Alexander Wan

alexwan@berkeley.edu

Education

University of California, Berkeley *September, 2021 – May, 2025*

B.A. in Computer Science, Statistics, and Mathematics (GPA: 3.792)

- Multivariable Calculus, Upper Division Linear Algebra, Data Structures, Intro to AI, Discrete Math and Probability, Probability Theory, Game Theory, Machine Structures, Intro to Machine Learning, Intro to Deep Learning*, Computational Models of Cognition*, Optimization Models in Engineering*, Intro to Analysis* (* = *in progress*)

Self Studied: Conv Nets for Visual Recognition (*Stanford CS231n*), NLP with Deep Learning (*Stanford CS224n*)

Experience

Berkeley NLP Group (Berkeley AI Research Lab) *April 2022 – Present*

Machine Learning Research Intern - Advised by Prof. Dan Klein

- Created concealed data poisoning attacks against large language models like BERT, improving efficacy from ~50% to 96% and created a software pipeline to allow for fast experimentation.
- Performed analyses of multi-task learning to develop better data poisoning techniques, reaching nearly perfect rates of misclassification in instruction-tuned language models [1].
- Investigated adversarial robustness in instruction-tuned LLMs by training massive 11 billion parameter models on hundreds of tasks utilizing both Google Cloud TPU and multi-GPU acceleration. (PyTorch/HF Transformers/Jax)

Michigan State University Heterogeneous Learning and Reasoning Group *June 2020 – Present*

Machine Learning Research Intern - Advised by Assist. Prof. Parisa Kordjamshidi

- Tested and designed deep learning constraint integration methods for use in a standard benchmark, taking into account metrics like constraint satisfaction rate and time complexity. Achieved 94% accuracy with only 5% of the full dataset on a weakly supervised task [2].
- Developed models that use the TypeNet ontology to perform fine-grained entity typing on nearly 2000 labels.
- Created sequence to sequence RNNs to study language acquisition in deep NLP. Designed algorithms to evaluate and improve the diversity of generations. (Python/PyTorch)

EleutherAI *February 2023 – May 2023*

Machine Learning Research Intern - Advised by Nora Belrose

- Investigated an unsupervised method of probing large language models using a consistency objective.
- Used prefix-tuning and projected gradient descent to investigate its robustness to adversarial perturbations.
- Optimized training and inference for use in a multi-GPU environment. (Python/PyTorch/HuggingFace Transformers)

University of Michigan Transportation Research Institute *Summer 2019 & 2020*

Machine Learning Research Intern - Advised by Dr. Daniel Park

- Built software package to automate labeling of body/face keypoints & alignment of 3D head shape models for use in passenger safety research, replacing previous manual process. (C#/Python)

Personal

- **Skills:** Java, C#, Web Dev, C++, Python (NumPy, OpenCV, Keras, PyTorch, Jax w/ GPUs and TPUs, HF Transformers)
- **Awards**: Michigan Math Prize Competition top 200/6000, 2018 & 2020 ISEF Finalist
- **Teaching:** InspiritAI Instructor, CS 198-126 Deep Learning for Computer Vision

Publications

[1] ***Poisoning Instruction-Tuned Language Models***

Alexander Wan*, Eric Wallace*, Sheng Shen, Dan Klein

International Conference on Machine Learning (ICML), 2023

[2] ***GLUECons: A Generic Benchmark for Learning Under Constraints***

Hossein Rajaby Faghihi, Aliakbar Nafar, Chen Zheng, Roshanak Mirzaee, Yue Zhang, Andrzej Uszok, Alexander Wan, Tanawan Premsri, Dan Roth, Parisa Kordjamshidi

AAAI Conference on Artificial Intelligence (AAAI), 2023