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

俞哲

浙江大学滨江研究院· 研究实习生
浙江传媒学院· 人工智能（工学学士在读）
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

当前任职

浙江大学滨江研究院 2025年11月– 至今
研究实习生；导师：邢文鹏、韩蒙| 实验室：IFRC-ZJU 杭州，中国

- 参与项目：多模态大模型安全系统研究与应用（广东省重点研发计划），2025年11月– 至今。
- 参与项目：基于区块链的新型信任体系（国家重点研发计划- 青年科学家项目），2025年11月– 至今。

教育背景

浙江大学滨江研究院 2025年11月– 至今
研究实习生（导师：邢文鹏、韩蒙）；实验室：IFRC-ZJU 杭州，中国

马来亚大学 2025年1月– 2025年2月
访问学生 吉隆坡，马来西亚

西湖大学 2024年3月– 2024年7月
访问学生（导师：张紫阳） 杭州，中国

浙江传媒学院 2023年– 至今（预计2027 年毕业）
人工智能专业，工学学士在读（导师：曾昊） 杭州，中国

研究概述

- 研究聚焦于可信大语言模型，围绕三条主线展开：（一）知识grounding 的内部机制——参数记忆与检索证据在生成过程中的交互，以及如何通过白盒监测与机制分析检测幻觉、记忆劫持与组合推理崩溃等失败模式；（二）推理与智能体系统中的表征–行动分离——探究模型何时在内部编码冲突信息却未能将其路由至下游决策；（三）可验证模型归属与去中心化信任——结合指纹、区块链与零知识证明构建可扩展的隐私保护归属与部署框架。

科研经历

浙江大学滨江研究院 2025年11月– 至今
研究实习生（导师：邢文鹏、韩蒙） 杭州，中国

- 参与广东省重点研发计划"多模态大模型安全系统研究与应用"，围绕Trustworthy LLMs、RAG 忠实性、白盒监测、智能体安全与医疗大模型安全开展研究，相关产出包括DISF、FIDES、LatentAudit、RETINA-SAFE / ECRT，五篇EMNLP ARR 在投稿件（FIDES 基于深层证据信号的检索-记忆冲突解码、Cordon-MAS 面向RAG 知识投毒的多智能体信息流控制防御、Composition Collapse 基于原子稳定性筛选的组合推理失败分析、Attribution Blind Spot 检测RAG 中参数记忆对检索上下文的劫持效应、Detecting Is Not Resolving 揭示多轮RAG 中的监控–控制缺口），以及Fingerprint Vector 面向向量加法的大模型指纹可扩展迁移，两篇NeurIPS 2026 在投稿件（表征–行动分离方向）。

- 主导DISF 的问题定义、方法设计、实验组织与论文撰写，面向parametric memory 与retrieved evidence 冲突导致的RAG 忠实性幻觉，提出dual-path internal-state forcing 白盒框架，显式刻画Conflict、Drift 与Instability 三类内部信号；ACL Findings 2026 已录用。
- 主导FIDES 的任务定义、解码器设计、实验评估与论文撰写，面向retrieval-memory conflict 提出基于深层证据信号的training-free 解码器，通过输出表面、隐藏表示与预测轨迹三个互补深度的内部信号探测冲突，并在每个解码步骤自适应调节干预强度；在三个基准与六个骨干网络（含70B 规模）的全部18 个设置中取得最优上下文忠实性，相比最强training-free 基线提升+3至+13 点。
- 主导LatentAudit 的监测任务设定、方法设计、跨模型实验与论文撰写，提出基于residual-stream geometry 与Mahalanobis distance 的real-time white-box monitor，在PubMedQA 上达到0.942 AUROC、额外开销0.77 ms，并验证fixed-point audit rule 可支撑Groth16 public verification。
- 主导RETINA-SAFE / ECRT 的benchmark 构建、task design、实验评估与论文撰写，围绕diabetic retinopathy 高风险决策建立12,522 样本的evidence-grounded benchmark，定义E-Align、E-Conflict、E-Gap 三类任务，并提出two-stage Evidence-Conditioned Risk Triage；Stage-1 balanced accuracy 相比外部uncertainty / self-consistency 基线提升0.15–0.19。
- 主导间接提示注入中表征–行动分离研究的问题定义、因果分析框架设计、实验组织与论文撰写，通过causal ladder（probes、activation patching、projection-out interventions）揭示Qwen-2.5-7B 与Llama-3.1-8B 在residual stream 早期已线性编码source role，但仅在late action-commitment band 才将其路由至工具调用决策；发现不同tool channel 的source-role 方向几乎正交且不可跨通道迁移；在AgentDojo-Slack 轨迹上评估；投稿至NeurIPS 2026。
- 主导CoT-Swap 范式设计、因果分析与论文撰写，发现当`<think>` 块包含不同于用户提问的Chain-of-Thought 时，7B–70B 推理微调模型在大多数情况下回答CoT 侧问题而非用户问题；通过class-balanced linear probe 确认source-conflict 信息线性可用但未被路由至answer policy；利用single-layer activation patching 将断裂定位到因果瓶颈，并通过rank-k learned-projection steering 恢复oracle 效果；trace-training 与consistency-training 实验揭示了训练侧诱因与缓解方向；投稿至NeurIPS 2026。
- 参与国家重点研发计划- 青年科学家项目“基于区块链的新型信任体系”，围绕区块链、零知识证明与链上/链下协同信任开展研究，相关工作包括ZK-FPE、ZK-VOT 与*Trusted Metadata-Coordinated Tiered Off-Chain Storage*。
- 在ZK-FPE 中主要负责实验改进与完善、图表绘制、结果执行与辅助撰写，围绕零知识证明与区块链模型指纹归属验证完成proof generation、verification cost 与链上开销实验，支撑隐私保护条件下的所有权验证结论。
- 参与ZK-VOT 与*Trusted Metadata-Coordinated Tiered Off-Chain Storage for Recovery-Safe and Low-Latency IoT Data Management* 的相关实验与投稿工作，聚焦链上/链下互信、零知识可验证传输、分层链下存储与恢复安全场景下的低时延数据管理。

