Yujun Zhou

Personal Website | Google Scholar | YujunZhou | yzhou25@nd.edu

Summary

I am a third-year PhD student at the University of Notre Dame starting from Spring 2023, under the supervision of Prof. Xiangliang Zhang, working on Large Language Models (LLMs).

My current research interests center around LLM reasoning and trustworthy LLMs. Broadly, I explore areas including LLM math reasoning [1], LLM logical reasoning [2], LLM safety [3,5], LLM benchmark evaluation [5], and LLM agents [3]. Please refer to my recent publications for additional details.

I am currently a research intern at Tencent America, working on LLM reasoning, especially on math reasoning tasks [1].

I am engaged in the ND-IBM Tech Ethics Lab Collaborative Project, where I aim to extend the trustworthiness of LLMs to real-world safety-critical applications, such as lab safety [3] and misalignment behaviors.

Education

2023 - Present	PhD in Computer Science at **University of Notre Dame, USA**	(GPA: 4.00/4.0)
2021 - 2022	Master's Degree in Computer Science at **KAUST, Saudi Arabia**	(GPA: 3.83/4.0)
2017 - 2021	Bachelor's Degree in Computer Science at **UESTC, China**	(GPA: 3.92/4.0)

Selected Publications

[1] **Yujun Zhou***, Zhenwen Liang*, Haolin Liu, Wenhao Yu, Kishan Panaganti, Linfeng Song, Dian Yu, Xiangliang Zhang, Haitao Mi, Dong Yu. "Evolving Language Models without Labels: Majority Drives Selection, Novelty Promotes Variation". In *arxiv preprint*. arXiv:2509.15194.

[2] **Yujun Zhou***, Jiayi Ye*, Zipeng Ling*, Yufei Han, Yue Huang, Haomin Zhuang, Zhenwen Liang, Kehan Guo, Taicheng Guo, Xiangqi Wang, and Xiangliang Zhang. "Dissecting Logical Reasoning in LLMs: A Fine-Grained Evaluation and Supervision Study". In: ***EMNLP Findings***. 2025.

[3] **Yujun Zhou**, Yufei Han, Haomin Zhuang, Taicheng Guo, Kehan Guo, Zhenwen Liang, Hongyan Bao, and Xiangliang Zhang. "Defending Jailbreak Prompts via In-Context Adversarial Game". In: ***EMNLP***. 2024.

[4] **Yujun Zhou**, Jingdong Yang, Kehan Guo, Pin-Yu Chen, Tian Gao, Werner Geyer, Nuno Moniz, Nitesh V Chawla, and Xiangliang Zhang. "LabSafety Bench: Benchmarking LLMs on Safety Issues in Scientific Labs". Accepted by ***Nature Machine Intelligence***.

[5] **Yujun Zhou***, Yufei Han*, Haomin Zhuang, Hongyan Bao, and Xiangliang Zhang. "Attack-free Evaluating and Enhancing Adversarial Robustness on Categorical Data". In: ***ICML***. 2024.

[6] Zhenwen Liang*, Ruosen Li*, **Yujun Zhou***, Linfeng Song, Dian Yu, Xinya Du, Haitao Mi, Dong Yu. "CLUE: Non-parametric Verification from Experience via Hidden-State Clustering". In *arxiv preprint*. arXiv:2510.01591.

[7] Xiangqi Wang, Yue Huang, Yanbo Wang, Xiaonan Luo, Kehan Guo, **Yujun Zhou**, Xiangliang Zhang. "AdaReasoner: Adaptive Reasoning Enables More Flexible Thinking". In ***Neurips as a Spotlight***. 2025

[8] Yue Huang, Chujie Gao, Siyuan Wu, Haoran Wang, Xiangqi Wang, **Yujun Zhou**, et.al. "On the Trustworthiness of Generative Foundation Models: Guideline, Assessment, and Perspective". In *arXiv preprint*. arXiv:2502.14296.

