Multilinguality after Quantization: Low-Bits, Calibration Languages, Mixed-Precisions, and Cross-Linguality

Muhammad Rafif Priyo Waskito[1], **Alfan Farizki Wicaksono**[1], **Alham Fikri Aji**[2], **Haryo Akbarianto Wibowo**[2]

[1]Faculty of Computer Science, Universitas Indonesia
[2]Departement of Natural Language Processing, MBZUAI
muhammad.rafif01@ui.ac.id, alfan@cs.ui.ac.id, {alham.fikri, haryo.wibowo}@mbzuai.ac.ae

Abstract

There are inconsistent conclusions about the impact of calibration data on Post-Training Quantization (PTQ) across studies, specifically regarding the factors that make the method effective, such as language that we explore in this work. We investigate this question by evaluating `GPTQ` and its mixed-precision variants across three low-bit settings, using language-targeted calibration sets across five language families. By evaluating performance across both perplexity and downstream tasks, we find that selecting an appropriate calibration language is generally beneficial, though model sensitivity varies with bit-width, target language, and task difficulty. We further show that mixed-precision techniques expand the representational capacity of low-bit models, occasionally allowing 2-bit models to approach 3-bit performance when paired with language-targeted calibration, while increased effective-bits remains small. Furthermore, we explore the correlation between these factors and observe that the calibration set sensitivity is influenced by different combinations of factors for each controlled variable, which can be explained by the downstream task results. Finally, while the benefit of language-targeted calibration transfers to sibling languages within the same family, the magnitude of this transfer does not reliably predicted by their closeness. Together, these findings show calibration data selection as a deliberate design choice in multilingual PTQ rather than an incidental one. Our code and result data can be accessed in https://github.com/rafifpriyo/multilingual-calibration-study.

1 Introduction

The impressive multilingual capabilities of Large Language Models (LLMs) scale with model size (Lin et al., 2022; Yang et al., 2025; Dang et al., 2024; Grattafiori et al.); as a result, deploying the model becomes expensive. To address this, Post-Training Quantization (`PTQ`) has become a widely used compression technique, as it offers a highly efficient compression method while retaining the original model's performance by conditioning on a small calibration dataset (e.g., 128 samples) (Gholami et al., 2021; Rakka et al., 2025). However, existing works primarily focus on English performance, thus making the field of quantization on low-resource languages underexplored, especially at extreme quantization levels (3- and 2-bit) (Marchisio et al., 2024; Zeng et al., 2025; Kurz et al., 2025; Choenni and Titov, 2025). Furthermore, since `PTQ` is a non-training approach, optimizing the model for a specific target language is challenging, leaving the single-pass calibration set as the primary adjustable hyperparameter.

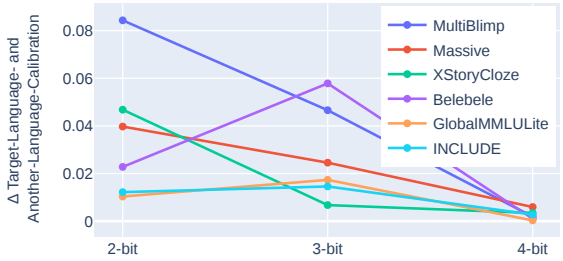


Figure 1: When evaluating on target language, Quantization using mono-target-language calibration set generally outperforms mono-another-language, while the sensitivity varies between bit-width and task difficulty.

Despite the intuitive potential to improve `PTQ` via calibration set composition, empirical studies report varying results across problem sets. Interestingly, the perceived impact of calibration sets has shifted chronologically, from negligible to substantial results. Rather than dismissing these diverse outcomes as random noise or coincidence, the underlying causes of these discrepancies demand closer analysis.

In this work, we offer a new perspective on

Findings of the Association for Computational Linguistics: IJCNLP-AACL 2026, pages 155–180
November 6-10, 2026 ©2026 Association for Computational Linguistics

the discrepant effects of calibration set by investigating multiple variables. We want to answer the following question: *How calibration set affects quantized model's performance in progressive representational power across task difficulties, i.e., smooth transition between bits through mixed-precision?* Specifically, we experiment using `GPTQ` (Frantar et al., 2023a) as the representative `PTQ` method, evaluating it across three low bit-levels (4-bit, 3-bit, and 2-bit). To smooth the transition between bit-levels, we incorporate both "light" and "heavy" mixed-precision variants (§3.1) (Huang et al., 2025; Xiao et al., 2025). Finally, we evaluate the effectiveness of language-targeted calibration across five language families, focusing on the relative improvement over other-language calibration.

Our primary contributions are as follows:

- We show that the effect of language-specific calibration is generally favorable relative to other-language calibrations. The advantage is clear in perplexity, although it is rarely substantial in downstream tasks.
- We observed that mixed-precision methods help stabilize the model's sensitivity to the calibration set. In some cases, they enable lower-bit models to compete with higher-bit counterparts, particularly when combined with language-targeted calibration, while only take less than 3% increase in effective-bits.
- We investigate four parameters that modulate the sensitivity of language-specific calibration: bit-levels, downstream task difficulty, mixed-precision variants, and target languages. Our findings remains explorative, due to the weak correlations.
- We examine the benefits of language-specific calibration transfer to sibling languages within the same family. Generally, the effect is amplified for lower-resource languages.

To the best of our knowledge, only a few studies address this specific problem, summarized in Table 1. Most notably, concurrent works by Kurz et al. (2025) and Chimoto et al. (2026) reach similar conclusion although in a different compression setting (§2).

2 Related Works

Quantization. Quantization is an optimization problem that seeks to preserve performance degradation despite the reduced representational capacity, motivating the development of several quantization algorithms (Gholami et al., 2021). These methods are broadly categorized into three types: Data-Free Quantization (e.g., BitsandBytes (Dettmers et al., 2022)), Post-Training Quantization (PTQ) (e.g., `GPTQ` (Frantar et al., 2023a) and AWQ (Lin et al., 2026)), and Quantization-Aware Training (QAT) (e.g., ternary-bit methods (Ma et al., 2024) and `EfficientQAT` (Chen et al., 2025)). Most PTQ and QAT methods are guided by calibration data, with QAT further requiring substantial retraining to recover performance loss. Additionally, some PTQ have succeeded in achieving extremely low-bit quantization, such as `BRECQ` (Li et al., 2021) and `QuIP` (Chee et al., 2024; Tseng et al., 2024). To further mitigate performance degradation, mixed-precision methods assign higher bit-widths to important weights (Rakka et al., 2025).

Effect of calibration set selection on PTQ. While most quantization research focuses on algorithmic improvements, a growing body of literature begins to analyze the impact of the calibration dataset on the quantized model. Early studies suggest that the effect of calibration set is negligible in general use cases (Paglieri et al., 2024; Williams and Aletras, 2024). We note that these works focus on 4-bit quantization, which is widely regarded as the bit that works in retaining performance (Dettmers and Zettlemoyer, 2023; Jin et al., 2024). However, it remains unclear whether calibration set can actively mitigate the performance drop of lower-bit models.

Recent studies have incorporated extreme-bits in their evaluations, revealing a much more pronounce effect of task-targeted calibration set selection. Liu et al. (2024) demonstrate substantial fluctuations when calibrating on the same task and subsequent task subject, while not consistently produce optimal result. Conversely, Liu et al. (2025a) demonstrate substantial improvements on heavy reasoning task, particularly when using `GPTQ`.

Calibration Sets and Multilinguality. While the impact of compression on multilingual LLMs has been the subject of ongoing study (Marchisio et al., 2024), the specific role of the calibration set has only recently been explored. For example, Zeng

Paper	Description	Methodology
Zeng et al. (2025)	During training, each language converges to its own local minimum, while higher-resource languages converge closer to English. Therefore, it should be similar in compression setting.	Pruning and Quantization (GPTQ, not specified; presumably 4-bit); construct a calibration set that reflects the language distribution of the model's training data
Chimoto et al. (2026) (Most Similar)	Study how monolingual, uniform, or randomly sampled multilingual calibration sets impact quantization	Focus on 4-bit quantization; compare `GPTQ`, `AWQ`, and `Any4`; also provide aggregated multilinguality and per-language evaluation
Choenni and Titov (2025)	`M-Wanda`: Improving language-agnostic `Wanda` via language-activation statistics for better-informed pruning decisions	Multi-sparsity pruning; multilingual calibration set
Kurz et al. (2025) (Most Similar)	Addresses the same problem as ours	Single-sparsity pruning (`Wanda` and `SparseGPT`); monolingual calibration set; internal analysis
Our paper	We systematically contrast direct- and cross-language performance on perplexity and broader task benchmarks across progressive representational power	Low-bit quantization; monolingual calibration set; `GPTQ` variants of mixed-precision; exploration on cross-lingual and factor analysis

Table 1: Comparison with closely related studies on calibration sets and multilinguality.

et al. (2025), Choenni and Titov (2025), and Chimoto et al. (2026) propose multilingual calibration sets to better preserve general multilingual capabilities.

Subsequent studies focus on language-targeted compression using calibration datasets. Kurz et al. (2025) investigate language-specific calibration through pruning, a compression technique that removes redundant model parameters. Similar to our PTQ setting, Chimoto et al. (2026) investigate how varying the language composition of calibration data affects outlier activation statistics, focusing on 4-bit precision and vanilla PTQ methods. Fundamentally, these studies arrive at a shared conclusion: the language distribution within the calibration set improves multilingual perplexity, while downstream performance gains are not guaranteed. Our work distinguishes itself through its analytical focus, broader quantization configurations, and preliminary cross-lingual analysis.

3 Experimental Setup

3.1 Quantization Techniques

Aside from a data-free calibration, SINQ (Müller et al., 2026), as a baseline, we primarily focus on PTQ because we want to observe the impact of calibration-set language. Specifically, we use `GPTQ` as a representative PTQ method, along with two mixed-precision variants that differ in how aggressively they protect important weights, enabling smooth transitions across bit-widths. The "light" variant (`SliM-LLM`, (Huang et al., 2025)) assigns important weights a higher bit-width than the rest, but still below 16 bits. The "heavy" variant (`TaCQ`, (Xiao et al., 2025)) goes further and keeps important weights in full FP16.

Under the hood, `SliM-LLM` assigns N+1-bit to the k-most important weight groups while assigns N-1-bit to the k-least important ones, with the number k is exhaustively searched to minimize quantization error. In contrast, `TaCQ` is inspired by activation patching in mechanistic interpretability. It treats vanilla `GPTQ` as a corrupted model and measures weight importance based on its deviation from the clean, unquantized model. For interested readers, we defer detailed discussions of these methods to Appendix A.

Although mixed-precision necessarily increases total effective bit-width relative to the target bit-width, the increase remains small. In our case, the effective-bit count increases by less than 3% (Appendix C). Ideally, mixed-precision models should improve memory efficiency and inference speed. However, our current choices, `SliM-LLM` and `TaCQ`, do not provide these benefits due to packing limitations (Appendix C).

3.2 Model, Calibration Dataset, and Languages

Model. We use open-source multilingual LLMs that have 8B parameter variants, namely `Qwen3 8B` (Yang et al., 2025), `Llama 3.1 Instruct 8B` (Grattafiori et al.), and `Aya Expanse 8B`[1] (Dang

[1]Although `Aya` primarily supports English, Indonesian, and Chinese (while `Llama` officially supports only English from the selections), we found that both downstream task per-

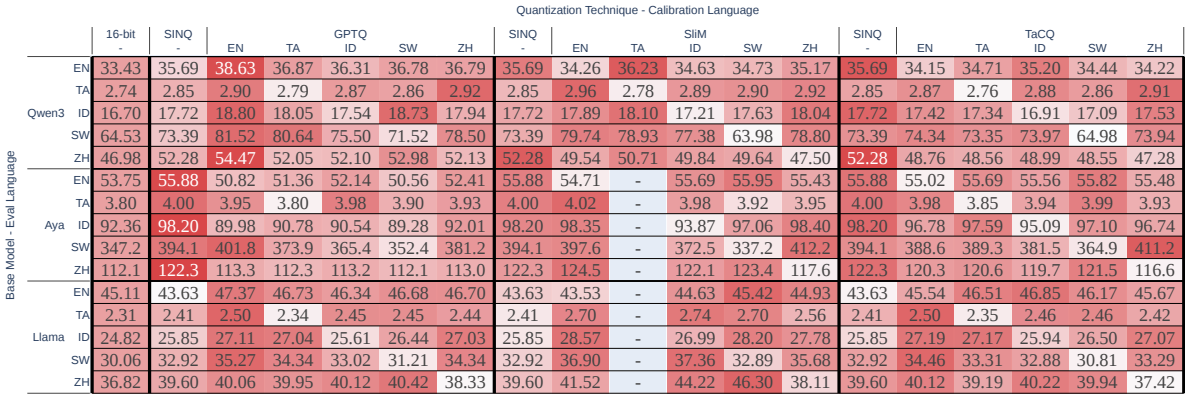


Figure 2: Multilingual Perplexity on 4-bit models. The local-diagonal cells represent the relative effect of Target-Calibration over Cross-Calibrations, i.e., a single row of each quantization method (lighter/lower is better). The relative advantage is clear even in the 4-bit settings, while we defer the rest lower-bits to Appendix D.1. `SliM LLM Llama` Tamil produce outlier, thus biasing the perplexity result.

et al., 2024). We then quantize each model into 3 bit-levels: 4-bit, 3-bit, and 2-bit. Although a 2-bit model is currently impractical, we seek to explore the relationships between variables to better understand how calibration impacts PTQ. Additionally, we evaluate `Qwen3-14B` under a subset of the calibration settings and report the results in Appendix E.

