

Sean Xuefeng Du

College of Computing and Data Science
Nanyang Technological University - NTU Singapore
50 Nanyang Ave, Block N 4, Singapore 639798
[Homepage](#), [Scholar](#), +1-608-720-4664
xuefeng.du@ntu.edu.sg, xfdu@cs.wisc.edu, xuefengdu1@gmail.com

Academic Position	<i>Assistant Professor</i> College of Computing and Data Science, Nanyang Technological University - Singapore	Sept. 2025-Present
Research Interests	Reliable machine learning, foundation model reliability, and related applications, specifically I work on algorithm and theory design for: <ul style="list-style-type: none">Out-of-distribution (OOD) research: designing adaptive and interpretable learning principles that help ML models detect and generalize on OOD samples, such as semantic/covariate shifts, adversarial and noisy samples.Reliability of foundation models: understanding the blindspots of Large Language Models (LLMs) and Multimodal LLMs through uncertainty estimation, such as model hallucination detection and mitigation.Security of foundation models: Exploring how foundation models can be strengthened against emerging threats, including prompt injection, jailbreak attacks, and distributional vulnerabilities.Other interested topics: LLM alignment, Agentic AI.Interdisciplinary research: Biometrics, AI-aided cryo-microscopy/protein structural analysis.	
Education	<i>Ph.D. in Computer Sciences</i> University of Wisconsin–Madison , Madison, WI <ul style="list-style-type: none">Ph.D. research in reliable machine learning.Advisor: Prof. Sharon Y. Li.Dissertation: Foundations of Unknown-aware Machine LearningThesis Committee Members: Profs. Jerry Zhu, Robert D. Nowak, Yong Jae Lee.	Jan. 2021-May 2025
	<i>B. Eng. in Electrical Engineering</i> Xi'an Jiaotong University , Xi'an, China <ul style="list-style-type: none">Overall GPA: 91.60/100(3.83/4.0), Rank: 1st/170.Honors: National Scholarships (2017, 2018).	Sept. 2016 – Jun. 2020
Awards	<ul style="list-style-type: none">AAAI-26 New Faculty Highlights Program.CS Ivanisevic Award in UW-Madison (\$3,000, awarding to one student in the department with research excellence), 2025.Rising Stars in Data Science, September, 2024 (worldwide 30 recipients).Jane Street Graduate Research fellowship (\$50,000 grant) (6 selected out of 600), April, 2023.NeurIPS Scholar Award, Novemeber, 2022.CS departmental research fellowship, UW-Madison, September, 2021.National Scholarship (2x),Ministry of Education in China, 2017-2018.	

