Measuring and Fixing Multilingual Accuracy Drops in Blackwell NVFP4 Weight Quantization

Azzam Al-solamy[1], Hesham Al-sadan[1], Noura Al-qasim[1], Hesham Banafa[2], Omer Nacar[1]
[1]Tuwaiq Academy, Riyadh, Saudi Arabia
[2]Resilience Co, Riyadh, Saudi Arabia
Correspondence: o.najar@tuwaiq.edu.sa

Abstract

Four-bit weight quantization is routine, but its cost is reported as one English-centric average and its calibration corpora are English by default. We show that the average hides where the loss lands, and that a calibration corpus does not buy knowledge of its own language. We run a controlled comparison of FP16, uncalibrated NVFP4, and two Arabic-calibrated AWQ variants over four instruction-tuned models and five languages — 56 paired cells per variant, scored on accuracy, KL divergence from FP16, perplexity, and weight reconstruction error. Three results stand out. First, the degradation is *not* concentrated in Arabic: Arabic MSA retains the most accuracy and Russian the least, even though Arabic shows the largest distributional shift under KL, so the metrics disagree at both extremes. Second, an all-Arabic corpus improves Arabic and non-Arabic evaluations by the same $+0.43$ pp within noise — an interaction of -0.001 pp — and only its non-Arabic KL gain reaches significance. Third, the benefit of calibration is proportional to the damage: the model quantization hurts most is the one calibration repairs most, while the least-damaged model gains nothing. And weight reconstruction error ranks the best variant last and is language-invariant by construction.

1 Introduction

Four-bit weight-only post-training quantization (PTQ) is now the default way to fit an instruction-tuned language model onto a single accelerator. GPTQ (Frantar et al., 2023) and AWQ (Lin et al., 2024) are standard practice, scaling analyses find four-bit parameters almost universally optimal for the trade-off between total model bits and zero-shot accuracy (Dettmers and Zettlemoyer, 2023), and hardware has followed: NVFP4, a four-bit floating-point format that pairs an FP8 scale on every 16-element block with a second, per-tensor scale, is now supported natively and exposed through vendor tooling (Alvarez et al., 2025; NVIDIA Corporation, 2025).

Two questions are usually left unanswered when such a model ships. First, *where does the loss land?* Aggregate quality is normally reported on English benchmarks, yet compression interacts with multilinguality: Marchisio et al. (2024) find that quantization disproportionately harms non-Latin-script languages, that automatic metrics underestimate the degradation relative to human judgement, and that the hardest tasks degrade fastest, while Mekala et al. (2025) report that quantization damage worsens when the input is not in English. Arabic — a morphologically rich, non-Latin-script language with a wide gap between its standard register and its regional dialects — is barely represented in this literature. Second, *what does the calibration corpus actually buy?* PTQ estimates its scales from a small calibration set (Hubara et al., 2021), and for language models the composition of that sample measurably changes the resulting model (Williams and Aletras, 2023). Chimoto et al. (2026) show that non-English and multilingual calibration sets reduce perplexity relative to English-only ones, and that matching the calibration language to the evaluation language yields the largest per-language gains. If a practitioner calibrates entirely on Arabic, do they buy Arabic quality at the expense of everything else?

We answer both with one controlled sweep: four instruction-tuned models, one 4-bit format, five evaluation languages, and a calibration corpus that is 100 % Arabic. Our contributions are:

1. A four-way comparison of FP16, uncalibrated NVFP4, and two Arabic-calibrated AWQ recipes over 56 paired cells and four metrics (§3, Table 1).
2. A per-language attribution showing that accuracy and KL disagree at the extremes — of the four comparable languages, Arabic MSA

retains the most accuracy while showing the largest distributional shift, and Russian the least while staying among the most faithful — so Arabic MSA is the most robust of the four on the metric users experience (§4.3, Table 2).

3. Evidence that the gain is delivered by AWQ's per-channel scale rather than by the Arabic corpus itself: with that scale, Arabic and non-Arabic evaluations gain indistinguishably; without it, the same corpus transfers nothing (§4.4, Table 3).
4. A demonstration that weight reconstruction error inverts the correct ranking of these variants and cannot attribute a loss to a language (§4.2, Tables 2 and 4).

2 Related Work

Post-training quantization. Weight-only PTQ replaces each weight tensor W with a quantized reconstruction $\widehat{W}$, leaving activations in higher precision. Outlier channels dominate the error budget (Dettmers et al., 2022), motivating either migration of the difficulty out of the activations and into the weights (Xiao et al., 2023) or a per-channel rescaling of weight columns derived from activation statistics, as in AWQ (Lin et al., 2024). Second-order weight updates (Frantar et al., 2023) and low-rank compensation of the quantization error (Yao et al., 2024) are complementary; see Zhu et al. (2024) for a survey and Kurtic et al. (2024) for the accuracy–performance trade-off at deployment scale. We study the AWQ family because it is the recipe the vendor toolkit exposes for NVFP4 (NVIDIA Corporation, 2025).

