浙江大学滨江研究院
BINJIANG INSTITUTE OF ZHEJIANG UNIVERSITY

Zhe Yu

Binjiang Institute of Zhejiang University · Research Intern
Communication University of Zhejiang · B.Eng. in Artificial Intelligence
Research Interest: Trustworthy LLMs / RAG Faithfulness

2th.l0ren17@gmail.com zyu@zju-if.com 235703223@stu.cuz.edu.cn
Phone/WeChat: +86 18257166408 (Yy7z_1) Homepage Google Scholar

Current Affiliation

Binjiang Institute of Zhejiang University Nov 2025 – Present
Research Intern; Advisors: Wenpeng Xing, Meng Han | Lab: IFRC-ZJU Hangzhou, China

- Participated in the project *Safety System Research and Applications for Multimodal Large Models* (Guangdong Provincial Key R&D Program), Nov 2025 – Present.
- Participated in the project *A New Trust Framework Based on Blockchain* (National Key R&D Program, Young Scientists Project), Nov 2025 – Present.

Education

Binjiang Institute of Zhejiang University Nov 2025 – Present
Research Intern (Advisor: Wenpeng Xing, Meng Han); Lab: IFRC-ZJU Hangzhou, China

University of Malaya Jan 2025 – Feb 2025
Visiting Student Kuala Lumpur, Malaysia

Westlake University Mar 2024 – Jul 2024
Visiting Student (Advisor: Ziyang Zhang) Hangzhou, China

Communication University of Zhejiang 2023 – Present (Expected 2027)
B.Eng. in Artificial Intelligence (Advisor: Hao Zeng) Hangzhou, China

Research Summary

- Research centers on trustworthy language models across three threads: (i) the internal mechanisms of knowledge grounding—how parametric memory and retrieved evidence interact during generation, and how to detect failures like hallucinations, memory hijacking, and compositional reasoning collapse through white-box monitoring and mechanistic analysis; (ii) representation–action dissociation in reasoning and agentic systems—probing when and why models encode conflict information internally yet fail to route it into downstream decisions; (iii) verifiable model ownership and decentralized trust—combining fingerprinting, blockchain, and zero-knowledge proofs to build scalable, privacy-preserving attribution and deployment frameworks.

Research Experience

Binjiang Institute of Zhejiang University Nov 2025 – Present
Research Intern (Advisors: Wenpeng Xing, Meng Han) Hangzhou, China

- Participated in the Guangdong Provincial Key R&D Program *Safety System Research and Applications for Multimodal Large Models*, with research spanning trustworthy LLMs, RAG faithfulness, white-box monitoring, agentic AI safety, and medical-LLM safety; representative outputs include DISF (received acceptance at ACL 2026), FIDES, LatentAudit, RETINA-SAFE / ECRT, five EMNLP ARR submissions (FIDES on token-level deep-evidence decoding for

retrieval-memory conflict, Cordon-MAS on knowledge-poisoning defense via information-flow control, Composition Collapse on compositional-reasoning failure after atomic-stability gating, Attribution Blind Spot on detecting parametric-memory hijacking in RAG, and Detecting Is Not Resolving on the monitoring–control gap in multi-turn RAG), Fingerprint Vector on scalable model-ownership transfer via vector addition, and two NeurIPS 2026 submissions on representation–action dissociation.

