

Arnav Chavan

Senior Applied Scientist, Amazon Lab126 | Incoming Ph.D. Student, UC Berkeley

✉ Email - arnavchavan04@gmail.com 🌐 Homepage 📄 Github 🎓 Google Scholar

Education

Fall 2026	University of California, Berkeley Incoming Ph.D. Student	Berkeley, CA, USA
	<i>Research interests: efficient architectures via Neural ODEs/SDEs, AI for scientific discovery, and continual learning.</i>	
Jul 2023 Jul 2018	Indian Institute of Technology, (ISM) Dhanbad Integrated Master of Technology in Mathematics and Computing	
	Course Work - Probability and Statistics, Linear Algebra, Statistical Inference, Numerical Methods, Machine Learning, Operating Systems, Computer Networks, Data Structures and Algorithms, High Performance Programming	

Experience

Present April 2025	Amazon Lab126 Senior Applied Scientist	San Francisco, USA
	Work on extreme hardware-software co-design , leading Generative AI capabilities on Amazon's custom silicon, a role directly enabled by the acquire of Nyun AI. Scaled model compression and hardware-aware optimization pipelines developed at Nyun AI to enable on-device GenAI for Alexa+ .	
March 2025 Aug 2023	Nyun AI Co-founder and CTO Advisor/CEO: Dr. Deepak Gupta	Bangalore, India
	Led the development of advanced compression pipelines for foundational models and optimized their deployment on edge devices while managing a multidisciplinary team to deliver tailored solutions for Fortune 5 enterprises. Acquired by Amazon Lab126 . Published work on model compression and efficiency[ACL'24 , TMLR'24 , '25, ICLR'24 , IJCAI'24].	
Feb 2023 Aug 2022	Microsoft Research Research Intern Advisor: Dr. Sunayana Sitaram , Suresh Parthasarathy	Bangalore, India
	Conducted research on large-scale HTML understanding models to improve generalization in information extraction from web content and investigated the implications of model compression on bias and fairness in large language models, published at [ACL'23].	
Apr 2022 Jan 2022	Google DeepMind Student Researcher Advisor: Dr. Pradeep Shenoy	Bangalore, India
	Executed a comprehensive empirical study on simplicity bias in deep neural networks, establishing methodologies to quantify and mitigate this bias, and developed synthetic datasets and experimental protocols to enhance the understanding of neural network performance dynamics.	
Dec 2021 Aug 2021	Mohammed Bin Zayed University of AI Research Assistant Advisor: Prof. Zhiqiang Shen	Abu Dhabi, UAE
	Innovated methods for compressing vision transformers by introducing sparsity across transformer modules within a unified framework, demonstrating significant advancements over existing compression techniques. Results published at [CVPR'22 , TMLR'25].	
May 2023 May 2020	Transmute AI Research Co-founding member, Senior Researcher Advisor: Dr. Deepak K. Gupta , Prof. Dilip K. Prasad	Tromsø, Norway
	Co-founded a research lab to foster research culture among peers and aspiring researchers, solicited funding by Texmin Hub and Bio-AI Lab of UiT Norway. Guided 10+ UG students from different colleges to pursue research. Worked on research problems around network compression, meta-learning and efficient ML leading to 4 publications so far [ICLR'21 , ICIP'21 , CVPR'22 , ICASSP'23].	

