TEXT SCORES DO NOT ESTABLISH PERFORMANCE ON LEXICALLY NON-DIAGNOSTIC SPEECH TASKS: A QWEN2-AUDIO QUANTIZATION CASE STUDY

Anonymous authors
Paper under double-blind review

ABSTRACT

Text-output scores alone do not show whether quantization preserves performance when the transcript does not reveal the target label. We evaluate fixed 6- and 7-bit Qwen2-Audio-7B-Instruct allocations on 508 English-to-German FLEURS utterances and on 512 RAVDESS emotion clips from 16 speakers. The provenance-bound FLEURS rerun has selected-minus-FP16 BLEU and chrF intervals that all include zero. On RAVDESS, where the same two sentences occur equally often with every emotion label, canonical-label accuracy is -3.71% for 6 bit and -1.17% for 7 bit relative to FP16. The 6-bit speaker interval excludes zero and an exact two-sided sign-flip test gives $p = 0.0148$; the 7-bit interval includes zero. Same-budget controls do not identify either selected allocation as best. This case study shows that a translation score does not establish performance on a speech task whose target cannot be recovered from the transcript.

1 INTRODUCTION

A good transcript does not establish performance on an emotion task after quantization. Two recordings can contain the same words but carry different emotion labels. The transcript therefore cannot determine the label. We call such a task *lexically non-diagnostic*. A model may preserve recognition or translation quality while its emotion accuracy changes. This task property does not identify which acoustic cues the model uses.

Post-training quantization (PTQ) studies commonly report text-task quality alongside memory or speed measurements (Dettmers et al., 2022; Frantar et al., 2022; Xiao et al., 2022; Lin et al., 2023). Recent speech-model work extends mixed-precision PTQ and compression to speech recognition, speech translation, and speech language models (Kang & Kim, 2025; Gu et al., 2025; Moslem, 2025; Liu et al., 2026; Palomino, 2026). Word error rate measures transcript accuracy, and translation scores measure text output. Neither evaluates a target label that cannot be recovered from the transcript.

We ask: *what evidence is needed before text-output scores are used to describe quantized-model performance on speech tasks beyond the transcript?* We report translation scores, emotion recognition on a task whose labels are not recoverable from the transcript, an FP16 baseline, and allocation controls with matched bit budgets. Figure 1 shows why these evaluations answer different questions.

We apply this protocol to Qwen2-Audio (Chu et al., 2024). The two selected allocation files were fixed using a disjoint speech-translation development split. We evaluate their complete, hash-bound FLEURS rerun and a seven-arm RAVDESS comparison with FP16, front-block, and stratified-uniform allocations at the same average bit budgets. The selected 6-bit allocation is below FP16 on RAVDESS, although its FLEURS translation intervals include zero. The 7-bit comparison remains unresolved against FP16. These observations make the endpoint boundary visible without asserting a general relation between translation and emotion recognition.

Contributions. This paper makes three contributions.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

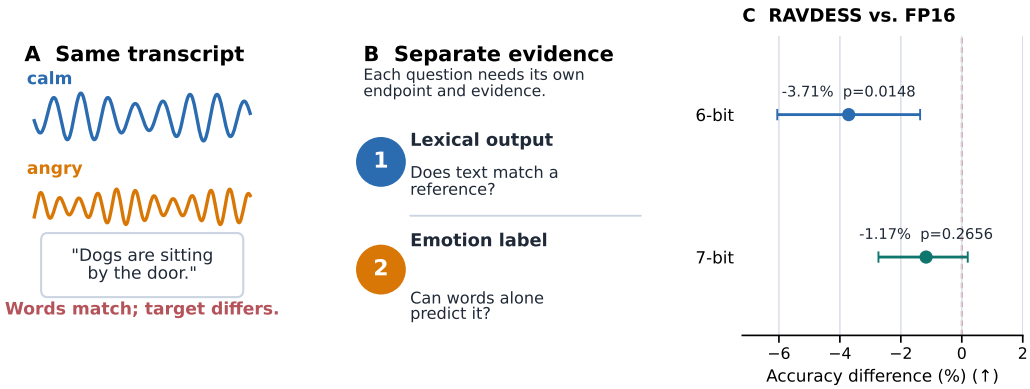


Figure 1: Why quantized speech models need more than a transcript score. **A:** Schematic waveforms illustrate two recordings with the same words but different emotion labels. **B:** Translation scores do not measure performance on labels that the transcript cannot determine. **C:** In the complete canonical-label RAVDESS comparison, the selected 6-bit allocation is below FP16 by -3.71% (95% speaker interval $[-6.05, -1.37]\%$; exact two-sided sign-flip $p = 0.0148$); the 7-bit interval includes zero.

1. We state an evaluation protocol that separates text output from performance on a task whose labels cannot be recovered from the transcript. The protocol reports an FP16 baseline, matched bit budgets, and uncertainty at the independent-unit level.

2. We provide a Qwen2-Audio case study with a provenance-bound FLEURS rerun and the same serialized 6- and 7-bit allocation files on RAVDESS, with complete matched-bit-budget controls.

