ONE TOKEN OWNS THE CALIBRATION HESSIAN: WHY LAYER-WISE QUANTIZATION CAN SILENTLY BREAK INSTRUCTION FOLLOWING

Anonymous authors
Paper under double-blind review

ABSTRACT

Quantized instruction-tuned models are usually validated with perplexity and multiple-choice accuracy, which can miss a collapse of instruction following. At 3 bits, reference GPTQ with act-order, calibrated on short documents as widely used libraries sample them by default, lowers the IFEval score of Llama-3.1-8B-Instruct from 0.768 to 0.155 and that of Qwen2.5-14B-Instruct from 0.820 to 0.426, below round-to-nearest (0.565, 0.697), while MMLU stays in the ordinary 3-bit range; GPTQModel's own column order avoids the collapse. We trace the failure to a blind spot of the layer-wise objective. At the down-projection that forms the attention sink, the first token owns the calibration Hessian, yet the loss is almost insensitive to its output. GPTQ's compensation, which we derive in closed form, therefore protects that token and writes its rounding error along the sink pattern into one matrix: a lesion that the chat-template tokens read out. Excising the lesion restores both models. The blind spot is present in all twenty-six models whose first token dominates a down-projection, and of the seventeen of these that we quantized, fourteen carry the lesion, yet only two collapse. Transplants indicate that the other matrices of a healthy network cancel the lesion, which would explain why the failure is rare and depends on how the calibration text is cut and which documents are drawn. Weighting each calibration token by its output-side sensitivity divided by its input energy removes the blind spot and the four collapses it was run on, at a cost of at most 5.6 points on healthy models.

1 INTRODUCTION

Weight-only post-training quantization is a standard way to deploy instruction-tuned language models, and GPTQ (Frantar et al., 2023) is among its most widely used methods. It rounds each linear layer column by column and compensates every rounding error on the remaining columns, using the curvature of the layer's reconstruction error on a small calibration set. The procedure is validated by perplexity and multiple-choice accuracy, and Fig. 2 shows what these metrics do not see. Quantized to 3 bits by reference GPTQ on the short-document calibration that widely used libraries sample by default, Llama-3.1-8B-Instruct keeps an MMLU score above its round-to-nearest (RTN) counterpart and a perplexity within 1.5 times RTN's, yet stops following instructions. Its IFEval score falls from 0.768 to 0.155, while RTN, which uses no calibration data, retains 0.565. Qwen2.5-14B-Instruct behaves the same way, falling from 0.820 to 0.426 against 0.697 for RTN.

Three properties explain why this failure has gone unreported. It is silent, which we attribute to the damage sitting on chat-template tokens that perplexity and multiple-choice prompts never contain. It is rare, affecting two of the 18 instruction-tuned models we tested under the short-document protocol. It also depends on how the calibration text is cut: the same model, code and sample count give 0.645 with 2,048-token windows in place of short documents. The layer-wise objective weights each calibration token by its squared activation and treats an output error as equally harmful at every position (Fig. 1). At the layer that forms the attention sink both assumptions fail at once, leaving the objective a blind spot: the beginning-of-sequence token owns the curvature, and the loss is nearly insensitive to that token's output while depending, on chat prompts, on the template tokens that attend to it. Separating that token, each compensation step has an exact closed form, and under its dominance the compensation turns this blind spot into what we call the lesion: an error

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.

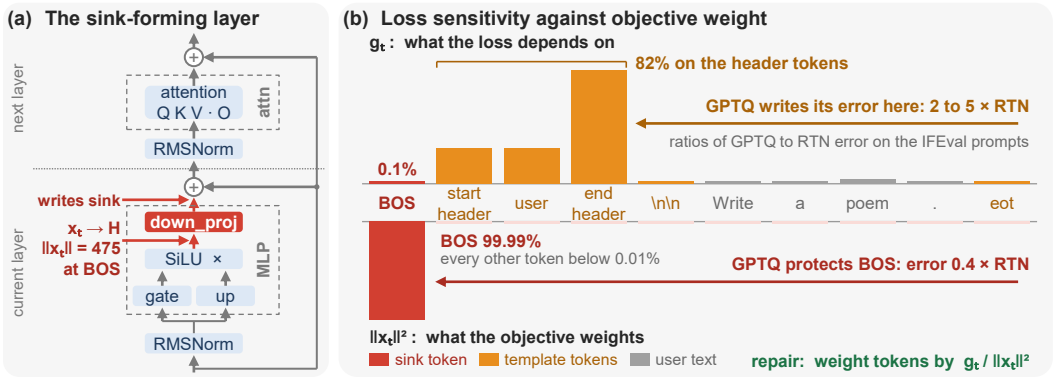


Figure 1: The mechanism, Llama-3.1-8B. **(a)** The sink-forming down-projection of layer 1: its calibration Hessian $\boldsymbol{H}$ is built from its inputs $\boldsymbol{x}_t$ alone, whose norm is 475 at BOS and about 1 elsewhere, and its output writes the sink. **(b)** An illustrative chat exchange at this matrix: the loss sensitivity $g_t = \|\partial L/\partial \boldsymbol{y}_t\|^2$ sits on the header tokens, the input energy $\|\boldsymbol{x}_t\|^2$, the only weight the objective uses, on BOS. On IFEval prompts the compensation protects BOS and raises the error at the header tokens (Section 5.3); weighting each token by $g_t/\|\boldsymbol{x}_t\|^2$ removes it (Section 6).

written into one matrix along the sink pattern and read out by the template tokens. The lesion can be excised without extra bits and transplanted between networks. It is necessary for the collapse but not sufficient: on Qwen it collapses a round-to-nearest network alone, on Llama it needs a weaker copy of itself one layer earlier, which only the short-document protocol makes BOS-heavy, and a healthy draw of Llama writes an equally fatal pair, which the other matrices of its first two layers cancel. Fourteen of seventeen sink models carry the lesion, so we attribute rarity and the dependence on the calibration draw to such cancellation. The direct remedy, weighting each token by its output-side sensitivity over its input energy, repairs the four collapses it was run on.

Our contributions are threefold. We describe a silent, protocol-dependent failure of layer-wise quantization, localize it to one or two matrices, and show that the lesion behind it is latent: healthy networks carry it too, and in Llama the other matrices of the first two layers cancel it (Sections 3 and 4). We derive the lesion from the objective in closed form: its direction and overshoot are confirmed in most sink models, its row concentration and its readout only in part, and its premise holds in all 26 models whose first token dominates a down-projection (Sections 5 and 7). We prove that a position-independent output factor cannot change the column-wise update while a token weight can, and we evaluate the corrected objective this motivates, with its costs (Section 6). The theory covers the formation of the lesion, not its cancellation. From the diagnosis on, most experiments were preceded by written predictions (refuted ones: Appendix C).

2 Background

The layer-wise objective. For a linear layer $\boldsymbol{W} \in \mathbb{R}^{d_{\text{out}} \times d_{\text{in}}}$ with calibration inputs $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T$, post-training quantization chooses $\hat{\boldsymbol{W}}$ to minimize

$$\operatorname{tr}\left[(\hat{\boldsymbol{W}} - \boldsymbol{W})\, \boldsymbol{H}\, (\hat{\boldsymbol{W}} - \boldsymbol{W})^\top\right], \qquad \boldsymbol{H} = \sum_{t=1}^{T} \boldsymbol{x}_t \boldsymbol{x}_t^\top. \tag{1}$$

Following GPTQ we call $\boldsymbol{H}$ the calibration Hessian. It is the Hessian of the reconstruction error, not of the loss, and Section 5 turns on this distinction: Optimal Brain Damage and Optimal Brain Surgeon (LeCun et al., 1990; Hassibi & Stork, 1993) used the loss Hessian, whereas OBC and GPTQ (Frantar & Alistarh, 2022; Frantar et al., 2023) replaced it by the input covariance. GPTQ rounds one column at a time and compensates each rounding error δ_j on the columns not yet rounded by $\Delta w_k = -\delta_j [\boldsymbol{H}^{-1}]_{jk} / [\boldsymbol{H}^{-1}]_{jj}$, after damping $\boldsymbol{H}$ by $\lambda \boldsymbol{I}$, $\lambda = \rho \operatorname{mean} \operatorname{diag} \boldsymbol{H}$, and, with act-order, sorting the columns by $\operatorname{diag} \boldsymbol{H}$. Round-to-nearest (RTN) skips the compensation. Later work refines the update (Li et al., 2025; Chen et al., 2026a; Kim et al., 2026), rotates outliers away or reconditions

the curvature first (Chee et al., 2023; Ashkboos et al., 2024; Zhang et al., 2026), learns rounding or clipping on a token-averaged layer or block loss (Nagel et al., 2020; Shao et al., 2024; Cheng et al., 2024), keeps sensitive weights at higher precision (Dettmers et al., 2024; Kim et al., 2024; Xiao et al., 2025) or rescales channels (Lin et al., 2024). Those that fit a reconstruction keep the two assumptions of Eq. (1): each token enters $\boldsymbol{H}$ with weight $\|\boldsymbol{x}_t\|^2$, and an output error is equally harmful at every position. Methods that reintroduce the loss weight the reconstruction by squared output gradients, per output in BRECQ and per token and channel group in GuidedQuant (Li et al., 2021; Kim et al., 2025), or use an output-channel curvature (Lee et al., 2026a) or a token saliency (Xue et al., 2025; Sung et al., 2025). By Proposition 1(b) a gradient weight alone can leave the sink token dominant, and a saliency read from attention concentration, as in RSQ, ranks it first.

Attention sinks and calibration. Decoder-only models route much of their attention to the first token, whose residual stream grows in the first layers to a norm hundreds of times larger than elsewhere (Xiao et al., 2024; Sun et al., 2024; An et al., 2025). This sink emerges with training (Gu et al., 2025), prevents over-mixing (Barbero et al., 2025), is written by a few weights of one early down-projection (Yu et al., 2024) and has a nearly constant direction (Chen et al., 2026c); because RMSNorm divides by the norm, it also passes almost no gradient back to itself (Chen et al., 2026b). Sink and spike tokens also complicate activation and KV-cache quantization (Bondarenko et al., 2023; Yang et al., 2024; Liu et al., 2024; Son et al., 2024). For a quantizer, one calibration token therefore contributes almost all of $\boldsymbol{H}$ at the sink-forming layer and almost nothing to the loss. GPTQ's sensitivity to its calibration set (Williams & Aletras, 2024), chat-formatted calibration (vLLM Project, 2025) and length normalization of $\boldsymbol{H}$ (Son et al., 2026) are known. Empirical studies find GPTQ well ahead of RTN at 3 bits (Huang et al., 2024; Zhou et al., 2026b), while likelihood metrics miss what compression does to task accuracy (Jaiswal et al., 2024; Lee et al., 2025), alignment (Xu et al., 2026), generation (Lee et al., 2026b) and early-layer computation at 2 bits (Zhou et al., 2026a). The failure studied here occurs at 3 bits, sits in one matrix although quantization damage is usually spread across layers (Hu & Ramachandran, 2026), and has its cause in the objective.

3 A SILENT, PROTOCOL-DEPENDENT FAILURE

Setup. The main experiments use Llama-3.1-8B-Instruct (Grattafiori et al., 2024), Qwen2.5-14B-Instruct (Qwen Team, 2025) and Mistral-7B-Instruct-v0.3 (Jiang et al., 2023); the census of Section 7 covers 18 instruction-tuned models from 1.7B to 32B. Weights are quantized to 3 or 4 bits (symmetric, group size 128, act-order, $\rho = 0.05$) and kept on the integer grid in 16-bit storage; Table 2 separates library defaults from our choices. GPTQModel's own default loop, which orders the columns group by group, collapses neither model, whereas act-order or the natural order inside it collapses Llama (Table 1). The failure is thus that of the reference loop under the libraries' sampling. The *short-document* protocol tokenizes 128 C4 documents (Raffel et al., 2020) separately (mean 420 tokens), as GPTQModel (ModelCloud, 2025) and LLM Compressor (vLLM Project, 2025) do by default. The *window* protocol takes the first 2,048 tokens of 128 documents at least that long, as AutoRound does, from C4 (c4win) or the Pile (Gao et al., 2021). IFEval (Zhou et al., 2023) is the mean of prompt- and instruction-level strict and loose accuracy over 541 prompts, with greedy decoding. A different calibration draw moves a non-collapsing arm by up to 4 points, which we treat as noise; an arm counts as collapsed when it falls more than 20 points below RTN; the nearest arms fall 11.7 and 21.5 points below. Multi-IF (He et al., 2024), GSM8K (Cobbe et al., 2021), MMLU (Hendrycks et al., 2021) and WikiText-2 perplexity (Merity et al., 2017) are described in Appendix B. On Llama and Mistral the chat template and the tokenizer each write a `BOS`, so every generation prompt carries two; a single-`BOS` control moves no conclusion (Appendix B).

Collapse that likelihood does not show. Round-to-nearest, the reference throughout, is GPTQ without the compensation step; the two also group the columns differently under act-order, but not under the natural order, which collapses Llama too (0.128), so the gap is the compensation's doing. Under the short-document protocol, 3-bit GPTQ leaves Llama at 0.155 and Qwen at 0.426, 41 and 27 points below RTN (Fig. 2, Table 5). The collapsed Llama answers most prompts with unformatted or truncated text, so GSM8K falls to 0.008 (RTN 0.280) and the loss persists over every turn of Multi-IF. Yet the same checkpoints have an MMLU above their RTN counterparts (0.524 and 0.739 against 0.416 and 0.698), and Qwen's perplexity is better than RTN's (7.66 against 9.26). Llama's perplexity (22.4 against 15.5) is the one likelihood signal that moves, and it moves far less than the

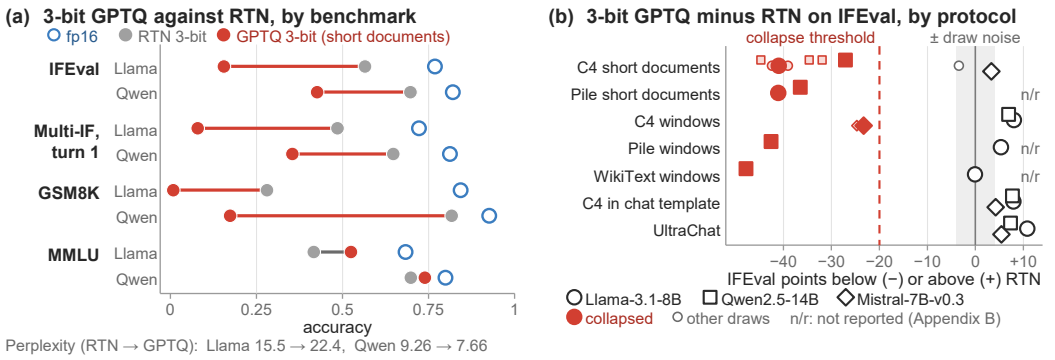


Figure 2: The collapse and its silence. **(a)** 3-bit GPTQ calibrated on short documents against RTN on the same grid, Llama-3.1-8B and Qwen2.5-14B: the compensation step loses 27 to 64 points on every generation benchmark and gains on MMLU. **(b)** IFEval of 3-bit GPTQ minus RTN under seven calibration protocols; red markers fall more than 20 points below RTN (collapse), the grey band is the 4-point draw noise, and small markers are the other calibration draws.

generation. We attribute the miss to where the damage sits, on the chat-template tokens that steer generation (Section 4), which neither WikiText nor a few-shot MMLU prompt contains.

Dependence on how the calibration text is cut. With 2,048-token windows in place of short documents, Llama recovers to 0.645 and Qwen to 0.766. Chat-formatted text, either the same documents in the chat template or UltraChat (Ding et al., 2023), also removes the collapse (Fig. 2b, Table 8). The cause is not sample length alone: Qwen collapses again under Pile and WikiText windows, and Mistral-7B-v0.3 collapses under C4 windows (0.187, RTN 0.420) but not under short documents. Nor is it the corpus, since 128 Pile documents tokenized separately collapse Llama and Qwen as C4 does, or the token count, since 1,024 short documents, more tokens than 128 windows, leave Qwen at 0.250 (Table 2). On Llama the effect of the protocol is visible in the calibration Hessians before any weight is rounded. The `BOS` token holds 0.266 of the trace at the layer-0 down-projection under short documents, against 0.069 under C4 windows, 0.063 under Pile windows and 0.073 under WikiText, so a second matrix becomes `BOS`-heavy exactly under the protocol that collapses (Section 4). Both sampling shapes are library defaults, and the collapse is reproducible across draws with one exception: Llama collapses in four of five draws (0.141 to 0.174; the fifth gives 0.531), Qwen in four of four, Mistral under C4 windows in both (Table 17). This exception shows that collapse depends on the calibration sample as well as on the model (Section 4).

