Rubric-Guided Fine-tuning of SpeechLLMs for Multi-Aspect, Multi-Rater L2 Reading-Speech Assessment

Aditya Kamlesh Parikh, Cristian Tejedor-Garcia, Catia Cucchiarini, Helmer Strik
Centre for Language Studies, Radboud University, The Netherlands
{aditya.parikh, cristian.tejedorgarcia, catia.cucchiarini, helmer.strik}@ru.nl

Abstract

Reliable and interpretable automated assessment of second-language (L2) speech remains a central challenge, as large speech-language models (SpeechLLMs) often struggle to align with the nuanced variability of human raters. To address this, we introduce a rubric-guided reasoning framework that explicitly encodes multi-aspect human assessment criteria: accuracy, fluency, and prosody, while calibrating model uncertainty to capture natural rating variability. We fine-tune the Qwen2-Audio-7B-Instruct model using multi-rater human judgments and develop an uncertainty-calibrated regression approach supported by conformal calibration for interpretable confidence intervals. Our Gaussian uncertainty modeling and conformal calibration approach achieves the strongest alignment with human ratings, outperforming regression and classification baselines. The model reliably assesses fluency and prosody while highlighting the inherent difficulty of assessing accuracy. Together, these results demonstrate that rubric-guided, uncertainty-calibrated reasoning offers a principled path toward trustworthy and explainable SpeechLLM-based speech assessment.

Keywords: SpeechLLM, L2 Reading Speech, Multi-Aspect Assessment, SpeechLLM Fine-tuning, Uncertainty Modeling

1. Introduction

Reading is a foundational skill for learning, communication, and participation in society. It includes multiple aspects: readers must pronounce words accurately (Newell et al., 2020), maintain a natural temporal flow to ensure fluency (Kuhn and Stahl, 2003), and convey appropriate phrasing and emphasis to reflect prosody (Schwanenflugel et al., 2004). These dimensions interact in complex ways. For instance, prosodic phrasing can mask or amplify word-level inaccuracies (Zechner et al., 2015).

Many readers, including children developing basic reading skills (Schwanenflugel et al., 2004), L2 learners acquiring phonological and rhythmic competence (Sleeman et al., 2022), and adults encountering unfamiliar vocabulary or orthographies (Chang et al., 2020), struggle to achieve fluent, accurate, and prosodically natural reading. Early, high-quality feedback is essential for enabling teachers and clinicians to identify learners' needs and adjust instruction (Grønli et al., 2024).

Nevertheless, expert assessment and feedback are costly, time-consuming, often inconsistent across raters (Smith and Paige, 2019), and suffer from variability in severity and scale usage (leniency, harshness, central tendency, and halo effects) and from drift over time due to fatigue or anchoring (Neittaanmäki and Lamprianou, 2024). Ensuring inter- and intra-rater reliability across raters and sessions remains challenging, especially for child and L2 speech, where accent variation and atypical prosody further reduce agreement (Ishikawa, 2023). In particular, raters tend to agree more on pronunciation accuracy than on fluency or prosody (van der Velde et al., 2024).

Diverse modeling approaches have been explored to develop automatic systems for speaking and reading assessment. Early work in the 1990s relied on statistical models such as Hidden Markov Models (HMMs) and Gaussian Mixture Models (GMMs), which primarily assessed pronunciation in reading speech using posterior probabilities or Goodness of Pronunciation (GOP) scores (Kim et al., 1997; Witt and Young, 2000; Witt, 2000). Subsequent studies broadened the scope to fluency and prosody, introducing timing- and pitch-based features (Wang et al., 2024).

With the advent of deep neural networks, pronunciation assessment research began emphasizing calibration to ensure that a model's predicted confidence aligns with the reliability of its assessments (Evanini et al., 2018). Later, text-based Language Models (LMs) such as BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019) were applied to automated scoring using Automatic Speech Recognition (ASR) transcripts; however, their dependence on textual transcriptions and lack of acoustic awareness limited their ability to capture fluency and prosody effectively. The emergence of pretrained acoustic models such as wav2vec 2.0 (Baevski et al., 2020) addressed this gap, improving performance on tasks like mispronunciation detection and diagnosis (Cao et al., 2024; Parikh et al., 2025a,b; Phan et al., 2025). In parallel, the rise of multimodal Large Language Models (LLMs) has extended natural language processing beyond text and inspired similar progress in the speech

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 10255–10265
11–16 May 2026.

domain. For instance, models such as SpeechLLaMA (Wu et al., 2023), SALMONN (Tang et al., 2023), VOILA (Wang et al., 2023), and Qwen-Audio (Chu et al., 2023) expanded text-based architectures by integrating acoustic and linguistic representations. These SpeechLLMs demonstrated strong performance on general audio understanding tasks such as speech recognition, translation, audio captioning, and spoken question answering, yet they still lacked the capability to follow detailed natural-language instructions.

A newer generation of instruction-tuned SpeechLLMs has recently emerged, including Qwen2-Audio-Instruct (Chu et al., 2024), GAMA (Ghosh et al., 2024), and Audio Flamingo 2 (Ghosh et al., 2025). These models combine large-scale audio–text pretraining with instruction tuning, allowing them to reason over spoken input and generate structured responses directly from raw speech. While such instruction-tuned models hold significant potential for rubric-guided and explainable assessment of L2 speech proficiency, they remain underexplored in the literature. Recent work has begun to highlight both their promise and limitations. For instance, Parikh et al. (2025c) and Ma et al. (2025) investigated the Qwen2-Audio-Instruct model under distinct settings. Parikh et al. (2025c) observed that rubric-based SpeechLLMs tend to produce overly generous scores in zero-shot conditions, reflecting a "niceness bias" inherited from instruction tuning that discourages low ratings even for poor-quality speech. Their study also introduced a multi-aspect assessment framework, allowing simultaneous scoring of complementary proficiency dimensions such as accuracy, fluency, prosody, and sentence completeness. In contrast, Ma et al. (2025) demonstrated that supervised fine-tuning can effectively mitigate this bias, yielding more consistent and reliable proficiency predictions. Nonetheless, their framework was limited to single-score regression and classification setups and did not account for inter-rater variability or predictive uncertainty, two critical factors for ensuring fairness and reliability in automated assessment.

