

# Lost in Compression: Calibration Bias and the Accent-WER Gap in Speech LLMs

Angad Ripudaman Singh Bajwa

Independent Researcher

angadisthere@gmail.com

## Abstract

GPTQ and similar calibration-based quantization methods tune their rounding to a small calibration sample, in practice usually convenient, native-accented read speech. Edge devices disproportionately run quantized models, so the speakers most likely to depend on the compressed model are often the ones calibration represented least. Compression is known to amplify bias in text LLMs and across languages in multilingual speech models, but whether it does so within a single language, between native and non-native accents, in a modern speech LLM, remains untested. We quantize the LLM backbone of Qwen2-Audio-7B-Instruct with GPTQ, hold the recipe fixed and vary only which EdAcc speakers fill the calibration set (native-only, balanced, L2-only), measuring the native-vs-L2 word error rate (WER) gap at matched overall WER. We expected native-only calibration to widen the gap and balanced calibration to close it. Instead, homogeneous calibration, native-only or L2-only, narrows the gap while balanced calibration does almost nothing. It recurs when we quantize the audio encoder, weakens at 3-bit and collapses at 2-bit.

## 1 Introduction

Post-training quantization (PTQ) methods such as GPTQ (Frantar et al., 2023) and AWQ (Lin et al., 2024) fit their scaling and channel-protection decisions to a small calibration sample, usually a few hundred examples. Whoever assembles that sample tends to reach for whatever is convenient, which usually skews toward majority, native-accented, read speech. That matters because resource-constrained and edge devices disproportionately run compressed models, so the group most likely to receive the compressed model is often the group calibration represented least.

Compression amplifying bias is already established for text LLMs (Marcuzzi et al., 2026; Rath and Maliakkal, 2026) and for language identity in

multilingual, multitask speech-model compression (Ferraz, 2024). Calibration-set composition is likewise known to affect PTQ and pruning accuracy in general (Williams and Aletras, 2024). What nobody has tested is a demographic axis within a single language, native versus non-native (L2) accent, in a speech LLM rather than a classical ASR system. Even if such a gap moves, it is worth asking why, whether it is the calibration population itself or just degradation from compression in general.

This paper does three things. First, we quantize Qwen2-Audio-7B-Instruct’s LLM backbone to 4-bit, 3-bit and 2-bit with GPTQ and track the native-vs-L2 WER gap on EdAcc (Sanabria et al., 2023), a corpus of accented conversational English. Second, holding the quantizer and the calibration size fixed, we swap only which speakers’ audio fills the calibration set (native-only, balanced, L2-only), so that matched overall WER across conditions serves as an aggregate-WER control: if overall WER does not move across conditions, whatever happens to the gap can be attributed to calibration composition rather than to one condition being compressed harder. Third, we characterise a small, directionally consistent effect at 4-bit and show that the same manipulation on the mechanistically distinct audio encoder points the same way, which we read as convergent evidence rather than a strict replication. The direction of the effect is the opposite of what we expected. That reversal is the central finding.

## 2 Related Work

The closest prior work is Ferraz (2024), who show that multitask, multilingual Whisper compression produces disparity across languages. We differ in modality (a speech LLM, not classical ASR), axis (demographic accent within one language, not language identity) and design (an explicit calibration-composition swap at matched overall WER, not a

compression-language correlation). [Marcuzzi et al. \(2026\)](#) and [Rath and Maliakkal \(2026\)](#) show quantization reshapes bias and alignment in text LLMs differently across precision levels, part of why we sweep bit-widths rather than testing one. [Williams and Aletras \(2024\)](#) shows calibration-data choice affects PTQ and pruning accuracy generally, supporting the mechanism we test, though without a fairness axis.