4 The Lesion

Before deriving a mechanism we locate the damage among the 224 quantized matrices, excise it and transplant it (Tables 10 and 11); we call its concentrated part the *lesion*.

Where it is. Comparing hidden states of the collapsed and the full-precision Llama singles out one position and one depth: at the output of layer 1 the relative error at position 4, the token `<|end_header_id|>` that closes the chat header, is 3.59 against 0.58 elsewhere. For Qwen it is the first newline of the template at the output of layer 4 (3.92). These are the depths at which the attention sink forms and positions that appear to act as secondary sinks, and the matrix that writes the sink state is the MLP down-projection of the sink-forming layer, layer 1 for Llama and layer 4 for Qwen. Re-rounding that matrix without compensation removes what the compensation wrote: with the other 223 matrices untouched, Llama rises from 0.155 to 0.622 and Qwen from 0.426 to 0.768. Keeping the attention projections of the first three layers (0.138) or the five largest rows of the matrix (0.174) at full precision does not help, so the damage lies in what the compensation wrote, not in attention or in the super weights themselves.

Necessary, not sufficient. Copied into the full-precision Llama the lesion costs almost nothing (0.764) and into the RTN network four points (0.528); Qwen's costs the full-precision Qwen 18 points (0.637), short of collapse. One matrix does not collapse a full-precision model on its own. Inside a GPTQ-quantized network it does, and not only in its own: copied into the healthy window-calibrated network (0.645) it collapses it (0.174). The window checkpoint's matrix, which its own network tolerates, does not cure the collapsed short-document network (0.152), whereas the RTN version of the same matrix does (0.622).

What else it needs. Not the layers quantized after it: rounding every matrix from layer k onward to nearest, for k from 2 to 24, leaves the collapse intact (0.13 to 0.18), whereas the same lesion inside an all-RTN network scores 0.528. Transplanting subsets of the first two layers into the RTN network (Table 15; single draws) isolates the one thing it needs, the down-projection of layer 0. The two down-projections together collapse the RTN network (0.139), as does every tested set containing them, including all fourteen matrices of the two layers (0.152). Every tested set without the layer-0 down-projection is healthy, among them the whole of layer 1 with the lesion inside it (0.572). This pair is Llama's minimal collapsing set. Each matrix is necessary, since removing either cures (0.595, 0.622) and neither alone does anything (0.543, 0.528 against 0.565). The layer-0 down-projection is a weaker copy of the same object (`BOS` share of its Hessian trace 0.266 against 0.998 at layer 1 and at most 0.006 elsewhere; overshoot and row concentration in Appendix D). It is `BOS`-heavy only under short documents, where a `BOS` token arrives every few hundred tokens (Section 3). Swapping in the window versions of both matrices cures the collapse (0.580); swapping only the layer-0 one does not (0.160), since a foreign GPTQ matrix is itself fatal. The set is model-dependent. On Qwen the sink-forming down-projection alone collapses the RTN network (0.320, below the fully quantized network's 0.426), the second matrix adds nothing, and excising both cures (0.763). On Mistral the window collapse needs the fourteen matrices of layers 0 and 1 (0.192) but neither the down-projections (0.423) nor layer 1 alone (0.417), a case we localize to two layers but not to a matrix.

Cancellation by the neighboring matrices. The calibration draw that leaves Llama healthy (0.531) writes a pair as fatal to the RTN network as the collapsing draw's (0.158 against 0.139). Its fourteen matrices of layers 0 and 1 together, however, score 0.525, its own value, where the collapsing draw's score 0.152. Tulu-3, the same base model under the same protocol, is healthy at 0.726 while its pair collapses its RTN network to 0.117 (RTN 0.637), and Qwen's window-calibrated sink matrix, from a protocol the model survives, collapses the RTN network to 0.208. The lesion, with its layer-0 copy on Llama, thus collapses each of these RTN networks, and, we infer, a GPTQ network survives it only when its other matrices cancel it. On Llama these are the twelve other matrices of the first two layers, since rounding everything later to nearest changes nothing; none of them sees the layer-1 lesion during quantization (Appendix B), so none adapts to it. On Qwen they are not yet localized. Protocol and draw thus act at different points: short documents decide whether Llama's layer-0 copy is `BOS`-heavy, and the draw decides whether its neighbors cancel the pair.

5 Theory: From Objective Mismatch to Error Misallocation

GPTQ and RTN differ above all in how they allocate rounding error: RTN leaves it where it occurs, GPTQ moves it onto the free columns to reduce the layer-wise objective. At the sink-forming down-projection, the objective's curvature is concentrated at `BOS`, to whose output the loss is nearly insensitive. GPTQ therefore suppresses the error at `BOS` by shifting it onto other positions, including those the loss depends on: an *objective mismatch* becomes an *error misallocation*.

Setup. Consider one linear layer $\boldsymbol{y}_t = \boldsymbol{W}\boldsymbol{x}_t$ with $\boldsymbol{W} \in \mathbb{R}^{d_{\text{out}} \times d_{\text{in}}}$, where $t = 0$ denotes the first position (`BOS` where present) and $\boldsymbol{x}_0$ its effective activation accumulated over the calibration samples. GPTQ quantizes the input coordinates sequentially using the damped input-side curvature $\boldsymbol{H}_\lambda = \boldsymbol{H} + \lambda \boldsymbol{I}$, $\boldsymbol{H} = \sum_t \boldsymbol{x}_t \boldsymbol{x}_t^\top$, $\lambda = \rho \operatorname{tr} \boldsymbol{H}/d_{\text{in}}$. We index the coordinates in quantization order, so that act-order is a permutation applied once before the loop; at step j the free set is the suffix $F_j = \{j, \dots, d_{\text{in}}\}$, and we write $\boldsymbol{x}_{t,j} = \boldsymbol{x}_t[F_j]$ and, since damping is applied once to the full curvature, $\boldsymbol{H}_{\lambda,j} = \boldsymbol{H}_\lambda[F_j, F_j] = \boldsymbol{H}[F_j, F_j] + \lambda \boldsymbol{I} \succ 0$. Column j is rounded from $\boldsymbol{w}_j = \boldsymbol{W}_{:,j}$ to

$\boldsymbol{q}_j$, leaving the residual $\delta_j = \boldsymbol{w}_j - \boldsymbol{q}_j \in \mathbb{R}^{d_{\text{out}}}$, and every free column $k > j$ is updated by

$$\Delta \boldsymbol{w}_k = -\,\delta_j \,\frac{[\boldsymbol{H}_{\lambda,j}^{-1}]_{jk}}{[\boldsymbol{H}_{\lambda,j}^{-1}]_{jj}}. \tag{2}$$

The step changes no column outside F_j, so it perturbs the output at position t by $\Delta \boldsymbol{y}_t^{(j)} = \Delta \boldsymbol{W}^{(j)}[:, F_j]\, \boldsymbol{x}_{t,j}$, and the next step rounds an already-modified column with the free set F_{j+1}. RTN rounds without Eq. (2), on groups of the original column order.

5.1 Objective mismatch: reconstruction protects the wrong positions

The layer-wise reconstruction objective measures how much a perturbation changes the output of the current layer, but not how costly that change is to the rest of the network. To make the missing output side explicit, let L be the loss of the network on the calibration set. Expanding L to second order in $\Delta \boldsymbol{W}$, and neglecting the first-order term as in OBS and cross-position blocks, gives, up to the factor $\frac{1}{2}$ of the expansion, the position-diagonal quadratic term

$$\mathcal{E}_{\text{down}}(\Delta \boldsymbol{W}) = \sum_t (\Delta \boldsymbol{W} \boldsymbol{x}_t)^\top \boldsymbol{G}_t (\Delta \boldsymbol{W} \boldsymbol{x}_t), \qquad \boldsymbol{G}_t = \frac{\partial^2 L}{\partial \boldsymbol{y}_t \, \partial \boldsymbol{y}_t^\top}. \tag{3}$$

Standard layer-wise reconstruction instead minimizes Eq. (1), $\mathcal{E}_{\text{layer}}(\Delta \boldsymbol{W}) = \sum_t \|\Delta \boldsymbol{W} \boldsymbol{x}_t\|^2 = \operatorname{tr}[\Delta \boldsymbol{W} \boldsymbol{H} \Delta \boldsymbol{W}^\top]$. Relative to Eq. (3), this objective replaces the output metric $\boldsymbol{G}_t$ by the same identity at every position, $\mathcal{E}_{\text{layer}} = \mathcal{E}_{\text{down}}$ with $\boldsymbol{G}_t \equiv \boldsymbol{I}$. It therefore retains the input-side geometry $\boldsymbol{x}_t \boldsymbol{x}_t^\top$ but ignores how strongly an output error at position t affects the loss. The substitution loses nothing for GPTQ when $\boldsymbol{G}_t$ is the same at every position (Proposition 1(a)), but can change where the error is placed when $\boldsymbol{G}_t$ varies across positions. We therefore compare the positional distributions of input-side curvature and downstream sensitivity through $a_t = \|\boldsymbol{x}_t\|^2$ and $g_t = \|\partial L / \partial \boldsymbol{y}_t\|^2 = \operatorname{tr}[(\partial L / \partial \boldsymbol{y}_t)(\partial L / \partial \boldsymbol{y}_t)^\top]$. Here g_t is the trace of the empirical-Fisher block of position t, a measurable proxy for the scale of $\boldsymbol{G}_t$ rather than the block itself; like $\boldsymbol{x}_0$, g_0 is accumulated over the calibration samples.

Observation 1 (Position mismatch)**.** *At the down-projection that the first position dominates most, a share $A_{\texttt{BOS}} = a_0 / \sum_t a_t \geq 0.50$ of the input-side curvature comes with almost none of the measured downstream sensitivity, $G_{\texttt{BOS}} = g_0 / \sum_t g_t \leq 0.01$; at the sink-forming down-projection $A_{\texttt{BOS}} = 0.998$ for Llama and 1.000 for Qwen. On chat prompts the sensitivity concentrates on the early template positions.*

The attention sink thus makes the two distributions nearly opposite (Fig. 3; Appendix D confirms by injection and activation patching that the loss depends on the template positions), so the layer-wise objective spends most of its curvature on a position the loss barely depends on. The mismatch is harmless for RTN, which does not consult the objective, but potentially harmful in GPTQ, which redistributes rounding error to reduce Eq. (1). Equation (3) motivates the weight and is not a quantitative model of the lesion's cost: on Llama the cross-position terms cancel most of its block-diagonal value (Appendix C), so the surrogate is validated by intervention and repair (Appendix D and Section 6), not by a curvature fit.

5.2 Compensation turns objective mismatch into error misallocation

To see how GPTQ improves its objective while raising the error at individual positions, split the free-block curvature of step j into its BOS part and the rest, and define

$$\begin{aligned} \boldsymbol{H}_{\lambda,j} &= \boldsymbol{x}_{0,j} \boldsymbol{x}_{0,j}^\top + \boldsymbol{M}_j, \qquad \boldsymbol{M}_j = \sum_{t \neq 0} \boldsymbol{x}_{t,j} \boldsymbol{x}_{t,j}^\top + \lambda \boldsymbol{I} \succ 0, \\ \boldsymbol{v} &= \boldsymbol{M}_j^{-1} \boldsymbol{x}_{0,j}, \qquad \kappa = \boldsymbol{x}_{0,j}^\top \boldsymbol{M}_j^{-1} \boldsymbol{x}_{0,j}, \qquad m_{kl} = [\boldsymbol{M}_j^{-1}]_{kl}, \end{aligned} \tag{4}$$

with the step index suppressed on $\boldsymbol{v}$, κ and m_{kl} and all indices running over F_j. $\boldsymbol{M}_j$ is the curvature of the ordinary tokens plus damping, and κ the dominance of BOS relative to it. Empirically, $\kappa \gg 1$ at the sink-forming matrices: before the first column is rounded it is 1.2×10^5 at Llama's and exceeds 10^3 in fifteen of the 17 models with a sink (per model in Table 16). It falls along the loop as the sink coordinates leave the free set (Appendix A.1).

Theorem 1 (Error redistribution with a first-token term). *Fix a step j: column j is rounded while the columns $k > j$ remain free. (a) The inverse curvature is exactly $\boldsymbol{H}_{\lambda,j}^{-1} = \boldsymbol{M}_j^{-1} - \boldsymbol{v}\boldsymbol{v}^\top/(1+\kappa)$, and the compensation of column $k > j$ is*

$$\Delta \boldsymbol{w}_k = \delta_j \, \frac{v_j v_k/(1+\kappa) - m_{jk}}{m_{jj} - v_j^2/(1+\kappa)}. \tag{5}$$

(b) Including the rounding of column j, the output perturbation at position t is

$$\Delta \boldsymbol{y}_t = -\,\delta_j \, \frac{[\boldsymbol{M}_j^{-1}\boldsymbol{x}_{t,j}]_j - v_j\, \boldsymbol{v}^\top \boldsymbol{x}_{t,j}/(1+\kappa)}{m_{jj} - v_j^2/(1+\kappa)}, \qquad \Delta \boldsymbol{y}_0 = -\,\delta_j \, \frac{v_j}{(1+\kappa)\, m_{jj} - v_j^2}. \tag{6}$$

(c) In the isotropic case $\boldsymbol{M}_j = \lambda \boldsymbol{I}$, with $\|\boldsymbol{x}_{0,j+1}\|^2 = \|\boldsymbol{x}_{0,j}\|^2 - x_{0j}^2$,

$$\begin{gathered} \Delta \boldsymbol{w}_k = \delta_j \, \frac{x_{0j} x_{0k}}{\lambda + \|\boldsymbol{x}_{0,j+1}\|^2}, \qquad \Delta \boldsymbol{y}_0 = -\,\delta_j x_{0j} \, \frac{\lambda}{\lambda + \|\boldsymbol{x}_{0,j+1}\|^2}, \\ \Delta \boldsymbol{y}_t = -\,\delta_j x_{tj} + \delta_j \, x_{0j} \, \frac{\boldsymbol{x}_{0,j+1}^\top \boldsymbol{x}_{t,j+1}}{\lambda + \|\boldsymbol{x}_{0,j+1}\|^2}. \end{gathered} \tag{7}$$

In the isotropic case of Eq. (7), the first term of $\Delta \boldsymbol{y}_t$, $-\delta_j x_{tj}$, is the rounding error RTN would also leave; the second is what the compensation adds. At BOS the compensation suppresses this error by a factor $1/(1+\kappa_{j+1})$, $\kappa_{j+1} = \|\boldsymbol{x}_{0,j+1}\|^2/\lambda$. When BOS dominates the curvature, GPTQ therefore protects BOS by redistributing its rounding error onto the remaining coordinates; which positions read it out, the theorem does not fix (Section 5.3).

In Eq. (5) the BOS-induced component is rank one along $\boldsymbol{v} = \boldsymbol{M}_j^{-1}\boldsymbol{x}_{0,j}$ and, under act-order, strongest in the early steps, when the sink-carrying coordinates are rounded first.

5.3 Signatures of the lesion

Under the isotropic idealization $\boldsymbol{M}_1 = \lambda \boldsymbol{I}$, Theorem 1 yields three directly testable signatures: where the error is written, which positions read it out, and what sets its amplitude.