Calibration Dataset and Languages. We use a machine translation dataset, Floresplus `dev` set (NLLB Team et al., 2024), as the source for our quantization calibration data. We do not enforce exact semantic alignment across calibration sets (except through random seed), as each language has varying token length for the same sentence, while calibration set has a sequence length requirement. We select English (EN), Tamil (TA), Indonesian (ID), Swahili (SW), and Chinese (ZH), which span diverse language families, with Tamil and Chinese using non-Latin scripts.

Evaluation Tools. We use `lm-eval-harness` (Gao et al., 2024) for downstream tasks, with `vLLM` (Kwon et al., 2023) as the inference backend for higher throughput. Token perplexity is computed using the Hugging Face evaluate library (Von Werra et al., 2022). We report the geometric mean across datasets, since perplexity values can become extremely large in the 2-bit setting, and the arithmetic mean would be dominated by these outliers.

Hyperparameter. Our hyperparameter, calibration set construction setup, and quantization failure can be seen in the Appendix C.

3.3 Multilingual Tasks

Perplexity. We use Floresplus `devtest` set (NLLB Team et al., 2024) to evaluate the multilingual perplexity of the quantized models. We use this data since it provides a fair comparative multilingual performance due to its parallel structure.

Multilingual Downstream Tasks. We evaluate on various task difficulties to explore the language-specific calibration effectiveness across configurations. We evaluate on minimal grammatical pair test with MultiBlimp and its language variants (Jumelet et al., 2025; Liu et al., 2025b; Leong et al., 2023; Kryvosheieva and Levy, 2025), scenario classification with MASSIVE (FitzGerald et al., 2022; Bastianelli et al., 2020), story completion with XStoryCloze (Lin et al., 2021). Furthermore, we evaluate on knowledge-explicit reading comprehension with Belebele (Bandarkar et al., 2024), and both knowledge-implicit question answering with Global MMLU Lite (Singh et al., 2025) and INCLUDE[2] (Romanou et al., 2024). All evaluations are conducted in a multiple-choice format without using thinking mode for any of models used.

For reasoning tasks that require thinking tokens, we include the multilingual grade-school math benchmark, MGSM (Shi et al., 2022; Peter et al.,

formance remains reasonably strong across the evaluated languages.

[2]We prioritize evaluation on the 5 languages, while some task may not covering these languages. Specifically for INCLUDE, we use aggregated german and dutch as the english evaluation, as they are from the Germanic language family.

Model	Quant	Bit	Perplexity Reduction TA	ID	SW	ZH
Qwen3 8B	GPTQ	4	3.9%	6.7%	12.3%	4.3%
		3	32.6%	21.1%	50.1%	10.8%
		2	99.98%	96.8%	96.6%	43.0%
	SliM	4	6.1%	3.8%	19.8%	4.1%
		3	42.2%	28.8%	55.7%	13.2%
		2	99.94%	91.4%	89.4%	99.97%
	TaCQ	4	3.9%	2.9%	12.6%	3.0%
		3	27.6%	16.9%	43.6%	9.8%
		2	78.9%	81.8%	82.4%	56.6%
Aya Expanse 8B	GPTQ	4	3.8%	−0.6%	12.3%	0.2%
		3	28.2%	17.8%	44.9%	27.1%
		2	99.98%	92.1%	92.8%	95.2%
	SliM	4	—	4.6%	15.2%	5.5%
		3	—	17.01%	46.4%	16.98%
		2	—	99.7%	94.7%	99.996%
	TaCQ	4	3.3%	1.7%	6.1%	3.1%
		3	14.9%	7.0%	28.1%	10.0%
		2	73.7%	62.8%	77.0%	61.3%

Table 2: Perplexity reduction when adopting C_{target} across quantization settings (higher reduction is less robust towards calibration). Calculated by $(C_{\text{en}} - C_{\text{target}})/C_{\text{en}}$. The green and red colored denote the least and most sensitive language towards C_{target}.

2025). We evaluate MGSM using a generate-until format with a maximum of 1024 thinking tokens and enable thinking mode for `Qwen` models. Due to the complex observed pattern and evaluation setting, we treat MGSM evaluation as exploratory and defer it to Appendix D.4.

3.4 Paper's Shorthand

For brevity, we introduce the following shorthand notation. C_{X} (X Calibration Language): An X-only calibration set. C_{target}, C_{cross}, C_{default}, or C_{free}: On evaluating a target language (X), using the same (C_{X}), different ($C_{\neg X}$), default (C_{en}), or without calibration set. Consequently, $Target$, $Cross$, $Default$, and $SINQ$: the impact of corresponding calibration set. `Unquantized`: The original 16-bit model.

4 Results and Analysis

4.1 The effect of Target-Calibration on Perplexity

Effects of C_{target} are Reflected in Multilingual Perplexity. Consistent with prior work by Kurz et al. (2025) and Chimoto et al. (2026), our results for `4-bit` models show that the C_{target} perplexity over unquantized baseline is small, typically within a 10% difference. However, C_{target} is generally better than C_{cross} across configurations. This pattern is clear on the diagonal entries of Figure 2. However, some exception do occurs, such as `GPTQ` on English. For extreme-bit, this effect intensifies, as elaborated in the next point; we defer the heatmaps to Appendix D.1. For the data-free quantization baseline, `SINQ`, C_{target} clearly outperforms `SINQ` in most configurations, except for the English evaluation. Meanwhile, `SINQ` still outperforms C_{cross}, particularly at 2-bit quantization, where the perplexity gap intensifies.

Effect of C_{target} becomes significant at extreme-bit widths. We compare the perplexity reductions when switching from C_{default} to C_{target} in Table 2. In general, the improvement made by C_{target} increases as bit-width decreases, amplify by the decreasing performance of C_{default}, although the magnitude varies across languages. Notably, Indonesian is the least sensitive to calibration language, while Swahili is the most sensitive, despite both languages using Latin script. Furthermore, sensitivity increases sharply when moving from 3-bit to 2-bit (e.g., `Qwen3 GPTQ` on Indonesian rises from 21.1% to 96.8%), which is far more severe than from 4-bit to 3-bit (from 6.7% to 21.1%). This contrast is more pronounced for non-Latin scripts (e.g., `SliM-LLM` on Tamil and Chinese). These results suggest a strong degradation in target-language linguistic capability under C_{default} at 2-bit quantization, particularly for non-Latin scripts. If this linguistic capability collapse can be mitigated, 2-bit performance may become comparable to 3-bit models.

Calibration Language and Mixed-Precision determine 2-bit model generation. To highlight this effect, we report generated Tamil sentences, which shows the most contrast between `GPTQ` and `TaCQ` (see Table 3). The outputs vary depending on whether Target-Calibration, Mixed-Precision, or both are used. In particular, C_{target} steers generation towards the target language, while mixed-precision improves overall coherence. Interestingly, the `TaCQ` 2-bit C_{en} model generates reasoning in the default language (i.e., English) but falls into repetitive looping behavior. This may indicate that C_{default} preserves reasoning ability but degrades broader capabilities, such as long-context comprehension and duplicate word penalty. Consequently, using both Mixed-Precision and C_{target} allows the 2-bit model to function as a targeted monolingual model, albeit generating overly simplistic sentences.

Prompt: இலையுதிர் காலத்தின் நடுவில் குடும் பத்துடன் சுற்றுலா செல்வது பற்றி ஒரு கவிதை எழுதுங்கள் Write me a poem about the picnic with family in the middle of the fall season
GPTQ 2-bit **English**-calibrated (48256.105) 矜toString Mid representations furthermore rematch convo阶 '\''riel整个_foot whereneau响应矩 ...
GPTQ 2-bit **Tamil**-calibrated (8.060) உயம், எனக் ஆக்ராஞ்சாடி வர்டின் ஒருங்கரூடி காரியை பயன்பாடு காக்காசேன் என்று ஒரு பெ ரிய ச Uyam, I am a big fan of Agrachandi Ward's Orungarudi Karikai Kakkasen.
TaCQ 2-bit **English**-calibrated (14.578) Okay, let's break down this sentence. First, the user is asking about the process of writing a sentence in a way that it's not a sentence, but ...
TaCQ 2-bit **Tamil**-calibrated (3.077) **கவித்துருவை நிறுத்துவது:** இலையுதிர் காலத்தின் நடுவில், குடும்பத்துடன் சுற்றுலா செல்லும் சிறு வி **Stopping the Poet:** In the middle of autumn, A little girl goes on a family trip.

Table 3: Contrasting the generation of Tamil when the model is calibrated using C_{en} and C_{ta} on `Qwen3 8B`. Number inside parenthesis is the model's Tamil perplexity. Higher-bit models adhere to the prompt.

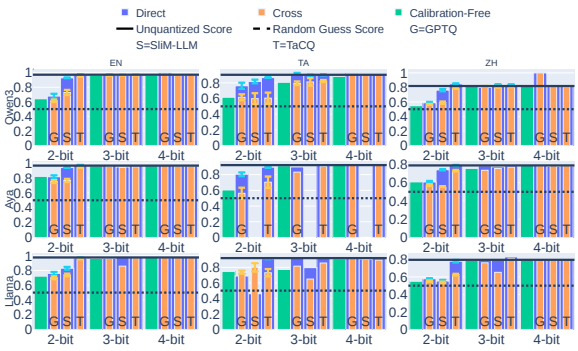


Figure 3: MultiBlimp: Contrasting accuracy (Higher is better) of *Direct* vs mean of 2 *Cross*. Notably, 2-bit can be comparable to 3-bit when use both mixed-precision and target-calibration. We omit the Swahili (SW) and Indonesian (ID) as their unquantized accuracy are near random guess (0.5).

4.2 The effect of Target-Calibration on downstream task performance

We observe that the effectiveness of C_{target} varies across task difficulties; it is critical for some while negligible for most. This suggests a fluctuating calibration set sensitivity across quantized configurations. We present results comparing C_{target}, C_{cross}, C_{free} as bar-charts.

2-bit models require mixed-precision and target-calibration on `MultiBlimp`. Figure 3 shows the MultiBlimp evaluation results (see Appendix D.2 for Massive and XStoryCloze), revealing four relationships between C_{target}, language resource level, and mixed-precision. Notice that `3-bit` models achieve near-lossless performance, while 2-bit models exhibit high variance, suggesting that MultiBlimp is a suitable benchmark for 2-bit quantization evaluation. First, 2-bit models can match 3-bit performance when using techniques such as mixed-precision and target-calibration; mixed-precision is primary for recovery, while C_{target} is secondary for performance gains (e.g., `Qwen3` on ZH). It is important to note that this performance gain comes from optimizing selected critical weights while keeping the total effective-bits unchanged for `SliM-LLM` or increasing it by less than 5% for `TaCQ`. Second, the behaviors of mixed-precisions and C_{target} differ between language resource level: low-resource language requires C_{target} for stability, hence mixed-precision is not enough on TA; while C_{target} remained secondary in high-resource languages (e.g., contrast `TaCQ` on TA and ZH). Third, the effect of light and heavy mixed-precisions are noticeably different. Fourth, `SINQ`'s performance is comparable to that of C_{cross} across settings in MultiBlimp. These four findings serve as foundational patterns that persist in subsequent, more difficult tasks. Finally, the delta performance between C_{target} and C_{cross} ranges from -0.05 to 0.27.

3-bit models sometimes show strong calibration sensitivity on `Belebele`, especially in low-resource languages. `Belebele` is a suitable benchmark for evaluating 3-bit models, due to the less informative result on 4-bit and 2-bit, as shown in Figure 4 (see Appendix D.3 for INCLUDE and Global MMLU Lite). The four patterns identified in the previous tasks are also observed here, although reduced in intensity. First, mixed-precision remains primary to be beneficial, as observed in the `TaCQ` results across 3-bit configurations. Furthermore, C_{target} improves perfor-



Figure 4: Belebele: Contrasting accuracy (Higher is better) of C_{target} vs mean of 4 C_{cross}. Overall, the relative advantage remains, while absolute improvements mostly occur on 3-bit and low-resource language. Notably, English is robust towards calibration language.

mance retention for low-resource languages (e.g., `Qwen3` on Tamil) and acts as a safeguard against unexpected quantization failure (e.g., `Llama` with `SliM-LLM` across all languages). Next, "light" and "heavy" mixed-precisions remain noticeably different. Lastly, `SINQ`'s performance is slightly lower than C_{Cross} performance and substantially lower than `TaCQ` (in contrast to INCLUDE and Global MMLU Lite, where its performance is comparable to C_{Cross}). Interestingly, 2-bit models can achieve decent performance when paired with `TaCQ` and Target-Calibration, although the outcome depends on the base model. This pattern suggests that stronger base model capabilities may enhance resilience to extreme quantization. However, these patterns becomes less pronounced in `Qwen3-14B` models, likely due to its increased representational capacity. Consequently, Belebele evaluation for `Qwen3-14B` becomes more simlar to MultiBlimp for `8B` models, as discussed in Appendix E. Overall, the delta performance at `3-bit` models ranges from -0.02 to 0.32.

