

CASK-Q: Centered Causal Attention Sketches for Joint Query–Key Quantization

The Decentralized Compute Company

October 12, 2026

Abstract

Low-bit quantization perturbs the interaction between query and key projections, not just each projection separately. We introduce CASK-Q, a calibration objective for jointly optimizing scalar-quantized query/key weights using randomized sketches of row-centered causal attention-logit error. An exact prefix-sum identity evaluates the sketch in $O(Ndr)$ arithmetic for sequence length N , head dimension d , and sketch width r , without constructing an $N \times N$ logit matrix. We characterize the estimator’s mean and variance, prove a sharp global attention-divergence bound, and derive an exact $O(Nd^2)$ prefix-moment comparator to distinguish the benefit of sketching from that of avoiding quadratic sequence storage. In a controlled 30-seed synthetic-attention experiment, CASK-Q reduces mean held-out attention-output normalized mean-squared error by 25.20% at nominal 2-bit and 28.13% at nominal 3-bit relative to independent query/key reconstruction, with identical serialized weight sizes and equal update counts. Separate equal-time experiments, six sketch widths, fixed-probe and uncentered controls, exported-weight distribution tests, and numerical checks further characterize the method. Dense joint calibration achieves similar accuracy. The results isolate a calibration component in a common reference quantizer; they do not evaluate compression or downstream quality of pretrained language models.

1 Introduction

Scalar weight quantization is attractive when a model must retain a simple, compact inference representation. A calibration objective determines which errors this representation should preserve. Independent projection reconstruction treats query and key weights separately, whereas attention depends on their product. A perturbation that is unfavorable for one projection can compensate for a perturbation in the other, and an individually small error can be amplified by their interaction.

Attention-aware calibration is already an established direction [1–3]. GSQ explicitly discusses joint query/key optimization, but its reported within-block staging begins with separate linear reconstruction because the joint objective is expensive [4, Appendix C]. This motivates a narrower question than designing a complete compression system: can an inexpensive objective preserve useful joint query/key structure while leaving the exported scalar representation unchanged?

CASK-Q answers this question with a sketch of the *joint, row-centered, causally masked logit error*. Centering removes row-constant shifts that softmax cannot observe. Fresh shared random probes yield an unbiased estimate of the centered squared error. A causal prefix identity evaluates these probes without materializing all pairwise logits. This operation is confined to calibration; the exported model does not execute the sketch.

The study makes three contributions. First, it specifies the centered causal operator, its implicit computation, and the precise scope of its stochastic guarantees. Second, it supplies an exact prefix-second-moment comparator, showing that randomization is not necessary for linear dependence

on sequence length: sketching instead reduces dependence on head dimension from quadratic to linear in a chosen width. Third, it provides a controlled empirical study with 30 teacher/data seeds per bit width, separate width and equal-time ablations, held-out distribution checks, and serialized-weight verification. The contributions are a particular calibration construction and its analysis, not a claim that attention coupling, random trace estimation, or causal scans are new in isolation.

2 Related work

Quantization and coupled calibration. GPTQ uses second-order reconstruction information to quantize pretrained weights [5]. GSQ learns discrete scalar-grid assignments and group scales through a Gumbel–Softmax relaxation [4]; the underlying continuous-relaxation constructions are due to Jang et al. [6] and Maddison et al. [7]. CASK-Q uses a small common straight-through scalar quantizer to isolate changes in the calibration objective. Its grid, initialization, optimizer, and model scope differ from the official GSQ implementation. Consequently, the independent condition in this paper is not a GSQ baseline.

APTQ incorporates attention sensitivity into mixed-precision quantization [1]. BoA and TurboBoA account for dependencies within attention modules through attention-aware quantization and error compensation [2, 3]. RCO addresses a complementary problem: optimizing coupled compression choices subject to explicit cost constraints [8]. CASK-Q does not perform mixed-precision allocation, and none of these systems is reimplemented as an empirical baseline here. Recent work on joint attention-output calibration also finds that local improvements need not survive integration with other quantized layers [9]. Our evaluation therefore separates local attention fidelity from downstream language-model performance.

Randomized estimation and causal computation. Frobenius-norm and trace estimation are standard tools in randomized numerical linear algebra [10, 11]. The estimator used here is the ordinary Rademacher quadratic-form estimator, not the Hutch++ algorithm. Associative products and prefix states also arise in linear attention [12]. CASK-Q differs in the object being computed: it sketches a *calibration error* between teacher and student causal logits, and does not approximate or replace softmax attention at inference. The exact prefix-moment identity below makes this distinction explicit and provides an additional deterministic reference.

3 Centered causal attention calibration

3.1 Objective and invariance

For one attention head, let $Q, K \in \mathbb{R}^{N \times d}$ be teacher activations and $\hat{Q}, \hat{K}$ their student counterparts, after positional encoding and any query/key normalization. For allowed positions $j \leq i$, define

$$e_{ij} = \frac{\hat{q}_i^\top \hat{k}_j - q_i^\top k_j}{\sqrt{d}}, \quad D_{ij} = e_{ij} - \frac{1}{i} \sum_{t=1}^i e_{it}; \quad D_{ij} = 0 \quad (j > i). \quad (1)$$

Indices are one-based. The unnormalized target is $\mathcal{L} = \|D\|_F^2$. Centering is the orthogonal projection of each allowed error row onto the complement of the constant vector. It removes exactly the component to which softmax is invariant, but it does not introduce a teacher-probability weighting.

The forward error retains the full interaction

$$\Delta(QK^\top) = \Delta Q K^\top + Q \Delta K^\top + \Delta Q \Delta K^\top. \quad (2)$$

In particular, the final term is not discarded as a first-order approximation. For batch size B and H heads, experiments divide the summed loss by $BHN(N+1)/2$, the number of allowed token pairs. This changes its scale but not its minimizers.

