

One Codebook for Every Architecture: Asymmetric NormalFloat for Calibration-Free KV-Cache Quantization

Anonymous authors

Paper under double-blind review

Abstract

Quantizing the key-value (KV) cache is the standard route to long-context LLM inference, and the data-free NormalFloat-4 (NF4) codebook (Dettmers et al., 2023) is a popular calibration-free choice. We show that **NF4 fails silently and catastrophically on an entire class of models**—those with high-ratio grouped-query attention (GQA) (Ainslie et al., 2023)—while appearing excellent on the Llama-family multi-head-attention (MHA) models that dominate published KV-quantization benchmarks. On Qwen2.5-7B (7:1 GQA), 4-bit NF4 KV quantization collapses LongBench-qasper from **43.8 (fp16) to 4.7** and WikiText-2 perplexity from **7.46 to 74.7**—degenerate repetition—whereas the same recipe is near-lossless on Llama-2-7B/13B and Mistral-7B. We decompose the failure into two measured factors. *Representation*: KV keys carry a large per-channel DC offset, so NF4’s zero-centred abs-max grid spends about half of its codes where there is no data—a claim we make exact and machine-check in Lean 4/Mathlib, which also shows the offset correction is optimal and worth a factor $2/\rho$ in worst-case error. *Error tolerance*: NF4’s key reconstruction error is *equal* on Qwen and Llama, so representation alone cannot explain the collapse; the high-GQA model routes each KV-head error into seven query heads and cannot absorb it. Equal error, unequal outcome—the two models differ in how they *read* the same perturbation. The fix follows from the decomposition: **asymmetric NF4 (asym-NF4)** adds a single per-channel zero-point, and is near-fp16 on *every* architecture we test, recovering the Qwen collapse to **41.9 qasper / 7.50 ppl** at an 11% metadata overhead (one extra fp16 scalar per 32-element group). We further show the offset is *RoPE-frequency-structured*, yielding a zero-point computable from weights and configuration alone—calibration-free in the strict sense—that matches calibrated asym-NF4. We release the method, a reproducible cross-model harness, the Lean development, and a practitioner’s decision guide.

1 Introduction

The KV cache dominates the memory footprint of long-context inference (Kwon et al., 2023), and 4-bit quantization of keys and values is now standard practice. Calibration-free codebooks are attractive because they require no data pass and no stored statistics: the 4-bit NormalFloat (NF4) grid, popularised by QLoRA (Dettmers et al., 2023), allocates levels by a unit Gaussian and is a common default for KV keys.

Most KV-quantization studies (Hooper et al., 2024; Liu et al., 2024; Zhang et al., 2023) benchmark on the Llama family (Touvron et al., 2023). We show that this practice hides a failure that is both catastrophic and silent: no exception, no warning, no obviously broken tensor—simply worse answers, and on one architecture class, degenerate ones.

The paper’s organising claim is a decomposition. A quantizer’s downstream damage is not a property of the quantizer alone; it is the product of *how badly the codebook represents the data* and *how strongly the consumer amplifies the resulting error*. Both factors are measurable, and separating them is what turns a mysterious collapse into a one-line fix. Our mechanism section measures each in turn and shows that the first is constant across the models that survive and the model that collapses—so the second is the whole story.

Contributions.

1. **A silent, catastrophic failure mode** of NF4 KV quantization on high-GQA models (Qwen2.5), invisible on the Llama-family models typically benchmarked (Section 4).
2. **A two-factor mechanism, measured:** NF4 costs $\sim 3\text{--}5\times$ the relative key reconstruction error of asymmetric uniform on DC-offset keys, *equally* on Qwen and Llama; the GQA ratio then sets the error tolerance. Bit-depth is not the axis—3-bit *uniform* beats 4-bit NF4 on Qwen (Section 5).
3. **Machine-checked geometry** (Section 6): the “wastes half its codes” claim, the optimality of the zero-point, and a $2/\rho$ error-ratio bound are proved in Lean 4/Mathlib—12 theorems, no **sorry**—and hold for any 16-level grid.
4. **The offset is RoPE-frequency-structured** (Section 7), which yields a zero-point computable from weights and config alone, with no calibration pass and no per-channel metadata, matching calibrated asym-NF4 on perplexity and task score.
5. **asym-NF4** (Section 8): a one-line per-channel zero-point that unifies the codebook choice across architectures at 11% metadata overhead.
6. A reproducible single harness, a cross-model matrix, the Lean development, and a practitioner decision guide (Section 11).

