LIGHTHOUSE CANTON
MARKET COMMENTARY · AI INFRASTRUCTURE · MEMORY SEMICONDUCTORS

DATE 26 Mar 2026
CATEGORY Thematic Note

Google's TurboQuant: Why It Changes Nothing for the Memory Trade

Addressing the sell-off in memory names following Google Research's KV cache compression paper

> *A year-old research paper with no production deployment does not alter the structural case for memory as AI's binding constraint.*

Context

On 25 March 2026, Google Research published a blog post highlighting **TurboQuant**, a compression algorithm that quantises the key-value (KV) cache used during large language model (LLM) inference down to 3 bits per channel, claiming a **6× reduction in KV cache memory** and up to an **8× speedup in attention-logit computation** on NVIDIA H100 GPUs — with zero accuracy loss. The announcement triggered a swift sell-off across memory semiconductor names: Samsung Electronics fell ~4.7%, SK Hynix dropped ~6.2%, and Micron declined ~3.4%, extending a five-session losing streak. Cloudflare's CEO called it "Google's DeepSeek moment." The internet compared it to Pied Piper from HBO's *Silicon Valley*.

Our view: **the market reaction is disproportionate to the substance of the development.** Below we lay out four reasons why TurboQuant does not change the structural investment case for memory.

Market Reaction Snapshot (25–26 March 2026)

COMPANY	TICKER	MOVE	NOTES
Samsung Electronics	005930.KS	−4.7%	KOSPI down >3% on the day
SK Hynix	000660.KS	−6.2%	Top traders net buying the dip
Micron Technology	MU	−3.4%	5th consecutive red session; compounded by capex concerns post-earnings
SanDisk / WD	SNDK / WDC	−1.6% to −3.5%	Sympathy sell-off
Lam Research	LRCX	−3.0%	Equipment sector contagion

Our View: Four Reasons This Changes Nothing

1 This Paper Is Almost a Year Old — Not a New Breakthrough

The underlying TurboQuant research first appeared on arXiv in April 2025. Its companion algorithms — PolarQuant and QJL — were published even earlier (AAAI 2025 and AISTATS 2026 respectively). What happened this week is simply that Google Research re-featured the paper on its blog ahead of the formal ICLR 2026 presentation in late April. The broader investment community picked up on the blog post and amplified it, but the science itself has been in the public domain for nearly twelve months. Crucially, Google has not released any official code, library, or integration — community implementations exist but remain early-stage and not production-ready. The technology has not been confirmed as running in any Google production system, whether Gemini, Google Search, or Cloud inference. If TurboQuant were truly game-changing, the question is: why has Google itself not deployed it widely in the year since publication?

2 Theoretical Compression ≠ Free Savings — There Are Real Trade-offs

Compression is never free. While TurboQuant reduces the memory footprint of the KV cache by quantising 32-bit floating-point values down to 3 bits, the compressed data must still be decompressed and reconstructed at inference time to compute attention scores. The TurboQuant pipeline involves a polar coordinate transformation (PolarQuant) followed by a Johnson-Lindenstrauss error-correction pass (QJL) — both adding computational overhead. The claimed "8× speedup" applies narrowly to attention-logit computation only, not to end-to-end inference throughput. Real-world wall-clock gains will be materially lower. Furthermore, Google's benchmarks were conducted exclusively on small models (≤ 8B parameters); whether the "zero accuracy loss" claim holds at 70B+ scale, on mixture-of-experts architectures, or at million-token context windows remains entirely undemonstrated. Any production deployment would also need to factor in the power and compute cost of running the decompression pipeline continuously, partially offsetting the memory savings.

3 KV Cache Is Only One Slice of the Memory Demand Stack

TurboQuant targets exclusively the key-value cache used during inference — the temporary memory that stores attention states for ongoing conversations. This is a genuine bottleneck for long-context serving, but it is only one component of the broader AI memory demand picture. It does not address:

Model weight storage: The parameters of a large model must be stored in full precision (or near-full precision) on HBM (high-bandwidth memory). A 70B-parameter model requires ~140 GB of HBM just for weights alone — TurboQuant offers zero relief here. As models continue to scale under prevailing scaling laws, weight storage demand only grows.

