YUANZHE HU

Homepage: hust-ai-hyz.github.io
Phone: (+1) 619-937-1812 ◇ Email: yzhu.ml@outlook.com

EDUCATION

University of California, San Diego (UCSD) *Sep. 2024 – Mar. 2026 (expected)*
MS. in Computer Science Engineering
GPA: 3.86/4.0

Huazhong University of Science and Technology (HUST) *Sep. 2020 – Jun. 2024*
BEng. in Artificial Intelligence (Honored Class, Qiming School)
GPA: 3.91/4.0

RESEARCH SUMMARY

Model Diagnostics and Optimization: Designed and improved layer-wise model diagnostic and hyperparameter optimization methods by applying mathematical tools from Random Matrix Theory, such as eigenspectrum analysis. These methods yield more accurate layer-wise hyperparameters for LLM pruning[1] and LLM fine-tuning[2].

Memory for LLM and Agents: Designed MemoryAgentBench[3], a comprehensive benchmark to evaluate memory agents and LLM across multiple core competencies. Contributed to the design of memory-augmented LLMs (M+)[4] that achieve scalable long-term knowledge retention by integrating a long-term memory.

PUBLICATIONS

[1] **Yuanzhe Hu**, K. Goel, V. Killiakov, and Y. Yang, "Eigenspectrum analysis of weight matrices without aspect ratio bias," in *ICML*, 2025.
[2] Z. Liu*, **Yuanzhe Hu***, T. Pang, Y. Zhou, P. Ren, and Y. Yang, "Model balancing helps low-data training and fine-tuning," in *EMNLP (Oral) (168/6105)*, 2024.
[3] **Yuanzhe Hu***, Y. Wang*, K. Lin, and J. McAuley, "Evaluating memory in llm agents via incremental multi-turn interactions," in *Submission to NeurIPS (D&B Track)*, 2025.
[4] Y. Wang, D. Krotov, **Yuanzhe Hu**, *et al.*, "M+: Extending memoryllm with scalable long-term memory," in *ICML*, 2025.

* Equal Contribution

RESEARCH AND PROJECT EXPERIENCE

Large-Scale Reasoning LLM Training and Causal Analysis July 2025 - Present
Supervisors: Shibo Hao, Zhoujun Cheng, and Prof. Zhiting Hu UCSD/MBZUAI IFM

- 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 multiple GPU cluster.
- Achieved good 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.

Memory LLM and Agents Benchmarking and Construction [3], [4] Oct 2024 - Jun 2025
CSE Research Course, Supervisors: Yu Wang and Prof. Julian McAuley UCSD

- Led the development of **MemoryAgentBench**, a comprehensive benchmark framework to systematically evaluate memory in LLM agents by simulating incremental, multi-turn interactions.
- Systematically evaluated a diverse range of memory agents across four core competencies: accurate retrieval, test-time learning, long-range understanding, and conflict resolution.

- Designed and constructed two novel datasets, **EventQA** and **FactConsolidation**, to assess previously under-evaluated agent capabilities such as temporal reasoning and knowledge updating.
- Designed and implemented the evaluation framework for M+, and conducted systematic benchmarking against baseline methods including LLaMA-3.1-8B with BM25 retrieval.
- Co-authored two research papers: one on the MemoryAgentBench framework (under review) and another on the M+ model (accepted to ICML 2025).

De-biased Model Diagnosis Based on Eigspectrum and MP-Law [1] Nov 2024 - Apr 2025
Supervisors: Prof. Yaoqing Yang Dartmouth College

- Theorized and empirically validated that existing eigenspectrum analysis is biased by weight matrix **aspect ratios**, a phenomenon explained by the **Marchenko-Pastur (MP) law**, leading to inaccurate layer diagnostics.
- Developed ***FARMS***, a novel subsampling method grounded in **Random Matrix Theory** (RMT), to normalize weight matrices to a fixed aspect ratio, enabling an unbiased, size-invariant evaluation of layer training quality.
- Validated the method's effectiveness across diverse domains (**LLM Pruning, CV, SciML**), consistently **outperforming state-of-the-art** layer-wise optimization methods.
- First-authored a research paper accepted to the **ICML** 2025.

Layer-wise Optimization on Low-data SFT and Training [2] Jun 2023 - Jun 2024
Supervisors: Dr. Pu Ren and Prof. Yaoqing Yang Dartmouth College

- Utilized the **spectral analysis of model weights**, based on Heavy-Tailed Self-Regularization (HT-SR) theory, to establish a quantitative link between data scarcity and imbalanced layer-wise training, providing a theoretical foundation for designing superior model diagnosis metrics and optimization methods.
- Developed a dynamic **layer-wise learning rate** scheduling algorithm to rebalance training quality across layers, overcoming limitations of prior optimizers in NLP training scenarios.
- Validated the method's effectiveness through extensive experiments on diverse models and benchmarks, improving LLM test accuracy **by 2-10%** in low-data SFT and leading to a **co-first** authored publication at EMNLP 2024 **(Oral Presentation)**.
- Investigated the framework's generalizability by extending the layer-wise method to tune **other hyperparameters** (e.g., weight decay, dropout), and validated its effectiveness on image classification with ResNet/ViT models on the CIFAR-100 dataset.
- Involved in the rebuttal process and supplementing experimental data for the paper accepted by NeurIPS 2023.

COMPETITION AWARDS

Second Prize at 17th National College Students Intelligent Car Race (7% among 2771 teams), awarded by Chinese Automation Association, Summer 2022.

SERVICE

Volunteer for ICML 2025
Reviewer for Workshops in ICML, ICLR, NeurIPS

TECHNICAL SKILLS

Programming Languages	Python, C/C++, SQL, Bash, Matlab
Machine Learning Tools	PyTorch, Hugging Face Transformers, LangChain, LLaMA Factory