Building upon these findings, we adopt the Qwen2-Audio-Instruct model as the foundation of our study, utilizing its instruction-following capacity with paired audio inputs and descriptive textual rubrics for assessment. The model is fine-tuned in a multi-aspect manner to infer holistic scores along three complementary dimensions: accuracy (degree of mispronunciation), fluency (smoothness and coherence of delivery), and prosody (intonation, rhythm, and stress). We will share our code repository upon acceptance.

To systematically examine whether incorporating additional configurations can improve robustness and alignment with human judgments, we design five state-of-the-art (SOTA) modeling strategies of increasing complexity based on the related literature. (1) Discrete Classification (**DiCl**) treats proficiency scoring as a categorical prediction task (Xi et al., 2012). However, this uniform treatment of errors disregards the ordinal relationship between categories, which is crucial in assessment tasks where the severity of misclassification depends on the distance between true and predicted proficiency levels; (2) Single-Rubric Regression (**SRR.M**) formulates the scoring as continuous prediction using mean squared error (MSE) (Chen et al., 2018), aligning better with human rating scales and capturing finer performance differences; (3) Multi-Rubric Regression (**MRR.M**) jointly predicts multiple rubrics simultaneously with MSE (Do et al., 2023), enabling shared representation learning in multiple aspects such as accuracy, fluency, and prosody; (4) Multi-Rubric Regression with Gaussian Negative Log-Likelihood (GNLL), with the acronym **MRR.G**, replaces MSE with GNLL to model prediction uncertainty (Kendall and Gal, 2017); and (5) Multi-Rubric Multi-Rater Regression with GNLL and Conformal Prediction (**MRR.GC**) incorporates multiple human ratings and applies Conformal Prediction (Angelopoulos and Bates, 2021; Braun et al., 2025) to generate calibrated confidence intervals. Among these, the last two configurations (MRR.G and MRR.GC) represent novel contributions to the field of automated L2 speech assessment. To the best of our knowledge, this is the first study to integrate Gaussian uncertainty modeling and conformal calibration within a multi-rater supervision framework for multi-aspect assessment using a rubric-guided fine-tuned SpeechLLM. This design advances the SOTA by jointly addressing score reliability, fairness, and explainability, three critical yet previously underexplored dimensions in SpeechLLM-based proficiency assessment.

This leads us to our research question (RQ): *To what extent can a SpeechLLM approximate human ratings in multi-aspect (accuracy, fluency, and prosody) performance assessment of L2 reading speech?*

2. Methodology

In this section, we describe the experimental setup for our sentence-level speech assessment framework, which leverages a multimodal (speech and text) SpeechLLM fine-tuned to predict rubric-based scores for multi-aspect (accuracy, fluency, and prosody) assessment. The task was framed either as classification (five discrete levels: Very Poor to Excellent) or as regression (continuous scores on a 1–10 scale). While classification provides interpretable, rubric-aligned decisions, regression enables finer granularity and captures subtle vari-

ations in human ratings (Xi et al., 2012). We fine-tuned the Qwen2-Audio-7B-Instruct[1] model using Low-Rank Adaptation (LoRA) (Hu et al., 2022) for parameter-efficient adaptation while preserving the model's pre-trained capabilities. The following subsections describe the model architecture, dataset, training setup, optimization loss functions, and the assessment protocol.

2.1. Model Architecture

We built upon the Qwen2-Audio-7B-Instruct model, a 7B-parameter multimodal transformer-based SpeechLLM pre-trained on large-scale audio–text pairs for conditional generation tasks. The model integrates an audio encoder with a text decoder, enabling it to process interleaved audio and textual instructions. For speech assessment, a lightweight scoring head was added to project the hidden representations from the final transformer layer to output predictions. We explored two variants: (i) a classification head trained with cross-entropy loss for discrete 5-level scoring, and (ii) a regression head trained with MSE or GNLL loss to predict continuous scores.

To enable parameter-efficient fine-tuning, LoRA was applied to all linear layers of the base model. The key hyperparameters were a rank of $r = 32$, a scaling factor of $\alpha = 32$, and a dropout rate of 0.1, with rank-stabilized LoRA (rsLoRA) enabled. This configuration resulted in approximately 10 million trainable parameters ($\approx 1.2\%$ of the total), focusing the optimization on low-rank update matrices while keeping the original model weights frozen.

2.2. Dataset

We used the publicly available dataset SpeechOcean762 (Zhang et al., 2021), a widely adopted benchmark for automatic pronunciation and speaking assessment for research. It contains 5000 English read-speech utterances, divided into 2500 for training and 2500 for testing. We fine-tuned our model on the training split and assessed the test split. The corpus includes recordings from both child and adult speakers whose native language is Mandarin Chinese (L1), reading English (L2). Each utterance was independently evaluated by five expert raters along sentence, word, and phoneme level aspects. Scores were assigned on a 1–10 scale following the official annotation protocol defined in the original dataset publication. This dataset is particularly suited for our study as it provides multi-rater, multi-aspect annotations, enabling analysis of both inter-rater variability and multi-dimensional scoring behavior.

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

In this work, we focus on sentence-level scoring (accuracy/fluency/prosody) to study rubric-guided utterance-level assessment under multi-rater supervision. We exclude Completeness due to an unclear rubric definition for our setup. Scores are skewed toward mid- to high ratings, potentially biasing predictions toward the central range.