4.3 The effect of Target-Calibration on within-family perplexity

We extend our study to cross-lingual intra-family scenarios, serving as a preliminary test of whether C_{target} preserves performance across similar, related languages. We analyze the effect on perplexity using all languages available in Floresplus `test` set (NLLB Team et al., 2024), leaving exhaustive downstream tasks evaluations for future work (see preliminary in Appendix G.1).

We limit our in-depth analysis to the `Austronesian` (Indonesian) language family. We investigate the correlation between the degree of transfer and its relationship to the target language based on sibling relationship (i.e. glottolog[3] (Hammarström et al., 2025).) and geographical proximity.

Austronesian (Indonesian) cross-lingual mainly applies to languages spoken in Indonesia, Malaysia, and Philippines. The Austronesian language family (Indonesian) includes seven phylogenetic subgroups, ranging from the highest level Austronesian node to the terminal node, Standard Indonesian. Almost all available Malay and Indonesian languages share a common ancestral subgroup within the second-highest (`Malayo-Polynesian`) branch, except for four Indonesian languages: Balinese (BAN), Buginese (BUG), Javanese (JAV), and Sundanese (SUN). Following our naming convention, the 13 languages meet at the `Malayo-Polynesian` level as distant-relative ($ID_{\text{gene,distant}}$), and the 5 more phylogenetically closer languages as close-relative ($ID_{gene,close}$).

From the 2-bit section in Figure 5, the cross-lingual transfer can be categorized into four levels: High, Moderate, Low, and Negligible. Specifically, we classify it from the transfer ratio percentage: 0-20% (C_{cross} is $5\times$ higher than C_{target}), 20-50%, 50-80%, and 80+%. These results suggest that transfer levels cannot be predicted solely by sibling or geographical proximity. For instance, while Indonesian and Malay exhibit a high transfer effect, not all language spoken in Indonesia or all Inner-Languages automatically exhibit a moder-

[3] https://glottolog.org/



Figure 5: Austronesian (Indonesian) cross-lingual transfer benefits aggregated across quantization methods. Values represent the perplexity ratio percentage between C_{id} and mean of 4 C_{cross}. Language are ordered (left to right) according to their phylogenetic proximity in Glottolog. Among the evaluated languages, only 5 languages from Indonesian (ID) to Acehnese (ACE) that are closer in siblingness, while the remaining languages meet at the highest node.

Variable	Order	Comment
Model's release data	Newer	Bias towards the newest
Quantization Technique	Heavier	See subsection 3.1
Bit Width	Higher	Omit 4-bit
Language Resource	Higher	See Table 5 in Appendix
Task Difficulty	Harder	Based on similar results
Delta Acc	Higher	$Target - Cross$

Table 4: Ordinal Setting for spearman correlation.

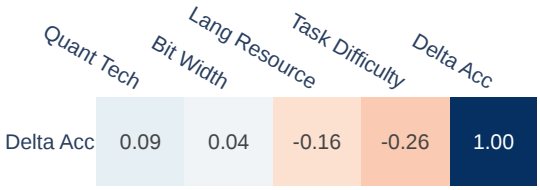


Figure 6: Which variable influence calibration sensitivity?

ate or high transfer effect. Specifically, Acehnese (ACE) ($ID_{\text{gene,close}}$) exhibits a low transfer effect, together with Buginese (BUG) ($ID_{\text{gene,distant}}$). Conversely, Javanese (JAV) and Sundanese (SUN) ($ID_{\text{gene,distant}}$) exhibit a moderate transfer effect. Geograhically, the former two originated outside Java Island, where the Indonesia's capital is located, while the latter two originated on the island itself. Interestingly, four languages spoken in Philippines are categorized as having a low transfer effect, while the remaining languages (Oceanic languages and Malagasy) exhibit a negligible transfer effect.

Other Language Family analysis.. We defer the rest intra-family transfer in Appendix G.1. In summary, we observe that high-resource languages tend to restrict benefits to themselves, yielding negligible or even negative effects for their sibling languages. Conversely, mid and low-resource languages, especially tamil that use different scripts, exhibit much stronger cross-lingual benefits. We note the limitation that these cross-lingual effects are observed only through perplexity and negligible except at 2-bit level.

4.4 Exploratory: Calibration Sensitivity Analysis

To uncover factors that influence model calibration sensitivity, we calculate the Spearman correlation between variables[4][5], as detailed in Table 4. We omit the correlations between variable settings because they only represent experimental co-occurrence. Importantly, these correlations do not directly indicate improvements in "quantization accuracy", but instead capture general calibration sensitivity, defined as $C_{\text{target}} - C_{\text{cross}}$. For example, a negative correlation with the quantization technique does not mean that using `TaCQ` reduces accuracy. Rather, it indicates that `TaCQ` is more robust than `GPTQ`, consistent with the downstream task results.

General Correlation. Figure 6 shows the overall Spearman correlations. We observe that task difficulty and language resource level have the strongest correlations with delta accuracy, albeit

[4]4-bit bias towards negative correlation.

[5]We used aggregated accuracy by bit-width as a proxy to rank the Task Difficulty, as shown in Figure 1. Therefore, from the easiest, MultiBlimp as 1; Massive and XstoryCloze as 2; Belebele as 3; INCLUDE and Global MMLU Lite as 4.

weak negative correlations. In contrast, bit width and quantization technique show relatively weak correlations with the delta, despite their apparent influence in the downstream task results. Therefor, we further analyze each variable in isolation to understand its individual relationships with calibration sensitivity.

Grouping by Bit. Bit Width provides the simplest perspective for interpreting calibration sensitivity: "If I want to quantize to N-bit, when should I care about language-targeted calibration?" Figure 7 shows the correlations for 2-bit and 3-bit models. At 2-bit, "easier" task difficulty ($\rho = -0.43$) and "heavier" quantization technique ($\rho = 0.30$) show the strongest correlations with calibration sensitivity, while language resource ($\rho = 0.10$) shows a weaker correlation. For 3-bit models, however, "lower"-resource language ($\rho = -0.41$) shows the strongest correlation with calibration sensitivity. These results indicate that calibration sensitivity is more strongly associated with "easier" task difficulty and "heavier" quantization technique at 2-bit, whereas language resource becomes more relevant at 3-bit. Lastly, the quantization technique correlation's magnitude decreases from $\rho = 0.30$ at 2-bit to $\rho = -0.14$ at 3-bit, indicating that the choice of mixed-precision technique becomes more robust at 3-bit. These patterns match with downstream task evaluations.

Grouping by others variables. We defer further analysis of interactions among the other variables to Appendix F. Overall, no single variable consistently influences calibration sensitivity across all configurations. Notably, for English, calibration sensitivity shows a weak negative correlation with task difficulty and bit width, which corresponds to the setting considered in most existing calibration studies.

5 Conclusion

Our work demonstrates that models quantized via `PTQ` *are affected* by the choice of language in the calibration set, leading to several key insights. First, we recommend using Target-Calibration for monolingual LLM applications. However, its improvements over Default- or Cross-Calibration are often marginal, while notable exceptions exist. Second, we find that a quantized model's sensitivity to the calibration set is primarily limited by the model's representational capacity. One implication is that low-resource languages have a greater

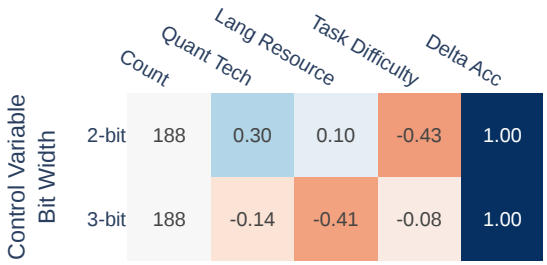


Figure 7: Calibration sensitivity across Bit Widths. Each row represents the correlation between delta accuracy and another variable.

potential to benefit from monolingual calibration, as supported by most downstream task evaluations. Another implication is that mixed-precision methods expand this representational capacity, occasionally allowing 2-bit models to perform comparable to their 3-bit counterparts without substantially increasing the effective bit count, whereas Target-Calibration maximizes performance within that capacity. Third, the effects of Target-Calibration do transfer across languages within the same language family (intra-family), although the degree depends on language resource level. Fourth, the correlation results support the downstream task findings, while no single parameter strongly influences calibration sensitivity across settings. Finally, our findings suggest that the conflicting calibration impacts reported in previous studies may be explained by the tendency of English calibration sensitivity to decrease as bit level and task difficulty increase.

Suggestion

We observe that assessing the true target-language capability of quantized LLMs remains challenging. Therefore, we encourage the NLP community to develop dedicated language capability benchmarks, analogous to CEFR language proficiency levels, that can quantitatively measure the linguistic competence of smaller models, including quantized LLMs. Such benchmarks may be more feasible to construct given that existing online language learning materials have been categorized according to CEFR proficiency levels.

Limitation

We limit our study to three `8B` models and one `14B`, thus the observed patterns may not generalize well to larger models. Second, although we use five lan-

guages spanning varying resource levels and language families, our selection is not exhaustive. In particular, non-Latin script languages warrant further investigation. Third, the distinction between simple tasks and more complex reasoning tasks is based on a quantitative accuracies aggregated by bit-width, although a more rigorous quantitative characterization of each task difficulty may be possible. Fourth, although recent studies suggest that GPTQ family exhibits higher sensitivity to calibration set, we have yet to evaluate whether this observation generalizes across a broader family of quantization methods. Future work should examine the complex relationships holds under more recent PTQ and QAT methods that have achieved strong performance on extreme-bit settings. In particular, evaluate on activation quantization techniques is interesting directions, becaue per-token activation outliers can introduce a different calibration sensitivity. Fifth, while Floresplus is a translation benchmark, we have not conducted an ablation study on the effect of semantic parallelism. Lastly, the cross-lingual analysis is based on geographic and sibling similarities, while more direct measures of linguistic similarity have yet to be explored, such as URIEL+ (Khan et al., 2025).

Ethical Considerations

LLM Usage: We use Gemini and ChatGPT to proofreading and correcting: grammar, word choice, and awkward phrasing.

Potential Risk: As we found that language-targeted calibration on 2-bit TaCQ can generate coherent response in target language, albeit simple, we afraid that it can cause unintended multilingual future risk, such as multilingual jailbreak (Deng et al., 2024). With assumption that the model can provide knowledgeable answer in extreme-bit setting on top of target language understanding.

Acknowledgement

This research is supported by University of Indonesia (UI) and Mohamed bin Zayed University of Artificial Intelligence (MBZUAI) that provide us with compute resources such as DGXA100.

References

Lucas Bandarkar, Davis Liang, Benjamin Muller, Mikel Artetxe, Satya Narayan Shukla, Donald Husa, Naman Goyal, Abhinandan Krishnan, Luke Zettlemoyer, and Madian Khabsa. 2024. The belebele benchmark: a parallel reading comprehension dataset in 122 language variants. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 749–775, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

Emanuele Bastianelli, Andrea Vanzo, Pawel Swietojanski, and Verena Rieser. 2020. SLURP: A spoken language understanding resource package. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7252–7262, Online. Association for Computational Linguistics.

Manuel Brack, Malte Ostendorff, Pedro Ortiz Suarez, José Javier Saiz, Iñaki Lacunza Castilla, Jorge Palomar-Giner, Alexander Shvets, Patrick Schramowski, Georg Rehm, Marta Villegas, and Kristian Kersting. 2024. Community OSCAR: A community effort for multilingual web data. In *Proceedings of the Fourth Workshop on Multilingual Representation Learning (MRL 2024)*, pages 232–235, Miami, Florida, USA. Association for Computational Linguistics.

Jerry Chee, Yaohui Cai, Volodymyr Kuleshov, and Christopher De Sa. 2024. Quip: 2-bit quantization of large language models with guarantees. *Preprint*, arXiv:2307.13304.

Mengzhao Chen, Wenqi Shao, Peng Xu, Jiahao Wang, Peng Gao, Kaipeng Zhang, and Ping Luo. 2025. Efficientqat: Efficient quantization-aware training for large language models. *Preprint*, arXiv:2407.11062.

