王可逸

wky32768@qq.com | (+86) 138 6765 1901

教育背景

复旦大学 2025.09 – 2028.06（预计）
计算机科学与技术，理学硕士在读 上海
– **研究方向**：大语言模型智能体（LLM Agent）、推理与能力评测、可控指令遵循。

杭州电子科技大学 2021.09 – 2025.07
计算机科学与技术，工学学士（荣誉） 杭州
– 卓越学院计算机科学英才班；GPA 4.25/5.00（87.5/100），专业排名前 2.7%。
– 媒体智能实验室；导师：余宙教授、黄孝喜副教授。

瑞典克里斯蒂安斯塔德大学 2023.08 – 2024.01
学校公派一学期交换项目 瑞典克里斯蒂安斯塔德

论文成果

[1] **SOPAdapt: Benchmarking LLM Agent Repair Under Procedural Blockages**
第一作者 *ACL ARR 在投*
提出基于需求标注 DAG 的流程修复评测基准，构建覆盖 **6 类阻塞、175 个实例**的数据集，设计 CRR 指标衡量偏离后对原流程的复用程度。评测 **7 种平台–模型配置、3,675 条轨迹**，发现相近任务成功率可对应显著不同的流程复用水平；自动评分与人工评分相关系数达 **0.920**。

[2] **Knowing but Not Doing: Discourse-level Constraints Adherence of Large Language Models**
第三作者 *Findings of EMNLP 2026* (CCF-B)
研究显式与隐式篇章级指令遵循，基于修辞结构理论设计 RSTsim 结构相似度指标，揭示约束识别与生成遵循之间的差距；通过引入结构奖励的 GRPO 训练，验证篇章级优化对约束遵循及隐式任务迁移的作用。

[3] **Immediate Inference: The Missing Foundation in Large Language Model Logical Reasoning**
第三作者 *ACL 2026, Main Conference (Oral)* (CCF-A)
构建包含 **5,284** 个样本的 IIBench，覆盖命题变换与对当关系，并以事实、反事实和匿名命题区分语义干扰。通过跨任务相关性与因果中介分析，研究基础即时推理对下游逻辑推理的影响，定位量词与否定算子的语义落地缺陷。

[4] **InsLogicBench: An Argumentation Logic Grounded Benchmark for Complex Insurance Claims Adjudication**
第三作者 *ACL 2026, Main Conference* (CCF-A)
基于嵌套 Toulmin 论证模型构建可控数据合成框架，通过分层真值分配与一致性校验生成可验证推理链；构建含 **110 份保单、2,621 个问答样本**的 InsLogicBench。对 8 个模型的评测揭示：正确判决不等于有效推理，例外条款处理与缺失条件验证仍是关键难点。

[5] **Measuring the Unmeasurable: Unveiling Latent Cognitive Capabilities of LLM**
第三作者 *AAAI 2026, Main Technical Track (Poster)* (CCF-A)
基于 ACT-R 理论建立涵盖 **5 类认知能力、16 项认知操作**的分层体系，提出 CogProbe 评测框架并构建中文、英文和西班牙语 CogEval 数据集（每种语言 **10,110** 个样本），从认知操作层面刻画模型能力与跨语言差异。

[6] **Is One Refusal Enough? Evaluating Robust Abstention in Multimodal Language Models**
第三作者 *ACL ARR 在投*
构建 EBI-Bench，包含 **1,107 组评测单元**，覆盖 4 类不可回答的视觉问题；通过同图问题改写与可回答对照，联合评估拒答一致性及解释质量。6 个多模态模型的完整单元成功率仅为 2.1%–33.3%；采用带具体拒答解释的 LoRA 微调后，3 个模型该指标平均提升 **39.2 个百分点**。

[7] **Reasoning-22: When Large Language Models Know the Rules but Miss the Route**
第四作者 *ACL ARR 在投*
参与构建经求解器验证的多规则可达性基准，包含三种语言表述下的 **1,080 对样本**，以单条规则修改翻转可达性，并用 PRA 指标检验成对判断。7 个模型的平均单题准确率为 61.1%，PRA 仅 23.1%；分阶段干预表明，在诊断条件下，闭源模型的主要瓶颈是从语言构建状态转移系统。

[8] **Readable but Underweighted: Logic and Belief in Language Model Decisions**
第四作者 *ACL ARR 在投*
研究逻辑有效性与常识信念冲突时模型的判断机制：将有效性方向分解为信念子空间内的混合分量 M 与正交残差 R，区分“信息可读”与“影响决策”。在 4 个系列 **7 个模型**上，冲突样本中 R 的有效性 AUC 均高于 M，但等范数编辑经 M 的平均效应更大；联合抑制 M、放大 R 可修复冲突错误，证据强度与敏感度的乘积在留出内容块上预测干预效果的 R^2 达**.504–.909**。

科研与项目经历

GenericAgent：通用自演化 LLM Agent 2026.03 – 2026.08
复旦大学知识工场实验室；技术报告共同作者 上海

- **项目简介**：GenericAgent 是一个通用、可自演化的 LLM Agent 框架，通过分层记忆、上下文压缩与经验（SOP）沉淀复用，使 Agent 能在多步任务执行中持续积累并复用经验。
- **个人工作**：负责 SOP-Bench 上的评测实验，设计对比方案并完成 GenericAgent 与 OpenClaw、Claude Code 等系统在多步流程任务上的对比评测。

荣誉奖项

- 美国大学生数学建模竞赛（MCM/ICM），**国际一等奖（Meritorious Winner）**，2023。
- 华为软件精英挑战赛，**杭厦赛区第 17 名（二等奖）**，2022。
- CCF CSP 计算机软件能力认证，**330/500（前 2.43%）**，2024。

专业技能

- **编程语言**：Python、C++；面向研究任务的数据处理、评测脚本与实验结果分析。
- **大模型评测**：Benchmark 构建、可控数据合成、成对 / 反事实样本设计、LLM-as-a-Judge、错误归因与多模型对比。
- **Agent**：SOP / DAG 流程建模、工具调用与执行轨迹分析，任务成功率、流程复用及 Token 效率评测；了解分层记忆、上下文压缩与经验复用机制。
- **模型训练与推理**：使用 LLaMA-Factory 进行 SFT 与 LoRA 微调；使用 vLLM 进行模型部署、批量推理及强化学习训练中的 rollout 采样。
- **语言能力**：英语 IELTS 7.0；中文（母语，普通话二甲）；法语（A1，大学法语四级通过）。