# WHEN PER-HEAD KV PRECISION HELPS: INTERACTIONS, ALLOCATION, AND SCALE TIMING

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

A common approach to mixed-precision KV caches allocates bits per attention head using isolated sensitivity scores: quantize one head, measure the loss increase, and assign more bits to the heads whose quantization increases loss most. We test three assumptions behind this recipe with paired, teacher-forced measurements on Qwen3-0.6B, Qwen3-1.7B, and Mistral-7B. First, joint low-bit damage is superadditive: when every head is quantized to two bits under a symmetric integer code, joint damage is about 3.3 times the sum of single-head damages on Qwen3-1.7B and 3.0 times on Mistral-7B, whereas four-bit damage on Qwen3-1.7B is nearly additive. Second, under a min-max code, whether keys or values are more sensitive depends on when quantization scales are computed. When scales use complete token groups, two-bit values cause much greater loss than two-bit keys; when scales use only entries already written, this value-side penalty disappears on all three models and the ordering reverses on both Qwen models. Third, the best allocator depends on the operating point: at about 3.3 average bits, measured selection on Qwen3-1.7B lowers loss by 0.044 nats relative to an equal-cost static first-layer rule, whereas at 2.75 bits a static first-16-layer rule gives the lowest held-out loss on Mistral-7B. In a serially filling cache, two-bit keys with four-bit values improve on uniform three bits by 0.16 nats on Mistral-7B at equal modeled cost. Per-head scores thus support useful allocation when combined with joint evaluation of the assembled cache, an explicit scale rule, and an equal-cost static baseline.

## 1 INTRODUCTION

The key-value (KV) cache is a major memory cost of long-context inference with large language models (LLMs) (Kwon et al., 2023; Hooper et al., 2024): it stores the keys and values of past tokens so that autoregressive decoding does not recompute them, and it grows linearly with the context length. Quantizing the cache reduces this cost, and mixed precision keeps more bits where they matter (Liu et al., 2024b; Hooper et al., 2024; Kang et al., 2024). In grouped-query attention (GQA; Ainslie et al., 2023), each KV head is shared by several query heads, so a (layer, KV head) pair is a natural allocation unit. A direct sensitivity measurement quantizes one such unit while all others stay in full precision, records the increase in next-token negative log-likelihood (NLL), and prioritizes the most affected units, the precision analogue of head-level budget allocation in cache eviction (Feng et al., 2024; Fu et al., 2024) and of sensitivity-aware layer allocation (Li et al., 2025). Figure 1(a) summarizes this recipe. We ask when such an isolated measurement guides precision allocation for an assembled low-bit cache, in which all heads are quantized at once.

Nonuniform cache precision is well developed (Li et al., 2025; Liu et al., 2026), yet the per-head recipe rests on three assumptions that are rarely tested separately (Figure 1(b)). First, isolated damages add: a head's damage measured against a full-precision background still describes its contribution when many heads are quantized. Second, key/value (K/V) sensitivity is a property of the model alone: whether keys or values need more bits does not depend on quantizer details. KVTuner observes that the quantization mode can change K/V importance (Li et al., 2025); we isolate one factor, scale timing. Third, measurement is worth its cost: ranking heads by measured damage beats a zero-calibration static rule, such as upgrading the first layers, at the same payload (average bits per cached scalar).



Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

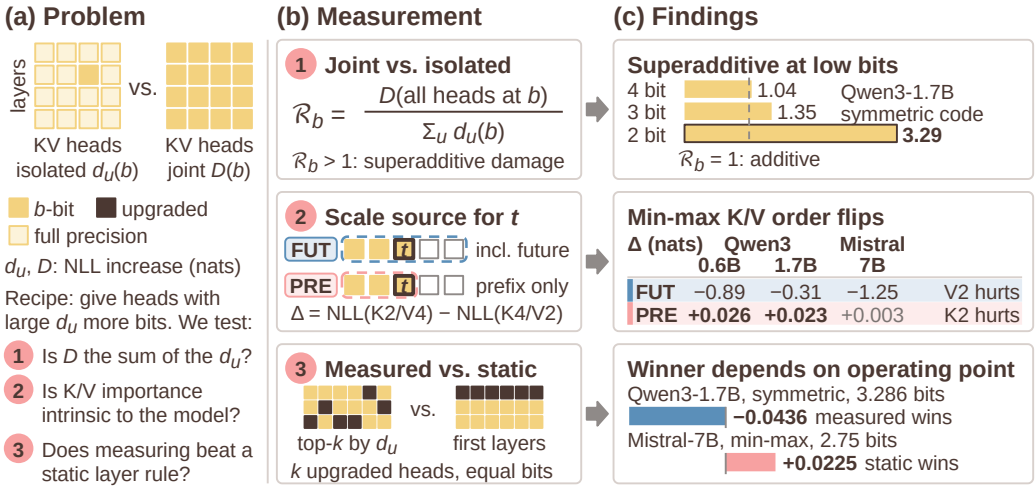



Figure 1: Per-head KV precision: the recipe, what we measure, and what we find. (a) A head score $d_u(b)$ is the NLL increase when only one (layer, KV head) unit $u$ is quantized to $b$ bits; the common recipe gives more bits to heads with large $d_u$. (b) Three tests of this recipe: (1) the ratio $\mathcal{R}_b$ of joint to summed isolated damage; (2) the K/V ordering $\Delta$ at a 3-bit average when each scale uses the full 32-token group, including future entries (FUT), or only the written prefix (PRE); (3) measured top-$k$ selection versus a static first-layer rule at equal upgrade count and payload. (c) Findings. Top: $\mathcal{R}_b$ on Qwen3-1.7B with the symmetric code (Section 4). Middle: mean min-max $\Delta$ over 18 (Qwen) or 12 (Mistral) documents; from FUT to PRE, the value-side penalty disappears on all three models and the ordering reverses on both Qwen models; on Mistral the PRE difference is small but positive (grey; Section 5). Bottom: measured-minus-static NLL at equal payload at two operating points; negative values favor measured selection (Section 6). Blue: FUT, measured; pink: PRE, static.

*Scale timing* specifies which cache entries contribute to a quantization scale, a choice not determined by bit width. A scale may use a complete 32-token group, including entries written after the one being quantized (FUT), only the prefix written so far (PRE), or a completed group converted later in a serial cache. Under FUT, a later group member can change how an earlier entry is quantized even under a causal attention mask. We test each assumption on Qwen3-0.6B, Qwen3-1.7B (Yang et al., 2025), and Mistral-7B (Jiang et al., 2023), with Qwen3-4B and Qwen3-8B as additional checks. The key design choice is to evaluate each assembled cache directly and to compare alternatives at equal payload on the same documents. We compare the joint damage of a quantized cache with the sum of isolated head damages; we compare 2-bit keys with 4-bit values (K2/V4) against the mirror configuration (K4/V2) under FUT and PRE scales, for a symmetric (absmax) and a min-max (affine) integer code; and we compare measured top-$k$ head selection with static layer rules that upgrade the same number of heads. All measurements are paired, teacher-forced NLL differences on book-derived PG-19 text (Rae et al., 2019) and on a same-source retrieval proxy. Quantization is simulated in floating point (quantize, then dequantize), which isolates the code's effect on model quality.

The measurements identify the conditions under which each assumption holds (Figure 1(c)). Isolated scores are a useful ranking signal: at moderate budgets, measured selection beats the equal-payload static first-layer rule. Summing them, however, understates the damage of an assembled low-bit cache, because joint damage is superadditive; joint evaluation captures it directly. The min-max K/V ordering follows the scale rule, and at 2.75 bits on Mistral-7B a zero-calibration static layer rule is the strongest tested allocator. In a serially filling cache, two-bit keys with four-bit values beat uniform three bits at equal cost. Three practices follow: evaluate assembled allocations jointly, state the scale rule with every K/V asymmetry claim, and use a static first-layer rule at equal payload as the first baseline. Our contributions are:

- **Joint damage is superadditive at low precision.** On Qwen3-1.7B with the symmetric code, joint damage rises from 1.04 times the sum of isolated damages at 4 bits to 3.29 times at 2 bits; on Mistral-7B the 2-bit ratio is about 3.0 under both codes (Section 4).
- **Scale timing sets the min-max K/V ordering.** With FUT scales, 2-bit values hurt more than 2-bit keys; with PRE scales this value-side penalty disappears on all three models and the ordering reverses on both Qwen models ($+0.026$ and $+0.023$ nats; $+0.003$ on Mistral-7B), while the symmetric-code ordering keeps its sign (Section 5).
- **The best allocator depends on the operating point (budget and code).** On Qwen3-1.7B with the symmetric code, measured top-64 selection beats a static first-eight-layer rule by 0.044 nats at 3.286 bits; on Mistral-7B with the min-max code at 2.75 bits, a static first-16-layer rule reaches 0.073 nats of damage on held-out documents versus 0.095 for measured selection (Section 6).
- **K2/V4 beats uniform 3 bits in a serial cache.** In a completed-group serial simulation of Mistral-7B with the min-max code, K2/V4 improves on uniform 3 bits by 0.16 nats at equal modeled cost, residual included (Section 7).

## 2 HEAD INTERVENTIONS, ALLOCATIONS, AND COST

Let $u = (\ell, h)$ index a layer and KV head, and let $N$ be the number of such units. For a fixed model, evaluation sample, and quantization protocol, let $L$ be mean next-token NLL, in nats. For a width vector $\mathbf{b} = (b_u)$, define

$$d_u(b) = L(\text{only unit } u \text{ at } b) - L(\text{full}), \qquad D(\mathbf{b}) = L(\mathbf{b}) - L(\text{full});$$

$D(b\mathbf{1})$ denotes every unit at $b$. Unless a K-only or V-only intervention is specified, a unit's K and V share one precision; K$a$/V$b$ denotes $a$-bit keys and $b$-bit values in every unit, and uniform-$b$ denotes $b$ bits for both. We compare joint damage at a uniform width with the isolated sum through

$$\mathcal{R}_b = \frac{D(b\mathbf{1})}{\sum_{u=1}^{N} d_u(b)}.$$

In every reported cell the summed isolated damage is positive, so $\mathcal{R}_b$ is well defined. A ratio above one means joint damage exceeds the isolated sum (superadditive damage) in the assembled cache.

Our measured allocator ranks $d_u(b_0)$ and upgrades the top $k$ units from $b_0$ to $b_1$. We evaluate the resulting $D(\mathbf{b})$ directly. The score is the isolated damage at the base width, rather than the improvement $d_u(b_0) - d_u(b_1)$ or the marginal improvement in a quantized background. Richer scoring and search variants are examined in Section 6.2 and Appendices A.3 and A.7. Random and bottom-ranked sets provide selection controls; first-layer blocks provide a zero-calibration alternative.

The payload budget averages over K and V scalars. With equal K/V head dimensions,

$$\bar{b} = \frac{1}{2N} \sum_{u=1}^{N} (b_u^K + b_u^V) = b_0 + \frac{k(b_1 - b_0)}{N}$$

for a joint K/V upgrade. Qwen3-1.7B has $N = 224$ units. Its top-64 upgrades therefore cost 3.286 bits for $3 \to 4$, and 4.571 bits for $4 \to 6$; top-128 $4 \to 6$ costs 5.143 bits. These are payload costs of the quantized K/V entries. The full set of isolated scores $\{d_u(b)\}$ (a head field) requires $N$ intervened forwards plus a shared full-precision reference per calibration batch. Computing one recorded six-document Qwen field took about 245 GPU-seconds in our harness, a one-time calibration cost.

## 3 WHAT IS MEASURED

### 3.1 CODE FAMILY AND SCALE TIMING

In the principal block experiments, a fixed token window (4,096 tokens unless stated) is evaluated in one teacher-forced pass whose attention-visible cache is quantized and dequantized in place. These experiments quantize keys after applying rotary position embedding (RoPE; Su et al., 2021), use groups of 32 token positions, and compare an absmax symmetric integer code with a min-max affine code (code points and scale axes: Appendix A.10). At 2 bits the symmetric code has three

Table 1: Joint damage divided by the isolated-head sum, $\mathcal{R}_b$, with every unit at $b$ bits under full-group (FUT) scaling; values above one are superadditive. Brackets: 90% document-bootstrap intervals. Qwen: six natural documents and 8,000 resamples; Mistral: 12 documents and 20,000 resamples.

| | | $\mathcal{R}_b$ at $b$ bits | | |
|---|---|---|---|---|
| Model | Code | $b = 4$ | $b = 3$ | $b = 2$ |
| Qwen3-1.7B | Symmetric | 1.04 [0.91, 1.18] | 1.35 [1.24, 1.46] | 3.29 [3.06, 3.51] |
| Mistral-7B | Symmetric | 1.27 [1.11, 1.43] | 1.78 [1.68, 1.89] | 3.00 [2.62, 3.39] |
| | Min-max | 1.30 [0.94, 1.72] | 1.70 [1.42, 2.03] | 3.01 [2.71, 3.34] |

represented integer levels, whereas min-max uses four. Appendices A.2 and A.9 report rotation and group-size controls.

**FUT: full-group scaling.** An entry uses a scale estimated over its complete group. During block evaluation, later group members can influence an earlier entry's quantization even though attention is causal. Most head fields and allocation comparisons use this offline intervention protocol.

**PRE: write-time prefix scaling.** Entry $t$ is quantized using only the written group prefix $[g_0, t]$, with that encoding frozen at write time. This removes access to unwritten scale-source entries. PRE is a write-time counterfactual inside the same one-pass block evaluation; the serial protocol below is simulated separately.

**Serial completed-group scaling.** The separate serial simulator leaves recent entries unquantized and converts a group only after it is complete and has left a residual window. It uses no unwritten tokens and differs from PRE in two ways: its scale uses the entire completed group, and conversion is delayed. Section 7 gives the residual and cost denominator. Quantized values are dequantized into floating-point tensors for attention in every experiment.

### 3.2 Samples, uncertainty, and implementation checks

Natural-text samples are book-derived PG-19 windows (Rae et al., 2019). The retrieval proxy combines a 128-token anchor, distractor text from another book, and a 256-token continuation. Both constructions use the same source collection. Most allocation cells have six documents per construction. The three-model FUT–PRE comparison uses 18 natural documents for each Qwen model and 12 for Mistral. Mistral's principal low-budget allocation comparison uses four calibration documents and eight disjoint evaluation documents. Additional models, 8k/16k windows, pool changes, and calibration-transfer results are reported in Appendices A.1–A.7 and A.9.

We use paired differences on the same documents and report 90% document-bootstrap intervals, with sample counts and resample counts given alongside each result. We call a difference resolved when its interval excludes zero. The full-precision manual forward agrees closely with the corresponding official-model loss (Appendix A.9).

## 4 Joint damage is superadditive at low precision

Table 1 gives the ratio of joint to summed single-head damage. On Qwen3-1.7B with the symmetric code, the ratio grows from 1.04 at 4 bits to 3.29 at 2 bits. Mistral shows substantial 2-bit amplification under both codes. At 3 bits, min-max is close to additive on Qwen3-1.7B (ratios 1.06 on natural text and 1.17 on the retrieval proxy; four-document point estimates, Appendix A.1) and superadditive on Mistral. In every row of Table 1, amplification grows as precision falls, and the precision at which it becomes large depends on model and code.

This finding motivates evaluating each assembled allocation instead of adding its isolated scores. Single-head scores remain a useful ranking signal: their ranking selects better upgrades (Section 6)

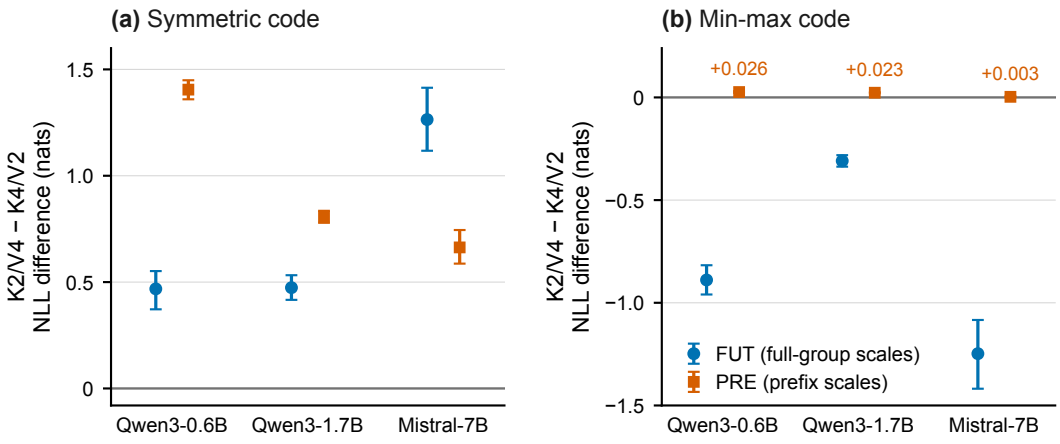



Figure 2: The same nominal bits can produce a different K/V ordering. Points show the mean paired NLL difference K2/V4 minus K4/V2 (3-bit average payload) on natural-text samples; bars are 90% document-bootstrap intervals (20,000 resamples), and some intervals are narrower than the markers. Qwen models use 18 documents each; Mistral uses 12. Circles use full-group scales (FUT), squares use write-time prefix scales (PRE); both arms use the same documents and differ only in scale contents. Positive values favor K4/V2. (a) With the symmetric code, both arms are positive on all three models. (b) With the min-max code, the value-side penalty under FUT disappears under PRE on all three models, and the ordering reverses on both Qwen models (Mistral: $+0.003$ nats).

Table 2: Leave-one-document-out stability of the min-max K/V ordering on the natural-text pools of Figure 2: mean NLL difference K2/V4 minus K4/V2 over all documents, and the range of the mean after omitting each document in turn.

| | | FUT difference (nats) | | PRE difference (nats) | |
|---|---|---|---|---|---|
| Model | Documents | All docs | Leave-one-out range | All docs | Leave-one-out range |
| Qwen3-0.6B | 18 | $-0.8889$ | $[-0.9105, -0.8733]$ | $+0.0262$ | $[+0.0250, +0.0272]$ |
| Qwen3-1.7B | 18 | $-0.3091$ | $[-0.3159, -0.2995]$ | $+0.0228$ | $[+0.0222, +0.0236]$ |
| Mistral-7B | 12 | $-1.2477$ | $[-1.2932, -1.1762]$ | $+0.0029$ | $[+0.0019, +0.0037]$ |

even when their sum misestimates total damage. Rescoring heads in a mixed-precision background produces a small held-out improvement in a study mixing 3- and 4-bit heads (Appendix A.3).

## 5 The K/V ordering changes with scale timing

Figure 2 compares K2/V4 with K4/V2 at the same three-bit average payload. A positive difference means that the configuration with two-bit keys has higher loss; a negative difference means that the configuration with two-bit values has higher loss. Under the symmetric code, the difference is positive for both FUT and PRE on all three models. Under min-max, the full-group comparison is negative, and the prefix comparison becomes positive: $+0.026$ nats on Qwen3-0.6B (90% interval $[+0.022, +0.030]$), $+0.023$ on Qwen3-1.7B ($[+0.020, +0.025]$), and $+0.003$ on Mistral-7B ($[+0.0002, +0.0057]$). The reversal is clear on both Qwen models; on Mistral the difference is small but also positive and resolved. On all three models, the large full-group min-max value-side sensitivity is therefore tied to the scale rule.

The reversal is also stable across documents in the recorded three-model pools. Table 2 recomputes the min-max K2/V4-minus-K4/V2 mean after omitting each document in turn. FUT remains negative and PRE remains positive throughout these ranges, including Mistral's small PRE difference.

The serial implementation adds a complementary view. It also uses only written entries, yet under the min-max code two-bit values remain more harmful than two-bit keys in the tested settings, as

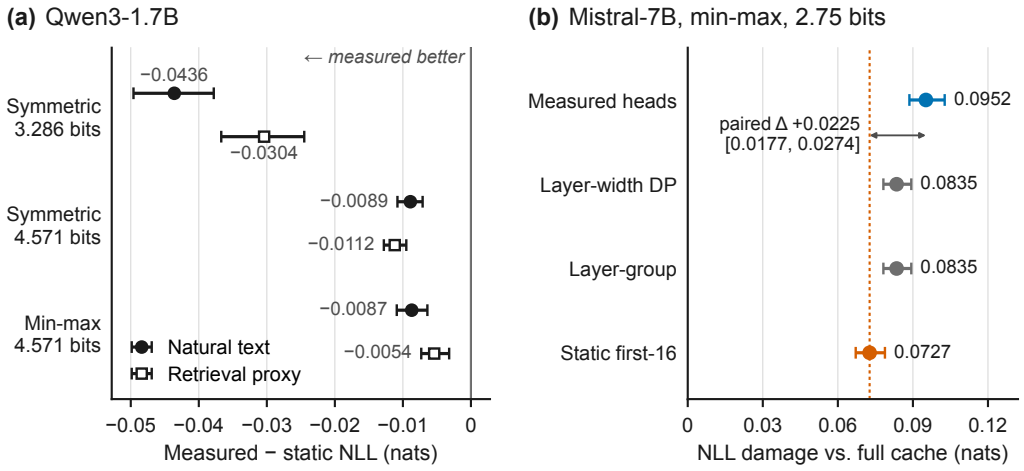



Figure 3: The best allocator depends on the operating point. (a) Qwen3-1.7B measured top-64 minus static first-eight-layer loss, using equal payload and six paired documents per construction; heads are scored on the evaluation documents. Negative values favor measured selection. Filled circles: natural text; open squares: retrieval proxy. (b) Mistral-7B min-max NLL damage at 2.75 payload bits, evaluated on eight documents disjoint from four calibration documents. The static first-16-layer rule attains the lowest loss, and the layer-group allocation and the layer-width dynamic program (DP) choose the same profile here. The dotted line marks the static first-16 mean; the arrow shows the paired measured-minus-static difference over the same eight documents ($+0.0225$ nats, 90% interval $[0.0177, 0.0274]$). Bars are 90% document-bootstrap intervals (20,000 resamples).

under FUT; it shares group-wide scale contents with FUT while differing in conversion time and residual treatment. Appendix A.9 reports residual, scale-window, and group-size controls.

At longer contexts, ordinal comparisons persist while several absolute damages change. On Qwen3-8B at a 2,048-token span and a 3.5-bit split, the min-max difference is likewise negative under FUT ($-0.025$ nats) and positive under PRE ($+0.004$; Appendix A.6).

## 6 When does measuring heads beat a static rule?

### 6.1 Measured selection wins at moderate payloads

For Qwen3-1.7B, static first-eight-layer allocation upgrades 64 KV heads, matching the count upgraded by measured top-64. Figure 3(a) compares their assembled losses. At 3.286 bits with the symmetric code, measured-minus-static is $-0.0436$ nats on natural text (90% interval $[-0.0496, -0.0378]$) and $-0.0304$ on the retrieval proxy. At 4.571 bits it is $-0.0089$ on natural text ($[-0.0108, -0.0071]$) and $-0.0112$ on the proxy. Min-max at the same 4.571-bit payload also favors measured selection. These results compare equal upgrade counts and equal code layouts, with heads scored on the evaluation documents; Appendices A.1 and A.3 replicate the related min-max selection batteries on disjoint held-out pools, including the 4.571-bit measured-versus-static contrast. A top-128 $4 \to 6$ allocation reaches small absolute damage at 5.143 bits, 14.3% below the uniform-6 payload.

Uniform, random, bottom-ranked, K-only, and cross-construction allocations are reported in Appendices A.1, A.3, and A.6. Table 3 collects the min-max selection controls at the base-3 budget across three Qwen3 sizes. At every size, mean damage orders as measured top-64 $<$ random-64 $<$ bottom-64 on both constructions, so measured selection has the lowest mean damage in every row of Table 3; within each battery uniform-3 damage is 1.4–2.2$\times$ the measured damage (for Qwen3-1.7B, see the note to Table 3). The measured-versus-random gap is interval-separated on both constructions at 0.6B and 1.7B and on natural text at 4B. Repeating these batteries on disjoint held-out pools leaves the resolved/unresolved status unchanged for 8/8, 8/8, and 7/8 of the pairwise allocator contrasts

Table 3: Selection controls under the min-max code on Qwen3. Every unit starts at 3 bits; the measured, random, and bottom-ranked allocations upgrade 64 units to 4 bits (payload in Bits). Heads are selected and evaluated on the same six documents per construction. Lower is better; lowest damage per row in bold. Paired intervals and held-out-pool replications: Appendix A.1.

| | | | Mean NLL damage (nats) | | | |
|---|---|---|---|---|---|---|
| Model | Construction | Bits | Measured top-64 | Random-64 | Bottom-64 | Uniform-3 |
| Qwen3-0.6B | Natural | 3.286 | **+0.109** | +0.185 | +0.221 | +0.217 |
| | Proxy | | **+0.086** | +0.130 | +0.157 | +0.168 |
| Qwen3-1.7B | Natural | 3.286 | **+0.049** | +0.084 | +0.093 | *+0.110*[†] |
| | Proxy | | **+0.044** | +0.063 | +0.074 | — |
| Qwen3-4B[‡] | Natural | 3.222 | **+0.048** | +0.071 | +0.074 | +0.088 |
| | Proxy | | **+0.040** | +0.045 | +0.052 | +0.057 |

[†] Italic: point estimate from the separate four-document min-max interaction run of Table A.1 (no proxy value). Within this six-document battery, uniform-3 damage is 2.2× / 1.9× the measured damage (Appendix A.1).
[‡] 288 units (224 for the smaller models); run in bfloat16.

