Renjie Gu

Email: jackyofgrj@gmail.com Linkedin HomePage

Education & Experience

Central South University — ChangSha China

Undergraduate of computer science and technology GPA : **90.4**/100 **3.92**/4.0 09/2022-now

Nanyang Technological University — Singapore

Research Assistant under the advice of Yiming Li 12/2024-now

- My current research focuses on building resilient safe diffusion model against **Downstream Fine-tuning.**
- Our paper has been submitted to NeurIPS 2025.I'm the second author,I spent lots of time studying defense of llm and DMs against harmful fintuning and the methods for Bilevel Optimization. I ialso mplanted all the codes.

SYNC **(https://sync-xyz.com)** **(a start-up for ai-company app)** — Beijing China

LLM and agent develop Intern 07/2024-08/2024

Researches

- **Secret because it's on submit to Nips 2025**
- This is the most interesting, elegant, and solid work I have ever done. **NeurIPS 2025 on submitting**
  - I have extensively researched methods for defending LLMs and diffusion models against harmful fine-tuning and methods for efficient Bilevel Optimization.
  - I designed the overall framework of our approach and finished the experiment codes.
- **MHALO: Evaluating MLLMs as Fine-grained Hallucination Detectors** **ACL 2025 findings**
  - I develop the benchmark with python and here is the github repository.
  - I proposed various prompt-based methods **enhancing Instrcution following of mllm** and metrics for better alignment with human judgement.
- **Course-Correction: Safety Alignment Using Synthetic Preferences.** **EMNLP 2024 Industry Track**
  - Fourth contribution
  - In this project, I was mainly responsible for the experimental part, during which I became proficient in using servers, managing environments with Conda, and collaborating via GitHub.
- AI Startup Research Experience
  - Researched **long-term memory mechanisms** for LLM agents, enabling personalized user memory retention across conversations.
  - Explored and applied **creativity enhancement** techniques (prompt tuning & lora fintuning) to improve LLM response diversity.

Projects

- A Job Recommendation and Competency Evaluation System.Responsible for researching and developing back-end

recommendation algorithms. Utilized **Python, PyTorch,** and Flask to implement several functions.

- An IOS App For traveling using SwiftUI and google firebase . This app allows you to publish travel guides, view guides shared by others, chat with people, receive AI-powered travel recommendations, and manage your budget efficiently.
- **RiseUp - AI-powered Morning Productivity Assistant：** The project sends personalized emails with gratitude reflections, goal reminders, weather updates, and optimized task lists to help users start their day with clarity. I **independently** designed and built the entire system, integrating AI content generation, multi-API data fetching, and automated delivery via GitHub Actions.

Skills

English: I received an IELTS score of 7.0 . I also participated in the Chinese Undergraduate English Competition and won a national second prize. I am proficient in reading research papers and communicating effectively with others.

Team-sorking: I have great experience in working as a team.

Tools and programming languages: Python,Pytorch,C++,Swift,Git,SQL,Conda,HTML/CSS