Yuanzhe Hu

+1-619-937-1812 | GitHub Profile
yuh127@ucsd.edu | Personal Website

Education

University of California, San Diego *Sep. 2024 – Mar. 2026*
MS. in Computer Science Engineering

Huazhong University of Science and Technology *Sep. 2020 – Jun. 2024*
BEng. in Artificial Intelligence (Innovation Experimental Honor Class) GPA: 3.91/4.0

Publications and Preprints

1. Y. Hu*, Y. Wang*, K. Lin, J. McAuley. "Evaluating Memory in LLM Agents via Incremental Multi-Turn Interactions", Submitted to NeurIPS 2025 (D&B Track), Paper Link.

2. Y. Hu, K. Goel, V. Killiakov, Y. Yang. "Eigenspectrum Analysis of Weight Matrices without Aspect Ratio Bias", ICML 2025, Paper Link.

3. Z. Liu*, **Y. Hu***, T. Pang, Y. Zhou, P. Ren, Y. Yang. "Model Balancing Helps Low-data Training and Fine-tuning", EMNLP 2024 (Oral) (168/6105), Paper Link.

4. H. Lu*, **Y. Hu***, Y. Zhou, T. Pang, X. Liu, P. Ren, Y. Yang. "Loss Landscape Analysis of Scientific Machine Learning Models", Under Review.

5. Y. Wang, D. Krotov, **Y. Hu**, Y. Gao, W. Zhou, J. McAuley, D. Gutfreund, R. Feris, Z. He. "M+: Extending MemoryLLM with Scalable Long-Term Memory", ICML 2025, Paper Link.

* Equal Contribution

Academic Experience

Large-Scale Reasoning LLM Training *Jun. 2025 – Now*
*Summer Research Intern, Advised by **Prof. Zhiting Hu*** La Jolla, CA

- Engineered and executed a large-scale supervised fine-tuning (SFT) pipeline for models up to 70B parameters (e.g., LLaMA-3.1-70B, Qwen2.5-32B) on a 64-node H200 GPU cluster.
- Achieved state-of-the-art results on challenging reasoning benchmarks by leveraging this pipeline, scoring 58.2 on LiveCodeBench (code generation) and 72.6 on AIME 2025 (math reasoning). My role included data preprocessing, experiment configuration, and model evaluation.

Constructing Benchmarks on LLMs' Long Context Understanding Ability *Oct. 2024 – May. 2025*
*CSE 298 Research Course, Advised by **Prof. Julian McAuley*** La Jolla, CA

- Contributed to the development of M+, a novel memory-augmented LLM that extends knowledge retention from under 20k to over 160k tokens, with findings published at **ICML 2025**.
- Built and deployed a comprehensive evaluation pipeline for memory-augmented LLM agents by aggregating and standardizing multiple datasets.
- Utilized this pipeline to systematically benchmark and analyze the performance of leading memory agent systems (e.g., Letta, Mem0, Cognee) and RAG systems, with findings submitted to **NeurIPS 2025**.

Model Diagnosis Based on Random Matrix Theory *Jul. 2023 – May. 2025*
*Research Intern, Advised by **Prof. Yaoqing Yang** and **Dr. Ren Pu*** Hanover, NH

- Developed a novel model diagnosis method using Random Matrix Theory to analyze neural network weight matrices, leading to a publication at **ICML 2025**.
- Engineered and evaluated a layer-wise LLM pruning strategy based on this method, demonstrating a clear path to reduce model parameters and improve inference efficiency on models like LLaMA-7B.
- Designed and implemented a layer-wise optimization algorithm that boosts model performance by **2% to 10%** in low-data fine-tuning scenarios for both NLP and SciML tasks. This work was accepted as an **Oral presentation at EMNLP 2024**.
- Develop a four-regime phrase transition analysis in the loss landscape of Neural Operators and PINNs.

Skills, Interests, Award

Programming Language: Python, C/C++, SQL, MATLAB
Field of Interest: Numerical Analysis in ML, SciML, Reason and Memory in LLM
Service: Reviwer for ICML 2025 Workshop / ICLR 2025 Workshop / NeurIPS 2024 Workshop