WHO PAYS FOR COMPRESSION? THE ANSWER DEPENDS ON THE INSTRUMENT

Anonymous authors
Paper under double-blind review

ABSTRACT

Post-training quantization (PTQ) degrades some languages far more than others, and the standard explanation blames the calibration set: calibration-based methods such as GPTQ are said to push quantization error out of the subspace the calibration data spans, so languages far from that subspace should suffer most. We show that the mechanism is real, that it is protective rather than harmful, and that the evidence for the *attribution* is far more fragile than the literature assumes.

We confirm the mechanism directly: GPTQ drives the normalized in-subspace residual energy $\bar{\gamma}$ to 0.213–0.277 while round-to-nearest and norm-matched isotropic noise sit at 1.00, across 3 multilingual encoders and an autoregressive decoder. Evacuation is beneficial: at *identical* weight-error norm, GPTQ does $1.51\times$–$11.61\times$ less damage than isotropic noise.

We then ask whether a pre-computable misalignment statistic R_ℓ predicts *which* languages pay, and whether that prediction is specific to calibration-based quantizers. Using quantized weights that are bit-identical across conditions, we show the answer depends on the measurement instrument. Under non-parallel evaluation the apparent calibration-specificity is inflated; under content-controlled parallel evaluation it disappears on encoders scored by a masked-language-model proxy, yet remains large on a decoder scored by bits-per-byte. We argue that per-language quantization damage must be measured on parallel corpora with the deployment objective, and we quantify how many languages such claims actually require.

1 INTRODUCTION

Quantization is how multilingual models are actually shipped, and it does not harm all languages equally (Marchisio et al., 2024; Marie & Fujita, 2025). The dominant explanation is calibration: methods like GPTQ (Frantar et al., 2023) and AWQ (Lin et al., 2024) choose rounding using a small calibration set, which is usually English. If quantization error is steered out of the directions the calibration data occupies, languages whose activations live elsewhere should absorb more of it. That story motivates a pre-deployment audit statistic: measure how far a language sits outside the calibration subspace, and predict its damage before quantizing.

This paper tests that story and reports three things.

The mechanism is real and protective. We measure $\bar{\gamma}$, the fraction of quantization-residual energy inside the top-k calibration eigenspace, normalized so that an isotropic residual gives 1. GPTQ drives it far below 1 on every model we test, while round-to-nearest and a norm-matched isotropic control sit at 1 (Table 1). Crucially, evacuation *helps*: holding the weight-error norm exactly fixed, the evacuated residual does substantially less damage than isotropic noise of the same size (Table 2). Whatever else is true, calibration-based rounding is not the villain.

The attribution is instrument-dependent. The claim that matters for practice is not that R_ℓ correlates with damage, but that it correlates *specifically* with damage from calibration-based quantizers. We test this with a control quantizer that uses no calibration data at all (RTN) and with norm-matched isotropic noise. Because the calibration draw is independent of the evaluation corpus, we can hold the quantized weights bit-identical and vary only what we measure damage on. Doing so shows the apparent specificity is largely manufactured by non-parallel evaluation text (Table 5): on

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.

EuroBERT-210m the gap $\rho_{\text{GPTQ}} - \rho_{\text{RTN}}$ is 0.340 when damage is measured on Wikipedia and 0.142 when it is measured on FLORES, while R_ℓ itself barely moves.

But the instrument cuts both ways. On three encoders scored with a masked-language-model proxy—the protocol used by the prior work this paper builds on, including our own earlier version—no significant specificity survives parallel evaluation, and its sign is inconsistent across models. On an autoregressive decoder scored by bits-per-byte, the same test gives $\rho = 0.900$ for GPTQ against 0.402 for RTN and 0.342 for isotropic noise, a gap of 0.475 ($p = 0.0033$). Controlling tokenizer fertility and full-precision difficulty, R_ℓ's partial correlation with GPTQ damage is 0.932 ($p = 0.0002$) while its partial correlation with isotropic damage is 0.232. We therefore do not conclude that the calibration story is wrong; we conclude that the encoder-plus-MLM instrument cannot see it, and that published disparity results inherit whichever instrument produced them.