### 3 Method

#### 3.1 Model and quantization targets

Qwen2-Audio-7B-Instruct ([Chu et al., 2024](#)) has 8.40B parameters, which by our count split into an audio encoder (`audio_tower`, 7.6%), a small projector (0.1%), an LLM decoder (`language_model`, 84.7%) and an output head (`lm_head`, 7.6%). Our primary condition quantizes only `language_model` (the backbone) and freezes everything else at FP16. A mirror-image condition instead quantizes `audio_tower` and the projector, the encoder condition, freezing the LLM decoder and `lm_head`. Because the frozen components still run during calibration, the backbone Hessian is computed from real accent-conditioned post-encoder activations rather than any hand-rolled embedding capture, and the same is true in reverse for the encoder condition. We run GPTQ ([Frantar et al., 2023](#)) through the `llm-compressor` library ([AI and vLLM Project, 2024](#)) with symmetric int weights and a group size of 128, at W4/W3/W2 for the backbone and W4 for the encoder. The library has no named preset for 3-bit or 2-bit, so we built the scheme by hand, keeping the same shape as the W4 preset and changing only the bit-width. We also use the library’s non-sequential calibration pipeline instead of its default error-propagating one, because the default pipeline’s graph-tracing calibration path is not compatible with this model’s forward pass.

#### 3.2 Data and grouping

We evaluate on the validation split of EdAcc ([Sanabria et al., 2023](#)), the corpus’s own dev portion, 9,848 utterances from 62 speakers speaking conversational English with a range of accents. Every speaker carries a first-language field (11), either one of 8 English varieties (Southern British English, Indian English, Jamaican English and so on) or one of 21 non-English languages (Spanish, Mandarin, Arabic and others). The two never over-

lap, so every speaker falls unambiguously on one side. We call a speaker native if their first language is one of the English varieties (23 speakers, 3,706 utterances) and L2 if English came later (39 speakers, 6,142 utterances). These labels are about language background, not skill, so an L2 speaker here is not assumed to be less fluent, only accented by a first language other than English. This native/L2 split, and the gap in transcription accuracy between the two groups, is the central quantity the paper measures.

Before scoring we drop 191 utterances that would otherwise distort the gap. Of these, 62 carry EdAcc’s `IGNORE_TIME_SEGMENT_IN_SCORING` marker and 129 have a reference transcript that is only a non-speech tag such as `<LAUGH>`. Scoring either kind would count as a spurious error, and if such tags occur more for one group than the other that noise would land unevenly and distort the native/L2 gap itself rather than adding harmless noise equally to both sides.

#### 3.3 Calibration conditions

For each condition, native, balanced and L2, we select 128 calibration utterances, sampling round-robin across speakers within the eligible group so that no single speaker dominates the calibration Hessian. The native and L2 conditions draw all 128 from their group. Balanced splits them 64/64. Everything else about the GPTQ setup is identical across conditions. Because overall WER is matched across the three W4 (or W3) conditions, that match acts as an aggregate-WER control, so any resulting gap movement cannot be attributed to one condition being compressed more severely than another. At W4 this is not a degradation control, since overall WER sits at or below FP16 (0.331–0.332 vs 0.336), so the conditions are matched to each other rather than degraded from full precision. Matched aggregate WER need not mean matched per-group behaviour, so we also report group-wise WER below.

#### 3.4 Metrics

We report corpus-level WER per group, the total edit distance over total reference words rather than a mean of per-utterance rates. Two quantities carry the analysis. The **gap** is  $WER(L2) - WER(native)$ , how much worse the model transcribes L2 speakers than native speakers for a given model. The  $\Delta gap$  is  $gap(quantized) - gap$

gap(FP16), how much quantization changes that gap relative to full precision, so a negative  $\Delta\text{gap}$  means quantization narrowed the native-vs-L2 gap.