In a multi-aspect assessment of speech, accuracy is measured as the pronunciation correctness of the spoken sentence. A score of 10 corresponds to excellent pronunciation with no noticeable mispronunciations (near-native articulation), whereas 1 indicates completely unintelligible speech or absence of voice. Fluency evaluates the temporal smoothness and coherence of speech, focusing on pauses, repetitions, and stammering. A score of 10 reflects coherent, uninterrupted delivery with natural pacing, while 1 denotes inability to read the sentence as a whole or no voice. Finally, Prosody measures the intonation, rhythm, and speaking rate, capturing the naturalness and expressiveness of speech. A score of 10 represents correct intonation with stable rhythm and speed, sounding natural and engaging, and 1 indicates speech too stammered to evaluate or the absence of voice.

The scoring rubrics for all three aspects follow the definitions provided in the SpeechOcean762 paper ((Zhang et al., 2021), page 3).

2.3. Training Procedure

Fine-tuning of the Qwen model was conducted using the Hugging Face Transformers library with a custom data collator and trainer (Wolf et al., 2020) as it provides flexibility in handling paired audio–text inputs and rubric-based labels. Each utterance was resampled to 16 kHz, converted to mono audio, and paired with a textual task prompt containing only rubric instructions (no target transcript), so the model assesses directly from audio rather than transcript comparison.

The input was formatted as a conversational prompt following the model's chat template, consisting of a user message that included the audio segment and a rubric-based instruction (e.g., "Score sentence-level accuracy on a scale from 1 to 10," accompanied by detailed rubric descriptions). Since the objectives included both regression and classification, no generation prompt was added, as the model was trained to produce direct scalar or categorical predictions rather than generated text. For optimization, the AdamW optimizer was used with a learning rate of 2×10^{-5}, a weight decay of 0.01, and a constant learning rate schedule (no warm-up). The per-device batch size was set to 1, with gradient accumulation steps of 1. Training was performed on a single GPU (NVIDIA A5000), with TF32 acceleration enabled for CUDA operations.

2.4. Optimization Loss Functions

To fine-tune the SpeechLLM for rubric-based speech assessment, five modeling strategies of increasing complexity were explored depending on how the task was formulated. Each formulation defines a distinct mapping between the model output and the target scores, influencing both learning behavior and interpretability. In what follows, we describe the loss functions used for the classification and regression variants of our experiments.

2.4.1. Discrete Classification (DiCl)

This method served as our baseline for sentence-level assessment. Each speech rubric was formulated as a five-class classification task, where every utterance was assigned one of five categorical levels: Very Poor, Poor, Fair, Good, or Very Good. To obtain these labels, we discretize the original 1–10 human rater scores into five ordinal categories using the rubric defined by Zhang et al. (2021). Scores of 1–2 were mapped to Very Poor, 3–4 to Poor, 5–6 to Fair, 7–8 to Good, and 9–10 to Very Good. A softmax output layer produced the probability distribution over the five classes. The model was optimized using the standard cross-entropy loss:

$$\mathcal{L}_{\text{DiCl}} = \frac{1}{N}\sum_{i=1}^{N}\left[-\sum_{c=1}^{C} y_{i,c}\log(\hat{y}_{i,c})\right]$$

where N is the number of utterances in the dataset, i indexes each utterance, C denotes the number of classes, $y_{i,c}$ is the one-hot ground-truth label, and $\hat{y}_{i,c}$ is the predicted probability for class c.

This loss treats all classes as independent and penalizes all misclassifications equally. For instance, predicting Very Good when the true label is Very Poor incurs the same penalty as predicting Good instead of Fair.

2.4.2. Single Rubric Regression with Mean Squared Error (SRR.M)

In this setting, the sentence-level assessment was formulated as a regression task, where the model predicts a continuous score within the range [1, 10] for each rubric independently. Compared to the classification approach, regression provides finer granularity, as the MSE loss penalizes predictions in proportion to their numerical deviation from the true score. For example, predicting 7.5 for a true score of 8 incurs a smaller penalty than predicting 5 for 8, thereby preserving ordinal relationships and avoiding discretization artifacts.

A separate regression head was used to predict a single continuous value for each aspect (accuracy, fluency, or prosody). The model was trained to minimize the MSE loss:

$$\mathcal{L}_{\text{SRR.M}} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$$

where N is the total number of utterances, i indexes each utterance, y_i is the gold mean score per aspect (averaged over the five human raters), and $\hat{y}_i$ is the model-predicted score for the same sample.

2.4.3. Multi Rubric Regression with Mean Squared Error (MRR.M)

In this configuration, the assessment task was extended to a multi-output regression problem, where the model simultaneously predicts continuous scores for all three rubrics. This was achieved using a shared encoder followed by three parallel regression heads, each producing one scalar output per aspect. It also improves computational efficiency, as a single model produces a structured evaluation vector $[\hat{y}_{\text{acc}}, \hat{y}_{\text{flu}}, \hat{y}_{\text{pro}}]$ without requiring separate fine-tuning for each rubric. The training objective minimizes the average MSE across all three rubrics:

$$\mathcal{L}_{\text{MRR.M}} = \frac{1}{3}\sum_{\text{aspect}}\frac{1}{N}\sum_{i=1}^{N}\left(y_{i,\text{aspect}} - \hat{y}_{i,\text{aspect}}\right)^2$$

where N is the total number of utterances, i indexes each utterance, and $y_{i,\text{aspect}}$ and $\hat{y}_{i,\text{aspect}}$ denote the gold and predicted mean scores, respectively, for each rubric.

2.4.4. Multi Rubric Regression with Gaussian Negative Log-likelihood (MRR.G)

Building upon the multi-head regression framework, this variant introduces uncertainty estimation by allowing each output head to predict not only the mean score μ_i but also the corresponding variance σ_i^2 for every rubric, accuracy, fluency, and prosody. This formulation enables the model to express both its prediction and its confidence for each utterance. Fine-tuning employed the GNLL loss, which penalizes large prediction errors while accounting for the predicted uncertainty.

