Naveenraj Kamalakannan

Brooklyn, NY · +1 914-490-3063 · naveenraj.k@nyu.edu
itsnav.com · github.com/therealnaveenkamal · linkedin.com/in/navz/

Education

New York University (NYU) — Brooklyn, NY
M.S. in Computer Engineering | ***GPA: 4.00 / 4.00*** — *Sep 2024 – May 2026 (Expected)*

- Coursework: High-Performance ML, Deep Learning, Computer Vision, Reinforcement Learning, Distributed Systems.
- Research interests: LLM inference & training systems, multimodal reasoning, human motion understanding, evaluation & interpretability.

Vellore Institute of Technology (VIT) — Vellore, India
B.Tech. in Electronics and Communication Engineering | *GPA: 8.79 / 10.0* — *Jul 2018 – May 2022*

Open Source Engineering (LLM Systems)

vLLM — Python, CUDA, PyTorch
Contributor | *PR #25103, PR #29769* — *Remote*

- Refactored **Multi-Head Latent Attention (MLA)** to decouple prefill/decode paths from a unified custom op, enabling `torch.compile` fusion and piecewise CUDA Graph capture, reducing Python overhead and making MLA more modular and easier to experiment with.

NVIDIA TensorRT-LLM — C++, Python
Contributor | *PR #7490* — *Remote*

- Integrated **Tree-of-Thought (ToT)** and **MCTS** controllers into the AutoDeploy scaffolding framework to enable multi-step reasoning flows and experimentation with search-based inference strategies.

Microsoft DeepSpeed — Python, Distributed Systems
Contributor | *PR #7302* — *Remote*

- Fixed a critical **ZeRO-3 CPU-offload gradient clipping bug**, ensuring global gradient norms correctly reflect clipped gradients during offload scenarios and improving training stability for large models.
- Collaborating with Olatunji Ruwase on PyTorch Core (Issue #158187) to implement Zip serialization support for **DeepNVMe**, improving I/O throughput and compatibility for NVMe-offloaded checkpoints.

Snowflake ArcticInference — Python, vLLM Plugin
Contributor | *PR #124* — *Remote*

- Integrated **FlashInfer** backend support into SwiftKV, optimizing high-throughput KV-cache-aware decoding and enabling more efficient large-scale LLM serving within the ArcticInference stack.

Research Experience

NYU Center for Data Science & NYU Langone — New York, NY
Research Assistant (Advisors: Prof. C. Fernandez-Granda, Prof. H. Schambra) — *Feb 2025 – Present*

- **Video Action Pipeline for Stroke Rehabilitation.** Studied whether current VLMs can recover clinically meaningful upper-limb motion primitives from rehab videos. Benchmarked SOTA VLMs (InternVL3, NVILA, LLaVa-OneVision) and engineered a pose-refined prompting pipeline that integrates YOLOv11 pose tracks with VLM context to extract sub-second motion primitives, achieving ∼67.75 Edit Score (ES) on structured upper-limb rehab tasks and exposing systematic failure modes on subtle hand-object interactions.
- **Fine-Grained Action Understanding w/ SAM 2 (WIP).** Building a hierarchical pipeline where **SAM 2** produces hand/object tokens encoding frame-level kinematics and contact maps, and **Qwen2.5-VL** operates over these token sequences for fast clinical QA,action detection and differencing on **GigaHands** and **HUMOTO** to model low-level motor signals and sub-second primitives.

Professional Experience

Snowflake AI Research Seattle, WA (Remote)
Incoming AI Research Intern (Collaborators: Aurick Qiao, Jeff Rasley, Olatunji Ruwase) *Jan 2026 – May 2026*

- Will work on open-source projects **ArcticInference** and **ArcticTraining**, focusing on large-scale model optimization, memory-efficient serving, and system efficiency for frontier LLM workloads.

J.P. Morgan (Asset & Wealth Management) New York, NY
AI & Data Science Associate Intern *Jun 2025 – Aug 2025*

- Architected an agentic platform using multi-stage retrieval pipelines (RAG), context engineering, and rerankers to reduce latency and improve relevance in financial LLM tools.
- Orchestrated multiple AI agents with advanced reasoning capabilities and a shared memory layer, integrated via **MCP servers** into production workflows for complex investment use cases.

Zeeco Middle East Dammam, Saudi Arabia
Control Engineer *Jul 2022 – Aug 2023*

- Engineered an anomaly detection system for circuit designs using neural networks, reducing manual verification effort by ∼**40%**, and designed industrial control strategies (PLC/DCS) with predictive models for Pressure, Temperature, and Flow optimization, improving efficiency by ∼**30%**.

Selected Publications

Li, V., **Kamalakannan, N.**, et al. "The Potential and Limitations of Vision-Language Models for Human Motion Understanding." *arXiv preprint arXiv:2511.17727*, 2025. [Link]

Kamalakannan, N., et al. "Exponential Pixelating Integral Transform with Dual Fractal Features for Enhanced Chest X-Ray Abnormality Detection." *Computers in Biology and Medicine*, Vol. 182, 2024. [Link]

Technical Skills

Languages: Python, C++, CUDA, C, Bash, SQL
ML Systems: vLLM, DeepSpeed, TensorRT-LLM, PyTorch Distributed (DDP/FSDP), FlashInfer, KV-cache optimization
Model Architectures: Transformers, Vision–Language Models (VLMs), Mixture of Experts (MoE), SAM-2
Concepts: Deep Learning, Optimization & Evaluation of LLM/VLM Systems, Mechanistic Interpretability, Reinforcement Learning, High-Performance Computing
Tools: Docker, Kubernetes, MLflow, Git, Weights & Biases, Linux