3. We find no resolved FLEURS difference from FP16 for either selected allocation. On RAVDESS, the canonical-label 6-bit difference from FP16 is -3.71% with a speaker interval excluding zero and exact sign-flip $p = 0.0148$; the 7-bit interval includes zero. Neither selected allocation leads both matched controls.

2 Related Work

Low-bit language and speech models. LLM.int8(), GPTQ, SmoothQuant, and activation-aware weight quantization establish widely used PTQ approaches for large language models (Dettmers et al., 2022; Frantar et al., 2022; Xiao et al., 2022; Lin et al., 2023). Speech-specific work studies mixed-precision PTQ for automatic speech recognition, model compression for speech translation, selected-layer codec compression, and PTQ for speech language models (Kang & Kim, 2025; Gu et al., 2025; Moslem, 2025; Palomino, 2026; Liu et al., 2026). GenPTQ studies mixed-precision ASR allocation and lexical recognition, whereas we audit fixed Qwen2-Audio allocations across text output and a transcript-nondiagnostic label endpoint without proposing an allocator.

Evaluating audio use. Audio-language evaluation distinguishes direct audio use from text-mediated answers and studies temporal grounding and token-level readout (Zhang et al., 2026; Dai et al., 2026; Kulkarni et al., 2026a; Ma et al., 2026; Kulkarni et al., 2026b). LISTEN intervenes on aligned and conflicting lexical and acoustic emotion cues across six audio-language models (Chen et al., 2026). Our balanced-sentence design instead holds empirical label frequencies equal across two sentences; it cannot isolate acoustic cues. We ask whether text-output evidence transfers to fixed, quantized Qwen2-Audio allocations evaluated against FP16 and same-budget controls, without identifying cue use or proposing a new audio-use benchmark.

3 EVALUATION PROTOCOL

How to read the study counts. The study has two test populations. FLEURS contributes 508 ordered English-to-German utterances. Each of FP16, selected 6 bit, and selected 7 bit supplies one output for every identity. RAVDESS contributes 512 clips from actors 1–16, with 32 clips from every speaker. The seven RAVDESS arms share this frozen list: FP16, selected, front-block, and stratified-uniform allocations at 6 and 7 bits. An *arm* is one allocation applied to a complete test list. The number of arms is therefore not a count of people or independently sampled datasets. Accuracy counts clips; uncertainty resamples speakers because clips from one actor are related. Appendix A records the analysis units and stored-output bindings.

3.1 TWO DIFFERENT QUESTIONS

Text-output quality. Automatic speech recognition (ASR) and speech translation (ST) evaluate lexical output. This study reports BLEU and chrF for English-to-German speech translation. These scores describe the generated text; they do not evaluate an emotion label that the transcript cannot determine.

A task whose label the transcript cannot determine. We report emotion recognition on a test set where the same two sentences occur with each of eight emotion labels. The words therefore do not determine the target, but this property does not establish that the model relies on acoustic cues; that would require transcript-only input or controlled audio perturbations. All low-bit alternatives use matched parameter-weighted bit budgets. We compare the translation-selected plan with front-block and stratified-uniform allocations that change which modules receive each bit width while holding the budget nearly fixed. The experiment applies dequantized per-channel round-to-nearest weights and does not measure a packed implementation or runtime efficiency.

What sentence balance establishes. Within the evaluated RAVDESS subset, both spoken sentences appear equally often with each of the eight labels. A rule that receives only the sentence identity therefore sees the same empirical label distribution for either sentence, and its best sentence-only accuracy is 12.5% on this balanced set. This is a property of the test design, not a measured transcript-only model baseline. Emotion accuracy above 12.5% shows that the evaluated system uses information beyond the identity of these two sentences, but the available outputs cannot identify whether that information came from acoustic emotion cues, speaker or recording characteristics, or another route. Figure 1 depicts this distinction. We consequently describe the task as lexically non-diagnostic and describe the measured outcome as emotion-label accuracy; we do not call it a causal test of waveform reliance.

Three comparisons with different meanings. The FP16 comparison measures the accuracy change for a specific quantized allocation under a fixed task and prompt. A same-budget comparison asks whether the placement of low and high precision matters beyond its average bit width; matching that average does not match each module's quantization error or guarantee equal activation behavior. The cross-task comparison asks whether a favorable translation result is sufficient to infer preserved performance on a task whose target is absent from the transcript. Its two tasks have different datasets, prompts, metrics, and output units, so their score differences cannot be subtracted or treated as a common effect size. The design supports a failure of inference from one endpoint to the other when the observed results disagree. It does not estimate the effect of changing the task while holding inputs or labels fixed.

3.2 QWEN2-AUDIO CASE STUDY

Model and allocation files. We use Qwen2-Audio-7B-Instruct at revision `0a095220c30b`. The selected 6- and 7-bit allocations were fixed using a separate 387-example speech-translation development split before either test evaluation. We neither re-estimate that workflow nor use RAVDESS to choose an allocation. Each plan starts modules at 4 bits, fixes the output head at 8 bits, and upgrades whole modules to 8 bits subject to its parameter-weighted budget.

