

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

Andrew H. Bond

June 28, 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 per-channel 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).

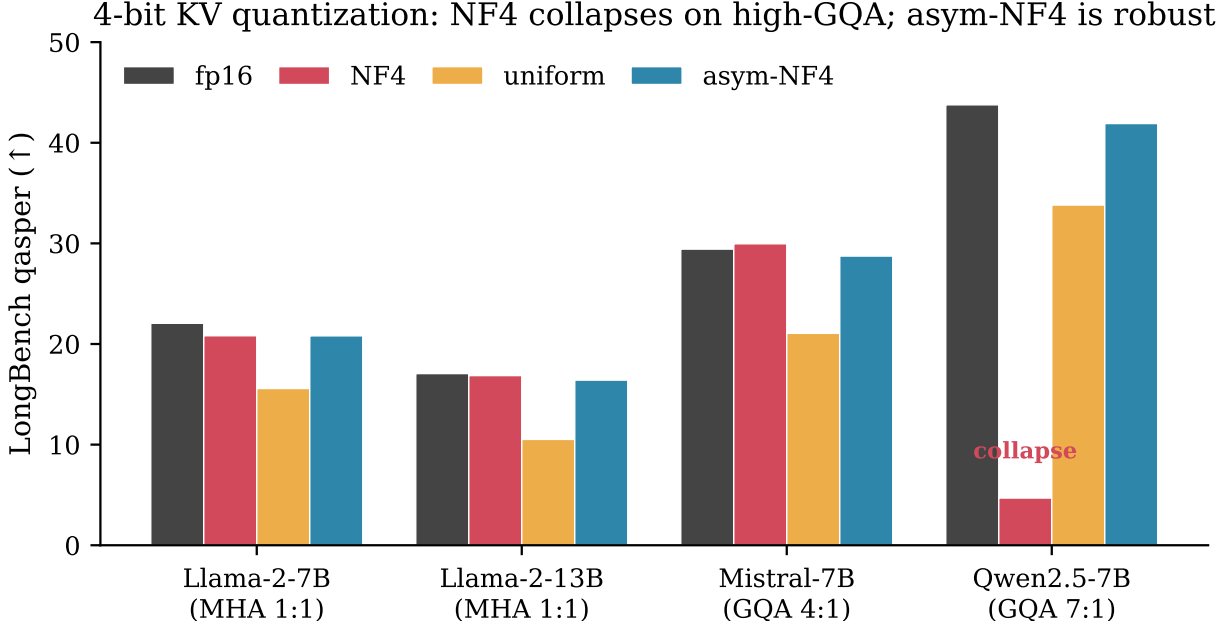


Figure 1: 4-bit KV quantization on the LongBench gasper task. NF4 (red) is competitive on the MHA / low-GQA models but **collapses** on Qwen2.5-7B (7:1 GQA). asym-NF4 (blue) is near-fp16 on every architecture.

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

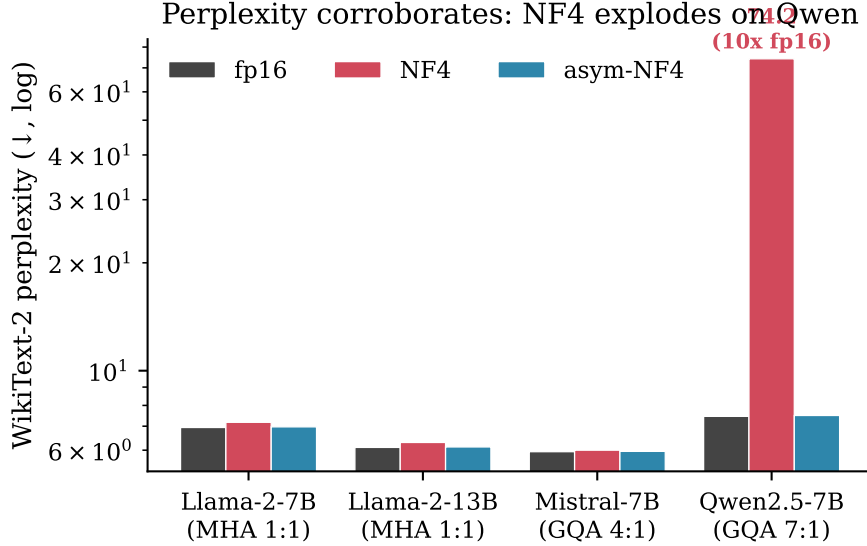
Figure 1 and Table 1 give the headline. NF4 is the best naive codebook on Llama-2-7B/13B and Mistral, but on Qwen2.5-7B it **collapses** from 43.8 (fp16) to 4.7—near-random, degenerate repetition. Asymmetric uniform never collapses but sacrifices 5–9 gasper points on the MHA models. asym-NF4 is best-or-tied everywhere. Perplexity tells the identical story (Fig. 2, Table 2): on Qwen, symmetric NF4 perplexity explodes $10\times$ ($7.46 \rightarrow 74.2$) 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	16.41
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 per-channel relative reconstruction error 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.

