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 (SAE) as Interpretability tools.
- Iterated through several formulations of novel SAE architectures designed to extract features along statistical complexity spectra.
- Proposed iterative supervision templates to map out latent features, enabling a more precise proxy for the extraction of semantic representations.
- Pivoted towards Tensor-based Interpretability approaches using μ-MoE related methods.

Sharon Li Laboratory ***Madison, WI***
Undergraduate Research Assistant *Summer 2023 - Present*

- Pioneered research investigation examining the vulnerability of Large Language Models (LLMs) to behavioral manipulation at the cause of Reinforcement Learning from Human Feedback (RLHF).
- Developed empirical analytical framework to quantitatively assess how alignment techniques systematically alter model behavioral distributions, potentially opening up vulnerabilities in adversarial attacks.
- Authored rigorous research paper, "How Does RLHF Alignment Shift Behavioral Distributions? Distinguishability and Steerability," submitted to the prestigious International Conference on Learning Representations (ICLR).

Matthew Banks Laboratory ***Madison, WI***
Undergraduate Research Assistant *2021 - 2023, Summer 2022*

- Designed and implemented advanced visualization techniques to uncover nuanced distinctions across brain regions and condition states, creating a generalizable data analysis pipeline.
- Demonstrated exceptional precision in data management, meticulously tracking intricate details across datasets, metadata, and computational workflows to ensure rigorous output validation.
- Proficiently analyzed complex neurometrics, including advanced techniques such as: gamma envelope correlations, alpha-weighted phase lag index, lemel-ziv complexity, diffusion map embeddings, effective dimensionality analysis, and multivariate autoregressive modeling
- Presented research findings at the prestigious Society for Neuroscience Conference in San Diego.
- Published research paper, "Clinical and intracranial electrophysiological signatures of post-operative and post-ictal delirium," in the peer-reviewed Clinical Neurophysiology Journal.

RESEARCH PAPERS

"How Does RLHF Shift Behavior Distributions? Distinguishability and Steerability" *Benjamin David Hayum, Quentin Feuillade Montixi, Yixuan Li*

"Clinical and Intracranial Electrophysiological Signatures of Post-operative and Post-ictal Delirium" *Matthew I Banks, Emily R. Dappen, Elie Matar, Benjamin D. Hayum, Michael H. Sutherland, Bryan M. Krause, Hiroto Kawasaki, Robert D. Sanders, Kirill V. Nourski*

PRESENTATIONS

"State of Frontier AI Safety" *Wisconsin AI Safety Initiative Speaker Event*
"A Conversation on AI and Society" *Panel with Professor Annette Zimmermann (Philosophy), Jerry Zhu (CS), and David Shaffer (Education)*
"Navigating the Frontier of AI Development: Balancing Transformative Potential and Emerging Risks" *Introduction for Seb Krier from Google DeepMind*
"Governing Frontier AI: Policy Challenges and Strategies" *Introduction for Caleb Withers from Center for a New American Security*
"Detecting Danger in AI: What, Why, and How" *Introduction for Thomas Broadley from METR* **"Generative AI Panel"** *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*
"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 Poster Session*
"Sex Differences and Serotonergic Mechanisms in the Behavioural Effects of Psilocin" *Banks Lab Journal Club*

HONORS AND EVENTS

The Curve

- Invited on a shortlist of ~200 to attend The Curve, a mini conference with thought leaders in AI Policy from Anthropic, OpenAI, DeepMind, several government bodies, and several academic institutions meant to encourage healthy bipartisan debate on the field's biggest questions.

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

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
- Upcoming Coursework: Mathematical Methods in Machine Learning, Topics in Mathematical Data Science: Nonparametric Methods, Theory of Reinforcement Learning.
- 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

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).

- Led engagements with CS Department Chair and UW-Madison Senior Vice Provost to strengthen student involvement in ML Safety, organize interdisciplinary AI events, and contribute to university strategy, including securing role for WAISI in the hiring process of new professors.

Machine Alignment Theory Scholar **Remote**
Summer Intern *Summer 2023*

- Decomposed and mathematically formalized important AI Alignment concepts from first principles such as preference stability, optimization/goodharting, value formation, and decision boundaries.
- Mapped out potential outcomes of AI Alignment efforts across diverse futures, analyzing the interplay between cultural variability, AI explainability, global coordination, and intelligence capability trajectories.
- Developed a model loss basin volume estimator in gradient space using estimations of the 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.