

Education

Jul 2019	Indian Institute of Technology, Bombay	CPI: 9.41/10
Aug 2024	B.Tech. + M.Tech. in Electrical Engineering Minor in Computer Science	Dept. Rank 7/70

Research Experience

Jul 2024	Google DeepMind	
Present	Pre-Doctoral Researcher Advisors: Dr. Prateek Jain, Dr. Aditya Kusupati, Dr. Karthikeyan Shanmugam	
Jul 2022	IIT Bombay	
Aug 2024	Undergraduate Researcher Advisors: Dr. Sharayu Moharir, Dr. Jaya Prakash Champati	
Jul 2022	MITACS, University of Calgary	
Aug 2024	Research Intern Advisor: Dr. Hatem Abou Zeid	
Jul 2022	Texas Instruments	
Aug 2024	Research Intern Advisor: Jawaharlal Tangudu	

Publications

S=In Submission, C=Conference, T=Technical Report, *=Equal Contribution

- [C.3] **Matryoshka Quantization** 
Puranjay Datta*, Pranav Nair*, Jeff Dean, Prateek Jain, Aditya Kusupati
International Conference on Machine Learning 2025
🏆 Oral Sparsity in LLMs Workshop, International Conference on Learning Representations 2025 [ICML'25]
- [C.2] **Online Learning with Stochastically Partitioning Experts** 
Puranjay Datta, Sharayu Moharir, Jaya Prakash Champati
Uncertainty in Artificial Intelligence 2025 [UAI'25]
- [C.1] **Regret Bounds for Online Learning for Hierarchical Inference** 
Ghina Al-Atat, Puranjay Datta, Sharayu Moharir, Jaya Prakash Champati
Proceedings of the Twenty-fifth International Symposium on Theory, Algorithmic Foundations, and Protocol Design for Mobile Networks and Mobile Computing [MobiHoc'24]
- [T.2] **Gemma 3n MatFormer Lab**  
Puranjay Datta, Aditya Kusupati, Ryan Mullins, Omar Sanseviero, Rakesh Shivanna, Gemma Team
- [T.1] **Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities** 
Gemini Team
- [S.1] **ROPES: Robotic Pose Estimation via Score-based Causal Representation Learning** 
Puranjay Datta*, Pranamya Kulkarni*, Emre Acartürk, Burak Varici, Karthikeyan Shanmugam, Ali Tajer
Embodied World Models for Decision Making Workshop Neurips 2025
Under review at International Conference on Learning Representations 2026 [In Submission]

Selected Research Projects

Elastic Continued Pretraining

Advisors: Aditya Kusupati, Prateek Jain

- Developed neuron reordering strategy for MatFormers, enabling efficient continual pretraining of smaller models.
- Delivered models powering Gemini Embeddings & Google AI Overviews, with over 100,000 internal downloads.
- Implemented Mix-n-Match for open-source Gemma 3n, allowing custom models via network parameter slicing.

Matryoshka Quantization

Advisors: Aditya Kusupati, Prateek Jain, Jeff Dean

- Proposed a novel technique allowing a single trained model to be served at multiple precisions (e.g., int8, int4, int2).
- MatQuant improved the int2 model quality by 4% on OmniQuant and 7% on QAT as base algorithms .
- Enabled layer-wise Mix-n-Match, generating a combinatorial number of models at no additional training cost.

Adaptive Chain of Thought

Advisors: *Aditya Kusupati, Prateek Jain*

- Introduced adaptive Chain of Thought (CoT) to control reasoning length and reduce model inference costs.
- Improved RL-training efficiency by enabling one model to support various reasoning budgets at 20% less compute.
- Tested various natural language prompt and special token injection strategies to control thought to answer leakages.

Future aware LLMs

Advisors: *Sravanti Addepalli, Karthikeyan Shanmugam*

- Proposed distilling future tokens during training to improve at long reasoning tasks without extra inference compute.
- Utilized the MatFormer architecture in a multi-task framework, where a harder future prediction objective on a larger model improves the performance of the core next-token prediction task of its sub-model.
- Future token joint optimization improves upon isotoken MatFormer baseline with 10% gains on perplexity evals.

Embodied World Model Understanding

Advisors: *Karthikeyan Shanmugam, Ali Tajer*

- Developed an unsupervised Causal Representation Learning (CRL) framework to encode robot poses without labels.
- Matched supervised baseline RoboPEPP's mse with much lower training compute in label-constrained settings.
- Scaled theoretical CRL concepts using JAX by stabilizing score matching for high-dimensional robotics data.

Hierarchical Inference at Edge

Advisors: *Sharayu Moharir, Jaya Prakash Champati*

- Explored optimal offloading strategy between small low-accuracy ML models on edge devices and large high-accuracy ML models on edge servers for inference tasks using EXP3 hedge online learning framework.
- Designed Hedge-G & AdaHedge-G in the expert advice framework for dynamic IID experts under adversarial losses.
- Proved Hedge-G achieves $\sqrt{T \log T}$ order-optimal regret, matching the theoretical lower bound for online experts.

Reconnaissance Blind Chess Agent

Advisor: *Shivaram Kalyanakrishnan*

- Modelled the agent as POMDP with Leela Chess Zero as a meta-learning expert and policy rollout of higher depths.
- Developed a Replay Buffer to assess blunders and move scores in order to highlight Fianchetto bot's limitations.
- Adapted a new consistent scoring system using $V(s)$ compared to $Q(s, a)$ with specialized heuristics for speed up.

Scholastic and Technical Accolades

- | | |
|--|------|
| → Secured All India Rank 132 in Kishor Vaigyanic Protsahan Yojana (KVPY) | 2018 |
| → Secured All India Rank 460 in IIT JEE-Advanced and All India Rank 300 in JEE Mains | 2019 |
| → Recipient of the MITACS Globalink Research Internship Fellowship | 2023 |
| → Received 2 Spot Bonuses at Google DeepMind for delivering high impact models | 2024 |
| → One of 12, of 20k applicants to be selected for the Google Pre-doctoral program 2024-26 | 2024 |
| → Oral acceptance at International Conference on Learning Representations Workshop on Sparsity in LLMs | 2025 |

Teaching and Academic Service

- TA for the Probability and Random Processes course at IIT Bombay.
- Backend Lead for NPTEL Digital Signal Processing and its Applications : Addressed queries on an online forum and online YouTube doubt sessions, providing assistance to more than 5k enrolled students.
- Volunteered for setting up inference servers for Gemini Models for pre-release product testing across diverse tasks.
- Presented our work Matryoshka Quantization in the Google-wide Cool Papers Reading group.

Key Courses Undertaken

- | | |
|-------------------------|--|
| Machine Learning | Probability and Random Processes, Stochastic Optimization, Convex Optimization, Markov Chains, Stochastic Control, Advanced Concentration Inequalities, Advanced ML, Intelligent & Learning Agents |
| Computer Science | Data Structures and Algorithms, Logic for Computer Science, Design and Analysis of Algorithms, Game Theory and Algorithmic Mechanism Design, Computer Networks |
| Mathematics | Calculus, Linear Algebra, Complex Analysis, Differential Equations, Number Theory and Cryptography |