Everlyn Asiko Chimoto, Mostafa Elhoushi, and Bruce A. Bassett. 2026. Calibrating beyond english: Language diversity for better quantized multilingual llm. *Preprint*, arXiv:2601.18306.

Rochelle Choenni and Ivan Titov. 2025. M-wanda: Improving one-shot pruning for multilingual llms. *Preprint*, arXiv:2505.21171.

John Dang, Shivalika Singh, Daniel D'souza, Arash Ahmadian, Alejandro Salamanca, Madeline Smith, Aidan Peppin, Sungjin Hong, Manoj Govindassamy, Terrence Zhao, Sandra Kublik, Meor Amer, Viraat Aryabumi, Jon Ander Campos, Yi-Chern Tan, Tom Kocmi, Florian Strub, Nathan Grinsztajn, Yannis Flet-Berliac, and 26 others. 2024. Aya expanse: Combining research breakthroughs for a new multilingual frontier. *Preprint*, arXiv:2412.04261.

Yue Deng, Wenxuan Zhang, Sinno Jialin Pan, and Lidong Bing. 2024. Multilingual jailbreak challenges in large language models. *Preprint*, arXiv:2310.06474.

Tim Dettmers, Mike Lewis, Younes Belkada, and Luke Zettlemoyer. 2022. Llm.int8(): 8-bit matrix multiplication for transformers at scale. *Preprint*, arXiv:2208.07339.

Tim Dettmers and Luke Zettlemoyer. 2023. The case for 4-bit precision: k-bit inference scaling laws. *Preprint*, arXiv:2212.09720.

Jack FitzGerald, Christopher Hench, Charith Peris, Scott Mackie, Kay Rottmann, Ana Sanchez, Aaron Nash, Liam Urbach, Vishesh Kakarala, Richa Singh, Swetha Ranganath, Laurie Crist, Misha Britan, Wouter Leeuwis, Gokhan Tur, and Prem Natarajan. 2022. Massive: A 1m-example multilingual natural language understanding dataset with 51 typologically-diverse languages. *Preprint*, arXiv:2204.08582.

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

Elias Frantar, Sidak Pal Singh, and Dan Alistarh. 2023b. Optimal brain compression: A framework for accurate post-training quantization and pruning. *Preprint*, arXiv:2208.11580.

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, and 5 others. 2024. The language model evaluation harness.

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. *Preprint*, arXiv: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. The llama 3 herd of models.

Harald Hammarström, Robert Forkel, Martin Haspelmath, and Sebastian Bank. 2025. glottolog/glottolog: Glottolog database 5.2.1.

Wei Huang, Haotong Qin, Yangdong Liu, Yawei Li, Qinshuo Liu, Xianglong Liu, Luca Benini, Michele Magno, Shiming Zhang, and Xiaojuan Qi. 2025. Slim-llm: Salience-driven mixed-precision quantization for large language models. *Preprint*, arXiv:2405.14917.

Renren Jin, Jiangcun Du, Wuwei Huang, Wei Liu, Jian Luan, Bin Wang, and Deyi Xiong. 2024. A comprehensive evaluation of quantization strategies for large language models. *Preprint*, arXiv:2402.16775.

Jaap Jumelet, Leonie Weissweiler, Joakim Nivre, and Arianna Bisazza. 2025. Multiblimp 1.0: A massively multilingual benchmark of linguistic minimal pairs. *Preprint*, arXiv:2504.02768.

Dhruva Kashyap, Chaitanya Murti, Pranav K Nayak, Tanay Narshana, and Chiranjib Bhattacharyya. 2026. Modhifi: Identifying high fidelity predictive components for model modification. *Preprint*, arXiv:2511.19566.

Aditya Khan, Mason Shipton, David Anugraha, Kaiyao Duan, Phuong H. Hoang, Eric Khiu, A. Seza Doğruöz, and En-Shiun Annie Lee. 2025. URIEL+: Enhancing linguistic inclusion and usability in a typological and multilingual knowledge base. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 6937–6952, Abu Dhabi, UAE. Association for Computational Linguistics.

Daria Kryvosheieva and Roger Levy. 2025. Controlled evaluation of syntactic knowledge in multilingual language models. In *Proceedings of the First Workshop on Language Models for Low-Resource Languages*, pages 402–413, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Simon Kurz, Jian-Jia Chen, Lucie Flek, and Zhixue Zhao. 2025. On the limitations of language targeted pruning: Investigating the calibration language impact in multilingual llm pruning. *Preprint*, arXiv:2408.14398.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.

Wei Qi Leong, Jian Gang Ngui, Yosephine Susanto, Hamsawardhini Rengarajan, Kengatharaiyer Sarveswaran, and William Chandra Tjhi. 2023. Bhasa: A holistic southeast asian linguistic and cultural evaluation suite for large language models. *Preprint*, arXiv:2309.06085.

Yuhang Li, Ruihao Gong, Xu Tan, Yang Yang, Peng Hu, Qi Zhang, Fengwei Yu, Wei Wang, and Shi Gu. 2021. Brecq: Pushing the limit of post-training quantization by block reconstruction. *Preprint*, arXiv:2102.05426.

Ji Lin, Jiaming Tang, Haotian Tang, Shang Yang, Wei-Ming Chen, Wei-Chen Wang, Guangxuan Xiao, Xingyu Dang, Chuang Gan, and Song Han. 2026. Awq: Activation-aware weight quantization for llm compression and acceleration. *Preprint*, arXiv:2306.00978.

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 language models. *Preprint*, arXiv:2112.10668.

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 T. Diab, and 2 others. 2021. Few-shot learning with multilingual language models. *CoRR*, abs/2112.10668.

Ruikang Liu, Yuxuan Sun, Manyi Zhang, Haoli Bai, Xianzhi Yu, Tiezheng Yu, Chun Yuan, and Lu Hou. 2025a. Quantization hurts reasoning? an empirical study on quantized reasoning models. *Preprint*, arXiv:2504.04823.

Yijun Liu, Yuan Meng, Fang Wu, Shenhao Peng, Hang Yao, Chaoyu Guan, Chen Tang, Xinzhu Ma, Zhi Wang, and Wenwu Zhu. 2024. Evaluating the generalization ability of quantized llms: Benchmark, analysis, and toolbox. *Preprint*, arXiv:2406.12928.

Yikang Liu, Yeting Shen, Hongao Zhu, Lilong Xu, Zhiheng Qian, Siyuan Song, Kejia Zhang, Jialong Tang, Pei Zhang, Baosong Yang, Rui Wang, and Hai Hu. 2025b. A systematic assessment of language models with linguistic minimal pairs in chinese. *Preprint*, arXiv:2411.06096.

Shuming Ma, Hongyu Wang, Lingxiao Ma, Lei Wang, Wenhui Wang, Shaohan Huang, Li Dong, Ruiping Wang, Jilong Xue, and Furu Wei. 2024. The era of 1-bit llms: All large language models are in 1.58 bits. *Preprint*, arXiv:2402.17764.

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.

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

Lorenz K. Müller, Philippe Bich, Jiawei Zhuang, Ahmet Çelik, Luca Benfenati, and Lukas Cavigelli. 2026. Sinq: Sinkhorn-normalized quantization for calibration-free low-precision llm weights. *Preprint*, arXiv:2509.22944.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, and 20 others. 2024. Scaling neural machine translation to 200 languages. *Nature*, 630(8018):841–846.

Davide Paglieri, Saurabh Dash, Tim Rocktäschel, and Jack Parker-Holder. 2024. Outliers and calibration sets have diminishing effect on quantization of modern llms. *Preprint*, arXiv:2405.20835.

Jan-Thorsten Peter, David Vilar, Tobias Domhan, Dan Malkin, and Markus Freitag. 2025. Mind the gap... or not? how translation errors and evaluation details skew multilingual results. *Preprint*, arXiv:2511.05162.

Irina Proskurina, Luc Brun, Guillaume Metzler, and Julien Velcin. 2024. When quantization affects confidence of large language models? *Preprint*, arXiv:2405.00632.

Mariam Rakka, Marios Fournarakis, Olga Krestinskaya, Jinane Bazzi, Khaled N. Salama, Fadi Kurdahi, Ahmed M. Eltawil, and Mohammed E. Fouda. 2025. Mixed-precision quantization for language models: Techniques and prospects. *Preprint*, arXiv:2510.16805.

Angelika Romanou, Negar Foroutan, Anna Sotnikova, Zeming Chen, Sree Harsha Nelaturu, Shivalika Singh, Rishabh Maheshwary, Micol Altomare, Mohamed A. Haggag, Snegha A, Alfonso Amayuelas, Azril Hafizi Amirudin, Viraat Aryabumi, Danylo Boiko, Michael Chang, Jenny Chim, Gal Cohen, Aditya Kumar Dalmia, Abraham Diress, and 40 others. 2024. Include: Evaluating multilingual language understanding with regional knowledge. *Preprint*, arXiv:2411.19799.

Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, Dipanjan Das, and Jason Wei. 2022. Language models are multilingual chain-of-thought reasoners. *Preprint*, arXiv:2210.03057.

Shivalika Singh, Angelika Romanou, Clémentine Fourrier, David I. Adelani, Jian Gang Ngui, Daniel Vila-Suero, Peerat Limkonchotiwat, Kelly Marchisio, Wei Qi Leong, Yosephine Susanto, Raymond Ng, Shayne Longpre, Wei-Yin Ko, Sebastian Ruder, Madeline Smith, Antoine Bosselut, Alice Oh, Andre F. T. Martins, Leshem Choshen, and 5 others. 2025. Global mmlu: Understanding and addressing cultural and linguistic biases in multilingual evaluation. *Preprint*, arXiv:2412.03304.

Shivalika Singh, Freddie Vargus, Daniel Dsouza, Börje F. Karlsson, Abinaya Mahendiran, Wei-Yin Ko, Herumb Shandilya, Jay Patel, Deividas Mataciunas, Laura OMahony, Mike Zhang, Ramith Hettiarachchi, Joseph Wilson, Marina Machado, Luisa Souza Moura, Dominik Krzemiński, Hakimeh Fadaei, Irem Ergün, Ifeoma Okoh, and 14 others. 2024. Aya dataset: An open-access collection for multilingual instruction tuning. *Preprint*, arXiv:2402.06619.

Albert Tseng, Jerry Chee, Qingyao Sun, Volodymyr Kuleshov, and Christopher De Sa. 2024. Quip#: Even better llm quantization with hadamard incoherence and lattice codebooks. *Preprint*, arXiv:2402.04396.

Leandro Von Werra, Lewis Tunstall, Abhishek Thakur, Sasha Luccioni, Tristan Thrush, Aleksandra Piktus,

Felix Marty, Nazneen Rajani, Victor Mustar, and Helen Ngo. 2022. Evaluate & evaluation on the hub: Better best practices for data and model measurements. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 128–136, Abu Dhabi, UAE. Association for Computational Linguistics.

Qianli Wang, Mingyang Wang, Nils Feldhus, Simon Ostermann, Yuan Cao, Hinrich Schütze, Sebastian Möller, and Vera Schmitt. 2026. Through a compressed lens: Investigating the impact of quantization on factual knowledge recall. *Preprint*, arXiv:2505.13963.

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)*, page 10100–10118. Association for Computational Linguistics.

Hanqi Xiao, Yi-Lin Sung, Elias Stengel-Eskin, and Mohit Bansal. 2025. Task-circuit quantization: Leveraging knowledge localization and interpretability for compression. *Preprint*, arXiv:2504.07389.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Hongchuan Zeng, Hongshen Xu, Lu Chen, and Kai Yu. 2025. Multilingual brain surgeon: Large language models can be compressed leaving no language behind. *Preprint*, arXiv:2404.04748.

A Background

A.1 Quantization

Model quantization aims to reduce memory and computational requirements while preserving the original performance. It seems to be a simple problem but a challenging one in practice. Doing this naively inevitably alters the weight distribution and therefore propagates the error to the final layer, leading to performance degradation. The existing methods range from simple lookup table, intricate mathematical optimizations, to resource-intensive retraining (Gholami et al., 2021; Rakka et al., 2025).

Despite their differences, quantization fundamentally seeks an optimal function that maps a model's parameters from a high-precision to a low-precision format while ensuring the mapped model maximally emulates the original behavior. Moreover, due to the inherent compromises of this trade-off, theoretically, quantization can be optimized to maintain general capabilities (which may result in a shared global performance loss) or to preserve a *niche* capability (which currently requires deliberate configuration). For example, one may prioritize preserving the model's broad multilingual proficiency or its proficiency in a single targeted-language. While most quantization methods focus on the former, for our problem, the latter is preferable.