3.2 Implicit random sketch

Let $R \in \mathbb{R}^{N \times r}$ contain independent entries $\pm 1/\sqrt{r}$. Write R_j for the column representation of its j th row. For either teacher or student separately, maintain

$$T_i = \sum_{j \leq i} k_j R_j^\top, \quad v_i = \sum_{j \leq i} k_j, \quad u_i = \sum_{j \leq i} R_j, \quad (3)$$

then set

$$z_i^\top = \frac{q_i^\top (T_i - v_i u_i^\top / i)}{\sqrt{d}}, \quad \hat{\mathcal{L}}_R = \sum_i \|\hat{z}_i - z_i\|_2^2. \quad (4)$$

Teacher and student use the *same* probes. The main method draws fresh probes independently of minibatch and quantizer randomness at each update.

Proposition 1 (Exact prefix identity). *The matrix whose rows are $(\hat{z}_i - z_i)^\top$ equals DR in real arithmetic.*

Proof. Expand $q_i^\top T_i = \sum_{j \leq i} (q_i^\top k_j) R_j^\top$. The correction subtracts the mean allowed teacher logit times $\sum_{j \leq i} R_j^\top$. Apply the same calculation to the student and subtract. The result is row i of Eq. (1) multiplied by R . $\square$

3.3 Calibration and export

Both query and key assignment logits and group scales receive gradients from Eq. (4). Every condition starts from the same scalar-grid initialization and uses the same validation criterion. The independent condition instead minimizes projection reconstruction; dense joint calibration evaluates $\|D\|_F^2$ directly. Thus the main comparison changes the objective and its computational realization, not the permitted representation.

Table 1: Calibration procedure for the primary equal-update experiment. The sketch is absent from exported inference weights.

-
- | | |
|---|---|
| 1 | Initialize scalar-grid assignments and FP16-representable group scales for W_Q, W_K . |
| 2 | Draw a minibatch from the calibration split with a dedicated random generator. |
| 3 | Form teacher and student queries and keys using identical positional encoding. |
| 4 | Draw fresh shared probes; compute the two centered prefix sketches. |
| 5 | Backpropagate their normalized squared difference into both quantizers. |
| 6 | Every 25 updates, retain the checkpoint with lowest validation output NMSE. |
| 7 | Pack integer codes and FP16 scales; reload and check exact tensor equality. |
| 8 | Evaluate the selected checkpoint on the held-out test split. |
-

4 Estimator analysis and attention fidelity

4.1 Mean, variance, and gradients

Proposition 2 (Fixed-error estimator). *For D independent of the fresh probes, with $A = D^\top D$,*

$$\mathbb{E}_R \hat{\mathcal{L}}_R = \|D\|_F^2, \quad \text{Var}_R(\hat{\mathcal{L}}_R) = \frac{2}{r} \sum_{j \neq k} A_{jk}^2. \quad (5)$$

If the same probes are shared across batch/head instances, replace A by $\sum_{b,h} D_{bh}^\top D_{bh}$ in the variance formula.

Proof. Write each probe column as $\varepsilon/\sqrt{r}$, where the entries of ε are independent signs. Since $\mathbb{E}[\varepsilon\varepsilon^\top] = I$, the quadratic form has mean $\text{tr}(A)$. Its diagonal contribution is constant, and its off-diagonal contribution is $2 \sum_{j < k} A_{jk} \varepsilon_j \varepsilon_k$. Different unordered pairs have zero covariance, giving variance $4 \sum_{j < k} A_{jk}^2 = 2 \sum_{j \neq k} A_{jk}^2$. Averaging r independent columns divides variance by r . With shared probes, summing quadratic forms first replaces A by the stated aggregate. $\square$

For differentiable continuous activations independent of the newly drawn R , differentiation commutes with the finite probe expectation. Their stochastic gradients therefore average to the gradient of the centered objective. This statement does *not* make a straight-through estimator for discrete assignments unbiased. Nor is the sketch an unbiased estimator of attention KL, output error, or language-model loss. A fixed low-width sketch can leave a null space that an optimizer could exploit; the fixed-probe ablation measures its behavior rather than assuming failure.

4.2 A sharp global divergence bound

Proposition 3 (Attention-distribution bound). *For teacher logits a , student logits b , distributions $p = \text{softmax}(a)$ and $q = \text{softmax}(b)$ on one allowed row, and $\delta_c = b - a - \text{mean}(b - a)\mathbf{1}$,*

$$\text{KL}(p\|q) \leq \frac{1}{4} \|\delta_c\|_2^2. \quad (6)$$

The constant $1/4$ is sharp uniformly over row length and logits.

Proof. Let $f(x) = \log \sum_j e^{x_j}$. Its Hessian is $\text{diag}(p) - pp^\top$. The absolute sum of row j is $2p_j(1 - p_j) \leq 1/2$, so its spectral norm is at most $1/2$. The Bregman divergence $f(b) - f(a) - \nabla f(a)^\top (b - a)$ equals $\text{KL}(p\|q)$. Taylor’s integral remainder bounds it by $\|b - a\|_2^2/4$. Shift invariance permits replacing $b - a$ by δ_c . For sharpness, take $a = (0, 0)$ and $b = (t, -t)$: the ratio is $\log \cosh(t)/(2t^2) \rightarrow 1/4$ as $t \rightarrow 0$. $\square$

Summing over rows gives $\sum_i \text{KL}(p_i\|q_i) \leq \|D\|_F^2/4 = \mathbb{E}_R \hat{\mathcal{L}}_R/4$. This is an upper bound for the exact objective or the sketch expectation, not for every random sample. For the normalized training loss, the corresponding factor $BHN(N + 1)/2$ must be restored.

