Calibrating Beyond English: Language Diversity for Better Quantized Multilingual LLMs

Everlyn Asiko Chimoto[1], **Mostafa Elhoushi**[2], **Bruce Bassett**[3,4,5]

[1]Lelapa AI, [2]Cerebras Systems, Inc , [3]University of the Witwatersrand,
[4]University of Cape Town, South Africa, [5]South African Astronomical Observatory
Correspondence: everlyn.asiko@lelapa.ai

Abstract

Quantization is an effective technique for reducing the storage footprint and computational costs of Large Language Models (LLMs), but it often results in performance degradation. Existing post-training quantization methods typically use small, English-only calibration sets; however, their impact on multilingual models remains underexplored. We systematically evaluate eight calibration settings (five single-language and three multilingual mixes) on two quantizers (GPTQ, AWQ) on data from 10 languages. Our findings reveal a consistent trend: non-English and multilingual calibration sets significantly improve perplexity compared to English-only baselines. Specifically, we observe notable average perplexity gains across both quantizers on Llama3.1 8B and Qwen2.5 7B, with multilingual mixes achieving the largest overall reductions of up to 3.52 points in perplexity. Furthermore, our analysis indicates that tailoring calibration sets to the evaluation language yields the largest improvements for individual languages, underscoring the importance of linguistic alignment. We also identify specific failure cases where certain language-quantizer combinations degrade performance, which we trace to differences in activation range distributions across languages. These results highlight that static one-size-fits-all calibration is suboptimal and that tailoring calibration data, both in language and diversity, plays a crucial role in robustly quantizing multilingual LLMs.

1 Introduction

Quantization – where the numerical precision of model parameters is reduced (e.g., from 32-bit floats to 8-bit or lower) – is a highly effective model compression technique and has become the de facto method for using Large Language Models (LLMs) on smaller, more accessible, and affordable infrastructure (Gray and Neuhoff, 1998; Gholami et al., 2021; Zhou et al., 2024). However, its effectiveness depends critically on the calibration set used (Hubara et al., 2021). Current practice overwhelmingly relies on English-only calibration sets such as C4 (Raffel et al., 2020) or Pile (Gao et al., 2020) (Frantar et al., 2023; Lin et al., 2024), even for multilingual models. This raises a key concern: does calibrating on English alone limit performance across other languages, and can we improve multilingual performance by broadening the choice of calibration set?

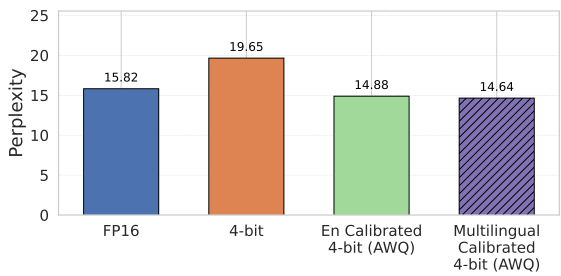


Figure 1: Average perplexity on 10 languages for Llama3.1 8B. Multilingual calibration achieves the lowest perplexity (14.64), illustrating that calibration language affects quantization quality.

This question matters most in settings where compute is scarce. In particular, techniques like quantization are often necessary for deploying models in low-resource language settings (Ahia et al., 2021). Sadly, prior work has shown that quantization disproportionately degrades performance in multilingual LLMs, with the most severe effects observed for low-resource, and non-Latin script languages – and as such, it provides limited utility for the long tail of low-resource languages (Marchisio et al., 2024).

Figure 1 provides a motivating example. It shows the perplexity score for Llama3.1 8B under 4-bit quantization with AWQ. While weight-only 4-bit quantization degrades performance relative to FP16, using a calibration set makes a clear differ-

Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics
Volume 1: Long Papers, pages 4822–4838
March 24-29, 2026

ence: English-only calibration reduces perplexity to 14.88, but a multilingual calibration set lowers it further to 14.64. The improvement is modest in absolute terms, yet consistent with a broader pattern across languages: calibration language and composition matter.

We hypothesize that non-English and mixed-language calibration sets can better preserve performance across diverse languages and might be an avenue for improvement for low-resourced languages. To test this, we compare English, non-English (French, Swahili, Chinese, isiXhosa), and multilingual calibration sets in post-training quantization. We evaluate these calibration sets on two similar sized LLMs (Llama3.1 8B, Qwen2.5 7B) using two Post-Training Quantization (PTQ) methods: (GPTQ, AWQ). Beyond performance, we probe the models' internal behavior (quantization error, weight updates, and activation distributions) when the calibration set language is varied. Our analysis also asks what constitutes an optimal calibration set in terms of token distribution and linguistic coverage.

We organize our investigation around three key arguments, each tied to a research question denoted as RQ:

- **Argument 1:** "One-Size-Fits-All" Calibration is Suboptimal

  RQ1: How does calibration-set language/composition affect quantization accuracy across languages?

  Result: We show that non-English and multilingual sets can outperform English-only sets (up to +3.52 ppl. on Llama-GPTQ).

- **Argument 2:** Calibration Sets Must Account for Rare and Extreme Data

  RQ2: Do outlier tokens or extreme activations in calibration data drive quantization error?

  Result: We demonstrate that including rare tokens and activation outliers reduces performance degradation.

- **Argument 3:** Calibration effectiveness is quantizer-dependent.

  RQ3: How do different calibration sets interact with GPTQ's Hessian-based updates versus AWQ's activation scaling?

  Result: GPTQ is more sensitive to calibration-language shifts, while AWQ is more robust due to its rescaling design.

2 Background: Calibration Sets for Quantization

Popular Post-Training Quantization (PTQ) techniques utilize calibration sets to determine how the weights are quantized. Examples include GPTQ, which changes weights in each channel and dynamically adjusts the quantized weights to compensate for the error, and AWQ which uses the calibration set to identify salient weights and scales them to keep their magnitude high. In this study, we focus our investigation on GPTQ (Frantar et al., 2023) and AWQ (Lin et al., 2024) for their prominence and representation in weight-only PTQ as well as widespread adoption in popular LLM libraries (e.g., GPTQModel[1], AutoGPTQ[2], AutoAWQ[3], LLM-AWQ[4]). We discuss them below:

GPTQ - General Post Training Quantization works in the following way: Given a weight matrix W and calibration data X, we obtain the quantized weight $\hat{W}$ by computing the inverse Hessian via Cholesky decomposition defined by:

$$H^{-1} = \left(2\,XX^T + \lambda I_d\right)^{-1},$$

where $\lambda > 0$ is a damping factor and I_d is the $d \times d$ identity matrix. We partition the columns of W into non-overlapping groups of size B. GPTQ then maps each row's real-valued weights into a small fixed set of integer levels defined by asymmetric min–max. This is referred to as the quantization grid. For each group, beginning at column index i, we quantize its columns $j = i, i+1, \ldots, i+B-1$ in sequence, by rounding each to the nearest number in the selected quantization grid, denoted by $Q[:,j]$. The rest of the weights are adjusted using error, E:

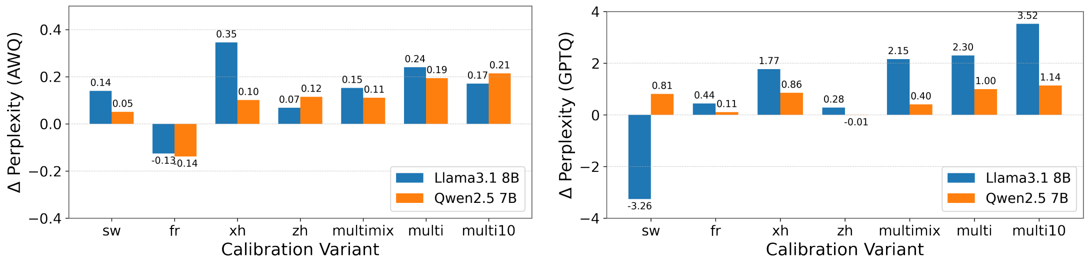
$$E[:, j-i] = \frac{W[:,j] - Q[:,j]}{H^{-1}[j,j]},$$

where i is the current column index, $H^{-1}[j,j]$ is the $[j,j]$ entry of H^{-1}. The updated weights are:

$$W[:, j : (i{+}B)] -= E[:, j{-}i] * H^{-1}[j, j : (i{+}B)].$$

This continues until all weights in each group have been updated. This procedure incrementally minimizes the quantization error by using the Hessian approximation to propagate residuals within each group.

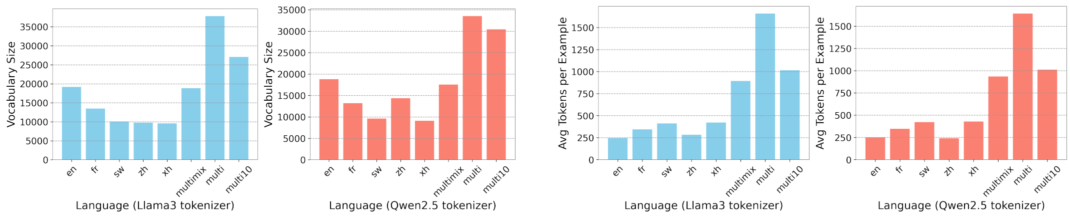
[1]https://github.com/ModelCloud/GPTQModel
[2]https://github.com/AutoGPTQ/AutoGPTQ
[3]https://github.com/casper-hansen/AutoAWQ
[4]https://github.com/mit-han-lab/llm-awq



(a) AWQ Delta

(b) GPTQ Delta

Figure 2: Δ Perplexity (higher = better) on the Wikipedia multilingual test set (English, French, Swahili, Xhosa, Chinese, Sesotho, Yoruba, Zulu, Hausa, Igbo), relative to an English-only calibration baseline. Calibration with non-English languages lead to better perplexity than English-only calibration. Multilingual variants(refer to Section 3.3 for details) provide the largest gains, showing that a linguistically diverse calibration set can outperform an English-only baseline.



