

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

Andrew H. Bond

June 30, 2026

Abstract

Quantizing the key-value (KV) cache is the standard route to long-context LLM inference, and the data-free NormalFloat-4 (NF4) codebook is a popular calibration-free choice. We show that **NF4 silently and catastrophically fails on an entire class of models**—those with high-ratio grouped-query attention (GQA)—while appearing excellent on the Llama-family multi-head-attention (MHA) models that dominate published benchmarks. On Qwen2.5-7B (7:1 GQA), 4-bit NF4 KV-cache 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 trace the failure to NF4’s *zero-centred, abs-max* scaling: KV keys carry a large per-channel DC offset, so a symmetric codebook wastes half of its codes on the empty side, and high-GQA models—which route each KV-head error into many query heads—cannot absorb the resulting error. The fix is a single per-channel zero-point: **asymmetric NF4 (asym-NF4)** centres the codebook on the data before applying the NF grid. asym-NF4 is one calibration-free codebook that is near-fp16 on *every* architecture we test—it ties symmetric NF4 where NF4 already works and recovers the Qwen collapse to **41.9 qasper / 7.50 ppl**—at no extra bit cost. We release the method, a reproducible cross-model harness, and a practitioner’s decision guide.

1 Introduction

The KV cache dominates the memory footprint of long-context inference, and 4-bit quantization of keys and values is now standard practice. Calibration-free codebooks are attractive because they require no data pass: the 4-bit NormalFloat (NF4) grid, popularised by QLoRA, allocates levels by a unit Gaussian and is a common default for KV keys. Most KV-quantization studies benchmark on the Llama family. We show that this hides a *silent, catastrophic* failure.

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 (Fig. 1, Fig. 2).
2. A measured mechanism: NF4 costs $\sim 4\times$ the relative reconstruction error of asymmetric uniform on DC-offset KV keys, and high-GQA error amplification turns that error into collapse. Bit-depth is *not* the axis—3-bit *uniform* beats 4-bit NF4 on Qwen (Fig. 3, Fig. 4).
3. **asym-NF4**: a one-line per-channel zero-point that unifies the codebook choice—best-or-tied on every architecture, calibration-free, no extra bits (Fig. 5).

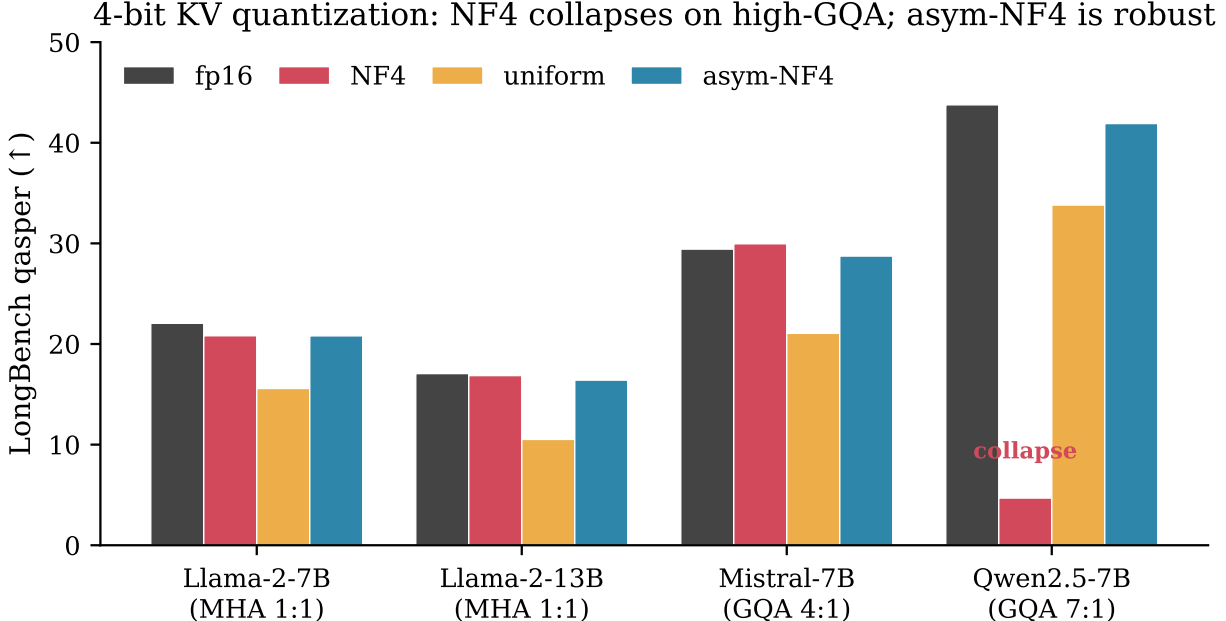


Figure 1: 4-bit KV quantization on the LongBench gasper task. NF4 (red) is competitive on the MHA / low-GQA models but **collapses** on Qwen2.5-7B (7:1 GQA). asym-NF4 (blue) is near-fp16 on every architecture.

