

HARSHIT TIMMANAGOUDAR

+1 (858) 405-1729 | htimmanagoudar@ucsd.edu | San Diego, CA, USA | linkedin.com/in/harshit-timmanagoudar-829b43212/ | github.com/harshitster | harshitst.netlify.app/

EDUCATION

University of California - San Diego

Master's, Computer Science

September 2024 - June 2026

GPA: 3.58

- Coursework: Data Systems for Machine Learning, Generative AI, Large Model Reasoning, Distributed Systems, Operating Systems, Database Principles.

PES University

Bachelor's, Computer Science

December 2020 - May 2024

GPA: 3.81

- Coursework: Deep Learning, Statistics for Data Science, Computer Vision, Linear Algebra, Data Structures, Cloud Computing, Software Engineering.

PROFESSIONAL EXPERIENCE

STABLE Lab, UC San Diego

Graduate ML Systems Researcher

San Diego, CA, USA

May 2025 - Present

- Architected pipeline-parallelized LLM inference with CXL shared memory and novel KV cache across GPU nodes, achieving constant 30 μ s decode latency and 30K tokens/sec throughput, outperforming HuggingFace's DynamicCache by 300-1600x on long-context inference.
- Enabled variable sequence length scaling with minimal fragmentation through block-based KV cache achieving O(1) complexity and robust performance, eliminating 20x slowdown with 16x lower memory overhead versus HuggingFace's KV cache implementations.
- Eliminated inter-GPU communication bottleneck through CXL shared memory architecture, enabling zero-copy tensor activations access across pipeline stages and removing network synchronization overhead in multi-node inference system.

Peptris Technologies

AI Research Intern

Bengaluru, KA, India

January 2024 - June 2024

- Developed end-to-end Graph Attention Variational Autoencoder pipeline with Gumbel-Sinkhorn operator for permutation-invariant 3D molecular conformation generation, training on 150M+ structures to produce 18K+ diverse conformations with 14% improved binding pose accuracy.
- Implemented Latent Diffusion Model on VAE latent space using score-matching with Langevin dynamics, achieving 38% inference speedup (2.4s to 1.5s per molecule) for high-throughput processing on multi-cloud infrastructure (GCS storage, AWS compute).
- Evaluated 6 graph architectures (GCN, MuchGCN, PIGVAE) and diffusion techniques (score matching with Langevin dynamics, Riemannian manifolds) under CTO/Chief Data Scientist supervision, establishing model selection framework for production deployment.

Peptris Technologies

Machine Learning Intern

Bengaluru, KA, India

June 2023 - August 2023

- Engineered Large Language Model inference pipeline for protein sequence generation, producing 2,400+ novel protein candidates achieving 87% structural validity for downstream screening in drug discovery process.
- Designed 3D U-Net model for voxel-based protein analysis, implementing 3D convolutional architecture to detect and segment binding sites in protein structures, enabling identification for downstream molecular docking in drug discovery workflows.

PROJECTS & PUBLICATIONS

Temporal Vector Database

August 2025 - Present

- Implementing versioned vector database with delta-based storage for time-series embeddings, using base snapshots plus incremental deltas for storage optimization while maintaining query performance.

StreamGrid - Video Storage System - [Link to project](#)

April 2025 - June 2025

- Built scalable video platform with RAFT consensus (etcd), consistent hashing for load balancing, and FFmpeg MP4-to-DASH transcoding via gRPC microservices with zero-downtime cluster rebalancing.

DistRAG - Distributed RAG Database - [Link to project](#)

June 2024 - January 2025

- Designed multi-node RAG platform with PostgreSQL Citus sharding, LlamaIndex/Gemini for NL-to-SQL via ChromaDB vector storage, Redis semantic caching, and automated WAL recovery

Enhancement of Malware Detection System using Mal-cGAN - [Link to project](#)

First Author

Bengaluru, KA, India

March 2023 - July 2023

- Developed conditional GAN architecture to enhance malware detection robustness through synthetic data generation, achieving 10.14% accuracy improvement (83.47% to 93.61%) and published in ISI, Springer Nature 2023.

Edge Map Extraction of High Resolution Facial Images - [Link to project](#)

First Author

Singapore

February 2023 - June 2023

- Formulated computer vision edge detection algorithm for facial images using Gaussian smoothing, directional gradients, dual-threshold detection, and Zhang-Suen thinning, outperforming Canny and HED algorithms on 10,000+ image validation set. Published at ISBM, Springer Nature 2023.

SKILLS

Languages & Frameworks: Python, C/C++, Go, Rust, CUDA, Triton, SQL, Bash, PyTorch, Tensorflow, JAX, Ray, DeepSpeed

Systems & Tools: Kubernetes, Docker, gRPC, Redis, AWS, GCP, Git, GitHub, Protocol Buffers, NCCL, etcd, Postgres, Pandas, Numpy, Scikit-learn, Modin

ML/AI: LLM, Transformers, AI Agents, Diffusion Models, CNNs, RNNs, VAEs, GANs, Graph Neural Networks