Table 4: Cross-construction transfer and allocation granularity on Qwen3-1.7B: mean NLL damage with 90% document-bootstrap intervals (2,000 resamples); symmetric code, units at 3 bits unless upgraded to 4, six documents per construction. Native sets are selected on the evaluated construction, transferred sets on the other; top-4 layers are those with the largest summed head scores. Lower is better; best at 3.143 bits in bold.

| | | Natural text | | Retrieval proxy | |
|---|---|---|---|---|---|
| Allocation | Bits | Damage (nats) | 90% interval | Damage (nats) | 90% interval |
| Uniform-3 | 3.000 | +0.176 | [0.165, 0.188] | +0.155 | [0.145, 0.166] |
| Bottom-32 heads | 3.143 | +0.178 | [0.166, 0.189] | +0.161 | [0.155, 0.168] |
| Random-32 heads | 3.143 | +0.157 | [0.150, 0.163] | +0.134 | [0.127, 0.141] |
| Top-4 layers (32 heads) | 3.143 | +0.130 | — | +0.112 | — |
| Top-32 heads, transferred | 3.143 | +0.104 | [0.098, 0.110] | +0.097 | [0.092, 0.101] |
| Top-32 heads, native | 3.143 | **+0.098** | [0.092, 0.103] | **+0.085** | [0.078, 0.093] |
| Top-64 heads, native | 3.286 | +0.076 | [0.071, 0.080] | +0.062 | [0.057, 0.067] |
| Uniform-4 | 4.000 | +0.026 | [0.022, 0.030] | +0.026 | [0.023, 0.029] |

(four pairs on two constructions) at 0.6B, 1.7B, and 4B; the one 4B contrast that changes becomes resolved (Appendix A.1).

The measured ranking also transfers across constructions and provides the largest benefit at the head level (Table 4). A top-32 set measured on one construction and deployed on the other reaches +0.104/+0.097 nats of damage (natural text / retrieval proxy) at 3.143 bits, close to the natively selected set and well below random-32, and it beats random selection on all six documents of both constructions. The head rankings from the two constructions are only moderately correlated (Spearman 0.45), yet the random-to-native margin is 4.1–9.3× the transfer-to-native gap. At the same payload, upgrading the four layers with the largest summed head scores reaches +0.130/+0.112 nats, against +0.098/+0.085 for the 32 highest-scoring heads, so head-level allocation uses the same bit budget more efficiently. The selected sets are also stable across document pools: in the disjoint-pool replications of the Table 3 batteries, the min-max top-64 sets share 54/44, 42/40, and 25/20 of 64 heads with the original selections at 0.6B, 1.7B, and 4B, against 18.29 (0.6B, 1.7B) and 14.22 (4B) shared heads expected by chance (Appendix A.1).

### 6.2 A STATIC LAYER RULE IS BEST AT 2.75 BITS

On Mistral, we start from K2/V3 and upgrade half of the value-head scalars to four bits, which gives 2.75 average payload bits. We compare measured head selection with random selection, inexpen-

Table 5: Serial Mistral-7B readout (min-max code) on the 12 PG-19 documents of Section 5. Damage is mean NLL minus the full-cache reference; $\Delta$ is the paired difference from the row above. Intervals: 90% document bootstrap. Bits is the modeled payload-plus-residual cost at the final cache length with a 16-bit residual (6.222 bits with the simulator's float32 residual), equal for all rows.

| Configuration | Bits | Damage (nats) | 90% interval | $\Delta$ (nats) | 90% interval |
|---|---|---|---|---|---|
| Uniform-3 | 4.444 | +0.1917 | [0.1261, 0.2667] | — | |
| K2/V4, completed groups | 4.444 | +0.0304 | [0.0163, 0.0443] | $-0.1612$ | [$-0.2351$, $-0.0994$] |
| K2/V4, windowed channel-K | 4.444 | +0.0271 | [0.0151, 0.0392] | $-0.0033$ | [$-0.0226$, +0.0144] |

sive sensitivity proxies, and static first-16-layer selection. The static rule reaches the lowest loss: 0.0727 nats of damage on the disjoint evaluation pool, against 0.0952 for measured head selection. The paired measured-minus-static difference over these same eight documents is $0.0225$ nats (90% interval $[0.0177, 0.0274]$, 20,000 resamples), so the static advantage is resolved at the document level. A layer-group allocation and a dynamic program (DP) over layer-wise value widths both reach 0.0835 nats (Figure 3(b)). The latter performs an additive search using measured layer effects, inspired by the layer-width allocation of KVTuner (Li et al., 2025).

The paired DP-minus-static difference is 0.0108 nats (90% interval $[0.0060, 0.0162]$), so the zero-calibration static rule also beats this richer search at this operating point. The DP and the layer-group rule select the same profile, so distinct procedures can converge on one assembled model. Appendix A.7 compares published gradient/degradation and range proxies with these controls.

The preferred allocator changes with budget and code. On Mistral with the symmetric code, measured top-128 lowers loss relative to an equal-payload static block at 3.5 bits (measured $-$ static $= -0.0054$ nats, $[-0.0065, -0.0044]$). Measured selection thus beats the equal-payload static rule at moderate budgets (Qwen3-1.7B at 3.286 bits with the symmetric code and at 4.571 bits with both codes; Mistral, on held-out documents, at 3.5 bits with the symmetric code), and the static rule is best at 2.75 bits on Mistral; Appendices A.3–A.7 report further budgets, codes, and transfers.

**Practical implication.** A static layer rule needs no calibration and is the natural first baseline; measurement adds value where it beats this rule on held-out documents at the target budget and code.

## 7 A SERIAL-CACHE READOUT

This section asks whether the K2/V4 split still beats uniform 3 bits when scales are formed only from already-completed groups of a serially filling cache, the tested setting closest to actual serving. The Mistral serial experiment uses the min-max code on the 12 PG-19 documents of Section 5, a 1,024-token prefix, and 128 teacher-forced continuation tokens. A recent residual of $R = 128$ entries remains full precision. The group rule quantizes a 32-token group only when the whole group has left that residual; between updates, up to 159 entries remain unquantized. The windowed channel-K comparator (per-channel key scales) uses a 512-token scale-source window and converts departed tokens individually. These procedures are implemented with float32 source and dequantized caches.

Table 5 summarizes the three configurations. The paired K2/V4-minus-uniform-3 difference is $-0.1612$ nats, with a 90% document interval $[-0.2351, -0.0994]$. Windowed channel-K matches K2/V4 within the interval: their difference is $-0.0033$ nats, with interval $[-0.0226, +0.0144]$. The channel-K variant follows the asymmetric quantization idea of KIVI (Liu et al., 2024b).

At the final length $S = 1152$, all three configurations have 1,024 quantized entries and 128 residual entries per channel, so their cost denominator $b_{\text{eff}} = (1024 \times 3 + 128 \times b_{\text{res}})/1152$ is common to all three at this final group-aligned length. Appendix A.9 extends this accounting to intermediate cache lengths, and Appendix A.8 reports 8-bit floating-point (FP8) and greedy-output controls.

# 8 RELATION TO EXISTING KV-CACHE METHODS

**The cache resource being allocated.** Multi-query and grouped-query attention reduce the number of stored heads (Shazeer, 2019; Ainslie et al., 2023). Heavy-hitter, sink, prompt-aware, layer-adaptive, and head-adaptive methods instead vary which tokens or heads receive cache capacity (Zhang et al., 2023; Xiao et al., 2023; Li et al., 2024; Cai et al., 2024; Feng et al., 2024; Fu et al., 2024; Xiao et al., 2024). Our primary allocation keeps the token positions and architecture fixed while changing bit widths. An isolated loss score is therefore a precision intervention, distinct from a retrieval-head label or an eviction score.

**Quantization geometry and nonuniform precision.** LLM quantization methods handle outliers, reconstruction error, activation scaling, and sparse exceptions in different ways (Dettmers et al., 2022; Xiao et al., 2022; Frantar et al., 2022; Lin et al., 2023; Kim et al., 2023). Hessian-aware mixed precision illustrates why a resource allocation needs an explicit sensitivity surrogate (Dong et al., 2019b;a). KVQuant, KIVI, GEAR, ZipCache, SKVQ, and QJL vary scale granularity, K/V treatment, residual correction, salience, recent-token retention, or projected representations (Hooper et al., 2024; Liu et al., 2024b; Kang et al., 2024; He et al., 2024; Duanmu et al., 2024; Zandieh et al., 2024). Rotations can further change quantization geometry (Ashkboos et al., 2024; Liu et al., 2024a; Saxena & Roy, 2025).

**Sensitivity-based allocation and practical execution.** KVTuner already combines sensitivity-aware mixed precision with final-model evaluation, supports layer or head granularity, and observes that quantization mode changes K/V importance (Li et al., 2025). PM-KVQ progressively lowers cache precision, assigns block-wise bits using gradient–error sensitivity and integer programming, and calibrates with positional interpolation (Liu et al., 2026). Our complementary contribution is a direct comparison of isolated and assembled loss, explicit scale timing, and strong same-code static controls; the layer-width dynamic program here is a specified comparator inspired by the KVTuner search. Efficient attention and cache paging address additional execution costs (Dao et al., 2022; Kwon et al., 2023); our measurements address model quality under a given bit allocation.

# 9 CONCLUSION

Per-head loss measurements guide useful precision allocation when accompanied by joint evaluation of the assembled low-bit cache, comparison with an equal-payload static baseline, and specification of the scale rule. Our experiments find moderate-budget gains for measured selection, a low-budget operating point (Mistral-7B, 2.75 bits) at which a zero-calibration static rule is best, and a min-max K/V ordering set by scale timing: from FUT to PRE scales, the value-side penalty disappears on all three models and the ordering reverses on both Qwen models. In the serial readout, K2/V4 improves on uniform 3 bits by 0.16 nats at equal residual-inclusive cost.

## 9.1 LIMITATIONS AND FUTURE WORK

Four directions would extend these findings. **(i) Broader evidence.** The measurements use fixed checkpoints, small book-derived document pools, and mainly teacher-forced NLL; more architectures, independent corpora, and downstream generation tasks would show how far the identified operating points carry. **(ii) Deployment.** Quantization is simulated in floating point, so packed kernels that store scale metadata should measure the memory and throughput consequences of the modeled payloads. **(iii) Allocation with guarantees.** Several Qwen comparisons select and evaluate on the same pool, as identified in the captions. Calibration pools disjoint from evaluation for every contrast, pre-specified equivalence margins, and search methods that model head interactions directly would turn the observed operating points into allocation rules that generalize across models and budgets. **(iv) Scale timing.** Controls that change when scales are computed without changing what they contain, together with mechanistic probes across attention types, would separate the effect of causality from scale geometry. Appendix A.11 discusses each direction in detail.

### Ethics Statement

This work evaluates publicly available language-model checkpoints using book-derived text and constructed evaluation tasks. It does not involve a human-subject study. Public corpora and pre-trained models can contain sensitive or biased content, so their use carries risks common to language-model research.

### Reproducibility Statement

The supplementary material provides code, configurations, launch scripts, analysis scripts, and recorded results, together with a README mapping tables and figures to relevant artifacts. Experimental settings are specified in the main text and the appendix.

### AI Use Statement

During the project, vibe coding tools (Codex and Claude Code) were used to help develop the code for the methods and experiments. All such work was conducted under human supervision and underwent human review. During manuscript preparation, AI tools were used to polish and revise the text and figures.

## References

Joshua Ainslie, James Lee-Thorp, Michiel de Jong, Yury Zemlyanskiy, Federico Lebrón, and Sumit Sanghai. GQA: Training generalized multi-query transformer models from multi-head checkpoints. *arXiv preprint arXiv:2305.13245*, 2023. URL https://arxiv.org/abs/2305.13245.

Saleh Ashkboos, Amirkeivan Mohtashami, Maximilian L. Croci, Bo Li, Pashmina Cameron, Martin Jaggi, Dan Alistarh, Torsten Hoefler, and James Hensman. QuaRot: Outlier-free 4-bit inference in rotated LLMs. *arXiv preprint arXiv:2404.00456*, 2024. URL https://arxiv.org/abs/2404.00456.

Zefan Cai, Yichi Zhang, Bofei Gao, Yuliang Liu, Yucheng Li, Tianyu Liu, Keming Lu, Wayne Xiong, Yue Dong, Junjie Hu, and Wen Xiao. PyramidKV: Dynamic KV cache compression based on pyramidal information funneling. *arXiv preprint arXiv:2406.02069*, 2024. URL https://arxiv.org/abs/2406.02069.

Tri Dao, Daniel Y. Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. FlashAttention: Fast and memory-efficient exact attention with IO-awareness. *arXiv preprint arXiv:2205.14135*, 2022. URL https://arxiv.org/abs/2205.14135.

Tim Dettmers, Mike Lewis, Younes Belkada, and Luke Zettlemoyer. LLM.int8(): 8-bit matrix multiplication for transformers at scale. *arXiv preprint arXiv:2208.07339*, 2022. URL https://arxiv.org/abs/2208.07339.

Zhen Dong, Zhewei Yao, Yaohui Cai, Daiyaan Arfeen, Amir Gholami, Michael W. Mahoney, and Kurt Keutzer. HAWQ-V2: Hessian Aware trace-Weighted quantization of neural networks. *arXiv preprint arXiv:1911.03852*, 2019a. URL https://arxiv.org/abs/1911.03852.

Zhen Dong, Zhewei Yao, Amir Gholami, Michael Mahoney, and Kurt Keutzer. HAWQ: Hessian AWare quantization of neural networks with mixed-precision. *arXiv preprint arXiv:1905.03696*, 2019b. URL https://arxiv.org/abs/1905.03696.

Haojie Duanmu, Zhihang Yuan, Xiuhong Li, Jiangfei Duan, Xingcheng Zhang, and Dahua Lin. SKVQ: Sliding-window key and value cache quantization for large language models. *arXiv preprint arXiv:2405.06219*, 2024. URL https://arxiv.org/abs/2405.06219.

Yuan Feng, Junlin Lv, Yukun Cao, Xike Xie, and S. Kevin Zhou. Ada-KV: Optimizing KV cache eviction by adaptive budget allocation for efficient LLM inference. *arXiv preprint arXiv:2407.11550*, 2024. URL https://arxiv.org/abs/2407.11550.

Elias Frantar, Saleh Ashkboos, Torsten Hoefler, and Dan Alistarh. GPTQ: Accurate post-training quantization for generative pre-trained transformers. *arXiv preprint arXiv:2210.17323*, 2022. URL https://arxiv.org/abs/2210.17323.

Yu Fu, Zefan Cai, Abedelkadir Asi, Wayne Xiong, Yue Dong, and Wen Xiao. Not all heads matter: A head-level KV cache compression method with integrated retrieval and reasoning. *arXiv preprint arXiv:2410.19258*, 2024. URL https://arxiv.org/abs/2410.19258.

Yefei He, Luoming Zhang, Weijia Wu, Jing Liu, Hong Zhou, and Bohan Zhuang. ZipCache: Accurate and efficient KV cache quantization with salient token identification. *arXiv preprint arXiv:2405.14256*, 2024. URL https://arxiv.org/abs/2405.14256.

Coleman Hooper, Sehoon Kim, Hiva Mohammadzadeh, Michael W. Mahoney, Yakun Sophia Shao, Kurt Keutzer, and Amir Gholami. KVQuant: Towards 10 million context length LLM inference with KV cache quantization. *arXiv preprint arXiv:2401.18079*, 2024. URL https://arxiv.org/abs/2401.18079.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7B. *arXiv preprint arXiv:2310.06825*, 2023. URL https://arxiv.org/abs/2310.06825.

Hao Kang, Qingru Zhang, Souvik Kundu, Geonhwa Jeong, Zaoxing Liu, Tushar Krishna, and Tuo Zhao. GEAR: An efficient KV cache compression recipe for near-lossless generative inference of LLM. *arXiv preprint arXiv:2403.05527*, 2024. URL https://arxiv.org/abs/2403.05527.

Sehoon Kim, Coleman Hooper, Amir Gholami, Zhen Dong, Xiuyu Li, Sheng Shen, Michael W. Mahoney, and Kurt Keutzer. SqueezeLLM: Dense-and-sparse quantization. *arXiv preprint arXiv:2306.07629*, 2023. URL https://arxiv.org/abs/2306.07629.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with PagedAttention. *arXiv preprint arXiv:2309.06180*, 2023. URL https://arxiv.org/abs/2309.06180.

Xing Li, Zeyu Xing, Yiming Li, Linping Qu, Hui-Ling Zhen, Wulong Liu, Yiwu Yao, Sinno Jialin Pan, and Mingxuan Yuan. KVTuner: Sensitivity-aware layer-wise mixed-precision KV cache quantization for efficient and nearly lossless LLM inference. *arXiv preprint arXiv:2502.04420*, 2025. URL https://arxiv.org/abs/2502.04420.

Yuhong Li, Yingbing Huang, Bowen Yang, Bharat Venkitesh, Acyr Locatelli, Hanchen Ye, Tianle Cai, Patrick Lewis, and Deming Chen. SnapKV: LLM knows what you are looking for before generation. *arXiv preprint arXiv:2404.14469*, 2024. URL https://arxiv.org/abs/2404.14469.

Ji Lin, Jiaming Tang, Haotian Tang, Shang Yang, Wei-Ming Chen, Wei-Chen Wang, Guangxuan Xiao, Xingyu Dang, Chuang Gan, and Song Han. AWQ: Activation-aware weight quantization for LLM compression and acceleration. *arXiv preprint arXiv:2306.00978*, 2023. URL https://arxiv.org/abs/2306.00978.

Tengxuan Liu, Shiyao Li, Jiayi Yang, Tianchen Zhao, Feng Zhou, Xiaohui Song, Guohao Dai, Shengen Yan, Huazhong Yang, and Yu Wang. PM-KVQ: Progressive mixed-precision KV cache quantization for long-CoT LLMs. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/afb8caec018d3c8f6ef8b81fa52386fe-Abstract-Conference.html.

Zechun Liu, Changsheng Zhao, Igor Fedorov, Bilge Soran, Dhruv Choudhary, Raghuraman Krishnamoorthi, Vikas Chandra, Yuandong Tian, and Tijmen Blankevoort. SpinQuant: LLM quantization with learned rotations. *arXiv preprint arXiv:2405.16406*, 2024a. URL https://arxiv.org/abs/2405.16406.

Zirui Liu, Jiayi Yuan, Hongye Jin, Shaochen Zhong, Zhaozhuo Xu, Vladimir Braverman, Beidi Chen, and Xia Hu. KIVI: A tuning-free asymmetric 2bit quantization for KV cache. *arXiv preprint arXiv:2402.02750*, 2024b. URL https://arxiv.org/abs/2402.02750.

Jack W. Rae, Anna Potapenko, Siddhant M. Jayakumar, and Timothy P. Lillicrap. Compressive transformers for long-range sequence modelling. *arXiv preprint arXiv:1911.05507*, 2019. URL https://arxiv.org/abs/1911.05507.

Utkarsh Saxena and Kaushik Roy. KVLinC: KV cache quantization with Hadamard rotation and linear correction. *OpenReview*, 2025. URL https://openreview.net/forum?id=FkaDML963W. Submitted to ICLR 2026.

Noam Shazeer. Fast transformer decoding: One write-head is all you need. *arXiv preprint arXiv:1911.02150*, 2019. URL https://arxiv.org/abs/1911.02150.

Jianlin Su, Yu Lu, Shengfeng Pan, Ahmed Murtadha, Bo Wen, and Yunfeng Liu. RoFormer: Enhanced transformer with rotary position embedding. *arXiv preprint arXiv:2104.09864*, 2021. URL https://arxiv.org/abs/2104.09864.

Guangxuan Xiao, Ji Lin, Mickael Seznec, Hao Wu, Julien Demouth, and Song Han. SmoothQuant: Accurate and efficient post-training quantization for large language models. *arXiv preprint arXiv:2211.10438*, 2022. URL https://arxiv.org/abs/2211.10438.

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient streaming language models with attention sinks. *arXiv preprint arXiv:2309.17453*, 2023. URL https://arxiv.org/abs/2309.17453.

Guangxuan Xiao, Jiaming Tang, Jingwei Zuo, Junxian Guo, Shang Yang, Haotian Tang, Yao Fu, and Song Han. DuoAttention: Efficient long-context LLM inference with retrieval and streaming heads. *arXiv preprint arXiv:2410.10819*, 2024. URL https://arxiv.org/abs/2410.10819.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025. URL https://arxiv.org/abs/2505.09388.

Amir Zandieh, Majid Daliri, and Insu Han. QJL: 1-bit quantized JL transform for KV cache quantization with zero overhead. *arXiv preprint arXiv:2406.03482*, 2024. URL https://arxiv.org/abs/2406.03482.

Zhenyu Zhang, Ying Sheng, Tianyi Zhou, Tianlong Chen, Lianmin Zheng, Ruisi Cai, Zhao Song, Yuandong Tian, Christopher Ré, Clark Barrett, Zhangyang Wang, and Beidi Chen. $H_2o$: Heavy-hitter oracle for efficient generative inference of large language models. *arXiv preprint arXiv:2306.14048*, 2023. URL https://arxiv.org/abs/2306.14048.

# APPENDIX

The detailed results below use the protocol attached to each cell. Unless explicitly marked PRE or serial, block tables use FUT (full-group) scales. Labels such as K-heavier or V-heavier refer to a loss ordering in that protocol, not intrinsic tensor sensitivity. Ratios above one denote superadditive joint damage. An interval including zero does not establish equivalence, and a same-source corpus transfer is not an independent-domain test. Payload is derived from the actual bit mapping of each allocation; the bit labels stored with some runs count one bit instead of two per unit for a joint K/V $4 \to 6$ upgrade and are not used.

This appendix preserves the measurement detail behind the main-text tables: cell provenance, per-configuration numbers, controls, and scope notes. Tables A.1–A.4 are the 4k anchor tables recorded at the time of each cell (n=4 point estimates or n=6/12 doc-bootstrap intervals as annotated); the main text draws on the identified per-document measurements and reports the relevant sample and protocol with each comparison. The full-precision comparisons with the official model implementation (Appendix A.9) are numerical diagnostics of the harness, not a validation of an official quantized execution.

**Notation and artifact names used in this appendix.** pg19 (or nat) and retri denote the natural-text and retrieval-proxy constructions (Appendix A.10). sym and mm denote the symmetric (absmax) and min-max codes. u$b$ (e.g., u3) is uniform-$b$, and k$a$v$b$ (e.g., $k2v4$, also written K@2/V@4) is K$a$/V$b$ of Section 2; K-2 and V-2 quantize only keys or only values to 2 bits. gs32 and gs128 are group sizes of 32 and 128 tokens; $T$ (or span) is the evaluated window length; front-$m$ and first-16 upgrade the first $m$ or 16 layers. $d(b)$, $d_b$, and $d^{(b)}$ denote the isolated head score $d_u(b)$ of Section 2, and d(4)-mm is that score under the min-max code. CE is the token-level cross-entropy, i.e., the NLL. The forward comparison (dC, parity) is the difference between the manual forward and a reference forward on the same input; 0.0 means identical. 4dp means four decimal places. n counts documents; CI denotes a 90% document-bootstrap interval; interval-separated (or resolved, separated) means that the interval excludes zero, and a tie means that it includes zero. seed-7, seed-8, and seed-9 name seeded document pools; cal and eval denote calibration and evaluation windows. A cell is one experimental configuration with its own result directory under `phq/runs/` in the supplementary material; cell names such as s34mm and s46mm (min-max selection batteries with base widths 3 and 4), static34 and static46 (static layer arms), s34mist (Mistral budget window), r10b, r11b, r13, and kvmm2b are those directory names (e.g., r11b is `r11b_base4`). The primary seed-7 store of a cell is its main result directory. Pinned result files are the result files listed in the table-to-artifact map of the supplementary material, from which the value generator derives every printed value; run annotations are summary values that the run scripts recorded at measurement time. A law is an empirical regularity of the anchor tables (e.g., the 2-bit floor or a K/V ordering); law cells and law runs are the cells that measure it. In appendix tables, — marks a quantity that was not measured or does not apply to that row.

