# Assessment of L2 Speech Global Dimensions using Large Audio Language Models

**Elsayed Issa**
Purdue University
West Lafayette, IN, USA
esissa@purdue.edu

**Mahmoud Ali**
University of Michigan
Ann Arbor, MI, USA
alimahm@umich.edu

## Abstract

Large audio language models (LALMs) integrate audio representations with large language models to enable unified understanding of spoken content. Their capabilities have been increasingly investigated across several benchmarks; however, the examination of their use in rating L2 speech is still in its infancy. This study explores the abilities of LALMs in scoring three L2 speech global dimensions: foreign accentedness, comprehensibility, and intelligibility. Ninety audio samples produced by L2 speakers were rated by ten native speaker raters as well as five LALM models. Model performance was evaluated against the human composite mean using Pearson $r$, Spearman $p$, mean absolute error (MAE), and systematic bias, with the human leave-one-out correlation ($r = .46-.73$ across dimensions) serving as an empirical performance benchmark. The results showed that no LALM reached human-level performance on any dimension. Only one model (i.e., Gemini) achieved a significant correlation with human ratings on comprehensibility ($r$ = .28, $p < .01$), while Qwen2-Audio showed modest correlation on intelligibility ($r$ = .32, p $< .01$). MAE ranged from 0.75 to 3.99 for accentedness (human: 1.24), 1.35 to 3.00 for comprehensibility (human: 1.24), and 12.03 to 15.43 for intelligibility (human: 8.49). All models exhibited systematic biases, with deviations ranging from -9.31 to +13.19 points. The paper concludes with a discussion of the implications for automated L2 speech assessment.

## 1 Introduction

Large audio language models (LALMs) integrate audio representations with large language models to enable unified understanding and interaction over spoken content, environmental sounds, and even music. Such integration pushes beyond traditional automatic speech recognition pipelines toward more general instruction following and audio question answering (Chu et al., 2024; Ding et al., 2025; Yang et al., 2024a; Tang et al., 2023). Recent models such as Qwen-Audio, Qwen2-Audio, Qwen2-Audio-7B-Instruct and Kimi-Audio demonstrate strong zero- and few-shot behavior across diverse audio tasks. These capabilities highlight how scale and multitask training can broaden audio reasoning and conversational capabilities (Chu et al., 2023, 2024; Ding et al., 2025).

As LALMs move from offline speech-to-text toward interactive voice assistants, the relevant notion of *understanding* increasingly encompasses not only speech but also non-speech audio, music, and other acoustic signals. These capabilities have been systematically investigated across a growing number of benchmarks, including AudioBench (Wang et al., 2025a), VoiceBench (Chen et al., 2024), SD-Eval (Ao et al., 2024), ADU-Bench (Gao et al., 2025), and MMAU (Wang et al., 2025b), each targeting different aspects of audio comprehension such as instruction following, spoken dialogue understanding, and multi-task audio reasoning. Previous work on L2 speech investigated GPT-4o capabilities for pronunciation assessment across phoneme, word, and sentence levels (Ahn and Nam, 2025), the potential of LALMs in language learning (Liu et al., 2026), multimodal LLM-based pronunciation scoring system (Fu et al., 2024), speech intelligibility (Pham, 2023), and ASR bias in relation to intelligibility, comprehensibility, and foreign accentedness (Issa et al., 2026). However, the investigation of L2 speech global dimension (i.e., intelligibility, comprehensibility, and foreign accentedness) is limited.

To address this issue, the current study aims to examine the extent to which LALMs can reliably and validly assess L2 speech listener-based global dimensions, namely intelligibility, comprehensibility, and foreign accentedness. These three constructs occupy a central place in L2 speech research (Munro and Derwing, 2015), yet the empirical in-



*Proceedings of the 21st Workshop on Innovative Use of NLP for Building Educational Applications*, pages 702–714

vestigation of LALMs effectiveness in measuring these dimensions remains unknown. The study is guided by the following research questions: 1) How foreign accented, comprehensible, and intelligible is L2 Arabic speech based on LALMs assessment? 2) To what extent do LALMs agree with human rating of intelligibility, comprehensibility, and foreign accentedness in L2 Arabic speech? 3) How LALMs-based ratings of intelligibility, comprehensibility, and foreign accentedness differ as a function of model category?

## 2 Related work

### 2.1 L2 speech global dimensions

In L2 pronunciation research, a shift has moved the focus from native-like pronunciation toward comprehensible and intelligible speech (Levis, 2005; Nagle et al., 2018). In line with this shift, recent research has centered on three global, perceptual dimensions of L2 speech: foreign-accentedness (deviation from native-speaker norms), comprehensibility (ease of understanding), and intelligibility (actual understanding) (Munro and Derwing, 1995). These constructs have been found to be distinct yet partially related, with the key finding being that foreign-accented speech is not necessarily less comprehensible or intelligible across multiple languages (e.g., Huensch and Nagle, 2021; Munro and Derwing, 1995; Ali, 2023).

These three constructs have been examined in the context of L2 pronunciation research based on trained or untrained human raters. Munro and Derwing (1995) examined L2 English speech and reported significant correlations between comprehensibility and intelligibility for 15 of their 18 listeners, with a mean correlation of $r = .51$, while only five listeners showed significant correlations between accentedness and intelligibility. In L2 Spanish, Huensch and Nagle (2021) found that comprehensibility is strongly linked to intelligibility, but foreign-accentedness is not, and that phonemic errors have a stronger negative association with listener ratings than grammatical errors. Ali (2023) extended the investigation to L2 Arabic by examining the relationships among foreign-accentedness, comprehensibility, and intelligibility and exploring the contribution of foreign-accentedness and/or comprehensibility ratings in intelligibility scores. The results supported the partial independence of the three dimensions. Recent meta-analytic evidence has provided a comprehensive synthesis of these dimensions relationships. Chau and Huensch (2025) analyzed 141 effect sizes from 49 studies spanning 1995 to 2023, documenting weighted mean correlations of $r = .75$ for foreign accentedness/comprehensibility, $r = .57$ for comprehensibility/intelligibility, and $r = .32$ for intelligibility/foreign accentedness. These findings establish a clear hierarchy in which intelligibility shows the weakest relationship with foreign accentedness and moderate associations with comprehensibility, while comprehensibility and foreign accentedness show stronger interdependence. The small correlation between intelligibility and foreign accentedness ($r = .32$, 95% CI [.08, .52]) provides the most compelling evidence that having a strong foreign accent does not necessarily impede actual understanding. Across studies in this line of research, human raters have been found to provide reliable and valid assessment of the three dimensions under investigation.

Efforts to automate the assessment of L2 speech have been examined, though most work has focused on specific speech dimensions such as segmental and suprasegmental features. Automatic speech scoring systems operationalize pronunciation differently and have been criticized for lack of transparency in how speech features are weighted (Cai et al., 2025). Traditionally, these systems rely on training procedures using neural networks. In addition, one of the limitations in these systems is the high error rates, particularly spontaneous L2 speech (Cai et al., 2025).