A.2 Post-Training Quantization

Utilizing a calibration set allows one to optimize the direction of quantization process; most notably, it enables the simulation of the quantized model's behavior. The primary metric to quantify the error during quantization is the layer-wise reconstruction error, defined in Equation A.2. The fundamental objective is to minimize the layer-wise output distance between the unquantized (original) model and its quantized counterpart, in the hope of achieving an effective yet efficient quantization process. Recent finding has proven that layer-wise optimization is linearly upper-bounded (at maximum linear multiplication) of the end result (Kashyap et al., 2026).

$$\ell(W, \hat{W}, l) = ||X_l W_l - X_l \widehat{W_l}||_2$$

The symbols in the equation are as follows: X_l represents the l-th layer's input batch, while W_l and $\widehat{W_l}$ represent the unquantized and quantized l-th layer's weight tensors, respectively. $||\cdot||_p$ denotes p-norm or, in simple term, a distance metric.

`GPTQ` (Frantar et al., 2023a) and its origin, `OBQ` (Frantar et al., 2023b), navigate layer-wise reconstruction optimization problem using the mathematical properties of Hessian matrix. In this context, the Hessian matrix (second-order derivatives) holds the sensitivity of each parameter to induce error when changed (e.g., during quantization). Specifically, `GPTQ` quantizes each parameter iteratively and optimizes the remaining unquantized weights through the inverse Hessian matrix, which informs *inverse* sensitivity to balance the backlash of point-quantization. **More importantly, each layer's Hessian matrix is calculated solely from the layer's input matrix, which intuitively made the calibration set important**. Although the original work states that `GPTQ` is robust to calibration set sampling, recent discoveries empirically prove otherwise.

A.3 Improvements: Mixed-Precision

Notice that, from the Hessian matrix, we can estimate which weights induce the most sensitive to error than the rest, implying that not every weight is equally important. It is an observed phenomena, such as outlier-inducing weights (Dettmers et al., 2022). Therefore, we can mix the precisions in at least two ways: maintaining the total bit count (such as `SliM-LLM` (Huang et al., 2025)) or quantizing all weights except the most important ones (such as `TaCQ` (Xiao et al., 2025)). For convenience, we call the former as Light Mixed Precision and the latter as Heavy Mixed Precision. Fundamentally, both variants introduce a weight-importance determination method step prior to quantization.

`SliM-LLM`, Salience-Driven Mixed-Precision Quantization (Huang et al., 2025) focuses on two aspects: structured mixed-precision and preserving the total bit count. First, structured mixed-precision ensures that the model can be inferred directly in low-precision through the currently supported efficient hardware operations, which necessitate weights having the same precision for each input dimensions. Second, by splitting $2D$ weights into groups of consecutive input dimensions (e.g., each group consisting of 128 parameters of output dimension), `SliM-LLM` quantizes the top-k most important groups to $N+1$ bits, the least-k important groups to $N-1$

bits, and the rest to the target bit, or N. The salience of each group is determined by averaging the group's Hessian matrix. We observe that `SliM-LLM` is prone to overfitting on calibration set due to the exhaustive nature of determining the k parameter.

`TaCQ`, Task-Circuit Quantization (Xiao et al., 2025) is inspired by Mechanistic Interpretability methods. Specifically, `TaCQ` uses absolute weight, the difference between unquantized (clean) and quantized (corrupted) weights, and calibration data's gradient to determine important weights, as expressed in Equation A.3. Moreover, the authors reported that task-specific calibration set improves related task accuracies by saving task-related weights in higher-precision. We observed that `TaCQ` is more robust than the other two `GPTQ` variants, probably due to its heavy mixed-precision and the use of gradient salience. However, the original `TaCQ` method does not yet support inference directly in low-bit due to its unstructured mixed-precision.

$$Salience = W \odot |W - \widehat{W}| \odot G$$

Here, W is the unquantized weight matrix (clean), $\widehat{W}$ is the vanilla `GPTQ` quantized weight matrix (corrupted), G is the gradient of chosen loss function, and $\odot$ represents the Hadamard (item-wise) multiplication operation.

B Expanded Related Works

B.1 Quantization Degrades Quality Beyond Downstream Metrics

Numerous studies suggest that PTQ can achieve "negligible performance loss" on standard benchmarks (Frantar et al., 2023a; Dettmers and Zettlemoyer, 2023; Jin et al., 2024). However, recent studies indicates that this claim does not hold true across more nuanced qualitative attributes. For instance, Marchisio et al. (2024) observe that while quantized models exhibit negligible performance degradation on automated generative tasks, they receive lower scores from human evaluators, particularly when generating non-Latin scripts. Wang et al. (2026) employ interpretability tools to show that the factual memorization abilities of quantized LLMs degrade modestly. Furthermore, Proskurina et al. (2024) highlight the shift in confidence score, resulting in overconfidence on incorrect answers. Finally, Liu et al. (2025a) find that extreme quantization substantially degrades the reasoning capability of 3-bit models.

C Experiment Setup

Hyperparameters Generally, we adopt the default parameters and scripts from the respective GitHub repositories: `GPTQModel`[6](ModelCloud.ai and qubitium@modelcloud.ai, 2024) (v4.2.0), `SliM-LLM`[7] (Aug 2024), and `TaCQ`[8] (June 2025). We set the random seed to `1234`, the number of calibration samples to `128`, and the sequence length to `2048`. Additionally, for `TaCQ`, we use `16-bit float datatypes` during backpropagation to mitigate GPU memory constraints.

Multi-seed Experiments Initially, we quantize and evaluate each configuration in a single run with a fixed random seed of 1234, cutting both budget and runtime. For selected configurations, specifically the 2-bit models and the 3-bit `Qwen3-8B` model, we conduct two additional random seeds, 1000 and 2000, to estimate variability. This is particularly important for the 2-bit models, which exhibit high volatility across random seeds, except for *Direct* setting. The results from additional seeds have been incorporated into all figures, with perplexity reported as mean ± std and downstream task results as error bars.

Calibration Set Construction We follow the methodologies of `SliM-LLM` and `TaCQ` to construct the calibration set. For each language, we first concatenate all dataset instances, separated by a double newline character (`\n\n`). Then, the concatenated text is tokenized using the corresponding model's tokenizer. Finally, we randomly select chunks of 2048 consecutive tokens as individual samples.

Library Modification: We locally extend the `SliM-LLM` and `TaCQ` libraries with additional model families and utilities for the Floresplus calibration dataset.

Quantization Failure `SliM-LLM Aya` experience numerical instability error, although we have tried multiple dampening coefficient.

Packing failure Unfortunately, both mixed-precision methods fail to pack the quantized models into lower-bit formats (e.g., storing weights natively in 2-bit); thus, we must simulate the quan-

[6] https://github.com/ModelCloud/GPTQModel
[7] https://github.com/Aaronhuang-778/SliM-LLM
[8] https://github.com/The-Inscrutable-X/TACQ

Language	Characteristic		Ranking	
	Script	Family	Order	Resource Level
English	Latin	Germanic	5	High
Chinese	Hans	Sino-Tibetan	4	High
Indonesian	Latin	Austronesian	3	Mid
Tamil	Tamil	Dravidian	2	Mid
Swahili	Latin	Niger-Congo	1	Low

Table 5: Language resource ranking

Target Bit	Avg bit (Higher Bit Count)		
	GPTQ	SliM	TaCQ
2-bit	2.0	2.0 (5.52%)	2,049 (0.35%)
3-bit	3.0	3.0 (5.21%)	3,046 (0.35%)
4-bit	4.0	4.0 (25.05%)	4,042 (0.35%)

Table 6: Average bits of each quantization techniques and target-bit levels. `SliM-LLM` maintains total-bits while optimizing its bit-allocation, allocating higher bit to $target + 1$. `TaCQ` keep its higher-bits in 16-bit precision. Total parameter = 6.9B

tized models in full precision during inference. For `SliM-LLM`, we encounter packing failure when executing it from the source code. `TaCQ`, on the other hand, was not originally designed to prioritize inference efficiency.

Language and Average Bits Figure 5 shows our work's language coverage. The resource level is taken from `Aya dataset` (Singh et al., 2024), while the order is taken from OSCAR[9] (Brack et al., 2024). Figure 6 shows the average bit of each quantization techniques. Notably, even though `TaCQ` allocates some weight into 16-bit, its average bit stay close to the `GPTQ` counterpart (less than 0.1 point).

D Expanded Results

D.1 Perplexity Results

Figure 13 shows perplexity results at 3-bit and 2-bit. The C_{target} advantage over C_{cross} persist in extreme-bit, even with a higher contrast.

D.2 Massive and XStoryCloze Task Results

We can observe that the substantial effects of C_{target} are still visible in Massive (shown in Figure 14) and less visible in XStoryCloze (shown in Figure 15). This can be attributed to the hypothesis that calibration set can have substantial effect when the potential of quantized model is present, i.e., the gap between unquantized and the quantized. While in XStoryCloze results, we can see that the gap between Unquantized score and random-guess threshold are close (<0.2 accuracy point on ID, SW, and ZH). Another way to explain it, the original score is close to the random-guess, therefore there is nothing left to calibrate.

D.3 Global MMLU Lite and INCLUDE Task Results

Figure 16 and Figure 17 show Global MMLU Lite and INCLUDE results, respectively. We observe that the C_{target} effect are visible in some configurations, in particular at 3-bit, similar to Belebele evaluation in Figure 4. The combination of mixed-precision and Target-Calibration can serve as a safeguard against extreme-bit quantization failure, such in `3-bit SliM-LLM` model on INCLUDE's ID. Another note is that German and Dutch evaluation at 3-bit in INCLUDE prefer Cross-Calibration rather than using English-Calibration. This can be attributed to the fact that English-Calibration has less effective transfers on Germanic languages as seen in Figure 28.

D.4 MGSM

Exploratory math reasoning benchmark using limited thinking tokens. Due to the limited time and computational budget before the conference, we restrict the maximum number of thinking and output tokens to 1024 and consider an answer correct only after the model produced its finalized, formatted answer. This means that the quantized model must either understand the required answer format (e.g., "Answer: [number]" in English) or stop thinking within the 1024-tokens. Moreover, due to limited storage for saving generated samples, we cannot analyze the ratio of completed answers to those stopped mid-thought. Therefore, our analysis only considers a coarse-grained accuracy: whether the model can completely solve the problem, regardless of the reason.

We analyze with the first, second, and fourth patterns observed in the main findings. Note that we exclude `Aya Expanse 8B` and `Llama 3.1 8B Instruct` from the analysis because their results are unnatural. For example, 4-bit `Aya` model produces zero score on Tamil, and C_{cross} occasionally substantially outperforms C_{target}. First, `Qwen3-8B` at 3-bit achieves performance comparable to 4-bit model with `TaCQ` on high-resource languages, while not necessarily using C_{target}. In contrasts, `Qwen3-14B` at 2-bit fails to achieve non-zero accuracy and cannot be brought closer to 3-bit perfor-

[9] https://huggingface.co/datasets/oscar-corpus/community-oscar

mance, except on higher-resource languages. Second, lower-resource languages benefit more from C_{target} than higher-resource languages. Fourth, most SINQ scores are low; in particular, `Qwen 8B` and `Qwen3 14B` produce close to zero across nearly all settings.

Rationale on why close to zero score at MGSM. First and formost, the 1024-token may be too little for grade school math questions. In particular, Liu et al. (2025a) experiment on longer reasoning tasks, allowing 32768 tokens. Their findings show that aggressive quantization (i.e., 3-bit quantization) produces longer outputs, whereas the output length of 4-bit models is comparable with the unquantized baseline.

D.5 Multi Seed Experiment

Quantization on three run across seeds. We repeated quantization experiments on all 8B model at 2-bit and `Qwen3-8B` at 3-bit precision across calibration languages and quantization methods. For each configuration, we used random seeds of 1000, 1234, and 2000.

Perplexity. Figure 13 shows the perplexity results across three quantization runs at 3- and 2-bit. The local-diagonal or C_{target} consistently remains the lowest, while C_{cross} occasionally widens, especially at the 2-bit settings. In particular, the standard deviations of C_{target} are substantially lower than those for C_{cross}.

E Qwen3 14B Results

E.1 Perplexity

Figure 19 and 20 show perplexity results of the `Qwen3 14B` model at 3- and 2-bit, respectively. The pattern of C_{target} outperforming C_{cross} persists, as clearly seen in the diagonal entries of the figures. However, `SINQ`'s performance is substantially lower than C_{cross}. This pattern is similar to that observed in `Qwen3 8B` models, but contrasts with `Aya Expanse 8B` and `Llama 3.1 8B Instruct`, for which `SINQ`'s performance is comparable to C_{cross}, as shown in Figure 13.

E.2 downstream tasks

We contrast whether the first, second, and fourth patterns observed in `8B` models also hold in `Qwen3-14B` model. Specifically, we examine whether 2-bit performance is comparable to 3-bit when using with mixed-precision and/or C_{target}, how performance differs between high- and low-resource languages, and how `SINQ` performs relative to C_{cross}. We divide the analysis into three parts: MultiBlimp, Belebele, and MGSM.