Corollary 1 (Signatures of sink-induced redistribution). *Assume $\boldsymbol{M}_1 = \lambda \boldsymbol{I}$, so that $\boldsymbol{M}_j = \lambda \boldsymbol{I}$ at every step, and let J be the leading block of sink-carrying columns. By Eq. (7), step j updates every later column $k > j$ by $\Delta \boldsymbol{w}_k^{(j)} = \gamma_j \, x_{0k} \, \delta_j$ with $\gamma_j = x_{0j}/(\lambda + \|\boldsymbol{x}_{0,j+1}\|^2)$.*

(i) Direction and row concentration. *For $k > \max J$, the accumulated sink-induced update satisfies $\Delta \boldsymbol{W}_{r,k}^{\text{sink}} = c_r x_{0k}$, $c_r = \sum_{j \in J} \gamma_j \delta_j^{(r)}$. Thus each affected row follows the sink pattern $\boldsymbol{x}_0$, while the accumulated rounding residuals c_r determine its magnitude across rows.*

(ii) Readout. *The step j updates the output at position t by $\gamma_j \left(\boldsymbol{x}_{0,j+1}^\top \boldsymbol{x}_{t,j+1}\right) \delta_j$: positions whose remaining activations overlap more with the sink pattern receive more redistributed error.*

(iii) Amplitude. *The coefficient of each update factorizes as $\gamma_j \, x_{0k} = \frac{\kappa_{j+1}}{1+\kappa_{j+1}} \cdot \frac{x_{0j}\, x_{0k}}{\|\boldsymbol{x}_{0,j+1}\|^2}$. The first factor is the share of the BOS rounding error that the step cancels. It is close to one when BOS dominates the remaining curvature and close to zero otherwise. The second factor is at most $|x_{0j}/x_{0k}|$. It can exceed one when $\boldsymbol{x}_0$ is concentrated on few coordinates, as the sink is (Sun et al., 2024). Together, the two factors can make the update larger than the residual it compensates and can push weights beyond their original range.*

What is measured. At Llama's sink matrix the direction of (i) and the overshoot of (iii) hold, its row concentration does not (Appendix D), and across the census most show all three (Section 7). The compensation pattern has cosine 0.52 to 0.93 with $\boldsymbol{x}_0$ in fourteen of seventeen sink matrices, and 23% of the template-position error lies in the super-weight row (Yu et al., 2024). That error is 2.25 for GPTQ against 0.43 for RTN, and 0.4 times RTN's at BOS (Fig. 1); the largest compensated entry is 7.3 where the largest original weight is 0.71. Signature (ii), the readout, fails in part: token by token, the overlap places the template positions on Qwen (recall 0.52) but not on Llama (0.11).

5.4 RESTORING THE OUTPUT SIDE ON THE TOKEN AXIS

The repair is to put the output side of Eq. (3) back where the compensation can see it. A position-independent output-channel factor cannot do this, including K-FAC (Martens & Grosse, 2015), whose curvature for a linear layer is the Kronecker product of $\boldsymbol{H}$ with an output factor averaged over positions.

Proposition 1 (Token-level weights act, channel-level factors do not)**.** *(a) Let $\boldsymbol{G} \succeq 0$ be independent of position. Replacing Eq. (1) by $\mathrm{tr}[\boldsymbol{G}\,\Delta\boldsymbol{W}\boldsymbol{H}\Delta\boldsymbol{W}^\top]$ leaves the column-wise compensation Eq. (2) unchanged; if $\boldsymbol{G}$ is diagonal with positive entries, it leaves the nearest-grid rounding within each row unchanged as well. (b) Let $\widetilde{\boldsymbol{H}} = \sum_t w_t \boldsymbol{x}_t \boldsymbol{x}_t^\top$ with $w_t = g_t/\|\boldsymbol{x}_t\|^2$. Then $\mathrm{tr}\,\widetilde{\boldsymbol{H}} = \sum_t g_t$, so the $\texttt{BOS}$ share of the weighted trace is $G_{\texttt{BOS}} \le 0.01$ rather than $A_{\texttt{BOS}}$, and the trace no longer singles out $\texttt{BOS}$. With $w_t = g_t$ alone the share is $g_0 a_0 / \sum_t g_t a_t$, which is 0.984 for Llama and 0.993 for Qwen: the dominant token survives.*

The weighted objective is $\mathrm{tr}[\Delta\boldsymbol{W}\widetilde{\boldsymbol{H}}\Delta\boldsymbol{W}^\top] = \sum_t g_t \,\|\Delta\boldsymbol{W}\boldsymbol{x}_t/\|\boldsymbol{x}_t\|\|^2$, that is, Eq. (3) with each $\boldsymbol{G}_t$ replaced by $w_t\boldsymbol{I}$. The normalization, a design choice rather than a consequence of Eq. (3), removes the activation-energy dominance that (b) shows g_t alone cannot remove (Appendix A.4). The weighted matrix $\widetilde{\boldsymbol{H}}$ is a sensitivity-aligned surrogate of the downstream Hessian, not that Hessian.

6 THE REPAIR: RESTORING THE OUTPUT SIDE

Each remedy found by trial targets one quantity of Theorem 1 and fails somewhere (Table 4), and the layer-1 signature itself persists in healthy networks (Appendix D). Proposition 1 gives a repair that acts on the objective instead: quantize with

$$\widetilde{\boldsymbol{H}} = \sum_t w_t\, \boldsymbol{x}_t \boldsymbol{x}_t^\top, \qquad w_t = \frac{g_t}{\|\boldsymbol{x}_t\|^2}, \tag{8}$$

with the weights rescaled by a constant per calibration sample (Appendix B.3) and everything else unchanged. The sensitivity g_t is measured once on the full-precision model with one backward pass per calibration sample, and the reweighting is one line in the Hessian accumulation of any quantizer that consumes $\boldsymbol{H}$ (Appendix B.3). Chat-formatted calibration, the practitioners' lever, also prevents the collapse on all three models (Mistral: 0.462 with C4 in the chat template, 0.474 with UltraChat). At 4 bits it recovers Mistral and Hermes-3 but lowers Qwen3-14B, scored in thinking mode (Appendix B), from 0.466 to 0.428, where the corrected objective gives 0.483.

Ablation. Table 6 (Appendix D) shows the two halves of Eq. (8) failing in complementary places. The corrected objective restores Llama to 0.638, Qwen to 0.736 and Mistral to the RTN level (0.433), and token normalization alone repairs Llama and Qwen but not Mistral (0.189). Output weighting alone, GuidedQuant's token weight with one channel group (Kim et al., 2025), repairs Mistral (0.475) but neither Llama (0.329) nor Qwen (0.580), where Proposition 1(b) predicts that the dominant token survives. Dropping the first two positions from $\boldsymbol{H}$ (before the loop correction) lifts Llama from 0.150 only to the RTN level (0.560; Appendix D). A constant weight on the first eight positions of every document, with no gradient, restores normalization alone to the RTN level on Mistral (0.405), so on the three main collapses the output side can be dropped (Llama, Qwen) or replaced by a position prior (Mistral); which half suffices on which model is not derived.

Cost. On instruction following the corrected objective matches re-rounding and heavy damping on Llama, trails them by a few points on Qwen and holds across the turns of Multi-IF (Table 13). It is not free (Table 12). Llama's GSM8K recovers only to 0.548 against 0.632 for re-rounding, and its MMLU falls to 0.514, below the collapsed checkpoint's 0.524; Qwen stays within 1.2 points of re-rounding on GSM8K and MMLU. The corrected objective removes the error the blind spot creates, not the ordinary 3-bit loss, so it is a repair for a model that validation shows collapsed, not a new default. At 4 bits neither model collapses, but across 18 models 4-bit GPTQ does not beat 4-bit RTN on IFEval (mean -0.1, negative on 11/18; per model in Table 16). The corrected objective recovers the losses on Mistral ($+6.3$) and Hermes-3 ($+5.9$) and moves Qwen3-14B within noise ($+1.7$).

7 UNIVERSAL BLIND SPOT, COMMON LESION, RARE COLLAPSE

All statements below but those on the lesion were predicted in writing before the models were run (Tables 14 and 16). The premise was measured on 28 instruction-tuned models; the quantization census covers 18 of them. All 26 models whose first token dominates a down-projection meet its bound at the most dominated one ($A_{\text{BOS}} \geq 0.50$, $G_{\text{BOS}} \leq 0.01$; Fig. 3), and neither gemma-2 model has such a layer (largest A_{BOS} 0.005). In 14 of the 17 census models with a sink, GPTQ's squared error at the worst template position of a first-token-dominated down-projection exceeds RTN's by 3.6 to 471 times; in 14 the largest row carries more than twice its RTN share, and in 16 the largest compensated entry exceeds twice the largest original one (Corollary 1(i), (iii)). The signature persists under GPTQModel's group-aware column order, 100 (Llama) and 7 (Qwen) times smaller in mass (Appendix B). Collapse does not follow: 2 of 18 models collapse under short documents and one of 18 under c4win, models with lesions as large as Llama-3-8B's or Qwen2.5-32B's do not, and the lesion's size neither sets a threshold nor ranks the losses well (Spearman -0.46). In this census the corrected objective never falls more than 4.6 points below RTN on 18 models, beyond noise only on gemma-2-9b, without a sink. It repairs the four 3-bit collapses it was run on (Table 6) and two of the three 4-bit losses, and on the 16 non-collapsing models it changes IFEval by $+0.3$ on average, -0.8 at the median (-5.6 to $+13.9$), losing beyond noise on three.

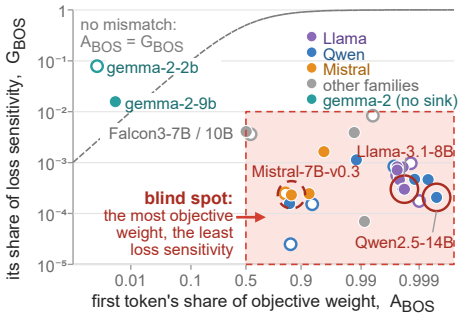


Figure 3: The blind spot across 28 models, at the most BOS-dominated down-projection of each. Rings: collapses; dashed, C4 windows; open, premise only.

Across quantizers. Observation 1 is a property of the objective, so it sorts quantizers by their objective (Table 9). Methods that compensate through the input-side curvature collapse, GPTQ on both models and GPTAQ on Llama, though not GPTQ under GPTQModel's group-aware column order, and rotating first (QuaRot) removes the collapse (0.648 / 0.709). AutoRound, a token-averaged block loss with free rounding offsets and no OBS step, collapses Qwen (0.181) and loses ten points to RTN on Llama (0.460) under short documents, though not under its own Pile windows (0.716 / 0.793). RTN, HQQ (Badri & Shaji, 2023) and AWQ, which fit no reconstruction objective or only per-channel scales, lose nothing. Theorem 1 covers the OBS family, not AutoRound.

8 CONCLUSION

The layer-wise reconstruction objective is not a neutral proxy for the loss. In models with an attention sink it is owned, at one layer, by a token whose output hardly affects the loss, and the compensation step converts that blind spot into an error written along the sink pattern and read out by the template tokens. The error is a matrix that can be excised and transplanted, its direction and overshoot match Corollary 1, it forms in most sink models but, we infer, collapses a network only where the rest fails to cancel it, and reweighting the tokens (Proposition 1) repairs the four collapses it was run on. Pruning with SparseGPT (Frantar & Alistarh, 2023) optimizes the same objective; whether it writes the same lesion is untested. In practice a quantized instruction-tuned model should be validated on free generation and, if it falls below its round-to-nearest counterpart, requantized with chat-formatted text or Eq. (8), or its sink matrix re-rounded to nearest.

Limitations and open questions. The theory is local to one matrix and exact for the steps that write the lesion. The columns rounded afterwards are measured, not derived, so the theory explains the lesion but not why one network or calibration draw cancels it and another does not. There is no cross-model criterion, and Mistral's window collapse is localized to two layers but not to a matrix. The corrected objective lowers Llama's MMLU by a point, recovers GSM8K only partly and costs up to 5.6 points on healthy models. Quantization is simulated, and the census stops at 32B. Thirty-seven advance predictions failed (Appendix C), among them a κ criterion and the collapse of GPTQModel's default loop; why its group-aware order confines the lesion is not derived.

Reproducibility statement

All quantization, evaluation and analysis code, the run index that maps every number in this paper to a checkpoint and a calibration protocol, and the job scripts whose headers state each batch's prediction before it ran are released at https://anonymous.4open.science/r/sink-hessian-ptq-D7BD and as supplementary material. Appendix B gives the calibration sampling of each protocol as verified against the library sources, the quantizer settings, the evaluation settings and the seeds; Appendix A gives the assumptions and proofs behind Section 5.

Ethics statement

This work studies a failure mode of a widely used compression method and proposes a repair. Its most direct consequence is that quantized instruction-tuned models validated only by perplexity may be deployed in a state in which they no longer follow instructions; we consider disclosure of the failure and of its cause to be in the interest of users. No human subjects or private data were involved.

AI use statement

The authors designed the study, wrote the advance predictions, derived the theory and its proofs, ran and checked every experiment and drew the conclusions. Large language model assistants helped the authors draft and edit the text and figures, draft and refactor code, and search the literature; every line of code was run and checked, every number traced to a run in the released index and every reference verified against its source by the authors, who take full responsibility for the content.

References

Yongqi An, Xu Zhao, Tao Yu, Ming Tang, and Jinqiao Wang. Systematic outliers in large language models. In *International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2502.06415.

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. In *Advances in Neural Information Processing Systems*, 2024. URL https://arxiv.org/abs/2404.00456.

Hicham Badri and Appu Shaji. Half-quadratic quantization of large machine learning models. https://github.com/mobiusml/hqq, 2023.

Federico Barbero, Álvaro Arroyo, Xiangming Gu, Christos Perivolaropoulos, Michael Bronstein, Petar Veličković, and Razvan Pascanu. Why do LLMs attend to the first token? *arXiv preprint arXiv:2504.02732*, 2025. URL https://arxiv.org/abs/2504.02732.

Yelysei Bondarenko, Markus Nagel, and Tijmen Blankevoort. Quantizable transformers: Removing outliers by helping attention heads do nothing. In *Advances in Neural Information Processing Systems*, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/edbcb7583fd8921dad78adecfe06a99b-Abstract-Conference.html.

Jerry Chee, Yaohui Cai, Volodymyr Kuleshov, and Christopher De Sa. QuIP: 2-bit quantization of large language models with guarantees. In *Advances in Neural Information Processing Systems*, 2023. URL https://arxiv.org/abs/2307.13304.

Jiale Chen, Yalda Shabanzadeh, Elvir Crnčević, Torsten Hoefler, and Dan Alistarh. The geometry of LLM quantization: GPTQ as Babai's nearest plane algorithm. In *International Conference on Learning Representations*, 2026a. URL https://arxiv.org/abs/2507.18553.

Yihong Chen, Zhouchen Lin, and Quanming Yao. Attention sinks induce gradient sinks: Massive activations as gradient regulators in transformers. *arXiv preprint arXiv:2603.17771*, 2026b. URL https://arxiv.org/abs/2603.17771.

Yung-Chin Chen, Chung Peng Lee, Ze-Wei Liou, and Naveen Verma. Massive spikes in LLMs are bias vectors: Mechanistic uncovering and spike-free quantization. *arXiv preprint arXiv:2606.02288*, 2026c. URL https://arxiv.org/abs/2606.02288.

Wenhua Cheng, Weiwei Zhang, Haihao Shen, Yiyang Cai, Xin He, Kaokao Lv, and Yi Liu. Optimize weight rounding via signed gradient descent for the quantization of LLMs. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, 2024. URL https://arxiv.org/abs/2309.05516.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

Tim Dettmers, Ruslan Svirschevski, Vage Egiazarian, Denis Kuznedelev, Elias Frantar, Saleh Ashkboos, Alexander Borzunov, Torsten Hoefler, and Dan Alistarh. SpQR: A sparse-quantized representation for near-lossless LLM weight compression. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2306.03078.

Ning Ding, Yulin Chen, Bokai Xu, Yujia Qin, Zhi Zheng, Shengding Hu, Zhiyuan Liu, Maosong Sun, and Bowen Zhou. Enhancing chat language models by scaling high-quality instructional conversations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 2023.

Elias Frantar and Dan Alistarh. Optimal brain compression: A framework for accurate post-training quantization and pruning. In *Advances in Neural Information Processing Systems*, 2022.

