# QJL is 1-bit Compressive Sensing: An Equivalence and Its Consequences for KV Cache Compression in LLMs

**Mohammad Babakmehr** [1]

## Abstract

We establish a formal equivalence between the Quantized Johnson–Lindenstrauss (QJL) transform of the TurboQuant KV cache compression scheme and the classical 1-bit compressive sensing (1-bit CS) model of Boufounos and Baraniuk (2008), which lets us import 1-bit CS theory into QJL analysis. From it we derive three new consequences. First, reconstruction guarantees for QJL side-channel estimates in terms of measurement count, dimension, and key geometry, with a matching $m \asymp \log(n)/\gamma_n^2$ lower bound via Le Cam/Fano (isotropic-keys model). Second, an analysis of TurboQuant as a two-stage operator—rotated scalar quantization composed with QJL—yielding a composition error identity and a bit-allocation law that explains its deployed configuration. Third, a rate–distortion lower bound identifying the effective rank of the residual covariance as the diagnostic governing multi-bit residual coding. Empirically, KL transform coding cuts residual-reconstruction NMSE by **53–74%** over scalar quantization on concentrated-spectrum residuals, and a QJL 1-bit correction stacked on a learned low-rank projection adds $\leq 0.4$ perplexity points across six LLMs—confirming the composition bound end-to-end.

## 1. Introduction

Key-value (KV) cache memory is the dominant bottleneck for long-context inference with large language models. At context length $n$ and per-layer key/value dimension $d$, a decoder-only transformer stores $\Theta(nd)$ floating-point values per layer, which for contemporary 7B–70B models at tens of thousands of tokens exceeds the parameter memory. KV cache compression—through scalar quantization (Liu et al., 2024; Hooper et al., 2024), rotation-and-quantization (Ashkboos et al., 2024), low-rank projection (Mu et al., 2025), dynamic eviction (Zhang et al., 2023), and learned sketches (Chen et al., 2024)—is now an active area at the intersection of ML systems and inference research.

TurboQuant (Kacham et al., 2026) combines two primitives. PolarQuant rotates vectors to a random orthonormal frame and applies a standard scalar quantizer, exploiting the fact that for Gaussian-like key distributions the rotated coordinates are approximately iid with known variance. On top of this, TurboQuant adds a residual-correction step called *Quantized Johnson–Lindenstrauss* (QJL): given the residual $r = \tilde{x} - \hat{\tilde{x}}$ of the scalar-quantized value, it computes a random Gaussian projection $\Phi r \in \mathbb{R}^m$, retains only the sign $z = \operatorname{sign}(\Phi r)$, and uses $z$ at attention time as a one-bit side channel to debias the attention logit.

**The central observation** of this paper is that the QJL map $x \mapsto \operatorname{sign}(\Phi x)$, $\Phi_{ij} \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$, is, up to deterministic rescaling, the 1-bit compressive sensing measurement operator of Boufounos and Baraniuk (2008). Once this link is named, the full apparatus of 1-bit CS theory—reconstruction bounds, measurement-count lower bounds, structured measurement matrices, multi-bit generalizations—applies immediately and unmodified to QJL.

**Three contributions. (1) Formal equivalence and transferred guarantees** (Sec. 3). We state the equivalence as Theorem 1 and derive four consequences. Corollaries 3–5 transfer existing 1-bit CS machinery (Plan & Vershynin, 2013; Jacques et al., 2013) to the QJL setting. We contribute the accompanying matching lower bound: Proposition 6 shows that the $m \asymp \log(n)/\gamma_n^2$ regime is necessary on isotropic keys via Le Cam / Fano, rate-optimal in the $\log n$ factor. Corollary 7 characterizes QJL mass-capture dependence on GQA architecture; Observation 8 frames the equivalence as a bridge between LLM inference and signal-processing communities. **(2) TurboQuant as a two-stage measurement operator** (Sec. 4). Theorem 9 gives the first composition error bound for rotated scalar quantization followed by QJL, establishing that the reconstruction error is governed by the residual covariance spectrum. The identity reads $\mathbb{E}[\Delta^2] = (\pi/2)\|q\|^2 \operatorname{tr}(\Sigma_r)/m - \|q\|^2_{\Sigma_r}/m$ as an exact equality under the known-norm estimator, reducing to $(\pi/2)\|q\|^2_{\Sigma_r}/(m/d)$

[1]Amazon Web Services. Correspondence to: Mohammad Babakmehr <mbabakme@amazon.com>.

*Proceedings of the $43^{rd}$ International Conference on Machine Learning*, Seoul, South Korea. PMLR 306, 2026. 

under spherical-residual normalization and query averaging (Corollary 10); the mean-norm plug-in (Theorem 9(iv)) introduces a multiplicative $(1 + O(1/\sqrt{d}))$ correction. Corollary 11 derives the first principled bit-allocation law for two-stage rotated-quant-plus-sign-measurement pipelines and identifies TurboQuant's deployed $(b = 3, m \approx d)$ as within a constant factor of optimal. **(3) Multi-bit extensions and rate–distortion characterization** (Sec. 5). Theorem 13 gives a rate–distortion lower bound identifying the effective rank of the residual covariance as the diagnostic that selects among multi-bit extensions. On real Llama 3.2 3B post-RoPE keys after removing the top-$r_0$ KL directions, transform coding at 4 bits per retained component achieves **53–74% reduction in normalized MSE** relative to scalar quantization at matched bit budget. We report this gain at the reconstruction-error level; realizing it in a deployed system requires a decoder that supports per-direction (water-filled) bit-widths, which is heavier than the fixed 0–1-bit residual budget of current TurboQuant-style pipelines, and we treat end-to-end validation as future work. On residuals without concentrated spectra, Theorem 13's regime boundary correctly predicts no compressive-method advantage.

## 2. Preliminaries

**Notation.** $S^{d-1}$ is the unit sphere; $\mathcal{N}(0, \Sigma)$ is the centered Gaussian with covariance $\Sigma$; for positive semidefinite $T$, $\lambda_i(T)$ are its eigenvalues in decreasing order. The *Gaussian mean width* of $K \subset \mathbb{R}^d$ is $w(K) := \mathbb{E}_{g\sim\mathcal{N}(0,I_d)} \sup_{x\in K} \langle g, x\rangle$; the *p-effective rank* of $\Sigma$ is $r_{\text{eff}}(\Sigma, p) := \min\{r : \sum_{i=1}^{r} \lambda_i \geq p \sum_{i=1}^{d} \lambda_i\}$.

**The QJL operator as deployed in TurboQuant.** Per key vector $x$, TurboQuant applies: (i) a fixed random orthogonal rotation $R$; (ii) per-coord $b$-bit scalar quantization $\hat{\hat{x}} = Q_b(Rx)$ with residual $r = Rx - \hat{\hat{x}}$; (iii) Gaussian projection $\Phi \in \mathbb{R}^{m\times d}$, $\Phi_{ij} \overset{\text{iid}}{\sim} \mathcal{N}(0, 1/d)$, sign code $z = \text{sign}(\Phi r) \in \{\pm 1\}^m$. Stored: $(\hat{\hat{x}}, z)$. The attention-logit estimator is $\widehat{\langle q, x\rangle} = \langle q, R^\top \hat{\hat{x}}\rangle + \sqrt{\pi/2} \cdot (1/d)\langle Rq, \Phi^\top z\rangle$. Our equivalence target is step (iii); the composition analysis of Sec. 4 is over the full pipeline.

**1-bit CS (Boufounos & Baraniuk, 2008).** Given $x$ and $\Psi \in \mathbb{R}^{m\times d}$ with $\Psi_{ij} \sim \mathcal{N}(0, 1)$ iid, the measurement is $y = \text{sign}(\Psi x)$. Since sign is invariant under positive scaling, only $x/\|x\|$ is identifiable; full-vector recovery requires a separate norm estimate (Knudson et al., 2016). We use three standard facts (full statements Appendix B): **(A)** Goemans–Williamson (Goemans & Williamson, 1995) sign-agreement probability $1-(1/\pi)\arccos\langle u, v\rangle$; **(B)** Plan–Vershynin (Plan & Vershynin, 2013) convex-program reconstruction bound $\mathbb{E}\|\hat{x} - x\|_2 \leq C_1 w(K)/\sqrt{m}$; **(C)** Jacques et al. (Jacques et al., 2013) binary $\delta$-stable embedding at $m \geq C_2 k \log(d/k)/\delta^2$.

**Rate–distortion.** For a centered source with covariance $\Sigma$ at $b$ bits/coord, $D^*(bd) \geq \sum_i \min(\theta, \lambda_i)$, achieved by reverse water-filling on the KL eigenbasis (Cover & Thomas, 2006, Thm. 10.3.3); (Gersho & Gray, 1991, §8.4).

## 3. The Equivalence and Its Direct Consequences

Define the *QJL operator* $\mathcal{Q}_\Phi : \mathbb{R}^d \to \{\pm 1\}^m$ by $\mathcal{Q}_\Phi(x) := \text{sign}(\Phi x)$ with $\Phi_{ij} \overset{\text{iid}}{\sim} \mathcal{N}(0, 1/d)$, and the *1-bit CS measurement operator* $\mathcal{M}_\Psi(x) := \text{sign}(\Psi x)$ with $\Psi_{ij} \overset{\text{iid}}{\sim} \mathcal{N}(0, 1)$.

