# Guoheng **Sun**

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

(240) 615-6444 | ghsun@umd.edu | scholar.google.com/citations?user=fMnmSXsAAAAJ | s1ghhh

## Research interests

My primary research focus is on **improving the efficiency of LLMs** or enhancing their capabilities through **training-free or training-less methods**. Additionally, I am also interested in the **privacy, safety alignment, and pretraining** of LLMs.

## Education

### University of Maryland, College Park

PH.D. STUDENT IN COMPUTER ENGINEERING — *8 / 2024 - 5 / 2029 (expected)*

Advisor: Prof. Ang Li

### Sichuan University

SCHOOL OF CYBER SCIENCE AND ENGINEERING — *9 / 2020 - 6 / 2024*

- GPA : 3.89 / 4.00 ( 91.79 / 100 )
- Rank : 3 / 189

## Publication

*Equal Contribution †Equal Supervision

**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]

**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]

**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"**. arXiv preprint arXiv:2505.18471, 2025. [Link]

**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]

**S. He*, G. Sun*, Z. Shen, A. Li, "What Matters in Transformers? Not All Attention is Needed"**. arXiv preprint arXiv:2406.15786, 2024. [Link] [Code]

**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]

**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]

**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]

**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

*Advisor: Zhewei Yao & Yuxiong He*

*3 / 2025 – now*

- 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.

**Research Internship at the LLM360 Team/MBZUAI** *Advisor:Hongyi Wang & Hector Liu*

*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.

**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**, 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 Undergraduate Graduates**, 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