Training memory requirements: Training runs for frontier models consume orders of magnitude more memory than inference, driven by activations, gradients, and optimiser states. TurboQuant is a purely inference-time technique and has no bearing whatsoever on the massive memory buildout required for training the next generation of models.

Agentic and multi-modal workloads: The shift toward agentic AI, multi-modal models processing video, images, and audio alongside text, and retrieval-augmented generation (RAG) pipelines are all driving explosive growth in memory demand well beyond the KV cache.

In short, even if TurboQuant delivered its full theoretical promise in production, it would only alleviate one narrow segment of memory demand while leaving the dominant drivers — weight storage, training, and next-generation workloads — entirely untouched.

4 Jevons' Paradox — Efficiency Drives Demand, Not Destruction

History is unambiguous on this point. When a resource becomes cheaper and more efficient to use, total consumption of that resource tends to increase, not decrease. This is Jevons' Paradox, and it has applied consistently across technology cycles: cheaper compute led to more computing, not less; cheaper storage led to more data, not less; cheaper bandwidth led to more streaming, not less.

If TurboQuant (or any successor compression technique) were to meaningfully reduce the cost of long-context inference, the immediate second-order effect would be a dramatic expansion of AI usage. Applications previously gated by memory cost — such as persistent-memory agents, real-time multi-user AI services, and local deployment on consumer hardware — would become economically viable, creating an entirely new layer of memory demand. Multiple sell-side analysts have already flagged this dynamic: Wells Fargo noted in an investor note that the Jevons' Paradox framework suggests TurboQuant could ultimately be a positive for memory companies, not a headwind. DS Investment & Securities echoed the view, arguing that technologies reducing memory usage tend to expand total demand by lowering the cost of AI adoption.

The DeepSeek analogy — which commentators have invoked — is itself instructive. When DeepSeek demonstrated cheaper training methods, the market initially sold off AI infrastructure names. What followed was a rapid acceleration in AI adoption and an even larger infrastructure buildout.

Conclusion

Our position is unchanged. Memory remains one of the most critical bottlenecks in the AI infrastructure stack. TurboQuant is a year-old research paper — not a deployed technology — that addresses only one narrow slice of memory demand (inference-time KV cache), introduces real computational trade-offs, and has not been adopted even by its own creator. The market sell-off in Samsung, SK Hynix, and Micron presents, in our view, a reaction disproportionate to the substance of the announcement. If anything, the fact that researchers are working hard to compress memory usage is itself evidence that **memory scarcity is the binding constraint** they are trying to engineer around. We continue to view memory as a core long-term beneficiary of the AI infrastructure cycle.

LIGHTHOUSE CANTON

Singapore

16 Collyer Quay #11-02
Collyer Quay Centre
Singapore 049318

+65 67130570

UAE

The Exchange Gate Village 11,
Unit 802 Dubai International
Financial Centre PO Box 507026
Dubai, UAE

+971 45 861500

UK

24 Hanover Square,
London W1S 1JD

+44 164 2843 487

India

Unit No-104A,Worldmark
2 Asset Delhi Aerocity,
New Delhi 110037

+91 9650473961

Unit No 507/508, A Wing,
INS Tower, G Block, BKC,
Mumbai– 400051

+91 9650473961

1st Floor, WeWork37,
Cunningham Cross Rd,
SRT Road Vasant Nagar,
Bengaluru-560001

+91 9900096873

RK Swamy Centre, Hansa
Building, Door No:3, Thousand
Lights,
Chennai-600006

+91 9650473961

Suite 502, Building 450, Central
Plaza, Genome Valley,
Shameerpet,
Hyderabad 500 078

+91 9650473961

Unit No FF-10, FF Floor, Pragya
Accelerator, Block 15T GIFT CITY,
Gandhinagar
Gujrat-382355

+91 9650473961

info@lighthouse-canton.com

Lighthouse Canton

DISCLAIMER