Yichen Wang

yichenwg@gmail.com • yichenzw.com

Education

Xi'an Jiaotong University (XJTU) — **Xi'an, China**
2020 - 2024 (Expected)

- B.S. in Computer Science
- **Computer Science Honors Program** (elite class for top 5% CS students)
- Overall GPA: 89.2/100 (**Ranking: 3rd/45**) , Junior Year: **95.1/100 (Ranking: 1st/45)**
- Outstanding Student (top 3%), Outstanding Undergrad Scholarship
- **Research Assistant, NLP**, Advisor: Prof. Chao Shen and Prof. Xiaoming Liu

University of California, Berkeley — **Berkeley, CA, US**
Jan. 2023 - June 2023

- Exchange Program in Computer Science, **GPA: 4.0/4.0**
- **A+** in CS288 NLP (Grad Level), A in CS188 Intro to AI
- **Intern in the Berkeley NLP Group**, Advisor: Prof. Dan Klein and Ph.D. Kevin Yang

Research Interests

Primarily focused on **machine-generated text detection**, **automatic evaluation** of semantics, and **controlled generation and planning** within the broader **NLP** field. I am also actively working on **AI safety concerns**, including **watermark** generation and **robustness** to attacks. Beyond these, I hold a keen interest in **agents**, and evaluation and enhancement for **human-LLM interaction**, such as **RLHF** in **alignment**.

Publications

Improving Pacing in Long-Form Story Planning

- **Yichen Wang**, Kevin Yang, Xiaoming Liu, Dan Klein. [paper link]
- ***EMNLP 2023*** *Findings*. Done while at UC Berkeley.

CoCo: Coherence-Enhanced Machine-Generated Text Detection Under Low Resource With Contrastive Learning

- Xiaoming Liu*, Zhaohan Zhang*, **Yichen Wang*(Equal Contribution)**, Hang Pu, Yu Lan, Chao Shen. [paper link]
- ***EMNLP 2023***.

Comparing Robustness of Machine-Generated Text Detectors Under Attacks: A Comprehensive Study

- **Yichen Wang***, Tianxing He*, Abe Bohan Hou, Xiao Pu, Shangbin Feng, Chao Shen, Xiaoming Liu, and Yulia Tsvetkov.
- In submission to ***ACL Rolling Review***.

Dialogue for Prompting: a Policy-Gradient-Based Discrete Prompt Optimization for Few-shot Learning

- Chengzhengxu Li, Xiaoming Liu, **Yichen Wang**, Yu Lan, Chao Shen. [paper link]
- ***AAAI 2024***.

DaCo: Few-Shot Machine-Generated Text Detection via Data-Augmentation Contrastive Learning

- Shengchao Liu*, **Yichen Wang*(Equal Contribution)**, Zehua Cheng, and Xiaoming Liu.
- In submission to ***ACL Rolling Review***.

SemStamp: A Semantic Watermark with Paraphrastic Robustness for Text Generation

- Abe Bohan Hou*, Jingyu Zhang*, Tianxing He*, **Yichen Wang**, Yung-Sung Chuang, Hongwei Wang, Lingfeng Shen, Benjamin Van Durme, Daniel Khashabi, and Yulia Tsvetkov. [paper link]
- In submission to ***ACL Rolling Review***.

Research Internships

University of Washington — **Seattle, WA, US**
June 2023 - Present

- **Research Intern**, TsvetShop Group of Paul G. Allen School of CSE
- Advisor: Prof. Yulia Tsvetkov and Dr. Tianxing He, *Paul G. Allen School of CSE*

University of Cambridge — **Cambridge, UK**
May 2022 - Nov. 2022

- **Research Intern** (remote), AI, Dept. of Computer Science and Technology
- Straight A's in Final Assessment (**1st/8**), Advisor: Prof. Pietro Lio', *AI Group*

Research Experience

Machine-Generated Text Detection with Coherence Graph Representation
Advisor: Prof. Xiaoming Liu, Prof. Chao Shen, EECS, XJTU *April 2022 - Feb. 2023*

- Proposed a novel machine-generated text detection framework including a graph-based coherence representation and a novel supervised contrastive learning method to handle low data situations, which I identified as a real-world scenario.
- Created news-type detection-task datasets using GPT3.5-DaVinci, simulating the versatility of origins and writing styles.
- Evaluated the detector's robustness to perturbation attacks; interpreted detection mechanisms using integrated gradients.
- Experiments revealed that newer LLMs (e.g., ChatGPT) were more detectable than older models, providing insights into the evolution of text generation technologies (follow-up research ongoing via the perspective of diversity).
- Independently led the project, proposed ideas in coherence, and took charge of coding (w/o any code base), conducting all experiments, dataset creation, graph evaluation, robust tests, and manuscript co-authoring and rebuttals.

Pacing: Enhancing Long-Form Story Planning
Advisor: Ph.D. Kevin Yang, Prof. Dan Klein, The Berkeley NLP Group *Jan. 2023 - June 2023*

- Developed a concreteness evaluator (a reference-free zero-shot model) to evaluate pacing (i.e., level of concreteness).
- Compiled hierarchical summarization datasets of stories using GPT3.5, and proposed a dynamic pairing training procedure.
- Devised a Concrete Outline Control system (Concoct), integrating a vague-first expansion strategy with a concrete children generation pipeline, improving the structure and flow of downstream story generation.
- Conducted human evaluations on both the outlines and the generated story excerpts. Found that Concoct's outputs were at least 57% more preferable in terms of pacing quality compared to baseline models, indicating a significant improvement.

