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 at ZhiCheng College and the School of Microelectronics at the Southern University of Science and Technology, where I am advised by Prof. Hao Yu. I also work as a Student Research Assistant at The University of Hong Kong, collaborating closely with Prof. Ngai Wong. My programming expertise spans Python, C++, and Java, with experience in developing efficient algorithms for complex systems. And I specialize in FPGA-based workflows, including digital front-end design with Verilog HDL, FPGA prototyping, chip layout, and compiler design. This diverse skill set allows me to bridge the gap between software development and hardware optimization effectively. My experience spans a wide range of tasks such as pretraining (from scratch and continued pretraining), supervised fine-tuning, evaluation, model compression (quantization, pruning, knowledge distillation, and low-rank decomposition), and deployment with a focus on edge devices. My research interests lie in Efficient and Low-resource Methods for NLP, Model Deployment on Edge, and AI Accelerators Design.

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).

INTERNSHIP

8/2024 - 2/2025 Part-Time
Next Gen AI(NGai) Lab of HKU — **Student Research Assistant**
Working approximately 20 hours per week, I proposed a novel architecture for pretraining and fine-tuning low-bit LLMs to enable efficient edge deployment while collaborating closely with Professor Ngai Wong.

6/2024 - 9/2024 Full-Time
High Performance Integrated Circuit Design Lab of SUSTech — **Engineering Intern**
Working approximately 40 hours per week, I implemented the complete pipeline for compressing, optimizing, and deploying quantized LLMs, successfully advancing high-performance AI solutions on edge.

PUBLICATIONS

1. Guan, Z., Huang, H., **Su, Y.**, Huang, H., Wong, N. and Yu, H. (2024, June). APTQ: Attention-aware post-training mixed-precision quantization for large language models. *In Proceedings of the 61st ACM/IEEE Design Automation Conference (pp. 1-6). Doi:* https://doi.org/10.1145/3649329.3658498.

2. **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 Internalizing 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.

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 six 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.