Harivallabha Rangarajan

⊙ Pittsburgh, PA ✉ hrangara@cs.cmu.edu ☎ (412)339-8783 ⦿ https://harivallabha.github.io/

in Hari-LinkedIn ⦾ Hari-Github

Education

Carnegie Mellon University *Aug 2024 – Dec 2025*
Master of Software Engineering, GPA: 4.05/4.00 *Pittsburgh, PA*

- **Highlighted Coursework:** Advanced Machine Learning Systems (PhD), Machine Learning With Large Datasets, Advanced Operating Systems and Distributed Systems (PhD), Special Topics in Databases: Query Optimization (PhD), Software Architectures, Software Project Management, Statistics, Formal Methods, Business and Marketing Strategy, Negotiations for Software Leaders.
- **Teaching Assistantships:** Prompt Engineering; Software Engineering for Executives; Machine Learning with Large Datasets.
- **Research Assistantships:** LLM Interpretability and Explainability with Prof. William Cohen (Google DeepMind x CMU): Understanding reasoning mechanisms of large language models through the lens of attention. Human-Centred Tools for Automated Prompt Engineering in Healthcare with Prof. Mayank Goel. Previously, with the AI for Social Good Lab: developed algorithms to tackle food insecurity, with fairness constraints and equitable resource allocation.
- **Sponsored Studio Project:** Architecting and building a LLM-powered Financial Insights platform. Lead a team of five engineers (sponsored by Troutwood LLC).

Birla Institute of Technology and Science (BITS), Pilani *Aug 2016 – May 2021*
B.E. Computer Science + M.Sc. Mathematics, GPA: 3.85/4.00 *Hyderabad, India*

- **Highlighted Coursework:** Operating Systems, Compiler Construction, Database Systems, Computer Networks, Programming Languages, Computer Architecture, Design and Analysis of Algorithms, Computational Geometry, Parallel Computing, Information Retrieval, Partial Differential Equations, Numerical Analysis, Graphs and Networks, Functional Analysis, Cryptography, Operations Research, Optimization.

Experience

Microsoft: OneDrive and SharePoint *Jul 2021 – Jul 2024*
Software Engineer *Hyderabad, India*

- Built and scaled Fluid, a framework for building distributed, and low-latency real-time collaborative applications. [Link] ↗
- Primarily worked on Scaling Fluid APIs (worked on an infrastructure to generate service load simulations by spinning up an AKS cluster that modelled client traffic at scale), and enabled media support in Fluid: extended the file format to interpret, process, and render video attachments.

Microsoft 365 Mobile: Intelligence Engineering

- Designed and deployed Copilots for Microsoft 365 (UnionCopilot, ScanCopilot) orchestrated using retrieval-augmented generation pipelines. Leveraged language models as orchestrators for vision agents, and architected prompt chains that balanced context retrieval, user intent disambiguation, and response quality under strict latency constraints. Built image agents for improved photo experiences. [Link] ↗
- Improved image search experiences under tight memory and compute constraints: computer vision on resource-constrained devices (auto-tagging, document classification, image segmentation).

Lawrence Berkeley National Lab | UC Berkeley *Dec 2020 – Jul 2021*
Research Affiliate *Berkeley, CA*

- Designed a Three-Parameter-Beta-Normal (TPBN) shrinkage prior for high-dimensional time series data, improving state-of-the-art sparse vector autoregression performance.
- **Published** findings at the NBER-NSF SBIES Conference, 2021. [PDF] ↗

Internships

MILA (Montreal-Quebec AI Institute)   *Jan 2021 – Mar 2021*
Research Collaborator   *Remote*
- ○ Investigated unsupervised representation learning for model-free, pixel-based reinforcement learning on procedurally generated environments.
- ○ Implemented Color Transformation Networks (CTNs) and Spatial Transformer Networks (STNs) and a cross-domain, cross-task transfer learning framework for improved sample efficiency.

King's College London   *Sep 2020 – Jan 2021*
Research Intern   *Remote*
- ○ Worked with the Systems Security Lab to develop adversarially robust algorithms for network intrusion detection systems (NIDS), with an emphasis on explainability.
- ○ Improved over current state-of-the-art autoencoder based NIDS, learning robust contrastive representations.

