Luning Wang

Email: wangluning2@gmail.com *Tel: (+1) 734-450-5036*

EDUCATION

University of Michigan **Ann Arbor, US**
M.S in Electrical and Computer Engineering *08/2024- 05/2026*

Tsinghua University **Beijing, China**
B.E. in Electronic Information Science and Technology (3.76/4.00) *09/2020- 06/2024*

INTERNSHIP EXPERIENCES

Infinigence AI **Beijing, China**
Algorithm Intern *02/2024- 06/2024*

- **Project: Training-Efficient Channel Shrinking for KV Cache in Long-Context Scenarios**
  - Independently designed and implemented an SVD-based channel reduction algorithm for KV cache in LLMs, which has achieved an overall compression ratio of 95% on multiple long-context tasks.
  - Responsible as the first author of the paper, which has been accepted by ENLSP NeurIPS Workshop 2024.

ByteDance Corporation **Beijing, China**
Algorithm Intern *09/2023- 01/2024*

- **Project: The Development of an Appeal Chatbot based on LLMs for TikTok Moderation System**
  - Developed the RAG component for the chatbot, and employed our RAG pipeline to enhance the generation of LLMs (Mistral, GPT-3.5, etc) for QA tasks, and achieved an improvement of 20% in accuracy on OpenBookQA dataset.
  - Contributed to the development of the explanation generation model. Implemented strategies (SFT, ICL, etc) with curated prompts, attaining an F1 score surpassing 70% in identifying violations within TikTok's moderation data.

RESEARCH EXPERIENCES

NICS Lab, Energy Efficient Computing Group (Tsinghua University) **Beijing, China**

- **Project: Evaluation of Quantized Large Language Models** *12/2023- 02/2024*
  - Responsible for experiments on evaluating the effect of quantization (Method: RTN, SmoothQuant, AWQ) on dialogue ability and trustworthiness of LLMs (LLaMA, Mistral, ChatGLM, etc), using popular benchmarks (MT-Bench, Adv-GLUE).
  - Responsible for the writing and rebuttal of the parts concerning dialogue ability and trustworthiness in our paper, which was accepted by ICML 2024.
- **Project: Low-Bit Quantization with Mixed Precision for Large Language Models** *03/2023- 09/2023*
  - Conducted sensitivity tests on LLMs (OPT, LLaMA, etc), gathering per-block and per-layer sensitivity data to guide subsequent mixed-bit quantization strategies.
  - Contributed to the experimental evaluation of our grouping and reordering quantization strategy, finally achieving an average bit-width of 2.8 bits without significant loss. Our paper was accepted by ENLSP NeurIPS Workshop 2023.

PUBLICATIONS

- [ENLSP NeurIPS Workshop'24] **"CSKV: Training-Efficient Channel Shrinking for KV Cache in Long-Context Scenarios"**. First Author.
- [ICML'24] **"Evaluating Quantized Large Language Models"**. Co-Author.
- [Arxiv'24] **"A Survey on Efficient Inference for Large Language Models".** Co-Author
- [ENLSP NeurIPS Workshop'23] **"LLM-MQ: Mixed-precision Quantization for Efficient LLM Deployment"**. Co-Author

SKILLS

- **Programming Languages：** Proficient in Python, Matlab. Have fundamental knowledge of C/C++, C#, Verilog, SQL, etc.
- **Software Tools：** Proficient in Linux, Git, PyTorch, Transformers, Latex, etc.

SELECTED HONORS & AWARDS

- Comprehensive Excellence Scholarship of Tsinghua University (Top 30% in major, 8000CNY) ***2022-2023***
- First Prize in the 5th 'Huiye Cup' Software Design Competition (Top 1, 5000 CNY) ***2021-2022***