TriAttention: 10.7x Memory Reduction for LLM Reasoning With No Accuracy Loss

Kabui, Charles

2026-04-08

Read at ToKnow.ai

Researchers from MIT, NVIDIA, and Zhejiang University released TriAttention, a KV cache compression method that reduces memory by 10.7x without hurting reasoning accuracy. The KV cache is the memory buffer LLMs use to store context during generation. It grows with sequence length, and when reasoning models produce 32K+ token chains of thought, the cache

can exceed the model weights in memory. Existing compression methods estimate token importance from recent attention scores, but those scores shift unpredictably due to RoPE (rotary position encoding, a standard way transformers track word order). TriAttention works in the pre-RoPE space instead, where query and key vectors cluster around fixed non-zero centers. These centers determine which token distances matter via a trigonometric series, giving a stable importance signal. On AIME25 with 32K-token generation, TriAttention matches full attention accuracy (40.8%) at 2.5x higher throughput, with peak speedups reaching 6.3x on MATH-500.

A 32B reasoning model running full attention can't fit on a consumer 24GB GPU once context exceeds a few thousand tokens. TriAttention changes that: the team demonstrated OpenClaw running a 32B model on a single RTX 4090 where full attention caused out-of-memory errors. The method plugs directly into vLLM as a transparent plugin with no code changes required. For anyone self-hosting reasoning models, this turns a multi-GPU setup into a single-card deployment.

Where Google's TurboQuant achieved 6x memory reduction through quantization, TriAttention gets 10.7x by exploiting geometric structure in the attention space itself. Both point the same direction: algorithmic efficiency, not bigger hardware, is driving the sharpest drops in inference cost. Training tricks come and go, but math that compresses without loss tends to stick.

Sources:

- TriAttention Paper (arXiv)
- TriAttention GitHub
- TriAttention Project Page
- HuggingFace Daily Papers