Matched controls. The front-block control visits layer blocks in model order. The stratified-uniform control visits block midpoints recursively across depth. The controls use the same 4/8-bit vocabulary as the selected files, but place high precision in different modules. The selected 6- and 7-bit files average 5.996166 and 6.996018 bits, respectively; their matched controls average 5.997935 and 6.994249 bits. The largest selected-control gap is 0.001769 bits. Thus a comparison with a control asks about the placement of precision at effectively fixed average precision, not simply about lowering precision.

Translation task and scoring. The FLEURS list contains 508 held-out English-to-German speech-translation examples (Conneau et al., 2022; Seamless Communication et al., 2023). We report SacreBLEU BLEU as the primary translation score and chrF as a secondary score. Both are computed from stored references and predictions with SacreBLEU 2.5.1 (Post, 2018; Popovič, 2015). The BLEU signature is `nrefs:1|case:mixed|eff:no|tok:13a|smooth:exp|version:2.5.1`. The chrF signature is

`nrefs:1|case:mixed|eff:yes|nc:6|nw:0|space:no|version:2.5.1`.

An earlier project archive contains additional FLEURS outputs without sufficient execution provenance. It is excluded from every table, figure, and result in this paper. The reported three-arm rerun records the manifest, serialized allocation files, raw predictions, scorer, and output hashes. This boundary makes the numerical evidence auditable, though it does not make the rerun a measurement of runtime efficiency.

RAVDESS emotion evaluation. RAVDESS contains acted speech recordings with eight emotion labels (Livingstone & Russo, 2018). We evaluate 512 clips from actors 1–16, with four clips for every actor–emotion pair. Each of the two spoken sentences occurs 32 times for every emotion, so sentence identity is non-diagnostic within this set. The prompt asks the model to choose exactly one of neutral, calm, happy, sad, angry, fearful, disgust, or surprised. Canonical scoring lowercases output, removes periods and commas, and maps the unambiguous morphological form "disgusted" to "disgust" before requiring exactly one listed label. All arms use greedy decoding, deterministic operations, seed 1729, and an eight-token limit. This balance does not rule out non-acoustic information outside the transcript or prove causal use of audio.

FLEURS paired intervals use 2,000 utterance bootstrap resamples with seed 1729. RAVDESS intervals use an equal-weight bootstrap over 16 speakers with 10,000 resamples. We additionally enumerate all 2^{16} sign assignments of the paired speaker means for a two-sided exact sign-flip test. This test is valid under the sharp null that a paired speaker-level difference is sign-exchangeable. We report its p values as a complement to the descriptive speaker bootstrap, not as a substitute for the interval or as a multiplicity-adjusted result.

Paired estimands and analysis units. Every reported difference has the order selected allocation minus a named comparator. For FLEURS, each bootstrap draw selects utterance identities with replacement and retains the paired prediction–reference records before recomputing BLEU and chrF. These intervals describe variation over the 508 stored outputs, conditional on this scoring procedure; they do not account for a new generation run. For RAVDESS, each draw selects the 16 speakers with replacement and retains all clips for a selected speaker. The resulting interval avoids treating repeated recordings from an actor as independent. Both intervals use the 2.5th and 97.5th percentiles. An interval including zero leaves that named comparison unresolved; it does not establish equal performance.

Provenance and multiplicity. The exact serialized files define the 6- and 7-bit interventions; verbal allocation descriptions do not. The paper binds the complete FLEURS and RAVDESS outputs to their manifests, plans, scorer, and analysis code. Two selected budgets and four matched-control comparisons do not form a family with a shared multiplicity adjustment. Therefore, a reported interval and sign-flip value describe their named pair of stored outputs. They are not a familywise guarantee or evidence that another selection procedure would choose the same allocation.

4 RESULTS

4.1 COMPLETE FLEURS OUTPUTS ARE NEAR FP16

Table 1 lists scores from the complete FLEURS rerun and RAVDESS canonical-label accuracy for the same selected files. The task outputs are generated separately because the prompts and targets differ. The table does not treat BLEU, chrF, and emotion accuracy as a common scale.

Table 1: Complete FLEURS rerun translation scores and RAVDESS canonical-label emotion accuracy for the listed allocation files. BLEU and chrF use SacreBLEU 2.5.1; RAVDESS shows exact-match accuracy (%). Reported differences use unrounded scores. The task outputs were generated separately using the same serialized allocation file. Shading identifies the translation-selected allocations.

Allocation	BLEU ↑	chrF ↑	Accuracy (%) ↑
FP16	20.36	49.73	70.12
Translation-selected, 6 bit	20.22	49.60	66.41
Translation-selected, 7 bit	20.33	49.68	68.95

For 6 bits, selected minus FP16 is -0.15 BLEU (paired 95% interval $[-0.51, +0.14]$) and -0.13 chrF ($[-0.41, +0.10]$). For 7 bits, it is -0.04 BLEU ($[-0.38, +0.26]$) and -0.05 chrF ($[-0.33, +0.19]$). Every FLEURS interval includes zero. Figure 2 and Table 2 show these results alongside the endpoint-specific RAVDESS comparisons. The intervals do not show that the outputs are interchangeable with FP16; they leave the measured differences unresolved on this fixed test list.