## A.1 THE ADDITIVITY STRUCTURE (4K ANCHOR TABLES)

**Table A.1.** Additivity ratios (@4/@3/@2 = joint-to-univariate damage ratio at 4/3/2 bits) and the uniform-2 floor (nats), by code family and model (per-doc paired 90% bootstrap CIs as annotated per row below; negative = lower damage = better).

| model | code | @4 | @3 | @2 (joint) | @2 (x-tensor) | uniform-2 floor |
|---|---|---|---|---|---|---|
| Qwen3-1.7B | sym | 1.04 [0.91,1.18] | 1.35 [1.24,1.46] | 3.29 [3.06,3.51] | — | +2.99/+2.75 |
| Qwen3-1.7B | min-max | near 1.0 | 1.06 pg19 / 1.17 retri | 1.09 – 1.17 | — | +0.68/+0.66 |
| Qwen3-0.6B | sym | 1.15/1.06 | 1.50/1.66 | 2.89/2.63 | — | +4.27/+3.89 |
| Mistral-7B | sym | 1.27 [1.11,1.43] | 1.78 [1.68,1.89] | 3.00 [2.62,3.39] | 1.61 [1.40,1.84] | +3.48 [+3.21,+3.78] |
| Mistral-7B | min-max | 1.30 [0.94,1.72] | 1.70 [1.42,2.03] | 3.01 [2.71,3.34] | 1.10 [0.97,1.25] | +1.43 [+1.31,+1.55] |

Qwen3-1.7B symmetric row: n=6 docs/corpus, 8000-resample 90% doc-bootstrap, both corpora. Joint column: joint$\div\Sigma_u d_u$; cross-tensor column: $u_2/(K-2+V-2)$. The 1.7B symmetric joint@2 cell includes cross-tensor 1.31/1.21. (retri cells: 4-bit 0.98 [0.87,1.10]; 3-bit 1.44 [1.29,1.62]; 2-bit 3.09 [2.96,3.19]). Mistral-7B rows: n=12 per-doc, 20k paired-bootstrap 90% CIs, gs32 protocol, all 256 single-head marginals per code and width. Unmarked: n=4 point estimates (pg19/retri). Note: the min-max 3-bit ratio pair (1.06 pg19 / 1.17 retri), the Mistral rows, and the 0.6B ratio@2/@3/@4 cells are values recorded in the run annotations (not re-derived here from per-document files); every

other value loads from the pinned result files through the value generator. The symmetric 3-bit ratio (1.35) and the accompanying uniform-3 absolutes, +0.176 (symmetric, `r13_cis` cell) and +0.110 (min-max, K3/V3 row of the `phq_kvmm2b` cell), are likewise read from the pinned result files. Provenance of each run is in App. A.10.

Readouts. (i) Uniform-4 is within document-level uncertainty of the per-head sum at Qwen scale (ratio CIs include 1.0); at Mistral scale the 4-bit point estimates exceed one (1.27/1.30); the symmetric interval excludes one, whereas the min-max interval includes it. (ii) At 3-bit the code-family attribution is model-strength-dependent: min-max restores near-additivity on the Qwen3 family (pg19 ratio 1.06 vs symmetric 1.35, and strictly cheaper: uniform-3 $+0.110$ vs $+0.176$ nats) but *not* on the Mistral-7B cross-check, where both codes show amplification (1.78 [1.68,1.89] symmetric vs 1.70 [1.42,2.03] min-max, both CI-excluding 1.0) — a code-independent component of the $\leq$3-bit superadditivity exists at 7B scale that the symmetric-clipping hypothesis alone (Appendix A.9) does not account for. (iii) At 2-bit the per-head ratios are numerically close on Mistral (3.00 symmetric / 3.01 min-max) — the "2-bit superadditivity is symmetric-specific" law does not transfer — while the cross-tensor amplification remains min-max-reduced (1.61 [1.40,1.84] symmetric vs 1.10 [0.97,1.25] min-max, near-additive, matching the Qwen min-max 1.09/1.17); the absolute floor magnitudes stay code-separated at every scale (last column). (iv) The Qwen 0.6B cross-scale check reproduces the whole curve with a harder absolute 2-bit floor in nats (+4.27 vs +2.99) and mildly weaker amplification (2.63–2.89 vs 3.09–3.29). The symmetric-code counterpart of the Mistral cross-tensor ratio is the Qwen 1.7B 2-bit joint superlinearity (separable K-sum + V-sum $1.33 + 0.95 = 2.28$ stays below the joint $+2.99$, a further amplification 1.31/1.21). The additive@4 result and the hard 2-bit floor are measured with the symmetric code; the min-max 2-bit behavior above is measured on Qwen3-1.7B (gs32, n=4 seed-7, in-run symmetric anchors reproduce the per-tensor law run to 4dp) with the Mistral CI companion completing the picture. Measured allocation remains useful under the better code: under min-max the pooled d3-selection still gives measured top $<$ random $<$ bottom at matched bits (pg19 $+0.049$/$+0.084$/$+0.093$, both steps interval-separated; retri $+0.044$/$+0.063$/$+0.074$, the random-vs-bottom interval [-0.0216,+0.0007] including 0) and the mixed 3.29-bit point reaches $+0.049$/$+0.044$ (pg19/retri), $2.2\times$ below min-max uniform-3 on pg19 — measured per-head precision is complementary to code quality, not redundant with it. The ordering is not a one-scale artifact: the identical min-max $\{3,4\}$ battery on Qwen3-0.6B-Base reproduces measured $<$ random $<$ bottom on both corpora with all four gaps interval-separated (pg19 [-0.0888,-0.0653] and [-0.0484,-0.0240]; retri [-0.0530,-0.0329] and [-0.0405,-0.0152]), the mixed point sitting 2.0/2.0$\times$ below min-max uniform-3 (pg19/retri). The ladder continues through Qwen3-4B-Base (run in bfloat16; protocol and the RoPE-precision check in App. A.1): each ordinal step CI-separates on exactly one corpus (measured-vs-random on pg19 [-0.0281,-0.0177], its retri interval including 0 ([-0.0115,+0.0029]); random-vs-bottom on retri [-0.0130,-0.0022], its pg19 interval including 0 ([-0.0100,+0.0030] — the mirror of the 1.7B retri leg)), while the primary contrasts measured $<$ uniform-3 ([-0.0455,-0.0340] / [-0.0250,-0.0082]) and measured $<$ bottom ([-0.0310,-0.0217] / [-0.0182,-0.0061]) are separated on both corpora, and the mixed 3.22-bit point sits 1.8/1.4$\times$ below min-max uniform-3 (pg19/retri) — the absolute headroom narrows with scale (2.0/2.0$\times$ at 0.6B, 2.2$\times$ at 1.7B, 1.8/1.4$\times$ at 4B). An early scalar-per-key-block variant (one scale per 128-key block, no per-channel axis) was measured and abandoned (its damage at global 4-bit deployment was catastrophic relative to the 2D-group code at the same width); it is not part of the reported laws. Precision additivity has the opposite sign to the interference we saw in preliminary per-head attention-window restrictions that are not reported in this paper (windows: joint usually below separable; precision: above). The 3-bit symmetric-code superadditivity, the 2-bit super-linearity and the K$>$V direction are all beyond doc noise at n=6 on both corpora.

### A.1 CONTINUED: ADDITIONAL INTERACTION RESULTS AND IMPLEMENTATION DETAILS

• Qwen3-1.7B symmetric cells (block 4096, 2D-group gs32): n=4 per-corpus point estimates (seed-7), also measured at n=6 docs/corpus with 8000-resample 90% doc-bootstrap CIs on both corpora (the CI rows of Table A.1); every n=4 point falls inside the n=6 CI. The 2-bit joint superlinearity (separable K+V below the joint cell) is read off the same symmetric n=4 cell.

• Qwen3-1.7B min-max family checks (n=4 seed-7; in-run symmetric anchors reproduce the per-tensor law run to 4dp): 3-bit joint ratio 1.06 vs symmetric 1.35; min-max uniform-2 $+0.68$/$+0.66$; cross-tensor 2-bit ratio 1.09/1.17.

• Qwen3-1.7B min-max {3,4} selection battery (the "survives the better code" cell): same gs32/T=4096 protocol, n=6 seed-7 per corpus, forward comparison $0.0$; the native per-corpus $d(3)$-mm ranking selects the top/bottom 64 of 224 units; random-64 is a fixed seed-20260912 draw shared across corpora; the three ordering contrasts are 20k paired 90% doc-bootstrap CIs. The cell reproduces bitwise on an independent single-GPU re-execution (same driver, same seeds; every per-doc value on both corpora and every bootstrap triple identical; artifact `runs/s34mm_1p7b_local/`). Pool-seed robustness on disjoint pools: a held-out seed-8 replication (the seed-7 ids are excluded from both corpora before the seeded pick; the 2-of-6 within-seed cross-corpus id sharing mirrors the seed-7 texture; identical arms, CI stream and the same fixed random-64 draw; artifact `runs/s34mm_1p7b_seed8/`, with the seed-7 damage maps preserved in `runs/s34mm_1p7b_sweep7/`, itself bitwise-identical to the primary seed-7 store) preserves the separation status of all four primary ladder contrasts on both corpora ($8/8$; measured-vs-random [-0.0372,-0.0270] / [-0.0307,-0.0200], random-vs-bottom [-0.0185,-0.0022] / [-0.0061,+0.0021] – the retri interval includes 0 under both seeds – measured-vs-bottom [-0.0506,-0.0362] / [-0.0327,-0.0228], measured-vs-uniform-3 [-0.0544,-0.0377] / [-0.0450,-0.0334]), and $11/12$ of the twelve stored contrasts overall; the single flip is the secondary bottom-vs-uniform-3 pg19 tail, which becomes a tie ([-0.0124,+0.0039]) while its retri counterpart stays separated ([-0.0180,-0.0055]). The selections replicate beyond chance: top-64 overlap $42/40$ of 64 against the null mean $18.29$ ($5.7 \times 10^{-14}$/$6.2 \times 10^{-12}$), and the full 224-unit damage rankings at Spearman $0.557/0.506$, comparable to the within-pool cross-content correlation $0.529$; the bottom-64 sets overlap less ($30/27$ of 64): the ranking's flat centre is pool-noisy, its heavy tail (which the selection reads) is stable. The uniform-3 headroom ratio over the measured point replicates at $2.0/2.0$ (seed-7: $2.2/1.9$).

• Qwen3-0.6B min-max {3,4} selection battery (cross-scale counterpart of the previous cell): identical driver, protocol and seed choices (gs32, T=4096, n=6 seed-7 per corpus, the same pg19/retri doc pools and the same seed-20260912 random-64 draw, 20k paired 90% doc-bootstrap CIs, forward comparison $0.0$), measured with the same single-document (batch-1) fp32 runner that accumulates the loss in chunks; arm means measured $<$ random $<$ bottom pg19 $+0.109/+0.185/+0.221$, retri $+0.086/+0.130/+0.157$; uniform-3 $+0.217$/$+0.168$ (pg19/retri). Pool-seed robustness on the same disjoint pools (artifacts `runs/s34mm_06b_seed8/` with its `stability.json` and the sweep-preserving seed-7 re-execution `runs/s34mm_06b_sweep7/`, itself bitwise-identical to the primary seed-7 store on every per-doc value and bootstrap triple): the four primary ladder contrasts preserve their separation status on both corpora ($8/8$; measured-vs-random [-0.0783,-0.0520] / [-0.0494,-0.0295]), $11/12$ of the twelve stored contrasts overall. The single changed cell is again the narrow bottom-vs-uniform-3 pg19 tail, moving in the *opposite* direction to the 1.7B cell's flip: the seed-7 tie ([-0.0072,+0.0167]) resolves to a separation ([-0.0271,-0.0042]), while on retri it stays a tie under both seeds ([-0.0268,+0.0032] / [-0.0267,+0.0011]). Set-level stability is the strongest measured in this family: top-64 overlap $54/44$ of 64 (null $18.29$; $3.7 \times 10^{-31}$/$3.2 \times 10^{-16}$), bottom-64 overlap $43/39$ ($4.5 \times 10^{-15}$/$5.5 \times 10^{-11}$), full-map Spearman $0.864/0.666$. The uniform-3 headroom ratio replicates at $2.1/1.8$ (seed-7 $2.0/2.0$).

• Qwen3-4B min-max {3,4} selection battery (third rung of the same ladder): identical corpora, seeds, bootstrap and driver (gs32, T=4096, n=6 seed-7 per corpus, the same seed-20260912 random-64 draw, here over $288$ units at $36$ layers; all three allocators stored at the same 3.22-bit budget), 20k paired 90% doc-bootstrap CIs, forward comparison $0.0$ (re-passed exactly at the run's own precision). Protocol deviation: execution in bfloat16 (fp32 needs $\sim$16 GiB, more than the memory of the GPU used for this cell), with CE accumulated in 512-token chunks and attention queries in 256-token chunks (floating-point associativity only), and the causal mask built by an in-place triu (bit-identical to the fp32-built mask after the forward's own cast to bfloat16). RoPE precision: both reported 4B cells build the runner's RoPE angle table in fp32. A table built in the working dtype accumulates, under bfloat16, a phase error of $\approx 4$ rad by position 4096; the internal forward comparison cannot detect this because both sides share the table, whereas an external Hugging Face transformers CE cross-check on one held-out prefix does. With the fp32 table the cross-check agrees to $3 \times 10^{-3}$ nats / 97.6% argmax, and the fp32 lineage is byte-identical (verified at 0.6B). The reported 4B stores use the fp32 table; the working-dtype runs are kept as validation artifacts (App. A.10; transcript `runs/_validation/hf_check_4b.py`/`.log`). With the fp32 table, arm means reproduce the full ordering on both corpora, measured $<$ random $<$ bottom pg19 $+0.048/+0.071/+0.074$, retri $+0.040/+0.045/+0.052$, with the per-step CIs as stated in the ladder sentence above; uniform-3 $+0.088$/$+0.057$ (pg19/retri). Pool-seed robustness on the same

disjoint pools (bf16 runs with the fp32 RoPE table; artifacts `runs/s34mm_4b_seed8/` with its `stability.json` and `runs/s34mm_4b_sweep7/` – the latter bitwise-identical to the primary seed-7 store, so the bf16 execution itself reproduces bitwise on this stack): $10/12$ of the stored contrasts preserve their separation status and $7/8$ of the eight primary ladder cells (measured-vs-random [-0.0290,-0.0159] on pg19; the rest of the seed-7 ladder texture reproduces otherwise). Two contrasts change status: the family's first primary one moves in the tie-to-separation direction — the retri measured-vs-random interval, included-0 at seed-7 ([-0.0115,+0.0029]), *resolves* to a separation at seed-8 ([-0.0157,-0.0061]), with point estimates retaining their signs; and the narrow bottom-vs-uniform-3 pg19 tail flips a third time (separation at seed-7 [-0.0176,-0.0095], tie at seed-8 [-0.0073,+0.0019]; retri stays a tie [-0.0095,+0.0038]) — this tail flips at every s34mm rung. At the map level the 3-bit field is flat at this scale: Spearman $0.223/0.182$, top-64 overlap $25/20$ of 64 (null $14.22$; $3.8 \times 10^{-4}/0.039$), and the bottom-64 overlaps are not distinguishable from chance ($15/18$; $0.46/0.13$). The uniform-3 headroom ratio replicates at $1.6/1.7$ (seed-7 $1.8/1.4$).

• Mistral CI companion (both code arms, the same pg19 window, n=12 per-doc, all 256 single-head marginals per code and width, 20k paired-bootstrap 90% CIs; forward difference dC=0.0): covers the Mistral rows of Table A.1 (3/4-bit from the law doc cell; 2-bit from the 2-bit companion cell). The 2-bit cell's k2/v2 rows come from a targeted run that leaves the other tensor (V in the K-2 row, K in the V-2 row) in full precision (a variant that encoded it at bits=0 gives NaN/degenerate rows and is not used); these rows are byte-identical to the standard harness, and the uniform-2/single-head/full rows come from the same cell.

### A.2 The floor, the K/V reversal, and rotation (4k anchor tables)

**Table A.2.** Per-tensor 2-bit damage (nats) by code family and model: the sensitive side flips from keys (symmetric) to values (min-max).

| model | code | K-2 (nats) | V-2 (nats) | sensitive side |
|---|---|---|---|---|
| Qwen3-1.7B | sym | +1.33/+1.32 | +0.95/+0.96 | K>V (n=6: +1.372/+0.917 pg19, +1.315/+0.963 retri; diff CI excl. 0 / [+0.28,+0.42]) |
| Qwen3-0.6B | sym | +2.12/+2.01 | +1.79/+1.55 | K>V |
| Mistral-7B | sym | +1.79 [+1.65,+1.93] | +0.45 [+0.39,+0.52] | K>V by +1.34 [+1.21,+1.47] |
| Qwen3-1.7B | min-max | ≈+0.13 | ≈+0.50 | V>K (≈3.9×) |
| Mistral-7B | min-max | +0.05 [+0.05,+0.06] | +1.32 [+1.15,+1.50] | V>K by -1.27 [-1.45,-1.10] |

Mistral rows: the CI-backed 2-bit companion (n=12 per-doc, 20k paired bootstrap 90% CIs). Qwen rows: n=4 per-tensor points (pg19/retri; in-run symmetric anchors reproduce the per-tensor law run to 4dp). The min-max K@2 near-freedom ($+0.05$ at Mistral) is precisely the premise of KIVI-style 2-bit keys. Protocol boundary (Appendix A.9(c)-(d)): the min-max V>K direction is a property of the group-wide min-max scale rather than of intrinsic K/V sensitivity or of future reading. Under a per-position (prefix) scale the min-max V-2 damage collapses 93–98% ($+0.44 \to +0.03$ at 1.7B, $+1.08 \to +0.05$ at 0.6B, $+1.32 \to +0.02$ at Mistral-7B) and the block ordering flips to weak K>V parity; but a write-once *serial* protocol with group-at-completion min-max scales preserves the V>K direction at moderated magnitudes (Qwen3-1.7B n=12, same battery: mm V-2 $+0.47$ vs mm K-2 $+0.07$ at R=0, $+0.43$ vs $+0.04$ at R=128) and likewise preserves the symmetric K>V direction (K-2 $+2.64/+1.29$ vs V-2 $+0.41/+0.36$), which also persists under the prefix scales on every model ($+2.64 \to +1.01$, $+4.34 \to +1.73$, $+1.78 \to +0.75$). 2-bit V is cheap exactly when its scale is per-position (the write-time / per-token rule) and costly when one group-wide scale spreads a 32-token value range. The reversal is cross-scale and magnitude-correlated with the asymmetry: stronger at 0.6B (V>K by $5.8\times/4.8\times$ at 2 bits, $5.9\times/5.8\times$ at 3 bits) and strongest on Mistral (V>K by $\approx 26\times$ at 2 bits; min-max uniform-2 still $\approx 6.1\times$ worse than min-max uniform-3); only the absolute 2-bit penalty scales with capacity ($\approx 2\times$ higher nats at 0.6B). Everywhere the separable K-2+V-2 sum stays below the joint uniform-2 cell (Mistral, CI'd in Table A.1: separable $+2.24$ [+2.07,+2.41] $<$ the joint uniform-2 $+3.48$ [+3.21,+3.78]) — the cross-tensor 2-bit amplification is symmetric-code-strong (ratio $1.61$ [1.40,1.84]) and min-max-reduced (ratio $1.10$ [0.97,1.25]; Sec 4).

**Mistral forward-difference gate (measured).** The Mistral manual forward has a position-0-concentrated absolute parity gap of 0.0526 nats on the 12-doc window, but the operative gate is delta-consistency: the recorded delta-consistency of the reported paired readouts is 0.00 nats (machine precision) and the qchunk 256 vs 8 variants give identical deltas (0.2323 each). A direct official-forward cell (same 12 docs, 4k window, fp32; transformers 4.44.2 full forward) tests whether the

manual forward matches the official implementation: the per-doc document-average loss of the manual full forward matches the official transformers full forward at four-decimal precision, with per-doc differences $\leq 5{\cdot}10^{-5}$ (paired 90% CI [-0.0000,-0.0000] at four decimals), and the delta conventions measured against the manual versus the official full baseline have paired 90% gap CIs of +0.0000, +0.0000, +0.0000, and +0.0000 (four decimals) for the four decisive configs (u3/k2v3/front8/k2v4): the per-doc delta readouts are invariant to the manual-vs-official denominator, a measured property of this gate (not an assumed cancellation theorem). Appendix A.9(a) marks all block rows as manual-harness interventions; the official-forward cell validates the differential readout against the official baseline rather than re-running the quantized legs inside the transformers forward.

### A.2 CONTINUED: PER-TENSOR RESULTS AND ROTATION CONTROLS

• Per-tensor sequence: the symmetric per-tensor cell at 1.7B (K-2/V-2 rows); the min-max code-family checks at 1.7B and 0.6B (in-run symmetric anchors reproduce the cross-model table to 4dp) and the third-family Mistral check (in-run symmetric copy reproduces the uniform law battery to 4dp); the CI 2-bit rows (per-tensor 2-bit companion, n=12).

• The rotation battery (in-harness normalized Hadamard on pre-quant k,q; rotation-exactness parity 0.0; extended to span-half gate controls): the rotation-survival of the symmetric 2-bit floor, the rotation-robustness of both K/V orderings, the half-span config-gap stability (k2v4-vs-k3v3 $+1.36$/$+1.86$ vs $+1.61$ full-span).

• Sources of the ratios and prefix-scale values in the A.2 narrative: the Mistral V>K 2-bit ratio ($\approx$26$\times$) is read from the measured law-doc cell; the min-max uniform-2/uniform-3 ratio ($\approx$6.1$\times$) from the measured rotation-window cell; the 0.6B prefix-collapse floor is $+1.08$; and the prefix-scale persistence triple uses the serial battery's r=0 column on the two Qwen models ($+2.64 \to +1.01$ at 1.7B, $+4.34 \to +1.73$ at 0.6B) and the pooled FUT/PRE battery on Mistral-7B ($+1.78 \to +0.75$), for which no serial battery was run.

### A.3 THE ALLOCATION MATRIX (4K ANCHOR TABLES)

**Table A.3.** Equal-mean-bit asymmetric K/V splits ($\Delta$, nats; negative = lower damage): the equal-bit win is Mistral-specific (strong V>K).

| config | mean bits | Mistral gs32 | Mistral group size 128 |
|---|---|---|---|
| K@2/V@4 ($k2v4$) | 3.0 | -0.165 [-0.252,-0.083] | -0.135 [-0.179,-0.094] |
| K@4/V@2 (mirror) | 3.0 | +1.25 [+1.08,+1.42] | (vs $k2v4$: +1.79 [+1.66,+1.91]) |
| K@2/V@3 ($k2v3$) | 2.5 | +0.022 [-0.026,+0.069] (straddles 0) | +0.300 [+0.203,+0.412] |
| K@3/V@4 ($k3v4$; vs uniform-4) | 3.5 | +0.008 [+0.005,+0.011] | +0.014 [+0.011,+0.017] |
| fractional V@4, 2.625 | 2.625 | -0.034 [-0.086,+0.011] | — |
| fractional V@4, 2.75 | 2.75 | -0.053 [-0.113,-0.001] | — |
| fractional V@4, 2.875 | 2.875 | -0.147 [-0.236,-0.064] | — |

Qwen-side cells of the same configurations:

| config | Qwen3-1.7B | Qwen3-0.6B |
|---|---|---|
| K@2/V@4 ($k2v4$), pg19 | +0.035 [+0.027,+0.042] | +0.018 [+0.008,+0.028] |
| K@2/V@4 ($k2v4$), retri | +0.052 [+0.040,+0.062] | +0.055 [+0.037,+0.072] |
| K@2/V@3 ($k2v3$) | +0.11 [+0.10,+0.13] | $\approx$2$\times$ worse than u3 |
| fractional V@4 | (no sub-3-bit win) | (no sub-3-bit win) |

Mistral cells: n=12 per-doc, 20k paired 90% bootstrap CIs, same 12-doc pool (the gs128 rerun leaves the fractional-V win unaffected); in-run anchors reproduce the split's sign and scale (the mirror-axis V-drop is +1.32 at n=4 and +1.25 [+1.08,+1.42] paired). Qwen cells: n=6 per-doc (1.7B, both corpora) and the 0.6B deployment cells (parity gate 0.0; nat n=12 seed-9 books, retri n=6). Readout. (i) The equal-bit asymmetric split pays precisely where min-max V>K is large (Mistral $\approx$26$\times$ at 2 bits; Qwen3-0.6B 4.8–5.9$\times$, 1.7B $\approx$3.9$\times$): $k2v4$ beats uniform-3 interval-separated on Mistral at equal bits, and at group size 128; on Qwen it loses interval-separated. (ii) $k3v4$ is a near-lossless interior point at 12.5% fewer bytes. (iii) The $k2v3$ 2.5-bit near-parity is a finer-granularity effect: it straddles zero at gs32 (+0.022 [-0.026,+0.069]) but is interval-separated worse at gs128 (+0.300 [+0.203,+0.412]; coarser 128-token key scales penalize aggressive K@2

when V stays at 3) and does not exist on Qwen (+0.11 [+0.10,+0.13], CI excluding the straddle) — on Qwen, the absence of a cheap 2-bit-key window at equal bits survives the better code at both scales. (iv) Below 3 bits the fractional V@4 family (upgrade some heads' V to 4 on a stride pattern, rest K@2/V@3; selection-free) beats uniform-3 interval-separated at 2.75 and 2.875 bits, monotonically, with a robust direction (all five free 128-head subsets at 2.75 bits beat uniform-3 by 0.05–0.16 nats) but a head-choice-dependent magnitude (2.41–2.98×, interval-separated from the arbitrary stride-even baseline; early-layer identity matters most, about 2.98× the baseline) — the apparent superiority of uniform-3 below three bits is strong-v>k-scoped, and a measured per-head V-field can only add on this family.

