刘傲凡

mailto:af.liu@stu.pku.edu.cn

电话：15799122068

四个月|立即到岗

GitHub 主页: github.com/Fab-Liu

教育背景

厦门大学（985） **软件工程** **工学学士** 2019.09-2023.06

- 综合排名：Top3%；加权成绩：94.4/100；雅思7.0

北京大学（985） **计算机应用技术(保研)** **工学硕士** 2023.09 - 至今

- 保研至北京大学
- 研究兴趣：大模型；多模态；大模型安全

工作经历

北京智源科技有限公司 2024.01-2024.06

北京 | 多模态组 **算法实习|**

- 参与大规模多模态数据集半自动标注技术研究
- 研究视觉对抗性示例绕过对齐LLMs的安全防护
- 参与优化少数有害语料库上的条件生成概率以及多模态模型安全研究

MiniMax 2023.08-2024.12

上海 | 对齐组 **算法实习|**

- 通过数据清洗，提高了数据集的整体质量，减少了模型训练过程中的噪声干扰
- 根据项目需求，对图像、文本和音频数据进行精确的标注，确保数据的准确性和一致性
- 优化了模型SFT、DPO训练流程，参与了模型的迭代改进，使模型在特定任务上的准确率提高了5%

北京大学 V2X 国家重点实验室 2023.01-2023.10

广东省 深圳市 | 国家重点实验室 **研究助理|**

- 使用 **AFLFuzz** 及 **LibFuzzer** 独立构建超 60+ CVE 的 Fuzz 数据库的构建
- 参与设计基于多头注意力机制导向定向 Fuzz 工具并完成 **AccuracyFuzz** 的部分实验部分

厦门大学 计算与数据科学实验室 2021.09-2022.01

福建省 厦门市 | 计算与数据科学实验室 **研究助理|**

- 参与**模型水印**，**图像隐写**的研究

科研经历

视觉模态输入对多模态大型语言模型（MLLMs）安全性影响研究 2024.04-2024.06

PiCo: Jailbreaking Multimodal Large Language Models via Pictorial Text and Code Instruction **第二作者|**

- 评估视觉输入对LLMs安全防护的弱点，探索视觉对抗性示例对LLMs的“越狱”能力。
- 尝试以视觉模态输出经典 Jailbreak Prompt DAN系列命令，并取得一定成果
- 通过实验设置评估攻击对不同VLMs（如MiniGPT-4, InstructBLIP, LLaVA）的效果
- 实施人工和自动化评估，以确定对抗性示例对模型输出的影响
- 对比视觉和文本攻击的优化损失和“越狱”效果，测试DiffPure等现有防御技术对抗视觉对抗性示例的能力

代码场景下多模态大模型安全基准分析 2024.01-2024.04

PiCo: Jailbreaking Multimodal Large Language Models via Pictorial Text and Code Instruction **第一作者|**

- 研究了越狱对齐LLMs的方法，包括提示注入、对抗性攻击、越狱和数据投毒等
- 类比 F1-Score, 提出了 **Toxicity and Helpfulness Evaluator**，用于基准化评测多模态大模型
- 专注于MLLMs的跨模态攻击，尤其是针对 Gemini-Pro 和 GPT-4 等高级MLLMs的安全性弱点
- PiCo在多个熟练的MLLMs上成功绕过了模型安全防护，对于Gemini Pro Vision的平均攻击成功率（ASR）为56.27%，对于GPT-4V为32.27%

大规模多模态数据集半自动标注技术研究 2024.02-2024.04

Research on Semi-Automatic Annotation Technology for Large-Scale Multi-Modal Datasets

- 参与构建可提示的视觉基础模型，采用一个模型即可分隔、识别、描述图像中的任意目标
- 参考 SAM 架构，基于混合监督大模型，构建人在回路的**协同标注**框架；
- 基于MSCoCo，CityScape，Mapillary 数据集构建半自动-交互标注引擎
- 标注效率提升1~2个数量级，构建了**50万张**高质量多模态数据集

AccuracyFuzz: 基于 CodeBert 的定向模糊测试工具 2023.08-2024.01

AccuracyFuzz: Targeted Fuzz Testing Tool Based on FineTuned Large Language Models **第三作者|**

