胡雨林



邮箱： ylhu@ir.hit.edu.cn　　**电话：** +86 15948553644
个人主页： https://yulinlp.github.io/　　**谷歌学术主页：** 链接
意向研究方向： Agent Memory · Personalized Agents · Agentic Post-Training
预计到岗时间： 2027 年 2 月　　**可实习时长：** 一年

教育背景

哈尔滨工业大学 ***2021.09—2025.06***
未来技术学院
计算机科学与技术 学士

哈尔滨工业大学 ***2025.09—至今***
计算学部 **社会计算与交互机器人研究中心（SCIR）**
计算机科学与技术 学术型硕士（已通过本校硕转博，2027年春季入读博士）
导师：赵妍妍教授　副导师：赵伟翔教授

科研成果

我的研究主要围绕智能体记忆、个性化交互与大模型安全对齐展开。已在 NeurIPS、ACL、AAAI 等会议及期刊发表或录用论文 **15 篇**，其中独立一作 **1 篇**、学生一作 **4 篇**、共同一作 **2 篇**，Google Scholar 总引用 **496 次**。另有 3 篇独立一作论文在投，分别研究多模态长期记忆评测、记忆检索与记忆使用。

智能体记忆、个性化

1. **Yulin Hu**, Zimo Long, Jiahe Guo, Xingyu Sui, Xing Fu, Weixiang Zhao, Yanyan Zhao, Bing Qin. OP-Bench: Benchmarking Over-Personalization for Memory-Augmented Personalized Conversational Agents. **EMNLP 2026**，独立一作。Memory Over-Personalization Benchmark

2. **Yulin Hu**, Yanyan Zhao, Zimo Long, Xing Fu, Xingyu Sui, Jiahe Guo, Weixiang Zhao, Bing Qin. SC-Embedding: Selective Contextualization for Long-Context Retrieval. **WWW 2027在投**，独立一作。Memory Retrieval Embedding Model

3. **Yulin Hu**, Yanyan Zhao, Zimo Long, Xing Fu, Mengtong Ji, Weixiang Zhao, Yutai Hou, Qianchao Wang, Dandan Tu. CUE-Mem: Benchmarking Long-Term User Memory via Implicit Cues in Multimodal Conversations. **AAAI 2027在投**，独立一作。Memory Benchmark Omni

4. **Yulin Hu**, Yanyan Zhao, Zimo Long, Jiahe Guo, Xingyu Sui, Xing Fu, Yutai Hou, Qianchao Wang, Xiao Zhao, Dandan Tu, Weixiang Zhao, Bing Qin. How Do Language Models Use Memory? From Internal Strategy Readout to Adaptive Control. **ICLR 2027在投**，独立一作。Memory Steering

5. Weixiang Zhao*, Xingyu Sui*, **Yulin Hu***, Jiahe Guo, Haixiao Liu, Biye Li, Yanyan Zhao, Bing Qin, Ting Liu. Teaching Language Models to Evolve with Users: Dynamic Profile Modeling for Personalized Alignment. **NeurIPS 2025**，共同一作。Personalization RL

6. Weixiang Zhao*, Xingyu Sui*, Jiahe Guo*, **Yulin Hu***, Yang Deng, Yanyan Zhao, Xuda Zhi, Yongbo Huang, Hao He, Wanxiang Che, Ting Liu, Bing Qin. Trade-offs in Large Reasoning Models: An Empirical Analysis of Deliberative and Adaptive Reasoning over Foundational Capabilities. **AAAI 2026**，共同一作。Reasoning Benchmark

安全对齐

7. Weixiang Zhao, **Yulin Hu**, Yang Deng, Tongtong Wu, Wenxuan Zhang, Jiahe Guo, An Zhang, Yanyan Zhao, Bing Qin, Tat-Seng Chua, Ting Liu. MPO: Multilingual Safety Alignment via Reward Gap Optimization. **ACL 2025, Oral & Panel (Top 0.3%)**, 学生一作。Multilingual Safety Alignment

8. Weixiang Zhao*, **Yulin Hu***, Yang Deng, Jiahe Guo, Xingyu Sui, Xinyang Han, An Zhang, Yanyan Zhao, Bing Qin, Tat-Seng Chua, Ting Liu. Beware of Your Po! Measuring and Mitigating AI Safety Risks in Role-Play Fine-Tuning of LLMs. **ACL 2025**, 学生一作。Role-playing Safety Tuning

9. Weixiang Zhao, **Yulin Hu**, Zhuojun Li, Yang Deng, Yanyan Zhao, Bing Qin, Tat-Seng Chua, Ting Liu. Towards Comprehensive and Efficient Post Safety Alignment of Large Language Models via Safety Patching. **IJCAI 2026**, 学生一作。Model Merging Safety Alignment

10. Weixiang Zhao, **Yulin Hu**, Xingyu Sui, Zhuojun Li, Yang Deng, Yanyan Zhao, Bing Qin, Wanxiang Che. The Gains Do Not Make Up for the Losses: A Comprehensive Evaluation for Safety Alignment of Large Language Models via Machine Unlearning. **Frontiers of Computer Science 2026**, 学生一作。Unlearning Benchmark

实习经历

华为终端 BG　研究型实习 ***2026.06—至今***

度小满金融科技　研究型实习 ***2024.12—2025.03***

项目经历

纵向课题:

1. **2026.01—至今**　核心骨干，国家自然科学基金项目，**人格具象化与安全防护双重建模的个性化共情对话技术研究**。
2. **2024.09—2025.12**　核心骨干，国家自然科学基金项目，**多模态细粒度情感计算关键技术研究**。

横向课题:

3. **2025.07—2026.06**　核心骨干，赛力斯企业合作项目，**情感计算项目**。

落地成果

巧板共情对话大模型　核心开发者

参与研发国内首个面向 K12 群体的共情对话大模型，负责个性化与记忆系统开发，以及 K12 共情对话模型训练。相关成果应用于科大讯飞、华为小艺和赛力斯。

哈工大 AI 辅导员“工小星”　核心负责人、核心开发者

负责面向大学生的情感对话模型训练与整体 RAG 框架、记忆系统建设，为哈工大一校三区师生提供校园咨询与个性化陪伴服务，用户规模达 25,402 人，累计完成问答 148,006 次。

获得荣誉

- 研究生国家奖学金（2026）
- 哈尔滨工业大学校级优秀学生标兵（2026）
- 哈尔滨工业大学特等学业奖学金（2025、2026）
- 哈尔滨工业大学优秀本科毕业论文（2025）