Teerth Sharma

Systems Researcher • High-Performance Computing • AI Infrastructure

Jaipur, India
teerths57@gmail.com
LinkedIn GitHub ResearchGate

SUMMARY

Systems-obsessed researcher specializing in **hardware-aware AI optimization**, **kernel development**, and **sparse attention mechanisms**. Contributor to major open-source ecosystems (**NVIDIA TensorRT-LLM**, **Meta xFormers**, **Google DeepMind Penzai**). Authored research on adaptive event-driven rendering achieving **up to 4.9× speedups** in LLM inference on consumer hardware.

TECHNICAL SKILLS

Languages
`Rust`, `C++` (CUDA), Python (PyTorch/MLX), Assembly, TypeScript

Systems & AI
NVIDIA Triton, TensorRT-LLM, Apple Metal, WebGPU, WASM, Docker

Core Competencies
Kernel Optimization, Sparse Attention, DSP, Latency Optimization

RESEARCH & PUBLICATIONS

A.E.T.H.E.R.: Adaptive Event-driven Threshold & Hybrid Entangled Rendering

PREPRINT

December 2025 • DOI: 10.13140/RG.2.2.22443.91687

Proposed a hierarchical sparse attention mechanism using a **"Geometric Gate"** based on Cauchy-Schwarz inequalities to prune redundant KV-cache blocks. Achieves **up to 80% block-level sparsity** without accuracy degradation.

250+ Reads Top Percentile Research Interest (2025)

OPEN SOURCE CONTRIBUTIONS

NVIDIA TensorRT-LLM • PR #10305

Contributor

Implemented dual-stage Triton kernel pipeline ("Event Radar" & "Selective Execution") for KV-cache bandwidth optimization.

Performance: `4.72×–4.90×` speedup on Llama-3-8B (16k context) on RTX 4060

`Python` `PyTorch` `Triton Kernels`

Meta xFormers • PR #1370

Contributor

Integrated AETHER geometric sparse attention operator with adaptive block pruning for memory-efficient transformers.

`Python` `CUDA` `Triton`

Google DeepMind Penzai • PR #134

Contributor

Implemented geometric sparse attention module with JAX-compatible primitives for neural network visualization.

`JAX` `Python`

ENGINEERING PROJECTS

NeuralWhisper: The Living Sanctuary

Creator

Research-grade neural audio synthesis with Gemini 1.5 Flash and Kokoro-82M. Custom Rust/WASM DSP stack (116× faster than JS) for real-time psychoacoustic analysis.

`Rust` `WebGPU` `WASM` `Gemini API`

AETHER-Link

Creator

High-performance I/O prefetch kernel for HFT and gaming. Predicts OS page cache bypass in sub-15ns. 65M ops/sec with zero allocations.

`Rust` `Systems Programming`

EXPERIENCE

Chief Technology Officer & Co-Founder — Jun 2025 – Present
KarmicSphere Media • Jaipur, India
Leading technical strategy, infrastructure, and Linux-based system deployments.

Writer (Self-Employed) — Jun 2021 – Jun 2025
Remote
Published writer specializing in creative and technical web content.

EDUCATION

B.Tech, Computer Science
Manipal University Jaipur
2024 – 2028 (Expected)
LLM Fine-tuning • Business Analysis • HPC

RECOGNITION

🚀 **Founding Engineer Offer**
Received founding engineer offer from a South East Asian unicorn startup

🌟 **Industry Recognition**
Research and GitHub work attracted recruiting interest from FAANG companies