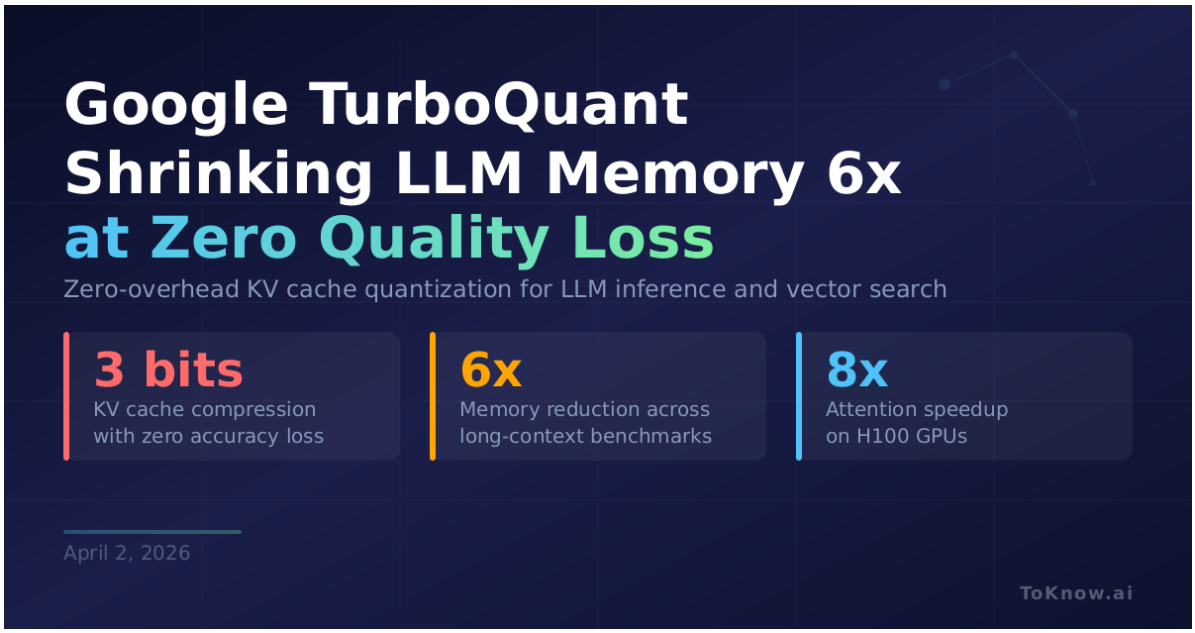
Google TurboQuant: Shrinking LLM Memory 6x at 3 Bits With Zero Quality Loss

Kabui, Charles

2026-04-02

Read at ToKnow.ai



Google Research introduced TurboQuant, a compression algorithm that shrinks the KV cache (the working memory LLMs use to store context during generation) to just 3 bits per value with zero accuracy loss and no retraining. It combines two techniques. PolarQuant converts key-value vectors from standard coordinates to polar coordinates, exploiting the predictable distribution of angles to skip the normalization step that normally adds 1-2 extra bits of

overhead. QJL then applies a 1-bit mathematical transform to the leftover error, correcting estimation bias at near-zero cost. Tested on Gemma and Mistral across LongBench, Needle-in-Haystack, RULER, and ZeroSCROLLS, TurboQuant achieves over 6x memory reduction while maintaining perfect scores on needle-in-haystack retrieval tasks. On H100 GPUs, 4-bit TurboQuant computes attention up to 8x faster than unquantized baselines. All three papers include theoretical proofs showing the methods operate within a 2.7x factor of information-theoretic lower bounds.

The KV cache is the main memory bottleneck when serving LLMs, especially with long context windows. A 6x reduction means a server handling 10 concurrent users can now handle 60, or a model that needed multiple GPUs fits on one. TurboQuant is data-oblivious and needs no fine-tuning, so it applies to any LLM deployment out of the box. The same techniques also speed up vector search index building, making large-scale semantic search cheaper.

The pattern here matches other recent efficiency work: mathematical structure, not bigger hardware, is unlocking the next round of deployment gains. Between Nemotron-Cascade 2's 20x parameter reduction through mixture-of-experts and TurboQuant's 6x memory compression, the cost of running frontier-level AI is dropping faster through algorithms than through chip improvements.

Sources: - Google Research Blog: TurboQuant - TurboQuant Paper (ICLR 2026) - PolarQuant Paper (AISTATS 2026) - QJL Paper (AAAI)
