YAO RONG

Room 218, Appelstraße 4, 30419 Hannover, Germany

CONTACT INFORMATION

Assistant Professor of Computer Science, Leibniz University Hannover, Germany
Homepage: https://yaorong0921.github.io/homepage/
Email: yao.rong@l3s.de
Google Scholar: Google Scholar/Yao Rong

RESEARCH INTERESTS

My research focuses on Trustworthy AI by building actionable XAI (Explainable AI) that helps users understand model behavior, identify problems, and determine how to improve the AI system. My current research focuses on Large Language Models, with an emphasis on safety alignment throughout training and inference, as well as interpretability. I ground this trustworthy AI research in high-stakes domains such as healthcare, where reliability and transparency are essential.

ACADEMIC EMPLOYMENTS

Leibniz University Hannover, Germany — *April 2026 - Present*
Assistant Professor
Department of Electrical Engineering and Computer Science

Rice University, USA — *September 2024 – March 2026*
Postdoctoral Fellow
Department of Computer Science

EDUCATION

Technical University of Munich, Germany — *April 2023 – July 2024*
Ph.D., Computer Science — Advisor: Prof. Dr. Enkelejda Kasneci

University of Tübingen, Germany — *September 2019 – March 2023*
Ph.D. Candidate, Computer Science (Transfer Out) — Advisor: Prof. Dr. Enkelejda Kasneci

Technical University of Munich, Germany — *October 2016 – June 2019*
M.Sc., Electrical and Computer Engineering

Tongji University, China &
Munich University of Applied Sciences, Germany — *September 2012 – September 2016*
B.Eng., Mechatronics *(Dual-degree)*

HONORS & AWARDS

- **EECS Rising Star** at MIT. 2025
- **Future Faculty Fellow** at Rice University School of Engineering and Computing. 2025 – 2026
- **Rice Academy of Fellows**, Two-year postdoctoral fellowship. Amount: **$149,000**. 2024
- Seed Fund for EU Project Coordination, Technical University of Munich. Amount: **€ 25,000**. 2023
- Travel grant from Cluster of Excellence – Machine Learning, Tübingen, Germany. 2022

- Master study *passed with distinction*, Technical University of Munich, Germany. 2019
- First Prize of the Undergraduate Student Design Competition of Electrical System, Delphi Technologies, China. 2015
- Student Scholarships awarded by Tongji University, China. 2013 – 2015

GRANT WRITING

- **NSF RITEL Proposal**, Rice University. 2024
  Co-authored: Technology-assisted frameworks for teaching teamwork skills to STEM students.
- **Google Academic Research Award Proposal**, Rice University. 2024
  Co-authored: Leveraging AI to make education systems more equitable and accessible.
- **EU Horizon Europe Proposal**, Technical University of Munich. 2023
  Coordinator: Human-friendly deployment of AI and related technologies.
- **DFG Proposal** (German Research Foundation), Technical University of Munich. 2023
  Contributor: AI methods for simulation-based learning in higher education.

PUBLICATIONS

[IEEE TLT'25] **Yao Rong**, Katharina Seßler, Ekin Gözlüklü, and Enkelejda Kasneci. "Benchmarking In-Context Learning Strategies of Large Language Models for Math Reasoning Tasks." *IEEE Transactions on Learning Technologies.*

[AAAI RDS'26] **Yao Rong**, Shuo Yang, Gjergji Kasneci, Enkelejda Kasneci. "Synthetic Data Generation with LLMs through Strategic Comparisons." In *RDS @ AAAI.*

[AAAI (Spring Symposia)'25] **Yao Rong** and Vaibhav Unhelkar. "The Need for Human-AI Collaborative Methods for Conducting Audits of Machine Learning Models." In *AAAI Spring Symposium Series.*

[TKDD'24] **Yao Rong**, Guanchu Wang, Qizhang Feng, Ninghao Liu, Zirui Liu, Enkelejda Kasneci, and Xia Hu. "Efficient GNN Explanation via Learning Removal-based Attribution." In *ACM Transactions on Knowledge Discovery from Data.*

