Nandan Kumar Jha

Ph.D., New York University (NYU)

Last updated: July 2026

Brooklyn, NY, USA
✉ *nj2049@nyu.edu*
nankj.com

Research Interests

I study representation learning and scaling laws in LLMs, focusing on how optimizers, not just architectures, shape representation geometry at scale. My recent work develops spectral telemetry to quantify realized capacity scaling, complementing classical loss-based scaling laws.

- **Representation Learning and Scaling Laws:**
  realized capacity, token-regime capacity allocation, optimizer–architecture co-design
- **High-Dimensional Learning Dynamics:**
  entropy dynamics, nonlinear eigenspectrum dynamics, spectral geometry
- **Systems and Hardware-Aware ML Efficiency:**
  privacy-preserving inference, roofline modeling, data-reuse-aware model optimization

Education

2020–2026 **Ph.D., New York University**, Brooklyn, NY, USA,
Electrical and Computer Engineering | GPA: 3.82/4 | Supervisor: Prof. Brandon Reagen.
- Thesis: Nonlinear Representation Dynamics: Spectral Scaling Laws and Applications to Private AI

2017–2020 **M.Tech., Indian Institute of Technology Hyderabad**, India,
Computer Science and Engineering | GPA: 9.27/10 | Supervisor: Dr. Sparsh Mittal.
- Thesis: Hardware-Aware Co-Optimization of Deep Convolutional Neural Networks

2009–2013 **B.Tech., National Institute of Technology Surat**, India,
Electronics and Communication Engineering | GPA: 8.20/10 | Supervisor: Dr. Upena Dalal.
- Thesis: Simulation and Analysis of Joint Source and Channel Coding for Video Transmission

Research and Publications

Peer-Reviewed Conference and Journal Proceedings

2026 NerVE: Nonlinear Eigenspectrum Dynamics in LLM Feed-Forward Networks
International Conference on Learning Representations (ICLR)
Nandan Kumar Jha, Brandon Reagen
arXiv, website, code

2025 Spectral Scaling Laws in Language Models: How Effectively Do Feed-Forward Networks Use Their Latent Space?
Empirical Methods in Natural Language Processing (EMNLP), ***Main Conference***
Nandan Kumar Jha, Brandon Reagen
arXiv, code

2025 Network and Compiler Optimizations for Efficient Linear Algebra Kernels in Private Transformer Inference (Invited Paper)
IEEE/ACM International Conference on Computer Aided Design (ICCAD)
Karthik Garimella, Negar Neda, Austin Ebel, **Nandan Kumar Jha**, Brandon Reagen
arXiv

2024 DeepReShape: Redesigning Neural Networks for Efficient Private Inference
Transactions on Machine Learning Research (TMLR)
Nandan Kumar Jha, Brandon Reagen
arXiv

2023 Characterizing and Optimizing End-to-End Systems for Private Inference
Architectural Support for Programming Languages and Operating Systems (ASPLOS)
Karthik Garimella, Zahra Ghodsi, **Nandan Kumar Jha**, Siddharth Garg, Brandon Reagen
arXiv

2021 DeepReDuce: ReLU Reduction for Fast Private Inference
International Conference on Machine Learning (ICML), ***Spotlight Talk***
Nandan Kumar Jha, Zahra Ghodsi, Siddharth Garg, Brandon Reagen
arXiv, Press release, TechXplore news, ScienceDaily news, **[100+ citations]**

2021 Circa: Stochastic ReLUs for Private Deep Learning
Neural Information Processing Systems (NeurIPS)
Zahra Ghodsi, **Nandan Kumar Jha**, Brandon Reagen, Siddharth Garg
arXiv

2020 ULSAM: Ultra-Lightweight Subspace Attention Module for Compact Convolutional Neural Networks
IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)
Rajat Saini*, **Nandan Kumar Jha***, Bedanta Das, Sparsh Mittal, C Krishna Mohan
arXiv (*Equal contributions.), **[100+ citations]**

2020 Modeling Data Reuse in Deep Neural Networks by Taking Data-Types into Cognizance
IEEE Transactions on Computers (TC)
Nandan Kumar Jha, Sparsh Mittal
arXiv

2020 DRACO: Co-Optimizing Hardware Utilization and Performance of DNNs on Systolic Accelerator
IEEE Computer Society Annual Symposium on VLSI (ISVLSI)
Nandan Kumar Jha, Shreyas Ravishankar, Sparsh Mittal, Arvind Kaushik, Dipan Mandal, Mahesh Chandra
arXiv

2020 E2GC: Energy-efficient Group Convolution in Deep Neural Networks
International Conference on VLSI Design (VLSID)
Nandan Kumar Jha*, Rajat Saini*, Subhrajit Nag, Sparsh Mittal
arXiv (*Equal contributions.)