For each utterance, the mean of the five rater scores was used as the gold standard, making this a single-target regression task that reflects the central human consensus. The GNLL loss for each aspect was defined as:

$$\mathcal{L}_{\text{MRR.G}}(i) = \frac{(\bar{y}_i - \mu_i)^2}{2\,\sigma_i^2} + \frac{1}{2}\log\sigma_i^2$$

where $\bar{y}_i$ denotes the averaged human rating, and (μ_i, σ_i^2) are the predicted mean and variance, respectively. The total loss was computed as the

mean across the three rubrics:

$$\mathcal{L}_{\text{total}} = \frac{1}{3} \sum_{\text{aspect}} \frac{1}{N} \sum_{i=1}^{N} \mathcal{L}_{\text{MRR.G}}(i)$$

This formulation enables the model to capture aleatoric uncertainty by adjusting its predicted variance σ_i^2 according to sample difficulty.

2.4.5. Multi Rubric Multi Rater Regression with Gaussian Negative Log-likelihood and Conformer Prediction (MRR.GC)

This configuration extends the uncertainty-aware regression framework by directly modeling all five human rater scores per utterance, rather than relying solely on their mean. For each rubric, accuracy, fluency, and prosody, the model predicts both a mean score μ_i and a variance σ_i^2, jointly capturing the central tendency and spread of human judgments. Fine-tuning again employed the GNLL loss, which in this case integrates the inter-rater variance term s_i^2 to account for disagreement among raters:

$$\mathcal{L}_{\text{MRR.GC}} = \frac{1}{3N} \sum_{\text{aspect}} \sum_{i=1}^{N} \left[\frac{(\bar{y}_i - \mu_i)^2 + s_i^2}{2\,\sigma_i^2} + \frac{1}{2} \log \sigma_i^2 \right]$$

where N is the total number of utterances, $\bar{y}_i = \frac{1}{R} \sum_{r=1}^{R} y_{i,r}$ is the mean rater score, $R = 5$ is the number of raters, and

$$s_i^2 = \frac{1}{R} \sum_{r=1}^{R} (y_{i,r} - \bar{y}_i)^2$$

represents the inter-rater variance for each utterance i. This formulation explicitly incorporates rater disagreement into the loss, enabling the model to reflect both prediction uncertainty and observed human variability.

After fine-tuning, conformal calibration was applied using a 5-fold split of the test set to empirically calibrate the predictive intervals. Normalized residuals $|y_i - \mu_i|/\sigma_i$ were computed on the calibration folds to estimate aspect-wise quantiles (q_{aspect}) corresponding to a target coverage of 90%. Final prediction intervals were then obtained as $[\mu_i - q_{\text{aspect}}\sigma_i,\ \mu_i + q_{\text{aspect}}\sigma_i]$.

This combination of multi-rater supervision and conformal prediction allows the model to capture both inter-rater variability and prediction uncertainty in a statistically interpretable manner. Conformal calibration further ensures empirical coverage, guaranteeing that a defined proportion of true scores fall within the predicted confidence intervals.

2.5. Evaluation Metrics

The model performance was assessed using a comprehensive set of metrics reflecting both categorical decision quality and numerical agreement with human ratings. For the DiCl baseline setup, performance was assessed using the Weighted F1-score and the Matthews Correlation Coefficient (MCC). Weighted F1 accounts for class imbalance by computing the F1-score per class and averaging it by class frequency, while MCC quantifies the overall correlation between predicted and true labels, providing a balanced measure even under uneven label distributions.

For all four regression-based methods (SRR.M, MRR.M, MRR.G, and MRR.GC), we report five complementary metrics: Weighted F1, MCC, Pearson Correlation Coefficient (PCC), Root MSE (RMSE), and Quadratic Weighted Kappa (QWK). Together, these capture both categorical agreement and continuous correlation with human ratings. The predicted continuous scores were rounded to the nearest integer on the 1–10 scale for computing Weighted F1 and MCC, ensuring comparability with discrete human ratings. PCC measures the linear association between predicted and gold scores, and RMSE quantifies the average prediction error. Quadratic Weighted Kappa (QWK) measures the agreement between two ratings on an ordinal scale, accounting for chance agreement and the distance between rating categories. It ranges from -1 (worse than chance) to 1 (perfect agreement), with 0 indicating random agreement. QWK penalizes larger score discrepancies more heavily than smaller ones, making it well-suited for ordinal rating tasks. For QWK (M–R), the agreement is computed between the model and each of the five human raters individually and reported as mean $\pm$ standard deviation across raters.

The regression results were analyzed under two assessment conditions: (1) a strict, exact-match setting without tolerance, and (2) a relaxed setting that allows a ± 1 score tolerance to account for natural rater variability. In practice, small differences, such as one rater giving a 7 and another an 8, are not considered true disagreements but normal variations in human judgment. Prior work in speaking and writing assessment (e.g., TOEFL, SpeechRater) reports inter-rater standard deviations of roughly 0.5–1.0 on a 10-point scale, supporting this tolerance as a realistic estimate of human rating variability. For the MRR.GC configuration, we also calculated the percentage of human scores falling within the model's predicted High–Low confidence range for each aspect. This coverage metric reflects how well the model's predictive intervals capture real human variability, serving as a direct indicator of calibration quality. Because QWK depends on exact ordinal matches, it is computed only under the strict setting, not with the ±1 tolerance.

3. Results

3.1. Inter-Rater Reliability QWK (R-R)

Before presenting the results of the five models, we first report the inter-rater reliability (QWK, R–R) in Table 1. The shown mean QWK values are averaged across all ten possible rater pairs, indicating overall moderate agreement among raters.

	Accuracy	Fluency	Prosody
QWK (R–R)	0.5585 ± 0.0671	0.5019 ± 0.1353	0.5021 ± 0.1153

Table 1: QWK (mean $\pm$ SD) across human raters.

3.2. Classification-Based Assessment

Table 2 summarizes the baseline classification performance (DiCl) across the three assessment rubrics: accuracy, fluency, and prosody. Overall, DiCl results demonstrate moderate discriminative ability across rubrics, providing a reference point for subsequent regression-based approaches that aim to capture finer score variations.

Rubrics	↑ F1	↑ MCC
Accuracy	0.558	0.341
Fluency	0.642	0.452
Prosody	0.670	0.468

Table 2: Classification performance across rubrics for DiCl. Arrows indicate that higher is better.