Confidence intervals come from a speaker-clustered bootstrap (2,000 resamples) that draws speakers with replacement rather than utterances, since utterance counts per speaker range from 27 to 434. The bootstrap is stratified by group. Each replicate draws the 23 native and the 39 L2 speakers with replacement separately, preserving each group’s size. Corpus-level edit distance and reference-word counts are then summed over the drawn speakers (a speaker drawn  $k$  times contributes  $k$  times). For  $\Delta\text{gap}$  and the direct condition contrasts we pair the resampling, applying the same speaker draw to both files, since both are scored on identical utterances. Pairing cancels speaker-specific noise common to both and gives a tighter estimate than treating them as independent.

The intervals do the inferential work here. The effects are small in absolute terms, the gap and its movement both on the order of a single WER point, comparable to the wobble a different draw of 62 speakers could produce on its own. A bare point estimate like  $\Delta\text{gap} = -0.010$  says nothing about how much it would shift under a different speaker sample, so throughout we treat an effect as real only when its 95% interval excludes zero rather than reading into the sign or size of a point estimate alone.

For reproducibility, GPTQ runs through the `llm-compressor` library (AI and vLLM Project, 2024) and WER is scored with `jiwer`. We report word error rate rather than the related MER or WIL measures (Morris et al., 2004). Decoding is greedy and the inference prompt is “Transcribe the English audio to text. Output only the transcription, with no punctuation or capitalization.”

## 4 Results

The FP16 baseline (Table 1) has overall WER 0.336 (native 0.330, L2 0.341), a small native-vs-L2 gap of +0.011 whose 95% CI  $[-0.038, +0.062]$  already crosses zero given only 23 native speakers. We therefore rely on the tighter paired  $\Delta\text{gap}$  rather than each condition’s raw gap CI for the headline comparisons. At W4 overall WER is tightly matched across the three backbone conditions (0.331–0.332) and the three encoder conditions (0.333–0.339), so the aggregate-WER control holds and any movement in the gap is not just one

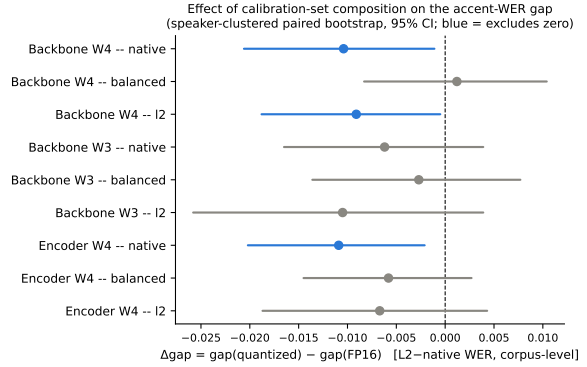


Figure 1:  $\Delta\text{gap}$  by calibration condition, target and bit-width. Blue intervals exclude zero.

condition being compressed harder. Figure 1 plots  $\Delta\text{gap}$  with its paired CI for every condition. Table 1 is the paper’s primary evidence. It reports the corpus-level gap and how much quantization moves it. Table 2 is diagnostic rather than a second headline result. It decomposes that same movement into each group’s own WER, native and L2 scored separately, to check what is driving the gap rather than just that it moved.

Two results stand out. At W4, native-only calibration narrows the gap relative to FP16 in both the backbone ( $-0.0104$ , CI  $[-0.0206, -0.0011]$ ) and the encoder ( $-0.0109$ , CI  $[-0.0202, -0.0021]$ ), the opposite of what we expected if the field’s default calibration population were what widens disparity. L2-only calibration moves the same way in the backbone ( $-0.0091$ , CI  $[-0.0188, -0.0005]$ ), with a similar but non-significant estimate in the encoder. Balanced calibration is the condition closest to no effect in both targets ( $+0.0012$  and  $-0.0058$ , both CIs crossing zero), which is not the fix we expected balancing to be. The pattern is not an artifact of one component, appearing in the same direction and roughly the same magnitude across two mechanistically different targets, a decoder stack and an acoustic encoder. Because those are different mechanisms rather than the same experiment re-run, we read the encoder result as convergent evidence for the backbone finding, not an independent replication. We treat the W4  $\Delta\text{gap}$  and the direct contrasts as primary analyses. The per-group breakdown and the macro-weighted check below are secondary evidence that the effect is not an artifact of corpus-level weighting or leveling-down.

