

Zixiang Xu

CONTACT INFORMATION

✉ zixiangx@usc.edu
🌐 github.com/xzx34
📧 x.com/xzx_34
in linkedin.com/in/zixiangx34
G Google Scholar
🏠 xzx34.github.io

Ph.D. Student
Computer Science
University of Southern California
Los Angeles, CA

EDUCATION

University of Southern California (USC)
Ph.D. Student in Computer Science · Advisor: Prof. Yue Zhao
Expected graduation: 2030.

Los Angeles, CA
Aug. 2026 – Present

RESEARCH EXPERIENCE

Tencent (WeChat AI)
Research Intern · Foundation Model Group
Post-trained large foundation models through supervised fine-tuning and alignment.

Beijing, China
Jun. 2026 – Aug. 2026

Alibaba Group
Research Intern

Beijing, China
Mar. 2026 – May 2026

Investigated representation-level mechanisms behind LLM-as-a-judge scoring bias and led a mechanistic interpretability study.

MBZUAI

Research Intern · Supervisor: Prof. Xiuying Chen

Abu Dhabi, UAE
Mar. 2025 – Aug. 2025

Led research on trustworthy LLMs and AI agents, including multilingual robustness, social reasoning, and dynamic agentic workflows; contributed to ACL, NeurIPS, and EMNLP papers.

MINE Lab, University of Notre Dame

Research Intern · Supervisor: Prof. Xiangliang Zhang

Remote
Oct. 2024 – Aug. 2025

Co-led research on robust reasoning, contextual reliability, and multilingual evaluation, resulting in publications at ACL and NeurIPS.

MIT Center for Transportation & Logistics

Summer Research Student · Supervisor: Prof. Elenna R. Dugundji

Massachusetts, USA
Aug. 2024

Led Temporal-IRL on port congestion and berth scheduling; selected for a TRB 2025 oral presentation.

Microsoft (Asia) Internet Engineering Institute

Research Intern · Supervisor: Wang Pan

Beijing, China
Jan. 2024 – Mar. 2024

Evaluated and analyzed LLMs and vision-language models, including GPT-4V assessment and capability reproduction.

HONORS AND AWARDS

USC Annenberg Fellowship

A competitive fellowship top-off awarded to selected Ph.D. students in the USC Viterbi School of Engineering.

2026

CCF Elite Collegiate Award

One of 100 junior and senior undergraduates selected nationwide; recognized for academics, research, and service. Invited speaker/session host at CNCC; team placed **3rd nationwide** in the CCSP Team Contest.

2024

Gold Medal, The 2022 ICPC Asia Jinan Regional Contest

Gold medal in the ICPC, the world's premier collegiate algorithmic programming competition.

2022

Silver Medal, National Olympiad in Informatics (NOI)

Silver medal in China's national informatics olympiad for secondary-school students, organized by the China Computer Federation (CCF).

2020

PUBLICATIONS

[1] SocialMaze: A Benchmark for Evaluating and Enhancing Social Reasoning in Large Language Models in Complex Social Environments

Zixiang Xu, Yanbo Wang, Yue Huang, Haomin Zhuang, Yujun Zhou, Jiayi Ye, et al.

Findings of EMNLP 2026; previously SocialSim @ **COLM 2025** Workshop (Spotlight).

[Project] [Paper] [Code] [Data]

[2] Cross-Lingual Pitfalls: Automatic Probing Cross-Lingual Weakness of Multilingual Large Language Models

Zixiang Xu*, Yanbo Wang*, Yue Huang*, Xiuying Chen, Jieyu Zhao, Meng Jiang, Xiangliang Zhang (*Equal contribution).