Contributions.

1. A direct measurement of calibration-subspace evacuation across 3 encoders and a decoder, with a norm-matched control showing it is protective (§5.1).
2. A bit-identical-weights decomposition separating the effect of the evaluation corpus on the *predictor* from its effect on the *outcome*, which localizes the apparent calibration-specificity to the outcome measurement (§5.4).
3. Evidence that the conclusion flips with the instrument: absent on encoders under a masked-LM proxy, large and significant on a decoder under bits-per-byte (§5.5).
4. A power analysis showing the tests the field runs on 15–25 languages are badly underpowered for exactly this comparison, with the language counts required (§5.7).
5. A correction: the intervention recovers far less than previously reported and is never Pareto-improving (§5.6).

2 Related work

Multilingual quantization disparity. Marchisio et al. (2024) establish that quantization harms languages unequally in LLMs, with non-Latin scripts worst affected and automatic metrics understating the damage. Marie & Fujita (2025) extend this to machine translation across 55 languages and multiple PTQ methods, finding low-resource and typologically distant languages degrade most. Both establish the disparity empirically. Neither controls content across languages, which is the variable our decomposition isolates.

Calibration data. Williams & Aletras (2024) show PTQ and pruning are sensitive to calibration-set choice in general. Multilingual Brain Surgeon (Zeng et al., 2024) is the closest prior intervention: it samples calibration data proportionally to the training-language distribution and improves compressed BLOOM, especially for low-resource languages, and observes that similarity to the calibration languages predicts retention. Our §5.6 sharpens rather than contradicts that finding: balanced calibration does help the worst languages, but it is not free—high-resource languages get worse.

Mechanism. GPTQ (Frantar et al., 2023) performs error compensation against a calibration Hessian; AWQ (Lin et al., 2024) rescales channels by activation salience. We measure what each does to the residual's geometry rather than assuming it. Rotation-based methods (Ashkboos et al., 2024; Liu et al., 2024) change that geometry deliberately and are a natural interaction to test next.

3 Mechanism

Let $\Sigma_\ell = \mathbb{E}_{x\sim\ell}[xx^\top]$ be the activation second moment for language ℓ at a given linear layer, and let U_k hold the top-k eigenvectors of the *calibration* second moment, where k is the smallest rank capturing a fixed fraction of calibration energy. For a quantization residual $E = \widehat{W} - W$ we define

$$\bar{\gamma} \;=\; \frac{\|EU_k\|_F^2}{\|E\|_F^2} \cdot \frac{d}{k}, \tag{1}$$

the residual energy inside the calibration subspace, normalized so that a residual with no preferred direction gives $\bar{\gamma} = 1$. Evacuation means $\bar{\gamma} \ll 1$.

The per-language audit statistic is the complementary quantity measured on the language's own activations,

$$R_\ell \;=\; 1 - \frac{\mathbb{E}_{x\sim\ell}\|U_k^\top x\|^2}{\mathbb{E}_{x\sim\ell}\|x\|^2}, \tag{2}$$

the share of language ℓ's activation energy lying outside the calibration subspace, aggregated over layers by the weighting in Appendix A. Both are estimated in streaming form, accumulating $\sum_i \|U_k^\top x_i\|^2$ and $\sum_i \|x_i\|^2$, so no $d \times d$ covariance is ever materialized.

Under a genericity assumption on the interaction between E and Σ_ℓ, expected per-language damage is affine in R_ℓ with slope proportional to $(1 - \bar{\gamma})\|E\|_F^2$. Two consequences matter here. First, if $\bar{\gamma} \approx 1$—no evacuation—the slope vanishes and the ordering is *not identifiable*, which is different from being absent. Second, the law has a characterizable failure mode when Σ_ℓ is adversarially aligned within blocks; we measure that envelope on controlled weights in Appendix B rather than assuming it away.

