When Does the Calibration Source Matter for Low-Bit GPTQ Quantization of LLMs? A Seed-Paired Study

Anonymous authors
Paper under double-blind review

Abstract

Earlier studies found that the calibration source has little effect on four-bit weight quantization of current language models, while a three-bit study of one reasoning model found GPTQ highly sensitive to the calibration domain. We measure both precisions on the same calibration draws. Qwen3-8B, Llama-3.1-8B and Mistral-7B-v0.3 are quantized with weight-only GPTQ at three and four bits using WikiText, random-token or prompted model-generated calibration text, which we call Diverse, with five calibration draws shared across sources and precisions and a budget of 32×128 tokens, and are evaluated on WikiText-2 perplexity and all 14,042 MMLU questions. On the recorded MMLU readout, the Diverse-minus-WikiText contrast is $+0.55$, $+0.29$ and $+0.80$ points at four bits and $+6.45$, $+3.84$ and $+1.99$ points at three bits, every three-bit seed-paired 95% interval lying above zero, while WikiText has the lower perplexity in all 15 model–seed pairs; the source-by-precision interaction averages 3.55 points, with a 95% interval of $[1.82, 5.28]$ over seed blocks. Choosing the source by WikiText perplexity thus forgoes 6.45, 3.92 and 1.99 MMLU points at three bits relative to the best source in hindsight and under one point at four bits, and choosing it by accuracy on held-out questions lowers this three-bit regret for Qwen3-8B and Llama-3.1-8B. Three-bit calibration draws vary 2.5–10 times more than four-bit draws. Registered replacement runs that score every prompt untruncated reproduce the three-bit gain in every seed of all three models, and on 15 fresh Qwen3-8B draws in three registered batches the gain and the interaction average $+4.95$ and $+4.33$ points with intervals above zero, as they do on ARC-Challenge for the ten of these draws scored on it.

1 Introduction

Post-training quantization (PTQ) lowers the numerical precision of a pretrained language model without retraining it. Weight-only methods such as GPTQ (Frantar et al., 2023) and AWQ (Lin et al., 2024) fit their quantized weights to a small calibration set: GPTQ uses the calibration activations to form a layer-wise Hessian, and AWQ uses them to choose channel scales. A few thousand calibration tokens thus decide how a model with billions of parameters behaves on every downstream input. Practitioners usually take this text from a pretraining-style corpus such as WikiText or C4 and judge the result by perplexity.

Three lines of prior work bear on this choice. First, at four bits the calibration source has small effects on current models: quantization is markedly less sensitive than pruning to repeated calibration sets drawn from five natural corpora (Williams & Aletras, 2024), modern models are largely robust to the calibration set (Paglieri et al., 2024), and text sampled from the model itself is a competitive calibration source for GPTQ and AWQ (Williams et al., 2025); recent calibration-selection methods are also developed and evaluated at four bits (He et al., 2025; Monaco et al., 2026; Shihab et al., 2026). Second, perplexity is known to miss capability changes under compression (Jaiswal et al., 2024; Dutta et al., 2024), and capability-oriented calibration curation can raise task accuracy while raising perplexity (He et al., 2025). Third, at three bits, GPTQ applied to a distilled reasoning model gained about ten points of average reasoning accuracy when calibrated on mathematics instead of WikiText, whereas AWQ gained under two (Liu et al., 2025). These observations come from different models, budgets, tasks and single precisions; a benchmark that shares calibration sets across two to

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.

four bits uses one draw per set (Liu et al., 2024a). Such comparisons, including that benchmark's two-seed GPTQ check, leave open the size of the source-by-precision interaction over calibration draws, and whether perplexity, the usual acceptance check for a quantized model, still points to a good source where the effect is large. Both questions call for one design with both precisions; an unpaired factorial could estimate the source-by-precision interaction too, but quantizing the same calibration draws at both precisions controls the variation that the two precisions share through the draw, which, as we show, is large at three bits.

We run such a design for GPTQ, described in Section 3. Three current 7–8B models are quantized at three bits (W3) and four bits (W4) using WikiText, random-token and prompted model-generated calibration text, called Diverse, with five calibration seeds shared across sources and precisions and a budget of 32×128 tokens. Every quantized model is scored on WikiText-2 perplexity and on all 14,042 MMLU questions (Hendrycks et al., 2021). Table 1 compares the design with the closest prior studies. We make three contributions.

1. **At a small budget, the calibration-source effect grows sharply from four to three bits, estimated within draws.** On the recorded MMLU readout, the Diverse-minus-WikiText contrast is $+0.55$, $+0.29$ and $+0.80$ points at W4, the size earlier four-bit studies report for general-purpose sources, and $+6.45$, $+3.84$ and $+1.99$ points at W3, positive in all 15 model–seed pairs. Paired by seed, the interaction averages 3.55 points, is positive in $14/15$ pairs and in all five seed blocks, with a 95% interval of $[1.82, 5.28]$ over the blocks and a per-model interval excluding zero for Qwen3-8B. On three batches of five fresh Qwen3-8B draws, each registered before it ran, the W3 gain is $+3.89$, $+5.13$ and $+5.82$ points, each with an interval above zero; over the 15 fresh draws the gain is $+4.95$ points and the interaction $+4.33$ points, both with intervals above zero, and over the ten of them scored on the challenge set of ARC, the AI2 Reasoning Challenge (Clark et al., 2018), both are positive in every draw with intervals above zero. The multi-point three-bit sensitivity previously seen with domain-matched data on a reasoning model thus also arises for general-purpose models with a task-agnostic source, as Sections 4.1, 4.2 and 4.4 show.
2. **Perplexity selects the wrong source where the choice matters.** Choosing among the three sources by WikiText perplexity selects WikiText in $15/15$ W3 seed triples, yet WikiText has the highest MMLU accuracy in $0/15$ of them. The gap to the best of the three sources in hindsight averages 6.45, 3.92 and 1.99 points and is positive in every triple; at W4 the same rule forfeits only 0.55, 0.48 and 0.80 points. In separate registered runs, choosing the source by accuracy on a disjoint half of the questions, with the other calibration draws, has a W3 regret on the remaining half $+6.46$ and $+6.97$ points below that of perplexity selection for Qwen3-8B and Llama-3.1-8B, with intervals above zero. Diverse also exceeds random-token calibration at W3 on Qwen3-8B and Mistral-7B-v0.3, although WikiText has the lowest perplexity of the three sources in every model–seed pair; Sections 4.1 and 4.3 give these results. We thus measure what the disagreement between perplexity and capability costs as a function of precision, and what validation on the deployment metric recovers.
3. **Three-bit comparisons call for several calibration draws.** For every model and source, the standard deviation of MMLU accuracy across calibration draws is 2.5–10 times larger at W3 than at W4, for example 2.69 versus 0.27 points for WikiText on Qwen3-8B, while three executions of each of eight Qwen3-8B configurations on identical calibration inputs agree exactly. As plug-in planning estimates from the observed SDs, a 95% interval of ± 2 points on a W3 source contrast takes about 12, 12 and 4 paired draws, whereas four-bit calibration studies use three to five (He et al., 2025; Shihab et al., 2026); Section 4.5 gives the details.

Audits of the instrument support these readings: the W3 mean gain remains positive on the questions whose prompts fit the 512-token window, and replacement runs that score every prompt untruncated reproduce it in every seed of all three models, as Section 4.4 reports.

2 Related Work

Calibration-based post-training quantization. GPTQ (Frantar et al., 2023) quantizes one column at a time and compensates each rounding error on the remaining columns through a Hessian estimated from calibration activations; its reference setup uses 128 random 2,048-token segments of C4. AWQ (Lin et al., 2024) and SmoothQuant (Xiao et al., 2023) use calibration activations to choose channel scales, and OmniQuant (Shao et al., 2024) learns clipping thresholds and equivalent

Table 1: Closest studies of calibration data for LLM weight quantization. "Draws": calibration sets per source and cell. "Paired": the same calibration draws are quantized at two or more precisions; with several draws, the source-by-precision interaction then has a draw-level interval. *A second calibration seed is checked for GPTQ.

Study	Quantizers; bits	Calibration sources	Budget	Draws	Paired	Evaluation
Williams & Aletras (2024)	GPTQ, SpQR; 4	five natural corpora	128×2048	10	no	10 zero-shot tasks, PPL
Williams et al. (2025)	GPTQ, AWQ; 4	C4, WikiText, vocabulary, Cosmopedia, self-generated	128×2048	5	no	10 zero-shot tasks
Liu et al. (2024a)	four methods; 2–4	C4, task datasets	128 examples	1*	yes	40+ datasets
Liu et al. (2025)	GPTQ, AWQ; 3	WikiText, Numina-Math	128×2048	1	no	5 reasoning tasks
He et al. (2025)	GPTQ, AWQ; 4	curated mixtures	128×2048	5	no	commonsense, math, code, PPL
Shihab et al. (2026)	GPTQ, AWQ; 4	outlier-coverage selection	$(32\text{–}256) \times 2048$	3	no	MMLU, 3 others, PPL
This work	GPTQ; 3, 4	WikiText, random tokens, prompted generation	32×128	5	yes	all of MMLU, WikiText-2 PPL

transforms on calibration data, so the quantizers differ in how much of the fitted model depends on the calibration text.

How much the calibration source matters. Williams & Aletras (2024) gave the first extensive study, with ten calibration sets from each of five natural corpora, GPTQ and SpQR at four bits and two pruning methods; quantization was markedly less sensitive to the calibration set than pruning. Paglieri et al. (2024) found current models largely robust to the calibration set under four-bit weight-only and eight-bit weight-activation quantization, and Liu et al. (2024a) built a benchmark of more than 40 datasets that varies the calibration distribution at two to four bits, calibrating 7B models on C4 or on task datasets with one draw of 128 calibration examples per set and a second calibration seed checked for GPTQ, and found that calibrating on the test distribution is not always best. Like ours, their design shares each calibration set across bit widths; repeating it over five draws turns the cross-precision difference into an interaction with a seed-level interval, as in Section 4.2, and exposes the three-bit draw variance that single-draw comparisons absorb, measured in Section 4.5. Pruning is more sensitive (Bandari et al., 2024; Ji et al., 2025), multilingual models benefit from language-diverse calibration (Chimoto et al., 2026), and concurrent work on financial models argues that calibration data matter in proportion to the task damage that compression causes, which was small for four-bit quantization (Ye et al., 2026). The clearest large source effect for quantization that we found is at three bits: on a distilled 1.5B reasoning model, switching GPTQ's calibration set from WikiText to mathematics raised average reasoning accuracy from 28.3% to 38.1%, while AWQ changed by 1.8 points (Liu et al., 2025). That comparison uses one calibration set per cell with three evaluation seeds, a task-matched source and no four-bit counterpart.

Model-generated calibration data. Self-calibration (Williams et al., 2025) samples calibration text from the model itself, starting from the beginning-of-sequence token, and finds it competitive with real data for GPTQ and AWQ at four bits, using five calibration sets of 128×2048 tokens on models including Llama-3.1-8B and Mistral-7B-v0.3, with random vocabulary the weakest source. Ji et al. (2025) synthesize calibration data by self-generation for pruning, FAQ (Xiao et al., 2026) regenerates calibration samples with a larger model of the same family, and LLM-QAT (Liu et al., 2024b) uses self-generated data for quantization-aware training. Our Diverse source conditions the generation on prompts: the model continues 32 short prompts from different domains. We measure how the effect of such a source depends on precision and how it should be evaluated, and Section 4.4 adds an equal-budget comparison with unprompted self-generation.

Selecting calibration data. COLA (He et al., 2025) analyses how calibration composition affects specific capabilities and selects samples for activation-space representativeness and diversity; ZipCal (Monaco et al., 2026) maximizes lexical diversity without running the model; COVERCAL (Shihab et al., 2026) selects samples that cover outlier channels; TASA (Wang et al., 2026) jointly chooses task-aware calibration data and bit allocation. The quantization experiments of COLA, ZipCal and COVERCAL are at four bits with default budgets of 128×2048 tokens; COLA and COVERCAL average over five and three seeds, and COVERCAL states that three bits lie outside its

experiments. We measure the three-bit regime, where calibration draws vary several times more than at four bits.

Perplexity versus downstream capability. Perplexity can miss capability losses of compressed models (Jaiswal et al., 2024), and similar accuracies can hide many changed answers (Dutta et al., 2024). For calibration data specifically, capability-oriented curation can raise task accuracy while raising perplexity (He et al., 2025), perplexity-based layer sensitivity can misidentify layers that matter for reasoning (Wang et al., 2026), and in financial models perplexity ranks quantizers differently from numerical question answering (Ye et al., 2026). We measure what this disagreement costs in calibration-source selection at each precision and what validation on held-out questions recovers.

Position of this study. This study estimates the source-by-precision interaction on shared calibration draws, confirms it on fresh draws, measures the accuracy cost of perplexity-based source selection at each precision, and quantifies how much more calibration draws vary at three bits than at four. Task-gradient saliency (Xiao et al., 2025) and the attribution of calibration effects are discussed in Appendix E.

3 Experimental Design

Unit of replication. The experimental unit is one model quantized with one calibration source and one calibration seed. Scoring more test questions sharpens the measurement of that fitted model but does not add calibration draws, so all uncertainty statements below are over seeds. Seeds 42–46 are shared across sources and across W3/W4, which makes both source contrasts and source-by-precision interactions seed-paired; wherever both were retained, in 21 of 45 combinations, the W3 and W4 calibration registries are identical token for token, as detailed in Appendix D.

Models and quantizers. We use Qwen3-8B (Yang et al., 2025), Llama-3.1-8B (Grattafiori et al., 2024) and Mistral-7B-v0.3 (Jiang et al., 2023), one fixed public release of each. GPTQ quantizes the weights only, at W3 and W4 with group size 128, symmetric quantization, dampening λ equal to 5% of the mean Hessian diagonal and activation ordering disabled; activations remain in BF16. This gives 3 models $\times$ 2 precisions $\times$ 3 sources $\times$ 5 seeds $= 90$ quantized models, all reported in Appendix A.

