Wenhao Zheng

(+1)9843630728 | shenmishajing@gmail.com | shenmishajing | The University of North Carolina at Chapel Hill

Education

PhD. of Computer Science — *Chapel Hill*
Department of Computer Science, **The University of North Carolina at Chapel Hill** — Sep. 2024 - May. 2029*

Master of Software Engineering — *Hangzhou*
College of Computer Science and Technology, **Zhejiang University** — Sep. 2021 - Jun. 2024
Class Rank: 4/48

Bachelor of Computer Science — *Hangzhou*
Chu Kochen Honors College, **Zhejiang University** — Sep. 2017 - Jun. 2021
GPA: 3.84/4.0, Last two years GPA: 3.92/4.0

Selected Publications [Google Scholar]

[1] **Wenhao Zheng**, Xinyu Ye, Peng Xia, Fang Wu, Linjie Li, Weitong Zhang, Lijuan Wang, Yejin Choi, Yun Li, Huaxiu Yao*. "The Agent's Marathon: Probing the Limits of Endurance in Long-Horizon Tasks," *The International Conference on Learning Representations*, 2026, under review.

[2] **Wenhao Zheng**, Jianshu She, Weitong Zhang, Yixiao Chen, Leshang Chen, Souvik Kundu, Eric P. Xing, Zhengzhong Liu, Qirong Ho, Hongyi Wang, Yun Li, Huaxiu Yao*. "CLEAR: A Cost-Aware Routing System for Edge-Cloud Language Model Collaborative Inference," *The International Conference on Learning Representations*, 2026, under review.

[3] **Wenhao Zheng**†, Yixiao Chen†, Weitong Zhang, Souvik Kundu, Yun Li, Zhengzhong Liu, Eric P. Xing, Hongyi Wang, Huaxiu Yao*. "CITER: Collaborative Inference for Efficient Large Language Model Decoding with Token-Level Routing," *Conference on Language Modeling*, 2025. [Link]

[4] **Wenhao Zheng**†, Liaoyaqi Wang†, Dongsheng Peng, Hongxia Xu, Hongtu Zhu, Tianfan Fu, Huaxiu Yao*. "LIFTED: Multimodal Clinical Trial Outcome Prediction via Large Language Models and Mixture-of-Experts," *The Conference on Empirical Methods in Natural Language Processing*, 2025. [Link]

[5] Jianshu She, **Wenhao Zheng**, Zhengzhong Liu, Hongyi Wang, Eric Xing, Huaxiu Yao, Qirong Ho*. "Token Level Routing Inference System for Edge Devices," *Annual Meeting of the Association for Computational Linguistics*, 2025. [Link]

[6] Zhaoyang Wang, Jinqi Jiang, Huichi Zhou, **Wenhao Zheng**, Xuchao Zhang, Chetan Bansal, Huaxiu Yao*. "Verifiable Format Control for Large Language Model Generations," *Annual Meeting of the Association for Computational Linguistics*, 2025. [Link]

[7] Tony Lee†, Haoqin Tu†, Chi Heem Wong†, **Wenhao Zheng**, Yiyang Zhou, Yifan Mai, Josselin Somerville Roberts, Michihiro Yasunaga, Huaxiu Yao, Cihang Xie, Percy Liang*. "VHELM: A Holistic Evaluation of Vision Language Models," *Conference on Neural Information Processing Systems*, 2024. [Link]

[8] **Wenhao Zheng**, Jintai Chen, Kai Zhang, Jiahuan Yan, Jinhong Wang, Yi Cheng, Bang Du, Danny Z. Chen, Honghao Gao*, Jian Wu, Hongxia Xu*. "Polygonal Approximation Learning for Convex Object Segmentation in Biomedical Images with Bounding Box Supervision," *IEEE Journal of Biomedical and Health Informatics*, 2023. [Link]

[9] Jinhong Wang†, Zhe Xu†, **Wenhao Zheng**†, Haochao Ying*, Tingting Chen, Zuozhu Liu, Danny Z. Chen, Ke Yao*, Jian Wu. "A Transformer-based Knowledge DistillationNetwork for Cortical Cataract Grading," *IEEE Transactions on Medical Imaging*, 2023. [Link]

[10] Tingting Chen†, **Wenhao Zheng**†, Haochao Ying, Xiangyu Tan, Kexin Li, Xiaoping Li, Danny Z. Chen, Jian Wu*. "A Task Decomposing and Cell Comparing Method for Cervical Lesion Cell Detection," *IEEE Transactions on Medical Imaging*, 2022. [Link]

[11] Yi Cheng, Haochao Ying*, Renjun Hu, Jinhong Wang, **Wenhao Zheng**, Xiao Zhang, Danny Z. Chen, Jian Wu. "Robust Image Ordinal Regression with Controllable Image Generation," *International Joint Conference on Artificial Intelligence*, 2023. [Link]