Four-bit formats. FP8 established narrow floating point as a practical inference format (Micikevicius et al., 2022); NVFP4 extends this to four bits with two levels of scaling: an FP8 scale on every 16-element block, and a per-tensor scale above that (Alvarez et al., 2025). Sensitivity is not uniform inside the network: Cim et al. (2026) find that MLP up- and down-projection layers dominate FP4 quantization sensitivity while attention projections are substantially less sensitive, and that the depth profile differs between NVFP4 and MXFP4 — early blocks can be highly sensitive under MXFP4 — which is one reason a single "4-bit" condition can hide large behavioural differences. We hold the format fixed at NVFP4 W4A16 and vary only the calibration recipe.

Calibration data and language. PTQ scales are estimated from a few hundred sequences (Hubara et al., 2021), and their composition matters (Williams and Aletras, 2023). Closest to our work, Chimoto et al. (2026) compare eight calibration settings — five single-language and three multilingual mixes — on two quantizers over data from ten languages, and report that multilingual mixes give the best average perplexity while language-matched sets give the best per-language results. Our study is complementary in three ways: we fix the calibration language at Arabic and vary only the evaluation language; we report task accuracy, distributional fidelity, perplexity and weight error per cell rather than perplexity alone; and we include a dialectal Arabic slice alongside the standard register.

3 Experimental Setup

Models. Four instruction-tuned models of comparable size, two multilingual and two Arabic-centric: Gemma-2-9B-it (Gemma Team, 2024), Llama-3.1-8B-Instruct (Grattafiori et al., 2024), ALLaM-7B-Instruct-preview (Bari et al., 2024) and Fanar-1-9B-Instruct (Fanar Team, 2025).

Variants. FP16 is the reference. `nvfp4` is NVFP4 with the quantization range read directly off the weights and *no* calibration data. `awq_clip` and `awq_lite` are NVFP4 plus AWQ (Lin et al., 2024), both calibrated on 512 Arabic samples: `awq_clip` clips the weight range only, while `awq_lite` additionally folds a learned per-channel scale into the weights. All three 4-bit variants are **W4A16** — weights quantized, activations in FP16 — so every contrast isolates the weight quantizer and the calibration data. Quantization is simulated with the vendor fake-quantization backend (NVIDIA Corporation, 2025) so that FP16 and quantized runs see identical inputs. This measures the *numerics* of NVFP4, not its execution: no kernel here runs in four bits and we report no latency or throughput, so deployment-cost claims below follow from where a variant places its scales, not from timing.

Calibration corpus. Calibration uses **512 Arabic samples of 512 tokens each — 262,144 calibration tokens** — half Modern Standard Arabic and half Gulf dialect, with no non-Arabic text. This deliberate mismatch — one calibration language, five evaluation languages — is what makes §4.4

answerable, and it varies language where Williams and Aletras (2023) varied domain and Chimoto et al. (2026) varied the language mixture.

Evaluation. Arabic MSA, English, Spanish and Russian are each scored on Global-MMLU (Singh et al., 2025), XNLI (Conneau et al., 2018) and XStoryCloze (Lin et al., 2022). Gulf Arabic uses two in-house dialectal sets, `arabic_cultural_qa` and `dialectal_mmlu`, because no public dialectal benchmark covers the same task formats; Gulf is therefore reported alongside, but never pooled with, the four comparable languages. Every cell is 1000 items, giving $4 \times 14 = 56$ paired cells per variant.

Metrics. Per cell we record task **accuracy**, and summarise a slice by its *retention* $R = \overline{\text{acc}_q}/\overline{\text{acc}_{\text{FP16}}}$, a ratio of means. **KL divergence** $D_{\text{KL}}(p_{\text{FP16}} \,\|\, p_q)$ between the FP16 and quantized next-token distributions on the same prompts measures how far the model moved from itself rather than whether it became wrong; we pair it with **top-1 agreement**, the fraction of positions where the two argmaxes coincide. **Perplexity** is reported as the ratio $r = \text{PPL}_q/\text{PPL}_{\text{FP16}}$ on identical text, so $r = 1$ is no loss. **Weight MSE** is $\frac{1}{|W|}\|W - \widehat{W}\|_2^2$ averaged over quantized linear layers, a function of the weights alone.