Calibration sources. Each source supplies 32 texts, each of which encodes to exactly 128 content tokens under the quantized model's tokenizer, 4,096 tokens in total; the quantizer receives the texts, and the Llama and Mistral tokenizers prepend a BOS token, giving 129 positions per text. *WikiText*: WikiText-2 training passages (Merity et al., 2017), shuffled by the seed and concatenated into a token stream; the shuffle acts on one passage list, so seed s fixes the same passage order for all three models, and Appendix D.1 lists the consumed passages and their replay. *Random*: token IDs drawn uniformly from the non-special vocabulary; a 128-token candidate is kept only if its decoded text re-encodes to exactly 128 tokens, so the source is uniform tokens conditioned on a length-preserving round trip. *Diverse*: the unquantized model continues 32 fixed prompts on code, SQL, mathematics, physics, history, biology, economics and other topics, including a letter, a recipe and a Spanish sentence, listed in Appendix D.4; each continuation has 160 new tokens at temperature 0.8 and top-p 0.9, with per-prompt seeds derived from the calibration seed, and prompts and completions are concatenated into one stream. Streams are cut into 128-token pieces whose decoded text re-encodes to exactly 128 tokens. The design equalizes the number of calibration tokens, not their information content or construction cost; Diverse additionally costs 32 generations of 160 tokens with the unquantized model. No evaluation data enter any calibration set.

The studied regime. The budget of 32×128 tokens is 64 times smaller than GPTQ's reference setup of 128×2048 tokens, and short sequences limit the context seen by each activation. All three models have hidden size 4,096, and their MLP down-projections have input dimension 12,288 for Qwen3-8B and 14,336 for Llama and Mistral, so for those layers the empirical input covariance built from at most 4,128 token positions is rank-deficient and GPTQ's inverse relies on dampening. Every accuracy conclusion below is stated for this low-budget regime. Registered W3 cells of Llama-3.1-8B at 8×128 and 128×128 tokens, three seeds each and listed in Table 12, keep the

perplexity ordering: WikiText has the lower perplexity in every seed, by $+1.73$ $[+1.37, +2.10]$ and $+2.43$ $[+2.11, +2.76]$ points. With the generator fixed to an equal-share Diverse builder, which gives each of the 32 prompts $1/32$ of the budget, the W3 MMLU gain of Qwen3-8B over WikiText in seeds 42–44 is $+6.15$ $[+0.44, +11.87]$ points at 32×128, $+4.79$ $[+2.43, +7.15]$ at 8×512, an approximately 4,096-token longer-context condition, and $+1.38$ $[+0.21, +2.56]$ at 128×128, where WikiText keeps a perplexity advantage of $+2.98$ $[+1.36, +4.59]$; the interaction is $+5.73$ $[+0.64, +10.82]$ and $+3.98$ $[+2.89, +5.07]$ at the two nominal 4,096-token budgets, all listed in Tables 15 and 16. The effect thus holds with 512-token contexts, and its point estimate is smaller at the larger budget.

Evaluation. *Perplexity*: the 1,833 lines of the raw WikiText-2 test split that are longer than 100 characters after whitespace stripping, the same filter that selects the training passages of the WikiText source, each tokenized independently and truncated to 256 tokens; the loss is averaged over all real next-token targets, 268,891, 263,553 and 286,770 targets for Qwen, Llama and Mistral. Contexts never cross passages, so values are comparable within a model but not with sliding-window WikiText perplexity; the replacement runs also record sliding-window perplexity on the same filtered passages, and for Qwen3-8B and Llama-3.1-8B it favours WikiText over every other source in every seed, as Table 13 lists. *MMLU*: all 14,042 test questions, zero-shot, formatted as "Question: … A. … D. … Answer:". The prediction is the largest of the four next-token logits for " A", " B", " C", " D", each a single distinct token, and prompts are truncated on the right at 512 tokens. The contrasts and intervals of Sections 4.1 to 4.3 and 4.5 are computed on this recorded readout; Section 4.4 restates the W3 contrast on the questions whose prompts fit the 512-token window and on untruncated re-scoring.

Estimands and inference. For model m, precision b, seed s and sources c, c', the paired source contrast and the source-by-precision interaction are

$$d_{ms}^{(b)} = y_{ms}^{(b,c)} - y_{ms}^{(b,c')},$$
$$\iota_{ms} = d_{ms}^{(3)} - d_{ms}^{(4)}.$$

Operationally, the source matters for a metric at a given precision and budget when the seed-paired contrast on that metric has an interval excluding zero, and a rule for choosing the source is judged by its regret on held-out questions. Primary inference is per model: means with two-sided 95% paired-t intervals over $n = 5$ seeds with $\mathrm{df} = 4$, counts of positive seeds, and Holm adjustment across the three models. Because the WikiText draw of a seed is shared by the three models, model–seed pairs with the same seed need not be independent. For cross-model statements we therefore treat each seed as a block, average ι_{ms} over models within the block, and report a paired-t interval over the five block means. All models of the main study were quantized and evaluated in one original set of runs, and every source and precision contrast outside the replacement runs of Section 4.4 re-analyses those original records; the three BF16 references of Table 3 and the replacement runs were evaluated afterwards with the same instruments, builders and quantizer settings and are never pooled with the original records. The per-model analyses, the truncation bound, the composition count and the dev/test selection analysis were recorded in author-maintained protocols after the original runs and before being computed according to those records; Table 9 lists the registered confirmatory tests in full. One registered rule was adapted: it required re-scoring the truncated MMLU items before any MMLU claim was kept; for the original records we instead state the W3 claim together with its intact-question bound, and the replacement runs re-score every prompt untruncated, as Section 4.4 reports. The seed-block and perplexity-selection analyses were added afterwards and are secondary.

4 Results

4.1 At three bits, calibration source reverses the metric ranking

Table 2 gives the primary contrast. At W3, Diverse calibration improves MMLU accuracy over WikiText by +6.45 points on Qwen3-8B, +3.84 on Llama-3.1-8B and +1.99 on Mistral-7B-v0.3, and every 95% interval lies above zero. The improvement holds in all 15/15 model–seed pairs; because the three models share each seed's WikiText order, cross-model inference counts the five seed blocks instead: averaged within blocks the W3 gain is 4.09 points, with 95% interval $[2.37, 5.82]$ and all five blocks positive. In the same quantized models WikiText calibration gives the lower WikiText

Table 2: Diverse-minus-WikiText contrasts under GPTQ, paired over five calibration seeds. Accuracy in percentage points, positive when Diverse is better; perplexity differences are positive when Diverse is worse. Intervals: two-sided 95% paired-t over seeds with df $= 4$; "+": seeds with a positive difference.

		ΔMMLU, pp			ΔPPL		
Model	Bits	Mean	95% CI	+	Mean	95% CI	+
Qwen3-8B	W3	$+6.45$	$[+2.56, +10.34]$	5/5	$+2.77$	$[+1.42, +4.11]$	5/5
Llama-3.1-8B	W3	$+3.84$	$[+0.08, +7.60]$	5/5	$+2.84$	$[+2.57, +3.11]$	5/5
Mistral-7B-v0.3	W3	$+1.99$	$[+0.46, +3.53]$	5/5	$+0.84$	$[+0.75, +0.93]$	5/5
Qwen3-8B	W4	$+0.55$	$[+0.23, +0.87]$	5/5	$+0.70$	$[+0.34, +1.05]$	5/5
Llama-3.1-8B	W4	$+0.29$	$[-0.56, +1.15]$	3/5	$+0.25$	$[+0.21, +0.29]$	5/5
Mistral-7B-v0.3	W4	$+0.80$	$[+0.24, +1.36]$	5/5	$+0.10$	$[+0.07, +0.13]$	5/5

perplexity, by +2.77, +2.84 and +0.84 points, in all 15/15 pairs. The two metrics therefore rank the two sources in opposite order for every model and every seed. Holm adjustment over the three models leaves every W3 accuracy contrast significant, with $p = 0.030$, 0.047 and 0.046.

The cost of using perplexity as the acceptance check can be measured against the best source in hindsight. For each model and seed we choose among the three sources the one with the lowest WikiText perplexity. At W3 this rule selects WikiText in $15/15$ seed triples, yet WikiText has the highest MMLU accuracy in $0/15$ of them. The gap to the most accurate of the three sources, which is chosen with knowledge of the test results and so bounds what any rule selecting among them could recover on this test set, averages 6.45, 3.92 and 1.99 points and is at least 0.24 points in every triple. At W4 the rule again selects WikiText in $15/15$ triples, and the gap is only 0.55, 0.48 and 0.80 points.

Validation on the deployment metric is a practical alternative. This registered analysis of the untruncated replacement runs of Section 4.4, listed in Table 14, uses WikiText and Diverse as candidate sources for Qwen3-8B and these two plus self-generated text and an external mixture for Llama-3.1-8B and Mistral-7B-v0.3; we split the questions within each subject into a dev and a test half. At W3 the perplexity rule has a test-half regret of 6.46 and 7.14 points on Qwen3-8B and Llama-3.1-8B; choosing the source with the best dev-half accuracy over the other four calibration draws, which picks Diverse in every Qwen3-8B seed and self-generated text in every Llama-3.1-8B seed, lowers it by $+6.46$ $[+2.96, +9.97]$ and $+6.97$ $[+2.56, +11.37]$ points, intervals that are conditional on one fixed question split and on selections that share draws. Validating each fitted model on its own dev half picks the test-best source in $5/5$, $3/5$ and $4/5$ seeds, whereas perplexity picks it in none.

The accuracy gain is large relative to the damage done by quantization. Averaged over seeds, W3 GPTQ with WikiText calibration reaches 49.8%, 39.2% and 44.5% MMLU on the three models, whereas W4 with the same source reaches 69.0%, 59.8% and 55.5%, as Table 3 shows. Diverse calibration closes 34%, 19% and 18% of this W3-to-W4 gap at identical bit width and token budget. Against the unquantized models at 71.3%, 63.5% and 58.2%, W3 WikiText calibration loses 21.5, 24.3 and 13.7 points, of which Diverse recovers 30.0%, 15.8% and 14.6%.

4.2 The source effect is concentrated at three bits

At W4 the same source choice moves MMLU by only +0.55, +0.29 and +0.80 points, and perplexity favours WikiText in every W4 pair by less than one point. Figure 1 shows the W3 and W4 contrast for each seed, and Table 4 gives their difference, the source-by-precision interaction ι_{ms}, for every model and seed. It is positive in 14/15 model–seed pairs, with mean 3.55 points. Treating each seed as a block, so that the shared WikiText draw is counted once, all five block means are positive; their mean is 3.55 points with 95% interval $[1.82, 5.28]$ and paired-t $p = 0.0047$. Per model, the interaction is largest for Qwen3-8B, $+5.90$ points with 95% interval $[+1.99, +9.81]$ and Holm-adjusted $p = 0.041$; its point estimate is $+3.55$ for Llama-3.1-8B, whose untruncated replacement cells in Section 4.4 give an interval above zero, and $+1.19$ for Mistral-7B-v0.3. The three models agree in direction and differ in magnitude, in the same order as their W4-to-W3 perplexity increase under WikiText calibration, 6.9, 5.4 and 2.0 points.

Table 3: Absolute GPTQ quality averaged over the same five calibration seeds; BF16 rows give the unquantized models on the same instruments. PPL and MMLU use the fixed within-model instruments; raw levels should not be compared as standardized cross-model difficulty. Paired contrasts are in Table 2.

		PPL ↓		MMLU, % ↑	
Model	Bits	WikiText	Diverse	WikiText	Diverse
Qwen3-8B	3	26.381	29.147	49.791	56.240
Qwen3-8B	4	19.501	20.199	69.031	69.583
Qwen3-8B	BF16	18.428		71.293	
Llama-3.1-8B	3	16.925	19.763	39.178	43.021
Llama-3.1-8B	4	11.523	11.770	59.809	60.103
Llama-3.1-8B	BF16	10.607		63.495	
Mistral-7B-v0.3	3	11.750	12.594	44.501	46.493
Mistral-7B-v0.3	4	9.758	9.857	55.489	56.291
Mistral-7B-v0.3	BF16	9.435		58.190	

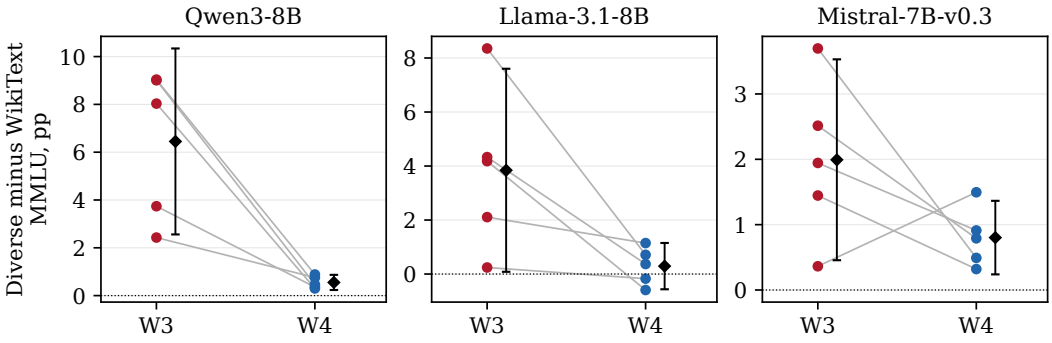

Figure 1: Diverse-minus-WikiText MMLU difference for each calibration seed at W3 in red and W4 in blue; grey lines join the same seed. Diamonds and bars: seed mean and 95% paired-t interval.

4.3 Generated text also outperforms random-token calibration

Random-token calibration controls for the presence of any linguistic content. On Qwen3-8B and Mistral-7B-v0.3, Diverse exceeds random tokens by +9.76 and +2.32 MMLU points at W3, with 95% intervals $[+8.25, +11.28]$ and $[+1.20, +3.45]$, and by +1.32 and +0.48 points at W4, with intervals above zero; Table 5 lists all rows, including Llama-3.1-8B, and Table 10 the WikiText-minus-Random contrasts. Perplexity ranks the sources differently: WikiText has the lowest perplexity of the three sources in every model–seed pair at both precisions, and at W3 it improves on random tokens by 5.06, 4.03 and 0.73 perplexity points, a larger margin than Diverse's. At this budget, then, encyclopedic text most improves the metric it resembles, and the largest accuracy gains come from the generated source. Diverse differs from WikiText in several respects at once, being model-generated, multi-domain and prompt-conditioned; Section 4.4 adds unprompted self-generation, an external multi-domain mixture and a narrower set of eight encyclopedic-prose prompts as controls.