2 Related work

Weight and activation quantization. Post-training quantization of weights is well developed—GPTQ (Frantar et al., 2023), AWQ (Lin et al., 2024)—as is the treatment of activation outliers (Dettmers et al., 2022; Xiao et al., 2023). NF4 itself originates in QLoRA’s weight quantization (Dettmers et al., 2023), where the unit-Gaussian level allocation matches weight distributions that are, by construction of the training process, close to zero-centred. Our finding is in part a statement about *transfer*: the distributional assumption that makes NF4 excellent for weights does not hold for KV keys, which carry a systematic per-channel offset (Section 7).

KV-cache quantization. KVQuant (Hooper et al., 2024) uses offline Fisher-weighted non-uniform quantization with per-channel key grids and pre-RoPE quantization; KIVI (Liu et al., 2024) is tuning-free and asymmetric, quantizing keys per channel and values per token at 2 bits. Both report on Llama-family models. KIVI’s asymmetry is the closest prior art to our fix and we regard our result as an explanation of *why* that design choice matters more than its authors’ evaluation could show: on the MHA models KIVI evaluates, the symmetric/asymmetric distinction is worth a few points, and on high-GQA models it is the difference between working and not. Eviction-based methods such as H2O (Zhang et al., 2023) and attention-sink retention (Xiao et al., 2024) are orthogonal—they choose *which* entries to keep, not how to encode them—and compose with our codebook.

Attention geometry. GQA (Ainslie et al., 2023) and its predecessor MQA (Shazeer, 2019) reduce KV memory by sharing key/value heads across query heads. Their cost is normally discussed in terms of model quality at training time. We show they additionally change a model’s *robustness to KV perturbation*, which is a deployment-time property that the KV-quantization literature has not accounted for. Rotary embeddings (Su et al., 2024) supply the structure behind the offset in Section 7.

3 Background and setup

We quantize keys per channel and values per token, with optional dense-and-sparse fp16 outliers (top 2% magnitude per channel, following Dettmers et al., 2022) and an attention sink (Xiao et al., 2024), in a deployable prefill-once cache that runs at $\sim$ fp16 speed. We compare four calibration-free key codebooks

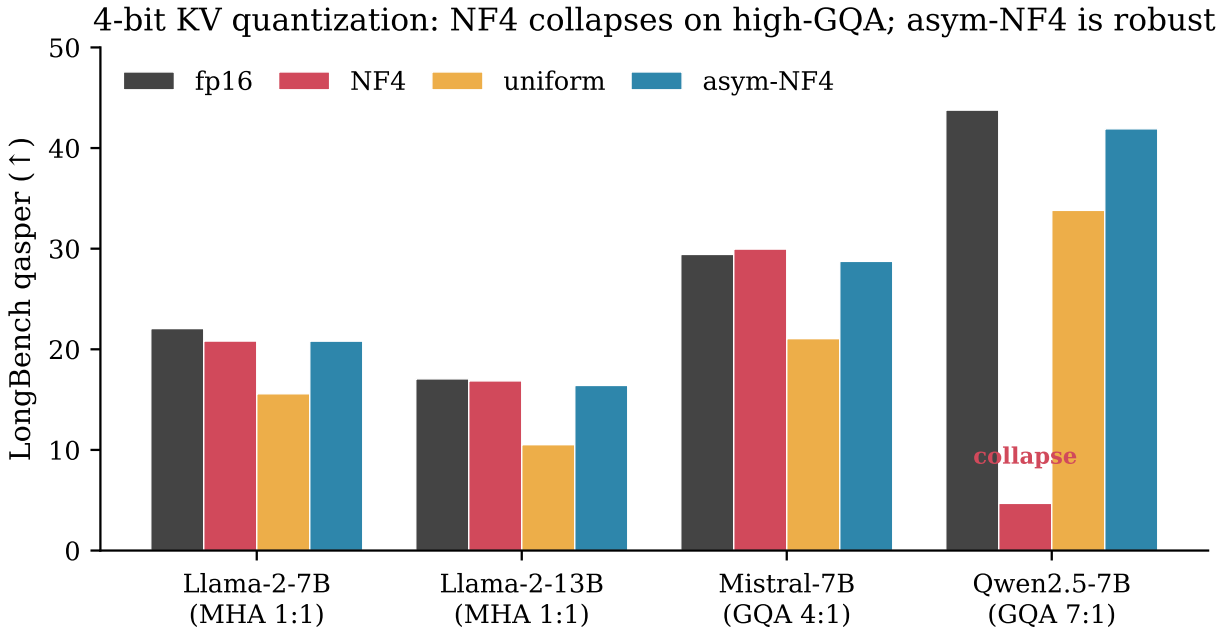


Figure 1: 4-bit KV quantization on the LongBench gasper task. NF4 (red) is competitive on the MHA / low-GQA models but **collapses** on Qwen2.5-7B (7:1 GQA). asym-NF4 (blue) is near-fp16 on every architecture.

at 4 bits: fp16 (reference), symmetric NF4 (abs-max scaling about zero), asymmetric uniform (per-group min/max), and our asym-NF4.