More recently, attention has turned to whether large language models and multimodal architectures can address this limitation. Ahn and Nam (2025) investigated the zero-shot capabilities of GPT-4o for pronunciation assessment across phoneme, word, and sentence levels. The study found that its scoring accuracy was significantly lower than tools that are specifically designed for that purpose, leading to the conclusion that multimodal LLMs may not yet replace purpose-built systems without fine-tuning. Similarly, Fu et al. (2024) proposed a multimodal LLM-based pronunciation scoring system and demonstrated competitive results against alignment-based baselines on the Speechocean762 dataset, though performance on spontaneous L2 speech and global dimensions remains unexamined.

Moreover, Pham (2023) investigated L2 English intelligibility using both the ASR transcription system (i.e., Otter) and 40 L2 English human

raters. The study found that while the ASR system achieved 79-91% transcription accuracy, human intelligibility scores were lower (M = 3.9-6.8 out of 10), revealing a critical divergence between ASR capabilities and human accuracy. In the same vein, Issa et al. (2026) examined linguistic bias in Whisper large-v3 performance on L2 Arabic speech, comparing ASR word error rate (WER) to human-based transcriptions (i.e., intelligibility) and ratings of comprehensibility and foreign accentedness. The findings revealed that intelligibility scores significantly predicted lower WER while foreign accentedness was associated with higher WER, whereas comprehensibility appear to be independent of WER. Finally, Liu et al. (2026) explored the potential of LALMs for interactive language learning by introducing L2-Arctic-plus, an English dataset annotated with detailed pronunciation error explanations and actionable suggestions for improvement. The study benchmarked both cascaded ASR-plus-LLM pipelines and existing LALMs on tasks including error detection and the generation of corrective feedback. Findings indicated that current models show promise for chat-based pronunciation training but still fall short of providing consistently accurate and pedagogically useful guidance.

### 2.2 LALMs benchmarks

A growing body of research has systematically evaluated LALM performance across diverse benchmarks. AudioBench (Wang et al., 2025a) provided a comprehensive evaluation framework spanning 8 task categories and 26 datasets while also examining prompt sensitivity effects in LALMs, where they found that LALMs could behave differently to different prompt templates. Across multiple audio modalities including speech, environmental sounds, and music, AIR-Bench (Yang et al., 2024b) assessed instruction-following competencies through a two-tier architecture: a foundation tier comprising 19 tasks with approximately 19,000 single-choice items, and a conversational tier containing roughly 2,000 open-ended question-answer pairs evaluated via GPT-4-as-judge methodology against reference responses. Likewise, MMAU (Tyagi et al., 2024) targeted audio understanding and reasoning across three domains (speech, sound, music) using exclusively multiple-choice formats to enable fully automated assessment without human annotation or LLM-based judging. Moreover, Dynamic-SUPERB (Huang et al., 2024b,a) implements a two-phase evaluation protocol for universal instruction-based speech models under zero-shot conditions. Results revealed that no single model achieves consistent performance across all tasks: SALMONN-13B demonstrates superiority in English ASR, whereas Qwen2-Audio-7B-Instruct exhibits stronger emotion recognition capabilities, yet neither generalizes effectively across the complete task spectrum.

Several benchmarks have targeted contextual and content-level comprehension. SD-Eval (Ao et al., 2024), for instance, addresses a more focused yet critical aspect of LALM functionality: spoken dialogue comprehension beyond surface-level transcription. It assesses models' capacity to simultaneously process content, speaking style, speaker emotion, and background acoustics—the multifaceted contextual understanding required for conversational assistants. Findings indicated inadequate contextual response generation, attributed to insufficient task formalization and appropriate evaluation datasets. Similarly, ADU-Bench (Gao et al., 2025) evaluates open-ended audio dialogue performance across 3 general scenarios, 12 competencies, 9 languages, and 4 ambiguity categories. Evaluation of 16 LALMs revealed superior performance on information-seeking queries relative to casual conversational exchanges, with systematic failures in mathematical reasoning, roleplay interactions, non-English language processing, and phonetic ambiguity resolution.

Additional benchmarks have specifically investigated audio hallucination phenomena. Ghosh et al. (2023) introduced Comp-A to assess compositional reasoning, demonstrating that contemporary models achieve only marginally above-chance performance when matching audio clips to captions with varied event orderings or attribute bindings. Sung-Bin et al. (2024) developed AVH-Bench, a cross-modal hallucination benchmark for audio-visual LLMs, revealing widespread difficulty in parsing fine-grained audio-visual relationships. Furthermore, AHa-Bench (Cheng et al., 2025) proposed a tripartite hallucination taxonomy—semantic, acoustic, and semantic-acoustic—and evaluated seven open-source LALMs, uncovering substantial limitations in models' joint interpretation of semantic and acoustic information. While these benchmarks have examined several capabilities of LALMs, they mainly focused on L1 English and L1 Chinese, leaving not only other L1s but also L2s unexplored, including Arabic.

## 3 Methods

### 3.1 Data

The L2 Arabic speech data came from 30 adult L1 English speakers (19M, 11F) who were enrolled in second- and third-year Arabic courses at a U.S. public university. All were rated at the Intermediate level of proficiency based on the American Council on the Teaching of Foreign Languages (ACTFL) scale. The speech samples were rated by 10 L1 Arabic speakers (9M, 1F) with minimal exposure to L2 Arabic speech and no teaching experience. No participants reported hearing or speech impediments (Ali, 2023).

*Speech Production:* The speech samples were as part of an Oral Proficiency Interview (OPI) administered by a certified OPI tester and recorded via Zoom (15–35 min) following an online language background questionnaire. Three speech samples per speaker were extracted from distinct OPI sections. Samples were amplitude-normalized (peak = 0.99) in Audacity (Audacity Team, 2020), yielding 90 samples (30 speakers x 3) averaging 13 seconds and 15 words.

*Speech Perception:* Ten L1 Arabic listeners completed three tasks on all 90 samples across counterbalanced sessions on separate days ( 3 hours total). Task 1 involved rating foreign-accentedness on 9-point scale where 0 means no foreign accent and 9 means strongly foreign accented. Task 2 involved rating comprehensibility on 9-point scale where 0 means very easy to understand and 9 means very difficult to understand. Finally, in task 3, listeners transcribed the speech samples as a measure intelligibility (Munro and Derwing, 1995). An intelligibility score was calculated as the percentage of correctly transcribed words. Tasks were blocked (35 samples/block) and presented via a custom Qualtrics interface, with listeners rating each sample after a single hearing following three practice items.

