

TAOLUE KURT YANG

taolue.yang@temple.edu — [Homepage](#) — [Google Scholar](#) — [GitHub](#) — [LinkedIn](#)

EDUCATION

Temple University , Philadelphia, PA Ph.D. in Computer Science Advisor: Prof. Sian Jin	2023 – July 2027 (expected)
Columbia University , New York, NY M.S. in Electrical Engineering	2022 – 2023
University of Glasgow , Glasgow, UK B.Eng. in Electrical Engineering, First Class Honours	2018 – 2022

RESEARCH INTERESTS

- High-performance computing and GPU-accelerated data compression
- Efficient AI systems: LLM quantization, KV-cache compression, efficient training and inference
- Bioinformatics and AI for science: pangenome data compression, genomic deep learning at scale
- Scientific data reduction: error-bounded lossy compression, error propagation, performance modeling
- Compressed-domain computing and data movement in parallel systems

RESEARCH EXPERIENCE

Graduate Research Assistant <i>Temple University, Advisor: Prof. Sian Jin</i>	2023 – Present <i>Philadelphia, PA</i>
• GPU-accelerated genomic data compression: designed CuVert-Q, GPUFASTQLZ, and GFaz, GPU compressors for sequencing reads and pangenome graphs that resolve the throughput-ratio tradeoff; built the SeqBench benchmark suite. All four are open source. • Computing over compressed data: pangenome analytics computed directly on grammar-compressed graphs, without decompression (under review at FAST'27). • Computation sharing for genomic deep learning: exact per-layer computation sharing for sequence models across thousands of haplotypes, $7.6\times$ faster than window deduplication and $584\times$ faster than standard inference on 1000 Genomes, with identical outputs. • LLM-aware lossy compression: co-developed KVComp, PackKV, and EmbdC, error-bounded compressors that shrink KV-cache and embedding memory for LLM inference while preserving accuracy. • Scientific data compression: TOPIQ (SC'26), Gaussian splatting particle compression (SC'26), point cloud compression (VLDB'26), cross-field prediction (HPDC'25), and performance modeling (DCC'25). • Interdisciplinary collaboration: work closely with Prof. Xinghua (Mindy) Shi (Temple, bioinformatics) on biological data compression (CuVert-Q, GFaz, SeqBench, DCC'26, FAST'27).	

INDUSTRY RESEARCH EXPERIENCE

Software Engineer Intern <i>Google, Cloud AI (Vertex AI)</i>	May 2026 – Aug 2026 <i>Sunnyvale, CA</i>
• Built a large-scale post-training quantization (PTQ) evaluation framework on GCP Slurm that benchmarks quantized LLMs for accuracy, KL divergence, and throughput on A3 and A4 GPUs. • Set hardware-aware PTQ deployment recipes for Vertex AI across GPU generations and model families. • Made quantization-aware RL training (GRPO) in NVIDIA NeMo-RL 70% faster by caching fake-quantized weights, and fixed a Megatron weight-folding bug in NVIDIA ModelOpt; both merged upstream. • Studied quantization across training, rollout, and the KV cache in RL post-training.	

Software Engineer Intern

Jul 2025 – Oct 2025

Google, AI and Infrastructure (MaxText)

Kirkland, WA

- Optimized Mixture-of-Experts (MoE) LLM inference throughput on TPUs, and implemented Mixture of Block Attention (MoBA) in JAX for training and inference.
- Built interleaved prefill and batch decode and an OpenAI-compatible API server for MaxText serving.
- Integrated **lm-evaluation-harness** and **Evalchemy** into MaxText for up to **30×** faster evaluation.

Software Engineer Intern

Jan 2025 – Jul 2025

Oracle, MySQL HeatWave

Redwood City, CA

- Developed a relaxed speculative decoding algorithm, **1.5×** faster LLM generation (patent filed).
- Built an 8-bit quantized LLM inference stack for AMD CPUs with AVX-512 and AMD BLIS int8 GEMM kernels, **4.5×** faster than llama.cpp on prefill.
- Designed a grouped 8-bit KV-cache layout (up to **20%** faster long-context prefill) and decode-phase weight quantization (**15%** higher throughput than Q4_K_M).

AI Engineer Intern

Jun 2023 – Sep 2023

Oortech

New York, NY

- Built a LangChain agent, data crawler, and data-loading pipeline for an LLM chatbot with semantic search.