4.4 Audits of the instrument and of the intervention

MMLU prompt length. We re-tokenized every MMLU prompt with each model's tokenizer, reproducing the evaluation's prompt construction. The median prompt has 75–85 tokens; 89, 76 and 143 of the 14,042 prompts of Qwen, Llama and Mistral, that is 0.54–1.02%, exceed 512 tokens, and right truncation removes the final "Answer:" cue from all of them. Because these items contribute at most a correct answers to either fitted model, a reported contrast d over all N questions implies a contrast of at least $(dN - a)/(N - a)$ on the $N - a$ questions whose prompts fit the window, as Table 7 lists. On those questions the mean W3 gain is at least +5.85, +3.32 and +0.98 points, positive in every model, and the per-seed bound is positive in 5/5, 4/5 and 4/5 seeds.

Table 4: Source-by-precision interaction $\iota_{ms} = d^{(3)}_{ms} - d^{(4)}_{ms}$ of the Diverse-minus-WikiText MMLU contrast for every model and calibration seed, in points. The last row averages the three models within each seed; these five block means are the units of the seed-block interval.

Model	Seed 42	Seed 43	Seed 44	Seed 45	Seed 46	Mean
Qwen3-8B	$+8.15$	$+8.55$	$+1.67$	$+7.73$	$+3.38$	$+5.90$
Llama-3.1-8B	$+3.96$	$+0.41$	$+4.77$	$+7.64$	$+0.96$	$+3.55$
Mistral-7B-v0.3	$+1.13$	-1.13	$+3.20$	$+1.03$	$+1.72$	$+1.19$
Seed block, mean over models	$+4.41$	$+2.61$	$+3.22$	$+5.47$	$+2.02$	$+3.55$

Table 5: Paired GPTQ source contrasts involving Diverse at W3 and W4. Each accuracy cell gives the mean MMLU difference in points and its 95% paired-t interval over five seeds; the row below gives the mean perplexity difference, negative when the first source is better.

Contrast	Qwen3-8B	Llama-3.1-8B	Mistral-7B-v0.3
W3			
Diverse − WikiText	$+6.45$ $[+2.56, +10.34]$	$+3.84$ $[+0.08, +7.60]$	$+1.99$ $[+0.46, +3.53]$
ΔPPL	$+2.77$	$+2.84$	$+0.84$
Diverse − Random	$+9.76$ $[+8.25, +11.28]$	$+4.48$ $[-0.05, +9.00]$	$+2.32$ $[+1.20, +3.45]$
ΔPPL	-2.30	-1.19	$+0.11$
W4			
Diverse − WikiText	$+0.55$ $[+0.23, +0.87]$	$+0.29$ $[-0.56, +1.15]$	$+0.80$ $[+0.24, +1.36]$
ΔPPL	$+0.70$	$+0.25$	$+0.10$
Diverse − Random	$+1.32$ $[+0.95, +1.70]$	$+0.48$ $[-0.54, +1.50]$	$+0.48$ $[+0.13, +0.83]$
ΔPPL	-0.54	-0.09	-0.01

Replacement runs. Under a registered protocol described in Appendix B, we re-ran all five seeds of the W3 and W4 WikiText and Diverse cells of the three models and of their W3 cells with two further sources, scoring every prompt untruncated. Every rebuilt WikiText and Diverse registry equals the retained original one, and 53 of 60 matched cells reproduce the original counts exactly; the other 7 differ by at most 1.55 points at W3 and 1.08 at W4 with identical inputs. On the untruncated instrument the W3 Diverse-minus-WikiText gain is $+6.45$ points $[+2.52, +10.37]$ for Qwen3-8B, $+4.27$ $[+1.16, +7.38]$ for Llama-3.1-8B and $+2.03$ $[+0.44, +3.62]$ for Mistral-7B-v0.3, positive in all five seeds of each model, and the source-by-precision interaction is $+6.06$ $[+2.46, +9.66]$ for Qwen3-8B and $+4.05$ $[+0.96, +7.15]$ for Llama-3.1-8B, positive in all five seeds of both; Table 11 lists all rows. As the seeds are regenerated, these runs confirm the instrument on the same calibration draws.

Fresh calibration draws. We registered three batches of five fresh Qwen3-8B seeds, 47–51, 52–56 and 57–61, never used by any builder, each before running it. The W3 gain is $+3.89$ points $[+0.25, +7.54]$, $+5.13$ $[+2.75, +7.52]$ and $+5.82$ $[+0.25, +11.39]$ in the three batches. Over the 15 fresh draws the gain is $+4.95$ $[+3.20, +6.70]$ and the interaction $+4.33$ $[+2.63, +6.03]$. Over all 20 draws, seeds 42–61, they are $+5.32$ $[+3.85, +6.79]$ and $+4.76$ $[+3.33, +6.19]$, with half-widths of 1.47 and 1.43 points; over seeds 42–56 the half-widths, 1.52 and 1.38, already meet the registered precision objective of two points. Tables 15, 16 and 21 list every batch. The second and third batches were also scored on all 1,172 ARC-Challenge questions with a zero-shot letter readout: over these ten draws Diverse gains $+8.41$ $[+4.50, +12.32]$ points at W3, with interaction $+7.74$ $[+3.95, +11.53]$, both positive in all ten, and on ARC-Easy $+6.66$ $[+1.99, +11.33]$ and $+6.42$ $[+1.69, +11.15]$, as Tables 17 and 22 list.

Other sources and prompt sets. The W3 gain also extends beyond our prompted construction: on Llama-3.1-8B, unprompted self-generated text (Williams et al., 2025) and an external multi-domain text mixture exceed WikiText by $+7.31$ $[+3.10, +11.52]$ and $+4.72$ $[+0.43, +9.00]$ points at the same budget, and on Qwen3-8B the mixture exceeds it by $+2.97$ $[+0.45, +5.49]$. The full prompt inventory is not needed for the gain: with generator, sampling and slot structure fixed, filling all 32 prompt slots of the equal-share builder with eight encyclopedic-prose prompts on history, geography,

Table 6: GPTQ results per model, precision and calibration source: mean ± sample SD over five calibration seeds.

		W3		W4	
Model	Source	PPL	MMLU, %	PPL	MMLU, %
Qwen3-8B	WikiText	26.381 ± 0.612	49.791 ± 2.692	19.501 ± 0.126	69.031 ± 0.268
Qwen3-8B	Random	31.443 ± 0.455	46.476 ± 0.991	20.744 ± 0.398	68.261 ± 0.386
Qwen3-8B	Diverse	29.147 ± 0.683	56.240 ± 1.345	20.199 ± 0.315	69.583 ± 0.376
Llama-3.1-8B	WikiText	16.925 ± 0.107	39.178 ± 2.841	11.523 ± 0.030	59.809 ± 0.571
Llama-3.1-8B	Random	20.955 ± 0.236	38.546 ± 2.030	11.856 ± 0.085	59.624 ± 0.411
Llama-3.1-8B	Diverse	19.763 ± 0.150	43.021 ± 2.943	11.770 ± 0.042	60.103 ± 0.432
Mistral-7B-v0.3	WikiText	11.750 ± 0.041	44.501 ± 0.678	9.758 ± 0.019	55.489 ± 0.272
Mistral-7B-v0.3	Random	12.483 ± 0.113	44.169 ± 1.869	9.864 ± 0.025	55.811 ± 0.282
Mistral-7B-v0.3	Diverse	12.594 ± 0.093	46.493 ± 1.548	9.857 ± 0.019	56.291 ± 0.403

biology and the history of mathematics gains $+7.53$ $[+3.44, +11.62]$ points over WikiText at W3 on Qwen3-8B, with interaction $+7.02$ $[+3.60, +10.45]$, positive in all four registered seeds of Table 20. This condition narrows the prompt set and the prose style; it does not isolate a single knowledge domain.

4.5 At three bits, runs with different calibration draws vary strongly

The seed-paired design also measures how much accuracy varies across calibration-seeded runs. The SD of the paired W3 contrast across seeds is 3.13, 3.03 and 1.24 points, against 0.26, 0.69 and 0.45 at W4. The per-source spread shows the same pattern for every model and source: the standard deviation of MMLU accuracy across the five calibration draws is 2.5 to 10 times larger at W3 than at W4, for example 2.69 versus 0.27 points for WikiText on Qwen3-8B in Table 6. For Qwen3-8B, a registered repetition separates the draw from the execution: two further executions of each of eight configurations, W3 and W4 with WikiText and Diverse for seeds 47 and 48, quantized the identical captured calibration inputs and reproduced the original correct-answer counts and perplexities exactly in all eight, against a between-draw SD of 1.98 points at W3 and 0.36 at W4 for the same model and sources over seeds 42–51, as Table 18 lists. Across the original and replacement executions of all three models, a conditional descriptive estimate under the assumptions stated in Appendix B puts the execution SD at 0.32 and 0.20 points at W3 and W4, about 2% and 25% of the between-draw variance. As plug-in planning estimates from the observed paired SDs, 95% half-widths of two and one points need 12, 12, 4 and 41, 38, 9 paired seeds for the three models. For comparison, the 20 Qwen3-8B draws of Section 4.4 reach half-widths of 1.47 and 1.43 points for the W3 contrast and the interaction.

5 Discussion and Conclusion

Practical implications. At 32×128 tokens, Diverse improves three-bit GPTQ MMLU accuracy by 2.0–6.4 points over WikiText, despite WikiText's lower perplexity; the gain persists with untruncated scoring, 15 fresh calibration draws and, for Qwen3-8B, ARC. Four-bit source differences are smaller. Building a Diverse set costs 32 generations of 160 tokens with the unquantized model. Validate calibration sources on held-out downstream questions at the deployed precision and repeat calibration draws; in the tested Qwen3-8B configurations, re-executing identical inputs reproduces the result exactly and so does not measure the between-draw variation.

Scope. The evidence concerns weight-only GPTQ on three 7–8B models, mainly at 32×128 tokens, with budget and context extensions in Tables 12, 15 and 16. AWQ, OmniQuant and other quantizers are outside its scope. The main runs disable activation ordering; Table 16 lists a three-seed Qwen3-8B block with it enabled. Diverse changes generation and prompting together; an eight-prompt encyclopedic-prose set also retains the gain, and the causal mechanism is left open. The supported conclusion is a precision-dependent calibration-source effect under the tested budgets and readouts, measured with paired draws and explicit evaluation controls.

AI USE STATEMENT

In this work, we used generative AI tools to design and improve experiments and to implement methods. We have not used generative AI tools to support qualitative and thematic data analysis, and formulating questions for surveys or interviews and transcribing recordings are not applicable to this work. Additionally, we used generative AI tools to edit the paper to improve readability and to create or edit software code. We have reviewed all AI-assisted work: AI-assisted code was tested before use, and every reported number is computed from recorded experimental outputs by analysis scripts. Separately from these uses, the open-weight models under study generate the Diverse and self-generated calibration text as part of the experimental design described in Section 3 and Appendix B. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This study involves no human subjects and no personal data. All three models, Qwen3-8B, Llama-3.1-8B and Mistral-7B-v0.3, are publicly released and used under their licenses, and all evaluation and calibration data are public benchmarks and corpora, namely WikiText-2, MMLU, ARC and the public datasets of the external mixture; the only new text is calibration text sampled from the studied models, which is released with the supplementary material for reproduction. The work measures how calibration data affect low-bit weight quantization. Its main foreseeable impact is on the reliability of compressed models: our results indicate that a calibration set validated only by perplexity can cost several points of downstream accuracy at three bits, and we recommend validating on the downstream metric with repeated calibration draws. We see no dual-use or fairness risk specific to this measurement beyond those of the underlying models, whose outputs may reflect biases of their training data; quantization does not add capabilities to a model.

REPRODUCIBILITY STATEMENT

The experimental design, calibration builders, quantizer settings and evaluation instruments are specified in Section 3 and Appendix D.1; the Diverse prompts and the contents of the supplementary material are described in Appendix D. Complete per-model results are in Appendix A, the registered replacement runs and their protocol in Appendix B, the assumptions and complete proof of the estimation bound in Appendix F, and every registered analysis, including the dev/test source selection, in Appendix G. The anonymous supplementary material contains the per-run records, per-question MMLU outputs of the replacement runs, the calibration token registries and captured quantizer inputs, the quantization, evaluation and analysis code, and scripts that recompute the reported tables and statistics from these records.

REFERENCES

Abhinav Bandari, Lu Yin, Cheng-Yu Hsieh, Ajay Kumar Jaiswal, Tianlong Chen, Li Shen, Ranjay Krishna, and Shiwei Liu. Is C4 dataset optimal for pruning? An investigation of calibration data for LLM pruning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 18089–18099, 2024. doi: 10.18653/v1/2024.emnlp-main.1004. URL https://aclanthology.org/2024.emnlp-main.1004/.

Sahil Chaudhary. Code Alpaca: An instruction-following LLaMA model for code generation. https://github.com/sahil280114/codealpaca, 2023.

Everlyn Asiko Chimoto, Mostafa Elhoushi, and Bruce Bassett. Calibrating beyond English: Language diversity for better quantized multilingual LLMs. In Vera Demberg, Kentaro Inui, and Lluís Marquez (eds.), *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 4822–4838, Rabat, Morocco, March 2026. Association for Computational Linguistics. doi: 10.18653/v1/2026.eacl-long.223. URL https://aclanthology.org/2026.eacl-long.223/.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? Try ARC, the AI2 reasoning challenge. *arXiv preprint arXiv:1803.05457*, 2018.

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. URL https://arxiv.org/abs/2110.14168.

Abhinav Dutta, Sanjeev Krishnan, Nipun Kwatra, and Ramachandran Ramjee. Accuracy is not all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, pp. 124347–124390, 2024. doi: 10.52202/079017-3950. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/e0e956681b04ac126679e8c7dd706b2e-Abstract-Conference.html.

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 (ICLR)*, 2023. URL https://arxiv.org/abs/2210.17323.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The Llama 3 herd of models. arXiv preprint arXiv:2407.21783, 2024. URL https://arxiv.org/abs/2407.21783.

Bowei He, Lihao Yin, Hui-Ling Zhen, Shuqi Liu, Han Wu, Xiaokun Zhang, Mingxuan Yuan, and Chen Ma. Preserving LLM capabilities through calibration data curation: From analysis to optimization. In *Advances in Neural Information Processing Systems*, volume 38, pp. 58531–58572, 2025. doi: 10.52202/085713-1961. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/54dcf25318f9de5a7a01f0a4125c541e-Abstract-Conference.html.