The four metrics do not share a denominator. Accuracy is scored on all 1000 items of every cell; KL and top-1 agreement on 200 comparison rows per cell, falling to 64 on the six cells where ALLaM-7B and Fanar-1-9B meet Arabic MSA (Limitations); perplexity on the 48 non-Gulf cells; and weight MSE once per model, identically across that model's 14 cells. Contrasts use paired t-tests over the 56 cells; the Arabic/non-Arabic interaction uses Welch's test. At 1000 items per cell one standard error is roughly 1.5 pp, so a difference of that size in a single cell is read as a tie.

4 Results

4.1 The cost of 4-bit, and what calibration returns

Every 4-bit variant lands within 2.3 pp of FP16 on the model average (Table 1), consistent with the four-bit operating point argued for by Dettmers and Zettlemoyer (2023). Two differences separate the variants. `awq_lite` is the only configuration with no catastrophic cell: it removes all four runs that fell more than three points (Table 1), every one of them Llama-3.1-8B (Table 4). And its improvement over uncalibrated `nvfp4` is significant ($+0.43$ pp, $p = 0.018$), whereas `awq_clip` is not distinguishable from no calibration at all ($+0.15$ pp, $p = 0.27$; Table 1).

Table 1: Aggregate results over all 56 paired cells. "Worst" is the largest single-cell accuracy drop against FP16; "< -3" counts cells falling more than three points below it. Paired t-tests against `nvfp4` over the 56 cells: `awq_lite` $+0.43$ pp ($p = 0.018$) and `awq_clip` $+0.15$ pp ($p = 0.27$) on accuracy; on KL, `awq_lite` improves 47 of 56 cells ($p = 0.033$) and `awq_clip` 27 of 56 ($p = 0.106$).

	Acc.	Ret.	KL	Top-1	Worst	< -3
FP16	60.89	100.00	—	—	—	—
`nvfp4`	59.87	98.33	0.0702	84.15	−7.2	4
`awq_clip`	60.03	98.58	0.0750	83.79	−7.8	4
`awq_lite`	**60.30**	**99.04**	**0.0627**	**85.23**	−2.9	**0**

Fidelity orders the variants the same way but more sharply. `awq_lite` lowers KL on all four models (Table 4) and on 47 of 56 cells ($p = 0.033$), while `awq_clip` improves only 27 of 56 and raises mean KL overall (Table 1). Since `awq_clip` and `awq_lite` consume the *same* Arabic corpus and differ only in whether a per-channel scale is written, this isolates the AWQ per-channel scale (Lin et al., 2024), not the calibration text, as the mechanism. Perplexity separates nothing: all three variants sit at 1.03–1.04× FP16 and the ordering flips between models — `awq_lite` is the best variant on Fanar-1-9B and the worst on ALLaM-7B (Table 4).

4.2 Weight MSE cannot rank or attribute

Weight reconstruction error is the cheapest available quality proxy, and it fails here twice. It *inverts* the ranking: `awq_lite` folds a learned per-channel scale into the weights and compensates in the activation path (Lin et al., 2024), so its tensors are deliberately further from the originals and its MSE is higher than `nvfp4` on every model, from 1.9× (ALLaM-7B) to 320× (Fanar-1-9B), including the model it helps most. `awq_clip`, which writes no such scale, tracks the uncalibrated baseline to within 0.06 %. And it is *language-invariant by construction*: the tensors are fixed before any evaluation text is seen, so the value is identical for all five languages (Table 2, last column). Nor does it rank the models: it catches the hardest one — Llama-3.1-8B, highest on MSE and worst on accuracy — but cannot separate the rest: Fanar-1-

Table 2: Uncalibrated `nvfp4` against FP16, by evaluation language. Accuracy and KL disagree at both extremes; over the four comparable languages the two orderings are only weakly related (Spearman $\rho = +0.40$ between retention and KL, where agreement would be negative; $n = 4$, $p = 0.60$, so descriptive only). Gulf uses different datasets (§3) and is listed apart. Weight MSE is identical everywhere (§4.2); the printed value is the full-precision one from the non-Arabic rows, because the Arabic rows are stored rounded to 10^{-6} (Limitations).

Language	Ret.	KL	Top-1	PPL	MSE
Arabic MSA	**99.03**	0.0773	81.42	1.043	8.3e-7
English	98.57	**0.0570**	**87.63**	**1.023**	8.3e-7
Spanish	98.06	0.0710	81.50	1.030	8.3e-7
Russian	97.52	0.0580	86.58	1.033	8.3e-7
Gulf Arabic	98.54	0.0965	83.38	n/a	8.3e-7
Arabic, all	98.83	0.0850	82.20	—	8.3e-7
non-Arabic, all	98.08	0.0620	85.24	—	8.3e-7

9B and Gemma-2-9B have effectively the same MSE (3.1e-7 against 3.2e-7) while Fanar loses $2.0\times$ Gemma's accuracy (Table 4). The MSE and accuracy-loss orderings agree no better than accuracy and KL agree across languages (Spearman $\rho = +0.40$ in both cases; $n = 4$, $p = 0.60$, descriptive only), so we draw neither a model-level nor a language-level conclusion from it.