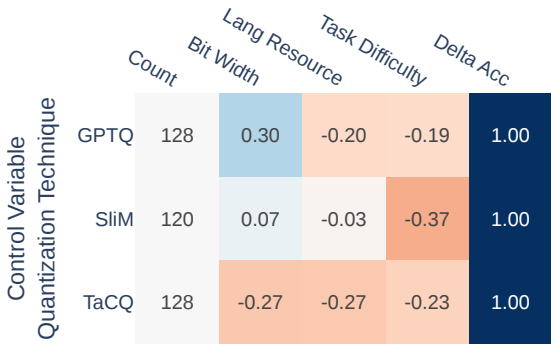


Figure 8: Calibration sensitivity across Quantization Techniques. Each row represents the correlation between delta accuracy and another variable.

The patterns exactly the same for MultiBlimp, MASSIVE, and XStoryCloze. Figure 21-23 show results of MultiBlimp, MASSIVE, and XStoryCloze, respectively, on `Qwen3 14B`. Again, we focus on 2-bit performance, while 3-bit performance is close to the unquantized baseline. The same three patterns persist in the `Qwen3 14B` model. First, 2-bit models can match 3-bit performance when using `TaCQ` and C_{target}. Specifically for MultiBlimp, even `GPTQ` 2-bit can match 3-bit performance. Second, the roles of `TaCQ` and C_{target} differ between high-resource and low-resource languages: `TaCQ` is necessary to achieve performance above random guess, while C_{target} is only necessary for lower-resource languages. Lastly, `SINQ`'s performance is comparable to C_{cross}, although it is occasionally performs worse on higher-resource languages.

The Target-Calibration effects shallower on Belebele, Global MMLU Lite, and INCLUDE.

Figure 24-26 show Belebele, Global MMLU Lite, and INCLUDE results, respectively, for `Qwen3 14B`. We observe that the choice of calibration becomes negligible at 3- and 2-bit, except for 2-bit `TaCQ`, where the difference remains substantial. The pattern becomes similar to MultiBlimp results of `Qwen3-14B` and `8B` models, but differs from the Belebele results for `8B` models. Therefore, the conclusions are unchanged with previous finding.

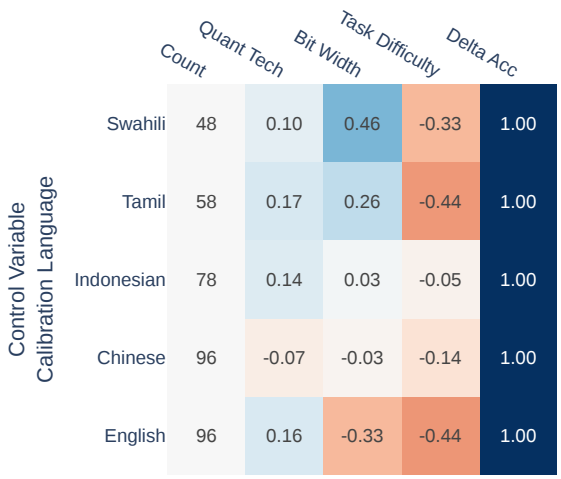


Figure 9: Calibration sensitivity across Calibration Languages. Each row represents the correlation between delta accuracy and another variable.

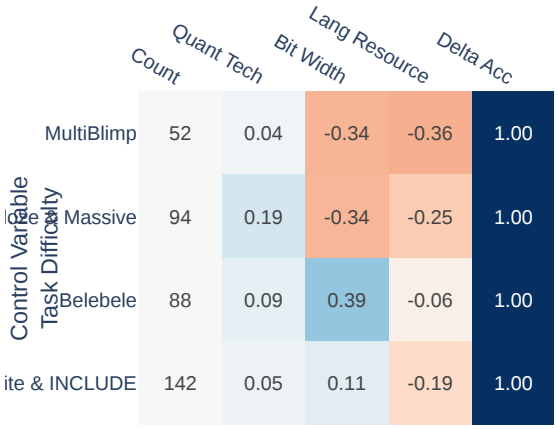


Figure 10: Calibration sensitivity across Task Difficulties. Each row represents the correlation between delta accuracy and another variable.

F Extended Calibration Sensitivity Analysis

F.1 Grouping by Quantization Technique

Figure 8 shows the correlations for `GPTQ`, `SliM-LLM`, and `TaCQ`. Bit width has an interesting relationship with quantization technique: calibration sensitivity tends to increase as the bit width increases for `GPTQ`, whereas the opposite trend is observed for `TaCQ`. Next, both `GPTQ` and `TaCQ` weakly correlate with "lower"-resource languages. Finally, all quantization techniques weakly correlate with "easier task, while SliM-LLM shows a stronger correlation with "easier" task.

F.2 Grouping by Language Resource

Figure 9 shows the correlations for each calibration language. Indonesian and Chinese exhibit the least variation across variables, with no clear correlation with calibration sensitivity. This is consistent with the downstream task results, where both languages show relatively low sensitivity to language-targeted calibration across settings. In contrast, Swahili and Tamil show stronger correlations with bit width and task difficulty, with higher calibration sensitivity at "higher" bit widths and "easier" tasks. Lastly, English shows an inverted sensitivity pattern with respect to bit-width and task difficutly, with $C_{\text{cross}} > C_{\text{en}}$ at higher bit widths and harder tasks. This pattern can be observed in the 3-bit results for Global MMLU Lite and INCLUDE, as shown in Figure 16-17.

F.3 Grouping by Task Difficulty

Figure 10 shows the correlations for "easier" and "harder" tasks. "Easier" tasks show a clear tendency toward "lower"-bit widths and "lower"-resource languages, consistent with the downstream task results. However, the correlation with quantization technique is very weak, despite the substantial delta accuracy observed for `SliM-LLM` and `TaCQ` and the negligible delta accuracy in `GPTQ` on "easier" tasks. This suggests that, for "easier" tasks, calibration sensitivity is more strongly associated with bit width than with quantization technique. In contrast, "harder" tasks show weaker sensitivity to the choice of calibration language.

G Expanded Cross-Lingual within family

Sino-Tibetan (Chinese) cross-lingual transfer may produce negative effect Based on the patterns in Figure 11, we categorize the cross-lingual effects into three levels: High, Low, and Detrimental (observed only at 2-bit setting). The high transfer effect are Chinese (ZH) and Yue Chinese (YUE), the two languages are the only Inner-Languages. Moderate effects are observed for Tibetan (BOD) and Dzongkha (DZO). The remaining languages fall into the detrimental effect, where a specific cross-lingual transfers yield significantly better perplexity. Upon closer inspection, all four in the Detrimental effect category tend to prefer calibration using a low-resource language: Swahili yields better perplexity for Southern Jinghpaw (KAC) and Mizo (LUS), while Tamil yields

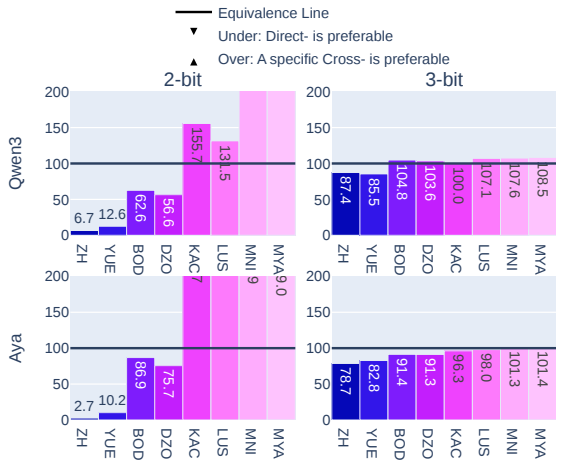


Figure 11: Sino-Tibetan (Chinese) cross-lingual transfer benefits aggregated across quantization methods. Values represent the perplexity ratio percentage between C_{zh} and mean of 4 C_{cross}. Languages are ordered (left to right) according to phylogenetic proximity in `Glottolog`. Among the evaluated languages, only 2 languages, ZH and Yue Chinese (YUE), that are closer in siblingness.

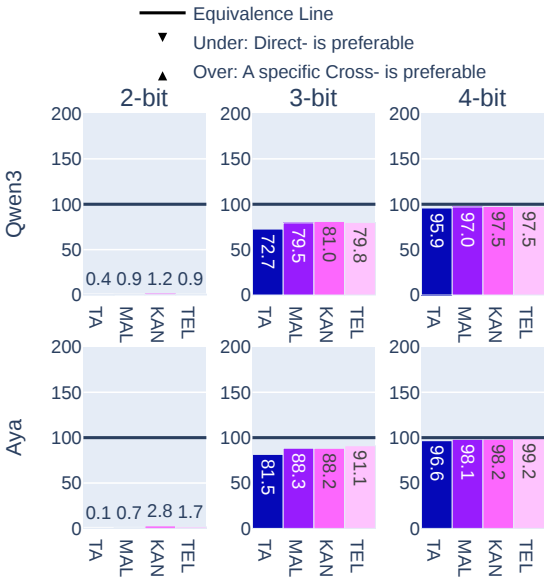


Figure 12: Perplexity of Dravidian (Tamil) on cross-lingual calibration. The score is calculated using $C_{\text{target}}/\overline{C_{\text{cross}}}$ (lower is preferring Tamil), aggregate across quantization methods.

better perplexity for Manipuri (MNI) and Burmese (MYA). Interestingly, all four of these languages are geographically close to Myanmar (Burmese). Notably, these suboptimal effects are observe only at the 2-bit level. This may suggest a nuanced role for low-resource languages in improving another low-resource languages.

Cross-lingual effects are negligible or negative for high-resource languages, limited for Indonesian, and significant for low-resource languages. Across these five language families, we can see a general pattern (see Appendix G.1 for the barplot). High-resource languages tend to restrict benefits to themselves, yielding negligible or even negative effects for their sibling languages. Conversely, lower-resource languages, especially those using different scripts, exhibit much stronger cross-lingual benefits. We note the limitation that these cross-lingual effects are observed only through perplexity remain and negligible except at 2-bit level; therefore, these findings do not override our general recommendation to use C_{target} for targeted language settings.

G.1 Cross-Lingual Perplexity Results

The pattern of high-resource languages tend to prefer transfer to self-language that can be seen in Figure 28, showing Germanic evaluations. We can see that at 2-bit, English is the only language that benefited from English-Calibration, even Eastern Yiddish (YDD) prefer Tamil language. The pattern of low-resources tend to share the benefit across language within the same family can be seen in Figure 12 and Figure 29, showing Dravidian (Tamil) and Niger-Congo (Swahili), respectively.

G.2 Cross-Lingual Downstream Task Results

Figure 30-33 show the general (barplot) and detailed (heatmap) evaluations of Cross-Lingual Calibration on Belebele, encompassing Austronesian (Indonesian) and Sino-Tibetan (Chinese). Due to the time constraint, we only evaluate on 3-bit for Austronesian and Qwen3 at three bit-levels for Sino-Tibetan. Overall, the transfer effect is negligible, clearly seen from the barplots, while the relative advantage varies according to the language family. Specifically, Austronesian still prefer Indonesian-Calibration (Swahili as secondary). As for Sino-Tibetan, evaluating on BOD and KAC have their unquantized's accuracy close to random-guess, so their evaluation should not be considered. As for Burmese (MYA), it prefers Tamil (Swahili as Secondary) for the extreme-bits. This pattern corroborates with the perplexity evaluations at Figure 11, although its belebele's delta accuracies are negligible.