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 (ICLR)*, 2021. URL https://arxiv.org/abs/2009.03300.

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 (ICLR)*, 2024. URL https://arxiv.org/abs/2310.01382.

Yixin Ji, Yang Xiang, Juntao Li, Qingrong Xia, Ping Li, Xinyu Duan, Zhefeng Wang, and Min Zhang. Beware of calibration data for pruning large language models. In *International Conference on Learning Representations (ICLR)*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/2ede933e10afa991a10b6f36b6522129-Abstract-Conference.html.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. Mistral 7B. arXiv preprint arXiv:2310.06825, 2023. URL https://arxiv.org/abs/2310.06825.

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)*, volume 6, pp. 87–100, 2024. URL https://proceedings.mlsys.org/paper_files/paper/2024/hash/42a452cbafa9dd64e9ba4aa95cc1ef21-Abstract-Conference.html.

Ruikang Liu, Yuxuan Sun, Manyi Zhang, Haoli Bai, Xianzhi Yu, Tiezheng Yu, Chun Yuan, and Lu Hou. Quantization hurts reasoning? An empirical study on quantized reasoning models. In *Conference on Language Modeling (COLM)*, 2025. URL https://arxiv.org/abs/2504.04823.

Yijun Liu, Yuan Meng, Fang Wu, Shenhao Peng, Hang Yao, Chaoyu Guan, Chen Tang, Xinzhu Ma, Zhi Wang, and Wenwu Zhu. Evaluating the generalization ability of quantized LLMs: Benchmark, analysis, and toolbox. arXiv preprint arXiv:2406.12928, 2024a. URL https://arxiv.org/abs/2406.12928.

Zechun Liu, Barlas Oguz, Changsheng Zhao, Ernie Chang, Pierre Stock, Yashar Mehdad, Yangyang Shi, Raghuraman Krishnamoorthi, and Vikas Chandra. LLM-QAT: Data-free quantization aware training for large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 467–484, 2024b. doi: 10.18653/v1/2024.findings-acl.26. URL https://aclanthology.org/2024.findings-acl.26/.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture models. In *International Conference on Learning Representations (ICLR)*, 2017. URL https://arxiv.org/abs/1609.07843.

Francesco Pio Monaco, Elia Cunegatti, Flavio Vella, and Giovanni Iacca. Frequency matters: Fast model-agnostic data curation for pruning and quantization. arXiv preprint arXiv:2603.16105 (to appear at EMNLP 2026), 2026. URL https://arxiv.org/abs/2603.16105.

Davide Paglieri, Saurabh Dash, Tim Rocktäschel, and Jack Parker-Holder. Outliers and calibration sets have diminishing effect on quantization of modern LLMs. arXiv preprint arXiv:2405.20835, 2024. URL https://arxiv.org/abs/2405.20835.

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 (ICLR)*, 2024. URL https://arxiv.org/abs/2308.13137.

Ibne Farabi Shihab, Sanjeda Akter, and Anuj Sharma. Coverage-based calibration for post-training quantization via weighted set cover over outlier channels. arXiv preprint arXiv:2604.24008, 2026. URL https://arxiv.org/abs/2604.24008.

Fei Wang, Chao Xue, Taoran Liu, Li Shen, Ye Liu, and ChangXing Ding. Beyond activation alignment: The alignment-diversity tradeoff in task-aware LLM quantization. arXiv preprint arXiv:2607.00908, 2026. URL https://arxiv.org/abs/2607.00908.

Johannes Welbl, Nelson F. Liu, and Matt Gardner. Crowdsourcing multiple choice science questions. In *Proceedings of the 3rd Workshop on Noisy User-generated Text*, pp. 94–106, 2017.

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, 2024. doi: 10.18653/v1/2024.acl-long.544. URL https://aclanthology.org/2024.acl-long.544/.

Miles Williams, George Chrysostomou, and Nikolaos Aletras. Self-calibration for language model quantization and pruning. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 10149–10167, 2025. doi: 10.18653/v1/2025.naacl-long.509. URL https://aclanthology.org/2025.naacl-long.509/.

Guangxuan Xiao, Ji Lin, Mickael Seznec, Hao Wu, Julien Demouth, and Song Han. SmoothQuant: Accurate and efficient post-training quantization for large language models. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 38087–38099. PMLR, 2023. URL https://proceedings.mlr.press/v202/xiao23c.html.

Haiyang Xiao, Weiqing Li, Jinyue Guo, Guochao Jiang, Guohua Liu, and Yuewei Zhang. FAQ: Mitigating quantization error via regenerating calibration data with family-aware quantization. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 21441–21460, San Diego, California, United States, July 2026. Association for Computational Linguistics. doi: 10.18653/v1/2026.findings-acl.1079. URL https://aclanthology.org/2026.findings-acl.1079/.

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 (COLM)*, 2025. URL https://arxiv.org/abs/2504.07389.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. arXiv preprint arXiv:2505.09388, 2025. URL https://arxiv.org/abs/2505.09388.

Junyi Ye, Mengjia Yu, Debapriya Hazra, and Guiling Wang. Damage predicts recovery: When calibration data matters in compressing financial LLMs. arXiv preprint arXiv:2609.26241, 2026. URL https://arxiv.org/abs/2609.26241.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. Personalizing dialogue agents: I have a dog, do you have pets too? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2204–2213, 2018.

Xiang Zhang, Junbo Zhao, and Yann LeCun. Character-level convolutional networks for text classification. In *Advances in Neural Information Processing Systems*, volume 28, 2015.

A Complete Results of the Main Study

All 90 GPTQ evaluations are summarised in Table 6, which gives means and sample SDs over the five calibration seeds for every model, precision and source, and in Table 5, which gives the paired contrasts involving Diverse. Every MMLU accuracy equals the saved number of correct answers divided by 14,042, and every perplexity equals the exponential of the saved mean token loss. The registered confirmatory analyses are listed in full in Appendix G.

The MMLU prompts longer than 512 tokens come from five subjects with long passages: professional law, the three high-school history subjects and college medicine. For Qwen3-8B the seed-level inference on the original records is robust to them: over every allocation of their answers compatible with the audit counts, the intact-question paired-t intervals stay within $[+1.76, +11.39]$ points for the W3 gain and $[+0.37, +11.83]$ for the interaction. These are sensitivity bounds on the reported fitted models; the replacement runs of Appendix B re-score these prompts untruncated.

Table 7: MMLU prompt-length audit and intact-question bound of the W3 Diverse-minus-WikiText contrast. a: prompts longer than 512 tokens; all of them lose the "Answer:" cue, and most also lose the "D." choice label. "Intact": lower end of the mean contrast in points on the $N - a$ questions that fit the window; "Seeds": seeds whose intact-question lower end is positive.

Model	a	No cue	No "D."	Reported	Intact	Seeds
Qwen3-8B	89 (0.63%)	89	68	$+6.45$	$+5.85$	5/5
Llama-3.1-8B	76 (0.54%)	76	60	$+3.84$	$+3.32$	4/5
Mistral-7B-v0.3	143 (1.02%)	143	116	$+1.99$	$+0.98$	4/5

B Replacement Runs with Untruncated Scoring

Protocol. The replacement runs were registered before any of their cells ran, with fixed cells, endpoints and a stop rule: no cell is rerun after its result is seen, except when an execution does not complete, which is reported. The registered analysis, with per-endpoint seed sets and completeness gates, was fixed before any block had completed and computes everything below. The runs use the original builders, quantizer settings and evaluation code and the same model releases as the original records. Relative to the original evaluation, the scoring code computes the four answer-letter logits and re-scores every prompt longer than 512 tokens without truncation, giving the "corrected" instrument; the "intact" instrument scores only prompts that fit the window. The released per-question records retain the predicted letters and the untruncated predictions for long prompts, but not the four logits or complete MMLU input IDs. Captured Qwen3-8B calibration inputs are included; they are distinct from evaluation-input records. A registered block enters the analysis only when all of its cells are complete. The analysis uses the complete registered blocks: the W3 source block for Llama-3.1-8B and Mistral-7B-v0.3 with 2 models $\times$ 4 sources $\times$ 5 seeds $= 40$ cells, their W4 WikiText and Diverse block of 20 cells, the Qwen3-8B block with W3 and W4 $\times$ WikiText and

Diverse $\times$ 5 seeds $= 20$ cells, its W3 self-generated and external-mixture block of 10 cells and the Qwen3-8B BF16 reference, together with the complete budget cells of Llama-3.1-8B at 8×128 and 128×128 with WikiText and Diverse in seeds 42–44, listed in Table 12, and of Qwen3-8B at 128×128 with W3 and W4, WikiText and equal-share Diverse in seeds 42–44, listed in Table 15. Table 23 lists the remaining registered cells. In all 20 Qwen3-8B WikiText and Diverse cells of seeds 42–46 the inputs prepared by the quantizer were recorded and equal the tokenizer IDs, and the W3 and W4 cells of each source and seed received identical inputs; because the quantizer re-tokenizes the decoded calibration text, in 4 Diverse cells, seeds 42 and 44 at both precisions, one of the 32 prepared sequences differs from the builder registry in 2 tokens.

Sources. WikiText and Diverse use the original builders. *Self-generated*: the unquantized model repeatedly samples 160 new tokens from the beginning-of-sequence token without a prompt, at temperature 1.0 and top-p 1.0 with a per-generation seed derived from the calibration seed; the generations are concatenated into one stream, which the Diverse chunker cuts; this follows the idea of self-calibration (Williams et al., 2025) but not its temperature schedule. *External mixture*: documents from news in AG News (Zhang et al., 2015), science passages from the SciQ supports (Welbl et al., 2017), code from the CodeAlpaca-20k outputs (Chaudhary, 2023), which were themselves generated by a language model, grade-school mathematics from GSM8K (Cobbe et al., 2021) and dialogue from Persona-Chat (Zhang et al., 2018), 600 per domain, shuffled with a fixed seed and consumed in domain rotation; token shares are not equalized. For Llama-3.1-8B and Mistral-7B-v0.3, rebuilding each stream reproduces all ten saved mixture registries token for token, and the consumed tokens are 7.6–11.1% news, 14.0–24.6% science, 8.2–17.6% code, 22.8–33.4% mathematics and 25.3–38.7% dialogue across the ten cells. The registration named arXiv abstracts for the science share; the pool, built before any cell ran, uses SciQ support passages instead. Because of the code share, this mixture combines human-written and model-generated text; it serves as a multi-domain comparison rather than as a human-versus-generated contrast. All sources supply 32×128 tokens.

Regeneration of original cells. For every WikiText and Diverse cell of the three models, the rebuilt calibration registry is identical to the retained original registry of the same model, source and seed. Of the 60 cells that have an original counterpart, 53 reproduce the original correct-answer count and saved perplexity exactly. The other 7 are four W3 cells, Llama-3.1-8B Diverse seeds 43 and 44 and WikiText seed 46 and Mistral-7B-v0.3 Diverse seed 46, and three W4 cells, Llama-3.1-8B WikiText seeds 42 and 43 and Qwen3-8B Diverse seed 44; with identical calibration inputs their MMLU accuracy differs from the original by $+1.55$, -1.12, -1.55 and -0.04 points at W3 and by $+1.08$, -0.72 and -0.85 points at W4. For the remaining original combinations, the W3 and W4 inputs are paired through the shared seed and the unchanged builder. The replacement contrasts are computed on the replacement cells only; regenerated seeds reproduce the original calibration draws and are not additional independent draws.

Execution variation. A registered descriptive analysis uses these matched cells to estimate execution variation. For each of the 30 W3 and 30 W4 cells with an original counterpart and identical calibration inputs, let e be the replacement-minus-original MMLU accuracy on the recorded instrument. If both executions carry centred noise, independent between executions and of the calibration draw, with equal variance across cells, $\hat{\sigma}^2_{\text{exec}} = \frac{1}{2}\overline{e^2}$, which gives $\hat{\sigma}_{\text{exec}} = 0.32$ points at W3, with 26 of 30 cells identical, and 0.20 at W4, with 27 of 30 identical. The between-draw SD of the original WikiText and Diverse cells, pooled as the root mean square of the per-model, per-source SDs over five seeds for the same six model–source groups at each precision, is 2.18 points at W3 and 0.40 at W4. Under these conditions execution accounts for about 2% of the observed variance across seeds at W3 and 25% at W4; the estimate rests on one repeat per cell and is a conditional descriptive estimate. Among the later registered executions, the repetition of Table 18 executes each of eight Qwen3-8B configurations twice more on the identical captured calibration input; all eight reproduce the original correct-answer counts, recorded and untruncated, and perplexities exactly, and the captured inputs of all eight equal the originals, so for these eight configurations the execution SD is zero at both precisions.

Later registered blocks. A further registration, appended before any of its cells ran, added Qwen3-8B blocks with the same builders, quantizer settings and evaluation, together with an ARC readout: all 1,172 ARC-Challenge and 2,376 ARC-Easy test questions (Clark et al., 2018), zero-shot, formatted

Table 8: Calibration-count sweep with the RMS-scaling quantizer, a different quantizer from the main study, at W3 with $\alpha = 0.5$, group 128 and seed 42. Perplexity on the first 50 WikiText-2 test passages: BF16 is absolute, all other columns are increases over it; numbered columns count 128-token calibration sequences. BF16 and RTN come from the same sweep as each row.

