段东杭

Donghang Duan

188 8132 2478
donghangduan@gmail.com
Blog:https://sowingg.space
SowingG2333
WeChat: Ddhang222 QQ: 3243856409

教育背景

23.09–27.07 **本科**, 电子科技大学, 英才实验学院 / 计算机科学与工程学院
计算机科学与技术（“成电英才计划”拔尖创新人才实验班）
智育成绩：91.49/100（排名 18/164, 前 11%） **GPA**：3.99/4.00

研究兴趣

- 大语言模型安全与对齐（LLM Safety & Alignment）
- 智能体安全（Agent Security）
- 可信机器学习（Trustworthy Machine Learning）

发表论文

[ACL 2026] **Donghang Duan**, Xu Zheng, Yuefeng He, Chong Mu, Leyi Cai, Lizong Zhang. “Look Twice before You Leap: A Rational Framework for Localized Adversarial Text Anonymization.” arXiv:2512.06713, 2025.

[IPCCC 2025] **Donghang Duan**, Xu Zheng, Yifu Zheng, Chong Mu, Ruozhou Wang, Ke Yan. “Exploiting Parasitic Dependency for Free-Rider Elimination in Blockchain-based Federated Learning.” In Proceedings of the IEEE International Performance Computing and Communications Conference, 2025.

项目经历

25.09–26.01 **基于理性代理的本地化文本匿名框架（RLAA）** 科研项目（第一作者）
LLM Safety / Natural Language Processing
- 围绕现有基于 LLM 的匿名化方法中存在的隐私悖论，以及本地朴素迁移后易出现效用崩塌的问题，设计并实现了无需训练的本地匿名化框架 RLAA；该框架构建了 Attacker-Arbitrator-Anonymizer 三方机制，并引入基于经济理性分析的仲裁与早停策略，成功抑制了由幻觉推断和边际收益递减带来的过度编辑。
- 本人负责整体方案设计、核心算法实现、实验评估与结果分析，并以第一作者完成论文撰写；该方法在多个主流模型上取得了优于强基线的隐私–效用权衡，在人工评测中相对现有方法的胜率达到 88.4%，相关成果已被 **ACL 2026 Findings** 接收。

25.03–25.07 **区块链联邦学习中搭便车攻击的防御机制（BFLGR）** 科研项目（第一作者）
Federated Learning / Blockchain
- 针对区块链联邦学习中高级搭便车者可利用链上历史更新伪造梯度、从而绕过传统异常检测的问题，设计了防御框架 BFLGR；该方法抓住攻击者对诚实节点更新的寄生依赖及其逐利动机，通过逆向拍卖与动态信誉机制，使欺骗性贡献在系统中变得不可持续。
- 本人负责算法设计、系统实现、实验评估与论文撰写；实验表明，该方法在高级搭便车攻击场景下可有效清除恶意节点并保持诚实节点收益，相关成果已发表于 CCF-C 类会议 **IEEE IPCCC 2025**。

所获表彰

- **电子科技大学 · 优秀学生奖学金**：2023–2024 学年
- **电子科技大学 · 专项奖学金**：2024–2025 学年

专业技能

- **英语能力**：CET-4（617 分），CET-6（607 分）
- **程序语言**：Python, C/C++, JavaScript, HTML/CSS, Verilog
- **机器学习**：PyTorch, Hugging Face Transformers
- **开发工具**：Linux, Docker, Kubernetes, Git, Vim