Ivory Yang

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

Education

Dartmouth College, Hanover, NH — **June 2025**
Masters of Science, Computer Science — **GPA 4.0/4.0**
Relevant Coursework: Machine Learning, Deep Learning, Data Mining
Honors/Awards: Guarini Merit Scholarship

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)

Experience

Supervised Program for Alignment Research (SPAR), Berkeley, CA — **Mar 2024-June 2024**
Machine Learning Researcher

- 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
- Co-authored a Mechanistic Interpretability paper titled "Scaling Laws for Contrastive Activation Addition with Refusal Mechanisms and Llama 2 Models", currently under review

Minds, Machines and Society Group, Hanover, NH — **Dec 2023-Present**
Machine Learning Research Assistant

- Collaborated with 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 working on corpus development for nüshu, a low-resource language, for use in task training such as sentence completion and word-sense disambiguation, so as to define underlying semantic and linguistic patterns of the language

Published papers

MentalManip: A Dataset for Fine-grained Analysis of Mental Manipulation in Conversations
ACL 2024 (Main, oral)

- Main conference paper accepted at the 2024 Association for Computational Linguistics (ACL) conference, additionally selected for oral presentation
- Creation of a dialogical corpus focused on the detection of mental manipulation in speech patterns, contributing to the development of novel methodologies to detect nuanced linguistic cues indicative of psychological manipulation within the realm of interpersonal communication
- Conducted experiments using multi-class classification tasks and LLMs (LLaMa-2-7B and 275 RoBERTa-large) to report prediction accuracy for each technique and vulnerability task

Enhanced Detection of Conversational Mental Manipulation Through Advanced Prompting Techniques
WiNLP @ EMNLP 2024

- Conducted extensive research on mental manipulation detection in dialogues using advanced prompting techniques such as Zero-Shot, Few-Shot and Chain-of-Thought (CoT) prompting, to show that CoT achieves highest accuracy by leveraging structured reasoning, but that Zero-Shot CoT results in a higher degree of false positives as model complexity increases
- Achieved significant improvements in detection accuracy, demonstrating the efficacy of CoT prompting combined with example-based learning, and highlighting the challenges of detecting subtle manipulative language with the evolution of model complexity from GPT-3.5 to GPT-4o

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