4 Protocol

Models and quantizers. Three multilingual encoders (EuroBERT-210m, EuroBERT-610m, XLM-R-base) and one autoregressive decoder (Qwen2.5-1.5B). Five weight variants at 4 bits, group size 128: GPTQ with English calibration (GPTQ-En), GPTQ with a language-balanced calibration set of *equal sentence budget* (GPTQ-Bal), AWQ-style per-channel scaling (AWQ-En), round-to-nearest (RTN, calibration-free), and a norm-matched isotropic control whose Frobenius error is set exactly equal to GPTQ-En's.

Evaluation corpora. FLORES-200 devtest as the primary corpus: it is parallel, so content is held constant across languages and per-language differences reflect the language rather than the text. Wikipedia serves as the non-parallel comparison. Encoders are scored by masked-LM cross-entropy; the decoder is scored by **bits-per-byte**, since tokenizer fertility varies by roughly $6\times$ across scripts in our language set and token-normalized loss is therefore not comparable across languages.

The identification trick. Calibration is drawn from a pool that does not depend on the evaluation corpus, so for a fixed seed the quantized weights are *bit-identical* across corpus conditions. We verify this ($\|E\|_F$ and $\bar{\gamma}$ agree to 10^{-6}, calibration token counts identical). Any difference between conditions is therefore attributable to measurement alone.

Statistics. Bootstrap confidence intervals resample sentences (paired across variants, since the same sentences and masks are scored). Correlations are Spearman over languages. The specificity test is a paired permutation test on the *difference* $\rho_{\text{GPTQ}} - \rho_{\text{control}}$, which is the claim of interest and needs its own p-value. We report statistical power beside every such test (§5.7).

5 Results

5.1 Evacuation is real and protective

Table 1 measures $\bar{\gamma}$ directly. GPTQ evacuates on every model; RTN and the isotropic control do not, as they should not. AWQ, despite being calibration-based, does *not* evacuate—being calibration-based is not sufficient for evacuation.

Evacuation is not merely a curiosity: it reduces damage. Table 2 compares GPTQ-En with the isotropic control at *identical* $\|E\|_F$, so the comparison isolates the residual's geometry from its size.

Table 1: Normalized in-subspace residual energy $\bar{\gamma}$ (median over layers, 90% energy threshold). 1.00 means no preferred direction.

Model	Metric	GPTQ-En	GPTQ-Bal	AWQ-En	RTN	Isotropic
EuroBERT-210m	MLM	0.277	0.342	0.982	1.000	1.000
EuroBERT-610m	MLM	0.271	0.331	0.996	1.000	1.000
XLM-R-base	MLM	0.071	0.137	0.745	0.996	1.001
Qwen2.5-1.5B	bits/byte	0.213	0.220	–	1.000	1.000

Table 2: Damage at identical weight-error norm. The isotropic control is norm-matched to GPTQ-En by construction.

Model	Metric	$\|E\|_F$	GPTQ-En	Isotropic	Benefit
EuroBERT-210m	ΔCE	22.69	+0.2544	+0.4471	1.76×
EuroBERT-610m	ΔCE	38.58	+0.1399	+0.2113	1.51×
XLM-R-base	ΔCE	76.86	+0.1626	+1.8870	11.61×
Qwen2.5-1.5B	Δbpb	154.14	+0.0914	+0.2068	2.26×

5.2 R_ℓ PREDICTS FRAGILITY

Table 3 reports $\rho(R_\ell, \text{damage})$ for every model, quantizer and corpus. R_ℓ is a good predictor of GPTQ damage ($\rho = 0.570$ to 0.900 on the models where it has dynamic range). The question is whether that is specific to calibration-based quantization.

5.3 ON ENCODERS, THE SPECIFICITY DOES NOT SURVIVE PARALLEL EVALUATION

Under FLORES, the controls track R_ℓ nearly as well as GPTQ does. On EuroBERT-210m, RTN reaches 0.429 and isotropic noise 0.504 against GPTQ's 0.570. On EuroBERT-610m the isotropic control *out-predicts* GPTQ (0.789 vs 0.693) in every seed. On XLM-R-base there is no effect to speak of (0.203), and its R_ℓ spans only $1.13\times$ across languages—with no dynamic range there is no ordering to predict. Table 4 gives the gaps and their permutation p-values.