4. A reproducible single harness, a cross-model matrix, and a practitioner decision guide.

2 Background and setup

We quantize keys per channel and values per token, with optional dense-and-sparse fp16 outliers (top 2% magnitude per channel) and an attention sink, in a deployable prefill-once cache that runs at $\sim$ fp16 speed. We compare four calibration-free key codebooks 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), Mistral-7B (GQA 4:1), Qwen2.5-7B (GQA 7:1). We evaluate on the LongBench English subset and WikiText-2 perplexity. All scores are produced by a single harness and are only compared intra-harness.

3 NF4 fails on high-GQA models

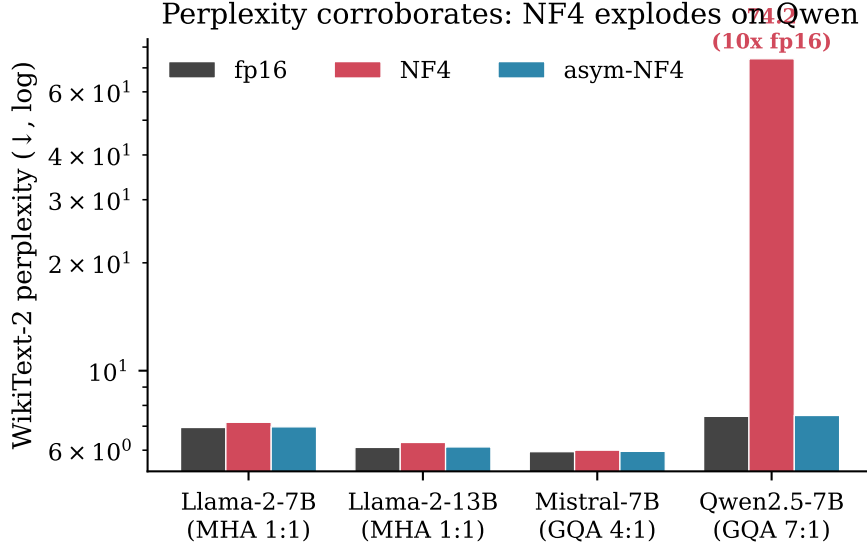
Figure 1 and Table 1 give the headline. NF4 is the best naive codebook on Llama-2-7B/13B and Mistral, but on Qwen2.5-7B it **collapses** from 43.8 (fp16) to 4.7—near-random, degenerate repetition. Asymmetric uniform never collapses but sacrifices 5–9 gasper points on the MHA models. asym-NF4 is best-or-tied everywhere. Perplexity tells the identical story (Fig. 2, Table 2): on Qwen, symmetric NF4 perplexity explodes $10\times$ ($7.46 \rightarrow 74.7$) while asym-NF4 holds at 7.50.

4 Anatomy of the collapse

It is the codebook shape, not the bit-depth. Figure 3 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—

Table 1: LongBench qasper ($\uparrow$), 4-bit keys, identical outliers/sink.

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

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.

Mechanism. On real pre-RoPE keys, NF4 incurs $\sim 3\text{--}5\times$ the relative reconstruction error (whole-key Frobenius) of asymmetric uniform—on *both* Qwen and Llama (Fig. 4)—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 (Fig. 5). Crucially, NF4’s key error is *equal* across the two models, so representation alone does not explain the collapse. The differentiator is **error tolerance**: Qwen’s 7:1 GQA routes each KV-head error into seven query heads, whereas Llama-2-7B (1:1 MHA) routes it into one. Qwen collapses above a key-relerr threshold between 0.04 (uniform-3b) and 0.08 (NF4-4b); Llama’s threshold exceeds 0.08.