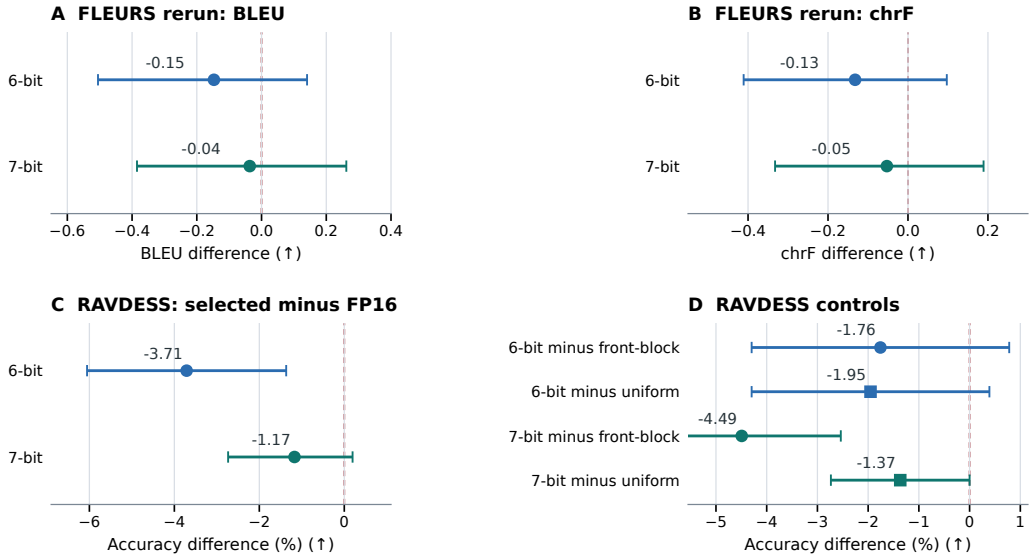


Figure 2: Selected allocation minus comparator with 95% intervals. Panels A and B show the complete FLEURS rerun, using paired utterance resampling (508 examples; 2,000 resamples). Panels C and D show canonical-label RAVDESS accuracy differences from FP16 and same-budget controls, using speaker-cluster resampling (16 speakers; 10,000 resamples). A negative value favors the comparator.

4.2 RAVDESS IDENTIFIES A 6-BIT FP16 DEFICIT

The RAVDESS columns in Table 1 use the same selected files as the FLEURS rerun. Figure 2 shows the complete seven-arm comparison. At 6 bits, the selected allocation reaches 66.41% accuracy. Selected minus FP16 is -3.71% (95% speaker interval $[-6.05, -1.37]\%$; two-sided exact sign-flip

Table 2: Later FLEURS rerun: selected allocation minus FP16 score differences. Paired 95% bootstrap intervals resample the 508 utterances 2,000 times with seed 1729. All intervals include zero.

Allocation	**ΔBLEU ↑ (95% CI)**	**ΔchrF ↑ (95% CI)**
6 bit	-0.15 $[-0.51, +0.14]$	-0.13 $[-0.41, +0.10]$
7 bit	-0.04 $[-0.38, +0.26]$	-0.05 $[-0.33, +0.19]$

$p = 0.0148$). At 7 bits, selected accuracy is 68.95%, with -1.17% versus FP16 ($[-2.73, +0.20]\%$; $p = 0.2656$). Thus the 6-bit comparison identifies a lower score for this allocation and endpoint, while the 7-bit comparison is unresolved.

What the controls show. At 6 bits, the selected allocation is -1.76% relative to front-block ($[-4.30, +0.78]\%$; $p = 0.2754$) and -1.95% relative to uniform ($[-4.30, +0.39]\%$; $p = 0.1836$). At 7 bits, it is -4.49% relative to front-block ($[-6.45, -2.54]\%$; $p = 0.0020$) and -1.37% relative to uniform ($[-2.73, 0.00]\%$; $p = 0.1465$). The selected plan has a lower point estimate than both controls at both budgets. Only the 7-bit front-block comparison has an interval that excludes zero. The controls therefore reject a reading in which the selected allocation is evidently best at its average bit budget, while leaving several control differences unresolved. The selected files were fixed before RAVDESS evaluation.

4.3 Speaker-level decomposition

The RAVDESS outcome has repeated clips from each actor, so the observed clip count should not be read as 512 independent pieces of evidence. Table 3 gives the complete per-speaker accuracies for FP16 and the two selected allocations. The 6-bit difference is negative for 12 actors, zero for two, and positive for two. The 7-bit difference is negative for five actors, zero for eight, and positive for three. This table is descriptive: it is included to expose the quantities resampled by the speaker bootstrap and sign-flip analysis, not to introduce an additional selection procedure.

The exact sign-flip calculation uses one paired mean difference per actor. Under its sharp null, each actor-level difference may be multiplied by either $+1$ or -1 with equal probability. Enumerating the $65{,}536$ possible signs gives $p = 0.0148$ for 6-bit selected minus FP16 and $p = 0.2656$ for 7-bit selected minus FP16. This calculation does not assume that clips are independent, but it does rely on sign-exchangeability of the actor-level differences under the null. The speaker bootstrap and exact test agree on the qualitative reading of the two FP16 contrasts: a resolved 6-bit deficit and an unresolved 7-bit difference.