**Allocator plateau and local-search evidence.** Every measured (bits/head, nats) point:

| point | bits | delta nats | PPL ratio |
|---|---|---|---|
| uniform-3 (gs128 / gs32) | 3.00 | +0.29 / +0.17 | 1.34× / 1.19× |
| base3+top64→4 | 3.29 | +0.12 / +0.10 / +0.11† | 1.13x |
| prefix-0.5 aging (gs32) | 3.00 | +2.6 | dominated |
| uniform-4 (gs128 / gs32) | 4.00 | +0.042 / +0.037 / +0.038 / +0.025† | 1.04x / 1.03x |

† values recorded in the run annotations (not re-derived from per-document files here; provenance in App. A.1). The uniform-3 granularity cells and the prefix-aging anchor load here from the pinned run stores (the fine/coarse-granularity additivity block and the prefix-aging battery, both pg19); PPL ratios are exp-delta arithmetic over the same deltas.

The base-3 + measured-top-64→4 point is strictly Pareto for bits $\in (3.0, 4.0)$: no uniform integer width or aging/position scheme can reach it, and it beats all 3.0-bit configs by 0.06..0.2 nats. But no per-head/position scheme reaches uniform-4 NLL within a 4-bit budget: uniform-4, with per-channel scaling, is already within a few $10^{-3}$ nats per head of the best tested per-head allocation — the allocator recovers a large fraction of the upgrade benefit at a fraction of the bit cost, yet PPL-parity bit-saving is not achievable with these codes. A saturation reading of this "4-bit plateau": at uniform-4 the per-head quantization errors fall below the residual-stream noise floor the next layer already injects (uniform-4 is $+0.026$ nats, PPL 1.03×), so no head selection can buy back damage that is already code-level; the interior structure reappears only once the upgrade width crosses the code's useful range again — at $4 \to 6$ the base-4 + top-64→6 rows regain 3–5× damage cuts vs uniform-4 (App. A.6), because uniform-4 is not free and a 6-bit head offers a measurable delta on top of it. This view predicts the $\{4\}$-regime plateau and the $\{4, 6\}$-regime structure simultaneously; it is hypothesis-level (the residual-stream floor is not directly measured). Two independent checks of this selected set follow, both at unchanged mean bits.

**(a) The basis (single-index curve) does not change the quality.** The allocator selects the 4-bit set by unconditional marginal damage $d_u^{(3)}$ (others full); we also measured the base-conditional residual benefit $g_u = \mathrm{NLL}(\text{all@3}) - \mathrm{NLL}(u\text{@4, others @3})$ and selected by $g$ instead. The two scoring rules select largely different sets (top-64 overlap only ∼0.34–0.42, per-head Spearman($d^{(3)}, g$) ≈-0.29..-0.11) yet the achieved joint quality is indistinguishable ($|g - d^{(3)}| \leq 0.0043$ nats both corpora): the 3.29-bit quality is a plateau across principled selection bases (consistent with the greedy-vs-marginal result). The diagnostic comparison: evaluating the assembled upgrade set (joint row delta from uniform-3) vs the separable prediction $\sum_u g_u$ gives by-$d^{(3)}$: $-0.102$ vs $-0.077$ (ratio 1.32, pg19; 0.81 retri), by-$g$: $-0.097$ vs $-0.232$ (ratio 0.42/0.36). Under low-bit superadditivity (Sec 4) the conditional benefit is a separable approximation of a non-separable objective: heads with the largest individual savings have strongly diminishing-return joint benefit (ratio ≈0.4) while the marginal-damage heads realize at least their modest prediction — so the unconditional marginal damage is the interaction-robust basis, and the worst-by-$g$ control ($+0.155 \approx$ uniform-3 $+0.155$) confirms the selection signal is real under both bases. A compact way to state this: on the

| 3.286-bit row | pg19 | retri |
|---|---|---|
| by-marginal $d^{(3)}$ (the allocator) | +0.075 | +0.061 |
| by-residual $g$ | +0.079 | +0.063 |
| random | +0.108 | +0.097 |
| worst-64 by $g$ (control) | +0.155 | +0.154 |

pg19 budget-development cell the separable-approximation error of the marginal basis is conservative (realized $\geq$ predicted, ratio 1.32), and greedy top-$k$ selection by it errs on the safe side under superadditivity there; the ratio is 0.81 on the retri window, so "conservative" is a claim about the cells where it was measured, not a general safety guarantee — the retri direction is reported since the under-claim bound is larger on the eval corpus. By contrast the conditional-benefit basis overpromises (realized $\ll$ predicted, ratio 0.42). The measured basis choice reduces to which direction the separable approximation may be wrong; the marginal ranking biases the allocator toward underestimating the benefit on the cells where its ratio $\geq 1$ was verified, which is conservative for the quality–memory trade-off in those cells.

**(b) Matched-bit equal-budget local search.** Starting from $S_0$ (top-64 by $d^{(3)}$), 11 equal-bit head swaps (9 singles + 2 doubles) from the $d^{(3)}$-vs-$g$ divergence (out: $g$-worst in $S_0$; in: $g$-best outside), each measured as a joint loss in the true mixed background (rest at 3-bit): against the $S_0$ calibration NLL, the best single improves calibration NLL by $-0.0024$, and the held-out (retri) point for that single conditionally-chosen set is $+0.0708$ vs $S_0$ $+0.0707$ — a $\sim$1e-3 calibration effect does not transfer, and no neighborhood-local-optimality or noise-level statement follows from one held-out point. With two independent calibration halves (pg19 split 3+3; a candidate must lower joint NLL on both, best on the first; held-out never participates in selection), one single swap survives consensus (subscripts give layer and KV head, $\ell_h$; out $0_5$, the $g$-worst in $S_0$; in $13_2$, a $g$-top outsider) scoring $-0.0012$ nats vs $S_0$ on held-out retri, and a second-order double swap (remove the two $g$-worst in $S_0$, $0_5$ and $5_3$; add the two $g$-top outsiders $13_2$, $11_4$) scores $\mathbf{-0.0022}$ nats at identical 3.286 bits — both survive calibration-half consensus and transfer (the single-batch version had picked a non-transferring candidate, $+0.0001$; the two single-swap consensus survivors reproduce their earlier per-half values to 4dp on the same docs, a measurement-reproducibility anchor). The evidence, all at unchanged 3.286 bits unless noted: The same protocol at base-4 + top-64→6 (4.5714 bits, $S_0$ the strict top-64 by $d^{(4)}$, in-job residual benefit $g_u^{(4)}$) finds 0/11 candidates negative on both halves (the second half was positive for every candidate), so $S_0$ stands there and no same-budget local improvement is found at 4.57 bits. Taken together: the measured 3.29-bit set is changed by equal-budget joint-loss local search only at the 3→4 point, by up to two head swaps worth $\approx -0.002$ nats (single held-out point per config, no CI; effect-size scale $\sim$1e-3), and is untouched at 4→6; these are protocol and magnitude statements, not optimality theorems. Side consistency throughout: transferred $S_0$ on retri $+0.0707$ vs native $+0.061$ is the $+0.010$ calibrated-transfer cost of Sec 6.

| local-search protocol | consensus | best $\Delta$ vs $S_0$ (cal half A / B) | held-out (retri) |
|---|---|---|---|
| single-batch, 11 swaps | 11/11 (cal only) | $-0.0024$ (best single) | $+0.0001$ (no transfer) |
| cross-doc, single swap $0_5 \to 13_2$ | 2/11 | $-0.0033$ / $-0.0003$ | $-0.0012$ |
| cross-doc, double swap ($0_5, 5_3 \to 13_2, 11_4$) | 5/11 | $-0.0057$ / $-0.0019$ | $-0.0022$ |
| cross-doc, single swaps @4.5714b (base-4→6) | 0/11 | — | $S_0$ stands (null) |

**Mixed-background validation (local-search control).** At 3.29 mean bits (base-3, upgrade 64 heads to 4), a randomized best-swap local search (RSLS; 24 rounds $\times$ 24 joint-evaluated single swaps, selected on pg19 d0-d5, held out on retri r0-r5, n=6 each) improves the joint loss by -0.0166 (calibration) / -0.0051 [-0.0082,-0.0024] (held-out) nats vs the one-shot unconditional-marginal set; the set-level fixed-point re-ranking reports $\sim$40-50% head churn without joint-damage consequence (per-set scores and trajectories are in the supplementary result files named below). At 4.57 mean bits (base-4, upgrade 64 to 6) the same search overfits: best set -0.0042 below one-shot on calibration but +0.0019 [+0.0005,+0.0031] worse on the held-out corpus. RSLS rounds/improving swaps: 13/9 and 11/7. Artifacts: `runs/phq_mixedref2/`.

### A.3 CONTINUED: ADDITIONAL ALLOCATION RESULTS

• Equal-bit split battery at Mistral (n=12 paired CIs; in-run anchors reproduce the split's sign and scale) and its gs128 KIVI-granularity rerun; the mirror-axis V-drop $k4v2$ (n=4 point + n=12 paired CI); the granularity scope cells ($k2v3$ at gs32 and gs128); the Qwen1.7B CI rows (n=6 both corpora) and the Qwen0.6B deployment rows (parity gate 0.0); the fractional-V@4 stride curve and the five-free-subset head-identity robustness cells (in-run stride-even anchor to 4dp).

• Measured-vs-static at $\{4, 6\}$ (static46; Qwen3-1.7B, n=6 per-doc, both corpora, equal bytes): reuses the measured per-doc arrays of the `r11b_base4` cell and freshly evaluates the two zero-calibration static layer arms on the same window with the base dict explicit (4-bit base over

all heads, layers upgraded to head width 6, exactly the labeled $4+2k/N$ bits). Static front-8-layers→6 (64 heads = 4.5714b) has higher damage than the measured top-64 set on both corpora (+0.0169 [+0.0161,+0.0178] / +0.0163 [+0.0145,+0.0181] vs +0.008/+0.005; measured-minus-static -0.0089 [-0.0108,-0.0071] on pg19, -0.0112 [-0.0128,-0.0095] on retri, negative = measured lower damage), and the measured top-128 set likewise beats static first-16l→6 at the wider 5.1429b interval-separated on both corpora (-0.0072 [-0.0094,-0.0052] pg19, -0.0095 [-0.0105,-0.0084] retri). Both static arms beat uniform-4 interval-separated (as does measured). The static arms here assign every unlisted head its base width, so all arms are at equal bytes. The directory `runs/static46_1p7b_unbalanced/` holds an earlier variant built as a sparse override-only dict, in which unlisted heads stay at full precision; it is not at equal bytes, shows near-zero static damage for that reason, and is not used. Artifact: `runs/static46_1p7b/static46.json` (bit labels follow the $4 + 2k/N$ convention; the `mean_bits` labels saved in `r11b_base4` add one bit per unit and are not used. Payload is derived from the actual $4 \to 6$ mapping; the recorded loss measurements are unchanged).

• Min-max $\{4, 6\}$ code-family scope (s46mm; Qwen3-1.7B, n=6 per-doc, both corpora, asym=True, same 4k window/doc_seed 7): full re-measurement of the $\{4, 6\}$ battery under the min-max code (fresh selection sweeps and arms; forward comparison 0.0). Measured top-64→6 still beats static front-8-layers→6 at equal 4.5714b (measured-minus-static -0.0087 [-0.0109,-0.0064] pg19 / -0.0054 [-0.0073,-0.0032] retri) and measured top-128→6 beats first-16-layers→6 at 5.1429b (-0.0103 [-0.0115,-0.0090] / -0.0052 [-0.0076,-0.0031]); both static arms beat uniform-4 interval-separated (as does measured). Arm means: uniform-4 0.0212/0.0170, measured top-64 0.0091/0.0075, static front-8 0.0178/0.0129, measured top-128 0.0029/0.0034, static first-16 0.0132/0.0086, uniform-6 0.0017/0.0013. Artifact: `runs/s46mm_1p7b/s46mm.json`. A re-execution with the single-document (batch-1) runner that accumulates the loss in chunks reproduces the original batched cell with every stored per-doc value and every bootstrap triple identical (underlying floats agreeing at fp rounding order, ce_selfcheck $1.2 \times 10^{-7}$; artifact `runs/s46mm_1p7b_local/`). Pool-seed robustness on the same disjoint pools as the s34mm cell (artifact `runs/s46mm_1p7b_seed8/`; the seed-7 damage maps are preserved in `runs/s46mm_1p7b_sweep7/`, itself bitwise-identical to both the batch-1 and the batched stores): the separation status of every stored contrast is preserved on both corpora ($14/14$; measured-vs-static [-0.0077,-0.0031] / [-0.0076,-0.0016] and [-0.0101,-0.0049] / [-0.0080,-0.0043], and the mixed point stays interval-separated above uniform-6 [+0.0053,+0.0077] / [+0.0040,+0.0084]). At the membership level the d(4)-mm damage field is much flatter per head than the d(3) one and the ranking churns accordingly: top-64 overlap $31/29$ of 64 against the null mean $18.29$ ($4.6 \times 10^{-5}$/$5.2 \times 10^{-4}$), top-128 overlap $80/76$ of 128 (null $73.14$; pg19 $0.042$ but retri $0.26$, not significant), and full-map Spearman $0.277/0.189$ (the seed-7 cross-content anchor is itself only $0.206$). The battery's claims therefore concern aggregated arm means – which replicate – rather than 4-bit selection membership; consistently with that picture, on the same documents the d(4) ranking tracks its larger-amplitude d(3) counterpart at Spearman $0.592/0.360$ (top-64 overlap $40/37$, $6.2\times10^{-12}$/$3.1\times10^{-9}$) – on pg19 more strongly than it tracks itself across pools, more weakly on the retrieval-proxy construction.

• Min-max $\{4, 6\}$ cross-scale counterpart at the small end (s46mm_06b; Qwen3-0.6B-Base, fp32): identical protocol family (asym=True, gs32, T=4096, n=6 seed-7 per corpus, the same pg19/retri pools, fresh d(4)-mm sweeps over $224$ units, equal 4.57b/5.14b measured-vs-static pairs, 20k paired 90% doc-bootstrap CIs). Strongest of the three rungs and reported separately: measured top-64→6 beats static front-8→6 interval-separated on both corpora ([-0.0307,-0.0195] / [-0.0109,-0.0025]), measured top-128→6 beats static first-16→6 on both ([-0.0242,-0.0171] / [-0.0147,-0.0085]), all four mixed-vs-uniform-4 contrasts are separated ([-0.0335,-0.0258] / [-0.0225,-0.0129] at 64; and +0.008/+0.005 vs +0.047/+0.035), and the mixed point is interval-separated *above* uniform-6 on both corpora ([+0.0136,+0.0168] / [+0.0140,+0.0191]) — the near-zero/negative retri damage of the 4B cell is therefore a 4B-specific phenomenon, not a ladder-wide one. The only loosened interval is static front-8 vs uniform-4 on pg19, which barely touches $0$ ([-0.0101,+0.0005]; retri [-0.0142,-0.0069] separated); static first-16 is separated on both ([-0.0231,-0.0136] / [-0.0250,-0.0128]). Arm means: static front-8 +0.042/+0.024, static first-16 +0.029/+0.016, measured top-64 +0.017/+0.018, uniform-6 +0.002/+0.001. Ladder-wide: 0.6B and 1.7B are separated on every paired contrast of the battery (including both static arms against uniform-4); at 4B every uniform-4 contrast separates on both corpora and only the two top-64-vs-static-front-8 intervals include $0$ (next cell). Execution-level selfchecks stored with the run: for-

ward comparison $0.0$; the chunked-CE evaluator equals the logits+CE reference path to $2.4 \times 10^{-7}$ (fp32 rounding order). Artifact: `runs/s46mm_06b/s46mm.json`. Pool-seed robustness on the same disjoint pools (artifacts `runs/s46mm_06b_seed8/` with its `stability.json` and `runs/s46mm_06b_sweep7/`, itself bitwise-identical to the primary seed-7 store): $13/14$ of the stored contrasts preserve their separation status — measured top-64 still beats static front-8 on both corpora ([-0.0249,-0.0123] / [-0.0229,-0.0141]) and the mixed point stays interval-separated above uniform-6 ([+0.0096,+0.0164] / [+0.0086,+0.0135]). The single changed cell is the pg19 static-front-8-vs-uniform-4 interval reported above as barely touching $0$: at seed-8 it *resolves* to a clean separation ([-0.0163,-0.0078]), while its retri counterpart stays separated ([-0.0113,-0.0008]). Membership level: top-64 overlap $43/31$ (null $18.29$; $4.5 \times 10^{-15}$/$4.6 \times 10^{-5}$), top-128 overlap $93/79$ (null $73.14$; pg19 $5.3 \times 10^{-8}$ but retri $0.072$, not significant), full-map Spearman $0.567/0.268$ (seed-7 cross-content anchor $0.428$). Cross-family at this scale the d(4)-mm ranking tracks the d(3)-mm one at $0.700/0.494$ with top-64 overlap $49/36$ ($7.9 \times 10^{-23}$/$2.0 \times 10^{-8}$) — its strongest reading in the family.

• Min-max $\{4, 6\}$ cross-scale counterpart at the large end (s46mm_4b; Qwen3-4B-Base, bfloat16; n=6 per-doc, both corpora, asym=True, same 4k window/doc seed 7, the same pg19/retri pools; $288$ units at 36 layers; fresh d(4)-mm selection sweeps and arms): the same family protocol with the protocol deviations of the 4B ladder (bfloat16 execution with the forward comparison re-passed exactly, $0.0$; CE in 512-token chunks and query attention in 256-token chunks; triu-built causal mask; and the chunked-CE evaluator reproduces the logits+CE reference path exactly on this model and precision — stored absolute difference $0.0$; like its $\{3,4\}$ counterpart this cell uses the fp32 RoPE table, see the RoPE-precision note in A.1). Outcome (a working-dtype-table run showed a retri-negative texture, an artifact of its phase error): measured top-64→6 and static front-8→6 tie within noise on both corpora ([-0.0052,+0.0011] / [-0.0059,+0.0008]); measured top-128→6 beats static first-16→6 interval-separated on both ([-0.0082,-0.0033] / [-0.0071,-0.0064]); against uniform-4, all eight contrasts (both measured budgets and both static arms) separate on both corpora ([-0.0071,-0.0018] / [-0.0091,-0.0029], [-0.0117,-0.0086] / [-0.0117,-0.0094], [-0.0041,-0.0008] / [-0.0048,-0.0018], [-0.0063,-0.0027] / [-0.0050,-0.0026]), and the mixed point is interval-separated *above* uniform-6 on both ([+0.0090,+0.0135] / [+0.0042,+0.0101]). Arm means: uniform-4 +0.018/+0.014, measured top-64 +0.014/+0.008, static front-8 +0.015/+0.011, measured top-128 +0.008/+0.003, static first-16 +0.014/+0.010, uniform-6 +0.002/+0.001; budgets 4.44b/4.89b. Artifact: `runs/s46mm_4b/s46mm.json`. Pool-seed robustness on the same disjoint pools (artifacts `runs/s46mm_4b_seed8/` with its `stability.json` and `runs/s46mm_4b_sweep7/`, itself bitwise-identical to the primary seed-7 store): $10/14$ of the stored contrasts preserve their separation status. Four contrasts change status: retri measured-64-vs-static-8 moves from a tie ([-0.0059,+0.0008]) to a separation ([-0.0052,-0.0021]) — the pg19 counterpart stays a tie ([-0.0052,+0.0002]), so the tie texture persists on pg19 while the second pool separates on retri, and on no point estimate, either seed, does static beat measured; retri measured-128-vs-static-16 moves from a narrow separation ([-0.0071,-0.0064]) to a tie ([-0.0032,+0.0022]), while on pg19 that separation strengthens ([-0.0092,-0.0073]); retri static-8-vs-uniform-4 ([-0.0048,-0.0018] → [-0.0013,+0.0009]) and pg19 static-16-vs-uniform-4 ([-0.0063,-0.0027] → [-0.0034,+0.0014]) move from separated to tied — the above-uniform-4 superiority of every arm therefore reads 6 of 8 cells at the disjoint pool (the retained cells include pg19 static-8 [-0.0036,-0.0001] and retri static-16 [-0.0086,-0.0010]), with every point estimate keeping sign, and the mixed point stays interval-separated above uniform-6 on both corpora ([+0.0091,+0.0116] / [+0.0061,+0.0096]) even as uniform-6's own damage collapses to $+0.0005/-0.0002$ nats on the second pool. At the map level the d(4)-mm field is near-chance at this scale: Spearman $0.159/0.074$ (seed-7 cross-content anchor $-0.024$), top-64 overlap $22/18$ of 64 (null $14.22$; $0.0079/0.13$), top-128 overlap $62/59$ of 128 (null $56.89$; $0.14/0.35$, both consistent with chance). Cross-family d(4)-vs-d(3) tracking drops to $0.223/0.180$ (top-64 overlap $24/22$; $1.1 \times 10^{-3}$/$0.0079$), completing the family-wide 0.6B $>$ 1.7B $>$ 4B gradient.

• Measured-vs-static at $\{3, 4\}$ (static34; Qwen3-1.7B, n=6 per-doc, both corpora, equal bytes): reuses the measured per-doc arrays of the `r10b_transfer` cell and freshly evaluates the same zero-calibration static layer arms on the equal-byte base-3 point (all heads at width 3, layers upgraded to head width 4). Static front-8-layers→4 (64 heads = 3.2857b) has higher damage than the measured top-64 set on both corpora (+0.1192 [+0.1153,+0.1230] / +0.0928 [+0.0881,+0.0987] vs +0.0757 [+0.0711,+0.0798]/+0.0624 [+0.0574,+0.0672]; measured-minus-

static -0.0436 [-0.0496,-0.0378] on pg19, -0.0304 [-0.0367,-0.0245] on retri), and the measured point also beats the wider static first-16-layers→4 (128 heads = 3.5714b) (-0.0137 [-0.0169,-0.0096] pg19, -0.0154 [-0.0226,-0.0085] retri), so the strictly-interior point survives the equal-byte positional comparator. The earlier sparse-dict variant (archived in `runs/static34_1p7b_unbalanced/`) is not used, for the same full-precision-leak reason as static46. Artifact: `runs/static34_1p7b/static34.json`.

Table A.1b consolidates the pool-seed robustness suite carried by the bullets above: every rung of both batteries was re-measured end-to-end on fully disjoint held-out pools, and the table below reports the complete set of comparisons.

**Table A.1b.** Pool-seed robustness consolidation for the equal-bit selection batteries (joint re-run on the held-out seed-8 pools; the seed-7 re-executions with preserved damage maps are bitwise-identical to the primary seed-7 stores in every battery). "preserved" = stored contrasts whose separation status matches the seed-7 store (same sign side); Spearman = rank correlation of the full 224/288-unit damage maps between the two pools; top-64 = overlap of the measured top-64 selections (hypergeometric null in parentheses); the flips count below names each status flip after the table (all point estimates keep sign; per-contrast CIs in the rung bullets).

| cell | preserved | Spearman | top-64 (null) | d(4)-vs-d(3) | flips |
|---|---|---|---|---|---|
| s34mm 0.6B | 11/12 | 0.864/0.666 | 54/44 (18.29) | — | 1 |
| s34mm 1.7B | 11/12 | 0.557/0.506 | 42/40 (18.29) | — | 1 |
| s34mm 4B | 10/12 | 0.223/0.182 | 25/20 (14.22) | — | 2 |
| s46mm 0.6B | 13/14 | 0.567/0.268 | 43/31 (18.29) | 0.700/0.494 | 1 |
| s46mm 1.7B | 14/14 | 0.277/0.189 | 31/29 (18.29) | 0.592/0.360 | none |
| s46mm 4B | 10/14 | 0.159/0.074 | 22/18 (14.22) | 0.223/0.180 | 4 |

Status changes by cell (each statistics column above gives the pg19/retri pair): s34mm 0.6B – bottom−uniform-3 pg19: tie→separated; s34mm 1.7B – bottom−uniform-3 pg19: separated→tie; s34mm 4B – measured−random retri: tie→separated; bottom−uniform-3 pg19: separated→tie; s46mm 0.6B – static8−uniform-4 pg19: borderline tie→separated; s46mm 4B – measured64−static8 retri: tie→separated; measured128−static16 retri, static8−uniform-4 retri, static16−uniform-4 pg19: separated→tie (uniform-4 reads 6/8).