Corollary 4 (Output bound with fixed values). *For fixed allowed value vectors v_j , let $\Delta_V = \max_{j,k} \|v_j - v_k\|_2$. Then*

$$\left\| \sum_j (p_j - q_j) v_j \right\|_2^2 \leq \frac{\Delta_V^2}{8} \|\delta_c\|_2^2. \quad (7)$$

Proof. The positive and negative masses of $p - q$ each sum to $\text{TV}(p, q)$. Their normalized weighted averages lie in the convex hull of the values, whose diameter is at most Δ_V . The output difference is therefore bounded by $\Delta_V \text{TV}(p, q)$. Squaring and applying Pinsker’s inequality, $\text{TV}^2 \leq \text{KL}(p\|q)/2$, followed by Proposition 3, gives the result. $\square$

4.3 Why the surrogate need not be monotone in quality

A global upper bound does not imply that every reduction of its right-hand side reduces its left-hand side. Consider teacher logits $a = (0, 0, -20)$. Perturbation $\delta_A = (0.1, -0.1, 0)$ has centered squared error 0.02 and attention KL approximately 4.99×10^{-3} . Perturbation $\delta_B = (0, 0, 1)$ has larger centered squared error $2/3$ but KL approximately 7.40×10^{-10} . The second perturbation mostly affects a negligible-probability key. This explicit counterexample motivates reporting actual attention outputs and retaining separate validation data. The fixed-value corollary likewise does not cover value quantization, residual paths, or downstream layers.

5 A deterministic prefix-moment comparator

Avoiding an $N \times N$ matrix is not unique to sketching. The same centered loss has an exact prefix-second-moment formulation. Define $2d$ -dimensional vectors

$$a_i = \begin{bmatrix} \hat{q}_i - q_i \\ q_i \end{bmatrix}, \quad b_j = \begin{bmatrix} \hat{k}_j \\ \hat{k}_j - k_j \end{bmatrix}, \quad s_i = \sum_{j \leq i} b_j, \quad M_i = \sum_{j \leq i} b_j b_j^\top. \quad (8)$$

Then $a_i^\top b_j = \hat{q}_i^\top \hat{k}_j - q_i^\top k_j$, including both first-order terms and the interaction term.

Proposition 5 (Exact causal second moments). *With Eq. (8),*

$$\|D\|_F^2 = \frac{1}{d} \sum_{i=1}^N \left(a_i^\top M_i a_i - \frac{(a_i^\top s_i)^2}{i} \right). \quad (9)$$

Proof. For each allowed row, expand $\sum_{j \leq i} (e_{ij} - \bar{e}_i)^2 = \sum_{j \leq i} e_{ij}^2 - i \bar{e}_i^2$ and substitute $e_{ij} = a_i^\top b_j / \sqrt{d}$. The squared-logit sum is $a_i^\top M_i a_i / d$, and the mean correction is $(a_i^\top s_i)^2 / (id)$. $\square$

The difference factorization avoids constructing the final result by subtracting three separately squared teacher/student products. It does not eliminate all finite-precision cancellation in Eq. (9). Tests compare values and gradients against the dense expression, including exact teacher/student equality.

Table 2: Per-head arithmetic and intermediate-storage order for the supplied vectorized implementations. All three compute calibration losses, not inference attention. Batch and head factors multiply storage.

Operator	Arithmetic	Intermediates	Objective
Dense logits	$O(N^2 d)$	$O(N^2)$	Exact
Prefix second moments	$O(N d^2)$	$O(N d^2)$	Exact
Centered sketch	$O(N d r)$	$O(N d r)$	Stochastic

The exact comparator shows that the sketch’s computational target is reduction of the d^2 prefix state to dr when $r \ll d$, not the exclusive ability to obtain linear dependence on N . Although the forward recurrences admit small running states, the supplied PyTorch autograd programs retain sequence-indexed intermediates. They are not constant-memory training implementations. A short sequence, wide sketch, or optimized dense kernel can reverse a runtime advantage; Section 7.5 measures this directly.

6 Experimental protocol

6.1 Teacher, data, and scalar representation

Every main seed generates an independent random single-head teacher, not a pretrained network. Input width is 32, head dimension is 16, and sequence length is 32. Query, key, and value weight entries are independent $\mathcal{N}(0, 1/32)$. A seed-specific input transform is formed as $C = G \text{diag}(\text{linspace}(0.5, 1.5, 32))$, where $G_{ij} \sim \mathcal{N}(0, 1/32)$. Independent Gaussian draws Z produce $X = ZC$: 128 training, 64 validation, and 128 test sequences. Conventional rotary position encoding is applied to Q and K. There is no Q/K normalization or cross-head key sharing.

Only Q/K weights are quantized. Values and activations remain FP32 in every condition. Each group of 16 weights uses a positive FP16-representable scale and 2^b odd integer levels: at 2-bit the grid is $\{-3, -1, 1, 3\}$. This symmetric mid-rise grid has no zero level and is intentionally documented rather than presented as a standard GSQ or GGUF format. Hard assignments select grid entries; straight-through soft categorical assignments provide optimization gradients. The quantizer and primary runner are unchanged from the earlier experiment artifact. Appendix A gives every training hyperparameter.

6.2 Primary comparison and uncertainty

The main experiment uses seeds 100–129, nominal bit widths 2 and 3, and four conditions: round-to-nearest initialization, independent Q/K reconstruction, dense joint centered-logit calibration, and fresh rank-8 CASK-Q. The three optimized conditions use 400 updates, batch size 8, identical minibatch streams, and the same quantizer-noise stream. A dedicated probe generator avoids consuming either stream. All methods retain their best validation output-NMSE checkpoint at 25-update intervals, including initialization as a fallback. Test metrics are never used for optimization or checkpoint selection; optimized checkpoints are tested after selection.

The primary metric is output NMSE, $\|\hat{O} - O\|_F^2 / \|O\|_F^2$, on held-out sequences with the shared teacher values. We also report teacher-to-student attention KL, centered-logit NMSE, and projection NMSE. Tables use sample standard deviations across seeds, not standard errors. Relative improvement is the ratio-of-means statistic $100(1 - \overline{\text{NMSE}}_{\text{CASK}} / \overline{\text{NMSE}}_{\text{independent}})$. Percentile confidence intervals resample paired seeds 20,000 times with bootstrap seed 20261012. They describe this synthetic population and exploratory protocol, not model-family uncertainty or a preregistered confirmatory test.