[12] Tingting Chen, **Wenhao Zheng**, Heping Hu, Chunhua Luo, Jintai Chen, Chunnv Yuan, Weiguo Lu, Danny Z. Chen, Honghao Gao* and Jian Wu*. "A Corresponding Region Fusion Framework for Multi-modal Cervical Lesion Detection," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 2022. [Link]

[13] Jinhong Wang†, Jingwen Wang†, Tingting Chen, **Wenhao Zheng**, Zhe Xu, Xingdi Wu, Wen Xu*, Haochao Ying*, Danny Z. Chen, and Jian Wu. "CTT-Net: A Multi-view Cross-token Transformer for Cataract Postoperative Visual Acuity Prediction," *IEEE International Conference on Bioinformatics and Biomedicine*, 2022. [Link]

[14] Tingting Chen†, Yi Cheng†, Jinhong Wang, Zhaoxia Yang, **Wenhao Zheng**, Danny Z. Chen, and Jian Wu*. "Automating Blastocyst Formation and Quality Prediction in Time-Lapse Imaging with Adaptive Key Frame Selection," *Medical Image Computing and Computer Assisted Intervention*, 2022. [Link]

Research Experience

Infinite Agentic Benchmark & Agent Scaling Laws *UNC - CH*

May 2025 - Now

- Motivated by the performance degradation of both single and multi-agent systems over long-term interactions, this project addresses the limitations of existing benchmarks, which often rely on human-crafted tasks requiring only a few, fixed-sequence tool invocations.
- Proposed and developed a novel framework to programmatically generate benchmarks of arbitrary length and complexity by constructing rule-based, tool-use dependency trees, enabling a more rigorous evaluation of agent capabilities in extended interaction scenarios.
- The generated benchmarks cover a diverse set of tasks, including document retrieval, image reasoning, and code analysis, to ensure comprehensive assessment.
- Investigated the scaling laws that govern agent performance as a function of interaction count, analyzing how these laws shift under varying per-interaction task difficulties.

Towards General Agentic Systems: From Verifiable to Unverifiable Domains *UNC - CH*

Mar. 2025 - Now

- While Reinforcement Learning (RL) has successfully trained multi-turn agents in domains with verifiable rewards (e.g., mathematics, code generation), its application is severely limited in domains where such clear reward signals are unavailable. To bridge this gap, this project proposes a novel framework that learns a reward model from verifiable domains to guide agent training in unverifiable ones.
- First, an agent model is trained using RL to act as an expert judge on tasks with verifiable ground truth. This "agent-judge" learns to effectively assess the quality and correctness of complex outputs. Subsequently, this trained agent-judge is utilized to generate reward signals for new, unverifiable tasks, enabling the extension of powerful RL-based agent training to a broader range of applications.
- The framework incorporates an iterative refinement loop, where the agent-judge is continually updated alongside the distributions of verifiable and unverifiable tasks to progressively enhance system performance.

Work Experience

GRIPS Agent: General Purpose Fine-tuned LLM for Scam Prevention Use-cases with Lifecycle Management Agent *Amazon*

May. 2025 - Aug. 2025

- Design a three-stage pipeline, including continued pretraining, supervised instruction fine-tuning, and quantization, for the LLM fine-tuning.
- Collect the IPS domain data for unsupervised continued pretraining and conduct the continued pretraining.
- Leverage the LoRA techniques and unified format prompt to conduct multi-task instruction fine-tuning.
- Quantized our model before developing it to further accelerate the inference process.
- Outperforms in-production baselines on most use-cases with only 0.67% inference cost.
- Design an automatic agent to refresh our grips model automatically, reducing over 150 hours human work per week with similar performance or achieve up to 57.40% performance gain over time compared to no refresh baselines all with no human work required.

1B Model Pretraining *LLM360*

Oct. 2024 - Feb. 2025

- Design the pretraining of a 1B parameter language model using Megatron-LM on the TxT360 dataset.
- Conduct experiments to determine optimal training configurations and debug issues to ensure stable training performance.
- Apply scaling laws to estimate the required token count for effective pretraining and utilize a smaller model to filter and curate the dataset, enhancing its quality and relevance.
- This project contributes to advancing the understanding of large-scale model training workflows and dataset preparation strategies for efficient and scalable machine learning development.

Awards and Honors

Nov. 2022	**Scholarship**: "Zhijun He Scholarship"	*Zhejiang University*
Dec. 2022	**Scholarship**: "First Class Academic Excellence Scholarship"	*Zhejiang University*
Oct. 2022	**Honorary Title**: "Miyoshi graduate student"	*Zhejiang University*
Oct. 2022	**Honorary Title**: "Outstanding graduate student"	*Zhejiang University*
Aug. 2019	**Award**: "1st Prize in The 12th World Robotics Sailing Championship(WRSC)"	*Zhejiang University*

Skills

Programming	Python, Pytorch, C/C++, CUDA
Drawing & Typesetting	Office, LaTeX, Beamer
Tools	Git, Vscode, Vim, Docker