RYAN WANG

Phone: (+1) 213-776-8218 ◇ Email: ryanywan@usc.edu
Homepage: ryanyxw.github.com
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

Pretraining with Undesired Data (Ongoing) Mar 2024 - Present
Supervisors: Prof. Robin Jia, Prof. Swabha Swayamdipta, Ph.D. Matt Finlayson USC

- Investigating training losses and architectural changes where models are pre-trained on undesired data (like toxic text) but are not going to generate it
- Continually pretrained the Olmo 1B model / pretrained from scratch variants of the Olmo model with architectural modifications

Developing Data Watermarks for Pretraining [1] Apr 2023 - Feb 2024
Supervisor: Prof. Robin Jia, Ph.D. Johnny Wei 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 various properties of random perturbations (like length or diversity)
- 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 Tsinghua

- 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 Matrixes 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: 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. T.-Z. Wei*, **Ryan Wang***, and R. Jia, *Proving membership in llm pretraining data via data watermarks*, 2024. arXiv: 2402.10892 [cs.CR]. [Online]. Available: https://arxiv.org/abs/2402.10892.

[2] **Wang, Ryan**, 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] **Wang, Ryan**, 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.

ACHIEVEMENTS

USC Provost Research Fellowship	*2024*
USC CURVE Undergraduate Research Fellowship	*2023*
USC CURVE Undergraduate Research Fellowship	*2022*
Viterbi Dean's List	*2022, 2023, 2024*

TEACHING EXPERIENCE

CSCI 467 Introduction to Machine Learning (Instructor: Robin Jia)	*Fall 23', Spring 24'*

PROFESSIONAL SERVICE

EMNLP Emergency Reviewer	*June 2024*

SKILLS

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