4.3 The loss is not concentrated in Arabic

On accuracy, Arabic MSA is the most robust of the four comparable languages, retaining 99.03 % of its FP16 score, while Russian retains the least at 97.52 %; grouped, Arabic retains 98.83 % against 98.08 % for non-Arabic (Table 2, Figure 1). On KL the ordering inverts: Arabic MSA has the highest divergence of the four (0.0773 against 0.0570 for English), Gulf Arabic is higher still (0.0965), and top-1 agreement follows KL rather than accuracy — 81.4 % on Arabic MSA against 87.6 % on English (Table 2). Perplexity puts Arabic worst by about two percent, a spread narrower than the per-cell noise (Table 2; Limitations).

Quantization therefore perturbs the Arabic output distribution more than any other language's, but that perturbation converts into *fewer* task errors in Arabic than in Russian or Spanish (Table 2). This is a concrete instance of the metric-disagreement warning of Marchisio et al. (2024), with the direction resolved: a practitioner tracking only KL would protect Arabic, while accuracy names Russian.

Table 3: Effect of the all-Arabic calibration data; both blocks are the variant minus uncalibrated `nvfp4`. Top: grouped by whether the evaluation language is the calibration language. Bottom: per language. Neither group accuracy gain is significant alone (paired t: Arabic $p = 0.09$, non-Arabic $p = 0.08$), so what survives is the pooled effect over all 56 cells (+0.43 pp, $p = 0.018$; Table 1) plus the interaction, which is −0.001 pp (95 % t CI $[-0.69, +0.69]$ pp, df $= 48.8$, $p = 0.999$). On KL only the non-Arabic `awq_lite` gain is significant ($p = 0.0005$; Arabic $p = 0.59$), and the gap between groups is not ($p = 0.64$). The last column shows `awq_clip` moving *away* from FP16 in all five languages.

		`awq_lite`			`awq_clip`
Group	**Cells**	**ΔAcc**	**ΔKL**	**ΔTop-1**	**ΔKL**
Arabic	20	+0.43	−0.0047	+1.60	+0.0064
non-Arabic	36	+0.43	−0.0090	+0.79	+0.0039
Arabic MSA	12	+0.08	−0.0015	+1.55	+0.0051
Gulf Arabic	8	+0.96	−0.0095	+1.69	+0.0083
English	12	+0.14	−0.0083	−0.25	+0.0090
Spanish	12	+0.25	−0.0051	+2.04	+0.0026
Russian	12	+0.90	−0.0138	+0.58	+0.0002

One caution follows from the same source: Marchisio et al. (2024) report that automatic metrics understate quantization damage against human judgement, and every number in Table 2 is automatic. "Arabic is the most robust" is thus a claim about multiple-choice accuracy, not about what an Arabic reader would notice, and the KL column warns that the understatement may be largest exactly where accuracy calls the language safest.

4.4 Arabic calibration transfers to every language

Arabic evaluations gain +0.43 pp and non-Arabic evaluations gain +0.43 pp (Table 3); the interaction between calibration language and evaluation language is −0.001 pp, with a Welch t 95 % confidence interval of $[-0.69, +0.69]$ pp (df $= 48.8$, $p = 0.999$; Table 3), which excludes any home advantage above 0.7 pp. The two largest per-language gains sit on opposite sides of the split: Gulf Arabic, in the calibration language, at +0.96 pp, and Russian, as far from it as this benchmark reaches, at +0.90 pp (Table 3, Figure 2). On KL the transfer is at least as strong as the home effect: non-Arabic evaluations move significantly closer to FP16 (-0.0090, $p = 0.0005$) while the Arabic improvement (-0.0047) is not significant on 20 cells, though that gap is not itself significant ($p = 0.64$). The largest single fidelity gain of any language is