### A.4 The deployment recipe and its scope (4K anchor tables)

**Table A.4.** Deployment scope: the front-V / layer-group recipe Δ vs full and vs uniform-3 [90% CI] across windows; the last row is the Qwen-family deployment gate (negative).

| stage (window) | config (bits) | Δ vs full | Δ vs u3 [CI] | note |
|---|---|---|---|---|
| budget development | front-8 (2.625) | +0.0795 | -0.169 [-0.279,-0.070] | (note 1) |
| budget development | first-16 (2.75) | +0.0727 | −0.176 [-0.284,-0.077] | (note 2) |
| budget development | measured layer-group top-16 (2.75) | +0.0835 | −0.165 [-0.279,-0.064] | (note 3) |
| out-of-pool transfer | front-8 (2.625) | +0.0703 | -0.242 [-0.360,-0.133] | (note 4) |
| 8192-token serving | front-8 (2.625) | +0.061 | -0.046 [-0.084,-0.014] | (note 5) |
| KIVI granularity (gs128) | front-8 (2.625) | — | -0.107 [-0.151,-0.067] | (note 6) |
| Qwen3-0.6B, pg19 | front-8 (2.643) | — | +0.148 [+0.124,+0.171] | (note 7) |
| Qwen3-0.6B, retri | front-8 (2.643) | — | +0.150 [+0.132,+0.170] separated; loses | (note 7) |

Row notes: (1) budget development (8-doc eval window): ≈96% of the 16-layer reduction; front beats same-bit scatter at every m; (2) budget development / first-16 (2.75): window argmin at 2.75 bits; (3) budget development / measured layer-group top-16 (2.75): trails first-16 by $+0.011$ [0.006,0.016] (interval-separated, no tie); ≈89% of the $k2v4$ benefit at 0.25 fewer bits; (4) out-of-pool transfer (16 new test books, n=16): recovers ≈77% of base damage on a harder window (u3 +0.3118); ≈97% 16-layer retention; (5) 8192-token serving (fresh out-of-pool subset, n=16): ordering survives serving length; margin compresses (u3 damage small); (6) KIVI granularity (same 12-doc pool, gs128): 0.375 fewer bits; trails the full 3.0-bit pipeline by only -0.026 [-0.030,-0.023]; beats same-bit scatter by +0.30 (deltas nominal; realized-cost note below); (7) Qwen3-0.6B pg19 / retri (n=6): uniform-3 dominates the whole family; this proxy does not support the Mistral allocation on Qwen.

Note on the KIVI row above: the quoted bit deltas are nominal — at identical nominal 3.0-bit payloads the u3/$k2v4$ granularity families realize 4.7164 / 6.8290 mean scored-cache bits (16-bit-residual / fp32-residual hypotheses) while the gs128 KIVI granularity realizes 4.5319 / 6.4173: the group-scale change alone shifts the realized accounting by 0.1846 / 0.4117 bits. The two granularity families are therefore not on a common realized bit axis; the KIVI arm is realized-cheaper than its nominal label, which is the conservative direction for every headline comparison (all quoted budgets remain nominal).

Details of the allocation procedure. (a) measured layer-group field: base K@2/V@3, measure the per-layer-group marginal $u_g = \mathrm{CE}(\mathrm{base}) - \mathrm{CE}(V_{\mathrm{layer}\,g} \to 4)$ over the 32 single-layer groups (8 KV heads each; 34 calibration forwards: 1 base + 32 groups + 1 assembled, on the 4-doc seed-7 window); the field is concentrated in early layers ($u_0 = +0.041$, $u_{2,3} = +0.027$, near zero past layer 8) and near-additive (assembled on-window gain $+0.121 \approx \sum_{g\in S} u_g = +0.127$, ratio 0.96). The selected static set $S = \{0, 2, 3, 4, 5, 6, 7, 8, 10, 11, 13, 15, 17, 20, 27, 31\}$ deploys on the disjoint eval window with an interval-separated out-of-sample win over uniform-3 at 2.75 bits (Table A.4, row 3: ≈89% of the full-$k2v4$ benefit at 0.25 fewer bits), trails first-16 by $+0.011$ [0.006,0.016] (interval-separated, so this contrast favors first-16; the two share the identical early-layer core) and dominates stride-even — per-(layer-)group measurement pays at sub-3-bit. An interaction-aware residual check confirms the one-shot marginal selection stands: Spearman(marginal, conditional) $= 0.57$ (L0 conditional contribution $+0.060$ vs marginal $+0.041$; then near-zero) and the best budget-preserving exchange (out 27, in 19) is worth $-0.0008$ [-0.0016,+0.0001] nats, CI straddling 0 — this tested swap has no resolved improvement (the evaluation table reproduces an earlier run of the same cell to 4dp). (b) fully calibration-free family: because the field concentrates in layers 0..8, the upgraded set is chosen purely from the budget — V of the first $m$ layers to 4, K@2 everywhere, $b(m) = 2.5 + m/64$, with the closed-form deployment map $m = \max(0, \min(32, \lfloor 32(2b - 5) \rfloor))$ (`phq_alloc_prefix.py`). front-8 at 2.625 bits is already an interval-separated sub-3-bit win over uniform-3 that captures ≈96% of the 16-layer reduction; the curve saturates past $m$=8 ($+0.0795 \to +0.0670 \to +0.0635$ as $m$:8 → 24 → 32 — how many layers is nearly irrelevant once $m \geq 8$) while their location matters: at equal bits front beats the same-budget scatter at every m (m=8: +0.0795 vs +0.1458; m=12: +0.0759 vs +0.1289; m=24: +0.0670 vs +0.0955). Scope: the budget-development rows (Table A.4, first three) are measured on the same 8-doc seed-7 eval window (no fresh holdout) — a deployment-quality curve with rows measured, not extrapolated. (c) KIVI-granularity experiment and the with-vs-without-allocator comparison (Table A.4, KIVI row): the equal-bit split positive and the V>K ordering reproduce at a gs=128 group scale in our own 2D-group code family (a granularity experiment, *not* a reproduction of KIVI's shipped per-channel-key/per-token-value+residual layout, which we measure separately in (g); same n=12 pool, forward difference dC=0.0, full-NLL identical to the gs32 window): matched-bit $k2v4$-vs-u3 -0.135 [-0.179,-0.094] (CIs overlap gs32 $-0.165$); the V>K asymmetry is stronger there ($k4v2$-vs-$k2v4$ +1.79 [+1.66,+1.91] vs +1.25 at gs32); the fractional-V allocator on the KIVI base beats uniform-3 at 0.375 fewer bits, trailing the full 3.0-bit pipeline by only -0.026 [-0.030,-0.023] and beating same-bit scatter by +0.30; and a K@2-specific gate control confirms the pos-0 causal-edge common mode does not drag the K@2 config deltas (front-8 +0.189 full vs +0.199/+0.178 on span halves; $k2v4$ +0.162/+0.164/+0.161; $k2v3$ +0.597/+0.655/+0.538).

(d) deployment transfers: the fixed front-8 survives the out-of-pool doc transfer (16 new PG-19 test books explicitly excluded from the selection pool, fresh seed, no calibration, no re-selection, n=16: -0.242 [-0.360,-0.133], recovering ≈77% of the base-2-bit damage ($+0.3515 \to +0.0703$) on a window where uniform-3 itself is harder (+0.3118 vs +0.2488); front-8 sits 0.013 [+0.012,+0.015] below the 3.0-bit $k2v4$ ceiling at 0.375 fewer bits) and the 8192-token serving-length check (fresh out-of-pool subset, prep at $T$=8192, n=16: -0.046 [-0.084,-0.014] with the front-placement-over-scatter preference in place and ≈95% of the 16-layer benefit retained). The 8k win margin compresses vs the 4k out-of-pool transfer because uniform-3's own damage is much smaller there (+0.107 vs +0.3118) while front-8's absolute damage is similar — an ordering statement (serving length), not a damage-magnitude context-invariance law. (e) measured family scope — the deployment gate: the same fixed front-8 does *not* transfer to Qwen3-0.6B (Table A.4, last row): the whole family sits between the $k2v3$ base and uniform-3 (the family recovers only 26% / 14% of the $k2v3$-to-uniform-3 span (pg19/retri), vs ≈77% of base damage on Mistral), even the full V-upgrade ceiling $k2v4$ loses (the equal-bit-split-loses law reproduces at 0.6B; Table A.3), and the front-vs-scatter placement preference inverts (-0.017 [-0.033,-0.001] pg19, -0.024 [-0.051,+0.002] straddling 0 on retri) vs the

interval-separated Mistral +0.048; and at the selection level the measured per-unit top-112 allocator (2.75 bits) is interval-separated worse than uniform-3 (+0.108 [+0.100,+0.115], at 0.25 fewer bits) and worse than random (+0.024 [+0.012,+0.036]) on the disjoint 8-doc window (App. A.7) — so the family loss is not confined to the static recipes — and the deployment gate is operational, not a free threshold: run the fixed front-V recipe against uniform-3 at matched bits on the calibration set and keep the recipe only if its interval-separated win reproduces there (Mistral: yes; Qwen3-0.6B: no, Table A.4 last row). The V>K sign alone is not the gate — Qwen3-0.6B also shows a per-tensor V>K reversal under min-max yet loses to uniform-3 — and no fixed asymmetry margin is claimed to separate the families without that direct matched-bit measurement; a cheap pre-computable discriminator that accepts Mistral and rejects Qwen before any NLL field is measured was not found in the statistics we ran, on both families (App. A.7: on Qwen3-0.6B the five offline statistics neither predict the measured ordering ($|\rho| \leq 0.078$ on the three key-side and composite statistics, -0.237 on the two value-side ones) nor separate the deployment outcome, and the measured allocator itself does not beat random at 2.75b), and is left as open work. This boundary is dataset-grounded (both directions measured at scale), not assumed; it is also consistent across metrics: the gold-hidden Qwen3-0.6B task cell (App. A.5) gives the ordering the NLL comparison suggests — full interval-separated best, uniform-3 ≈ k2v4, and front-8 the worst task config — so no task-vs-NLL divergence qualifies that comparison (the gold-visible protocol suggested such a divergence; App. A.5 explains the difference). (f) measured precision VS free block at the budget: the per-head marginal top-64 at the same 2.625 bits (258 calibration forwards vs 0) is interval-separated worse than the calibration-free front-8 (+0.1061 vs +0.0795; -0.027 [-0.032,-0.021]) with 258 calibration forwards rather than zero, read off the selection itself (only 28/64 heads in layers 0..7; rank-64 marginal ≈0.0000): the single-head ranking performs poorly at this budget, while the front-8 layer block is the best tested static structure at 2.625 bits (a standard single-batch marginal basis is gauged, not an interaction-aware head field). (g) KIVI-as-shipped layout head-to-head (a faithful in-harness reproduction of the published 2-bit-key / 4-bit-value layout — per-channel K (one min-max scale per head-dim over the whole quantized prefix), per-token V (one min-max scale per (token, head)), and the last $R$=128 tokens kept bf16 (KIVI's residual cache; the residual boundary is the block's cutting plane — the block-NLL analogue of KIVI's sliding most-recent-R window at the last inference step) — is run on the same in-pool 12-doc window under the same differential gate (dC=0.0), with the 2D-group min-max references as anchors (all reproduce: front-8 $+0.0814$, k2v4 $+0.0671$ vs full, interval-separated better than u3). The KIVI-layout cells land at $+1.71$/$+1.72$ nats vs full — interval-separated $\approx 1.49$ nats worse than uniform-3 (k2v3-style 2.92 bits: $-1.482$ [-1.567,-1.392]; k2v4-style 3.41 bits: $-1.486$ [-1.561,-1.405]) — and the V width barely matters (k2v3 vs k2v4 differ by $0.005$): the damage is concentrated in the per-channel 2-bit KEY scale, which saturates on the wide full-prefix key range (one (scale, offset) per channel over 3968 tokens vs 124 gs32 blocks). Two scoping caveats: (i) this is teacher-forced block NLL, where the entire 4k block contributes to each per-channel key scale — a 2-bit per-channel key code is adversarial here, and the streaming cell with windowed scales and a refreshed residual (Sec 7) shows the catastrophe is specific to the block cutting plane: online write-once $W$=512 refresh lands at $+0.027$ (overlapping front-8/k2v4); (ii) the single residual cutting plane under-protects earlier query positions relative to streaming. Implication for the anchor-table regularities: the near-free min-max K@2 result ($+0.05$ at Mistral, App. A.2) is a 2D-group-granularity result and does *not* transfer to the per-channel key layout under the block plane; the absence of a cheap 2-bit-key window at equal bits holds across key-scale geometries at the block plane (gs32 near-parity → gs128 worse → per-channel catastrophic) but does *not* extend to streaming decode (online windowed refresh keeps per-channel 2-bit K within the recipe's cost, Sec 7); a fixed-prefix causality cell further shows the block per-channel reading embeds 0.5453 nats of future-reading scale (Appendix A.9(b)); and the 2-bit cross-tensor behavior under the shipped layout is K-dominated — the superadditivity magnitude is code-granularity-dependent. Scope of (a)–(g): the claims are offline intervention NLL curves, doc-level inside-corpus transfers, and the measured model dependence — no kernel/serving or cross-corpus statements.

**2.75-bit granularity at the deployment budget.** The natural next question — at the same 2.75-bit budget (128 of 256 units' V upgraded 3→4 on the k2v3 base), does per-head V-selection by the measured per-head field beat the measured layer-group top-16? Reusing the already-measured per-head field (256 heads) and the measured layer-group selection, evaluated on the same disjoint 8-doc window: per-head top-128 +0.0932 vs layer-group top-16 +0.0835 is interval-separated worse (-0.0097 [-0.0145,-0.0045]); per-head top-128 also loses to the static first-16 layer block (-0.0205 [-0.0261,-0.0128]); and layer-group top-16 trails first-16 (-0.0108 [-0.0163,-0.0060]). The per-head

top-128 readout here (+0.0932) is a separate re-measurement that reuses the earlier per-head field; the measured-selection cell of Section 6.2 (Figure 3(b)) and App. A.7 records $+0.0952$ for the same budget and window, and both readouts place per-head selection behind layer-group top-16 and first-16. The 2.75-bit V@4 ordering is therefore first-16 +0.0727 $<$ layer-group +0.0835 $<$ per-head +0.0932, with block anchors front-8 +0.0795, layer-group +0.0835, and first-16 +0.0727; granularity ranked Spearman(per-head marginal, layer-group marginal) $= 0.434$. Interpretation: per-head marginal granularity is less effective than the tested static rule at this budget — the measured field settles granularity and code questions (= the instrument), while the zero-calibration front layer-block remains the best tested selector at 2.625–2.75 bits (the endpoint).

**Layer-width search (KVTuner-style reference).** A published-width-search shape gives the per-layer V width $w_l \in \{2, 3, 4\}$ on the k2 base the freedom to fund V@4 upgrades by V@2 downgrades, at the same 2.75-bit budget (total width $\sum_l w_l = 112$). V@4 layer marginals come from the measured layer-group field; V@2 downgrade marginals are measured freshly against the same base (32 calibration forwards). The exact 32-layer DP optimum under marginal independence uses 16 V@4 and 0 V@2 layers, and its assembled readout on the same disjoint 8-doc window is +0.0835. This DP optimum equals the layer-group recipe (+0.0835) to four decimals ([+0.0000,+0.0000] paired gap) and is interval-separated worse than the static first-16 block ([-0.0162,-0.0060] for static minus DP). The additive objective predicts an improvement of 0.1266 from the K2/V3 base; +0.0835 is instead absolute damage relative to full precision, so these two numbers are not directly comparable. The width freedom buys nothing distinguishable at this budget: the static comparison remains favorable at layer granularity too. Scope: single budget, single pool, min-max code.

**Budget window above 2.75 bits (s34mist cell).** The same question one budget step up, on the same 12-doc pool with the identical seed-7 cal-4/disjoint-eval-8 split, under both INT code families: fresh per-head $d_3$ damage sweeps on the calibration window select the top-64/128 units (upgrades $3\to4$ on both tensors, $+1$ bit/unit, so 3.25 and 3.5 mean bits on the 256-unit geometry), and the static front-8/first-16 blocks are the zero-calibration comparators at identical unit counts (parity $7.9\times10^{-5}$ CE vs official on the 256-token prefix; per-doc paired 90% CIs, 8-doc window; artifact `phq/runs/s34mist_7b/llm_s34mist.json`). Symmetric code: uniform-3 +0.1633; at 3.25 bits measured +0.0601 vs static +0.0625 (-0.0025 [-0.0060,+0.0011], unresolved); at 3.5 bits the measured top-128 +0.0384 interval-separated beats the static first-16 +0.0438 (-0.0054 [-0.0065,-0.0044]); uniform-4 +0.0206. Min-max code: uniform-3 +0.2488; at 3.25 bits measured +0.0297 vs static front-8 +0.0287 (+0.0010 [-0.0031,+0.0047], unresolved); at 3.5 bits measured +0.0239 vs static first-16 +0.0207 (+0.0031 [-0.0004,+0.0067], unresolved, static numerically lower); uniform-4 +0.0112. The measured selection is real on both codes — measured-minus-random [-0.0624,-0.0494]/ [-0.4042,-0.0984] and bottom-minus-measured +0.0730/ +0.2716 — but on Mistral under min-max the static front-8 recipe numerically *is* the measured optimum (differing by +0.0010 nats at 3.25 bits), consistent with the front-loaded V-side min-max field of App. A.4(a): the zero-calibration positional recipe captures the same structure the measurement finds. Cross-code: top-64 selection overlap 31 of 64 (16 by chance), Spearman($d_3^{\mathrm{sym}}, d_3^{\mathrm{mm}}$) = +0.2875 — the two codes' per-head fields weakly agree at this damage level, while the full-field rank comparison decorrelates (App. A.6), so the selection sets are not interchangeable even here. Interpretation: the measured top-$k$ advantage re-opens on Mistral at 3.5 bits under the symmetric code only; both min-max budgets stay unresolved ties against the positional block.

### A.4 CONTINUED: CONFIGURATION DETAILS

• Layer-group field (vfield): 34-calibration-forward $u_g$ field on the 4-doc seed-7 window, top-16 static set, out-of-sample eval on the disjoint 8-doc window.

• Interaction residual check: the conditional-field residual battery on the assembled background (Spearman 0.57; best exchange $-0.0008$, straddles 0).

• Calibration-free front-V curve: front-$m$ budget family on the same 8-doc eval window (zero calibration; endpoints reused from the identical-window run, gated on per-doc full-parity $\leq 5\cdot10^{-4}$).

• Out-of-pool doc transfer: the fixed front-8 on 16 new PG-19 test books (books excluded from every earlier document list, no re-selection, n=16).

• 8192-token serving check: fresh out-of-pool subset prep at $T{=}8192$, n=16; the full-precision comparison runs in smaller batches to fit memory, which does not change it (it is batch-invariant).

• 0.6B deployment scope: the same fixed recipe on Qwen3-0.6B (n=6 per corpus, forward comparison 0.0).

• Per-head-field check: the per-head marginal top-64 at the same 2.625 bits (258 forwards; eval reproduces the shared window to 4dp).

• KIVI-scale with-vs-without allocator (gs128, same 12-doc pool, n=12 paired CIs; in-run anchors).

• KIVI-as-shipped layout head-to-head (same in-pool 12-doc pool, R=128 residual, dC=0.0): per-channel-K (min-max) × per-token-V (min-max) layout vs the 2D-group min-max references (u3/k2v3/front8/k2v4 – anchors reproduce); byte accounting per config. The layout cells are ≈1.48–1.49 nats worse than u3 (interval-separated), damage concentrated in the per-channel 2-bit key scale; the read is granularity- not streaming-protocol-scoped.

• Streaming/decode AR cell (same in-pool 12-doc pool, P=1024/G=128/R=128/W=512, dC=0.0; Sec 7): cache-serial write-once quantization as cache tokens leave the last-R residual window, teacher-forced CE of the gold continuation + greedy divergence arm. The equal-bit k2v4-vs-u3 damage ratio widens under decode ($3.5\times \rightarrow 6.3\times$, point estimates; the absolute gap is unchanged, $+0.165$ [+0.084,+0.255] block vs $+0.161$ [+0.099,+0.235] stream); per-channel-K windowed refresh lands $+0.027$ (vs $+1.71$ block, catastrophe is block-cutting-plane-specific); FP8-E4M3 ≡ full (18% greedy divergence).

### A.5 Task cells with gold-hidden entries

**Task-level retrieval cells.** A single-needle retrieval cell (needle-in-a-haystack style) at 4096 tokens was evaluated for full / uniform-3 / front-8 / $k2v4$ with the deployment harness. Exact top-1 accuracy is ≈0 for every config including full precision (a metric-brittleness fact of the random-filler stress, not a recipe failure; the discriminative metric is the $-\log P$ residual). Task cells use one of two entry protocols. In the gold-visible protocol the harness scores the gold token at its own input position, so the gold token can enter the group-quantization scales; the early task-cell numbers below use this protocol and are listed as reference values. In the gold-hidden protocol the gold is fully excluded from the input, masks are trimmed, the last visible query predicts the held-out answer, pairing is same-seed, and each build is identity-hashed, so the model must retrieve the answer; the primary real-text cell and all cells below labeled gold-hidden use it.

Gold-hidden real-text 4k cell (Mistral-7B, 8-book pool × 24 builds × 5 depths; book-clustered paired 90% CIs, n=8 books): with a real pg19 haystack the true-retrieval signal is large for every config (control 11.876–12.235; answered-value $-\log P$: $k2v4$ 8.757 < front-8 8.866 < full 9.150 < u3 9.695; residuals front-8 -3.14 /$k2v4$ -3.12 deeper than full -3.09, u3 -2.39). u3 is interval-separated worse than full (+0.55 [+0.12,+0.98] — the measured retrieval cost of uniform-3 the recipe avoids); front-8 and $k2v4$ are interval-separated better than u3 (+0.83 [+0.35,+1.34] / +0.94 [+0.39,+1.58]); vs full they are point-estimate below full with CIs straddling zero (front-8 -0.28 [-0.70,+0.15]; $k2v4$ -0.39 [-0.99,+0.21]; config minus full, negative = lower NLL) and front-8 vs $k2v4$ tied. So the recipe-favorable direction holds under both protocols; the strongest CI-backed claims are the uniform-3 retrieval cost and both recipe configs beating uniform-3, while "below full" is a point estimate at the n=8-book clustering power; exact accuracy remains ≈0 on the real-text cells (metric brittleness there; the synthetic cell below is the exception). A denoising-by-coarse-V-quantization mechanism is a hypothesis, not measured. The Qwen-side companion cell is a built-in generator-artifact control: the spliced needle there harms full precision (+0.93 under the gold-visible protocol; gold-hidden cell below), so a shared generator artifact cannot explain the Mistral-side ordering.

Gold-hidden synthetic-4k cell (needle-in-a-haystack style with a random-filler haystack; n=24 seed-level paired 20k-bootstrap 90% CIs): with the gold held out, the synthetic single-needle copy task is cleanly quantization-sensitive — full answers best (8.06), uniform-3 second (8.73), and both recipe configs worst (front-8 9.71 = k2v4 9.71); every quantized config is interval-separated worse than full (u3 +0.67 [+0.42,+0.94]; front-8 +1.65 [+1.25,+2.06]; k2v4 +1.65 [+1.27,+2.03]), and the recipe configs are interval-separated worse than u3 (front-8 +0.97 [+0.55,+1.44]; k2v4 +0.98

[+0.62,+1.36]); front-8 vs k2v4 tied (-0.00 [-0.21,+0.21]); and in this gold-hidden cell exact accuracy mildly discriminates, the only task cell where it does (full 0.125 > u3 0.067 > front-8 0.042 > k2v4 0.025, peaks below 0.4). The gold-visible reading that front-8 matches full on absolute value (9.25 vs 9.27) does *not* hold under the gold-hidden protocol: on this redundancy-free random-filler stress the 2–3-bit V-heavy configs pay a real retrieval cost. The synthetic cell therefore documents task sensitivity under true retrieval rather than recipe near-parity; it does not disturb the corpus NLL deployment cells (Tables A.3–A.4).

Gold-hidden real-text-8k cell (the same real-text protocol at 8192 on the out-of-pool 8k documents; n=24 seed-level paired CIs): the retrieval signal survives serving length for every config (residuals full $-2.06$ / u3 $-1.52$ / front-8 $-1.32$ / k2v4 $-1.41$ vs controls $\approx$12.0), and the uniform-3 retrieval cost persists (u3 vs full $+0.50$ [+0.28,+0.71] interval-separated); but the recipe's $\geq$full property does *not* persist: front-8 and k2v4 are interval-separated worse than full on absolute answered-value (front-8 $+0.63$ [+0.36,+0.91]; k2v4 $+0.52$ [+0.29,+0.75]) and tie uniform-3 (gaps $+0.13$ [-0.12,+0.37] / $+0.02$ [-0.20,+0.26], straddling 0) — the recipe-favorable 8k ordering seen under the gold-visible protocol does not hold under the gold-hidden protocol. Net task readout: the recipe's retrieval property is 4k-real-text-scoped (point-estimate no-worse than full there, better than u3 interval-separated); the uniform-3 retrieval cost is CI'd at both 4k and 8k; the synthetic cell is a clean quantization-sensitivity datum; the NLL deployment cells (Tables A.3–A.4) are unaffected.

