github.com/ttttttris

Jiayi Tian

+1 (805) 245 0298 | jiayi_tian@ucsb.edu | linkedin.com/in/jiayi-tian-32b9652a5/

Focus on Efficient LLM Training & Inference, Efficient CoT Reasoning via Algorithm-System Co-Design.

EDUCATION

University of California, Santa Barbara, *Ph.D. in Computer Engineering* | CA, USA **3.93/4.0** Fall 2023 - ongoing
University of California, Santa Barbara, *M.S. in Computer Engineering* | CA, USA **3.93/4.0** Fall 2023 - Fall 2025
Nanjing University, *B.Eng. in Electrical Engineering* | China **4.50/5.0** Fall 2019 - Fall 2023

INDUSTRIAL EXPERIENCE

Intel Corporation, *Research Intern 2025* | Hillsboro, OR | **Mentor**: Souvik, Kundu Jun 2025 – Sep 2025

- Proposed SkipKV, a training-free KV-cache compression framework featuring sentence-level selective eviction and dynamic thoughts-type generation control for efficient chain-of-thought reasoning in multi-batch serving scenarios.
- Designed a semantic similarity metric to evict redundant sentences from the cache, improving reasoning coherence.
- Introduced a dynamic latent thoughts-type steering mechanism to enable concise and stable inference.
- Demonstrated strong results on long-reasoning tasks with LLMs: up to 26.7% higher accuracy, 1.6× shorter generation, and 1.7× higher throughput vs. SOTA under equal compression.
- Resulting paper accepted to MLSys, 2026.

Intel Corporation, *Research Intern 2024* | Hillsboro, OR | **Mentor**: Hai Li Jun 2024 - Sep 2024

- Proposed a tensor-compressed Transformer training accelerator on FPGA, optimizing compute ordering, dataflow, and memory allocation for LLMs.
- Designed a bidirectional tensor contraction scheme enabling substantial reduction in intermediate memory and compute cost during long-sequence training and inference.
- Built an HLS-based training engine achieving up to 51× memory efficiency and 4× energy efficiency compared with an Nvidia RTX 3090 GPU.
- Resulting paper accepted to IEEE TCAD, 2025.

AMD-Xilinx Technology, *Co-Op/Intern* | Beijing, China Jun 2023 - Sep 2023

- Developed a C++/HLS Transformer training framework with custom tensorized linear layers and nonlinear operations for LLM acceleration, achieved 30× $\sim$ 52× saving in model size for end-to-end Transformer training.

SKILLS & RESEARCH INTERESTS

Languages & Tools Python, PyTorch, Huggingface, vLLM, C/C++

ML & NLP Efficient Large Language Models (LLMs) Training/Inference, Efficient Large Reasoning Models (LRMs) (Model Compression, KV Cache Compression, Pruning, Low-rank decomposition, Early Exit, Knowledge Distillation, Quantization, Speculative Decoding/Reasoning), Efficient Agentic System

SELECTED PUBLICATIONS & PREPRINTS

RankGuide: Tensor-Rank-Guided Routing and Steering for Efficient Reasoning [paper] [code]
Jiayi Tian, Yupeng Su, Ryan Solgi, Souvik Kundu, Zheng Zhang, accepted to Third Conference on Language Modeling (COLM), 2026.

SkipKV: Selective Skipping of KV Generation and Storage for Efficient Inference with Large Reasoning Models [paper] [code]
Jiayi Tian, Seyedarmin Azizi, Yequan Zhao, Erfan Baghaei Potraghloo, Sean McPherson, Sharath Nittur Sridhar, Zhengyang Wang, Zheng Zhang, Massoud Pedram, Souvik Kundu, Ninth Annual Conference on Machine Learning and Systems (MLSys), 2026.

FLAT-LLM: Fine-grained Low-rank Activation Space Transformation for Large Language Model Compression [paper] [code]
Jiayi Tian, Ryan Solgi, Jinming Lu, Yifan Yang, Hai Li, Zheng Zhang, Findings of EACL, 2026.

Ultra Memory-Efficient On-FPGA Training of Transformers via Tensor-Compressed Optimization [paper]
Jiayi Tian, Jinming Lu, Hai Li, Xiangwei Wang, Cong (Callie) Hao, Ian Young, Zheng Zhang, IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems (TCAD), 2025.

BEBERT: Efficient and robust binary ensemble BERT [paper] [code]
Jiayi, Tian, Chao Fang, Haonan Wang, and Zhongfeng Wang, IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2023.

RESEARCH PROJECTS

Context Compression for Long-Horizon Coding Agents Apr 2026 – ongoing

- Developed ReFold, a training-free, reversible context-folding layer for long-horizon LLM coding agents that folds inter-turn redundancy in the rendered history while leaving the agent's memory, tools, and prompt untouched.
- As a drop-in plugin to an unmodified scaffold, it cuts the median instance's token cost by 34–49% and peak context by up to 63% across five coding benchmarks with no loss in task resolve.

Tensor-Rank-Guided Steering and Routing for High-performance SRM Reasoning Jan 2026. – Mar 2026.

- Proposed RankGuide, a framework that leverages tensor-rank signals from hidden states to accelerate LRMs reasoning, which can yield up to 1.75× and 1.36× latency benefit compared to LRM and SoTA collaborative inference framework, while maintaining or improving the accuracy.
- Designed a tensor-rank scoring metric on step-level hidden states to detect low-quality reasoning steps and selectively route them to larger models, improving the accuracy–latency trade-off.
- Developed a rank-based calibration pipeline that filters low-rank samples to construct high-quality steering vectors for inference-time hidden-state intervention, encouraging concise and stable reasoning.

Structural Pruning for Efficient LLM Inference via Low-rank Decomposition Aug 2024 - May 2025

- Developed FLAT-LLM, a training-free, fine-grained compression method that leverages the low-rank structure of the activation space to transform and compress the model weights.
- Introduced a novel training-free rank selection algorithm that allocates ranks using a greedy redistribution strategy and can be integrated with existing low-rank LLM compression pipelines.
- Achieved strong performance on LLaMA-2, 3 and Mistral models with minimal calibration overhead (within minutes), validated across language modeling and downstream tasks.

Binary-Quantized Ensemble LLM for Fast and Robust Language Model Inference Apr 2021 - Jun 2023

- Developed BEBERT, a novel quantization-ensemble strategy enabling efficient and accurate 1-bit BERT inference.
- Leveraged efficient knowledge distillation strategy for high training efficiency.
- Achieved 13× model size reduction and 15× compute savings over standard BERT with minimal accuracy loss.
- Proposed early-exit inference variant, further cutting compute by 20% $\sim$ 40% on GLUE benchmark.