Arnab Sen Sharma

sensharma.a@northeastern.edu

arnab-api.github.io

github.com/arnab-api

About me

I am a PhD candidate in the Interpretable Neural Networks lab at Northeastern University. I am interested in understanding the inner workings of large language models.

Education

Ph.D. Candidate, Computer Science September, 2022 - Present

MS, Computer Science (4.00/4.00) December, 2024

Khoury College of Computer Sciences, Northeastern University, Boston

BSc(Engg.) in Computer Science and Engineering (3.91/4.00) November, 2018

Shahjalal University of Science and Technology

Selected Publications

- **Arnab Sen Sharma**, Giordano Rogers, Natalie Shapira, David Bau. "LLMs Process Lists With General Filter Heads". Under review for ICLR 2026. `filter.baulab.info`
- **Arnab Sen Sharma**, David Atkinson, and David Bau. "Locating and Editing Factual Associations in Mamba", COLM-2024. `romba.baulab.info`.
- Evan Hernandez*, **Arnab Sen Sharma***, Tal Haklay, Kevin Meng, Martin Wattenberg, Jacob Andreas, Yonatan Belinkov, and David Bau. "Linearity of Relation Decoding in Transformer Language Models", ICLR-2024 (**Spotlight**). `lre.baulab.info`. (* Equal contribution)
- Eric Todd, Millicent L Li, **Arnab Sen Sharma**, Aaron Mueller, Byron C Wallace, and David Bau. "Function Vectors in Large Language Models", ICLR-2024. `functions.baulab.info`.
- Kevin Meng, **Arnab Sen Sharma**, Alex Andonian, Yonatan Belinkov, and David Bau. "Mass-Editing Memory in a Transformer", ICLR-2023 (top 25% paper). `memit.baulab.info`.
- Prakash et al. (including **Arnab Sen Sharma**). "Language Models use Lookbacks to Track Beliefs". Under review for ICLR 2026. `belief.baulab.info`
- Mueller et al. (including **Arnab Sen Sharma**). "The Quest for the Right Mediator: A History, Survey, and Theoretical Grounding of Causal Interpretability"
- Fiotto-Kaufman et al. (including **Arnab Sen Sharma**). "NNsight and NDIF: Democratizing Access to Foundation Model Internals", ICLR-2025.

Teaching Experience

Northeastern University

Graduate TA

- DS 4440 - Practical Deep Networks Spring, 2024

Shahjalal University of Science and Technology

Lecturer July, 2019 - July, 2022

- **Taught Courses:** Database Management Systems, Software Engineering and Design Patterns, Web Tech. See details in my faculty profile.
- Mentoring undergrads working on Data Science and Machine Learning projects.
- Took workshops on advanced data structures and algorithms to train competitive programmers. Also served as coach for competitive programming teams.

Samsung R&D Institute, Bangladesh

- Conducted KT (knowledge transfer) sessions on BIT, hashing, and other data structures to coach my colleagues for Samsung's professional certificate test.

Professional Experience

Samsung R&D Institute, Bangladesh
Software Engineer I — September, 2018 - June, 2019

- ◇ Implemented core features for adapting chromium for Samsung wearables operating on Tizen OS. I specifically focused on *multimedia* components and V8 JavaScript engine performance.
- ◇ Implemented cross-platform file sharing protocols in Samsung Files app, enabling seamless network transfers between Android devices.
- ◇ Designed and developed apps for Android smartphones, focusing on sensor data collection and analysis.
- ◇ Developed gesture-based authentication and action-binding systems for Android and Tizen wearables.

ML Alignment and Theory Scholars (MATS 7.0)
Apprentice — October, 2024 - November, 2024

- ◇ Worked on a project aimed at enabling a Language Model to explain its thinking state by introspecting its own inner representations — turning the LM into a *talkative* and *interactive* probe into its own inner reasonings. [code, report]

Honors and Awards

- **Northeastern Graduate Assistantship**, Northeastern University. 2022-Present
- **ACCESS Compute Credits — Discover Tier** for *Investigating Compositional Reasoning with Parametric Knowledge in LLMs*. July, 2025 to July 2026
- **Khoury Startup Fund**, Northeastern University PhD fellowship worth **$5000**. 2022
- Co-authored and served as the PI on a project entitled *"A Deep Learning Approach for Hate Speech Detection in Bangla Text and a Benchmark Dataset in Bangla"*. The proposed project won a research grant worth **2,30,000 BDT** from SUST Research Center (Project ID: AS/2020/1/26). [report] 2020
- 3^{rd} place in SAMSUNG Research ROBOT HACKATHON-2018, awarded **80,000 BDT**. 2018
- Monetary award of **30,000 BDT** for passing Samsung's Professional Certificate Test. 2018
- **Notable performances on onsite competitive programming contests**
  - ◇ ACM ICPC ASIA Regional, Dhaka Site 2017, Rank: 21, (SUST_PeakSeeker)
  - ◇ ACM ICPC ASIA Regional, Dhaka Site 2016, Rank: 11, (SUST_DeapThunder)
  - ◇ IUBAT IUPC, Rank: 4 , (SUST_DeapThunder)
  - ◇ BSCCL National Programming Contest (Hackathon) at UITS, Rank: 6, (SUST_1)
  - ◇ LU Inter University Programming Contest 2016, Rank: 4, (SUST_BlackJAM)

Skills Summary

- **Languages**: Python, C++, java
- **ML Libraries**: PyTorch, TensorFlow
- **Web Development**: ReactJS, Flask, Django
- **Others**: GIT, SQL, NoSQL(MongoDB), Android Studio, Unity2D, JASP

Service

- Served as a reviewer for conferences ICLR (2024, 2025, 2026), NeurIPS (2024, 2025), COLM (2024, 2025), ICML (2025), BlackboxNLP (2025).
- Problem Setter, judge, and alternate solution writer for several programming contests; including NEUB Junior Inter University Programming Contest-2017, Varendra University Inter University programming contest-2017, and SUST Intra University Programming Contest-2017.
- Served as a member of Undergrad Final Year Thesis/Project Evaluation Committee, Session 2019-20, 2020-21.
- Served as a core member of the Technical Sub-committee, SUST Admission Process in years 2020 and 2021.

Open Source Contributions: Contributed to `baukit` and `nnsight`.