Jason Vega

javega3@illinois.edu · +1 (925) 481-4210 · jason-vega.github.io

Education

University of Illinois Urbana-Champaign Urbana, IL
Ph.D. Computer Science September 2022 - May 2028 (expected)
University of California, San Diego La Jolla, CA
B.S. Computer Science *GPA: 3.916 (magna cum laude)* September 2018 - June 2022

Research Experience

Safety of Large Language Models (LLMs) Urbana, IL
Graduate Researcher August 2023 - Present

- **Advisor:** UIUC Prof. Gagandeep Singh. **Topic:** Investigating vulnerabilities of safety-trained open-source autoregressive LLMs such as Llama 2, e.g. efficiently circumventing safety-training. **Paper:** Vega, J., Chaudhary, I., Xu, C., & Singh, G. (2023). Bypassing the Safety Training of Open-Source LLMs with Priming Attacks. arXiv preprint arXiv:2312.12321.

Common Corruption Robustness Urbana, IL
Graduate Researcher August 2022 - Present

- **Advisor:** UIUC Prof. Gagandeep Singh. **Topic:** Investigating training-time methods to empirically improve the robustness of image classifiers against common corruptions.

Interpretability Robustness (Honors Thesis / McNair Project) La Jolla, CA
Undergraduate Researcher January 2021 - June 2022 (Remote)

- **Advisor:** UCSD Prof. Tsui-Wei (Lily) Weng. **Topic:** Formulating defenses for training robust image classification neural networks against adversarial attacks on various interpretation methods.
- **Contributions:** Implemented defense, verification and attack frameworks, and ran experiments to obtain preliminary robustness results of a 465x improvement compared to standard training.
- **Recognition:** Selected to give a plenary talk to represent the field of Engineering at UCSD's 34th annual Undergraduate Research Conference. (Recording available on conference website.)

Neural Representation Learning for Scribal Hands of Linear B La Jolla, CA
Undergraduate Researcher October 2020 - June 2021 (Remote)

- **Advisor:** UCSD Prof. Taylor Berg-Kirkpatrick. **Topic:** Applying neural networks to learn features (glyph shape and writing style) of the ancient Greek script Linear B.
- **Contributions:** Cropped 2,171 symbols from book scans to help create a dataset of Linear B symbols. Investigated using a neural object detection model to automate the cropping process.
- **Paper:** Srivatsan, N., Vega, J., Skelton, C., & Berg-Kirkpatrick, T. (2021, September). Neural Representation Learning for Scribal Hands of Linear B. In International Conference on Document Analysis and Recognition (pp. 325-338). Springer, Cham.

Text Line Extraction for Printed Historical Documents La Jolla, CA
Undergraduate Researcher October 2019 - June 2020, October 2020 - June 2021 (Remote)

- **Advisor:** UCSD Prof. Taylor Berg-Kirkpatrick. **Topic:** investigating statistical and neural methods to improve text line extraction for degraded printed historical documents.
- **Contributions (first year):** proposal writing, poster presenting, created tools for performance evaluation and ground truth generation, and training+qualitatively evaluating a neural network.
- **Leadership:** Served an additional project management role in the first year's team of four undergrads. Contributed only as a mentor during second year to a new team of four undergrads.

Work Experience

UCSD Computer Science & Engineering Department La Jolla, CA
Course Tutor January 2022 - March 2022

- Tutored in an introductory data structures and object-oriented design course of ∼ 600 students.
- Provided student support through lab interactions and an online classroom forum.

- Managed a pod of 18 students, regularly checking their progress in the course, grading their assignments and intervening when noticing signs of struggle.

Microsoft Bellevue, WA
Software Engineering and Program Management Intern June 2020 - September 2020 (Remote)

- Worked on the new Digital Marketing Center online platform from Microsoft Ads in both program management (weeks 1-6) and software engineering (weeks 7-12) roles.
- Produced a 22 page specification document proposing a new feature, supported by observations from real customer data and with a competitive analysis of four major competitors.
- Implemented a new home page component, including CSS, display logic, responsive layout integration, E2E testing and refractoring of existing code to improve responsive behavior.

Hackingtons Code School Pleasant Hill, CA
Assistant Instructor August 2019 - September 2019

- Provided guidance to students ages 8-15 learning web development (HTML, CSS, JavaScript) and game development (C#/Unity, Scratch) in a flipped classroom learning environment.
- Answered students' technical questions and individually checked in with students to evaluate their progress on coding projects.

Diablo Valley College Pleasant Hill, CA
College for Kids Instructional Assistant June 2019 - July 2019

- Engaged with approximately 100 students entering the 4th-9th grade over four "Coding & Robotics" course sections utilizing the BBC Micro Bit, Microsoft MakeCode and MicroPython.
- Assisted students with debugging and circuit setup, resolved student disputes and helped to develop and teach two major robotics projects utilizing accelerometer and ultrasonic sensors.

Awards

Sloan Scholar University of Illinois Urbana-Champaign
Alfred P. Sloan Foundation's Minority Ph.D. (MPHD) Program (institutional match). Sept. 2022

Wing Kai Cheng Fellowship University of Illinois Urbana-Champaign
A one-year department fellowship graciously sponsored by the Wing Kai Cheng estate. Sept. 2022

Alumni Leadership Scholarship University of California, San Diego
A two-year scholarship awarded for college-level academic and campus leadership. Aug. 2020

Violet and Matthew Lehrer Scholarship University of California, San Diego
A two-year scholarship awarded for college-level academic and campus leadership. Aug. 2020

Academic Services

MLSys 2023 .... Emergency Reviewer

Extracurricular Activites

UCSD ACM AI - Event & Social Lead October 2020 - June 2022
Designed and led educational workshops, organized and gave research talks, led research paper reading group sessions and organized social activities for UCSD undergrads interested in artificial intelligence.

Skills

Python, PyTorch, NumPy, Unix, Conda, Docker