ACL 2025. [Paper] [Code]

- [3] **Adaptive Distraction: Probing LLM Contextual Robustness with Automated Tree Search**
Yanbo Wang*, **Zixiang Xu***, Yue Huang*, Chujie Gao, Siyuan Wu, Jiayi Ye, et al. (*Equal contribution).
NeurIPS 2025. [Paper] [Code]
- [4] **DyFlow: Dynamic Workflow Framework for Agentic Reasoning**
Yanbo Wang, **Zixiang Xu**, Yue Huang, Xiangqi Wang, Zirui Song, Lang Gao, Chenxi Wang, et al.
NeurIPS 2025. [Paper] [Code]
- [5] **Inside the Unfair Judge: A Mechanistic Interpretability Account of LLM-as-Judge Bias**
Zixiang Xu, Sixian Li, Huaxing Liu, Xiang Wang, Shuai Li, Zirui Song, Xiuying Chen.
Preprint; currently under review. [Project] [Paper]
- [6] **GTA: Graph Theory Agent and Benchmark for Algorithmic Graph Reasoning with LLMs**
Zixiang Xu, Yanbo Wang, Chenxi Wang, Lang Gao, Zirui Song, Yue Huang, Zhaorun Chen, Xiangliang Zhang, Xiuying Chen.
Preprint, 2026. [Code]
- [7] **Do LLMs “Feel”? Emotion Circuits Discovery and Control**
Chenxi Wang, Yixuan Zhang, et al., **Zixiang Xu**, et al.
ICML 2026. [Paper] [Code]
- [8] **Audio Jailbreak: An Open Comprehensive Benchmark for Jailbreaking Large Audio-Language Models**
Zirui Song, Qian Jiang, et al., **Zixiang Xu**, et al.
ACL 2026. [Paper] [Code]
- [9] **ManipLVM-R1: Reinforcement Learning for Reasoning in Embodied Manipulation with Large Vision-Language Models**
Zirui Song, Guangxian Ouyang, et al., **Zixiang Xu**, et al.
AAAI 2026. [Paper]
- [10] **Temporal-IRL: Modeling Port Congestion and Berth Scheduling with Inverse Reinforcement Learning**
Yikuan Hu*, **Zixiang Xu***, Wei Zhang*, et al. (*Equal contribution).
TRB 2025 (Oral).
- [11] **Under the Shadow of Babel: How Language Shapes Reasoning in LLMs**
Chenxi Wang, Yixuan Zhang, Lang Gao, **Zixiang Xu**, et al.
EMNLP 2025 Findings. [Paper]
- [12] **GenUDC: High-Quality 3D Mesh Generation with Unsigned Dual Contouring Representation**
Ruowei Wang, Jiaqi Li, et al., **Zixiang Xu**, et al.
ACM MM 2024. [Paper]
- [13] **Beyond Survival: Evaluating LLMs in Social Deduction Games with Human-Aligned Strategies**
Zirui Song, Yuan Huang, Junchang Liu, et al., **Zixiang Xu**, et al.
arXiv preprint, 2025. [Paper]
- [14] **Evaluate Bias without Manual Test Sets: A Concept Representation Perspective for LLMs**
Lang Gao, Kaiyang Wan, Wei Liu, et al., **Zixiang Xu**, et al.
arXiv preprint, 2025. [Paper]
- [15] **MTFusion: Reconstructing Any 3D Object from Single Image Using Multi-word Textual Inversion**
Yu Liu, Ruowei Wang, Jiaqi Li, **Zixiang Xu**, Qijun Zhao.
PRCV 2024 (Oral). [Paper]
- [16] **Text-Based Face Retrieval: Methods and Challenges**
Yuchuan Deng, Qijun Zhao, Zhanpeng Hu, **Zixiang Xu**.
CCBR 2023 (Oral).
- Conferences:** AAAI 2027; COLM 2026; ICLR 2026; NeurIPS 2026; NLPCC 2025.
- ACL Rolling Review:** February and May 2025 cycles.
- Workshops:** GroundLM and REALM at EMNLP 2026; Social Simulation with LLMs at COLM 2025.