Zachary Coalson

Corvallis, OR | coalsonz@oregonstate.edu | zachcoalson.com

Summary

I am a first-year PhD student at Oregon State University working on trustworthy and socially responsible AI under the supervision of Prof. Sanghyun Hong.

Education

Oregon State University, Corvallis, OR Sept 2020 – June 2025

Honors B.S. in Computer Science, Minor in Mathematics (*Summa Cum Laude*)

Honors Thesis: On the Robustness of Neural Architecture Search to Data Poisoning Attacks

Academic advisor: Prof. Sanghyun Hong

Honors and Awards

NSF GRFP, National Science Foundation	2025
GEM Fellowship, National GEM Consortium	2025
ARCS Foundation Oregon Scholar Award, Oregon State University	2025
Dean's List, Oregon State University	2020 – 2025
Finley Academic Excellence Scholarship, Oregon State University	2020

Publications

Conference Publications

- **[ICCV '25]** Dongwoo Kang, Akhil Perincherry, **Zachary Coalson**, Aiden Gabriel, Stefan Lee, and Sanghyun Hong, "Harnessing Input-adaptive Inference for Efficient VLN".
  [acceptance rate: 24.0%]
- **[NeurIPS '23] Zachary Coalson**, Gabriel Ritter, Rakesh Bobba, Sanghyun Hong, "BERT Lost Patience Won't Be Robust to Adversarial Slowdown", https://openreview.net/forum?id=TcG8jhOPdv.
  [acceptance rate: 26.1%]

Preprints

- **[arXiv '25] Zachary Coalson**, Juhan Bae, Nicholas Carlini, Sanghyun Hong, "IF-Guide: Influence Function-Guided Detoxification of LLMs", https://arxiv.org/abs/2506.01790.
- **[arXiv '24] Zachary Coalson**, Jeonghyun Woo, Shiyang Chen, Yu Sun, Lishan Yang, Prashant Nair, Bo Fang, Sanghyun Hong, "PrisonBreak: Jailbreaking Large Language Models with Fewer Than Twenty-Five Targeted Bit-flips", https://arxiv.org/abs/2412.07192.
- **[arXiv '24] Zachary Coalson**, Huazheng Wang, Qingyun Wu, Sanghyun Hong, "Hard Work Does Not Always Pay Off: Poisoning Attacks on Neural Architecture Search", https://arxiv.org/abs/2405.06073.

Research Experience

Influence Functions to Reduce Large Language Model Toxicity Dec 2024 – May 2025

- Created a method that uses influence functions to attribute and suppress toxicity-promoting training data.
- Evaluated the method on four open-source large language models across three datasets.
- Achieved a 5–10× reduction in LLM toxicity in both pre-training and fine-tuning settings.

Bit-Flip Attacks to Jailbreak Large Language Models April 2024 – Nov 2024

- Created a comprehensive bit-flip attack pipeline.
- Evaluated the pipeline on eight open-source large language chat models across two harmful tasks.
- Demonstrated state-of-the-art attack success while flipping minimal bits.

Data Poisoning on Neural Architecture Search Dec 2023 – May 2024

- Developed a gradient-based clean-label poisoning attack to audit the robustness of NAS algorithms.
- Evaluated the attack on two representative NAS algorithms and one computer vision dataset.

- Discovered that such algorithms are surprisingly robust to practical poisoning attacks.

Slowdown Attacks on Input-Adaptive NLP Models Aug 2022 – Dec 2023

- Designed an objective function and two slowdown attacks based on the state-of-the-art text attacks.
- Performed an evaluation of the attacks on three input-adaptive NLP models across seven datasets.
- Demonstrated 100% attack success and proposed potential countermeasures such as input sanitization.

Professional Academic Activities

Conference Presentations

- NeurIPS '23 Poster: *BERT Lost Patience Won't Be Robust to Adversarial Slowdown* Dec 2023