Gold-hidden Qwen3-0.6B real-text cell (same real-text protocol, n=24 seed-level paired CIs): the needle helps every config (residuals full $-1.38$ / u3 $-0.91$ / front-8 $-2.29$ / k2v4 $-1.62$ vs config-level no-needle controls; exact accuracy $\approx$0 everywhere), and the absolute answered-value ordering is full 11.83 < u3 12.68 < k2v4 12.78 < front-8 13.11 — every quantized config interval-separated worse than full (u3 $+0.84$ [+0.54,+1.15]; front-8 $+1.28$ [+0.94,+1.65]; k2v4 $+0.95$ [+0.65,+1.24]), uniform-3 $\approx$ k2v4 (straddle-0), and front-8 interval-separated worse than k2v4 ($+0.33$ [+0.03,+0.65]) and point-worse than u3. The gold-visible reading — the needle harms full, task residuals mildly favoring the V-heavy configs (a task-vs-NLL divergence) — does *not* hold: the gold-hidden Qwen cell is consistent with the NLL family gate (full best, uniform-3 the best quantized configuration, front-8 the worst task config), and the copy task remains the weak cell in absolute terms (band 11.8–13.1 vs Mistral's gold-hidden 8.8–9.7).

Gold-hidden variable-tracking cell (updating bindings plus distractors; same-seed pairing, n=24): remains a clean near-null under true gold-holding — residuals vs control are all $\approx$0 (full $+0.000$, u3 $-0.05$, front-8 $-0.02$, k2v4 $+0.01$; $\leq$0.05 in magnitude), every config-vs-config paired CI straddles zero, and full's answered value is bit-identical to its control (residual exactly $+0.000$: the VT answer is not made more likely by the task context at full precision). Exact accuracy $\approx$0 everywhere. Variable-tracking does not discriminate at this granularity under either protocol — a protocol near-null, reported for completeness, with no bearing on the corpus NLL deployment cells (gold-visible VT values: $|r| \leq 0.17$). The gold-hidden task battery comprises: synthetic-4k (clean quantization cost), real-text 4k (recipe $\geq$ full point-estimate, > u3 interval-separated), real-8k (u3 cost persists, recipe ties u3), Qwen3 (consistent with the NLL gate), and VT (near-null). These floors are properties of the symmetric INT code at 2-bit, not of the allocator.

### A.5 CONTINUED: TASK CONSTRUCTION DETAILS

- Gold-visible cells (reference values): synthetic single-needle 4k and 8k, the variable-tracking cell, and the real-text 4k/8k/Qwen cells (including the 24-seed 4k run with per-seed CIs).
- Gold-hidden entry (gold fully excluded, trimmed masks, last-visible query predicts the held-out answer, same-seed pairing, per-build identity hashing): real-text 4k (8-book pool $\times$ 24 builds, book-clustered paired CIs), synthetic-4k niah, real-text 8k, Qwen3-0.6B real-text, and variable-tracking, all with seed-paired CIs (n=24). Their readouts are given in the task cells above.

### A.6 CALIBRATION TRANSFER, GRANULARITY, PER-TENSOR, AND THE 16K EXPANSION

All cells in this section: Qwen3-1.7B, gs=32, block 4096, n=6 docs/corpus, per-doc NLL deltas with 2000-resample 90% bootstrap CIs — the same protocol as the point estimates in A.1–A.4.

**(a) A measured protect set transfers across corpora.** Per-head ranking differs across pg19 vs retri (top-16/32/64 set Jaccard 0.23/0.33/ 0.44, Spearman 0.45), yet a top-32 set measured on one corpus and deployed on the other scores $+0.104/+0.097$ nats (pg19/retri) vs native $+0.098/+0.085$, random $+0.157/+0.134$ and bottom $+0.178/+0.161$ at matched bits (3.143); deployed-transfer beats random at 6/6 docs on both corpora, with transfer-vs-native gap ($+0.006..+0.012$ mean) about 4.1–9.3x smaller than the top-vs-random margin. A set calibrated on one corpus is therefore usable on the other.

**(b) Document-bootstrap intervals (6 documents).** All adjacent pairs in the top $<$ transfer $<$ random $<$ bottom ordering are CIs-disjoint; the 4-point effects of Sec 6 are not doc-noise.

| config (mean bits) | pg19 $\Delta$ [90% CI] | retri $\Delta$ [90% CI] |
|---|---|---|
| uniform-3 (3.00) | +0.176 [0.165, 0.188] | +0.155 [0.145, 0.166] |
| base3+top32→4 (3.143) | +0.098 [0.092, 0.103] | +0.085 [0.078, 0.093] |
| transfer top32 (3.143) | +0.104 [0.098, 0.110] | +0.097 [0.092, 0.101] |
| rand32 (3.143) | +0.157 [0.150, 0.163] | +0.134 [0.127, 0.141] |
| bottom32 (3.143) | +0.178 [0.166, 0.189] | +0.161 [0.155, 0.168] |
| base3+top64→4 (3.286) | +0.076 [0.071, 0.080] | +0.062 [0.057, 0.067] |
| uniform-4 (4.00) | +0.026 [0.022, 0.030] | +0.026 [0.023, 0.029] |

**(c) Per-head granularity beats per-layer at matched bits.** top-2- layer$-$(3.07b) $+0.147/+0.131$ and top-4-layer$-$(3.14b) $+0.130/+0.112$ vs per-head top16 (3.07b) $+0.113/+0.118$ and top32 (3.14b) $+0.098/+0.085$ — head-level allocation is the efficient unit.

**(d) Per-tensor K-only upgrades do not dominate (negative result).** K-only base3+top-$k$($\rightarrow$4 on K, V stays 3) costs half the bits per head but leaves the V penalty everywhere: at matched mean bits, K+V rows win or draw (top32-K-only 3.07b +0.142/+0.114 vs K+V top16 3.07b +0.113/+0.118; top128-K-only 3.29b +0.120/+0.091 vs K+V top64 3.29b +0.076/+0.062). The K$>$V asymmetry is a (symmetric-code) 2-bit-floor phenomenon, not a 3$\rightarrow$4-margin saving.

**The sensitivity ranking itself is code-determined (calibration-field code swap, Mistral-7B).** On the same 4-doc seed-7 cal window the per-head V@4 marginal on the $k2v3$ deployment base was measured under the symmetric and the min-max INT code (256 heads each, differential forward difference dC=0.0). The two fields are essentially uncorrelated: Spearman(rank $d^{\mathrm{sym}}, d^{\mathrm{mm}}$) $= -0.04$, Pearson $0.05$, Kendall $-0.04$; the top-k head sets share a fraction $0.41/0.28/0.47$ of their entries at $k$=$32/64/128$ (shared-fraction overlap $|\cap|/k$, against the fixed-size chance levels $k/256 = 0.125/0.25/0.50$ of the 256-head field), and even the layer-aggregated orderings disagree (rank Spearman $-0.11$): symmetric concentrates the V-marginal in the early layers ($u^{\mathrm{sym}}_{L0} = +0.049$) while min-max spreads it flatter with a deep-layer component ($u^{\mathrm{mm}}_{L0} = +0.012$, layers 30/31 in the top 6 by layer mean). Interpretation: one calibration field is *not* reusable across code families – deployment recalibrates per code (complementing, not contradicting, the within-code cross-corpus transfer above). Within one code a layer-only ranking captures only a moderate share of the per-head order (head-vs-layer-mean rank Spearman $0.44/0.45$ sym/mm): layer calibration is a partial proxy, and the calibration-free front-block recipe (Section 6.2) remains the cheapest option (it needs no field at all). This is the strongest form of the code-family law – not a tensor-sign flip but a full re-ranking at head and layer level – and it is consistent with (though not itself a direct probe of) the correlated-clipping mechanism.

**(Third-family replication, SmolLM2-1.7B, a Llama-type model with multi-head attention (MHA); measured on a single GPU, fp32.)** A law-only replication of the same 12-document PG-19 token-stream pool protocol (span $S$=2048, gs32, both scale arms; the fixed Qwen-BPE stream reduced modulo the SmolLM2 vocabulary, touching 4.8% of positions identically for every arm, so all comparisons stay paired) reproduces both headline directions where the damage regime matches the Qwen law table: the symmetric K$>$V direction persists per-document (+0.363 at FUT, 12 of 12 docs doc-positive; +0.207 under the prefix scale, 12 of 12), and the min-max V$>$K direction reproduces (-0.165 at FUT, only 1 of 12 docs doc-positive; -0.061 under prefix, 3 of 12). The Qwen-family prefix inversion of the min-max gap (a weak K$>$V flip, Appendix A.9(c)) does *not* reproduce on this family: the gap only shrinks toward zero under the prefix scale. At the saturated 2-bit equal split every scale arm blows up on this degraded stream (symmetric absolute CE 10.95–24.20 nats, min-max 9.77–12.19, both envelopes over the two scale arms), so the sign readout was taken at the moderate

3.5-bit split ($k3v4$/$k4v3$), which matches the Qwen law-table damage regime. The uniform-width ladder on this family stays monotone ($u2>u3>u4$ mean damage in all four code × arm combinations; sym-FUT +7.227>+0.562> +0.059, 12 of 12 docs doc-ordered at each step), while the saturated 2-bit asymmetric pair reads $k2v4= +2.063 < k4v2= +4.820$ under sym-FUT – the opposite of the moderate-regime K>V readout – so the regime exclusion above is itself measured, not assumed. Readout: the code-family laws replicate on a third model family at matched damage regime through both code families, while the prefix inversion turns out to be family-graded rather than Qwen-specific: strongest on Qwen (+0.025, 12 of 12 doc-positive at the 2-bit table; still present at the moderate split, +0.002 with 10 of 12 on the CPU window), weakly present on Mistral-7B (+0.003, 8 of 12), and absent on SmolLM2 (the -0.061, 3 of 12 above). For reference, the Qwen family itself at the same pool, window, split, and scale arms (1.7B, GPU fp32) reads the symmetric pair at -0.010 at pool level under the full-group scale (3 of 12 docs doc-positive) while the write-time prefix scale keeps it weakly positive (+0.025, 11 of 12); the min-max pair keeps V>K solidly at FUT (-0.054, 0 of 12) and flips weakly positive under prefix (+0.002, 9 of 12) – the family-graded prefix inversion above. The SmolLM2 symmetric-arm strength is therefore family behaviour under a matched mild-damage split, not a property of the split choice; since the per-cell damages differ per family (sym FUT +0.086/+0.097 Qwen vs +0.457/+0.094 SmolLM2), the comparison reads the pair gaps, whose structures differ qualitatively, not the cell levels; the min-max reads line up where they overlap. All span-2048 cells quoted here are measured on a single device and pair only same-device numbers (full-precision CE is device-invariant at $1 \times 10^{-4}$ nats here; quantized rows carry a documented per-device offset, so doc-level signs within 0.017/0.031 nats of zero are read at pool level only). The 2-bit cells of the original batched battery at span 4096, re-measured on a single device at the same pool and window, recover gap means to 0.0037/0.0124 nats at FUT and 0.0198/0.0001 under prefix (the prefix-inversion magnitudes agreeing to 0.0001).

**(In-family 8B rung of the asymmetric-pair law, Qwen3-8B-Base, GQA $32\mathbf{q}/8\mathbf{kv}$, 36 layers; measured on CPU in fp32, same 12-document PG-19 pool/span and split as the reference above, paired against the 1.7B same-device CPU cells.)** At the matched moderate 3.5-bit split the 1.7B CPU reference reads the FUT symmetric pair at -0.015 at pool level (2 of 12 docs doc-positive) and the min-max V>K pair solidly negative (-0.052, 0 of 12), with the prefix side at +0.029/+0.002. The 8B rung at the same split and pool reads: symmetric +0.008 at FUT (7 of 12 docs doc-positive; +0.029 under prefix, 12 of 12), and min-max -0.025 at FUT (0 of 12; +0.004 under prefix, 9 of 12). The FUT symmetric pair, read at the declared pool level rather than zeroed by the doc-level band, does not stay put across the step: it is resolvably negative at 1.7B (-0.015, 2 of 12 docs doc-positive, i.e. 10 of 12 negative, one-sided sign test $p$=0.019) and flips to majority-positive at 8B (+0.008, 7 of 12), with the cross-rung flip itself resolved (paired doc-level difference +0.023, 10 of 12 docs in the flip direction, $t= +3.48$, two-sided sign test $p$=0.039), while the min-max V>K readout stays consistent across documents on both, its mean magnitude attenuating by a factor of 0.480 across the $4.8\times$ parameter step, while the symmetric prefix-side mean is essentially unchanged between rungs (+0.029 vs +0.029, both 12 of 12 docs doc-positive). The min-max prefix-side onset, by contrast, does not attenuate (pool mean +0.004 at 8B vs +0.002 at 1.7B; 9 vs 10 of 12 docs doc-positive; the raw ratio of the two pool means, 2.00×, is not resolvable at $n$=12 – paired 90% interval on the ratio [0.87,3.13]× from the same paired doc-level device, whose lower end excludes the FUT-side attenuation magnitude 0.480; the paired doc-level difference +0.002 is not resolved at $n$=12, 8 of 12 in the onset direction, two-sided sign test $p$=0.39, so we claim the absence of attenuation rather than a scale factor). Doc-level signs within the 0.017/0.031 nat device band of zero are read at pool level, as flagged above. Readout: the family's min-max asymmetry is sign-stable through a $4.8\times$ parameter increase at matched mild-damage split, with magnitude attenuation by scale, and its prefix-side onset replicates on the larger rung; the symmetric arm dependence (prefix > full-group) also carries to 8B; and the FUT symmetric cell undergoes a resolved sign reversal across the step rather than staying at zero.

**(Saturated 2-bit split on the 8B rung; same store and protocol, rows k2v4/k4v2, complete for all 48 row–document entries.)** At the saturated 2-bit asymmetric split the symmetric pair reads +1.124 at FUT (12 of 12 docs doc-positive) and the min-max pair reads -0.154 at FUT (0 of 12). Within the same rung the symmetric FUT cell amplifies from +0.008 at 3.5-bit (7 of 12) to +1.124 at the saturated split (all 12 docs), while the min-max FUT ladder continues and deepens, from -0.025 (0 of 12) to -0.154 (0 of 12): the symmetric K-over-V direction on this family is budget-gated, not scale-gated, and the min-max V>K direction stays consistent across documents at the stronger split. The 1.7B same-protocol GPU cells are sign-consistent on both arms (+0.202, 11 of 12 at FUT;

-0.373, 0 of 12); the cross-rung comparison is sign-level only (CPU fp32 at 8B vs GPU at 1.7B), while every 2-bit-vs-3.5-bit comparison above is same-device. Prefix-side rows of the 2-bit grid (complete for all 48 entries) replicate the qualitative PRE direction on both arms: pooled relative PRE +1.147 (sym, 12 of 12) and +0.013 (min–max, 10 of 12) are both positive, and the 1.7B GPU cells are sign-consistent (+0.602, 12 of 12; +0.025, 12 of 12): the prefix side carries no sign change across the parameter step at the saturated budget.

**(e) Calibration-count sensitivity of the field, and a cost bracket (differential forward difference dC=0.0).** On the same $k2v3/k2v4$ 2.75-bit V-side budget, a fresh $n$=16 pg19 window (seed-7/T-4096, excluding every doc in the Qwen evidence footprint, $n_{\text{cal}}$=8/$n_{\text{eval}}$=8) gives the per-doc per-unit V@4 marginals and re-measures selected sets built from the first-$\ell$ calibration docs ($\ell \in \{1, 2, 3, 4, 6, 8\}$) on the disjoint eval window. Three reads. (i) The field ordering is calibration-count-sensitive: per-doc marginals are shallow (per-doc sd 0.0061 vs mean-|marginal| 0.0055), rank Spearman vs the full-cal field grows 0.32 (1 doc) $\rightarrow$ 0.65 (4) $\rightarrow$ 1.00 (8), top-half shared fraction ($|\cap|/112$ of the 112-unit selected set) is 0.76 at $\ell$=4 and reaches 1.00 only at $\ell$=8 (24% membership churn between the cal-4 and full-cal selections; the cal-count ordering metrics are computed from the per-doc marginal arrays stored with the `llm_qwen17calsweep_r2` cell), so the head-level field calibrated on 4 docs is not a stable ordering. (ii) Yet held-out allocation usefulness is flat in calibration size: the measured top-112 damage is +0.167/+0.171/ +0.163/+0.166 (mean nats) at $\ell$=1/2/4/8 with per-doc paired 90% CIs vs the cal-size-independent random selection straddling zero at 4/6 sizes (marginally below at $\ell$=4 $[-0.012, -0.002]$ and $\ell$=8 $[-0.0079, -0.0002]$) – no monotone improvement with more calibration. Within-run equal-bit anchors reproduce the established weak-V>K family readout (uniform-3 +0.094 < k2v4 +0.132; measured +0.166 $\approx$ random +0.170). (iii) Cost bracket: the entire cell – cal-8 field, 8-doc held-out eval, and the six calibration-size selections – took 430.7 s ($\approx$7.2 GPU-min) on one GPU; the field-marginal share is bounded below that, so the $\approx$4 GPU-minute per-field cost stated in App. A.10 holds as a bracket. Interpretation: on the weak-V>K family the per-head field is a noisy instrument that stabilizes only near full calibration, but this gates no practical recommendation – the offline allocation recipe (Section 6.2) is calibration-free. Where the per-head field is primary – the code-family swap readouts above – the calibration-count caveat is acknowledged as a field-reliability boundary, not a cost one.

**(f) Law directions are length-stable at 16384 ($n$=18 long books).** Do the per-tensor laws (Table A.2) survive 16$\times$ the calibration length? The same 2-bit/3-bit/4-bit rows and the per-tensor $k2v4/k4v2$ split were scored per-doc under both INT codes with the same 2D-group gs32 protocol as the 4k law tables (forward comparison against the reference harness `pkvr_measure.py`: 0.0 per doc) on a long-book pg19 window ($T$=16384): an $n$=4 seed-7 sample extended to $n$=6 and set beside a fresh disjoint seed-9 sample of $n$=12 books (all excluding the Qwen evidence footprint), pooled $n$=18; the same 18 books were then re-scored at $T$=4096 so the length comparison is within-doc and paired. Reads. *(i) The observed orderings persist at 16k and are doc-unanimous*: the symmetric 2-bit floor ($u2$ +4.375 vs $u3$ +0.298, $\approx$15$\times$; $u4$ +0.034) and the min-max graceful degradation ($u2$ +0.614 vs sym +4.375, $\approx$7$\times$ cheaper; $u4$ +0.018); the code-family K/V reversal in *every* book (symmetric $k2v4$ +3.019 $\gg$ $k4v2$ +0.866 = keys more sensitive, paired gap +2.153 [+2.071,+2.236]; min-max $k4v2$ +0.377 $\gg$ $k2v4$ +0.150 = values more sensitive, gap +0.226 [+0.203,+0.249]); and the equal-bit readout that no 2-bit per-tensor split beats uniform-3 at 3.0 mean bits on either code (symmetric $k2v4$ vs $u3$ +2.722 [+2.641,+2.805], $k4v2$ vs $u3$ +0.569 [+0.530,+0.607]; min-max $k4v2$ vs $u3$ +0.293 [+0.271,+0.314], $k2v4$ vs $u3$ +0.066 [+0.062,+0.070]) — the "no cheap 2-bit window" readout is length-stable. *(ii) Not a small-sample artifact*: the two disjoint pools agree within $\approx$0.12 nats on every emitted row pair (worst case sym $k2v4$: $|3.098 - 2.980|$; sym $u2$ +4.390 vs +4.368, $k2v4$ +3.098 vs +2.980; mm $k4v2$ +0.412 vs +0.359) and the pooled paired CIs are about 2.6$\times$ narrower than the $n$=4 readout (e.g. the symmetric K/V-gap CI narrows from [+2.03,+2.46] at $n$=4 to [+2.071,+2.236] at $n$=18). *(iii) Length-scaling is within-doc paired and asymmetric.* At 4$\times$ context on the same books, symmetric damage grows sublinearly per token ($u2$ +3.022 $\rightarrow$ +4.375, paired excess +1.353 [+1.284,+1.423]; $u3$ +0.166 $\rightarrow$ +0.298, +0.131 [+0.119,+0.143]; $k2v4$ +1.353 $\rightarrow$ +3.019, +1.666 [+1.606,+1.729]) and the growth concentrates in the symmetric 2-bit floor: $u4$ is flat (+0.009), the values-heavy split is length-flat ($k4v2$ +0.879 $\rightarrow$ +0.866, −0.013 [-0.040,+0.011]), and the min-max rows are length-stable or marginally cheaper ($u3$ −0.010 [-0.014,-0.007], $k4v2$ −0.059 [-0.077,-0.043]). So 16384 is a length extension — the symmetric K-head damage, not the general quantization error, is what context amplifies, and the K>V asymmetry intensifies with length. *(iv) The length-stable readout reproduces on the second scale (Qwen3-0.6B, identical head geometry).* The same 24 books re-scored

on 0.6B at both lengths give the same observed orderings, doc-unanimous again (sym K>V 18/18 and 6/6, mm V>K 18/18 and 6/6; equal-bit split worse than uniform-3 at 3.0 bits 18/18 and 6/6 on both codes), pooled readings of the same character (sym $u2$ $+5.337$ vs $u3$ $+0.487$; retri $u2$ $+5.208$; sym K/V gap $+2.025$ [+1.923,+2.125]; mm V/K gap $+0.742$ [+0.674,+0.810]), and the same within-doc $4\times$-context growth shape: symmetric 2-bit-K damage grows ($u2$ $+4.243$ $\rightarrow$ $+5.337$, paired $+1.094$ [+0.999,+1.189]; $k2v4$ $+2.141$ $\rightarrow$ $+3.627$, $+1.486$ [+1.372,+1.602]) while the values-heavy split is length-flat ($k4v2$ $-0.070$ [-0.117,-0.025]) and the min-max rows are length-stable or cheaper. The larger 0.6B magnitudes sit on top of its larger 4k floors (uniform-2 $+4.27$ vs 1.7B's $+2.99$, Table A.1). Scope: two scales (Qwen3-1.7B and 0.6B), gs32, block 16 384; the block tables stay 4k anchors and this cell extends their claim to 16k directionally.

**(g) Several ordinal comparisons persist on the retrieval proxy at 16384.** Re-score the same five rows $\times$ both codes on the retrieval-proxy corpus (long anchor$\rightarrow$continuation pairs, 4k/6k distractor gaps; $n$=6, built deterministically from the same PG-19 token pool with the evidence footprint excluded) at $T$=16384 — a book-derived proxy construction rather than a validated retrieval task. Every law direction persists: the symmetric 2-bit floor ($u2$ $+3.891$ vs $u3$ $+0.358$) with the symmetric K-heavier side ($k2v4$ $+2.842$ $\gg$ $k4v2$ $+1.030$, paired gap $+1.812$ [+1.773,+1.858]), the min-max V-heavier swap ($k4v2$ $+0.436$ $\gg$ $k2v4$ $+0.166$, gap $+0.270$ [+0.239,+0.300]), and "no cheap 2-bit window" under both codes ($k4v2$ vs $u3$ +0.672 [+0.640,+0.701] symmetric, +0.353 [+0.326,+0.379] min-max). The within-doc $4\times$-context growth pattern reproduces on retri ($u2$ $+2.845$ $\rightarrow$ $+3.891$, paired excess $+1.047$ [+0.996,+1.097]; $k2v4$ $+1.414$ $\rightarrow$ $+2.842$, $+1.428$ [+1.373,+1.482]), with one distributional difference: even the values-heavy symmetric split rises mildly on retri ($k4v2$ $+0.046$ [+0.018,+0.071], flat on natural text). Pooled retri magnitudes shift vs natural text ($u3$ higher, $+0.358$ vs $+0.298$; $u2$ lower, $+3.891$ vs $+4.375$) — absolute damage levels are corpus-dependent, while the listed ordinal comparisons persist within these tested cells. Scope: retri is a deterministic retrieval-proxy construction over PG-19 tokens (not a new download), one model; the 4k anchor tables remain unchanged.

