# Yu Fu

https://fyyfu.github.io/

Email : yfu093@ucr.edu
Mobile : +1-951-973-1929

## Research Interests

- **Summarization**, **Long-Context Generation**, **AI Safety**, **Reinforcement Learning**

## Education

- **University of California, Riverside** Riverside, CA
  *Ph.D in Computer Science (Advisor: Yue Dong)* *Sep. 2023 – Present*
- **Tianjin University** Tianjin, China
  *Master of Computer Technology (Advisor: Deyi Xiong)* *Sep. 2020 – Jun. 2023*
- **Xidian University** Xi'an, China
  *Bachelor of Software Engineering* *Sep. 2016 – June. 2020*

## Experience

- **Microsoft** Seattle, WA
  *Research Intern* *Jun. 2024 - Sep. 2024*
  - **Project**: KV-cache Compression for Long-Context Generation
- **University of California, Riverside** Riverside, CA
  *Research Intern* *Sep. 2022 - Aug. 2023*
  - **Project**: Inverse Reinforcement Learning for Summarization
- **LangBoat** Beijing, China
  *Research Intern* *Jan. 2022 - Sep. 2022*
  - **Project**: Neural Machine Translation with Translation Memory

## Publications

- [**1**]: **Yu Fu**, Zefan Cai, Abedelkadir Asi, Wayne Xiong, Yue Dong, Wen Xiao (2024). Not All Heads Matter: A Head-Level KV Cache Compression Method with Integrated Retrieval and Reasoning. *preprint*
- [**2**]: **Yu Fu**, Yufei Li, Wen Xiao, Cong Liu, Yue Dong (2023). Safety Alignment in NLP Tasks: Weakly Aligned Summarization as an In-Context Attack. *Association for Computational Linguistics (ACL 2024)*
- [**3**]: Erfan Shayegani, Md Abdullah Al Mamun, **Yu Fu**, Pedram Zaree, Yue Dong, Nael Abu-Ghazaleh. Survey of Vulnerabilities in Large Language Models Revealed by Adversarial Attacks. *preprint. (ACL 2024 tutorial)*
- [**4**]: **Yu Fu**, Deyi Xiong, Yue Dong (2023). Watermarking Conditional Text Generation for AI Detection: Unveiling Challenges and a Semantic-Aware Watermark Remedy. *Association for the Advancement of Artificial Intelligence (AAAI2024)*
- [**5**]: **Yu Fu**, Deyi Xiong, Yue Dong (2023). Inverse Reinforcement Learning for Text Summarization. *Findings of Empirical Methods of Natural Language Processing (Findings of EMNLP2023)*

## Honors & Awards

- **UCR Dean's Distinguished FellowShip** Riverside, CA
  *Personal* *Sep. 2023 – Sep. 2024*

## Programming Skills

- **Programming**: Python, Pytorch, Faiseq, Transformers, RL for Summarization
- **Natural Languages**:Mandarin Chinese (Native), English (Proficient)

## Program Committee & Reviewer

- ACL Rolling Review, EMNLP2024, NAACL-HLT/ACL/EMNLP Tutorials 2025