

TurboQuant-Pro: Keep the Geometry

Task-Geometry-Aware Compression for Embeddings and KV Caches,
with Certificates Instead of Promises

Andrew H. Bond

San José State University

`andrew.bond@sjsu.edu` ORCID 0009-0003-2599-6158

Working draft — v0.1

Abstract

Vector compression is usually evaluated by reconstruction fidelity and shipped on a promise. This paper describes TurboQuant-Pro, an open-source compression system for embedding retrieval and LLM KV caches, organized around a single principle learned the hard way and then proved: *decompose scale from direction, and keep whichever part carries the downstream task’s geometry — then certify it, don’t promise it*. The system contributes, on the practice side: a PCA-Matryoshka + scalar-quantization retrieval pipeline with compressed-domain search ($27\times$ storage at high recall); the discovery that per-vector polar quantization of attention *keys* passes reconstruction benchmarks while destroying generation (perplexity $\sim 10^4$ at cosine 0.995) and the per-channel architecture that fixes it; asymmetric NormalFloat (asym-NF4), one calibration-free codebook that is near-fp16 on every architecture tested where symmetric NF4 silently collapses on high-ratio GQA models; the measurement that the key DC offset absorbing that zero-point is RoPE-frequency-structured (Spearman 0.91–0.98 against rotary wavelength; 96–99% of offset mass in channels whose wavelength exceeds the window), yielding a *deterministic, zero-calibration zero-point* computed from weights and config alone that matches calibrated quality on perplexity and beats it on LongBench; and a fused decode kernel family that computes attention directly on compressed codes — including, new here, per-channel keys with dense-sparse fp16 outliers fused as sparse score deltas (exact to $8e-8$, $2-6\times$ over decompress-then-attend). On the guarantees side, imported from a companion theory paper: distribution-free *rank certificates* (Kendall $\tau \geq 1 - 2\hat{\mu}(\kappa)$ for any corpus, with vacuity as a principled “rerank required” signal), and a consumer-metric probe that reproduces the keys catastrophe at calibration time. Two further empirical laws — a density-quotient dissociation (quotient density only when it is nuisance for the task truth; confirmed out-of-sample on public data) and a spectral boundary for bit-tapering (effective rank vs. bit-head width) — complete a methodology in which every production incident becomes an installed instrument, every claim carries an evidence level, and negative results are published next to the wins they bound.

1 Introduction: the reconstruction-metric fallacy

The central failure this system is built around is easy to state: a quantizer was near-perfect by every reconstruction metric — 0.095 mean relative error, cosine similarity 0.995 — and raised a 7B model’s perplexity from 12.2 to 10,643. The tensor was the attention *key* cache; the quantizer preserved each key’s norm and quantized its direction; and the consumer, $\text{softmax}(QK^\top)$, reads per-channel scale structure that per-vector normalization deletes. Reconstruction fidelity and task fidelity are not merely loosely coupled — on keys they were *anti-correlated* across our quantizer family (§3).

The constructive lesson generalizes. Compression is the choice of a quotient: some coordinates are kept, some are discarded or coarsened. The correct quotient is determined not by the data’s statistics but by the *downstream metric* — what we call the task geometry. For cosine retrieval over isotropic embeddings, direction is the geometry and per-vector scale is disposable (store it, cheaply, exactly). For attention keys, per-channel scale *is* the geometry. For graph geodesics — the setting of the companion theory paper [1] — the angular coordinate carries everything and the radius is provably degree noise. One principle covers all three: **keep the geometry; sometimes that is the angle, sometimes angle plus norm, sometimes per-channel scale**. The companion paper makes the boundary precise (its condition (A2): a scale-discarding quotient is safe exactly when the consumer’s metric is carried by the tangential part of the displacement); this paper is the practice half — the systems, the discoveries the principle organizes, and the instruments that keep the next violation from shipping.

Contributions.

