Feiyue Xu

Shanghai Jiao Tong University
linuxwolfie@sjtu.edu.cn https://xotaichi.github.io/FeiyueXu/ github.com/XOTaichi

Education

Shanghai Jiao Tong University **Sep. 2022 – Current**

Cyber Science ,The School of Electronics, Information and Electrical Engineering *Shanghai, China*

Undergraduate student with a total score of **91.93/100**, ranked **5th** out of 87 students.

Passed the CET-6 (College English Test Band 6) with a score of 690 (Total Score 710).

Proficient in C++, Python, SQL, Pytorch, LangChain, Docker, Django, Web Application Development, Git

Expertise in NLP, LLM/MLLM Application, trustworthy boundaries for large models, and privacy protection.

Honors and Awards

- National Scholarship **(Rank 1/87)** **2023-2024**
- Shanghai Jiao Tong University Undergraduate A Scholarship **(Rank 1/87)** **2023-2024**
- Global AI Attack-Defense Challenge: Third-place Team **(Top 5)** **2024**
- Second prize in National Undergraduate Mathematical Modeling Competition, Shanghai Division **2023**
- Second prize in the eighth 'Freshmen Cup' Science and technology innovation competition **2023**
- Outstanding League Cadre, Shanghai Jiao Tong University **2023-2024**
- Outstanding Student Cadre, Shanghai Jiao Tong University **2022-2023**
- Outstanding Social Practice Individual, School of Electronic Information and Electrical Engineering **2023**

Experience

LLM Safety Benchmark **August 2024 – Current**

Undergraduate research Intern, supervised by Prof. Shuo Wang *Shanghai JiaoTong University*

- Developed **a train-free baseline** for the LLM Safety Benchmark, integrating 10+ red teaming strategies and customized defenses on a unified evaluation platform.
- Investigated jailbreak resistance evolution in LLMs, assessing the effectiveness of old techniques on newer models.
- **Quantified misalignments** between LLM-based judgment and human values.
- Analyzed jailbreak attack consistency across models, **identifying vulnerabilities** related to hidden space clustering, attention shifts, and security alignment.

Privacy Protection in Large Language Models **Feb 2024 – Sep 2024**

Undergraduate research Intern, supervised by Prof. Liyao Xiang *ShangHai JiaoTong University*

- **Developed a localized privacy detection system**, eliminating reliance on cloud-based models, suitable for large model training corpus cleaning and everyday privacy protection.
- Explored and compared SOTA **triplet extraction techniques** (NER, OpenIE, LLM), ultimately selecting supervised fine-tuned small models for knowledge graph construction.
- Extended the knowledge graph with **RAG techniques** and performed **end-to-end inference** using a small language model, surpassing Microsoft Presidio's performance and approaching the leading inference model O1 (68%) on datasets like PersonaChat and Reddit.
- Co-authored the paper **"Graph-based Contextual Privacy Leakage Detection for User-Generated Texts"**, currently under submission.

Brain-Computer Interface Summer Program **July 2024- Aug 2024**

Summer Workshop, supervised by Prof. Andrei Kozlov *Top Talent Development Program*

- Conducted in-depth research on the latest advancements in brain-computer interfaces (BCI) for signal acquisition (e.g. endovascular electrodes) and the influence of computational neuroscience on model architecture design.
- Compiled recent research on BCIs for human enhancement, analyzing how BCI technologies for people with disabilities can be optimized and adapted for healthy individuals from both signal acquisition and control device perspectives.
- Authored and published the academic paper **"BCI for Exoskeletons and Prostheses: From Rehabilitation to Human Enhancement"** at an international conference.

T2I-Fuzzer: Jailbreaking Text-to-Image Models with Hybrid Strategies **Oct 2024**

- Developed a strategy library for bypassing post-image security, including adversarial canvas separation, with **discrete weight** searching to optimize element bypass.
- Implemented pre-text bypass strategies using knowledge base retrieval, generating multiple prompts with **a genetic algorithm** for mapping embeddings to harmless token-IDs for text jailbreaking.
- Designed a zero-sum game-breaking module combining **RAG, GAN, and COT** for prompt optimization, integrating HITL to blend human experience with ChatGPT's generative capabilities for iterative jailbreaking.