4.4 Endpoint-specific interpretation

The FLEURS and RAVDESS comparisons share selected allocation files but not inputs, reference types, prompts, or metric scales. A FLEURS utterance has an English speech input and a German text reference; BLEU and chrF aggregate properties of generated translations. A RAVDESS clip has an emotion label and a constrained eight-label output. Its accuracy is a discrete agreement score. The observed numbers therefore cannot be averaged, rank-normalized, or turned into a single cross-task effect. Their useful relationship is logical: an evaluation on text output does not supply the missing outcome for an endpoint whose label is not determined by text.

The two datasets also support different forms of uncertainty. FLEURS intervals resample utterance identities because the translated output is scored at the utterance level. RAVDESS intervals resample speakers because the endpoint is measured repeatedly within actor. Treating every RAVDESS clip as an independent resampling unit would produce a more optimistic interval by ignoring actor clustering. Conversely, treating the 16 actors as a large population sample would overstate scope: the experiment remains an acted-emotion study with one checkpoint and a fixed prompt. The paper reports both the interval and exact finite-cluster calculation so the numerical conclusion is transparent about these choices.

The selected allocations were chosen on a speech-translation development split, so RAVDESS was not used to tune their bit placements. This separation makes the RAVDESS comparison a test of

Table 3: Canonical-label RAVDESS accuracy by actor for the complete FP16 and selected-allocation outputs. Each actor contributes 32 clips. The final two columns are selected minus FP16 accuracy differences for that actor. These actor-level quantities are the units used by the speaker bootstrap and exact sign-flip analysis.

Actor	FP16 (%) ↑	Selected 6 (%) ↑	Selected 7 (%) ↑	6 − FP16 (%) ↑	7 − FP16 (%) ↑
01	56.25	53.12	53.12	−3.12	−3.12
02	75.00	65.62	75.00	−9.38	0.00
03	84.38	78.12	84.38	−6.25	0.00
04	68.75	62.50	68.75	−6.25	0.00
05	71.88	68.75	71.88	−3.12	0.00
06	65.62	62.50	65.62	−3.12	0.00
07	81.25	78.12	84.38	−3.12	+3.12
08	68.75	56.25	62.50	−12.50	−6.25
09	65.62	65.62	62.50	0.00	−3.12
10	59.38	62.50	62.50	+3.12	+3.12
11	75.00	81.25	75.00	+6.25	0.00
12	71.88	59.38	62.50	−12.50	−9.38
13	65.62	62.50	62.50	−3.12	−3.12
14	68.75	68.75	68.75	0.00	0.00
15	68.75	65.62	68.75	−3.12	0.00
16	75.00	71.88	75.00	−3.12	0.00

the transferred allocation files. It does not make the selected plans universally translation-optimal, or give a causal explanation for their RAVDESS behavior. The matched controls answer a narrower question: whether the particular placement selected by the development workflow has better observed RAVDESS accuracy than other placements under nearly the same average bit budget. Their adverse point estimates are informative even where their intervals include zero, because an unresolved comparison is not support for superiority.

4.5 Diagnostic checks and decision rules

Several deterministic checks constrain the interpretation before uncertainty is calculated. First, each complete output file must cover the 512 manifest indices exactly once; an arm with a missing prediction or an extra index is excluded instead of being scored on an available subset. Second, the manifest is balanced by actor and emotion: each of eight labels appears four times per actor, and each sentence has the same empirical label distribution. Third, the canonical label function is fixed before comparing arms. It accepts one prompted label after lowercasing and punctuation removal, with the single morphological mapping from "disgusted" to "disgust". Applying this rule uniformly prevents a model whose outputs use this ordinary English variant from being penalized differently across conditions.

The RAVDESS comparisons use three decision rules with separate purposes. An FP16 difference asks whether one specified quantized allocation changes accuracy relative to its full-precision reference. A matched-control difference asks whether the selected placement has a better observed outcome than a different placement at effectively identical average precision. The cross-endpoint reading asks whether a translation output is sufficient evidence for an outcome that its reference text cannot determine. The last is not a numerical hypothesis test: it follows from the mismatch between what a translation metric evaluates and what an emotion label measures.

We call a comparison resolved in the direction of its point estimate only when its speaker bootstrap interval excludes zero. The exact sign-flip value is reported for the same speaker-level contrast to expose its finite-cluster support. We do not use an interval including zero to label allocations equivalent, and we do not use a positive point difference to rank a selected allocation above a control. This conservative reading matters here because the 6-bit selected allocation is clearly lower than FP16 while multiple control contrasts remain unresolved. Reporting all complete controls therefore avoids presenting a favorable allocation narrative that the available RAVDESS evidence cannot sustain.

The same discipline applies to FLEURS. Its paired bootstrap retains utterance identities and recomputes corpus scores in every draw. It does not transform BLEU or chrF into a per-speaker statistic, because speakers are not the registered independent unit in this translation analysis. The FLEURS results can establish the observed translation scores and their paired-output uncertainty. They cannot supply a missing RAVDESS measurement, establish model deployment efficiency, or identify the source of an audio-language decision. The paper treats these as separate evidentiary boundaries, each connected to a recorded output and a stated analysis unit.

