Zhuoxin Zhan

Vancouver, BC, Canada · (+1) 778-683-2849 · zhuoxin_zhan@sfu.ca
zxzhan.github.io · Google Scholar · GitHub · LinkedIn

RESEARCH INTERESTS

AI safety and adversarial robustness of large language models: red-teaming of LLM and computer-use agents, indirect prompt injection, jailbreaking, and safety alignment.

EDUCATION

Simon Fraser University, Burnaby, Canada — Sep 2022 – Present
Ph.D. in Computing Science, advised by Prof. Ke Wang · GPA **4.21**/4.33
- **Awards:** Special Graduate Entrance Scholarship, Graduate Fellowship (×6).

Shenzhen University, Shenzhen, China — Sep 2018 – Jun 2022
B.E. with Honors in Artificial Intelligence · GPA **3.96**/4.5 (Rank **4**/103)
- **Awards:** Outstanding Graduate, Outstanding Bachelor Thesis, Star of SZU Scholarship, MCM Meritorious Winner.

RESEARCH EXPERIENCE

RBC Borealis (Borealis AI), Vancouver, Canada — Jan 2026 – May 2026
Machine Learning Researcher (Intern) · Computer-Use Agent Safety
- Introduced ***multi-step indirect prompt injection***, a new attack class that decomposes an adversarial goal into innocuous sub-steps spread across a chain of web pages.
- Built an automatic LLM-driven pipeline for adversarial goal decomposition.
- Released the CUA safety benchmark ***StepJack***, and evaluated six state-of-the-art computer-use agents.
- Raised attack success rate on GPT-5.4-mini from **41.7%** at single-step to **72.9%** at three-step.
- Research paper submitted, with patent filed and code open-sourced (see Preprint [1])

PUBLICATIONS & PREPRINTS

[1] **Zhuoxin Zhan**, A. Rafiey, A. Ma, L. Pishdad, L. El Asri. **StepJack: Benchmarking Computer-Use Agent Safety Against Multi-Step Indirect Prompt Injection.** *Preprint, arXiv 2026.* [Paper] [Code]

[2] **Zhuoxin Zhan**, K. Wang, P. Xiong, L. Li. **Benign Prompts Can Jailbreak Large Language Models.** *Under submission, 2026.*
- Demonstrated that benign prompts can trigger harmful responses in safety aligned LLMs.
- Proposed ***Benign Prompt Attack Paradigm*** for boosting existing jailbreak attacks, lifting attack success rates by **35–45%**.
- Introduced adaptive safety alignment by incorporating (Benign prompt, Harmful response) conversations to mitigate the new attacks.

[3] **Zhuoxin Zhan**, K. Wang, P. Xiong. **Accelerating Adversarial Training on Under-Utilized GPU.** *IJCAI 2025.* [Paper] [Code]
- Proposed ***AttackRider*****,** consisting of a subset selection method and an example packing method, to accelerate adversarial training.
- Achieved **2–4**× speedup at comparable robust accuracy across NN, CNN, and Transformer models on image and tabular tasks.

[4] **Zhuoxin Zhan**, M. He, W. Pan, Z. Ming. **TransRec++: Translation-Based Sequential Recommendation with Heterogeneous Feedback.** *Frontiers of Computer Science, 2022.* [Paper]

[5] **Zhuoxin Zhan**, L. Zhong, J. Lin, W. Pan, Z. Ming. **Sequence-Aware Similarity Learning for Next-Item Recommendation.** *The Journal of Supercomputing, 2021.*

[6] Y. Ni, **Zhuoxin Zhan**, W. Pan, Z. Ming. **Asymmetric Pairwise Preference Learning for Heterogeneous One-Class Collaborative Filtering.** *ICONIP 2020.*

[7] J. Lin, **Zhuoxin Zhan**, W. Pan, Z. Ming. **Single-Behavior Sequential Recommendation.** *Book chapter, Intelligent Recommendation Technology, Tsinghua University Press, 2022 (in Chinese).* [Book]

SERVICE & LEADERSHIP

Conference Reviewer: NeurIPS ’26, ICML ’26, EMNLP ’26, WWW ’26, WSDM ’25, CIKM ’25/’24, KDD ’24, BigData ’24.

Student Chair, Artificial Intelligence Club, Shenzhen University — Sep 2019 – Jul 2021
- Collaborated with sponsors (Tencent, IBM, Baidu) to organize events such as IBM summer camps and Baidu seminars.

TECHNICAL SKILLS

Languages: Python, C/C++, Java, SQL
ML / AI: PyTorch, HuggingFace Transformers, vLLM, scikit-learn, NumPy, Pandas, Weights & Biases