- Publications** See full list in my google scholar page, I publish under the name "Xuefeng Du".
 * indicates Equal Contribution
- 48. Limited Preference Data? Learning Better Reward Model with Latent Space Synthesis
 Leitian Tao, **Xuefeng Du**, Yixuan Li.
 NeurIPS 2025.
 - 47. How to Steer LLM Latents for Hallucination Detection?
 Seongheon Park, **Xuefeng Du**, Min-Hsuan Yeh, Haobo Wang, Yixuan Li.
 ICML 2025.
 - 46. Understanding Multimodal LLMs Under Distribution Shifts: An Information-Theoretic Approach
 Changdae Oh, Zhen Fang, Shawn Im, **Xuefeng Du**, Yixuan Li.
 ICML 2025.
 - 45. Challenges and Future Directions of Data-Centric AI Alignment
 Min-Hsuan Yeh, Jeffrey Wang, **Xuefeng Du**, Seongheon Park, Leitian Tao, Shawn Im, Yixuan Li.
 ICML 2025
 - 44. Robust Palmprint Recognition via Multi-stage Noisy Label Selection and Correction
 Huikai Shao, Siyu Shi, **Xuefeng Du**, Dan Zeng, Dexing Zhong.
 IEEE Transactions on Image Processing.
 - 43. Out-of-Distribution Learning with Human Feedback
 Haoyue Bai, **Xuefeng Du**, Katie Rainey, Shibin Parameswaran, Yixuan Li
 TMLR 2025.
 - 42. Safety-Aware Fine-Tuning of Large Language Models
 Hyeong Kyu Choi, **Xuefeng Du**, Yixuan Li
 Advances in Neural Information Processing Systems 2024, Safe Generative AI Workshop.
 - 41. HaloScope: Harnessing Unlabeled LLM Generations for Hallucination Detection
Xuefeng Du, Chaowei Xiao, Yixuan Li
 Advances in Neural Information Processing Systems 2024, spotlight.
 - 40. Exploring Transition States of Protein Conformational Changes via Out-of-Distribution Detection in the Hyperspherical Latent Space
 Bojun Liu, Jordan G Boysen, Ilona Christy Unarta, **Xuefeng Du**, Yixuan Li, Xuhui Huang
 Nature Communications, 2024
 - 39. When and How does In-distribution Label Help Out-of-distribution Detection?
Xuefeng Du, Yiyou Sun, Yixuan Li
 International Conference on Machine Learning (ICML) 2024.
 - 38. How does Unlabeled Data Provably Help Out-of-distribution Detection?
Xuefeng Du*, Zhen Fang*, Ilias Diakonikolas, Yixuan Li
 International Conference on Learning Representations (ICLR) 2024.
 - 37. Feed Two Birds with One Scone: Exploiting Wild Data for Both Out-of-Distribution Generalization and Detection
 Haoyue Bai, Gregory Canal, **Xuefeng Du**, Jeongyeol Kwon, Robert D Nowak, Yixuan Li
 International Conference on Machine Learning (ICML) 2023.

36. Dream the Impossible: Outlier Imagination with Diffusion Models
Xuefeng Du, Yiyou Sun, Jerry Zhu, Yixuan Li
Advances in Neural Information Processing Systems (NeurIPS) 2023.
35. OpenOOD v1.5: Enhanced Benchmark for Out-of-Distribution Detection
Jingyang Zhang, Jingkang Yang, Pengyun Wang, Haoqi Wang, Yueqian Lin, Haoran Zhang, Yiyou Sun, **Xuefeng Du**, Kaiyang Zhou, Wayne Zhang, Yixuan Li, Ziwei Liu, Yiran Chen, Hai Li
Journal of Data-centric Machine Learning Research.
34. Non-parametric Outlier Synthesis
Leitian Tao, **Xuefeng Du**, Jerry Zhu, Yixuan Li
International Conference on Learning Representations (ICLR) 2023.
33. Noise-robust Graph Learning by Estimating and Leveraging Pairwise Interactions
Xuefeng Du, Tian Bian, Yu Rong, Bo Han, Tongliang Liu, Tingyang Xu, Wenbing Huang, Yixuan Li, Junzhou Huang
Transactions on Machine Learning Research (TMLR) 2023
32. SIREN: Shaping Representations for Detecting Out-of-distribution Objects
Xuefeng Du, Gabriel Gozum, Yifei Ming, Yixuan Li
Advances in Neural Information Processing Systems (NeurIPS) 2022.
31. OpenOOD: Benchmarking Generalized Out-of-Distribution Detection
Jingkang Yang, Pengyun Wang, Dejian Zou, Zitang Zhou, Kunyuan Ding, Wenxuan Peng, Haoqi Wang, Guangyao Chen, Bo Li, Yiyou Sun, **Xuefeng Du**, Kaiyang Zhou, Wayne Zhang, Dan Hendrycks, Yixuan Li, Ziwei Liu
Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track, 2022.
30. Unknown-Aware Object Detection: Learning What You Don't Know from Videos in the Wild
Xuefeng Du, Xin Wang, Gabriel Gozum, Yixuan Li
IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) 2022, oral.
29. Performance-Aware Mutual Knowledge Distillation for Improving Neural Architecture Search
Pengtao Xie, **Xuefeng Du**
IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) 2022.
28. VOS: Learning What You Don't Know by Virtual Outliers Synthesis
Xuefeng Du, Eric Wang, Mu Cai, Yixuan Li
International Conference on Learning Representations (ICLR) 2022.
27. Learning Diverse-Structured Networks for Adversarial Robustness
Xuefeng Du, Jingfeng Zhang, Bo Han, Tongliang Liu, Yu Rong, Gang Niu, Junzhou Huang, Masashi Sugiyama
International Conference on Machine Learning (ICML) 2021.
26. How to save your annotation cost for Panoptic Segmentation?
Xuefeng Du, Chenhan Jiang, Hang Xu, Gengwei Zhang, Zhenguo Li
The AAAI Conference on Artificial Intelligence (AAAI) 2021.
25. Active learning to classify macromolecular structures *in situ* for less supervision in cryoelectron tomography
Xuefeng Du, Haohan Wang, Zhenxi Zhu, Xiangrui Zeng, Yi-Wei Chang, Jing Zhang, Eric Xing, Min Xu
Bioinformatics, 2021