Seeds 100–109 had been examined previously and are rerun here. Seeds 110–129 are a new extension. Every primary result uses the current environment; earlier values are not pooled into the main table. Cohort-separated results preserve the distinction between repetition and new evidence. The main comparison is equal-update, not equal-time; Section 7.3 adds a separate time-controlled track.

6.3 Actual storage cost

A Q/K pair contains 1,024 quantized weights. At nominal 2-bit, codes occupy 256 bytes and FP16 scales 128 bytes; nominal 3-bit requires 384 and 128 bytes. The pair’s custom headers add 208 bytes in both cases, giving 592 and 720 bytes respectively. Codes plus scales thus consume 3 and 4 effective bits per Q/K weight; including headers gives 4.625 and 5.625 bits. The total excludes V and any other model parameters. All optimized methods have identical byte counts at the same nominal precision, and exported tensors are checked against the selected in-memory tensors. The format is an experimental scalar-code container, not a deployment-kernel benchmark.

7 Results

7.1 Thirty-seed attention fidelity

Table 3: Held-out attention-output NMSE over 30 synthetic teacher/data seeds (100–129). Entries are mean $\pm$ sample standard deviation. Lower is better; nominal bits refer to weight codes, not scale-inclusive storage.

Method	2-bit	3-bit
Round to nearest	0.21791 ± 0.03152	0.03776 ± 0.00491
Independent	0.08600 ± 0.01051	0.02601 ± 0.00362
Dense joint	0.06429 ± 0.00904	0.01883 ± 0.00214
CASK-Q ($r = 8$)	0.06432 ± 0.00874	0.01870 ± 0.00256

Table 3 shows that CASK-Q reduces mean output error by 25.20% at 2-bit and 28.13% at 3-bit relative to independent reconstruction. Paired 95% bootstrap intervals are [22.48%, 28.13%] and [24.50%, 31.47%], respectively; CASK-Q improves 30/30 and 30/30 paired seeds. The new 20-seed cohort also improves every seed at both widths (Table 4).

Table 4: Paired reduction in mean NMSE relative to Independent. Percentile intervals use 20,000 paired seed resamples. The extension was specified after earlier results were observed.

Cohort	Bits	n	Reduction	95% interval	Wins
Re-execution: 100–109	2	10	25.53%	[22.35, 28.68]%	10/10
Re-execution: 100–109	3	10	26.55%	[20.25, 32.59]%	10/10
New extension: 110–129	2	20	25.05%	[21.32, 29.13]%	20/20
New extension: 110–129	3	20	28.86%	[24.52, 32.82]%	20/20
All: 100–129	2	30	25.20%	[22.48, 28.13]%	30/30
All: 100–129	3	30	28.13%	[24.50, 31.47]%	30/30

Dense joint calibration and CASK-Q have nearly the same mean output error: dense is slightly lower at 2-bit, whereas CASK-Q is slightly lower at 3-bit. These small differences do not establish a systematic accuracy advantage of randomization. The supported interpretation is that joint calibration is useful in this experiment and the sketch retains that usefulness.

Table 5: Secondary held-out metrics, averaged over the same 30 seeds. Attention KL is teacher-to-student mean row KL; the two NMSE columns use separate teacher-energy normalizers.

Bits	Method	Attention KL	Logit NMSE	Projection NMSE
2	Independent	0.06416	0.14074	0.07682
2	Dense joint	0.04999	0.10994	0.12796
2	CASK-Q ($r = 8$)	0.05009	0.10929	0.12720
3	Independent	0.01903	0.04152	0.02224
3	Dense joint	0.01433	0.03138	0.03019
3	CASK-Q ($r = 8$)	0.01422	0.03120	0.03036

The secondary metrics provide a mechanistic check. Joint calibration can improve actual attention outputs while increasing projection reconstruction error (Table 5). This is consistent with preserving the interaction rather than separately minimizing each projection’s error. It does not imply that larger projection errors are intrinsically beneficial, nor that a block-local gain transfers to a complete transformer.

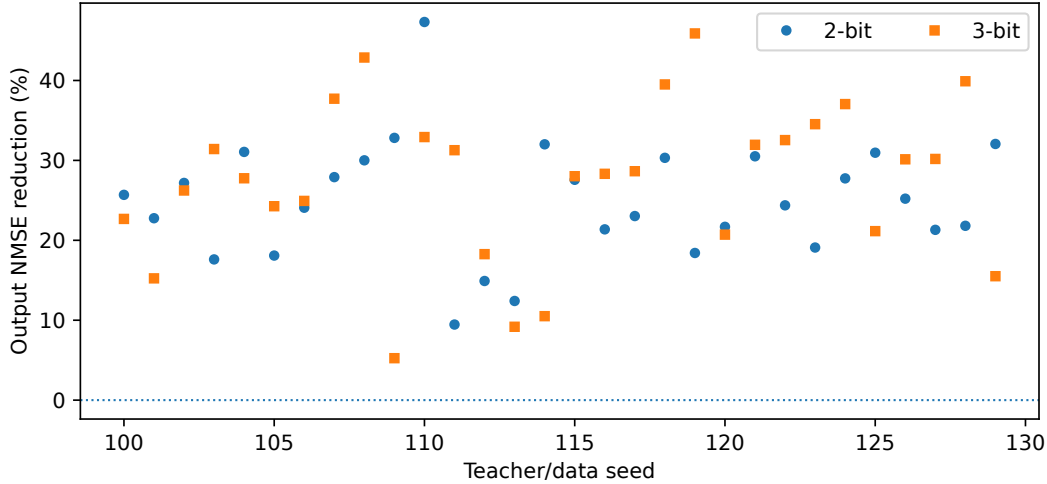


Figure 1: Paired output-NMSE reductions for all 30 primary seeds. Each point compares CASK-Q with the independent condition on the same teacher and test data; positive values favor CASK-Q. These are per-seed reductions, whereas reported aggregate improvements are ratios of means.

7.2 Sketch width, fixed probes, and centering

Table 6: Optimization ablations on separate seeds 300–304; mean $\pm$ sample standard deviation of test NMSE. Every condition uses the same scalar grid, 400 updates, data, and validation selection rule. No rank is selected using these test scores.