- Led problem formulation, method design, experimentation, and manuscript preparation for DISF, targeting RAG faithfulness hallucinations caused by conflict between parametric memory and retrieved evidence; proposed a white-box dual-path internal-state forcing framework that models Conflict, Drift, and Instability; received acceptance at ACL 2026.
- Led task definition, decoder design, experimental evaluation, and writing for FIDES, a training-free decoder that reads three internal signals probing retrieval-memory conflict at complementary depths—output surface, hidden representations, and prediction trajectory—and fuses them to govern intervention strength at each decoding step; achieves the best context fidelity in all 18 settings across three benchmarks and six backbones up to 70B, outperforming the strongest training-free baseline by +3 to +13 points.
- Led monitor formulation, method design, cross-model evaluation, and manuscript preparation for LatentAudit, a real-time white-box monitor based on residual-stream geometry and Mahalanobis distance; achieved 0.942 AUROC on PubMedQA with 0.77 ms overhead, generalized across five model families and four stress settings, and preserved 99.8% of FP16 AUROC under fixed-point quantization for Groth16-based public verification.
- Led benchmark construction, task design, experimental evaluation, and manuscript preparation for RETINA-SAFE / ECRT in medical-LLM safety, building a 12,522-sample evidence-grounded benchmark for diabetic retinopathy with E-Align, E-Conflict, and E-Gap tasks, and proposing a two-stage Evidence-Conditioned Risk Triage framework whose Stage-1 balanced accuracy exceeds external uncertainty and self-consistency baselines by 0.15–0.19.
- Led problem formulation, causal-analysis design, experiments, and manuscript preparation for a study on representation–action dissociation in indirect prompt injection; established a causal ladder (probes, activation patching, projection-out interventions) showing that Qwen-2.5-7B and Llama-3.1-8B agents linearly encode source role early in the residual stream but route it into tool-use decisions only at a late action-commitment band; revealed that source-role directions for controlled tool outputs, Slack, and persistent memory are channel-conditioned and nearly orthogonal; evaluated on AgentDojo-Slack trajectories; under review at NeurIPS 2026.
- Led paradigm design, causal analysis, and manuscript preparation for CoT-Swap, showing that when the `<think>` block targets a different question than the user prompt, reasoning-tuned LMs (7B–70B) default to the CoT-side answer in the majority of cases; used class-balanced linear probes to confirm that source-conflict information is linearly available yet not routed into the answer policy; localized the break to a causal bottleneck via single-layer activation patching and recovered the oracle effect through rank-k learned-projection steering; trace-training and consistency-training experiments identified training contributors and mitigations; under review at NeurIPS 2026.
- Participated in the National Key R&D Program (Young Scientists Project) *A New Trust Framework Based on Blockchain*, focusing on blockchain-, zero-knowledge-, and on-chain/off-chain trust mechanisms; related work includes ZK-FPE, ZK-VOT, and *Trusted Metadata-Coordinated Tiered Off-Chain Storage*.
- For ZK-FPE, contributed by refining experiments, producing figures, running result-generating evaluations, and supporting manuscript preparation; the work studies blockchain- and zero-knowledge-based model-fingerprint attribution for privacy-preserving ownership verification.
- Also participated in experiments and submission support for ZK-VOT and *Trusted Metadata-Coordinated Tiered Off-Chain Storage for Recovery-Safe and Low-Latency IoT Data Manage-*

ment, focusing on on-chain/off-chain mutual trust, zero-knowledge-verifiable transmission, tiered off-chain storage, and recovery-safe low-latency data management.

University of Malaya Jan 2025 – Feb 2025
Visiting Student Kuala Lumpur, Malaysia

- Completed structured study and academic exchange in an English-medium international setting, covering natural language processing, robotics, machine learning, and computer vision, with regular presentations and technical discussions.
- Implemented Python-based robotics exercises for locomotion, turning, and image capture, and participated in medical-data modeling exercises conducted in collaboration with the National University of Singapore and National University Hospital (NUH), Singapore.

Westlake University Mar 2024 – Jul 2024
Visiting Student, Optical Laboratory (Advisor: Ziyang Zhang) Hangzhou, China

- Developed a MATLAB-based quantitative image-analysis workflow for beam-deflector simulations in integrated optics, combining ROI selection, centerline suppression, grayscale smoothing, Otsu thresholding, Canny edge detection, and Hough-line fitting to recover the principal beam trajectory and estimate deflection angles from simulated field images.
- Performed batch angle extraction, mean-statistics analysis, and trend visualization across simulated image sets, identifying a deflection-angle range of approximately 5.46 to 7.92 degrees to support scan-pattern characterization, parameter calibration, and experimental workflow updates.

Selected Publications

1. **Zhe Yu***, Wenpeng Xing*, Wenjie Luo, Weize Xu, Lingtong Huang, Yourong Chen, Changting Lin, Meng Han†. *DISF: Detecting Hallucinations in Retrieval-Augmented Generation via Dual-path Internal State Forcing Framework*. ACL 2026; PDF.