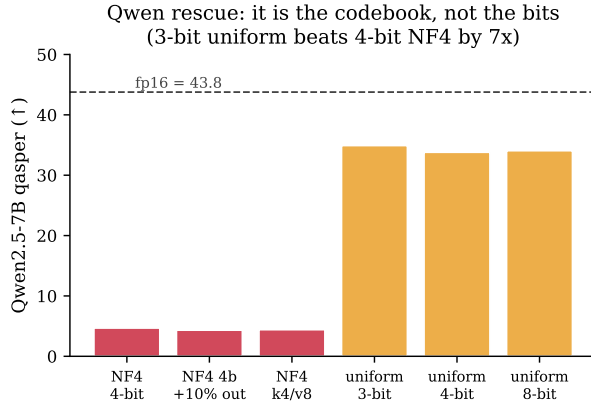
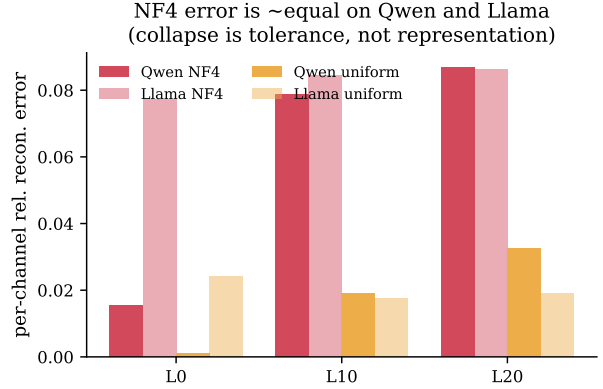
5 Asymmetric NF4

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:

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

Table 2: WikiText-2 perplexity ($\downarrow$), same recipe.

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:** Qwen rescue sweep. The codebook—not the bit budget—decides survival.**Figure 4:** NF4’s reconstruction error is $\sim$ equal on Qwen and Llama; the collapse is error *tolerance*, not representation.

This stores one extra fp32 scalar per channel (μ) beyond NF4’s abs-max—negligible (4-bit compression ratio $7.9\times$, identical to NF4). It keeps NF4’s nonlinear levels (so it *beats* uniform) while centring the grid on the data (so it *does not collapse*). Figure 5 shows the intuition. Tables 1–2 show asym-NF4 is best-or-tied on every model.

Breadth. Across seven further LongBench tasks (single/multi-doc QA, multi-hop, summarization, dialogue), asym-NF4 rescues Qwen’s NF4 collapse on *every* task: the 7-task average rises from **5.3** (NF4) to **37.3** (asym-NF4) against 41.1 (fp16). It is near-fp16 on the QA, multi-hop, and dialogue tasks; the residual gap to fp16 concentrates entirely on two tasks, which the next section shows is a generation-length effect, not a task-type one.

6 A general limitation: long-generation degradation

asym-NF4 is still 4-bit, so a small per-channel key error remains. On Qwen its per-task gap to fp16 is negligible on six of the seven breadth tasks but large on `gov_report` and `multi_news`. These are *not* “the summarization tasks”: `samsun` is also summarization and is fine (gap 0.9). They are the two tasks with `max_new_tokens`=512; every other task generates ≤ 128 . The gap tracks **generation length**, not task type (Table 3).

Isolating generation length. The cleanest test holds the task and model fixed and varies only `max_new_tokens`. On Qwen2.5-7B `gov_report` ($n = 200$), sweeping `max_gen` $\in \{64, 128, 256\}$ (Table 4), asym-NF4 is near-lossless at 64 tokens (1.7% relative) but the gap to fp16 widens *monotonically* to 20% at 128 and 35% at 256. fp16 keeps converting the longer budget into higher

