Zimin **Zhang**

NLP Researcher at UIUC

(+1) 424-440-0696 | zhang110297@gmail.com | ziminz19.github.io | ziminz19

Research Interests

My research focuses on natural language processing, particularly **mechanistic interpretability** and **LLM post-training** for enhancing LLM reasoning. I aim to uncover the internal representations and causal pathways that underlie reasoning behavior in LLMs, and leverage these insights to guide post-training optimization and inference-time scaling. My research investigates how LLMs encode and organize internal knowledge representations and how model components interact to produce reasoning behaviors, while simultaneously developing post-training and test-time scaling methods to enhance reasoning accuracy, robustness, and efficiency.

Education

University of Illinois Urbana-Champaign   *Champaign, IL*
MASTER OF COMPUTER SCIENCE   *Aug. 2024 - Present*

- Graduate researcher under the supervision of Prof. Hao Peng.

University of California, Los Angeles   *Los Angeles, CA*
B.S. IN MATHEMATICS OF COMPUTATION & STATISTICS AND DATA SCIENCE (DOUBLE MAJOR)   *Sep. 2020 - Jun. 2024*

- Undergraduate researcher at UCLA Data Mining Lab and UCLA NLP Group, under the supervision of Prof. Yizhou Sun and Prof. Kai-Wei Chang.

Publications

(*: Equal contributions)

- **The Unreasonable Effectiveness of Entropy Minimization in LLM Reasoning**
  Shivam Agarwal, **Zimin Zhang**, Lifan Yuan, Jiawei Han, Hao Peng
  *In Proceedings of the 39th Annual Conference on Neural Information Processing Systems (**Neurips 2025**)*

- **How Post-Training Reshapes LLMs: A Mechanistic View on Knowledge, Truthfulness, Refusal, and Confidence**
  Hongzhe Du, Weikai Li, Min Cai, Karim Saraipour, **Zimin Zhang**, Himabindu Lakkaraju, Yizhou Sun, Shichang Zhang
  *In Proceedings of the Conference on Language Modeling of 2025 (**COLM 2025**)*

- **Automated Molecular Concept Generation and Labeling with Large Language Models**
  Zimin Zhang*, Qianli Wu*, Botao Xia*, Fang Sun, Ziniu Hu, Yizhou Sun, Shichang Zhang
  *In Proceedings of the 31st International Conference on Computational Linguistics (**COLING 2025**)*

- **OfficeBench: Benchmarking Language Agents across Multiple Applications for Office Automation**
  Zilong Wang, Yuedong Cui, Li Zhong, **Zimin Zhang**, Da Yin, Bill Yuchen Lin, Kai-Wei Chang, Jingbo Shang
  Arxiv Preprint

Research Experience

UIUC Hao's Group   *Champaign, IL*
GRADUATE RESEARCHER (SUPERVISED BY PROF. HAO PENG)   *September 2024 - Present*

- Conducted research on enhancing LLM's reasoning capabilities in math and coding through **entropy minimization–based test-time scaling**.
- Designed **unsupervised reward signals** from LLM's internal confidence and entropy for reinforcement learning to improve model performance without labeled supervision.
- Implemented and conducted LLM Post-training with **Reinforcement Learning** framework using **Verl, PyTorch FSDP, OpenRLHF, and DeepSpeed** for large-scale distributed experiments.

UCLA Data Mining Lab   *Los Angeles, CA*
UNDERGRADUATE RESEARCHER (SUPERVISOR: PROF. YIZHOU SUN)   *September 2023 - February 2025*

- Developed the framework **Automated Molecular Concept Generation**, which leverages the knowledge in Large Language Models to automatically generate molecular concepts that are predictive and label them for each molecule.
- Research on knowledge locating and editing for LLMs to improve model performance and explainability.
- Explored using inverse linear relational embedding to improve LLM's ability in inverse search.

UCLA NLP Group

Los Angeles, CA

RESEARCH ASSISTANT (SUPERVISOR: PROF. KAI-WEI CHANG)

December 2023 - August 2024

- Building a realistic, multi-domain office agent benchmark that evaluates current LLM agents' ability to perform office-related tasks with multiple softwares.
- Developing the framework of office-agent benchmark's testing environment and implementing APIs for agents to interact with multiple office softwares using **Python**.

Industry Experience

China National Software & Services Co.

Beijing, Chiina

MACHINE LEARNING ENGINEER INTERN

June 2023 - September 2023

- Implemented the Knowledge Graph construction page and Entity/Relation extraction page using **Vue.js**.
- Delivered portable pre-trained entity/relation recognition models to clients on different operating systems and computer architectures using **Java** and **Docker**, with 53ms average response time.
- Fine-tuned ChatGLM-6B with personalized dataset using **Huggingface Accelerate** and **DeepSpeed** to achieve 27% increase in recognition accuracy than base model.

Beijing Hiqos System

Beijing, China

MACHINE LEARNING ENGINEER INTERN

January 2021 - May 2021

- Improved and trained YOLOv3 and **Mask R-CNN** with PyTorch on a custom dataset of GUI components; compared models based on mAP and determined Mask R-CNN to be the better choice with 12% higher precision.
- Developed a fully automated software testing application with integrated refined Mask R-CNN model using Java, Python, and Appium API, with 15% higher testing coverage than benchmark.

Projects

MemoMuse

MACHINE LEARNING ENGINEER

June 2024 - April 2025

- **Project Overview:** Designed and developed an AI-powered iOS voice memo app that automatically transcribes and summarizes daily voice memos into organized journals and calendar tasks, helping users connect fragment but related events in their life and stay productive. Collaborated with a 5-person team from concept to deployment.
- Developed an **LLM-driven personal memory system for transcribed voice memos** using **PostgreSQL + pgvector**, indexing over **90K memo entries (0.5M tokens/user, 140 users)** and implementing hybrid semantic–keyword–recency retrieval with **HNSW** vector indexing, achieving **P95 retrieval latency under 300 ms**.
- Developed an intelligent summarization and reminder generation pipeline with **LangChain** and deployed self-hosted **vLLM** inference using **PagedAttention**, enabling **<600 ms TTFT** and **2–4 s** response times for seamless, real-time journaling and personalized reminder experiences.

Skills

ML Framework	PyTorch, Verl, OpenRLHF, DeepSpeed, vLLM, LangChain
Programming	Python, R, Rust, Java, LaTeX
Languages	English, Chinese