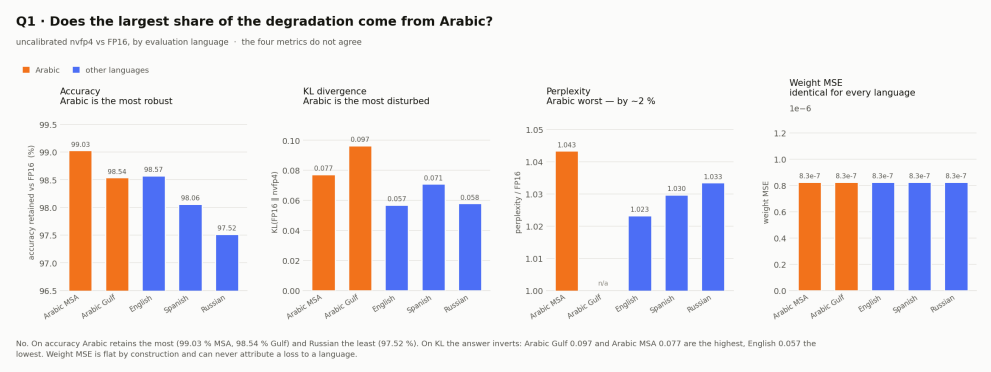

Figure 1: The four metrics by evaluation language, uncalibrated `nvfp4` against FP16 (orange: Arabic). Accuracy makes Arabic the most robust language, KL makes it the most disturbed, perplexity separates nothing, and weight MSE is flat by construction. Numbers in Table 2.

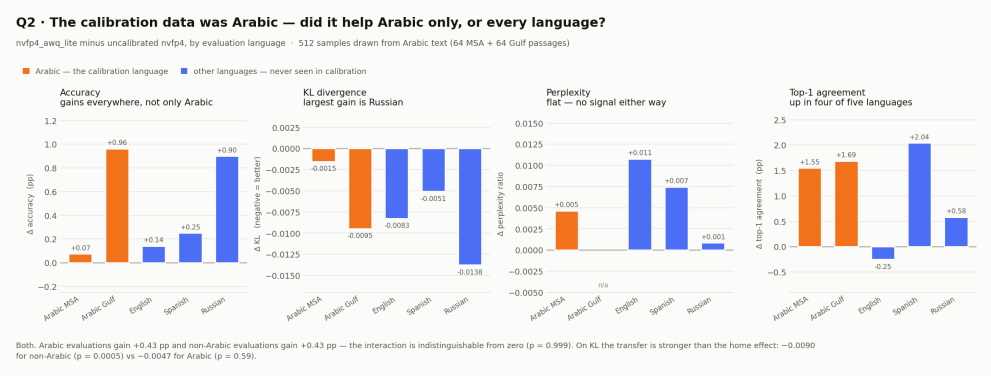

Figure 2: Change in each metric after Arabic calibration (`awq_lite` minus `nvfp4`), by evaluation language; orange is the calibration language. The gains are not concentrated on it, and the largest KL gain belongs to Russian. Numbers in Table 3.

again Russian (Table 3). No language regresses on accuracy (Table 3), and perplexity shifts by less than 0.011 in ratio terms everywhere, in the same direction, with no Arabic/non-Arabic pattern (Figure 2).

What the Arabic corpus buys `awq_lite` is therefore not Arabic knowledge but a better estimate of each weight column's activation scale, which is dominated by which channels are outliers; outlier channels are largely shared across languages (Dettmers et al., 2022; Xiao et al., 2023), so the correction generalises. `awq_clip` supports this by contrast: same Arabic text, no per-channel scale, and KL rises in all five evaluation languages (Table 3, last column). Our result is compatible with Chimoto et al. (2026), who find language-matched calibration best *per language* while multilingual mixes win on average: at W4A16 with a per-channel scale, we find the home advantage shrinks to zero on accuracy and reverses on fidelity, which suggests the size of any language-matching effect depends on how much of the estimator's capacity is language-specific.

4.5 Damage predicts repair

Llama-3.1-8B is the model quantization hurts most, on every metric that tracks task quality: it loses 2.26 pp — roughly three times the next-worst model — retains 96.14 %, agrees with its own FP16 output on only 75.4 % of positions, and carries the highest weight MSE (Table 4). It also owns all four cells that fall more than three points (Table 1), the worst of them −7.2 pp on English Global-MMLU. Two metrics dissent instructively: KL names Fanar-1-9B, whose divergence is highest at 0.0922 but which loses only 0.79 pp (Table 4), so a model can move far in distribution without crossing decision boundaries; and perplexity produces a third ordering again (§4.1).

The same model gains most from calibration: +1.11 pp and −0.0147 KL, more than double the next model on both, lifting its retention from 96.14 % (Table 4) to 98.03 %. Gemma-2-9B, barely damaged to begin with, gains nothing

Table 4: Per-model results. Left: damage from uncalibrated `nvfp4` relative to FP16. Right: what the Arabic calibration corpus recovers, as `awq_lite` minus `nvfp4`, with `awq_clip` for contrast. The ordering of repair follows the ordering of damage. Weight MSE is taken from the non-Arabic rows, where it is stored at full precision (Limitations).