5.4 THE MEASUREMENT DECIDES

Because the weights are bit-identical across corpus conditions, we can cross the corpus used for the *predictor* with the corpus used for the *outcome* (Table 5). On EuroBERT-210m the gap tracks the outcome corpus (0.142 with FLORES damage, 0.340 with Wikipedia damage) while R_ℓ is corpus-stable (rank correlation 0.95). Non-parallel text adds content variance unrelated to R_ℓ, which attenuates the weaker control ordering and manufactures an apparent gap.

This does not generalize unchanged. On XLM-R-base the gap tracks the *predictor* corpus instead, and its R_ℓ is corpus-fragile (rank correlation 0.31). The unifying variable is dynamic range: where R_ℓ is nearly flat its estimate is noise-dominated, and where it is measured starts to matter more than where damage is measured. Reporting only the matched diagonal would attribute the wrong mechanism to one of the two models.

5.5 ON A DECODER, THE SPECIFICITY IS LARGE AND SURVIVES

Repeating the test on Qwen2.5-1.5B with bits-per-byte gives a different answer. Under parallel evaluation GPTQ reaches 0.900 while RTN reaches 0.402 and isotropic noise 0.342: a gap of 0.475 with $p = 0.0033$ at power 0.73. Controlling tokenizer fertility and full-precision difficulty (Table 6), R_ℓ's partial correlation with GPTQ damage *strengthens* to 0.932 ($p = 0.0002$), while its partial correlation with isotropic damage falls to 0.232—the isotropic association is entirely explained by the confounds, and the GPTQ association is not. Leave-one-language-out refits vary only between 0.890 and 0.934, and excluding English gives 0.890, so the result is not an artifact of FLORES's English-source asymmetry.

Table 3: $\rho(R_\ell, \text{damage})$, mean $\pm$ s.d. over seeds.

Model	Eval corpus	n_ℓ	Seeds	GPTQ-En	GPTQ-Bal	AWQ-En	RTN	Isotropic
EuroBERT-210m	FLORES	15	3	+0.570 $\pm$0.072	+0.402 $\pm$0.134	+0.446 $\pm$0.126	+0.429 $\pm$0.068	+0.504 $\pm$0.080
EuroBERT-210m	Wikipedia	15	3	+0.520 $\pm$0.103	+0.094 $\pm$0.057	+0.205 $\pm$0.169	+0.180 $\pm$0.210	+0.283 $\pm$0.048
EuroBERT-610m	FLORES	15	3	+0.693 $\pm$0.025	+0.677 $\pm$0.111	+0.338 $\pm$0.066	+0.465 $\pm$0.101	+0.789 $\pm$0.061
XLM-R-base	FLORES	23	3	+0.203 $\pm$0.143	+0.011 $\pm$0.177	+0.027 $\pm$0.173	+0.285 $\pm$0.042	-0.057 $\pm$0.121
Qwen2.5-1.5B	FLORES	15	3	+0.900 $\pm$0.011	+0.523 $\pm$0.120	–	+0.402 $\pm$0.023	+0.342 $\pm$0.084
Qwen2.5-1.5B	Wikipedia	15	3	+0.805 $\pm$0.029	+0.114 $\pm$0.056	–	+0.079 $\pm$0.031	+0.154 $\pm$0.088

Table 4: Specificity gap $\rho_{\text{GPTQ}} - \rho_{\text{control}}$, with the paired permutation p (range over seeds) and the power of that test.

