL, Firebase, Linux, Git, CI/CD
Core Concepts: LLM Post-Training, Activation Checkpointing, Quantization (NF4/AWQ), Homomorphic Encryption, Federated Learning, Semi-Supervised Learning

Software & ML Engineering Experience

Carnegie Mellon University **Remote**
Research Collaborator, LLM Safety & Reinforcement Learning (Advisor: Dr. Chhavi Yadav) *Aug 2026 – Present*

- Co-authored *"If-Taught Reasoners Forget How to be Safe"* (Under Review), quantifying catastrophic safety drift across **10-round continual LLM training** trajectories and engineering a dual-mechanism mitigation using a **safety-aware verifier** and **alignment replay buffer**.
- Evaluated **Llama-3.1-8B** and **Qwen2.5-7B** across mathematical reasoning and safety benchmarks, significantly reducing aggregate **Attack Success Rate (ASR)** while preserving peak reasoning performance in self-improving models.
- Architecting a **multi-turn RL framework** to quantify dynamic social engineering and alignment faking, evaluating how policies mirror user personas to exploit **RL from Verifiable Rewards (RLVR)** against a secondary **LLM User Simulator**.
- Deploying full fine-tuning pipelines for 7B/8B models on a **2x NVIDIA A100 (40GB) DGX** node via **ZeRO-2 sharding**, **CPU-offloaded fp32 AdamW**, and per-block **activation checkpointing** (**32,768-token context**), alongside detection-budget evaluations to isolate **stealth-gap** ASR spikes.

Carnegie Mellon University & Harvard Medical School **Remote**
Research Intern, Medical Image Analysis *Jan 2026 – Present*

- Engineered a cascaded **3D Siamese U-Net** pipeline on an **8x NVIDIA RTX 3090** cluster aligning longitudinal brain MRI scans for a **515-patient ADNI cohort (70 GB)**; resolved cross-patient gradient leakage via **Smooth-L1 affine isolation**, achieving **0.9971 INTRA Elastic Dice** and holding INTER deformation below **1e-4**.
- Architected a physics-informed segmentation framework (**1 mm to 0.4 mm isotropic lattice**) featuring **Entropy-Gated Cross-Pseudo Supervision (EG-CPS)** and a custom **SyncedElasticAugmenter** (**50/50 labeled/unlabeled batches**) to decouple boundary mixing from epistemic uncertainty.
- Formulated an **Asymmetric Temporal Consistency loss** modeling diagnosis-specific cortical thinning (**4e-3/yr for AD**) and enforced laminar ordering via **Betti matching**; sustained multi-day training on CMU's LTI Babel cluster via **SSH ProxyJump** containers under **20 GB CUDA quotas**.

University of Texas at San Antonio & University of the Arts London **Remote**
Research Collaborator, Decentralized Edge-AI Systems (First Author) *May 2026 – Present*

- Built a decentralized semantic synchronization architecture for **50-node edge swarms** in **DDIL networks**, serving **Llama-3-8B** across **8 vLLM A100 endpoints** to compress 201-byte raw sensor telemetry into **110-byte invariant tokens** (**1.8x reduction**).

- Achieved **14.1x lower bandwidth and energy** than Epidemic routing using a sender-side **Invariant Preservation Score (IPS) gate** and **EMA link-memory 2-hop relaying** over a Gilbert-Elliott channel, sustaining a **99.6% Decision Preservation Rate** under **80% packet loss**.

Indian Institute of Technology (BHU) **Varanasi, India**
Research Intern, Privacy-Preserving Machine Learning *Nov 2025 – May 2026*

- Built a Python execution pipeline using **TenSEAL** to run neural inference directly on **CKKS-encrypted data** and implemented **Secure Aggregation protocols** for multi-party federated model updates without raw client data exposure.
- Developed **HEAT (Token-Condensed Homomorphic Attention)** for **Vision Transformers**, pruning encrypted token sequences during self-attention to drastically cut cryptographic memory overhead and inference latency.

SRM Institute of Science and Technology **Chennai, India**
Undergraduate Researcher, Graph Neural Networks (First Author) *Aug 2024 – Present*

- Built the **PyTorch** training, **UMAP/HDBSCAN** clustering, and evaluation pipeline for **DST-GAT** (Dynamic Spatiotemporal Graph Attention Networks) to identify neurobiological subtypes in autism from fMRI neuroimaging data, achieving **0.74 AUC**.

Publications

- [1] A. Sharma et al., *"If-Taught Reasoners Forget How to be Safe: Safety Drift in Continual LLM Training,"* **Under Review (CMU, 2026)**.
- [2] A. Sharma et al., *"Task-Oriented Decentralized Semantic Synchronization for Swarm Resilience in Extreme DDIL Environments,"* **InCIS 2027 (First Author, Under Review)**.
- [3] A. Sharma et al., *"Discovering Neurobiological Subtypes in Autism via Dynamic Spatiotemporal Graph Attention Networks,"* **HICSS 2026 (First Author, ESCI-indexed)**.
- [4] A. Sharma et al., *"HEAT: Token-Condensed Homomorphic Attention for Privacy-Preserving Vision Transformers,"* **Under Review (IIT BHU, 2026)**.

Engineering Projects

SkillSync: AI Career Intelligence Platform | *Python, Flask, Next.js, TypeScript, scikit-learn, Vercel* *Mar 2026*

- Built a full-stack AI career platform featuring **ML regression models** for salary prediction and a **time-series demand forecasting** simulator to evaluate multi-year skill ROI across tech roles.
- Engineered a low-latency **Flask REST API** backend for real-time model inference and deployed a responsive **Next.js / TypeScript** frontend on **Vercel** with automated CI/CD workflows.

CipherTrain: Encrypted ML Training Framework | *Python, PyTorch, TenSEAL, Cryptography* *Mar 2026*

- Engineered a simulation framework modeling **CKKS homomorphic encryption** noise accumulation during neural network forward passes without incurring full cryptographic execution overhead.
- Implemented modular **PyTorch** layers for **learnable polynomial activations** and automated **bootstrapping gates**, accelerating rapid prototyping for privacy-preserving deep learning.

zk-SafeGuard: Verifiable Privacy-Preserving ML Backend | *Python, PyTorch, ZK-SNARKs, NF4 QuantizationJan 2026*

- Built a verifiable cryptographic inference backend combining **Zero-Knowledge Proofs (ZK-SNARKs)** and **Homomorphic Encryption** to guarantee model execution integrity on private client inputs.
- Optimized GPU memory footprints using **4-bit NormalFloat (NF4) quantization**, sustaining **12+ tokens/sec** inference throughput on **NVIDIA T4 GPUs** despite verification overhead.

NeuraTwin: Digital Twin Web App | *React, Firebase, Python, REST APIs* *Feb 2026*

- Developed a full-stack **digital-twin web application** for longitudinal wellness and behavioral state modeling, integrating an interactive conversational AI agent (**AURA**).
- Architected a real-time state synchronization and authentication backend on **Firebase**, supporting low-latency telemetry ingestion and persistent multi-session user tracking.

Leadership & Achievements

Cintel Association (Campus AI Society) **Chennai, India**
Head of Research & Development | ***Full Merit Scholarship Recipient (SRMIST)*** *Aug 2025 – Aug 2026*

- Direct the technical roadmap of the campus AI society, mentoring **15+ student developers** to architect, test, and deploy end-to-end full-stack machine learning applications.