Condition	2-bit	3-bit
Independent	0.08703 $\pm$ 0.01651	0.02520 $\pm$ 0.00170
Dense joint	0.06044 $\pm$ 0.00623	0.01781 $\pm$ 0.00149
CASK-Q, fresh $r = 1$	0.06272 $\pm$ 0.00623	0.01718 $\pm$ 0.00300
CASK-Q, fresh $r = 2$	0.05815 $\pm$ 0.00908	0.01764 $\pm$ 0.00218
CASK-Q, fresh $r = 4$	0.05684 $\pm$ 0.00506	0.01824 $\pm$ 0.00225
CASK-Q, fresh $r = 8$	0.05844 $\pm$ 0.00640	0.01774 $\pm$ 0.00227
CASK-Q, fresh $r = 16$	0.06034 $\pm$ 0.00714	0.01748 $\pm$ 0.00128
CASK-Q, fresh $r = 32$	0.05966 $\pm$ 0.00452	0.01801 $\pm$ 0.00245
Fixed probes, $r = 8$	0.06200 $\pm$ 0.00782	0.01890 $\pm$ 0.00230
Uncentered, fresh $r = 8$	0.05773 $\pm$ 0.00564	0.01819 $\pm$ 0.00299

Ten optimization conditions on separate seeds 300–304 test widths $r \in \{1, 2, 4, 8, 16, 32\}$, fixed rank-8 probes, and uncentered rank-8 sketches, alongside independent and dense controls. All 100 runs use the same grid, 400 updates, data construction, and validation rule. No width is selected using these test scores, and the main rank remains 8.

Every joint condition has lower mean test output NMSE than the independent condition at both widths (Table 6). Increasing rank does not monotonically improve optimized test error. The fixed-probe condition has higher mean error than fresh rank-8 probes in both columns, but five seeds do not establish universal failure of fixed sketches. Removing centering improves the 2-bit mean and worsens the 3-bit mean relative to centered rank 8. Thus centering has an exact invariance justification, but these results do not establish universal empirical superiority over the uncentered control.

A distinct frozen-matrix diagnostic draws 4,096 sketches per width. Its relative estimator standard deviation closely follows the $r^{-1/2}$ prediction of Eq. (5) (Figure 2). This validates a fixed-error estimator property, not monotonic convergence of the quantization optimizer.

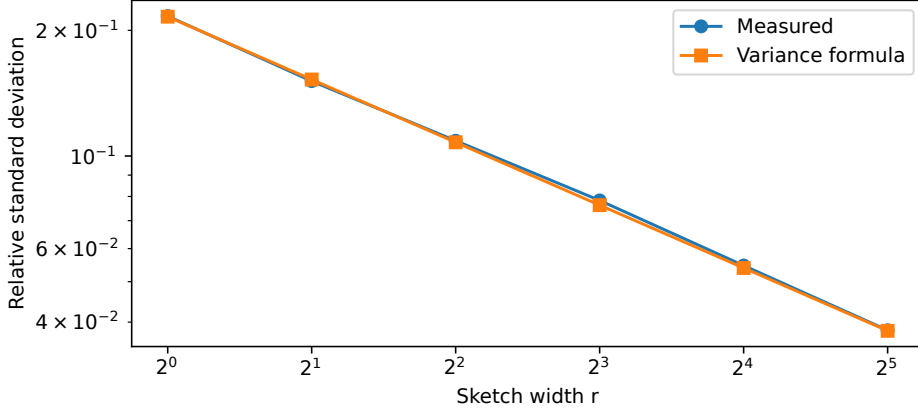


Figure 2: Frozen-error estimator relative standard deviation over 4,096 independent draws per width, compared with the exact variance formula. This diagnostic holds D fixed and is separate from the five-seed optimization-width ablation.

7.3 Equal-time calibration

The independent objective is cheaper per update at the primary short context. To examine this confound, a separate track uses seeds 400–409 and a common 2-second calibration target on one CPU thread for each optimized condition. Temperature and noise follow the same endpoints as the main experiment, but progress according to elapsed time rather than update index. Validation targets occur at eighths of the budget, with initialization and final validation included in selection. Update work, probes, intermediate and final validation, and checkpoint copies are timed; data/optimizer initialization and export are outside the timed region. One update and final validation can slightly exceed the target, so actual timings and counts are reported. Condition order is shuffled within each seed/bit, after warming all operators on a disjoint case.

Table 7: Separate 2-second calibration-target experiment, seeds 400–409. NMSE is mean $\pm$ sample standard deviation; updates and actual timed seconds are means. The timed region includes validation and can slightly overshoot the target.

Bits	Method	Output NMSE	Updates	Seconds
2	Independent	0.07616 ± 0.00731	1182.4	2.0039
2	Dense joint	0.05433 ± 0.00593	1037.2	2.0038
2	CASK-Q ($r = 8$)	0.05474 ± 0.00488	874.9	2.0044
3	Independent	0.02411 ± 0.00239	1023.9	2.0044
3	Dense joint	0.01765 ± 0.00308	913.3	2.0041
3	CASK-Q ($r = 8$)	0.01761 ± 0.00293	761.6	2.0041

This track is not pooled with the equal-update experiment: both the annealing schedule and the validation cadence differ. At matched time targets, CASK-Q reduces mean output error by 28.12% at 2-bit and 26.96% at 3-bit relative to independent reconstruction, with 10/10 and 10/10 wins (Table 7). Paired bootstrap intervals are [23.47%, 32.38%] and [21.04%, 32.62%], respectively. The measurement concerns this reference calibration loop, not end-to-end compression or serving throughput. Raw timings, complete update counts, validation records, and selected exported tensors accompany all 60 conditions.

Table 8: Post-export distribution tests on all 30 main seeds, with 128 new sequences per distribution and no parameter or checkpoint updates. Positive reduction favors CASK-Q.