Elias Frantar and Dan Alistarh. SparseGPT: Massive language models can be accurately pruned in one-shot. In *International Conference on Machine Learning*, 2023. URL https://arxiv.org/abs/2301.00774.

Elias Frantar, Saleh Ashkboos, Torsten Hoefler, and Dan Alistarh. GPTQ: Accurate post-training quantization for generative pre-trained transformers. In *International Conference on Learning Representations*, 2023. URL https://arxiv.org/abs/2210.17323.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, Shawn Presser, and Connor Leahy. The Pile: An 800GB dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*, 2021.

Aaron Grattafiori et al. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Xiangming Gu, Tianyu Pang, Chao Du, Qian Liu, Fengzhuo Zhang, Cunxiao Du, Ye Wang, and Min Lin. When attention sink emerges in language models: An empirical view. In *International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2410.10781.

Babak Hassibi and David G. Stork. Second order derivatives for network pruning: Optimal brain surgeon. In *Advances in Neural Information Processing Systems 5*, 1993.

Yun He, Di Jin, Chaoqi Wang, Chloe Bi, Karishma Mandyam, Hejia Zhang, Chen Zhu, Ning Li, Tengyu Xu, Hongjiang Lv, Shruti Bhosale, Chenguang Zhu, Karthik Abinav Sankararaman, Eryk Helenowski, Melanie Kambadur, Aditya Tayade, Hao Ma, Han Fang, and Sinong Wang. Multi-IF: Benchmarking LLMs on multi-turn and multilingual instructions following. *arXiv preprint arXiv:2410.15553*, 2024. URL https://arxiv.org/abs/2410.15553.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *International Conference on Learning Representations*, 2021.

Jundong Hu and Shekar Ramachandran. The structure of quantization damage in LLMs: Why the next bit should be spent globally. *arXiv preprint arXiv:2609.01587*, 2026. URL https://arxiv.org/abs/2609.01587.

Wei Huang, Xingyu Zheng, Xudong Ma, Haotong Qin, Chengtao Lv, Hong Chen, Jie Luo, Xiaojuan Qi, Xianglong Liu, and Michele Magno. An empirical study of LLaMA3 quantization: From LLMs to MLLMs. *Visual Intelligence*, 2(36), 2024. URL https://arxiv.org/abs/2404.14047.

Ajay Jaiswal, Zhe Gan, Xianzhi Du, Bowen Zhang, Zhangyang Wang, and Yinfei Yang. Compressing LLMs: The truth is rarely pure and never simple. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2310.01382.

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.

Jaemin Kim, Sungkyun Kim, Junyeol Lee, and Jiwon Seo. Robust ultra low-bit post-training quantization via stable diagonal curvature estimate. In *EuroMLSys*, 2026. URL https://arxiv.org/abs/2604.13806.

Jinuk Kim, Marwa El Halabi, Wonpyo Park, Clemens J. S. Schaefer, Deokjae Lee, Yeonhong Park, Jae W. Lee, and Hyun Oh Song. GuidedQuant: Large language model quantization via exploiting end loss guidance. In *International Conference on Machine Learning*, 2025. URL https://arxiv.org/abs/2505.07004.

Sehoon Kim, Coleman Hooper, Amir Gholami, Zhen Dong, Xiuyu Li, Sheng Shen, Michael W. Mahoney, and Kurt Keutzer. SqueezeLLM: Dense-and-sparse quantization. In *International Conference on Machine Learning*, 2024. URL https://arxiv.org/abs/2306.07629.

Yann LeCun, John S. Denker, and Sara A. Solla. Optimal brain damage. In *Advances in Neural Information Processing Systems 2*, 1990.

Donghyun Lee, Yuhang Li, Ruokai Yin, and Priyadarshini Panda. KronQ: LLM quantization via Kronecker-factored Hessian. In *Conference on Language Modeling*, 2026a. URL https://arxiv.org/abs/2607.07964.

Jemin Lee, Sihyeong Park, Jinse Kwon, Jihun Oh, and Yongin Kwon. Exploring the trade-offs: Quantization methods, task difficulty, and model size in large language models from edge to giant. In *International Joint Conference on Artificial Intelligence*, 2025. URL https://arxiv.org/abs/2409.11055.

Jung Hyun Lee, June Yong Yang, Jungwook Choi, and Eunho Yang. LFQ: Logit-aware final-block quantization for boosting the generation quality of low-bit quantized LLMs. In *International Conference on Machine Learning*, 2026b. URL https://arxiv.org/abs/2605.29756.

Yuhang Li, Ruihao Gong, Xu Tan, Yang Yang, Peng Yu, Qi Zhang, Fengwei Yu, Wei Wang, and Shi Gu. BRECQ: Pushing the limit of post-training quantization by block reconstruction. In *International Conference on Learning Representations*, 2021.

Yuhang Li, Ruokai Yin, Donghyun Lee, Shiting Xiao, and Priyadarshini Panda. GPTAQ: Efficient finetuning-free quantization for asymmetric calibration. In *International Conference on Machine Learning*, 2025. URL https://arxiv.org/abs/2504.02692.

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 on-device LLM compression and acceleration. In *Proceedings of Machine Learning and Systems*, 2024. URL https://arxiv.org/abs/2306.00978.

Ruikang Liu, Haoli Bai, Haokun Lin, Yuening Li, Han Gao, Zhengzhuo Xu, Lu Hou, Jun Yao, and Chun Yuan. IntactKV: Improving large language model quantization by keeping pivot tokens intact. In *Findings of the Association for Computational Linguistics: ACL 2024*, 2024. URL https://arxiv.org/abs/2403.01241.

James Martens and Roger Grosse. Optimizing neural networks with Kronecker-factored approximate curvature. In *International Conference on Machine Learning*, 2015.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture models. In *International Conference on Learning Representations*, 2017.

ModelCloud. GPTQModel. https://github.com/ModelCloud/GPTQModel, 2025.

Markus Nagel, Rana Ali Amjad, Mart van Baalen, Christos Louizos, and Tijmen Blankevoort. Up or down? adaptive rounding for post-training quantization. In *International Conference on Machine Learning*, 2020.

Qwen Team. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2025.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140), 2020.

Wenqi Shao, Mengzhao Chen, Zhaoyang Zhang, Peng Xu, Lirui Zhao, Zhiqian Li, Kaipeng Zhang, Peng Gao, Yu Qiao, and Ping Luo. OmniQuant: Omnidirectionally calibrated quantization for large language models. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2308.13137.

Seungwoo Son, Wonpyo Park, Woohyun Han, Kyuyeun Kim, and Jaeho Lee. Prefixing attention sinks can mitigate activation outliers for large language model quantization. In *Conference on Empirical Methods in Natural Language Processing*, 2024. URL https://arxiv.org/abs/2406.12016.

Seungwoo Son, Ingyu Seong, Junhan Kim, Hyemi Jang, and Yongkweon Jeon. On the importance of a multi-scale calibration for quantization. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2026. URL https://arxiv.org/abs/2602.07465.

Mingjie Sun, Xinlei Chen, J. Zico Kolter, and Zhuang Liu. Massive activations in large language models. In *Conference on Language Modeling*, 2024. URL https://arxiv.org/abs/2402.17762.

Yi-Lin Sung, Prateek Yadav, Jialu Li, Jaehong Yoon, and Mohit Bansal. RSQ: Learning from important tokens leads to better quantized LLMs. *arXiv preprint arXiv:2503.01820*, 2025. URL https://arxiv.org/abs/2503.01820.

vLLM Project. LLM Compressor. https://docs.vllm.ai/projects/llm-compressor/en/latest/, 2025.

Miles Williams and Nikolaos Aletras. On the impact of calibration data in post-training quantization and pruning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, 2024. URL https://arxiv.org/abs/2311.09755.

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient streaming language models with attention sinks. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2309.17453.

Hanqi Xiao, Yi-Lin Sung, Elias Stengel-Eskin, and Mohit Bansal. Task-circuit quantization: Leveraging knowledge localization and interpretability for compression. In *Conference on Language Modeling*, 2025. URL https://arxiv.org/abs/2504.07389.

Bruce Changlong Xu, Adarsh Kumarappan, and Mu Zhou. Alignment collapse under KV cache quantization: Diagnosis and mitigation. *arXiv preprint arXiv:2606.09864*, 2026. URL https://arxiv.org/abs/2606.09864.

Yufei Xue, Yushi Huang, Lunjie Zhu, Jiawei Shao, and Jun Zhang. SalQ-VLM: Fine-grained saliency-guided quantization for vision-language models. *arXiv preprint arXiv:2508.03351*, 2025. URL https://arxiv.org/abs/2508.03351.

Jaewoo Yang, Hayun Kim, and Younghoon Kim. Mitigating quantization errors due to activation spikes in GLU-based LLMs. *arXiv preprint arXiv:2405.14428*, 2024. URL https://arxiv.org/abs/2405.14428.

Mengxia Yu, De Wang, Qi Shan, Colorado J. Reed, and Alvin Wan. The super weight in large language models. *arXiv preprint arXiv:2411.07191*, 2024. URL https://arxiv.org/abs/2411.07191.

Jinhao Zhang, Yunquan Zhang, Zicheng Yan, Boyang Zhang, Jun Sun, and Daning Cheng. HeRo-Q: A general framework for stable low bit quantization via Hessian conditioning. *arXiv preprint arXiv:2601.21626*, 2026. URL https://arxiv.org/abs/2601.21626.

Chenxi Zhou, Pengfei Cao, Jiang Li, Bohan Yu, Jinyu Ye, Jun Zhao, and Kang Liu. From signal degradation to computation collapse: Uncovering the two failure modes of LLM quantization. In *Findings of the Association for Computational Linguistics: ACL 2026*, 2026a. URL https://arxiv.org/abs/2604.19884.

Chenxi Zhou, Pengfei Cao, Jinyu Ye, Bohan Yu, Haida Yu, Jiang Li, Jun Zhao, and Kang Liu. Quantization degradation in large language models: A signal-noise perspective. *arXiv preprint arXiv:2608.08188*, 2026b. URL https://arxiv.org/abs/2608.08188.

Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. Instruction-following evaluation for large language models. *arXiv preprint arXiv:2311.07911*, 2023. URL https://arxiv.org/abs/2311.07911.

A PROOFS

Notation is that of Section 5: at step j the free set is $F_j = \{j, \dots, d_{\text{in}}\}$, $\boldsymbol{x}_{t,j} = \boldsymbol{x}_t[F_j]$, $\boldsymbol{H}_{\lambda,j} = \boldsymbol{H}_\lambda[F_j, F_j] = \boldsymbol{x}_{0,j}\boldsymbol{x}_{0,j}^\top + \boldsymbol{M}_j$ with $\boldsymbol{M}_j \succ 0$, $\boldsymbol{v} = \boldsymbol{M}_j^{-1}\boldsymbol{x}_{0,j}$, $\kappa_j = \boldsymbol{x}_{0,j}^\top \boldsymbol{v}$ (written κ in the main text when the step is clear, and $\kappa = \kappa_1$ before the first column is rounded) and $m_{kl} = [\boldsymbol{M}_j^{-1}]_{kl}$. The BOS block of $\boldsymbol{H}$ is exactly $\boldsymbol{x}_0\boldsymbol{x}_0^\top$ on Llama and Mistral: the first token attends only to itself, so its activation is the same vector $\boldsymbol{x}$ in every calibration sample and $\boldsymbol{x}_0 = \sqrt{n}\,\boldsymbol{x}$ for n samples (leading-eigenvalue share 1.000 in the run index). Qwen has no BOS token and $t = 0$ is the first document token; its block has leading-eigenvalue share 0.9995, and the share is at least 0.90 at every sink matrix of the census.

A.1 THE LOOP

The objective $E(\Delta\boldsymbol{W}) = \operatorname{tr}[\Delta\boldsymbol{W}\boldsymbol{H}_\lambda\Delta\boldsymbol{W}^\top]$ separates over rows. If before step j the free coordinates F_j minimize E given the rounded ones, fixing the rounding of column j and re-minimizing over F_{j+1} gives, by the block-inverse identity, exactly Eq. (2) with $\boldsymbol{H}_{\lambda,j} = \boldsymbol{H}_\lambda[F_j, F_j]$, and the new point is again a minimizer; this is the OBQ form of GPTQ (Frantar & Alistarh, 2022; Frantar et al., 2023). Each step therefore sees the principal block $\boldsymbol{H}_{\lambda,j+1} = \boldsymbol{x}_{0,j+1}\boldsymbol{x}_{0,j+1}^\top + \boldsymbol{M}_{j+1}$ with $\boldsymbol{M}_{j+1} = \boldsymbol{M}_j[F_{j+1}, F_{j+1}] \succ 0$, so Theorem 1 applies at every step, with δ_j the residual of column j as the loop reaches it. Since $\kappa_j = \max_{\boldsymbol{u}\in\mathbb{R}^{F_j}}\{2\boldsymbol{u}^\top\boldsymbol{x}_{0,j} - \boldsymbol{u}^\top\boldsymbol{M}_j\boldsymbol{u}\}$ and restricting $\boldsymbol{u}$ to F_{j+1} gives κ_{j+1}, κ only falls along the loop; if $\boldsymbol{M}_j = \lambda\boldsymbol{I}$, then $\boldsymbol{M}_{j+1} = \lambda\boldsymbol{I}$ and $\kappa_{j+1} = \|\boldsymbol{x}_{0,j+1}\|^2/\lambda$.

A.2 PROOF OF THEOREM 1

Proof. *(a)* Since $\boldsymbol{M}_j \succ 0$ and $1 + \kappa > 0$, Sherman–Morrison gives

$$\boldsymbol{H}_{\lambda,j}^{-1} = \left(\boldsymbol{M}_j + \boldsymbol{x}_{0,j}\boldsymbol{x}_{0,j}^\top\right)^{-1} = \boldsymbol{M}_j^{-1} - \frac{\boldsymbol{v}\boldsymbol{v}^\top}{1+\kappa},$$

so that

$$[\boldsymbol{H}_{\lambda,j}^{-1}]_{jk} = m_{jk} - \frac{v_j v_k}{1+\kappa}, \qquad [\boldsymbol{H}_{\lambda,j}^{-1}]_{jj} = m_{jj} - \frac{v_j^2}{1+\kappa} > 0,$$

the last because $\boldsymbol{H}_{\lambda,j} \succ 0$. Substituting into Eq. (2) gives Eq. (5).

(b) The rounding of column j contributes $-\delta_j x_{tj}$ and the updates $\sum_{k>j}\Delta\boldsymbol{w}_k x_{tk}$; with Eq. (2) the two combine into a sum over all $k \in F_j$,

$$\Delta\boldsymbol{y}_t = -\,\delta_j\,\frac{\sum_{k\in F_j}[\boldsymbol{H}_{\lambda,j}^{-1}]_{jk}\,x_{tk}}{[\boldsymbol{H}_{\lambda,j}^{-1}]_{jj}} = -\,\delta_j\,\frac{[\boldsymbol{H}_{\lambda,j}^{-1}\boldsymbol{x}_{t,j}]_j}{[\boldsymbol{H}_{\lambda,j}^{-1}]_{jj}}.$$

By (a), $[\boldsymbol{H}_{\lambda,j}^{-1}\boldsymbol{x}_{t,j}]_j = [\boldsymbol{M}_j^{-1}\boldsymbol{x}_{t,j}]_j - v_j\,\boldsymbol{v}^\top\boldsymbol{x}_{t,j}/(1+\kappa)$, which is Eq. (6). At $t=0$, $[\boldsymbol{M}_j^{-1}\boldsymbol{x}_{0,j}]_j = v_j$ and $\boldsymbol{v}^\top\boldsymbol{x}_{0,j} = \kappa$, so the numerator is $v_j/(1+\kappa)$ and the second form of Eq. (6) follows.