Models span the GQA spectrum: Llama-2-7B/13B (MHA, 1:1) (Touvron et al., 2023), Mistral-7B (GQA 4:1) (Jiang et al., 2023), Qwen2.5-7B (GQA 7:1) (Qwen Team, 2025). We evaluate on the LongBench English subset (Bai et al., 2024) and WikiText-2 perplexity (Merity et al., 2017).

Protocol note. Two WikiText-2 perplexity protocols appear in this paper and are *not* mutually comparable. The **matrix protocol** (Table 2; `kvquant_matrix/wikitext_ppl.py`) scores every non-overlapping 2048-token chunk of the WikiText-2 test split with the full recipe of Table 1: 4-bit keys *and* 4-bit per-token values, group 32, 2% fp16 outlier channels, 4 sink tokens, a 128-token fp16 recent window, the settled region quantized once. The **zero-point protocol** (Section 7; `benchmarks/deterministic_zeropoint.py`) isolates the key codebook: three non-overlapping 512-token windows, keys quantized at group 32, values left in fp16, with the same sink and recent-window handling. Short windows raise absolute perplexity, which is why Qwen2.5-7B fp16 reads 7.46 under the first and 9.06 under the second; every comparison we draw is made strictly within one protocol, and the LongBench numbers in Section 7 use the matrix protocol’s harness unchanged.

4 NF4 fails on high-GQA models

Figure 1 and Table 1 give the headline. NF4 is the stronger of the two naive codebooks on Llama-2-7B/13B and Mistral, but on Qwen2.5-7B it **collapses** from 43.8 (fp16) to 4.7—near-random, and qualitatively degenerate: the model emits repetitive fragments rather than answers. Asymmetric uniform never collapses but sacrifices 5–9 gasper points on the models where NF4 works. asym-NF4 is best-or-tied on three of four models and within 1.2 points of NF4 on the fourth, while being the only codebook that never collapses.

Perplexity tells the identical story (Figure 2, Table 2): on Qwen, symmetric NF4 perplexity explodes $10\times$ ($7.46\rightarrow 74.7$) while asym-NF4 holds at 7.50. That the two metrics agree matters: perplexity is continuous and task-free, so the collapse is not an artifact of LongBench’s answer matching.

Table 1: LongBench gasper ($\uparrow$), 4-bit keys, identical outliers/sink. **Bold** marks our method, not the per-row maximum; the per-row maximum is underlined. Differences below ~ 0.5 points are within run-to-run variation of this harness.

model	attn (Q:KV)	fp16	NF4	uniform	asym-NF4
Llama-2-7B	MHA 1:1	<u>22.06</u>	20.82	15.58	20.81
Llama-2-13B	MHA 1:1	17.06	16.86	10.52	<u>17.34</u>
Mistral-7B	GQA 4:1	<u>29.43</u>	29.96	21.06	28.74
Qwen2.5-7B	GQA 7:1	<u>43.77</u>	4.69	33.81	41.91

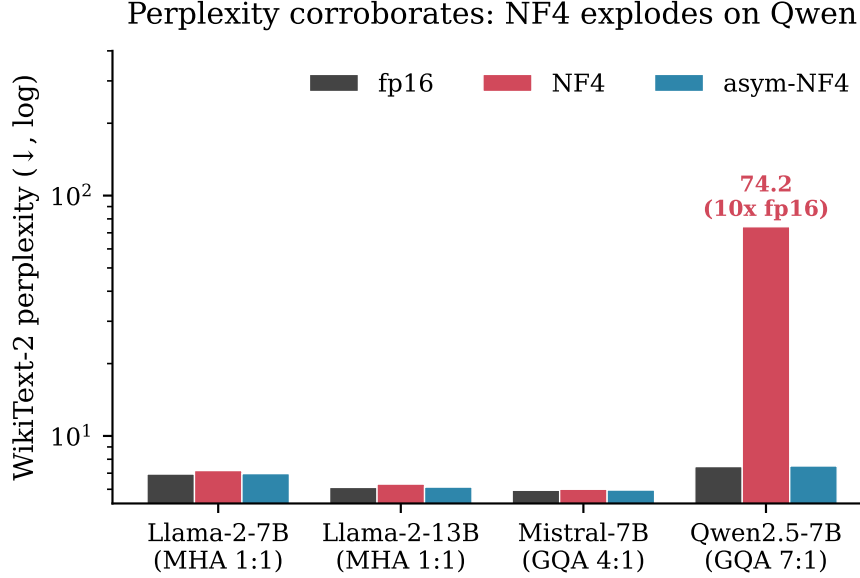


Figure 2: WikiText-2 perplexity ($\downarrow$, log scale). Symmetric NF4 explodes $10\times$ on Qwen2.5-7B; asym-NF4 sits on fp16 everywhere.

