Xiaomeng Hu Ph.D. student

(+852) 53931720
greghxm@link.cuhk.edu.hk
https://gregxmhu.github.io/

Education

The Chinese University of Hong Kong — Hong Kong, China
Ph.D. in Computer Science and Engineering — 2023 - 2027 *(expected)*
- Advisor: Prof. Tsung-Yi Ho and Dr. Pin-Yu Chen (IBM Research)
- Research area: LLMs & Trustworthy Machine Learning

Northeastern University — Shenyang, China
B.E. in Artificial Intelligence — 2019 - 2023
- GPA: 4.22/5.00, Rank: 1/123.

Publications

1. **Xiaomeng Hu**, Pin-Yu Chen, Tsung-Yi Ho. Token Highlighter: Inspecting and Mitigating Jailbreak Prompts for Large Language Models. *Under Review*, 2024.
2. **Xiaomeng Hu**, Pin-Yu Chen, Tsung-Yi Ho. Gradient Cuff: Detecting Jailbreak Attacks on Large Language Models by Exploring Refusal Loss Landscapes. *NeurIPS 2024*, 2024.
3. **Xiaomeng Hu**, Pin-Yu Chen, Tsung-Yi Ho. RADAR: Robust AI-Text Detection via Adversarial Learning. *NeurIPS 2023*, 2023.
4. **Xiaomeng Hu**, Shi Yu, Chenyan Xiong, Zhenghao Liu, Zhiyuan Liu, Ge Yu. P^3 Ranker: Mitigating the Gaps between Pre-training and Ranking Fine-tuning with Prompt-based Learning and Pre-finetuning. *SIGIR 2022*, 2022.

Projects

Jailbreak Prompt Detection
Detecting Jailbreak prompts via checking the refusal loss landscape. 2023.08 - 2023.12

AI-Text Detection
Training a robust ai-text detector via adversarial learning 2023.03 - 2023.05

Prompt-based document reranking
Using prompt-based learning to achieve few-shot document reranking. 2022.08 - 2022.12

Awards and Honors (selected)

- **NeurIPS Scholar Award**, NeurIPS 2024 — 2024.10
- **Full Postgraduate Scholarship**, The Chinese University of Hong Kong, — 2023.08
- **First Class Scholarship**, Northeastern University — 2022.09
- **Outstanding Student Leader**, Northeastern University — 2021.09
- **First Prize Award**, The Chinese Mathematics Competitions — 2020.12
- **Second Prize Award**, The Chinese Physics Olympiad — 2018.09

Skills

Languages: Chinese, English, Cantonese.
Programming: `Python`, `C++`, `MATLAB`, `Latex`.

Academic Services

Reviewers for: *ICLR 2025*, *NeurIPS Adv-ML 2023*

Invited speaker for: *AIS 2022*