### 3.2 Experiments

Five LALMs were purposively selected due to their multilingual capabilities, including Arabic. These models are Qwen2-Audio[1], Qwen2-Audio-7B-Instruct[2] (Chu et al., 2024), Ultravox[3], Gemini-2.5-Pro (Google DeepMind, 2025) and GPT-4o-audio-preview (OpenAI, 2024) (see Table 4 for more information about the open-source models). We probed each model using a standardized zero-shot prompt designed to elicit numerical ratings on each of the three global speech dimensions. Models are prompted independently per dimension using the prompts shown in Table 5. To evaluate the extent to which these five LALMs can simulate human perception of L2 speech dimensions under investigation, a series of statistical analyses were conducted in Python using the pingouin library (Vallat, 2018).

Our analyses map onto the three research questions as follows. For RQ1, descriptive statistics (mean, SD, range, 95% CI) summarize how LALMs rate each dimension. For RQ2, we evaluate LALM–human agreement using complementary metrics that capture distinct facets of agreement: Pearson $r$ and Spearman $p$ index linear and rank-order correspondence; mean absolute error (MAE) and root mean square error (RMSE) quantify the magnitude of per-item disagreement; and signed bias captures systematic over- or underestimation. We additionally fit each LALM as an 11th rater alongside the ten human raters and recompute the intraclass correlation, testing whether a LALM behaves as an exchangeable member of the rater panel. A leave-one-out (LOO) analysis, in which each human rater's ratings are correlated with the mean of the remaining nine, provides an empirically grounded benchmark for human-level performance against which LALM agreement can be interpreted.

For RQ3, we compare per-item MAE between open- and closed-source model families using a Wilcoxon signed-rank test (paired by item), with Cohen's $d$ as a measure of effect size. The Wilcoxon test was chosen over a paired t-test because per-item MAE is bounded at zero and skewed, violating the normality assumption. We measure Intraclass Correlation Coefficient (ICC) variants (Shrout and Fleiss, 1979; McGraw and Wong, 1996). On the one hand, ICC2 is the single-rater absolute-agreement coefficient. It indicates how closely one randomly selected rater (human or LALM) matches another in absolute terms. ICC2k, on the other hand, is the corresponding average-measures coefficient, showing the reliability of the mean rating across the full panel. ICC3 and ICC3k are the consistency analogs (two-way mixed effects), which discount systematic between-rater

[1] https://huggingface.co/Qwen/Qwen2-Audio-7B

[2] https://huggingface.co/Qwen/Qwen2-Audio-7B-Instruct

[3] https://huggingface.co/fixie-ai/ultravox-v0_5-llama-3_1-8b

shifts in mean; we report these for completeness. ICC2 is our primary metric because the research question is whether a LALM produces the same numerical rating a human would, not merely a rank-consistent one. This same logic motivates reporting MAE and signed bias alongside Pearson $r$: Pearson correlation is invariant to additive shifts that absolute-agreement metrics correctly penalize.

The five LALMs differ systematically along a dimension that is theoretically relevant to L2 Arabic assessment: training scale and Arabic representation. The closed-source models (Gemini 2.5-Pro, GPT-4o-audio-preview) are trained on substantially larger and more multilingual corpora than the open-source models in our sample, which are built on Whisper-large-v2 or v3-turbo encoders paired with 7–8B parameter language model backbones (Qwen-7B, Llama-3.1-8B; see Table 4). Arabic, and L2 Arabic in particular, is plausibly underrepresented in the training data of the open-source models. Prior benchmark work has shown that LALMs struggle in non-English language environments (Gao et al., 2025), making the open- versus closed-source comparison a useful probe of whether scale and broader multilingual coverage partially compensate for the demands of low-resource L2 perceptual rating. We therefore examine whether per-item agreement with human raters differs systematically between the two model families, recognizing that with five models this comparison is hypothesis-generating rather than definitive.

## 4 Results

Table 1 presents the descriptive statistics for LALMs ratings across the three dimensions. LALMs rated foreign accentedness toward the higher end of the 9-point scale (M = 5.70, SD = 2.40), comprehensibility ratings fell near the midpoint of the scale (M = 4.78, SD = 2.47), and intelligibility showed the broadest range among the three measures (M = 77.08, SD = 28.85, range = 20–100) (see Figure 3 for distribution of comprehensibility and foreign-accentedness ratings and intelligibility scores by human raters and LALMs).

Table 2 presents Pearson $r$, mean absolute error (MAE), and systematic bias for each model across the three dimensions, alongside the human LOO analysis. Across all dimensions and all five models, no LALM reached the human LOO correlation threshold. All models showed statistically significant deviations from the human agreement composite mean (Wilcoxon signed-rank, all $p < .001$; see Table 7). For intelligibility, Qwen2-Audio was the only model to significantly correlate with human ratings ($r = .322$, $p = .002$). On the other hand, Gemini and GPT showed no significant correlations.

Ultravox and Qwen2-Audio-Instruct provided the same scores/ratings to all speech samples across all the three dimensions. From statistical analysis perspective, we acknowledge that these are considered statistical artifacts. However, from model performance evaluation, this is also indicative of model hallucinations which will be discussed later.

As for comprehensibility, while the association between Gemini and human agreement composite mean was significant ($r = .281$, $p = .007$), the correlation between GPT and the human agreement composite mean was not ($r = .180$, $p = .090$). Open-source models did not correlate meaningfully. Finally, for foreign accentedness, none of the LALMs significantly correlated with human-based ratings. It is worth noting that Ultravox and Qwen2-Audio-Instruct rated speech samples with constant scores across constructs, resulting in meaningless correlations with those of humans.

Table 3 summarizes the results that examine whether model category influenced model performance. a Wilcoxon signed-rank test was adopted for each dimension comparing per-item mean absolute error between the open- and closed-source families. Per-item error of open-sourced models was significantly larger than that of closed-sourced models for intelligibility (open MAE = 29.02, SD = 5.02; closed MAE = 13.76, SD = 10.80) and for comprehensibility (open MAE = 2.49, SD = 0.57; closed MAE = 2.10, SD = 1.05) while there was no significance difference between the two model categories for foreign accentedness (open MAE = 2.07, SD = 0.42; closed MAE = 2.10, SD = 0.87).

