

MUKUL RANJAN

☎ +1 6462471239 ✉ mukulranjan0@gmail.com [in mr54](#) [mukul54](#) [G Google Scholar](#)

EDUCATION

Cornell University

PhD in ECE

Aug. 2026 – Present

NYC, USA

Mohamed Bin Zayed University of AI (MBZUAI)

Aug. 2024 – Jun. 2026

MS in Machine Learning

Abu Dhabi, UAE

Indian Institute of Technology, Guwahati

Jul. 2017 – Jun. 2021

B.Tech. in ECE

Guwahati, India

PUBLICATION

- **Flash-dLLM: IO-Aware KV Caching and Parallel Decoding for Fast, Memory-Efficient Diffusion LLMs** | [Link](#)
Quan Nguyen-Tri, Mukul Ranjan, and Zhiqiang Shen.
Published at NeurIPS 2026.
- **Attention Is All You Need for KV Cache in Diffusion LLMs** | [Link](#)
Mukul Ranjan*, Quan Nguyen-Tri*, and Zhiqiang Shen.
Published at ICLR 2026.
- **Time Blindness: Why Video-Language Models Can't See What Humans Can?** | [Link](#)
Mukul Ranjan*, Ujjwal Upadhyay*, Zhiqiang Shen, and Mohamed Elhoseiny.
Published at CVPR 2026.
- **Mobile-MMLU: A Mobile Intelligence Language Understanding Benchmark** | [Link](#)
Mukul Ranjan*, Sondos Mahmoud Bsharat*, Aidar Myrzakhan*, et al.
Published at DMLR 2026.
- **KITAB-Bench: A Comprehensive Multi-Domain Arabic OCR Benchmark** | [Link](#)
Mukul Ranjan*, Ahmed Heakl*, Muhammad Abdullah Sohail*, Rania Elbadry*, et al.
Published at ACL 2025.
- **Deep-ASPECTS: A Segmentation-Assisted Model for Stroke Severity Measurement** | [Link](#)
Ujjwal Upadhyay, Mukul Ranjan, et al.
Published at ECCV 2022, MCV Workshop.
- **An Evaluation Framework for Sanskrit-English Google Translate** | [Link](#)
Akshat Shukla, Chaarvi Bansal, Sushrut Badhe, Mukul Ranjan, and Rohitash Chandra.
Published at Natural Language Processing Journal, 2023.

UNDER SUBMISSION

- **Beyond Size: How Gradients Shape Pruning Decisions in Large Language Models**
Rocktim Jyoti Das*, Mukul Ranjan*, Mingjie Sun*, Liquan Ma, and Zhiqiang Shen.
Under Review at TMLR 2026 (AISTATS 2026 Reject with (6,5,4,2) out of 7).
- **One-for-All: Generalized LoRA for Parameter-Efficient Fine-tuning**
Arnav Chavan*, Mukul Ranjan*, Zhuang Liu, Deepak Gupta, Eric Xing, and Zhiqiang Shen.
Pending Submission for IEEE TPAMI (NeurIPS 2025 Reject with (4,4,4) out of 6) .

ACHIEVEMENTS

- Awarded the **Irwin Jacobs Fellowship (2026)** by the School of ECE, Cornell University.
- Received the **Department Chair's Award (2026)** for the outstanding research contribution at MBZUAI.
- Placed 1st in the 2024 DGE Elite Hackathon for Cybersecurity at GITEX.
- Got 1st position in the Insilico-Medicine AI-driven Drug Discovery Hackathon 2024.
- **Undergraduate Samsung Fellowship Award (2021)** for research on automated movie colorization.
- Secured **Gold medal** in Inter-IIT Tech Meet (2021) for multilingual sentiment analysis and headline generation.
- **KVPY Fellowship (2017)**: selected for the national research fellowship by the Govt of India.
- State Examination Board (2015): Secured 2nd **position** among 1.5M candidate

* denotes equal contribution, random order

SERVICES

Reviewer: NeurIPS (2024, 2025, 2026), AISTATS (2025, 2026, 2027), ICML 2025, AAAI (2026,2027), ICLR (2026, 2027)