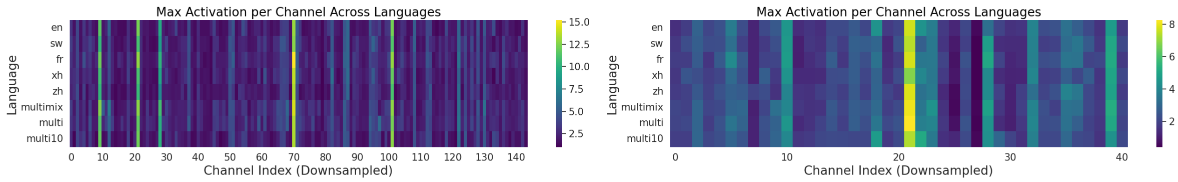
(a) Unique vocabulary size per calibration set

(b) Mean number of tokens per calibration example

Figure 3: Comparison of calibration-set statistics across different languages. (a) The total unique tokens in each calibration set; (b) The average token count per example. Multilingual sets exhibit both larger vocabularies and longer examples, indicating broader coverage of token contexts. This richer distribution correlates with the improved performance shown in Figure 2.

AWQ - Activation Aware Quantization works by identifying salient channels based on activation magnitudes. It runs a calibration set through the model to collect per-input-channel activation magnitudes. The channels with the largest average activations are deemed most salient. For each salient channel $j \in \mathcal{S}$, we scale its corresponding weights in $W[:, j]$ by a factor $\alpha_j > 1$ (e.g., proportional to $\max |W[:, j]|$), yielding a scaled weight $w'_{r,j}$. AWQ then applies standard grouped PTQ (e.g., 3- or 4-bit) on each row or group of size g. This two-step process, salient-channel scaling followed by grouped grid quantization, preserves the most influential weights while compressing the remainder.

Comparison Studies - While a growing body of work has examined quantization for LLMs, only a few have systematically investigated the role of calibration data in shaping post-training quantization outcomes. Williams and Aletras (2024) conducted a large-scale study of calibration data, showing that performance varies markedly across calibration sources and even across sets from the same source. Unlike their focus on English-only web data, we investigate multilingual calibration sets and their role across quantizers. Marchisio et al. (2024) analyzes multilingual models, finding that quantization disproportionately harms non-Latin script languages, that automatic metrics underestimate degradation, and that reasoning tasks degrade fastest. While they highlight disparities, we explicitly compare multilingual versus monolingual calibration sets to measure how calibration choice influences these effects. Williams et al. (2025) propose self-calibration, using the model itself to generate synthetic calibration sets, which often match or surpass real data. Elhoushi and Johnson (2025) showed that calibrating using single diverse curated sample could outperform a dataset. Our work differs by isolating language composition as the key experimental variable in quantization calibration. While keeping everything else constant—the model, quantizer, calibration size, and domain—we vary only the languages used for calibration. To our knowledge, this is the first systematic study showing that multilingual calibration is a primary control factor for quantization stability, not just a secondary consideration. We test this across different



(a) Gate projection max channel activations

(b) Query projection max channel activations

Figure 4: Heatmaps of maximum activations after AWQ quantization of Llama3.1 8B for (a) the MLP gate projection and (b) the attention query projection in layer 31 (selected for its large quantization errors; see Appendix A). Rows denote different calibration language variants. Across languages, the same salient channels dominate, but their peak magnitudes shift—showing that AWQ rescales fixed channels rather than changing which channels matter, which helps explain the modest perplexity deltas.

quantizers and model families. Beyond measuring accuracy, we analyze quantization error, activation distributions, and Hessian statistics to explain why language-aware calibration improves robustness, especially for low-resource languages.

3 Experimental Set-up

3.1 Models

We assess the impact of multilingual calibration sets using two pre-trained multilingual LLMs: Llama3.1 8B (instruction-tuned)[5] and Qwen2.5 7B (instruction-tuned)[6]. They are 8B and 7B-parameter respectively. We selected these models for their multilingual properties, each being trained on 8 (Grattafiori et al., 2024; Meta AI, 2024) and 29 (Qwen et al., 2025; Qwen Team, 2024) languages respectively. To verify that the observed calibration-language effects are not specific to these model families, we additionally include BLOOMZ-7B1-MT[7] as a multilingual-heavy baseline, trained on 46 languages (Muennighoff et al., 2023). Due to space constraints, BLOOMZ-7B1-MT results are reported in the appendix (Appendix A.2). All experiments were performed on a single NVIDIA A100 40GB GPU. We utilised the official pre-trained weights available on the Hugging Face Hub for both models.

3.2 Quantization Technique

We primarily apply two post-training weight quantization methods in our experiments: GPTQ and AWQ (as discussed in the Background). To assess whether calibration-language effects generalize beyond these widely used approaches, we additionally evaluate Any4, a recent calibration-based 4-bit quantization method (Elhoushi and Johnson, 2025); detailed Any4 results are reported in Appendix A.4. For GPTQ, we follow an existing open-source implementation ModelCloud and qubitium (2024) adapted from (Frantar et al., 2023) which uses 4-bit quantization and group size 128. We set the calibration batch size to 2 in our GPTQ procedure. Similarly, for AWQ, we use 4-bit quantization and group size 128 (Lin et al., 2024). Prior work has shown that 4-bit quantization offers a favorable balance between model size and performance (Dettmers and Zettlemoyer, 2023); thus, we fixed the bit width to 4 bits for all quantization methods in our experiments.

3.3 Calibration Data Sources and Preparation

We construct eight calibration sets for each language or strategy, designed to isolate the effect of *calibration language composition* under a fixed calibration budget: (i) *English baseline*: the original English calibration sets—C4 for GPTQ and the Pile validation split for AWQ. For AWQ, we use 512 examples of 512 tokens each; for GPTQ, we use 1024 examples of 1042 tokens each.

Across all experiments, we explicitly hold the total calibration token budget constant to decouple the effect of calibration language composition from calibration size. Prior work in post-training quantization demonstrates that, beyond a moderate budget, performance is substantially more sensitive to the *distributional properties* of calibration data than to the absolute number of tokens (e.g., GPTQ, AWQ, SmoothQuant; Williams and Aletras, 2024). Accordingly, our goal is not to optimize calibration size, but to conduct controlled experiments that disentangle the impact of linguistic diversity under a fixed budget.

(ii) *Single-language translated*: translations of the English calibration set into French, Swahili,

[5]HuggingFace ID: `meta-Llama/Llama3.1 8B-instruct`
[6]HuggingFace ID: `qwen/Qwen-2.5-7B-Instruct`
[7]HuggingFace ID: `bigscience/bloomz-7b1-mt`

Chinese, and isiXhosa using the `deep-translator` library[8]. These sets provide a controlled mechanism for isolating language effects while holding lexical content constant. (iii) *C4*: randomly sampled examples from the C4 training split in each target language. (iv) *Wikipedia*: randomly sampled articles from the Wikipedia training split, excluding the first 5,000 examples reserved for evaluation.

Unlike translated calibration sets, C4 and Wikipedia provide *native multilingual data* with naturally occurring lexical, syntactic, and orthographic variation, allowing us to verify that observed effects are not artifacts of machine translation. Full results are reported in Appendix A.8.

(v) *C4 Multilingual,* `multimix`: random samples drawn from the C4 multilingual split without language balancing. (vi) *C4 Multilingual,* `multi10`: an equal allocation of $\frac{N}{10}$ examples from each of ten languages (`en`, `fr`, `sw`, `zh`, `xh`, `st`, `zu`, `yo`, `ig`, `ha`). (vii) *C4 Multilingual,* `multi`: N examples drawn uniformly from all 112 C4 languages[9]. (viii) *Code/Math variants (*`code-*`*,* `math-*`*,* `codemath-*`*)*: to assess whether rare or structured tokens (e.g., numerals, symbols, identifiers) aid quantization, we augment each base calibration set X with uniformly sampled data from DeepMind's `mathematics_dataset` (Saxton et al., 2019) and CodeParrot's `github-code-clean`[10]. All variants preserve the original calibration token budget via uniform mixing.

3.4 Evaluation Datasets and Languages

We evaluate quantization performance across multiple languages and evaluation datasets. The primary evaluation languages are English (`en`), French (`fr`), Swahili (`sw`), Chinese (`zh`), and isiXhosa (`xh`). In addition, we also evaluate 5 more low-resource languages: Sotho (`st`), Zulu (`zu`), Yoruba (`yo`), Igbo (`ig`), and Hausa (`ha`).

For each language, we measure performance on: (i) Wikipedia validation; first 5,000 samples of the available split, (ii) C4 validation set. For each model/quantizer/calibration combination, we report:

- Absolute perplexity on FP16 (float-16 precision), 4-bit weight-only baseline, vs. quantized model.

  FP16 represents a standard full-precision baseline, while "4-bit weight-only" indicates that all weights have been rounded to 4 bits but activations remain in FP16. Comparing these shows the accuracy loss from quantization.
- Δ perplexity relative to the English-only calibration baseline.

  We focus on perplexity in the main text. Any downstream task probes are deferred to the appendix and do not affect our primary conclusions.
- Downstream task accuracy on XNLI (Conneau et al., 2018), XStoryCloze (Lin et al., 2022), and Global MMLU (Singh et al., 2025). XNLI covers sentence-level natural language inference in `en`, `fr`, `sw`, and `zh`; XStoryCloze evaluates paragraph-level causal reasoning in `en`, `sw`, and `zh`; and Global MMLU provides a multilingual knowledge and reasoning benchmark spanning diverse languages and domains. Due to computational constraints, downstream evaluations are reported for the languages supported in the experiments.
- Layer-wise quantization error and activation statistics.

Calibration	XNLI Avg (N=5)	XStoryCloze Avg (N=4)	GlobalMMLU Avg (N=5)
en	45.51 ± 0.45	64.97 ± 0.61	48.90 ± 1.07
sw	$\mathbf{46.19 \pm 0.45}$	62.99 ± 0.62	44.30 ± 1.07
fr	45.43 ± 0.45	65.36 ± 0.61	48.25 ± 1.08
xh	45.18 ± 0.45	64.91 ± 0.61	47.15 ± 1.09
zh	45.15 ± 0.45	65.48 ± 0.60	49.25 ± 1.07
multi	45.54 ± 0.45	65.56 ± 0.61	$\mathbf{50.45 \pm 1.08}$
multi10	45.49 ± 0.45	$\mathbf{65.82 \pm 0.61}$	50.20 ± 1.09
multimix	44.96 ± 0.45	65.06 ± 0.61	49.95 ± 1.07

Table 1: Task-level performance (macro-averaged) using GPTQ quantization. Avg denotes an unweighted macro-average over languages available for each task (XNLI: 5, XStoryCloze: 4, MMLU: 5). Error bars are propagated from per-language standard errors assuming independence.

