

教育背景

吉林大学, 软件工程, 硕士 (推免)

2024.09 – 2027.09

吉林大学, 软件工程, 学士

2020.09 – 2024.09

实习经历

 **Trust & Application AI Lab (TAI Lab)**, 香港科技大学 (广州), 科研助理 2024.09 – 至今

TAI Lab 由戴恩炎教授主导, 围绕新一代可信人工智能展开, 涵盖大语言模型 (LLM) 可信边界、安全机制与复杂系统中的风险治理。实习期间主要研究 LLM 服务新型威胁, 在传统后门攻击、对抗样本与安全对齐研究的基础上, 探索从训练到部署的各层级安全, 包括模型知识产权保护 (EACL 2026) 与 LLM 路由系统安全 (ACL 2026)。

科研经历

已发表论文

[1] **CCF-A ACL 2026 Main R²A: Directing LLM Routers to Expensive Models via Adversarial Suffix Optimization**
(项目负责人 & 第一作者) [Arxiv] 2025.09-2026.01

- **研究动机:** 大语言模型 (LLM) 路由通过动态将请求分配至不同能力与成本的模型以降低推理开销, 但也引入了新的攻击面; 现有路由攻击依赖模型权重、梯度等白盒信息, 难以适配真实商业黑盒系统。
- **核心方法:** 提出混合集成代理路由器 (Hybrid Ensemble Surrogate Router) 与对抗后缀优化 (Adversarial Suffix Optimization) 串联的黑盒攻击框架: 仅凭少量 (用户输入, 路由结果) 反馈构建代理模型, 近似目标路由器的决策边界; 再通过归一化梯度聚合在代理模型上优化对抗后缀, 稳定诱导请求转向高成本模型。
- **实验验证:** 在多个开源路由器与多数据集上验证泛化性; 首次在 GPT-Auto 与 OpenRouter 等真实商业黑盒系统上成功触发高成本模型调用, 在 OpenRouter 上造成实际推理成本 $2.7\times-2.9\times$ 的上升。

[2] **NLP 顶会 EACL 2026 Findings DuFFin: A Dual-Level Fingerprinting Framework for LLMs IP Protection**
(核心参与 & 第二作者) [Arxiv] 2024.09-2025.05

- **方法概述:** 面向大语言模型知识产权保护, 提出黑盒指纹框架 DuFFin, 从触发响应与知识边界两个层面验证模型所有权; 无需修改模型权重, 可在微调、量化和安全对齐等派生场景下识别疑似盗用模型。
- **主要贡献:** 完成从知识边界提取模型所有权模块, 构建多领域知识问题作为私有密钥, 利用模型答案序列形成知识层指纹; 负责模块实现、领域敏感性分析与鲁棒性评估。

[3] **CCF-A KDD 2026 General Protein Pretraining or Domain-Specific Designs? Benchmarking Protein Modeling on Realistic Applications** [Arxiv] 2025.03-2025.07

- **主要贡献:** 构建面向真实蛋白质应用的系统性基准; 负责部分基线的实现与评测接入和模型适用性分析。

在审论文

[4] **D-Miner: Extracting Post-Training Data from Diffusion Language Models**
(项目负责人 & 第一作者) 2026.02-2026.06

- **研究动机:** 扩散语言模型 (DLM) 依赖双向去噪从纯 [MASK] 输入重建文本, 与自回归模型左到右续写不同。该机制使后训练数据信号在无任何前缀条件下即可被触发, 形成现有提取方法未覆盖的新型隐私面。
- **核心方法:** 提出 D-Miner, 仅以 [MASK] 序列为输入, 无需目标前缀或辅助文本。以掩码自重建打分从模型自身生成的候选池中识别类成员候选; 将候选对比信号转化为隐状态引导方向; 在迭代去噪中施加方向引导并抑制重复模式, 定向引导生成向后训练数据区域收敛。
- **实验结论:** 覆盖监督微调 (SFT) 与强化学习 (RL) 共 8 个 DLM 目标, 在覆盖度与生成质量上优于无引导基线; 首次从掩码输入中恢复结构化思维链 (CoT) 数据, 揭示 SFT 与 RL 后训练策略在 DLM 中呈现不同泄露特征。

项目经历

- 参与 **StudyClawHub** 开发与维护, 构建面向学生的 AI Agent, Skill 生态入口: 支持用户浏览可复用 Agent、安装任务型 Skill、提交自定义 Skill, 并通过 GitHub 驱动的轻量化注册机制完成社区发布与维护。
- 围绕“中西医多模态大模型”方向, 参与**国家重大科技专项课题**申报, 承担申报书核心撰写与技术路线凝练, 梳理多模态数据融合、重症预警建模与临床验证方案, 相关课题已获批立项。
- 参与学术审稿工作, 曾以审稿人或子审稿人身份服务于 WWW 2025、WSDM 2025、CIKM 2025 和 KDD 2026。