WORK EXPERIENCE

Mohamed Bin Zayed University of AI

Aug. 2024 – Now

Teaching & Research Assistant / Supervisor: [Prof. Zhiqiang Shen](#)

Abu Dhabi, UAE

- Supported ML7101 Probabilistic and Statistical Inference course for 148 students through TA sessions, quiz creation, and assignment evaluation; focused instruction on time series modeling and hypothesis testing.
- Developing reinforcement learning based methods to improve the quantitative reasoning of **diffusion models**.
- Developed different applied, theoretical and empirical machine learning methods and benchmarks including: **Time-Blindness, Elastic-Cache, Kitab-Bench** and **Mobile-MMLU**.

University of Illinois Urbana-Champaign

Jul. 2025 – Oct. 2025

Research Scholar / Supervisor: [Prof. Deming Chen](#)

Urbana-Champaign, USA

- Engineered heterogeneous inference algorithm for Hybrid LLM models (Hymba, Jamba, Qwen-Next) across AMD SoC components (iGPU, NPU, CPU), achieving 1.5× speedup on AMD Ryzen AI 7 350 vs. iGPU-only baseline
- Profiled and optimized hybrid attention based large language models for NPU-specific constraints in initial PoC; continuing development remotely

Meesho

Feb. 2022 – Aug. 2024

Data Scientist / Supervisor: [Abhay Shukla](#) and [Debdoot Mukherjee](#)

Bangalore, India

- Architected deep learning based personalized ranking systems across different components of the applications to bring **2.7% NMV per user** impact. Received **best team award**.
- Developed a quality-based re-ranking framework, reducing low-rated catalog views by 3% and post-order low ratings by 16%, leading to **0.40% retention** and **5% NMV per user growth** in 1.5 years.
- Devised a representation learning model, utilizing multi-modal textual and visual inputs to learn long-term user and item representation, which **doubled the relevance** and boosted **net conversion rate by 1.1%**.
- Personalized different widgets and widget groups of homepage feed to improve the **net conversion rate by 2%**. Received **best team award**.

Qure AI

Dec. 2021 – Feb. 2022

AI Research Scientist / Supervisor: [Sasank Chilamkurthy](#)

Bangalore, India

- Architected Vision Transformers (ViT) based segmentation framework for **automated ASPECT score prediction** in stroke severity assessment, deployed at multiple hospitals across globe.
- Improved MCA anatomy and infarct segmentation accuracy from 34% to 68%, achieving radiologist-level performance, published in **ECCV 2022**.

UNSW Sydney

May 2021 – Dec. 2021

Machine Learning Researcher (remote) / Supervisor: [Prof. Rohitash Chandra](#)

Sydney, Australia

- Designed a comprehensive framework for the thematic analysis of the Bhagavad Gita and the Upanishads using sentence transformer models like MPNet, and DistilBert. Published research findings in PLOS ONE Journal.
- Led the development of a framework for the evaluation of Google Translate for Sanskrit language.

Wadhwani AI

May 2020 – Jul. 2020

Research Intern / Supervisor: [Dr. Alpan Raval](#) and [Dr. Rahul Panicker](#)

Mumbai, India

- Developed a multi-task-learning framework for automated diagnosis of Tuberculosis using chest ultrasound videos.
- Optimized different CNN models based backbone for multi-task architecture, achieving **97% ROC-AUC**

University of Illinois Urbana-Champaign

May 2019 – Jul. 2019

Research Intern / Supervisor: [Prof. Rohit Bhargava](#)

Urbana-Champaign, USA

- Developed framework for **automated annotation of spectral tissue images** using Mutual Information and CNN based Image Registration methods.
- Formulated supervised and unsupervised learning based methods for automated rejection framework for distorted bounding box.

TECHNICAL SKILLS

Languages: Python, CUDA, HIP, MLIR-AIE, Scala, C/C++, SQL, MATLAB, Bash

Frameworks: PyTorch, TensorFlow, Django, Flask, FastAPI, Spark, PySpark, Hadoop, OpenCV

Developer Tools: Git, Linux, CI/CD, Agile, AWS, Google Cloud, Kafka, Wandb, MLflow