5 Anatomy of the collapse: representation $\times$ tolerance

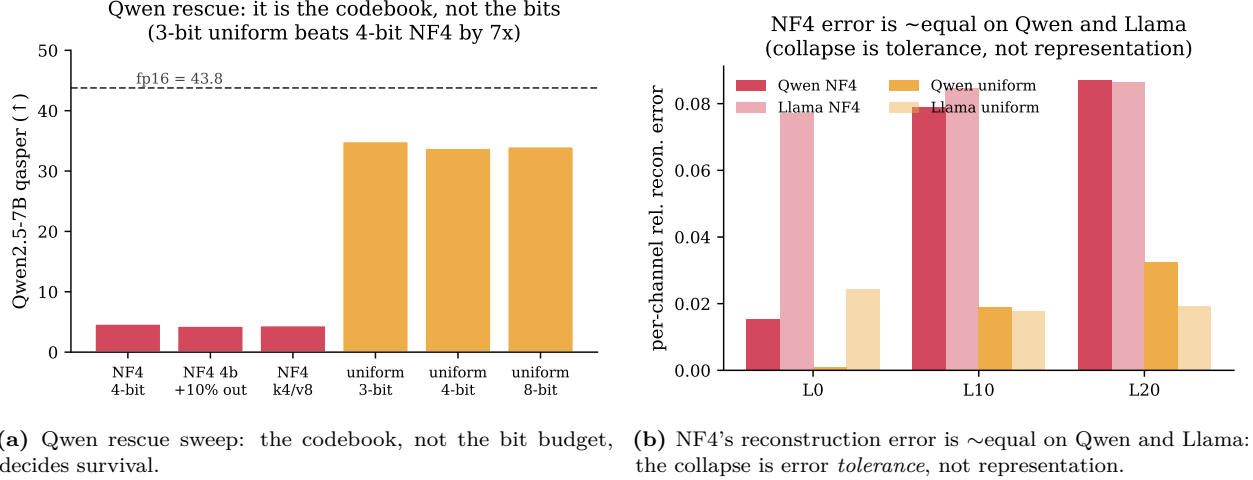
It is the codebook shape, not the bit-depth. Figure 3a isolates the cause on Qwen. Holding the model fixed, 4-bit NF4 scores 4.7; 4-bit *uniform* scores 33.8; and 3-bit *uniform* scores 34.9—lower precision, $7\times$ better. Adding 10% fp16 outliers, shortening the context, or raising values to 8-bit all leave NF4 collapsed. Only changing the codebook helps. Whatever is wrong is not a shortage of bits.

Factor 1: representation. On real pre-RoPE keys, NF4 incurs $\sim 3\text{--}5\times$ the relative reconstruction error (whole-key Frobenius) of asymmetric uniform, because KV keys carry a per-group DC offset (range/abs-max ≈ 0.9) and NF4’s zero-centred grid spends about half of its codes on the empty side (Figure 4). Section 6 makes both halves of that sentence exact: a range/abs-max ratio below 1 *entails* one-sided data, and the dead fraction of the grid span is exactly $(2 - \rho)/2 = 0.55$ at the measured $\rho = 0.9$.

Factor 2: tolerance—and it is the whole difference. The representation cost above is measured on *both* Qwen and Llama, and it is *equal* across them (Figure 3b). This is the pivot of the paper: if the two models see the same key error and only one collapses, no property of the codebook can explain the collapse. The differentiator is what the model does with that error. Qwen2.5-7B has 28 query heads and 4 KV heads, so each KV-head perturbation is read by seven query heads; Llama-2-7B (1:1 MHA) routes it into one. Empirically Qwen collapses above a key-relerr threshold between 0.04 (uniform-3b) and 0.08 (NF4-4b), while Llama’s threshold exceeds 0.08.