Figure 1 shows the aggregated confusion matrix across all rubrics for DiCl. Predictions are largely concentrated along the main diagonal, indicating good alignment between the model's output and human ratings. The Good category dominates the predictions. Misclassification mainly occurs between adjacent levels (see Figure 1).

3.3. Regression-Based Assessment

In this section, we compare four regression-based fine-tuning configurations: SRR.M, MRR.M, MRR.G, and MRR.GC. Performance is reported for each rubric (accuracy, fluency, and prosody) under both strict (exact-match) and lenient evaluation settings (± 1 tolerance or High–Low calibration).

3.3.1. Exact-Match Regression Results (without tolerance)

Table 3 presents the regression-based results under exact-match evaluation. Performance is reported separately for the three rubrics: accuracy, fluency, and prosody, to analyze aspect-specific trends before summarizing overall behavior across model variants.

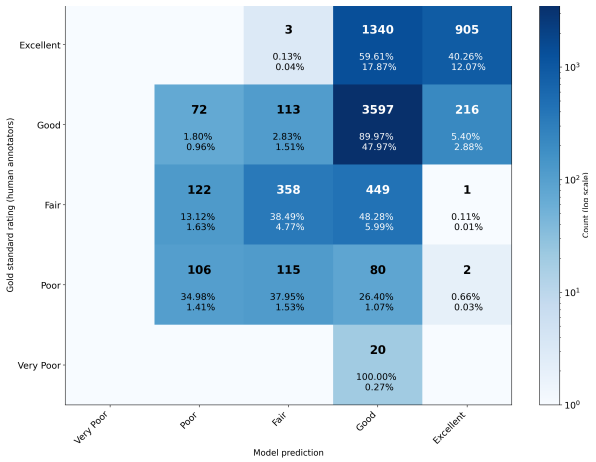

Figure 1: Aggregated confusion matrix across all rubrics for DiCl. Rows show human (gold-standard) ratings. Columns show model predictions. Each cell displays the count (top), percentage within the true class (middle), and percentage across all samples (bottom).

	Rubrics	↑ F1	↑ MCC	↑ PCC	↓ RMSE	↑ QWK (M–R)
SRR.M	Accuracy	0.2479	0.0843	0.7390	1.1854	0.432 ± 0.046
	± 1	0.7966	0.7543	0.9417	0.5456	
	Fluency	0.3717	0.2416	0.7956	0.9326	0.445 ± 0.081
	± 1	0.9137	0.9000	0.9571	0.4218	
	Prosody	0.4073	0.2838	0.7728	0.9163	0.468 ± 0.093
	± 1	0.9216	0.9117	0.9558	0.4357	
MRR.M	Accuracy	0.2424	0.0710	0.6969	1.2894	0.497 ± 0.051
	± 1	0.7561	0.6830	0.7454	0.6221	
	Fluency	0.4571	0.3161	0.7522	0.8910	0.501 ± 0.078
	± 1	0.9378	0.9286	0.8890	0.4462	
	Prosody	0.3435	0.1918	0.7463	1.0342	0.527 ± 0.052
	± 1	0.8809	0.8536	0.8457	0.4553	
MRR.G	Accuracy	0.2285	0.0705	0.7352	1.1762	0.464 ± 0.049
	± 1	0.7785	0.7428	0.8444	0.5402	
	Fluency	0.3931	0.2657	0.8003	0.9237	0.463 ± 0.054
	± 1	0.9179	0.9078	0.8489	0.4308	
	Prosody	0.4095	0.2821	0.7933	0.9082	0.494 ± 0.079
	± 1	0.9226	0.9133	0.8526	0.4185	
MRR.GC	Accuracy	0.4014	0.2654	0.7649	1.0159	0.505 ± 0.039
	High–Low Cal	0.8707	0.8539	0.8716	0.5467	
	Fluency	0.4589	0.3397	0.8521	0.7768	0.521 ± 0.022
	High–Low Cal	0.9276	0.9176	0.9164	0.4212	
	Prosody	0.4767	0.3458	0.8361	0.7903	0.496 ± 0.085
	High–Low Cal	0.9319	0.9222	0.9119	0.3952	

Table 3: Regression-based results across all model families. Arrows indicate the direction of improvement. Tolerance (± 1) and High–Low Calibration indicate lenient evaluation settings where predictions within the tolerance or predicted uncertainty range are considered acceptable.

Across all three rubrics, Table 3 results indicate consistent improvements with increasing model complexity (i.e. from top to bottom). The multi-head models (MRR.M, MRR.G, MRR.GC) generally outperform single-head regression (SRR.M), showing the benefit of shared representation learning across rubrics. Introducing the GNLL objective (MRR.G)

further improves robustness by jointly modeling prediction uncertainty, and the full MRR.GC configuration achieves the strongest overall performance and highest alignment with human ratings across all evaluation metrics.

Figure 2 presents the confusion matrices for all regression configurations. Overall, all models show tight clustering around mid-level utterances but limited separability at the lowest and highest scores. Predictions across models are concentrated in the mid-range (scores 5–8), while the extremes (1–2 and 10) are rarely or never predicted.

3.3.2. Regression Results with Tolerance and Calibration

Table 3 also summarizes the regression-based results evaluated under a relaxed setting that allows a tolerance of ± 1 score point and, in the final configuration, incorporates conformal calibration. Under this evaluation, predictions within one score point of the human rating are considered acceptable, reflecting natural variability among human raters. When lenient settings are applied, the overall performance increases across all rubrics. The ± 1 tolerance evaluation substantially boosts F1, MCC, and PCC scores, reflecting stable prediction behavior within one rating level of the human mean. In the uncertainty-aware MRR.GC model, conformal calibration achieves a comparable improvement by explicitly modeling confidence intervals instead of tolerance bands.