Finally, we examined agreement reliability model fit with human-based agreement as baseline. Each LALM was fitted as an additional (11th) rater, and ICC2 was recomputed for the augmented panel and compared against the ten-human baseline. As shown in Figure 1, across all three dimensions, model fitting declined by adding any of the LALMs. For intelligibility, the decline was most pronounced for Gemini (baseline ICC2 = .49; with Gemini = .39) followed with GPT (baseline ICC2 = .49; with GPT = .43), and finally Qwen2-Audio (baseline ICC2 = .49; with Qwen2 = .44). As for comprehensibility, the decline was shown for Qwen2-Audio

| | M | SD | Min | Max | 95% CI |
|---|---|---|---|---|---|
| Foreign-accentedness | 5.70 | 2.40 | 1 | 9 | 5.47-5.92 |
| Comprehensibility | 4.78 | 2.47 | 1 | 9 | 4.55-5.00 |
| Intelligibility | 77.08 | 28.85 | 20 | 100 | 74.40-79.75 |

Table 1: Descriptive statistics for LALMs' ratings of Foreign-accentedness, Comprehensibility, and Intelligibility

| | | Foreign Accentedness | | | Comprehensibility | | | Intelligibility | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | *r* | MAE | Bias | *r* | MAE | Bias | *r* | MAE | Bias |
| Open | Qwen2-Audio | .15 | 3.99 | 3.90 | .16 | 3.00 | 1.32 | .32** | 12.03 | +8.80 |
| | Qwen2-Instruct | —† | 0.75 | +0.48 | —† | 1.35 | +1.11 | —† | 12.28 | 9.31 |
| | Ultravox | —† | 1.48 | +1.48 | —† | 1.35 | +1.11 | —† | 12.28 | 9.31 |
| Closed | Gemini | .15 | 1.56 | +0.39 | .28** | 2.11 | 1.40 | .12 | 15.43 | +13.19 |
| | GPT | .04 | 2.64 | 2.58 | .18 | 2.09 | 2.06 | .07 | 12.08 | +10.02 |
| | Human LOO | .46 | 1.24 | — | .57 | 1.24 | — | .73 | 8.49 | — |

MAE = mean absolute error. Bias = model mean minus human mean. Human LOO = leave-one-out benchmark: mean Pearson r and MAE of each human rater correlated against the remaining nine raters, averaged across raters. † Zero-variance output where correlation is undefined. ** $p < .01$. See Table 9 for individual rater LOO correlations and MAE across dimensions

Table 2: Per-model agreement with human ratings across the three constructs (N = 90).

(baseline ICC2 = .336; with Qwen2 = .219), GPT (baseline ICC2 = .336; with GPT = .27), and Gemini (baseline ICC2 = .336; with Gemini = .29). Finally, the decline is pronounced for Qwen2-Audio (baseline ICC2 = .209; with Qwen2 = .122), followed by GPT (.15) and Gemini (.18) respectively for foreign-accentedness (see Table 8 for more on model fit).

## 5 Discussion

### 5.1 LALMs assessment of global speech dimensions

LALMs rated L2 Arabic speech as moderately to strongly foreign accented (M = 5.70, SD = 2.40), moderately difficult to understand (M = 4.78, SD = 2.47), and mostly highly intelligible (M = 77.08, SD = 28.85). However, aggregate similarity in mean ratings is not, by itself, evidence that LALMs reproduce human rating behavior. The within-dataset evidence presented in Sections 5.2 and 5.3, item-level agreement, distributional patterns, and the effect of substituting LALMs into the rater panel, yields a less favorable assessment. Prior work has documented limitations of multimodal LLMs in segmental and sentence-level pronunciation assessment (Ahn and Nam, 2025; Fu et al., 2024); the present results extend the scope of these limitations to global perceptual dimensions, where the gap between LALM and human raters, quantified against the human LOO benchmark, is more pronounced.

### 5.2 Global speech dimensions: LALMs-based versus human-based

Three sources of within-dataset evidence converge on the conclusion that LALMs do not reliably reproduce human rating behavior on our corpus.

**Item-level agreement:** No LALM reached the human LOO correlation benchmark on any dimension (Table 3). The human LOO correlations averaged $r = .73$ for intelligibility, .57 for comprehensibility, and .46 for foreign accentedness, whereas the strongest LALM correlations were $r = .32$ (Qwen2-Audio, intelligibility), .28 (Gemini, comprehensibility), and .15 (Qwen2-Audio and Gemini, foreign accentedness). Per-item MAE values likewise exceeded the human LOO MAE on every dimension and for every model. Even the closest match, Qwen2-Instruct on foreign accentedness (MAE = 0.75), is misleading, as that value reflects a constant output rather than item-sensitive rating.

**Distributional patterns:** Although LALM mean ratings approximate human means at the aggregate level, Figure 2 shows that the underlying distributions diverge. LALM ratings exhibit larger standard deviations and, in two cases (Ultravox and Qwen2-Instruct), collapse to a single value across all 90 samples. This distributional mismatch indicates that aggregate similarity in central tendency masks substantial item-level disagreement.

**Effect on inter-rater reliability:** When each LALM is fitted as an 11th rater alongside the

| Construct | Open MAE | Closed MAE | W | $p$ | $d$ |
|---|---|---|---|---|---|
| Foreign Accentedness | 2.07 | 2.10 | 1934.5 | .781 | 0.03 |
| Comprehensibility | 2.49 | 2.10 | 1355.0 | .008 | 0.27 |
| Intelligibility | 29.02 | 13.76 | 263.0 | $< .001$ | 1.11 |

Table 3: Open-source vs. closed-source comparison of per-item mean absolute error across the three constructs. $W$ and $p$ are from Wilcoxon signed-rank tests (paired by item); $d$ is Cohen's $d$. The Wilcoxon test was chosen because per-item MAE is bounded at zero and right-skewed, violating the normality assumption of the paired t-test.

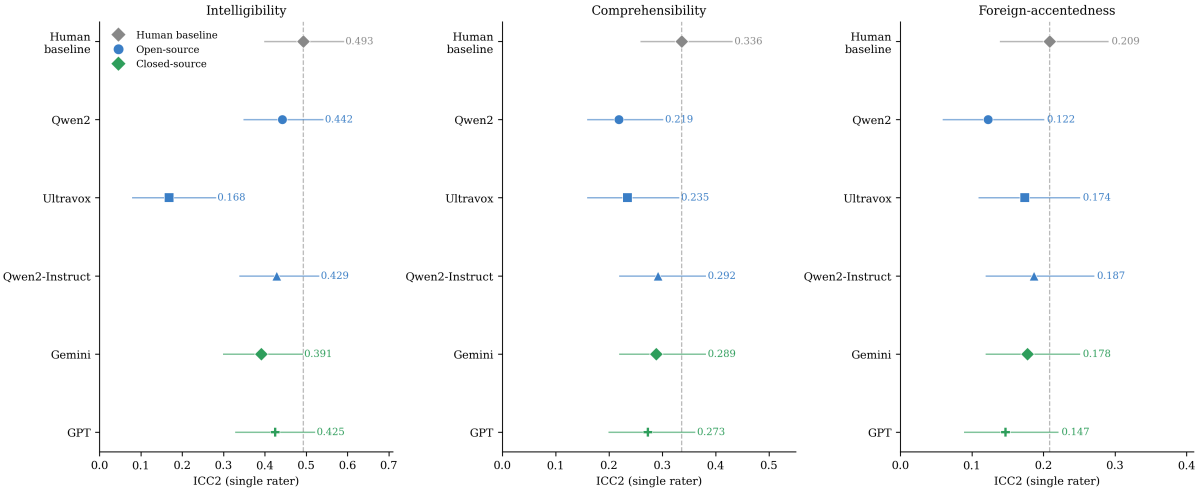



Figure 1: ICC2 when each model is added as an 11th rater to the human panel (dashed line = 10-human baseline; error bars = 95% CI)

ten human raters, the panel's single-rater absolute-agreement coefficient (ICC2) declines on every dimension and for every model (Figure 1; Table 8). The decline is most pronounced for intelligibility with Gemini (.49 → .39), for comprehensibility with Qwen2-Audio (.34 → .22), and for foreign accentedness with Qwen2-Audio (.21 → .12). Adding any LALM to the rater panel reduces, rather than augments, the reliability of the rating panel which is a direct evidence that LALMs do not function as exchangeable raters.