24. Node Classification on Graphs with Few-Shot Novel Labels via Meta Transformed Network Embedding
Lin Lan, Pingui Wang, **Xuefeng Du**, Kaikai Song, Jing Tao, Xiaohong Guan
Advances in Neural Information Processing Systems (NeurIPS) 2020.
23. A deep biometric hash learning framework for three advanced hand-based biometrics.
Huikai Shao, Dexing Zhong, **Xuefeng Du**
IET Biometrics.
22. Few-shot learning for palmprint recognition via meta-siamese network
Huikai Shao, Dexing Zhong, **Xuefeng Du**, Shaoyi Du, Raymond NJ Veldhuis
IEEE transactions on instrumentation and measurement.
21. Deep distillation hashing for unconstrained palmprint recognition
Huikai Shao, Dexing Zhong, **Xuefeng Du**
IEEE transactions on instrumentation and measurement.
20. Cross-domain palmprint recognition via regularized adversarial domain adaptive hashing
Xuefeng Du, Dexing Zhong, Huikai Shao
IEEE Transactions on Circuits and Systems for Video Technology.
19. Effective deep ensemble hashing for open-set palmprint recognition
Huikai Shao, Dexing Zhong, **Xuefeng Du**
Journal of Electronic Imaging.
18. Cross-domain palmprint recognition based on transfer convolutional autoencoder
Huikai Shao, Dexing Zhong, **Xuefeng Du**
IEEE International Conference on Image Processing (ICIP) 2019.
17. Continual palmprint recognition without forgetting
Xuefeng Du, Dexing Zhong, Huikai Shao
IEEE International Conference on Image Processing (ICIP) 2019.
16. Building an active palmprint recognition system
Xuefeng Du, Dexing Zhong, Huikai Shao
IEEE International Conference on Image Processing (ICIP) 2019.
15. Low-shot palmprint recognition based on meta-siamese network
Xuefeng Du, Dexing Zhong, Pengna Li
IEEE International Conference on Multimedia and Expo (ICME) 2019.
14. A hand-based multi-biometrics via deep hashing network and biometric graph matching
Dexing Zhong, Huikai Shao, **Xuefeng Du**
IEEE Transactions on Information Forensics and Security.
13. Decade progress of palmprint recognition: A brief survey
Dexing Zhong, **Xuefeng Du**, Kuncai Zhong
Neurocomputing.
12. Classification in Cryo-Electron Tomograms
Ilja Gubins, Gijs van der Schot, Remco C Veltkamp, Friedrich Förster, **Xuefeng Du**, Xiangrui Zeng, Zhenxi Zhu, Lufan Chang, Min Xu, Emmanuel Moebel, Antonio Martinez-Sanchez, Charles Kervrann, Tuan M Lai, Xusi Han, Genki Terashi, Daisuke Kihara, Benjamin A Himes, Xiaohua Wan, Jingrong Zhang, Shan Gao, Yu Hao, Zhilong Lv, Xiaohua Wan, Zhidong Yang, Zijun Ding, Xuefeng Cui, Fa Zhang
Eurographics 2019.