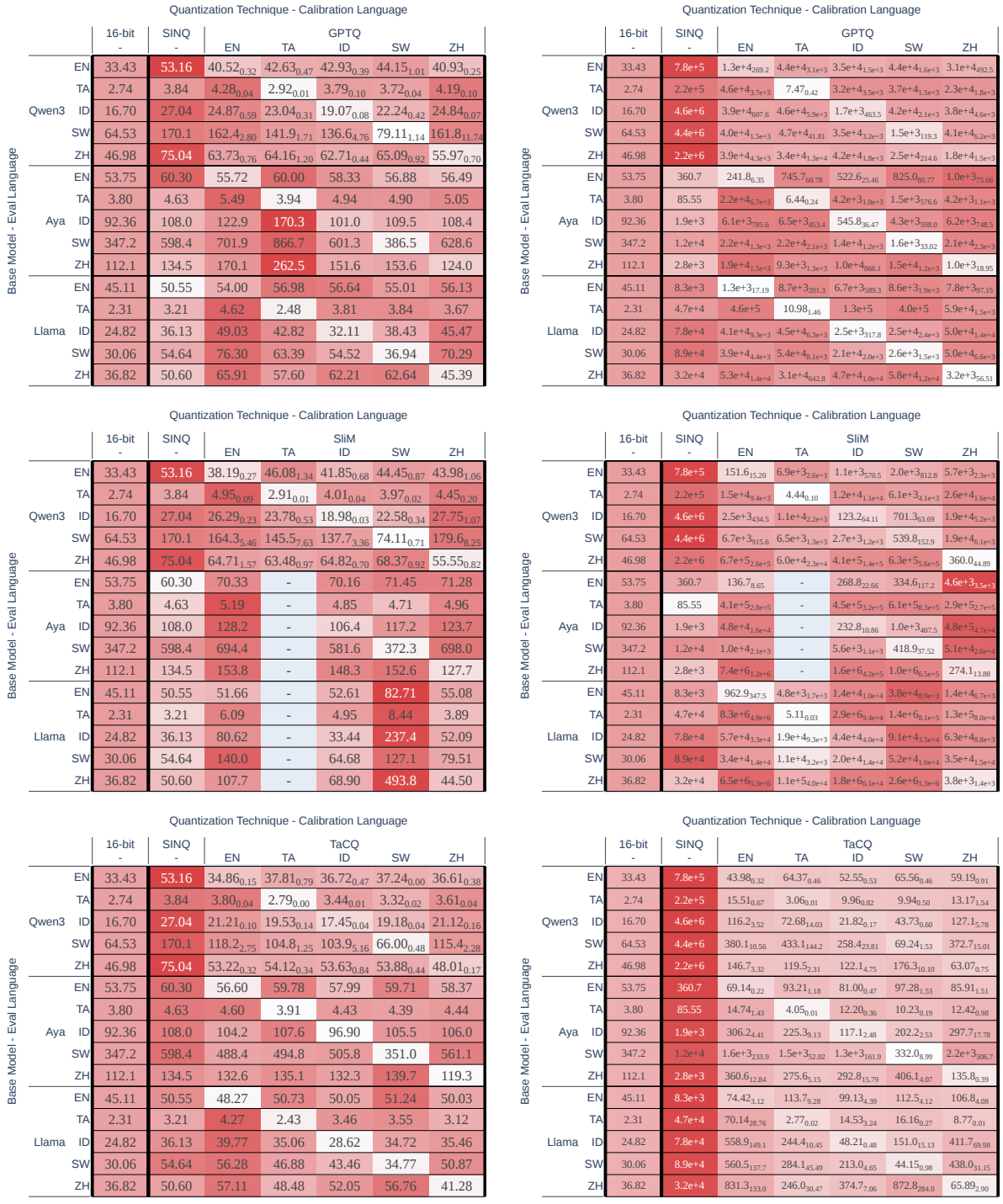


Figure 13: Ablation perplexity with a recent data-free Quantization (SINQ) and 3 random seed (Qwen3 and Aya Expanse) - 3-bit (Left) and 2-bit (Right) - GPTQ (Top), SliMLLM (Middle), and TaCQ (Bottom) - lighter or lower is better.

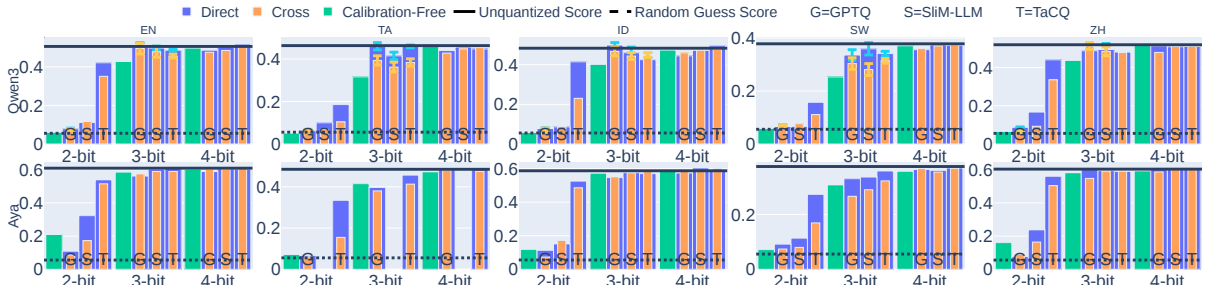


Figure 14: Massive Barplot.



Figure 15: XStoryCloze Barplot.



Figure 16: Global MMLU Lite Barplot.



Figure 17: INCLUDE Barplot. INCLUDE does not have English evaluation, so we aggregate Dutch and German evaluations as a proxy for English.

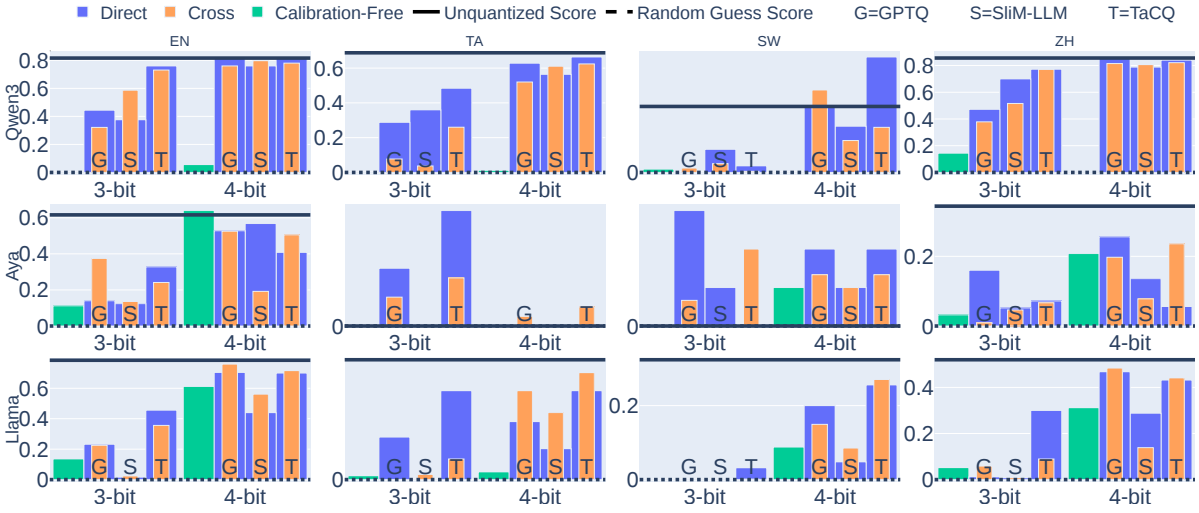


Figure 18: MGSM: Contrasting finalized accuracy within the allocated 1024 thinking tokens (Higher is better) of C_{target} vs mean of 3 C_{cross}. The evaluation only considers the final response after the complete reasoning chain has terminated, i.e., formatted answer.

Quantization Technique - Calibration Language

Model - Eval Language		16-bit -	SINQ -	GPTQ EN	GPTQ TA	GPTQ ID	GPTQ SW	GPTQ ZH	SINQ -	TaCQ EN	TaCQ TA	TaCQ ID	TaCQ SW	TaCQ ZH
Qwen3 14B	EN	31.91	65.45	36.57	39.74	38.12	37.60	37.18	65.45	33.21	39.80	34.56	34.83	34.47
	TA	2.53	3.09	3.43	2.65	3.11	3.03	3.25	3.09	3.34	2.56	2.92	2.84	3.10
	ID	15.14	22.25	19.30	18.60	16.84	18.03	19.42	22.25	18.20	16.79	15.55	16.30	17.86
	SW	46.73	140.3	106.3	101.0	87.25	53.11	102.7	140.3	83.34	74.30	66.94	46.64	81.96
	ZH	44.59	58.40	54.28	53.33	54.87	54.54	49.93	58.40	48.42	48.44	49.45	49.46	45.36

Figure 19: Qwen3-14B Floresplus 3-bit.

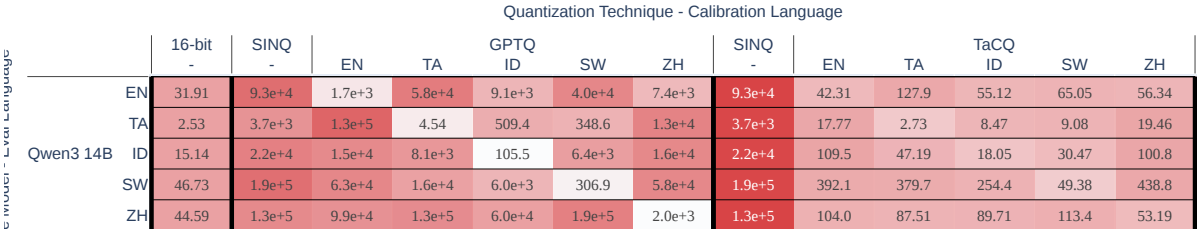


Figure 20: Qwen3-14B Floresplus 2-bit.

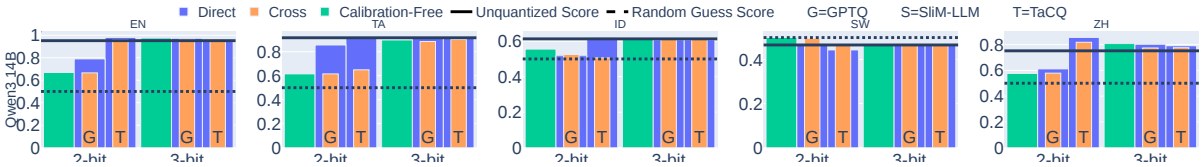


Figure 21: Qwen3-14B MultiBlimp Barplot.

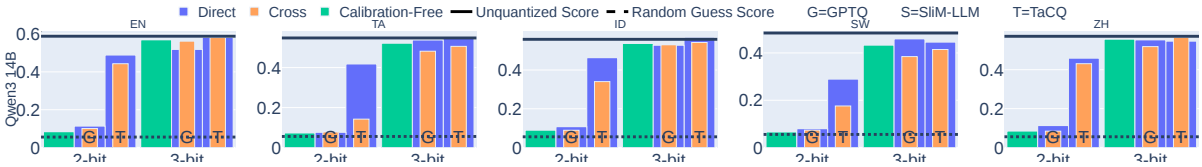


Figure 22: Qwen3-14B Massive Scanario Barplot.

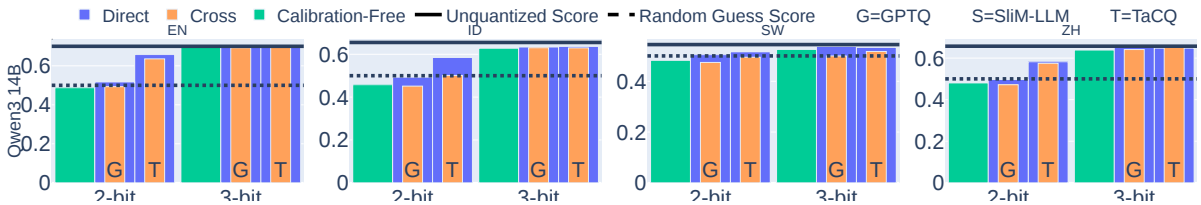


Figure 23: Qwen3-14B XStoryCloze Barplot.

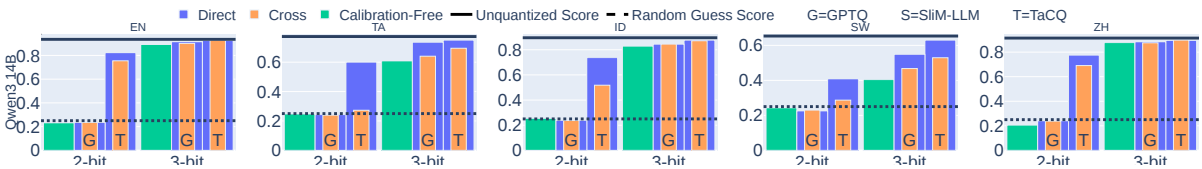


Figure 24: Qwen3-14B Belebele Barplot.

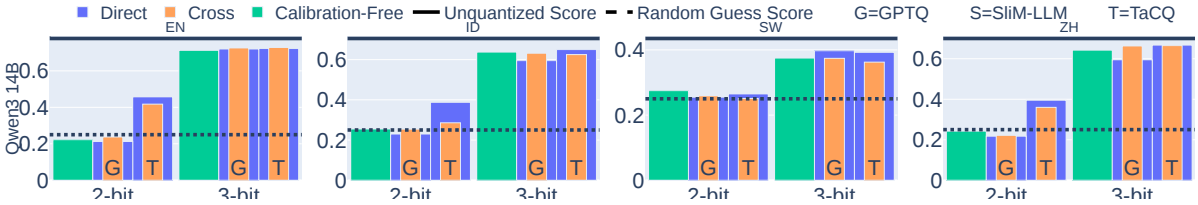


Figure 25: Qwen3-14B Global MMLU Lite Barplot.

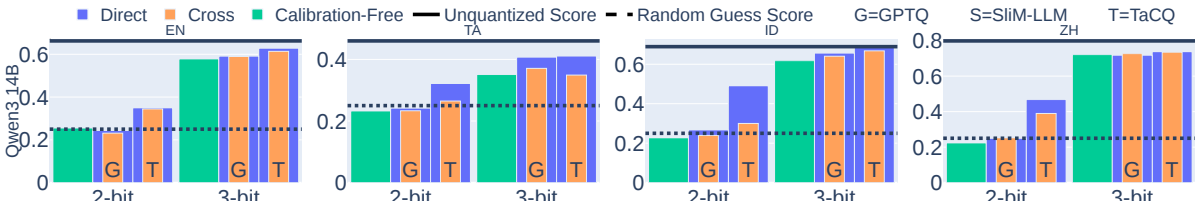


Figure 26: Qwen3-14B INCLUDE Barplot. INCLUDE does not have English evaluation, so we aggregate Dutch and German evaluations as a proxy for English.

