

Yunze Xiao

☎ 4123200250 | ✉ lyxiao@cmu.edu | 📄 algoxyolo | 🌐 yunzexiao | 🐦 @lrzneedresearch

Education

Carnegie Mellon University

Doha, Qatar

BS COMPUTER SCIENCE

July 2022 - May 2025

- Advisor: **Houda Bouamor**
- Minor in Computational Ethics
- GPA 3.2/4.0
- **Research Interest:** Computational Social Science, Natural Language Processing

Carnegie Mellon University

Pittsburgh, PA

MS LANGUAGE TECHNOLOGY

Aug. 2025 - Aug. 2026

- GPA 4.08/4.33
- Advisor: **Mona Diab**
- **Research Interest:** Natural Language Processing, Human-AI interaction, Anthropomorphism

Research Experience

Carnegie Mellon University

Pittsburgh, PA

ADVISOR: PROF. MONA DIAB

Jan 2025 - Present

- Co-authored two position papers proposing theory-grounded evaluation frameworks for LLMs: (1) a four-part taxonomy for culture in benchmarks (knowledge, preference, performance, bias), and (2) a multi-level framework operationalizing anthropomorphism via perceptual, linguistic, behavioral, and cognitive cues.
- Audited 21 cultural benchmarks to identify six recurring design pitfalls, and proposed actionable repair guidelines to improve construct validity and reduce spurious correlations.
- Outlined cross-disciplinary recommendations for participatory, context-sensitive benchmark design and cue-capability alignment, with emphasis on evaluation protocols that better reflect real user-facing interactions.

Singapore University of Technology and Design

Singapore

ADVISOR: PROF. ROY KA-WEI LEE

May 2024 - Jul 2024

- Developed robustness-oriented evaluation for Chinese offensive language detection under homophone and emoji perturbations, targeting real-world adversarial and obfuscation attacks.
- Implemented and evaluated multiple LLM and classifier baselines with macro-F1, emphasizing controlled comparisons and reproducible evaluation protocols.
- Quantified degradation under perturbations and surfaced concrete model limitations, motivating data augmentation and defense strategies for more reliable deployment.

Qatar Computing Research Institute

Doha, Qatar

ADVISOR: DR. FIROJ ALAM

May 2023 - Oct 2023

- Contributed to the ArAIEval 2023 shared task on Arabic propaganda and disinformation detection, focusing on robust evaluation under domain shift and class imbalance.
- Fine-tuned BERT-based transformer models and benchmarked GPT-4 in zero-shot and few-shot settings, comparing error patterns and performance tradeoffs across modeling choices.
- Ran systematic ablations and hyperparameter sweeps to improve macro-F1 on imbalanced data, and documented practical guidance for stable training and evaluation.

Carnegie Mellon University Qatar

Doha, Qatar

ADVISOR: PROF. HOUDA BOUAMOR

May 2023 - Oct 2023

- Led a structured survey of Chinese offensive language detection, synthesizing benchmarks, datasets, modeling paradigms, and evaluation practices.
- Analyzed linguistic and cultural factors that challenge existing OLD systems (e.g., implicit toxicity, euphemisms, code-switching), and identified concrete research gaps and evaluation failure modes.
- Designed a labeling task for subversive Chinese phrases and curated a 6k+ entry dataset to support fine-grained analysis and improved detection robustness.

Publications

PUBLISHED

- Yunze Xiao**. 2022. A Transformer-based Attention Flow Model for Intelligent Question and Answering Chatbot. In *the proceeding of 14th International Conference on Computer Research and Development (ICCRD)*, Dec 2022.
- Yunze Xiao**, Firoj Alam. 2023. Nexus at ArAIEval shared task: Persuasion techniques detection: an interdisciplinary approach to identifying manipulative strategies. In *proceeding of The First Arabic Natural Language Processing Conference (ArabicNLP 2023)*, Dec 2023.
- David R. Mortensen*, Valentina Izrailevitch*, **Yunze Xiao**, Hinrich Schütze and Leonie Weissweiler. Verbing Weirds Llamas: Evaluation of English Zero-Derivation in Large Language Models. In *The proceedings of 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*
- Xintao Wang, **Yunze Xiao**, Quan Tu, Yaying Fei, Ziang Leng, Wei Wang, Jianjie Chen, Cheng Li, Rui Xu, Haoran Guo, and Yanghua Xiao. **InCharacter**: Do Role-Playing Agents Accurately Capture Characters' Personalities? In *the proceedings of 62nd Annual Meeting of the Association for Computational Linguistics (ACL 2024)*
- Yunze Xiao***, Yujia Hu*, Kenny Tsu Wei Choo, Roy Ka-wei Lee. ToxiCloakCN: Evaluating Robustness of Offensive Language Detection in Chinese with Homophonic and Emoji Perturbations In *proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP 2024)*
- Jiarui Liu, Yueqi Song, **Yunze Xiao**, Mingqian Zheng, Lindia Tjuatja, Jana Schaich Borg, Mona Diab, Maarten Sap. 2025. Synthetic Socratic Debates: Examining Persona Effects on Moral Decision and Persuasion Dynamics. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP 2025)*.
- Weihao Xuan, Rui Yang, Heli Qi, Qingcheng Zeng, **Yunze Xiao**, Yun Xing, Junjue Wang, Huitao Li, Xin Li, Kunyu Yu, Nan Liu, Qingyu Chen, Douglas Teodoro, Edison Marrese-Taylor, Shijian Lu, Yusuke Iwasawa, Yutaka Matsuo, and Irene Li. 2025. MMLU-ProX: A Multilingual Benchmark for Advanced Large Language Model Evaluation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP 2025)*.
- Yunze Xiao**, Lynnette Hui Xian Ng, Jiarui Liu, Mona T. Diab. 2025. Humanizing Machines: Rethinking LLM Anthropomorphism Through a Multi-Level Framework of Design. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP 2025)*. **Oral**
- Gordon Dai*, **Yunze Xiao***. 2025. Embracing Contradiction: Theoretical Inconsistency Will Not Impede the Road of Building Responsible AI Systems. In *Proceedings of the 39th Conference on Neural Information Processing Systems (NeurIPS 2025)*.
- Yunze Xiao***, Tingyu He*, Lionel Z. Wang*, Yiming Ma, Xingyu Song, Xiaohang Xu, Mona Diab, Irene Li, and Ka Chung Ng. 2026. JiraiBench: A Bilingual Benchmark for Evaluating Large Language Models' Detection of Human Self-Destructive Behavior Content in Jirai Community. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2026)*, Main Conference. **Oral**
- Mai AlKhamissi*, **Yunze Xiao***, Badr AlKhamissi, and Mona T. Diab. 2026. Hire Your Anthropologist! Rethinking Culture Benchmarks Through an Anthropological Lens. In *Findings of the Association for Computational Linguistics: EACL 2026*.
- Shu Yang, Junchao Wu, Xin Chen, **Yunze Xiao**, Xinyi Yang, Derek F. Wong, and Di Wang. 2025. Understanding Aha Moments: from External Observations to Internal Mechanisms. In *Transaction of Association for Computational Linguistics: TACL*.

