# Mayank Mishra

Research Engineer
MIT-IBM Watson AI Lab

**Email**: mayank31398@gmail.com
**GitHub**: https://github.com/mayank31398
**LinkedIn**: https://www.linkedin.com/in/mayank31398
**Google Scholar**: https://scholar.google.com/citations?user=YsbtW6cAAAAJ&hl=en

## ACADEMIC DETAILS

- B.Tech, Electrical Engineering, Indian Institute of Technology Delhi (IIT Delhi)
- Secured All India Rank 921 in JEE Advanced 2016 (out of 1.6 million candidates)
- Secured All India Rank 682 in JEE Mains 2016 (out of 15 million candidates)

## WORK EXPERIENCE

**MIT-IBM Watson AI Lab, IBM Research** *Jan 2024 - present*

- Worked on creating Dolomite-Engine, a repository for training both Dense and Mixture of Experts (MoE) models on thousands of GPUs with 4D parallelism. Dolomite Engine has been tested on 6k H100s with good performance scaling
- Working on new LLM architectures that are better suited for both training and serving at scale. Have published works like Cross-Layer Attention, Ladder Residual etc that are more efficient than a standard transformer model
- Trained SOTA Granite Code Models and Granite Language Models for IBM using Dolomite-Engine, have published papers and released SOTA code models ranging from 3B to 34B scales

**IBM Research** *Aug 2020 - Jan 2024*

- Worked on large-scale distributed training of Language Models and optimizing the LLMs for training and inference efficiency
- Created an optimized serving framework for serving large transformers for inference to researchers within IBM, some CPU-side systems optimization here are now a part of HuggingFace TGI

## PUBLICATIONS

- William Brandon*, **Mayank Mishra***, Aniruddha Nrusimha, Rameswar Panda, Jonathan Ragan-Kelley, Reducing Transformer Key-Value Cache Size with Cross-Layer Attention, Accepted for publication at **NeurIPS 2024**
- Gaurav Pandey, Yatin Nandwani, Tahira Naseem, **Mayank Mishra**, Guangxuan Xu, Dinesh Raghu, Sachindra Joshi, Asim Munawar, Ramn Fernandez Astudillo, BRAIn: Bayesian Reward-conditioned Amortized Inference for natural language generation from feedback, **ICML 2024**
- **Mayank Mishra**, Prince Kumar, Riyaz Bhat, Rudra Murthy, Danish Contractor, Srikanth Tamilselvam, Prompting with Pseudo-Code Instructions, **EMNLP 2023**
- BigScience Group, BLOOM: A 176b-parameter open-access multilingual language model, **JMLR 2024**
- BigCode Group, StarCoder: may the source be with you!, **TMLR**
- BigCode Group, SantaCoder: don't reach for the stars!, **TMLR**
- **Mayank Mishra**, Danish Contractor, Dinesh Raghu, Joint Reasoning on Hybrid-knowledge sources for Task-Oriented Dialog, **EACL 2023**
- **Mayank Mishra**, Dhiraj Madan, Gaurav Pandey, Danish Contractor, Variational Learning for Unsupervised Knowledge Grounded Dialogs, **IJCAI 2022**
- Prathosh A. P.*, Varun Srivastava*, **Mayank Mishra***, Adversarial Approximate Inference for Speech to Electroglottograph Conversion, **IEEE TASLP**

## PRE-PRINTS

- Granite Team, Granite 3.0 Language Models, Preprint
- Yikang Shen, Matthew Stallone, **Mayank Mishra**, Gaoyuan Zhang, Shawn Tan, Aditya Prasad, Adriana Meza Soria, David D. Cox, Rameswar Panda, Power Scheduler: A Batch Size and Token Number Agnostic Learning Rate Scheduler, Arxiv Preprint
- **Mayank Mishra***, Matt Stallone*, Gaoyuan Zhang* et al., Granite Code Models: A Family of Open Foundation Models for Code Intelligence, Arxiv Preprint
- Matt Stallone, Vaibhav Saxena, Leonid Karlinsky, Bridget McGinn, Tim Bula, **Mayank Mishra**, Adriana Meza Soria, Gaoyuan Zhang, Aditya Prasad, Yikang Shen, Saptha Surendran, Shanmukha Guttula, Hima Patel, Parameswaran Selvam, Xuan-Hong Dang, Yan Koyfman, Atin Sood, Rogerio Feris, Nirmit Desai, David D. Cox, Ruchir Puri, Rameswar Panda, Scaling Granite Code Models to 128K Context, Arxiv Preprint
- Achintya Kundu, Rhui Dih Lee, Laura Wynter, Raghu Kiran Ganti, **Mayank Mishra**, Enhancing Training Efficiency Using Packing with Flash Attention, Arxiv Preprint
- IBM Infrastructure team, The infrastructure powering IBM's Gen AI model development, Arxiv Preprint
- Bowen Pan, Yikang Shen, Haokun Liu, **Mayank Mishra**, Gaoyuan Zhang, Aude Oliva, Colin Raffel, Rameswar Panda, Dense Training, Sparse Inference: Rethinking Training of Mixture-of-Experts Language Models, Arxiv preprint
- Aniruddha Nrusimha, **Mayank Mishra**, Naigang Wang, Dan Alistarh, Rameswar Panda, Yoon Kim, Mitigating the Impact of Outlier Channels for Language Model Quantization with Activation Regularization, Arxiv preprint
- Taishi Nakamura*, **Mayank Mishra***, Simone Tedeschi* et al., Aurora-M: The First Open Source Multilingual Language Model Red-teamed according to the U.S. Executive Order, Arxiv preprint
- BigCode Group, StarCoder 2 and The Stack v2: The Next Generation, Arxiv preprint
- Vinay Kyatham*, **Mayank Mishra***, Tarun Kumar Yadav, Deepak Mishra, Prathosh AP, Variational Inference with Latent Space Quantization for Adversarial Resilience, Arxiv preprint

## BLOGS

- Mayank Mishra, Saving Memory Using Padding-Free Transformer Layers during Finetuning
- Mayank Mishra, Aurora-M: The First Open Source Biden-Harris Executive Order Red teamed Multilingual Language Model

## OPEN-SOURCE PROJECTS

- dolomite-engine: Created dolomite-engine for training LLMs on thousands of GPUs with 4D parallelism
- transformers-bloom-inference: Created a serving frameworks for LLMs before popular frameworks like vLLM/TGI

## PATENTS

**EdgeEGG - A system and method for hand-held electrode free elctroglottograph using neural networks on programmable controllers**

- Proposed a safe contact-free ElectroGlottoGraph which provides an accurate estimate of EGG signal
- Proposed a cost-effective and efficient mechanism with integrated speech sensors to allow edge computation of EGG
- Designed a resource efficient hardware device optimizing both energy consumption and prediction latency, by performing computations on very low power micro-controllers