Shuvendu Roy

LinkedIn | Webpage | bikash.shuvendu@gmail.com | +1 (343) 580-8376

SUMMARY

AI Scientist with 8+ years of research and industry experience in data science and machine learning, with:

- Expertise in large language models (LLMs), generative models, and multi-modal learning.
- Proven track record of leading AI projects from R&D to deployment at Google Research, Borealis AI, and Vector Institute.
- Experience collaborating with cross-functional teams to integrate AI into enterprise applications.
- Strong skills in designing AI solutions using Python, PyTorch, large-scale datasets, and foundation models.
- Research excellence with publications in top-tier venues, including ICLR, AAAI, TMLR, and ICASSP.

Work Experience

Machine Learning Research Intern Jan 2025 – Present
Borealis AI Toronto, ON, Canada

- Developed an efficient LLM inference pipeline which reduced the inference latency by 90% while improving performance by 3.5 points.
- Optimized foundation model for real-time AI applications, improving scalability and robustness in production.
- Improved inference efficiency for long-context tasks like retrieval-augmented generation (RAG), reducing compute costs while maintaining performance.

Applied Machine Learning Intern Jan 2024 – Dec 2024
Vector Institute for AI Toronto, ON, Canada

- Led the development of a multi-modal foundation model for healthcare with a focus on learning from limited paired data, significantly reducing data annotation costs. - GitHub
- Contributed to the creation of a framework for training multi-modal models and conducted benchmarking of existing methods to identify optimal approaches. - GitHub
- Proposed a novel few-shot tuning approach for vision-language models, achieving superior performance in low-resource learning scenarios compared to existing methods, published in TMLR'25.
- Achieved state-of-the-art performance in medical foundation models across multiple downstream tasks and in few-shot tuning, resulting in five publications.

Student Researcher May 2023 – Oct 2023
Google Research Montreal, QC, Canada

- Designed a strategic sampling method for self-supervised learning, cutting training costs by 80% while boosting accuracy by 2% on IMU-based activity recognition.
- Developed a few-shot class-incremental learning framework that enhanced model adaptability and stability in continual learning scenarios, resulting in a publication in TMLR'24.

Applied ML Researcher Nov 2019 – Jul 2021
Robi Axiata Limited Dhaka, Bangladesh

- Designed and deployed large-scale ML systems for personalized recommendations based on user behaviour patterns, increasing user engagement by 15%.
- Built and deployed high-performance ML models for churn prediction and usage drop detection, with 85% accuracy.
- Developed end-to-end ML pipelines, covering data collection, labelling, validation, model development, deployment, and monitoring.

Jr. Software Engineer Mar 2019 – Oct 2019
REVE Systems Ltd. Dhaka, Bangladesh

- Developed and deployed LLM-powered chatbots with retrieval-augmented generation (RAG), enhancing context awareness and response relevance for domain-specific applications.
- Contributed to the development of one of the first LLM for Bengali language, and the application of spell and grammar correction.

Research Collaborations

Workday Inc

- Collaborated with Workday on developing LLM agents for financial workflows.
- Contributed to the development of efficient LLM inference pipeline with prompt compression, resulting in 11× faster inference and two publications: one submitted to ICML'25 (under review) and one accepted at AAAI'25.

EDUCATION

Jun 2025 (expected) - **PhD** in Electrical and Computer Engineering, Queen's University, Canada

- Thesis: Unsupervised Representation Learning: Downstream Adaptation and Continual Tuning
- Developed methods for training large foundation models with limited labelled data and fine-tuning them for robust generalization under data scarcity and distribution shifts, resulting in 15+ publications in top-tier venues (e.g. ICLR, AAAI, TMLR).

Dec 2021 - **MASc** in Electrical and Computer Engineering, Queen's University, Canada

- Thesis: Unsupervised Visual Representation Learning
- Promoted to the PhD program for outstanding research contributions and academic performance.

Jan 2019 - **BSc** in Computer Science and Engineering, Khulna University of Engineering & Technology, Bangladesh

- Thesis: Facial Emotion Recognition Using Transfer Learning in Deep CNN

TECHNICAL PROFICIENCY

Programming Languages	Python, C++, R, Java
DL/ML Frameworks	PyTorch, TensorFlow, JAX, Keras, NumPy, SciPy, Scikit-learn
Version Control & Exp. track	Git, Weights & Biases, TensorBoard
Computing, Cloud, and HPC	AWS, Google Colab, SLURM
Database and Deployment	MySQL, Oracle, Apache Spark, Docker
ML expertise	LLM training, fine-tuning, RAG, deployment, multi-modal learning, domain adaptation, test-time adaptation, reinforcement learning, computer vision, recommendation system, semi-supervised learning, self-supervised learning.
Soft Skills	Technical writing (35+ publications), team collaboration, mentoring

Professional Services and Awards

Awards	First prize in the *IEEE Research Excellence Award (PhD)*, 2024; Vocational Scholarship from KUET (Academic year: 2014/15 and 2017/18)
Competition	Second place in the *System Development Project Competition* at KUET.
Reviewing Activities	Program committee member for top-tier venues, including CVPR, ICLR, ICML, NeurIPS, AAAI, ICCV, ECCV, and IEEE TPAMI.
Teaching Assistant	Courses: *Artificial Intelligence*, *Introduction to Programming*; Queen's University.

SELECTED PUBLICATIONS

A full list of publications is available at my Google Scholar Profile.

- [TMLR'25] **S Roy**, E Dolatabadi, A Afkanpour, A Etemad, 'Consistency-Guided Asynchronous Contrastive Tuning for Few-Shot Class-Incremental Tuning of Foundation Models'. - GitHub
- [AAAI'25] B Liskavets, M Ushakov, **S Roy**, M Klibanov, A Etemad, S Luke, 'Prompt Compression with Context-Aware Sentence Encoding for Fast and Improved LLM Inference'. - GitHub
- [ICLR'24] **S Roy**, A Etemad, 'Consistency-guided Prompt Learning for Vision-Language Models'. - GitHub
- [AAAI'24] **S Roy**, A Etemad, 'Scaling Up Semi-supervised Learning with Unconstrained Unlabelled Data'. - GitHub
- [TMLR'24] **S Roy**, C Park, A Fahrezi, A Etemad, 'A Bag of Tricks for Few-Shot Class-Incremental Learning'.
- [IEEE-TAFFC'24] **S Roy**, A Etemad, 'Exploring the Boundaries of Semi-Supervised Facial Expression Recognition: Learning from In-Distribution, Out-of-Distribution, and Unconstrained Data'. ***Invited paper****/ACII'22*, - GitHub
- [NeurIPS'23-W] **S Roy**, A Etemad, 'Learning Through Consistency for Prompt Tuning'. ***Spotlight*** - GitHub
- [ICASSP'23] **S Roy**, A Etemad, 'Temporal Contrastive Learning with Curriculum'.