Bishwajit Saha

Mchenry, IL, USA

bsaha205@gmail.com (+1) 5188059331 Linkedin Google Scholar Website

RESEARCH SUMMARY

Research Scientist at IBM Research with first-author publications at **ICLR 2026**, **NeurIPS 2024 Workshop** and **ICML 2023**, and two submissions under review at **ICLR 2027**. Research focus on energy-based language models, associative memory, Mixture-of-Experts, clustering and large-scale distributed training of billion-parameter LLMs. Four years of prior production-systems engineering at scale.

EDUCATION

Rensselaer Polytechnic Institute — **Troy, NY, USA**
Ph.D. in Computer Science — *January, 2021 - January, 2025*

Rensselaer Polytechnic Institute — **Troy, NY, USA**
M.Sc in Computer Science — *January, 2021 - December, 2022*

Bangladesh University of Engineering & Technology — **Dhaka, Bangladesh**
B.Sc in Computer Science and Engineering — *February, 2013 - September, 2017*

RESEARCH & PUBLICATIONS

Nima Dehmamy*, Benjamin Hoover*, **Bishwajit Saha***, Leo Kozachkov, Jean-Jacques Slotine and Dmitry Krotov*, "NRGPT: An Energy-based Alternative for GPT", **ICLR 2026**. Paper.

- Proposed NRGPT, a language model architecture integrating associative memory dynamics that reconceptualizes GPT inference as a dynamical process on an energy landscape, demonstrating strong performance across diverse tasks from algebraic reasoning to open-domain text generation.

Bharat Runwal, **Bishwajit Saha**, Benjamin Hoover, Mauro Martino, Nima Dehmamy, "Boltzmann Routing for Energy-Compatible Mixture of Experts", **NFAM Workshop @ICLR 2026**. Paper.

- Introduced Boltzmann Routing, an auxiliary-loss-free method that derives MoE expert selection from a free-energy functional, proving exact energy compatibility and a formal attention–routing duality within Energy Transformers.

Bharat Runwal, **Bishwajit Saha**, Benjamin Hoover, Dmitry Krotov, Mauro Martino, Nima Dehmamy. "Latent Test-Time Computation with Energy Transformer Components." **Under review at ICLR 2027**.

- Identified two distinct recurrent-transformer computation modes through a 9-architecture factorial study at 400M–1B scale — distributed refinement (energy attention) vs. concentrated construction (standard attention) — enabling 12.8–19% adaptive-halting FLOP savings at no quality loss.

Bishwajit Saha, Bharat Runwal, Benjamin Hoover, Mauro Martino and Nima Dehmamy, "Boltzmann-MoE: Energy-based Mixture-of-Experts with Native Routing", **Under review at ICLR 2027**.

- Derived a Mixture-of-Experts transformer from a single energy function, so attention, expert layers and the router all follow from first principles; matches a same-size Switch-MoE within 0.1–1.3 points of zero-shot accuracy, with 3–4× more consistent layer updates.

"Scaling NRGPT: Energy-based Language Models at Billion-Parameter Scale", Work-in-progress, **IBM Research**, 2026.

- Scaling NRGPT (ICLR 2026) to multi-billion parameters via distributed HPC training, targeting GPT-competitive performance while preserving energy-based associative memory dynamics.

Bishwajit Saha, Dmitry Krotov, Mohammed J. Zaki, and Parikshit Ram. "Deep Clustering with Associative Memories." **Workshop @ICLR 2025 & Workshop @NeurIPS 2024**. Paper.

- Integrated autoencoder and associative memories together to perform end-to-end differentiable deep clustering (**DClAM**) which simultaneously learns effective representations and clusters in the latent space with a simplified deep clustering formulation and achieves significant improvement.

Bishwajit Saha, Dmitry Krotov, Mohammed J. Zaki, and Parikshit Ram. "End-to-end Differentiable Clustering with Associative Memories." **ICML 2023**. Paper.

- Uncovered a novel connection between the Associative Memories (AM) dynamics and the inherent discrete assignment necessary in clustering to propose a novel unconstrained continuous relaxation of the discrete clustering problem, enabling end-to-end differentiable clustering with AM (**ClAM**).

Rownak Ara Rasul, Promy Saha, Diponkor Bala, S. M. Karim, Ibrahim Abdullah, and **Bishwajit Saha**. "An Evaluation of Machine Learning Approaches for Early Diagnosis of Autism Spectrum Disorder." **Healthcare Analytics 2023**. Paper.

Ibtehaz, Nabil, Shafayat Ahmed, **Bishwajit Saha**, M. Sohel Rahman, and Md Shamsuzzoha Bayzid. "NORTH: a highly accurate and scalable Naive Bayes based ORTHologous gene clustering algorithm." bioRxiv (2019). Paper.

WORK EXPERIENCE

Research Scientist — **Cambridge, MA, USA**
IBM Research — *February 2025 - Present*

Research Topic: Energy-based Language Models, Large-scale Distributed Training, Mixture-of-Experts.

Graduate Research Assistant — **Troy, NY, USA**
Rensselaer Polytechnic Institute — *January 2021 - January 2025*

Research Topic: Clustering, Associative Memories, Machine Learning.

Summer AI Researcher (2 internships) — **Cambridge, MA, USA**
IBM Research — *May 2022 & May 2023*

Research: Differentiable Clustering with Associative Memories (ICML 2023, NeurIPS 2024 Workshop).

Senior Engineer — **Dhaka, Bangladesh**
R&D, Technology Division, ***Bkash Ltd*** — *May 2019 - January 2021*

- Built loan and payment systems for bKash, Bangladesh's largest fintech platform (65M+ users).

Senior Software Engineer — **Dhaka, Bangladesh**
REVE Systems Ltd — *October 2017 - April 2019*

- Developed government-scale web platforms (OISF, BCC Billing) using Java Spring Boot and MySQL.

Lecturer, Department of Computer Science and Engineering — **Dhaka, Bangladesh**
Bangladesh University of Business and Technology — *August 2017 - September 2017*

TECHNICAL SKILLS

Languages	Python, C, C++, Java, CUDA, Dart, Matlab, Prolog, Haskell, Scala
ML & DL	PyTorch, JAX, TensorFlow, Scikit-learn, FSDP
Infrastructure	Multi-node GPU Training, LSF/Slurm, HPC Clusters, AWS, Docker
Models & Methods	Transformers, Energy-based Models, Associative Memories, MoE, LoRA

PROFESSIONAL SERVICE & HONORS

- **Reviewer**: NeurIPS 2023–2026, ICML 2024–2026, ICLR 2025–2027, AAAI 2026, AISTATS 2025–2026.
- Undergraduate Merit Position **25**/150, Department of CSE, BUET. Govt. Board **Scholarship** 2007–2017.