Taken together, these three lines of evidence indicate that the apparent similarity between LALM and human ratings at the aggregate level does not extend to the item level, the distributional level, or the level of rater interchangeability. This pattern is consistent with prior findings of divergence between automated and human speech assessment (Pham, 2023) and suggests that LALMs do not simulate the underlying mechanisms human raters draw on when evaluating L2 speech.

### 5.3 Open- versus closed-model assessment

The by-category comparison reveals a dimension-dependent pattern that aligns with the theoretical motivation outlined in Section 3.2. Closed-source models outperformed open-source models on intelligibility (open MAE = 29.02, closed MAE = 13.76; $d = 1.11$, $p < .001$), modestly outperformed them on comprehensibility (open MAE = 2.49, closed MAE = 2.10; $d = 0.27$, $p = .008$), and showed no advantage on foreign accentedness (open MAE = 2.07, closed MAE = 2.10; $d = 0.03$, $p = .781$). Figure 2 and Table 10 corroborate this gradient via Bland-Altman statistics. The dimension-dependent gradient is informative about what kind of capability drives performance on each construct. Intelligibility is operationalized as the proportion of words correctly transcribed, so it depends most directly on ASR quality, a capability that scales with training data volume and multilingual coverage, where closed-source models have a clear advantage. Comprehensibility is a perceptual judgment of processing effort that is less directly tied to ASR quality, which may explain the smaller advantage. Foreign accentedness requires phonetic comparison against L1 Arabic norms, a capability that appears to be poorly developed in both model families and is consistent with Arabic being underrepresented in training data across the board. Read this way, the open/closed comparison is not

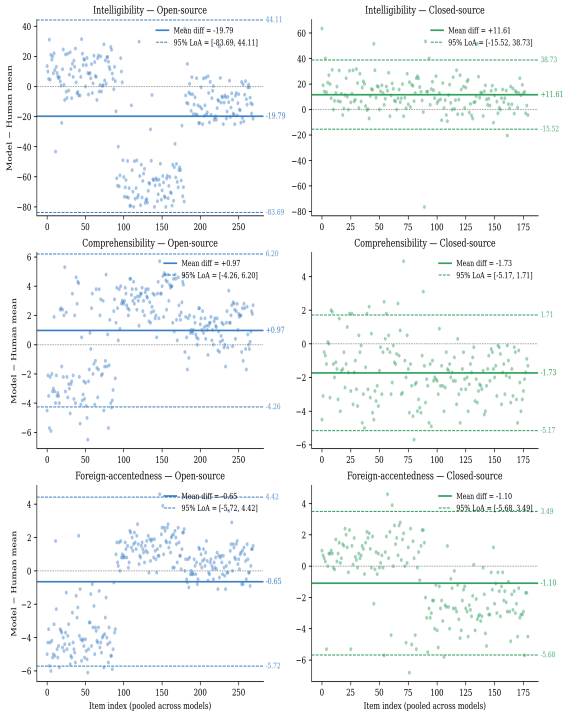



Figure 2: Model category ratings vs. human composite mean

merely a performance comparison between model families but evidence that scale and multilingual coverage partially compensate for low-resource L2 settings on transcription-driven dimensions, while leaving phonetic-norm-based judgments largely unaddressed.

The constant-output behavior observed in Ultravox and Qwen2-Audio-Instruct warrants attention. Both models produced identical ratings across all 90 samples on every dimension, which is reflected in their undefined Pearson correlations (Table 3) and in the spike-shaped distributions in Figure 3. This pattern indicates that the models did not engage with item-level acoustic or linguistic variation when prompted for perceptual ratings, defaulting instead to a fixed response. This behavior is reported as hallucination phenomena documented in LALM benchmarks (Cheng et al., 2025; Sung-Bin et al., 2024), in which models generate plausible but unfounded outputs. This raises the question of whether instruction-tuned LALMs are prone to substituting heuristic responses for genuine acoustic analysis on perceptual rating tasks. Possible explanations include insufficient exposure to L2 Arabic in pretraining/instruction-tuning that biases the model toward modal scale-point responses regardless of input, or failure of the audio encoder to extract dimension-relevant features from the signal.

### 5.4 Implications for automated L2 speech assessment

The findings of this study carry several implications for the use of LALMs in L2 speech assessment. First, current LALMs may not be used in high-stakes assessment contexts such as proficiency certification or placement testing. This aligns with the broader consensus in language assessment that automated scoring systems must accumulate validity evidence comparable to that expected of human raters before they can be justified for operational use, especially in high-stakes decisions (Chapelle and Voss, 2021). Second, the finding that certain models showed partial sensitivity to specific dimensions suggests that LALMs may be amenable to domain-specific fine-tuning for targeted dimension assessment. Rather than relying on general-purpose multitask training, future work could fine-tune LALMs on L2 speech data annotated for specific global dimensions, potentially improving model–human agreement on dimensions where zero-shot performance has the greatest potential. Finally, the results highlight the need for closer collaboration between the speech technology researchers and applied linguistics communities. Current LALM benchmarks (Wang et al., 2025a; Tyagi et al., 2024; Wang et al., 2025b; Sung-Bin et al., 2024) evaluate models primarily on L1 speech, but not L2 speech in their assessment criteria. One benchmark, (Gao et al., 2025), indicated that LALMs struggle in non-English language environments. Our findings demonstrate that strong performance on standard audio understanding tasks does not transfer to the perceptual demands of L2 speech rating. Future benchmark development should incorporate L2 speech assessment tasks across multiple languages and proficiency levels.

## 6 Conclusion

This study provided an evaluation of LALMs as automated raters across three global dimensions of L2 speech. Current LALMs cannot simulate human ratings with sufficient accuracy or reliability to serve as raters in L2 speech assessment. This conclusion rests on three converging within-dataset findings: LALM agreement with human ratings falls below the human LOO benchmark on every dimension; LALM rating distributions diverge from human distributions despite similar means; and adding any LALM to the rater panel reduces rather than augments panel reliability.

