

Guoheng Sun

Second-year PhD student at Department of Electrical & Computer Engineering, University of Maryland, College Park

☎ (240) 615-6444 | ✉ ghsun@umd.edu | 🏠 s1ghhh.github.io | 🌐 [s1ghhh](https://s1ghhh.github.io)

Research interests

My primary research focus is on **improving the efficiency of LLMs/VLMs/VLAs**. In addition, I am interested in **trustworthy AI** (e.g., transparency in AI service billing), **VLA models** (improving spatial reasoning, design efficient architectures), and **post-training techniques for LLMs** (e.g., enhancing reasoning ability, multi-turn interaction, and tool calling).

Education

University of Maryland, College Park

PH.D. STUDENT IN COMPUTER ENGINEERING

8/2024 - 5/2028 (expected)

Advisor: [Prof. Ang Li](#)

Sichuan University

SCHOOL OF CYBER SCIENCE AND ENGINEERING

9/2020 - 6/2024

• GPA : 3.89 / 4.00

Selected Publications

*Equal Contribution †Equal Supervision

[VLA / Efficiency] [G. Sun](#), T. Du, K. Feng, C. Luo, X. Ding, Z. Shen, Z. Wang, Y. He, A. Li, “ROCKET: Residual-Oriented Multi-Layer Alignment for Spatially-Aware Vision-Language-Action Models”. arXiv preprint arXiv:2602.17951, 2026. [\[Link\]](#)

[Security] [G. Sun*](#), Y. Fu*, H. Yang, J. Huang, R. Zhang, H. Wang, “Enhancing the Security of Large Character Set CAPTCHAs Using Transferable Adversarial Examples”. IEEE Transactions on Dependable and Secure Computing (TDSC), 2025. [\[Link\]](#)

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

[Reasoning LLM Audit] [G. Sun*](#), Z. Wang*, X. Zhao, B. Tian, Z. Shen, Y. He, J. Xing, A. Li, “Invisible Tokens, Visible Bills: The Urgent Need to Audit Hidden Operations in Opaque LLM Services”. ResponsibleFM Oral @ NeurIPS 2025. [\[Link\]](#)

[Reasoning LLM Audit / RL] Z. Wang*, [G. Sun*](#), Y. He, Z. Shen, B. Tian, A. Li, “Predictive Auditing of Hidden Tokens in LLM APIs via Reasoning Length Estimation”. arXiv preprint arXiv:2508.00912, 2025. [\[Link\]](#)

[Efficiency] S. He*, [G. Sun*](#), Z. Shen, A. Li, “What Matters in Transformers? Not All Attention is Needed”. Transactions on Machine Learning Research (TMLR). [\[Link\]](#) [\[Code\]](#)

[Reasoning / Code / RL] Y. Wang*, [G. Sun*](#), W. Ye, G. Qu, A. Li, “VeriReason: Reinforcement Learning with Testbench Feedback for Reasoning-Enhanced Verilog Generation”. arXiv preprint arXiv:2505.11849, 2025. [\[Link\]](#)

[Reasoning / Code / RL] Z. Yao*, [G. Sun*](#), L. Borchmann, Z. Shen, M. Deng, B. Zhai, H. Zhang, A. Li, Y. He, “Arctic-Text2SQL-R1: Simple Rewards, Strong Reasoning in Text-to-SQL”. arXiv preprint arXiv:2505.20315, 2025. [\[Link\]](#) [\[Code\]](#)

[Efficiency] X. Zhao*, [G. Sun*](#), R. Cai*, Y. Zhou*, P. Li*, P. Wang, B. Tan, Y. He, L. Chen, Y. Liang, B. Chen, B. Yuan, H. Wang†, A. Li†, Z. Wang†, T. Chen†, “Model-GLUE: Democratized LLM Scaling for A Large Model Zoo in the Wild”. NeurIPS 2024 Datasets and Benchmarks [\[Link\]](#)