Distribution	Bits	Independent	CASK-Q	Reduction	Wins
Matched, new samples	2	0.08561	0.06429	+24.91%	30/30
Matched, new samples	3	0.02597	0.01880	+27.62%	30/30
Permuted covariance	2	0.18529	0.15389	+16.95%	28/30
Permuted covariance	3	0.04898	0.04121	+15.87%	27/30
Radial t_3	2	0.15282	0.13116	+14.17%	21/30
Radial t_3	3	0.05545	0.04639	+16.34%	22/30

7.4 Post-export activation-distribution tests

All 180 optimized primary checkpoints are reloaded from their serialized Q/K weights. Their original held-out metrics match the recorded metrics exactly in the current environment; 60 round-to-nearest metric reconstructions are also checked. Each of these 240 conditions is then evaluated on 128 fresh sequences under three distributions, producing 720 additional evaluations without optimization or checkpoint changes.

The first distribution retains the input covariance. The second cyclically permutes input coordinates, changing covariance relative to the learned projections. The third uses a covariance-matched radial t_3 mixture with heavy tails; Appendix A defines the normalization. CASK-Q retains positive mean improvements under all three tests, but covariance and heavy-tail shifts introduce individual regressions (Table 8). In particular, the radial- t_3 win counts fall to 21/30 and 22/30. Mean improvements therefore coexist with meaningful sensitivity to the activation distribution.

7.5 Calibration-operator timing

Table 9: Calibration-loss forward/backward time in milliseconds on one AMD EPYC 9V74 CPU thread. Medians of nine repetitions after three warmups; $d = 64$, $r = 8$. All required teacher-dependent work is recomputed.

N	Dense	Exact prefix moments	CASK sketch
128	0.64	58.98	1.66
512	4.90	441.35	5.76
2048	112.80	1872.01	23.29
4096	608.41	4064.07	48.08

The timing experiment compares dense logits, exact prefix moments, and the sketch at $N \in \{128, 512, 2048, 4096\}$, $d = 64$, and $r = 8$. It times forward and backward through each calibration loss on one AMD EPYC 9V74 CPU thread, with three warmups and nine measured repetitions. Method order rotates and reverses across repetitions. Required teacher-dependent work is recomputed in all conditions; probe generation, model projections, optimizer steps, and inference are not timed. The exact moment loss is numerically checked against the dense loss at each shape before timing.

The sketch is slower than dense computation at $N = 128$ and 512, and faster at 2048 and 4096 (Table 9). The supplied exact-moment implementation is slower than both throughout this range despite its linear-in- N arithmetic. Its large materialized second-moment tensors and execution pattern make asymptotic notation an inadequate predictor of these wall-clock times. These measurements compare reference PyTorch realizations, not optimized kernels or the best achievable implementations. Peak memory, GPU speed, and inference throughput are not measured.

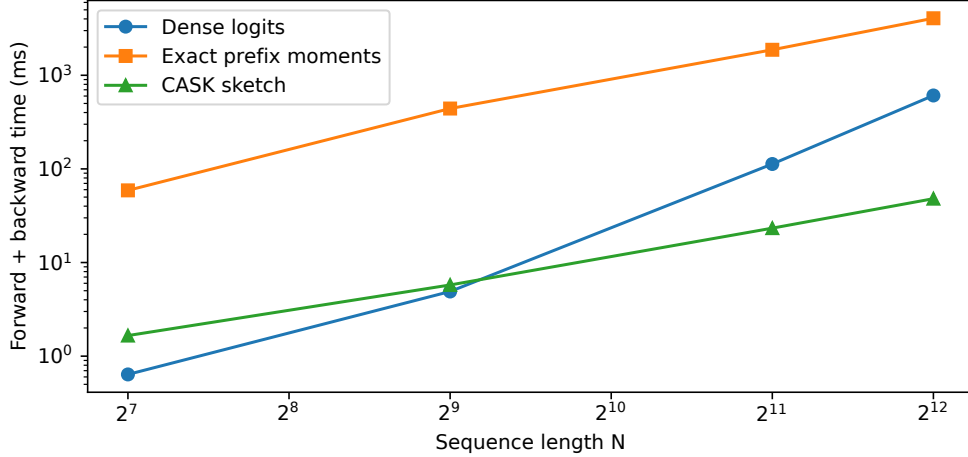


Figure 3: Median calibration-loss forward/backward time from Table 9. Logarithmic axes emphasize the observed crossover and the cost of the vectorized exact-moment comparator. The figure does not measure inference speed.

7.6 Numerical and implementation checks

Twenty-six tests cover the original core and added analysis. They include dense/prefix value and gradient agreement, shift invariance, causality, exhaustive small-probe mean and variance checks, shared-probe batch aggregation, softmax bounds, packed-code round trips, and the exact second-moment comparator. The added ablation runner reproduces the original runner exactly on three control conditions in the current environment, including byte-identical exported weights.

At sequence length 257, the largest tested FP64 prefix-versus-dense absolute difference is 1.60×10^{-14} , and the FP32 relative difference is 7.80×10^{-8} . These are measured numerical checks, not universal floating-point error bounds. Tests pass on the reported environment; they are not a formal verification of every architecture, precision, or possible input.

8 Limitations and scope

The experiment uses synthetic single-head teachers and a specific reference scalar quantizer. It does not evaluate pretrained language models, full-model perplexity, downstream tasks, value or MLP quantization, mixed-precision allocation, or actual deployment kernels. GQA, Q/K normalization, quantized-prefix activations, sliding windows, packed-document masks, and long-context model calibration require additional integration and testing. The current helper supports a standard full causal mask with shared teacher/student positions.

The independent reconstruction condition is a controlled objective ablation, not the strongest available independent quantizer. Objective scaling and clipping interact with a shared learning-rate schedule; no claim of optimizer-independent superiority follows. The equal-time track addresses one compute confound within the reference loop but does not include candidate generation, budget search, or kernel engineering in a full quantization system. Width and distribution comparisons are exploratory and involve multiple comparisons without confirmatory significance claims.