1. **A two-track compression system (§2)**: embedding/vector-DB compression with compressed-domain (ADC) search, and KV-cache compression with an asymmetric-by-quantizer key/value architecture; 528 tests, evidence-laddered claims, reproducible harnesses.
2. **The keys finding (§3)**: reconstruction metrics cannot gate key quantization; per-channel scale is the consumer geometry.
3. **asym-NF4 and the GQA collapse mechanism (§4)**: one calibration-free codebook robust across MHA and high-GQA architectures, with the failure it fixes traced end-to-end.
4. **The RoPE-frequency structure of key offsets and the deterministic zero-point (§5)**: the offset lives in config-identifiable slow rotary channels; a zero-point computed from weights + config alone ties calibrated quality on perplexity and beats it on LongBench-qasper (43.36 vs. 42.35; fp16 43.77).
5. **Certificates instead of promises (§6)**: distribution-free rank-agreement floors from measured distortion and corpus concentration; a calibration-time consumer-metric probe; a streaming drift guardrail — the theory paper’s results shipped as product features.
6. **Two empirical laws with their boundaries (§7)**: the density-quotient dissociation (confirmed out-of-sample) and the spectral effective-rank boundary for bit-tapering (a synthetic win that honestly reversed on public data).
7. **Compute-on-codes decode (§8)**: a fused attention kernel family over compressed KV, culminating in per-channel keys with dense-sparse outliers fused as warp-cooperative sparse score deltas — no reconstruction, no scatter, no dense-loop branching.
8. **A methodology (§9)**: evidence ladder, pre-registration, incident-to-instrument.

2 The system

Track 1 (embeddings). A PCA-Matryoshka rotation (PCA recovers the ordered eigenfunctions Matryoshka training targets [6]) makes non-Matryoshka models truncatable without retraining; a random-rotation scalar quantizer (TurboQuant [3]) codes the unit direction; the norm is stored beside the codes. Retrieval runs *compressed-domain*: an ADC index scores queries directly against

codes (AVX2 kernel), with exact reranking at a fixed oversample as the standard protocol. Headline operating point: $27\times$ storage at high recall@10, beating RaBitQ on recall and tying OPQ at $4\text{--}20\times$ lower index-build cost, under one CI-tested harness on public data with provided ground truth. The honest scope is part of the claim: PCA truncation wins on concentrated-spectrum encoders and loses to PQ/OPQ on compact descriptors; at matched bytes the scalar quantizer alone still wins or ties.

Track 2 (KV caches). A two-tier streaming cache (fp16 hot window, compressed cold pages) that is *asymmetric by quantizer*: values through the polar flow (near-lossless), keys through PerChannelKV (§3–5), with dense-sparse fp16 outliers and an attention sink. Model-aware configuration (AutoConfig) selects bits, RoPE-aware boosting, and — since the zero-point work — the key zero-point mode per family. A fused CUDA decode family computes attention on the codes (§8).

3 The keys finding

Quantizing post-RoPE keys with the per-vector polar flow (rotation, L2 normalization, Lloyd-Max codebook) yields the best reconstruction error of our key-quantizer family and the worst generation quality by three orders of magnitude (Qwen2.5-1.5B, WikiText-2: ppl 10,643 vs. fp16 12.24), while a per-channel uniform quantizer with $1.5\times$ the reconstruction error is near-fp16. The mechanism is geometric: attention logits are bilinear in the keys, so what must survive is per-channel scale structure across the token axis, precisely the coordinate per-vector normalization quotients away. Two system-level consequences: (i) PerChannelKV — per-channel asymmetric scales over tokens, optional non-uniform levels, dense-sparse outliers — is the shipped key path; (ii) no reconstruction metric gates any key change in this project; generation (perplexity, then task scores) is the gate. The finding is also a boundary result for the companion theory: keys are the counterexample to “always keep the angle,” cited there as the scope of the polar move [1].

4 asym-NF4 and the GQA collapse

The calibration-free NF4 codebook, a common default, *silently collapses* on high-ratio GQA models: Qwen2.5-7B (7:1) drops LongBench-qasper from 43.8 (fp16) to 4.7 and WikiText-2 perplexity from 7.46 to 74.7, while the same recipe is near-lossless on Llama-2 and Mistral — invisible if one benchmarks only the Llama family. The causal chain, each link measured: KV keys carry a large per-channel DC offset, so NF4’s zero-centred abs-max grid wastes half its codes on the empty side (representation); NF4’s key error is *equal* on Qwen and Llama, so the differentiator is error *tolerance* — a 7:1 GQA model routes each KV-head’s error into seven query heads (amplification). The fix is one per-channel zero-point: **asym-NF4** centres the grid before applying the NF levels — best-ordered on every model tested at no extra bit cost (Qwen qasper recovered to 41.9), calibration-free, and a dead heat with KVQuant’s Fisher-calibrated NUQ on LongBench. Details and the cross-model matrix are in the satellite paper [2].