4 Results and Analysis

Argument 1: One-Size-Fits-All Calibration Set is Suboptimal.

Using a non-English calibration set generally leads to better post-quantization performance than using an English-only calibration set.

In Figure 2, we plot the change in perplexity: $\Delta\text{PPL} = \text{PPL}_{\text{English}} - \text{PPL}_{\text{Other}}$ on a 5000-sentence Wikipedia validation set for each language. Here, a positive delta means a reduction in perplexity (lower average loss per token), which corresponds to better model fit on the evaluation data. We see that almost all of the non-English cali-

[8] https://pypi.org/project/deep-translator
[9] https://huggingface.co/datasets/allenai/c4
[10] `https://huggingface.co/datasets/codeparrot/github-code-clean`

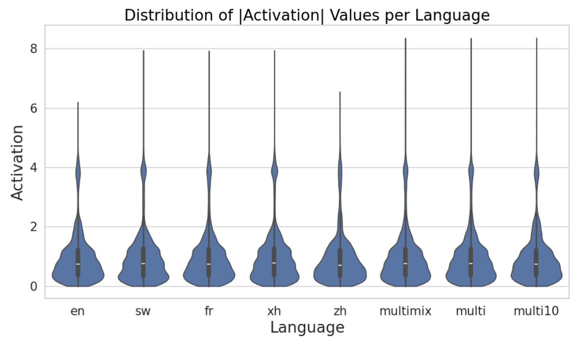


Figure 5: Activation distributions from the unquantized Llama model across different calibration sets. Violin plots compare absolute activations per set: the three multilingual variants (`multimix`, `multi`, `multi10`) exhibit longer upper tails than single-language sets (`en`, `fr`, `sw`, `zh`, `xh`), capturing higher-magnitude outliers. This broader coverage suggests that multilingual calibration is better suited to handle unseen extremes at test time.

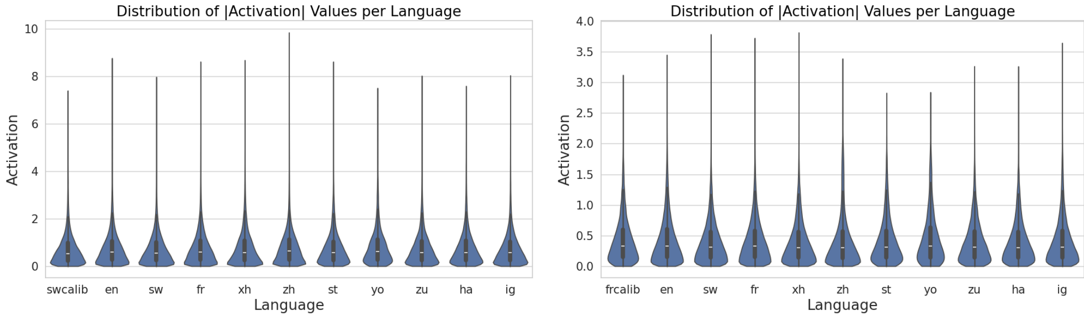
bration sets yield a positive Δ perplexity compared to the English calibration baseline. Notable exceptions are the French calibration set under AWQ and the Swahili calibration set under GPTQ on the Llama model, where performance decreases slightly relative to using the English set. We examine these two outlier cases in our analysis in Figure 6 and see that Swahili and French calibration sets exhibit narrower ranges in their activation distribution compared to the broader distribution encountered during inference. This may result in poor representation of critical outliers as out-of-distribution values are clipped to the boundaries seen during calibration, resulting in substantial quantization error. Overall, the trend supports our hypothesis that English-only calibration is suboptimal for multilingual models.

Calibration Aligned with Evaluation Language Outperforms Static Sets. Using a calibration set that matches the evaluation language yields significant gains on that language's performance for Llama3.1 8B. As shown in Table 2 and Table 3, language-specific calibration sets produce the largest perplexity reductions for their corresponding languages under AWQ. For instance, the Swahili calibration set yields an improvement on Swahili benchmarks, and similarly, the French, Chinese, and isiXhosa calibration sets achieve the highest gains for their respective languages. Moreover, calibration data can benefit closely related languages: the isiXhosa calibration set improves performance for both isiXhosa and its linguistic relatives Sotho and Zulu. This cross-language uplift is most pronounced for Llama3.1 using AWQ, where isiXhosa calibration reduces perplexity by on Sotho and on Zulu (see Table 2). To verify that the observed calibration-language effects are not model-specific, we additionally conducted experiments on BLOOMZ-7B1-MT, a natively multilingual language model. While BLOOMZ-7B1-MT exhibits greater robustness to calibration-language choice compared to LLaMA, non-English calibration sets still consistently outperform English-only calibration, confirming the generality of our findings (see Appendix A.2).

Perplexity gains from non-English calibration extend to downstream tasks. To assess whether perplexity improvements translate into functional gains, we evaluate GPTQ-quantized Llama3.1 8B on zero-shot XNLI, XStoryCloze, and Global MMLU (Table 1). Downstream performance broadly follows the perplexity trends observed above, albeit with smaller margins. On XNLI, calibration aligned with the evaluation language performs best, with Swahili calibration achieving the highest mean accuracy. In contrast, XStoryCloze favors calibration diversity: the `multi10` calibration set attains the highest average accuracy, outperforming English-only calibration. Global MMLU exhibits a similar pattern, with multilingual calibration consistently surpassing the English baseline. Across all benchmarks, both language-matched and multilingual calibration strategies outperform English-only calibration, indicating that calibration choice affects downstream task behavior rather than merely language modeling loss. We report a rank-correlation analysis between perplexity and downstream accuracy based on the XStoryCloze and Global MMLU results in Appendix A.7. While limited to these two benchmarks, which offer consistent multilingual coverage and comparable evaluation structure, the analysis reveals a stable negative correlation, supporting the conclusion that calibration strategies which reduce perplexity also tend to yield higher downstream accuracy.

Argument 2: Diverse Calibration Sets Account for Rare and Extreme Data.

Multilingual calibration sets offer the highest average performance among all settings. On Llama3.1 8B (Wikipedia), the multilingual mixes (`multimix`, multi, and `multi10`) consistently sit at the top of the Avg column, with `math-multi` essentially tying or edging the best plain multilingual average; under GPTQ, `multi10` is the best plain mix and `codemath-multi10` is the best overall average, highlighting GPTQ's sensitivity to broad



(a) Activation distributions - Swahili calibration set (swcalib) v.s. Wikipedia test data.

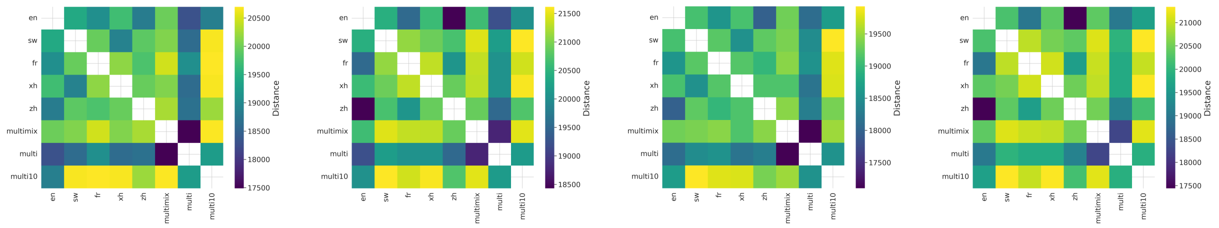
(b) Activation distributions - French calibration set (frcalib) vs. Wikipedia test data.

Figure 6: Activation-range mismatch between calibration and test sets in the last Llama3.1 8B layer. (a) With Swahili calibration, activations span up to $\approx$7.5, while most test languages exceed 8. (b) With French calibration, activations top out at $\approx$3.2, while test languages extend beyond 3.3. Missing these outliers leads to overly conservative quantization thresholds, limiting perplexity gains on test languages with broader activation ranges.

Quantization	Calibration	en	sw	fr	xh	zh	st	yo	zu	ha	ig	Avg
FP16	–	7.327	6.510	5.698	69.300	9.526	17.579	10.148	13.410	11.013	7.645	15.816
Uniform INT4	–	8.327	7.836	6.323	84.210	11.085	23.266	12.551	17.053	15.670	10.151	19.647
AWQ	**en**	7.695	5.936	5.939	66.536	9.989	15.941	8.822	10.198	9.915	7.816	14.879
AWQ	**sw**	7.679	**5.859**	5.944	65.762	9.983	15.702	8.804	10.103	9.859	7.695	14.739
AWQ	**fr**	7.693	5.950	**5.913**	67.328	9.975	16.127	8.841	10.365	9.997	7.855	15.004
AWQ	**xh**	7.703	5.914	5.951	**64.254**	10.005	**15.524**	8.721	**9.811**	9.815	**7.630**	**14.533**
AWQ	**zh**	7.683	5.924	5.938	66.004	**9.888**	15.901	8.887	10.126	9.982	7.772	14.810
AWQ	**multimix**	7.683	5.913	5.929	65.662	9.970	15.702	8.789	10.050	9.851	7.724	14.727
AWQ	**multi**	**7.675**	5.894	**5.913**	64.981	9.932	15.843	8.757	9.962	**9.753**	7.676	14.639
AWQ	**multi10**	7.688	5.874	5.926	65.773	9.994	15.590	**8.706**	10.040	9.788	7.695	14.707
AWQ	**code-multi**	7.663	5.901	5.918	65.792	9.932	15.675	8.770	9.982	9.762	7.648	14.704
AWQ	**math-multi**	7.674	5.889	5.915	65.154	9.934	15.574	8.718	9.951	9.746	7.673	14.623
AWQ	**codemath-multi**	7.664	5.896	5.921	65.460	9.937	15.499	8.730	10.001	9.811	7.683	14.660
AWQ	**codemath-multi10**	7.675	5.873	5.918	65.944	9.975	15.679	8.745	10.048	9.790	7.711	14.736
GPTQ	**en**	8.300	8.730	6.685	89.283	12.242	28.453	13.022	17.829	18.999	12.696	21.624
GPTQ	**sw**	9.286	9.633	7.908	95.637	15.054	35.093	15.035	19.274	23.086	18.786	24.879
GPTQ	**fr**	8.329	8.519	6.410	86.817	12.490	28.205	12.907	17.306	18.068	12.756	21.181
GPTQ	**xh**	8.612	7.844	6.693	79.847	12.538	25.564	12.671	15.714	17.028	11.990	19.850
GPTQ	**zh**	8.489	8.523	6.596	89.869	11.228	25.902	12.816	18.194	17.925	13.854	21.340
GPTQ	**multimix**	8.189	7.870	6.291	81.481	11.220	24.050	12.294	16.270	15.664	11.359	19.469
GPTQ	**multi**	8.211	7.604	6.260	82.564	11.050	24.725	11.890	16.304	14.545	10.060	19.321
GPTQ	**multi10**	**8.184**	**7.222**	**6.194**	**77.444**	**10.787**	**20.693**	**11.329**	**15.176**	**14.296**	**9.714**	**18.104**
GPTQ	**code-multi**	8.191	7.496	6.214	79.968	10.950	22.969	11.547	15.836	13.895	9.699	18.676
GPTQ	**math-multi**	8.190	7.523	6.188	80.772	10.851	23.110	11.600	16.039	14.069	9.846	18.819
GPTQ	**codemath-multi**	8.174	7.495	6.193	80.949	10.744	22.026	11.610	16.203	13.838	9.534	18.677
GPTQ	**codemath-multi10**	8.159	7.176	6.183	76.654	10.695	20.648	11.237	15.083	13.803	9.464	17.910