Exact local identities do not imply unbiased discrete training or monotonic downstream improvement. The counterexample and distribution regressions show why output- or model-level validation is necessary. Cross-environment re-execution also does not produce identical nonconvex optimization outcomes; Appendix B reports the discrepancy rather than attributing it to an

unverified cause. The appropriate next evaluation is a same-format, same-budget intervention in a real quantization pipeline, retaining the original pipeline as a fallback and comparing actual model outcomes.

9 Conclusion

CASK-Q uses row-centered causal sketches to jointly calibrate scalar-quantized queries and keys without an explicit sequence-by-sequence loss matrix. Its exact prefix identity, estimator variance, attention-divergence bound, and deterministic second-moment comparator clarify what the method computes and what it does not guarantee. Across 30 controlled synthetic seeds per bit width, it retains the output-fidelity benefit of dense joint calibration while matching the independent method’s serialized weight size. Additional compute, width, distribution, and numerical experiments delimit that result. The evidence supports CASK-Q as a characterized calibration component, with pretrained-model effectiveness remaining an empirical question.

Reproducibility and AI assistance

The accompanying ancillary material contains executable code, generated tables, machine-readable results, selected quantized weights, environment records, and an integrity manifest. All reported experimental conditions, including regressions and timing samples, are retained. Generative AI tools assisted mathematical exposition, implementation, experiment execution, and manuscript preparation. The numerical checks and executable artifacts document the verification performed; they do not substitute for independent scientific review.

References

- [1] Ziyi Guan, Hantao Huang, Yupeng Su, Hong Huang, Ngai Wong, and Hao Yu. APTQ: Attention-aware post-training mixed-precision quantization for large language models, 2024. URL <https://arxiv.org/abs/2402.14866>.
- [2] Junhan Kim, Ho-young Kim, Eulrang Cho, Chungman Lee, Joonyoung Kim, and Yongkweon Jeon. BoA: Attention-aware post-training quantization without backpropagation, 2024. URL <https://arxiv.org/abs/2406.13474v3>. Revised 2025.
- [3] Junhan Kim, Yeo Jeong Park, Seungwoo Son, Chungman Lee, Ho-young Kim, Joonyoung Kim, and Yongkweon Jeon. TurboBoA: Faster and exact attention-aware quantization without backpropagation, 2026. URL <https://arxiv.org/abs/2602.04929>.
- [4] Alireza Dadgarnia, Soroush Tabesh, Mahdi Nikdan, Michael Helcig, Eldar Kurtić, Maximilian Kleinegger, and Dan Alistarh. GSQ: Highly-accurate low-precision scalar quantization for LLMs via gumbel-softmax sampling, 2026. URL <https://arxiv.org/abs/2604.18556v2>. Version 2, May 15, 2026.
- [5] Elias Frantar, Saleh Ashkboos, Torsten Hoefer, and Dan Alistarh. GPTQ: Accurate post-training quantization for generative pre-trained transformers, 2022. URL <https://arxiv.org/abs/2210.17323>.
- [6] Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with Gumbel-Softmax, 2017. URL <https://arxiv.org/abs/1611.01144v5>. arXiv:1611.01144v5.

- [7] Chris J. Maddison, Andriy Mnih, and Yee Whye Teh. The concrete distribution: A continuous relaxation of discrete random variables, 2017. URL <https://arxiv.org/abs/1611.00712v3>. arXiv:1611.00712v3.
- [8] Michael Helcig and Dan Alistarh. Model compression with exact budget constraints via riemannian manifolds, 2026. URL <https://arxiv.org/abs/2605.00649v3>. arXiv:2605.00649v3, September 27, 2026.
- [9] Nafiseh HosseinpourFardi, Negar Alihadi, Mahmoudreza Babaei, Milad Hosseini, and Adrian Weller. From attention sensitivity to layer role: Revisiting mixed-precision quantization of transformers, 2026. URL <https://arxiv.org/abs/2609.34866>.
- [10] Michael W. Mahoney. Randomized algorithms for matrices and data, 2011. URL <https://arxiv.org/abs/1104.5557>.
- [11] Raphael A. Meyer, Cameron Musco, Christopher Musco, and David P. Woodruff. Hutch++: Optimal stochastic trace estimation, 2021. URL <https://arxiv.org/abs/2010.09649v5>.
- [12] Angelos Katharopoulos, Apoorv Vyas, Nikolaos Pappas, and François Fleuret. Transformers are RNNs: Fast autoregressive transformers with linear attention. In *Proceedings of the 37th International Conference on Machine Learning*, 2020. URL <https://arxiv.org/abs/2006.16236>.

A Complete calibration and evaluation settings

Scalar-grid optimization. For each weight group, the initial scale is its maximum absolute weight divided by the maximum integer level, clamped positive. Assignment logits initialize to -1.5 times squared distance between the scaled weight and each grid entry. Hard forward assignments use categorical argmax, and straight-through soft probabilities carry the gradient. Positive scales are represented as exponentiated log scales and rounded through FP16 with a straight-through gradient; exported and evaluated scales are therefore FP16-representable.

Adam uses learning rate 0.035 for assignment logits and 0.004 for log scales, with default $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 10^{-8}$, and no weight decay. For the main 400-update schedule, $f = t/399$, temperature is $\tau = 1.5(0.12/1.5)^f$, and Gumbel-noise amplitude is $0.25(1 - f)$. Gradients over both quantizers are clipped to total norm 1.0. Minibatches sample eight training sequences with replacement. Validation is checked every 25 updates; initialization remains an eligible checkpoint. No best-width or best-objective condition is selected from test metrics.

The primary seed sets, their extension/repetition split, and the statistical rule are stored in `results/submission_protocol.json`. The equal-time track uses the same initialization, grid, optimizer, and endpoints, but replaces f by elapsed timed-region duration divided by 2 seconds. Seven intermediate validation targets occur at $k/8$ of the time target, $k = 1, \dots, 7$, and a final validation follows the last update. The target can be exceeded by at most the work finishing at the cutoff plus final validation and checkpoint copying; the empirical overshoot is recorded for every condition, rather than assumed zero.

