Wenbin Hu

Year-1 PhD Student
Department of Computer Science and Engineering
Hong Kong University of Science and Technology, HK

☎ +852-63405208
✉ whuak@connect.ust.hk
GitHub Profile
LinkedIn Profile

Education

Hong Kong University of Science and Technology, HK *Feb 2025 - Now*
Ph.D in Computer Science

- **Scholarship**: Full Scholarship.

Hong Kong University of Science and Technology, HK *Sept 2020- Jul 2024*
B.Eng in Computer Science

- **Awards**: Dean's List (Term GPA > 3.7, Top 10 Percentile).

Technical University of Munich, Munich *Apr 2023 - Aug 2023*
Exchange Programme in TUM Informatics

- **Awards**: Graded the highest level of academic performance: 'Very Good'

Fudan University, Shanghai *Jun 2022 - Aug 2022*
Summer Exchange Programme

- **Research Topic**: Bayesian Inference and Learning.

Publications

- **Context Reasoner: Incentivizing Reasoning Capability for Contextualized Privacy and Safety Compliance via Reinforcement Learning**
  Wenbin Hu, *Haoran Li, Huihao Jing, Qi Hu, Ziqian Zeng, Sirui Han, Heli Xu, Tianshu Chu, Peizhao Hu, Yangqiu Song. 2025, Arxiv Preprint.* [paper]

- **MCIP: Protecting MCP Safety via Model Contextual Integrity Protocol**
  Huihao Jing, Haoran Li, ***Wenbin Hu****, Qi Hu, Heli Xu, Tianshu Chu, Peizhao Hu, Yangqiu Song. 2025, Arxiv Preprint.* [paper]

- **PrivaCI-Bench: Evaluating Privacy with Contextual Integrity and Legal Compliance**
  Haoran Li,* ***Wenbin Hu*****, Huihao Jing*, Yulin Chen, Qi Hu, Sirui Han, Tianshu Chu, Peizhao Hu, Yangqiu Song. ACL 2025 Main.* [paper]

- **Node Level Graph Autoencoder: Unified Pretraining for Textual Graph Learning**
  Wenbin Hu**, Huihao Jing*, Qi Hu*, Haoran Li, Yangqiu Song. 2024, Arxiv Preprint.* [paper]

- **Mitigating the Alignment Tax of RLHF**
  Yong Lin, Hangyu Lin, Wei Xiong, Shizhe Diao, Jianmeng Liu, Jipeng Zhang, Rui Pan, Haoxiang Wang, ***Wenbin Hu****, Hanning Zhang, Hanze Dong, Renjie Pi, Han Zhao, Nan Jiang, Heng Ji, Yuan Yao, Tong Zhang. EMNLP 2024 Main.* [paper]

- **Attacking by Aligning: Clean-Label Backdoor Attacks on Object Detection**
  Yize Cheng,* ***Wenbin Hu*****, Minhao Cheng. 2023, Arxiv preprint.* [paper]

(* represents equal contribution.)

Core Courses

- **AI Courses**: Machine Learning, Deep Learning, Computer Vision, Natural Language Processing, Graph Machine Learning, Reinforcement Learning.

- **CS Courses**: C++, OOP, Computer Architecture, Operation System, Algorithm, Computer Networking, Software Engineering, Computer Graphics, Web Search Engine.

- **MATH Courses**: Multivariable Calculus, Linear Algebra, Abstract Algebra, Probability Theory, Convex Optimization, Mathematic Analysis, Ordinary Differential Equation.