5 DISCUSSION

The clean FLEURS rerun and RAVDESS study answer different empirical questions. In FLEURS, the selected-minus-FP16 intervals include zero for both BLEU and chrF at both budgets. In RAVDESS, the selected 6-bit allocation is below FP16 under a speaker-resampled interval and a finite exact analysis; the 7-bit FP16 comparison is unresolved. These results do not show a monotonic relation between translation and emotion accuracy. They show why a translation endpoint cannot establish performance on a task whose target is absent from the transcript.

The matched controls sharpen that interpretation. FP16 is a useful reference for the loss incurred by one allocation, but it does not show whether the selected placement is distinguished from other placements at the same average bit width. The control outcomes identify lower selected-plan point estimates at both budgets and a resolved adverse 7-bit front-block comparison. They leave the remaining control pairs unresolved. This combination is more informative than a single reported text score, while remaining conditional on one checkpoint, two selected budgets, a fixed prompt, and 16 speakers.

RAVDESS sentence balance makes the endpoint lexically non-diagnostic, but it does not prove that the model uses emotion-related acoustic cues. Speaker identity, recording conditions, or other non-lexical information may also contribute. The study applies dequantized weights and has no packed-model measurement, so nominal bit width does not establish memory savings, latency, or throughput. These boundaries motivate task-specific evaluation; they do not describe deployment behavior.

A practical reading of a text score. A BLEU or chrF value is a statement about translation under its stored outputs and scorer. Extending it to a lexically non-diagnostic endpoint requires that endpoint's own outputs, an appropriate baseline, and controls that separate an allocation from its budget. Here, sentence identity cannot predict the RAVDESS labels, and the speaker-level analysis identifies a 6-bit FP16 deficit. Neither result says that every quantized speech model will lose emotion accuracy, or that translation and emotion must trade off. It establishes an evaluation requirement for this case study.

What a stronger test would change. A transcript-only condition would measure whether information outside the words improves emotion-label predictions under the same setup. Audio perturbations that preserve spoken words while altering candidate cues could test which information the model uses. Replication across checkpoints, naturalistic emotion data, and fully recorded packed implementations would separately test scope, perceptual validity, and efficiency. These experiments answer distinct questions and are not implied by the current results.

6 LIMITATIONS

The experiments use one Qwen2-Audio checkpoint and one acted North American English emotion set. Results may differ for spontaneous emotion, long-form speech, multilingual prosody, or other tasks whose labels cannot be recovered from the transcript. The FLEURS intervals are conditional on one complete generation record and do not represent execution-to-execution variability. The RAVDESS analysis has 16 independent speaker clusters, and its exact sign-flip test uses a sharp sign-exchangeability null. The balanced sentences do not exclude non-acoustic metadata routes or establish causal audio use. No packed-model or runtime efficiency result is reported.

No human listening or human emotion judgment was performed for this study. Exact-match accuracy measures agreement with the dataset's acted emotion labels under the specified model prompt;

it does not establish perceived emotion quality or usefulness to listeners. Human evaluation would be a separate future step if the method were presented as preserving perceptual emotion understanding. The present finding is confined to label accuracy, translation text scores, and the comparisons for which complete outputs exist.

7 CONCLUSION

Text-output scores alone do not establish quantized-model performance on a task whose labels cannot be recovered from the transcript. In this Qwen2-Audio study, the complete FLEURS rerun leaves the selected-plan differences from FP16 unresolved at 6 and 7 bits. On RAVDESS, the canonical-label selected 6-bit allocation is below FP16, while the 7-bit interval includes zero; neither selected allocation leads both same-budget controls. These results motivate reporting text quality and lexically non-diagnostic task performance separately, with conclusions restricted to the recorded outputs and endpoint.

For this type of study, a compact reporting sequence follows directly from the analysis. First, state whether the text task and the additional endpoint have the same target and whether the text can determine that target. Second, bind every reported score to a complete output set, manifest, scorer, and allocation file. Third, report the full-precision reference and controls that preserve the relevant resource budget. Fourth, choose an uncertainty unit that matches the dependence in the data, such as speakers for repeated recordings from an actor. Finally, distinguish a resolved difference, an unresolved comparison, and a missing outcome. This sequence does not require an allocation method; it applies when a paper evaluates an existing quantization procedure as well as when it proposes one.

The present case also separates three statements that are often compressed into one. A near-zero translation difference is evidence about translation under the stated generation and scoring record. A negative RAVDESS difference is evidence about canonical-label accuracy under a constrained emotion prompt. A matched-control comparison is evidence about this particular placement of precision relative to other placements at the same average bit budget. None of these statements establishes the others. Keeping them separate avoids turning a convenient lexical metric into a proxy for an endpoint it does not evaluate, and makes the remaining uncertainty visible to a reader who needs to decide whether a fixed low-bit allocation is suitable for a particular speech task.

The numerical pattern also illustrates why the outcome should be reported at the level of the allocation and task. The 6-bit selected plan is lower than FP16 for most actors, but it is not lower for every actor; the 7-bit selected plan has more zero actor-level differences and a wider unresolved comparison with FP16. A single aggregate accuracy percentage would conceal this structure. Likewise, the two matched controls are not replications of one another: they deliberately place the same high-precision capacity in different module sets. Their results help identify that average bit width alone is not a sufficient description of the intervention. They do not reveal which modules are causally responsible for the observed accuracy differences.