(c) For $\boldsymbol{M}_j = \lambda\boldsymbol{I}$,

$$\boldsymbol{v} = \frac{\boldsymbol{x}_{0,j}}{\lambda}, \qquad m_{kl} = \frac{\delta_{kl}}{\lambda}, \qquad 1+\kappa = \frac{\lambda + \|\boldsymbol{x}_{0,j}\|^2}{\lambda}, \qquad \|\boldsymbol{x}_{0,j}\|^2 = \|\boldsymbol{x}_{0,j+1}\|^2 + x_{0j}^2.$$

The denominator of Eqs. (5) and (6) becomes $(\lambda+\|\boldsymbol{x}_{0,j+1}\|^2)/(\lambda(\lambda+\|\boldsymbol{x}_{0,j}\|^2))$ and the numerator of Eq. (5) $x_{0j}x_{0k}/(\lambda(\lambda+\|\boldsymbol{x}_{0,j}\|^2))$, which gives $\Delta\boldsymbol{w}_k$. For Eq. (6), using $\boldsymbol{x}_{0,j}^\top\boldsymbol{x}_{t,j} - x_{0j}x_{tj} = \boldsymbol{x}_{0,j+1}^\top\boldsymbol{x}_{t,j+1}$,

$$\Delta\boldsymbol{y}_t = -\,\delta_j\,\frac{x_{tj}(\lambda+\|\boldsymbol{x}_{0,j}\|^2) - x_{0j}\,\boldsymbol{x}_{0,j}^\top\boldsymbol{x}_{t,j}}{\lambda+\|\boldsymbol{x}_{0,j+1}\|^2} = -\,\delta_j x_{tj} + \delta_j x_{0j}\,\frac{\boldsymbol{x}_{0,j+1}^\top\boldsymbol{x}_{t,j+1}}{\lambda+\|\boldsymbol{x}_{0,j+1}\|^2},$$

and at $t=0$ the last fraction is $\|\boldsymbol{x}_{0,j+1}\|^2/(\lambda+\|\boldsymbol{x}_{0,j+1}\|^2)$, which gives $\Delta\boldsymbol{y}_0$. □

A.3 Proof of Corollary 1

Proof. By assumption $\boldsymbol{M}_1 = \lambda\boldsymbol{I}$ and Appendix A.1, $\boldsymbol{M}_j = \lambda\boldsymbol{I}$ at every step, so Eq. (7) applies at each step with its own residual δ_j, and step j updates a column $k>j$ by $\Delta\boldsymbol{w}_k^{(j)} = \gamma_j x_{0k}\delta_j$.

(i) Column k receives the updates of the steps $j<k$. Those of the steps in J sum to $x_{0k}\sum_{j\in J,\,j<k}\gamma_j\delta_j$, which is $x_{0k}\boldsymbol{c}$ for $k>\max J$ and a partial sum for k inside J.

(ii) $\sum_{k>j}\Delta\boldsymbol{w}_k^{(j)}x_{tk} = \gamma_j\delta_j\sum_{k>j}x_{0k}x_{tk} = \gamma_j(\boldsymbol{x}_{0,j+1}^\top\boldsymbol{x}_{t,j+1})\delta_j$. For general $\boldsymbol{M}_j$, the term of Eq. (6) that involves $\boldsymbol{v}$ depends on t only through $\boldsymbol{v}^\top\boldsymbol{x}_{t,j} = \boldsymbol{x}_{0,j}^\top\boldsymbol{M}_j^{-1}\boldsymbol{x}_{t,j}$, so the readout becomes an overlap with $\boldsymbol{v} = \boldsymbol{M}_j^{-1}\boldsymbol{x}_{0,j}$ in place of $\boldsymbol{x}_{0,j}$, and the per-step directions of (i) differ, leaving a coherent component along the sink pattern.

(iii) For $\boldsymbol{x}_{0,j+1}\neq\boldsymbol{0}$ (otherwise $x_{0k}=0$ and the update vanishes),

$$\gamma_j x_{0k} = \frac{\|\boldsymbol{x}_{0,j+1}\|^2}{\lambda+\|\boldsymbol{x}_{0,j+1}\|^2}\cdot\frac{x_{0j}x_{0k}}{\|\boldsymbol{x}_{0,j+1}\|^2} = \frac{\kappa_{j+1}}{1+\kappa_{j+1}}\cdot\frac{x_{0j}x_{0k}}{\|\boldsymbol{x}_{0,j+1}\|^2},$$

with $\kappa_{j+1}\le\kappa$ by Appendix A.1. By Eq. (7) the step leaves $\Delta\boldsymbol{y}_0 = -\delta_j x_{0j}/(1+\kappa_{j+1})$ where rounding alone leaves $-\delta_j x_{0j}$, so the first factor is the share of the BOS rounding error it cancels. Since $k\in F_{j+1}$, $\|\boldsymbol{x}_{0,j+1}\|^2\ge x_{0k}^2$, so the second factor is at most $|x_{0j}/x_{0k}|$. □

A.4 Proof of Proposition 1

Proof. (a) At step j write the objective in the row deviations $\boldsymbol{\epsilon}_r\in\mathbb{R}^{F_j}$ from the current minimizer as $E_{\boldsymbol{G}} = \sum_{r,r'}G_{rr'}\boldsymbol{\epsilon}_r^\top\boldsymbol{A}\boldsymbol{\epsilon}_{r'}$, $\boldsymbol{A} = \boldsymbol{H}_{\lambda,j}$ (the induction of Appendix A.1 goes through because the minimizer below is the same at every step). Let $\boldsymbol{u} = \boldsymbol{A}^{-1}\boldsymbol{e}_j/[\boldsymbol{A}^{-1}]_{jj}$, so $u_j=1$ and $\boldsymbol{A}\boldsymbol{u} = \boldsymbol{e}_j/[\boldsymbol{A}^{-1}]_{jj}$, and write $\boldsymbol{\epsilon}_r = c_r\boldsymbol{u}+\boldsymbol{z}_r$ with $c_r = -\delta_j^{(r)}$ fixed by the rounding and $z_{r,j}=0$. Then $\boldsymbol{u}^\top\boldsymbol{A}\boldsymbol{z}_{r'} = z_{r',j}/[\boldsymbol{A}^{-1}]_{jj} = 0$, so $E_{\boldsymbol{G}} = (\sum_{r,r'}G_{rr'}c_rc_{r'})\boldsymbol{u}^\top\boldsymbol{A}\boldsymbol{u} + \mathrm{vec}(\boldsymbol{Z})^\top(\boldsymbol{G}\otimes\boldsymbol{A})\,\mathrm{vec}(\boldsymbol{Z})$. The first term is fixed, the second is ≥ 0 since $\boldsymbol{G}\succeq 0$, $\boldsymbol{A}\succ 0$, and vanishes at the feasible point $\boldsymbol{Z}=\boldsymbol{0}$; so $\boldsymbol{\epsilon}_r = -\delta_j^{(r)}\boldsymbol{u}$ minimizes for every $\boldsymbol{G}$ (uniquely if $\boldsymbol{G}\succ 0$), which is Eq. (2). A diagonal $\boldsymbol{G}$ scales each row's objective by a positive constant, which does not change any nearest grid point. (b) $\mathrm{tr}\,\widetilde{\boldsymbol{H}} = \sum_t w_t\|\boldsymbol{x}_t\|^2 = \sum_t g_t$, with BOS term g_0; with $w_t = g_t$ the trace is $\sum_t g_t a_t$ with BOS term g_0a_0. The identity stated after the proposition: $\mathrm{tr}[\Delta\boldsymbol{W}\widetilde{\boldsymbol{H}}\Delta\boldsymbol{W}^\top] = \sum_t w_t\|\Delta\boldsymbol{W}\boldsymbol{x}_t\|^2 = \sum_t g_t\|\Delta\boldsymbol{W}\boldsymbol{x}_t/\|\boldsymbol{x}_t\|\|^2 = \sum_t(\Delta\boldsymbol{W}\boldsymbol{x}_t)^\top(w_t\boldsymbol{I})(\Delta\boldsymbol{W}\boldsymbol{x}_t)$. □

B Protocols and Implementation Details

B.1 Calibration sampling in the reference implementations

We verified the sampling of each protocol against the source of the implementation that uses it (September 2026).

- **GPTQ reference code** (Frantar et al., 2023): draws documents of at least 2,048 tokens and cuts one 2,048-token window from each at a random offset, so that a window rarely begins with `BOS`. Our c4win protocol keeps its corpus and sample length but takes the first 2,048 tokens of each document, `BOS` first, as AutoRound does; the random-offset variant was not run.
- **AutoRound** (Cheng et al., 2024): its default calibration set is `NeelNanda/pile-10k`, 128 samples of 2,048 tokens, each the start of a document at least that long (shorter documents are dropped). Our Pile protocol reproduces this sampling, and c4win applies it to C4.
- **GPTQModel** (ModelCloud, 2025), `utils/calibration.py`: tokenizes each calibration row separately, does not truncate below the model's maximum length and does not concatenate rows by default; the README example takes 1,024 rows of C4. Its `QuantizeConfig` defaults are symmetric quantization, group size 128, `damp_percent` 0.05, `true_sequential` on and `desc_act` unset. Our short-document protocol reproduces its sampling with 128 rows; with 1,024 rows and the same damping our act-order loop leaves Qwen at 0.250 (0.435 at $\rho = 0.10$) and Llama at 0.144 (0.148 at $\rho = 0.10$; both Llama arms ran before the loop correction of this appendix), whereas GPTQModel's own default loop leaves them at 0.665 and 0.745 (Table 1).
- **LLM Compressor** (vLLM Project, 2025), `transformers/data/base.py`: truncates each sample to `max_seq_length` and does not concatenate by default, the same sampling shape; its documented example for instruction-tuned models uses chat-formatted data, which is one of the remedies of Corollary 1. Its `GPTQModifier` (`modifiers/gptq`) defaults to `actorder` static, which processes the columns in the order of $\operatorname{diag} \boldsymbol{H}$, as our loop does, but keeps every quantization group on consecutive columns of the original order (no `g_idx`), whereas our loop forms the groups along the sorted order.

The WikiText protocol takes consecutive, non-overlapping 2,048-token chunks of the WikiText-2 training split. In every protocol a calibration draw is a block of consecutive eligible documents of the corpus stream, and draw k skips the first k blocks, so that draws are disjoint but not random samples. Our quantizer matches the reference loop of GPTQ with act-order and $\rho = 0.05$ and collects the Hessians of the seven matrices of a block from the block's unquantized input (no true-sequential pass). RTN uses the same group size and symmetric grid, with its groups formed of consecutive columns in the original order, whereas act-order forms them along the sorted order and computes each group's scale from the compensated weights, as the reference loop does; GPTQ and RTN therefore also differ in the group partition under act-order, but not under the natural order, which collapses Llama as well (0.128). The loop runs layer by layer: the seven matrices of layer l are fitted on the output of the already quantized layers $0, \ldots, l-1$ passed through the unquantized layer l, and are then quantized together. On Llama the six other matrices of layer 0 are therefore fitted on unperturbed activations, those of layer 1 on activations that carry the layer-0 copy of the lesion, and none of the twelve on activations that carry the layer-1 lesion, which only layers 2 and later see; the cancellation of Section 4 is therefore not an adaptation of these matrices to the lesion.

The library's own loop. An advance prediction that GPTQModel's end-to-end default path collapses both models failed (Table 1). With nothing but bits and group size set, gptqmodel 5.6.12 on the README recipe (the first 1,024 or 128 rows of one C4 shard, tokenized by the library one row at a time, `BOS` added) leaves Llama at 0.665 / 0.646 and Qwen at 0.745 / 0.759, above RTN. Changing one setting at a time on Llama with 128 rows places the difference in the column order: act-order (`desc_act`) collapses (0.166), the natural order with the library's group-aware reordering switched off collapses harder (0.113), and our loop's three settings together collapse (0.164), whereas collecting the Hessians without the true-sequential pass (0.623), fixing the damping (0.644), feeding the paper's document stream (0.641) or pre-tokenized rows (0.663) change nothing. GPTQModel's current default, group-aware reordering, sorts the columns inside each 128-column group by $\operatorname{diag} \boldsymbol{H}$ and ranks the groups, keeping every group contiguous; the reference code's act-order sorts all columns globally, and the natural order sorts none. The same holds inside our loop: with the group-aware order in place of act-order and nothing else changed, Llama scores 0.614 and Qwen 0.746 (Table 1), and the natural order collapses Llama (0.128). The mechanism statistics of the sink matrix (Section 5.3) show what the order changes. Under act-order the compensation displaces the

weights of Llama's sink matrix by 32% of their squared norm, protects BOS (output error 0.4 times RTN's) and raises the root-mean-square error at every template position 2 to 10 times; under the group-aware order the displacement is 0.3%, BOS is protected a hundredfold (0.01 times RTN's) and the error rises mainly at one position, the token that closes the header (6.6 times), the others staying within 1.8 times RTN's; under the natural order, which leaves the BOS coordinates free for the first 2,400 steps, the same position reaches 41 times RTN's. Qwen behaves alike (displacement 17% against 2.3%, the worst template position 22 against 5.6 times). The lesion signature of Section 7 (squared error at the worst template position at least 3 times RTN's) is therefore present under every order, 105 / 43 / 1,700 on Llama and 471 / 32 on Qwen for act-order / group-aware / natural, and what the order sets is how much of the sink's rounding error is redistributed and over how many positions; why the group-wise order confines it is not derived (Section 8). The collapse of this paper is that of the reference loop and of the natural order under the libraries' sampling shape, not of GPTQModel's current default; LLM Compressor, whose default processes the columns in our loop's global order but groups them as RTN does, was not run end to end, so whether any library's complete default collapses is open.

Table 1: GPTQModel end to end (gptqmodel 5.6.12, 3 bits, group 128, IFEval). Default: `QuantizeConfig` with only bits and group size set (`desc_act` unset, `act_group_aware` on, `true_sequential` on, `damp_percent` 0.05 with auto-increment), the README calibration recipe. Each Llama row after the first two changes one setting, except the one that applies our loop's three settings at once; the last two rows are our loop with the column order changed. RTN: 0.565 (Llama), 0.697 (Qwen); our loop: 0.155 / 0.426.

Arm	Llama-3.1-8B	Qwen2.5-14B
Default, 1,024 rows	0.665	0.745
Default, 128 rows	0.646	0.759
act-order on (`desc_act`)	0.166	–
group-aware reordering off (natural order)	0.113	–
our loop's settings: act-order, no true-sequential, no group-aware	0.164	–
true-sequential pass off	0.623	–
damping auto-increment off	0.644	–
the paper's C4 document stream in place of the README rows	0.641	–
README rows pre-tokenized by us, BOS first	0.663	–
our loop, group-aware order in place of act-order	0.614	0.746
our loop, natural order	0.128	–

Table 2: What in the main arms is a library default and what is our choice.

