Benjamin D. Hayum

bhayum@wisc.edu • 82 Farbrook Drive, Short Hills, NJ 07078 • Mobile: (973) 590-1596

RESEARCH EXPERIENCE

Sharon Li Laboratory ***Madison, WI***

Undergraduate Research Assistant *Summer 2023 - Present*

- Spearheaded technical research initiative to investigate the susceptibility of Large Language Models (LLMs) to negative behavioral steering via Reinforcement Learning from Human Feedback. This work culminated in a paper titled "How Does RLHF Alignment Shift Behavioral Distributions? Distinguishability and Steerability," submitted to the International Conference on Learning Representations.
- Leading the development of a benchmark for measuring the robustness of human preferences elicitation in Large Language Models through two-turn conversation. Evaluating the benchmark on frontier models with a multifaceted approach that includes system prompting, representation engineering, and goal-directed dialogue training.

Banks Laboratory ***Madison, WI***

Undergraduate Research Assistant *2021 - 2023, Summer 2022*

- Visualized neurometrics across various plot forms in interpretable manners to reveal distinctions across condition states with generalizable code to be used for any future dataset brought into our brain data analysis pipeline.
- Gained deep experience in tracking across many minute details in the data, metadata, and code pipeline to ensure that the code was producing the intended output.
- Interpreted cutting edge neurometrics such as gamma envelope correlations, alpha weighted phase lag index, Lemel-Ziv complexity, diffusion map embeddings, effective dimensionality, and multivariate autoregressive models.
- Attended Society for Neuroscience Conference in San Diego to present a poster of my research titled "Intracranial electrophysiological signatures of delirium in neurosurgical patients".

RESEARCH PAPERS

"How Does RLHF Shift Behavior Distributions? Distinguishability and Steerability" - Sharon Li Laboratory
"Clinical and Intracranial Electrophysiological Signatures of Post-operative and Post-ictal Delirium" - Banks Laboratory

PRESENTATIONS

"Navigating the Frontier of AI Development: Balancing Transformative Potential and Emerging Risks" Introduction for Seb Krier from Google DeepMind – Wisconsin AI Safety Initiative Speaker Event
"Governing Frontier AI: Policy Challenges and Strategies" Introduction for Caleb Withers from Center for a New American Security – Wisconsin AI Safety Initiative Speaker Event
"Detecting Danger in AI: What, Why, and How" Introduction for Thomas Broadley from ARC Evals – Wisconsin AI Safety Initiative Speaker Event
Generative AI Panel Discussion - Wisconsin School of Business
"Evaluating Language Model Agents on Realistic Autonomous Tasks" - AI Alignment Paper Reading Group
"Fundamental Limitations of Alignment in Large Language Models" – AI Alignment Paper Reading Group
"ChatGPT and the Mind" Introduction for Professor Gary Lupyan – Wisconsin AI Safety Initiative Speaker Event
Sparks of AGI: Early Experiments with GPT-4 – Wisconsin AI Safety Initiative Discussion
Intro to AI Safety (2x) – Wisconsin AI Safety Initiative Opening Event
Intracranial Electrophysiological Signatures of Delirium in Neurosurgical Patients – Society for Neuroscience Conference
Post-operative Delirium and Brain Complexity Fluctuations – Biology Independent Project Poster Session
Sex differences and serotonergic mechanisms in the behavioural effects of psilocin – Banks Lab Journal Club Neurobiology Paper Discussion

HONORS AND EVENTS

University of Wisconsin - Madison

- Dean's List - Fall 2020, Spring 2021, Fall 2021, Spring 2022, Fall 2022, Spring 2023, Spring 2024
- Merit-based Near Full Scholarship
- Invited to speak on Business Generative AI panel by the Associate Dean of the School of Business
- Invited to and Attended the Society for Neuroscience Conference in San Diego to present a poster on my research
- Invited to and Attended Wisconsin's Psychedelic Symposium Event

Millburn High School

- Graduated with High Honors

EDUCATION

University of Wisconsin-Madison ***Madison, WI***
Computer Sciences PhD Student *2024-Present*

- Committed to the University of Wisconsin-Madison for a 5 year PhD in Computer Sciences

University of Wisconsin - Madison ***Madison, WI***
Bachelor of Science in Data Science and Neurobiology, Certificate in Computer Science *2020 - 2024*

- Selected among 30K+ applicants to receive merit-based full scholarship
- Cumulative GPA: 3.919/4.00
- Programming Languages: Python, MATLAB, R, Julia, Java, SQL, Shell, Docker
- Past & Present Relevant Coursework: Deep Learning and Generative Models, Intro to Artificial Intelligence, Intro to Optimization, Intro to Big Data Systems, Proof Based Linear Algebra, Intro to Theory of Probability, Data Modeling II, Data and Algorithms: Ethics and Policy

Millburn High School ***Millburn, NJ***

- Weighted GPA: 4.23/4.00 *2016 - 2020*
- ACT Superscore: 35/36

LANGUAGE ABILITIES

English: Fluent native speaker
Spanish: Semi-fluent speaker

OTHER EXPERIENCE

Wisconsin AI Safety Initiative (WAISI) ***Madison, WI***
Director and Founder *2023-Present*

- Led strategic development, built organizational infrastructure, and designed leadership framework, resulting in a high-functioning, sustainable student organization that will educate and build skills for talented students in AI Safety for years to come at UW-Madison.
- Organized outreach campaigns, leading to 331 interest form sign ups and around 115 students taking our "AI Safety Fundamentals" introductory program, including 35 of whom I've personally taught, across only 3 semesters.
- Led development of "Safety Scholars" advanced program of 36 students, including 6 PhD students and 3 master's students, reading and discussing new frontier AI papers together weekly and collaborating on new safety research agendas.
- Currently leading development of "Research Network" program to support the research of our Safety Scholars and build infrastructure for members to seek feedback and collaborators
- Spoke on Generative AI Panel with invite by the Associate Dean of the Wisconsin School of Business to give voice to the side of AI Safety against industry professional consultant from VelocityAI and lawyer from Intuit Mailchimp.
- Received University Group Organizer Fellowship grant from Open Philanthropy of $22,300 between group and personal funding

Stanford Existential Risk Initiative **Virtual Program**
Machine Alignment Theory Scholar *Summer 2023*

- Participated in a month and a half of workshops through John Wentworth's Agent Foundations SERI-MATS
- Decomposed and mathematically formalized important AI Alignment concepts from first principles such as information channels, stable equilibria, optimization, models, boundaries, and control
- Tackled coding a model basin volume estimator in gradient space using the parameter's Hessian matrix in computationally tractable forms

Ultraspeaking **Virtual Program**
Public Speaking Course Student *Summer 2023*

- Learned and executed strategies over 4 weeks for concise and clear speaking, direct and empathetic communication, gaining influence and buy-in, as well as preparing and delivering effective presentations
- Practiced improvising compelling speeches in response to randomly generated prompts under pressure and observation

Effective Altruism Student Organization ***Madison, WI***
Leadership Team *2021 - Present*

- Deepened understanding and knowledge of AI Catastrophic Risk by attending multiple conferences including the Stanford Existential Risk Initiative Conference and EA Global Bay Area Conference which featured speakers from OpenAI, Anthropic, Redwood Research, Open Philanthropy, Far AI, among others, and networking with other over 20 attendants.
- Hosted weekly meditation group to encourage club members to care for and better their mental wellbeing.