Junsheng Huang

☎ (+86)13798300044
✉ jh103@illinois.edu ✉ junsheng.21@intl.zju.edu.cn
github.com/no-touch-fish

Education

ZheJiang University-University of Illinois at Urbana-Champaign Expected June 2025
Dual Bachelor of Electrical and Computer Engineering (UIUC GPA: 4.00 / 4.00) (ZJU GPA: 3.97 / 4.00)

- **Relevant Courses:** Computer System Engineering(ECE391), Data Structure(CS225), Linear Algebra(MATH257),Probability with Engineering Applications(ECE313), Artificial Intelligence(CS440), Algorithm(CS374), Machine Learning(CS446), Applied Parallel Programming(ECE408)

Research Experience

Random Augmentations Cheaply Break LLM Safety Alignment
Instructor: *Gagandeep Sigh(UIUC Professor)*, *Jason Vega(UIUC Ph.D)*

- **Motivation:** Current jailbreak methods are rather costly or involve a non-trivial amount of creativity and effort. Since that, we investigate how simple random augmentations to the input prompt affect safety alignment effectiveness in LLMs from different dimensions.
- **Result:** We show that low-resource and unsophisticated attackers can significantly improve their chances of bypassing alignment with just 25 random augmentations per prompt.
- **Contribution:** I research and implement different simple data augmentations, including string level and character level. At the same time, I help to check the evaluation metric of the project and do case study as well as labeling the experimental result manually to see how LLM classification align with human evaluation.

Teaching Large Language Models to Handle the Composition of Multiple Problems Simultaneously
Instructor: *May Fung(HKUST Professor)*

- **Motivation:** Current evaluation of LLM hallucination only focus on single problem setting. Because of this, we investigate how LLM perform and deal with hallucination under multiple problem setting.
- **Result:** We propose a novel fine-tune method called **M**ultiple **A**nswers and **C**onfidence Stepwise **Tuning** (**MAC-Tuning**) with up to 12% improvement comparing with baseline and up to 40% improvement comparing with Zero-shot model under multiple problem setting.
- **Contribution:** I conduct the entire process of data collection and building the project code. At the same time, I tested various approaches, like LLM-Judge and keyword extraction, to assess the accuracy of LLM-generated outputs. Furthermore, I experimented diverse evaluation metrics including accuracy, AP score and MAP to comprehensively evaluate model performance.

Publication

- Jason Vega, **Junsheng Huang**, Gaokai Zhang, Hangoo Kang, Minjia Zhang, Gagandeep Singh. *Stochastic Monkeys at Play: Random Augmentations Cheaply Break LLM Safety Alignment*. Submitted to ICLR 2025 (under review)

- **Junsheng Huang**, Zhitao He, Sandeep Polisetty, May Fung. *Teaching Large Language Models to Handle the Composition of Multiple Problems Simultaneously*. Submitted to NAACL 2025 (under review)

Projects

391 OS System

- **Basic Functionalities**: Implemented an operating system supporting basic functionalities like scheduling, interrupts, system calls, exceptions, and file systems
- **Self-designed features**: ATA drivers to support writable file system, command history, changeable color and auto complete

LLM Attack Based on Gradient Method
Instructor: *Gagandeep Sigh(UIUC Professor),Jason Vega(UIUC Ph.D)*

- **Motivation:** If we can decide the very first output part of LLM generation (which is ”**prefilling attack**”), we can easily bypass the safety training of LLMs. One of the easiest way to do so is utilizing the Greedy Coordinate Gradient (**GCG**) attack to find ”ignore string” to ignore the ”ending token” that separates the input prompt and LLM generation. Also, comparing with random token in **GCG** attack, we can briefly give an explanation to the random string.
- **Result:** We attack LLaMA2-7B and LLaMA2-13B with 97% attack successful rate (**ASR**).
- **Contribution:** I develop the code to find ”ignore string” based on **GCG** attack and try different loss functions as well as different place to insert the string.

Technical Skills

Programming: C, C++, Python, MATLAB, x86 assembly, Unreal Engine 5, Pytorch frame, CUDA frame, VLLM, PEFT
Spoken Languages: English (Proficient), Mandarin (Native), Cantonese (Native)

Honors

Honorable Mention of Mathematical Contest of Modeling (**May 2023**)

ZJU-UIUC Institute Dean’s List in **Semaphore Year**

ZJU-UIUC Institute Third-Class Academic Excellence Award for **Semaphore Year** and **Junior Year**

UIUC Grainger Engineering Department Dean’s List for two semesters in **Junior Year**

Teaching and Leadership

Course Assistant for ECE391 (Computer System Engineering)	Jan 2024 - May 2024
Teaching Assistant for MATH241 (Calculus III)	Sep 2024 - Dec 2024
Student Representative for ECE Major in ZJUI	Sep 2021 - Jun 2022, Sep 2024 - Jun 2025