Model	Source	BF16	RTN	1	4	16	32	64	128
Qwen3.5-2B	WikiText	24.39	159.19	42.14	33.78	34.09	34.13	33.99	31.40
Qwen3.5-4B	WikiText	20.44	60.80	25.28	22.33	23.22	21.81	20.93	20.84
Qwen3.5-9B	WikiText	15.35	25.28	9.25	9.19	8.76	8.62	8.10	8.01
Qwen3.5-9B	Random	15.35	25.28	12.59	12.26	11.80	11.55	11.07	11.24
Qwen3.5-9B	Diverse	15.34	25.28	10.89	9.86	9.10	9.17	8.35	8.77
Qwen3.5-27B	WikiText	12.62	4.65	4.23	3.59	3.62	3.32	3.18	2.86

as "Question: … A. … B. … Answer:" with the choices relabelled A, B, C, … in dataset order, the prediction being the largest next-token logit among the letters of the question's choices, prompts not truncated. The blocks are a second batch of fresh seeds 52–56 with WikiText and Diverse at W3 and W4, 20 cells, with a precision objective fixed before any of its results was seen: a 95% half-width of at most two points for the W3 contrast and the interaction over seeds 42–56; activation ordering on, with WikiText and Diverse in seeds 42–44, 12 cells; the equal-share Diverse builder at 32×128 and, with WikiText, at 8×512 in seeds 42–44, 18 cells; the composition control of Table 20, 10 cells; two further executions of the eight WikiText and Diverse configurations of seeds 47 and 48 with the calibration input taken from the original runs, so that every execution quantizes the identical input, 16 cells; and the coverage analysis of Table 19, which uses unquantized forward passes on the exact W3 quantizer inputs. One composition cell was executed a second time on the same calibration input, as the registered stop rule permits for an incomplete execution. All registered endpoints of these blocks are listed in Tables 16 to 20. A last registration, appended before any of its cells ran, added a third batch of fresh seeds 57–61 with WikiText and Diverse at W3 and W4, 20 cells, with the same builders, settings and readouts, a size fixed at five seeds and endpoints over this batch, over all fresh seeds 47–61 and over seeds 42–61; all 20 cells completed, the W3 and W4 cells of each source and seed quantized identical captured inputs, and Tables 21 and 22 list every endpoint.

Results. Table 11 lists every registered contrast that the complete blocks allow, on all three instruments. The registered paired item bootstrap, with 10,000 resamples of the 14,042 questions for each seed-paired pair of fitted models on the recorded instrument, measures how precisely one pair of fitted models is compared, separately from calibration-draw uncertainty: the half-width of its 95% interval is 0.81–1.06 points for every W3 pair of Llama-3.1-8B and Mistral-7B-v0.3 behind Table 11 and 0.92–0.98 points at W3 and 0.59–0.63 at W4 for the Qwen3-8B Diverse–WikiText pairs. On the replacement cells, Diverse recovers 29.8% of the 21.6-point loss of W3 WikiText calibration relative to BF16 for Qwen3-8B on the corrected instrument, the registered BF16 gap closure. Averaged over the three models, the W3 gain of the external mixture over WikiText is 69% of the gain of Diverse in the registered mixture-to-Diverse comparison on the corrected instrument; the direct Diverse-minus-self-generated contrasts are listed in Table 11.

C Earlier Exploratory Studies

Before the controlled main study, we explored a quantizer that scales channels by their activation root mean square, RMS, in the style of SmoothQuant (Xiao et al., 2023) as a calibration-light alternative. These explorations are single-seed and serve as context for the main study. The one exploratory sweep that bears on the calibration budget is Table 8: with the RMS-scaling quantizer, one seed and 50 test passages, most of the RTN-to-calibrated perplexity improvement across four Qwen3.5 sizes is reached with 4–16 sequences, and moving from 32 to 128 sequences changes the perplexity increase by at most 2.7 points. Its scope is perplexity under a different quantizer. On Qwen3-32B, the same quantizer with WikiText calibration reduced the perplexity increase over BF16 from 237.2 under RTN to 7.3, in a single seed.

D Implementation and Reproducibility Details

D.1 Calibration builders and quantizer configuration

WikiText. Because the tokenizers differ, a model consumes 20–31 passages of the shuffled list; following the actual cut positions of the stream, in every seed the shortest consumed list is a prefix of the other two, with 20–26 shared passages, and rebuilding the sets reproduces all 26/26 retained WikiText calibration files token for token. *Random.* The round-trip condition is a length condition, not ID equality. Re-running the sampler reproduces all 26/26 saved Random calibration files; its acceptance rate is 1.1–1.8% of candidates for Qwen, 1.4–2.7% for Llama and 9.3–11.2% for Mistral. *Quantizer.* Every GPTQ model of the main study and of the replacement runs uses bits $\in \{3, 4\}$, group size 128, symmetric quantization without zero point, activation ordering off and dampening equal to 5% of the mean Hessian diagonal; all linear layers of the transformer blocks are quantized, the embeddings and the output head stay in BF16, and the quantizer receives the decoded calibration texts, 32 at the main budget. The registered extensions change one setting each: activation ordering is enabled in its block of Table 16, and the budget cells of Tables 12, 15 and 16 use 8×128, 128×128 or 8×512 tokens. The configuration is stored in every run record, and the supplementary material documents the command that re-runs any single cell from the public model weights and data files.

D.2 Supplementary material and replay

The reported numbers are reproduced by recomputation from archived records, and each record type supports a specific part. The original main-study cells are archived as result records with correct-answer counts and mean token losses, from which their accuracies, perplexities and all tables built on them are recomputed. The replacement, budget and later-block cells additionally carry the predicted answer to every MMLU and ARC question, from which their counts are recomputed. The coverage table is recomputed from the archived per-set coverage statistics. The four answer-letter logits, the complete evaluation input IDs and the quantized weights are not released; re-running a cell from the public model weights with the included code is an independent reproduction beyond this recomputation. The order of registrations and results is reported from the authors' protocol records.

The supplementary material contains one machine-readable row per model with model, quantizer, precision, source, seed, perplexity and number of correct MMLU answers, and the per-run result records with model and quantizer configuration; besides the 90 GPTQ models analysed here it also holds 12 AWQ records from preliminary runs outside the scope of this paper. The three BF16 references of Table 3 are stored with their per-question predictions, which reproduce their MMLU counts. Their evaluation calls the original perplexity function and re-implements the original MMLU scoring, with the same prompt, 512-token right truncation and letter-logit readout, to also save per-question outputs; the Qwen3-8B reference, evaluated with the replacement runs of Appendix B, additionally records its inputs. They score the same perplexity targets as the quantized models, 268,891, 263,553 and 286,770, and every saved prompt length matches a re-tokenization. All runs, including the BF16 references, use the same public release of each model; the tokenizer files of Llama-3.1-8B and Mistral-7B-v0.3 used in the runs are included, and the Qwen3-8B tokenizer is fetched from its public repository. It further contains the calibration token registries where retained, 78 sets of 32×128 IDs as stored by the builders, a script that rebuilds the public MMLU and WikiText-2 files exactly as used and fetches the Qwen3-8B tokenizer, and the quantization, evaluation and analysis code. For every replacement and budget cell of Appendices B and G it contains the result record with recorded and corrected MMLU counts and perplexity, the predicted answer to every MMLU question, including the untruncated prediction of every long prompt, and the calibration registry; the Qwen3-8B cells also contain the inputs prepared by the quantizer. A registry is the builder's token list: the cut stream slice, the accepted random candidate, or the re-encoded IDs of an adaptive cut; the quantizer receives the decoded texts, and the builders require each text to re-encode to exactly 128 tokens, a length condition that the replay confirms for all 2,496 retained sequences. Decoding is deterministic, so identical registries give identical quantizer inputs.

The supplementary material separates two kinds of evidence. *Replayable from the supplementary material and the public files it fetches*, by an audit script that loads the original calibration builders unchanged: the MMLU prompt-length counts of Table 7; the Random sampler, which reproduces all 26/26 retained Random registries and their acceptance rates; the WikiText builder, which reproduces all 26/26 retained WikiText registries, 22 GPTQ and four from the preliminary runs, and, following

the actual cut positions, the consumed-passage counts and shared prefixes of Appendix D.1; and the prompt coverage of every retained Diverse set. An analysis script then recomputes from the run records and these counts every number of Tables 2 to 7 and 9, the statistics quoted in the text and the sensitivity envelope of Appendix A, and compares all 137 with the values reported in the paper; four further scripts recompute from the replacement and budget records all 513 values of Appendix B, Tables 11, 12, 14 and 15 and the replacement statements of Sections 4.1, 4.4 and 4.5, a fifth recomputes Table 10 from the original records, a sixth Table 13 from the replacement and budget records, and two more the later Qwen3-8B blocks of Tables 16 to 22; all compare with the paper in the same way. *Saved records*: the quantized-model evaluations themselves, the BF16 evaluations, and the Diverse text, whose sampling is not replayed. For the 21 source–model–seed combinations whose W3 and W4 registries are both retained, the two are identical token for token. All inference in the main text is over calibration seeds, the relevant replication unit.

D.3 Prompts consumed by the Diverse source

The 32 prompts and their 160-token completions form a stream of more than 5,000 tokens, of which the quantizer consumes the first 4,096. Matching the first 24 normalized characters of each prompt in the decoded calibration IDs of the 26 Diverse cells where they are available, we find 25–28 of the 32 prompts, median 26: 25 for Qwen3-8B, 27–28 for Llama-3.1-8B and 25–26 for Mistral-7B-v0.3; prompts 29, 30, 31, 32, on the TCP handshake, photosynthesis, the capital of Japan and quantum computing, lie beyond the budget. Diverse is thus effectively a 25–28-prompt source, and fixed-budget comparisons of generated sources should report which prompts are consumed.

D.4 Diverse calibration prompts

The 32 prompts in stream order; the symbol ↩ marks a line break inside a prompt. Prompts 29–32 never reach the 32×128 calibration budget, as Appendix D.3 shows.

1. `The history of mathematics began`
2. `In quantum physics, the wave function`
3. `def fibonacci(n):↩if n <= 1:`
4. `SELECT * FROM users WHERE`
5. `Once upon a time in a small village`
6. `The stock market crashed because`
7. `import torch↩class Model(nn.Module):`
8. `Dear Sir, I am writing to`
9. `The recipe for chocolate cake requires`
10. `In 1969, humans first walked on`
11. `The chemical formula for water is`
12. `According to the theory of relativity`
13. `function quickSort(arr) {`
14. `The population of China is approximately`
15. `In machine learning, gradient descent`
16. `The mitochondria is the powerhouse`
17. `#!/bin/bash↩for i in $(seq 1 100)`
18. `The French Revolution started in`
19. `To solve this differential equation`
20. `In psychology, cognitive behavioral`
21. `The DNS protocol operates on port`
22. `La temperatura del sol es`
23. `The Fourier transform of a signal`
24. `In economics, supply and demand`
25. `The human genome contains approximately`
26. `To train a neural network`
27. `The Great Wall of China was built`
28. `In organic chemistry, benzene has`
29. `The TCP three-way handshake involves`
30. `Photosynthesis converts sunlight into`
31. `The capital of Japan is Tokyo`
32. `Quantum computers use quantum bits`

E Attributing a Calibration Effect

Two tempting explanations of a calibration gain assign it to the wrong cause. Both apply to any calibration study and both can be checked with simple algebra; the GPTQ source comparisons of the main text are measured directly.

Activation-aware scaling is also a data-free weight transform. A common way to show that calibration data matter is to compare activation-aware scaling with plain round-to-nearest, RTN. With per-channel activation RMS a_j and weight range $w_j = \max_i |W_{ij}|$, the scales, the scaled quantizer and RTN are

$$\begin{aligned} s_j &= a_j^{\alpha} / w_j^{1-\alpha}, \\ S &= \mathrm{diag}(s_j), \\ \text{scaled:} \quad & Q(WS)S^{-1}, \\ \text{RTN:} \quad & Q(W). \end{aligned}$$

At $\alpha = 0$ the scaled quantizer does not reduce to RTN: it becomes $Q(W\mathrm{diag}(1/w))\mathrm{diag}(w)$, a range normalisation that uses no calibration data at all. For the one-row, one-group W3 example $W = (1, 1/5)$, RTN gives $(1, 1/3)$ with squared error $4/225$, while the $\alpha = 0$ transform reproduces W exactly, as Appendix E.1 shows. A scaled-versus-RTN improvement therefore mixes a data-free transform with the information in the activations. Comparing with the $\alpha = 0$ endpoint removes the RTN confound but not all of it, because changing α also changes the weight exponent: with constant activations $a_j \equiv 1$ the scale is $s_j = w_j^{\alpha-1}$, which still varies with α. The contribution of the activation statistics is isolated by holding α fixed and replacing a_j with a constant or a permutation of itself. The source comparisons of the main text avoid this confound entirely because all sources share one quantizer and one configuration.

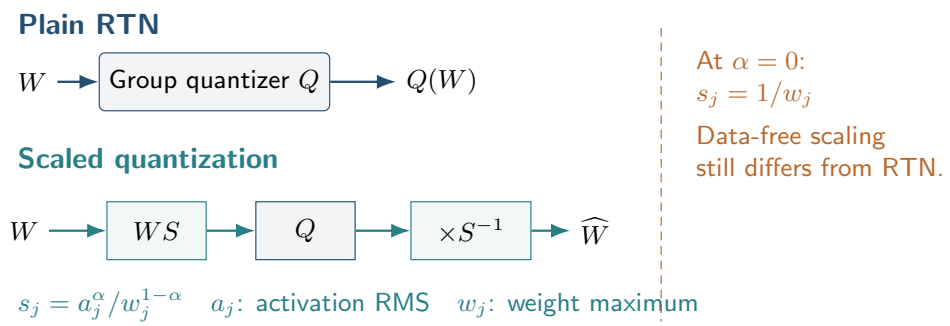

Figure 2: A scaled-versus-RTN comparison changes two things at once: a data-free weight normalisation and the activation information. At $\alpha = 0$ the scaled quantizer is already a data-free transform that differs from RTN; isolating the activation statistics requires a fixed α with constant or permuted activation statistics.

The answer-position gradient is a partial derivative. A second route attributes accuracy changes to how quantization moves the answer margin

$$m(W) = z_y(W) - \log \sum_{a \neq y} \exp z_a(W).$$

For a linear module ℓ shared across positions t, with input $x_{\ell t}$ and output adjoint $g_{\ell t} = \partial m / \partial o_{\ell t}$, the first-order change under the quantization step $\Delta W_\ell = Q(W_\ell) - W_\ell$ is

$$Dm(W)[\Delta W] = \sum_{\ell} \sum_{t} g_{\ell t}^{\mathsf{T}} \Delta W_\ell \, x_{\ell t}. \tag{1}$$

The sum runs over every position, because earlier positions reach the answer through attention; keeping only the answer position, $J_{\mathrm{ans}} \Delta W$, omits the term $J_{\mathrm{other}} \Delta W$. If the margin is twice

differentiable along the segment $W + t\Delta W$ with Hessian norm at most H there, Taylor's theorem gives

$$\left|m(W + \Delta W) - m(W) - J_{\text{ans}}\Delta W\right| \le \underbrace{\|J_{\text{other}}\|_2\|\Delta W\|_2}_{\text{omitted positions}} + \underbrace{\tfrac{1}{2}H\|\Delta W\|_2^2}_{\text{curvature}},$$

which separates the two errors of an answer-position proxy: omitted positions and finite-step curvature. Small values of both terms suffice for a small error bound; this upper bound does not make either term a necessary condition. Symmetric finite differences on the same inputs can check the approximation. The margin sign is also not the correctness event: logits $(1, 0.9, 0.8, 0.7)$ answer correctly yet give $m = -0.902$. Task-gradient saliency methods such as Task-Circuit Quantization (Xiao et al., 2025) use this first-order term; equation 1 is its full shared-weight form.