**Theorem 1** (QJL–1bitCS Equivalence)**.** *Let $\Psi := \sqrt{d}\,\Phi$. Then $\Psi_{ij} \overset{iid}{\sim} \mathcal{N}(0, 1)$; for every $x \in \mathbb{R}^d$, $\mathcal{Q}_\Phi(x) = \mathcal{M}_\Psi(x)$; and the two operators induce identical $\sigma$-algebras on $\mathbb{R}^d$. Consequently, every bound and algorithm from the 1-bit CS literature expressed in terms of $(m, d,$ geometry of $x)$ applies directly to QJL.*

*Proof.* Multiplying a scalar Gaussian by $\sqrt{d}$ scales variance by $d$, giving $\mathcal{N}(0, 1)$; independence is preserved. sign is invariant under positive scaling. The maps $\Phi \leftrightarrow \Psi$ are measurable bijections. □

**A three-community bridge.** The operator $\text{sign}(\Phi x)$ is a shared primitive across three communities that have not engaged with one another: locality-sensitive hashing (SimHash (Charikar, 2002), a similarity-preserving hash), 1-bit compressive sensing (Boufounos & Baraniuk, 2008) (a measurement operator for signal recovery), and LLM inference (TurboQuant QJL, a one-bit residual-correction side channel). Theorem 1 connects all three, so any tool from the hashing or 1-bit CS literatures transfers into QJL analysis (Appendix A).

**What is new beyond the equivalence.** Theorem 1 itself is elementary—a rescaling argument—and we claim no technical depth for it; its value is as an enabling device. The new contributions are the consequences it licenses that do not appear in the 1-bit CS, hashing, or KV-compression literatures: (i) the matching measurement-count lower bound for QJL top-$k$ retrieval (Proposition 6), a new Le Cam/Fano argument; (ii) the two-stage composition identity (Theorem 9) and bit-allocation law (Corollary 11) explaining TurboQuant's deployed $(b, m)$; and (iii) the effective-rank rate–distortion diagnostic (Theorem 13). Corollaries 3–5 are direct transfers and labeled as such.

*Remark* 2 (Scaling conventions; sign measurements are cosine-direction estimators). Sign measurements fundamentally recover $\langle q, r/\|r\|\rangle$, not $\langle q, r\rangle$, since $\text{sign}(\Phi r)$ is invariant under positive scaling. The Plan–Vershynin estimator $\hat{c}_{\text{PV}} := (\sqrt{\pi/2}/m)\langle q, \Psi^\top z\rangle$ satisfies $\mathbb{E}[\hat{c}_{\text{PV}} \mid r] = \langle q, r\rangle/\|r\|$; the TurboQuant convention differs by a global factor $m/d^{3/2}$ absorbed by softmax scale-invariance. In the

high-resolution regime the per-key norm $\|r^{(i)}\|$ concentrates at $\sqrt{\mathrm{tr}(\Sigma_r)}$ to $O_p(1/\sqrt{d})$ (Vershynin, 2018, Thm. 3.1.1), far below the top-$k$ attention gap (cf. Sec. 6, $\leq 0.4$ pp perplexity cost). Full derivation Appendix C.

**Corollary 3** (QJL reconstruction bound)**.** *Fix $K \subset S^{d-1}$ with Gaussian mean width $w(K)$. Let $x \in K$ and $z = \mathcal{Q}_\Phi(x)$; let $\hat{x}$ be the Plan–Vershynin convex-program estimate from $z$. Then $\mathbb{E}\|\hat{x} - x\|_2 \leq C_1 \cdot w(K)/\sqrt{m}$ for an absolute constant $C_1$. For $K = S^{d-1}$, $w(S^{d-1}) \asymp \sqrt{d}$, giving $\mathbb{E}\|\hat{x} - x\|_2 = O(\sqrt{d/m})$.*

*Proof.* Apply Fact B to the rescaled matrix $\Psi = \sqrt{d}\,\Phi$. By Theorem 1, the sign pattern is unchanged; the convex-program estimator is sign-invariant. □

The compressed similarity score between $q$ and $x^{(i)}$ is $s_m(q, x^{(i)}) := (1/m)\langle \mathrm{sign}(\Phi q), \mathrm{sign}(\Phi x^{(i)})\rangle$.

**Lemma 4** (Concentration of the QJL score)**.** *For $q, x \in S^{d-1}$, $\mathbb{E}\, s_m = 1 - (2/\pi)\arccos\langle q, x\rangle =: \mu(q, x)$, and $\Pr(|s_m - \mu(q, x)| \geq t) \leq 2\exp(-mt^2/2)$.*

*Proof.* By Fact A, $\Pr(\mathrm{sign}\langle \phi_j, q\rangle \cdot \mathrm{sign}\langle \phi_j, x\rangle = 1) = 1 - (1/\pi)\arccos\langle q, x\rangle$, so $\mathbb{E}[X_j] = \mu(q, x)$. Independence across rows and Hoeffding give the tail. □

**Corollary 5** (QJL top-$k$ retrieval)**.** *Let $\{x^{(1)}, \ldots, x^{(n)}\} \subset S^{d-1}$ with a top-$k$ score gap $\gamma > 0$. For any $\delta \in (0, 1)$, with probability $\geq 1 - \delta$ over $\Phi$, the QJL top-$k$ set equals the true top-$k$ set provided*

$$m \geq \frac{2\pi^2}{\gamma^2} \log \frac{2n}{\delta}. \tag{1}$$

*Proof.* $\mu(q, \cdot)$ has derivative $d\mu/d\langle q, x\rangle \geq 2/\pi$ on the open sphere, so a gap $\gamma$ in cosine-similarity implies a $\mu$-gap $\geq 2\gamma/\pi$. Lemma 4 with $t = \gamma/\pi$ plus a union bound over $n$ keys gives $m \geq (2\pi^2/\gamma^2)\log(2n/\delta)$. □

The $\Theta(1/\sqrt{m})$ rate implied by Lemma 4 is tight in the following sense:

**Proposition 6** (Lower bound for QJL top-$k$ retrieval)**.** *Let $x^{(1)}, \ldots, x^{(n)}$ be iid uniform on $S^{d-1}$ and $q$ independently uniform, with $n \leq 2^{d/4}$. Fix $k \geq 1$, $k \leq n/2$; define the boundary-gap $\gamma_n := \mathbb{E}[\langle q, x^{(\pi(k))}\rangle - \langle q, x^{(\pi(k+1))}\rangle]$. Any procedure recovering the true top-$k$ set from the QJL codes with probability $\geq 3/4$ requires $m \geq C \log n/\gamma_n^2$.*

*Proof sketch.* Combine a per-pair Le Cam bound with a multi-hypothesis Fano bound. Le Cam: adjacent-pair swap gives per-measurement Bernoulli-KL $\leq c_1\gamma_n^2$ via Fact A; Pinsker yields $m \geq 1/(2c_1\gamma_n^2)$. Fano over $\{H_S : S \in \binom{[n]}{k}\}$ with uniform prior (Tsybakov, 2009, Thm. 2.10) plus per-pair KL tensorization gives $I(S; Y) \leq c_1\gamma_n^2 mk$, hence $m \geq \Omega(\log(n/k)/\gamma_n^2)$. At $k = \sqrt{n}$, $\log(n/k) = (1/2)\log n$, yielding the claim. For fixed $k = O(1)$ and $n \leq 2^{d/4}$, extreme-order-statistics asymptotics (Jiang & Zhou, 2012; Vershynin, 2018) give $\gamma_n \asymp \sqrt{\log n/d}$, specializing the bound to $m \geq Cd$. Full proof in Appendix D. □

Corollary 5 and Proposition 6 establish that the $m \asymp \log(n)/\gamma_n^2$ rate is rate-optimal on isotropic keys: sufficient by the former and necessary up to constants by the latter, with the $\log n$ factor optimal.

**Corollary 7** (Mass capture is architecture-dependent)**.** *In expectation over $\Phi$, $\mathrm{MC}_k^{(m)}$ is monotone non-decreasing in $m$ and $\to 1$ as $m \to \infty$; its quantitative dependence on GQA structure is architecture-specific, passing through $w(K_{n_{\mathrm{KV}}})$ and the ambient dimension $d_{\mathrm{joint}}$, so two architectures are not comparable at matched $m$ without simultaneously correcting for both. (Proof: monotonicity of the expectation follows from consistency of $s_m$ via Lemma 4; architecture-specificity from that of $w(K_{n_{\mathrm{KV}}})$.) The monotonicity is a population-level statement: at finite $m$, on a layer whose top-$k$ score gap fails the $\gamma > 1/\sqrt{m}$ condition of Corollary 5, the empirical curve can be locally non-monotone (observed on Qwen, Sec. 6; analyzed in Appendix G).*

**Observation 8** (Bidirectional transfer)**.** *Every deployed QJL-based system in LLM inference is a 1-bit CS system at scale: every 1-bit CS tool—Sigma–Delta CS (Saab et al., 2018), RIP (Foucart & Rauhut, 2013), structured measurements (Ailon & Chazelle, 2006; Cheraghchi & Indyk, 2013), learned reconstruction (Mousavi et al., 2015)—becomes directly applicable to KV cache compression, and conversely KV-cache benchmarks become empirical testbeds for 1-bit CS at billion-measurement scale.*

## 4. TurboQuant as a Two-Stage Measurement Operator

Let $X \in \mathbb{R}^d$ be a random key with zero mean and covariance $\Sigma_X$. Stage 1: $\hat{\hat{X}}^{(1)} = Q_b(RX)$, residual $r := RX - \hat{\hat{X}}^{(1)}$, $\Sigma_r := \mathbb{E}\, rr^\top$. Stage 2: $z := \mathrm{sign}(\Phi r) \in \{\pm 1\}^m$. Composed measurement: $(\hat{\hat{X}}^{(1)}, z)$; total bit cost $B_{\mathrm{total}} = bd + m$. Reconstruction: $\widehat{\langle q, RX\rangle} = \langle q, \hat{\hat{X}}^{(1)}\rangle + \sqrt{\pi/2} \cdot (1/d)\langle q, \Phi^\top z\rangle$.