[xAI'24] **Yao Rong**, David Scheerer, and Enkelejda Kasneci. "Faithful Attention Explainer: Verbalizing Decisions Based on Discriminative Features." In *Proceedings of the 2nd World Conference on eXplainable Artificial Intelligence.*

[AAAI'24] **Yao Rong**, Peizhu Qian, Vaibhav Unhelkar, and Enkelejda Kasneci. "I-CEE: Tailoring Explanations of Image Classification Models to User Expertise." In *AAAI Conference on Artificial Intelligence.*

[ACL Findings'24] Shuo Yang, Chenchen Yuan, **Yao Rong**, Felix Steinbauer, and Gjergji Kasneci. "P-TA: Using Proximal Policy Optimization to Enhance Tabular Data Augmentation via Large Language Models." In *Findings of the Association for Computational Linguistics.*

[ETRA'24] Süleyman Özdel, **Yao Rong**, Berat Mert Albaba, Yen-Ling Kuo, Xi Wang, and Enkelejda Kasneci. "Gaze-Guided Graph Neural Network for Action Anticipation Conditioned on Intention." In *Proceedings of the ACM Symposium on Eye Tracking Research and Applications.*

[ETRA'24] Süleyman Özdel, **Yao Rong**, Berat Mert Albaba, Yen-Ling Kuo, Xi Wang, and Enkelejda Kasneci. "A Transformer-Based Model for the Prediction of Human Gaze Behavior on Videos." In *Proceedings of the ACM Symposium on Eye Tracking Research and Applications.*

[TPAMI'23] **Yao Rong**, Tobias Leemann, Thai-Trang Nguyen, Lisa Fiedler, Peizhu Qian, Vaibhav Unhelkar, Tina Seidel, Gjergji Kasneci, and Enkelejda Kasneci. "Towards Human-Centered Explainable AI: User Studies for Model Explanations." In *IEEE Transactions on Pattern Analysis and Machine Intelligence.*

[NeurIPS XAI'23] Tobias Leemann, **Yao Rong**, Thai-Trang Nguyen, Enkelejda Kasneci, and Gjergji Kasneci. "Caution to the Exemplars: On the Intriguing Effects of Example Choice on Human Trust in XAI." In *XAI in Action @ NeurIPS.*

[CVPRW'23] **Yao Rong**, Xiangyu Wei, Tianwei Lin, Yueyu Wang, and Enkelejda Kasneci. "DynStatF: An Efficient Feature Fusion Strategy for LiDAR 3D Object Detection." In *IEEE/CVF Conference on Computer Vision and Pattern Recognition.*

[UAI'23] Tobias Leemann, Michael Kirchhof, **Yao Rong**, Enkelejda Kasneci, and Gjergji Kasneci. "When are Post-hoc Conceptual Explanations Identifiable?" In *Conference on Uncertainty in Artificial Intelligence.*

[ICML'22] **Yao Rong**, Tobias Leemann, Vadim Borisov, Gjergji Kasneci, and Enkelejda Kasneci. "A Consistent and Efficient Evaluation Strategy for Attribution Methods." In *International Conference on Machine Learning.* **(Spotlight)**

[PACMHCI'22] **Yao Rong**, Naemi-Rebecca Kassautzki, Wolfgang Fuhl, and Enkelejda Kasneci. "Where and What: Driver Attention-based Object Detection." In *Proceedings of the ACM on Human-Computer Interaction.* **(Oral Presentation** at the *ACM Symposium on Eye Tracking Research and Applications.)*

[CHI TRAIT'22] **Yao Rong**, Nora Castner, Efe Bozkir, and Enkelejda Kasneci. "User Trust on an Explainable AI-Based Medical Diagnosis Support System." In *TRAIT Workshop at the ACM Conference on Human Factors in Computing Systems.*