11. Open-set Recognition of Unseen Macromolecules in Cellular Electron Cryo-Tomograms by Soft Large Margin Centralized Cosine Loss
Xuefeng Du, Xiangrui Zeng, Bo Zhou, Alex Singh and Min Xu
British Machine Vision Conference (BMVC) 2019, **Spotlight**.
10. Semi-supervised Macromolecule Structural Classification in Cellular Electron Cryo-Tomograms using 3D Autoencoding Classifier
Siyuan Liu, **Xuefeng Du**, Rong Xi, Fuya Xu, Xiangrui Zeng, Bo Zhou and Min Xu
British Machine Vision Conference (BMVC) 2019.
9. Efficient Deep Palmprint Recognition via Distilled Hashing Coding Huikai Shao, Dexing Zhong and **Xuefeng Du**
IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops 2019.
8. Palm vein recognition with deep hashing network
Dexing Zhong, Shuming Liu, Wenting Wang, **Xuefeng Du**
Pattern Recognition and Computer Vision 2019.
7. Palmprint recognition using siamese network
Dexing Zhong, Yuan Yang, **Xuefeng Du**
13th Conference of Biometric Recognition, CCBR 2018.

Preprints

6. VLMGuard: Defending VLMs against Malicious Prompts via Unlabeled Data
Xuefeng Du, Reshma Ghosh, Robert Sim, Ahmed Salem, Vitor Carvalho, Emily Lawton, Yixuan Li, Jack W. Stokes.
arXiv preprint arXiv:2410.00296.
5. The Ghanaian NLP Landscape: A First Look
Sheriff Issaka, Zhaoyi Zhang, Mihir Heda, Keyi Wang, Yinka Ajibola, Ryan De-Mar, **Xuefeng Du**
arXiv preprint arXiv:2405.06818.
4. Learning by Passing Tests, with Application to Neural Architecture Search
Xuefeng Du, Pengtao Xie
arXiv preprint arXiv:2011.15102.
3. Skillearn: Machine Learning Inspired by Humans' Learning Skills
Pengtao Xie, **Xuefeng Du**, Hao Ban
arXiv preprint arXiv:2012.04863.
2. Small-Group Learning, with Application to Neural Architecture Search
Xuefeng Du, Pengtao Xie
arXiv preprint arXiv:2012.12502.
1. Towards efficient unconstrained palmprint recognition via deep distillation hashing
Huikai Shao, Dexing Zhong, **Xuefeng Du**
arXiv preprint arXiv:2004.03303.

Patents

- VisionGuard: Detecting Multimodal AI Systems from Indirect Prompt Injections (pending)
Xuefeng Du, Reshma Ghosh, Vitor Carvalho, Robert Sim, Emily Lawton, Jay Stokes, Lukas Wutschitz, Reshma Ghosh, Ahmed Salem
- A kind of cross-platform palm grain identification method
Dexing Zhong, Huikai Shao, Xuefeng Du, Runzhao Yao

- A kind of vehicle early warning vehicle intelligent system based on multisensor
Dexing Zhong, Huikai Shao, Xuefeng Du

Grants

- NTU CCDS, Start-up grant, *Reliable Foundation Models: Theory and Algorithms for Detecting and Mitigating Hallucinations in LLMs and Multimodal Systems*, S\$150,000.
- MOE Academic Research Fund (AcRF) Tier 1 Seed Funding, *From Output to Process: Calibrated Confidence Estimation for Large AI Models*, S\$100,000.

Service

Area Chair for ICLR 2026, ICML 2026

Reviewer for

- NeurIPS, ICML, ICLR
- CVPR, ECCV, ICCV
- WACV, AAAI, IJCAI, MICCAI
- TMLR, Nature Communications, IJCV, IEEE TCSVT, TMM, TIP, TIFS

University Service

- UM-Madison, Admission Committee, Master for Data Science Program, 2024.
- “Learning with AI” evaluation panel, NTU CCDS, 2025.

