GRACE PROEBSTING

proebsting.g@northeastern.edu

Research Interests

I am interested in LLM oversight, interpretability, and debugging.

- **Oversight**: How can we scalably discover LLM behaviors and errors?
- **Interpretability**: How do LLMs "work" at the latent level?
- **Debugging**: How can we use interpretability tools to audit and edit models?

Education

Northeastern University Sep 2025 – Present
Ph.D. in Computer Science — Advisor: David Bau

Haverford College Aug 2021 – May 2025
B.S. in Computer Science — GPA: 4.0
Selected coursework: Machine Learning, Real Analysis, Linear Algebra, Information Theory, Linear Optimization, Probability Theory, Statistical Methods, Syntax, Theory of CS, Multivariable Calculus, Math for ML, Compiler Design

Research Experience

Northeastern University – PhD Student – Advisor: David Bau Sep 2025 – Present

- Researching ML interpretability in the Bau Lab.

Microsoft Research – Research Intern – AI Frontiers May 2025 – Aug 2025

- Research intern for Microsoft Research AI Frontiers.

Haverford College – Undergrad Thesis – Advisor: Sorelle Friedler Sep 2024 – May 2025

- Wrote a thesis on how interpretability techniques (e.g. linear probes) could be used to debias LLMs.
- Performed a literature review on probing representations of LLMs and feature importance metrics.

Microsoft Research – Research Intern – Advisors: Adam Fourney and Gagan Bansal May 2024 – Aug 2024

- Studied whether LLMs can accurately detect errors made by LLM-based agents (e.g. on GAIA benchmark tasks).
- Created an LLM-based automatic error identification and visualization tool, used by MSR's Human AI-eXperiences (HAX) team to perform ablation studies on their LLM-based agents.
- Solicited human annotations of LLM-agent errors from expert developers within MSR to establish a task baseline.

Bryn Mawr NLP Lab at Bryn Mawr College – Research Assistant – Advisor: Adam Poliak Aug 2023 – May 2024

- Investigated whether LLM-generated and human crowd-sourced NLI datasets had similar spurious annotation artifacts.
- Wrote a first-authored paper (accepted to COLING 2025) showing severe artifacts in LLM-generated NLI.
- Developed dataset generation pipeline to create NLI datasets using four LLMs and fine-tuned classification models.

University of Maryland, College Park – Research Assistant – Advisor: Marine Carpuat June 2023 – Aug 2023

- Worked under Drs. Marine Carpuat and Eleftheria Briakou to improve Wikipedia translations for African languages.
- Fine-tuned a Transformer-based token-level quality estimation (QE) model for English to Hausa Wikipedia translations.
- Developed a data processing pipeline to convert Wikipedia translations and edit histories to a synthetic QE dataset.

Publications & Preprints

Grace Proebsting, Oghenefejiro Isaacs Anigboro, Charlie M Crawford, Danaé Metaxa, and Sorelle A. Friedler. Identity-related Speech Suppression in Generative AI Content Moderation. *ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization (EAAMO)*. 2025.

Grace Proebsting and Adam Poliak. Biases in Large Language Model-Elicited Text: A Case Study in Natural Language Inference. *Proceedings of the 31st International Conference on Computational Linguistics (COLING)*. 2025.

Grace Proebsting and Adam Poliak. Hypothesis-only Biases in Large Language Model-Elicited Natural Language Inference. *arXiv preprint*. 2024. https://arxiv.org/pdf/2410.08996. (Short version of the COLING paper above.)

Adam Fourney, Gagan Bansal, Hussein Mozannar, Cheng Tan, Eduardo Salinas, Erkang (Eric) Zhu, Friederike Niedtner, Grace Proebsting, Griffin Bassman, Jack Gerrits, Jacob Alber, Peter Chang, Ricky Loynd, Robert West, Victor Dibia, Ahmed Awadallah, Ece Kamar, Rafah Hosn, Saleema Amershi. Magentic-One: A Generalist Multi-Agent System for Solving Complex Tasks. *arXiv preprint*. 2024. https://arxiv.org/pdf/2411.04468

Teaching Experience

Haverford College – Teaching Assistant – Machine Learning (CMSC 360) Jan 2024 – May 2024

Haverford College – Teaching Assistant – Foundations of Data Science (CMSC 260) Aug 2023 – Dec 2023

Service

Secondary Reviewer – EMNLP (with Prof. Adam Poliak) July 2024

Work Experience

Effective Altruism Communications Fellowship – Fellow May 2022 – Aug 2022

- Proposed, researched and drafted blog posts for Giving What We Can.
- Revised the "Introduction to Effective Altruism" curriculum used by EA university groups internationally.

Casa Alitas – Intern May 2020 – Mar 2021

- Created web applications for Casa Alitas, a charity that serves asylum-seeking migrants.
- Proposed and designed a web-based medical questionnaire for migrants who speak Mam or K'iche'.

Skills & Extracurriculars

Outreach Math Inclusion and Diversity Committee (Lead Organizer)
Libraries & Frameworks Hugging Face, PyTorch, Pandas, NumPy, TransformerLens, scikit-learn
Hobbies Lindy Hop swing dance, meditation, arthouse film, and blogging about ML!