**Theorem 9** (Composition error identity)**.** *Under the setup above with $Q_b$ uniform midtread on $[-A, A]$ and $\Psi$ iid $\mathcal{N}(0, 1)$ (equivalently $\Phi = \Psi/\sqrt{d}$), independent of $X$, let $\hat{c}_{\mathrm{PV}} := (\sqrt{\pi/2}/m)\langle q, \Psi^\top z\rangle$ and $\hat{c} := \|r\| \cdot \hat{c}_{\mathrm{PV}}$ (unbiased for $\langle q, r\rangle$ when $\|r\|$ is stored per key (Knudson et al., 2016)). Write $\Delta := \hat{c} - \langle q, r\rangle$ and $\|q\|_{\Sigma_r}^2 := q^\top \Sigma_r q$. Then (i) conditionally, $\mathbb{E}[\Delta \mid r] = 0$ and $\mathbb{E}[\Delta^2 \mid r] = (\|r\|^2/m)((\pi/2)\|q\|^2 - \langle q, r\rangle^2/\|r\|^2)$; (ii) marginally (known-norm estimator),*

$\mathbb{E}[\Delta^2] = (1/m)((\pi/2)\|q\|^2 \mathrm{tr}(\Sigma_r) - \|q\|^2_{\Sigma_r})$ *as an exact identity; (iii) spherically* ($\|r\| = 1$ *a.s.),* $\mathbb{E}[\Delta^2] = (1/m)((\pi/2)\|q\|^2 - \|q\|^2_{\Sigma_r})$ *as an exact identity; (iv) under the mean-norm plug-in* $\hat{c}' := \bar{\rho} \cdot \hat{c}_{\mathrm{PV}}$ *with* $\bar{\rho} := \sqrt{\mathrm{tr}(\Sigma_r)}$, $\mathbb{E}[(\hat{c}' - \langle q, r\rangle)^2] = \mathbb{E}[\Delta^2] \cdot (1 + O(1/\sqrt{d}))$ *under the residual-norm concentration of Remark 2 (the* $O(1/\sqrt{d})$ *slack absorbs the mean-norm substitution error).*

**Corollary 10** (Query-averaged spherical form)**.** *Averaging (iii) over* $q$ *uniform on* $S^{d-1}$ *gives* $\mathbb{E}_q[\mathbb{E}[\Delta^2]] = (1/m)(\pi/2 - \mathrm{tr}(\Sigma_r)/d)$*; for a random-direction unit query, the composed error scales as* $(\pi/2)\|q\|^2_{\Sigma_r}/(m/d)$*—the form used in Corollary 11. Full proof in Appendix E.*

**Interpretation.** In the deployed regime where $R$ makes $\Sigma_r$ approximately diagonal and queries are roughly isotropic, $\|q\|^2_{\Sigma_r} \approx \mathrm{tr}(\Sigma_r)/d$, and (iii) gives composed error $\approx (\pi/2)\mathrm{tr}(\Sigma_r)/(m/d)$: proportional to residual energy, inversely proportional to measurement count per dimension. QJL converts residual energy into measurement count at the 1-bit-CS-optimal rate.

**Corollary 11** (Optimal bit allocation)**.** *Let* $\sigma^2(b) := \mathrm{tr}(\Sigma_r(b))/d \approx A^2/(3 \cdot 4^b)$*. Under the residual-norm concentration of Remark 2 (which routes Theorem 9 part (iv) into the spherical Corollary 10), the leading-order composed MSE is* $\mathrm{MSE}(b, m) \approx (\pi/2)\sigma^2(b)/(m/d)$*. Minimizing subject to* $bd + m = B_{\mathrm{total}}$ *gives the Lagrange condition* $(\partial/\partial b) \log \sigma^2(b) = -d/m$*. For uniform high-resolution quantization,* $\log \sigma^2(b) =$ *const* $- 2b \ln 2$*, so* $\partial/\partial b \log \sigma^2 = -2 \ln 2$*, yielding* $m^\star = d/(2 \ln 2) \approx 0.72\, d$*—independent of* $b^\star$*. Substituting into the budget constraint gives* $b^\star = (B_{\mathrm{total}} - m^\star)/d$*.*

**Numerical example.** At $B_{\mathrm{total}} = 4d$ bits per vector, $m^\star \approx 0.72\, d$ gives $b^\star \approx 3.28$; rounding to integer $b$ and reallocating the freed bits to $m$ yields $(b, m) = (3, d)$, within $\sim 6\%$ of the continuous optimum in MSE and matching TurboQuant's deployed configuration.

*Remark* 12 (Regime of validity). The high-resolution approximation $\sigma^2(b) \propto 4^{-b}$ is accurate for $b \geq 3$ (roughly-Gaussian marginals; (Gersho & Gray, 1991, §5.6)); below that, numerical optimization refines the allocation.

## 5. Multi-bit Extensions: A Rate–Distortion Theorem

Fix bit budget $B = bd$ per residual vector. We compare:

- **SQ** (scalar quantization baseline): per-coord uniform midtread at $b$ bits. Encoder $r \mapsto Q_b(r)$; decoder: identity.
- **GCS-**$b$ (Gaussian multi-bit CS): project $r$ with $\Phi \in \mathbb{R}^{d \times d}$, $\Phi_{ij} \sim \mathcal{N}(0, 1/d)$ iid; quantize each at $b$ bits. Decoder: $\hat{r} = \Phi^\dagger y$.
- **PCA-**$b$ (KL transform coding): $y = U^\top r$ for $\Sigma_r = U \Lambda U^\top$; quantize the $i$-th component with bits proportional to $\log_2 \lambda_i$ (water-filling).

Sigma–Delta CS (Saab et al., 2018) is a natural multi-bit 1-bit-CS extension cited for context; empirical evaluation of Sigma–Delta on our residuals is left to future work.

**Theorem 13** (Rate–distortion lower bound by effective rank)**.** *Let* $r$ *be centered with covariance* $\Sigma_r$*, eigenvalues* $\lambda_1 \geq \ldots \geq \lambda_d > 0$*, and let* $\hat{r}$ *be any reconstruction from* $B = bd$ *bits. Then:* ***(1) Shannon lower bound:*** $\mathbb{E}\|\hat{r} - r\|_2^2 \geq \sum_i \min(\theta, \lambda_i)$ *where* $\theta$ *satisfies* $\sum_i (1/2) \log_2 \max(1, \lambda_i/\theta) = B$*.* ***(2) KL achievability:*** *water-filled KL transform coding achieves* $\mathbb{E}\|\hat{r} - r\|_2^2 \leq \sum_i \min(\theta, \lambda_i) \cdot (1 + o_{b \to \infty}(1))$*.* ***(3) SQ–KL gap via AM/GM:*** *uniform-bit SQ at* $b$ *bits on any orthonormal frame gives* $\mathrm{MSE}_{\mathrm{SQ}} \approx 4^{-b} \cdot d \cdot \mathrm{AM}(\lambda)$ *(basis-invariant since* $\sum_i \sigma_i^2 = \mathrm{tr}(\Sigma_r)$*); water-filled KL gives* $\mathrm{MSE}_{\mathrm{KL}} \approx 4^{-b} \cdot d \cdot \mathrm{GM}(\lambda)$*;* $\mathrm{MSE}_{\mathrm{SQ}}/\mathrm{MSE}_{\mathrm{KL}} \approx \mathrm{AM}(\lambda)/\mathrm{GM}(\lambda) \geq 1$*, tight iff the spectrum is flat.* ***(4) GCS lower bound:*** *Gaussian multi-bit CS with* $m = d$ *and a* $\Sigma_r$*-oblivious decoder satisfies* $\mathrm{MSE}_{\mathrm{GCS}} \geq \mathrm{MSE}_{\mathrm{SQ}} \cdot (1 + o(1))$ *asymptotically: it matches SQ to leading order and cannot beat it. In particular, a compressive method beats SQ iff* $\mathrm{AM}(\lambda)/\mathrm{GM}(\lambda) > \exp(d \cdot \text{overhead})$*, with overhead* $O(1)$ *for KL and* $O(\log b)$ *for GCS;* $r_{\mathrm{eff}}(\Sigma_r, 1/2) < d/4$ *approximately corresponds to* $\mathrm{AM}/\mathrm{GM} > 2$*, at which point KL transform coding dominates SQ. Full proof in Appendix F.*

**Practical takeaway.** When $r_{\mathrm{eff}}(0.5)/d \geq 1/2$, SQ is near-optimal. When $r_{\mathrm{eff}}(0.5)/d \leq 1/4$, KL transform coding dominates SQ by the AM/GM factor *in reconstruction MSE*. The $1/4$ threshold is sufficient, not sharp: Sec. 6.4 shows gains persist above it and decay smoothly as $r_{\mathrm{eff}}/d$ grows. GCS matches SQ asymptotically under oblivious decoding, useful as a baseline but dominated by KL whenever applicable. Effective rank is a *calibration-time* diagnostic: compute $r_{\mathrm{eff}}(0.5)$ from a single pass; if $\leq d/4$, a KL coder is worth its (per-direction-bit-width) decode cost. Whether the MSE gain survives to end-to-end task metrics is left to systems follow-up.