	Damage: `nvfp4` vs FP16						**Repair: `awq_lite` vs `nvfp4`**		**`awq_clip` vs `nvfp4`**	
Model	FP16	ΔAcc	Ret.	KL	Top-1	MSE	ΔAcc	ΔKL	ΔAcc	ΔKL
Gemma-2-9B	63.49	−0.39	**99.38**	**0.0296**	**90.79**	3.2e-7	−0.04	−0.0033	+0.34	+0.0050
Fanar-1-9B	61.75	−0.79	98.72	0.0922	84.29	3.1e-7	+0.51	−0.0023	+0.26	+0.0101
ALLaM-7B	59.82	−0.62	98.96	0.0799	86.14	9.9e-7	+0.14	−0.0095	−0.07	−0.0002
Llama-3.1-8B	58.49	−2.26	96.14	0.0793	75.39	1.7e-6	+1.11	−0.0147	+0.09	+0.0045

(−0.04 pp; Table 4). Across the four models the ordering of benefit follows the ordering of damage one for one — −2.26 → +1.11, −0.79 → +0.51, −0.62 → +0.14, −0.39 → −0.04 (Table 4). Calibration is best understood as a repair whose size is bounded by the damage, not as a general upgrade — consistent with the AWQ account in which the per-channel scale exists to protect salient weight channels (Lin et al., 2024) and with the model-dependence of outlier structure (Dettmers et al., 2022). `awq_clip` recovers little on accuracy and raises KL on three of four models (Table 4).

4.6 Model choice depends on the target language

Over all five languages Gemma-2-9B leads at 63.06 %, 1.6 pp clear of Fanar-1-9B (Table 5); it is also the strongest model at FP16 (Table 4), so its 4-bit ranking is largely inherited rather than earned by quantization behaviour. On the Arabic slice the ordering flips: ALLaM-7B leads at 59.70 % with Fanar-1-9B 0.3 pp behind, and Gemma-2-9B falls to third — it leads non-Arabic by 3.6 pp and trails on Arabic by 2.3 pp (Table 5, Figure 3). Within Arabic the two Arabic-centric models split by register: Fanar-1-9B is ahead on MSA (60.00 against 59.73) and ALLaM-7B on Gulf dialect (59.65 against 58.50), a distinction a single pooled Arabic score would hide (Table 5).

Both Arabic-centric models finish at or above their own FP16 Arabic scores (101.2 % and 100.4 %) while Gemma-2-9B does not (98.9 %; Table 5); given a per-cell standard error near 1.5 pp(§3) we read this as no measurable Arabic loss rather than as quantization improving a model. Fidelity ranks the models differently again: Gemma-2-9B has the lowest KL and the lowest perplexity ratio on both slices (Table 5), so a fidelity-driven choice and an Arabic-accuracy-driven choice select different systems.

5 Discussion

Four practical conclusions follow. **Ship `awq_lite` for weight-only 4-bit**: it retains 99.04 % of FP16 accuracy, has the lowest KL, produces no catastrophic cell (Table 1), and should cost nothing extra at inference because its per-channel scale is folded into the weights rather than applied at runtime (Lin et al., 2024) — a structural consequence, not a timing we measured. **Do not default to `awq_clip`**: it is indistinguishable from no calibration on accuracy and degrades fidelity on three of four models (Tables 1, 4) despite consuming the same corpus — calibration data is only as useful as the estimator that consumes it. **Calibrate in proportion to expected damage**, since the benefit tracks the damage one for one (Table 4). **An Arabic corpus is a safe multilingual default**: grouped, non-Arabic gained the same average accuracy as Arabic, though per language the gains run +0.08 to +0.96 pp (Table 3).

For evaluation practice, report accuracy *and* a distributional metric and expect them to disagree: KL calls Arabic the most disturbed language where accuracy calls it the most robust (Table 2), and KL names the wrong model as most damaged (Table 4). Both quantities are real; only one is what a user experiences. Weight MSE should not be reported as a quality signal at all (§4.2).

6 Conclusion

Across 56 paired cells, NVFP4 W4A16 weights cost about one point of accuracy, and Arabic-calibrated AWQ with a per-channel scale recovers roughly half of it while eliminating every catastrophic cell (Table 1). The loss is not concentrated in Arabic — on accuracy Arabic is the most robust language and Russian the least, though KL orders them the other way (Table 2) — and an all-Arabic calibration corpus improves the languages it never

Table 5: Standings after quantization and Arabic calibration (`awq_lite` unless noted). "Ar. ret." is Arabic accuracy retention against the same model's FP16 Arabic score. The overall and Arabic rankings disagree.