Workshop Organization

- Co-organizer, *special session of Responsible Foundation Models in the Wild* at IJCNN 2025.
- Co-organizer, *Socially Responsible and Trustworthy Foundation Models (ResponsibleFM)* workshop at NeurIPS 2025.

Students and Mentees

Current Students

5. Junlin Fang (PhD student, CCDS, Spring 2026-)
4. Jianxiong Zhang (Visiting PhD student, CCDS, Fall 2025-Spring 2026)
3. Shunzhang Chen (MSAI, CCDS, Fall 2025-)
2. Wenyu Chen (Msc in Modeling and Simulation, SPMS, Fall 2025-)
1. Yeqing Tang (Undergraduate student, CCDS, Fall 2025-)

Past Students

Qualification Exam Committees Served On

1. Yuzhou Cao (PhD student 2025, NTU CCDS)

Exam topic: Convex, Smooth, and Consistent Surrogate Losses with Linear Surrogate Regret for General Prediction Tasks
Advisor: Bo An

Dissertation Committees Served On

2. Zonglin Yang (PhD 2025, NTU CCDS)

Thesis topic: From Inductive Reasoning to Automated Scientific Discovery: Foundations and Methods with Large Language Models
Advisor: Erik Cambria
1. Teng Yao Long (PhD 2029, NTU CCDS)

Thesis topic:
Advisor: Bo An

Final-year project (FYP)

2. Raphael Liew Jin Cheng (2025)
Role: Examiner (Supervisor: Erik Cambria)
Project title: AI-Driven Analyst Report Generation: A Multi-Agent LLM Approach to Investment Guidance
1. Hazim Khoiruddin (2025)
Role: Examiner (Supervisor: Oh Hong Lye)
Project title: One-stop knowledge center for course modules

Mentees before tenure-track

9. [Seongheon Park](#), PhD student at UW-Madison, now PhD student at UW-Madison.
8. [Sheriff Issaka](#), undergraduate student at UW-Madison, 2024, now PhD student at UCLA.
7. [Leitian Tao](#), undergraduate student at WHU, 2022, now PhD student at UW-Madison.
6. [Tian Bian](#), PhD student at CUHK, 2021, now PhD student at CUHK.
5. [Gabriel Gozum](#), undergraduate student at UW-Madison, 2021, now perception engineer at Applied Intuition.
4. [Eric Wang](#), undergraduate student at UW-Madison, 2021, now research assistant at UCSD.
3. Zhenxi Zhu, undergraduate student at BUPT, 2019, now PhD student at Nanjing University.
2. [Alex Singh](#), undergraduate student at CMU, 2019, now AI Engineer at Tesla.
1. [Pengna Li](#), undergraduate student at XJTU, 2019, now PhD student at XJTU.

Teaching

Instructor for

- NTU, SC 5005 NLP, LLM and Applications (Spring 2026)

Exam Moderator for

- NTU, SC 4000 / CZ4041 Machine Learning (Fall 2025)

Teaching Assistant for

- [CS762](#) (graduate course), UW-Madison: Advanced Deep Learning, Fall 2022.
–Designing lectures on topics in trustworthy machine learning, grading students' written papers, holding office hours, leading discussions and answering questions.
- [CS540](#) (undergraduate course), UW-Madison: Intro to AI, Spring 2021.
–Designing lectures on topics in unsupervised learning, deep learning, and reinforcement learning. Preparing all the homework related to the topic of deep learning, and design final exams, guiding students on final projects, holding office hours, and delivering discussion sections.