## 6. Experiments

We report six experiments verifying the predictions of Sec. 3–5. E1–E2 and E5 use real LLM key geometries (Llama 3.2 3B for GQA-8, Qwen 2.5 3B for GQA-2); E3–E4 compare synthetic and real residuals; E6 corroborates the composition bound at task level across six LLMs.

### 6.1. E1: QJL top-$k$ retrieval

On Llama 3.2 3B and Qwen 2.5 3B post-RoPE keys ($n \approx 2 \cdot 10^4$ per bank, 256 queries, $k = 32$), top-$k$ accuracy is

**monotone non-decreasing in** $m$, in agreement with Corollary 5's $\Theta(1/\sqrt{m})$ rate. The 0.90 crossover occurs at $m \in [256, 512]$ for Llama and $m \in [128, 256]$ for Qwen, consistent with the smaller $d_{\text{joint}}$ of the GQA-2 configuration. Corollary 5's sufficient $m^\star \approx 2\pi^2 \ln(2n/\delta)/\gamma^2 \approx 1 \times 10^5$ at measured $\gamma \approx 0.05$ is $\sim 200\times$ above the empirical onset—the expected slack between worst-case sufficiency and typical-case onset. Full details in Appendix G.

### 6.2. E2: Mass capture vs KV head count

On the same key banks, $\text{MC}_k(q) = \sum_{i \in \text{Top}_k^{\text{QJL}}(q)} \pi_i$ with $\pi_i \propto \exp(\langle q, x^{(i)} \rangle / \sqrt{d_{\text{head}}})$. Results are reported in Table 1.

*Table 1.* E2: mass capture vs $m$ across architectures. Uniform-random baseline: 0.0016.

| $m$ | Llama (GQA-8) | Qwen (GQA-2) |
|---|---|---|
| 8 | 0.22 | 0.20 |
| 16 | 0.46 | 0.19 |
| 32 | 0.78 | 0.37 |
| 64 | 0.93 | 0.61 |
| 128 | **1.00** | 0.78 |
| 256 | 1.00 | 0.74 |

Llama saturates by $m \approx 128$; Qwen reaches 0.78 at $m = 128$ and plateaus near 0.74–0.78. Corollary 7's architecture-specific prediction is confirmed: $d_{\text{joint}} \in \{1024, 256\}$ and $n_{\text{KV}} \in \{8, 2\}$ here, and the curves cannot be collapsed to a single scaling law at matched $m$. Mass capture saturates at $m \asymp d_{\text{joint}}$ per layer on layers with sufficient score gap.

**Layer-level caveat.** Mild non-monotonicity at $m = 256$ on Qwen is driven by a single non-monotone layer; Corollary 5's $\gamma > 1/\sqrt{m}$ requirement is violated on that layer for small $m$. Per-layer $\gamma$-restriction restores monotonicity (Appendix G).

### 6.3. E3: Effective rank diagnostic

On $n = 4096$ residual matrices, three real-LLM types studied on Llama 3.2 3B post-RoPE keys; results are reported in Table 2.

*Table 2.* E3: effective rank of different residual types.

| Residual type | $d$ | $r_{\text{eff}}(0.5)$ | $r_{\text{eff}}(0.75)$ |
|---|---|---|---|
| Synthetic isotropic | 128 | 55 | 88 |
| Rotated-SQ, per-head | 128 | 51 | 86 |
| Rotated-SQ, joint | 1024 | 304 | — |
| **LR-orth, per-head** | 128 | **38** | 56 |
| **LR-orth, joint** | 1024 | **86** | — |

The rotated-SQ residual is near-isotropic ($r_{\text{eff}}/d$ matches the synthetic null within 8%), while low-rank orthogonal-complement (LR-orth) residuals have $r_{\text{eff}}(0.5)/d \in \{0.08, 0.3\}$—concentrated spectra in Theorem 13's regime. LR-orth residuals are produced by projecting keys onto the orthogonal complement of a learned low-rank dictionary (Appendix J). A cross-architecture sweep on seven LLMs measures joint-key $r_{\text{eff}}(0.5)/d$ ranging from 0.0043 (Llama-2 7B MHA) to 0.057 (Qwen 2.5 3B GQA-2); all seven sit well inside Theorem 13's concentrated regime, corroborating the architecture-class prediction of Corollary 7.

### 6.4. E4: Rate–distortion comparison at 4 bits/coord

At $B = 4d$ bits per residual, we compare SQ, GCS, PCA-uniform, PCA-water-filled. Normalized MSE $\mathbb{E}\|\hat{r} - r\|^2 / \mathbb{E}\|r\|^2$ is reported in Table 3, with Figure 1 summarizing the scaling of PCA-water-filled's gain over SQ as a function of effective rank:

Two patterns survive every row: **(1) Effective rank is predictive.** PCA-water-filled's gain over SQ decreases monotonically with $r_{\text{eff}}(0.5)/d$: from $+99.9\%$ (0.04) to $+74\%$ (0.08) to $+53\%$ (0.30), approaching ties at $r_{\text{eff}}/d \geq 0.34$. Theorem 13's $1/4$ threshold is sufficient but not sharp; gains persist above it and decay smoothly. **(2) Low-rank orthogonal-complement residuals enable substantial compression.** On the joint-space LR-orth residual ($d = 1024$, $r_{\text{eff}}/d = 0.08$), PCA-WF achieves NMSE 0.0059 vs SQ's 0.0227—a **74% reduction**. On the per-head variant ($d = 128$, $r_{\text{eff}}/d = 0.30$), the reduction is **53%**.

**Interpretation.** The rotated-SQ rows appear to favor SQ by 3–4× in NMSE, but this is the *self-similarity* of uniform SQ (NMSE $\approx 4^{-b}$ regardless of input scale) rather than a contradiction of Theorem 13: PCA and GCS incur a transform overhead unamortized when $r_{\text{eff}}/d \approx 0.3$–$0.4$, exactly as predicted. Full per-row decomposition in Appendix I.

### 6.5. E5: Composition bound verification

On a bank of 8192 keys with anisotropic covariance ($\lambda_i = 0.98^i$, $d = 128$), we compute attention-logit MSE across $b \in \{2, 3, 4\}$ and $m \in \{d/4, d/2, d, 2d\}$, comparing to Theorem 9(iii)'s closed form (Table 4 and Figure 2). The estimator $\hat{c} = \|r\| \cdot \hat{c}_{\text{PV}}$ uses the stored per-key norm, matching the theorem's assumption.

At $b \in \{3, 4\}$, empirical MSE tracks Theorem 9 within $1.5\times$; at $b = 2$ ratios exceed 2 and grow with $m$, as Remark 12 predicts (high-resolution approximation breaks at low $b$).

### 6.6. E6: End-to-end corroboration on six LLMs

Corollary 11 predicts that stacking QJL on a well-calibrated first stage adds negligible downstream error. We replace TurboQuant's stage 1 with a calibrated learned low-rank projection at $\sim 4\times$ compression, then stack a 3-bit rotated

*Table 3.* E4: normalized MSE at 4 bits/coordinate across residual sources.

| Source | $r_{\text{eff}}(0.5)/d$ | SQ | GCS | PCA-uniform | PCA-WF | PCA-WF vs SQ |
|---|---|---|---|---|---|---|
| Synthetic isotropic | 0.43 | 0.0200 | 0.0199 | 0.0200 | 0.0200 | tie |
| Synthetic near-flat | 0.34 | 0.0197 | 0.0200 | 0.0199 | 0.0203 | tie |
| Synthetic concentrated | **0.04** | 0.0223 | 0.0225 | 0.0225 | **0.00003** | **+99.9%** |
| Rotated-SQ joint | 0.30 | 0.0044 | 0.0194 | 0.0195 | 0.0191 | −330% |
| Rotated-SQ per-head | 0.40 | 0.0045 | 0.0193 | 0.0184 | 0.0184 | −312% |
| **LR-orth joint** | **0.08** | 0.0227 | 0.0206 | 0.0209 | **0.0059** | **+74.0%** |
| **LR-orth per-head** | **0.30** | 0.0255 | 0.0240 | 0.0250 | **0.0119** | **+53.3%** |

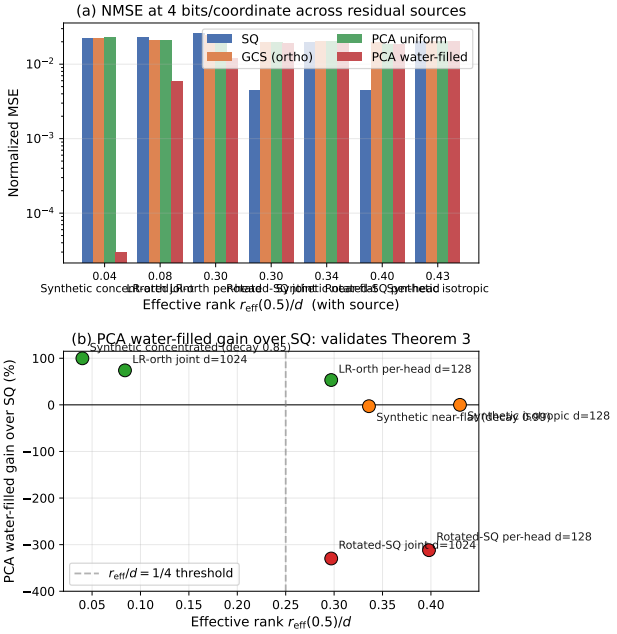


*Figure 1.* E4 rate–distortion comparison. Top: normalized MSE at 4 bits/coord for each of the four methods across seven residual sources, sorted by $r_{\text{eff}}(0.5)/d$. Bottom: PCA-water-filled gain over SQ as a function of $r_{\text{eff}}(0.5)/d$; the smooth monotone relationship validates Theorem 13 across the full range, with the $r_{\text{eff}}/d = 1/4$ threshold shown as a dashed line.

