Jun Yan

yanjun@usc.edu junyann.github.io

RESEARCH INTERESTS

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

EDUCATION

University of Southern California — **Los Angeles, CA, U.S.**
Ph.D. in Computer Science — *08/2019 - present*
- Advisor: Prof. Xiang Ren

Tsinghua University — **Beijing, China**
B.Eng. in Electronic Engineering — *08/2015 - 07/2019*
- Advisor: Prof. Zhiyuan Liu

PUBLICATIONS AND PREPRINTS

(* indicates equal contribution)

Alignment and Safety

- 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
  arXiv:2307.16888 [Preprint]

- 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
  To appear in *Proceedings of ICLR'24* [Paper]

- How Susceptible are Large Language Models to Ideological Manipulation?
  Kai Chen, Zihao He, **Jun Yan**, Taiwei Shi, Kristina Lerman
  arXiv:2402.11725 [Preprint]

- Test-Time Backdoor Mitigation for Black-Box Large Language Models with Defensive Demonstrations
  Wenjie Mo, Jiashu Xu, Qin Liu, Jiongxiao 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
  arXiv:2307.10558 [Preprint]

- 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 Dual Retrieval Module for Semi-Supervised Relation Extraction
  Hongtao Lin, **Jun Yan**, Meng Qu, Xiang Ren
  In *Proceedings of TheWebConf'19* [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]

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]

WORK EXPERIENCE

Research Intern @ Samsung Research America — **Mountain View, CA, U.S.**
Knowledge and Dialogue Team — *05/2023 – 08/2023*

- Mentors: Vikas Yadav, Shiyang Li
- Projects: Data Poisoning Attack on Instruction Tuning; Quality-Guided Data Filtering for Efficient Instruction Tuning; Instruction-Following Evaluation for LLMs

Research Intern @ Meta — **Seattle, WA, U.S.**
Language Understanding and Question Answering Team — *05/2022 – 08/2022*

- Mentors: Asish Ghoshal, Scott Wen-tau Yih, Asli Celikyilmaz, Pedro Rodriguez
- Project: Generalizable Parameter-Efficient Finetuning for Natural Language Generation

Applied Scientist Intern @ Amazon — **Seattle, WA, U.S. (Remote)**
Product Graph Team — *06/2020 – 11/2020*

- Mentors: Nasser Zalmout, Yan Liang, Xin Luna Dong
- Project: Multi-Attribute Value Extraction from Product Profiles

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, IEEE TNNLS.

KEY SKILLS

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