5 The offset is RoPE geometry; the zero-point is free

Where does the DC offset live? Measuring post-RoPE keys across Qwen2.5 1.5B/7B/14B and Mistral-7B: the per-channel offset magnitude correlates with the channel’s rotary wavelength at pooled Spearman 0.91–0.98, with 96–99% of the offset mass in the channels whose wavelength

exceeds the context window (67% of channels at $\theta=10^6$, 52% at 10^4). The mechanism: a rotary channel slower than the window is position-near-invariant within it, so any pre-RoPE per-channel mean survives rotation as DC, while fast channels average it away. This closes the §4 chain (structural DC $\rightarrow$ codebook waste $\rightarrow$ GQA amplification) and makes the zero-point computable *without data*: pushing the model’s own `k_proj` bias through the position-averaged rotation gives a zero-point from weights + config alone. Validated twice: WikiText-2 perplexity 12.533/9.179 (Qwen2.5-1.5B/7B) vs. 12.534/9.155 calibrated; and LongBench-qasper **43.36** vs. 42.35 calibrated in the same harness (fp16 43.77, symmetric NF4 4.69) — the zero-calibration zero-point *beats* dense calibration on the collapse-sensitive task. A second lean mode stores calibrated means only on the config-identified slow channels (one-third less metadata, 42.66 qasper). Both ship in `PerChannelKV` (`zero_point="bias"/"sparse"`), plumbed through the streaming cache with per-block absolute positions, and “bias” is the Qwen-family default in `AutoConfig`. Scope: the bias route requires key-projection biases (the Qwen family); bias-free models keep the calibrated or sparse modes.

6 Certificates instead of promises

Track 1’s quality story was, until recently, measured recall plus cosine — promises. The companion theory paper’s rank proposition converts to a shipped floor: for any corpus, measure the robust distortion κ (97.5/2.5 percentile ratio of compressed to exact distance) and the corpus’s distance-ratio concentration $\hat{\mu}(\kappa)$ (one pass, $O(P \log P)$); then *with no distributional assumptions*, Kendall $\tau \geq 1 - 2\hat{\mu}(\kappa)$ and Spearman $\rho \geq 1 - 3\hat{\mu}(\kappa)$ [1, 7]. The floor’s *vacuity* is itself the product feature: on distance-concentrated corpora no finite distortion certifies rank — exactly when single-stage retrieval dies — so “exact reranking required” becomes a derived, per-corpus signal rather than a fixed oversampling menu; `autotune` reports $\kappa/\hat{\mu}/\tau$ -floor per operating point. Guarding the other track: an (A2) *consumer-metric probe* quantizes a calibration batch under each candidate family’s information quotient and compares rank agreement of the declared consumer’s scores (cosine, L2, attention logits) — it reproduces the §3 catastrophe on synthetic key statistics as a unit test — and a streaming tangential-fraction statistic in the quality monitor turns norm-dominated drift into a dashboard alert. The probe work surfaced a distinction worth stating: there are *two* failure classes — radial displacement (hubness; the streaming statistic sees it) and direction concentration (the keys regime; displacements are fully tangential yet squeezed below the quantizer’s cell size — only the end-to-end probe sees it).

7 Two empirical laws, with their boundaries

Density dissociation. A per-vector density quotient on compressed-domain candidates (the retrieval analogue of the theory paper’s $\alpha=1$ normalization) helps *iff* density is nuisance for the task truth: +15.7 recall points — beating exact reranking, which cannot remove the nuisance — when an anisotropic common mean is excluded from the ground truth, and −28 points when it is part of it. The public-data replication (BGE-M3 over WikiText-2, observed-space ground truth) confirmed both predictions out-of-sample and isolated the shippable form: correct only the *compression-induced* density shift ($\mu_{\text{recon}} - \mu_{\text{orig}}$, nuisance by construction, computable at index build): +0.3 recall points where the plain quotient loses 5.9. **Spectral boundary for bit-tapering.** A pre-registered prediction from the theory paper’s heat-filter theorem — exponentially tapered bit schedules beat hard truncation at matched total bits — held on a synthetic power-law corpus (recall +1.6 pts at constrained budgets; certified τ -floor non-vacuous at half the budget) and *reversed* on real BGE-M3 embeddings. The reversal is quantified: the synthetic spectrum has effective rank

13.6 (tail dims near-noise; cheap taper bits free), BGE has 136.7 (mid-spectrum dims carry real variance; coarse tail bits inject real noise). The rule — *taper wins iff spectral effective rank is far below the bit-budget’s head width* — is a one-number diagnostic from the covariance the pipeline already estimates.