Table 2: Llama3.1 8B perplexity on Wikipedia (lower is better). Natural-language calibration: **bold** = best, underline = second-best. AWQ (incl. code/math): best, second-best. AWQ benefits from language-matched calibration locally, whereas GPTQ prefers diverse multilingual calibration (`multi10`) for robust averages.

coverage (see Table 2). All code/math variants are constructed by *uniformly sampling* from the respective pools (code, math, or code+math) to fill the same calibration budget as other settings, with identical quantization hyperparameters—so the gain comes from *content*, not data volume or tuning. We add code and math *specifically* to test our rare/unique-token hypothesis: these domains contain high-rarity symbols, numerals, operators, and identifier-like subwords that broaden coverage beyond natural text. While AWQ occasionally rewards language-matched calibration with the very best per-language scores (e.g., for `xh`), the multilingual (+ code/math) variants remain the most reliable choice across languages. The same qualitative pattern holds for Qwen2.5 7B in the Appendix (Table 3)—multilingual `multi`/`multi10` rank at or near the top for both AWQ and GPTQ—reinforcing the general principle that broader, more diverse calibration improves average performance across



(a) Layer 0 (Llama3.1 8B) (b) Layer 31 (Llama3.1 8B) (c) Layer 0 (Qwen2.5 7B) (d) Layer 27 (Qwen2.5 7B)

Figure 7: Heatmaps show inverse-Hessian distances (higher = larger change) between GPTQ calibrations for Llama3.1 8B (layers 0, 31) and Qwen2.5 7B (layers 0, 27). Among them, `multi10` shows the largest shifts, `multimix` moderate, and `multi` the smallest. Some monolingual pairs such as `en`–`zh` remain relatively close.

models, datasets, and quantizers.

Broader token/structure coverage produces longer activation tails at calibration time, reducing clipping and quantization error. Multilingual mixes exhibit the largest unique vocabularies and higher tokens per example (Figure 3), exposing quantizers to more subwords, numerals, punctuation, and script-specific patterns. Their vocabularies also overlap more with each single language—≈1.5k–3.6k shared types vs. ≈1k–3k for direct language–to–language pairs— and three-way overlaps are larger when a multilingual set is included (Figure 8). This richer lexical/structural diversity yields heavier upper activation tails (Figure 5), the very region where under-calibrated scales would otherwise clip unseen extremes. Adding code/math amplifies this effect by injecting rare symbols and bracketed/numeric patterns, further extending the tails and lowering post-quantization error—which is exactly what we observe in Table 2.

Argument 3: Calibration effectiveness is quantizer-dependent: the same calibration set can help or hurt differently depending on the algorithm's mechanics.

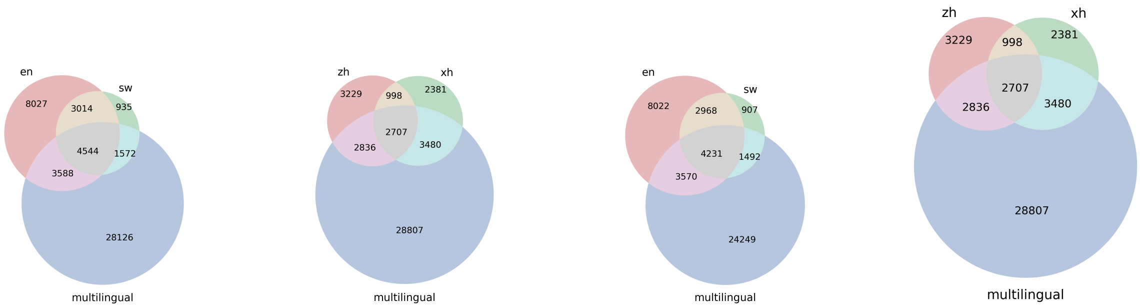
In Figure 2, GPTQ shows large swings across calibration sets (up to 3.52 ppl), while AWQ varies only slightly (max ≈ 0.35 ppl), echoing our claim that calibration is not language-agnostic and must be chosen with intent. Concretely, multilingual mixes (especially `multi10`) deliver the most robust averages for GPTQ, whereas AWQ's gains are steadier and often maximized by language-matched calibration for a given target language.

To understand this difference, we examine how each quantization algorithm interacts with the calibration data at specific layers. In particular, we focus on the extremity layers — those most affected by quantization noise. For Llama3.1 8B, these are layers 0 and 31; for Qwen2.5 7B, layers 0 and 27. These layers consistently exhibit the highest quantization error magnitudes (see Appendix A), and thus provide a focused lens through which to study how calibration data influences quantization behavior.

AWQ's activation scaling keeps channel identity stable, so calibration language mainly adjusts magnitudes, not which channels matter. In Figure 4, we plot the maximum per-channel activations at layer 31 (gate and query projections). Across all calibration languages, AWQ consistently selects the same "salient" channel, but the absolute activation peaks differ. This explains AWQ's modest perplexity deltas: when channel identity is stable, language choice fine-tunes the rescaling but rarely causes large distributional mismatches. Practically, if you must optimize a single language with AWQ, match the calibration language; if you need cross-language robustness, multilingual remains near-optimal.

GPTQ's Hessian-based error compensation is acutely sensitive to the calibration distribution, but diverse calibration stabilizes it. Because GPTQ propagates quantization residuals using an inverse-Hessian estimated from calibration activations, changing the calibration language changes the second-order statistics the algorithm relies on. Figure 7 visualizes large inverse-Hessian differences across single-language sets, aligning with GPTQ's larger Δ ppl swings; by contrast, `multi` calibrations cluster tightly in this space, indicating that linguistically diverse calibration dampens GPTQ's sensitivity. These results support our core claim: robust quantization needs linguistically diverse calibration.

To further test quantizer generality, we additionally evaluate Any4, a recent activation-aware post-training quantizer, on Qwen 2.5 (see Appendix A.4). Despite its different mechanics, Any4 exhibits the same qualitative behavior: multilingual



(a) Llama3.1 8B (b) Llama3.1 8B (c) Qwen2.5 7B (d) Qwen2.5 7B

Figure 8: Overlap analysis between the multilingual calibration set and single-language sets. Each language overlaps more with the multilingual set (≈1.5k–3.6k activations) than with another single language (≈1k–3k). The shared region across single languages and multilingual (≈2.7k–4.5k) is also larger than direct language–language intersections. This shows that multilingual calibration covers a broader slice of the activation tail and reliably captures extremes that individual languages encounter at inference time.

and multilingual-mixture calibration consistently yield lower average perplexity across languages, while monolingual calibration provides only localized improvements. This confirms that the benefits of language-diverse calibration are not specific to GPTQ or AWQ, but reflect a general interaction between calibration data and activation distributions across quantization families.

Practically, we recommend balanced multilingual calibration for GPTQ and language-matched calibration (with a multilingual fallback) for AWQ. Understanding these algorithm-specific interactions is essential for selecting calibration data that best synergizes with each quantizer's mechanics.

5 Conclusion and Future Work

Effective quantization is essential for deploying multilingual LLMs in GPU-poor environments. We show that appropriate calibration is a valuable tool to minimise loss of performance. In this paper, we presented the first systematic study of multilingual calibration sets for LLM quantization, challenging the convention of using English-only calibration sets. We observed that non-English and mixed-language calibration sets consistently outperform the English-only baseline, achieving up to a 3.52 perplexity reduction on Llama3.1 using GPTQ. English-only calibration leaves substantial accuracy gaps for many target languages. We also demonstrate that multilingual calibration sets exhibit longer activation tails (Figure 5), indicating better coverage of outlier tokens and activation values. This broader coverage directly translates into lower quantization error and more robust performance across languages. Lastly, we show that algorithm mechanics matter with GPTQ's Hessian-based error compensation being more sensitive to the calibration language choice than AWQ's activation-aware scaling (Figure 2). This interaction implies that calibration data must be chosen in harmony with the quantizer's algorithmic design.

These findings demonstrate that *calibration set design* is a pivotal factor in preserving multilingual LLM accuracy under low-precision constraints. From our experiments, we distill three practical guidelines:

- Use a single multilingual calibration set (e.g., multi10) for general multilingual deployment and reserve language-matched calibration only for single-language–dominated systems, as an optional optimization rather than a requirement.
- Targeted AWQ: If you are optimizing for one language, choose a matching calibration language dataset, with a multilingual fallback for broader robustness.
- Targeted GPTQ: Prefers balanced multilingual calibration; adding special tokens (code/math) can further stabilize second-order statistics.

Limitations

In this work, we focus on varying the language aspect of a calibration set, evaluating on perplexity. However, evaluating only perplexity provides a limited view of real-world utility; extending our framework to downstream tasks e.g., translation would more directly quantify the practical benefits of non-English and multilingual calibration. We also limit our analysis to 10 languages, two LLMs (Llama3.1 8B, Qwen2.5 7B), and two post-training quantization methods (GPTQ, AWQ) due to computational resource limits; broadening this scope to include additional languages, model families, model sizes, and quantizers is warranted to strengthen our argument. Lastly, we did not explore dynamic calibration sample selection based on quantization error signals; future work should explore this area which may lead to further reduction in quantization loss.