## Limitations

Several limitations should be acknowledged. First, the sample of 90 recordings represents a relatively small corpus of L2 speech. We recognize this not only as a limitation but also as an invitation to L2 researchers to evaluate LALMs on their own datasets and to propose strategies for improving LALM performance. Second, we acknowledge that Arabic may be underrepresented in the training data of the open-source models examined in the current study. That said, this presents an opportunity to the broader L2 community to extend this line of inquiry to other languages in order to validate our findings. Third, models were evaluated in a zero-shot configuration using a single standardized prompt per dimension, without few-shot examples, chain-of-thought reasoning, et cetera. We chose this design deliberately to establish a baseline of out-of-the-box LALM behavior on L2 speech rating, which mirrors how non-expert end users (e.g., learners, instructors) are most likely to deploy these systems in practice. We acknowledge, however, that this design cannot disentangle inherent model limitations from prompt-induced underperformance, and some portion of the observed model–human gap may be attributable to suboptimal prompting rather than representational shortcomings. That said, two patterns in our results suggest prompting alone is unlikely to fully close the gap. First, the constant-output behavior of Ultravox and Qwen2-Audio-Instruct across all 90 samples points to a failure to engage with the acoustic signal rather than a calibration issue that anchors could resolve. Second, the magnitude of systematic bias observed for several models (up to +13.19 points on intelligibility) and the dimension-specific divergence between open- and closed-source families suggest underlying differences in audio representation quality, not merely scale-mapping artifacts. Future work should systematically vary prompting strategy to quantify the share of variance attributable to prompt design versus model capability.

## Ethics Statement

We adhered to ethical principles to ensure responsible and respectful collection of the data. Participants were fully informed about the purpose of the study, the nature of the data being collected, and their right to withdraw at any time without consequence. No personally identifiable information was collected (see Ali, 2023 for more information).

## Data Availability

The L2 Arabic speech corpus and human ratings are not yet publicly available due to ongoing research, but the authors plan to release them in the near future. Access may be granted to researchers upon reasonable request.

## Acknowledgment

We would like to thank the anonymous reviewers for their valuable comments, which helped to improve this manuscript; all remaining errors are our own. We also gratefully acknowledge the financial support and resources provided by the School of Languages and Cultures at Purdue University.

## References

Taekyung Ahn and Hosung Nam. 2025. English pronunciation evaluation without complex joint training: Lora fine-tuned speech multimodal llm. *arXiv preprint arXiv:2509.02915*.

Mahmoud M. E. Ali. 2023. The foreign-accentedness, comprehensibility, and intelligibility of l2 arabic speech. *Language Teaching Research*, 0(0).

Junyi Ao, Yuancheng Wang, Xiaohai Tian, Dekun Chen, Jun Zhang, Lu Lu, Yuxuan Wang, Haizhou Li, and Zhizheng Wu. 2024. Sd-eval: A benchmark dataset for spoken dialogue understanding beyond words. *Advances in Neural Information Processing Systems*, 37:56898–56918.

Audacity Team. 2020. Audacity.

Danwei Cai, Ben Naismith, Maria Kostromitina, Zhongwei Teng, Kevin P Yancey, and Geoffrey T LaFlair. 2025. Developing an automatic pronunciation scorer: Aligning speech evaluation models and applied linguistics constructs. *Language Learning*, 75(S1):170–203.

Carol A Chapelle and Erik Voss. 2021. Introduction to validity argument in language testing and assessment. *Validity argument in language testing: Case studies of validation research*, pages 1–16.

Tuc Chau and Amanda Huensch. 2025. The relationships among l2 fluency, intelligibility, comprehensibility, and accentedness: A meta-analysis. *Studies in Second Language Acquisition*, 47(1):282–307.

Yiming Chen, Xianghu Yue, Chen Zhang, Xiaoxue Gao, Robby T Tan, and Haizhou Li. 2024. Voicebench: Benchmarking llm-based voice assistants. *arXiv preprint arXiv:2410.17196*.

Xize Cheng, Dongjie Fu, Chenyuhao Wen, Shannon Yu, Zehan Wang, Shengpeng Ji, Siddhant Arora, Tao Jin, Shinji Watanabe, and Zhou Zhao. 2025. Ahabench: Benchmarking audio hallucinations in large

audio-language models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhifang Guo, Yichong Leng, Yuanjun Lv, Jinzheng He, Junyang Lin, and 1 others. 2024. Qwen2-audio technical report. *arXiv preprint arXiv:2407.10759*.

Yunfei Chu, Jin Xu, Xiaohuan Zhou, Qian Yang, Shiliang Zhang, Zhijie Yan, Chang Zhou, and Jingren Zhou. 2023. Qwen-audio: Advancing universal audio understanding via unified large-scale audio-language models. *arXiv preprint arXiv:2311.07919*.

Ding Ding, Zeqian Ju, Yichong Leng, Songxiang Liu, Tong Liu, Zeyu Shang, Kai Shen, Wei Song, Xu Tan, Heyi Tang, and 1 others. 2025. Kimi-audio technical report. *arXiv preprint arXiv:2504.18425*.

Kaiqi Fu, Linkai Peng, Nan Yang, and Shuran Zhou. 2024. Pronunciation assessment with multi-modal large language models. *arXiv preprint arXiv:2407.09209*.

Kuofeng Gao, Shu-Tao Xia, Ke Xu, Philip Torr, and Jindong Gu. 2025. Benchmarking open-ended audio dialogue understanding for large audio-language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4763–4784.

Sreyan Ghosh, Ashish Seth, Sonal Kumar, Utkarsh Tyagi, Chandra Kiran Reddy Evuru, S Sakshi, Oriol Nieto, Ramani Duraiswami, Dinesh Manocha, and 1 others. 2023. Compa: Addressing the gap in compositional reasoning in audio-language models. In *The Twelfth International Conference on Learning Representations*.

Google DeepMind. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. Technical report, Google DeepMind.

Chien-yu Huang, Wei-Chih Chen, Shu-wen Yang, Andy T Liu, Chen-An Li, Yu-Xiang Lin, Wei-Cheng Tseng, Anuj Diwan, Yi-Jen Shih, Jiatong Shi, and 1 others. 2024a. Dynamic-superb phase-2: A collaboratively expanding benchmark for measuring the capabilities of spoken language models with 180 tasks. In *The Thirteenth International Conference on Learning Representations*.

Chien-yu Huang, Ke-Han Lu, Shih-Heng Wang, Chi-Yuan Hsiao, Chun-Yi Kuan, Haibin Wu, Siddhant Arora, Kai-Wei Chang, Jiatong Shi, Yifan Peng, and 1 others. 2024b. Dynamic-superb: Towards a dynamic, collaborative, and comprehensive instruction-tuning benchmark for speech. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 12136–12140. IEEE.

Amanda Huensch and Charlie Nagle. 2021. The effect of speaker proficiency on intelligibility, comprehensibility, and accentedness in l2 spanish: A conceptual replication and extension of munro and derwing (1995a). *Language Learning*, 71(3):626–668.

