

# 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); and a decision-level equivalence metric with a noise floor answering the *illusion of equivalency*, which de-confounded at matched bits reverses a published projection-sensitivity claim — value/output weights are  $2.3-6\times$  more behaviorally sensitive than query/key, the keys finding’s mirror image under weight rather than activation quantization; a compiler pass that *infers* the observer from the model graph (structural rules plus `torch.fx`) rather than requiring it declared; and the (A2) sensitivity boundaries for the two operators beyond attention — MoE routing, carried by the selection margin not the logit magnitude, and state-space decay, whose slow channels quantized in a log-time-constant basis cut real-Mamba perplexity from  $10^{10}$  to near-baseline. 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). The consumer-relative synthesis [2] sharpens this into a coding theorem and, crucially, *demonstrates* the mechanism: a compressor that *omits* a direction the consumer reads incurs an irreducible downstream *distortion floor* that no bit budget removes — now shown on a trained Llama attention layer, where the omitted read direction costs a softmax divergence  $800\text{--}5700\times$  (median 0.10 nats) that of a rate-matched consumer-aware coder, sealed and confirmed out-of-sample. The keys catastrophe is that floor. 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. **Behavior is the gate** (§6): a decision-level equivalence metric with a noise floor (answering the “illusion of equivalency”), and a de-confounded experiment that *reverses* a published projection-sensitivity ranking — at matched bits value/output weights are functionally  $2.3\text{--}6\times$  more sensitive than query/key across three models, the keys finding’s mirror image under the opposite operator.

6. **Inferring the observer** (§7): a compiler pass (structural rules + `torch.fx`) that maps each tensor to its operator regime and (A2) discipline from the model alone — the declared consumer becomes an inferred one, so bit allocation on a novel architecture needs no human in the loop.
7. **Operator sensitivity beyond attention** (§8): the measured (A2) boundary for MoE routing (selection is carried by the margin, not the magnitude) and state-space decay (slow channels fragile; a log-time-constant basis cuts real-Mamba perplexity from  $10^{10}$  to near-baseline).
8. **Certificates instead of promises** (§9): 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.
9. **Two empirical laws with their boundaries** (§10): 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).
10. **Compute-on-codes decode** (§11): 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.
11. **A methodology** (§12): 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 [7]) makes non-Matryoshka models truncatable without retraining; a random-rotation scalar quantizer (TurboQuant [4]) 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 (§11).

## 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 [3].

## 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 → codebook waste → 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 Behavior is the gate: the illusion of equivalency and the flipped projection

One rung above the reconstruction-metric fallacy sits a task-metric fallacy: a quantized model can match the full-precision model’s *accuracy* while getting a different *set* of items right — aggregate scores preserved, individual decisions churned (the “illusion of equivalency” [9]). The remedy is

the same one level up: gate on the consumer’s decisions, with a control. We ship a decision-level instrument, **behavioral\_agreement**: a symmetric flip rate (McNemar regressions *and* recoveries, whose sum is the churn that accuracy nets to zero), a prediction-level agreement (do the two models return the same answer, right or wrong), and a *noise floor* — the churn between two near-lossless requantizations — so drift is reported as excess over floor with a  $z$ -score. This repairs two defects in the Correctness Agreement of [9]: it is bounded above by  $\min(\text{acc})$  and blind to base-wrong→quant-right recoveries, and it has no control for the free flip rate of near-threshold items. On Qwen2.5-1.5B (next-token on WikiText-2, all weights per-output-channel fake-quantized) a 4-bit run whose accuracy falls  $\sim 10$  points hides that **41% of token predictions changed** (agreement 0.591; churn 16.6%  $\gg$  the 9.5% accuracy delta),  $\sim 28\sigma$  over a 0.014 floor. On the source paper’s own Mistral-7B a 4-bit run with a 4.3-point accuracy drop hides 23% churn ( $33\times$  floor), while 8-bit Correctness Agreement equals base accuracy — so the 14-point “lossless-Q8” gap that paper reports (Llama-3.2-3B) is near-threshold item brittleness, which the floor isolates rather than attributes to quantization.

