Yan Yang

+86 18918821601 • yanyang.nlp@gmail.com • yannnnnny.github.io

Education

Shanghai University of Finance and Economics — **Shanghai, China**
School of Computer and Artificial Intelligence — *Sep. 2023 — present*

- Master Candidate of Software Engineering
- GPA: 3.54/4.00
- Advisor: Professor Yun Chen

Shanghai University of Finance and Economics — **Shanghai, China**
School of Information Management and Engineering — *Sep. 2019 — Jun. 2023*

- Bachelor of Electronic Business
- GPA: 3.52/4.00, Rank: 7/44

Research Interest

Large Language Models, Model Merge, Model Compression, LLM Safety.

Publications

ImPart: Importance-Aware Delta-Sparsification for Improved Model Compression and Merging in LLMs Paper Code

- Yan Yang, Yixia Li, Hongru Wang, Xuetao Wei, Jianqiao Yu, Yun Chen, Guanhua Chen
- In Proceedings of **ACL 2025**.

SeqAR: Jailbreak LLMs with Sequential Auto-Generated Characters Paper Code

- Yan Yang, Zeguan Xiao, Xin Lu, Hongru Wang, Xuetao Wei, Hailiang Huang, Guanhua Chen, Yun Chen.
- In Proceedings of **NAACL 2025**.

Distract Large Language Models for Automatic Jailbreak Attack Paper Code

- Zeguan Xiao, Yan Yang, Guanhua Chen, Yun Chen.
- In Proceedings of **EMNLP 2024**.

Pi-SQL: Enhancing Text-to-SQL with Fine-Grained Guidance from Pivot Programming Languages

- Yongdong Chi, Hanqing Wang, Yun Chen, Yan Yang, Zonghan Yang, Xiao Yan, Guanhua Chen.
- In Findings of **EMNLP 2025**.

Selected Research Experience

ImPart: Importance-Aware Delta-Sparsification for Improved Model Compression and Merging in LLMs

- Motivated by the observation that singular vectors with larger singular values encode more important task-specific information, we developed ImPart, which **assigns variable sparsity ratios to singular vectors based on their corresponding singular values**.
- We conducted extensive experiments on several task-specific LLMs, including **WizardMath** for mathematical reasoning, **WizardCoder** for code generation, and **LLaMA-2-Chat** for instruction following. Experimental results show that ImPart demonstrates **2× higher compression efficiency than baselines**.

- To further showcase its versatility, ImPart was combined with existing **quantization** and **model merging** methods. When used with quantization, ImPart achieved **near-lossless performance** by compressing the delta parameter to just **1/32** of its original size, while also **enhancing model merge performance**.
- In Proceedings of **ACL 2025** (1^{st} Author).

SeqAR: Jailbreak LLMs with Sequential Auto-Generated Characters

- Building on existed character simulation methods for jailbreaks, SeqAR **optimizes multiple characters** and **prompts LLMs to respond sequentially as these characters** in a single output, thereby **further distracting LLMs** and **expanding the applicable area** of the generated jailbreak prompt.
- We conducted extensive experiments on **open-source models (e.g., LLaMA-2 and LLaMA-3)** and **proprietary models (e.g., GPT-3.5-Turbo, GPT-4, GPT-4o, and Gemini)**.
- SeqAR achieved **state-of-the-art jailbreak performance**, with **attack success rates > 90% on LLaMA-2** and > **85% on GPT-3.5 series models**. SeqAR also exhibits strong transferability, and existing defense methods are insufficient, underscoring the **widespread and critical nature of the identified vulnerabilities**.
- In Proceedings of **NAACL 2025** (1^{st} Author).

Distract Large Language Models for Automatic Jailbreak Attack

- Leveraging the observation that **irrelevant context can distract large language models and diminish their performance**, we proposed DAP, which employs specially designed jailbreak templates embedded with irrelevant context to **conceal malicious content** and **iteratively refines** these templates using an LLM **memory-reframing** mechanism.
- We conducted rigorous experiments across both **open-source models (e.g., Vicuna and LLaMA-2)** and **proprietary models (e.g., GPT-3.5-Turbo and GPT-4)**.
- DAP demonstrated **robust jailbreak performance** with a **100% attack success rate on Vicuna** and **nearly 80% on GPT-3.5 series models**, highlighting both the severity and pervasiveness of this safety vulnerability. Moreover, when combined with other jailbreak techniques, DAP's attack performance is further enhanced.
- In Proceedings of **EMNLP 2024** (2^{nd} Author).

Internship

Toursun Synbio — **Shanghai, China**
Research Intern — *Jun. 2022 – Feb. 2023*

- Host: Yuguang Wang, Yiqing Shen
- Research Topic: Multimodal Medical Classification

Awards

- Outstanding Graduate of Shanghai University of Finance and Economics (top 10%). *Jun. 2023*
- Tailong Commercial Bank Scholarship (top 15%). *Jan. 2023*
- Renming Scholarship, 3^{rd} Prize (top 15%). *Sep. 2020 – Jun. 2023*

Others

- Core Technical Team Member of the N.O.P.E. Robotics Club at Shanghai University of Finance and Economics. *Sep. 2022 – Sep. 2024*
- Leader of the Academic Department of the College Student Union. *Sep. 2020 – Jun. 2021*