**The narrowing is not leveling-down.** A gap can close two ways, the worse-off group improving or

| Target   | Condition | Bits | Overall WER | Gap (L2–native) | $\Delta$ gap vs. FP16 [95% CI] |
|----------|-----------|------|-------------|-----------------|--------------------------------|
| –        | FP16      | 16   | 0.336       | +0.011          | –                              |
| Backbone | native    | 4    | 0.332       | +0.000          | –0.0104 [–0.0206, –0.0011]     |
| Backbone | balanced  | 4    | 0.331       | +0.012          | +0.0012 [–0.0083, +0.0104]     |
| Backbone | L2        | 4    | 0.332       | +0.002          | –0.0091 [–0.0188, –0.0005]     |
| Backbone | native    | 3    | 0.339       | +0.005          | –0.0062 [–0.0165, +0.0039]     |
| Backbone | balanced  | 3    | 0.347       | +0.008          | –0.0027 [–0.0136, +0.0077]     |
| Backbone | L2        | 3    | 0.341       | +0.000          | –0.0105 [–0.0258, +0.0039]     |
| Backbone | native    | 2    | 1.913       | +0.131          | +0.1167 [–0.1516, +0.3708]     |
| Backbone | balanced  | 2    | 2.515       | +0.173          | +0.1587 [–0.2668, +0.5670]     |
| Backbone | L2        | 2    | 1.801       | +0.031          | +0.0129 [–0.2349, +0.2601]     |
| Encoder  | native    | 4    | 0.333       | +0.000          | –0.0109 [–0.0202, –0.0021]     |
| Encoder  | balanced  | 4    | 0.335       | +0.005          | –0.0058 [–0.0145, +0.0027]     |
| Encoder  | L2        | 4    | 0.339       | +0.004          | –0.0067 [–0.0187, +0.0043]     |

Table 1: All conditions. Gap and  $\Delta$ gap are corpus-level WER(L2)–WER(native); CIs are a paired, speaker-clustered bootstrap (2,000 resamples). W2 rows are reported for completeness but reflect uninterpretable transcription collapse, not a meaningful accent-specific effect.

the better-off group being dragged down to meet it. Table 1 shows only the net effect, gap(condition) minus gap(FP16), and cannot tell these two apart. Table 2 can. It gives each group’s own WER and its paired bootstrap change from FP16. At W4 the gap closes the good way: L2 WER falls (backbone native –0.0088, backbone L2 –0.0080, encoder native –0.0072, all 95% CIs excluding zero) while native WER does not significantly rise (all three  $\Delta$ Native CIs cross zero). Homogeneous calibration improves the worse-off group rather than equalising by regression. The one place both groups worsen together is W3 balanced, consistent with the shakier W3 aggregate-WER control.

**Direct contrasts between conditions.** Testing each condition against FP16 leaves the key comparison implicit, so we also contrast the conditions directly as gap(A) minus gap(B) under the same paired bootstrap. At W4 backbone each homogeneous condition differs significantly from balanced. Native calibration narrows the gap by 0.0119 more than balanced (CI [–0.0199, –0.0042]) and L2 calibration by 0.0102 more than balanced (CI [–0.0182, –0.0030]), while native and L2 do not differ from each other (0.0015, CI crossing zero). What moves the gap is calibration homogeneity, not which population it is homogeneous in. At W3 and on the encoder no direct contrast clears zero, consistent with the effect weakening away from the W4 backbone. The direction is also unchanged under equal-weight-per-speaker (macro) weighting, where the FP16 macro gap (+0.0237) is about twice the corpus gap (+0.0110), so the per-speaker disparity is larger than the word-weighted headline suggests.