Figure 2 also illustrates the confusion matrices under lenient evaluation. In the first three panels (SRR.M, MRR.M, and MRR.G), the red boxes mark the ± 1 tolerance region around the diagonal, highlighting predictions within one score point of the human gold standard. The bottom-right panel (MRR.GC) shows calibrated boundaries from conformal prediction, with red contours marking the median empirical high–low range per score bin.

3.3.3. Coverage Analysis under Conformal Calibration

To evaluate how well the calibrated prediction intervals align with the empirical variability among human ratings, we present in Table 4 the percentage of utterances for which a given number of human raters (0–5) fall within the model's predicted high–low interval, obtained from conformal calibration in the MRR.GC configuration.

$\leq N$ raters	Accuracy (%)	Fluency (%)	Prosody (%)
5	6.68	0.96	1.92
4	25.92	8.60	12.72
3	60.04	34.24	44.16
2	83.92	67.20	77.56
1	93.84	91.52	94.68

Table 4: Cumulative percentage with at most N raters within the model's prediction interval (monotonic increasing with N).

4. Discussion

In this section, we first discuss the main findings and then provide an answer to our RQ. To address the RQ, we fine-tuned the Qwen2-Audio-7B-Instruct SpeechLLM using rubric-guided data for multi-aspect assessment of L2 read speech under five different configurations. We observe that performance gradually improves from SRR.M to MRR.M, MRR.G, and MRR.GC (Table 3). Multi-rubric regression improves stability and captures cross-rubric dependencies. Incorporating the GNLL objective in MRR.G further improves calibration by modeling variability in human ratings, in line with computer vision research (Kendall and Gal, 2017). The final MRR.GC, achieves the best overall alignment with human judgments by combining multi-rater supervision and conformal calibration, as proven in other research fields (Braun et al., 2025), which produces adaptive confidence intervals that reflect empirical rater disagreement.

Allowing a tolerance of ± 1 score point improves F1, MCC, and PCC, indicating that minor discrepancies fall within normal perceptual variability (Table 3, red boxes in Fig. 2). However, this margin spans 20% of the 10-point scale, so part of the gain stems from more lenient evaluation. In contrast, conformal calibration in MRR.GC provides a principled way for quantifying uncertainty: interval widths adapt to local variability, yielding statistically valid confidence bounds consistent with human rating behavior. As shown in Table 4, higher coverage for accuracy suggests closer human–model agreement, while narrower intervals for fluency and prosody indicate more consistent rater behavior and tighter calibration around consensus scores.

Human raters show moderate agreement among themselves (Table 1). Furthermore, we observed higher inter-rater reliability for accuracy compared to fluency and prosody (Table 1) and the percentages of human ratings that fall within the model's predicted high–low intervals (Table 4). On the other hand, Table 3 shows that the regression-based results are better for fluency and prosody than for accuracy. Possible explanations for these differences might lie in (a) the definitions of the three aspects of accuracy, fluency, and prosody and (b) the operationalization of these three constructs. As to (a), while the definition of accuracy seems rather straightforward, the correctness of pronunciation, the definitions of fluency and prosody are somewhat confusing. As a matter of fact, they seem to mix several features. For example, speaking