Acknowledgments

We thank Benjamin Rosman, Jade Abbott and the anonymous reviewers for the comments and contributions to the draft.

References

Orevaoghene Ahia, Julia Kreutzer, and Sara Hooker. 2021. The low-resource double bind: An empirical study of pruning for low-resource machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3316–3333, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485, Brussels, Belgium. Association for Computational Linguistics.

Tim Dettmers and Luke Zettlemoyer. 2023. The case for 4-bit precision: k-bit inference scaling laws. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 7750–7774. PMLR.

Mostafa Elhoushi and Jeff Johnson. 2025. any4: Learned 4-bit numeric representation for LLMs. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 15215–15231. PMLR.

Elias Frantar, Saleh Ashkboos, Torsten Hoefler, and Dan Alistarh. 2023. Gptq: Accurate post-training quantization for generative pre-trained transformers. *Preprint*, arXiv:2210.17323.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, Shawn Presser, and Connor Leahy. 2020. The Pile: An 800gb dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*.

Amir Gholami, Sehoon Kim, Zhen Dong, Zhewei Yao, Michael W. Mahoney, and Kurt Keutzer. 2021. A survey of quantization methods for efficient neural network inference. *CoRR*, abs/2103.13630.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

R.M. Gray and D.L. Neuhoff. 1998. Quantization. *IEEE Transactions on Information Theory*, 44(6):2325–2383.

Itay Hubara, Yury Nahshan, Yair Hanani, Ron Banner, and Daniel Soudry. 2021. Accurate post training quantization with small calibration sets. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 4466–4475. PMLR.

Ji Lin, Jiaming Tang, Haotian Tang, Shang Yang, Wei-Ming Chen, Wei-Chen Wang, Guangxuan Xiao, Xingyu Dang, Chuang Gan, and Song Han. 2024. Awq: Activation-aware weight quantization for llm compression and acceleration. In *MLSys*.

Xi Victoria Lin, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig, Myle Ott, Naman Goyal, Shruti Bhosale, Jingfei Du, Ramakanth Pasunuru, Sam Shleifer, Punit Singh Koura, Vishrav Chaudhary, Brian O'Horo, Jeff Wang, Luke Zettlemoyer, Zornitsa Kozareva, Mona Diab, and 2 others. 2022. Few-shot learning with multilingual generative language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9019–9052, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Kelly Marchisio, Saurabh Dash, Hongyu Chen, Dennis Aumiller, Ahmet Üstün, Sara Hooker, and Sebastian Ruder. 2024. How does quantization affect multilingual LLMs? In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 15928–15947, Miami, Florida, USA. Association for Computational Linguistics.

Meta AI. 2024. Meta Llama 3.1 8B Instruct Model on Azure Marketplace. `https://github.com/marketplace/models/azureml-meta/Meta-Llama-3-1-8B-Instruct`. Accessed: 2025-05-20.

ModelCloud and qubitium. 2024. Gptqmodel. `https://github.com/modelcloud/gptqmodel`. Contact: qubitium@modelcloud.ai.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng Xin Yong, Hailey Schoelkopf, Xiangru Tang, Dragomir Radev, Alham Fikri Aji, Khalid Almubarak, Samuel Albanie, Zaid Alyafeai, Albert Webson, Edward Raff, and Colin Raffel. 2023. Crosslingual generalization through multitask finetuning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15991–16111, Toronto, Canada. Association for Computational Linguistics.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, and 25 others. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Qwen Team. 2024. Qwen 2.5 release blog. `https://qwenlm.github.io/blog/qwen2.5/`. Accessed: 2025-05-20.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(1).

David Saxton, Edward Grefenstette, Felix Hill, and Pushmeet Kohli. 2019. Analysing mathematical reasoning abilities of neural models. *ArXiv*, abs/1904.01557.

Shivalika Singh, Angelika Romanou, Clémentine Fourrier, David Ifeoluwa Adelani, Jian Gang Ngui, Daniel Vila-Suero, Peerat Limkonchotiwat, Kelly Marchisio, Wei Qi Leong, Yosephine Susanto, Raymond Ng, Shayne Longpre, Sebastian Ruder, Wei-Yin Ko, Antoine Bosselut, Alice Oh, Andre Martins, Leshem Choshen, Daphne Ippolito, and 4 others. 2025. Global MMLU: Understanding and addressing cultural and linguistic biases in multilingual evaluation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 18761–18799, Vienna, Austria. Association for Computational Linguistics.

Miles Williams and Nikolaos Aletras. 2024. 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)*, pages 10100–10118, Bangkok, Thailand. Association for Computational Linguistics.

Miles Williams, George Chrysostomou, and Nikolaos Aletras. 2025. 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)*, pages 10149–10167, Albuquerque, New Mexico. Association for Computational Linguistics.

Zixuan Zhou, Xuefei Ning, Ke Hong, Tianyu Fu, Jiaming Xu, Shiyao Li, Yuming Lou, Luning Wang, Zhihang Yuan, Xiuhong Li, Shengen Yan, Guohao Dai, Xiao-Ping Zhang, Yuhan Dong, and Yu Wang. 2024. A survey on efficient inference for large language models. *Preprint*, arXiv:2404.14294.

A Appendix

A.1 Cross-Model Check on Wikipedia (Qwen2.5 7B)

On a different model (Qwen2.5 7B on C4), multilingual calibration again delivers the strongest averages. Table 3 mirrors the Llama–Wikipedia trend: multilingual `multi`/`multi10` rank at or near the top for both AWQ and GPTQ, while some language-matched settings win per-language minima. This cross-model replication strengthens our main claim that broad linguistic coverage in calibration improves average robustness.

A.2 Additional Model Family: BLOOM

To further test whether our findings depend on English-dominant pretraining, we also evaluate BLOOMZ-7B1-MT, a multilingual model trained with a more globally balanced pretraining mixture and weaker English dominance than Llama 3.1 or Qwen 2.5. Table 4 reports AWQ perplexity results for BLOOMZ-7B1-MT across calibration strategies. We observe the same qualitative pattern as in earlier experiments: language-matched calibration yields localized improvements for the corresponding or closely related languages, while multilingual calibration remains competitive on average. These results indicate that the benefits of multilingual and language-aware calibration extend beyond English-centric pretraining regimes and are not an artifact of a single model family.

A.3 Additional Downstream Results

Table 5 reports full language-specific downstream results for Llama3.1 8B using GPTQ quantization across XNLI, XStoryCloze, and Global MMLU.

A.4 Additional Quantizer: Any4

To assess whether our findings extend beyond GPTQ and AWQ, we additionally evaluate Any4 (Elhoushi and Johnson, 2025), a recent activation-aware post-training quantizer, on Qwen 2.5 under 4-bit quantization. Table 6 reports language-specific perplexity across calibration strategies. We observe the same pattern seen for GPTQ and AWQ: multilingual and multilingual-mixture calibration sets achieve the lowest average perplexity across languages, while single-language calibration primarily yields localized gains. The relative ordering of calibration strategies remains stable, confirming that language-diverse calibration improves global robustness even under a newer, activation-aware quantization algorithm.

A.5 Complete Per-Language Results for Llama3.1 8B on Wikipedia

Full Llama3.1 8B on Wikipedia results reinforce the main pattern: multilingual mixes win on averages, while language-matched AWQ often yields the best per-language minima. Table 8 lists AWQ runs (plus FP16 and uniform 4-bit). Multilingual `multi`/`multi10` are consistently strong on the Avg column; language-matched (e.g., `xh`) often edges the best per-language scores; and code/math-enriched multilingual sets are competitive or best on average—supporting the "rare tokens help" claim.

For GPTQ, broad coverage matters most, and code/math boosts help. In Table 9, `multi10` is the best plain mix, and `codemath-multi10` often yields the best overall average, underscoring GPTQ's sensitivity to calibration diversity and to tokens that expand activation tails (digits, brackets, symbols).

A.6 Cross-Lingual Transfer

Our results provide evidence of language-family transfer as we cover languages within the same family: English, French (Indo-European), Swahili, isiXhosa, Sotho, Zulu(Bantu families). As shown in Table 10, when calibrating using isiXhosa (xh), we observe improvements not only for isiXhosa itself, but also for structurally related Bantu languages (Zulu and Sotho), compared to English-calibrated baselines. We also see the same within-family transfer pattern under GPTQ.

A.7 Perplexity–Downstream Correlation

To directly quantify the relationship between language modeling quality and task-level behavior, we analyze the rank correlation between perplexity (PPL) and downstream accuracy using the XStoryCloze results reported in the main paper and the Global MMLU results introduced in this revision. For each language where both metrics are available, we compute Spearman's rank correlation coefficient across calibration strategies. As shown in Table 7, we observe a consistent negative correlation across languages and tasks, indicating that lower perplexity reliably corresponds to higher downstream accuracy. This confirms that calibration-induced PPL improvements translate

Quantization	Calibration	en	sw	fr	xh	zh	st	yo	zu	ha	ig	Avg
FP16	–	6.835	13.163	6.227	20.296	9.948	33.537	12.608	16.983	24.070	19.754	16.642
4-bit	–	7.562	14.766	6.639	21.599	10.769	35.509	13.723	18.828	26.601	21.456	17.845
AWQ	**en**	**7.212**	14.097	6.436	20.928	10.594	33.497	12.828	17.398	25.240	20.663	16.689
AWQ	**sw**	7.267	13.793	6.462	20.737	10.592	33.569	12.737	17.194	25.275	20.754	16.638
AWQ	**fr**	7.231	14.132	6.378	20.946	10.579	34.085	12.797	17.432	25.554	21.137	16.727
AWQ	**xh**	7.235	**13.752**	6.427	20.701	10.635	33.428	12.771	**17.055**	25.183	20.695	16.598
AWQ	**zh**	7.235	13.902	6.456	20.781	**10.467**	**32.996**	**12.637**	17.267	25.289	20.716	16.475
AWQ	**multimix**	7.243	13.900	6.415	20.743	10.563	33.173	12.738	17.279	**25.123**	20.611	16.479
AWQ	**multi**	**7.212**	13.842	**6.350**	**20.599**	10.568	32.997	**12.637**	17.088	25.213	**20.448**	**16.395**
GPTQ	**en**	**7.324**	14.958	6.617	21.665	11.661	37.648	13.779	18.789	26.789	21.717	17.995
GPTQ	**sw**	7.570	**13.703**	6.677	20.996	11.805	35.089	13.286	17.898	25.020	20.816	17.186
GPTQ	**fr**	7.489	14.715	6.453	21.516	11.772	36.549	13.606	18.705	26.928	22.139	17.707
GPTQ	**xh**	7.648	13.974	6.770	**20.627**	11.650	**34.379**	13.439	**17.343**	25.395	21.158	17.178
GPTQ	**zh**	7.473	14.973	6.683	21.543	11.689	36.945	13.767	18.705	27.305	21.935	17.802
GPTQ	**multimix**	7.451	14.385	6.562	21.047	13.147	35.451	13.335	18.007	25.605	21.935	17.494
GPTQ	**multi**	7.482	14.121	6.465	21.126	10.712	34.550	13.134	17.994	**24.841**	**20.523**	16.895
GPTQ	**multi10**	7.424	13.732	**6.451**	**20.627**	**10.520**	34.390	**12.976**	17.424	25.378	20.616	**16.563**