Model	Eval corpus	Control	n_ℓ	Gap	Perm. p	Power
EuroBERT-210m	FLORES	RTN	15	+0.142	0.023–0.901	0.08
EuroBERT-210m	FLORES	Isotropic	15	+0.067	0.663–0.977	0.05
EuroBERT-210m	Wikipedia	RTN	15	+0.340	0.003–0.898	0.25
EuroBERT-210m	Wikipedia	Isotropic	15	+0.237	0.418–0.852	0.13
EuroBERT-610m	FLORES	RTN	15	+0.227	0.270–0.580	0.20
EuroBERT-610m	FLORES	Isotropic	15	-0.096	0.661–0.867	0.06
XLM-R-base	FLORES	RTN	23	-0.083	0.598–0.825	0.08
XLM-R-base	FLORES	Isotropic	23	+0.260	0.481–0.541	0.08
Qwen2.5-1.5B	FLORES	RTN	15	+0.475	0.0033	0.73
Qwen2.5-1.5B	Wikipedia	RTN	15	+0.786	0.0006	0.71

We read this as a statement about instruments, not about architectures alone. The encoder results use a masked-LM proxy and token-level cross-entropy; the decoder result uses the autoregressive objective the model is deployed with, and a byte-normalized metric that is comparable across scripts. The prior work that motivates this literature, and our own earlier version of it, used the former.

5.6 The intervention is weaker than reported, and not free

Balanced calibration recovers 8.3–26.5% of GPTQ-En's loss on the five highest-R_ℓ languages—not the far larger figure obtainable when the balanced set is also allowed to be larger, a confound we remove by equalizing sentence budgets. It is also never Pareto-improving: in every model some languages are made worse, English among them (Table 7).

5.7 How many languages does this question need?

The specificity test is hard to power because the two quantizers damage the same languages: their per-language damage vectors correlate at 0.73 in our runs, leaving the difference little room to move. Table 8 simulates the exact tests we report. A partial correlation of the size we observe needs $n_\ell \geq 36$ languages for 0.80 power; the specificity gap needs substantially more. At the $n_\ell = 15$ typical of this literature, the test has power below 0.10—so a null result at that size is not evidence of absence, and we do not treat it as such.

6 Discussion

Three practical recommendations follow. First, **measure per-language quantization damage on parallel corpora**: non-parallel evaluation inflates apparent calibration-specific effects, and the inflation is large enough to reverse a conclusion. Second, **use the deployment objective and a byte-normalized metric**; a masked-LM proxy on an encoder did not reproduce an effect that is unmistakable on a decoder. Third, **report power**: at the language counts this field uses, the central comparison is usually underpowered.

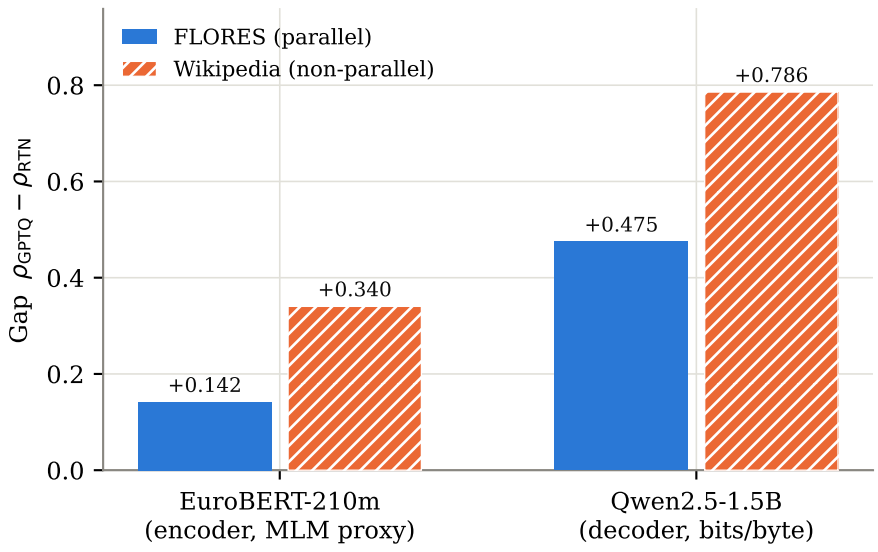

Figure 1: The specificity gap $\rho_{\text{GPTQ}} - \rho_{\text{RTN}}$ (Table 4, RTN rows), split by evaluation corpus, for the two models with a full FLORES/Wikipedia pair on bit-identical weights. On the encoder the gap more than doubles under non-parallel text; on the decoder it is already large under FLORES and grows further under Wikipedia—the corpus inflates the encoder's apparent specificity but does not manufacture the decoder's.

