Zihan Wang

Homepage: https://zihanwang314.github.io
Email : zihanw@u.northwestern.edu
Mobile : +1-217-308-3456

Education Experience

- **Northwestern University, IL** Sep 2024 - 2029 (expected)
  PhD Candidate
  Advisor: Manling Li
- **Renmin University, China** (Ranked 25th globally on CSRankings 2024) Sep 2020 - Jun 2024
  B.S. Engineering
  Advisor: Zhicheng Dou
- **University of California at Berkeley, CA** Aug 2022 - Dec 2022
  Exchange Student
  GPA: 4.00/4.00 - CS282A Deep Learning, CS 285 Reinforcement Learning*, CS 188 Artificial Intelligence

Research Experience

- **DeepSeek AI, China** Feb 2024 - July 2024
  - **Mentors**: Deli Chen, Damai Dai, Yu Wu.
  - **Topic**: **A.** We leverage sparse architectures to train specialized language models. Near-full performance can be achieved by picking the best 5% experts for specialized LLM tuning. **First-author paper presented at EMNLP 2024**.
    B. We develop DeepSeek-V2, a 236-billion Parameter model using Multihead Latent Attention to compress the LLM efficiency bottleneck KV-cache, cutting costs by 42.5%, boosting generation 5.76x, with **3.5k GitHub stars and 100k users**.
- **University of Illinois at Urbana-Champaign, IL** Jun 2023 - Sep 2023
  - **Advisor**: Heng Ji and Xingyao Wang.
  - **Topic**: We introduce the MINT benchmark comprising eight diverse tasks to evaluate LLMs in multi-turn interactions, focusing on tool use and language feedback. It highlights the misalignment of current LLMs between single-turn and multi-turn capabilities. **Co-first author paper presented at ICLR 2024.**
- **Renmin University, China** (Undergraduate Research) Jun 2021 - Jun 2024
  - **Advisor**: Zhicheng Dou
  - **Topic**: We investigate critical challenges in generative retrieval and retrieval augmented generation. **2 first-author papers at CIKM and CCL** about learning representations like document identifiers (DocID) for generative retrieval, and **1 open-source project with 200 Stars at GitHub** about retrieval augmented generation.

Research Interest

Language Agents:
Interaction with Language Feedback, Retrieval-Augmented Generation, Long-Context Understanding, Scaling Agents

Model Efficiency:
Mixture-of-Experts Models, Parameter-Efficient Fine-Tuning, Sparse Attention Mechanisms, Video Temporal Search

Publication List (Scholar Page)

1. **Zihan Wang**, Deli Chen, Damai Dai, Runxin Xu, Zhuoshu Li, Yu Wu. Let the Expert Stick to His Last: Expert-Specialized Fine-Tuning for Sparse Architectural Large Language Models, **EMNLP 2024**. (Paper and Code. **Star: 140+**)
2. Xingyao Wang*, **Zihan Wang***, Jiateng Liu, Yangyi Chen, Lifan Yuan, Hao Peng, Heng Ji. MINT: Evaluating LLM in Multi-turn Interaction with Tools and Language Feedback, **ICLR, 2024**. (Paper and Code. **Star: 100+**)
3. DeepSeek AI (157 authors including **Zihan Wang**). DeepSeek-V2: A Strong, Economical, and Efficient Mixture-of-Experts Language Model. (Paper and Code. **Star: 3,500+**)
4. **Zihan Wang**, Yujia Zhou, Yiteng Tu, Zhicheng Dou. NOVO: Learnable and Interpretable Document Identifiers for Model-Based IR, **CIKM 2023**. (**Oral presentation**. Paper)
5. Jiongnan Liu, Jiajie Jin, **Zihan Wang**, Jiehan Cheng, Zhicheng Dou, Ji-Rong Wen. RetaLLM: A Retrieval-Augmented Large Language Model Toolkit. (Paper and Code. **Star: 200+**)
6. **Zihan Wang**, Hongjin Qian, Zhicheng Dou. Learning on Structured Documents for Conditional Question Answering, *CCL 2023.* (Paper)

Honors and Awards

- McCormick School of Engineering Fellowship (**$45,000**), *Northwestern University.* 2024
- Outstanding Reviewer, *EMNLP Conference.* 2024
- Baosteel Outstanding Students (**CNY 10,000. 7 among 30,000 students** in RUC). *Baosteel Education Fund.* 2023
- Honorable Awards (Top 10%). *CCF Big Data and Computing Intelligence Contest.* 2022
- Academic Excellence Award (Top 3% GPA). *Renmin University.* 2021
- Merit Student Award. *Renmin University.* 2021
- First Prize. *Contemporary Undergraduate Mathematical Contest in Modeling.* 2021
- Honorable Mention. *Mathematical Contest in Modeling and Interdisciplinary Contest in Modeling.* 2021

Professional Service

- **Reviewer**: EMNLP 2024 (Outstanding Reviewer) , ICLR 2025 (external)
- **Session Organization**: Session of Language Models & Agents in Chinese R Conference 2023, BoF and Affinity Group at EMNLP 2024.
- **Academic Mentor**: National University Student Innovation Program (2023, No. 3)
- **Article Translation:** English (Unveiling DeepSeek: A Story of Even More Radical Chinese Technological Idealism), Chinese (COS Interview with Donald B. Rubin, Core Views on AI Safety: When, Why, What, and How)

Invited Talks and Presentations

LLM Agents with Language Feedback
Chinese R Conference 2023.11

Retrieval Augmented Language Models and Applications
RUC Science and Technology Fair 2023.05

Large Language Models and Applications
Capital of Statistics 2023.03

Pre-trained Language Models and Applications
RUC Mingli College 2023.01