8 Compute on the codes: fused decode

One idea serves both tracks: *compute on the codes, rotate at the boundaries*. In retrieval it is the ADC index. In decode it is a kernel family that computes one attention step directly from compressed K/V — per-query LUT, online (flash) softmax over ADC scores, value accumulation in code space, a single inverse rotation — validated *exact* against the dequantize-then-attend path at every milestone and 1.8–13.3× faster at 2k–32k context (split-K, one warp per (head, key-split), shuffle reductions). New in this draft, the recommended *key* format is fused too: since keys enter attention only through $q \cdot k$, the whole per-channel format becomes score terms —

$$\text{score}_s = \underbrace{q \cdot \mu(h)}_{\text{zero-point bias, folded once}} + \underbrace{\sum_j (q_j a_j) \text{grid}[c_{js}]}_{\text{dense, branch-free}} + \underbrace{\sum_{j \in \text{out}(s)} q_j \Delta_{sj}}_{\text{sparse CSR deltas}}$$

with outliers stored token-major as *deltas* against the dense dequant, applied by a warp-cooperative pass (lanes stride the ~ 3 -entry row, one shuffle reduction folds the correction in before the softmax update). Memory contiguity comes from a one-off build-time channel-major $\rightarrow$ token-major transpose; branch divergence is bounded to the ragged tail of a short list and never touches the dense loop. All three zero-point modes fold into the same bias term — the kernel never branches on the mode. Exact to 8×10^{-8} against decompress-then-attend across every key variant, and 2–6× faster end-to-end even with per-call CSR rebuild (the cache integration amortizes it per cold flush).

9 Methodology as a contribution

Three habits made the results above auditable. **Evidence ladder:** every headline claim carries a level (L1 public one-click reproduction through L5 unit-tested engineering) and links a runnable notebook; L4/L5 is a disclosure, not a weakness. **Pre-registration:** directional predictions (the taper test, the dissociation replication) are written down before the run, so a reversal is a documented boundary rather than a quiet omission. **Incident to instrument:** the keys catastrophe became the consumer-metric probe; the GQA collapse became asym-NF4 and then the deterministic zero-point; the cosine-vs-recall divergence became rank certificates. The system’s guarantees section exists because its failures were kept.

10 Related work

TurboQuant (online vector quantization) [3] supplies the rotation-and-scalar-quantize primitive; QLoRA’s NF4 [4] the codebook asym-NF4 repairs; KVQuant [5] the calibrated per-channel NUQ that asym-NF4 matches calibration-free; RaBitQ/OPQ/PQ the retrieval baselines under one harness. Varici et al. [6] justify the PCA-Matryoshka rotation. Hubness correction in high-dimensional kNN motivates §7’s quotient, with the dissociation supplying the missing applicability condition. The companion theory paper [1] provides the transfer theorems, the rank proposition, the (A2) condition, and the heat-filter prediction tested here; the KV write-up [2] details §4–5.

11 Limitations and open problems

KV results cover four models $\leq 14\text{B}$ on LongBench-English/WikiText-2; the bias zero-point is Qwen-family by construction; the fused per-channel kernel awaits its cache-dispatch integration and the nuq grid stays decompress-then-attend. The rank certificate is worst-case and typically far below measured recall (a loose floor, honestly priced). The delta-centering retrieval option and the effective-rank taper diagnostic are validated on one public corpus each. The GQA-tolerance threshold rests on four architecture points. Deployment-grade format freezing (conformance tests, fuzzing) is roadmapped but not complete.

References

- [1] A. H. Bond. Keep the angle: a geometry-preserving basis in spectral embeddings. Working draft v0.8, <https://github.com/ahb-sjsu/the-angular-observer>, 2026.
- [2] A. H. Bond. One codebook for every architecture: asymmetric NormalFloat for calibration-free KV-cache quantization. Draft, [paper/kv_tmlr/](#), 2026.
- [3] A. Zandieh et al. TurboQuant: online vector quantization with near-optimal distortion. ICLR, 2026.
- [4] T. Dettmers, A. Pagnoni, A. Holtzman, and L. Zettlemoyer. QLoRA: efficient finetuning of quantized LLMs. NeurIPS, 2023.
- [5] C. Hooper et al. KVQuant: towards 10 million context length LLM inference with KV cache quantization. NeurIPS, 2024.
- [6] B. Varici et al. On the theory of Matryoshka representations and PCA recovery of ordered eigenfunctions. 2025.
- [7] H. E. Daniels. Rank correlation and population models. *J. Royal Stat. Soc. B*, 12(2):171–181, 1950.