2019 DeepPeep: Exploiting Design Ramifications to Decipher the Architecture of Compact DNNs
ACM Journal on Emerging Technologies in Computing Systems (JETC)
Nandan Kumar Jha, Sparsh Mittal, Binod Kumar, Govardhan Mattela
arXiv

2019 Data-type Aware Arithmetic Intensity for Deep Neural Networks
IEEE International Conference on Computer Design (ICCD), (accepted as work in progress)
Nandan Kumar Jha, Sparsh Mittal, Sasikanth Avancha
ResearchGate

2019 The Ramifications of Making Deep Neural Networks Compact
International Conference on VLSI Design (VLSID)
Nandan Kumar Jha, Sparsh Mittal, Govardhan Mattela
arXiv

Manuscripts Under Review

2026 Same Architecture, Different Capacity: Optimizer-Induced Spectral Scaling Laws
Under review at NeurIPS 2026
Nandan Kumar Jha, Brandon Reagen
arXiv, website, code, blog

2026 AERO: Entropy-Guided Framework for Private LLM Inference
Under review at EMNLP 2026
Nandan Kumar Jha, Brandon Reagen
arXiv

Peer-Reviewed Workshop Proceedings

2025 Regularizing the Entropy Landscape of Self-Attention: Towards a Soft Inductive Bias in LLMs
The 17th International Workshop on Optimization for Machine Learning (OPT), NeurIPS
Nandan Kumar Jha, Brandon Reagen
Webpage

2025 A Random Matrix Theory Perspective on the Learning Dynamics of Multi-head Latent Attention
The 3rd Workshop on High-dimensional Learning Dynamics (HiLD), ICML
Nandan Kumar Jha, Brandon Reagen
arXiv, News article

2025 Spectral Scaling Laws in Language Models: *How Effectively Do Feed-Forward Networks Use Their Latent Space?*
Workshop on Actionable Interpretability (AIW), ICML
Nandan Kumar Jha, Brandon Reagen
Webpage

2025 Entropy-Guided Attention for Private LLMs
The 6th Workshop on Privacy-Preserving Artificial Intelligence (PPAI), AAAI
Nandan Kumar Jha, Brandon Reagen
arXiv, code, Press release, LinkedIn article

2024 ReLU's Revival: On the Entropic Overload in Normalization-Free Large Language Models
The 2nd Workshop on Attributing Model Behavior at Scale (ATTRIB), NeurIPS
Nandan Kumar Jha, Brandon Reagen
arXiv, code

2021 Sisyphus: A Cautionary Tale of Using Low-Degree Polynomial Activations in Privacy-Preserving Deep Learning
Privacy Preserving Machine Learning Workshop (PPML), ACM CCS
Karthik Garimella, **Nandan Kumar Jha**, Brandon Reagen
arXiv

Preprints (arXiv)

2024 TruncFormer: Private LLM Inference Using Only Truncations
Patrick Yubeaton, Jianqiao Cambridge Mo, Karthik Garimella, **Nandan Kumar Jha**, Brandon Reagen, Chinmay Hegde, Siddharth Garg
arXiv

2021 CryptoNite: Revealing the Pitfalls of End-to-End Private Inference at Scale
Karthik Garimella, **Nandan Kumar Jha**, Zahra Ghodsi, Siddharth Garg, Brandon Reagen
arXiv

2020 On the Demystification of Knowledge Distillation: A Residual Network Perspective
Nandan Kumar Jha*, Rajat Saini*, Subhrajit Nag, Sparsh Mittal
arXiv (*Equal contributions.)

Talks

Invited Talks and Discussions

2026 **Under the Hood of AI**
Panelist, Expert Panel on AI Infrastructure, NYU School of Law
Event page

2025 **Entropy and Private Language Models**
CILVR Seminar Series, NYU Center for Data Science
Seminar page, Video

2025 **Entropy-Guided Attention for Private LLMs**
Ploutos AI Fireside Chat, hosted by Cecile Tamura
Video

Conference Presentations

2021 DeepReDuce: ReLU Reduction for Fast Private Inference
Spotlight Talk at International Conference on Machine Learning (ICML)
Video

Industry Experience

2015–2017 **Seagate Technology HDD (India) Private Limited**, Bangalore, India.

- Designation: *Electrical Design Engineer*
- Job role: Design and verification of Solid State Drives (SSDs); Electrical characterization of DRAM and NAND; Signal integrity verification of NAND and DRAM datapath

Technical Skills

ML PyTorch, NumPy, scikit-learn, Hugging Face Transformers, evaluation pipelines
Training Multi-GPU training (PyTorch DDP), mixed-precision training, Slurm, Weights & Biases
Tools Python, Git, Docker, Linux, Bash, LaTeX
Systems Roofline modeling, data-reuse analysis, Verilog, Synopsys EDA tools

Honors and Awards

2025 **ECE Student Research Poster Day Award**
New York University

2021–2022 **Ernst Weber PhD Fellowship**
New York University

2019 **Certificate of Appreciation in Research**
Indian Institute of Technology Hyderabad

Reviewing

Conferences NeurIPS (2023–2026), ICLR (2024–2026), ICML (2024–2026), CVPR 2024, ICCV 2025, AISTATS 2025, AAAI 2025

Journals TMLR (2025–2026), TIFS 2025, JETC 2020

Teaching and Outreach

2024 **Guest Instructor, K12 Machine Learning Summer School**, *New York University*.
Taught LLM fundamentals and guided 30 students in practical Hugging Face-based projects.

2023 **Lead Instructor, K12 Machine Learning Summer School**, *New York University*.
Led three student cohorts through two-week ML curricula and mentored hands-on projects.

2019 **Mentor, Artificial Intelligence Summer School**, *IIT Hyderabad*.
Mentored student capstone projects in recommendation systems and imbalanced classification.