**The 4–6-bit domain, payload accounting, and the FP8 comparator. Measured allocation improves quality** (Qwen3-1.7B, gs=32, n=6 docs; same protocol and per-doc 90% bootstrap CIs.)

| config (mean bits) | pg19 d [90% CI] | retri d [90% CI] |
|---|---|---|
| uniform-4 (4.00) | +0.026 [0.022, 0.030] | +0.026 [0.023, 0.029] |
| base4+top32$\rightarrow$6 (4.2857) | +0.013 | +0.010 |
| base4+top64$\rightarrow$6 (4.57) | +0.008 [0.006, 0.010] | +0.005 [0.004, 0.006] |
| base4+top128$\rightarrow$6 (5.14) | +0.002 [0.001, 0.003] | $-0.001$ [-0.002, 0.000] |
| uniform-6 (6.00) | +0.002 [0.001, 0.003] | +0.003 [0.002, 0.003] |
| base4+rand32$\rightarrow$6 (4.2857) | +0.023 | +0.020 |
| base4+bottom32$\rightarrow$6 (4.29) | +0.029 | +0.022 |

Selection ordering (top $<$ random $<$ bottom) and cross-corpus transfer (reproduce at base-4: transfer-top64 $+0.013$/$+0.013$ vs native $+0.008$/$+0.005$, still $\ll$ random) survive in the $\{4, 6\}$ domain. (Rows base4+top32$\rightarrow$6, rand and bottom at 4.2857 bits are batch point estimates; per-doc CIs were computed for the decisive cells only.) The decisive cells: base-4 + top-64$\rightarrow$6 (4.57 bits) has lower damage than uniform-4 on both corpora — a 3.3–5.1$\times$ damage reduction with $+14.3\%$ more payload bits than uniform-4 — and base-4 + top-128$\rightarrow$6 (5.14 bits) has small absolute damage at $14.3\%$ fewer payload bits than uniform-6. The one slightly-negative retri cell ($-0.001$, CI $[-0.002, 0.000]$) reads as doc noise at the uniform-6 level rather than as quantization-induced regularization; both high-precision cells have small residual errors and their sign flips within a $10^{-3}$ nats band. This is a measured interior allocation at a higher budget, not evidence that uniform-4 or the mixed allocation is globally optimal. Two negative results: (a) the K-only base-4$\rightarrow$6 matched-bit test is measured and negative — at the same mean bits K-only loses to K+V on both corpora (K-only top-128 @4.5714 bits $+0.014$/$+0.016$ vs K+V top-64 @4.5714 bits $+0.008$/$+0.005$; K-only top-64 @4.2857 bits $+0.015$/$+0.018$ vs K+V top-32 $+0.013$/$+0.010$), extending the per-tensor negative from 2-bit and 3$\rightarrow$4 to the $\{4, 6\}$ domain; the s46mm experiment reports per-document min-max top-64/top-128 and static comparisons on both constructions (Appendix A.3), extending these batch-only controls; (b) a doc-ensemble protect set (rank by per-doc mean curve, or union of per-doc top-64 sets over the 6 calibration docs) transfers worse to the held-out corpus than the single-

corpus batch selection (doc-mean-top64 $+0.020$ vs batch-transfer $+0.013$ on retri; union-115-head $+0.009$ at 5.03 bits) — batch calibration carries the doc-average better than ensemble constructions on this scale.

*(cost fact) Payload accounting for the $\{4, 6\}$ cells.* Mean payload is $4 + 2k/224$ bits/unit when $k$ of the 224 KV units are upgraded k-and-v from 4 to 6 bits: $k = 32/64/128$ give **4.29 / 4.57 / 5.14** bits/unit (precisely $4.2857/4.5714/5.1429$), and the last point is $5.14/6 = 85.7\%$ of a uniform-6 cache, i.e. a $14.2857\%$ payload saving relative to uniform-6; similarity of NLL does not establish equivalence. (Bit labels of the form $4+k/224$, stored with some runs, undercount the payload by one bit per upgraded unit and are not used.) The byte denominator is part of the claim: under the paper's own gs32 symmetric code the per-block fp32 scale metadata (App. A.10, $+1.0$ bit/value) is identical on both sides and cancels in an equal-layout comparison, so the equal-layout cache-byte saving is also $12.2449\%$ ($(5.142857+1)/(6+1)$); if scales are kept fp16 ($+0.5$ bit/value) it is $13.1868\%$; and charged per-value against a no-metadata 6-bit reference it is $14.2857\%$. Any cross-layout byte comparison (e.g. against the FP8 rows below) must fix this denominator explicitly. *(practical regime)* The $\{4, 6\}$-domain increment becomes decision-relevant in exactly two measured places: (i) as a payload statement, $5.14$ bits/unit with small measured NLL damage and a $14.3\%$ value-level payload saving (byte-denominator line above) — a byte-budget-decision input at $\geq$4-bit serving where the window is already near-lossless on NLL; and (ii) on the retrieval-tail evaluation window, where uniform-4's small damage is *not* uniformly negligible: the sub-5-bit-vs-uniform-4 gap separates with $n$=19 per-doc CIs there (tail $-0.020$ [-0.028,-0.012] at 4.57 bits, App. A.7), the same tail where the 3.29-bit-vs-u4 gap concentrates. The tail instrument describes loss concentration on the selected evaluation window; it does not show that a sub-0.03-nat change improves an application metric or increases usable batch or context length. (An anchor-control (App. A.7) shows this tail concentration is end-of-document robustness, not long-distance-retrieval value — no retrieval claim is attached to the tail instrument.) No latency or throughput statement is attached to either.

*(FP8 comparator)* Production FP8 KV (uniform E4M3/E5M2, 8 bits, no per-element scale metadata) is measured on the same window, against uniform-3/4/6 (same Mistral 12-doc pool, forward difference dC=0.0): uniform FP8-E4M3 is $+0.0032$ vs full — interval-separated better than uniform-4 ($+0.0090$ [+0.0071,+0.0108]) and only $-0.0029$ [-0.0034,-0.0023] off uniform-6 ($+0.0003$) — i.e. an 8-bit, zero-calibration, zero-measured-allocation cache is essentially uniform-6 quality; uniform FP8-E5M2 is merely $\approx$ uniform-4 ($+0.0111$). Framework read: the mixed allocation must be judged against both uniform widths and this non-allocated baseline — the measured 5.14-bit point saves $14.3\%$ over 6-bit INT and $\approx$36% over 8-bit FP8-E4M3 at the value level (byte denominators as fixed in the cost fact above; the saving uses the payload ratio $5.1429/8$ payload line), while a plain 8-bit FP8-E4M3 cache delivers near-uniform-6 quality with no calibration or allocation (caveats: E4M3's narrow max-exponent range needs per-tensor/group scale handling in production, and the FP8 rows here are value-bit-only). These comparisons concern nominal payload and NLL. They do not establish matched quality, actual packed-cache compression, or a serving advantage over FP8.

### A.6 CONTINUED: 16K EXPERIMENT DETAILS

• Natural-text leg (App. A.6 (f)): rows u3/u2/u4/k2v4/k4v2 $\times$ codes {sym, min-max}, 2D-group gs32, per-doc NLL deltas. Pool A: seed-7, 6 long books (books 0,44,53,77,80,90; extends a four-book seed-7 run to six), pool B: seed-9, 12 fresh long books (24,48,49,50,65,75,85,86,94,95,96,99), both disjoint from the 38-doc Qwen evidence footprint. Measured in two jobs: seed-7 pool A documents d0–d3, then d4/d5 together with seed-9 pool B in full. Pooling: phq/scripts/phq_law16k_pool.py (keyed by doc id, per-doc bootstrap resample, pooled llm_law16k_natural.json n=18).

• 4k paired companion: the same 18 natural + 6 retri books re-scored at span 0,4096 $\rightarrow$ within-doc (d16k - d4k) paired CIs (phq_law16k_scaling.py; phq/runs/llm_law16k_scaling.json). All four legs' per-doc JSONs carry the forward comparison 0.0/doc.

• Retri leg (App. A.6 (g)): retrieval-proxy corpus at T=16384, n=6 (books 23,24,29,82,90,95; anchor$\rightarrow$continuation pairs with 4k/6k distractor gaps; pkvr/scripts/pkvr_build_retri_corpus.py seed-19, evidence footprint excluded). Same rows x codes protocol; pooled llm_law16k_retri.json n=6.

• Cross-scale leg (App. A.6 (f)(iv)): the same 24 books (18 natural + 6 retri) re-scored on Qwen3-0.6B-Base (identical geometry and tokenizer) at T=16384 and T=4096, both codes → second-scale pooled + within-doc scaling (phq/runs/llm_law16k_06b_{natural,retri,scaling}.json). Document-family separation between natural (d/b stems) and retri (r% stems) is enforced in pooling/scaling because the retri builder reuses anchor book ids (24, 90, 95) that also occur as natural docs – the two are distinct document constructions and must not share per-doc keys.

### A.7 Free proxies, the retrieval tail, and the strong-statistics comparison

(Qwen3-1.7B, gs=32, n=6 docs, calibration corpus pg19; free statistics from one full forward.) Spearman(measured d3, proxy) over the 224 heads: key-channel absmax over the block $0.19$ (weak), sink-token attention mass $-0.02$ (nil). At matched 3.286 bits the proxy allocators lose most of the measured gain (pg19/retri): measured-top64 $+0.076$/$+0.071$ vs key-absmax-top64 $+0.140$/$+0.113$ and sink-top64 $+0.114$/$+0.100$ (random-32 control at 3.14 bits: $+0.157$/$+0.129$). So the per-head NLL signal is *not* redundant with these free statistics: that measurement buys signal over free heuristics is an empirical result on this model, not an assumption.

Marginal-vs-joint basis: a greedy incremental allocation (stage-1 top32 by measured d3, then joint re-measurement of every other head's 3→4 upgrade in the presence of stage-1, stage-2 top32) reaches $+0.072$/$+0.070$ vs the one-shot marginal top64 $+0.076$/$+0.071$ — a small margin in the greedy set's favor on pg19 ($-0.0037$ [-0.0051,-0.0022], $< 0$ = greedy gentler), so the two bases are materially exchangeable here. Under the (mild, gs32 ratio 1.35) 3-bit superadditivity the marginal-vs-joint tension does not change the chosen set materially; a single-stage measured-damage ranking suffices.

Task-level translation (synthetic copy/retrieval cell; Qwen3-1.7B, 24 seeds, mid-vocab gold tokens, no-gold control): a repeated gold-pair cue 4090 tokens from the query lowers $-\log P(\text{target})$ by ≈4 nats (control $10.8$→$6.2$–$6.8$) — genuine retrieval signal — but the cell is saturated at every tested width: uniform-3, uniform-4, the 3.286-bit and 4.57-bit points all preserve the copy as well as full precision (residuals $-4.0$..$-4.2$ nats vs the matched control grid; per-seed sd $2.1 \gg$ inter-config spread $0.6$). A single-cue variant at 300 tokens shows no retrieval under any config (residuals within $\pm 0.3$ nats of control; the model's copy threshold sits between the two). The synthetic task-success cell is therefore saturated at exactly the scale the frontier pairs differ; the at-margin discrimination is furnished instead by the retrieval-tail NLL metric below, which resolves those differences with per-doc CIs.

*(retrieval-tail at-margin cell)* Concentrate the metric on the retri corpus's final 256 tokens — the true continuation after the anchor/distractor gap, whose content depends on the anchor book (an anchor-control below separates that content-dependence from the quantized ordering). At the data-limited maximum $n$=19 (fresh build, seed-7/T-4096 protocol; per-doc 90% bootstrap CIs), the frontier pairs separate on the tail: The 3.29b-vs-u4 quality difference — the tightest frontier pair,

| pair (damage delta) | whole-span gap [CI] | tail-256 gap [CI] | tail/whole |
|---|---|---|---|
| 3.29b mixed − uniform-4 | $+0.031$ [+0.028,+0.034] | $+0.040$ [+0.029,+0.051] | 1.29 |
| 4.5714b mixed − uniform-4 | $-0.022$ [-0.023,-0.020] | $-0.020$ [-0.028,-0.012] | 0.9 |

where the synthetic needle was flat — is interval-separated on the tail and $1.29\times$ concentrated there (+0.031→+0.040), and the sub-5-bit point's advantage over uniform-4 is maintained on the tail ($-0.020$ [-0.028,-0.012], ≈$3.9\times$ lower damage; the directional n=6 margin-narrowing did not reproduce on this corpus, so the interval-separated n=19 reading is the reportable one; at n=6 the same pairs showed only 11–12% concentration). Cross-scale check (Qwen3-0.6B-Base, same retri $n$=19 seed-7/T-4096 build, parity maxdiff 0.00e+00): the separation reproduces at 0.6B — 3.29b-vs-u4 tail gap $+0.042$ [+0.030,+0.055] and 4.5714b-vs-u4 $-0.032$ [-0.040,-0.025], same sign and near-magnitude as 1.7B — so the at-margin metric's deliverable is scale-invariant. The amplification is not: at 0.6B the 3.29b pair shows no tail concentration (tail/whole 1.007 vs 1.29 at 1.7B; 0.6B whole-span gaps are themselves larger, $+0.042$ vs $+0.031$, so the tail window adds no additional discrimination), while the sub-5-bit advantage turns tail-concentrated (1.30) instead of slightly diluted (0.9). Scope: the tail-256 cell separates across scale; whether it amplifies the boundary pairs or merely mirrors the whole-span reading is scale-dependent. Two scope notes bound the instrument: it is evaluation-side — selecting the 4-bit set by the per-head tail marginal instead of the whole-span one does not improve the tail (nor the whole span) at $n$=19 (tail row ≈+0.017 worse on the whole

span, $\sim$2 sd, within noise on the tail; per-head 256-token tail marginals are noisier than whole-span ones, shared-fraction overlap vs the whole-selected set 0.58 (top-64 set; chance $64/224 \approx 0.29$)), so the at-margin signal separates aggregated frontier pairs but does not redeem per-head selection; and it is retri-specific and NLL-based (a task-success instrument remains open).

*(anchor-control, same $n$=19 pool, parity 0.0.)* Whether the tail separation actually flows through retaining the anchor: we replace the first 128 anchor tokens with (i) the next document's anchor (swap) and (ii) seeded random tokens (delete), leaving distractor and true continuation identical, and re-measure the frontier tail gaps per-doc paired 20k-bootstrap 90% CIs. The measured separation persists under both interventions: the 3.29b-vs-u4 tail gap is $+0.040$ [+0.029,+0.051] intact, $+0.043$ [+0.028,+0.056] swap, $+0.044$ [+0.032,+0.056] delete, and the per-condition differences straddle 0 ($+0.003$ [-0.011,+0.016] swap, $+0.004$ [-0.006,+0.014] delete); the sub-5-bit-vs-u4 tail gap likewise intact ($-0.020$ [-0.028,-0.012] vs $-0.023$/$-0.020$ under swap/ delete). These intervals include zero but also permit nonzero changes; no equivalence margin was specified, so neither statistical equivalence nor independence from anchor availability is established. The observed nonzero separation after replacement weakens a strict anchor-necessity account, but does not exclude partial mediation or isolate a single end-of-document mechanism. The data therefore support the descriptive sensitivity-concentration instrument, without a retrieval-value claim, rather than a "requires-retaining-the-anchor" interpretation. The frontier means and CIs above use the stored equal-document paired arrays. Separate whole/tail aggregate fields stored with the runs give equal weight to chunks of 6/6/6/1 documents; they are not the inputs to the reported CIs, which use the equal-document reanalysis. All CIs in this cell are the equal-document per-doc $n$=19 bootstrap family (a stored tabulation with an $n$=12 resample is not used).

*(strong-statistics extension, Mistral 12-doc pool.)* The weak free proxies above leave open whether the offline sensitivity statistics the quantization literature actually uses (KVQuant-style per-channel outlier ratios, dynamic-range manifestations) would redeem selection. We measure five such statistics on the calibration window from full-precision post-RoPE key and value activations: per-channel peak-to-average ($\max/\mathrm{mean}$), per-channel dynamic range ($(\max-\min)/\mathrm{mean}$), and a mean+2$\sigma$ outlier composite ($\sum_c \mathrm{frac}_c \times \mathrm{magnitude}$) for K, for V, and K+V — each as a 2.75-bit head allocator on the same disjoint 8-doc window as Table A.4. Result: the measured per-head field is interval-separated better than every strong proxy at the equal budget — measured $+0.0952$ vs p2a-K / range-K $+0.1074$ ($-0.012$ [-0.017,-0.007]), vs the outlier composite $+0.128$ ($-0.033$ [-0.045,-0.023]), vs random $+0.111$ ($-0.016$ [-0.023,-0.009])); Spearman(proxy, measured field) is $\approx$0 in magnitude throughout ($-0.08$ to $+0.11$; K-side $-0.07$, V-side $+0.11$, outlier composite $-0.08$) — no proxy carries a correlation that could order heads; and the V-side p2a/range statistics are actively harmful at equal bits ($+0.1855$, interval-separated worse than random $+0.111$; the two coincide — per-channel $|v|$ minima are $\approx$0, so range = peak-to-average for V — and both pick the wrong heads by value-magnitude concentration). Interpretation: the strongest sensitivity statistics from the KV-quant literature are uninformative for per-head V-allocation at the deployment budget; "measurement is required" is thereby an empirical result on this model, not an absence-of-effort claim. Scope: Mistral-7B single budget (2.75 bits) and window; proxies are the natural post-RoPE calibration-window statistics (a second-order/Hessian proxy remains unmeasured).

*(cross-family check, Qwen3-0.6B, same cal-4/ev-8 protocol and forward difference dC=0.0 / parity 0.0.)* On the Qwen3 family the picture *reverses* at the same 2.75-bit budget on the disjoint 8-doc window: the measured per-unit top-112 allocator is interval-separated *worse* than random ($+0.335$ vs $+0.311$ nats vs full; measured-minus-random $+0.024$ [+0.012,+0.036]) — the calibration-window field does not transfer to heads that help out-of-sample on this family — while the peak-to-average and dynamic-range proxies (K-side and V-side alike) are only mildly better than random ($+0.037$ [+0.028,+0.045] K and $+0.027$ [+0.017,+0.036] V) and the outlier composite straddles random — a $\approx$0.03-nat margin on a $\approx$0.3-nat constant that does not predict the deployment outcome — and Spearman(proxy, measured field) is $\approx$0 in magnitude ($-0.24..+0.08$; p2a/range on V $= -0.24$, the outlier composite $+0.08$; the V-side proxies anti-correlate with the measured field). Two-family read: on the strong-V>K family (Mistral) the measured field is the decisive signal and CI-separately beats every offline statistic and random; on the weak-V>K family (Qwen) neither the measured field nor these statistics yields a universally useful selector — at 2.75b the measured allocator is interval-separated worse than uniform-3 ($+0.108$ [+0.100,+0.115], at fewer bits), consistent with the family scope of App. A.4(e). For the pre-computable-discriminator question raised in App. A.4(e) this is therefore a two-family negative: no offline statistic among those tested predicts the measured order-

ing or the deployment outcome on either family, so the measured matched-bit comparison remains the deployment calibration on the cells we measured.

### A.8 SERIAL-CACHE PROTOCOL DETAILS

The Mistral-7B experiment uses 12 selected PG-19 documents, a 1024-token prefix, a 128-token gold continuation, and a recent-token residual of 128. At step $t$, group-quantized entries are converted only when a complete 32-token group has left the residual. Thus as many as 159 recent tokens remain unquantized between group updates. Per-channel-key and FP8 variants convert an individual departed token. The simulator retains float32 raw and attention-visible caches; its residual is not numerically bf16. This online conversion uses only written tokens, unlike the offline block interventions. Teacher forcing measures loss under gold histories, while the greedy arm measures disagreement with full-cache outputs, not task quality.

We recompute 90% percentile intervals from the 12 recorded document means using 20,000 paired document resamples with shared indices across methods. The stored means were rounded to four decimals, which also limits the precision of this reanalysis. The intervals stored with the runs resample within-document tokens as independent observations, which is too narrow for a document-level estimand, and use the opposite sign convention; they are not used.

| **Configuration** | **Mean loss damage** | **90% document interval** |
|---|---|---|
| uniform-3 | 0.1917 | [0.1261,0.2667] |
| K2/V3 | 0.2229 | [0.1596,0.2950] |
| front-8 | 0.0470 | [0.0323,0.0609] |
| K2/V4 | 0.0304 | [0.0163,0.0443] |
| windowed channel-K, W=512 | 0.0271 | [0.0151,0.0392] |
| running channel-K | 0.0506 | [0.0382,0.0629] |
| FP8-E4M3 | -0.0001 | [-0.0012,0.0010] |

Memory side of the same seven configurations (one denominator; modeled end-cache accounting):

| **Configuration** | **end-cache bits, 16-bit resid.** | **end-cache bits, fp32 resid.** |
|---|---|---|
| uniform-3 | 4.444 | 6.222 |
| K2/V3 | 4.000 | 5.778 |
| front-8 | 4.111 | 5.889 |
| K2/V4 | 4.444 | 6.222 |
| windowed channel-K, W=512 | 4.444 | 6.222 |
| running channel-K | 4.444 | 6.222 |
| FP8-E4M3 | 8.889 | 10.667 |

At equal modeled payload-plus-residual cost, the paired K2/V4 minus uniform-3 difference is -0.1612 [-0.2351,-0.0994] nats. Front-8 minus uniform-3 is -0.1447 [-0.2177,-0.0860]. Windowed channel-K minus K2/V4 is -0.0033 [-0.0226,+0.0144], so their difference remains unresolved. Running minus windowed channel-K is +0.0235 [+0.0133,+0.0337]. These are conditional intervals for this fixed model and selected pool, not cross-model or new-search uncertainty.

The memory columns apply one denominator to every configuration: at the end S=1152, 1024 entries per channel have quantized payload and 128 remain high precision. They exclude scale/offset metadata and the simulator's retained full-precision scale-source buffers, so they are not total or measured memory. At intermediate steps, group rows also retain an incomplete group. In particular, K2/V4 is not a 3-bit total-cache alternative to a 4.44-bit channel-K configuration: both have the same end residual cost. The block and serial experiments score different windows with different cache construction, so the ratio of their loss gaps does not isolate an effect of error compounding. FP8's small loss delta likewise does not establish equivalence of generated outputs (18% greedy-token divergence).

### A.9 PROTOCOL COMPARISONS IN FULL, AND THE MECHANISM SKETCH

**Cross-scale serial replication (Qwen3-0.6B).** The same serial comparison on the identical 12-doc pool and protocol reproduces every scale-rule readout of the serial and scale-timing results

at 0.6B: the symmetric K>V direction persists under both the full-group serial loop (gap +3.680 [+3.624,+3.742] at $R$=0) and the prefix rule (+1.414); the min-max V>K direction persists under the full-group loop (gap -1.111 [-1.230,-1.000]) and inverts to an interval-separated weak K>V gap under the per-position prefix scale (+0.026 [+0.022,+0.030]). The block-FUT to block-PRE split is the same sign structure with larger magnitude at the smaller scale (sym +0.501→+1.409; mm -0.841→+0.025 [+0.019,+0.030]).

(a) The official-forward comparison evaluates the full-precision baseline on 12 fixed documents using both transformers and the manual implementation. Agreement of document-average loss to the recorded precision supports that baseline comparison. Both reported quantized deltas reuse the same manual quantized numerator; agreement after changing only the denominator does not establish equivalence to a quantized official-model execution. Nor does chunk-size invariance imply that a common logit discrepancy cancels through nonlinear cross entropy. All block results are therefore identified as manual-harness loss interventions with this limitation.

(b) The recorded suffix experiment fixes a 2048-token prefix, aligned to the 32-token quantization groups. The full-span, prefix-only, and replaced-suffix measurements address influence from outside that prefix. Group scales still use later tokens inside each group: for a token before the group's end, future keys or values can alter the quantization of its visible predecessors. An aligned-prefix result cannot establish per-token causality. Whole-span channel-key scales show additional measured suffix dependence (+0.5453 [+0.4912,+0.5990] nats, full-span vs prefix-only scale readout), while the serial-cache design in Sec 7 avoids using unwritten tokens. This distinction changes the interpretation of the block tables: they describe offline block interventions, not verified causal serving laws.