Table 3: Per-language perplexity for Qwen2.5 7B (lower is better) on the Wikipedia evaluation set, with per-language scores and the overall **Avg**. The first two rows show the FP16 baseline and an uncalibrated 4-bit quantization result. The remaining rows report results using AWQ (top block) and GPTQ (bottom block) with calibration data drawn from a single language (`en`, `sw`, `fr`, `xh`, `zh`) or from one of three multilingual calibration sets (`multimix`, `multi`, `multi10`). Bold values indicate the best perplexity per language, while underlined values indicate the second-best. For AWQ, the multilingual `multi` configuration yields the best or second-best performance in 7 out of 10 languages, outperforming most language-specific variants and suggesting a strong generalization benefit from multilingual calibration. For GPTQ, `multi10` delivers the best results in 5 out of 10 languages and is second-best in 4 others, demonstrating that diverse calibration data offers consistent improvements across languages, but not as uniformly as in AWQ.

Calibration	en	sw	zh	xh	st
en	**14.3781**	18.2803	22.9508	120.1768	**37.4696**
sw	14.4567	**18.1266**	22.9426	**119.5649**	37.7456
xh	14.4829	18.2722	22.9723	120.3146	38.2460
zh	14.4376	18.3402	**22.8780**	120.7325	37.5867
multi10	14.4096	18.1300	22.9440	120.4858	37.8344
multimix	14.4057	18.2377	22.9303	120.5694	37.6794
multi	14.5457	18.3571	22.9825	120.9074	37.9532

Table 4: AWQ perplexity results for BLOOMZ-7B1-MT under different calibration strategies. Language-matched calibration provides localized gains, while multilingual calibration remains competitive across languages.

into functional task gains, rather than reflecting purely distributional artifacts.

A.8 Translated vs. Native Calibration Data

Several single-language calibration sets in our experiments are constructed by machine-translating the English calibration corpus. This design enables a controlled isolation of language effects while holding lexical content and calibration budget fixed. To ensure that our conclusions are not driven by artifacts of machine translation, we additionally perform calibration using *native* multilingual corpora—specifically, language-specific subsets of C4 and Wikipedia—where calibration data is drawn from naturally occurring text. Corresponding results are reported in Tables 11 to 14. Across both Llama3.1 8B and Qwen2.5 7B, and for both GPTQ and AWQ, we observe consistent qualitative behavior between translated and native calibration sources. In particular, language-diverse calibration consistently outperforms English-only baselines, and within-family transfer effects (e.g., among Bantu languages) remain stable. These findings indicate that the benefits of language-specific calibration reflect genuine linguistic and distributional properties of the calibration data, rather than effects specific to machine-translated text.

A.9 Controlled Calibration Design: Same-Corpus Calibration for GPTQ and AWQ

Our primary calibration choices—C4 for GPTQ and the Pile validation split for AWQ—follow the default preprocessing pipelines and recommended usage in the respective method papers and reference implementations. To verify that the behavioral differences observed between GPTQ and AWQ are not artifacts of calibrating on different corpora, we conduct controlled experiments in which

Calibration	XNLI					XStoryCloze				Global MMLU				
	en	fr	sw	zh	hi	en	sw	zh	hi	en	sw	fr	xh	yo
en	**53.94 ± 1.00**	51.81 ± 1.00	38.11 ± 0.97	39.20 ± 0.98	44.50 ± 1.00	78.62 ± 1.06	53.74 ± 1.28	65.19 ± 1.23	62.34 ± 1.25	64.00 ± 2.39	37.75 ± 2.42	57.50 ± 2.47	53.75 ± 2.48	31.50 ± 2.31
fr	53.29 ± 1.00	52.45 ± 1.00	37.55 ± 0.97	37.07 ± 0.97	46.79 ± 1.00	78.69 ± 1.05	54.93 ± 1.28	65.06 ± 1.23	62.74 ± 1.24	63.00 ± 2.39	39.00 ± 2.44	55.50 ± 2.49	51.00 ± 2.49	32.75 ± 2.32
sw	52.45 ± 1.00	52.05 ± 1.00	**40.20 ± 0.98**	**39.56 ± 0.98**	46.67 ± 1.00	77.70 ± 1.07	53.01 ± 1.28	62.74 ± 1.24	58.50 ± 1.27	61.00 ± 2.38	37.75 ± 2.43	44.00 ± 2.48	50.00 ± 2.49	28.75 ± 2.26
zh	53.01 ± 1.00	**53.01 ± 1.00**	38.76 ± 0.98	35.30 ± 0.96	45.66 ± 1.00	**79.81 ± 1.03**	54.00 ± 1.28	64.59 ± 1.23	63.53 ± 1.24	**66.50 ± 2.35**	37.75 ± 2.44	56.50 ± 2.46	53.25 ± 2.49	32.25 ± 2.32
xh	51.93 ± 1.00	52.41 ± 1.00	39.48 ± 0.98	37.91 ± 0.97	44.18 ± 1.00	78.49 ± 1.06	54.60 ± 1.28	64.26 ± 1.23	62.28 ± 1.25	62.50 ± 2.41	39.00 ± 2.45	49.75 ± 2.48	51.50 ± 2.50	33.00 ± 2.35
multi	52.97 ± 1.00	51.77 ± 1.00	39.92 ± 0.98	37.43 ± 0.97	45.62 ± 1.00	78.76 ± 1.05	54.73 ± 1.28	64.73 ± 1.23	**64.00 ± 1.24**	66.25 ± 2.34	39.00 ± 2.43	56.75 ± 2.46	**56.00 ± 2.49**	**34.25 ± 2.35**
multi10	52.89 ± 1.00	50.64 ± 1.00	39.16 ± 0.98	36.99 ± 0.97	**47.75 ± 1.00**	78.89 ± 1.05	**55.72 ± 1.28**	**65.32 ± 1.22**	63.34 ± 1.24	65.00 ± 2.37	**42.25 ± 2.46**	58.25 ± 2.46	53.75 ± 2.50	31.75 ± 2.31
multimix	53.17 ± 1.00	52.85 ± 1.00	38.92 ± 0.98	34.38 ± 0.95	45.50 ± 1.00	78.16 ± 1.06	53.47 ± 1.28	64.86 ± 1.23	63.73 ± 1.24	65.50 ± 2.37	36.50 ± 2.40	**60.00 ± 2.44**	54.50 ± 2.49	33.25 ± 2.31

Table 5: Full language-specific downstream accuracy (%) for Llama3.1 8B. The table reports per-language results for XNLI (en, fr, sw, zh, hi), XStoryCloze (en, sw, zh, hi), and Global MMLU (en, sw, fr, xh, yo).

Quant.	Calibration	en	sw	fr	xh	st	yo	zu	ha	ig	Avg
FP16	**None**	9.29	28.69	8.10	53.91	38.93	20.88	40.20	36.64	22.89	28.84
4-bit	**None**	10.06	32.68	8.85	59.99	41.96	22.78	44.36	39.54	24.59	31.64
Any4	**en**	**9.82**	31.77	8.65	58.27	41.14	22.12	43.25	38.71	24.15	30.88
Any4	**sw**	9.89	**31.57**	8.66	59.16	41.37	22.19	43.75	38.97	24.55	31.12
Any4	**fr**	9.84	31.80	8.65	**58.03**	41.44	22.19	43.35	38.86	24.36	30.95
Any4	**xh**	9.91	31.59	8.68	58.12	41.29	21.96	43.19	**38.53**	24.29	**30.84**
Any4	**zh**	9.88	32.04	8.68	58.47	41.82	22.22	43.51	38.99	24.39	31.11
Any4	**multimix**	9.87	31.82	**8.64**	58.55	**40.99**	**21.95**	43.21	38.88	24.19	30.90
Any4	**multi**	9.88	31.91	8.68	58.91	41.01	22.17	43.67	38.90	24.38	31.06
Any4	**multi10**	9.88	31.78	8.67	58.14	41.21	22.06	**43.14**	38.66	**24.12**	30.85

Table 6: Perplexity results for Any4 4-bit quantization on Qwen2.5 across calibration strategies. Multilingual and multilingual-mixture calibration yield consistently lower perplexity across languages compared to monolingual calibration.

Language	XStoryCloze ρ	Global MMLU ρ
English (en)	-0.21	-0.70
Swahili (sw)	-0.46	-0.29
Chinese (zh)	-0.14	-0.20

Table 7: Spearman rank correlation (ρ) between perplexity (PPL) and downstream accuracy across calibration strategies. Negative values indicate that lower perplexity corresponds to higher task performance.