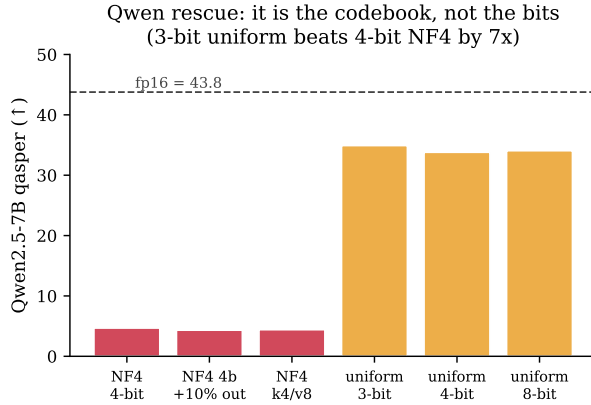
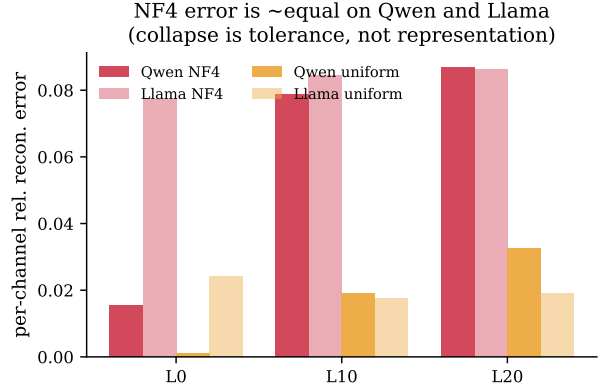
5 Asymmetric NF4

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	74.19	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 fp16 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 tasks; summarization is the one soft spot, where asym-NF4 trails fp16 more but remains $\sim 3\times$ above NF4’s collapse. (Full per-task breadth in the released results; Llama/Mistral breadth in progress.)

6 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.

7 Limitations

Four models $\leq 13B$; LongBench English and WikiText-2; no MoE or $> 13B$ models. Calibrated KVQuant/KIVI are compared in-harness on Llama-2-7B only; a broader calibrated comparison is future work. The GQA-ratio-tolerance hypothesis is tested on four points.

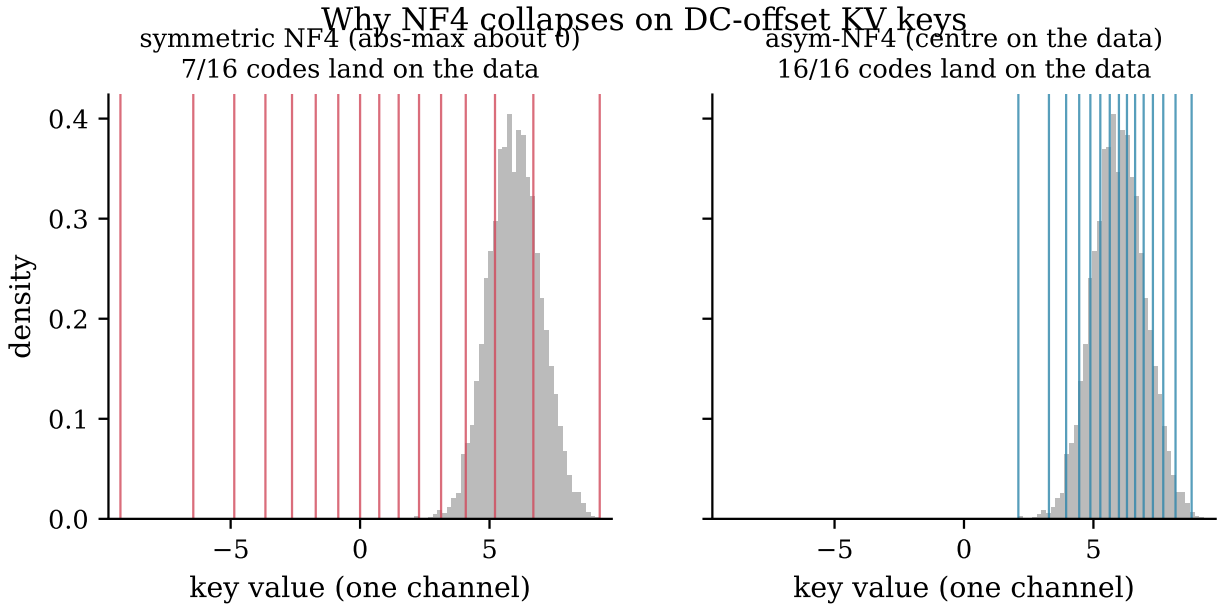


Figure 5: On a DC-offset key channel, symmetric NF4 (left) places most of its 16 codes where there is no data; asym-NF4 (right) centres the grid, so every code is used.

8 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/`).