Zheyu Shen

Contact Information

A. James Clark School of Engineering
University of Maryland, College Park
College Park, Maryland 20742

E-mail: `zyshen@umd.edu`
Phone (+1) 202-306-3697
`linkedin.com/in/zheyushen/`

Research

Machine Learning, Systems.

Education

PhD in Electrical and Computer Engineering 2023-present
A. James Clark School of Engineering,
University of Maryland, College Park

MS in Computer Science 2019-2020
Viterbi School of Engineering, University of Southern California

Awards

Graduate School Dean's Fellowship 2023.

Publications

EdgeLoRA: An Efficient Multi-Tenant LLM Serving System on Edge Devices.
Z. Shen, Y. He, Z. Wang, Y. Zhang, G. Sun, W. Ye, A. Li.
MobiSys 2025

Flora: Federated Fine-Tuning Large Language Models with Heterogeneous Low-Rank Adaptations.
Z. Wang, **Z. Shen**, Y. He, G. Sun, H. Wang, L. Lyu, A. Li.
NeurIPS 2024

SHED: Shapley-Based Automated Dataset Refinement for Instruction Fine-Tuning.
Y. He, Z. Wang, **Z. Shen**, G. Sun, Y. Dai, Y. Wu, H. Wang, A. Li.
NeurIPS 2024

What Matters in Transformers? Not All Attention Is Needed.
S. He, G. Sun, **Z. Shen**, A. Li.
arXiv preprint arXiv:2406.15786 (2024)

Domino: Eliminating Communication in LLM Training via Generic Tensor Slicing and Overlapping.
G. Wang, C. Zhang, **Z. Shen**, A. Li, O. Ruwase.
arXiv preprint arXiv:2409.15241 (2024)

One Communication Round Is All It Needs for Federated Fine-Tuning Foundation Models.
Z. Wang, B. Tian, Y. He, **Z. Shen**, L. Liu, A. Li.
arXiv preprint arXiv:2412.04650 (2024)

Fair Diagnosis: Leveraging Causal Modeling to Mitigate Medical Bias.
B. Tian, Y. He, M. Liu, Y. Dai, Z. Wang, S. He, G. Sun, **Z. Shen**, W. Ye, Y. Wu, and others.
arXiv preprint arXiv:2412.04739 (2024)

Prada: Black-Box LLM Adaptation with Private Data on Resource-Constrained Devices.
Z. Wang, Y. He, **Z. Shen**, Y. Li, G. Sun, M. Lee, A. Li.
arXiv preprint arXiv:2503.14932 (2025)

SymRTLO: Enhancing RTL Code Optimization with LLMs and Neuron-Inspired Symbolic Reasoning.
Y. Wang, W. Ye, P. Guo, Y. He, Z. Wang, B. Tian, S. He, G. Sun, **Z. Shen**, S. Chen, and others.
arXiv preprint arXiv:2504.10369 (2025)

CoIn: Counting the Invisible Reasoning Tokens in Commercial Opaque LLM APIs.
G. Sun, Z. Wang, B. Tian, M. Liu, **Z. Shen**, S. He, Y. He, W. Ye, Y. Wang, A. Li.
arXiv preprint arXiv:2505.13778 (2025)

Invisible Tokens, Visible Bills: The Urgent Need to Audit Hidden Operations in Opaque LLM Services.
G. Sun, Z. Wang, X. Zhao, B. Tian, **Z. Shen**, Y. He, J. Xing, A. Li.
arXiv preprint arXiv:2505.18471 (2025)

Arctic-Text2SQL-R1: Simple Rewards, Strong Reasoning in Text-to-SQL.
Z. Yao, G. Sun, L. Borchmann, **Z. Shen**, M. Deng, B. Zhai, H. Zhang, A. Li, Y. He.
arXiv preprint arXiv:2505.20315 (2025)

CogniPair: From LLM Chatbots to Conscious AI Agents-GNWT-Based Multi-Agent Digital Twins for Social Pairing-Dating & Hiring Applications.
W. Ye, S. Chen, Y. Wang, S. He, B. Tian, G. Sun, Z. Wang, Z. Wang, Y. He, **Z. Shen**, and others.
arXiv preprint arXiv:2506.03543 (2025)

Services

Reviewer of ICLR, NeurIPS.

Talks

EdgeLoRA: An Efficient Multi-Tenant LLM Serving System on Edge Devices.
MobiSys 2025

Teaching

Teaching Assistant, Cryptography, UMD	Spring 2025
Teaching Assistant, Advanced Computer Communication, USC	Fall 2021

Industry Experience

Snowflake. Summer 2025
Research intern.
I developed Arctic-Text2SQL-R1 boosting the throughput of reinforce learning training on text2sql task.

Tencent. 2021-2023
Research Development Engineer.
I developed a TensorFlow-based sparse-model online training system to replace Tencent's self-developed ML framework, boosting usability and performance. I engineered a distributed, trillion-parameter model inference system with a parameter server.