Elsayed Issa, Mahmoud Ali, and Kevin Hirschi. 2026. Measuring linguistic bias in ASR: Whisper large-v3 on non-native speech versus human perception. *Procedia Computer Science*, 275:692–699.

John M Levis. 2005. Changing contexts and shifting paradigms in pronunciation teaching. In *Pronunciation*, pages 265–272. Routledge.

Hongfu Liu, Zhouying Cui, Xiangming Gu, and Ye Wang. 2026. Unlocking large audio-language models for interactive language learning. *arXiv preprint arXiv:2601.14744*.

Kenneth O. McGraw and S. P. Wong. 1996. Forming inferences about some intraclass correlation coefficients. *Psychological Methods*, 1(1):30–46.

Murray J Munro and Tracey M Derwing. 1995. Foreign accent, comprehensibility, and intelligibility in the speech of second language learners. *Language learning*, 45(1):73–97.

Murray J Munro and Tracey M Derwing. 2015. A prospectus for pronunciation research in the 21st century: A point of view. *Journal of Second Language Pronunciation*, 1(1):11–42.

Charles Nagle, Rebecca Sachs, and Germán Zárate-sández. 2018. Exploring the intersection between teachers' beliefs and research findings in pronunciation instruction. *The Modern Language Journal*, 102(3):512–532.

OpenAI. 2024. GPT-4o System Card. Technical report.

Trang Minh Thi Pham. 2023. The intelligibility of vietnamese-accented english to artificial intelligence software and asian listeners. Technical report, TESOL Working Paper Series.

Patrick E. Shrout and Joseph L. Fleiss. 1979. Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, 86(2):420–428.

Kim Sung-Bin, Oh Hyun-Bin, JungMok Lee, Arda Senocak, Joon Son Chung, and Tae-Hyun Oh. 2024. Avhbench: A cross-modal hallucination benchmark for audio-visual large language models. In *The Thirteenth International Conference on Learning Representations*.

Changli Tang, Wenyi Yu, Guangzhi Sun, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun MA, and Chao Zhang. 2023. Salmonn: Towards generic hearing abilities for large language models. In *The Twelfth International Conference on Learning Representations*.

Utkarsh Tyagi, Sonal Kumar, Ashish Seth, Ramaneswaran Selvakumar, Oriol Nieto, Ramani Duraiswami, Sreyan Ghosh, and Dinesh Manocha. 2024. Mmau: A massive multi-task audio understanding and reasoning benchmark. *arXiv preprint arXiv:2410.19168*.

Raphael Vallat. 2018. Pingouin: statistics in python. *Journal of Open Source Software*, 3(31):1026.

Bin Wang, Xunlong Zou, Geyu Lin, Shuo Sun, Zhuohan Liu, Wenyu Zhang, Zhengyuan Liu, AiTi Aw, and Nancy F. Chen. 2025a. AudioBench: A universal benchmark for audio large language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4297–4316, Albuquerque, New Mexico. Association for Computational Linguistics.

Dingdong Wang, Jincenzi Wu, Junan Li, Dongchao Yang, Xueyuan Chen, Tianhua Zhang, and Helen Meng. 2025b. Mmsu: A massive multi-task spoken language understanding and reasoning benchmark. *arXiv preprint arXiv:2506.04779*.

Dongchao Yang, Haohan Guo, Yuanyuan Wang, Rongjie Huang, Xiang Li, Xu Tan, Xixin Wu, and Helen Meng. 2024a. Uniaudio 1.5: Large language model-driven audio codec is a few-shot audio task learner. *Advances in Neural Information Processing Systems*, 37:56802–56827.

Qian Yang, Jin Xu, Wenrui Liu, Yunfei Chu, Ziyue Jiang, Xiaohuan Zhou, Yichong Leng, Yuanjun Lv, Zhou Zhao, Chang Zhou, and Jingren Zhou. 2024b. AIR-bench: Benchmarking large audio-language models via generative comprehension. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1979–1998, Bangkok, Thailand. Association for Computational Linguistics.

## A Data Annotation Guidelines

The annotation procedure followed established guidelines from the L2 pronunciation research literature, specifically the protocol introduced by Munro and Derwing (1995) and subsequently adopted in numerous studies of L2 speech global dimensions (e.g., Huensch and Nagle, 2021; Chau and Huensch, 2025). The guidelines operationalize each of the three constructs as follows: foreign accentedness was rated on a 9-point scale (1 = no foreign accent, 9 = strong foreign accent); comprehensibility was rated on a 9-point scale (1 = very easy to understand, 9 = very difficult to understand); and intelligibility was operationalized as the proportion of correctly transcribed words relative to the total number of words actually spoken (0–100). The full operational definitions provided to raters are reported in Table 5 of the manuscript and are identical to the prompts used to elicit ratings from the LALMs, ensuring comparability between human and model ratings. The complete annotation protocol, including listener recruitment, task instructions, scale anchors, and session structure, is documented in (Ali, 2023).

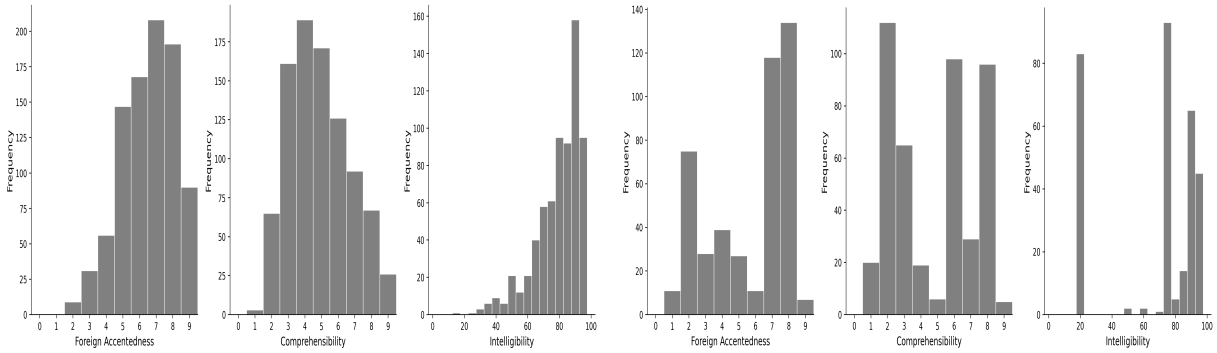
## B Model Information

| **Qwen2-Audio** |
|---|
| Qwen2-Audio builds upon its predecessor Qwen-Audio (Chu et al., 2023), integrating audio and text inputs through a unified encoder-decoder architecture to generate textual outputs. The model processes audio inputs using a Whisper-large-v2 audio encoder paired with a Qwen-7B large language model backbone. It introduces two distinct audio interaction modes: audio analysis and voice chat. |
| **Qwen2-Audio-Instruct** |
| Qwen2-Audio-7B-Instruct is the instruction-tuned variant of Qwen2-Audio, built on the same Whisper-large-v2 encoder and Qwen-7B backbone architecture but further optimized for interactive dialogue through supervised fine-tuning on spoken dialogue data and instruction-following datasets. |
| **Ultravox** |
| Ultravox v0.5 is a multimodal speech language model that integrates a pretrained Llama-3.1-8B-Instruct backbone with the encoder component of Whisper-large-v3-turbo to process both speech and text inputs and generate textual outputs. |

Table 4: Open-source model information

## C Appendix

Appendix C reports supplementary materials referenced throughout the paper and the zero-shot prompts.




