Jacob E. Brinton

jbrin@bu.edu · jebrinton.github.io · Boston, Massachusetts
linkedin.com/in/jbrinton · github.com/jebrinton

EDUCATION

Boston University *May 2026*
B.A./M.S. in Mathematics and Computer Science GPA: 3.89 (B.A.); 4.0 (M.S.)
- **Relevant coursework:** Interpretable Machine Learning, Directed Study in LLM Translation Interpretability, Principles of ML, Natural Language Processing, Stochastic Processes, Causal Inference, Abstract Algebra, Advanced Algorithms, Intro to AI, Foundations of Mathematics, Logic and Programming Languages, Theory of Computation

Utah State University *Aug 2022 – May 2023*
- **Relevant coursework:** Honors Calculus II & III, Chemistry, Physics, Programming in C, CAD GPA: 4.0

EXPERIENCE

Technical Fellowship Reviewer *May 2026 – Present*
Constellation Remote Contractor
- Screened and reference checked 200+ applicants to the Anthropic Fellows Program and the Astra Fellowship.

Technical Interviewer *March 2026 – Present*
FAR.AI Remote Contractor
- Interviewed research scientist and engineer candidates across a variety of ML, math, and programming questions.

Executive Director *May 2025 – Present*
Boston University AI Safety & Alignment Boston, MA
- Increased membership by 10x; members accepted to ERA, LASR, MARS, SPAR; Harvard and MIT graduate programs.
- Presented on a wide range of AI risks to over a dozen courses, reaching a total enrollment of 1,000+ students.
- Negotiated hackathon partnerships; organized 20 events, from guest speakers to interuni. socials and research meetings.
- Wrote and facilitated technical AI safety curriculum; organized 6 discussion cohorts and 2 upskilling programs.

AI Safety Fieldbuilding Mentor *Dec 2025 – May 2026*
Pathfinder Fellowship (Kairos) Remote
- Guided Columbia University's AI safety group directors on outreach strategy, lecture series, and leadership transitions.

Research Fellow *Dec 2025 – Present*
Mentorship for Alignment Researchers (Cambridge AI Safety Hub) Cambridge, UK
- Developed model diffing interpretability methods for model safeguard checkpoints; mentored by Prof. Zhijing Jin.

Teaching Fellow, Course Assistant *Aug 2024 – Dec 2025*
Boston University College of Arts and Sciences, College of Engineering Boston, MA
- Taught in lab sections and office hours for linear algebra, computer systems, and calculus courses.

Machine Learning Research Intern *May 2023 – Aug 2023*
NSF Research Experiences for Undergraduates, Montana State University Bozeman, MT
- Designed UAV flight experiment, collecting 3 hours of audio; applied signal processing and ML methods for detection.

Undergraduate Researcher *Dec 2022 – May 2023*
DIRECT Laboratory (Utah State University) Logan, UT
- Wrote Python-based publishers and subscribers in ROS for robot pathfinding and camera movement.

PROJECTS & PUBLICATIONS

The Interlingua Hypothesis: LLMs Translate via a Latent Task-agnostic Feature Space *Mar 2025 – Present*
- Experimented on whether LLMs translate via shared multilingual representations as opposed to pairwise mappings.
- Used gradient-based methods to identify and then causally validate components implicated in translation.
- Quantified multilinguality of features' receptive fields and output spaces, and representation similarity across layers.
- Presented posters at the New England Mechanistic Interpretability Workshop and Boston University's UROP Symposium.

Cross-Language Synonym Analysis and Scoring *Sep 2024 – Jan 2025*
- Devised and empirically computed metrics for cross-language synonym closeness and translation flexibility.

Assessing LWIR and Visual Sensor Modalities for UAV Detection and Tracking *May 2024*
2024 IEEE Intermountain Engineering, Technology and Computing (IETC), pp. 141–146.
- Coauthored paper assessing long-wave infrared (LWIR) and visual sensor modalities for UAV detection.

Tetris Reinforcement Learning Agent *Apr 2024 – May 2024*
- Designed Q-learning agent with heuristic-guided exploration and reward shaping; placed 2nd in 150-person course.

SKILLS

Technical: Python, PyTorch, HuggingFace, nnsight, Weights & Biases, Claude Code, OCaml, C.
Languages: English (Native), Spanish (Fluent).