Qiusi Zhan

(+1) 217-607-4841 | qiusiz2@illinois.edu | zqs1943.github.io | ZQS1943

Summary

First-year Ph.D. student at the University of Illinois Urbana-Champaign, with a broad research interest in Natural Language Processing, currently centered on making large language models more robust and trustworthy.

Looking for research internship opportunities for 2024 summer.

Education

University of Illinois Urbana-Champaign — *Illinois, U.S.A.*

PH. D. IN COMPUTER SCIENCE — *Aug. 2023 - May 2027 (Expected)*

- Advisor: Prof. Daniel Kang

University of Illinois Urbana-Champaign — *Illinois, U.S.A.*

M.ENG. IN ELECTRICAL & COMPUTER ENGINEERING — *Aug. 2021 - Dec. 2022*

- Advisor: Prof. Heng Ji

Peking University — *Beijing, China*

B.S. IN COMPUTER SCIENCE — *Sept. 2017 - July 2021*

- Advisor: Prof. Sujian Li

Publications

InjecAgent: Benchmarking Indirect Prompt Injections in Tool-Integrated Large Language Model Agents
Qiusi Zhan, Zhixiang Liang, Zifan Ying, Daniel Kang
arXiv preprint, 2024

LLM Agents can Autonomously Hack Websites
Richard Fang, Rohan Bindu, Akul Gupta, Qiusi Zhan, Daniel Kang
arXiv preprint, 2024

Removing RLHF Protections in GPT-4 via Fine-Tuning
Qiusi Zhan, Richard Fang, Rohan Bindu, Akul Gupta, Tatsunori Hashimoto, Daniel Kang
NAACL, 2024

GLEN: General-Purpose Event Detection for Thousands of Types
Qiusi Zhan*, Sha Li*, Kathryn Conger, Martha Palmer, Heng Ji, Jiawei Han
Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, 2023, Singapore

User Simulator Assisted Open-ended Conversational Recommendation System
Qiusi Zhan, Xiaojie Guo, Heng Ji, Lingfei Wu
Proceedings of the 5th Workshop on NLP for Conversational AI (NLP4ConvAI 2023), 2023, Toronto, Canada

EA^{2}E: Improving Consistency with Event Awareness for Document-Level Argument Extraction
Qi Zeng*, Qiusi Zhan*, Heng Ji
Findings of the Association for Computational Linguistics: NAACL 2022, Seattle, WA, United States, July 10-15, 2022, 2022

ConFiguRe: Exploring Discourse-level Chinese Figures of Speech
Dawei Zhu, Qiusi Zhan, Zhejian Zhou, Yifan Song, Jiebin Zhang, Sujian Li
Proceedings of the 29th International Conference on Computational Linguistics, COLING 2022, Gyeongju, Republic of Korea, October 12-17, 2022, 2022

Academic Experience

University of Illinois Urbana-Champaign — *Illinois, U.S.A.*

RESEARCH ASSISTANT. ADVISOR: PROF. DANIEL KANG — *Aug. 2023 - Present*

- Constructed a benchmark for indirect prompt injections in tool-integrated LLM agents.
- Built LLM agents that can autonomously hacking websites.
- Removed the RLHF protections in GPT-4 with a 95% success rate by fine-tuning at low costs.

University of Illinois Urbana-Champaign — *Illinois, U.S.A.*

RESEARCH ASSISTANT. ADVISOR: PROF. HENG JI — *Sept. 2021 - June 2023*

- Created a wide-coverage event detection dataset, with 205K event mentions covering 3,465 types. Developed a novel multi-stage event detection model.
- Enhanced argument consistency in document-level event extraction by training the model to consider event relations during both the training and inference stages.

Peking University *Beijing, China*

Research Assistant. Advisor: Prof. Sujian Li *Oct. 2020 - May 2021*

- Created a novel benchmark for the identification of Chinese figures of speech

University of California Santa Barbara *California, U.S.A.*

Research Assistant. Advisor: Prof. Xifeng Yan *July 2020 - Sept. 2020*

- Improved the robustness of text-to-SQL systems by augmenting the training data using conditional paraphrasing

Industrial Experience

JD.com Silicon Valley Labs *U.S.A.*

Research Scientist Intern *Jan. 2022 - May 2022*

- Developed a user simulator (US) for interacting with a Conversational Recommendation System (CRS) based on user preferences. Applied reinforcement learning to enhance open-ended CRS training through CRS-US interactions.

ByteDance *Beijing, China*

Applied Scientist Intern *Apr. 2021 - July 2021*

- Developed an effective AI system with sequence labeling and information extraction to solve K12 math problems.

Skills

Programming Languages Python, C/C++, Java, JavaScript, SQL, HTML, MATLAB

Frameworks & Toolkits: PyTorch, PyTorch Lighting, Scikit-learn, Pandas, Numpy, Transformers, spaCy