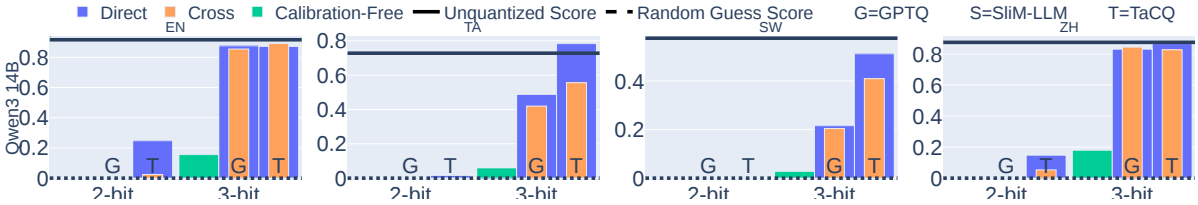


Figure 27: Qwen3-14B MGSM Barplot. Contrasting finalized accuracy within the allocated 1024 thinking tokens (Higher is better) of C_{target} vs mean of 3 C_{cross}. The evaluation only considers the final response after the complete reasoning chain has terminated, i.e., formatted answer.



Figure 28: Perplexity of Germanic (English) on cross-lingual calibration. The score is calculated using $C_{\text{target}}/\overline{C_{\text{cross}}}$ (lower is preferring English), aggregate across quantization methods.

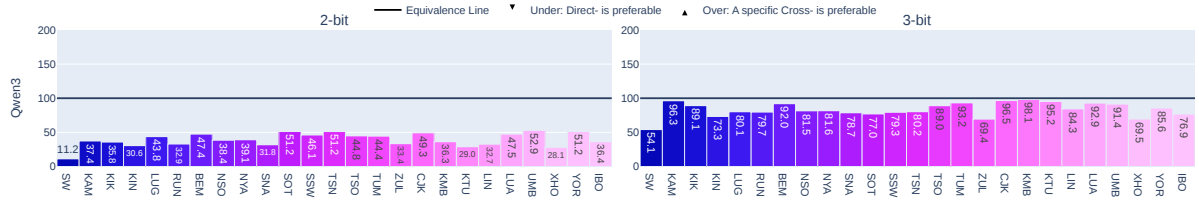


Figure 29: Perplexity of Niger-Congo (Swahili) on cross-lingual calibration. The score is calculated using $C_{\text{target}}/\overline{C_{\text{cross}}}$ (lower is preferring Swahili), aggregate across quantization methods.

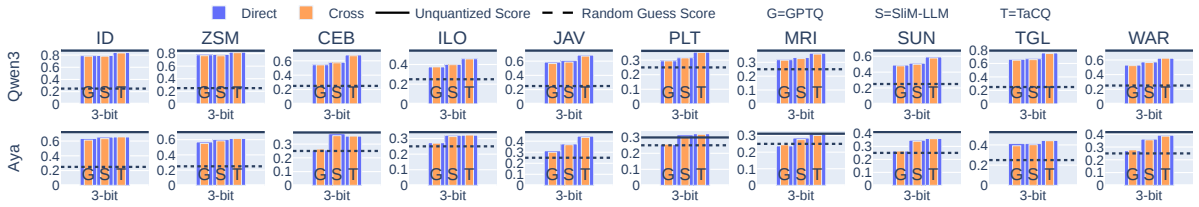


Figure 30: Belebele Austronesian (Indonesian) accuracy.

Quantization Technique - Calibration Language - Base Model

Eval Language		16-bit	Gptq					Slimllm					Tacq				
		-	EN	TA	ID	SW	ZH	EN	TA	ID	SW	ZH	EN	TA	ID	SW	ZH
Qwen3	ID	0.87	-0.14	-0.13	-0.08	-0.10	-0.08	-0.14	-0.10	-0.08	-0.10	-0.12	-0.04	-0.07	-0.03	-0.06	-0.06
	ZSM	0.85	-0.13	-0.12	-0.06	-0.10	-0.10	-0.18	-0.12	-0.06	-0.11	-0.11	-0.04	-0.06	-0.02	-0.03	-0.05
	CEB	0.74	-0.25	-0.20	-0.20	-0.12	-0.24	-0.25	-0.17	-0.16	-0.14	-0.19	-0.10	-0.09	-0.06	-0.08	-0.11
	ILO	0.54	-0.18	-0.19	-0.17	-0.10	-0.19	-0.18	-0.13	-0.14	-0.13	-0.16	-0.09	-0.09	-0.08	-0.05	-0.13
	JAV	0.75	-0.28	-0.20	-0.16	-0.16	-0.26	-0.29	-0.18	-0.14	-0.16	-0.21	-0.13	-0.12	-0.06	-0.07	-0.11
	PLT	0.36	-0.09	-0.06	-0.07	-0.05	-0.08	-0.07	-0.02	-0.05	-0.05	-0.06	-0.01	-0.02	-0.01	-0.01	-0.02
	MRI	0.38	-0.11	-0.05	-0.07	-0.05	-0.10	-0.09	-0.05	-0.05	-0.08	-0.07	-0.05	-0.03	-0.02	-0.04	-0.03
	SUN	0.67	-0.26	-0.20	-0.18	-0.15	-0.25	-0.27	-0.18	-0.16	-0.15	-0.22	-0.17	-0.11	-0.08	-0.07	-0.14
	TGL	0.79	-0.21	-0.18	-0.13	-0.13	-0.18	-0.23	-0.12	-0.12	-0.13	-0.18	-0.08	-0.05	-0.03	-0.04	-0.09
	WAR	0.73	-0.26	-0.22	-0.20	-0.16	-0.25	-0.23	-0.18	-0.15	-0.15	-0.17	-0.09	-0.10	-0.10	-0.07	-0.14
Aya	ID	0.72	-0.11	-0.23	-0.09	-0.17	-0.12	-0.14	-	-0.07	-0.16	-0.16	-0.07	-0.04	-0.07	-0.09	-0.04
	ZSM	0.70	-0.15	-0.29	-0.13	-0.18	-0.15	-0.16	-	-0.10	-0.17	-0.17	-0.08	-0.06	-0.08	-0.11	-0.08
	CEB	0.38	-0.10	-0.12	-0.13	-0.08	-0.12	-0.06	-	-0.01	-0.06	-0.08	-0.03	-0.01	-0.03	-0.04	-0.01
	ILO	0.34	-0.09	-0.07	-0.07	-0.07	-0.09	-0.06	-	-0.02	-0.04	-0.06	-0.04	-0.01	-0.02	-0.00	-0.01
	JAV	0.48	-0.24	-0.24	-0.17	-0.16	-0.21	-0.13	-	-0.11	-0.11	-0.16	-0.09	-0.05	-0.04	-0.08	-0.08
	PLT	0.30	-0.04	-0.01	-0.04	-0.04	-0.05	-0.03	-	0.02	-0.05	-0.03	0.01	0.02	0.02	0.01	0.01
	MRI	0.31	-0.05	-0.04	-0.07	-0.06	-0.06	-0.07	-	-0.03	-0.06	-0.06	-0.03	-0.03	0.00	-0.05	-0.03
	SUN	0.41	-0.15	-0.16	-0.15	-0.14	-0.13	-0.09	-	-0.06	-0.10	-0.11	-0.07	-0.06	-0.05	-0.05	-0.04
	TGL	0.53	-0.15	-0.25	-0.11	-0.14	-0.18	-0.15	-	-0.11	-0.14	-0.18	-0.12	-0.09	-0.08	-0.08	-0.08
	WAR	0.42	-0.14	-0.13	-0.15	-0.09	-0.13	-0.09	-	-0.05	-0.07	-0.08	-0.06	-0.03	-0.02	-0.04	-0.06

Figure 31: Belebele Austronesian (Indonesian) accuracy: 3-bit, lighter or higher is better.

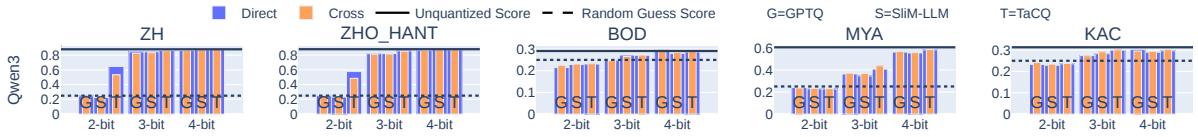


Figure 32: Belebele Sino-Tibetan (Chinese) accuracy. ZHO_HANT is Tradiational Chinese

Quantization Technique - Calibration Language - Base Model

Eval Language		16-bit	Gptq					Slimllm					Tacq				
		-	EN	TA	ID	SW	ZH	EN	TA	ID	SW	ZH	EN	TA	ID	SW	ZH
Qwen3	ZH	0.88	-0.01	-0.01	-0.02	-0.01	-0.02	-0.01	-0.01	-0.01	-0.01	-0.01	-0.00	-0.01	-0.00	0.00	-0.00
	BOD	0.29	0.00	-0.00	-0.01	-0.01	0.01	-0.01	0.00	-0.01	0.00	-0.01	-0.01	0.00	0.00	0.01	-0.00
	KAC	0.31	-0.02	-0.03	-0.02	-0.02	-0.01	-0.02	-0.01	-0.02	-0.02	-0.02	0.00	-0.01	-0.01	-0.01	-0.02
	MYA	0.60	-0.03	-0.02	-0.06	-0.02	-0.04	-0.05	-0.03	-0.05	-0.04	-0.05	-0.04	-0.01	-0.01	-0.01	-0.02
	ZHO	0.88	-0.01	-0.03	-0.02	-0.02	-0.01	-0.02	-0.01	-0.01	0.00	-0.00	-0.01	-0.01	-0.00	-0.00	-0.01

Quantization Technique - Calibration Language - Base Model

Eval Language		16-bit	Gptq					Slimllm					Tacq				
		-	EN	TA	ID	SW	ZH	EN	TA	ID	SW	ZH	EN	TA	ID	SW	ZH
Qwen3	ZH	0.88	-0.07	-0.07	-0.07	-0.08	-0.03	-0.05	-0.05	-0.06	-0.08	-0.03	-0.01	-0.04	-0.02	-0.02	-0.02
	BOD	0.29	-0.04	-0.05	-0.06	-0.04	-0.04	-0.04	-0.02	-0.04	-0.03	-0.02	-0.01	-0.02	-0.02	-0.02	-0.02
	KAC	0.31	-0.05	-0.04	-0.03	-0.04	-0.04	-0.03	0.01	-0.02	-0.01	-0.03	-0.03	-0.01	0.00	-0.01	-0.01
	MYA	0.60	-0.29	-0.22	-0.24	-0.16	-0.24	-0.28	-0.13	-0.22	-0.24	-0.26	-0.19	-0.09	-0.16	-0.11	-0.20
	ZHO	0.88	-0.06	-0.08	-0.08	-0.08	-0.06	-0.06	-0.08	-0.06	-0.09	-0.06	-0.04	-0.04	-0.02	-0.03	-0.02

Quantization Technique - Calibration Language - Base Model

Eval Language		16-bit	Gptq					Slimllm					Tacq				
		-	EN	TA	ID	SW	ZH	EN	TA	ID	SW	ZH	EN	TA	ID	SW	ZH
Qwen3	ZH	0.88	-0.65	-0.64	-0.64	-0.64	-0.63	-0.65	-0.63	-0.66	-0.65	-0.66	-0.34	-0.51	-0.36	-0.63	-0.23
	BOD	0.29	-0.05	-0.08	-0.05	-0.05	-0.08	-0.04	-0.07	-0.06	-0.06	-0.06	-0.06	-0.04	-0.06	-0.06	-0.06
	KAC	0.31	-0.09	-0.05	-0.07	-0.04	-0.08	-0.08	-0.07	-0.07	-0.08	-0.08	-0.08	-0.05	-0.09	-0.08	-0.08
	MYA	0.60	-0.37	-0.36	-0.38	-0.38	-0.36	-0.37	-0.37	-0.36	-0.38	-0.38	-0.38	-0.36	-0.38	-0.38	-0.38
	ZHO	0.88	-0.62	-0.63	-0.64	-0.63	-0.63	-0.65	-0.64	-0.64	-0.65	-0.65	-0.37	-0.54	-0.40	-0.63	-0.30

Figure 33: Belebele Austronesian (Chinese) accuracy: 4-bit (Top), 3-bit (Middle), and 2-bit (Bottom), lighter or higher is better. ZHO is Traditional Chinese.