马来亚大学 2025年1月– 2025年2月

访问学生 吉隆坡，马来西亚

- 在英文授课与国际化交流环境中系统学习人工智能相关内容，训练主题覆盖自然语言处理、机器人技术、机器学习与计算机视觉，并完成跨文化学术展示与讨论。
- 基于Python 完成机器人行走、转向与拍照控制实践；同时参与与新加坡国立大学（National University of Singapore）及新加坡国立大学医院（National University Hospital, NUH）合作的医疗数据分析建模练习。

西湖大学 2024年3月– 2024年7月

光学实验室访问学生（导师：张紫阳） 杭州，中国

- 围绕集成光学中光束偏转器角度仿真的定量分析问题，基于MATLAB 设计图像特征提取流程，通过ROI 裁剪、中轴线干扰抑制、灰度平滑、Otsu 阈值分割、Canny 边缘检测与霍夫直线拟合，对仿真光场中的主传播轨迹进行自动识别并反演偏转角。
- 对多组仿真图像进行批量测角、均值统计与趋势可视化，得到偏转角约5.46 至7.92 度的变化区间，用于分析扫描规律、辅助参数标定与实验流程更新。

已录用论文

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 Findings 2026 已录用; OpenReview.

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) 已录用.

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]

代表性在投稿件

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

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

3. **Zhe Yu***, Wenpeng Xing*, Meng Han†. *LatentAudit: Real-Time White-Box Faithfulness Monitoring for Retrieval-Augmented Generation with Verifiable Deployment*. CoLM 2026 在投; 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*. ARR / EMNLP 在投; [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*. ARR / EMNLP 在投; [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*. ARR / EMNLP 在投; [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*. ARR / EMNLP 在投; [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*. ARR / EMNLP 在投; [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*. ARR / EMNLP 在投.

10. **Zhe Yu***, Wenpeng Xing*, Meng Han†. *From Retinal Evidence to Safe Decisions: RETINA-SAFE and ECRT for Hallucination Risk Triage in Medical LLMs.* MICCAI 2026 在投；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. Blockchain: Research and Applications* 在投.
12. *ZK-VOT: Establishing On-Chain/Off-Chain Mutual Trust via Zero-Knowledge Verifiable Oracle Transmission.* 投稿至WASA 2026.

* 表示共同一作；† 表示通讯作者。

专利

1. 韩蒙, **俞哲**, 胡佳妍, 邢文鹏, 林昶廷, 李荣昌, 陈友荣, 洪榛. 一种基于双路径内部状态强迫的检索增强生成幻觉检测方法. 中国发明专利, 申请号: 2026104125719, 申请日: 2026-03-31. (实质审查中)

企业研发经历

BoostEngine 2025年7月– 2025年10月
研发实习生（全栈开发/ AI Agents） 杭州，中国（与纽约团队远程协作）

- 面向中国跨境品牌与美国本土品牌共同使用的TikTok Shop 增长业务，负责多品牌Manifest V3 Chrome Extension（BoostEngine / ProBoost / Kaloboost）研发，基于React、TypeScript、Vite 与Chrome APIs 实现OAuth/验证码登录、in-page request relay、content script 数据采集与dashboard 模块，覆盖20 个区域站点与9 种语言环境。
- 围绕团队累计合作8000+ 达人、累计操盘数十万美元投放的业务场景，负责Spring Boot + MyBatis-Plus + MySQL/Redis 数据分析后台建设，开发达人、商品、店铺、视频等核心KPI 接口、TikTok Creator 指标聚合、报表与AI 邀约相关模块，支撑杭州研发团队与纽约团队跨时区协同的跨境增长决策。
- 独立设计并实现FastAPI + SQLAlchemy + Redis + React + Docker 的飞书多维表格/MySQL 双向同步系统，构建5 分钟哈希防环、队列合并、冲突处理、24 小时成功率监控，以及队列积压1000 条和失败率1% 的告警阈值，提升跨团队数据同步的稳定性与可运维性。

荣誉奖项

1. **“建行杯”浙江省国际大学生创新大赛（2024）银奖**，项目《脊柱侧弯检测AI 系统》项目骨干，融合计算机视觉与机器学习技术，实现脊柱侧弯风险评估与辅助诊断。

技术能力

- **研究栈**： Python, PyTorch, Hugging Face Transformers, scikit-learn, NumPy, pandas, FAISS, vLLM, Weights & Biases (W&B), MATLAB.
- **工程栈**： TypeScript/JavaScript, Java, React, FastAPI, Spring Boot, SQLAlchemy, MyBatis-Plus, MySQL, Redis, Docker, Chrome Extension（Manifest V3）.
- **研究主题**： Trustworthy LLMs, RAG faithfulness, hallucination detection, risk-guided decoding, medical LLM safety, verifiable attribution.

推荐人

韩蒙，研究实习导师，浙江大学研究员，mhan@zju.edu.cn.
邢文鹏，研究实习导师，浙江大学博士后研究员，wpxing@zju.edu.cn.
张紫阳，西湖大学导师，zhangziyang@westlake.edu.cn.
曾昊，浙江传媒学院导师，hao.zeng@cuz.edu.cn.