Table 2: WikiText-2 perplexity ($\downarrow$), same recipe and protocol as Table 1.

model	fp16	NF4	asym-NF4
Llama-2-7B	6.94	7.18	6.97
Llama-2-13B	6.11	6.30	6.13
Mistral-7B	5.94	6.00	5.96
Qwen2.5-7B	7.46	74.66	7.50

**Figure 3:** Isolating the two factors behind the Qwen collapse.

Equal error, unequal outcome. Stated generally: the damage a quantizer does is not a property of its reconstruction error alone. It is that error read through the consumer that will use it, and two models with the same nominal error can sit on opposite sides of a usability cliff. The practical consequence for evaluation is immediate and, we think, the most transferable claim in this paper: a reconstruction-error budget calibrated on one architecture does not transfer to another, and a scalar error target is not a sufficient acceptance test for a KV codebook. A companion preregistered study on the same model reaches the same conclusion from the opposite direction—by *equalising* a scalar uncertainty statistic across two structurally different KV-eviction scorers and finding that decode quality still differed by 0.359 ± 0.068 (5.3 standard errors), the equality prediction it had registered in advance being refuted (Bond, 2026). Two independent routes, one lesson: match the error *structure*, not its magnitude.

6 The geometry, machine-checked

Two claims in Section 5 are geometric rather than empirical, and we state them as theorems proved in Lean 4 (de Moura & Ullrich, 2021) against Mathlib (The mathlib Community, 2020). The development is `paper/kv_tmlr/lean/AsymNF4.lean`: 12 theorems, no axioms beyond Mathlib, no `sorry`. The statements hold for *any* 16-level grid—they concern where a grid is *placed*, not how its levels are spaced—so they apply to NF4, to uniform, and to any other codebook.

Write a channel’s data range as $[a, b]$ with $b = \text{absmax} > 0$, and let $\rho := (b - a)/b$ denote the measured range-to-abs-max ratio.

Theorem 1 (The diagnostic; `range_div_absmax_lt_one`). *If $0 < a \leq b$ then $\rho < 1$; if the data are symmetric about zero ($a = -b$) then $\rho = 2$. Hence a measured $\rho < 1$ entails one-sided data.*

Theorem 2 (Wasted codes; `dead_code`). *Let every datum satisfy $x \geq a$ and let $g_1 < g_2 \leq a$ be grid points. Then $|x - g_2| < |x - g_1|$ for every such x . No code strictly below the data minimum is ever emitted.*

Theorem 3 (Dead span; `dead_span_fraction`). *For one-sided data on $[a, b]$, a symmetric grid spanning $[-b, b]$ has a dead fraction $(2 - \rho)/2$ of its span. At the measured $\rho = 0.9$ this is 0.55.*

Theorem 4 (Optimality of the zero-point; `absmax_sub_ge_halfRange`). *For any offset c , $\max(|a - c|, |b - c|) \geq (b - a)/2$, with equality at $c = (a + b)/2$. No offset achieves a smaller scale than the half-range.*

Theorem 5 (Error ratio; `error_ratio`). *With a common normalized rounding bound δ , the symmetric quantizer (scale b) and the asymmetric one (scale $(b - a)/2$) have worst-case error bounds in ratio $2b/(b - a) = 2/\rho$, which exceeds 2 for every strictly one-sided channel and equals 2.22 at $\rho = 0.9$.*

Theorem 1 licenses reading $\rho \approx 0.9$ as evidence of offset rather than as a heuristic. Theorems 2 and 3 make “wastes half its codes” exact. Theorem 4 shows the zero-point is *optimal*, not merely helpful—no cleverer offset exists. Theorem 5 accounts for a $2.22\times$ error factor from placement alone.

What this does and does not settle. The measured ratio is $3\text{--}5\times$ (Figure 3b), and the proved bound is $2.22\times$; the remainder is attributable to NF4’s Gaussian level spacing applied to residuals that are not Gaussian, which these theorems deliberately do not model. Nothing here bears on Factor 2: the GQA amplification is measured, not proved, and we make no formal claim about it. We report the decomposition because separating the proved part from the measured part is more useful than a single unexplained ratio.

7 Where the offset lives: RoPE-frequency structure

The per-channel offset that asym-NF4’s zero-point absorbs is not an arbitrary per-channel statistic—it is concentrated in the *slow rotary channels* of RoPE (Su et al., 2024). Measuring post-RoPE keys over WikiText-2 windows (512 tokens, 8 passages, all layers and KV heads; `benchmarks/rope_offset_frequency.py`), the per-channel mean magnitude $|\mu_c|$ correlates strongly with the channel’s rotary wavelength $2\pi \theta^{2(c \bmod d/2)/d}$: pooled Spearman 0.91/0.96/0.94 on Qwen2.5-1.5B/7B/14B (per-layer median 0.74–0.82), with 99%/99%/98% of the total offset mass in the channels whose wavelength exceeds the context window (67% of channels at $\theta=10^6$). The effect crosses model families: Mistral-7B ($\theta=10^4$) gives Spearman 0.98 (per-layer median 0.95), with 96% of the mass in 52% of channels.