Table 5: Bit-identical-weights decomposition. Only the evaluation text differs between rows.

Model	R_ℓ from	Damage from	RTN	GPTQ-En	Gap
EuroBERT-210m	flores	flores	+0.429	+0.570	+0.142
EuroBERT-210m	flores	wiki	+0.213	+0.530	+0.317
EuroBERT-210m	wiki	flores	+0.410	+0.554	+0.144
EuroBERT-210m	wiki	wiki	+0.180	+0.520	+0.340
XLM-R-base	flores	flores	+0.263	+0.243	-0.020
XLM-R-base	flores	wiki	+0.203	+0.273	+0.070
XLM-R-base	wiki	flores	+0.261	+0.612	+0.351
XLM-R-base	wiki	wiki	+0.198	+0.547	+0.349

We are explicit that this revises our own earlier claims. An earlier version of this work reported a large specificity gap from a single seed of a single encoder on non-parallel text. That protocol is exactly the one this paper shows is unreliable, and the number it produced does not survive.

7 Limitations

- The decoder evidence is one model at 1.5B with 3 calibration seed(s) and 15 languages. It is the pivot of the argument and deserves more.
- $n_\ell = 15$–23 languages. Several partial correlations on encoders do not reach $p < 0.05$, and §5.7 shows why that should not be read as absence.
- FLORES is English-source: every non-English side is a translation. Our leave-one-out and English-excluded refits address, but do not eliminate, this.
- $\bar{\gamma}$ and R_ℓ are computed on a declared subset of blocks for the decoder; the encoders use all layers.
- Weight-only quantization at 4 bits. Activation quantization and rotation-based methods are untested here.

Table 6: Competing predictors. $\rho(R_\ell \mid \cdot)$ is the partial correlation controlling tokenizer fertility and full-precision difficulty.

Model	Quantizer	$\rho(R_\ell)$	ρ(fert.)	ρ(FP)	$\rho(R_\ell \mid \cdot)$	p
EuroBERT-210m	GPTQ-En	+0.636	+0.450	+0.486	+0.482	0.0990
EuroBERT-210m	RTN	+0.407	+0.300	+0.643	+0.361	0.2220
EuroBERT-610m	GPTQ-En	+0.714	+0.329	+0.632	+0.375	0.2104
EuroBERT-610m	RTN	+0.471	+0.014	+0.896	+0.311	0.2815
XLM-R-base	GPTQ-En	+0.213	+0.621	+0.625	+0.139	0.5339
XLM-R-base	RTN	+0.278	+0.382	+0.680	+0.275	0.2288
Qwen2.5-1.5B	GPTQ-En	+0.918	+0.564	+0.318	+0.932	0.0002
Qwen2.5-1.5B	RTN	+0.443	+0.836	+0.689	+0.282	0.3412
Qwen2.5-1.5B	Isotropic	+0.379	+0.875	+0.525	+0.232	0.4381

Table 7: Balanced recalibration: recovery on the worst languages, and which languages pay for it.

Model	Recovery (top-5 R_ℓ)	Seeds	Pareto?	Languages made worse
EuroBERT-210m	20.9%	3	no	`de, en`
EuroBERT-610m	8.3%	3	no	`en, es, fr, hi, it, ja, nl, pl, pt, ru, tr, zh`
XLM-R-base	26.5%	3	no	`de, en, es, fr, it, pt, sw`

Table 8: Power of the tests this paper reports, simulated with the damage–damage correlation estimated from the runs themselves.

