Tianrong Zhang

zhangtianrong@psu.edu   tianrong-zhang-753939150   ZhangTianrong@GitHub
https://zhangtianrong.github.io/profile/#en

Employment History

2021 – 2021 **Software Engineering Intern,** Tencent

Education

2022 – · · · · **Ph.D. Informatics,** Pennsylvania State University
Advisor: Dr. Amulya Yadav & Dr. Jinghui Chen & Dr. Prasenjit Mitra
Field of interest: computational linguistics, model trustworthiness & robustness

2020 – 2022 **M.S. Computer Science,** University of Michigan
GPA: 4.0

2018 – 2020 **B.S. Computer Science,** University of Michigan
GPA: 3.875
Capstone: OoO computation unit design w/ branch predicting features.

2016 – 2020 **B.S. Electric & Computer Engineering,** Shanghai Jiao Tong University
GPA: 3.25

Research Publications

1. **T. Zhang**, B. Cao, Y. Cao, L. Lin, P. Mitra, and J. Chen, "WordGame: Efficient & effective LLM jailbreak via simultaneous obfuscation in query and response," in *Findings of the Association for Computational Linguistics: NAACL 2025*, L. Chiruzzo, A. Ritter, and L. Wang, Eds., Albuquerque, New Mexico: Association for Computational Linguistics, Apr. 2025, pp. 4779–4807, ISBN: 979-8-89176-195-7. URL: https://aclanthology.org/2025.findings-naacl.269/.

2. Y. Cao, **T. Zhang**, B. Cao, *et al.*, "Personalized steering of large language models: Versatile steering vectors through bi-directional preference optimization," in *Advances in Neural Information Processing Systems*, A. Globerson, L. Mackey, D. Belgrave, *et al.*, Eds., vol. 37, Curran Associates, Inc., 2024, pp. 49 519–49 551. URL: https://proceedings.neurips.cc/paper_files/paper/2024/file/58cbe393b4254da8966780a40d023c0b-Paper-Conference.pdf.

3. Z. Yin, M. Ye, **T. Zhang**, *et al.*, "Vqattack: Transferable adversarial attacks on visual question answering via pre-trained models," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, 2024, pp. 6755–6763.

4. **T. Zhang**, Z. Xi, T. Wang, P. Mitra, and J. Chen, "PromptFix: Few-shot backdoor removal via adversarial prompt tuning," in *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, K. Duh, H. Gomez, and S. Bethard, Eds., Mexico City, Mexico: Association for Computational Linguistics, Jun. 2024, pp. 3212–3225. DOI: 10.18653/v1/2024.naacl-long.177.

5. M. Ye, Z. Yin, **T. Zhang**, *et al.*, "Unit: A unified look at certified robust training against text adversarial perturbation," in *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL: https://openreview.net/forum?id=oGxE2Nvlda.

6. Z. Yin, M. Ye, **T. Zhang**, *et al.*, "VLATTACK: Multimodal adversarial attacks on vision-language tasks via pre-trained models," in *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL: https://openreview.net/forum?id=qBAED3u1XZ.

Miscellaneous Experience

Teaching

2023 - 2024	**TA, Applied Data Sciences**, Pennsylvania State University.
	TA, Data Mining: Techniques and Applications, Pennsylvania State University.
2020 - 2021	**TA, Database Management**, University of Michigan.
2019 - 2020	**TA, Introductino to Algorithms**, University of Michigan.

References

Available on Request