Gaokai Zhang

+1 217-974-0847 | gaokaiz@andrew.cmu.edu |
github.com/yourusername | linkedin.com/in/yourlinkedin

Education

Carnegie Mellon University Aug 2025 – May 2027 (Expected)
M.S. in Intelligent Information Systems (MIIS), Language Technologies Institute (LTI) *Pittsburgh, PA*

University of Illinois at Urbana-Champaign Aug 2021 – May 2025
B.S. in Computer Engineering, GPA: 3.89/4.0 *Urbana, IL*

Zhejiang University Aug 2021 – May 2025
B.Eng. in Electrical and Computer Engineering, GPA: 3.95/4.0 *Haining, China*

Experience

Research Intern – System and Networking Group Jul 2024 – Jul 2025
Microsoft Research Asia (MSRA) *Beijing, China*

- Initiated the development of an efficient Reinforcement Learning recipe for long-context reasoning with LLMs.
- Contributed to the LongRoPE2 Research project, extending LLM context length to millions of tokens while preserving short-context capabilities; accepted as poster at ICML 2025.
- Delivered context-extended, downstream-ready LLMs to internal teams, including Microsoft Asia-Pacific R&D.
- Designed a scalable pipeline for curating large-scale, high-quality supervised fine-tuning (SFT) datasets.

Research Intern – LLM Systems & Cloud Optimization Nov 2024 – Present
University of Illinois Urbana-Champaign *Urbana, IL*

- Designed and evaluated cost-efficient LLM training/inference strategies across heterogeneous accelerators (A100, H100, TPU) using Megatron-LM on CloudLab.
- Contributed to an automated planner for optimal parallelism and deployment configurations under dynamic SLOs; co-authoring a system paper.

Research Intern – LLM Safety & Robustness Mar 2024 – Oct 2024
University of Illinois Urbana-Champaign *Urbana, IL*

- Quantified LLM robustness to stochastic attacks (word- and character-level augmentations) using SORRY-Bench, with confidence bounds based on Hoeffding and Clopper-Pearson methods.
- Co-authored a paper under review at *Transactions on Machine Learning Research (TMLR).*

Technical Skills

Languages: Python, C/C++, SQL, x86 Assembly
Frameworks & Libraries: PyTorch, Hugging Face Transformers, Megatron-LM
Tools: Git, Slurm, MySQL, QEMU, CloudLab, Docker
Areas: LLMs, NLP, Distributed Systems, Long-Context Learning, Reinforcement Learning