(c) Within-group future-scale control. We bound the within-group coupling in (b) with a write-time causal counterfactual (PRE) of the same 2D-group rows: every entry's per-channel scale (absmax for the symmetric code; min-max for the asymmetric code) is formed from its group's written prefix $[g_0, t]$ only, so no future member enters any encoding a prediction can read. The block-table rows use the future-inclusive (FUT) protocol; PRE re-encodes the same bit width with only causally available scale data. The FUT arm reproduces the stored per-doc anchor deltas to $5 \times 10^{-5}$ on all 24 documents (both codes), which anchors the FUT reconstruction. PRE deliberately changes the scale rule. Qwen3-1.7B at 4k, natural pool (n=18), gs32, per-doc mean damage in nats:

| code | row | FUT (full-group, natural n=18) | PRE (write-time prefix) | PRE − FUT |
|---|---|---|---|---|
| sym | uniform-3 | 0.1664 | 0.0692 | [-0.1020,-0.0920] |
| sym | uniform-2 | 3.0218 | 1.7788 | [-1.3040,-1.1800] |
| sym | K@2/V@4 | 1.3532 | 0.9773 | [-0.4130,-0.3400] |
| sym | K@4/V@2 | 0.8793 | 0.1700 | [-0.7430,-0.6770] |
| min-max | uniform-3 | 0.0944 | 0.0118 | [-0.0880,-0.0780] |
| min-max | uniform-2 | 0.6178 | 0.0793 | [-0.5750,-0.5030] |
| min-max | K@2/V@4 | 0.1270 | 0.0513 | [-0.0790,-0.0720] |
| min-max | K@4/V@2 | 0.4361 | 0.0285 | [-0.4370,-0.3780] |

Changing from full-group to prefix scales substantially changes several reported magnitudes (28–94% of the 2-bit rows: value-side 80–94%, key-side 28–60%; and 58–87% of uniform-3), so the absolute damage figures in Tables A.1–A.4 are labeled as FUT-scale measurements. Their numerical differences from PRE are not general upper bounds for causal kernels. The ordinal ordering readouts split: the symmetric K>V direction persists and strengthens (2-bit-key over 2-bit-value gap +0.4739 [+0.4176,+0.5328] → +0.8073 [+0.7792,+0.8356]); the 2-bit floor (uniform-2 ≫ uniform-3) persists on both codes (symmetric ratio $18\times \rightarrow 26\times$, min-max $6.5 \rightarrow 6.7\times$, both $\approx 6.6\times$); and the equal-bit-split-loses (k2v4>u3 at 3.0 mean bits) persists on both codes (+1.1868→+0.9081 symmetric, +0.0326→+0.0395 min-max). The min-max code's V>K reversal flips to weak K>V parity under the prefix scale (+0.0228 [+0.0201,+0.0254]; V-2 damage collapses ≈93%, 0.4361→0.0285) — 2-bit V is cheap precisely when its scale is per-position — but this is a per-position-scale content effect, not an isolated estimate of causality or cache refolding: the write-once serial protocol with group-at-completion scales preserves both orderings on the same battery (Appendix A.9(d)), so the reported min-max V>K (Table A.2) is a property of the group-wide min-max scale geometry rather than an intrinsic ordering of K/V sensitivity; future reading alone does not explain it. The PRE − FUT term is uniform across within-group offsets (interior per-offset means within $\pm 25\%$ of the pooled value, same sign everywhere) and flat along the sequence apart from a modest early-region enhancement, so it is not a boundary artifact of the 32-token grid. The same directions and comparable term magnitudes hold on the retrieval-proxy pool (n=6). The control reproduces on the other

two models: at 4k, the symmetric K>V gap stays interval-separated under PRE on Qwen3-0.6B ($+0.497 \to +1.402$ [+1.367,+1.437], same 24-doc battery) and on Mistral-7B ($+1.264 \to +0.663$ [+0.587,+0.745], n=12); the min-max V-2 damage collapses ($+1.08 \to +0.05$ and $+1.32 \to +0.02$; the ordering crosses into weak K>V near-parity, $+0.025$ [+0.021,+0.029] at 0.6B and $+0.003$ [+0.0002,+0.0057] at Mistral); and the 2-bit floor persists on both. Notably the Mistral equal-bit $k2v4$-beats-uniform-3 gap of Table A.3 (-0.165, n=12) flips to +0.018 [+0.017,+0.020] under PRE, yet the serial-cache simulation (Sec 7), which delays quantization to group completion while keeping the full-group scale, preserves the win (-0.1612): the operative ingredient of that row is the full-group scale, not causality itself. A 16k repeat on the seed-7 natural (n=6) and retrieval-proxy (n=6) windows reproduces every direction: the symmetric K>V gap grows under PRE (+2.1964→+2.7090 natural, +1.8120→+2.4598 retri), uniform-2 remains $\gg$ uniform-3 (symmetric ratio 14.7× →21.9× and 10.9× →16.1×), each pool's 2-bit-V ($k4v2$) row collapses toward uniform-3 under PRE (+0.9014→+0.1793 natural, +1.0302→+0.2003 retri; uniform-3 minus $k4v2$ within $\pm 0.02$ nats of parity: -0.0145 [-0.0240,-0.0050] natural, +0.0027 [-0.0040,+0.0110] retri), and the min-max ordering flips to a weak K>V gap (natural +0.0366 [+0.0340,+0.0390]; retri +0.0355 [+0.0320,+0.0390]); the length-extension readout of App. A.6 (f) carries the same FUT-scale boundary.

**Consistency across pools (the Appendix A.9(c) keys-vs-values gap on three pools).** The 2-bit-key over 2-bit-value damage gap ($k2v4$ minus $k4v2$) keeps one sign structure on every pool we ran: symmetric keys>values persists and strengthens from full-group (fut) to write-time causal (pre) scales, while min-max inverts from values>keys (fut) to a small keys>values (pre). Two independent document pools at 4k — the block battery's natural pool and the serial battery's nat+retri pool (the 12-document battery of item (d) below) — plus the retrieval-proxy pool give the same conclusion, so no single pool carries the readout:

| pool | symmetric fut | symmetric pre | min-max fut | min-max pre |
|---|---|---|---|---|
| block battery, natural (n=18) | +0.4739 [+0.4176,+0.5328] | +0.8073 [+0.7792,+0.8356] | -0.3091 [-0.3374,-0.2819] | +0.0228 [+0.0201,+0.0254] |
| block battery, retrieval-proxy (n=6) | +0.4295 | +0.8676 | -0.3312 | +0.0245 |
| serial battery, nat+retri (n=12, item (d)) | +0.471 [+0.422,+0.526] | +0.843 [+0.809,+0.877] | -0.338 [-0.352,-0.321] | +0.025 [+0.020,+0.029] |

The serial row is computed on the serial write-once loop (item (d) below), the two block rows on the batch refold; absolute magnitudes are therefore not pool-for-pool comparable, but the sign structure is.

(d) Serial simulation versus block evaluation. The block rows (FUT, PRE) are batch refolds: the quantized cache is replayed forward once, so its errors are closed under the quantized loop. A deployed write-once kernel is a causal unfold: each token is quantized once at group completion (full-group scale) or at write (prefix scale), with $R$ recent tokens kept full precision and no unwritten token entering any scale. On the same 12-doc battery (Qwen3-1.7B, 4k, gs32; teacher-forced over the full span) the two loops agree on orderings but differ in magnitude:

| | block-FUT | block-PRE | serial full-group | | serial prefix | | cost (16-bit |
|---|---|---|---|---|---|---|---|
| row | (n=12) | | $R$=0 | $R$=128 | $R$=0 | $R$=128 | residual) |
| sym uniform-3 | +0.1641 | +0.0697 | +0.0965 | +0.0591 | +0.0708 | +0.0059 | 3.0 / 3.41 |
| sym uniform-2 | +2.9479 | +1.7718 | +3.2007 | +1.9487 | — | — | 2.0 / 2.44 |
| sym K@2/V@4 | +1.4085 | +1.0230 | +2.6427 | +1.2936 | +1.0124 | +0.7458 | 3.0 / 3.41 |
| sym K@4/V@2 | +0.9378 | +0.1801 | +0.4090 | +0.3565 | +0.1803 | +0.0040 | 3.0 / 3.41 |
| min-max uniform-3 | +0.0937 | +0.0116 | +0.0684 | +0.0617 | +0.0116 | +0.0006 | 3.0 / 3.41 |
| min-max uniform-2 | +0.6429 | +0.0846 | +0.5647 | +0.4940 | — | — | 2.0 / 2.44 |
| min-max K@2/V@4 | +0.1305 | +0.0533 | +0.0698 | +0.0441 | +0.0526 | +0.0067 | 3.0 / 3.41 |
| min-max K@4/V@2 | +0.4682 | +0.0284 | +0.4674 | +0.4344 | +0.0286 | +0.0008 | 3.0 / 3.41 |

The two K/V orderings reproduce under the serial loop (symmetric K-2 $+2.64/+1.29$ vs V-2 $+0.41/+0.36$ at R=0/R=128; min-max V-2 $+0.47/+0.43$ vs K-2 $+0.07/+0.04$, and mm K-2 stays near-free as in the Mistral serial window of Sec 7). Magnitudes move between the refold and the unfold: the block understates the serial symmetric-2-bit-key damage (sym K-2 $+1.41$ block vs $+2.64$ serial R=0) and uniform-2 (sym u2 $+2.95$ vs $+3.20$), while overstating uniform-3 ($+0.164$ vs $+0.096$); the residual cut R=128 removes most of the remaining recent-token damage on the prefix rows but only modestly on the catastrophic symmetric-2-bit rows. The block tables therefore report refold magnitudes (self-consistent batch interventions); the serial rows above and in Sec 7 are

the causal-decode readout, and Sec 7's stream reanalysis remains the Mistral-pool evidence that serial K2/V4 beats uniform-3 ($-0.161$ [-0.235,-0.099]) while leaving its difference unresolved against a windowed per-channel-key baseline ($-0.003$ [-0.023,+0.014]).

**Mechanism probes and unsupported explanations.** The observed joint-to-isolated ratio does not identify pairwise interactions, attention fan-in, or a particular residual-stream mechanism. Several probes constrain these interpretations without isolating a causal mediator. In one uniform-3 setting, extending the attended span from 256 to 1024 tokens changes damage from +0.189 to +0.201; the 2048- and 4096-token extensions give +0.192 and +0.196. The sublinear, partly nonmonotone response does not support damage proportional to attention fan-in.

Code and scale changes also alter the result. Full-group min-max reduces the measured K-side penalty by 10–32$\times$ relative to the symmetric code, and the original min-max 3-bit joint/sum ratio is 1.06. Under PRE, the value-side damage changes from +0.44 to +0.03, while the symmetric K-side gap changes from +1.37 to +1.00. These controls vary code points, scale contents, and error magnitude together. They cannot uniquely attribute the ordering to saturation, future access, or error decorrelation.

The direct error-vector probe uses two PG-19 documents, 1024 tokens, and KV heads 0–3 in every layer. Here "pre-quantization" refers to captured unquantized keys and queries, not to the PRE scale protocol, unless a probe row states otherwise. Errors are recomputed with the recorded group-size-32 quantizer as $e = k - Q(k)$. Their relative magnitude is $\langle|e|\rangle/\langle|k|\rangle = 0.54$ for symmetric 2-bit and 0.24 for min-max, a 2.3$\times$ gap. The key alignment is -0.21 versus -0.03. With this error convention a negative projection onto $k$ means that $Q(k)$ overshoots along the key direction; it does not mean that quantization removes that direction. Query alignment is approximately zero under every tested code and width, and mean absolute inter-head error cosine is approximately 0.08. Thus the proposed attention-aligned, coherent-across-heads clipping account is not supported by this probe. Large error magnitude remains an observation, not an identified dominant cause of the downstream loss.

Rotation leaves a substantial symmetric 2-bit loss floor: +3.48 changes to +2.57, still about 14$\times$ the rotated 3-bit damage. A proposed common explanation linked that floor with 3-bit nonadditivity across models. The recorded floor/ratio pairs are +4.27/1.50 for Qwen3-0.6B, +2.99/1.35 for Qwen3-1.7B, and +3.48/1.78 for Mistral. Two model pairs are concordant and one is discordant: the floor decreases from 0.6B to Mistral while the ratio rises. The all-pairs monotone prediction is therefore unsupported. The Mistral floor also exceeds the 1.7B floor, contradicting a monotone parameter-count interpretation. These are descriptive comparisons across different model protocols, not paired uncertainty tests. The input values and pairwise calculations are in `runs/mechanism_crossmodel_input_v228.json` and `runs/mechanism_pairwise_rank_v2.json`.

Table 2 (Section 5) gives the leave-one-document-out sensitivity of the full-group versus prefix reversal.

**End-cache cost versus intermediate lengths.** Section 7 gives the common cost denominator of the three serial configurations at the final length $S = 1152$. For a cache of length $S$, residual threshold $R = 128$, and group size $G = 32$, completed-group conversion leaves $Q_g(S) = 32\lfloor(S - 128)/32\rfloor$ entries quantized, whereas individual conversion leaves $Q_i(S) = S - 128$; completed groups therefore retain up to 31 extra floating-point entries between group boundaries. At three average payload bits the nominal cost is $[3Q(S) + b_{\mathrm{res}}(S - Q(S))]/S$. Over the 129 integer lengths from 1024 through 1152, a 16-bit residual gives completed-group costs in [4.444, 4.959] bits and individual-conversion costs in [4.444, 4.625] bits; weighting these nominal bits by the corresponding cache lengths gives 4.713 and 4.529 bits, respectively. The length-weighted accounting thus separates the two conversion rules even though they share the final footprint. This arithmetic is not a measurement of actual storage or attention memory traffic; the simulator uses floating-point source and dequantized buffers, and scale/offset metadata remains outside this payload model.

### A.10 Reproducibility, corpora, compute, storage, artifacts

**Protocol.** The block experiments use per-head fake quantization in a manual transformer forward. A 32-token group uses a scale over its token axis for each channel, with either symmetric integer levels or asymmetric min-max levels. The grouping can couple a prediction to later tokens inside the

group, even under a causal attention mask. Loss deltas therefore quantify that offline intervention. The serial-cache experiment instead waits until all members of a group have been written and have left the residual; it provides the direct evidence for online cache operation in this study.

**Quantization execution.** Each block row replaces the post-RoPE attention-visible cache in place: for every selected (layer, kv-head unit) and every 32-token group, the per-channel scale (absmax for the symmetric INT code; scale and offset for the asymmetric code) is read from the group's full-precision values, entries round to the $b$-bit grid and dequantize to fp32, and the forward continues on the dequantized cache with unchanged masks and arithmetic. Scale metadata (fp32; one per (group, channel), plus offset under min-max) is accounted separately from the value grid, so no table measures a kernel or a memory-saved total; equivalence to a shipped quantized model is not presumed (Appendix A.9(a)). The write-time causal counterfactual (Appendix A.9(c)) replaces each group scale with its running prefix statistic over $[g_0, t]$ — the only scale data a write-once kernel would have for a still-open group — and its future-inclusive (FUT) arm reproduces all 24 stored per-doc anchor deltas to within $5 \times 10^{-5}$ on both codes.

Document counts, windows, and quantizer conventions vary across the reported panels and are specified with their results. The streaming reanalysis uses 12 paired document means and 20,000 resamples (Sec 7). Checks of unquantized logits and chunk-size sensitivity are diagnostic observations; they do not establish quantized official-model equivalence. Mistral uses its own tokenizer, so matched book streams across model families do not imply identical token IDs.

**Corpora (definition).** (i) `pg19`: the first 4096-token stream of n=6 seeded PG-19 books (raw token ids, no wrapper filtering). (ii) `retri`: a retrieval-proxy corpus built from PG-19 books — each 4096-token doc lays down an anchor slice (128 tokens) followed, after 1.1–1.9k tokens of other-book distractor prose, by the anchor's true continuation (256 tokens), so the tail's content only makes sense with the anchor. (An anchor-control (App. A.7) finds that the measured quantization ordering persists after replacing the anchor; this does not establish equivalence or the absence of partial anchor dependence.) (iii) `mid8k`: the middle third of each of n=4 PG-19 books (excluding Gutenberg header/trailer boilerplate), truncated to 8192 tokens. All three are token-id `.npy` slices; doc ids/picks are deterministic via `default_rng(doc_seed)`; spans are fixed 0–4096 (8k rows use mid8k at 0–8192). Note on "corpora": pg19/retri/mid8k are differently constructed windows over the same PG-19 book data (book-head stream, interleaved-retrieval-with-anchor, and mid-book), not cross-domain corpora; "cross-corpus" transfer in the text is within-PG-19 construction transfer, and the only cross-model-construction claim is the measured model dependence (Sec 5).

**Compute.** Unless stated otherwise, jobs run on a single GPU in fp32 with 4096-token 6-doc batches. A single-width per-head damage field is exactly $(n_{\text{heads}} + 1)$ forward passes per doc-batch: 225 for Qwen3-1.7B's 224 KV heads. Measured (per-phase timers, identical harness): the 224-head d(4) sweep on a 6-doc batch took 245s ($\approx$1.1s/forward), a per-doc singleton 0.19s, and the complete battery (2 sweeps + 1344 per-doc singletons + 17 rows + per-doc CIs, incl. model loads and parity) finished in 770s wall clock on one A100-class GPU. Calibration cost is therefore $\approx$4 GPU-minutes per width-field; a deployment needs exactly one field on one calibration corpus per model, and the full $\{3, 4\}$ and $\{4, 6\}$ frontier batteries are both reproduced within a single 13–28-minute GPU job.

*Storage scope.* Offline payload ratios describe bit arithmetic for specified layouts. Actual serving additionally requires residual entries, scale/offset metadata, any scale-source buffers, and implementation working memory. Sec 7 applies the same residual denominator to all serial configurations. No packed mixed-width kernel, actual memory saving, or speed measurement is provided, and the cost of implementing such a kernel is not established.

**Artifacts.** The supplementary material stores one JSON result file per cell under `phq/runs/`, holding per-document values for every row so that all intervals can be recomputed. Its `README.md` gives a table-to-artifact map from each table, figure, and appendix section to the exact result directories, the scripts that produced them, and the analysis scripts that derive the printed values; it also gives the corpus specifications and the environment. The stored cells comprise: the attended-token scaling complement at S=2048 (CPU-anchored, with the full-precision comparison re-anchored in the store); the SmolLM2-1.7B replication (llama-type multi-head attention, same 12-document PG-19 pool protocol at span 2048); the Qwen3-1.7B matched-window reference cells (a CPU/GPU device pair at span 2048 and the span-4096 attribution cell, same 12-document PG-19 pool and

rows, gs32); the Qwen3-8B-Base cell (36 layers, 32 query and 8 KV heads, same pool at span 2048, CPU fp32, rows k3v4/k4v3 $\times$ {sym, mm} $\times$ {FUT, PRE}, with complete 2-bit FUT and prefix rows for all 48 row–document entries); the per-cell directories of the 2-bit, kv, group, 8k, prefix, c06b, transfer, base4, cis, asymalloc, cis-retri, mistral, cond, refine, tail, taskrow, konly6, kvmm2, mistral-split, and split-ci cells; the equal-bit selection batteries (s34mm and s46mm at 0.6B, 1.7B, and 4B, each with a seed-8 pool-seed store and a sweep-preserving seed-7 re-execution, plus the static34 and static46 layer arms); two validation stores for the RoPE-precision check of App. A.1; the proxy allocator; the synthetic tasks; the KIVI-layout, streaming, and rotation cells; the dual-code and calibration-count cells; and the serial, FUT/PRE, and law families. The selection, static, proxy, task, KIVI-layout, dual-code, calibration-count, law, and validation cells carry a run log with a completion marker; the twelve robustness-suite cells instead store sweep sidecars (plus stability files on the six seed-8 stores) and an embedded cross-entropy self-check field; and the serial and FUT/PRE families store their result files together with their analysis counterparts. Earlier attempts that no reported value uses are kept with their logs for provenance and are marked as such in the README. One wrapper script per cell (`phq/scripts/wrapper_phq_*.sh`) reproduces it in a single run, and `phq/scripts/phq_emit_values.py` derives the printed values from the result files; the manuscript figure script recomputes paired document intervals and payload budgets. Seeds and protocols are fixed within each cell (doc seed 7, spans 0–4096 in the 4k panels, gs=32), with per-panel document counts specified at the claims.

**Scope of the code-family claims.** Quantization here is per-channel and 2D-group symmetric INT on post-RoPE keys and values with no rotation compensation; the measured K$>$V ordering and the $\leq$3-bit superadditivity are claimed for that (unrotated, symmetric) code family only — confirmed across three models (Qwen3-1.7B/0.6B, Mistral-7B), with the K$>$V sign consistent but its strength family-dependent. Rotation-compensated or outlier-aware schemes (QuaRot-style) redistribute key-range outliers and may change both conclusions; checking this would require reimplementing rotated forwards outside the current harness, and is left as explicit future work. (The cache-path K/Q rotation with a full-precision comparison *is* measured in App. A.2; only a full weight+activation rotation is untested.)

*Analytical decode-cost model for mixed per-head widths (no kernel benchmark claimed).* The byte numbers above count K+V value payload only. On top sits the 2D-group scale metadata: one fp32 absmax scale per (group-size $\times$ channel) block, i.e. $4/\mathrm{gs}$ bytes per value ($+4/\mathrm{gs}$ for the min-max zero-point), and – for per-head mixed widths – a shared per-unit width header (a constant 0.25 KiB per token at the 256-unit Mistral layout, independent of the group size; not a per-value term). Mistral has 65536 K+V values/token (256 units $\times$ 128 dims $\times$ 2 tensors; the value-only wide row is uniform-4 $=$ 32.0 KiB, 3.29-bit $=$ 26.3 KiB as reported above). Adding the metadata: at the paper's gs32 symmetric code, uniform-4 becomes $65536 \times 0.625 = 40.0$ KiB and the 3.286-bit point $65536 \times 0.536 = 34.3$ KiB — a 14.3% cut, i.e. the visible $\approx$17.9% value-level cut dilutes because the constant per-value metadata compresses the ratio; at the group scale KIVI ships (gs128) uniform-4 is 34.0 KiB vs the mixed point 28.3 KiB (16.8% cut, near the value-level ratio); min-max adds a zero-point (gs128: 36.0 vs 30.3 KiB, 15.9%). These are byte-arithmetic identities on the paper's own accounting; variable-unpack/header decode is one extra branch per head per block and is *not* measured as latency anywhere — no throughput claim attaches to this model.

### A.11 Extended discussion of limitations

This section expands the four directions of Section 9.1 and refers back to the sections that contain the underlying evidence. The first two paragraphs below correspond to directions (i) and (ii), the statistical-scope and allocation paragraphs to direction (iii), and the scale-timing paragraph to direction (iv).

**Evidence scale and metrics.** The findings come from fixed checkpoints, small book-derived pools (6–18 documents per construction; four calibration and eight evaluation documents at 2.75 bits), two integer code families, and primarily teacher-forced NLL. The checkpoints, corpora, code families, and pool sizes are enumerated in Section 3 and Appendix A.10; downstream task quality is outside the primary evidence. The retrieval proxy shares the natural-text source, so it is neither an independent domain nor a downstream retrieval benchmark. A small FP8 loss delta accompanies about 18% greedy-token disagreement (Appendix A.8), so closeness in NLL does not imply closeness of generated outputs. Wider-context and model controls supply counterexamples as well

as replications, rather than a universal scaling law: the low-budget static regime (Section 6.2), the calibration-transfer and 16k checks (Appendix A.6), and the leave-one-out ranges (Table 2) each describe where an effect holds or fails on the tested pools, and none licenses extrapolation to untested models, corpora, or context lengths.

**Simulation and deployment cost.** All quantization is simulated in floating point: packed memory, kernel latency, generation throughput, and KIVI's released kernel are not measured, and the 245 GPU-second field cost is not a serving speedup. Nominal payload arithmetic is not a memory-access trace, and the byte arithmetic in Appendix A.10 is an accounting identity on the paper's own bit mapping, not a latency model. Payload bits exclude residual storage and scale metadata, which would reduce the 14.3% top-128 saving in a byte model; the serial cost also excludes scale-source buffers and simulation storage, so it is modeled rather than a measured memory reduction. Between group boundaries, completed-group conversion keeps up to 31 more floating-point entries, so a common final footprint is not a common cost trajectory (Appendix A.9).

**Statistical scope.** Intervals condition on the checkpoint and chosen allocation and exclude uncertainty from selecting a new calibration set or running a different search; the paired document intervals attached to the allocation and serial readings (Sections 6–7) are within-pool statements about measurement noise on those documents. An interval covering zero is not an equivalence test, and no equivalence margin was pre-specified (e.g., for windowed channel-K versus K2/V4). The Qwen allocation contrasts (Section 6.1, Figure 3(a), Table 3) select heads on the evaluation pool and are therefore in-sample. In Table 3, measured-versus-random is unresolved on the 4B proxy, and random-versus-bottom is unresolved on the 1.7B proxy and on 4B natural text. The 3-bit Qwen3-1.7B min-max ratios have no intervals. Table 2 is a descriptive influence calculation on the same pools, neither a new-domain replication nor an interval incorporating calibration or model uncertainty. The full-precision agreement with the official model does not validate an official quantized implementation.

**Interactions and allocation.** Superadditivity ratios describe assembled loss, not allocation optimality or the causal interaction mechanism, and support neither an exact additive model at 4 bits nor a universal code-specific boundary. A ranking success at one budget does not make the isolated score an accurate model of low-bit interactions; the top-$k$ rule has no optimality guarantee, and convergence or global optimality is not established. Mixed-background rescoring improves calibration loss but worsens held-out loss in the $\{4, 6\}$ study (Appendix A.3). At 2.75 bits on Mistral, measured head selection has the highest damage of the four allocators in Figure 3(b), and the richer DP search also stays above the static rule (DP-minus-static $0.0108$ nats). On Mistral, measured-versus-static comparisons are unresolved at 3.25 bits with the symmetric code and under min-max at 3.25 and 3.5 bits. The 2.75-bit recipe transfers poorly to Qwen3-0.6B, where it can be worse than random selection or uniform-3, and cross-code rankings and document-ensemble selectors do not transfer reliably (Appendices A.3–A.7); the evidence identifies operating points, not a model-independent allocation law. Published-form gradient/degradation and range proxies offer no clear advantage over these controls on the tested cells (Appendix A.7). These restricted proxies, the layer-width DP, and our code and group-size controls do not reproduce the complete published systems (e.g., the full KVTuner search). Similar loss at 5.143 bits does not establish equivalence to uniform-6.

**Scale timing.** FUT and PRE are two specified quantizers: PRE also changes scale contents and granularity, so their difference does not isolate the benefit of causality from scale geometry, and PRE is not a serial serving protocol. On Mistral-7B the PRE difference ($+0.003$ nats) is negligible in magnitude. A fixed, group-aligned prefix test addresses only dependence on text outside that prefix, since later positions inside a group can still affect earlier quantization. Under additional residual, scale-window, and group-size controls the K2/V4 benefit can disappear or invert (Appendix A.9), and group-size and rotation controls do not isolate a single error mechanism. On SmolLM2-1.7B, a multi-head-attention model, the min-max gap only shrinks toward zero under PRE instead of reversing (Appendix A.6). The error-vector probe of Appendix A.9 does not validate a universal cross-model mechanism relating interaction strength to residual error alignment. Different windows and cache construction prevent reading the block-to-serial loss change as an isolated compounding effect.