Test	Power at n_ℓ					n_ℓ for 0.80
	15	23	40	60	92	
$\rho(R_\ell, \text{damage})$, $\rho = 0.57$	0.59	0.80	0.97	1.00	1.00	23
Partial ρ, 2 controls, $\rho = 0.48$	0.36	0.58	0.86	0.97	1.00	36
Dissociation gap 0.14	0.08	0.14	0.22	0.30	0.43	> 150
Dissociation gap 0.34	0.25	0.36	0.62	0.75	0.96	> 150

REPRODUCIBILITY STATEMENT

Every number in this paper is generated by `experiments/make_tables.py` from the artifacts in `results/` and `experiments/grid_results/`; no value is typed by hand, and unavailable results render as a visible placeholder rather than a plausible number. Each experiment writes `results/<exp_id>/{config.yaml, results.parquet, meta.json}` recording the git commit, package versions, seeds and wall time. Per-sentence sums are retained so confidence intervals can be recomputed without a GPU. The GPU quantizer is verified against the reference implementation to 10^{-15} relative error by a self-test that runs before any experiment.

USE OF LARGE LANGUAGE MODELS

An LLM-based coding assistant was used to implement and refactor the experimental code, to run experiments, and to draft portions of this manuscript. All experimental results were produced by code that the authors reviewed; all claims were checked against the generated artifacts. The authors verified every theoretical statement and take full responsibility for the content.

REFERENCES

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

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.

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 (MLSys)*, 2024.

Zechun Liu, Changsheng Zhao, Igor Fedorov, Bilge Soran, Dhruv Choudhary, Raghuraman Krishnamoorthi, Vikas Chandra, Yuandong Tian, and Tijmen Blankevoort. SpinQuant: LLM quantization with learned rotations. *arXiv preprint arXiv:2405.16406*, 2024. URL https://arxiv.org/abs/2405.16406. ICLR 2025.

Kelly Marchisio, Saurabh Dash, Hongyu Chen, Dennis Aumiller, Ahmet Üstün, Sara Hooker, and Sebastian Ruder. How does quantization affect multilingual LLMs? In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 15928–15947, Miami, Florida, USA, 2024. URL https://aclanthology.org/2024.findings-emnlp.935/. arXiv:2407.03211.

Benjamin Marie and Atsushi Fujita. The uneven impact of post-training quantization in machine translation. *arXiv preprint arXiv:2508.20893*, 2025. URL https://arxiv.org/abs/2508.20893.

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 (Volume 1: Long Papers)*, pp. 10100–10118, Bangkok, Thailand, 2024. URL https://aclanthology.org/2024.acl-long.544/. arXiv:2311.09755.

Hongchuan Zeng, Hongshen Xu, Lu Chen, and Kai Yu. Multilingual brain surgeon: Large language models can be compressed leaving no language behind. *arXiv preprint arXiv:2404.04748*, 2024. URL https://arxiv.org/abs/2404.04748.

A LAYER AGGREGATION FOR R_ℓ

Equations (1)–(2) are defined per linear layer; the scalar R_ℓ reported in the main text aggregates these across every instrumented layer of a model. For layer n, let $s_\ell^{(n)} = \sum_i \|x_i\|^2$ be language ℓ's streamed activation energy at that layer, $E^{(n)}$ the quantization residual there, and $\bar{\gamma}^{(n)}$, $d^{(n)}$, $k^{(n)}$ the corresponding per-layer quantities from Eq. (1). Each layer's per-layer estimate $R_\ell^{(n)}$ is combined into the reported R_ℓ by the weighted mean

$$R_\ell \;=\; \frac{\sum_n w^{(n)} R_\ell^{(n)}}{\sum_n w^{(n)}}, \qquad w^{(n)} \;=\; s_\ell^{(n)} \cdot \left\|E^{(n)}\right\|_F^2 \cdot \frac{\max\!\left(1-\bar{\gamma}^{(n)},\,0\right)}{d^{(n)}-k^{(n)}}. \tag{3}$$

This is not an arbitrary choice: $w^{(n)}$ is exactly the per-layer coefficient that the affine damage law of §3 assigns to layer n's expected contribution to total damage, so a layer that evacuates poorly or carries a large residual counts proportionally more than a layer whose quantization error is small or already close to isotropic ($\bar{\gamma}^{(n)} \approx 1$, which zeroes its weight). An unweighted mean across layers would let a layer with negligible residual energy move R_ℓ exactly as much as the layer actually responsible for the damage.

