Xinyuan Li

No.866, Yuhangtang Road, Xihu District, Hangzhou, Zhejiang Province, China
+86 130 3529 0600   xinyuanli2327@outlook.com   Website   GitHub   Blog

Educational Experience

Zhejiang University

Hangzhou, Zhejiang

B.E. in Automation Engineering, Minor in ACEE Honor Class of CKC;

Sept 2023 - Jun 2027

- Ranked in the top 5 % of the major in the first academic year.
- GPA: 3.95/4.00

Research Experience

CAD&CG Lab

Jan 2025 - Jul 2025

- Gained exposure to foundational research practices by assisting with a project on Gaussian Splatting in CV, supervised by Prof. Sida Peng.

MLL Lab

Jul 2025 -

- Focus on improving the spatial reasoning ability and spatial understanding of language models.

Related Course & Projects

Deep Learning for Computer Vision, Umich EECS498.008

Jan 2025 - Feb 2025

- Finished all videos and code assignments, [Code repo].
- Developed a basic understanding of deep learning.
- Strengthened programming skills and software engineering practices.

LLM from scratch

Apr 2025

- BUilt a toy GPT from scratch, following the instruction of this guidance. [Code repo]
- Enhanced my understanding of LLMs

NeRF Replication

Jun 2025

- Replicated NeRF given a rough infrastructure. [Code repo]
- Greatly improved my programming skills and software engineering capabilities.
- Deeper understanding of experimental infrastructures.

Mathematical Foundations of Reinforcement Learning

Apr 2025 - May 2025

- Focused on the mathematical theory behind key RL algorithms, providing a math basis for future pratical implementation. [Course Notes]

CMU 11-868: Large Language Model Systems

Jul 2025 - Present

- Developed a comprehensive understanding of the full-stack LLM system, from low-level GPU kernel optimization to large-scale distributed training and efficient model serving. [Code Repo]
- Implemented a deep learning framework, including a custom auto-differentiation engine, to grasp the core mechanics of modern LLM systems.
- Grained extensive hands-on experience in high-performance computing by writing and optimizing CUDA kernelsfor key Transformer components, such as softmax, attention and layernorm on GPUs.

- Mastered distributed training paradigms for scaling large models, implementing both **Data Parallelism** and **Model Parallelism** (Tensor/Pipeline), and studied communication-efficient algorithms.
- Explored SOTA techniques for LLM inference and serving, including model **quantization**, MoE architecture, advanced KV Cache management, and serving system leveraging **PagedAttention** (vLLM) and **SGLang**.