Component	Used here	Default?
Calibration sampling	128 C4 documents tokenized separately	GPTQModel and LLM Compressor sampling shape; GPTQModel's README example uses 1,024 rows (Qwen: 0.250 in our loop, 0.745 in GPTQModel's; Llama: 0.144 in our loop before its correction, 0.665 in GPTQModel's)
Grid	symmetric, group size 128, 3 or 4 bits	GPTQModel default
Damping	$\rho = 0.05$	GPTQModel default
Column order	act-order, groups along the sorted order	an option in the GPTQ reference code; LLM Compressor's default (`actorder` static) uses the same order with the original groups; GPTQModel 5.6 defaults to group-aware reordering (`act_group_aware`), `desc_act` unset
Hessian source	block input, no true-sequential pass	GPTQModel default is true-sequential
Storage	fake-quant on the integer grid	ours (shared inference path)

B.2 A CORRECTION TO OUR LAYER-WISE LOOP

An earlier version of our loop passed the keyword arguments captured from the model's forward call, including an automatically created key-value cache, into the per-layer calls. Because each layer is called twice per sample (once to accumulate the Hessian, once to propagate after quantization), the second call attended the layer's pre-quantization keys and values. Every GPTQ-family number in this paper comes from the corrected loop unless it is marked as run before the correction, and the methods that do not use our loop (RTN, AutoRound, HQQ, and GPTQModel with its GPTAQ) are unaffected; we re-ran under the corrected loop every GPTQ-family arm that the paper reports without that mark. We validated the corrected loop by comparing the activations it propagates with a full forward pass of the quantized model: on Llama, Qwen and Mistral the two are bitwise identical at the input of every layer (largest relative error 0, against a pass tolerance of 5×10^{-2} set for bf16 recomputation noise). The fix keeps the cache out of every layer call, which the check asserts, and is not specific to a model; the other census models were not checked individually. Sixteen models were quantized only before the correction and are outside the quantization census: Llama-3.2-1B, Hermes-2-Pro, R1-Distill-Llama-8B, Qwen2-7B, Qwen2.5-0.5B, Qwen2.5-1.5B, Qwen3-4B, R1-Distill-Qwen-14B, Mistral-7B-v0.1, Mistral-7B-v0.2, Mistral-Small-24B, Falcon3-10B, SmolLM3-3B, gemma-2-2b, Yi-1.5-9B and OLMo-2-7B (ten of them are among the 28 models of Fig. 3, drawn open there). None fell more than 20 points below RTN (the lowest, Hermes-2-Pro, 3.7 points below; OLMo-2 and Mistral-Small-24B have no RTN arm, and the latter lost 6 points to fp16), and the corrected-loop runs planned for them were not made. Mistral under Pile windows, a protocol of Fig. 2b, was also run only before the correction and collapsed (0.179, RTN 0.420); the figure does not report it (n/r), nor the two protocols never run on Mistral. The two collapses, their cures and the census statements were unchanged; two results moved. Llama under Pile windows went from 0.402, 16 points below RTN, to a healthy 0.618, so the protocol dependence for Llama is now carried by the short-document protocol alone. And the transplant of the short-document lesion into the window-calibrated network went from a mild loss (0.567) to a collapse (0.174): under the defective loop the later layers had been compensated against partially unquantized activations and were less tightly fitted to their own sink-forming matrix, which is why a foreign matrix was tolerated. We report the corrected values and draw the conclusion of Section 4 from them. Numbers from before the correction are kept only where they are marked, as a record of what was run: besides those above, the Llama arms with 1,024 short documents (Table 2), the single-BOS control (below), the arm that drops the first positions (Section 6), three lesion statistics of healthy arms (Appendix D) and the refuted predictions so marked in Table 3. The census, the lesion-size correlation of Section 7, the arms nearest the collapse cutoff and the four repairs use no number from before the correction.

B.3 THE CORRECTED OBJECTIVE IN PRACTICE

The corrected objective was run on four collapsing model–protocol pairs, Llama and Qwen under 128 short C4 documents, Qwen under Pile windows and Mistral under C4 windows (Table 6), and repairs all four; it was not run on the other collapsing arms of this paper (Qwen under WikiText windows, 128 short Pile documents and 1,024 short C4 documents for either model, Llama under the natural column order), so the claims of Sections 6 and 7 cover these four pairs. $g_t = \|\partial L/\partial \boldsymbol{y}_t\|^2$ is measured on the full-precision model before quantization: one forward and one backward pass per calibration sample with L the summed next-token cross-entropy of that sample, hooks on the input and on the output gradient of every linear layer, parameters frozen (activation gradients only). As implemented, g_t is normalized to unit mean over the calibration set per layer, and token normalization rescales every input $\boldsymbol{x}_t$ of calibration sample s to that sample's root-mean-square token norm r_s, so that the weight entering $\widetilde{\boldsymbol{H}}$ is $w_t = r_s^2\, g_t/\|\boldsymbol{x}_t\|^2$: within a sample this is Eq. (8) up to a constant, and normalization alone ($w_t = r_s^2/\|\boldsymbol{x}_t\|^2$) leaves each sample's share of $\operatorname{tr}\boldsymbol{H}$ unchanged while equalizing its tokens. No cap or floor is applied; the capped variant of Table 6 is the only exception. Because the damping is $\lambda = \rho \operatorname{mean}\operatorname{diag}\widetilde{\boldsymbol{H}}$, it follows the reweighted curvature: at the sink-forming layer the BOS term no longer dominates $\operatorname{tr}\widetilde{\boldsymbol{H}}$, so λ is ρ times the mean reweighted curvature of the ordinary tokens, whereas under GPTQ it also contains the BOS mass; ρ was not retuned (Table 17 lists the sensitivity to ρ: 0.572 at $\rho = 0.01$, 0.610 at $\rho = 0.2$ against 0.638 at 0.05, Llama). The sensitivity is that of the full-precision network; whether it should be re-measured on the partially quantized network as the layer-wise loop proceeds is left open. Cost on one NVIDIA H200 GPU, 128 short documents, quantization only: on Llama-3.1-8B GPTQ takes

173 s and the corrected objective 206 s, with peak memory 18.8 GB against 28.0 GB; on Qwen2.5-32B 1,351 s against 1,406 s and 83.6 GB against 94.6 GB. The extra cost is one backward pass per calibration sample and the storage of g_t, which grows with the layer width while the weights grow with its square, so the relative memory overhead falls with model size (49% at 8B, 13% at 32B). On Qwen2.5-14B the corrected objective takes 467 s and 46.5 GB against 288 s and 31.3 GB for GPTQ, but the two were measured in different submissions and possibly on different nodes, so only the 8B and 32B pairs are within-batch comparisons. The gradient is not needed at every layer. Observation 1 concerns, and the lesion of Theorem 1 forms at, sink-dominated down-projections, which the forward pass that builds $\boldsymbol{H}$ already reveals through $A_{\texttt{BOS}}$; hooking only those layers changes the storage but not the time, since the backward pass is the same whether one layer or all are read from it, and re-rounding those layers without compensation (Section 4) needs no gradient at all. The position prior of Section 6, a constant weight on the first positions of every sample, is the gradient-free surrogate: with token normalization it repairs Llama and Mistral without a gradient (normalization alone repairs Qwen); whether it holds on a model whose g_t sits away from the first positions is untested.

B.4 Evaluation

Every generation benchmark uses the model's own chat template and greedy decoding in bf16. IFEval is the 541-prompt set of Zhou et al. (2023) with the official checker, at most 1,280 new tokens, reported as the mean of prompt-level strict, instruction-level strict, prompt-level loose and instruction-level loose accuracy; Table 5 and the census report this mean. The chat templates are applied with their defaults, which for the Qwen3 models enables the thinking mode: with at most 1,280 new tokens many of their responses end inside the reasoning trace, so their IFEval scores are low in absolute terms and are only compared within a model, the 4-bit Qwen3-14B results of Section 6 included. Multi-IF is the English subset of He et al. (2024), three turns, each turn appended to the running conversation and scored with the IFEval instruction registry against that turn's response; we report prompt-level strict accuracy per turn. GSM8K is the 1,319-problem test set, zero-shot with a chain-of-thought instruction in the chat template, at most 512 new tokens, the answer being the number after `####` or, failing that, the last number in the response. MMLU is five-shot with the standard harness prompt, scored by comparing the logits of the four answer letters. Perplexity is on the WikiText-2 raw test set, non-overlapping 2,048-token windows, token-weighted mean negative log-likelihood. On Llama and Mistral the chat template writes a `BOS` token and the tokenizer adds another, so every generation prompt in this paper carries two; all arms share this, and the single-`BOS` control moves the full-precision Llama from 0.768 to 0.808 and a collapsed checkpoint from before the loop correction from 0.150 to 0.158, so no conclusion depends on it. The position indices of Section 4 count both. The chat-formatted calibration sets (C4 in the chat template, UltraChat) and the chat prompts of the mechanism statistics carry the same two `BOS` tokens on Llama and Mistral; their inputs to the sink-forming down-projection have the same norm (475 on Llama), so the share of $\boldsymbol{H}$ that one `BOS` holds under short documents is split between positions 0 and 1 (0.5 each at Llama's layer 1). One arm is one checkpoint; IFEval takes about twenty minutes per 8B checkpoint on one H200 and the quantization itself the times of Appendix B.3. The 4-point noise threshold of Section 3 is the largest movement between two calibration draws of any non-collapsing arm we ran, rounded up: the corrected objective on Llama moves from 0.638 to 0.669 (Table 17), the other three pairs by at most one point; the threshold therefore rests on four pairs of draws, and every census model was quantized with one draw. The arms nearest the 20-point cutoff are Qwen with output weighting alone (11.7 points below RTN; Table 6) and Qwen at $\rho = 0.1$ (21.5 below, counted as collapsed; Table 4), the latter within one bootstrap interval (below) of the cutoff; beyond it follow Mistral's window collapse with its variants and Llama with output weighting alone (22.8 to 24.7 below; Table 6). The template positions of the mechanism statistics are the first eight positions of a chat prompt without its sink tokens (on Llama positions 2 to 7: the header and the first words of the default system prompt); the chat scaffold of Appendix D is every prompt token the template inserts. The 20-point rule is relative to RTN, and five census models have an RTN score below 0.35 (Qwen2.5-3B, SmolLM2, Mistral-Nemo, Qwen3-8B, Hermes-3), where a drop of 20 points would reach the floor at which collapsed arms land; there the rule can hardly fire, so the census rate is closer to two collapses among the thirteen models in which one could register. A paired bootstrap over the 541 prompts (10,000 resamples) puts the 95% interval of a difference between two arms at about ± 3 to ± 4 points: -41.0 $[-45.0, -36.8]$ for Llama's GPTQ against RTN and -27.2 $[-31.5, -22.9]$ for Qwen's, -4.6 $[-8.0, -1.3]$ for gemma-2-9b's corrected objective

against RTN, $+0.7$ $[-2.1, +3.6]$ for Qwen2.5-32B's against GPTQ, and $+3.1$ $[-0.4, +6.7]$ between the two draws of Llama's corrected objective that set the noise threshold.

C ADVANCE PREDICTIONS THAT FAILED

From the diagnosis onward, most batches of experiments were preceded by a written prediction, recorded in the header comment of the job script before the batch ran; the scripts are released with the code. The table lists the refuted predictions, grouped by what they were meant to establish, with what each refutation removed from the account; those of batches run before the loop correction of Appendix B are marked: their numbers were not re-run, so what they rule out is established for the loop as it then was.

Table 3: Refuted advance predictions.

Predicted	Observed	Ruled out
What predicts collapse		
Collapse is predictable from the template-position error of the lesion (a threshold)	No threshold, and at most a weak ranking: Spearman -0.46 across the census (-0.50 on 16 unseen models, pre-correction loop)	A lesion-size criterion
κ orders the protocols by collapse (Corollary 1(iii))	Qwen collapses at $\kappa = 2.6 \times 10^5$ and at 87, is healthy at 2.7×10^5	κ as a collapse criterion
The share of the trace held by `BOS` at the second matrix orders collapse across models	Tulu-3 and Llama share the same value (0.266) at 0.726 and 0.155; Qwen2.5-32B is healthy at 0.963	The count of `BOS`-dominated matrices as a cross-model criterion
The share of newline, special and document-initial tokens in the calibration draw predicts collapse	The draw that leaves Llama healthy has a lower share than one that collapses it; window protocols with ten times the share collapse or not	Calibration composition as a collapse predictor
Collapse under the short-document protocol is deterministic given the model	A second draw of 128 documents leaves Llama at 0.531	Collapse as a property of the model alone
4-bit is immune	Mistral-7B-v0.3, Qwen3-14B and Hermes-3 lose 7.2, 6.2 and 4.6 points to RTN4 (Table 16)	Bit-width as a safe boundary
GPTQModel's own default loop collapses Llama and Qwen as our loop does	Neither collapses (0.646 / 0.759); inside the library, act-order (0.166) or the natural order (0.113) collapses Llama, its group-aware order does not, and the same holds in our loop (0.614 / 0.746; Table 1)	The sampling shape as a sufficient condition; the column order sets the lesion's mass and breadth, and that is not derived
Scaling the `BOS` input by ten inside $\boldsymbol{H}$ collapses Qwen2.5-7B, Llama-3.2-3B, Mistral-7B-v0.3 and 4-bit Llama-3.1-8B (before the loop correction)	None collapses: 0.653, 0.493, 0.439 and 0.745 against 0.672, 0.577, 0.468 and 0.745 unscaled	More `BOS` dominance as a trigger where κ is already large
Per-channel quantization ($g = -1$) collapses Qwen2.5-32B, Hermes-3 and Mistral-Nemo (before the loop correction)	GPTQ stays above RTN at $g = -1$: 0.583, 0.201 and 0.238 against 0.175, 0.150 and 0.182	A lesion-size threshold under a coarser grid
The Llama-3.1 fine-tunes Tulu-3, Hermes-3 and R1-Distill-Llama-8B collapse as their base model does (before the loop correction)	None does: 0.725, 0.306 and 0.335 against RTN 0.637, 0.339 and 0.295	Collapse as a property of the base model
Where the lesion is and what it needs		
Injecting the lesion's direction into a healthy model at the lesion's amplitude collapses it	No collapse up to four times the amplitude	Fragility of the full-precision network

continued on the next page

Table 3, continued

Predicted	Observed	Ruled out
Two sink matrices are needed wherever the collapse occurs	Qwen's sink-forming down-projection alone collapses the RTN network (0.320); the second adds nothing (0.347)	A fixed size for the minimal collapsing set
Mistral's window collapse is carried by the down-projections	One or two of them leave the RTN network unchanged (0.411, 0.399 against 0.420)	One lesion site for every phenotype
Mistral's window collapse sits in one layer or in the down-projections	Layer 1 alone (0.417) and the four early down-projections (0.423) leave RTN unchanged; the fourteen matrices of layers 0–1 collapse it (0.192)	A single-matrix carrier for that phenotype
Substituting the window-calibrated version of a lesion disarms it	The collapse survives at layer 0 (0.160) as it does at layer 1, since a foreign matrix is itself fatal	Substitution as a test of what the lesion contributes
The background and its absorption		
The amplification sits in layers 2–4, where the error first grows	Whole-layer transplants of layers 2–4 change neither outcome (pre-correction loop)	A layer-local background
The protocol dependence lives in the background, not the lesion: a foreign lesion is tolerated by the window-calibrated network	Under the corrected loop the foreign lesion collapses it (0.174)	Separating lesion from background by transplantation between GPTQ checkpoints
The background is the co-adaptation of the layers quantized after the lesion	Rounding every layer from the second onward to nearest keeps the collapse (0.13 to 0.18); the lesion alone in an all-RTN network scores 0.528	The later layers as the background; it lies in the other matrices of the first two layers
The other matrices of the first two layers amplify the lesion	Every set without the layer-0 down-projection is healthy, including all of layer 1 with the lesion (0.572); the other twelve cancel the pair in the healthy draw and fail to in the collapsing one (Section 4)	Amplification by the neighbors; what varies between networks is cancellation
The healthy calibration draw writes a benign pair	Its two down-projections collapse the RTN network (0.158); its fourteen matrices of layers 0–1 do not (0.525)	The lesion as what varies between draws; it is the neighbors
A healthy model of the same family carries a benign pair	Tulu-3's own pair collapses its RTN network (0.117 against 0.637) while the full GPTQ network scores 0.726	Sufficiency of the lesion as a model-level property
The healthy calibration draw writes a smaller lesion	Same row readout on the template tokens (15.9 against 0.61 on ordinary tokens; 15.5/0.60 for the collapsing draw)	Lesion size as what varies between draws
The readout and the direction		
The lesion accumulates along the response	Response tokens barely affected	Damage during generation
Template tokens are the high-overlap tail of the isotropic readout $\lvert \boldsymbol{x}_0^\top \boldsymbol{x}_t \rvert$	The tail carries 2.5 times the error of the bottom quintile, but template tokens are 0.03 of it (Llama)	The isotropic form of Corollary 1(ii) as the placement of the template tokens
The whitened overlap $\boldsymbol{x}_0^\top \boldsymbol{M}^{-1} \boldsymbol{x}_t$ places the damaged tokens	Worst of the three readouts on Qwen (recall 0.33 against 0.55 for the row readout)	The general form of Corollary 1(ii) as the readout
Restoring Llama's special-token positions alone repairs the collapse	0.144; the whole scaffold, which on Llama is mostly its system block, 0.636	The special tokens, rather than the template tokens as a class, as the carriers
Removing the `BOS` token from $\boldsymbol{H}$ changes the direction of the compensation on the sink column	Direction only partly changed (cosine 0.05 to 0.87 against 0.52 to 0.93 with `BOS`); the amplitude falls by 4 to 500, in step with κ	The dominant token as the sole source of the direction; it sets the amplitude
The block-diagonal restriction of Eq. (3) holds on Llama	Full curvature 0.35 of the block-diagonal one, and negative; 1.09 on Qwen	The two-sided expansion as a model of the lesion's cost
The repair		

continued on the next page

Table 3, continued

Predicted	Observed	Ruled out
Capping the output weights rescues output weighting alone on Llama	The capped arm falls further (0.329 to 0.160)	Dynamic range as the reason output weighting alone fails
Output weighting alone repairs all three collapses	Not Llama (0.329) or Qwen (0.580)	Sufficiency of g_t
The extreme weights of g_t are the working part of the repair	Capping at $100\times$ the median changes nothing (0.629); the earlier 0.160 was output weighting alone	Dynamic range of g_t as the mechanism of the repair
The repair needs the weight on the first positions of each sample	Removing it leaves 0.637 on Llama	Document-initial positions as the carriers of g_t on Llama
The corrected objective needs the weight on the first positions on Mistral	Removing it costs two points (0.412); adding a constant weight to them repairs normalization alone (0.405)	The gradient as the only route to the output side
Chat-formatted calibration recovers at most half of the 4-bit loss	Full recovery on Mistral and Hermes-3, a further loss on Qwen3-14B	Superiority of the corrected objective where both apply
Capping each token's norm at ten times the root-mean-square, then the median, token norm cures Llama and Qwen without the cost of normalization on the Mistral family (before the loop correction)	Llama stays collapsed (0.155, 0.153); Qwen is cured by the median cap only (0.375, 0.742); Mistral still loses (0.414 against 0.468)	A soft cap as a substitute for normalization
The corrected objective moves by at most 3 points between Pile and C4 calibration on six models (before the loop correction)	Four of the six move more: Falcon3-7B $+6.0$, Qwen2.5-7B $+4.7$, Tulu-3 $+4.0$, Mistral-Nemo -3.6	Independence of the repair from the calibration corpus
GPTQ that keeps 1% of the weights, chosen by OBS saliency, in full precision still collapses (SpQR-style, before the loop correction)	It does not: 0.678 on Llama, 0.788 on Qwen	The large inverse curvature of the BOS coordinates as hiding them from outlier selection