PUBLICATIONS

Peer-Reviewed Conference Papers

- [C1] **Taolue Yang**, Y. Liu, B. Jiang, C. Li, X. Shi, S. Jin. “CuVert-Q: Resolving the Throughput-Ratio Tradeoff in Genomic Sequence Data Compression via GPU Acceleration.” *17th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics (ACM-BCB)*, 2026. **Best Paper Award.**
- [C2] **Taolue Yang**, Y. Liu, B. Jiang, X. Shi, S. Jin. “GFAz: State-of-the-Art Graphical Fragment Assembly Compression.” *40th ACM International Conference on Supercomputing (ICS)*, pp. 650–661, 2026.
- [C3] **Taolue Yang**, Y. Liu, C. Li, X. Shi, S. Jin. “SeqBench: A Benchmark Suite for Lossless and Lossy Compression of Sequence Data.” *15th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics (ACM-BCB)*, 2024.
- [C4] Y. Liu, B. Jiang, **Taolue Yang**, S. Di, R. Underwood, S. Jin. “TOPIQ: Statistical Error Propagation for Quantity-of-Interest Prediction under Lossy Compression.” *International Conference for High Performance Computing, Networking, Storage, and Analysis (SC)*, 2026.
- [C5] B. Jiang, Y. Liu, **Taolue Yang**, S. Di, S. Jin. “3D Gaussian Splatting for Scientific Particle Data Compression and Rendering.” *International Conference for High Performance Computing, Networking, Storage, and Analysis (SC)*, 2026.
- [C6] Y. Liu, L. Zhang, R. Li, B. Jiang, **Taolue Yang**, K. Zhao, S. Di, E. Dragut, S. Jin. “Error-Bounded Point Cloud Compression Using Truncated Octahedron Quantization.” *Proceedings of the VLDB Endowment (PVLDB)*, 19(11): 3105–3118, 2026.
- [C7] B. Jiang, **Taolue Yang**, Y. Liu, X. He, S. Di, S. Jin. “PackKV: Reducing KV Cache Memory Footprint through LLM-Aware Lossy Compression.” *IEEE International Parallel and Distributed Processing Symposium (IPDPS)*, 2026.
- [C8] B. Jiang, **Taolue Yang**, Y. Liu, X. He, S. Di, S. Jin. “EmbdC: Error-Bounded Lossy Video Embedding Compression for On-Device LLM Inference.” *26th IEEE International Symposium on Cluster, Cloud and Internet Computing (CCGrid)*, 2026.

- [C9] Y. Liu, W. Jia, **Taolue Yang**, B. Jiang, M. Yin, S. Jin. “Advancing Scientific Data Compression via Cross-Field Prediction.” *34th International Symposium on High-Performance Parallel and Distributed Computing (HPDC)*, 2025.

Peer-Reviewed Workshop Papers and Short Papers

- [W1] **Taolue Yang**, Y. Liu, B. Jiang, C. Li, X. Shi, S. Jin. “GPUFast-Q: Novel High-Throughput Sequence Data Compression on GPUs Using Fine-Grained Parallelism.” *Data Compression Conference (DCC)*, p. 479, 2026.
- [W2] **Taolue Yang**, Y. Liu, B. Jiang, S. Jin. “GPUFASTQLZ: An Ultra Fast Compression Methodology for Fastq Sequence Data on GPUs.” *SC24 Workshops (SC-W)*, 2024.
- [W3] Y. Liu, **Taolue Yang**, S. Jin. “Accurate Performance Modeling and Uncertainty Analysis of Lossy Compression in Scientific Applications.” *Data Compression Conference (DCC)*, p. 390, 2025.
- [W4] Y. Liu, W. Jia, **Taolue Yang**, M. Yin, S. Jin. “Enhancing Lossy Compression through Cross-Field Information for Scientific Applications.” *SC24 Workshops (SC-W)*, 2024.

Under Review

- [U1] **Taolue Yang**, Y. Liu, B. Jiang, X. Shi, S. Jin. “Computing over Compression: Pangenome Analytics Directly on Grammar-Compressed Graphs.” Submitted to *USENIX Conference on File and Storage Technologies (FAST)*, 2027.

In Preparation

- [M1] **Taolue Yang**, S. Jin. “Exact Per-Layer Computation Sharing for Convolutional Sequence Models Across Haplotypes.” In preparation, 2026.

Preprint

- [P1] B. Jiang, **Taolue Yang**, Y. Liu, C. Zhang, X. He, S. Jin. “KVComp: A High-Performance, LLM-Aware, Lossy Compression Framework for KV Cache.” *arXiv:2509.00579*, 2025.

Patents

- [T1] Relaxed speculative decoding for accelerated LLM token generation. Oracle Corporation, patent application filed, 2025.

HONORS AND AWARDS

Best Paper Award , ACM-BCB 2026 (for CuVert-Q)	2026
First Class Honours , University of Glasgow	2022

TEACHING EXPERIENCE

Teaching Assistant , Temple University	2024
• CIS 1068: Program Design and Abstraction – Spring 2024 (2 sections), Fall 2024 (1 section) First required programming course in the CS major: Java, OOP, searching and sorting. <i>Duties:</i> led weekly lab sections, taught algorithm topics (e.g., searching and sorting) and Java programming in lab, graded assignments and exams, and held weekly office hours. • CIS 2109: Database Management Systems – Fall 2024 (1 section) Undergraduate data modeling, SQL, and database application development. <i>Duties:</i> led weekly SQL and database design labs, graded assignments, and held weekly office hours.	

Teaching Interests

Undergraduate: introductory programming and data structures; computer systems and architecture; parallel programming; database systems.

Graduate: LLM systems (efficient training and inference, quantization, KV-cache optimization); HPC and GPU/CUDA programming; data compression for scientific and AI workloads.

OPEN-SOURCE CONTRIBUTIONS

NVIDIA NeMo-RL (upstream contributions)

- #3226: FP8 KV-cache scale propagation in the colocated refit path.
- #3441: Avoiding redundant weight fake-quantization during QAT logprob computation.
- #3556: Caching fake-quantized weights across QAT training microbatches.

NVIDIA ModelOpt

- #2131: Debugging `fold_weight` behavior with Megatron tied embeddings.

Research software: CuVert-Q, GFaz, GPUFASTQLZ (GPU compressors); SeqBench (benchmark).

Other projects: MaxText, vLLM, lm-evaluation-harness.

ACADEMIC SERVICE

Conference Reviewer (7 papers reviewed)

- ACDSA 2026 and 2027 (5 papers): Intl. Conf. on AI, Computer, Data Sciences, and Applications
- ICECET 2026 (2 papers): Intl. Conf. on Electrical, Computer and Energy Technologies

TECHNICAL SKILLS

Languages: C++, CUDA, Python **Frameworks:** PyTorch, JAX, Megatron, vLLM, SGLang