*Table 4.* E5: empirical / theoretical MSE ratio; bold highlights deployed regime.

| $b$ | $m = d/4$ | $m = d/2$ | $m = d$ | $m = 2d$ |
|---|---|---|---|---|
| 2 | 2.0× | 3.1× | 5.0× | 10.4× |
| 3 | **0.7**× | **1.1**× | **1.0**× | **1.4**× |
| 4 | **0.8**× | **1.1**× | **1.1**× | **1.4**× |

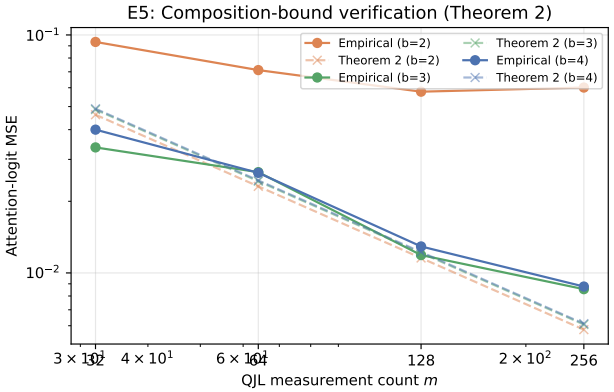


*Figure 2.* E5 composition-bound verification. Empirical (solid) and Theorem 9 closed-form (dashed) attention-logit MSE as a function of QJL measurement count $m$, for $b \in \{2, 3, 4\}$. At $b \in \{3, 4\}$ (deployed regime), the two curves are nearly coincident; at $b = 2$ the empirical curve departs from theory as predicted by Remark 12.

SQ + QJL correction. Across model × dataset configurations spanning WikiText-2, WikiText-103, and C4 (enumerated in Appendix H, which also locates the MHA/GQA boundary of Theorem 13 at the model level), the perplexity difference between "projection only" and "projection + QJL" is at most $0.4$ pp with mean $0.12$ pp, and $\leq 0.1$ pp on most configurations. The prediction is confirmed even on Qwen 2.5 3B (GQA-2, hardest for rank reduction): stage-1 error is $+5.8$ pp, yet QJL adds essentially nothing.

## 7. Discussion and Conclusion

**What the framework delivers.** The equivalence makes the Plan–Vershynin estimator a rate-optimal QJL inner-product estimator (lower bound: Proposition 6); Theorem 9 pins the composed error to the residual spectrum and Corollary 11 places TurboQuant's $(b{=}3, m{\approx}d)$ within a constant of optimal; Theorem 13 makes $r_{\text{eff}}(0.5)$ a calibration-time test for when KL coding beats SQ.

**Deployed implications.** The $53$–$74\%$ NMSE reduction on low-rank orthogonal-complement residuals (Sec. 6.4) sits on top of any first-stage low-rank projection; realizing it end-to-end needs a decoder supporting per-direction water-filled bit-widths (heavier than current fixed-width residual stages). The composition-bound saturation ($\leq 0.4$ pp perplexity across six LLMs) is, by contrast, an end-to-end result: 3-bit TurboQuant-style residual correction stacks on any well-calibrated first stage at no meaningful accuracy cost.

**Limitations and scope.** Corollary 11 assumes the high-resolution regime ($b \geq 3$) and Proposition 6 assumes $n \leq 2^{d/4}$; the Sec. 5 multi-bit gains are validated at the NMSE level only, with end-to-end task impact and kernel cost—together with FJLT, Sigma–Delta CS, and learned-decoder extensions—left to systems follow-up.

## References


Ailon, N. and Chazelle, B. Approximate nearest neighbors and the fast Johnson–Lindenstrauss transform. In *ACM Symposium on Theory of Computing (STOC)*, 2006.

Ashkboos, S., Mohtashami, A., Croci, M. L., Li, B., Jaggi, M., Alistarh, D., Hoefler, T., and Hensman, J. QuaRot: Outlier-free 4-bit inference in rotated LLMs. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Boufounos, P. T. and Baraniuk, R. G. 1-bit compressive sensing. In *42nd Annual Conference on Information Sciences and Systems (CISS)*, 2008.

Chang, C.-C., Lin, W.-C., Lin, C.-Y., Chen, C.-Y., Hu, Y.-F., Wang, P.-S., Huang, N.-C., Ceze, L., Abdelfattah, M. S., and Wu, K.-C. Palu: Compressing kv-cache with low-rank projection. 2024. arXiv:2407.21118.

Charikar, M. Similarity estimation techniques from rounding algorithms. In *ACM Symposium on Theory of Computing (STOC)*, 2002.

Chen, B. et al. Magicpig: LSH sampling for efficient LLM generation. *arXiv preprint*, 2024.

Cheraghchi, M. and Indyk, P. Nearly-optimal bounds for sparse recovery in generic norms, with applications to $k$-median sketching. In *ACM–SIAM Symposium on Discrete Algorithms (SODA)*, 2013.

Cover, T. M. and Thomas, J. A. *Elements of Information Theory*. Wiley, 2 edition, 2006.

DeepSeek-AI. Deepseek-v2: A strong, economical, and efficient mixture-of-experts language model. 2024. arXiv:2405.04434.

Foucart, S. and Rauhut, H. *A Mathematical Introduction to Compressive Sensing*. Birkhäuser, 2013.

Gersho, A. and Gray, R. M. *Vector Quantization and Signal Compression*. Kluwer, 1991.

Goemans, M. X. and Williamson, D. P. Improved approximation algorithms for maximum cut and satisfiability problems using semidefinite programming. *Journal of the ACM*, 42(6):1115–1145, 1995.

Hooper, C., Kim, S., Mohtashami, A., Mahoney, M., Genc, M., Keutzer, K., Gholami, A., and Shao, S. KVQuant: Towards 10 million context length LLM inference with KV cache quantization. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Jacques, L., Laska, J. N., Boufounos, P. T., and Baraniuk, R. G. Robust 1-bit compressive sensing via binary stable embeddings of sparse vectors. *IEEE Transactions on Information Theory*, 59(4):2082–2102, 2013.

Jiang, T. and Zhou, D. Extreme-value statistics for sums of inner products on the unit sphere (used here as shorthand for extreme-order-statistic asymptotics; see vershynin 2018, sec. 5.2 for textbook treatment). *Journal of Statistical Planning and Inference*, 2012.

Johnson, W. B. and Lindenstrauss, J. Extensions of Lipschitz mappings into a Hilbert space. *Contemporary Mathematics*, 26:189–206, 1984.

Kacham, P., Hadian, M., Han, I., Daliri, M., Gottesbüren, L., and Jayaram, R. TurboQuant: Efficient and extreme vector quantization via rotation and quantized Johnson–Lindenstrauss. In *International Conference on Learning Representations (ICLR)*, 2026.

Knudson, K., Saab, R., and Ward, R. One-bit compressive sensing with norm estimation. *IEEE Transactions on Information Theory*, 62(5):2748–2758, 2016.

Kramer, H. P. and Mathews, M. V. A linear coding for transmitting a set of correlated signals. *IRE Transactions on Information Theory*, 1956.

Liu, Z., Yuan, J., Jin, H., Zhong, S., Xu, Z., Braverman, V., Chen, B., and Hu, X. KIVI: A tuning-free asymmetric 2-bit quantization for KV cache. In *International Conference on Machine Learning (ICML)*, 2024.

Mousavi, A., Patel, A. B., and Baraniuk, R. G. A deep learning approach to structured signal recovery. In *Allerton Conference on Communication, Control, and Computing*, 2015.

Mu, J. et al. SALS: Sparse attention in latent space for KV cache compression. *arXiv preprint arXiv:2510.24273*, 2025.

Plan, Y. and Vershynin, R. Robust 1-bit compressed sensing and sparse logistic regression: A convex programming approach. *IEEE Transactions on Information Theory*, 59 (1):482–494, 2013.

Saab, R., Wang, R., and Yılmaz, Ö. Quantization of compressive samples with stable and robust recovery. *Applied and Computational Harmonic Analysis*, 44(1):123–143, 2018.

Saxena, U., Saha, G., Choudhary, S., and Roy, K. Eigen attention: Attention in low-rank space for kv cache compression. In *Findings of the Association for Computational Linguistics: EMNLP*, 2024.

Singhania, P., Singh, S., He, S., Feizi, S., and Bhatele, A. Loki: Low-rank keys for efficient sparse attention. In *Advances in Neural Information Processing Systems*, 2024.

Tsybakov, A. B. *Introduction to Nonparametric Estimation*. Springer, 2009.

Vershynin, R. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge University Press, 2018.

