# Yi Cai

## Msc in Computer Science

🔗 https://caiy0220.github.io/ https://github.com/caiy0220

✉ yi.cai@fu-berlin.de ☎ +49 152 28524403

🌐 Chinese (Native); English (IELTS 7.5); German (DSH2);

## EDUCATION

**PhD Candidate**, *Freie Universität Berlin*, Germany *April 2022 - Present*

Department of Mathematics and Computer Science, Cybersecurity and AI Group
Supervisor: Prof. Dr.-Ing. Gerhard Wunder

**M.Sc. in Computer Science**, *Leibniz Universität Hannover*, Germany *April 2016 - July 2018*

Faculty of Electrical Engineering and Computer Science
Thesis title: Clustering-based semi-supervised learning over streams, Grade: 1.0
Supervisor: Prof. Dr. Eirini Ntoutsi, Prof. Dr. Ralph Ewerth
**Overall grade:** ***1.7***

**B.Sc. in Computer Science and Technology**, *Xi'dian University*, China *August 2011 - July 2015*

School of Computer Science and Technology
Thesis title: Particle swarm optimization based software and hardware design for SoC, Grade: B
**Overall grade:** ***82/100***

## PUBLICATIONS

- **Yi Cai**, Thibaud Ardoin, Mayank Gulati, Gerhard Wunder (2026). "Rethinking Explanation Evaluation under the Retraining Scheme." Accepted at *Proceedings of the AAAI Conference on Artificial Intelligence*.
- Thibaud Ardoin, **Yi Cai**, Gerhard Wunder (2025). "Where Confabulation Lives: Latent Feature Discovery in LLMs." In: *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics.
- **Yi Cai**, Thibaud Ardoin, and Gerhard Wunder (2025). "GEFA: A General Feature Attribution Framework Using Proxy Gradient Estimation." In: *Proceedings of the 42nd International Conference on Machine Learning*. PMLR, pp. 5360–5382.
- **Yi Cai**, Arthur Zimek, Eirini Ntoutsi, and Gerhard Wunder (2024). "Transparent Neighborhood Approximation for Text Classifier Explanation by Probability-Based Editing". In: *IEEE 11th International Conference on Data Science and Advanced Analytics (DSAA)*. IEEE, pp. 1–10.
- **Yi Cai**, Gerhard Wunder (2024). "On Gradient-like Explanation under a Black-box Setting: When Black-box Explanations Become as Good as White-box". In: *Proceedings of the 41st International Conference on Machine Learning*. PMLR, pp. 5360–5382.
- **Yi Cai**, Arthur Zimek, Gerhard Wunder, and Eirini Ntoutsi (2022). "Power of Explanations: Towards automatic debiasing in hate speech detection". In: *IEEE 9th International Conference on Data Science and Advanced Analytics (DSAA)*. IEEE, pp. 1–10.
- **Yi Cai**, Arthur Zimek, and Eirini Ntoutsi (2021). "XPROAX-Local explanations for text classification with progressive neighborhood approximation". In: *IEEE 8th International Conference on Data Science and Advanced Analytics (DSAA)*, IEEE, pp. 1–10.

# PROJECTS

**Opportunities and risks of generative AI in cybersecurity (AIgenCY)** *January 2024 - Present*

Funded by the Federal Ministry for Education and Research, AIgenCY aims to research the methods of generative AI in relation to new types of attacks in cyberspace. At the same time, suitable measures are to be developed to improve the detection of and defence against such cyber attacks.

🔗 *https://www.forschung-it-sicherheit-kommunikationssysteme.de/projekte/aigency*

**Center for Trustworthy AI (ZVKI)** *May 2022 - December 2023*

Funded by the Federal Ministry for the Environment, Nature Conservation, Nuclear Safety and Consumer Protection, ZVKI aims to monitor the developments associated with AI and explore the legal and policy measures needed to protect people from possible negative effects of AI.

🔗 *https://www.zvki.de/*

**Responsible Artificial Intelligence** *November 2020 - April 2022*

Funded by the Ministry of Science and Culture of Lower Saxony, ResponsibleAI's goal is to design and apply AI systems in a reliable, transparent, secure, and legally acceptable way.

🔗 *https://verantwortungsvolleki.de/en/*

**KISWind** *April 2021 - February 2022*

Funded by the Federal Ministry for Economy and Energy, KISWind aims to contribute to the further development of automated damage detection in wind energy turbines based on acoustic emission testing (AET) and machine learning.

🔗 *https://www.l3s.de/research-at-l3s/all-projects/project-archive/kiswind/*)

# EXPERIENCE

**Freie Universität Berlin**, *Researcher Assistant* *April 2022 - Present*

- Active research in AI ethics and trustworthiness;
- Supervise B.Sc and M.Sc theses;
- Teaching assistant for: Cybersecurity and AI I (Winter semester 2022);
- Teaching assistant for: Cybersecurity and AI II (Summer semester 2023, Summer semester 2025);

**Leibniz Universität Hannover**, *Researcher Assistant* *November 2020 - April 2022*

- Active research in explainable artificial intelligence and application of generative AI;
- Supervise B.Sc and M.Sc theses.
- Teaching assistant for: Artificial Intelligence and Machine Learning (Winter semester 2021);

**Academic Service**

- IEEE Transactions on Neural Networks and Learning Systems (*TNNLS*), Reviewer;
- International Conference on Learning Representations (*ICLR'25*) Reviewer;
- Conference on Neural Information Processing Systems (*NeurIPS'25*) Reviewer;

**Ruijie Networks Co., Ltd.**, *Software Engineer* *July 2018 - July 2020*

- Designed the software framework of automatic guided vehicles (transport robots) for warehousing;
- Developed a SLAM algorithm with loose-coupling of visual and lidar solutions;
- Developed an auto-labeling system for a specific visual object detection task;
- Deep learning models deployment with edge computing