Xiaofei Wen

Davis, CA | xfwe@ucdavis.edu | (279)766-9275 | Homepage | Linkedin | Github

Bio

Xiaofei Wen is a Computer Science Ph.D. student at University of California, Davis, advised by Prof. Muhao Chen.

His research interest lies in improving the reliability and behavior of language and video models in general-purpose applications.

Publication

[a] Towards Policy-Compliant Agents: Learning Efficient Guardrails For Policy Violation Detection Oct 2025

Xiaofei Wen, Wenjie Jacky Mo, Yanan Xie, Peng Qi, Muhao Chen

Preprint [paper] [webpage]

[b] REDCODER: Automated Multi-Turn Red Teaming for Code LLMs July 2025

Wenjie Jacky Mo, Qin Liu, **Xiaofei Wen**, Dongwon Jung, Hadi Askari, Wenxuan Zhou, Zhe Zhao, Muhao Chen

Preprint [paper] [code]

[c] Diagnosing and Mitigating Modality Interference in Multimodal Large Language Models May 2025

Rui Cai, Bangzheng Li, **Xiaofei Wen**, Muhao Chen, Zhe Zhao

Preprint [paper] [code]

[d] ThinkGuard: Deliberative Slow Thinking Leads to Cautious Guardrails Feb 2025

Xiaofei Wen, Wenxuan Zhou, Wenjie Jacky Mo, Muhao Chen

ACL 2025 Findings [paper] [code]

[e] Red Teaming Language Models for Processing Contradictory Dialogues Nov 2024

Xiaofei Wen, Bangzheng Li, Tenghao Huang, Muhao Chen

EMNLP 2024 [paper] [code]

[f] Personalized Topic Selection Model for Topic-Grounded Dialogue Aug 2024

Shixuan Fan, Wei Wei, **Xiaofei Wen**, Xianling Mao, Jixiong Chen, Danyang Chen

ACL 2024 Findings [paper]

[g] Sequential Topic Selection Model with Latent Variable for Topic-Grounded Dialogue Dec 2022

Xiaofei Wen, Wei Wei, Xianling Mao

EMNLP 2022 Findings [paper]

Education

University of California, Davis, Ph.D. in Computer Science Sept 2024 - Present

- **GPA:** 3.95/4.00

Huazhong University of Science and Technology, M.S. in Computer Science Sept 2021 – June 2024

- **GPA:** 3.90/4.00
- **Awards:** First-Class Graduate Scholarship

Dalian University of Technology, B.Eng in Software Engineering Sept 2017 – June 2021

- **GPA:** 3.89/4.00
- **Awards:** Mitsubishi Scholarship, Outstanding Scholarship, Meritorious Winner of 2019 ICM