D ADDITIONAL RESULTS

The sensitive positions (Section 5.1). A synthetic error along the sink direction written into the RTN-quantized Llama at the template positions costs 0.565→0.291 at eight times the template norm, while the same error at six ordinary positions costs nothing (0.558). Restoring the collapsed model's hidden states after the lesion layer at the chat scaffold alone, the 35 (Llama) or 28 (Qwen) prompt tokens outside the user's text, repairs both collapses, 0.155→0.636 and 0.426→0.755, while restoring as many user-text positions changes nothing (0.141, 0.435); the template positions alone give 0.144 on Llama and 0.750 on Qwen, BOS alone 0.153, the whole prompt 0.664.

Verification of Corollary 1 (Section 5.3). At Llama's sink-forming down-projection the quantized matrix's error on template positions is 2.25 for GPTQ against 0.43 for RTN and 0.4 times RTN's at BOS, largest on template and newline tokens and smallest on the response (Fig. 4a,b); 23% of it sits in the super-weight row (Yu et al., 2024), against 30% of RTN's error, so that Llama fails the row-concentration test of Section 7, as do Qwen2.5-7B and Falcon3-7B; that row's readout tracks the block's error (Fig. 4c), and the largest compensated entry of the matrix is 7.3 where the original matrix has 0.71 (5.4 against 0.62 within that row): weights were pushed past their group range and clipped. The lesion statistic of Section 7 is the squared GPTQ/RTN error ratio at the worst template position, the template positions being the first eight positions of the chat prompt without its sink tokens (those whose input to the matrix is more than ten times the median norm of the eight: positions 0 and 1 on Llama and Mistral, position 0 on Qwen2.5-32B), taken at the first-token-dominated down-projection ($A_{\text{BOS}} \geq 0.50$) where it is largest. It is at least 3.6 in 14 of the 17 sink models,

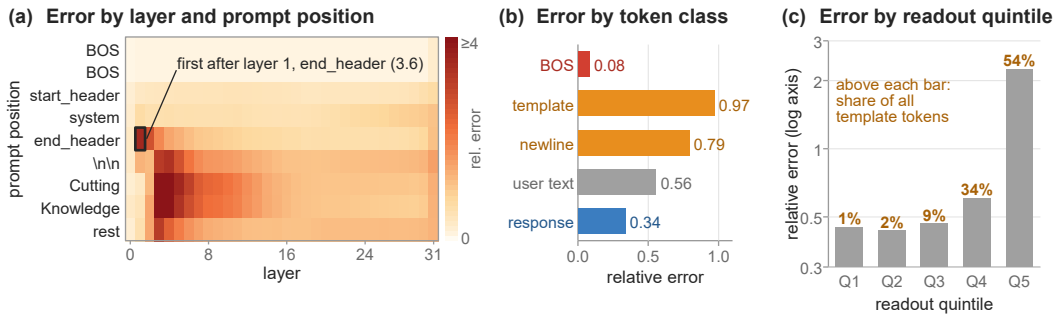


Figure 4: Anatomy of the lesion (Llama-3.1-8B-Instruct, the collapsed short-document GPTQ3 checkpoint against fp16). (a) Relative hidden-state error $\|\boldsymbol{h}_t^{\mathrm{q}} - \boldsymbol{h}_t\|/\|\boldsymbol{h}_t\|$ at the output of each layer (indexed from 0, as in the text) by prompt position, over 32 IFEval prompts: it appears after layer 1, the sink-forming layer, at the token that closes the chat header, and spreads from layer 3. The two BOS rows are the chat template's and the tokenizer's (Appendix B), *Cutting* and *Knowledge* are the first words of Llama 3.1's default system prompt, and *rest* is the mean over the remaining positions. (b) The same error after layer 1 by token class, over 16 IFEval prompts extended by 128 greedy fp16 tokens (the response). (c) The error of (b) by quintile of $\boldsymbol{e}_r^\top \boldsymbol{x}_t$, where $\boldsymbol{e}_r$ is the row of the quantized-minus-original sink matrix that carries the most template-position error; the number above each bar is the share of all template tokens that fall in the quintile. Panel (c) decomposes the measured error by the row that writes it; it is not the readout predicted by Corollary 1(ii), which on Llama does not place the template tokens (Appendix C). No panel has an RTN reference; the ratios to RTN are in Section 5.3.

Table 4: The remedies found by trial, the quantity of Theorem 1 each is designed to move, and what it does (IFEval, 3 bits, Llama-3.1-8B / Qwen2.5-14B under short documents). The last row is the repair of Section 6. Token normalization is $w_t = 1/\|\boldsymbol{x}_t\|^2$, the corrected objective $w_t = g_t/\|\boldsymbol{x}_t\|^2$, each up to a constant per calibration sample (Appendix B.3); Mistral-7B-v0.3 values are under C4 windows. The second column was not measured under every remedy, and a remedy that removes the collapse need not remove the layer-1 signature (Appendix D).

Remedy	Quantity targeted	Llama / Qwen	Limitation
GPTQ, 3 bits	—	0.155 / 0.426	
RTN, 3 bits	no compensation	0.565 / 0.697	
Damping, $\rho = 5$	λ in $\boldsymbol{M}_j$ rises, κ falls	0.662 / 0.770	fails at $\rho = 0.2$ (Llama 0.169) and $\rho = 0.1$ (Qwen 0.482); Llama GSM8K 0.604 vs 0.632
Token normalization	every $\|\boldsymbol{x}_t\| \to r_s$: no dominant token	0.636 / 0.758	fails on Mistral (0.189 vs RTN 0.420)
Chat calibration	template activations enter $\boldsymbol{M}_j$	0.644 / 0.774	fails on Qwen3-14B at 4 bits (0.428 from 0.466; thinking mode, Appendix B)
Rotation (QuaRot)	$\boldsymbol{x}_0$ spread out	0.648 / 0.709	changes the inference path; tested on Llama and Qwen only
Re-rounding sink matrix	Eq. (2) removed at one matrix	0.622 / 0.768	needs the matrix located (A_{BOS}, no gradient; Appendix B.3)
Corrected objective	output side restored (Proposition 1)	0.638 / 0.736	Llama MMLU 0.514, GSM8K 0.548 vs 0.632

in all of them at the most BOS-dominated down-projection of Table 16 (Qwen2.5-32B there 55, at its template position 1), and 1.1 to 2.3 in Qwen2.5-7B, Falcon3-7B and SmolLM2, which exceed 3 only at down-projections the first token does not dominate (A_{BOS} 0.14, 0.36 and 0.008); the Spearman correlation of Section 7 is over all seventeen. Llama's second BOS-heavy matrix, the layer-0

Table 5: Two collapses under the short-document protocol (3 bits, group 128). IFEval is the mean of prompt- and instruction-level strict and loose accuracy; Multi-IF is turn 1 of three. MMLU of the collapsed checkpoints is above their RTN counterparts' and Qwen's perplexity is lower (better); only Llama's perplexity worsens.

Model	Arm	IFEval	GSM8K	Multi-IF	MMLU	PPL
Llama-3.1-8B-Instruct	fp16	0.768	0.843	0.722	0.683	7.2
	RTN, 3-bit	0.565	0.280	0.485	0.416	15.5
	GPTQ, 3-bit	0.155	0.008	0.079	0.524	22.4
	GPTQ, 4-bit	0.756	0.816	0.714	0.659	9.7
Qwen2.5-14B-Instruct	fp16	0.820	0.926	0.812	0.799	5.70
	RTN, 3-bit	0.697	0.817	0.647	0.698	9.26
	GPTQ, 3-bit	0.426	0.173	0.354	0.739	7.66
	GPTQ, 4-bit	0.808	0.933	0.778	0.787	6.20

Table 6: The corrected objective and its two halves (IFEval, 3 bits; second calibration seed in parentheses). Short documents: 128 C4 documents tokenized separately; windows: the first 2,048 tokens of 128 documents (Appendix B). Token normalization (tnorm) is $w_t = 1/\|\boldsymbol{x}_t\|^2$, output weighting (out. w.) is $w_t = g_t$, the corrected objective $w_t = g_t/\|\boldsymbol{x}_t\|^2$, each up to a constant per calibration sample (Appendix B.3); output weighting alone collapses the healthy Pile-window Llama (0.140 against 0.618), and a dash is an arm not run; "capped" caps g_t at one hundred times its median inside the corrected objective; the position prior is token normalization with the first eight positions of every document weighted $\times 10$ / $\times 100$, no gradient.

Model	Calib.	GPTQ	RTN	tnorm	out. w.	corrected	corr. capped	prior $\times 10/\times 100$
Llama-3.1-8B	short docs	0.155	0.565	0.636	0.329	0.638 (0.669)	0.629	0.623/0.602
Llama-3.1-8B	Pile win.	0.618	0.565	0.675	0.140	0.626	–	–
Qwen2.5-14B	short docs	0.426	0.697	0.758	0.580	0.736 (0.730)	0.751	–
Qwen2.5-14B	Pile win.	0.271	0.697	–	0.589	0.782	–	–
Mistral-7B-v0.3	C4 win.	0.187	0.420	0.189	0.475	0.433 (0.424)	0.425	0.405/0.418
Mistral-Nemo	short docs	0.478	0.285	0.537	0.541	0.540	–	–

Output weighting alone with the cap, Llama short docs: 0.160. Damping of the corrected objective, Llama short docs: $\rho = 0.01$ 0.572, 0.05 0.638, 0.2 0.610. Corrected objective with the first eight positions' weight removed: Llama 0.637, Mistral 0.412.

down-projection of Section 4, is a weaker copy: range overshoot 3.7 against 8.7 at layer 1, one row with 0.21 of its error energy against 0.23. The sending pattern of the loop itself, the row of the Cholesky factor of $\boldsymbol{H}_\lambda^{-1}$ in act-order along which the rounded sink column compensates the later ones, has cosine 0.52 to 0.93 with $\boldsymbol{x}_0$ in fourteen of seventeen sink matrices (0.02 to 0.70 with $\boldsymbol{v}$) against 0.03 to 0.43 on ordinary columns and 0.01 for a random direction; removing the BOS token from $\boldsymbol{H}$ lowers the cosine with $\boldsymbol{x}_0$ in fourteen of the seventeen (to 0.05 to 0.87; Llama 0.83 to 0.43) but not in Qwen2.5-7B, Qwen2.5-14B and Qwen2.5-32B, whose ordinary columns already compensate along $\boldsymbol{x}_0$, and removes the amplitude in step with κ: the pattern shrinks by about 1 where $\kappa \leq 6 \times 10^3$ (granite, Qwen2.5-3B, Falcon3), by 4 to 7 times in the Llama family ($\kappa \approx 10^5$) and by 150 to 500 times in the Qwen2.5 family ($\kappa \geq 2.6 \times 10^5$), the overshoot falling from 5 to 12 to 1.7 times the largest original entry. The dominant token therefore sets the amplitude and only partly the direction. Of the readouts in (ii), the isotropic overlap $\boldsymbol{x}_0^\top \boldsymbol{x}_t$ holds at the extremes, the top quintile of a prompt's tokens carrying 2.5 times the error of the bottom one (Qwen 3.9), but does not place the template tokens (recall 0.11 on Llama, 0.52 on Qwen); the row readout $\boldsymbol{e}_r^\top \boldsymbol{x}_t$ of the lesion row ($\boldsymbol{e}_r$ its row of the quantized-minus-original matrix), a decomposition of the measured error rather than a prediction, does (recall 0.54 on Llama and 0.55 on Qwen, against 0.22 and 0.33 for the whitened overlap $\boldsymbol{x}_0^\top \boldsymbol{M}^{-1} \boldsymbol{x}_t$; Appendix C). In the activations the error is one direction, cosine 0.88 with the BOS hidden state at the lesion position, largest on template and newline tokens (0.97 against 0.56) and smallest on the response (0.34).

Table 7: Calibration protocols. All use 128 samples; they differ in how the text is cut, and therefore in the share of the curvature that the beginning-of-sequence token receives.

Protocol	Corpus	Sample	Tokens	Default of
short-document	C4	one document, no concatenation	420 (mean)	GPTQModel, LLM Compressor
c4win	C4	first 2,048 tokens of a document	2,048	corpus and length of the GPTQ reference code, which cuts at a random offset
Pile	Pile	first 2,048 tokens of a document	2,048	AutoRound
WikiText	WikiText-2	consecutive 2,048-token chunk	2,048	
c4chat	C4	one document in the chat template	430	
UltraChat	UltraChat	one conversation	2,048	

Table 8: IFEval of 3-bit GPTQ (our act-order loop) under different calibration protocols. The short-document sampling shape is the default of GPTQModel and LLM Compressor, the Pile windows that of AutoRound, and the C4 windows take the corpus and length of the GPTQ reference code with AutoRound's cut at the start of a document (Table 7); Mistral-7B-v0.3 collapses under the window protocol instead. The corrected objective was run on the collapses under short-document C4 and Pile windows and on Mistral's, not on those under short-document Pile or WikiText windows (Appendix B.3).

Protocol	Tokens / sample	Llama-3.1-8B	Qwen2.5-14B	Mistral-7B-v0.3
RTN (no calibration)	–	0.565	0.697	0.420
short-document C4, 128 docs	420	0.155	0.426	0.453
short-document Pile, 128 docs	–	0.154	0.332	–
C4 windows, 128 × 2,048	2,048	0.645	0.766	0.187
Pile windows, 128 × 2,048	2,048	0.618	0.271	–
WikiText windows, 128 × 2,048	2,048	0.564	0.219	–
C4 documents in the chat template	430	0.644	0.774	0.462
UltraChat conversations	2,048	0.673	0.770	0.474

Healthy networks carry the lesion (Sections 4 and 6). The layer-1 signature of Llama, the largest squared GPTQ/RTN error ratio at template positions 2 to 7 of the sink matrix, is 105 in the collapsed network, 163 in the healthy second draw and 43 under the group-aware order; under the loop before its correction it was 301 with C4 windows, 84 with chat-formatted calibration and 33 at $\rho = 5$, all healthy arms. The lesion is thus written whether or not the network collapses: heavy damping and chat-formatted calibration shrink it but leave it far above RTN's, and chat-formatted calibration removes the layer-0 copy instead (end-header ratio 0.009 against 3.6 in the collapsed network), in line with the pair of Section 4, while heavy damping, the group-aware order and C4 windows leave that copy at 1.1 to 1.9. The healthy second draw carries the layer-0 copy as well (3.4), so the two draws differ neither in the lesion nor in its copy, and their pairs are equally fatal in the RTN network (Section 4); what separates them lies in the other matrices, which is the cancellation we infer. Removing the first two positions of every calibration sample from $\boldsymbol{H}$ at every layer (`--hess-drop-pos 2`, before the loop correction), the one direct intervention on the dominant token, lifted Llama from 0.150 to 0.560, the RTN level, against 0.648 for token normalization in the same loop; in this single run, down-weighting the first token did better than dropping it.