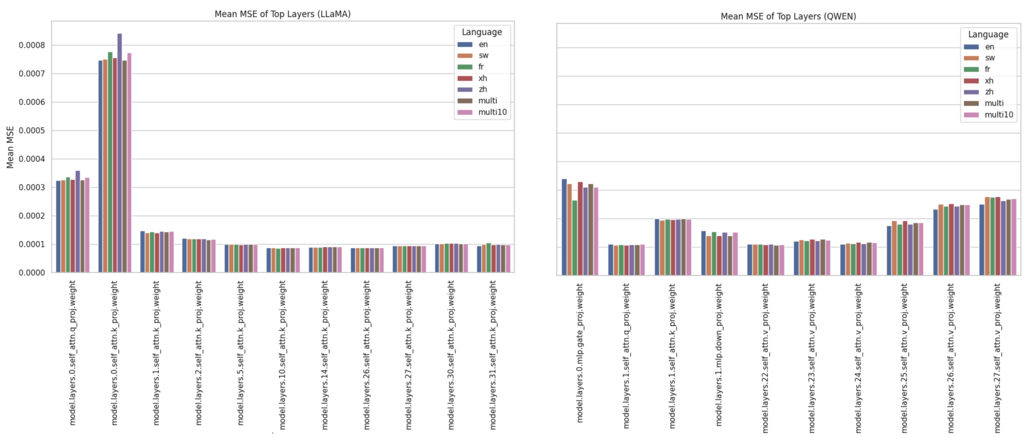
both quantizers are calibrated on the *same* underlying corpus, while varying only the calibration language composition. Concretely, we evaluate (i) Wikipedia→Wikipedia calibration and (ii) C4→C4 calibration for both GPTQ and AWQ, across two model families (Llama3.1 8B and Qwen2.5 7B). Results in Tables 11 to 14 show that the qualitative differences between GPTQ and AWQ persist under identical calibration corpora: multilingual and multilingual-mixture calibration improves global robustness, while language-matched calibration yields localized gains. This confirms that the quantizer-specific effects reported in the main text arise from genuine algorithmic differences rather than corpus selection.

A.10 Impact of Quantization on Layers

Quantization error concentrates in the extremal layers, and its magnitude varies with the calibration set. Beyond perplexity,we measure weight MSE and inspect activation statistics across layers to see where calibration matters most. For both models, the largest errors appear in the first and last transformer blocks—layers 0 and 31 for Llama3.1 8B and layers 0 and 27 for Qwen2.5 7B—providing a focused lens for algorithm–data interactions. In Figure 9, bars give the mean-squared error between original and quantized weights for the most error-prone tensors, color-coded by calibration set.

Quantization	Calibration	en	sw	fr	xh	zh	st	yo	zu	ha	ig	Avg
FP16	–	7.327	6.510	5.698	9.526	69.300	17.579	10.148	13.410	11.013	7.645	15.816
Uniform INT4	–	8.327	7.836	6.323	11.085	84.210	23.266	12.551	17.053	15.670	10.151	19.647
AWQ	**en**	7.695	5.936	5.939	9.989	66.536	15.941	8.822	10.198	9.915	7.816	14.879
AWQ	**sw**	7.679	**5.859**	5.944	9.983	65.762	15.702	8.804	10.103	9.859	7.695	14.739
AWQ	**fr**	7.693	5.950	**5.913**	9.975	67.328	16.127	8.841	10.365	9.997	7.855	15.004
AWQ	**xh**	7.703	5.914	5.951	10.005	**64.254**	**15.524**	8.721	**9.811**	9.815	**7.630**	**14.533**
AWQ	**zh**	7.683	5.924	5.938	9.888	66.004	15.901	8.887	10.126	9.982	7.772	14.810
AWQ	**multi**	**7.675**	5.894	**5.913**	9.932	64.981	15.843	8.757	9.962	**9.753**	7.676	14.639
AWQ	**multi10**	7.688	5.874	5.926	9.994	65.773	15.590	**8.706**	10.040	9.788	7.695	14.707
AWQ	**multimix**	7.683	5.913	5.929	**9.970**	65.662	15.702	8.789	10.050	9.851	7.724	14.727
AWQ	**code**	7.666	5.926	5.935	9.959	66.480	15.737	8.744	10.171	9.925	7.735	14.828
AWQ	**code-en**	7.658	5.927	5.925	9.949	65.907	15.715	8.768	10.034	9.878	7.776	14.754
AWQ	**code-sw**	7.676	5.876	5.945	9.988	66.097	15.959	8.814	10.171	9.928	7.689	14.814
AWQ	**code-fr**	7.675	5.924	**5.905**	9.959	66.631	15.933	8.820	10.255	9.858	7.730	14.869
AWQ	**code-xh**	7.688	5.906	5.944	9.994	**64.516**	**15.374**	8.743	**9.846**	9.802	7.758	14.557
AWQ	**code-zh**	7.677	5.927	5.928	**9.881**	65.805	15.761	8.888	10.086	9.952	7.808	14.771
AWQ	**code-multi**	7.663	5.901	5.918	9.932	65.792	15.675	8.770	9.982	**9.762**	**7.648**	14.704
AWQ	**code-multi10**	**7.657**	**5.864**	5.924	9.966	65.541	15.482	**8.742**	10.020	9.763	7.691	**14.665**
AWQ	**code-multimix**	7.667	5.928	5.917	9.959	65.846	15.695	8.790	10.071	9.900	7.746	14.752
AWQ	**math**	7.743	6.013	5.982	10.091	66.868	16.055	8.892	10.287	10.077	7.887	14.990
AWQ	**math-en**	7.685	5.918	5.925	9.981	66.263	15.884	8.775	10.181	9.903	7.799	14.831
AWQ	**math-sw**	7.691	**5.872**	5.942	9.988	65.941	15.716	8.747	10.142	9.863	**7.640**	14.754
AWQ	**math-fr**	7.675	5.935	**5.908**	9.970	65.880	15.916	8.886	10.093	9.787	7.726	14.778
AWQ	**math-xh**	7.694	5.922	5.945	10.012	**64.327**	15.505	8.777	**9.818**	9.800	7.719	**14.552**
AWQ	**math-zh**	7.687	5.930	5.937	**9.886**	66.490	15.865	8.886	10.182	9.956	7.737	14.856
AWQ	**math-multi**	**7.674**	5.889	5.915	9.934	65.154	15.574	**8.718**	9.951	**9.746**	7.673	14.623
AWQ	**math-multi10**	7.694	5.883	5.937	9.986	65.614	**15.441**	8.762	10.035	9.815	7.689	14.686
AWQ	**math-multimix**	7.690	5.928	5.934	9.972	65.834	15.758	8.812	10.100	9.767	7.702	14.750
AWQ	**codemath**	7.672	5.934	5.936	9.964	67.009	15.994	8.785	10.197	9.901	7.759	14.915
AWQ	**codemath-en**	7.681	5.917	5.927	9.968	66.284	15.805	8.731	10.225	9.944	7.818	14.830
AWQ	**codemath-sw**	7.694	**5.873**	5.947	9.991	66.149	15.849	8.758	10.140	9.864	7.686	14.795
AWQ	**codemath-fr**	7.672	5.934	**5.909**	9.976	66.358	15.906	8.823	10.165	9.871	7.747	14.836
AWQ	**codemath-xh**	7.687	5.908	5.946	9.998	**64.352**	**15.405**	8.751	**9.827**	9.800	7.704	**14.538**
AWQ	**codemath-zh**	7.684	5.937	5.935	**9.879**	65.811	15.685	8.830	10.090	9.900	7.737	14.749
AWQ	**codemath-multi**	**7.664**	5.896	5.921	9.937	65.460	15.499	**8.730**	10.001	9.811	**7.683**	14.660
AWQ	**codemath-multi10**	7.675	**5.873**	5.918	9.975	65.944	15.679	8.745	10.048	**9.790**	7.711	14.736
AWQ	**codemath-multimix**	7.686	5.915	5.933	9.966	66.047	15.573	8.853	10.091	9.808	7.716	14.759

Table 8: Llama3.1 8B Per-language perplexity for AWQ (plus FP16 and uniform INT4 baselines) on the Wikipedia evaluation set (lower is better). Within each block (natural / code / math / codemath), **bold** marks the best and underline marks the second-best per column. Across all AWQ rows, marks the best overall per column and marks the second-best overall; global-best cells use **bold+purple**. Multilingual calibration (`multi`/`multi10`) achieves top or near-top averages; language-matched AWQ often wins per-language minima; and code/math additions to multilingual sets provide small extra gains on average.



(a) Llama3.1 8B

(b) Qwen2.5 7B

Figure 9: Bars give the mean-squared error (MSE) between original and quantized weights for the most error-prone tensors in (a) Llama3.1 8B and (b) Qwen2.5 7B, colour-coded by calibration set (`en`, `sw`, `fr`, `xh`, `zh`, `multi`, `multi10`). Errors are more prominent in on layer 0 and for Qwen last layers.