**Robustness across calibration draws.** Everything above rests on one calibration draw. We re-drew the 128-utterance set under two more seeds and re-ran GPTQ at W4 backbone. The narrowing direction holds: both homogeneous-vs-FP16  $\Delta$ gaps stay negative in all three draws (means –0.0077 for native, –0.0060 for L2). Its strength does not hold. Each is significant in only one draw of three, and the direct homogeneity-vs-balanced contrast, significant in two draws, reverses sign in the third. With only 23 native speakers the movement is about the size of the draw-to-draw wobble it rides on. We read this as directional consistency without draw-level robustness, which is the main reason we frame the paper as a case study.

At W3 the same direction holds (all three point estimates stay negative) but no condition’s  $\Delta$ gap CI excludes zero, and the control is shakier (overall WER 0.339–0.347, balanced the worst). At W2 all three conditions fall apart: overall WER climbs past 1.0, meaning more edits than reference words, and the transcripts are incoherent, partly non-English output unrelated to the audio (a reference of ‘BYE’ decoded as multilingual noise). That uniform collapse has nothing to do with calibration composition, so we read nothing causal into its gap numbers.

## 5 Discussion

We frame this as a case study rather than a general causal claim. Within this one model, quantizer and recipe, calibration-set composition moves the accent-WER gap at matched overall WER, in the same direction across two different quantized components. What we do not claim is that the direction,

| Target   | Cond     | Bits | Native WER | L2 WER | $\Delta$ Native vs. FP16 [95% CI] | $\Delta$ L2 vs. FP16 [95% CI] |
|----------|----------|------|------------|--------|-----------------------------------|-------------------------------|
| –        | FP16     | 16   | 0.3297     | 0.3407 | –                                 | –                             |
| Backbone | native   | 4    | 0.3316     | 0.3319 | +0.0019 [−0.0056, +0.0091]        | −0.0088 [−0.0159, −0.0020]    |
| Backbone | balanced | 4    | 0.3232     | 0.3354 | −0.0065 [−0.0125, −0.0009]        | −0.0053 [−0.0133, +0.0016]    |
| Backbone | l2       | 4    | 0.3307     | 0.3327 | +0.0010 [−0.0038, +0.0066]        | −0.0080 [−0.0160, −0.0015]    |
| Backbone | native   | 3    | 0.3357     | 0.3406 | +0.0060 [−0.0021, +0.0152]        | −0.0001 [−0.0075, +0.0060]    |
| Backbone | balanced | 3    | 0.3425     | 0.3505 | +0.0128 [+0.0051, +0.0197]        | +0.0098 [+0.0022, +0.0172]    |
| Backbone | l2       | 3    | 0.3406     | 0.3409 | +0.0109 [−0.0015, +0.0244]        | +0.0002 [−0.0085, +0.0079]    |
| Encoder  | native   | 4    | 0.3334     | 0.3335 | +0.0037 [−0.0029, +0.0102]        | −0.0072 [−0.0144, −0.0016]    |
| Encoder  | balanced | 4    | 0.3323     | 0.3376 | +0.0026 [−0.0038, +0.0087]        | −0.0031 [−0.0099, +0.0030]    |
| Encoder  | l2       | 4    | 0.3366     | 0.3409 | +0.0069 [−0.0022, +0.0170]        | +0.0002 [−0.0066, +0.0060]    |

Table 2: Group-wise WER by condition, with each group’s own paired speaker-clustered bootstrap change from FP16 (2,000 resamples). A gap that narrows via rising native WER ( $\Delta$ Native > 0) is leveling-down. A gap that narrows via falling L2 WER ( $\Delta$ L2 < 0) is a genuine L2 improvement. At W4,  $\Delta$ L2 is negative with a CI excluding zero in three conditions while the paired  $\Delta$ Native CI crosses zero in all of them.

or a clean composition-only attribution, generalises beyond this setting, or that we have identified the mechanism. Our best guess is that a homogeneous calibration set, whichever population it draws from, produces lower-variance per-layer activation statistics and therefore more consistent GPTQ rounding, while a mixed set splits the difference between two distributions and specialises in neither. We present that as a hypothesis to test, not something we have shown.