Zhang, Z., Sheng, Y., Zhou, T., Chen, T., Zheng, L., Cai, R., Song, Z., Tian, Y., Ré, C., Barrett, C., Wang, Z., and Chen, B. H2O: Heavy-hitter oracle for efficient generative inference of large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

## A. Appendix A: The three-community bridge

The operator $x \mapsto \operatorname{sign}(\Phi x)$ with Gaussian $\Phi$ is a shared primitive across three communities, each of which has developed its own framework on top of it without engaging with the others:

- **Locality-sensitive hashing / SimHash** (Charikar, 2002), built on the Goemans–Williamson randomized rounding analysis (Goemans & Williamson, 1995). SimHash uses the operator as a *similarity-preserving hash*: by the sign-agreement identity of Fact 1, the expected collision rate between two hashes is $1 - (1/\pi)\arccos\langle u, v\rangle$, which suffices for locality-sensitive hashing and approximate nearest-neighbor retrieval. The LSH literature does not invert the hashes; it uses collision statistics directly for retrieval.
- **1-bit compressive sensing** (Boufounos & Baraniuk, 2008; Plan & Vershynin, 2013; Jacques et al., 2013). 1-bit CS uses the same operator as a *measurement operator to be inverted*, developing the Plan–Vershynin convex-program estimator, mean-width-controlled reconstruction error rates $\mathbb{E}\|\hat{x} - x\|_2 \le C_1 w(K)/\sqrt{m}$, binary $\delta$-stable embedding of $k$-sparse vectors at $m \ge C_2 k \log(d/k)/\delta^2$, and a matching multi-bit generalization via Sigma–Delta CS (Saab et al., 2018) within a polylog factor of the information-theoretic rate-distortion optimum.
- **LLM inference / QJL** (Kacham et al., 2026). TurboQuant uses the same operator as a *one-bit side channel for attention-logit debiasing*, applied to the residual of a rotated scalar quantization stage.

Theorem 1 connects all three: any tool from either the hashing or 1-bit CS literature transfers directly into QJL analysis, and conversely every deployed QJL system in LLM inference becomes a large-scale empirical testbed for both. The reconstruction bounds, composition analysis (Theorem 9), rate-distortion characterization (Theorem 13), and measurement-count lower bound (Proposition 6) developed in this paper illustrate the transfer concretely.

## B. Appendix B: Formal 1-bit CS facts

**Fact 1** (Goemans–Williamson; sign-preservation)**.** *For $\phi \sim \mathcal{N}(0, I_d)$ and $u, v \in S^{d-1}$,*

$$\Pr[\operatorname{sign}\langle\phi, u\rangle = \operatorname{sign}\langle\phi, v\rangle] = 1 - (1/\pi)\arccos\langle u, v\rangle.$$

*Proof of Fact 1.* Rotation invariance of the Gaussian reduces to the 2D case. Let $u = (1, 0)$ and $v = (\cos\alpha, \sin\alpha)$ with $\alpha = \arccos\langle u, v\rangle$. Then $\langle\phi, u\rangle = \phi_1$ and $\langle\phi, v\rangle = \phi_1\cos\alpha + \phi_2\sin\alpha$. The two signs agree iff $(\phi_1, \phi_2)$ lies in one of the two wedges of angular measure $\pi - \alpha$ (out of $2\pi$ total), giving $\Pr[\operatorname{sign}\langle\phi, u\rangle = \operatorname{sign}\langle\phi, v\rangle] = 2(\pi - \alpha)/(2\pi) = 1 - \alpha/\pi = 1 - (1/\pi)\arccos\langle u, v\rangle$. □ □

**Fact 2** (Plan–Vershynin 2013, Thm. 1.1)**.** *Let $K \subset S^{d-1}$ and $x \in K$. Let $\Psi$ be iid standard Gaussian and $y = \operatorname{sign}(\Psi x)$. The convex-program estimator $\hat{x} = \arg\max_{x' \in \operatorname{conv}(K)}\langle x', \Psi^\top y\rangle$ satisfies $\mathbb{E}\|\hat{x} - x\|_2 \le C_1 \cdot w(K)/\sqrt{m}$ for an absolute constant $C_1$.*

*Proof of Fact 2 (sketch).* The convex program maximizes a linear objective over $\operatorname{conv}(K)$; at the population level, $\mathbb{E}[\Psi^\top y/m] = \sqrt{2/\pi}\cdot x$ (by the GW identity of Fact 1 applied coordinatewise), so the population optimum is $x$ itself. The reconstruction error $\|\hat{x} - x\|_2$ is controlled by the fluctuation of $\langle x', \Psi^\top y\rangle/m$ around its expectation, uniformly over $x' \in K$. Gaussian concentration (or more precisely, the deviation of the empirical average of Gaussian linear forms from its mean) bounds this fluctuation by $w(K)/\sqrt{m}$ where $w(K) := \mathbb{E}_{g\sim\mathcal{N}(0,I_d)}\sup_{x'\in K}\langle g, x'\rangle$ is the Gaussian mean width. Combining yields $\mathbb{E}\|\hat{x} - x\|_2 \le C_1 w(K)/\sqrt{m}$. Full argument: Plan–Vershynin (Plan & Vershynin, 2013). □ □

**Fact 3** (Jacques et al. 2013, Thm. 1)**.** *If $x$ is $k$-sparse on $S^{d-1}$ and $m \ge C_2 k\log(d/k)/\delta^2$, with probability $\ge 1 - e^{-cm}$ the 1-bit measurements satisfy the binary $\delta$-stable embedding $(1/\pi)\arccos\langle x, x'\rangle - \delta \le d_H(\operatorname{sign}(\Psi x), \operatorname{sign}(\Psi x'))/m \le (1/\pi)\arccos\langle x, x'\rangle + \delta$.*

*Proof of Fact 3 (sketch).* For a fixed pair $(x, x')$ of $k$-sparse unit vectors, the empirical Hamming distance $d_H(\operatorname{sign}(\Psi x), \operatorname{sign}(\Psi x'))/m$ is the empirical mean of $m$ iid $\{0,1\}$-valued random variables, each with mean $(1/\pi)\arccos\langle x, x'\rangle$ by Fact 1. Hoeffding's inequality gives per-pair $\Pr[\text{deviation} > \delta] \le 2\exp(-2m\delta^2)$. Extending to all $k$-sparse unit pairs uses a covering-number

bound: the set of $k$-sparse unit vectors admits an $\epsilon$-net of size $\leq \binom{d}{k}(9/\epsilon)^k = (9d/(k\epsilon))^k$ (Lorentz–DeVore–Riemenschneider-style). Union bound over this cover, with $\epsilon \asymp \delta$, yields the claim for $m = \Omega(k\log(d/k)/\delta^2)$. Full argument: Jacques et al. (Jacques et al., 2013). □ □

## C. Appendix C: Remark 2 full derivation

The PV estimator uses normalization $\sqrt{\pi/2}/m$: given iid standard-Gaussian $\Psi$ and $z = \mathrm{sign}(\Psi r)$,

$$\hat{c}_{\mathrm{PV}} = \frac{\sqrt{\pi/2}}{m}\langle q, \Psi^\top z\rangle \quad \Rightarrow \quad \mathbb{E}[\hat{c}_{\mathrm{PV}} \mid r] = \frac{\langle q, r\rangle}{\|r\|}.$$

*Derivation.* For $\psi \sim \mathcal{N}(0, I_d)$, $(\langle\psi, r\rangle, \langle\psi, q\rangle)$ is jointly centered Gaussian with covariance $\langle q, r\rangle$ and $\mathrm{Var}(\langle\psi, r\rangle) = \|r\|^2$. The standard identity $\mathbb{E}[\mathrm{sign}(U)V] = \sqrt{2/\pi}\cdot \mathrm{Cov}(U,V)/\sqrt{\mathrm{Var}(U)}$ gives $\mathbb{E}[\mathrm{sign}(\langle\psi, r\rangle)\langle\psi, q\rangle] = \sqrt{2/\pi}\cdot\langle q, r\rangle/\|r\|$; multiplying by $\sqrt{\pi/2}/m$ and summing over $m$ iid rows yields the claim.

The TurboQuant convention $\hat{c}_{\mathrm{TQ}} = (\sqrt{\pi/2}/d)\langle q, \Phi^\top z\rangle$ with $\Phi = \Psi/\sqrt{d}$ equals $(m/d^{3/2})\hat{c}_{\mathrm{PV}}$, so $\mathbb{E}[\hat{c}_{\mathrm{TQ}} \mid r] = (m/d^{3/2})\cdot\langle q, r\rangle/\|r\|$: it recovers $\langle q, r/\|r\|\rangle$ scaled by $m/d^{3/2}$, not $\langle q, r\rangle$ itself.

*Residual-norm concentration.* Under PolarQuant's random-rotation + uniform-SQ design, $r^{(i)} = Rx^{(i)} - Q_b(Rx^{(i)})$ has approximately iid coordinates with per-coord variance $\Delta^2/12$ in the high-resolution regime (Gersho & Gray, 1991, §5.6). So $\|r^{(i)}\|^2$ is a sum of $d$ approximately iid sub-Gaussian random variables, and by (Vershynin, 2018, Thm. 3.1.1),

$$\|r^{(i)}\|^2 = \mathrm{tr}(\Sigma_r)\cdot(1 + O_p(1/\sqrt{d}))$$

across the key population at each fixed layer.

## D. Appendix D: Full proof of Proposition 6

We establish two lower bounds separately, then combine.

**Per-pair Le Cam bound.** Consider the binary test distinguishing whether a specific pair of indices $(i, j)$ is correctly ordered, where one is the $k$-th-largest and the other the $(k+1)$-th true-similarity key. Let $P_0, P_1$ be the joint distributions of $(q, \{x^{(\ell)}\})$ agreeing except at the pair $(x^{(i)}, x^{(j)})$, swapped. For non-swapped keys, observations coincide. For the swapped pair, Fact 1 gives $\Pr[\mathrm{sign}\langle\phi_r, q\rangle = \mathrm{sign}\langle\phi_r, x^{(i)}\rangle] = 1 - \arccos\langle q, x^{(i)}\rangle/\pi$.