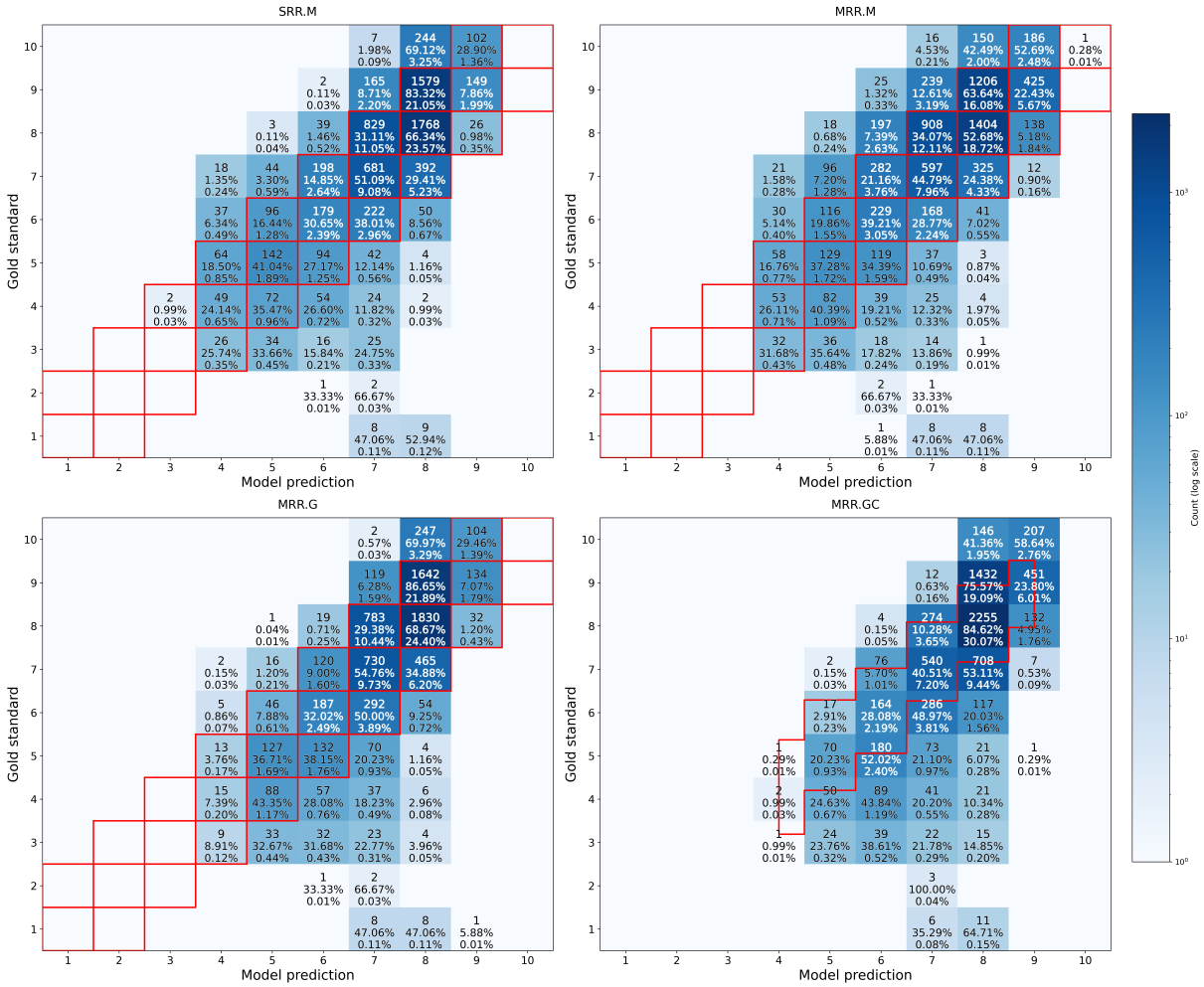

Figure 2: Aggregated confusion matrices across regression methods (SRR.M, MRR.M, MRR.G, MRR.GC). Red boxes in the first three panels indicate the ± 1 tolerance region, while in the bottom-right panel, red contours denote the median calibrated range from conformal prediction. Each cell shows the total count (top), the percentage within the true class (middle), and the percentage relative to all samples (bottom).

rate could just as well be part of fluency instead of prosody, based on the definition of fluency as temporal smoothness. As to (b), the instructions for fluency and prosody may be easier to operationalize for LLMs than those for accuracy. Temporal aspects have long been recognized as easier to compute automatically than features related to segmental quality that may involve multiple dimensions. Previous research on automatic assessment of non-native speech revealed that ASR systems are better at capturing temporal-related aspects of non-native speech than those related to segmental quality (Cucchiarini et al., 2000b,a, 2002).

The main limitations of our work concern data imbalance and generalization. Model predictions tend to cluster around scores 6 to 8, reflecting the skewed distribution of human ratings and reducing sensitivity to extreme proficiency levels. This mid-range bias inflates global performance metrics and constrains the model's ability to generalize to underrepresented proficiency extremes.

Finally, in response to our RQ, the results presented in the current paper demonstrate that the proposed models align closely with human judgments across all rubrics, indicating that a well-designed SpeechLLM can effectively support multi-aspect automatic assessment. Among them, the MRR.GC model performs best, offering the additional advantage of capturing not only mean human ratings but also their variability.

5. Conclusion

This study investigated whether a rubric-guided SpeechLLM can approximate human judgments in the assessment of L2 reading speech. To systematically examine whether incorporating additional configurations can improve robustness and alignment with human judgments, we developed five SOTA modeling strategies of increasing com-

plexity for fine-tuning the Qwen2-Audio-7B-Instruct SpeechLLM, based on insights from related literature. Among the evaluated strategies, MRR.GC achieved the strongest overall alignment with human raters, with aggregated performance across Accuracy, Fluency, and Prosody of PCC ≈ 0.81, RMSE ≈ 0.83, and QWK ≈ 0.50. These results demonstrate that incorporating multi-rater supervision, Gaussian uncertainty modeling, and conformal calibration yields reliable and interpretable scoring for multi-aspect L2 reading speech assessment. However, the model behaves conservatively at score extremes, highlighting the need to address mid-range bias and extend the framework to diagnostic feedback and error localization to improve assessment validity. Future work will extend beyond scoring to diagnostic feedback and error localization for actionable learner guidance.

6. Acknowledgements

This publication is part of the project Responsible AI for Voice Diagnostics (RAIVD) with file number NGF.1607.22.013 of the research programme NGF AiNed Fellowship Grants which is financed by the Dutch Research Council (NWO).

7. Bibliographical References

Anastasios N Angelopoulos and Stephen Bates. 2021. A gentle introduction to conformal prediction and distribution-free uncertainty quantification. *arXiv preprint arXiv:2107.07511*.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460.

Sacha Braun, Eugène Berta, Michael I. Jordan, and Francis Bach. 2025. Multivariate Conformal Prediction via Conformalized Gaussian Scoring. *arXiv preprint arXiv:2507.20941*.

Xinwei Cao, Zijian Fan, Torbjørn Svendsen, and Giampiero Salvi. 2024. A Framework for Phoneme-Level Pronunciation Assessment Using CTC. In *Interspeech 2024*, pages 302–306.

Ya-Ning Chang, JSH Taylor, Kathleen Rastle, and Padraic Monaghan. 2020. The relationships between oral language and reading instruction: Evidence from a computational model of reading. *Cognitive Psychology*, 123:101336.

Lei Chen, Klaus Zechner, Su-Youn Yoon, Keelan Evanini, Xinhao Wang, Anastassia Loukina, Jidong Tao, Lawrence Davis, Chong Min Lee, Min Ma, et al. 2018. Automated scoring of nonnative speech using the speechrater sm v. 5.0 engine. *ETS Research Report Series*, 2018(1):1–31.

Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhifang Guo, Yichong Leng, Yuanjun Lv, Jinzheng He, Junyang Lin, et al. 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*.

Catia Cucchiarini, Helmer Strik, and Lou Boves. 2000a. Different aspects of expert pronunciation quality ratings and their relation to scores produced by speech recognition algorithms. *Speech Communication*, 30(2-3):109–119.

Catia Cucchiarini, Helmer Strik, and Lou Boves. 2000b. Quantitative assessment of second language learners' fluency by means of automatic speech recognition technology. *The Journal of the Acoustical Society of America*, 107(2):989–999.

Catia Cucchiarini, Helmer Strik, and Lou Boves. 2002. Quantitative assessment of second language learners' fluency: Comparisons between read and spontaneous speech. *the Journal of the Acoustical Society of America*, 111(6):2862–2873.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1*, pages 4171–4186.

Heejin Do, Yunsu Kim, and Gary Geunbae Lee. 2023. Score-balanced loss for multi-aspect pronunciation assessment. In *Interspeech 2023*, pages 4998–5002.