For a practitioner, the immediate implication is operational. If the intended use involves a speech endpoint beyond transcription or translation, evaluate that endpoint under the intended prompt and output rule before choosing a quantized allocation. Retain a full-precision reference when possible, and include controls that isolate the role of the allocation placement from the role of the budget. Report the missing outcomes as missing rather than replacing them with a different task or partial run. This procedure produces narrower statements than a broad score comparison, but it gives a reader enough information to assess whether the demonstrated evidence applies to the task at hand.

This reporting discipline also makes negative and unresolved outcomes useful. The 6-bit deficit identifies a configuration that requires caution for the evaluated RAVDESS endpoint. The 7-bit interval and several control intervals identify questions the present outputs cannot settle. Neither type of observation is wasted evidence: together they specify which allocation–task pairs were measured, which comparisons remain uncertain, and what future experiment would be needed to broaden the conclusion. Such a record is more informative for model selection than a text score presented without the target endpoint, comparison budget, or analysis unit.

AI USE STATEMENT

We used generative AI tools to help find related work, draft and edit text, review the evaluation and its interpretation, and revise figure code and layout. We checked citations, numerical results, and the rendered paper against the cited sources and stored experiment outputs. The authors take responsibility for the final manuscript and artifacts.

REPRODUCIBILITY STATEMENT

Section 3, the two figures, and Tables 1–2 specify the complete manifests, allocation files, scoring rules, analyses, and reported outputs.

REFERENCES

Jingyi Chen, Zhimeng Guo, Jiyun Chun, Pichao Wang, Andrew Perrault, and Micha Elsner. Do audio LLMs really LISTEN, or just transcribe? measuring lexical vs. acoustic emotion cues reliance. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 5848–5877. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.eacl-long.274. URL https://aclanthology.org/2026.eacl-long.274/.

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. Qwen2-Audio technical report, 2024. URL https://arxiv.org/abs/2407.10759.

Alexis Conneau, Min Ma, Simran Khanuja, Yu Zhang, Vera Axelrod, Siddharth Dalmia, Jason Riesa, Clara Rivera, and Ankur Bapna. FLEURS: Few-shot learning evaluation of universal representations of speech. In *2022 IEEE Spoken Language Technology Workshop*, pp. 798–805, 2022. doi: 10.1109/SLT54892.2023.10023141.

Yuhang Dai, Xin Shu, Zengxi Li, Lei Xie, Xiangang Li, and Jianwei Yu. TAG-Bench: Benchmarking temporal audio grounding in large audio language models, 2026. URL https://arxiv.org/abs/2609.01542.

Tim Dettmers, Mike Lewis, Younes Belkada, and Luke Zettlemoyer. LLM.int8(): 8-bit matrix multiplication for transformers at scale, 2022. URL https://arxiv.org/abs/2208.07339.

Elias Frantar, Saleh Ashkboos, Torsten Hoefler, and Dan Alistarh. GPTQ: Accurate post-training quantization for generative pre-trained transformers, 2022. URL https://arxiv.org/abs/2210.17323.

Tianteng Gu, Bei Liu, Haoyu Wang, and Yanmin Qian. Ultra-low bit post-training quantization of large speech models via k-means clustering and mixed precision allocation. In *Interspeech 2025*, 2025. URL https://www.isca-archive.org/interspeech_2025/gu25b_interspeech.html.

Beom Jin Kang and Hyun Kim. GenPTQ: Green post-training quantization for large-scale ASR models with mixed-precision bit allocation. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 10704–10718. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.findings-emnlp.566. URL https://aclanthology.org/2025.findings-emnlp.566/.

Apoorva Kulkarni, Kaousheik Jayakumar, Sreyan Ghosh, Utathya Aich, Ramani Duraiswami, and Dinesh Manocha. TEMPO: Temporally-grounded multi-task post-training for large audio-language models. In *Proceedings of the 2026 Conference on Empirical Methods in Natural Language Processing*, 2026a. URL https://openreview.net/forum?id=LoXjHBlPEd.

Apoorva Kulkarni, Kaousheik Jayakumar, Sreyan Ghosh, Sarah Wiegreffe, Ramani Duraiswami, and Dinesh Manocha. A closer look at failure modes in temporal understanding of large audio-language models, 2026b. URL https://arxiv.org/abs/2606.17417.

Ji Lin, Jiaming Tang, Haotian Tang, Shang Yang, Wei-Ming Chen, Wei-Chen Wang, Guangxuan Xiao, Xingyu Dang, Chuang Gan, and Song Han. AWQ: Activation-aware weight quantization for LLM compression and acceleration, 2023. URL https://arxiv.org/abs/2306.00978.

Danni Liu, Sai Koneru, and Jan Niehues. Diet-KIT: Post-training quantization for speech LLMs. In *Proceedings of the 23rd International Conference on Spoken Language Translation*, pp. 189–196,

2026. doi: 10.18653/v1/2026.iwslt-1.21. URL https://aclanthology.org/2026.iwslt-1.21/.

Steven R. Livingstone and Frank A. Russo. The Ryerson audio-visual database of emotional speech and song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in north american english. *PLOS ONE*, 13(5):e0196391, 2018. doi: 10.1371/journal.pone.0196391.