The multi-seed check sharpens this caution. Across three draws the narrowing direction is consistent, but its magnitude is on the order of the variation a fresh draw introduces. The test that most directly encodes “homogeneity beats balancing” reverses in one draw of three. The practical takeaway is not to calibrate on a single group. It is a caution in both directions, that calibrating on the majority is not a neutral choice and that balancing is not a dependable fix.

The most motivated next step is covariate-matched calibration sets that hold speaker count, duration and word budget fixed, separating composition from draw-to-draw wobble. A mechanistic probe of per-layer activation variance would test the lower-variance hypothesis. Extending past one model, quantizer and dataset remains the broader open question.

## Limitations

Our evidence comes from a single setting, one model (Qwen2-Audio-7B-Instruct), one quantizer (GPTQ) and one calibration size of 128 examples. The narrowing direction is stable across three calibration draws, but individual-seed significance is not and the homogeneity-vs-balanced

contrast reverses in one draw, so the effect may not generalise. EdAcc’s 62 speakers are a convenience sample where native/L2 grouping may confound with recording channel, session and style, and with only 23 native speakers the per-condition CIs are wide, so headline claims use the paired  $\Delta$ gap. Two design points remain. The calibration utterances are drawn from the same split used for evaluation, but the overlap is benign, since removing each condition’s 128 calibration utterances moves the W4 backbone gap by at most 0.0021. Speaker-level overlap remains by design, and a fully speaker-disjoint split is future work. Because only 23 native speakers exist, the conditions are also not matched on covariates beyond native/L2 composition (unique speakers 23/62/39, duration 733/930/911 s, words 2209/2438/2370), so covariate-matched calibration sets are the most important next step.

## Ethics Statement

We study accent-based transcription disparity under quantization to surface risk, not to prescribe fixes. Balanced calibration is not a validated mitigation here.

## References

- Red Hat AI and vLLM Project. 2024. [LLM Compressor](#).
- Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhifang Guo, Yichong Leng, Yuanjun Lv, Jinzheng He, Junyang Lin, Chang Zhou, and Jingren Zhou. 2024. [Qwen2-Audio technical report](#). *Preprint*, arXiv:2407.10759.
- Thomas Palmeira Ferraz. 2024. [Efficient compression](#)

of multitask multilingual speech models. *Preprint*, arXiv:2405.00966. Master’s thesis.

Elias Frantar, Saleh Ashkboos, Torsten Hoefer, and Dan Alistarh. 2023. GPTQ: Accurate post-training quantization for generative pre-trained transformers. In *International Conference on Learning Representations (ICLR)*.

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)*.

Federico Marcuzzi, Xuefei Ning, Roy Schwartz, and Iryna Gurevych. 2026. [How quantization shapes bias in large language models](#). In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 363–404, Rabat, Morocco. Association for Computational Linguistics.

Andrew Morris, Viktoria Maier, and Phil Green. 2004. [From WER and RIL to MER and WIL: improved evaluation measures for connected speech recognition](#). In *Proc. Interspeech*.

Plawan Kumar Rath and Rahul Maliakkal. 2026. Quantization undoes alignment: Bias emergence in compressed LLMs across models and precision levels. In *2026 IEEE Cloud Summit*. IEEE.

Ramon Sanabria, Nikolay Bogoychev, Nina Markl, Andrea Carmantini, Ondrej Klejch, and Peter Bell. 2023. [The Edinburgh international accents of English corpus: Towards the democratization of English ASR](#). In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.

Miles Williams and Nikolaos Aletras. 2024. [On the impact of calibration data in post-training quantization and pruning](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10100–10118, Bangkok, Thailand. Association for Computational Linguistics.