Selected Publications

* denotes equal contribution

- [C.9] **S2D: Selective Spectral Decay for Quantization-Friendly Conditioning of Neural Activations** [🔗]
Arnav Chavan*, Nahush Lele*, Udbhav Bamba, Sankalp Dayal, Aditi Raghunathan, Deepak Gupta
The IEEE/CVF Conference on Computer Vision and Pattern Recognition 2026 [CVPR'26]
- [C.10] **DOT-MoE: Differentiable Optimal Transport for MoEification** [🔗]
Udbhav Bamba*, Arnav Chavan*, Aryamaan Thakur, Steve Teig, Deepak Gupta
The 43rd International Conference on Machine Learning [ICML'26]
- [W.2] **Jacobian-guided Noise Injection for Quantization Robustness in Large Language Models**
Deepanshu Pandey*, Nahush Lele*, Arnav Chavan*, Sankalp Dayal, Deepak Gupta
ICML 2026 Workshop on Adaptive Foundation Models (AdaptFM) [ICML-W'26]
- [S.1] **One-for-All: Generalized LoRA for Parameter-Efficient Fine-tuning** [🔗]
Arnav Chavan*, Zhuang Liu, Deepak Gupta, Eric Xing, Zhiqiang Shen*
Under Submission at Transactions on Machine Learning Research
- [J.3] **Rethinking the Value of Training-Free Structured Pruning of LLMs** [🔗]
Arnav Chavan*, Nahush Lele*, Aryamaan Thakur, Deepak Gupta
Transactions on Machine Learning Research [TMLR'25]
- [J.2] **Pushing the Limits of Gradient Descent for Efficient Learning on Large Images** [🔗]
Arnav Chavan*, Deepak K Gupta*, Gowreesh Mago*, Dilip K Prasad, Rajat Thomas
Transactions on Machine Learning Research [TMLR'24]
- [C.8] **Surgical Feature-Space Decomposition of LLMs: Why, When and How?** [🔗]
Arnav Chavan, Nahush Lele, Deepak Gupta
The 62nd Annual Meeting of the Association for Computational Linguistics [ACL'24]
- [C.7] **Faster and Lighter LLMs: A Survey on Current Challenges and Way Forward** [🔗]
Arnav Chavan, Raghav Magazine, Shubham Kushwaha, M  rouane Debbah, Deepak Gupta
The 33rd International Joint Conference on Artificial Intelligence [IJCAI'24]
- [C.6] **Rethinking Compression: Reduced Order Modelling of Latent Features in Large Language Models** [🔗]
Arnav Chavan, Nahush Lele, Deepak Gupta
The Twelfth International Conference on Learning Representations. [Tiny Paper @ ICLR'24]
- [C.5] **A Comparative Study of Model Compression Techniques on Fairness in Language Models** [🔗]
Krithika Ramesh, Arnav Chavan, Shrey Pandit, Sunayana Sitaram
The 61st Annual Meeting of the Association for Computational Linguistics [ACL'23]
- [W.1] **RCV2023 Challenges: Benchmarking Model Training and Inference for Resource-Constrained Deep Learning** [🔗]
Arnav Chavan*, Rishabh Tiwari*, Deepak K. Gupta* et. al.
ICCV 2023 Workshop on Resource-Efficient Deep Learning for Computer Vision [ICCV-W'23]
- [C.4] **Dynamic Kernel Selection for Improved Generalization and Memory Efficiency in Meta-learning** [🔗]
Arnav Chavan*, Rishabh Tiwari*, Udbhav Bamba, Deepak K. Gupta
The IEEE/CVF Conference on Computer Vision and Pattern Recognition 2022 [CVPR'22]
- [C.3] **Vision Transformer Slimming: Multi-Dimension Searching in Continuous Optimization Space** [🔗]
Arnav Chavan*, Zhiqiang Shen*, Zhuang Liu, Zechun Liu, Kwang-Ting Cheng, Eric Xing
The IEEE/CVF Conference on Computer Vision and Pattern Recognition 2022 [CVPR'22]
- [C.2] **Chipnet: Budget-aware pruning with heaviside continuous approximations** [🔗]
Arnav Chavan*, Rishabh Tiwari*, Udbhav Bamba*, Deepak K. Gupta
The Ninth International Conference on Learning Representations [ICLR'21]
- [C.1] **Rescaling CNN through learnable repetition of network parameters** [🔗]
Arnav Chavan*, Udbhav Bamba*, Rishabh Tiwari*, Deepak K. Gupta
The 28th IEEE International Conference on Image Processing [ICIP'21]
- [J.1] **Deep learning for detection and segmentation of artefact and disease instances in gastrointestinal endoscopy** [🔗]
Sharib Ali, Arnav Chavan, Jens Rittscher et. al.
Medical Image Analysis [MIA]

Selected Honors and Awards

➤ **Kaggle Competitions Master:** Became one of the youngest Indian Kaggle Competitions Master at 19 in 2020. [Profile]

- > **Winner** of first ever national level **Amazon ML Challenge 2021** with over 3k+ participating teams, received an internship offer and a cash price of 100k INR.
- > **Received Fellowship Award** Kishore Vaigyanik Protsahan Yojana (KVPY), 2016; awarded to top 1000 high school students with science majors across India.

Notable Positions of Responsibility

- > **Core Workshop Organizer**
 - > [Resource-Adaptive Foundation Model Inference \(AdaptFM\)](#) ICML'26
 - > [Resource Efficient Deep Learning for Computer Vision](#) ICCV'23
- > **Competition Organizer**
 - > Designed and ran the [Efficient LLM Challenge](#) (176 teams) and authored the [results writeup](#) AdaptFM @ ICML'26
- > **Open Source Dev Lead**
 - > **Nyuntam** is Nyun AI's advanced suite designed to optimize and accelerate a wide array of deep learning models.
- > **Mentorship**
 - > Aryamaan Thakur, Nahush Lele, Raghav Magazine 2023-24
Research Intern, Nyun AI
 - > Saksham Aggarwal, Taneesh Gupta, Pawan Kumar Sahu 2022-23
Research Intern, Transmute AI Research (w/ Dr. Deepak K. Gupta)
- > **Reviewer** at NeurIPS '24 '23, ICML '24, CVPR '24 '23 '22, ICCV '23, PRL Journal 2022-24
- > **Student Coordinator** at Cyber Labs, the official cyber society of IIT (ISM), Dhanbad 2020-23