

# Jun Yan

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

## RESEARCH INTERESTS

---

Large Language Model (LLM) Agents, Robustness, and Safety

## PROFESSIONAL EXPERIENCE

---

**Research Scientist @ Google** **Sunnyvale, CA, U.S.**  
*Cloud AI Research Team* *07/2024 – Current*

- Conduct research on building LLM agents and improving LLM trustworthiness.
- Improve Gemini agentic capabilities and robustness.

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

- Formulated data poisoning attacks and studied their defenses for LLM instruction tuning.
- Developed quality-guided data selection for effective and efficient instruction tuning.
- Developed an instruction-following evaluation protocol based on verbalizer manipulation.

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

- Developed parameter-efficient tuning with adaptive capacity for improved generalization.

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

- Developed a multi-task model to enable multi-attribute value extraction from product profiles.

**Research & Teaching Assistant @ USC** **Los Angeles, CA, U.S.**  
*Intelligence and Knowledge Discovery (INK) Lab* *08/2019 – 05/2023*

- Conducted research on interpretable reasoning, model robustness, and safety vulnerabilities.

## EDUCATION

---

**University of Southern California** **Los Angeles, CA, U.S.**  
*Ph.D. in Computer Science* *08/2019 - 08/2024*

- 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)

## Technical Report

- Gemini 2.5: Pushing the Frontier with Advanced Reasoning, Multimodality, Long Context, and Next Generation Agentic Capabilities  
Gemini Team, Google  
*arXiv:2507.06261* [[Preprint](#)]

## LLM Agents

- In Prospect and Retrospect: Reflective Memory Management for Long-term Personalized Dialogue Agents  
Zhen Tan, **Jun Yan**, I-Hung Hsu, Rujun Han, Zifeng Wang, Long T. Le, Yiwen Song, Yanfei Chen, Hamid Palangi, George Lee, Anand Iyer, Tianlong Chen, Huan Liu, Chen-Yu Lee, Tomas Pfister  
In *Proceedings of ACL'25* [[Paper](#)]
- Magnet: Multi-turn Tool-use Data Synthesis and Distillation via Graph Translation  
Fan Yin, Zifeng Wang, I-Hung Hsu, **Jun Yan**, Ke Jiang, Yanfei Chen, Jindong Gu, Long T. Le, Kai-Wei Chang, Chen-Yu Lee, Hamid Palangi, Tomas Pfister  
In *Proceedings of ACL'25* [[Paper](#)]

## 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  
In *Findings of NAACL'25* [[Paper](#)]
- 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

---

- Best Paper Runner-up at ICLR 2024 Workshop on Secure and Trustworthy Large Language Models (SeT LLM), 2024.
- Annenberg Fellowship, University of Southern California. 2019.
- Excellent Graduate, Tsinghua University. 2019.
- Samsung/JJWorld/Evergrande Scholarship, Tsinghua University. 2016/2017/2018.

## SERVICES

---

- Action Editor: ACL Rolling Review.
- PC/Reviewer: ACL Rolling Review, ACL, EMNLP, NAACL, NeurIPS, IEEE TNNLS.

## KEY SKILLS

---

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