

# Chenxi Wang

[chenxi.wang@mbzuai.ac.ae](mailto:chenxi.wang@mbzuai.ac.ae) | [Google Scholar](#) | [GitHub](#) | [Personal Website](#)

## Education

<b>Mohamed bin Zayed University of Artificial Intelligence (MBZUAI), UAE</b>	Aug 2024 - Jun 2026
MSc in Natural Language Processing	GPA: 95.2/100
<b>Xi'an Jiaotong University (C9 League), China</b>	Sep 2019 - Jul 2023
BEng in Computer Science and Technology	GPA: 87.6/100

## Research Interests

## Personal Website

- My research focuses on **awakening** the knowledge and behaviors in LLMs that do not spontaneously manifest under normal inference. Through interpretability-driven post-training methods, I uncover the internal mechanisms of LLMs and enable user-aligned activation of specific model capabilities at inference time.
- My current work explores personalized AI systems that fulfill human emotional needs and evolve with their users.

## Publications

## Google Scholar

First-/Co-first-author papers:

- [1] **Do LLMs "Feel"? Emotion Circuits Discovery and Control** *Submitted to ARR (October 2025)*  
Chenxi Wang, Yixuan Zhang, Ruiji Yu et al., Gus Xia, Huishuai Zhang, Dongyan Zhao, Xiuying Chen†.
- [2] **Under the Shadow of Babel: How Language Shapes Reasoning in LLMs** *Findings of EMNLP 2025*  
Chenxi Wang, Yixuan Zhang, Lang Gao, Zixiang Xu et al., Xiuying Chen†.
- [3] **Word Form Matters: LLMs' Semantic Reconstruction under Typoglycemia** *Findings of ACL 2025*  
Chenxi Wang, Tianle Gu, Zhongyu Wei, Lang Gao, Zirui Song, Xiuying Chen†.
- [4] **Decoding echo chambers: LLM-powered simulations revealing polarization in social networks** *COLING 2025*  
Chenxi Wang\*, Zongfang Liu\*, Dequan Yang, Xiuying Chen†.
- [5] **Autonomous Agents for Collaborative Task under Information Asymmetry** *NeurIPS 2024*  
Wei Liu\*, Chenxi Wang\*, Yifei Wang, Zihao Xie, Rennai Qiu et al., Chen Qian†.

Selected Co-authored papers:

- [1] **Evaluate Bias without Manual Test Sets: A Concept Representation Perspective for LLMs** *Under review at ICLR 2026*  
Lang Gao, Kaiyang Wan, Wei Liu, Chenxi Wang, Zirui Song, Zixiang Xu, Yanbo Wang, Veselin Stoyanov, Xiuying Chen†
- [2] **When Personalization Tricks Detectors: Feature-Inversion Trap in Machine-Generated Text Detection** *Submitted to ARR (October 2025)*  
Lang Gao, Xuhui Li, Chenxi Wang et al., Preslav Nakov, Xiuying Chen†
- [3] **ManipLVM-R1: Reinforcement Learning for Reasoning in Embodied Manipulation with LVLMs** *Under review at AAAI 2026*  
Zirui Song, Guangxian Ouyang, Mingzhe Li, Yuheng Ji, Chenxi Wang, Zixiang Xu, Zeyu Zhang, Xiaoqing Zhang, Qian Jiang, Zhenhao Chen, Zhongzhi Li, Rui Yan, Xiuying Chen†
- [4] **DyFlow: Dynamic Workflow Framework for Agentic Reasoning** *NeurIPS 2025*  
Yanbo Wang, Zixiang Xu, Yue Huang, Xiangqi Wang, Zirui Song, Lang Gao, Chenxi Wang, Xiangru Tang, Yue Zhao, Arman Cohan, Xiangliang Zhang†, Xiuying Chen†.
- [5] **Data-Efficient Fine-Grained Cross-Cultural Transfer of Commonsense Reasoning in LLMs** *Findings of EMNLP 2025*  
Saeed Almheiri, Rania Elbadry, Mena Attia, Chenxi Wang, Fajri Koto, Timothy Baldwin, Preslav Nakov†.

## Experience

---

### State Key Laboratory of General Artificial Intelligence, Peking University

Beijing, China

Research Intern (Supervisor: Prof. Dongyan Zhao)

Jun 2025 – Sep 2025

- **Research Focus:** Discovering and controlling emotion circuits in LLMs.
- **Responsibilities:** Constructed a controllable dataset; designed and conducted experiments to extract context-agnostic emotion directions and identify neurons and attention heads underlying diverse emotional expressions, integrating them into coherent emotion circuits via causal interventions; achieved controllable emotional modulation in LLMs through these circuits; led the project from idea conception to paper writing.
- **Achievements:** An open-source and generalizable framework, EmotionCircuits-LLM (<https://github.com/Aurora-cx/EmotionCircuits-LLM>), for discovering and controlling emotion circuits; paper submitted to ARR (October 2025).

### THUNLP Lab, Tsinghua University

Beijing, China

Research Intern (Supervisor: Prof. Zhiyuan Liu)

Mar 2024 – Sep 2024

- **Research Focus:** Investigated how LLM-powered agents collaborate under information asymmetry in multi-agent systems by simulating human-like information exchange behaviors.
- **Responsibilities:** Constructed task-specific datasets; implemented multi-agent interaction modules and memory systems; contributed to the InfoNav reasoning mechanism and paper writing.
- **Achievements:** A well-maintained and publicly available open-source project, iAgents (<https://thinkwee.top/iagents/>), supporting real-time agent interaction; the project paper was accepted to NeurIPS 2024.

### Trust & Safety ML Team, Xiaohongshu (RedNote / Little Red Book)

Shanghai, China

NLP Intern

Jul 2023 – Mar 2024

#### **Project 1: Time Series Forecasting for Risk Merchant Alerting**

- Designed a hybrid model combining Transformer-based time series forecasting and binary classification to forecast merchants' violation risk, integrating temporal signals with merchant-specific features.
- Achieved 75% precision and 82% recall; daily alerts successfully flagged over 1,000 merchants.

#### **Project 2: Fraudulent Merchant Identification via XGBoost**

- Built and optimized an XGBoost classifier to identify low-quality merchants; conducted feature selection using MIC and forward search to reduce noise and overfitting; handled severe class imbalance using SMOTE;
- Achieved 85% precision and 80% recall; identified 20,000+ high-risk accounts; deployment in production led to a 39.79% drop in complaint rate within 15 days.

## Selected Honors and Awards

---

Top 10, Xiaohongshu Annual Internal Hackathon (themed on LLM Agents)

Apr 2024

Outstanding Graduate Award (Top 5%), Xi'an Jiaotong University

Jun 2023

Academic Excellence Scholarship (Top 10%, awarded twice), Xi'an Jiaotong University

2021-2022

Outstanding Student Leader Award (awarded twice), Xi'an Jiaotong University

2021-2022

## Technical & Research Skills

---

- **Model Expertise:** Familiar with open-source LLMs (e.g., LLaMA, Qwen); skilled in using interpretability methods to uncover and control internal model mechanisms (e.g., circuits); capable of building multi-agent interaction systems.
- **Programming:** Proficient in Python; skilled in PyTorch and HuggingFace Transformers; strong in data processing, experiment automation, and result visualization; maintain open-source repositories such as [EmotionCircuits-LLM](#).
- **Research Skills:** Experienced in independent research ideation, experimental design, and scientific writing with LaTeX; familiar with publication workflows of top-tier NLP/ML venues (e.g., ACL, NeurIPS).