Quantization	Calibration	en	sw	fr	xh	zh	st	yo	zu	ha	ig	Avg
FP16	–	7.327	6.510	5.698	9.526	69.300	17.579	10.148	13.410	11.013	7.645	15.816
Uniform INT4	–	8.327	7.836	6.323	11.085	84.210	23.266	12.551	17.053	15.670	10.151	19.647
GPTQ	en	8.3	8.73	6.685	12.242	89.283	28.453	13.022	17.829	18.999	12.696	21.624
GPTQ	sw	9.286	9.633	7.908	15.054	95.637	35.093	15.035	19.274	23.086	18.786	24.879
GPTQ	fr	8.329	8.519	6.41	12.49	86.817	28.205	12.907	17.306	18.068	12.756	21.181
GPTQ	xh	8.612	7.844	6.693	12.538	79.847	25.564	12.671	15.714	17.028	11.99	19.85
GPTQ	zh	8.489	8.523	6.596	11.228	89.869	25.902	12.816	18.194	17.925	13.854	21.34
GPTQ	multimix	8.189	7.87	6.291	11.22	81.481	24.05	12.294	16.27	15.664	11.359	19.469
GPTQ	multi	8.211	7.604	6.26	11.05	82.564	24.725	11.89	16.304	14.545	10.06	19.321
GPTQ	multi10	**8.184**	**7.222**	**6.194**	**10.787**	**77.444**	**20.693**	**11.329**	**15.176**	**14.296**	**9.714**	**18.104**
GPTQ	code	8.268	7.768	6.364	11.089	82.555	**22.846**	12.059	16.544	14.834	10.293	19.262
GPTQ	code-en	8.302	8.476	6.656	11.806	88.28	28.43	12.585	17.729	18.043	13.355	21.366
GPTQ	code-sw	8.436	**7.45**	6.583	11.738	81.903	26.098	12.403	16.265	16.904	11.466	19.925
GPTQ	code-fr	8.238	7.88	6.245	11.32	82.009	24.961	12.126	16.246	15.536	11.112	19.567
GPTQ	code-xh	8.472	7.967	6.617	12.206	78.896	25.842	12.316	**15.621**	16.928	12.123	19.699
GPTQ	code-zh	8.292	8.161	6.459	11.061	84.575	25.325	12.224	16.834	16.647	11.946	20.152
GPTQ	code-multi	**8.191**	7.496	**6.214**	**10.95**	**79.968**	22.969	**11.547**	15.836	**13.895**	**9.699**	**18.676**
GPTQ	code-multi10	8.462	7.971	6.515	11.742	83.681	26.239	11.951	16.459	17.113	11.695	20.183
GPTQ	code-multimix	8.286	8.289	6.393	11.498	84.651	25.529	12.32	16.907	17.342	11.769	20.298
GPTQ	math	9.064	10.196	7.378	13.528	100.6	32.712	14.731	20.347	24.035	16.717	24.931
GPTQ	math-en	8.263	9.037	6.593	11.809	90.716	29.414	12.837	18.077	19.416	12.431	21.859
GPTQ	math-sw	8.813	8.45	7.188	13.721	90.164	30.551	13.625	18.066	20.476	14.857	22.591
GPTQ	math-fr	8.216	7.882	6.193	11.356	80.508	24.129	12.167	16.069	15.69	11.156	19.337
GPTQ	math-xh	8.334	**7.295**	6.381	11.295	**73.26**	**20.281**	**11.52**	**14.515**	**13.904**	**9.528**	**17.631**
GPTQ	math-zh	8.338	8.28	6.471	10.894	87.605	27.126	12.385	17.504	17.501	12.747	20.885
GPTQ	math-multi10	8.398	7.805	6.458	11.575	82.704	24.904	11.855	16.362	17.17	11.6	19.883
GPTQ	math-multi	8.19	7.523	**6.188**	**10.851**	80.772	23.11	11.6	16.039	14.069	9.846	18.819
GPTQ	math-multimix	**8.178**	7.963	6.303	11.379	81.328	24.029	12.102	16.25	16.273	10.924	19.473
GPTQ	codemath	8.349	8.225	6.438	11.217	84.687	25.21	12.347	17.313	16.325	11.408	20.152
GPTQ	codemath-en	8.199	8.134	6.414	11.424	83.765	25.346	12.374	16.824	17.464	11.452	20.14
GPTQ	codemath-sw	8.418	7.557	6.643	12.108	82.33	26.268	12.616	16.257	16.798	12.135	20.113
GPTQ	codemath-fr	8.341	8.367	6.334	11.606	86.197	27.064	12.713	17.342	17.631	12.189	20.778
GPTQ	codemath-xh	8.348	7.571	6.43	11.375	**76.623**	23.131	12.08	15.219	15.316	11.045	18.714
GPTQ	codemath-zh	8.287	8.013	6.38	**10.677**	83.715	24.378	12.067	16.521	15.908	10.579	19.652
GPTQ	codemath-multi	8.174	7.495	6.193	10.744	80.949	22.026	11.61	16.203	13.838	9.534	18.677
GPTQ	codemath-multi10	8.159	**7.176**	**6.183**	10.695	76.654	**20.648**	**11.237**	**15.083**	**13.803**	**9.464**	**17.91**
GPTQ	codemath-multimix	**8.141**	7.672	6.28	11.052	78.737	21.988	11.715	15.668	14.431	10.089	18.577

Table 9: Llama3.1 8B per-language perplexity for GPTQ (plus FP16 and uniform INT4 baselines) on the Wikipedia evaluation set (lower is better). Within each block (Natural / Code / Math / Codemath), **bold** marks the best and underline marks the second-best per column. Across all GPTQ rows (all blocks combined), marks the best overall per column and marks the second-best overall. `multi10` is the strongest plain multilingual mix; code/math-enriched multilingual sets (e.g., `codemath-multi10`) deliver the best averages overall, highlighting GPTQ's gain from diverse, outlier-rich calibration.

Calibration Language	AWQ			GPTQ		
	xh	en	Improvement (↓)	xh	en	Improvement (↓)
isiXhosa (xh)	64.254	66.536	2.282	79.847	89.283	9.436
Zulu (zu)	9.811	10.198	0.387	15.714	17.829	2.115
Sotho (st)	15.524	15.941	0.417	25.564	28.453	2.889

Table 10: Perplexity comparison of calibration languages for AWQ and GPTQ across evaluation languages for Llama3.1 8B. Lower values indicate better performance.

Quantizer	Calibration	en	sw	fr	zh	st	xh	yo	zu	ha	ig	Avg
AWQ	en	7.641	5.911	5.921	9.961	15.823	65.947	8.772	10.109	9.850	7.756	14.769
AWQ	fr	7.671	5.916	5.905	9.974	15.666	65.653	8.762	10.061	9.838	7.692	14.714
AWQ	sw	7.691	5.864	5.944	9.978	15.699	65.921	8.728	10.121	9.946	7.672	14.756
AWQ	xh	7.707	5.913	5.948	10.017	15.423	63.682	8.729	9.768	9.848	7.666	14.470
AWQ	zh	7.691	5.917	5.936	9.861	15.851	66.348	8.818	10.187	9.978	7.811	14.840
GPTQ	en	7.994	7.658	6.286	11.623	23.507	82.570	11.884	16.187	14.979	10.876	19.356
GPTQ	sw	9.131	9.905	7.444	15.010	39.161	98.595	14.577	19.890	25.060	19.789	25.856
GPTQ	fr	8.109	7.708	6.137	11.107	23.952	82.231	11.896	16.373	14.901	10.726	19.314
GPTQ	xh	8.364	7.692	6.519	11.855	25.236	78.434	12.009	15.468	16.669	11.515	19.376
GPTQ	zh	8.510	8.516	6.671	11.383	27.828	87.892	12.454	17.846	17.557	13.385	21.204

Table 11: Wikipedia→Wikipedia controlled calibration results (perplexity) for Llama3.1 8B, where both GPTQ and AWQ are calibrated on the same corpus.

Quantizer	Calibration	en	sw	fr	zh	st	xh	yo	zu	ha	ig	Avg
AWQ	en	7.194	13.923	6.435	10.639	33.023	20.639	12.744	17.078	25.076	20.531	16.728
AWQ	fr	7.212	13.842	6.350	10.568	32.997	20.599	12.637	17.088	25.213	20.448	16.695
AWQ	sw	7.241	13.675	6.440	10.615	33.807	20.820	12.812	17.245	25.006	20.856	16.852
AWQ	xh	7.272	13.727	6.452	10.713	33.737	20.619	12.824	17.036	25.200	20.852	16.843
AWQ	zh	7.223	13.941	6.422	10.349	32.936	20.716	12.748	17.234	25.216	20.497	16.728
GPTQ	en	13.440	26.310	8.108	13.693	51.723	50.512	25.755	29.092	28.213	16.910	26.376
GPTQ	sw	13.718	23.996	8.083	14.251	47.954	47.362	24.417	27.794	26.809	16.675	25.106
GPTQ	fr	13.644	26.118	7.877	14.253	50.867	49.204	24.704	28.817	28.021	16.949	26.046
GPTQ	xh	13.748	24.457	8.112	13.159	46.910	46.787	24.149	27.339	26.919	16.383	24.796
GPTQ	zh	13.937	26.027	8.271	12.704	50.932	49.071	25.080	28.879	27.735	16.661	25.930

Table 12: Wikipedia→Wikipedia controlled calibration results (perplexity) for Qwen2.5 7B, where both GPTQ and AWQ are calibrated on the same corpus.

Quantizer	Calibration	en	sw	fr	zh	xh	st	yo	zu	ha	ig	Avg
AWQ	en	12.496	11.695	8.893	13.072	–	24.153	18.451	22.376	15.726	10.723	15.287
AWQ	fr	12.471	–	8.830	13.016	31.623	24.044	18.307	22.053	15.600	10.683	17.403
AWQ	sw	12.552	11.449	8.892	13.050	31.623	23.619	18.200	–	15.649	10.680	16.190
AWQ	xh	12.514	11.575	8.885	12.966	30.999	23.517	17.946	21.697	15.531	10.580	16.621
AWQ	zh	12.517	11.660	8.892	12.894	31.712	23.906	18.498	22.267	15.889	10.718	16.895
GPTQ	en	15.473	16.319	10.767	17.861	43.863	40.179	23.115	32.704	25.826	14.551	24.066
GPTQ	sw	17.573	15.267	11.428	19.616	43.076	41.675	23.090	33.459	26.192	15.428	24.680
GPTQ	fr	15.289	14.028	9.843	16.236	39.774	32.883	21.540	30.626	22.193	13.211	21.562
GPTQ	xh	16.066	14.154	10.492	16.645	38.736	35.667	21.620	30.256	22.644	14.102	22.038
GPTQ	zh	16.038	14.845	10.488	15.082	40.798	33.604	20.983	32.145	23.598	13.642	22.122

Table 13: C4→C4 controlled calibration results (perplexity) for Llama3.1 8B. Missing entries indicate unavailable measurements in the provided runs.

Quantizer	Calibration	en	sw	fr	zh	xh	st	yo	zu	ha	ig	Avg
AWQ	en	13.524	24.613	7.905	12.554	31.807	36.384	26.660	24.429	29.204	17.006	22.408
AWQ	fr	13.502	24.341	7.801	12.594	31.568	36.417	26.515	24.291	29.009	17.083	22.312
AWQ	sw	13.597	23.786	7.914	12.546	31.391	36.418	26.480	24.271	28.983	17.300	22.269
AWQ	xh	13.577	23.965	7.888	12.553	30.989	35.982	26.772	24.008	29.070	17.108	22.191
AWQ	zh	13.528	24.286	7.922	12.308	31.445	36.263	26.846	24.178	29.196	17.071	22.304
GPTQ	en	13.440	26.310	8.108	13.693	51.723	50.512	25.755	29.092	28.213	16.910	26.376
GPTQ	sw	13.718	23.996	8.083	14.251	47.954	47.362	24.417	27.794	26.809	16.675	25.106
GPTQ	fr	13.644	26.118	7.877	14.253	50.867	49.204	24.704	28.817	28.021	16.949	26.046
GPTQ	xh	13.748	24.457	8.112	13.159	46.910	46.787	24.149	27.339	26.919	16.383	24.796
GPTQ	zh	13.937	26.027	8.271	12.704	50.932	49.071	25.080	28.879	27.735	16.661	25.930

Table 14: C4→C4 controlled calibration results (perplexity) for Qwen2.5 7B, where both GPTQ and AWQ are calibrated on the same corpus.