

# Yan Yang

☎ +86 18918821601 • ✉ [yanyang.nlp@gmail.com](mailto:yanyang.nlp@gmail.com) • 🌐 [yannnnny.github.io](https://yannnnny.github.io)

## Education

---

### Shanghai University of Finance and Economics

School of Computing and Artificial Intelligence

Shanghai, China

Sep. 2023 — present

- Master Candidate of Software Engineering
- GPA: 3.55/4.00
- Advisor: Professor [Yun Chen](#)

### Shanghai University of Finance and Economics

School of Information Management and Engineering

Shanghai, China

Sep. 2019 — Jun. 2023

- Bachelor of Electronic Business
- GPA: 3.54/4.00, Rank: 7/44

## Research Interest

---

Large Language Models, Efficient Machine Learning, LLM Alignment.

## 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 [📄 Paper](#) [</> Code](#)

- [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

- Proposed ImPart, an importance-aware delta-sparsification method that **assigns variable sparsity ratios to singular vectors based on their corresponding singular values**.
- Conducted extensive experiments on several task-specific LLMs, including **WizardMath** for mathematical reasoning, **WizardCoder** for code generation, and **LLaMA-2-Chat** for instruction following. Results revealed that ImPart performs **2× higher compression efficiency than the baselines**.

- Combined ImPart with existing quantization and model merging methods to demonstrate its versatility. 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<sup>st</sup> Author).

### SeqAR: Jailbreak LLMs with Sequential Auto-Generated Characters

- **Optimized multiple characters** and prompted LLMs to **sequentially respond as these characters in a single output**, thereby **further distracting LLMs** and **expanding the applicable area** of the generated jailbreak prompt.
- 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 exhibited strong transferability, while existing defense methods proved insufficient, underscoring **the widespread and critical nature of the identified vulnerabilities**.
- In Proceedings of **NAACL 2025** (1<sup>st</sup> Author).

### Distract Large Language Models for Automatic Jailbreak Attack

- Noticed 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 this context** using an **LLM memory-reframing mechanism**.
- Executed 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).
- Achieved an **exceptional jailbreak performance** with DAP — **100% attack success rate on Vicuna** and **nearly 80% on GPT-3.5 series models**, highlighting both **the severity and pervasiveness of safety vulnerability**. Moreover, when combined with other jailbreak techniques, DAP's attack performance is further enhanced.
- In Proceedings of **EMNLP 2024** (2<sup>nd</sup> Author).

## Awards

---

- Nominee, “Academic Star” Award, Shanghai University of Finance and Economics (top 1%). *Nov. 2025*
- National Scholarship, Postgraduate (top 1%). *Nov. 2025*
- Outstanding Graduate of Shanghai University of Finance and Economics (top 10%). *Jun. 2023*
- Tailong Commercial Bank Scholarship (top 15%). *Jan. 2023*
- Renming Scholarship, 3<sup>rd</sup> Prize (top 15%). *Sep. 2020 – Jun. 2023*

## Internship

---

### Toursun Synbio

Research Intern

Shanghai, China

*Jun. 2022 – Feb. 2023*

- Host: [Yuguang Wang](#), [Yiqing Shen](#)
- Research Topic: Multimodal Medical Classification

## Service & Activities

---

- ACL ARR Reviewer *2025*
- 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*