

SU YUPENG

EE/CS PhD Applicant of 2025 Fall.

 [Personal Website](#)  [Google Scholar](#)
 +86 183 9018 3270  [Linkedin Profile](#)
 Shen Zhen, China  yupengsu06@gmail.com

Hi there! I am a senior student from [ZhiCheng College](#) and [School of Microelectronics](#) in Southern University of Science and Technology, advice by [Prof. Hao Yu](#). I also work at [Hong Kong University](#) as a Student RA and collaborate with [Prof. Ngai Wong](#) closely. My skills are not only proficient in implementing algorithms design using Python, C++ or Java, but also the entire IC design process, including digital front-end design using verilog HDL, chip layout drawing, and some compiler design. I have plentiful experiences in Pretraining(from scratch/continue pretrain), Supervised Fine-tuning, Evaluation, Model Compression(Quantization, Pruning, Knowledge Distillation and Low-rank Decomposition) and Deployment(mainly focus on edge). My research interests includes Efficient/Low-resource methods for NLP, Model Edge Deployment and AI accelerator.

EDUCATION

9/2021 - 6/2025 expected **Southern University of Science and Technology** **Bachelor of Microelectronics Science and Engineering**
Cumulative Grade Point Average (CGA/GPA): 3.9/4.0, Major Rank: 1/92, 153 scores have been gained.
Grad course Microelectronics Innovations & Technology Leadership by [Prof. Kai Chen](#): A+ (Top 1),
Deep Learning on Chip by [Prof. Hao Yu](#): A (Top 1), Data Structures & Algorithm Analysis: A+ (Top 1).

PUBLICATIONS

- Guan, Z.; Huang, H.; **Su, Y.**; Huang, H.; Wong, N. and Yu, H. (2024). APTQ: Attention-aware Post-Training Mixed-Precision Quantization for Large Language Models. Accepted by the 2024 61th ACM/IEEE Design Automation Conference (DAC): <https://arxiv.org/abs/2402.14866>.
- Su, Y.**; Guan, Z.; Liu, X.; Jin, T.; Wu, D.; Chesi, G.; Wong, N. and Yu, H. (2024). LLM-Barber: Block-Aware Rebuilder for Sparsity Mask in One-Shot for Large Language Models. Preprint Version: <https://arxiv.org/abs/2408.10631>.

RESEARCH EXPERIENCES

10/2024 - 2/2025 Senior **LLMs Knowledge Distillation for Internalize CoT Reasoning** **Research Assistant of Ngai Wong's Lab of HKU**
We will explore an alternative reasoning approach: instead of explicitly producing the chain of thought reasoning steps, we use the language model's internal hidden states to perform implicit reasoning.

8/2024 - 2/2025 Senior **Low Bit MatMulFreeLM Pretraining and Finetuning** **Research Assistant of Ngai Wong's Lab of HKU**
We will propose a novel architecture for pretraining a low bit MatMul-Free LLM for edge deployment.

6/2024 - 10/2024 Senior **LLMs Compilation and Edge Deployment** **High Performance Integrated Circuit Design Lab of SUSTech**
We have implemented the complete process from compression, compilation to edge deployment, successfully inferring the 4-bit quantized chatglm3-8b model on the Xilinx VCU128 FPGA.

2/2024 - 8/2024 Junior **LLMs Post-Training Pruning and Sparsity** **High Performance Integrated Circuit Design Lab of SUSTech**
We built LLM-Barber, a novel method for efficiently pruning LLMs by rebuilding the sparsity mask in a one-shot fashion, without any retraining or weight reconstruction. Code is available at [this URL](#).

6/2023 - 12/2023 Junior **LLMs Mix-Precision Post-Training Quantization** **High Performance Integrated Circuit Design Lab of SUSTech**
We propose APTQ (Attention-aware Post-Training Mixed-Precision Quantization), which considers not only the second-order information of each layer's weights, but also the nonlinear effect of attention outputs.

2/2022 - 9/2022 Freshman **Ultra-High Vacuum (UHV) Experimentation** **Research Assistant of Quantum Academy of SUSTech**
Perform precision instrument calibration, maintenance and assisting in observational study preparations.

WORK EXPERIENCES

9/2023 - 6/2025 Junior - Senior **Peer Mentor for Academic Advisory Program** **Student Affairs Department of SUSTech**
I have accumulated nearly a hundred hours of one-on-one consultation experience with fellow students.

9/2022 - 6/2025 Sophomore - Senior **Instructor for Undergraduate Course Calculus I/II** **Student Development Center of SUSTech**
I have instructed nearly a thousand of fellow students in [Calculus I/II review courses](#) over four semesters.

ACADEMIC PROJECTS

Hisilicon **HiBao: Your Artificial Intelligent Voice Assistant** **Hisilicon Embedded Chip and System Design Competition Link**
We have designed an AI voice assistant "Hibao" that can recognize family members and provide customized conversational Q&A in the platforms of Hisilicon Pegasus and Taurus, advancing the application of LLMs deployment within the Hisilicon ecosystem. This project won Second prize in the South Division.

Please visit my [personal website](#) for more detailed informations.