Jason Vega

javega3@illinois.edu · linkedin.com/in/jason-vega/ · 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

Adaptive Prefilling Attacks and Defenses on Large Language Models Urbana, IL
Graduate Researcher May 2025 - Present

- **Topic:** Investigating safety vulnerabilities of deep safety alignment for LLMs, and proposing more robust fine-tuning techniques for deep safety alignment. **Advisor:** UIUC Prof. Gagandeep Singh.
- **Paper: Vega, J.** and Singh, G. (2024). Matching Ranks Over Probability Yields Truly Deep Safety Alignment. arXiv preprint arXiv:2512.05518.

Random Augmentation Attacks on Large Language Models Urbana, IL
Graduate Researcher March 2024 - January 2025

- **Topic:** Investigated the intersection of safety under random augmentations with multiple dimensions: augmentation type, model size, quantization, fine-tuning-based defenses, and decoding strategies (e.g., sampling temperature). **Advisor:** UIUC Prof. Gagandeep Singh.
- Showed that SoTA aligned LLMs, such as Llama 3.1 Instruct, Phi 3 and Qwen 2, can have their safety alignment bypassed in as few as 25 simple random augmentations per harmful prompt.
- **Paper: Vega, J.**, et al. (2024). Stochastic Monkeys at Play: Random Augmentations Cheaply Break LLM Safety Alignment. arXiv preprint arXiv:2411.02785.

Prefilling Attacks on Large Language Models Urbana, IL
Graduate Researcher August 2023 - December 2023

- **Topic:** Co-authored the first publication on a simple jailbreaking technique, now known as the *prefilling attack*, to easily bypass LLM alignment. **Advisor:** UIUC Prof. Gagandeep Singh.
- Prefilling is increasingly being used as a baseline attack for evaluating LLM safety; e.g., our paper has been cited in work from Meta (Zhang et al., 2024) and Anthropic (Marks et al., 2025).
- **Paper: Vega, J.***, Chaudhary, I.*, Xu, C.* and Singh, G., Bypassing the Safety Training of Open-Source LLMs with Priming Attacks. In *The Second Tiny Papers Track at ICLR 2024*.

Common Corruptions Robustness of Image Classifiers Urbana, IL
Graduate Researcher November 2022 - May 2024

- **Topic:** Investigated training-time methods to improve the robustness of ResNets and Vision Transformers against image corruptions. **Advisor:** UIUC Prof. Gagandeep Singh.
- **Investigated techniques include:** Self-supervised learning (SimCLR) robustness evaluation, regularization of Class Activation Mapping explanations, training a CycleGAN for data augmentation, diversity regularization of ensembles, distillation for robust smaller models

Interpretability Robustness of Image Classifiers 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. Talk available on conference website.

Glyph Extraction for Ancient Greek Script La Jolla, CA
Undergraduate Researcher October 2020 - June 2021 (Remote)

- **Topic:** Applying neural networks to learn features (glyph shape and writing style) of the ancient Greek script Linear B. **Advisor:** UCSD Prof. Taylor Berg-Kirkpatrick.

- **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)

- **Topic:** investigating statistical and neural methods to improve text line extraction for degraded printed historical documents. **Advisor:** UCSD Prof. Taylor Berg-Kirkpatrick.
- **Contributions:** 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.

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

ICLR 2025 .... Reviewer
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, Matplotlib, Unix, Conda, LaTeX