Toward Robust Detection: A Benchmark on Machine-Generated Text Detection with Attacks
Advisor: Dr. Tianxing He, Prof. Yulia Tsvetkov, Paul G. Allen School of CSE *Work in process*

- Developed a comprehensive benchmark encompassing 10 machine-generated text detectors across 5 categories.
- Introduced a novel composite metric designed to assess attacks, facilitating a more equitable comparison of attack strategies.
- Conducted an extensive study on the robustness of detectors against 15 distinct attack methods imperceptible to humans. These methods spanned a range of tactics, including editing, paraphrasing, prompting, sampling, and fine-tuning attacks.
- Identified and documented key weaknesses in current detection systems and explored initial interpretation.
- Proposed some initial out-of-the-box defense strategies and corrective measures against these attacks.
- The proposal I helped write based on this project is awarded the **Postdoc Research Award, UW (with Dr. Tianxing He)**.

Watermark on Large Language Model via Semantic Parsing Graph
Advisor: Dr. Tianxing He, Prof. Yulia Tsvetkov, Paul G. Allen School; Prof. Xiaoming Liu, XJTU *Work in process*

- Proposed a sentence-level semantic watermark, demonstrating superior paraphrastic robustness. (Submitted to ARR)
- Further enhanced approach by encoding watermarks into texts using the geometric structure of Semantic Parsing Graphs. Planned to leverage the selection and sequencing of entities and the interaction between sentences as fingerprints.
- Aimed to improve watermark robustness against editing attacks, synonymous substitution, and paraphrasing.

Reinforcement Learning Discrete Prompt Optimization with Dialogue
Advisor: Prof. Xiaoming Liu, Prof. Chao Shen, EECS, XJTU *April 2023 - Sep. 2023*

- Proposed an efficient prompt-set generation method based on pre-selected sample candidates to reduce human involvement; Aligned the characteristics of prompts with the few-shot sample set by designing a multi-round dialogue system.
- Utilized reinforcement learning to optimal prompt, matching each individual input with the most suitable prompt.

Graph-based Semantic Representation and Graph Attention Network
Advisor: Prof. Pietro Lio', Ph.D. Kehai Qiu, University of Cambridge *May 2022 - Nov. 2022*

- Proposed a framework of semantic relation-aware graph attention network (GAN) for text classification.
- Introduced static feature analysis of complex networks to our graph-based text representation to evaluate distinguishability.

Out-of-Distribution Text Detection in the Open World
Advisor: Prof. Xinyu Dai, NLP, Nanjing Univ. & Postdoc O.Yawen, Tsinghua Univ. *May 2022 - Sep. 2022*

- Conducted a survey on existing out-of-distribution (OOD) detection methods, followed by a meticulous reproduction.
- Constructed an architecture for the Stream Emerging New Class Problem, grounded in class-incremental learning.
- Received the **Best Project Award (1st out of 12 projects)** for 2022.

Fast Question Answering via Novel Local Sensitive Hashing Sketch
Advisor: Prof. Pinghui Wang, Prof. Jing Tao, EECS, XJTU *Mar. 2021 - Feb. 2022*

- Approached question answering by framing it as a retrieval problem, aimed at a sublinear-time algorithm, inspired by the group's previous work on '*Bidirectionally Densifying LSH Sketches with Empty Bins*' presented at SIGMOD 2021.
- Led the application of sign random projections and local sensitive hashing techniques within the ColBERT framework.
- Earned the annual **Rising Star Undergrad Researcher Award (top 3%)** for 2021 .

Honors & Awards

- **Postdoc Research Award 2023 (with Dr. Tianxing He)**, *UW*. *Nov. 2023*
- **Dr. Shidi Lu Scholarship (3/6100)**, *XJTU*. *Sep. 2023*
- **Social Enterprise Scholarship (first-class)**, *RaySilicon & XJTU*. *Sep. 2022*
- **Best Project Award (1/12)**, *NLP Group, Nanjing University*. *Sep. 2022*
- **Rising Star Undergrad Researcher (top 3%)**, *Lab for Intelligent Networks, XJTU*. *Jan. 2022*
- **Best Developer Award**, *Computer Network Association, XJTU*. *Dec. 2021*

Leadership & Community Engagement

- **Conference Reviewer: EMNLP23, AAAI24, ARR24 Apr.-Oct.** *2023 - Present*
- **Presenter**, *Chinese Academy of Sciences & Beijing Academy of Artificial Intelligence*. *Dec. 2023*
- **Product Manager, AI Service Dialogue System**, *Computer Network Association, XJTU*. *2021 - 2022*
- **Volunteer Leader, Bridge China Program**, *Bridge China Charitable Foundation, Hong Kong*. *2021 - 2022*
- **Organizer, Volunteer Service Center**, *XJTU*. *2020 - 2022*

Skills

- **Programming:** Python (PyTorch), C++, C, Java, MATLAB, JavaScript, HTML, NODE, RISC_V
- **Language:** Mandarin (native), English (advanced), German and Japanese (elementary)