The mechanism is geometric. A rotary channel whose wavelength exceeds the window is position-near-invariant within it, so any pre-RoPE per-channel mean survives rotation as a DC component; fast channels average their mean away over positions. The offset is therefore a structural property of RoPE at the model’s θ and window, not an accident of a particular checkpoint—which is why it appears at every scale of the Qwen family and, with a different base, in Mistral.

A zero-point with no calibration at all. This structure yields something stronger than a cheap fix: a zero-point computable from weights and configuration alone. Pushing the model’s own `k_proj` bias through the position-averaged rotation gives a per-channel offset requiring no calibration data and no stored per-channel metadata. It recovers WikiText-2 perplexity 12.533/9.179 on Qwen2.5-1.5B/7B versus 12.534/9.155 for calibrated asym-NF4 (fp16 12.234/9.057; symmetric NF4 30.96/241.3)—a dead heat with the calibrated zero-point at zero calibration cost.¹ Restricting calibrated zero-points to the config-identified long-wavelength channels alone matches dense calibration (12.484/9.150) with one-third less metadata.

The result holds at the task level. On LongBench-qasper (Qwen2.5-7B, full 200-sample split, one harness), the bias-derived zero-point scores **43.36** and the config-sparse zero-point 42.66, against 42.35 for calibrated asym-NF4 in the same run (fp16 43.77; symmetric NF4 4.69). Both calibration-lean variants match or exceed dense calibration on the collapse-sensitive task, though at margins (1.0 and 0.3 points) within this harness’s run-to-run variation; we claim parity, not superiority. The bias route applies to models with key-projection biases (the Qwen family); bias-free models such as Mistral carry their offsets in hidden-state statistics and keep the calibrated or sparse variants.

¹Absolute values here use the zero-point study’s protocol and are not comparable to Table 2; see the protocol note in Section 3.

Why NF4 collapses on DC-offset KV keys

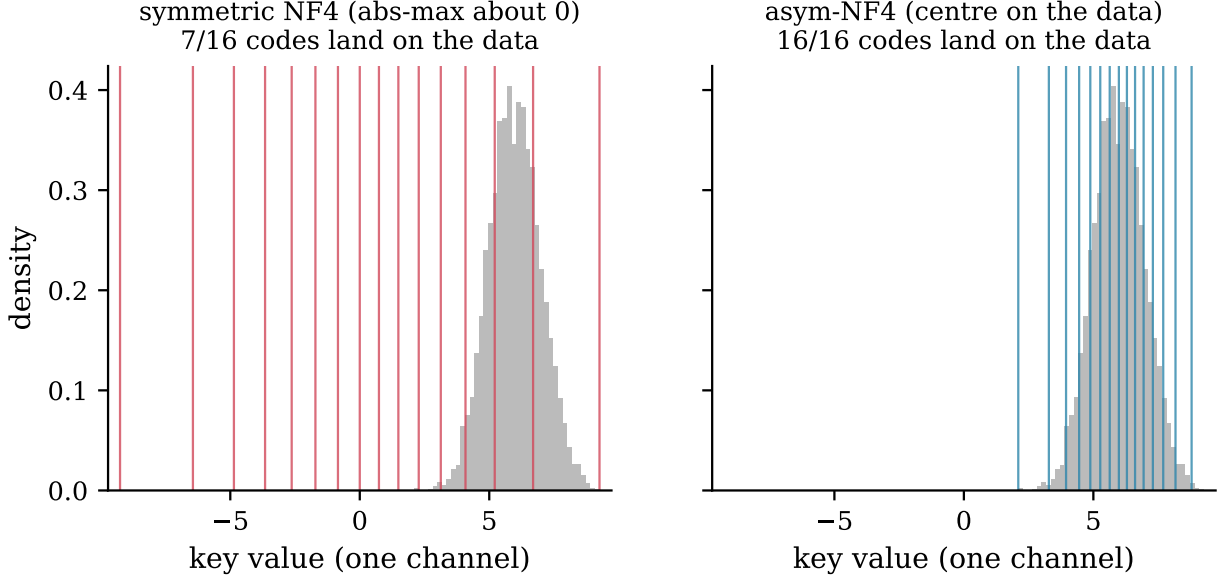


Figure 4: On a DC-offset key channel, symmetric NF4 (left) places most of its 16 codes where there is no data; asym-NF4 (right) centres the grid so the codes cover the data.

8 Asymmetric NF4

asym-NF4 adds a per-channel zero-point. For each channel we subtract the mean μ , NF4-quantize the residual scaled by its abs-max, and add μ back at decode:

$$\hat{x} = \mu + \text{absmax}(x - \mu) \cdot \text{NF4}\left(\frac{x - \mu}{\text{absmax}(x - \mu)}\right),$$

where $\text{NF4}(\cdot)$ denotes nearest-level rounding onto the 16-level NormalFloat grid on $[-1, 1]$ followed by de-quantization. By Theorem 4 the mean is within a bounded factor of the optimal offset and the midpoint attains the optimum exactly; we use the mean because it is what a single streaming pass already computes.