2. Weiping Yu, Weihan Wang, Mingyuan Yan, Keyang He, **Zhe Yu**, Wenpeng Xing, Liyuan Liu, Meng Han†. *Trusted Metadata-Coordinated Tiered Off-Chain Storage for Recovery-Safe and Low-Latency IoT Data Management*. *Electronics* (MDPI).

3. F. Zhou, C. Chang, Q. Chang, H. Zhang, **Zhe Yu**, W. Liu, J. Li, J. Yang. *Orthogonal salinity and temperature detection via paralleled dual all-fiber interferometers*. *Optics Communications*, 583 (2025): 131688. [DOI]

4. **Zhe Yu**, H. Zeng, Y. Zhao, X. Zhang, Z. Wang, Y. Tao, M. Yuan, X. Sun. *Bibliometric analysis of physical education research in China from 2014 to 2024*. In *Proceedings of the 2024 7th International Conference on Educational Technology Management*. ACM, 2025, pp. 128–132. [DOI]

Manuscripts Under Review

1. **Zhe Yu**, Wenpeng Xing, Zhenhua Xu, Xingxing Yang, Meng Han†. *Knowing Is Not Acting: Representation–Action Dissociation in Indirect Prompt Injection*. Under review at NeurIPS 2026.

2. **Zhe Yu**, Wenpeng Xing, Zhenhua Xu, Ruiqi Zhang, Meng Han†. *Whose Thoughts? Chain-of-Thought Override in Reasoning-Tuned Language Models*. Under review at NeurIPS 2026.

3. **Zhe Yu***, Wenpeng Xing*, Meng Han†. *LatentAudit: Real-Time White-Box Faithfulness Monitoring for Retrieval-Augmented Generation with Verifiable Deployment*. Under review at CoLM 2026; arXiv:2604.05358.

4. **Zhe Yu**, Wenpeng Xing, Yunzhao Wei, Hongzhi Wang, Xuyang Teng, Meng Han†. *Composition Collapse: Stable Factual Knowledge Does Not Imply Compositional Reasoning*. Under review at ARR / EMNLP; [PDF] [arXiv].

5. **Zhe Yu**, Wenpeng Xing, Chen Ye, Xuyang Teng, Bo Yang, Changting Lin, Meng Han†. *Detecting Is Not Resolving: The Monitoring–Control Gap in Retrieval-Augmented LLMs*. Under review at ARR / EMNLP; [PDF] [arXiv].

6. **Zhe Yu**, Wenpeng Xing, Bo Yang, Chen Ye, Gaolei Li, Yunzhao Wei, Meng Han†. *The Attribution Blind Spot: Language Models Cannot Distinguish Reading from Remembering*. Under review at ARR / EMNLP; [PDF] [arXiv].

7. **Zhe Yu**, Wenpeng Xing, Gaolei Li, Shuguang Xiong, Hongzhi Wang, Xuyang Teng, Meng Han†. *Cordon-MAS: Defending RAG against Knowledge Poisoning via Information-Flow Control*. Under review at ARR / EMNLP; [PDF] [arXiv].

8. **Zhe Yu***, Wenpeng Xing*, Tiancheng Zhao, Mohan Li, Changting Lin, Meng Han†. *FIDES: Faithful Inference via Deep Evidence Signals for Retrieval-Memory Conflict in RAG*. Under review at ARR / EMNLP; [PDF].

9. Zhenhua Xu, Qichen Liu, Zhebo Wang, **Zhe Yu**, Xixiang Zhao, Wenpeng Xing, Dezhang Kong, Mohan Li, Meng Han. *Fingerprint Vector: Enabling Scalable and Efficient Model Fingerprint Transfer via Vector Addition*. Under review at ARR / EMNLP.

10. **Zhe Yu***, Wenpeng Xing*, Meng Han†. *From Retinal Evidence to Safe Decisions: RETINA-SAFE and ECRT for Hallucination Risk Triage in Medical LLMs*. Under review at MICCAI 2026; arXiv:2604.05348.