- 基于 Transformer 的方法，在更细粒度的线路级别预测漏洞
- 使用预训练的 CodeBERT 模型和自注意力机制来实现更高的准确性和效率
- 基于大模型评测定向对软件函数脆弱位置进行模式测试
- 该方法在**功能级预测**和**线路级定位**方面显着优于现有方法，提供更精确且更具成本效益的漏洞检测

比赛与项目经历

2022-2023 年第十八届“花期杯”金融创新应用大赛国家一等奖 2022.06–2023.04

- 通过 Solidity 语言编写画册存储程序
- 利用 HTML/CSS、JavaScript 编写前端客制化星图片生成程序
- 参与图像风格艺术化生成深度学习程序的部分编写以及调试

第八届中国国际互联网大学生创新创业大赛国家级铜奖(队长) 2022.04-2022.10

- 参与 商业计划书框架制定 组织商科同学有序完成商业计划书撰写与 PPT 制作
- 通过 回归分析 和 加权平均 等方式 计算得出 智能护膝 的首发城市和全国门店扩展示意图
- 利用 PEST 模型 和 Ansoff 矩阵模型 对 智慧医疗行业 进行 潜力 分析和 风险 分析

社团和组织经历

NASA 编程挑战赛 | 北美 | 队长 2022.02-2022.04

- 协调来自中国，巴基斯坦，英国，印度的 **4 名**队友; 书写 **7000+** 单词的项目说明文档以及APP介绍文档; 与队友一起在 72 小时时间内限制内使用 **Kotlin** 语言开发了一款手机 APP

厦门大学区块链协会 | 副部长 | 活动部 2021.09-2023.09

- 参与并策划由厦门大学和区块链协会联合主办的区块链相关系列讲座，对于主流代币的运行机制具有了初步认识
- 参加校内“区块链+金融”研讨会，以某区块链应用平台为例，探讨非同质化代币、社区代币在低碳经济中的应用

AIESEC 国际志愿者&诺丁汉大学 | 马来西亚 | 国际志愿者 2021.08-2021.10

- 为周边国家的难民子女提供 **20** 节英文授课的通识教育; 为班级中 **80+** 名来自全球各地的学生提供课后作业辅导以及作业评分; 协调安排来自全世界各地的 **100+** 名志愿者在整个活动中的课程时间表

专业技能

编程语言

- 熟练 Python 编程语言 3.x 版本
- 开发和维护 Web 应用程序，具备 **Django** 或 **Flask** 等 Web 框架的经验
- 熟悉 Python 标准库以及第三方库和框架，如 **NumPy**、**Pandas**、**Django**、**Flask** 等

开发环境

- 熟悉 **Linux/Unix** 操作系统，包括基本的命令行操作和系统管理
- 使用 Git 进行版本控制和团队协作，熟悉 GitHub 或 GitLab 等平台，掌握 **Docker** 容器化应用程序

数据挖掘和爬虫

- 熟练使用 **Requests** 库进行HTTP请求
- 使用 **BeautifulSoup** 或 **LXML**进行HTML/XML的解析
- 熟悉JavaScript渲染的页面，使用 **Selenium** 工具进行数据抓取
- 能够将抓取的数据存储到数据库中，如使用SQLite、MySQL、MongoDB等

获奖经历

- 花旗杯金融应用创新大赛 | 国赛一等奖(负责 编程/设计) 2023.02-2023.06
- 美国大学生数学建模竞赛（MCM/ICM）| 国赛一等奖(负责 建模/编程) 2023.02-2023.02
- 高教社杯全国大学生数学建模竞赛 | 国赛二等奖(负责 建模/编程) 2022.11-2022.11
- 第八届中国国际“互联网+”大学生创新创业大赛 | 国赛铜奖(负责人) 2022.04-2022.10
- 第七届中国国际“互联网+”大学生创新创业大赛 | 国赛银奖 2021.07-2021.10

技能与特长

兴趣爱好：水肺潜水、视频剪辑（PR，剪映）、文稿撰写

语言能力：中文（母语）；英文（雅思7.0）；粤语（基础）；马来语（基础）