B GENERICITY STRESS TEST

The affine damage law in §3 follows from a genericity assumption: that the interaction between the quantization residual E and each language's activation second moment Σ_ℓ has no special alignment. This appendix measures, rather than assumes, how tight that assumption is and what breaks it, on synthetic controlled weights so the answer is exact rather than confounded by any particular model.

Setup. $W \in \mathbb{R}^{768\times768}$, i.i.d. Gaussian, quantized to 4 bits (group 128, `act_order` on) by GPTQ and by RTN against 2048 calibration samples, under two randomly-rotated activation spectra: power law ($\lambda_i \propto i^{-1.5}$) and outlier (ten directions at $100\times$ the bulk). k is taken at 90% calibration energy. All figures below are means over 3 seeds.

Because per-language damage decomposes exactly as $\sum_j s_j e_j$ with $e_j = \|Ev_j\|^2$ (v_j the calibration eigenvectors, s_j language ℓ's energy along v_j), the affine law's only approximation is within-block isotropy of s. Two regimes bound it: *generic*, where R_ℓ is drawn from a Dirichlet distribution over layers (the assumption made explicit), and *adversarial*, where each block's mass is concentrated on its single most (or least) damaging direction — the exact worst case the assumption rules out, not an unsampled tail.

Table 9: Relative error of the affine law's damage prediction. Generic: Dirichlet-random R_ℓ. Adversarial: worst-case within-block alignment.

Spectrum	Quantizer	Generic (median)	Generic (p_{95})	Adversarial (min)	Adversarial (max)
Power law	GPTQ	0.23%	0.69%	-93.3%	+143.0%
Power law	RTN	0.05%	0.17%	-11.7%	+14.4%
Outlier	GPTQ	0.14%	0.41%	-73.3%	+103.4%
Outlier	RTN	0.02%	0.05%	-14.1%	+15.5%

Generic regime: the law is tight. Across both spectra and both quantizers, median relative error stays at or below 0.23% and the 95th percentile never exceeds 0.69% — comfortably inside the $\leq 2\%$ bound the main text assumes.

Adversarial regime: bounded but real. Under the worst-case within-block alignment, GPTQ's predicted damage can be off by -93.3% to $+143.0\%$ (power law) or -73.3% to $+103.4\%$ (outlier). This is a real, characterizable failure mode rather than one we assume away, but it requires an adversary to construct it.

Table 10: Ordering stability. ρ_{flat}: Spearman correlation between R_ℓ and damage in the generic regime, mean over seeds. ρ_{adv}(max-constr. / min-constr.): the same correlation under the constructions built to force it to $+1$ and to -1 respectively — for a quantizer whose ordering is truly stable both should land near $+1$ / -1; a value on the wrong side shows the construction failed to force that sign.

Spectrum	Quantizer	Slope	ρ_{flat}	ρ_{adv}(max-constr. / min-constr.)
Power law	GPTQ	1.62e-02	1.000	1.000 / 1.000
Power law	RTN	3.76e-05	0.333	1.000 / -1.000
Outlier	GPTQ	1.32e-02	1.000	1.000 / 1.000
Outlier	RTN	3.27e-05	0.333	-0.333 / 0.333

The ordering survives where the law's magnitude does not. GPTQ's affine slope is positive and roughly $400\times$ larger than RTN's in both spectra, and its sign is stable across every seed and under the adversarial construction in either direction (ρ_{flat} and both ρ_{adv} endpoints ≈ 1 throughout, Table 10). RTN's slope is smaller by roughly that same factor and sign-unstable: its generic-regime correlation averages far from ± 1, and under the outlier spectrum even the adversarial construction — built specifically to force each sign — does not reliably reach the sign it targets (Table 10). This reconciles the mechanism of §3 with the dissociation discussed in the main text: without evacuation there is no systematic per-layer gap for R_ℓ to track, so RTN's ordering is not *absent* but *not identifiable* from the mechanism alone. On real models its sign is decided by incidental activation geometry, which is neither adversarial nor fully generic — which is exactly why we measure it empirically (§5.1) rather than derive it.