Yujian Ma, Jinqiu Sang, Ruizhe Li, Jiaao Yu, and Ang Li. From semantics to readout: Mechanistic understanding of audio tokens after fine-tuning for temporal audio grounding, 2026. URL https://arxiv.org/abs/2607.25355.

Yasmin Moslem. Efficient speech translation through model compression and knowledge distillation, 2025. URL https://arxiv.org/abs/2505.20237.

Alonso Palomino. Selected-layer codec compression for compact speech translation models: An IWSLT 2026 english-to-chinese submission. In *Proceedings of the 23rd International Conference on Spoken Language Translation*, pp. 40–46, 2026. doi: 10.18653/v1/2026.iwslt-1.4. URL https://aclanthology.org/2026.iwslt-1.4/.

Maja Popovič. chrF: Character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pp. 392–395, Lisbon, Portugal, 2015. Association for Computational Linguistics. doi: 10.18653/v1/W15-3049. URL https://aclanthology.org/W15-3049/.

Matt Post. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pp. 186–191, Brussels, Belgium, 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-6319. URL https://aclanthology.org/W18-6319/.

Seamless Communication, Loic Barrault, Yu-An Chung, Mariano Cora Meglioli, David Dale, Ning Dong, Paul-Ambroise Duquenne, Hady Elsahar, et al. SeamlessM4T: Massively multilingual and multimodal machine translation, 2023. URL https://arxiv.org/abs/2308.11596.

Guangxuan Xiao, Ji Lin, Mickael Seznec, Hao Wu, Julien Demouth, and Song Han. SmoothQuant: Accurate and efficient post-training quantization for large language models, 2022. URL https://arxiv.org/abs/2211.10438.

Yizhou Zhang, Wangjin Zhou, Xin Gu, Yichi Wang, Wei Tan, Yi Zhao, Zhi Gong, Keisuke Imoto, and Tatsuya Kawahara. Direct or mediated? task-dependent audio information routing in large audio language models, 2026. URL https://arxiv.org/abs/2608.27026.

A Study Identities and Analysis Units

This appendix separates a test example, a generated output, an allocation arm, and an independent analysis unit. The FLEURS rerun contains 508 ordered English-to-German utterances. FP16 and both selected allocation files have one stored prediction for every identity. Paired resampling retains the prediction and reference for an identity together. The output hashes, manifest hash, and SacreBLEU signatures are stored with the scoring record. The paper uses only this complete three-arm rerun; historical FLEURS output archives without sufficient execution provenance are excluded.

RAVDESS contains 512 clips from actors 1–16. Every actor has 32 clips, including four clips for every actor–emotion pair. The seven compared arms share the exact same ordered manifest, task prompt, model revision, generation seed, and maximum generation length. A clip is the unit for accuracy, while an actor is the unit for the speaker bootstrap and the exact sign-flip calculation. The sign-flip analysis first computes the mean paired difference within each actor, then enumerates all $2^{16} = 65{,}536$ sign assignments of these actor means. Its two-sided value is the fraction of assignments with absolute mean at least as large as the observed absolute mean.

B COMPLETE-OUTPUT RAVDESS RESULTS

Table 4: Canonical-label RAVDESS results for all complete arms. Differences are selected allocation minus comparator. Intervals are 95% equal-weight speaker bootstrap intervals over 16 actors with 10,000 resamples; p is a two-sided exact sign-flip value over the 16 paired actor means. Reported differences use unrounded scores.

Allocation	**Accuracy (%)** ↑	**Selected − comparator (%)** ↑	**Sign-flip** p
FP16	70.12	6 bit: -3.71 $[-6.05, -1.37]$	0.0148
Selected, 6 bit	66.41	6 bit: Reference	—
Front-block, 6 bit	68.16	6 bit: -1.76 $[-4.30, +0.78]$	0.2754
Stratified-uniform, 6 bit	68.36	6 bit: -1.95 $[-4.30, +0.39]$	0.1836
Selected, 7 bit	68.95	7 bit: Reference	—
Front-block, 7 bit	73.44	7 bit: -4.49 $[-6.45, -2.54]$	0.0020
Stratified-uniform, 7 bit	70.31	7 bit: -1.37 $[-2.73, 0.00]$	0.1465

The table gives each complete arm once and states the comparison orientation explicitly. The 6-bit selected-minus-FP16 interval and exact test both support a negative observed difference for this configuration. The sign-flip analysis is not a correction across the several reported contrasts. Its role is to make the small-cluster uncertainty calculation inspectable alongside the speaker bootstrap.

C SCORING AND OUTPUT SCOPE

FLEURS uses SacreBLEU 2.5.1 with the BLEU and chrF configurations specified in Section 3; the scoring record stores the full signatures. RAVDESS scoring lowercases generated text, removes periods and commas, maps "disgusted" to "disgust", and counts a prediction only when exactly one prompted label remains. This canonicalization is applied identically to every complete arm.

The reported evidence concerns fixed stored outputs for one checkpoint, the specified prompts, and the stated evaluation populations. It does not establish a packed implementation property, causal acoustic-cue use, human-perceived emotion quality, or a conclusion across other speech models and tasks.