Lingjie (Jason) Chen

+86 173-173-22856 | ljchen21@m.fudan.edu.cn | https://lingjiechen2.github.io/

Summary

Research Interests: My research primarily focuses on **Trustworthy LLMs** and **Mechanistic Interpretability**. My ultimate goal is to **develop more transparent and controllable LLMs by unifying trustworthiness with mechanistic interpretability**. Currently, I am devising methods that leverage mechanistic interpretability to analyze models' internal structures. I also actively follow advancements in the trustworthy LLM domain to address emerging challenges.

Highlights: 3 years of programming experience; 2 years of research experience in LLMs, with a solid mathematical and practical foundation. Experienced in deploying various LLMs.

Relevant Courses: Artificial Intelligence(A), Neural Network and Deep Learning(A), Natural Language Processing(A), Numerical Linear Algebra(A), Statistical Machine Learning (A), Time Series(A-), Convex Optimization(A), Computer Vision(A-), Computer Architecture(A), Algorithms and Data Structures(A).

Education

Fudan University Sep. 2021 - Jun. 2025 (expected)
B.S. in Data Science *Shanghai, China*

- Major GPA: **3.84/4.0** (ranking **top 5** in the department); Overall GPA: 3.74/4.0

University of California, Berkeley Aug. 2023 - Jan. 2024
Exchange Student, Statistics *California, USA*

Publication

- **[C3] WAPITI: A Watermark for Finetuned Open-Source LLMs.**
  Lingjie Chen*, Ruizhong Qiu*, Siyu Yuan, Zhining Liu, Tianxin Wei, Hyunsik Yoo, Zhichen Zeng, Deqing Yang, Hanghang Tong
  Under review for ICLR 2025. [Paper]

- **[C2] A Mechanistic View of Intrinsic Multilingualism in Large Language Models.**
  Lingjie Chen, Fukang Zhu, Ningyu Xu, Xuyang Ge, Junxuan Wang, Zhengfu He, Xipeng Qiu
  Under review for ACL 2025.

- **[C1] "A good pun is its own reword": Can Large Language Models Understand Puns?**
  Zhijun Xu, Siyu Yuan, Lingjie Chen, Deqing Yang
  EMNLP 2023. [Paper]

Research Experience

IDEA Lab, University of Illinois Urbana-Champaign Apr. 2024 – Oct. 2024
Research topics: Trustworthy LLM, Watermark, Model Intervention *Illinois, USA*
Advisor: Prof. **Hanghang Tong**

- **Watermarking Fine-tuned Large Language Models[C3]**
  - Identified and validated the incompatibility between existing watermarking techniques and fine-tuned models.
  - Proposed a training-free, parameter-based watermarking method with thorough theoretical derivation.
  - Designed experiments to demonstrate the effectiveness and generalizability of our method.
  - Performed an in-depth analysis of our method, offering insights into its effectiveness.

OpenMoss, Fudan University Jan. 2024 – Present
Research topics: Interpretability, Multilingual LLM *Shanghai, China*
Advisor: Prof. **Xipeng Qiu**

- **Exploration of intrinsic and transferred multilingualism[C2]**
  - Synthesized custom datasets to investigate the model's 'thinking state' during multilingual processing.
  - Designed cross-SAE patching experiments to examine the relationships within the feature space of LLMs.

- Explored the internal mechanisms of multilingual models, revealing meaningful internal processes.

Shanghai Key Laboratory of Data Science Dec. 2022 – Dec. 2023
Research topics: Evaluation Methodology, Dataset *Shanghai, China*
Advisor: Prof. **Deqing Yang**

- **Evaluation of Large Language Models for Pun Understanding[C1]**
  - Conducted a systematic evaluation of eight different LLMs' capabilities in three pun-related tasks.
  - Designed and implemented novel pipelines for pun explanation and generation.
  - Improved the state-of-the-art performance of LLMs in pun understanding from 72% to 83%.

Project Portfolio (Selected)

Sparse AutoEncoder Framework Jan. 2024 – Present
Founder & Developer. [Code] *Shanghai, China*

- Provide a general codebase for conducting dictionary-learning-based mechanistic interpretability research
- Provides tools for analyzing and visualizing the learned dictionaries.

BERT-based Chinese QA English Oct. 2023 – Dec. 2023
Founder & Developer. [Code] *California, USA*

- Evalute the BERT-based model's performance on Chinese QA and provide a comparison with SOTA LLMs.

Attendece Checking Miniprogram Mar. 2023 – Jun. 2023
Founder & Developer. [Code] *Shanghai, China*

- Developed a WeChat Mini Program that enables attendance checks for both the teacher and student sides.

Academic Services

Reviewer International Conference on Learning Representations **(ICLR)**, 2024, 2025

Reviewer Empirical Methods in Natural Language Processing **(EMNLP)**, 2024

Honors & Awards (Selected)

National Natural Science Fund for Youth Science (130 recipients in China)	**2024**
Fudan University Scholarship (Top 10%)	**2021-2024**
Sou-Bin Scholarship (Top-performing students in Shanghai)	**2019-2024**
Second Prize, CUMCM	**2024**

Skills

Languages: Mandarin(Native speaker), English(TOFEL L30 R30 W21 S28)
Programming: Python, C/C++, LaTeX, MATLAB, Linux, R, SQL, Bash
Frameworks: Pytorch, Numpy, Anaconda, MySQL, Git, OpenCV