RYAN WANG

Phone: (+1) 213-776-8218 ◇ Email: ryanywan@usc.edu
Homepage: ryanyxw.github.io
Google Scholar ◇ Github ◇ LinkedIn

EDUCATION

University of Southern California (USC) *Aug 2021 - May 2025 (expected)*
B.S. in Computer Science
GPA: 3.98/4.0
Related courses: Introduction to Machine Learning, Advanced Topics in NLP, History of Language and Computing, Mathematics of Machine Learning, Probability Theory, Mathematical Statistics

RESEARCH EXPERIENCE

Teaching Models to Understand (but not Generate) High-risk Data Mar 2024 - Present
Supervisors: Prof. Robin Jia, Prof. Swabha Swayamdipta USC

- High-risk data like toxic text are filtered out of pre-training corpora to prevent the model from outputting toxic generations. However, models need exposure to toxic texts in order to robustly identify and mitigate them. This research explores new training paradigms that allow models to pre-train on high-risk data without learning to generate them.
- Under submission to COLM 2025

Hubble: A Model Suite for Memorization (Ongoing) Oct 2024 - Present
Supervisors: Prof. Robin Jia USC

- Tasked with training from scratch a suite of Olmo-like models of up to 7B parameters on natural and synthetic data of up to 500B tokens using a NVIDIA DGX cluster
- Synthetic data is designed to study memorization risks from pre-training data, including copyright and privacy risks

Developing Data Watermarks for Pretraining [1] Apr 2023 - Feb 2024
Supervisor: Prof. Robin Jia USC

- Developed data watermarks for membership inference on pretraining data with statistical guarantees
- Trained models up to 1.4B parameters on up to 12B tokens using the GPT-NeoX pretraining library
- Investigated relationships between LLM memorization and duplication/length of sequences in a model's pre-training corpus
- Performed a post-hoc analysis of data watermarks on BLOOM 176B and found that robust detection can be made with data watermarks that occur at least 90 times throughout the entire pretraining corpus.

Mesh-based Visual Localization and Pose Tracking Aug 2022 - Apr 2023
Group: Network Systems Lab USC

- Ran experiments that applied Superglue for mesh-based visual localization.
- Investigated factors of 3D meshes that impacts pose tracking

Learning Genetic Regulatory Grammar May 2022 - Aug 2022
Group: Center for Synthetic & Systems Biology USC Viterbi Tsinghua Summer Research

- Learned to use the Tomtom tool under the MEME suite along with a Shannon Entropy filter to identify genetic subsequences with high information content
- Applied U-nets onto Probability Weight Matrices of genetic sequences to predict genetic motifs, achieving a test accuracy of 76%.

Predicting COVID-19 Severity using Genomic Patterns [2] [3] June 2020 - May 2021
Supervisor: Prof. Manolis Kellis

- Developed a haplotype-block based algorithm to identify genetic hotspots that led to an increased severity of COVID-19
- Trained random forest and various neural network architectures to predict patient susceptibility to COVID-19 using identified genetic hotspots

PUBLICATIONS

[1] J. Wei*, **R. Wang***, and R. Jia, "Proving membership in LLM pretraining data via data watermarks," in *Findings of the Association for Computational Linguistics ACL 2024*, L.-W. Ku, A. Martins, and V. Srikumar, Eds., Bangkok, Thailand and virtual meeting: Association for Computational Linguistics, Aug. 2024, pp. 13 306–13 320. DOI: 10.18653/v1/2024.findings-acl.788. [Online]. Available: https://aclanthology.org/2024.findings-acl.788.

[2] **R. Wang**, T. Qinsong Guo, L. Guanhua Li, and J. Yutian Jiao, "Using gwas snps to determine association between covid-19 and comorbid diseases," in *2020 IEEE 14th International Conference on Big Data Science and Engineering (BigDataSE)*, 2020, pp. 36–40. DOI: 10.1109/BigDataSE50710.2020.00013.

[3] **R. Wang**, T. Q. Guo, L. G. Li, J. Y. Jiao, and L. Y. Wang, "Predictions of covid-19 infection severity based on co-associations between the snps of co-morbid diseases and covid-19 through machine learning of genetic data," in *2020 IEEE 8th International Conference on Computer Science and Network Technology (ICCSNT)*, 2020, pp. 92–96. DOI: 10.1109/ICCSNT50940.2020.9304990.

* Equal Contribution.
[2] and [3] are part of a high school summer research program

OTHER EXPERIENCE

White & Case LLP Law Internship July 2023 - Aug 2023
Supervisor: Alex Zhang

- Summarized arbitration cases and contributed to high-tech legal case preparations
- Gave talks on technical aspects of language models and engaged in discussions on AI policy and applications of data watermarks

AI for Urban Planning Sep 2021 - Apr 2022
USC Center for AI in Society

- Built tools using natural language processing techniques to allow residents of resource-limited neighborhoods to actively participate in urban planning

VMware Internship June 2020 - Aug 2020
Supervisor: Layne Peng

- 6-week internship at VMWare R&D
- Setup AWS servers to use the KubeFate module to create a FATE Cluster that models a Vertical Federated Machine Learning process

TEACHING EXPERIENCE

Teaching Assistant (CSCI 467) *Fall 23', Spring 24', Spring 25'*
Instructor: Robin Jia

- Teaching assistant for CSCI 467 Introduction to Machine Learning
- Hosted recitation sessions, office hours, mentored course projects, prepared Midterm / Final exams

ACHIEVEMENTS

NSF Graduate Research Fellowship	*2025*
CRA Outstanding Undergraduate Researcher Award Honorable Mention	*2025*
USC Provost Research Fellowship	*2024 Spring, Summer, Fall*
USC CURVE Undergraduate Research Fellowship	*2022, 2023*
Viterbi Dean's List	*2022, 2023, 2024*

PROFESSIONAL SERVICE

EMNLP Emergency Reviewer for 2 Papers	*June 2024*

SKILLS

Programming Languages	Python, C/C++, Java
Machine Learning Tools	Huggingface, Pytorch, Slurm, Pandas, Numpy, Seaborn, Sklearn