[Efficiency] Z. Shen, Y. He, Z. Wang, Y. Zhang, [G. Sun](#), W. Ye, A. Li, “EdgeLoRA: An Efficient Multi-Tenant LLM Serving System on Edge Devices”. 23rd ACM International Conference on Mobile Systems, Applications, and Services (MobiSys), 2025. [\[Link\]](#)

[Efficiency] S He, T Ge, [G Sun](#), et al. “Router-Tuning: A Simple and Effective Approach for Enabling Dynamic-Depth in Transformers”. EMNLP (Main) 2025. [\[Link\]](#)

[Security] Z Wang, Z Shen, Y He, [G Sun](#), et al. “Flora: Federated fine-tuning large language models with heterogeneous low-rank adaptations”. NeurIPS 2024. [\[Link\]](#)

[Efficiency] Y He, Z Wang, Z Shen, G Sun, et al. "SHED: Shapley-Based Automated Dataset Refinement for Instruction Fine-Tuning". NeurIPS 2024. [\[Link\]](#)

Research Experience

Research Internship at [Snowflake AI Research](#)

Mentor: [Zhouwei Yao & Yuxiong He](#)

5 / 2025 – 8 / 2025

- Verl, Ray etc.
- Using reinforcement learning to train a SOTA text-to-SQL model. Arctic-text2sql-r1-32B achieved Top 1 on the BIRD-SQL Leaderboard (as of June 2025).
- Enhancing LLM text-to-SQL capability through reinforcement learning by introducing tool calls in multi-turn dialogues.
- Ongoing project: Building a multi-turn dialogue and tool-calling Text2SQL model with online knowledge injection, enabling the model to execute tools during multi-turn conversations and to learn knowledge it has never been trained on before.

Research Internship at the LLM360 Team/MBZUAI

Mentor: [Hongyi Wang](#)

9 / 2024 – 12 / 2024

- Megatron, TorchTitan, etc.
- Reproduce the scaling law of LLaMA 3 and explore scaling laws under overtraining scenarios. Contribute to model size selection, FLOPs calculation, fitting IsoFLOPs curves, and predicting model loss on test sets as well as performance on downstream tasks.
- Process mathematical data, including deduplication, format modification, and fixing formulas and tables.
- The outcomes have been consolidated and presented in [\[Slides\]](#).

What Matters in Transformers? Not All Attention is Needed

Advisor: [Ang Li](#)

6 / 2024 – 10 / 2024

- Exploring redundancy in the transformers architecture.
- We found that the more computationally intensive attention layers are more redundant, especially in the middle layers. Simply dropping these layers does not result in significant performance degradation.
- For 70B LLMs, discarding about half of the attention layers can increase speed by approximately 50% while maintaining over 95% of the performance.

Model-GLUE: Democratized LLM Scaling for A Large Model Zoo in the Wild

Advisor: [Ang Li](#)

2 / 2024 – 6 / 2024

- Benchmark the existing model merging methods and model mixture methods.
- By effectively combining merging and model mixture techniques, we can reuse models with different architectures, varied initializations, and models that are difficult to merge.
- Our proposed Model-GLUE method achieves an average performance improvement of 5.61% compared to the best models in the model zoo.

Skills

Languages Python, C#, C++, C, JAVA, JavaScript, etc.

Tool Kits Git, Bash, Docker, MySQL, etc.

Others PyTorch, TensorFlow, \LaTeX , Unity, Sklearn, Blender, etc.

Awards

2025	Qualcomm Innovation Fellowship [Link] , Qualcomm
2024	The Dean's Fellowship , University of Maryland, College Park
2022	National Scholarship , Ministry of Education (the highest honor scholarship in China)
2022&2023	First-class University Annual Scholarship , Sichuan University
2023	Outstanding Graduate , Sichuan University
2021&2022	Outstanding Student , Sichuan University
2022	National 1st Prize , China International College Students "Internet+" Internet innovation and Entrepreneurship Competition
2022	National 3rd Prize , "China Software Cup" College Student Software Design Competition