

Jun Yan

✉ junyann@google.com 🌐 junyann.github.io

RESEARCH INTERESTS

Alignment, Safety, and Robustness of Large Language Models (LLMs), Machine Reasoning, Information Extraction

PROFESSIONAL EXPERIENCE

Research Scientist @ Google

Cloud AI Research

Sunnyvale, CA, U.S.

07/2024 – Current

- Conducted applied research on LLM safety and robustness.

Research Intern @ Samsung Research America

Knowledge and Dialogue Team

Mountain View, CA, U.S.

05/2023 – 08/2023

- Mentors: Vikas Yadav, Shiyang Li
- Formulated data poisoning attacks and studied their defenses for LLM instruction tuning.
- Developed quality-guided data selection approach for effective and efficient instruction tuning.
- Developed an instruction-following evaluation protocol based on verbalizer manipulation.

Research Intern @ Meta

Language Understanding and Question Answering Team

Seattle, WA, U.S.

05/2022 – 08/2022

- Mentors: Asish Ghoshal, Scott Wen-tau Yih, Asli Celikyilmaz, Pedro Rodriguez
- Developed parameter-efficient tuning with adaptive capacity for improved generalization.

Applied Scientist Intern @ Amazon

Product Graph Team

Seattle, WA, U.S. (Remote)

06/2020 – 11/2020

- Mentors: Nasser Zalmout, Yan Liang, Xin Luna Dong
- Developed a model that automates value extraction of multiple attributes from product profiles.

EDUCATION

University of Southern California

Ph.D. in Computer Science

Los Angeles, CA, U.S.

08/2019 - 08/2024

- Advisor: Prof. Xiang Ren

Tsinghua University

B.Eng. in Electronic Engineering

Beijing, China

08/2015 - 07/2019

- Advisor: Prof. Zhiyuan Liu

PUBLICATIONS AND PREPRINTS

(* indicates equal contribution)

Safety and Alignment

- Rethinking Backdoor Detection Evaluation for Language Models
Jun Yan, Wenjie Jacky Mo, Xiang Ren, Robin Jia
arXiv:2409.00399 [[Preprint](#)]
- How Susceptible are Large Language Models to Ideological Manipulation?
Kai Chen, Zihao He, **Jun Yan**, Taiwei Shi, Kristina Lerman
In *Proceedings of EMNLP'24* [[Paper](#)] **Best Paper Runner-up at SeT LLM @ ICLR 2024**
- Backdooring Instruction-Tuned Large Language Models with Virtual Prompt Injection
Jun Yan, Vikas Yadav*, Shiyang Li*, Lichang Chen, Zheng Tang, Hai Wang, Vijay Srinivasan, Xiang Ren, Hongxia Jin
In *Proceedings of NAACL'24* [[Paper](#)]
- AlpaGasus: Training A Better Alpaca with Fewer Data
Lichang Chen*, Shiyang Li*, **Jun Yan**, Hai Wang, Kalpa Gunaratna, Vikas Yadav, Zheng Tang, Vijay Srinivasan, Tianyi Zhou, Heng Huang, Hongxia Jin
In *Proceedings of ICLR'24* [[Paper](#)]
- Test-Time Backdoor Mitigation for Black-Box Large Language Models with Defensive Demonstrations
Wenjie Mo, Jiashu Xu, Qin Liu, Jiong Xiao Wang, **Jun Yan**, Chaowei Xiao, Muhao Chen
arXiv:2311.09763 [[Preprint](#)]
- BITE: Textual Backdoor Attacks with Iterative Trigger Injection
Jun Yan, Vansh Gupta, Xiang Ren
In *Proceedings of ACL'23* [[Paper](#)]

Robustness and Evaluation

- Instruction-Following Evaluation through Verbalizer Manipulation
Shiyang Li, **Jun Yan**, Hai Wang, Zheng Tang, Xiang Ren, Vijay Srinivasan, Hongxia Jin
In *Findings of NAACL'24* [[Paper](#)]
- GPT-4V(ision) as a Generalist Evaluator for Vision-Language Tasks
Xinlu Zhang*, Yujie Lu*, Weizhi Wang*, An Yan, **Jun Yan**, Lianke Qin, Heng Wang, Xifeng Yan, William Yang Wang, Linda Ruth Petzold
arXiv:2311.01361 [[Preprint](#)]
- On the Robustness of Reading Comprehension Models to Entity Renaming
Jun Yan, Yang Xiao, Sagnik Mukherjee, Bill Yuchen Lin, Robin Jia, Xiang Ren
In *Proceedings of NAACL'22* [[Paper](#)]
- RockNER: A Simple Method to Create Adversarial Examples for Evaluating the Robustness of Named Entity Recognition Models
Bill Yuchen Lin, Wenyang Gao, **Jun Yan**, Ryan Moreno, Xiang Ren
In *Proceedings of EMNLP'21* [[Paper](#)]

Information Extraction

- AdaTag: Multi-Attribute Value Extraction from Product Profiles with Adaptive Decoding
Jun Yan, Nasser Zalmout, Yan Liang, Christan Grant, Xiang Ren, Xin Luna Dong
In *Proceedings of ACL'21* [[Paper](#)]
- Learning from Explanations with Neural Execution Tree
Ziqi Wang*, Yujia Qin*, Wenxuan Zhou, **Jun Yan**, Qinyuan Ye, Leonardo Neves, Zhiyuan Liu, Xiang Ren
In *Proceedings of ICLR'20* [[Paper](#)]
- Learning Dual Retrieval Module for Semi-Supervised Relation Extraction
Hongtao Lin, **Jun Yan**, Meng Qu, Xiang Ren
In *Proceedings of TheWebConf'19* [[Paper](#)]

Machine Reasoning

- Learning Contextualized Knowledge Structures for Commonsense Reasoning
Jun Yan, Mrigank Raman, Aaron Chan, Tianyu Zhang, Ryan Rossi, Handong Zhao, Sungchul Kim, Nedim Lipka, Xiang Ren
In *Findings of ACL'21* [[Paper](#)]
- Scalable Multi-Hop Relational Reasoning for Knowledge-Aware Question Answering
Yanlin Feng*, Xinyue Chen*, Bill Yuchen Lin, Peifeng Wang, **Jun Yan**, Xiang Ren
In *Proceedings of EMNLP'20* [[Paper](#)]

Language Modeling

- Language Modeling with Sparse Product of Sememe Experts
Yihong Gu*, **Jun Yan***, Hao Zhu*, Zhiyuan Liu, Ruobing Xie, Maosong Sun, Fen Lin, Leyu Lin
In *Proceedings of EMNLP'18* [[Paper](#)]

HONORS AND AWARDS

- Annenberg Fellowship, University of Southern California. 2019.
- Excellent Graduate, Tsinghua University. 2019.
- Samsung/JJWorld/Evergrande Scholarship, Tsinghua University. 2016/2017/2018.

SERVICES

- PC/Reviewer: ACL Rolling Review (ARR), ACL, EMNLP, NAACL, NeurIPS, IEEE TNNLS.

KEY SKILLS

Programming Languages	Python, C, C++, Java, MATLAB
Machine Learning Libraries	PyTorch, Scikit-Learn, NumPy, Pandas