[9] Yue Huang*, Zhengqing Yuan*, **Yujun Zhou***, Kehan Guo, Xiangqi Wang, Haomin Zhuang, Weixiang Sun, Lichao Sun, Jindong Wang, Yanfang Ye, Xiangliang Zhang. "Exposing and Patching the

Flaws of Large Language Models in Social Character Simulation". In ***COLM***. 2025.

[10] Yicheng Lang, Kehan Guo, Yue Huang, **Yujun Zhou**, Haomin Zhuang, Tianyu Yang, Yao Su, Xiangliang Zhang. "Beyond Single-Value Metrics: Evaluating and Enhancing LLM Unlearning with Cognitive Diagnosis". In ***ACL Findings***. 2025.

[11] Kehan Guo, Bozhao Nan, **Yujun Zhou**, Taicheng Guo, Zhichun Guo, Mihir Surve, Zhenwen Liang, Nitesh V Chawla, Olaf Wiest, and Xiangliang Zhang. "Can LLMs Solve Molecule Puzzles? A Multimodal Benchmark for Molecular Structure Elucidation". In ***Neurips as a Spotlight***. 2024

[12] Zhenwen Liang, Kehan Guo, Gang Liu, Taicheng Guo, **Yujun Zhou**, Tianyu Yang, Jiajun Jiao, Renjie Pi, Jipeng Zhang, and Xiangliang Zhang. "SceMQA: A Scientific College Entrance Level Multimodal Question Answering Benchmark". In: ***ACL***. 2024.

[13] Hongyan Bao, Yufei Han, **Yujun Zhou**, Yun Shen, and Xiangliang Zhang. "Towards understanding the robustness against evasion attack on categorical data". In: ***ICLR***. 2022.

Research Experiences

LLM Reasoning

- **On the Evolution of Language Models without Labels.** We developed EVOL-RL to solve "Entropy Collapse", a core failure mode in label-free RL where models lose reasoning diversity and fail to generalize. Our evolution-inspired framework balances consensus with a novelty reward, achieving state-of-the-art results by boosting pass@1 and pass@16 accuracy, preventing in-domain diversity collapse and improving out-of-domain generalization. (first-authored paper submitted to ICLR 2026)
- **Dissecting Logical Reasoning in LLMs: A Fine-Grained Evaluation and Supervision Study.** We proposed FineLogic, a fine-grained 3-D evaluation of LLM reasoning (accuracy, stepwise soundness, representation alignment) with a comparative evaluation of four supervision formats; we show that natural language supervision generalizes best and offer practical guidance for future directions to improve logical reasoning in LLMs with RL. (first-authored paper in EMNLP Findings 2025)

Trustworthy LLMs

- **Defending against Jailbreak Prompts via an In-Context Adversarial Game.** We propose to apply LLM agent learning to conduct adversarial training for LLMs without fine-tuning to defend against jailbreak attacks. (first-authored paper in EMNLP 2024)
- **ND-IBM Tech Ethics Lab Collaborative Project.** We propose LabSafety Bench, a specialized evaluation framework designed to assess the reliability and safety awareness of LLMs in laboratory environments. This extends the trustworthiness of LLMs to real-world safety-critical applications, such as lab safety, beyond the traditional focus areas. (first-author paper, accepted in principle for publication in Nature Machine Intelligence)

Adversarial Machine Learning

- **Improving Adversarial Robustness for categorical data.** We propose a method to suppress the adversarial risk by smoothing the attributional sensitivity of features and the classifier's decision boundary during training with a theoretical guarantee. (first-authored paper in ICML 2024)

Working Experiences

Tencent Research Internship (2025.5-2025.12)

- **LLM Reasoning.** I'm currently a full-time research intern at Tencent America LLC, focusing on reasoning in large language models (LLMs) as part of the AGI (Audio/Speech/Language & Multimodal) team.

  We propose EVOL-RL that enables language models to evolve through label-free reinforcement learning, effectively addressing the issue of entropy collapse during training.

  Meanwhile, we are also exploring ways to expand the model's problem-solving boundaries, allowing it to tackle tasks that were previously entirely unsolvable.

Services

Invited Conference/Journal Reviewer: ICLR, NeurIPS, TDSC, ARR, COLM, AAAI, KDD