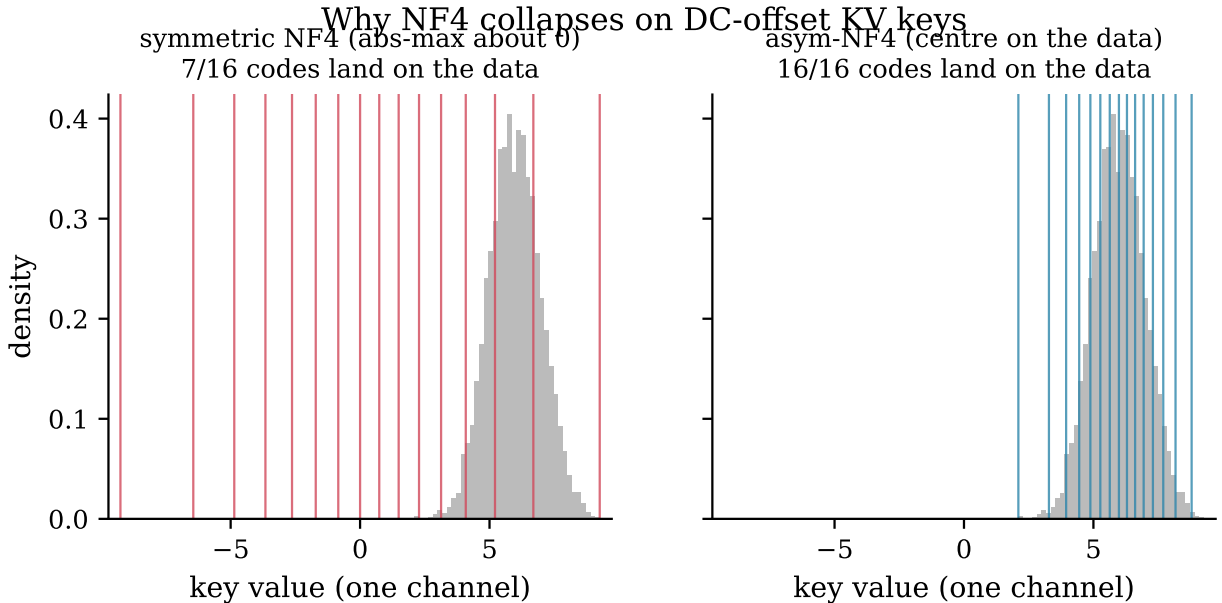


Figure 5: On a DC-offset key channel, symmetric NF4 (left) places most of its 16 codes where there is no data; asym-NF4 (right) centres the grid, so every code is used.

Table 3: Qwen2.5-7B: asym-NF4’s gap to fp16 tracks generation length, not task type.

task	max_gen	fp16	asym-NF4 (gap)
2wikimqa / hotpotqa	32	47.1 / 57.7	45.2 / 56.3 (1.4–1.9)
multifieldqa_en	64	52.6	52.0 (0.6)
narrativeqa	128	28.8	28.1 (0.7)
samsum (<i>summ.</i>)	128	46.0	45.1 (0.9)
gov_report	512	31.8	18.1 (13.7)
multi_news	512	24.0	16.5 (7.5)

ROUGE (14.8→27.3) while asym-NF4 saturates (14.6→17.7) and plateaus toward the 512-token point in Table 3. Same task, same model—only the decode length changes—so the degradation is unambiguously a generation-length effect, not a task-type artifact.

It is not GQA-specific. The same effect appears on Llama-2-7B—the error-*tolerant* MHA model that shrugs off quantization on every short task: `gov_report` drops from 26.8 (fp16) to 14.6 (asym-NF4), a **12.1** gap matching Qwen’s 13.7. So long-generation degradation is a *general* property of 4-bit KV quantization under long autoregressive decoding—distinct from the GQA-specific codebook collapse, and not fixable by the codebook.

Mechanism: compounding. Each decode step reads the quantized prefill with a small residual error; over hundreds of steps, with the model conditioning on its own increasingly drifted output, the error compounds. A `gov_report` generation makes this concrete: the asym-NF4 summary opens faithfully (“*This report examines the use of Multiyear Procurement...*”) and **degenerates into token soup by the tail** (“*for ((, real ((((the...*”), whereas fp16 stays coherent throughout—and the degenerate output fails to emit EOS, running to the full length. The remedy is orthogonal

Table 4: Controlled `max_new_tokens` sweep, single task (Qwen2.5-7B, `gov_report`, $n = 200$, ROUGE $\uparrow$). With task and model fixed, asym-NF4’s gap to fp16 grows monotonically with generation length: fp16 keeps gaining from the longer budget while asym-NF4 saturates as its 4-bit residual error compounds over the decode.

<code>max_gen</code>	fp16	asym-NF4	gap (rel.)	
64	14.81	14.56	0.25	(1.7%)
128	20.57	16.38	4.19	(20%)
256	27.31	17.71	9.60	(35%)

to the codebook (a larger fp16 window or ≥ 5 bits for long-output workloads), which we leave to future work.

7 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- vs. 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.

8 Limitations

Four models ≤ 13 B; LongBench English and WikiText-2; no MoE or > 13 B models. *Same-harness competitor baselines are not yet faithful*. Our in-harness KVQuant implementation (offline Fisher-weighted per-channel NUQ, pre-RoPE) runs end-to-end but reaches only **8.16** qasper on Llama-2-7B at 4 bits—far short of the published ~ 21 —so it is not a credible reproduction; the KVQuant/KIVI comparison here therefore relies on the vendors’ *published* figures rather than a controlled same-harness run, and a faithful in-harness calibrated comparison remains future work. The GQA-ratio-tolerance hypothesis is tested on four points. The long-generation degradation (§6) is characterized on `gov_report/multi_news` and confirmed by a controlled `max_new_tokens` sweep on a single task (Table 4); a remedy (larger fp16 window or ≥ 5 -bit values for long-output workloads) is left to future work.

9 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/`).