Model	All 5	All 5 (`awq_clip`)	Arabic	Ar. ret.	Arabic MSA	Gulf	non-Arabic	KL	PPL
Gemma-2-9B	**63.06**	**63.44**	57.36	98.93	57.53	57.10	**66.23**	**0.026**	**1.012**
Fanar-1-9B	61.46	61.21	59.40	**101.23**	**60.00**	58.50	62.61	0.090	1.024
ALLaM-7B	59.34	59.13	**59.70**	100.44	59.73	**59.65**	59.14	0.070	1.084
Llama-3.1-8B	57.34	56.33	50.92	97.51	50.73	51.20	60.91	0.065	1.034

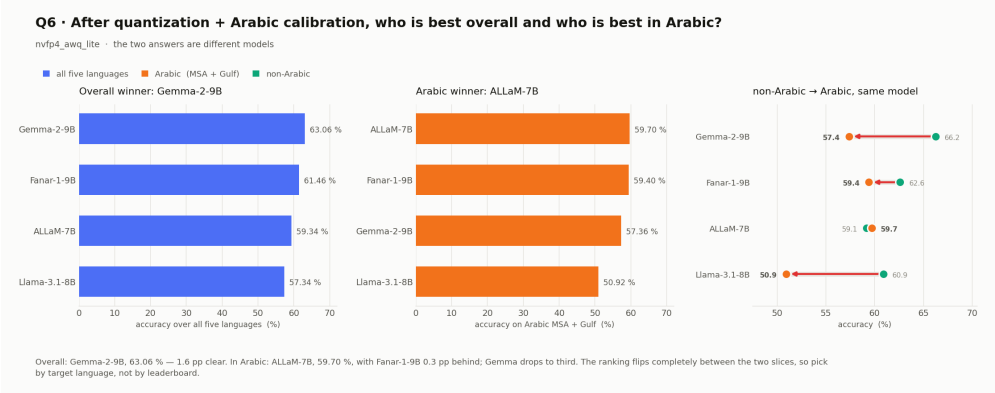

Figure 3: Standings under `awq_lite` over all five languages (left) and on Arabic only (centre); right, each model's move from its non-Arabic score to its Arabic score. The ranking is not preserved between the two slices. Numbers in Table 5.

contained as much as the one it did — but only when AWQ's per-channel scale is there to consume it (Table 3). The model quantization hurts most is the model calibration repairs most (Table 4), and the best multilingual model after quantization is not the best Arabic one (Table 5). Practitioners should choose the estimator deliberately, choose the model by target language, and never rank either by weight reconstruction error.

Limitations

Single run. One standard error on a single 1000-item cell is about 1.5 pp, so a difference of that size *within one cell* — the above-FP16 Arabic retention in Table 5 — is a tie. The rule does not extend to paired contrasts: over 56 matched cells +0.43 pp carries an SE of 0.18 pp and is significant ($p = 0.018$, Table 1), while per-language gains resting on 8–12 cells are not (Gulf +0.96, Russian +0.90; both $p = 0.07$). A new seed would change nothing — scoring is argmax and the quantizers are deterministic. The randomness that exists is the draw of the calibration sample and of each cell's 1000 items; bootstrapping over items and redrawing that sample would bound both at no GPU cost.

KL rests on a subsample, unevenly. KL and top-1 use 200 comparison rows per cell, and only 64 on six of them: ALLaM-7B and Fanar-1-9B on the three Arabic MSA datasets, in the calibrated variants. Six of the 20 Arabic cells therefore carry a noisier estimate — one reason the Arabic KL gain in Table 3 is under-powered.

Gulf perplexity is excluded. The artifact is confined to the calibrated variants: the Gulf ratio is 1.054 under `nvfp4`, in line with the 1.033 non-Gulf average, but 1.184 under both `awq_lite` and `awq_clip`, whose Gulf runs scored 10–15 % fewer tokens than the FP16 reference they divide by. The cause is undetermined: an early-EOS shift on dialect prompts and a scoring-side truncation both fit, and separating them needs per-item lengths we did not retain. Gulf perplexity is therefore omitted from Tables 1, 2 and 3. Accuracy, KL and top-1 are per-item and unaffected.

Gulf datasets are not public. The two dialectal sets are in-house, so the Gulf slice is never pooled with the four public-benchmark languages (Singh et al., 2025; Conneau et al., 2018; Lin et al., 2022).

One calibration corpus. We vary the calibration *language* but hold the corpus fixed at one 262k-token sample of Arabic conversational text, so "Arabic calibration transfers" may partly be "this corpus transfers"; separating them needs several corpora per language, as in Chimoto et al. (2026).