The instrument then settles a mechanism claim. [9] reports the query/key projections as *more* quantization-sensitive than value/output and advises compressing V/O harder — but measures sensitivity as weight-distribution drift under `llama.cpp` K-quant, whose schemes give V and O *more* bits than Q/K (Q2\_K: Q,K at 2-bit, V at 4, O at 3): “Q/K drift more” is entangled with “Q/K got fewer bits.” De-confounded — the *same* per-output-channel  $b$ -bit quantizer on all four projections, drift read both in weight space and functionally (the movement of the output distribution,  $\text{KL}(p_{\text{base}}\|p_{\text{quant}})$ , from perturbing one projection in every layer) — the ranking inverts. Weight-space drift is *equal* across Q/K/V/O (relative-Frobenius ratio  $\approx 1$  at every bit-width), while functionally **V/O move the output 2.3–6 $\times$  more than Q/K** across three architectures, including the source paper’s own Mistral-7B ( $3.4\times$  at 4-bit; per-projection top-1 flip places V and O above Q and K). The advice is backwards for weight PTQ: at matched bits V/O are the projections whose error the model tolerates *least*; Q/K only draw level at destructive 2-bit, where the model is already broken.

The payoff is the paper’s principle sharpened. This finding and the keys finding (§3) look contradictory — there keys are the fragile side, here keys are the robust side — and they reconcile through one architectural fact read under two operators.  $\text{softmax}(QK^\top)$  mediates Q and K, so a *shared* weight perturbation to  $W^K$  is partly absorbed by attention’s relative comparison, while  $W^V, W^O$  write the residual stream almost linearly and pass their error on; but an *activation*/KV perturbation on the score path corrupts each cached key’s per-channel scale directly, which the softmax cannot absorb (§3). Same geometry, opposite fragile coordinate, selected by the operator: keys under activation quantization, V/O under weight quantization. Neither ordering is visible in weight-distribution or reconstruction statistics — skewness, kurtosis, cosine, the source paper’s own tools — only in a consumer-metric readout. That is the (A2) discipline of §9 applied to weight PTQ, and one more incident-to-instrument: the metric ships (**behavioral\_agreement**, scipy-free, unit-tested; L5), the finding is a reproducible experiment (**experiments/matched\_bit\_projection\_sensitivity.py**; L4), and the scope is part of the claim — fake-quantization, a clean per-output-channel absmax control (not any vendor kernel), three models on a WikiText-2 next-token sample, relative and directional. Llama-3.2-3B (gated), the downstream multiple-choice tasks, and the authors’ own `llama.cpp` pipeline forced to equal bit-width are the open confirmations.

## 7 Inferring the observer

Every result so far names its observer by hand: a human knew keys feed  $\text{softmax}(QK^\top)$  and V/O write the residual, and passed that to the (A2) probe (§9) as a *declared* consumer. That is the last manual step, and `operator_trace` removes it. Given only a model, it maps every parameter tensor to the operator its output flows into — one of {score path, linear residual, gated selection, state decay, norm} — and from the regime to the (A2) discipline, with no declaration. Two front-ends combine. A *structural* pass reads module type and name (a `LayerNorm` is a norm regardless of name; a router `gate` is selection, but SwiGLU’s `gate_proj` is not); it is always available but blind to renamed operators. A best-effort `torch.fx` pass symbolically traces the graph, finds the sink operators — softmax, top- $k$ , cumulative scan — and backtraces to the `Linear` layers feeding them, tagging the score/gate/state tensors *even when the names are obfuscated*: on a module whose query/key projections are named `left/right`, the graph pass alone recovers them as the score path. Where `fx` has positive evidence it overrides the structural guess; otherwise the structural baseline stands, and an untraceable model degrades gracefully rather than failing.

The discipline is a function of the regime *and the quantization target*, because the sensitive coordinate flips between the two — the reconciliation of §3 and §6 made mechanical. For the score path, the projection’s *weights* are the robust, symmetric side (softmax absorbs a shared perturbation) while its *cached keys* are the fragile per-channel-plus-zero-point side; for the residual path, the weights are the sensitive side (V/O, §6) and the cached values are cheap polar. An unclassified tensor defaults to the conservative per-channel discipline — never discard scale on a coordinate whose observer is unknown. This is the point at which “keep the geometry the observer sees” stops needing a human to name the observer.