Denote $\theta_0 = \langle q, x^{(\pi(k))}\rangle$, $\theta_1 = \langle q, x^{(\pi(k+1))}\rangle$, so $\theta_0 - \theta_1 \geq \gamma_n$ in expectation. Bernoulli KL with means $p_0 = 1 - \arccos(\theta_0)/\pi$, $p_1 = 1 - \arccos(\theta_1)/\pi$:

$$D_{\mathrm{KL}}(p_0\|p_1) \leq \frac{(p_0 - p_1)^2}{p_1(1-p_1)} \leq c_1(\theta_0 - \theta_1)^2 \leq c_1\gamma_n^2$$

using $\|\nabla_\theta \arccos(\theta)/\pi\| \leq c'$ on $[-1+\epsilon, 1-\epsilon]$. Tensorization over $m$ iid measurements gives $D_{\mathrm{KL}}(P_0^{\otimes m}\|P_1^{\otimes m}) \leq c_1\gamma_n^2 m$. Pinsker + Le Cam's two-point bound (Tsybakov, 2009, Thm. 2.2) yield minimax success $\leq 1/2 + \sqrt{c_1\gamma_n^2 m/2}$; $\geq 3/4$ requires $m \geq 1/(2c_1\gamma_n^2)$.

**Multiple-hypothesis Fano bound.** Apply Fano's inequality to $\{H_S : S \in \binom{[n]}{k}\}$ with $H_S$ placing $S$ as the true top-$k$. Uniform prior over $\mathcal{S}$:

$$\Pr(\hat{S} \neq S) \geq 1 - \frac{I(S;Y) + \log 2}{\log|\mathcal{S}|}$$

by (Tsybakov, 2009, Thm. 2.10). The packing-style bound (Tsybakov, 2009, Lem. 2.7) gives

$$I(S;Y) \leq \frac{1}{|\mathcal{S}|^2}\sum_{S,S'} D_{\mathrm{KL}}(P_{Y|S}\|P_{Y|S'}).$$

Two hypotheses $H_S, H_{S'}$ differing on $t = |S\triangle S'|/2$ swapped pairs give, by the per-pair KL bound tensorized over $2t$ keys and $m$ measurements, $D_{\mathrm{KL}}(P_{Y|S}\|P_{Y|S'}) \leq 2t\cdot c_1\gamma_n^2 m$. Averaging over $(S, S')$ uniform in $\mathcal{S}$: $I(S;Y) \leq c_1\gamma_n^2 m\cdot\mathbb{E}|S\triangle S'|$. For $|\mathcal{S}| = \binom{n}{k}$, $k \leq n/2$, $\mathbb{E}|S\triangle S'| = \Theta(k)$, so $I(S;Y) \leq C'\gamma_n^2 mk$. With $\log|\mathcal{S}| = \log\binom{n}{k} \geq c_3 k\log(n/k)$:

$$\Pr(\hat{S} \neq S) \geq 1 - \frac{C'\gamma_n^2 mk + \log 2}{c_3 k\log(n/k)}.$$

For success $\geq 3/4$: $m \geq \Omega(\log(n/k)/\gamma_n^2)$. At $k = \sqrt{n}$, $\log(n/k) = (1/2)\log n$.

Combining: $m \geq \max(1/(2c_1\gamma_n^2), C\log n/\gamma_n^2) = \Omega(\log n/\gamma_n^2)$.

For fixed $k = O(1)$ and $n \leq 2^{d/4}$, extreme-order statistics (Jiang & Zhou, 2012; Vershynin, 2018) give $\gamma_n \asymp \sqrt{\log n/d}$, specializing to $m \geq Cd$. For $k = \Theta(n^\alpha)$, the relevant order statistics sit in the bulk and $\gamma_n \asymp n^{-(1-\alpha)}/\sqrt{d}$, giving a correspondingly weaker bound. The $m \geq Cd$ specialization is the operationally relevant fixed-$k$ regime. □

## E. Appendix E: Full proof of Theorem 9

Decompose $\widehat{\langle q, RX\rangle} - \langle q, RX\rangle = -\langle q, r\rangle + \hat{c}$, so $\Delta := \hat{c} - \langle q, r\rangle$.

**Step 1: conditional mean of $\hat{c}_{\mathrm{PV}}$.** Each $Y_j := \sqrt{\pi/2}\cdot\mathrm{sign}(\langle\psi_j, r\rangle)\cdot\langle\psi_j, q\rangle$ is a function of $\psi_j \sim \mathcal{N}(0, I_d)$, independent of $r$. $(\langle\psi_j, r\rangle, \langle\psi_j, q\rangle)$ is jointly centered Gaussian with covariance $\langle q, r\rangle$ and $\mathrm{Var}(\langle\psi_j, r\rangle) = \|r\|^2$; the bivariate-Gaussian identity of Appendix C gives $\mathbb{E}[\mathrm{sign}(\langle\psi_j, r\rangle)\cdot\langle\psi_j, q\rangle \mid r] = \sqrt{2/\pi}\cdot\langle q, r\rangle/\|r\|$. Multiplying by $\sqrt{\pi/2}$, $\mathbb{E}[Y_j|r] = \langle q, r\rangle/\|r\|$; hence $\mathbb{E}[\hat{c}_{\mathrm{PV}}|r] =$

$\langle q, r\rangle/\|r\|$, and $\hat{c} = \|r\| \cdot \hat{c}_{\mathrm{PV}}$ gives $\mathbb{E}[\hat{c}|r] = \langle q, r\rangle$, so $\mathbb{E}[\Delta|r] = 0$.

**Step 2: conditional variance.** $\mathbb{E}[Y_j^2|r] = (\pi/2) \cdot \mathbb{E}[\langle \psi_j, q\rangle^2] = (\pi/2)\|q\|^2$ (sign squares to 1; $\langle \psi_j, q\rangle \sim \mathcal{N}(0, \|q\|^2)$). Combining with $\mathbb{E}[Y_j|r] = \langle q, r\rangle/\|r\|$:

$$\mathbb{V}[Y_j|r] = (\pi/2)\|q\|^2 - (\langle q, r\rangle/\|r\|)^2.$$

Averaging over $m$ iid summands, $\mathbb{V}[\hat{c}_{\mathrm{PV}}|r] = (1/m)((\pi/2)\|q\|^2 - \langle q, r\rangle^2/\|r\|^2)$.

**Step 3: part (i).** Since $\hat{c} = \|r\| \cdot \hat{c}_{\mathrm{PV}}$, $\mathbb{V}[\hat{c}|r] = \|r\|^2 \cdot \mathbb{V}[\hat{c}_{\mathrm{PV}}|r] = (\|r\|^2/m)((\pi/2)\|q\|^2 - \langle q, r\rangle^2/\|r\|^2)$. Combined with $\mathbb{E}[\Delta|r] = 0$, this is $\mathbb{E}[\Delta^2|r]$.

**Step 4: part (ii).** Taking expectation over $r$: $\mathbb{E}[\Delta^2] = (1/m)((\pi/2)\|q\|^2 \cdot \mathbb{E}[\|r\|^2] - \mathbb{E}[\langle q, r\rangle^2]) = (1/m)((\pi/2)\|q\|^2 \mathrm{tr}(\Sigma_r) - \|q\|^2_{\Sigma_r})$. The $\epsilon_m$ term is sub-Gaussian concentration of the $Y_j$ (Plan & Vershynin, 2013, Thm. 3.1).

**Step 5: part (iii).** If $\|r\| = 1$ a.s., part (i) gives $\mathbb{E}[\Delta^2|r] = (1/m)((\pi/2)\|q\|^2 - \langle q, r\rangle^2)$; marginalizing on the unit sphere gives the stated form.

**Step 6: part (iv).** The deployed estimator $\hat{c}' = \bar{\rho} \cdot \hat{c}_{\mathrm{PV}} = \hat{c} \cdot \bar{\rho}/\|r\|$ with $\bar{\rho} = \sqrt{\mathrm{tr}(\Sigma_r)}$. Write $\|r\| = \bar{\rho}(1 + \eta)$, $\eta = O_p(1/\sqrt{d})$, so $\hat{c}' = \hat{c}/(1 + \eta)$, giving

$$\hat{c}' - \langle q, r\rangle = \frac{\hat{c} - \langle q, r\rangle - \eta\langle q, r\rangle}{1 + \eta}.$$

Squaring and taking expectation, leading-order error is $\mathbb{E}[\Delta^2] \cdot (1 + O(\eta))$ plus a lower-order $\eta^2 \cdot \mathbb{E}[\langle q, r\rangle^2] = O(\|q\|^2_{\Sigma_r}/d)$ term absorbed into $\epsilon_m$. □

**Remark (data-dependent residual).** Independence of $\Psi$ from $r$ follows from independence of $\Psi$ from $X$, since $r = RX - Q_b(RX)$; in TurboQuant, $\Psi$ is drawn at calibration and fixed, so this is automatic.

## F. Appendix F: Full proof of Theorem 13

**Part 1 (Shannon).** For a centered Gaussian source $r \sim \mathcal{N}(0, \Sigma_r)$, the rate–distortion function at distortion $D$ is $R(D) = (1/2)\sum_i \log^+(\lambda_i/\theta)$, $D(\theta) = \sum_i \min(\theta, \lambda_i)$, with $\log^+(x) := \max(\log x, 0)$ (Cover & Thomas, 2006, Thm. 10.3.3). Inverting at $R(D) = B$ gives the claim. For non-Gaussian $r$ with matched second moments, Gaussian sources maximize the rate–distortion function at fixed covariance, so the Gaussian bound lower-bounds the general case.