What estimation theory predicts. For fixed tail parameters, minimum second moment and failure probability, $O(\epsilon^{-2} \log d_{\text{in}})$ independent observations suffice to estimate all per-channel activation second moments to relative accuracy ϵ, by Proposition 1. This is an independent-observation bound, whereas GPTQ uses the full input covariance, whose informative directions depend on what the calibration text contains. The bound concerns the number of observations; the ranking of calibration sources is an empirical question, addressed by the experiments of the main text.

E.1 Worked example of data-free scaling

Take one weight row $W = (1, 1/5)$, one quantization group and a symmetric three-bit quantizer with largest positive level 3 and step $\max_j |W_j|/3$. *Round-to-nearest.* The step is $1/3$, so

$$Q(W) = (1,\ 1/3),$$
$$\|Q(W) - W\|_2^2 = (1/3 - 1/5)^2 = 4/225.$$

Scaling at $\alpha = 0$. The scaling rule gives $s_j = 1/w_j$, i.e. $S = \text{diag}(1, 5)$, and

$$\begin{aligned} WS &= (1,\ 1) && \text{represented exactly,} \\ Q(WS)S^{-1} &= (1,\ 1)\,\text{diag}(1, 1/5) \\ &= (1,\ 1/5) = W. \end{aligned}$$

The error falls to zero without any calibration activation. The example shows that a scaled-versus-RTN gain can arise from range normalisation alone. A channel with $w_j = 0$ has zero weights, so setting $s_j = 1$ there leaves $Q(WS)S^{-1}$ unchanged.

F Second-Moment Estimation Bound

This appendix states how many calibration observations suffice to estimate all per-channel activation second moments, the statistics used by activation-aware scaling. It concerns the number of observations, and the empirical analyses of the main text are self-contained.

A random variable Y with mean zero is sub-exponential with parameters (ν, b) if

$$\mathbb{E}e^{\theta Y} \le e^{\theta^2\nu^2/2} \qquad \text{for all } |\theta| < 1/b.$$

Proposition 1 (Simultaneous second-moment estimation)**.** *Let $\mathbf{X}_1, \ldots, \mathbf{X}_T \in \mathbb{R}^{d_{\text{in}}}$ be independent, with a common per-channel second moment $\mu_j = \mathbb{E}X_{t,j}^2$ for all t, $\mu_{\min} = \min_j \mu_j > 0$, and $X_{t,j}^2 - \mu_j$ sub-exponential with parameters (ν, b) for all t, j. Let $\hat{\mu}_j = T^{-1}\sum_t X_{t,j}^2$. There is a universal constant C such that for $\epsilon, \delta \in (0, 1)$,*

$$T \ge C \ln\Big(\frac{2d_{\text{in}}}{\delta}\Big) \max\Big\{\frac{\nu^2}{\epsilon^2\mu_{\min}^2}, \frac{b}{\epsilon\mu_{\min}}\Big\}$$

implies

$$\Pr\big[\exists j : |\hat{\mu}_j - \mu_j| > \epsilon\mu_j\big] \le \delta,$$

and on the complementary event $|\sqrt{\hat{\mu}_j}/\sqrt{\mu_j} - 1| \le \epsilon$ for all j.

Proof. Step 1, one channel. Bernstein's inequality for averages of independent sub-exponential variables gives

$$\Pr[|\hat{\mu}_j - \mu_j| > u] \le 2\exp\bigl(-cT\min\{u^2/\nu^2, u/b\}\bigr).$$

Step 2, all channels. Take $u = \epsilon\mu_j \ge \epsilon\mu_{\min}$, apply a union bound over the d_{in} channels and solve for T with $C = 1/c$.

Step 3, square roots. On the good event $\hat{\mu}_j/\mu_j \in [1-\epsilon, 1+\epsilon]$, and $|\sqrt{1\pm\epsilon}-1| \le \epsilon$ for $\epsilon \in (0,1)$. □

The channel dimension enters only logarithmically for fixed tail parameters and $\mu_{\min}$. Tokens within a sequence are dependent. The proposition does not justify replacing their count by an effective independent sample size; such a reduction requires additional dependence assumptions and a separate concentration result. GPTQ's Hessian and AWQ's scale search depend on the calibration data beyond diagonal second moments, so the proposition covers only methods that use these moments alone.

G Registered Confirmatory Analyses

This section lists every registered confirmatory analysis in full: the analyses of the original records in Tables 9 and 10; all registered endpoints that the complete blocks of the replacement runs of Appendix B allow in Table 11; the registered budget cells in Table 12; the secondary sliding-window perplexity in Table 13; the dev/test source selection on the per-question outputs of the replacement runs in Table 14; the fresh-draw and larger-budget cells of Qwen3-8B in Table 15; the later Qwen3-8B blocks in Tables 16 to 20; the third fresh batch in Tables 21 and 22; and the remaining registered cells in Table 23.

Table 9: Registered confirmatory analyses, registration E1, analyses 1–2: Diverse-minus-WikiText MMLU in points over five seeds. p: two-sided paired-t with df $= 4$; Holm: adjusted over the three models. †Registered exact sign-flip test over 2^{15} assignments. It treats the 15 model–seed pairs as independent, which the seed shared by the three models does not satisfy; it is not used for inference, and cross-model statements use the seed-block interval of Section 4.2 instead.

Endpoint	Model	Mean	95% CI	p	Holm
W3 contrast	Qwen3-8B	$+6.45$	$[+2.56, +10.34]$	0.010	0.030
W3 contrast	Llama-3.1-8B	$+3.84$	$[+0.08, +7.60]$	0.047	0.047
W3 contrast	Mistral-7B-v0.3	$+1.99$	$[+0.46, +3.53]$	0.023	0.046
Interaction	Qwen3-8B	$+5.90$	$[+1.99, +9.81]$	0.014	0.041
Interaction	Llama-3.1-8B	$+3.55$	$[-0.12, +7.22]$	0.055	0.110
Interaction	Mistral-7B-v0.3	$+1.19$	$[-0.75, +3.13]$	0.164	0.164
Interaction	pooled, 15 pairs†	$+3.55$	–	3.7×10^{-4}	–

Table 10: Registered analysis E1.3, WikiText-minus-Random part; the Diverse-minus-Random part is in Table 5. Paired GPTQ contrasts of the original records: recorded MMLU in points and WikiText-2 perplexity, negative when WikiText is better; mean and two-sided 95% paired-t interval over seeds 42–46 with df $= 4$, and seeds with a positive MMLU difference.

Contrast	Qwen3-8B	Llama-3.1-8B	Mistral-7B-v0.3
W3			
WikiText − Random	$+3.31$ $[-0.37, +7.00]$ 5/5	$+0.63$ $[-4.42, +5.69]$ 3/5	$+0.33$ $[-1.42, +2.09]$ 3/5
ΔPPL	-5.06 $[-6.04, -4.08]$	-4.03 $[-4.42, -3.64]$	-0.73 $[-0.90, -0.57]$
W4			
WikiText − Random	$+0.77$ $[+0.25, +1.29]$ 5/5	$+0.19$ $[-0.83, +1.20]$ 2/5	-0.32 $[-0.61, -0.03]$ 0/5
ΔPPL	-1.24 $[-1.86, -0.63]$	-0.33 $[-0.44, -0.22]$	-0.11 $[-0.14, -0.07]$

Table 11: Registered endpoints of the complete blocks of the replacement runs, GPTQ, 32×128 tokens, seeds 42–46 paired. MMLU contrasts in points, mean and two-sided 95% paired-t interval over seeds. Recorded: the original readout, with prompts right-truncated at 512 tokens; corrected: every longer prompt re-scored untruncated; intact: questions whose prompt fits 512 tokens. Replacement runs are reported separately from the original records and are not pooled with them.

Model	Contrast	n	Recorded	Corrected	Intact
Qwen3-8B	W3 Diverse−WikiText	5	$+6.45$ [$+2.56$, $+10.34$]	$+6.45$ [$+2.52$, $+10.37$]	$+6.49$ [$+2.56$, $+10.42$]
Qwen3-8B	W4 Diverse−WikiText	5	$+0.38$ [-0.05, $+0.82$]	$+0.39$ [-0.05, $+0.83$]	$+0.40$ [-0.03, $+0.82$]
Qwen3-8B	Interaction W3−W4	5	$+6.07$ [$+2.50$, $+9.64$]	$+6.06$ [$+2.46$, $+9.66$]	$+6.09$ [$+2.47$, $+9.72$]
Qwen3-8B	W3 Self-gen.−WikiText	5	$+2.33$ [-1.02, $+5.68$]	$+2.31$ [-1.05, $+5.68$]	$+2.33$ [-1.01, $+5.67$]
Qwen3-8B	W3 Ext. mixture−WikiText	5	$+2.95$ [$+0.46$, $+5.43$]	$+2.97$ [$+0.45$, $+5.49$]	$+2.97$ [$+0.47$, $+5.47$]
Qwen3-8B	W3 Diverse−Self-gen.	5	$+4.12$ [-0.47, $+8.71$]	$+4.13$ [-0.51, $+8.78$]	$+4.16$ [-0.47, $+8.79$]
Llama-3.1-8B	W3 Diverse−WikiText	5	$+4.24$ [$+1.15$, $+7.32$]	$+4.27$ [$+1.16$, $+7.38$]	$+4.27$ [$+1.19$, $+7.34$]
Llama-3.1-8B	W4 Diverse−WikiText	5	$+0.22$ [-0.81, $+1.25$]	$+0.22$ [-0.80, $+1.24$]	$+0.22$ [-0.82, $+1.25$]
Llama-3.1-8B	Interaction W3−W4	5	$+4.02$ [$+0.96$, $+7.08$]	$+4.05$ [$+0.96$, $+7.15$]	$+4.05$ [$+1.00$, $+7.10$]
Llama-3.1-8B	W3 Self-gen.−WikiText	5	$+7.28$ [$+3.11$, $+11.45$]	$+7.31$ [$+3.10$, $+11.52$]	$+7.31$ [$+3.12$, $+11.50$]
Llama-3.1-8B	W3 Ext. mixture−WikiText	5	$+4.68$ [$+0.45$, $+8.92$]	$+4.72$ [$+0.43$, $+9.00$]	$+4.72$ [$+0.43$, $+9.00$]
Llama-3.1-8B	W3 Diverse−Self-gen.	5	-3.04 [-6.83, $+0.74$]	-3.04 [-6.87, $+0.79$]	-3.04 [-6.84, $+0.76$]
Mistral-7B-v0.3	W3 Diverse−WikiText	5	$+1.99$ [$+0.45$, $+3.52$]	$+2.03$ [$+0.44$, $+3.62$]	$+2.01$ [$+0.44$, $+3.59$]
Mistral-7B-v0.3	W4 Diverse−WikiText	5	$+0.80$ [$+0.24$, $+1.36$]	$+0.84$ [$+0.30$, $+1.38$]	$+0.83$ [$+0.27$, $+1.40$]
Mistral-7B-v0.3	Interaction W3−W4	5	$+1.18$ [-0.75, $+3.12$]	$+1.19$ [-0.79, $+3.18$]	$+1.18$ [-0.80, $+3.16$]
Mistral-7B-v0.3	W3 Self-gen.−WikiText	5	$+1.19$ [-0.31, $+2.69$]	$+1.21$ [-0.36, $+2.78$]	$+1.20$ [-0.32, $+2.71$]
Mistral-7B-v0.3	W3 Ext. mixture−WikiText	5	$+1.15$ [-2.13, $+4.42$]	$+1.15$ [-2.04, $+4.35$]	$+1.16$ [-2.10, $+4.43$]
Mistral-7B-v0.3	W3 Diverse−Self-gen.	5	$+0.79$ [-1.40, $+2.99$]	$+0.82$ [-1.42, $+3.07$]	$+0.82$ [-1.41, $+3.04$]

Table 12: Registered calibration-budget cells, endpoint B1, Llama-3.1-8B, GPTQ W3, Diverse−WikiText, paired over seeds. MMLU in points and WikiText-2 perplexity difference, positive when Diverse is worse; mean and two-sided 95% paired-t interval with $\mathrm{df} = n - 1$. The 32×128 rows are the replacement cells of Table 11, the first restricted to seeds 42–44. All cells use the replacement-run evaluation of Appendix B; recorded, corrected and intact as in Table 11.

Budget	Seeds	n	Recorded	Corrected	Intact	ΔPPL
8×128	42–44	3	-1.45 [-6.06, $+3.16$]	-1.48 [-6.16, $+3.20$]	-1.46 [-6.14, $+3.22$]	$+1.73$ [$+1.37$, $+2.10$]
32×128	42–44	3	$+3.06$ [-0.10, $+6.22$]	$+3.09$ [-0.18, $+6.35$]	$+3.10$ [-0.09, $+6.28$]	$+2.69$ [$+1.70$, $+3.69$]
32×128	42–46	5	$+4.24$ [$+1.15$, $+7.32$]	$+4.27$ [$+1.16$, $+7.38$]	$+4.27$ [$+1.19$, $+7.34$]	$+2.71$ [$+2.35$, $+3.07$]
128×128	42–44	3	$+6.05$ [-4.13, $+16.23$]	$+6.12$ [-4.20, $+16.44$]	$+6.07$ [-4.14, $+16.28$]	$+2.43$ [$+2.11$, $+2.76$]
8×512	42–44	0	not run: the Diverse builder yields fewer than eight 512-token pieces			

Table 13: Registered secondary readout of the replacement and budget runs: conventional WikiText-2 sliding-window perplexity with context 2048 and stride 512, next to the primary passage-wise perplexity of Section 3. Each row: source minus WikiText, mean and two-sided 95% paired-t interval over seeds with $\mathrm{df} = n - 1$, and seeds in which the source has the higher perplexity.