Microsoft   *May 2020 – Jul 2020*
Software Developer Intern   *Hyderabad, India*
- ○ Prototyped image segmentation algorithms for lightweight, on-device inference within Office Lens (document scanning).
- ○ Validated critical model improvements enhancing document boundary detection in real-world scenarios.

Cerenaut.ai   *Dec 2019 – May 2020*
Research Engineer (Remote)   *Melbourne, Australia*
- ○ Built causal ML models and explainable deep learning architectures.
- ○ Developed a bio-inspired, self-organizing neural network framework emulating interactions among neocortical regions. [PDF] ↗
- ○ **Co-authored** paper accepted at AJCAI 2020 and Cold Spring Harbor Laboratory's NAISys 2020.

Technische Universität Kaiserslautern   *Apr 2019 – Oct 2019*
Research Intern   *Remote*
- ○ Enhanced sparse Mahalanobis metric learning for gradient-based kernel regression with Tikhonov regularization.
- ○ GPU parallelized with CUDA, and performed automatic differentiation (AD) with the in-house C++ reverse-mode AD framework, CoDiPack.

Selected Projects

GPU-Accelerated Meshfree Solver for Inviscid Compressible Flows   *hari/meshfree-solver* ↗
- ○ Developed q-LSKUM based meshfree solver for aerodynamic shape optimization, in C++. GPU Parallelized the solver with CUDA. [PDF] ↗
- ○ The solver employs a least squares based spatial discretization of partial derivatives, for the numerical solution of Euler equations that govern inviscid compressible fluid flows [Deshpande et. al].

LSH-Based Plagiarism Detector
- ○ Developed an approximate nearest-neighbor pipeline using locality-sensitive hashing. Evaluated performance across Manhattan, Euclidean, Jaccard, and Cosine similarity metrics.

Simple and Explainable Machine-Learning based Proactive Autoscaling for Kubernetes
- ○ Implemented proactive Kubernetes autoscaling using the interpretable N-BEATS time-series forecasting model, enabling resource efficiency and reducing compute costs. [PDF] ↗

Latent Dirichlet Allocation based Probabilistic Search Engine
- ○ Developed a probabilistic search engine for financial tweets using LDA with Collapsed Gibbs Sampling inference; Hellinger Distance–based topic scoring, and combined sentiment analysis to predict positive or negative trends.

Technical Skills

Programming Languages: C++, Python, Rust, C, Java, C#, SQL, Typescript

Technologies: Azure, AWS, Docker, Kubernetes, Kafka, Prometheus, Grafana, Redis

Frameworks: PyTorch, Tensorflow, CUDA, Spark, React

Teaching and Leadership Experience

Carnegie Mellon University, Pittsburgh
- Teaching Assistant for Machine Learning with Large Datasets (Fall '25).
- Teaching Assistant for Prompt Engineering: The science of designing and engineering prompts (Spring '25).
- Teaching Assistant and Course Leader: Software Engineering for Executives (Spring '25).
- Research Assistant: AI for Social Good Lab (Fall '24).
- Research Assistant: Prof. William Cohen's Lab (Summer '25).
- Research Assistant: Prof. Mayank Goel's Lab (Summer '25).
- **President**: Master of Software Engineering Student Government.
- **Admissions Committee Member**: Master of Software Engineering.

Birla Institute of Technology and Science (BITS), Pilani
- Teaching Assistant for Real Analysis (Fall '19).
- Core Member of Chess Team (FIDE rating: 1742).

Awards and Recognition

- **National Supercomputing Mission (NSM) GAH-2020:** Selected for GPU-accelerated HPC program.
- **Indian Academy of Sciences Summer Research Fellowship (2019):** Selected participant (declined).
- **Indian National Olympiad in Informatics (x2):** Twice qualified for the Indian national round.
- **Google Summer of Code (2019):** Selected participant, contributing to open-source ML frameworks.

A bit about me

Currently: I am a Software Engineering graduate student at Carnegie Mellon University, Institute for Software Research. Experienced at working on large-scale software engineering projects, with **200M+ Monthly Active Users** and **Exascale** data volumes, interfacing with multiple teams across **OneDrive and SharePoint**. I've worked on **0→ 1** ideas (Microsoft365 Copilot), and **1→ ∞** products (Microsoft Fluid) - and I thrive on doing it again. Passionate about driving the latest research to a state of realization in the industry.