## 8 Beyond attention: gates and state decay

Hybrid and state-space models add two operators whose sensitive coordinate is neither per-channel DC (keys) nor symmetric residual (V/O). `operator_sensitivity` supplies each one’s (A2) analysis, and each boundary was measured before it was believed.

**Gates: selection is carried by the margin, not the magnitude.** A top- $k$  router reads only the *order* of its logits  $\ell = W_g x$ , so it is invariant to a common-mode shift — a common-mode quantization error is free, and only the *differential* component (the part orthogonal to the all-ones direction) can flip a route. That is the tangential/(A2) split restated for selection, and `differential_fraction` is its statistic, the routing analog of the tangential fraction. A token survives iff the perturbation stays below its *margin* (the gap between the  $k$ -th and  $(k+1)$ -th logit), so the fragile tokens are the low-margin ones and the sensitive quantity is the margin distribution, not the logit magnitude. On a synthetic 16-expert router, top-1 flips concentrate on low-margin tokens by  $1.5\times$ ,  $6.3\times$ , and  $88\times$  at 2-, 3-, and 4-bit gate quantization respectively — almost every surviving flip at usable precision is a low-margin token. **On real Mixtral-8x7B-Instruct-v0.1** (top-2 routing, WikiText-2, 131,072 routing decisions), the top-2 boundary margins are substantial (median logit gap 0.485, tenth percentile 0.076)—unlike a high- $k$  router such as OLMoE (top-8 of 64), whose near-zero boundary margins (median 0.0015) saturate under any coarse quantization. Per-tensor gate-weight quantization flips the top-2 expert set for 16.6% of tokens at 4-bit and 34.0% at 3-bit, and the damage is margin-concentrated exactly as predicted: below-median-margin tokens flip  $10.7\times$  more often than above-median at 4-bit, and  $2.7\times$  at 3-bit. A controlled differential-logit perturbation isolates the mechanism further—at a quarter of the median margin the low-margin flip rate exceeds the high-margin rate by over  $2000\times$ , collapsing toward  $1\times$  only as the perturbation

grows past the margins. The absolute flip rate tracks the harshness of the gate quantizer, but the *margin-concentration* is the invariant, and it is far stronger on the substantial-margin top-2 router than the saturated top-8 one.<sup>1</sup> The discipline: size the gate budget against a low percentile of the margin, and protect the differential component; unlike keys, no DC term is needed — relative order is the geometry.

**Recurrences: the slow channels are fragile, and the error compounds.** For a per-channel linear recurrence  $h_t = a h_{t-1} + b_t$  the steady-state gain is  $1/(1-a)$  and the memory length is  $\sim 1/(1-a)$ , so a fixed error in the decay  $a$  is amplified far more in slow (long-memory) channels ( $a \rightarrow 1$ ) and accumulates over the sequence — the recurrent analog of the RoPE-slow-channel key finding, where the slow *rotary* channels carried the fragile DC offset. `decay_sensitivity` returns the per-channel coefficient  $\sum_t t a^{t-1}$  (steady state  $1/(1-a)^2$ ), which grows toward  $a \rightarrow 1$  and with sequence length; on a synthetic recurrence a fixed decay error is amplified 28–52 $\times$  more in slow than fast channels. The fix follows the geometry: quantize the decay in a *log-time-constant* basis ( $\tau = -\log a$ , then a uniform grid on  $\log \tau$ ), concentrating levels where  $a \rightarrow 1$  — the SSM analog of NF4’s non-uniform key levels. **On real Mamba-790m**, quantizing the continuous decay  $A = -e^{A_{\log}}$  on a linear grid raises WikiText-2 perplexity from a baseline 11.65 to  $1.0 \times 10^{10}$  at 3-bit — the slow channels, whose  $A$  ranges over many orders of magnitude, are annihilated — while quantizing in Mamba’s native  $A_{\log}$  (log) basis keeps it at 14.44: a gap of nearly nine orders of magnitude ( $\sim 7 \times 10^8$ ) from choosing the basis the geometry dictates.<sup>2</sup> The synthetic 5–6 $\times$  becomes catastrophic-versus-fine on a real model precisely because the real decay range is enormous.