				Passage-wise		Sliding window	
Block, seeds	Model	Bits	Source	ΔPPL	+	ΔPPL	+
Replacement, 42–46	Qwen3-8B	W3	Diverse	$+2.77$ [$+1.42$, $+4.11$]	5/5	$+1.39$ [$+1.11$, $+1.67$]	5/5
Replacement, 42–46	Qwen3-8B	W3	self-generated	$+4.46$ [$+2.83$, $+6.10$]	5/5	$+1.58$ [$+0.99$, $+2.16$]	5/5
Replacement, 42–46	Qwen3-8B	W3	external mixture	$+2.40$ [$+0.94$, $+3.87$]	5/5	$+0.83$ [$+0.56$, $+1.11$]	5/5
Replacement, 42–46	Qwen3-8B	W4	Diverse	$+0.77$ [$+0.37$, $+1.17$]	5/5	$+0.17$ [$+0.09$, $+0.25$]	5/5
Replacement, 42–46	Llama-3.1-8B	W3	Diverse	$+2.71$ [$+2.35$, $+3.07$]	5/5	$+1.91$ [$+1.52$, $+2.31$]	5/5
Replacement, 42–46	Llama-3.1-8B	W3	self-generated	$+2.93$ [$+2.27$, $+3.59$]	5/5	$+1.62$ [$+1.09$, $+2.15$]	5/5
Replacement, 42–46	Llama-3.1-8B	W3	external mixture	$+2.18$ [$+2.00$, $+2.36$]	5/5	$+1.49$ [$+1.04$, $+1.93$]	5/5
Replacement, 42–46	Llama-3.1-8B	W4	Diverse	$+0.25$ [$+0.20$, $+0.29$]	5/5	$+0.14$ [$+0.12$, $+0.15$]	5/5
Replacement, 42–46	Mistral-7B-v0.3	W3	Diverse	$+0.87$ [$+0.74$, $+0.99$]	5/5	$+0.05$ [-0.93, $+1.04$]	3/5
Replacement, 42–46	Mistral-7B-v0.3	W3	self-generated	$+0.83$ [$+0.74$, $+0.92$]	5/5	-0.02 [-0.80, $+0.76$]	3/5
Replacement, 42–46	Mistral-7B-v0.3	W3	external mixture	$+0.73$ [$+0.63$, $+0.84$]	5/5	-0.19 [-0.92, $+0.54$]	3/5
Replacement, 42–46	Mistral-7B-v0.3	W4	Diverse	$+0.10$ [$+0.07$, $+0.13$]	5/5	$+0.03$ [$+0.03$, $+0.04$]	5/5
Fresh, 47–51	Qwen3-8B	W3	Diverse	$+3.86$ [$+2.79$, $+4.93$]	5/5	$+1.62$ [$+1.30$, $+1.95$]	5/5
All, 42–51	Qwen3-8B	W3	Diverse	$+3.31$ [$+2.53$, $+4.09$]	10/10	$+1.51$ [$+1.32$, $+1.70$]	10/10
Fresh, 47–51	Qwen3-8B	W4	Diverse	$+0.53$ [$+0.09$, $+0.97$]	5/5	$+0.17$ [$+0.11$, $+0.24$]	5/5
All, 42–51	Qwen3-8B	W4	Diverse	$+0.65$ [$+0.40$, $+0.89$]	10/10	$+0.17$ [$+0.13$, $+0.21$]	10/10
8×128, 42–44	Llama-3.1-8B	W3	Diverse	$+1.73$ [$+1.37$, $+2.10$]	3/3	$+1.32$ [$+1.02$, $+1.62$]	3/3
128×128, 42–44	Llama-3.1-8B	W3	Diverse	$+2.43$ [$+2.11$, $+2.76$]	3/3	$+1.28$ [$+0.95$, $+1.60$]	3/3
128×128, 42–44	Qwen3-8B	W3	equal-share Diverse	$+2.98$ [$+1.36$, $+4.59$]	3/3	$+1.13$ [$+0.50$, $+1.76$]	3/3
128×128, 42–44	Qwen3-8B	W4	equal-share Diverse	$+1.05$ [-0.85, $+2.94$]	3/3	$+0.19$ [-0.09, $+0.47$]	3/3

Table 14: Registered dev/test source selection on the replacement runs with untruncated scoring. The 14,042 questions are split into a dev and a test half, within each subject as the primary split, labelled within, or by subject as the secondary split. Regret: test-half accuracy of the best available source for the seed minus that of the chosen source, in points, mean over five seeds. Rules: PPL, lowest WikiText-2 perplexity of the cell; Dev, same: highest dev-half accuracy of the same draw; Dev, other: highest dev-half accuracy averaged over the other four draws. Δ: the PPL regret minus the Dev, other regret, with a paired-t 95% interval over seeds, df $= 4$. Best: seeds in which PPL / Dev, same / Dev, other picks the test-best source. Picks: the source each rule picks in seeds 42–46. Sources: W WikiText, D Diverse, S self-generated, X external mixture.

				Regret, pp						
Model	Bits	Split	Sources	PPL	Dev, same	Dev, other	Δ	95% CI	Best	Picks
Qwen3-8B	W3	within	WD	6.46	0.00	0.00	$+6.46$	$[+2.96, +9.97]$	0/5/5	WWWWW/DDDDD/DDDDD
Qwen3-8B	W3	by subject	WD	5.42	0.00	0.00	$+5.42$	$[+1.21, +9.63]$	0/5/5	WWWWW/DDDDD/DDDDD
Llama-3.1-8B	W3	within	WDSX	7.14	0.17	0.17	$+6.97$	$[+2.56, +11.37]$	0/3/3	WWWWW/SSSSS/SSSSS
Llama-3.1-8B	W3	by subject	WDSX	6.53	0.00	0.11	$+6.43$	$[+2.67, +10.18]$	0/5/4	WWWWW/SSXSS/SSSSS
Mistral-7B-v0.3	W3	within	WDSX	2.82	0.05	2.41	$+0.41$	$[-2.37, +3.19]$	0/4/0	WWWWW/SXDXX/DDXDD
Mistral-7B-v0.3	W3	by subject	WDSX	2.76	0.00	2.37	$+0.39$	$[-1.70, +2.48]$	0/5/0	WWWWW/SXDXX/DDXDD
Qwen3-8B	W4	within	WD	0.49	0.05	0.05	$+0.44$	$[-0.17, +1.05]$	1/4/4	WWWWW/DDDDD/DDDDD
Qwen3-8B	W4	by subject	WD	0.42	0.04	0.04	$+0.38$	$[-0.16, +0.92]$	1/4/4	WWWWW/DDDDD/DDDDD
Llama-3.1-8B	W4	within	WD	0.54	0.00	0.63	-0.09	$[-0.25, +0.06]$	2/5/0	WWWWW/WDWDD/DWDWW
Llama-3.1-8B	W4	by subject	WD	0.46	0.00	0.29	$+0.17$	$[-0.58, +0.92]$	2/5/1	WWWWW/WDWDD/DWDDW
Mistral-7B-v0.3	W4	within	WD	0.98	0.02	0.02	$+0.96$	$[-0.11, +2.03]$	1/4/4	WWWWW/DDDDD/DDDDD
Mistral-7B-v0.3	W4	by subject	WD	0.80	0.10	0.10	$+0.70$	$[-0.28, +1.68]$	1/4/4	WWWWW/DDDDD/DDDDD

Table 15: Registered endpoints on fresh calibration draws and at a larger budget, Qwen3-8B, GPTQ, replacement-run evaluation. Fresh: seeds 47–51, never used by any builder before these runs; All: the replacement seeds 42–46 together with 47–51; 128×128: WikiText versus the equal-share Diverse builder, which gives each of the 32 prompts $1/32$ of the budget, seeds 42–44. MMLU contrasts, Diverse minus WikiText, in points at W3 and W4 and their W3$-$W4 interaction, mean and two-sided 95% paired-t interval over seeds with df $= n - 1$; instruments as in Table 11; $+$: seeds with a positive corrected difference. ΔPPL: WikiText-2 perplexity difference, positive when Diverse is worse. Reference: the original Diverse builder at 32×128 in the replacement cells of seeds 42–44, a descriptive comparison for the 128×128 rows, which use a different builder.

Contrast	Recorded	Corrected	Intact	$+$	ΔPPL
Fresh, 32×128*, seeds 47–51*					
W3	$+3.92$ $[+0.31, +7.53]$	$+3.89$ $[+0.25, +7.54]$	$+3.93$ $[+0.29, +7.57]$	4/5	$+3.86$ $[+2.79, +4.93]$
W4	$+0.74$ $[-0.13, +1.61]$	$+0.73$ $[-0.15, +1.60]$	$+0.76$ $[-0.12, +1.64]$	4/5	$+0.53$ $[+0.09, +0.97]$
W3$-$W4	$+3.18$ $[+0.11, +6.25]$	$+3.17$ $[+0.09, +6.25]$	$+3.17$ $[+0.09, +6.25]$	4/5	–
All, 32×128*, seeds 42–51*					
W3	$+5.18$ $[+2.93, +7.43]$	$+5.17$ $[+2.90, +7.44]$	$+5.21$ $[+2.94, +7.48]$	9/10	$+3.31$ $[+2.53, +4.09]$
W4	$+0.56$ $[+0.16, +0.96]$	$+0.56$ $[+0.16, +0.95]$	$+0.58$ $[+0.18, +0.98]$	8/10	$+0.65$ $[+0.40, +0.89]$
W3$-$W4	$+4.62$ $[+2.51, +6.73]$	$+4.61$ $[+2.49, +6.73]$	$+4.63$ $[+2.50, +6.77]$	9/10	–
128×128*, seeds 42–44*					
W3	$+1.43$ $[+0.22, +2.64]$	$+1.38$ $[+0.21, +2.56]$	$+1.39$ $[+0.23, +2.55]$	3/3	$+2.98$ $[+1.36, +4.59]$
W4	$+0.63$ $[-0.62, +1.87]$	$+0.63$ $[-0.54, +1.81]$	$+0.64$ $[-0.63, +1.91]$	3/3	$+1.05$ $[-0.85, +2.94]$
W3$-$W4	$+0.80$ $[-1.55, +3.16]$	$+0.75$ $[-1.48, +2.98]$	$+0.75$ $[-1.56, +3.07]$	2/3	–
Reference, 32×128*, seeds 42–44*					
W3	$+6.82$ $[-2.63, +16.28]$	$+6.82$ $[-2.74, +16.39]$	$+6.87$ $[-2.68, +16.43]$	3/3	$+2.13$ $[-0.13, +4.38]$
W4	$+0.42$ $[-0.80, +1.63]$	$+0.42$ $[-0.80, +1.65]$	$+0.43$ $[-0.77, +1.62]$	2/3	$+0.68$ $[-0.35, +1.71]$
W3$-$W4	$+6.41$ $[-1.97, +14.78]$	$+6.40$ $[-2.09, +14.89]$	$+6.44$ $[-2.07, +14.95]$	3/3	–

Table 16: Registered endpoints of the later Qwen3-8B blocks, GPTQ, replacement-run evaluation with the instruments of Table 11. Fresh: the second batch of fresh seeds 52–56, never used by any builder before these runs; All: the replacement seeds 42–46 with the first two fresh batches, df $= 14$; Activation ordering on: GPTQ with activation ordering enabled, next to the ordering-off replacement cells of the same seeds, registered as descriptive with $n = 3$; Equal-share Diverse: the equal-share builder of Table 15 against WikiText at the same budget and seeds, where 8×512 denotes eight target 512-token sequences. For Diverse, decoding and re-tokenizing the joined pieces yields 4,094, 4,096 and 4,095 actual tokens for seeds 42–44, respectively, with some boundary IDs changed in each seed; WikiText retains 4,096 tokens. This is a near-budget-matched longer-context condition, not a token-identical context-only intervention. MMLU contrasts, Diverse minus WikiText, in points at W3 and W4 and their W3−W4 interaction, mean and two-sided 95% paired-t interval over seeds with df $= n - 1$; +: seeds with a positive corrected difference; ΔPPL: WikiText-2 perplexity difference, positive when Diverse is worse.

Contrast	Recorded	Corrected	Intact	+	ΔPPL
Fresh, 32×128*, seeds 52–56*					
W3	+5.10 [+2.74, +7.46]	+5.13 [+2.75, +7.52]	+5.15 [+2.77, +7.52]	5/5	+4.09 [+2.54, +5.64]
W4	+0.78 [−0.09, +1.65]	+0.78 [−0.15, +1.72]	+0.78 [−0.12, +1.68]	5/5	+0.86 [+0.48, +1.23]
W3−W4	+4.32 [+2.67, +5.97]	+4.35 [+2.68, +6.02]	+4.36 [+2.70, +6.03]	5/5	–
All, 32×128*, seeds 42–56*					
W3	+5.16 [+3.65, +6.66]	+5.16 [+3.64, +6.68]	+5.19 [+3.67, +6.71]	14/15	+3.57 [+2.93, +4.21]
W4	+0.63 [+0.31, +0.96]	+0.63 [+0.30, +0.97]	+0.65 [+0.31, +0.98]	13/15	+0.72 [+0.53, +0.90]
W3−W4	+4.52 [+3.15, +5.89]	+4.53 [+3.15, +5.90]	+4.54 [+3.16, +5.93]	14/15	–
Activation ordering on, 32×128*, seeds 42–44*					
W3	+5.69 [−2.95, +14.32]	+5.73 [−3.04, +14.50]	+5.74 [−3.00, +14.48]	3/3	+4.21 [+1.81, +6.61]
W4	−0.29 [−1.09, +0.50]	−0.29 [−1.05, +0.46]	−0.28 [−1.07, +0.50]	0/3	+0.71 [−0.02, +1.45]
W3−W4	+5.98 [−2.43, +14.39]	+6.02 [−2.55, +14.60]	+6.02 [−2.50, +14.55]	3/3	–
Activation ordering off, replacement cells, seeds 42–44					
W3	+6.82 [−2.63, +16.28]	+6.82 [−2.74, +16.39]	+6.87 [−2.68, +16.43]	3/3	+2.13 [−0.13, +4.38]
W4	+0.42 [−0.80, +1.63]	+0.42 [−0.80, +1.65]	+0.43 [−0.77, +1.62]	2/3	+0.68 [−0.35, +1.71]
W3−W4	+6.41 [−1.97, +14.78]	+6.40 [−2.09, +14.89]	+6.44 [−2.07, +14.95]	3/3	–
Equal-share Diverse, 32×128*, seeds 42–44*					
W3	+6.14 [+0.56, +11.72]	+6.15 [+0.44, +11.87]	+6.19 [+0.46, +11.91]	3/3	+2.94 [−1.11, +6.98]
W4	+0.46 [−0.14, +1.06]	+0.42 [−0.20, +1.05]	+0.44 [−0.17, +1.06]	3/3	+0.68 [+0.06, +1.31]
W3−W4	+5.68 [+0.70, +10.66]	+5.73 [+0.64, +10.82]	+5.74 [+0.63, +10.85]	3/3	–
Equal-share Diverse, 8×512*, seeds 42–44*					
W3	+4.77 [+2.41, +7.14]	+4.79 [+2.43, +7.15]	+4.82 [+2.47, +7.17]	3/3	+4.55 [+3.14, +5.95]
W4	+0.85 [−0.61, +2.32]	+0.81 [−0.66, +2.28]	+0.83 [−0.66, +2.33]	3/3	+0.72 [+0.05, +1.39]
W3−W4	+3.92 [+2.80, +5.04]	+3.98 [+2.89, +5.07]	+3.98 [+2.92, +5.05]	3/3	–

Table 17: Registered ARC readout, endpoints A1 and A2, of the second fresh batch, Qwen3-8B, GPTQ, seeds 52–56: Diverse-minus-WikiText accuracy in points on the 1,172 ARC-Challenge and 2,376 ARC-Easy questions, mean and two-sided 95% paired-t interval with df $= 4$, and seeds with a positive difference; last four rows: mean accuracies in percent of the four configurations, with the corrected MMLU accuracy of the same fitted models.