(a) Human raters (b) LALMs

Figure 3: Distribution of of comprehensibility and foreign-accentedness ratings and intelligibility scores by human raters and LALMs.

| **Foreign Accentedness** |
|---|
| **Definition:** The degree to which L2 speech is similar to or different from L1 speech in terms of pronunciation as perceived by a native L1 listener with no familiarity with L2 speech.<br>**Operationalization:** 9-point scale 1=Very similar to L1 speech in pronunciation (no foreign accent at all) 9-Very different to L1 speech (strong foreign accent) |
| **Comprehensibility** |
| **Definition:** The degree to which L2 speech is easy or difficult o understand as perceived by a native L1 listener with no familiarity with L2 speech.<br>**Operationalization:** 9-point scale 1=Very easy to understand (processing effort is zero to little) 9-Very difficult to understand (processing effort is maximum) |
| **Intelligibility** |
| **Definition:** The degree to which L2 speech is correctly recognized as transcribed by a native L1 listener with no familiarity with L2 speech.<br>**Operationalization:** 0-100 The proportion of correctly transcribed words as compared to the total number of words actually spoken. |

Table 5: Zero-shot prompts for rating foreign accentedness, comprehensibility, and intelligibility.

| ICC type | Foreign Accentedness | | | Comprehensibility | | | Intelligibility | | |
|---|---|---|---|---|---|---|---|---|---|
| | ICC | 95% CI | $p$ | ICC | 95% CI | $p$ | ICC | 95% CI | $p$ |
| ICC2 | .209 | [.14, .29] | $< .001$ | .336 | [.26, .43] | $< .001$ | 493 | [.40, .59] | $< .001$ |
| ICC2k | .725 | [.62, .81] | $< .001$ | .835 | [.78, .88] | $< .001$ | .907 | [.87, .94] | $< .001$ |
| ICC3 | .725 | [.62, .81] | $< .001$ | .361 | [.28, .45] | $< .001$ | .549 | [.47, .63] | $< .001$ |
| ICC3k | .784 | [.71, .84] | $< .001$ | .850 | [.80, .89] | $< .001$ | .924 | [.90, .95] | $< .001$ |

Table 6: Intraclass correlation coefficients for the ten human raters under a two-way random-effects model (Shrout and Fleiss, 1979). ICC2 = single-rater absolute agreement; ICC2k = average-rater absolute agreement (reliability of the panel mean); ICC3 = single-rater consistency (two-way mixed); ICC3k = average-rater consistency.

| Model | Foreign Accentedness | | Comprehensibility | | Intelligibility | |
|---|---|---|---|---|---|---|
| | W | $p$ | W | $p$ | W | $p$ |
| Qwen2 | 19.5 | <.001 | 1144.0 | <.001 | 508.0 | <.001 |
| Qwen2-Instruct | 787.0 | <.001 | 321.5 | <.001 | 443.5 | <.001 |
| Ultravox | 4.0 | <.001 | 0.0 | <.001 | 0.0 | <.001 |
| Gemini | 956.5 | <.001 | 704.5 | <.001 | 162.5 | <.001 |
| GPT | 27.0 | <.001 | 24.5 | <.001 | 381 | <.001 |

Table 7: Wilcoxon signed-rank test results for each model against the human composite mean across three dimensions.

| Configuration | Foreign Accedntedness | | Comprehensibility | | Intelligibility | |
|---|---|---|---|---|---|---|
| | ICC2 | ICC2k | ICC2 | ICC2k | ICC2 | ICC2k |
| human raters (baseline) | .209 | .725 | .336 | .835 | .493 | .907 |
| +Qwen2 | .122 | .605 | .219 | .755 | .442 | .897 |
| +Qwen2-Instruct | .187 | .716 | .292 | .819 | .429 | .892 |
| +Ultravox | .174 | .698 | .235 | .771 | .168 | .690 |
| +Gemini | .178 | .704 | .289 | .817 | .391 | .876 |
| +GPT | .147 | .655 | .273 | .805 | .425 | .891 |

Table 8: ICC2 (single rater) and ICC2k (average raters) when each model is added as an 11th rater to the human panel.

| Rater | INT r | INT MAE | COMP r | COMP MAE | FA r | FA MAE |
|---|---|---|---|---|---|---|
| R01 | .699 | 8.64 | .660 | 0.76 | .650 | 0.94 |
| R02 | .747 | 7.75 | .587 | 1.41 | .491 | 1.45 |
| R03 | .688 | 8.56 | .577 | 1.07 | .526 | 1.09 |
| R04 | .736 | 8.02 | .473 | 1.37 | .397 | 1.26 |
| R05 | .756 | 7.85 | .645 | 1.06 | .392 | 1.41 |
| R06 | .714 | 10.27 | .438 | 1.56 | .499 | 1.32 |
| R07 | .714 | 8.77 | .602 | 1.39 | .430 | 0.94 |
| R08 | .743 | 7.21 | .534 | 1.45 | .435 | 1.74 |
| R09 | .723 | 10.64 | .597 | 1.04 | .494 | 1.18 |
| R10 | .726 | 7.17 | .564 | 1.34 | .317 | 1.10 |
| Mean | .725 | 8.49 | .568 | 1.24 | .463 | 1.24 |

Table 9: Individual human rater leave-one-out (LOO) correlations and mean absolute error across dimensions. INT, COM, and FA refer to intelligibility, comprehensibility and foreign accentedness, respectively.

| | Mean diff | SD | LoA lower | LoA upper | Range | W |
|---|---|---|---|---|---|---|
| Intelligibility — Open | 21.09 | 32.76 | 85.29 | +43.11 | 128.40 | 508.0 |
| Intelligibility — Closed | +11.61 | 13.84 | 15.52 | +38.73 | 54.25 | 381.5 |
| Comprehensibility — Open | +0.97 | 2.67 | 4.26 | +6.20 | 10.47 | 1144.0 |
| Comprehensibility — Closed | 1.73 | 1.75 | 5.17 | +1.71 | 6.87 | 704.5 |
| Foreign-accentedness — Open | 0.65 | 2.59 | 5.72 | +4.42 | 10.14 | 19.5 |
| Foreign-accentedness — Closed | 1.10 | 2.34 | 5.68 | +3.49 | 79.17 | 956.5 |

Table 10: Bland-Altman statistics