What the released run index adds. Every number above and in the main text is one row of the run index released with the code, which also holds the arms that the text cites but the tables do not tabulate: the per-model Theorem 1 statistics of Section 5.3 (κ, $\kappa_{\text{iso}} = \|\boldsymbol{x}_0\|^2/\lambda$, the rank-one share of the `BOS` block, the sending pattern's cosine with $\boldsymbol{x}_0$ and $\boldsymbol{v}$ on `BOS` and on ordinary columns, with and

Table 9: Quantizers at 3 bits (group 128, short-document calibration where calibration is used, except the second AutoRound row, which uses AutoRound's own default, Pile windows), sorted by objective. IFEval.

Objective	Method	Llama-3.1-8B	Qwen2.5-14B
input-side curvature (OBS)	GPTQ	0.155	0.426
	GPTQ in GPTQModel, default group-aware order	0.646	0.759
	GPTAQ	0.150	0.752
	GPTQ after rotation	0.648	0.709
token-averaged block loss	AutoRound	0.460	0.181
	AutoRound, Pile windows	0.716	0.793
none	RTN	0.565	0.697
	HQQ	0.599	0.749
	AWQ (channel scales)	0.623	0.719

Table 10: Excising the lesion. Starting from the collapsed 3-bit GPTQ checkpoint, one matrix is re-rounded with RTN (no calibration, no extra bits) or a set of matrices is kept at full precision.

Model	Change to the collapsed checkpoint	IFEval
Llama-3.1-8B-Instruct	none	0.155
	sink-forming down-projection (layer 1) re-rounded	0.622
	down-projections of layers 0–1 re-rounded	0.632
	attention projections of layers 0–2 at fp16 (control)	0.138
	five largest rows of the lesion matrix at fp16 (control)	0.174
Qwen2.5-14B-Instruct	none	0.426
	sink-forming down-projection (layer 4) re-rounded	0.768

without the `BOS` token, and at two control matrices per model; κ across protocols for the three main models); the readout quintiles of Corollary 1(ii) for the isotropic overlap, the whitened overlap and the row readout, with the template share and recall of each quintile; the position-matched injection and the activation-patching arms of Section 5.1 (0.144, 0.197, 0.153, 0.664, 0.636, 0.141 on Llama; 0.750, 0.441, 0.755, 0.435 on Qwen); the position-weight arms of Section 6 (0.623, 0.602, 0.637; Mistral 0.405, 0.418, 0.412); the per-position error profiles behind Section 4; the other quantizers under the window protocols and the single-block anatomy of AutoRound's failure; and every transplant of Section 4. It also holds the arms of batches run before the loop correction of Appendix B, which the paper cites only where it marks them, and two void AutoRound rows (`m_l_ar3`, `m_q14_ar3`), for which the library's loader kept three calibration samples (`jobs/w56b_autoround.sh`).

Table 11: Transplanting the lesion. The sink-forming down-projection of a donor checkpoint replaces the same matrix of a target network; everything else in the target is unchanged. IFEval of the result; the target's own score in parentheses. The last Llama row is a transplant into an already collapsed network: it tests whether the tolerated matrix cures, not whether it collapses.

Target network (own score)	Donor of the sink-forming down-projection	IFEval
Llama fp16 (0.768)	short-document GPTQ3 (the lesion)	0.764
Llama RTN3 (0.565)	short-document GPTQ3	0.528
Llama GPTQ3, window calibration (0.645)	short-document GPTQ3	0.174
Llama GPTQ3, short-document (0.155)	window-calibrated GPTQ3 (tolerated by its own network; RTN in its place cures, 0.622)	0.152
Qwen fp16 (0.820)	short-document GPTQ3 (layer 4)	0.637

Table 12: What each repair costs (3 bits, short-document calibration). Re-rounding needs the sink matrix located, which $A_{\texttt{BOS}}$ does in the forward pass that builds $\boldsymbol{H}$ (Appendix B.3); damping $\rho = 5$ needs a per-model search; the corrected objective needs one backward pass per calibration sample.

Model	Arm	IFEval	GSM8K	MMLU	PPL
Llama-3.1-8B	fp16	0.768	0.843	0.683	7.2
	GPTQ	0.155	0.008	0.524	22.4
	sink down-projections re-rounded	0.632	0.632	0.556	9.9
	damping $\rho = 5$	0.662	0.604	0.543	21.0
	corrected objective	0.638	0.548	0.514	12.4
Qwen2.5-14B	fp16	0.820	0.926	0.799	5.70
	GPTQ	0.426	0.173	0.739	7.66
	sink down-projection re-rounded	0.768	0.889	0.728	7.81
	damping $\rho = 5$	0.770	0.880	0.744	7.57
	corrected objective	0.736	0.877	0.727	7.80

Table 13: Multi-IF, three turns (turn 1 / 2 / 3, prompt-strict), 3 bits, short-document calibration.

Model	fp16	GPTQ	RTN	corrected objective
Llama-3.1-8B	0.722 / 0.559 / 0.409	0.079 / 0.010 / 0.001	0.485 / 0.222 / 0.094	0.513 / 0.284 / 0.142
Qwen2.5-14B	0.812 / 0.566 / 0.409	0.354 / 0.106 / 0.021	0.647 / 0.349 / 0.163	0.680 / 0.452 / 0.266

Table 14: What generalizes, with the number of models behind each statement. Per-model values are in Appendix D; differences are computed from unrounded scores.

Statement	Models	Outcome
Blind spot (Observation 1) at the sink-forming layer	26 with a sink, 2 without	26/26; 0/2
Lesion signature at a first-token-dominated down-projection (template error of GPTQ over RTN ≥ 3; measured, not predicted)	17 with a sink	14/17
Row concentration and range overshoot (Corollary 1(i) and (iii))	17	14/17; 16/17
3-bit collapse under the short-document protocol	18	2
3-bit collapse under C4 windows	18	1 (Mistral-7B-v0.3)
3-bit collapse under any of five protocols (three main models)	3	3: Llama and Qwen (short documents; Qwen also Pile and WikiText windows), Mistral (C4 windows; Pile windows run only before the loop correction, collapsed)
4-bit GPTQ minus 4-bit RTN	18	mean -0.1; negative on 11/18
Corrected objective minus RTN, worst case	18	-4.6 (gemma-2-9b, no sink)
Corrected objective minus GPTQ on models that did not collapse	16	mean $+0.3$ (-5.6 to $+13.9$)

Table 15: Which matrices the collapse needs (Section 4). Fixed-matrix transplants: the named matrices of a donor checkpoint replace the same matrices of a target network, everything else in the target unchanged; (ℓ, down) is the down-projection of layer ℓ. IFEval, 3 bits, group 128.

Transplanted set		IFEval
Llama-3.1-8B, collapsed short-document GPTQ3 network (0.155), layers k onward re-rounded to nearest		
$k = 2$ / 4 / 8 / 16 / 24	0.152 / 0.146 / 0.133 / 0.147 / 0.178	
Into the RTN3 Llama (0.565), from the collapsing draw (seed 0)		
(1, down)	the lesion alone	0.528
(0, down)	the layer-0 down-projection alone	0.543
(0, down) + (1, down)	the pair	0.139
layer 0 all + (1, down)	8 matrices	0.140
layers 0–1 all	14 matrices	0.152
layer 1 all	7 matrices, lesion inside	0.572
layer 1 MLP	gate, up, down	0.522
layer 1 attention + (1, down)		0.567
layer 0 attention + (1, down)		0.577
layer 0 gate, up + (1, down)		0.536
Into the RTN3 Llama, from the healthy draw (seed 1, 0.531)		
(0, down) + (1, down)	the pair	0.158
layers 0–1 all	14 matrices	0.525
Into the collapsed short-document Llama, from the window-calibrated (c4win) network		
(0, down)		0.160
(1, down)		0.152
(0, down) + (1, down)		0.580
Excision inside the collapsed short-document Llama (re-rounded to nearest, everything else GPTQ)		
(0, down) / (1, down) / both	0.595 / 0.622 / 0.632	
Into the RTN3 Tulu-3 (0.637; its GPTQ3 network is healthy, 0.726)		
(0, down) + (1, down)	its own pair	0.117
Into the RTN3 Qwen2.5-14B (0.697)		
(4, down)	the lesion, short documents	0.320
(4, down)	the same matrix from the window-calibrated network	0.208
(4, down) + (5, down)		0.347
both excised	in the GPTQ network	0.763
Into the RTN3 Mistral-7B-v0.3 (0.420), from its collapsed window-calibrated network (0.187)		
(1, down)		0.411
(0, down) + (1, down)		0.399
layers 0–3 down	4 matrices	0.423
layer 1 all	7 matrices	0.417
layers 0–1 all	14 matrices	0.192
layers 0–3 all	28 matrices	0.191

Table 16: The census, one row per model (Section 7): all 28 models for the premise, the first 18 for quantization. Sink layer, $A_{\texttt{BOS}}$ and $G_{\texttt{BOS}}$ are measured on the full-precision model at its most BOS-dominated down-projection (128 C4 documents); $\kappa = \boldsymbol{x}_0^\top \boldsymbol{M}^{-1} \boldsymbol{x}_0$ of Theorem 1 at that matrix. IFEval (mean of four accuracies) at 3 bits for fp16, RTN, GPTQ under short documents, the corrected objective under short documents and GPTQ under C4 windows, then 4-bit GPTQ and RTN; group size 128 throughout. At 4 bits the corrected objective scores 0.536 on Mistral-7B-v0.3, 0.589 on Hermes-3, 0.483 on Qwen3-14B, the three models that lose more than 4 points to RTN4 (Section 6), and 0.762 on Llama-3.1-8B and 0.829 on Qwen2.5-32B. The gemma-2 models have no sink-dominated layer; their rows give the largest $A_{\texttt{BOS}}$ found. The last ten models were quantized only before the loop correction of Appendix B and are outside the quantization census. RTN and fp16 values come from runs that the loop correction does not affect; a dash is a run we do not have. †Evaluated in the default thinking mode, in which many responses end inside the reasoning trace (Appendix B); only comparisons within the row are meaningful. $G_{\texttt{BOS}}$ is given in units of 10^{-3}, so the bound $G_{\texttt{BOS}} \leq 0.01$ of Observation 1, as pre-registered, reads 10 here; the largest value is SmolLM3-3B's. Differences quoted in the text are computed from the unrounded scores of the run index (gemma-2-9b: corrected objective 0.6645, RTN3 0.7109, -4.6 points).

Model	Size	Layer	$A_{\texttt{BOS}}$	$10^3 G_{\texttt{BOS}}$	κ	fp16	RTN3	GPTQ3	Corr.	c4win	GPTQ4	RTN4
Llama-3.1-8B-Instruct	8B	1	0.998	0.30	1.2×10^5	0.768	0.565	0.155	0.638	0.645	0.756	0.760
Qwen2.5-14B-Instruct	14B	4	1.000	0.21	2.6×10^5	0.820	0.697	0.426	0.736	0.766	0.808	0.803
Mistral-7B-Instruct-v0.3	7B	1	0.859	0.23	3.9×10^3	0.552	0.420	0.453	0.415	0.187	0.473	0.545
Qwen2.5-7B-Instruct	7B	3	0.850	0.16	6.4×10^3	0.765	0.500	0.666	0.629	0.674	0.767	0.747
Qwen2.5-32B-Instruct	32B	4	0.998	0.30	5.3×10^5	0.838	0.774	0.784	0.791	0.791	0.839	0.846
Qwen2.5-3B-Instruct	3B	2	0.988	1.1	121	0.658	0.130	0.440	0.460	0.431	0.621	0.568
Mistral-Nemo-Instruct	12B	2	0.925	0.25	8.2×10^4	0.651	0.285	0.478	0.540	0.515	0.618	0.619
gemma-2-9b-it	9B	0	0.005	16	—	0.767	0.711	0.712	0.664	0.713	0.754	0.757
Llama-3-8B-Instruct	8B	1	0.998	0.83	1.2×10^5	0.758	0.521	0.651	0.596	0.644	0.720	0.733
Llama-3.2-3B-Instruct	3B	1	0.998	0.82	6.9×10^4	0.750	0.538	0.575	0.565	0.566	0.685	0.716
Falcon3-7B-Instruct	7B	3	0.501	4.1	5.7×10^3	0.770	0.608	0.672	0.679	0.682	0.753	0.764
SmolLM2-1.7B-Instruct	1.7B	7	0.987	3.9	1.6×10^5	0.543	0.185	0.245	0.322	0.239	0.551	0.451
Hermes-3-Llama-3.1-8B	8B	1	0.998	0.46	1.2×10^5	—	0.339	0.276	0.415	0.407	0.530	0.577
Ministral-8B-Instruct	8B	2	0.957	1.6	1.0×10^4	—	0.493	0.481	0.503	0.502	0.577	0.573
Qwen3-14B†	14B	6	0.999	0.47	3.5×10^5	—	0.418	0.504	0.481	0.446	0.466	0.528
Llama-3.1-Tulu-3-8B	8B	1	0.998	0.72	1.2×10^5	—	0.637	0.726	0.673	0.725	0.827	0.793
Qwen3-8B†	8B	6	0.999	0.47	2.7×10^5	—	0.332	0.392	0.386	0.392	0.416	0.425
granite-3.1-8b-instruct	8B	5	0.991	0.070	169	—	0.446	0.554	0.542	0.592	0.686	0.665
Llama-3.2-1B-Instruct	1B	1	0.999	0.18	—	—	—	—	—	—	—	—
Hermes-2-Pro-Llama-3-8B	8B	1	0.998	0.56	—	—	—	—	—	—	—	—
R1-Distill-Llama-8B	8B	1	0.999	0.97	—	—	—	—	—	—	—	—
Qwen2-7B-Instruct	7B	3	0.933	0.15	—	—	—	—	—	—	—	—
Qwen3-4B	4B	6	0.997	0.84	—	—	—	—	—	—	—	—
R1-Distill-Qwen-14B	14B	4	0.855	0.025	—	—	—	—	—	—	—	—
Mistral-7B-Instruct-v0.2	7B	1	0.828	0.25	—	—	—	—	—	—	—	—
Falcon3-10B-Instruct	10B	3	0.545	3.6	—	—	—	—	—	—	—	—
SmolLM3-3B	3B	1	0.994	8.4	—	—	—	—	—	—	—	—
gemma-2-2b-it	2B	1	0.003	79	—	—	—	—	—	—	—	—

Table 17: Calibration draws. IFEval of the three collapses under independent draws of the calibration set (seed 0 is the draw used everywhere else), and of the corrected objective under two draws; the corrected objective's damping sensitivity on Llama. Every draw that collapses does so by 23 to 45 points below RTN; the one that does not (Llama, seed 1) is the case analyzed in Section 4.

Model, protocol	Arm	seed 0	seed 1	seed 2	seed 3	seed 4
Llama-3.1-8B, short documents	GPTQ3	0.155	0.531	0.149	0.141	0.174
	corrected	0.638	0.669			
Llama-3.1-8B, Pile windows	corrected	0.626	0.626			
Qwen2.5-14B, short documents	GPTQ3	0.426	0.351	0.378	0.250	
	corrected	0.736	0.730			
Mistral-7B-v0.3, C4 windows	GPTQ3	0.187	0.173			
	corrected	0.433	0.424			
Corrected objective, Llama, short documents, by damping: $\rho = 0.01$ 0.572, $\rho = 0.05$ 0.638, $\rho = 0.2$ 0.610.						