Xiaosen Zheng

Computer Science · Ph.D. Candidate

(+65) 8393 3781 | xszheng.2020@phdcs.smu.edu.sg | xszheng2020 | xszheng2020

Summary

I am a Ph.D. candidate in the School of Computing and Information System at Singapore Management University. My recent interest focuses on **Data-Centric AI** and **AI Safety**.

Education

Singapore Management University — *Singapore*

Ph.D. Candidate in Computer Science — *Sept. 2020 - ongoing*

Supervisor: Prof. Jing Jiang

Central South University — *Changsha, China*

B.Eng. in Software Engineering — *Sept. 2015 - June 2019*

Working Experience

Research Associate — *Singapore*

Sea AI Lab, supervised by Dr. Tianyu Pang — *Jan. 2025 - ongoing*

- **Trustworthy AI**

Research Intern — *Singapore*

Sea AI Lab, supervised by Dr. Tianyu Pang — *May. 2023 - Oct. 2024*

- **Training Data Attribution on Diffusion Models/Large Language Models**
- **Jailbreaking (Multimodal) Large Language Models**

Research Engineer — *Singapore*

Singapore Management University, supervised by Prof. Jing Jiang — *Sept. 2019 - Aug. 2020*

- **Post-hoc Interpretability for NLP**

Publications

(* indicates equal contribution)

[ACL 2022] An Empirical Study of Memorization in NLP
***Xiaosen Zheng**, Jing Jiang*

[ICLR 2024] Intriguing Properties of Data Attribution on Diffusion Models
***Xiaosen Zheng**, Tianyu Pang, Chao Du, Jing Jiang, Min Lin*

[ICML 2024] Agent Smith: A Single Image Can Jailbreak One Million Multimodal LLM Agents Exponentially Fast
Xiangming Gu, **Xiaosen Zheng***, Tianyu Pang*, Chao Du, Qian Liu, Ye Wang, Jing Jiang, Min Lin*

[NeurIPS 2024] Improved Few-Shot Jailbreaking Can Circumvent Aligned Language Models and Their Defenses
***Xiaosen Zheng**, Tianyu Pang, Chao Du, Qian Liu, Jing Jiang, Min Lin*

[ICLR 2025] RegMix: Data Mixture as Regression for Language Model Pre-training (Spotlight)
Qian Liu **Xiaosen Zheng***, Niklas Muennighoff, Guangtao Zeng, Longxu Dou, Tianyu Pang, Jing Jiang, Min Lin*

[ICLR 2025] Cheating Automatic LLM Benchmarks: Null Models Achieve High Win Rates (Oral)
***Xiaosen Zheng**, Tianyu Pang*, Chao Du, Qian Liu, Jing Jiang, Min Lin*

Skills

Machine Learning: LLMs, Multimodal LLMs, Diffusion Models, Data Attribution, etc.

Libraries & Tools: PyTorch, Transformers, Diffusers, Peft, Captum, Scikit-Learn, NLTK, etc.

Honors & Awards

2024 **Awarded**, Presidential Doctoral Fellowship (Singapore Management University) — *Singapore*
2022 **Awarded**, Presidential Doctoral Fellowship (Singapore Management University) — *Singapore*
2019 **Top-1%**, Elo Merchant Category Recommendation Competition (Kaggle) — *Online*