Yuanfan Chen

646-963-8426 | yuanfan0504@gmail.com | github | linkedin | scholar

Education

Cornell Tech, Cornell University — New York, NY
Master of Engineering in Computer Science (GPA: 3.9) — *Aug. 2025 – May. 2026*

University of Toronto — Toronto, Canada
Honours Bachelor of Science, Computer Science Specialist (GPA: 3.9) — *Sep. 2020 – May. 2024*

Publications

- **Beyond Prediction: Tail-Aware Scheduling for LLM Inference**. Yueying Li, **Yuanfan Chen**, Jiayang Chen, Esha Choukse, Haoran Qiu, G. Edward Suh, Rodrigo Fonseca, Ziv Scully, Udit Gupta. **ICML 2026**. arXiv:2606.18431.
- **Making Sense of DPU Performance for Cloud Data Processing**. Jiasheng Hu, Chihan Cui, **Yuanfan Chen**, Philip A. Bernstein, Jialin Li, Qizhen Zhang. **SoCC 2026** (17th ACM Symposium on Cloud Computing). arXiv:2504.05536.

Open Source

vLLM | Production bug fix: Fixed uninitialized padded fp8 MoE expert weights that silently corrupted outputs (#55251).

LMCache | Observability: Per-tier cache-hit metrics (#4734, #4962, #5842); co-developed a prefix-caching optimization (#5383).

AITER (ROCm) | Attention kernel tuning: Proposed a prefill attention config for Gemma-4 head-512 layers on MI325X (#5649, #5650).

SGLang-Omni | Multimodal inference optimization: Same-GPU weight sharing across replicas (#1124); placement planner for multi-stage speech serving (#2134).

Experience

AI Infra Engineer — Aug 2026 – Present
*Character.AI: **Inference Optimization*** — *Bay Area, CA*

- Production inference for multi-turn, long-context chat at **10K+ QPS** on **thousands of GPUs** including AMD Instinct, across models that differ in architecture, attention mechanism, and length distribution.
- Implemented **APC-aware KV-cache lookup** in LMCache, skipping lookup and L2 prefetch for prefixes already held in vLLM's automatic prefix cache and lowering **p50 TTFT by 16–25%**.
- Profiled AITER attention kernels with **AMD ROCm profilers** and traced Gemma-4 tail TTFT to head-512 attention layers at **∼6% MFU** (one workgroup exhausting a CU's 64 KB LDS, collapsing occupancy); retuned block/tile/warp configs for a **4.3× faster prefill kernel**, **32% faster cold prefill**, and **42–48% lower P99 TTFT** at equal load, doubling SLO-bound capacity per replica.
- Built a **production traffic replay tool** that reproduces KV-cache hit rate within the pod's own variance (metrics within **±3–10%**) by **replaying requests in real arrival time** with **session-aware multi-turn replay**; reduced external-KV warmup from **12 h to 3 h** with prefill-only closed-loop admission, providing the validation basis for every optimization that followed.

Machine Learning System Researcher – S4AI Lab — Sep 2025 – Present
Cornell University — *New York, NY*

- Designed a **measurement-driven placement planner** for multi-stage speech serving that sets stage placement, memory split, replica count, and GPU-sharing mode from two short probes per plan instead of a full sweep; **+40%** and **+60%** over shipped configurations.
- Applied γ-Boost queuing theory to design a **prediction-free, tail-aware scheduler** with KV-cache-aware preemption; reduced **tail TTFT/TTLT by 37–60%** while improving throughput by **1.01–1.12×** on Azure production traces (**ICML 2026**).
- Designed **PTA** (Prefill→Think→Answer) three-stage disaggregation on SGLang, giving each stage its own GPU pool, **DVFS frequency**, and **KV-cache budget**; versus prefill–decode disaggregation it cuts **P90 time-between-answer-tokens by 49–71%** and keeps decode queueing near zero under a 40% power cap (**12.5× lower end-to-end latency**) for ∼0.1% extra power.
- Formalized bounded user latency slack as a **Flex service tier** (the "slow mode" commercial APIs already sell) and built a **KKT-based online solver** that converts that slack into per-stage frequency and GPU settings within 7.7 ms of each budget change; at a 30% power cap it sustains **78.3% online goodput vs. 47.6%** for the best baseline (1.64×) with latency-critical tails within 1.3× nominal.

AI Infra Engineer — May 2025 – Jul 2025
*Tencent: **Distributed Training and Inference*** — *Beijing, CN*

- Designed **request-level admission control** with **batch-size stepwise profiling** for latency estimation and **SLA-aware priority scheduling**; maintained **100% goodput** under **2–10× QPS bursts** by enforcing per-class latency budgets across mixed-priority traffic.
- Optimized CUDA attention kernels for **GQA** (A100) and **DeepSeek FlashMLA** sparse attention (H200); profiled with **Nsight Compute** to localize memory-hierarchy bottlenecks and redesigned thread-block tiling, shared-memory layout, and warp scheduling; achieved **∼20% throughput gain** on GQA and **∼15%** on FlashMLA over baselines.

Undergraduate Student Researcher – Far Data Lab — Jan 2024 – Feb 2025
University of Toronto — *Toronto, CA*

- Co-designed **dpBento**, a modular benchmark framework that standardizes performance evaluation across heterogeneous DPUs (**BlueField-2/3**, **OCTEON TX2**), from micro-level primitives to end-to-end cloud database offloading.
- Profiled **RocksDB** compaction and decompression paths end-to-end; offloaded compute-heavy stages to DPU, reducing **host CPU utilization by ∼40%** and achieving **∼3× decompression throughput** via hardware acceleration.

Academic Service

- **Invited Reviewer**: IEEE Journal on Selected Areas in Information Theory (JSAIT) – Special Issue on Energy and Data Efficiency in AI (2026). Reviewed submissions focusing on cutting-edge LLM inference systems.