YULIANG SUN

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

EDUCATION

BeiHang University, Beijing, China Sep.2021 – Present

Fourth year undergraduate in Computer Science and Technology(CS)

GPA 3.68/4.0, 88.5/100

SKILLS

- **Programming Languages**: Python, C, C++, Java, Verilog, JavaScript ,HTML/CSS, SQL
- **Frameworks**: Git, Flask, Pytorch

PROJECT EXPERIENCES

AirQA Civil Aviation Passenger Question and Answer System Jan.2023 – Apr.2023

Data Preprocess Supervisor: Prof.Richong Zhang, Beihang University

- Constructed a data processing system so that it can classify the original messy data of the airlines and generate question-answer pairs.
- Used the template method and docT5query model to generate question and answer data for regular and irregular data respectively.
- Used a new model trained on the civil aviation data set based on RocketQA pre-training model(Baidu Inc.) to generate retrieval-style questions and answers.

A Compiler based on SysY Sep.2023 – Dec.2023

Individual Project Beihang University Compilation Technology Course Project

Project Address: https://github.com/intrepidLi/BUAA_Compilers_2023

- Completed the compiler with Java to convert source code in SysY(a sub set of the C language) to LLVM IR intermediate code.
- System components: lexical analysis, syntax analysis, semantic analysis to generate abstract syntax tree (AST), intermediate code generation.
- Used automata theory to manually perform lexical analysis, and use recursive descent to perform syntactic analysis to generate an abstract syntax tree (AST), while performing error handling. Then build a new LLVM IR syntax tree and generate LLVM IR code.

BUAA Price Comparison Flea Market Sep.2023 – Dec.2023

Front-end and Interaction Beihang University Database System Introduction Course Project

Project Address: https://github.com/intrepidLi/BUAA_DataBase_2023

- A front-end and back-end interactive website using ReactJS, Flask framework and Huawei Cloud Database, including user system, comment system, commodity trading system and other functions.
- Mainly responsible for front-end and interaction work, including using Material-UI for website design, using the axios library for front-end and back-end interaction and interaction debugging improvements.

BattleByte Real-time Code Battle Platform Mar.2024 – Jun.2024

Front-end and Interaction Beihang University Software Engineering Course Project

Project Address: https://github.com/HelloWorldSE/BattleByte-frontend

- Developed using VueJs, SpringBoot framework, and Tencent Cloud servers, this online battle website enables real-time programming battles with multiple concurrent users, featuring user systems, battle systems, and question database systems.
- Primarily responsible for front-end and interaction tasks, including website design using Ant Design UI and interfacing with the backend using the axios library.

PUBLICATIONS

WaterBench: Towards Holistic Evaluation of Watermarks for Large Language Models

Shangqing Tu*, **Yuliang Sun***, Yushi Bai, Jifan Yu, Lei Hou, Juanzi Li (*Equal Contribution)

- Passage Link: 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(*Equal Contribution)

- Preprint: https://arxiv.org/abs/2406.11682
- Under Review

RESEARCH EXPERIENCES

LLM Watermarking Benchmark and Methods for LLM Jailbreak and Defense

Research Assistant, Supervisor: Prof.Juanzi Li, Tsinghua University May.2023 – Jun.2024

- Study and understand different LLM watermarking algorithms.
- Design a comprehensive benchmark for evaluating LLM watermark algorithms called WaterBench. This benchmark constructs nine different tasks (across five categories) as the evaluation dataset, using methods such as GPT4-Judge and human evaluation to assess the text. In this work, the main responsibilities include hyperparameter search, conducting evaluation experiments, data processing, and writing parts of the paper.
- Learn about methods for LLM jailbreaking and defense, and collaborate with senior students to complete Knowledge-to-jailbreak, a new algorithm that utilizes knowledge within different domains to perform domain-specific Jailbreak on LLMs.

LLM Watermarking Benchmark and Methods for LLM Jailbreak and Defense

Trustworthy AI and Multi-Modal Watermarking

Research Assistant, Supervisor: Prof. Irwin King, Chinese University of Hong Kong Mar.2024 – Sep.2024

- Learn more about the security and watermarking of large language models.
- Design a new method for transferring language model watermarks to image-to-text models, making the new watermark more robust against various attacks.

HONORS

Scholarships for Outstanding Second-class Learning at Beihang University	2022
Scholarships for Outstanding Second-class Competitions at Beihang University	2022,2023
Second-class Prize of the 15th Lanqiao Cup Python Undergraduate A Group	2024

MISCELLANEOUS

- Language：English（CET6, IELTS:6.5），Mandarin - Native speaker
- GitHub: https://github.com/intrepidLi