11. Zhiguo Ma*, Wenpeng Xing*, **Zhe Yu***, Yourong Chen, Meng Han†. *ZK-FPE: Blockchain-Verifiable Model Fingerprinting with Zero-Knowledge Privacy for Ownership Attribution*. Under review at *Blockchain: Research and Applications*.

12. *ZK-VOT: Establishing On-Chain/Off-Chain Mutual Trust via Zero-Knowledge Verifiable Oracle Transmission*. Submitted to WASA 2026.

* co-first authorship; † corresponding author.

Patents

1. Meng Han, **Zhe Yu**, Jiayan Hu, Wenpeng Xing, Changting Lin, Rongchang Li, Yourong Chen, Zhen Hong. *A Retrieval-Augmented Generation Hallucination Detection Method Based on Dual-Path Internal State Forcing*. Chinese Patent Application, App. No.: 2026104125719, filed Mar 31, 2026. (Under Review)

2. Meng Han, **Zhe Yu**, Jiayan Hu, Rongchang Li, Wenpeng Xing, Jingyi Yu, Zhen Hong, Bin Wang, Hongting Feng, Jing Xiong. *A Post-Processing Method, System, Device, and Medium for Hallucination Detection in Large Language Models Based on Adaptive Order Statistics Aggregation*. Chinese Patent Application, App. No.: 2026107898102, filed Jun 3, 2026. (Pending)

3. **Zhe Yu**, Jiayan Hu, Jingyi Yu, Weihang Yu, Wenpeng Xing, Jing Xiong, Yourong Chen, Zhen Hong, Changting Lin, Meng Han. *A Hallucination Detection Method, System, and Device for Large Language Models Based on Multi-Dimensional Heterogeneous Feature Fusion*. Chinese Patent Application, App. No.: 2026107899270, filed Jun 3, 2026. (Pending)

Industry Experience

BoostEngine Jul 2025 – Oct 2025
R&D Intern, Full-Stack Development / AI Agents Hangzhou, China

- Owned a multi-brand Manifest V3 Chrome extension for TikTok Shop operations using React, TypeScript, Vite, and Chrome APIs; shipped OAuth/SMS auth, request relays, content-script capture, and dashboard modules across 20 regional domains and 9 language packs.
- Owned a Spring Boot analytics backend with MyBatis-Plus, MySQL, and Redis for 8,000+ creator collaborations and several hundred thousand U.S. dollars in managed ad spend; delivered KPI APIs, creator-metrics aggregation, reporting, and AI outreach modules for cross-time-zone work with the New York team.
- Built a Feishu Bitable-MySQL synchronization system with FastAPI, SQLAlchemy, Redis, React, and Docker; implemented a 5-minute anti-loop window, queue coalescing, conflict resolution, 24-hour success-rate monitoring, and alerts for backlog above 1,000 or sync failure above 1%.

Awards & Honors

1. **Silver Award, CCB Cup Zhejiang Provincial International College Students' Innovation Competition (2024)**, for the project *"Scoliosis Detection AI System."*

Technical Skills

- **Research Stack:** Python, PyTorch, Hugging Face Transformers, scikit-learn, NumPy, pandas, FAISS, vLLM, Weights & Biases (W&B), MATLAB.
- **Engineering Stack:** TypeScript/JavaScript, Java, React, FastAPI, Spring Boot, SQLAlchemy, MyBatis-Plus, MySQL, Redis, Docker, Manifest V3 Chrome Extensions.
- **Themes:** Trustworthy LLMs, RAG faithfulness, hallucination detection, risk-guided decoding, medical LLM safety, and verifiable ownership attribution.

References

Dr. Meng Han, Intern Supervisor, Researcher, Zhejiang University, mhan@zju.edu.cn.
Dr. Wenpeng Xing, Intern Supervisor, Postdoctoral Researcher, Zhejiang University, wpxing@zju.edu.cn.
Dr. Ziyang Zhang, Westlake University, zhangziyang@westlake.edu.cn.
Dr. Hao Zeng, Communication University of Zhejiang, hao.zeng@cuz.edu.cn.