Alexander Wan

Curriculum Vitae

Berkeley, CA — alexwan@berkeley.edu — wanalex.com

EDUCATION

B.A. in Computer Science *September 2021 - May 2024*
University of California, Berkeley (GPA: 3.857)

Coursework: Multivariable Calculus, Abstract Linear Algebra, Data Structures, Algorithms, Discrete Math and Probability, Probability Theory, Statistics, Game Theory, Machine Structures, Intro to AI, 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 - February 2024*
Undergraduate Research *with Prof. Dan Klein*

- Demonstrated the vulnerability of instruction-tuned models to data poisoning, showing that models can be manipulated to consistently misclassify samples or produce degenerate outputs across hundreds of tasks [1].
- Investigated the adversarial robustness of instruction-tuned LLMs by training 11-billion parameter models on hundreds of tasks utilizing both Google Cloud TPU and multi-GPU acceleration.
- Designed a benchmark to study how retrieval-augmented models judge the credibility of websites. Created a dataset consisting of 4000 pages over more than 400 search queries.
- Conducted sensitivity analyses to determine how in-the-wild differences in websites can bias RAG models. Used these insights to demonstrate that RAG models significantly differ from humans when determining the credibility of text, making them vulnerable to misinformation [2].
- Studied the robustness of AI text detection software, including designing adversarial attacks using gradient-based optimization to circumvent likelihood-based metrics.

Michigan State University Heterogeneous Learning & Reasoning *June 2020 - December 2023*
Research Intern *with 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 [3].
- Created sequence-to-sequence RNNs to study language acquisition through the lens of deep learning. Designed algorithms to evaluate and improve the diversity of generations.
- Implemented a constraint-satisfaction algorithm based on inference-time gradient steps into DomiKnowS, a deep learning library enabling the use of domain-knowledge during training and inference.

EleutherAI *February 2023 – May 2023*
Research Intern with *Nora Belrose*

- Investigated the robustness of probing methods to adversarial attacks. Optimized training and inference for a multi-GPU environment.
- Received a $2000 scholarship from the Center for AI Safety to conduct this research.

PUBLICATIONS

[1] **Alexander Wan***, Eric Wallace*, Sheng Shen, and Dan Klein. Poisoning language models during instruction tuning. In *International Conference on Machine Learning (ICML)*, 2023.

[2] **Alexander Wan**, Eric Wallace, and Dan Klein. What evidence do language models find convincing? In *Association for Computational Linguistics (ACL)*, 2024.

[3] Hossein Rajaby Faghihi, Aliakbar Nafar, Chen Zheng, Roshanak Mirzaee, Yue Zhang, Andrzej Uszok, **Alexander Wan**, Tanawan Premsri, Dan Roth, and Parisa Kordjamshidi. GLUECons: A generic benchmark for learning under constraints. In *Association for the Advancement of Artificial Intelligence (AAAI)*, 2023.

INVITED TALKS

USC ISI Natural Language Seminar *November 2023*
Manipulating Large Language Model Predictions Through Data (Link)
Hosted by Justin Cho and Prof. Jon May

REVIEWING

NeurIPS Workshop on Instruction Tuning and Instruction Following *October 2023*

MISCELLANEOUS

Machine Learning at Berkeley *October 2021 - Present*
Member, Former Education & Research Officer
Developed course content and gave lectures for CS 198-126 (Modern Computer Vision and Deep Learning) and an internal machine learning course for new members.

InspiritAI *Feb 2023 - May 2023*
Instructor
Introduced AI concepts like natural language processing and computer vision, along with Scratch programming, to 5th/6th graders.