Ivory Yang

Hanover, NH · 347-251-0374 · ivory.yang.gr@dartmouth.edu
linkedin.com/in/ivoryayang/ · https://ivoryayang.github.io/

Education

Dartmouth College, Hanover, NH **June 2025**
Masters of Science, Computer Science **GPA 4.0/4.0**
Relevant Coursework: Machine Learning, Deep Learning, Artificial Intelligence
Honors/Awards: Guarini Merit Scholarship, Thomas D. Sayles Ethics Award

University of Michigan, Ann Arbor, MI **May 2020**
Bachelors of Business Administration, Bachelors of Science (Cognitive Science) **GPA 3.82/4.0**
Honors/Awards: UM Pan-Asia Scholar, James B. Angell Scholar, Global Experience Scholar, University Honors
Activities: Michigan Stocks and Bonds Organization, Equestrian Team, Alpha Omicron Pi Sorority, HEC Paris

Graduate Coursework, San Francisco, CA **Dec 2022**
Stanford University - Computer Organization & Systems (CS107) **GPA 4.0/4.0**
Harvard University - Introduction to CS (CS50), Data Structures & Algorithms (CS124)
UC San Diego - Discrete Math (CSE-41243), Linear Algebra (CSE-40023), Intermediate Programming with Objects (CSE-40477)

Research Interests

As a researcher, I am passionate about the application of NLP and ML to address socially impactful challenges. My current research areas of focus are:

- Generalization capabilities of LLMs; culturally aware adaptation for low-resource languages and tasks
- Resilience of large AI system safety features against persuasion-based activation steering
- Revitalization of endangered / underrepresented languages with AI, with an emphasis on indigenous / ancient languages

Publications

NüshuRescue: Revitalization of the Endangered Nüshu Language with AI

Ivory Yang, Weicheng Ma, Soroush Vosoughi
In *The 31st International Conference on Computational Linguistics (COLING 2025)* [Main]

MentalManip: A Dataset for Fine-grained Analysis of Mental Manipulation in Conversations

Yuxin Wang, **Ivory Yang**, Saeed Hassanpour, Soroush Vosoughi
In *The 62nd Annual Meeting of the Association for Computational Linguistics (ACL 2024)* [Main, Oral*]

Enhanced Detection of Conversational Mental Manipulation Through Advanced Prompting Techniques

Ivory Yang, Xiaobo Guo, Sean Xie, Soroush Vosoughi
In *The Eighth Widening NLP Workshop* at *The 2024 Conference on Empirical Methods in Natural Language Processing (WiNLP @ EMNLP 2024)*

Is it Navajo? Accurate Language Detection in Endangered Athabaskan Languages

Ivory Yang, Weicheng Ma, Chunhui Zhang, Soroush Vosoughi
Under Review at *The 2025 Annual Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics (NAACL 2025)*

Communication is All You Need: Persuasion Dataset Construction via Multi-LLM Communication

Weicheng Ma, Hefan Zhang, **Ivory Yang**, Shiyu Ji, Joice Chen, Farnoosh Hashemi, Shubham Mohole, Ethan Gearey, Michael Macy, Saeed Hassanpour, Soroush Vosoughi
Under Review at *The 2025 Annual Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics (NAACL 2025)*

Research & Work Experience

Minds, Machines and Society Group, Hanover, NH **Jan 2024-Present**
Machine Learning Research Assistant

- Supervised by Professor Soroush Vosoughi as part of the Minds, Machines and Society Group, conducting research in the field of machine learning so as to develop computational tools that offer new perspectives on social systems and issues
- Engaged in natural language processing (NLP) and machine learning research, specifically exploring large language models (LLMs) to detect manipulation tactics in speech, so as to harness findings to develop automatic systems to properly handle and mitigate verbal mental manipulation

- Currently exploring LLM generalization capabilities for revitalization of endangered and low-resource languages, with a focus on Nüshu and Native American languages under the Athabaskan family (Navajo, Apache etc.), to enhance linguistic data accessibility and preservation.

GraphMind Group, Hanover, NH **Sep 2024-Present**
Machine Learning Research Assistant
- Worked with Professor Yujun Yan on incorporating graph representations into LLM role-play debates to analyze the structure and patterns of persuasive interactions, so as to unveil novel insights into discourse structure and strategies for social AI applications
- Currently developing a heterogeneous graph explainer to enhance the transparency and interpretability of role-play debates, leveraging graph neural networks to model and explain persuasive dynamics in LLM interactions

Supervised Program for Alignment Research (SPAR), Berkeley, CA **Mar 2024-June 2024**
Machine Learning Research Intern
- Conducted alignment research with a focus on AI safety and mechanistic interpretability, contributing to the understanding of activation steering vectors and further development of LLM defense mechanisms
- Conducted technical experiments such as testing of refusal dataset with Contrastive Activation Addition (CAA) using LLaMa-2 models, so as to determine the optimal layer for inserting steering vectors to improve model defense performance

TEACHING EXPERIENCE

Machine Learning (COSC 274), Dartmouth College **Jan 2025**
Teaching Assistant (Incoming), Instructor: Dr. Soroush Vosoughi

Artificial Intelligence (COSC 276), Dartmouth College **Sep 2024-Dec 2024**
Teaching Assistant, Instructor: Dr. Soroush Vosoughi

Ross Integrative Semester (RIS), University of Michigan **Sep 2019-Dec 2019**
Teaching Assistant

HONORS & AWARDS

Guarini Merit Scholarship ($33,000/yr), Dartmouth College

Thomas D. Sayles Ethics Research Grant ($3500), Dartmouth College Ethics Institute

Travel Grant ($2000), Widening Natural Language Processing (WiNLP) at EMNLP 2024

Travel Grant ($1000, $1300), Dartmouth Women in CS (WiCS)

Guarini Travel Grant ($1000), Dartmouth College

Alumni Research Award ($1000) - *Pending*, Dartmouth College

Pan-Asia Scholar ($4000), University of Michigan

Global Experience Scholar ($1500), University of Michigan

James B. Angell Scholar, University of Michigan

University Honors, University of Michigan

SKILLS & INTERESTS

Programming Languages/Tech: C, C++, Python
Languages: Mandarin (Fluent), French (Conversational), Korean (Conversational)
Lived in six countries, took a gap year before college to backpack across Asia