

Hongli Shen

hlshen19@gmail.com • honglishen.com • github.com/renmiamu • +86 13059998867

Education

Southern University of Science and Technology (SUSTech)

Bachelor of Engineering in Computer Science and Technology

GPA: 3.89/4.00 Average Score: 93.06/100

Sep. 2023 – Present

Shenzhen, China

Rank: 12/152

Relevant Coursework: Algorithm Design and Analysis (99), Introduction to Mathematical Logic (H) (98), Advanced Programming (97), Principles of Database Systems (95), Natural Language Processing (94), Deep Learning (93), Discrete Mathematics (93), Object-Oriented Analysis and Design (92), Linear Algebra (90), Probability and Statistics for Engineering (90), Artificial Intelligence (88).

National University of Singapore (NUS)

School of Computing Summer Workshop

Grade: A-

May 2025 – July 2025

Singapore

Research Interests

Large Language Models, Large Reasoning Models, Multimodal Large Language Models, AI Agents.

Publications

1. [Preprint] **Hongli Shen***, Shaopeng Fu*, Qinbo Zhang, Jian Li, and Di Wang. “Dual-Adversarial Safety Alignment: Cultivating Intrinsic Threat Comprehension in LRMs.” *arXiv preprint arXiv:2608.09542*, 2026. [Code]

2. [Paper] Yan Hu, Mingdao Gong, Zhongxi Qiu, Jiabao Liu, **Hongli Shen**, Mingzhen Yuan, Xiaoqing Zhang, Heng Li, Hai Lu, and Jiang Liu. “COPh100: A Comprehensive Fundus Image Registration Dataset from Infants Constituting the “RIDIRP” Database.” *Scientific Data*, 12, 99, 2025.

*Equal contribution.

Research Experience

Safety Alignment for Large Reasoning Models

PRADA Lab, KAUST | Advisor: Prof. Di Wang

Feb. 2026 – Aug. 2026

Thuwal, Saudi Arabia

- Proposed **AdvSafe**, a dual-adversarial safety alignment framework that trains large reasoning models to internalize harmful intent and adversarial mechanisms rather than memorizing surface-level refusal patterns.
- Developed an **agentic jailbreak synthesis pipeline** and a teacher-guided adversarial extraction process to construct structured safety reasoning traces, and fine-tuned student LRMs using LoRA-based supervised fine-tuning.
- Improved jailbreak robustness and out-of-distribution safety generalization while preserving general reasoning utility.
- First-authored the resulting paper, *Dual-Adversarial Safety Alignment: Cultivating Intrinsic Threat Comprehension in LRMs*, currently under review at **NeurIPS 2026**.

Ophthalmic Image Registration

iMed Lab, SUSTech | Advisor: Prof. Jiang Liu

Aug. 2024 – Nov. 2024

Shenzhen, China

- Conducted research on fundus image processing and retinal image registration, focusing on challenging infant ophthalmic images with substantial variations in image quality and acquisition conditions.
- Contributed to the construction and evaluation of **COPh100**, a comprehensive infant fundus image registration benchmark consisting of 100 eyes and 491 image pairs from the RIDIRP database.
- Contributed to dataset annotation, preprocessing, and benchmark evaluation for retinal image registration; the resulting work was published in **Scientific Data** in 2025.

Ongoing Research Projects

Chain-of-Thought Reconstruction for Large Reasoning Models

PRADA Lab, KAUST | Advisor: Prof. Di Wang

June 2026 – Present

Thuwal, Saudi Arabia

- Investigating chain-of-thought reconstruction in large reasoning models, with a focus on recovering and understanding reasoning trajectories to improve model interpretability and reliability.

Environment Scaling for LLM Agents

RUC | Advisor: Shaolei Zhang

June 2026 – Present

Beijing, China

- Developing scalable interactive workspace environments for training and evaluating autonomous LLM agents on diverse and long-horizon tasks.
- Investigating how scaling the diversity and complexity of interactive environments can improve agent capabilities and generalization.

Honors & Awards

SUSTech Merit Student Scholarship (First Class, Top 2%)

Star of Sports, Shuli College

Provincial Third Prize, China Undergraduate Mathematical Contest in Modeling (CUMCM)

SUSTech Outstanding Freshman Scholarship

Nov. 2024, Nov. 2025

Dec. 2024, Dec. 2025

Sep. 2024

Dec. 2023

Leadership & Activities

Captain, [Shuli College Football Team](#)

Member, SUSTech Varsity Football Team

Skills

Programming

Machine Learning & LLMs

Development Tools

Hardware

Languages

English Proficiency

Python, C/C++, Java, SQL, JavaScript

PyTorch, LoRA, Supervised Fine-Tuning, LLM and Agent Development

Git, FastAPI, PostgreSQL, React

RISC-V, Verilog, FPGA

Chinese (Native), English (Fluent)

TOEFL: 5.5 (Reading: 4.5, Listening: 5.5, Speaking: 6.0, Writing: 5.5)