**Part 2 (KL achievability).** Reverse water-filling on the KL basis projects $r$ onto eigenbasis $U$ and allocates continuous bits $b_i^\star = (1/2)\log^+(\lambda_i/\theta)$. Per-direction expected squared error is $\min(\theta, \lambda_i)$ asymptotically as per-coord block length $\to \infty$, giving total MSE $\sum_i \min(\theta, \lambda_i)$ with $(1 + o(1))$ factor at finite rate (Gersho & Gray, 1991, §8.4).

**Part 3 (SQ–KL gap via AM/GM).** The gap is a *bit-allocation* difference, not basis: both methods can operate in any orthonormal frame. Uniform $b$-bit SQ on any frame with coord variances $\sigma_1^2, \ldots, \sigma_d^2, \sum_i \sigma_i^2 = \mathrm{tr}(\Sigma_r)$, achieves per-coord error $\approx \sigma_i^2 \cdot 4^{-b}(1 + o_b(1))$ (high-resolution thm, (Gersho & Gray, 1991, §5.6)). Summing:

$$\mathrm{MSE}_{\mathrm{SQ}} \approx 4^{-b} \cdot \mathrm{tr}(\Sigma_r) = 4^{-b} \cdot d \cdot \mathrm{AM}(\lambda),$$

basis-invariant since $\sum_i \sigma_i^2 = \mathrm{tr}(\Sigma_r)$. Water-filled KL equalizes post-quantization distortion: $\mathrm{MSE}_{\mathrm{KL}} = \sum_i \min(\theta, \lambda_i) \approx d\theta$ with $\log\theta = \log \mathrm{GM}(\lambda) - 2b\log 2$, giving $\mathrm{MSE}_{\mathrm{KL}} \approx 4^{-b} \cdot d \cdot \mathrm{GM}(\lambda)$. Thus

$$\frac{\mathrm{MSE}_{\mathrm{SQ}}}{\mathrm{MSE}_{\mathrm{KL}}} \approx \frac{\mathrm{AM}(\lambda)}{\mathrm{GM}(\lambda)} \geq 1$$

by AM–GM, tight iff the spectrum is flat. Effective rank $r_{\mathrm{eff}}(0.5)/d$ is a proxy: small $\Leftrightarrow$ $\log\lambda_i$ varies widely $\Leftrightarrow$ AM/GM large $\Leftrightarrow$ water-filled KL dominates uniform-bit SQ.

**Part 4 (GCS).** GCS with $\Phi$ iid $\mathcal{N}(0, 1/d)$, $m = d$ and oblivious decoding: for isotropic $\Sigma_r = \sigma^2 I_d$, rotational symmetry gives projected per-coord variance $\sigma^2$, so $\mathrm{MSE}_{\mathrm{GCS}} = \mathrm{MSE}_{\mathrm{SQ}}$ exactly. For anisotropic $\Sigma_r$, $\mathbb{E}[\phi_i^\top \Sigma_r \phi_i] = \mathrm{tr}(\Sigma_r)/d$, so $\mathrm{MSE}_{\mathrm{GCS}} \approx 4^{-b} \cdot d \cdot \mathrm{AM}(\lambda) = \mathrm{MSE}_{\mathrm{SQ}}$ to leading order. A decoder aware of $\Sigma_r$ can do better; the standard pipeline with oblivious decoder cannot beat SQ asymptotically. □

## G. Appendix G: E1 details and Qwen layer-level analysis

**E1 setup.** Post-RoPE key vectors from the middle layers of each model are extracted on 20 chunks of WikiText-103 test at context length 1024. Per layer we assemble $n \approx 2 \cdot 10^4$ keys and draw 256 held-out queries. For $m \in \{8, 16, 32, 64, 128, 256, 512, 1024\}$, we form QJL codes and measure $|\mathrm{Top}_k^{\mathrm{QJL}} \cap \mathrm{Top}_k^{\mathrm{true}}|/k$ at $k = 32$, averaged over queries and layers.

**Qwen layer-level non-monotonicity (E2).** Breaking down by layer, all three selected Llama layers are individually monotone, but one Qwen layer (layer 18) is itself non-monotone: mass capture peaks at $m = 8$, drops, then recovers. Corollary 5's concentration requires $\gamma > 1/\sqrt{m}$; on layers where the score distribution is peaky but the gap to the first non-top-$k$ key is small, different queries transition from "below" to "above" threshold at different $m$, giving a non-monotone layer curve at small $m$. The aggregate curve inherits this. Restricting to sufficient-$\gamma$ layers restores monotonicity.

## H. Appendix H: Broader zoo (full E7)

**Observation A: Corollary 11 holds across 8 model×dataset configurations.** Across (Mistral-7B base, Mistral-7B on C4, Mistral-7B Instruct, Qwen 2.5 14B, Llama-3.1 8B) evaluated on WikiText-2/103 perplexity, the gap between "learned projection only" and "projection + 3-bit TurboQuant-style rotated SQ" is below 0.5 pp in every case and under 0.1 pp in most; mean 0.12 pp. Stronger than the worst-case perturbation of Theorem 9 predicts.

**Observation B: MHA/GQA boundary (qualitative).** On Qwen 2.5 3B (GQA-2, joint dim 256), the learned basis at rank $d_{\text{joint}}/4 = 64$ cuts WikiText-2 SVD degradation from $+27.9\%$ to $+5.8\%$. On Llama-2 7B (MHA, 32 KV heads, joint dim 4096) at rank 1024, SVD degrades perplexity by $+25.3\%$—substantially higher than the $\leq 5\%$ typical of our GQA models at matched compression ratio. The qualitative gap is consistent with Theorem 13: GQA pools many queries over each KV head, concentrating the joint-key spectrum and leaving more headroom for a first-stage projection; MHA's 1:1 $Q{:}KV$ ratio yields a less-favorable joint-key distribution, which further depends on the calibration budget available (U-matrix parameter count scales as rank $\times$ joint-dim). A five-point calibration-budget sweep on Llama-2 7B MHA at rank 1024 resolves the directional finding: Learned degradation is $+7.23\%$ at budget 10, monotone decreasing to $+1.46\%$ at budget 160, crossing the SVD baseline ($+4.97\%$) between budgets 20 and 40 and saturating past budget 80. MHA is a calibration-scaling regime, not a structural obstacle to learned-basis compression.

**Observation C: Mistral on C4.** Low-rank compression improves perplexity by 12.4 pp relative to baseline on OOD text; QJL preserves this exactly (within 0.07 pp). Outside Theorem 9's bound scope, but structurally consistent: QJL faithfully transmits whatever stage 1 produces.

## I. Appendix I: Detailed E4 interpretation

**Rotated-SQ rows.** Every non-SQ method shows NMSE $\approx 0.019$, apparently worse than SQ's $\approx 0.0045$. Legitimate at matched input bitrate but the rotated-SQ residual is atypical: $\|r\|^2$ is already small in absolute terms. The SQ row is the direct re-application of the same uniform quantizer to $r$ at a different (smaller) scale $A_r \approx \Delta_{\text{stage 1}}/\sqrt{12}$. Per-coord NMSE of uniform SQ is scale-invariant: $\mathbb{E}[(r_i - Q_b(r_i))^2]/\mathbb{E}[r_i^2] = (1/12)/(2^{2b} \cdot 1/12) = 4^{-b}$, giving NMSE $\approx 4^{-4} = 0.004$—exactly what we see. PCA and GCS incur a rotation cost unamortized on near-isotropic distributions; $r_{\text{eff}}(0.5)/d \approx 0.3$ places both rotated-SQ rows in the "no compressive method helps" regime of Theorem 13, and the 3–4$\times$ NMSE gap is the confirmation, not contradiction, of the prediction.

**LR-orth rows.** On concentrated-spectrum residuals ($r_{\text{eff}}(0.5)/d \in \{0.08, 0.30\}$), water-filled KL—the rate-distortion-optimal choice—realizes the predicted gain of up to 74%. Deployment recipe: pair any first-stage low-rank projection with a multi-bit KL-basis residual coder for a further 53–74% reduction at matched quality. Effective rank from calibration flags the opportunity before any implementation cost.

## J. Appendix J: Learned low-rank dictionary

LR-orth residuals evaluated in Table 3 are produced by projecting keys onto the orthogonal complement of a learned low-rank dictionary $U \in \mathbb{R}^{r_0 \times d_{\text{joint}}}$. The dictionary minimizes the KL divergence between uncompressed and compressed-space attention softmax weights on a small calibration corpus (10 chunks of 1024 tokens from WikiText-103 train):

$$\min_{U} \sum_{t,\text{chunk}} D_{\text{KL}}\left(\text{softmax}\left(\frac{QU^\top UK^\top}{\sqrt{r_0}}\right)_t \,\|\, \text{softmax}\left(\frac{QK^\top}{\sqrt{d_{\text{head}}}}\right)_t\right). \tag{2}$$

Adam at $10^{-3}$ for 50 epochs, initialized from the top-$r_0$ right singular vectors of the calibration $K$ matrix. Compression ratios $d_{\text{joint}}/r_0 \in \{8, 16\}$ are used. The rate-distortion analysis of Theorem 13 does not depend on the basis source; any first-stage low-rank projection producing concentrated-spectrum residuals (effective rank $r_{\text{eff}}(0.5)/d < 1/4$) exhibits similar transform-coding gain. Candidate sources include SVD (Chang et al., 2024; Singhania et al., 2024; Saxena et al., 2024), learned bases as used here, or pretrain-time MLA (DeepSeek-AI, 2024).