Cost. This stores one extra scalar per channel-group beyond NF4’s abs-max. Every experiment in this paper uses the harness’s configuration of group size 32 with fp16 metadata, so symmetric NF4 costs $4 + 16/32 = 4.5$ bits per key element and asym-NF4 costs $4 + 32/32 = 5.0$: compression of the quantized key region falls from $3.56\times$ to $3.20\times$ against fp16 ($7.11\times$ to $6.40\times$ against fp32), an **11% relative overhead**. The overhead halves at group 64 and halves again at 128; the fp16 outlier channels and the recent-token window are common to all arms and excluded from these ratios. Section 7 removes the stored zero-point altogether for models with key-projection biases, where it is derived from weights.

asym-NF4 keeps NF4’s nonlinear levels (so it beats uniform on the models where level shape matters) while centring the grid on the data (so it does not collapse where placement matters). Figure 4 shows the intuition; Section 6 proves it.

Breadth. Table 3 extends the Qwen2.5-7B comparison to seven further LongBench tasks (single- and multi-document QA, multi-hop, summarization, dialogue) spanning generation lengths from 32 to 512 tokens. asym-NF4 is within 1 point of fp16 on average (mean gap 0.8; $41.2 \rightarrow 40.4$), with no task losing more than 1.9 points as recorded (3.75 on 2wikimqa when re-measured at $n=40$), and the two long-generation tasks show *no* residual gap. On the one long-generation cell where symmetric NF4 was re-measured under the same conditions it scores 5.19 against fp16 31.83—the collapse of Section 4, deepening under long decoding.

Table 3: Qwen2.5-7B, seven further LongBench tasks ($\uparrow$), fp16 vs asym-NF4. Provenance: the five short-generation rows are the recorded full-split values ($n=200$; `results_longgen.json`); the two 512-token rows are the $n=40$ re-validation values (`REVAL-2026-08-08`) that replace the retracted record, whose fp16 arm matched the full-split fp16 to within 0.3 on both tasks. The 2wikimqa gap re-measured at 3.75 ($n=40$) against the recorded 1.9. Means are over the seven rows as shown.

task	max_gen	fp16	asym-NF4	gap
2wikimqa	32	47.1	45.2	1.9
hotpotqa	32	57.7	56.3	1.4
multifieldqa_en	64	52.6	52.0	0.6
narrativeqa	128	28.8	28.1	0.7
samsum	128	46.0	45.1	0.9
gov_report ($n=40$)	512	31.83	32.14	-0.31
multi_news ($n=40$)	512	24.24	23.73	0.51
mean		41.18	40.37	0.81

The table’s provenance is deliberately mixed and is stated in the caption: the five short-generation cells are the recorded full-split values, and the two 512-token cells are the $n=40$ re-validation values that replaced the retracted record (Section 10). An earlier version of this paragraph quoted a 7-task aggregate of 37.3 for asym-NF4 built on the retracted cells; that number, and the accompanying claim of a long-generation limitation, are withdrawn. A single end-to-end re-run of all seven tasks under the harness’s config-sidecar guard is the clean replacement and had not been completed at the time of writing.

9 Recommendations

Use asym-NF4 by default; in the released library this is `TurboQuantKVCache.robust()`. Symmetric NF4 should not be shipped for unknown or high-GQA models. Asymmetric uniform is a safe but lower-quality fallback. Pre- versus post-RoPE quantization is a model-specific micro-optimization (it helps Llama-2-7B but hurts Llama-2-13B and Mistral) and is not a recommended default.

More generally, and independent of our codebook: **do not accept a KV quantizer on reconstruction error measured on a different architecture**. Section 5 shows the same error is benign at 1:1 and fatal at 7:1. Acceptance testing should run on the deployment architecture, and should include at least one task sensitive to degenerate generation—the Qwen collapse is invisible in a spot check of short answers.

10 Limitations

Four models ≤ 13 B; LongBench English and WikiText-2; no MoE or >13 B models. The GQA-ratio-tolerance hypothesis is tested at four points (1:1, 1:1, 4:1, 7:1), all ≤ 13 B and none above 7:1; the mechanism is consistent with the data, but the ratio is not independently varied within a fixed architecture, which would be the decisive experiment and is future work.

Same-harness competitor baselines are not yet faithful. Our in-harness KVQuant (Hooper et al., 2024) implementation (offline Fisher-weighted per-channel NUQ, pre-RoPE) runs end to end but reaches only **8.16** gasper on Llama-2-7B at 4 bits—far short of the published ~ 21 —so it is not a credible reproduction. The KVQuant and KIVI (Liu et al., 2024) comparisons here therefore rest on those papers’ published figures rather than a controlled same-harness run, and a faithful in-harness calibrated comparison remains future work.

