Emily Bunnapradist

+1 424-284-6633 embunna@stanford.edu linkedin.com/in/ebunnapradist

EDUCATION

Stanford University **Palo Alto, CA**
Master's in Computer Science (Artificial Intelligence), GPA: 4.03 / 4.0 *Sept 2023 – Dec 2024*
Bachelor's in Mathematics & Symbolic Systems (Neuroscience) *Sept 2020 – June 2024*

TECHNICAL SKILLS

Languages: Python, C, C++, Javascript, MatLab; **ML Frameworks**: PyTorch, Tensorflow, Jax, HF, SLURM
ML Skills: Training, Fine-tuning, Model Editing, Agentic Frameworks, Preference Optimization, Evaluation
Research Areas: NLP, NLU, Reliability, Explainability & Interpretability, Applied Math, CompNeuro, Phil of Mind

SELECTED RESEARCH PROJECTS

Heterogeneity of Preference Datasets for AI Alignment **Sept 2024 – Present**
Affiliated with Stanford Trustworthy AI Research (STAIR), advised by Prof. Sanmi Koyejo *Paper in Progress*

- Quantitatively measuring the diversity of preferences expressed utilizing the ideal point model framework by to learn k prototypical "individuals", proposing a new metric beyond one-dimensional inter-annotator disagreement.

Investigating Memory Mechanisms of Toy Language Models **Mar 2024 – Present**
Affiliated with the Stanford NLP Group, advised by Profs. Chris Manning and Chris Potts *Paper in Progress, Poster*

- Conducting interpretability research on state-space models representations through systematic ablation studies.
- Developing comparative analysis framework for memory characteristics across RNNs, Transformers, and SSMs.

Scaled-Up Social Learning through Reinforcement Learning **June 2023 – Sept 2023**
Affiliated with the Stanford AI Lab (CoCoLab), advised by Prof. Noah Goodman *Poster*

- Implemented multi-armed bandit RL algorithms to model multi-generational knowledge transmission.
- Designed and ran human experiments using Dallinger and Javascript to validate computational findings.

Modeling of Cortical Network Dynamics for Neuroscience **May 2022 – August 2022**
Affiliated with the NYU Courant Institute of Applied Mathematics, advised by Dr. Lai-Sang Young *Presentation*

- Modeled cortical network dynamics with LSTMs to predict multiple firing events from high-dimensional neural data.
- Fully funded through NSF-RTG in Modeling & Simulation grant and presented research at the AM-SURE symposium.

SELECTED SOFTWARE PROJECTS

Visualizations and Analyses of Real-Time Neural Data for Vision **Sept 2022 – Sept 2023**
Affiliated with the Stanford School of Medicine (Chichilnisky Lab), advised by Prof. E.J. Chichilnisky *Github Repo*

- Contributed to the Stanford Artificial Retina Project, a research initiative which aims to develop an epiretinal implant that reproduces high-fidelity vision for blind people affected by incurable retinal disease with a novel electronic device.
- Developed a GUI to analyze and visualize compressed neural data recordings in real time, increasing speed by 40%.

RELEVANT COURSEWORK

Graduate-Level AI/ML: CS224N (NLP), CS224U (NLU), CS229 (Statistical ML), CS231N (Computer Vision), CS329H (ML from Human Preferences), CS339N (ML for Neural Data), CS362 (AI Alignment)
Cognitive Science: CS428A (Probabilistic Cognition), MATSCI 384 (NeuroTech), PSYC124 (Brain Plasticity), PSYCH45 (Learning & Memory), PSYCH209 (NN Models of Cognition), PSYCH240A (Curiosity in AI)
Mathematics: CS157 (Logic), CS205L (Math for ML), MATH51-53 (Lin. Alg/MVC/DiffEQs), MATH110-115 (Combinatorics/Number Theory/Real Analysis), PHIL150/151 (Metalogic), STATS116 (Probability Theory)

HONORS & AWARDS

Stanford Award of Excellence Recipient **Jun 2024**
Designed to recognize the top 10% of the class for their impact on the university. *More Info*

Google CS Research Mentorship Program (CSRMP) Award Recipient **Sept 2023**
Designed to support the pursuit of computing research for students from historically marginalized groups. *More Info*

SELECTED TECHNICAL PROJECTS

Interchange Interventions in Vision Models **Mar 2024 – Jun 2024**
CS 231N: Deep Learning for Computer Vision Final Project, with Profs. Eshan Adeli and Fei-Fei Li *Paper, Poster*

- Implemented interchange interventions, leveraging Distributed Alignment Search (DAS), for vision models.
- Fine-tuned ResNet architecture achieving 100% classification accuracy revealing key insights about feature importance.
- Created custom image datasets to evaluate model reliance on color, shape, and background features.

Investigating Internal Representations of Garden Path Sentences in LLMs **Mar 2023 – Jun 2023**
CS 224U: Natural Language Understanding Final Project, with Prof. Chris Potts *Paper, Github Repo*

- Compared how different language models process complex linguistic structures, specifically analyzing BERT and GPT models' handling of garden path sentences, using Manhattan distance, cosine similarity, and surprisal metrics.
- Demonstrated that GPT-2's processing aligns more closely with human comprehension patterns, while BERT's bidirectional architecture shows different processing characteristics.

Modeling Attachment Theory through Reinforcement Learning **Mar 2023 – Jun 2023**
PSYCH 240A: Curiosity in Artificial Intelligence Final Project, with Prof. Nick Haber *Paper, Github Repo*

- Developed a custom reinforcement learning environment using Python and Minigrid to model parent-child attachment behaviors, integrating developmental psychology principles into AI training.
- Collaborated in a 3-person team using Git for version control, managing codebase on GitHub, and iterating through multiple framework options (Multigrid, TF Agents) to optimize performance.

Are LLMs Smarter than a 5th Grader? Mathematical Cognition in LLMs **Jan 2023 – Mar 2023**
PSYCH 209: Neural Network Models of Cognition Final Project, with Prof. Jay Mcclelland *Paper*

- Designed and conducted empirical study evaluating GPT-3's mathematical reasoning capabilities through analysis of over 5,600 model responses to novel mathematical operations to assess mathematical comprehension vs. pattern matching.
- Quantified significant performance degradation (up to 68% accuracy drop) when models encounter higher magnitude numbers and complex explanatory requirements.

EI's in Disguise: Classification of Retinal Ganglion Cells **Jan 2023 – Mar 2023**
CS 339N: ML Methods for Neural Data Analysis Final Project, with Prof. Scott Linderman *Paper*

- Designed and implemented CNN architectures achieving 70.5% accuracy in classifying retinal ganglion cell types using only electrical imaging data, improving upon prior benchmarks of 63%.
- Built data preprocessing pipeline for normalizing spatiotemporal voltage patterns across multiple retinal recordings.

TEACHING & OUTREACH

Teaching Assistant for Stanford Computer Science Department **Mar 2022 – Present**
CS 224N: Natural Language Processing with Deep Learning (Winter 2025) *Course Website*
CS 224V: Conversational Virtual Assistants with Deep Learning (Fall 2024) *Course Website*
CS 181: Computers, Ethics, and Public Policy (Spring 2023) *Course Website*
CS 184: Bridging Technology and Policy Through Design (Spring 2022)

Co-Instructor for CS 25: Transformers United **Mar 2024 – Jun 2024**
Speaker series focused on research in Transformers, reached an audience of 400+ each week. *Course Website*