Naeem Khoshnevis

naeemkhoshnevis.com nkhshnvs naeemkh Cambridge, MA

Summary

Senior ML Research Engineer at Kempner Institute at Harvard University with expertise in large-scale ML systems, GPU computing, and AI infrastructure. Proven track record designing scalable HPC and AI solutions, including co-architecting Harvard's Kempner AI Cluster (TOP500, 2024) and leading NeuroAI research codebases for scalable RL-based brain simulation. Developed globally adopted open-source packages (50K+ downloads), advancing causal inference and health/climate studies. Skilled in distributed training (DDP, FSDP, TP) on multi-GPU clusters, GPU profiling, and PyTorch-based ML solutions. Experienced in mentoring researchers, architecting reproducible ML frameworks, and collaborating with industry leaders (NVIDIA, WEKA, VAST) to advance AI scaling.

Professional Experience

Harvard University, Kempner Institute **2024 – Present**

Senior ML Research Engineer *Cambridge, MA*

- Co-architected the Kempner AI Cluster (TOP500, 2024) with non-blocking Infiniband design, optimized workflows, and orchestrated an AI/HPC benchmarking tests for GPU compute, memory, and communication performance.
- Refactored the OLMo codebase into minOLMo, a minimalist LLM research framework enabling single-GPU experimentation and reproducible transformer research for Kempner researchers.
- Led ML engineering efforts on large-scale NeuroAI projects, including a modular RL-based brain simulation framework integrating PyTorch, Gym, and MiniWorld; mentored junior engineers on ML system design.
- Contributed to development of the Transformer Research Codebase (TMRC), a reproducible and researcher-friendly platform for foundational ML experiments.
- Implemented a spatial decomposition approach in a JAX-based simulation project to redistribute computational workloads efficiently across multiple GPU devices, improving scalability and performance.
- Conducted ML research on consistency models, focusing on scalability and reducing training time for one-step noise-to-data mappings (work in progress).
- Tailored benchmarking tools to measure intra- and inter-node GPU communication, requiring significant setup and job customization for the Kempner AI Cluster; collaborated with NVIDIA solution architects and AI storage vendors (WEKA, VAST) to optimize cluster performance.
- Organized and taught the institute's first Compute Workshop on AI at scale, authored the Kempner Computing Handbook, and consulted researchers on distributed ML training (DDP, FSDP, TP) and profiling.

Harvard University **2021 – 2024**

Senior Research Engineer *Cambridge, MA*

- Engineered scalable HPC and cloud-based solutions for large-scale health and climate data, enabling secure CMS data analysis on AWS and Harvard's largest compute clusters.
- Developed and maintained open-source statistical learning packages (e.g., `CausalGPS`, `GPCERF`, and `CRE`) with 50K+ global downloads, advancing causal inference research and powering ArcGIS's commercial tools.
- Applied advanced scientific software engineering, algorithmic development, and statistical learning methods to build reproducible research software for diverse scientific domains.
- Consulted and trained 50+ researchers across life sciences, medical, and engineering fields through workshops, office hours, and troubleshooting sessions on HPC, scientific programming, and big data workflows; additionally taught courses in R/Python, HPC, and statistical learning.
- Mentored junior engineers and scientists on best practices in reproducible ML research software and scalable data systems.
- Ported TensorFlow 1.0 to TensorFlow 2.0 to support Harvard's TinyML course and consulted on ML deployment for Harvard edX students.
- Contributed to Machine Learning Systems (MIT Press, forthcoming), in collaboration with Harvard faculty, providing a comprehensive reference for edge devices and TinyML applications.

Projects

Big Little Brain (BLB): Modular NeuroAI Framework | *PyTorch, RL, HPC, GPU Scaling* **2024 – Present**

- Engineering lead on the Big Little Brain (BLB) project, A simulation framework for modeling a virtual mouse equipped with a brain model that interacts with its environment.
- Designed modular and reproducible codebase architecture supporting reinforcement learning agents, neural modules, and environment orchestration, enabling flexible experimentation and rapid iteration.

- Rigourously investigated corner cases in RL agent–environment interaction (terminated and truncated episodes), and provided a design to reduce the likelihood of future bugs for developers.
- Mentored junior engineers and researchers on ML engineering best practices, reproducibility, and performance profiling in large-scale RL and GPU-based ML systems.
- Collaborated with interdisciplinary teams to define brain-inspired computational modules (e.g., dendritic models) and integrated them into the RL ecosystem for scalable NeuroAI research.

minOLMo: Minimalist LLM Research Codebase | *PyTorch, LLM, Transformer, GPU* **2024**

- Re-designed the AllenAI OLMo codebase into minOLMo, a lightweight and researcher-friendly LLM framework tailored for single-GPU experimentation and educational use.
- Simplified distributed training dependencies to allow rapid prototyping on commodity GPUs while retaining reproducibility and extensibility for advanced ML experiments.
- Adopted by Kempner Institute researchers for studying LLM internals and pedagogy, accelerating exploration of transformer models in constrained GPU environments.

Open-Source Statistical Learning Packages (e.g., CausalGPS) | *R, Python, C++, HPC* **2021 – 2024**

- Developed and maintained CausalGPS, an R package implementing generalized propensity score (GPS) matching and weighting for continuous exposures to support causal inference in health and climate research.
- Optimized computationally intensive routines via C++ and OpenMP parallelization for shared-memory HPC environments, significantly improving performance and scalability.
- Enabled estimation workflows covering GPS calculation, pseudo-population generation, covariate balance testing, and outcome modeling (parametric and non-/semi-parametric), advancing reproducible causal analysis.
- Packaged with robust logging, unit testing, CI integration, and reproducibility features, and published on CRAN with full documentation and vignettes for widespread adoption.
- Extended the NSAPH Software ecosystem with complementary packages such as GPCERF (Gaussian process exposure–response estimation) and CRE (causal-rule ensemble models), contributing to scalable, interpretable tools for large-scale environmental health studies.

Leadership & Community Contributions

- Official member of MLCommons Working Groups (since Apr 2025)
- Reviewer of the Journal of Open Source Software (since Dec 2022)

Education

University of Memphis, Memphis, TN 2020
M.Sc. in Computer Science

University of Memphis, Memphis, TN 2018
Ph.D. in Geophysics

Technical Skills

Programming Languages: Python, C, C++, R, CUDA, Bash
ML & AI Frameworks: PyTorch, TensorFlow, Nvidia NeMo, Hugging Face, Triton, Ray
Distributed & Scaling: PyTorch DDP/FSDP/TP/MP, Model Sharding, SLURM, MPI, NCCL
GPU Computing & Profiling: NVIDIA Nsight Systems/Compute, CUDA, NCCL, FP16/FP8 Precision, GPU-to-GPU Communication, HPC Benchmarks (HPL, HPL-MxP, HPCG, STREAM)
Systems & Infrastructure: Linux (HPC/Cluster Environments), Docker, Podman, Singularity, AWS Cloud (EC2, S3), Kubernetes (basic), GitHub Actions, CI/CD
Data & Statistical Methods: Causal Inference (potential outcome framework), Statistical Learning