Alexander Wan

alexwan@berkeley.edu

Education

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

B.A. in Computer Science, Statistics, and Mathematics (GPA: 3.840)

- 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

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

- Demonstrated the vulnerability of instruction-tuned models to data poisoning, showing how models can be manipulated to consistently misclassify samples or produce degenerate outputs across hundreds of tasks [1].
- Investigated the 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)
- Designed a benchmark to study how retrieval-augmented models like Bing Chat judge the credibility of websites. Created a dataset consisting of 82k webpages across 5k search queries.
- Conducted sensitivity analysis to determine how in-the-wild differences in websites can bias RAG models. Used these insights to demonstrate that RAG models are vulnerable to adversarial manipulation through retrieved text.

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

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

- Used constraint-integration methods to train models on a weakly-supervised classification task. Achieved 94% accuracy with only 5% of the full dataset. Contributed findings as a part of a benchmark for integrating domain knowledge into deep learning models through constraints [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.
- Implemented a constraint-satisfaction algorithm based on inference-time gradient steps into DomiKnowS, a library enabling knowledge integration in deep learning models through logical constraints. (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)

Personal

- **Skills:** Java, C#, Web Dev, C++, Python (NumPy, OpenCV, Keras, PyTorch, Jax w/ GPUs and TPUs, HF Transformers)
- **Teaching:** InspiritAI Instructor, CS 198-126 Deep Learning for Computer Vision
- **Talks:** "Manipulating Large Language Model Predictions Through Data" at USC NLG, 2023

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