Weight MSE precision. The Arabic result files

stored weight MSE rounded to 10^{-6}; Table 4 takes the full-precision value from the non-Arabic rows of the same model, which is exact because the quantity is language-independent (§4.2).

Scope. Four models at 7–9B, one format, weights only. We make no claim about activation quantization, other model scales, other 4-bit formats such as MXFP4 (Cim et al., 2026), or generative quality beyond the four metrics of §3.

References

Eduardo Alvarez, Omri Almog, Eric Chung, Simon Layton, Dusan Stosic, Ronny Krashinsky, and Kyle Aubrey. 2025. Introducing NVFP4 for efficient and accurate low-precision inference. NVIDIA Technical Blog. Accessed: 2026-08-18.

M. Saiful Bari, Yazeed Alnumay, Norah A. Alzahrani, Nouf M. Alotaibi, and 1 others. 2024. ALLaM: Large language models for Arabic and English. *arXiv preprint arXiv:2407.15390*.

Everlyn Asiko Chimoto, Mostafa Elhoushi, and Bruce Bassett. 2026. Calibrating beyond English: Language diversity for better quantized multilingual LLMs. In *Proceedings of EACL*.

Musa Cim, Burak Topcu, and Mahmut Taylan Kandemir. 2026. Diagnosing FP4 inference: A layer-wise and block-wise sensitivity analysis of NVFP4 and MXFP4. In *ICLR Workshop on Scientific Methods for Understanding Deep Learning*.

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 EMNLP*.

Tim Dettmers, Mike Lewis, Younes Belkada, and 1 others. 2022. GPT3.int8(): 8-bit matrix multiplication for transformers at scale. *Advances in Neural Information Processing Systems*, 35:30318–30332.

Tim Dettmers and Luke Zettlemoyer. 2023. The case for 4-bit precision: k-bit inference scaling laws. In *Proceedings of ICML*, pages 7750–7774.

Fanar Team. 2025. Fanar: An Arabic-centric multimodal generative AI platform. *arXiv preprint arXiv:2501.13944*.

Elias Frantar, Saleh Ashkboos, Torsten Hoefler, and 1 others. 2023. GPTQ: Accurate post-training quantization for generative pre-trained transformers. In *Proceedings of ICLR*.

Gemma Team. 2024. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, and 1 others. 2024. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Itay Hubara, Yury Nahshan, Yair Hanani, and 1 others. 2021. Accurate post training quantization with small calibration sets. In *Proceedings of ICML*, pages 4466–4475.

Eldar Kurtic, Alexandre Marques, Shubhra Pandit, and 1 others. 2024. "Give me BF16 or give me death"? Accuracy-performance trade-offs in llm quantization. *arXiv preprint arXiv:2411.02355*.

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 on-device LLM compression and acceleration. In *Proceedings of Machine Learning and Systems (MLSys)*.

Xi Victoria Lin, Todor Mihaylov, Mikel Artetxe, and 1 others. 2022. Few-shot learning with multilingual generative language models. In *Proceedings of EMNLP*, pages 9019–9052.

Kelly Marchisio, Saurabh Dash, Hongyu Chen, and 1 others. 2024. How does quantization affect multilingual LLMs? In *Findings of EMNLP*, pages 15928–15947.

Anmol Mekala, Anirudh Atmakuru, Yixiao Song, and 1 others. 2025. Does quantization affect models' performance on long-context tasks? *arXiv preprint arXiv:2505.20276*.

Paulius Micikevicius, Dusan Stosic, Neil Burgess, and 1 others. 2022. FP8 formats for deep learning. *arXiv preprint arXiv:2209.05433*.

NVIDIA Corporation. 2025. NVIDIA TensorRT model optimizer. V0.45.0.

Shivalika Singh, Angelika Romanou, Clémentine Fourrier, and 1 others. 2025. Global MMLU: Understanding and addressing cultural and linguistic biases in multilingual evaluation. In *Proceedings of ACL*, pages 18761–18799.

Miles Williams and Nikolaos Aletras. 2023. On the impact of calibration data in post-training quantization and pruning. *arXiv preprint arXiv:2311.09755*.

Guangxuan Xiao, Ji Lin, Mickael Seznec, and 1 others. 2023. SmoothQuant: Accurate and efficient post-training quantization for large language models. In *Proceedings of ICML*.

Zhewei Yao, Xiaoxia Wu, Cheng Li, Stephen Youn, and Yuxiong He. 2024. Exploring post-training quantization in LLMs from comprehensive study to low rank compensation. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Xunyu Zhu, Jian Li, Yong Liu, and 1 others. 2024. A survey on model compression for large language models. *TACL*.