[BMVC'21] **Yao Rong**, Wenjia Xu, Zeynep Akata, and Enkelejda Kasneci. "Human Attention in Fine-Grained Classification." In *British Machine Vision Conference.*

[ITSM'21] **Yao Rong**, Chao Han, Christian Hellert, Antje Loyal, and Enkelejda Kasneci. "Artificial Intelligence Methods in In-Cabin Use Cases: A Survey." In *IEEE Intelligent Transportation Systems Magazine.*

[ITSC'20] **Yao Rong**, Zeynep Akata, and Enkelejda Kasneci. "Driver Intention Anticipation Based on In-Cabin and Driving Scene Monitoring." In *IEEE International Conference on Intelligent Transportation Systems.*

[FG'20] Okan Köpüklü, Thomas Ledwon, **Yao Rong**, Neslihan Kose, and Gerhard Rigoll. "Drivermhg: A Multi-Modal Dataset for Dynamic Recognition of Driver Micro Hand Gestures and a Real-Time Recognition Framework." In *IEEE International Conference on Automatic Face and Gesture Recognition.*

[ICCVW'19] Okan Köpüklü, **Yao Rong**, and Gerhard Rigoll. "Talking with Your Hands: Scaling Hand Gesture Recognition with CNNs." In *IEEE/CVF International Conference on Computer Vision Workshops.*

Preprint and Under Review

[Under Review'25] Harrison Huang, **Yao Rong**, Peizhu Qian, and Vaibhav Unhelkar. "OOPS: Out-of-Distribution Policy Summarization." *Under Review.*

[Under Review'25] **Yao Rong** and Vaibhav Unhelkar. "Formalizing Audits of ML Models as a Sequential Decision-Making Problem." *Under Review.*

[Preprint'25] Franka Exner, Xufan Lu, Shaoming Zhang, Enkelejda Kasneci, and **Yao Rong**. "Leveraging AI to Predict and Explain Disease Incidence from Climate Data." *Preprint.*

[Preprint'25] Zilong Zhao, **Yao Rong**, Dongyang Guo, Emek Gözlüklü, Emir Gülboy, and Enkelejda Kasneci. "Stepwise Self-Consistent Mathematical Reasoning with Large Language Models." *arXiv Preprint.*

[Preprint'24] Enkelejda Kasneci, Hong Gao, Suleyman Ozdel, Virmarie Maquiling, Enkeleda Thaqi, Carrie Lau, **Yao Rong**, Gjergji Kasneci, Efe Bozkir. "Introduction to eye tracking: A hands-on tutorial for students and Practitioners." *arXiv Preprint.*

INVITED TALKS

Seminar, Computer Science Department, University of Houston *2025*
Title: "Actionable XAI for Understanding, Auditing, and Improving Models."

Chair of Hardware for Artificial Intelligence, Technical University of Darmstadt, Germany *2025*
Title: "Actionable XAI for Understanding, Auditing, and Improving Models."

Chair of Psychology of Action and Automation, Technical University of Berlin, Germany *2025*
Title: "Human Factors in Interpretable AI."

ECE Department, Leibniz University Hannover, Germany (Virtual) *2025*
Title: "Human-Centered Explainability: Bringing AI Closer to Human Reasoning."

Samsung Electronics America, Monthly Machine Learning Forum (Virtual) *2024*
Title: "Human-Centered Explainability: Bringing AI Closer to Human Reasoning."

Graduate Research Seminar in Machine Learning, Rice University *2024*
Title: "Promoting Human-Centered AI by Integrating Human Factors into Model Design."

TEACHING EXPERIENCE

Guest Lecturer, Department of Data Science, Rice University *Spring 2025*
Lecture: "Artificial Intelligence."

Guest Lecturer, Department of Psychological Sciences, Rice University *Fall 2024*
Lecture: "Human-Computer Interaction."