**Past / Upcoming
Talks**

- Talk at AAAI-26 new faculty highlights program, 1/2026
- Talk at Huawei Singapore Research Center, 11/2025
- Talk at University of Queensland, 6/2025
- Talk at MBZUAI, 5/2025
- Talk at Microsoft Research, 4/2025
- Talk at Pennsylvania State University, 3/2025

- Talk at University of Texas at Arlington, 3/2025
- Talk at Oregon State University, 2/2025
- Talk at UC Riverside, Statistics and CS departments, 2/2025 and 3/2025
- Talk at Rochester Institute of Technology, 2/2025
- Talk at Johns Hopkins University, 2/2025
- Talk at OpenAI, 1/2025
- Talk at NTU CCDS, 1/2025
- Talk at HKUST CSE, 12/2024
- Talk at MLOPT Seminar, UW-Madison, 12/2024
- Talk at UCSD, 11/2024
- Talk at Microsoft, 08/2024
- Talk at Jane Street, 04/2023
- Talk at [AI talks](#), 02/2023
- Talk at [MLOPT Seminar](#), UW-Madison, 05/2022
- Talk at Google Research, 01/2022
- Talk at Adobe Research, 12/2021
- Talk at Microsoft Research, 12/2021

Media Coverage

- Author Interview on VOS: [Youtube video](#), [Author interview](#)
- Graduate Fellowship: [Department News](#)
- CS Ivanisevic Award: [Department News](#)
- Research on LLM safety: [Wisconsin Engineer Magazine](#)

Research Experience

<i>Research Intern</i>		June 2024 – Sept. 2024
Microsoft Research, Redmond, WA, USA		
<ul style="list-style-type: none"> • Hosts: Dr. Robert Sim, Jay Stokes and Reshma Ghosh. • Research on <i>Malicious prompt detection for vision language models</i> 		
<i>Student Researcher</i>		June 2022 – Sept. 2022
Google Research, Sunnyvale, CA, USA		
<ul style="list-style-type: none"> • Hosts: Dr. Zizhao Zhang, Ting Chen and Han Zhang. • Research on <i>Open-vocabulary object detection with language models</i> 		
<i>Research Intern</i>		Mar. 2021 – June 2021
Tencent AI Lab, Shenzhen, China		
<ul style="list-style-type: none"> • Supervised by Dr. Yu Rong and Junzhou Huang. • Research on <i>Robust graph neural networks against noisy labels</i> . 		
<i>Research Assistant</i>		Oct. 2020 – Jan. 2021
Department of Computer Science, Hong Kong Baptist University, Hong Kong		
<ul style="list-style-type: none"> • Supervised by Dr. Bo Han. • Research on <i>Effective network architecture for adversarial robustness</i>. 		
<i>Research Engineer Intern</i>		Aug. 2020 -Sept. 2020
AI Lab, ByteDance Inc., Beijing, China		
<ul style="list-style-type: none"> • Supervised by Dr. Changhu Wang. 		

- Research on *Fine-grained image classification*.

Research Intern

AI Theory Group, Noah's Ark Lab, Shenzhen, China

- Supervised by Dr. [Hang Xu](#) and [Chenhan Jiang](#).
- Research on *Hybrid supervised panoptic segmentation*.

Dec. 2019 -July 2020

Student Intern

Department of Computational Biology, CMU, Pittsburgh, PA, USA

- Supervised by Dr. [Min Xu](#) and Dr. [Haohan Wang](#).
- Research on *Deep learning for cellular electron cryo-tomography analysis*.

July 2018 – Feb. 2020

Research Assistant

Intelligent Networks and Network Security Lab, XJTU, Xi'an, China

- Supervised by Dr. [Pinghui Wang](#).
- Research on *Few-shot node classification with meta learning*.

Nov. 2018 – Jun. 2020

Research Assistant

Institute of Automatic Control, XJTU, Xi'an, China

- Supervised by Dr. [Dexing Zhong](#).
- Research on *Machine learning for hand-based biometrics*.

Jun. 2017 - Apr. 2019

Additional Information

Language skills

- Native speakers of Mandarin with professional English speaking capability .

Programming skills

- Proficient with Python, TensorFlow, PyTorch. Familiar with C, Matlab, C++.