Both boundaries are measured on synthetic operators, and state decay now end-to-end on a real SSM; the formal sensitivity theorems for gated selection and selective recurrences are the theory paper’s work, not claimed here.

## 9 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  $\kappa$  (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, 8]. 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).

<sup>1</sup>Measured by `benchmarks/validate_mixtral_routing.py` and `validate_olmoe_routing.py`; raw data in the corresponding `results_*.routing.json`.

<sup>2</sup>Measured by `benchmarks/validate_mamba_decay.py` (fp32 eval so the collapsed variant reports a finite perplexity rather than overflowing); raw data in `results_mamba_decay.json`.

## 10 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  $\tau$ -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.

## 11 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  $\sim 3$ -entry row, one shuffle reduction folds the correction in before the softmax update). Memory contiguity comes from a one-off build-time channel-major→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  $8 \times 10^{-8}$  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).

## 12 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.

## 13 Related work

TurboQuant (online vector quantization) [4] supplies the rotation-and-scalar-quantize primitive; QLoRA’s NF4 [5] the codebook asym-NF4 repairs; KVQuant [6] the calibrated per-channel NUQ that asym-NF4 matches calibration-free; RaBitQ/OPQ/PQ the retrieval baselines under one harness. Varici et al. [7] justify the PCA-Matryoshka rotation. Hubness correction in high-dimensional kNN motivates §10’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 [3] details §4–5. Rababah et al. [9] introduce the equivalency framing and Correctness Agreement that §6 sharpens with a noise floor and whose projection-sensitivity ranking it reverses at matched bits. Selective state-space models (Mamba [10]) and sparse mixture-of-experts routing (Mixtral [11]) motivate §8; their quantization has been studied empirically, but the operator-level sensitivity boundary — the selection margin for gates, the log-time-constant basis for decays — is the (A2) contribution here rather than a benchmarking sweep.

## 14 Limitations and open problems

KV results cover four models  $\leq 14$ 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. The projection-sensitivity reversal (§6) is fake-quantization on three models over a next-token sample, directional rather than absolute, and still owes the source paper’s multiple-choice tasks and the gated Llama-3.2-3B. The operator-regime tracer (§7) is heuristic — name/type rules plus a best-effort `torch.fx` pass, defeated by dynamic control flow and models `fx` cannot symbolically trace, where it falls back to the conservative structural default. The gate and state-decay boundaries (§8) are confirmed end-to-end on real models — routing on Mixtral-8x7B-Instruct (top-2) and OLMoE-1B-7B (top-8), state decay on Mamba-790m — but on WikiText-2 only, and the routing evidence is margin- concentration (whose absolute flip rate is gate-quantizer-dependent) rather than a single calibrated deployment recipe; the broader multi-model sweep and the formal sensitivity theorems for gated selection and selective recurrences are still open. 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. Observation Theory: geometry, distortion, and reliability as properties of observation. <https://doi.org/10.5281/zenodo.21423315>, 2026.

- [3] A. H. Bond. One codebook for every architecture: asymmetric NormalFloat for calibration-free KV-cache quantization. Draft, [paper/kv\\_tmlr/](#), 2026.
- [4] A. Zandieh et al. TurboQuant: online vector quantization with near-optimal distortion. ICLR, 2026.
- [5] T. Dettmers, A. Pagnoni, A. Holtzman, and L. Zettlemoyer. QLoRA: efficient finetuning of quantized LLMs. NeurIPS, 2023.
- [6] C. Hooper et al. KVQuant: towards 10 million context length LLM inference with KV cache quantization. NeurIPS, 2024.
- [7] B. Varici et al. On the theory of Matryoshka representations and PCA recovery of ordered eigenfunctions. 2025.
- [8] H. E. Daniels. Rank correlation and population models. *J. Royal Stat. Soc. B*, 12(2):171–181, 1950.
- [9] B. Rababah, C. G. Akcora, and C. K. Leung. The illusion of equivalency: statistical characterization of quantization effects in LLMs. arXiv:2607.08734, 2026.
- [10] A. Gu and T. Dao. Mamba: linear-time sequence modeling with selective state spaces. COLM, 2024.
- [11] A. Q. Jiang et al. Mixtral of experts. arXiv:2401.04088, 2024.