

Kristen Bestavros

📍 Boston, MA ✉ kbest@bu.edu ☎ 508-499-9112 🔗 [linkedin.com/in/kristen-bestavros](https://www.linkedin.com/in/kristen-bestavros)
🐙 github.com/kristenbestavros

RESEARCH INTERESTS

Adversarial robustness of large language models, particularly the tension between safety and helpfulness; automated red-teaming methods for identifying and addressing model vulnerabilities; novel approaches to making LLMs more robust against jailbreak attacks. Broader interests in emergent behavior in multi-agent systems, computational modeling of information dynamics, and fairness and value alignment in AI.

EDUCATION

Boston University

Boston, MA

Master of Science in Data Science

Expected May 2027

GPA: 3.78

Relevant Coursework: Deep Learning for Data Science; Theory and Applications of Large Language Models; Stochastic Methods for Data Science Algorithms; Tools for Data Science

Boston University, Kilachand Honors College

Boston, MA

Bachelor of Science in Data Science, Minor in Statistics

May 2025

GPA: 3.71

Relevant Coursework: Machine Learning; Reinforcement Learning; Algorithms; Statistics; Data Mechanics; Data Science Ethics; Practicums in Civic Tech & Software Engineering

CURRENT RESEARCH

Adversarial LLM Safety Training Project

Spring 2026 – present

Boston University Center for Computing & Data Sciences

Studying LLM robustness under adversarial pressure using an automated red-teaming loop where an attacker and defender co-evolve through LoRA fine-tuning, with LlamaGuard as an independent evaluator.

- Attacker improves via in-context learning, expert iteration, and iterative self-critique; defender improves via incremental fine-tuning with validation mixing to prevent over-refusal
- Ablating improvement mechanisms and comparing three defender strategies (refusal, reinterpretation, selective omission) for robustness under adaptive attack

RESEARCH EXPERIENCE

Computational Social Science Intern

May 2024 – May 2025

Boston University, Center for Computing and Data Sciences | Mentor: Prof. Wesley Wildman

- Developed agent-based computational simulations modeling how ideological beliefs form, spread, and polarize within populations, incorporating theoretical frameworks from social psychology and cognitive science.
- Produced literature review published in *Acta Sociologica* analyzing ideological dynamics in Norway (2014–2023); developed new simulations incorporating these theoretical insights.

UROP Student Researcher

Sep 2021 – Aug 2022

Boston University, Department of Physics | Mentor: Prof. David Campbell

- Conducted computational research on nonlinear dynamical systems (Fermi–Pasta–Ulam–Tsingou problem); wrote Bash scripts for parallel computation on BU's shared computing cluster.

PROJECTS

Computational Modeling of Political Polarization – Senior Capstone Project Spring 2025

Kilachand Honors College Keystone Project

Built an agent-based model simulating information dissemination and ideological evolution using a cognitive-motivational framework (Jost et al.); studied how individual-level cognitive biases produce emergent population-level polarization patterns.

Affordable Housing Applications – CHAPA Client Project Fall 2025

SPARK, Boston University Center for Computing & Data Sciences

Analyzed affordable housing application data for CHAPA, a Massachusetts-based housing nonprofit, to study geographic mobility, demographic representation, and access disparities using statistical and geospatial methods.

Anagrammer – Personal Project Spring 2026

Inspired by BU CDS class, Theory and Applications of LLMs

Implemented character-level Markov chain trained on real names to generate name-like anagrams; features configurable temperature, name templates, and sampling options via command-line interface.

WORK EXPERIENCE

Teaching Assistant – Data, Society, and AI Ethics Fall 2025 – present

Boston University, Center for Computing and Data Sciences

- Facilitate discussion sections covering fairness and bias in ML systems, value alignment, automated decision-making, and the ethical and policy dimensions of AI deployment.
- Mentor students on ethical reasoning; evaluate essays; co-develop interactive course materials.

TECHNICAL SKILLS

AI & Machine Learning: Transformer architectures; LLM fine-tuning (LoRA); reinforcement learning; deep learning; regression and classification; stochastic methods; interpretability

Languages: Python, SQL, R, Java, Rust

Mathematics: Statistics; probability; linear algebra; calculus; stochastic optimization

Tools & Frameworks: PyTorch; scikit-learn; NumPy; pandas; Docker; Podman; Git; GitHub; Azure

Additional Experience: Agent-based simulation modeling (AnyLogic); geospatial analysis; data visualization

PUBLICATIONS

McDonnell Maayan, Gabriel; **Bestavros, Kristen**; Wildman, Wesley. "Woke/Anti-Woke Dynamics in Norway, 2014-2023." *Acta Sociologica*. doi:10.1177/00016993261432785

Bestavros, Kristen. "Deciphering FPUT Metastable State: An Investigation of Resonance in the FPUT Spectra." Poster, 25th Annual Undergraduate Research Symposium, Boston University, Oct 2022.

Bestavros, Kristen. "Game Development to Gameplay: Women in the Video Game Industry." *Kaleidoscope*, vol. 1, 2022, p. 42.