Instructor, Department of Educational Sciences, Technical University of Munich *Summer 2024*
Seminar: "Recent Advances in Human-Computer Interaction."

Instructor, Department of Educational Sciences, Technical University of Munich *Summer 2024*
Lecture-Tutorial: "Learning through Digitally Supported Instructional Designs."

Instructor, Department of Educational Sciences, Technical University of Munich *Fall 2023*
Lecture-Tutorial: "Human-AI Interaction."

Instructor, Department of Computer Science, University of Tübingen *Fall 2022*
Lecture-Tutorial: "Human-AI Interaction."

Instructor, Department of Computer Science, University of Tübingen *Fall 2021*
Seminar: "Advanced Topics in Human-Computer Interaction."

Instructor, Department of Computer Science, University of Tübingen *Fall 2020*
Seminar: "Introductory Topics in Human-Computer Interaction."

Guest Lecturer, Department of Computer Science, University of Tübingen *Fall 2020*
Lecture: "Multimodal Human-Computer Interaction."

SELECTED MENTORSHIP

Ph.D. Students

Tobias Kalmbach, Leibniz University Hannover *2026 – Present*
Project: LLM Safety Alignment against Multi-Turn Jailbreaks

Christian Kalfar, Leibniz University Hannover *2026 – Present*
Project: Adversarially Robust Training of JEPA-Based World Models

Harrison Huang, Rice University *2025 – Present*
Project: Interpreting Reinforcement Learning Policies through Explainable AI

Graduate Students

Janhavi Sathe, Rice University *March 2025 – May 2025*
Project: User Study on Machine Learning Application Audits

Mary Nam, Rice University *November 2024 – January 2025*
Project: Interpreting Saliency Maps using Multimodal Language Models

Isabel Schorr and Mira Trouvain, Technical University of Munich *January 2024 – June 2024*
Project: Simulating Human-Centered User Experience in XAI using LLMs

Thai Trang Nguyen, University of Tübingen *January 2023 – June 2023*
Project: Model Faithfulness and Preconceptions in Subjective Ratings of Explanations

Jacqueline Hirch, University of Tübingen *June 2022 – December 2022*
Project: Improving Interactive Medical Support System Performance with Knowledge Distillation

Naemi-Rebecca Kassautzki, University of Tübingen *January 2022 – June 2022*
Project: Driver Attention-Based Object Detection

David Scheerer, University of Tübingen *May 2021 – December 2021*
Project: Verbalizing Classification Decisions Based on Model Explanation

Undergraduate Students

Mohammed Abbas Ansari, India *March 2024 – July 2024*
Project: Semi-Supervised Learning Techniques for Scanpath Prediction

Carolin Niedermaier, Claudia Guadarrama Serrano, Letizia Wörrlein, Shaoming Zhang, Franka Exner, and Xufan Lu, Technical University of Munich *2024*
Project: Designing Human–AI Interaction for Speech-Based Educational Applications

Thai Trang Nguyen, University of Tübingen *May 2020 – December 2020*
Project: Human Attention in Fine-Grained Classification

ACADEMIC SERVICES

Organizing Committee:

- Co-Chair, Session on Equity in Distributed Digital Education, German-American Frontiers of Engineering Symposium, 2025
- Organizer, Workshop *GenEAI: Generative AI Meets Eye Tracking*, 2025

- Diversity & Accessibility Chair, ACM Symposium on Eye Tracking Research and Applications (ETRA), 2022 – 2025.

Program Chair:
ACM Symposium on Eye Tracking Research and Applications (ETRA), 2024 – 2025.

Student Advisory Service: Department of Computer Science, University of Tübingen, 2020 – 2022.

Program Committee Member/Reviewer:
Conferences: ICML, NeurIPS, ICLR, AISTATS, WACV, AAAI, ACM MM, CHI, HRI, etc.
Journals: TPAMI, TNNLS, IJHCI, ACM Computing Surveys, Information Systems Frontiers, etc.