孙玉粮

(+86) 139-4510-8536 · 21371245@buaa.edu.cn · https://intrepidLi.github.io/

教育背景

北京航空航天大学, 计算机科学与技术, 在读大四学生 2021.9 至今

GPA 3.68/4.0, 88.5/100

技术能力

- **编程语言**: Python, C, C++, Java, Verilog, JavaScript ,HTML/CSS, SQL
- **工程方面**: Git, Flask, Pytorch

项目经历

AirQA 民航旅客问答系统 | 北航 ACT 实验室，指导人：张日崇教授 2023.1-2023.4

- 负责构造整套数据处理系统使之能够对航司原始杂乱数据进行分类生成问答对
- 采用模板化方法和 docT5query 模型分别对规律化和非规律化数据进行问答数据生成
- 采用基于百度公司的 RocketQA 预训练模型在民航数据集上训练出的新模型进行检索式问答生成

基于 SysY 的编译器 | 北航编译技术课程设计 2023.9-2023.12

项目地址：https://github.com/intrepidLi/BUAA_Compilers_2023

- 使用 Java 完成编译器，可对 SysY（C 语言的一个子集）进行从源码到 LLVM IR 中间代码的转换
- 系统组成：词法分析，语法分析，语义分析生成抽象语法树（AST），中间代码生成
- 利用自动机理论手动进行词法分析，利用递归下降进行语法分析生成抽象语法树（AST），同时进行错误处理。之后使用“一切皆 Value”的思想构建新的 LLVM IR 的语法树，生成 LLVM IR 代码
  本项目是北航编译技术的课程设计，通过了目标代码为 LLVM IR 编译器的所有测试。

北航比价跳蚤市场 | 北航数据库系统概论课程设计 2023.9-2023.12

项目地址：https://github.com/intrepidLi/BUAA_DataBase_2023

- 使用 ReactJS ,Flask 框架和华为云数据库完成的前后端交互网站，包括用户系统，评论系统，商品交易系统等功能
- 主要负责前端和交互工作，包括使用 Material-UI 进行网站设计，并用 axios 库进行前后端交互和前后端的 Debug 完善

BattleByte 实时代码对战平台 | 北航软件工程课程设计 2024.3-2024.6

项目地址：https://github.com/HelloWorldSE/BattleByte-frontend

- 使用 VueJS ,SpringBoot 框架和腾讯云服务器完成的可前后端交互的，支持多人并发实时编程的在线对战网站，包括用户系统，对战系统，题库系统等功能
- 主要负责前端和交互工作，包括使用 Ant Design UI 进行网站设计，并用 axios 库进行与后端的对接

论文发表

WaterBench: Towards Holistic Evaluation of Watermarks for Large Language Models

Shangqing Tu*, **Yuliang Sun***, Yushi Bai, Jifan Yu, Lei Hou, Juanzi Li(* 相同贡献)

- 文章链接: https://aclanthology.org/2024.acl-long.83.pdf

- *The Proceedings of 62nd Annual Meeting of the Association for Computational Linguistics*

Knowledge-to-Jailbreak: One Knowledge Point Worth One Attack

Shangqing Tu*, Zhuoran Pan*, Wenxuan Wang, Zhexin Zhang, **Yuliang Sun**, Jifan Yu, Hongning Wang, Lei Hou, Juanzi Li(* 相同贡献)

- 预印本: https://arxiv.org/abs/2406.11682
- 在投

研究经历

LLM Watermarking Benchmark and Methods for LLM Jailbreak and Defense

研究助理 | 清华大学知识工程研究室，指导人: 李涓子教授 2023.5-2024.6

- 学习和了解不同的 LLM 水印算法
- 设计一个可以用于全面评估 LLM watermark 算法的 Benchmark——WaterBench，该 Benchmark 构建了九个不同的任务（五个类别）作为评估数据集，使用 GPT4-Judge 和人工评价的方法对文本进行评估。在这个工作中主要负责超参数搜索，评估实验的进行，数据处理以及部分论文的编写
- 学习关于 LLM 越狱和防护的方法，合作完成 Knowledge-to-jailbreak，这是一种利用不同领域内知识对 LLM 进行针对领域的 Jailbreak 新算法

Trustworthy AI and Multi-Modal Watermarking

研究助理 | 香港中文大学 MISC Lab，指导人: Prof. Irwin King 2024.3-2024.9

- 学习更多关于大语言模型安全和水印方面的知识
- 设计一种将语言模型水印迁移到图生文模型上的新方法，使得新水印在应对各种攻击时检测更加鲁棒

竞赛/荣誉

- 北航 2021-2022 学年学习优秀二等奖学金
- 北航 2021-2022 学年，2022-2023 学年学科竞赛二等奖学金
- 第十五届蓝桥杯 Python A 组二等奖，2024 年 4 月
- 第十四届蓝桥杯 C/C++ 研究生组三等奖，2023 年 4 月

其他

- 语言：英语（CET6，IELTS:6.5），汉语
- GitHub: https://github.com/intrepidLi