

Lingjie (Jason) Chen

☎ +1(217)-607-3395 | ✉ lingjie7@illinois.edu | 🌐 <https://lingjiechen2.github.io/>

Research Interests: My research focuses on **LLM post-training** and **mechanistic interpretability**. I aim to **develop transparent and controllable LLMs** by bridging trustworthiness with interpretability. Currently, I investigate the traits of reasoning models and their mechanisms of knowledge utilization in large reasoning models (LRMs).

🎓 EDUCATION

University of Illinois Urbana-Champaign Sep. 2025 - Present
Ph.D in Computer Science Urbana, IL

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

• Major GPA: **3.84/4.0**

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

📖 PUBLICATION

- [C5] **On the Role of Reasoning Traces in Large Reasoning Models.**
Yijie Hao*, [Lingjie Chen*](#), Ali Emami, Joyce C. Ho
Submitted to ICLR 2026
- [C4] **VLMs Can Aggregate Scattered Training Patches.**
Zhanhui Zhou, [Lingjie Chen](#), Chao Yang, Chaochao Lu
Neurips 2025 [[Paper](#)]
- [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
Submitted to ICLR 2026
- [C2] **Llama Scope: Extracting Millions of Features from Llama-3.1-8B with Sparse Autoencoders.**
Zhengfu He, Wentao Shu, Xuyang Ge, [Lingjie Chen](#), Junxuan Wang, Yunhua Zhou, Frances Liu, Qipeng Guo, Xuanjing Huang, Zuxuan Wu, Yu-Gang Jiang, Xipeng Qiu
Available on Arxiv [[Paper](#)]
- [C1] **“A good pun is its own reword”: Can Large Language Models Understand Puns?**
Zhijun Xu, Siyu Yuan, [Lingjie Chen](#), Deqing Yang
EMNLP 2024. [[Paper](#)]

🏢 RESEARCH EXPERIENCE

IDEA Lab, University of Illinois Urbana-Champaign Apr. 2024 – Present
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 – July. 2024
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)

DLLM Trainer

Sep. 2025 - Present

Lead Developer. [[Github URL](#)]

- Developed a lightweight training framework tailored for diffusion language models (DLLMs).
- Implemented modular components for noise scheduling, model wrapping, and logging.
- Enabled flexible experimentation with minimal boilerplate for research and prototyping.

Sparse AutoEncoder Framework

Jan. 2024 – July. 2024

Lead Developer. [[HGithub URL](#)]

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

🏛️ ACADEMIC SERVICES

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

Reviewer Annual Meeting of the Association for Computational Linguistics (**ACL**), 2025

Reviewer Empirical Methods in Natural Language Processing (**EMNLP**), 2025

🏆 HONORS & AWARDS (SELECTED)

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 W25 S28)

Programming: Python, Linux, \LaTeX , MATLAB, R, SQL, C/C++

Frameworks: Pytorch, Transformers, DeepSpeed, Accelerate