Secondary metrics. Attention KL is the mean, over batch and causal rows, of teacher-to-student categorical KL. Centered-logit NMSE divides the summed squared centered error by the teacher’s summed squared centered logits. Projection NMSE divides the sum of squared Q and K projection errors by the sum of their teacher energies. All output comparisons use the same value activations and causal softmax for teacher and student. These normalized losses are not percentages of language-model accuracy.

Fresh-distribution construction. For each main teacher, all methods use the same three additional input sets, generated independently of training, validation, and original test data. Matched samples have $X = ZC$ for a new standard Gaussian Z . The covariance-shift case cyclically permutes the columns of C by seven positions before applying it to Gaussian samples. For the radial heavy-tail case, each token draws $U \sim \chi_3^2$ independently and uses $X = (Z/\sqrt{U})C$. Since $\mathbb{E}[U^{-1}] = 1$ for three degrees of freedom, this is the variance-normalized t_3 mixture: it retains the Gaussian construction’s covariance while lacking a finite fourth moment. A fixed 128-sequence sample can therefore produce unstable squared-error estimates. These tests are distribution probes, not substitutes for naturally occurring text.

Environment. Primary, ablation, equal-time, and new timing experiments use Python 3.13.5, PyTorch 2.10.0+cpu, NumPy 2.3.5, and one PyTorch thread on an AMD EPYC 9V74 host. No GPU is available in this run. Full PyTorch build information is recorded. CPU frequency and other host activity are not controlled, so timing results are local measurements, not portable performance guarantees.

B Re-execution, artifacts, and complete seed results

The original core and primary experiment source files retain their input-artifact hashes. The preceding verification release’s 265-file integrity manifest was checked before modification. The current release has a new manifest, so additions are not confused with the original artifact. New scripts add ablations, an exact-moment operator, time-controlled calibration, reload checks, and table generation without changing the primary quantizer or core.

Ten previously examined seeds are re-executed on the current AMD environment; the preceding run used an Intel Xeon Platinum 8370C host. The largest absolute condition-level output-NMSE difference across these repeated cases is 0.010809. The cause is not isolated. Earlier and current full-precision values are retained separately in the machine-readable comparison. Consequently, the main table is computed entirely from the current 30-seed run, not a mixture of the most favorable values from different executions.

The primary study comprises 180 optimization runs and 60 initialization evaluations; width/centering controls add 100 optimization runs, and the time-controlled track adds 60. The three optimized tracks export 680 Q/K files, excluding six parity-check files and six operator-warmup files. Every such experimental export is reloaded immediately for exact tensor comparison. A separate metric reconstruction checks all 400 primary, ablation, equal-time, and initialization conditions from exported weights where applicable, including their selected validation scores. Three fresh-distribution evaluations of each of the 240 primary conditions give 720 further metric records. Tests, diagnostic repetitions, and warmups are not counted as independent experimental seeds.

Regenerating the results. The ancillary root contains the experiment scripts and recorded outputs. `reproduce.py` supports a checks-only mode and explicit experiment tracks. `analyze_submission.py` and `analyze_compute.py` regenerate the paper’s tables from JSON files; `make_figures.py` regenerates its figures. Full reruns are computational experiments, not expected to be bit-identical across different CPU libraries or hosts. The original JSON and exported weights remain the record underlying this manuscript. No downloads, access tokens, or external training data are required for the synthetic experiments.

Table 10: Complete per-seed held-out NMSE for seeds 100–129. Full-precision machine-readable results include all secondary metrics.

Seed	2b Indep.	2b Dense	2b CASK	3b Indep.	3b Dense	3b CASK
100	0.08778	0.06028	0.06523	0.02709	0.01862	0.02094
101	0.07695	0.05507	0.05943	0.02295	0.01693	0.01945
102	0.07553	0.05607	0.05500	0.02137	0.01606	0.01577
103	0.09423	0.07163	0.07764	0.03105	0.01804	0.02129
104	0.06614	0.04774	0.04560	0.02068	0.01538	0.01494
105	0.09428	0.08012	0.07722	0.02658	0.01877	0.02013
106	0.08164	0.06184	0.06199	0.02572	0.02071	0.01930
107	0.07101	0.05455	0.05119	0.02069	0.01529	0.01289
108	0.09544	0.06818	0.06680	0.03098	0.01957	0.01770
109	0.09278	0.05947	0.06233	0.02059	0.01913	0.01951
110	0.10996	0.06987	0.05793	0.02666	0.02089	0.01788
111	0.07535	0.06188	0.06823	0.02456	0.01521	0.01688
112	0.08718	0.07112	0.07419	0.02711	0.01952	0.02216
113	0.09088	0.06837	0.07960	0.02520	0.02106	0.02288
114	0.08200	0.05790	0.05575	0.02362	0.02072	0.02114
115	0.10034	0.08892	0.07267	0.03311	0.02321	0.02383
116	0.07326	0.05216	0.05760	0.02087	0.01820	0.01496
117	0.08448	0.06760	0.06502	0.02480	0.01721	0.01769
118	0.09723	0.07315	0.06775	0.02817	0.02042	0.01704
119	0.08846	0.06780	0.07217	0.02922	0.01910	0.01581
120	0.07246	0.05404	0.05675	0.02306	0.01810	0.01828
121	0.08580	0.06792	0.05962	0.02747	0.02292	0.01869
122	0.07483	0.05622	0.05659	0.02792	0.01789	0.01883
123	0.10124	0.07141	0.08191	0.03346	0.02166	0.02190
124	0.09660	0.07125	0.06980	0.03030	0.02031	0.01908
125	0.08898	0.06529	0.06143	0.02375	0.01937	0.01873
126	0.08854	0.06495	0.06622	0.02767	0.01926	0.01933
127	0.07206	0.05506	0.05670	0.02346	0.01669	0.01638
128	0.08571	0.05678	0.06701	0.02764	0.01619	0.01661
129	0.08876	0.07206	0.06031	0.02465	0.01839	0.02083