GUANZHENG CHEN

Homepage gc.chen@u.nus.edu
Google Scholar Github

EDUCATION

National University of Singapore Singapore
PhD Student (Second Year), School of Computing Aug 2024 - Present
Advisor: Dr. Michael Qizhe Shieh

Sun Yat-sen University Guangzhou, China
Master Student, Computer Science and Engineering Sep 2021 - Jun 2024

Chongqing University Chongqing, China
Bachelor Degree, Computer Science Sep 2017 - Jun 2021

RESEARCH INTERESTS

- Amplify the long-context capabilities of LLMs
  - explore highly efficient and scalable architectures for long-context modeling
  - advance long-context post-training
  - develop agentic memory systems.
- Enhance long-term memory by test-time training & continual learning

PUBLICATIONS

- **LongRLVR: Long-Context Reinforcement Learning Requires Verifiable Context Rewards**
  Guanzheng Chen, Michael Qizhe Shieh, Lidong Bing
  Accepted at Fourteenth International Conference on Learning Representations (***ICLR 2026***).
- **RAPID: Long Context Inference with Retrieval-Augmented Speculative Decoding**
  Guanzheng Chen*, Qilong Feng*, Jinjie Ni, Xin Li, Michael Qizhe Shieh
  Accepted at Forty-Second International Conference on Machine Learning (***ICML 2025, Spotlight***).
- **LongPO: Long Context Self-Evolution of Large Language Models through Short-to-Long Preference Optimization**
  Guanzheng Chen, Xin Li, Michael Qizhe Shieh, Lidong Bing
  Accepted at The Thirteenth International Conference on Learning Representations (***ICLR 2025***).
- **CLEX: Continuous Length Extrapolation for Large Language Models**
  Guanzheng Chen, Xin Li, Zaiqiao Meng, Shangsong Liang, Lidong Bing
  Accepted at The Twelfth International Conference on Learning Representations (***ICLR 2024***).
- **Mitigating Object Hallucinations in Large Vision-Language Models through Visual Contrastive Decoding**
  Sicong Leng, Hang Zhang, **Guanzheng Chen**, Xin Li, Shijian Lu, Chunyan Miao, Lidong Bing
  Accepted at Conference on Computer Vision and Pattern Recognition 2024 (***CVPR 2024***).

- **Revisiting Parameter-Efficient Tuning: Are We Really There Yet?**
  Guanzheng Chen, Fangyu Liu, Zaiqiao Meng, and Shangsong Liang
  Accepted at The 2022 Conference on Empirical Methods in Natural Language Processing (***EMNLP 2022, Oral Presentation***).

- **DISCO: Distributed Long Context Scaling for Large Language Models**
  Guanzheng Chen, Viet Dac Lai, Subhojyoti Mukherjee, Branislav Kveton, Seunghyun Yoon, Franck Dernoncourt, Michael Qizhe Shieh, Trung Bui
  Submitted to Forty-Third International Conference on Machine Learning (***ICML 2026***).

RESEARCH PROJECTS

- **SeaLLMs: Large Language Models for Southeast Asia** Singapore & Hangzhou, China
  - **Research Objective**: Worked on a family of pretrained large language models for Southeast Asian (SEA) languages, which mainly consists of Vietnamese, Indonesian, Thai, along with those in English and Chinese. Our customized training process helps enhance our models' ability to understand, respond, and serve communities whose languages are often neglected by previous English-dominant LLMs, while outperforming existing polyglot LLMs, like BLOOM or PolyLM.
  - **Outcome**: The family of SeaLLMs would be released and publicly available, contributing to the community of polyglot LLMs.
  - **Personal Contribution**: Design the strategy for low-quality data filtering, using the FastText and Ken-LM, to prepare a cleaned corpus for the pertaining of SeaLLMs.
- **VideoLLaMA 2 & 3** Singapore & Hangzhou, China
  - **Research Objective & Outcome**: A series of multimodal foundation models with frontier image and video understanding capacity.
  - **Personal Contribution**: Design the strategy for synthetic caption generation.
- **Mirothinker: Open-source Deep Research Agent** Singapore
  - **Research Objective**: Train open-source models targeting advanced deep research capabilities.
  - **Outcome**: A family of deep research models (Mirothinker) along with agentic workflow (MiroFlow).
  - **Personal Contribution**: (1) Accelerating deep research agents via post-hoc sparse attention — without retraining the base model, using MInference for prefill and learned auxiliary gating modules (SeerAttention-R) for decoding. (2) Integrate these sparse attention kernels into SGLang to achieve practical speedup.

RESEARCH EXPERIENCE

Adobe Research San Jose & California, United State
Research Intern, Mentor: Dr. Trung Bui Nov 2025 - Jan 2026

MiroMind AI Singapore
Research Intern, Mentor: Dr. Lidong Bing April 2025 - Nov 2025

Alibaba DAMO Academic Singapore & Hangzhou, China
Research Intern, Mentor: Dr. Xin Li Apr 2023 - May 2025

Tencent AI Lab Shenzhen, China
Research Intern, Mentor: Dr. Yong Dai Oct 2022 - Feb 2023

HONORS & AWARDS

- National Scholarship (<1%) Sep 2022
- NUS Research Scholarship 2024 - 2028

TEACHING & REVIEW SERVICE

Teaching Assistant:

- *Machine Learning*, Sun Yat-sen University Fall 2022
- Instructor: Shangsong Liang

Reviewer:

- ACL Rolling Review, NeurIPS 2024, ICLR 2025, ICML 2025

Courses and Skills

- **Selected Courses:**
  - Mathematics: Advanced Mathematics, Linear Algebra, Probability & Mathematical Statistics, Discrete Mathematics, Mathematical and Interdisciplinary Modeling
  - Machine Learning: Machine Learning, Pattern Recognition
  - Computer Science: Computer Networks, Operating Systems, Computer Composition Principle
- **Programming languages & machine learning tools:**

  C++, Python, Pytorch, Triton
- **Languages:**

  Mandarin, English