Readout	Contrast	Mean [95% CI]	+
ARC-Challenge	W3	+8.45 [+2.22, +14.67]	5/5
ARC-Challenge	W4	+0.68 [−0.51, +1.88]	4/5
ARC-Challenge	W3−W4	+7.76 [+2.02, +13.51]	5/5
ARC-Easy	W3	+6.06 [−1.32, +13.44]	5/5
ARC-Easy	W4	+0.20 [−0.15, +0.55]	4/5
ARC-Easy	W3−W4	+5.86 [−1.50, +13.22]	5/5
Accuracy	Diverse W3	ARC-C 76.74, ARC-E 87.91	MMLU 56.59
Accuracy	Diverse W4	ARC-C 89.42, ARC-E 95.47	MMLU 69.66
Accuracy	WikiText W3	ARC-C 68.29, ARC-E 81.85	MMLU 51.46
Accuracy	WikiText W4	ARC-C 88.74, ARC-E 95.27	MMLU 68.87

Table 18: Registered fixed-input repetition, Qwen3-8B, GPTQ, 32×128: for each configuration, the corrected MMLU accuracy in percent and the WikiText-2 perplexity of the original execution and of two further executions on the identical captured calibration input, and whether the three executions agree exactly in recorded and untruncated correct-answer counts and in perplexity. Between-draw SD over seeds 42–51 for the same model and sources: 1.98 points at W3 and 0.36 at W4.

Source	Bits	Seed	MMLU, three executions	PPL, three executions	Identical
WikiText	W3	47	52.95, 52.95, 52.95	26.279, 26.279, 26.279	yes
WikiText	W3	48	48.76, 48.76, 48.76	25.314, 25.314, 25.314	yes
WikiText	W4	47	69.23, 69.23, 69.23	19.957, 19.957, 19.957	yes
WikiText	W4	48	68.54, 68.54, 68.54	19.373, 19.373, 19.373	yes
Diverse	W3	47	52.21, 52.21, 52.21	30.185, 30.185, 30.185	yes
Diverse	W3	48	55.55, 55.55, 55.55	30.096, 30.096, 30.096	yes
Diverse	W4	47	69.52, 69.52, 69.52	19.984, 19.984, 19.984	yes
Diverse	W4	48	69.88, 69.88, 69.88	20.153, 20.153, 20.153	yes

Table 19: Registered coverage analysis, Qwen3-8B, unquantized forward passes on the exact W3 quantizer inputs. For each of 24 probed linear-layer inputs, namely the attention, output-projection, MLP and down-projection inputs of layers 0, 7, 14, 21, 28 and 35, with calibration second moment S_c, second moment S_m over every fourth MMLU prompt and $\lambda = 0.05\,\overline{\operatorname{diag} S_c}$, the coverage is $\kappa = 1 - \lambda \operatorname{tr}((S_c + \lambda I)^{-1} S_m)/\operatorname{tr} S_m$; $\bar{\kappa}$ averages the 24 inputs. M1: per-seed Kendall τ_b between $\bar{\kappa}$ and W3 corrected MMLU accuracy over the four sources, averaged over seeds, with the exact one-sided permutation p-value; M2: paired Diverse-minus-WikiText $\bar{\kappa}$ difference with 95% paired-t interval, df $= 9$; M3: Pearson correlation of the per-seed $\bar{\kappa}$ and accuracy differences with Fisher-z 95% interval. Further rows: descriptive $\bar{\kappa}$ by calibration set, and by input type at 32×128, averaged over the seeds of each source and over the six layers: attention, output-projection, MLP and down-projection inputs.

Endpoint	Sets	Statistic	
M1	seeds 42–46, four sources, W3	mean Kendall τ_b 0.20	one-sided exact $p = 0.226$
M2	seeds 42–51, Diverse−WikiText	$\Delta\kappa = +0.0004$ [−0.0029, +0.0037]	6/10 positive
M3	seeds 42–51, $\Delta\kappa$ vs ΔMMLU	W3 $r = 0.57$ [−0.10, 0.88]	W4 $r = -0.15$ [−0.71, 0.53]
$\bar{\kappa}$	equal-share Diverse, 128×128, $n = 3$	mean 0.8819	range 0.8793–0.8866
$\bar{\kappa}$	Diverse, 32×128, $n = 10$	mean 0.7597	range 0.7552–0.7651
$\bar{\kappa}$	external mixture, 32×128, $n = 5$	mean 0.7598	range 0.7589–0.7609
$\bar{\kappa}$	Random, 32×128, $n = 5$	mean 0.6599	range 0.6578–0.6616
$\bar{\kappa}$	self-generated, 32×128, $n = 5$	mean 0.7378	range 0.7253–0.7584
$\bar{\kappa}$	WikiText, 128×128, $n = 3$	mean 0.8906	range 0.8885–0.8919
$\bar{\kappa}$	WikiText, 32×128, $n = 10$	mean 0.7593	range 0.7555–0.7628
Input type	WikiText, $n = 10$	attention 0.8480, output 0.8237	MLP 0.8122, down 0.5534
Input type	Diverse, $n = 10$	attention 0.8538, output 0.8226	MLP 0.8103, down 0.5522
Input type	self-generated, $n = 5$	attention 0.8303, output 0.8046	MLP 0.7845, down 0.5317
Input type	external mixture, $n = 5$	attention 0.8521, output 0.8273	MLP 0.8122, down 0.5477
Input type	Random, $n = 5$	attention 0.7851, output 0.6670	MLP 0.7111, down 0.4762

Table 20: Registered composition control, Qwen3-8B, GPTQ, 32×128, seeds 42–45, corrected MMLU instrument. Broad: the equal-share Diverse builder with all 32 prompts; narrow: the same builder, generator, sampling settings, per-slot token share and sequence construction, with the 32 prompt slots filled by the eight encyclopedic-prose prompts, four slots each, each slot with its own seed. The eight prompts still cover several topics: narrow refers to the prompt set and prose style, not a single knowledge domain. K1: broad-minus-narrow W3 contrast; K2: its W3−W4 interaction; also each composition minus the replacement WikiText cells of the same seeds, and the WikiText-2 perplexity differences. Mean and two-sided 95% paired-t interval with df $= 3$; +: seeds with a positive difference.

Contrast	Mean [95% CI]	+
K1: broad−narrow, W3	-0.99 [-2.93, $+0.95$]	0/4
broad−narrow, W4	-0.14 [-0.70, $+0.41$]	2/4
K2: broad−narrow, W3−W4	-0.84 [-2.42, $+0.73$]	1/4
broad−WikiText, W3	$+6.54$ [$+3.31$, $+9.78$]	4/4
broad−WikiText, W4	$+0.36$ [-0.02, $+0.75$]	4/4
broad−WikiText, W3−W4	$+6.18$ [$+3.15$, $+9.20$]	4/4
narrow−WikiText, W3	$+7.53$ [$+3.44$, $+11.62$]	4/4
narrow−WikiText, W4	$+0.51$ [-0.24, $+1.26$]	4/4
narrow−WikiText, W3−W4	$+7.02$ [$+3.60$, $+10.45$]	4/4
ΔPPL broad−narrow, W3	$+0.56$ [-0.81, $+1.93$]	3/4
ΔPPL broad−WikiText, W3	$+3.10$ [$+0.92$, $+5.28$]	4/4
ΔPPL broad−WikiText, W4	$+0.72$ [$+0.37$, $+1.06$]	4/4
ΔPPL narrow−WikiText, W3	$+2.54$ [$+1.20$, $+3.89$]	4/4
ΔPPL narrow−WikiText, W4	$+0.72$ [-0.31, $+1.75$]	4/4

Table 21: Registered endpoints of the third fresh batch, Qwen3-8B, GPTQ, 32×128, replacement-run evaluation with the instruments of Table 11. Third fresh batch: seeds 57–61, never used by any builder before these runs, batch size fixed at five; all fresh draws: seeds 47–61, df $= 14$; all: the replacement seeds 42–46 with the three fresh batches, df $= 19$. MMLU contrasts, Diverse minus WikiText, in points at W3 and W4 and their W3−W4 interaction, mean and two-sided 95% paired-t interval over seeds with df $= n - 1$; +: seeds with a positive corrected difference; ΔPPL: WikiText-2 perplexity difference, positive when Diverse is worse.

Contrast	Recorded	Corrected	Intact	+	ΔPPL
Third fresh batch, 32×128*, seeds 57–61*					
W3	$+5.77$ [$+0.17$, $+11.36$]	$+5.82$ [$+0.25$, $+11.39$]	$+5.83$ [$+0.23$, $+11.44$]	5/5	$+3.89$ [$+2.98$, $+4.80$]
W4	$+0.39$ [-0.24, $+1.02$]	$+0.36$ [-0.28, $+1.00$]	$+0.38$ [-0.27, $+1.02$]	4/5	$+0.61$ [-0.18, $+1.40$]
W3−W4	$+5.38$ [-0.41, $+11.17$]	$+5.46$ [-0.32, $+11.25$]	$+5.46$ [-0.37, $+11.29$]	5/5	–
All fresh draws, 32×128*, seeds 47–61*					
W3	$+4.93$ [$+3.19$, $+6.67$]	$+4.95$ [$+3.20$, $+6.70$]	$+4.97$ [$+3.22$, $+6.72$]	14/15	$+3.95$ [$+3.44$, $+4.45$]
W4	$+0.64$ [$+0.29$, $+0.98$]	$+0.62$ [$+0.27$, $+0.98$]	$+0.64$ [$+0.29$, $+0.99$]	13/15	$+0.67$ [$+0.42$, $+0.91$]
W3−W4	$+4.29$ [$+2.60$, $+5.98$]	$+4.33$ [$+2.63$, $+6.03$]	$+4.33$ [$+2.62$, $+6.04$]	14/15	–
All, 32×128*, seeds 42–61*					
W3	$+5.31$ [$+3.84$, $+6.77$]	$+5.32$ [$+3.85$, $+6.79$]	$+5.35$ [$+3.87$, $+6.82$]	19/20	$+3.65$ [$+3.15$, $+4.15$]
W4	$+0.57$ [$+0.31$, $+0.84$]	$+0.56$ [$+0.29$, $+0.84$]	$+0.58$ [$+0.31$, $+0.85$]	17/20	$+0.69$ [$+0.50$, $+0.88$]
W3−W4	$+4.74$ [$+3.31$, $+6.16$]	$+4.76$ [$+3.33$, $+6.19$]	$+4.77$ [$+3.33$, $+6.21$]	19/20	–

Table 22: Registered ARC readout of the third fresh batch, seeds 57–61, and of all ARC-scored fresh draws, seeds 52–61, Qwen3-8B, GPTQ: Diverse-minus-WikiText accuracy in points on the 1,172 ARC-Challenge and 2,376 ARC-Easy questions, mean and two-sided 95% paired-t interval with df $= n - 1$, and seeds with a positive difference.

Readout	Contrast	Mean [95% CI]	+
Third fresh batch, seeds 57–61			
ARC-Challenge	W3	$+8.38$ [$+0.32$, $+16.44$]	5/5
ARC-Challenge	W4	$+0.67$ [-0.06, $+1.39$]	4/5
ARC-Challenge	W3−W4	$+7.71$ [-0.31, $+15.73$]	5/5
ARC-Easy	W3	$+7.26$ [-2.34, $+16.85$]	5/5
ARC-Easy	W4	$+0.27$ [-0.29, $+0.83$]	3/5
ARC-Easy	W3−W4	$+6.99$ [-2.83, $+16.81$]	5/5
All ARC-scored fresh draws, seeds 52–61			
ARC-Challenge	W3	$+8.41$ [$+4.50$, $+12.32$]	10/10
ARC-Challenge	W4	$+0.67$ [$+0.14$, $+1.21$]	8/10
ARC-Challenge	W3−W4	$+7.74$ [$+3.95$, $+11.53$]	10/10
ARC-Easy	W3	$+6.66$ [$+1.99$, $+11.33$]	10/10
ARC-Easy	W4	$+0.24$ [-0.02, $+0.49$]	7/10
ARC-Easy	W3−W4	$+6.42$ [$+1.69$, $+11.15$]	10/10

Table 23: Remaining registered cells and analyses, which were not run. A block enters the analysis only when all of its cells are complete, so the endpoints listed here are outside the reported analyses.

Block	Models	Cells	n	Endpoints
Diverse 8×512 budget	Llama-3.1-8B	GPTQ W3, seeds 42–44	3	B1 at 8×512
AWQ sources	all three	AWQ W4	33	W1
Coverage of the composition sets	Qwen3-8B	W3 inputs of broad and narrow, seeds 42–45	8	K3