A retracted result, reported. An earlier version of this work reported a long-generation degradation of asym-NF4 (LongBench `gov_report` ROUGE falling by 13.7 points at 512 generated tokens, with a monotone sweep across generation lengths). Independent re-validation using the same harness and LongBench’s own metrics (`REVAL-2026-08-08`, $n=40$) measures a gap of -0.31 on that cell; the `multi_news` and Llama-2 rows fail to reproduce identically. The most probable cause is arm-label contamination during off-repo aggregation

of back-to-back `nf4/nf4a` sweeps. **The claim is withdrawn**, and Table 3 reports the corrected cells. A real and larger long-generation collapse (26.64 on the same cell) exists under *symmetric* NF4—that is, it is the collapse of Section 4 deepening under long decoding, not a codebook-independent limitation. The harness now rejects unknown codebook identifiers, writes a per-shard config sidecar, and refuses single-arm aggregation when sidecars disagree.

11 Reproducibility

All results come from a single harness with JSON-recorded numbers and a notebook offering a single-GPU subset and a multi-GPU full mode (`benchmarks/kvquant_matrix/`). Every reported arm carries a config sidecar naming the codebook actually executed, and the aggregator refuses to report an arm whose sidecars disagree. The Lean development of Section 6 builds against Mathlib with no `sorry`.

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. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2023. arXiv:2305.13245.
- Yushi Bai, Xin Lv, Jiajie Zhang, Hongchang Lyu, Jiankai Tang, Zhidian Huang, Zhengxiao Du, Xiao Liu, Aohan Zeng, Lei Hou, Yuxiao Dong, Jie Tang, and Juanzi Li. LongBench: A bilingual, multitask benchmark for long context understanding. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2024. arXiv:2308.14508.
- Andrew H. Bond. Consumer-relative access width on a serving stack: preregistration GO-P-2026-077, 2026. Preregistered study, sealed 2026-08-12; governed run seed 20260812, Qwen2.5-7B-Instruct, $n = 64$. Registry: <https://github.com/ahb-sjsu/geometric-observation>.
- Leonardo de Moura and Sebastian Ullrich. The Lean 4 theorem prover and programming language, 2021. Conference on Automated Deduction (CADE).
- Tim Dettmers, Mike Lewis, Younes Belkada, and Luke Zettlemoyer. LLM.int8(): 8-bit matrix multiplication for transformers at scale. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. arXiv:2208.07339.
- Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. QLoRA: Efficient finetuning of quantized LLMs. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. arXiv:2305.14314.
- Elias Frantar, Saleh Ashkboos, Torsten Hoefer, and Dan Alistarh. GPTQ: Accurate post-training quantization for generative pre-trained transformers. In *International Conference on Learning Representations (ICLR)*, 2023. arXiv:2210.17323.
- 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.
- Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, et al. Mistral 7B. *arXiv preprint arXiv:2310.06825*, 2023.
- 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. In *ACM Symposium on Operating Systems Principles (SOSP)*, 2023. arXiv:2309.06180.
- Ji Lin, Jiaming Tang, Haotian Tang, Shang Yang, Xingyu Dang, Chuang Gan, and Song Han. AWQ: Activation-aware weight quantization for on-device LLM compression and acceleration. In *Conference on Machine Learning and Systems (MLSys)*, 2024. arXiv:2306.00978.

- 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. In *International Conference on Machine Learning (ICML)*, 2024. arXiv:2402.02750.
- Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture models. In *International Conference on Learning Representations (ICLR)*, 2017. arXiv:1609.07843.
- Qwen Team. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2025.
- Noam Shazeer. Fast transformer decoding: One write-head is all you need. *arXiv preprint arXiv:1911.02150*, 2019.
- Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. RoFormer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568, 2024. arXiv:2104.09864.
- The mathlib Community. The Lean mathematical library, 2020. Certified Programs and Proofs (CPP).
- Hugo Touvron, Louis Martin, Kevin Stone, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.
- Guangxuan Xiao, Ji Lin, Mickael Seznec, Hao Wu, Julien Demouth, and Song Han. SmoothQuant: Accurate and efficient post-training quantization for large language models. In *International Conference on Machine Learning (ICML)*, 2023. arXiv:2211.10438.
- Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient streaming language models with attention sinks. In *International Conference on Learning Representations (ICLR)*, 2024. arXiv:2309.17453.
- 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. H2O: Heavy-hitter oracle for efficient generative inference of large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. arXiv:2306.14048.