Benjamin D. Hayum

bhayum@wisc.edu • LinkedIn • GitHub • WAISI • Personal Website

RESEARCH EXPERIENCE

Grigoris Chrysos Laboratory ***Madison, WI***
Independent PhD Research Assistant *Summer 2024 - Present*

- Executed comprehensive literature review on Sparse Autoencoders as Interpretability tools.
- Iterated through several project idea formulations.
- Pivoted towards Tensor-based Interpretability approaches using μ-MoE related methods.

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 ICLR
- 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.

Matthew 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

"A Conversation on AI and Society" Panel with Professor Annette Zimmermann (Philosophy), Jerry Zhu (CS), and David Shaffer (Education) – Wisconsin AI Safety Initiative Panel Speaker Event

"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 METR – 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

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
- Relevant Coursework: Linear Optimization, Mathematical Methods in Data Science

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
- 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***
Founder and Research Lead *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 ML Safety for years to come at UW-Madison.
- Organized outreach campaigns, leading to over 150 students taking our "AI Safety Fundamentals" introductory programs, including 35 of whom I've personally taught, across only 4 semesters.
- Led development of "Safety Scholars" advanced programs of 50 students, including 13 PhD students, reading and discussing new frontier ML Safety, Transparency, Fairness, and Reliability papers together weekly.
- Currently leading development of "Research Network" program of office hours and co-working hours to support the research of our Safety Scholars and build interfaces for mutual benefit.
- Hosted speaker events with staff from Anthropic, Google DeepMind, and METR as well as with Professors Zimmermann (Philosophy), Zhu (CS), Shaffer (Education), Kennedy (History), and Lupyan (Psychology).

Machine Alignment Theory Scholar **Virtual Program**
Summer Intern *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 preference stability, optimization/goodharting, value formation, and decision boundaries.
- 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