刘傲凡

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

电话：13418788132

区块链、量化爱好者

GitHub 主页: github.com/ghLcd9dG

教育背景

厦门大学（985） **软件工程** **工学学士** 2020.09-2024.06

- 综合排名：2/57；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** 2020.09 - 至 今

- 保研至北京大学 | 研究兴趣：多模态检索、AI安全性和隐私保护

工作经历

字节跳动 2024.09-2025.03

深圳 | Stone 团队 **算法实习|**

- 在**NextOnCall**项目中，内部AI客服拦截率达到73%，显著提高了客户服务的自动化水平。
- 基于**RAG**（Retrieval-Augmented Generation）技术，优化知识库的检索机制，提升了AI客服的响应质量与效率。
- 用多智能体框架进行客服问题的规划与调度，涵盖**知识库检索**和**人工转单**等流程，确保问题高效处理。
- 通过模型反馈分析与**错误归因**，识别系统潜在问题与性能瓶颈，为持续优化提供数据支持。
- 设计并优化Prompt改写策略，有效提升了**检索精度与效率**，确保AI客服在响应过程中具备更高的准确性与及时性。

粤港澳大湾区人工智能研究院(IDEA) 2024.07-2024.09

深圳 | 多模态组 **算法实习|**

- 针对现有代码检索领域缺乏标准化基准的现状，设计并引入了 **RepoAlign-Bench** 数据集。
- 开发了一种基于**反思的 RAG** 对齐的双塔模型，用于自然语言查询与代码的语义匹配。
- 利用**抽象语法树（AST）**技术和上下文增强，改进了代码和文档表示的精度。

北京智源人工智能研究院 (BAAI) 2023.10-2024.06

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
- 基于 **LLaMa Factory** 参与了模型 **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%

月之暗面 2023.05-2023.08

北京 | 对齐组 **数据标注|**

- 参与用于训练大语言模型的海量数据标注管道开发，设计并实现了一套高效的数据清洗和预处理模块；
- 开发了基于 NLP 技术的智能标注辅助工具，帮助识别常见标签分类错误，显著减少了人工审校的时间；

科研经历

基于上下文扩增仓库级别的NLPL-Probing任务研究 (ACL CCF-A) 2024.09-2024.11

Research on NLPL Probing Tasks Based on Context Augmentation **第一作者|**

- 探索通过**扩增代码中的注释部分**来提升语言模型（LLM）在自然语言理解代码任务中的能力
- **评估Context扩增对NLPL模型理解能力**的提升，通过引入丰富的上下文信息来测试语言模型在多模态任务中的表现。
- 利用包含丰富代码注释的大型开源代码库，并设计任务，评估上下文扩增的效果。
- 对比扩增上下文与传统方法（如仅使用代码结构或注释部分）的性能差异

RefleXGen: 基于反思的可控代码生成 (ICASSP CCF-B) 2024.06-2024.10

RefleXGen: The Unexamined Code Is Not Worth Using **学生一作|**

- 结合了**检索增强生成（RAG）技术**与大语言模型（LLM）的**自反思机制**
- 在不需要对模型进行微调或创建专门安全数据集的情况下，显著提高了代码生成的安全性。
- 通过整合历史反馈和安全代码片段，构建动态的知识库，优化代码生成提示并改善模型生成安全代码的能力。

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

VisualDAN: Exposing Vulnerabilities in VLMs with Visual-Driven DAN Commands **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 (ICME CCF-B) 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

项目经历

基于 MCP 和树莓派的情感陪伴智能硬件开发项目：轻量化流式对话 2024.11–2025.01

- 选用树莓派作为核心硬件，引入 **Model Context Protocal（MCP）**大模型交互协议。
- 通过语音识别模块实现用户语音输入，显示屏模块实现信息输出，在开发过程中反复测试对话效果，不断优化代码与硬件连接。
- 运用 Python 编写程序，借助语音传感器采集用户语音，利用 MCP 与大模型交互，使虚拟女友能理解用户情感意图并生成恰当回应。

MAS: 一种基于MultiAgent的的智能信息搜集与分析系统 2024.03–2023.06

- 参与了系统的架构设计，负责将MindSearch框架与前端技术（Gradio）集成，以实现用户友好的交互界面
- 开发了多个智能体，模拟人类的思维过程，优化信息搜集策略，提高信息搜集的准确性和效率。
- 通过多智能体框架和LLM的结合，实现了深度知识探索，为用户提供了更全面的答案

社团和组织经历

北京大学团委 |青年研究中心 | 学生骨干 2024.09-2025.02

- 负责北京大学校级社团活动的统筹与规划，协调50余个学生社团，确保活动有序开展并达到预期效果。主导校级大型活动（“深研院青年与未来交流沙龙”）的策划与执行，吸引超过1000名师生参与；推动活动品牌化运营，提升了团委活动在校内外的影响力。

汇丰商学院QTA协会 | 学术研究部 | 部长 2022.02-2022.04

- 参与股票量化选股策略开发，基于多因子模型构建因子库，结合市值、估值、动量等因子筛选优质股票组合，显著提升策略收益率。处理海量金融数据，包括行情数据和财务报表数据，优化数据清洗与预处理流程，确保模型训练数据的准确性与时效性。

专业技能

编程语言

- 开发和维护 Web 应用程序，具备 **Django** 或 **Flask** 等 Web 框架的经验
- 了解 Solidity 语言，能够编写高效、安全的智能合约，熟悉 **ERC-20、ERC-721** 等标准协议。

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

语言能力：中文（母语）；英文（雅思7.0）；

兴趣爱好：攀岩、水肺潜水、文稿撰写、视频剪辑（PR，剪映）