PRE-PRINTS

- Yunze Xiao**, Houda Bouamor, and Wajdi Zaghouni, 2024. Chinese Offensive Language Detection: Current Status and Future Directions.
- Qingyang Wu, Ying Xu, Tingsong Xiao, **Yunze Xiao**, Yitong Li, Tianyang Wang, Yichi Zhang, Shanghai Zhong, Yuwei Zhang, Wei Lu, Yifan Yang. 2024. Surveying Attitudinal Alignment Between Large Language Models Vs. Humans Towards 17 Sustainable Development Goals.
- Chiyuan Fu*, Lyuhao Chen*, **Yunze Xiao***, Weihao Xuan, Carlos Busso, and Mona Diab. 2026. Sentipolis: Emotion-Aware Agents for Social Simulations. *arXiv preprint arXiv:2601.18027*.
- Weihao Xuan, Qingcheng Zeng, Heli Qi, **Yunze Xiao**, Junjue Wang, and Naoto Yokoya. 2026. The Confidence Dichotomy: Analyzing and Mitigating Miscalibration in Tool-Use Agents. *arXiv preprint arXiv:2601.07264*.
- Zhihao Yuan*, **Yunze Xiao***, Ming Li*, Weihao Xuan, Richard Tong, Mona Diab, and Tom Mitchell. 2026. Towards Valid Student Simulation with Large Language Models. *arXiv preprint arXiv:2601.05473*.

Rui Yang, Huitao Li, Weihao Xuan, Heli Qi, Xin Li, Kunyu Yu, Yingjian Chen, Rongrong Wang, Jacques Behmoaras, Tianxi Cai, Bibhas Chakraborty, Qingyu Chen, Lionel Tim-Ee Cheng, Marie-Louise Damwanza, Chido Dzinotiwei, Aosong Feng, Chuan Hong, Yusuke Iwasawa, Yuhe Ke, Linah Kitale, Taehoon Ko, Jisan Lee, Irene Li, Jonathan Chong Kai Liew, Hongfang Liu, Lian Leng Low, Edison Marrese-Taylor, Yutaka Matsuo, Isheanesu Misi, Yilin Ning, Jasmine Chiat Ling Ong, Marcus Eng Hock Ong, Enrico Petretto, Hossein Rouhizadeh, Abiram Sandralegar, Oren Schreier, Iain Bee Huat Tan, Patrick Tan, Daniel Shu Wei Ting, Junjue Wang, Chunhua Weng, Matthew Yu Heng Wong, Fang Wu, **Yunze Xiao**, Xuhai Xu, Qingcheng Zeng, Zhuo Zheng, Yifan Peng, Douglas Teodoro, and Nan Liu. 2026. Toward Global Large Language Models in Medicine. *arXiv preprint arXiv:2601.02186*.

Presentations

Yunze Xiao, Malak Ibrahim, Madhavi Ganapathiraju. Classical and Modern Language Modeling Techniques for Transmembrane Helix Prediction. Extended Abstract Poster presentation in ISMB 2024

Yunze Xiao. Embracing Contradiction: Theoretical Inconsistency Will Not Impede the Road of Building Responsible AI Systems. CMU FEAT Reading Group

Yunze Xiao. Humanizing Machines: Rethinking LLM Anthropomorphism Through a Multi-Level Framework of Design. MLNLP Seminar

Teaching

F2023 **15-110: Principles of Computing**, Course Assistant

S2024 **15-112: Fundamentals of Programming and Computer Science**, Course Assistant

S2024 **98-031: NLP Ethics in a Nutshell**, Student Instructor

Outreach & Professional Development

SERVICE

2023 **Student Majilis**, Head of Academics

Doha, Qatar

2024 **The Web Conference 2024**, Volunteer

Singapore

2025 **AAAI Student Abstract and Poster Program**, Program Chair Member

Remote

2025 **PersonaNLP 2025 @ NeurIPS 2025**, Lead Organizer

CDMX, Mexico

2026 **LTI Teaching Track Search Committee**, Student Member

Pittsburgh, PA

PEER REVIEW

Workshops: RecSys 2025, AAAI 2026 SAPP, EAACL SRW 2026

Conferences: CSCW 2025, ICWSM 2025, ICLR 2025/2026, ACL 2025, EMNLP 2025, WWW 2026, ICASSP 2026, ACL 2026

Journals: TNNLS, TMLR