Keelan Evanini, Matthew Mulholland, Rutuja Ubale, Yao Qian, Robert A Pugh, Vikram Ramanarayanan, and Aoife Cahill. 2018. Improvements to an Automated Content Scoring System for Spoken CALL Responses: the ETS Submission to the Second Spoken CALL Shared Task. In *Interspeech 2018*, pages 2379–2383.

Sreyan Ghosh, Zhifeng Kong, Sonal Kumar, S Sakshi, Jaehyeon Kim, Wei Ping, Rafael Valle, Dinesh Manocha, and Bryan Catanzaro. 2025. Audio flamingo 2: An audio-language model with long-audio understanding and expert reasoning abilities. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 19358–19405. PMLR.

Sreyan Ghosh, Sonal Kumar, Ashish Seth, Chandra Kiran Reddy Evuru, Utkarsh Tyagi, S Sakshi, Oriol Nieto, Ramani Duraiswami, and Dinesh Manocha. 2024. GAMA: A large audio-language model with advanced audio understanding and complex reasoning abilities. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 6288–6313, Miami, Florida, USA. Association for Computational Linguistics.

Karianne Megard Grønli, Bente Rigmor Walgermo, Erin M McTigue, and Per Henning Uppstad. 2024. Teachers' feedback on oral reading: A critical review of its effects and the use of theory in research. *Educational Psychology Review*, 36(4):121.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3.

Shin'ichiro Ishikawa. 2023. Effects of raters' l1, assessment experience, and teaching experience on their assessment of l2 english speech: A study based on the icnale global rating archives. *LEARN Journal: Language Education and Acquisition Research Network*, 16(2):411–428.

Alex Kendall and Yarin Gal. 2017. What uncertainties do we need in bayesian deep learning for computer vision? *Advances in neural information processing systems*, 30.

Yoon Kim, Horacio Franco, and Leonardo Neumeyer. 1997. Automatic pronunciation scoring of specific phone segments for language instruction. In *Eurospeech*, pages 645–648.

Melanie R Kuhn and Steven A Stahl. 2003. Fluency: A review of developmental and remedial practices. *Journal of educational psychology*, 95(1):3.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach.

Rao Ma, Mengjie Qian, Siyuan Tang, Stefano Bannò, Kate M. Knill, and Mark J.F. Gales. 2025. Assessment of L2 Oral Proficiency using Speech Large Language Models. In *Interspeech 2025*, pages 5078–5082.

Reeta Neittaanmäki and Iasonas Lamprianou. 2024. Communal factors in rater severity and consistency over time in high-stakes oral assessment. *Language Testing*, 41(3):584–605.

Kirsten W Newell, Robin S Codding, and Tara W Fortune. 2020. Oral reading fluency as a screening tool with english learners: A systematic review. *Psychology in the Schools*, 57(8):1208–1239.

Aditya Kamlesh Parikh, Cristian Tejedor-Garcia, Catia Cucchiarini, and Helmer Strik. 2025a. Enhancing GOP in CTC-Based Mispronunciation Detection with Phonological Knowledge. In *Interspeech 2025*, pages 5068–5072.

Aditya Kamlesh Parikh, Cristian Tejedor-Garcia, Catia Cucchiarini, and Helmer Strik. 2025b. Evaluating Logit-Based GOP Scores for Mispronunciation Detection. In *Interspeech 2025*, pages 2405–2409.

Aditya Kamlesh Parikh, Cristian Tejedor-Garcia, Catia Cucchiarini, and Helmer Strik. 2025c. Zero-Shot Speech LLMs for Multi-Aspect Evaluation of L2 Speech: Challenges and Opportunities. *Proc. SLaTE 2025*, pages 11–15.

Nhan Phan, Mikko Kuronen, Maria Kautonen, Riikka Ullakonoja, Anna von Zansen, Yaroslav Getman, Ekaterina Voskoboinik, Tamás Grósz, and Mikko Kurimo. 2025. Mispronunciation Detection Without L2 Pronunciation Dataset in Low-Resource Setting: A Case Study in Finland Swedish. In *Interspeech 2025*, pages 2435–2439.

Paula J Schwanenflugel, Anne Marie Hamilton, Melanie R Kuhn, Joseph M Wisenbaker, and Steven A Stahl. 2004. Becoming a fluent reader: reading skill and prosodic features in the oral reading of young readers. *Journal of educational psychology*, 96(1):119.

Mike Sleeman, John Everatt, Alison Arrow, and Amanda Denston. 2022. The identification and classification of struggling readers based on the simple view of reading. *Dyslexia*, 28(3):256–275.

Grant S Smith and David D Paige. 2019. A study of reliability across multiple raters when using the naep and mdfs rubrics to measure oral reading fluency. *Reading Psychology*, 40(1):34–69.

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. *arXiv preprint arXiv:2310.13289*.

Max van der Velde, Bo Molenaar, Bernard P Veldkamp, Remco Feskens, and Jos Keuning. 2024. What do they say? assessment of oral reading fluency in early primary school children: A scoping review. *International Journal of Educational Research*, 128:102444.

Kuo Wang, Xin Qiao, George Sammit, Eric C Larson, Joseph Nese, and Akihito Kamata. 2024. Improving automated scoring of prosody in oral reading fluency using deep learning algorithm. In *Frontiers in Education*, volume 9, page 1440760. Frontiers Media SA.

Tianrui Wang, Long Zhou, Ziqiang Zhang, Yu Wu, Shujie Liu, Yashesh Gaur, Zhuo Chen, Jinyu Li, and Furu Wei. 2023. Viola: Unified codec language models for speech recognition, synthesis, and translation. *arXiv preprint arXiv:2305.16107*.

Silke Maren Witt. 2000. *Use of speech recognition in computer-assisted language learning*. Ph.D. thesis, University of Cambridge.

Silke Maren Witt and Steve J Young. 2000. Phone-level pronunciation scoring and assessment for interactive language learning. *Speech communication*, 30(2-3):95–108.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Jian Wu, Yashesh Gaur, Zhuo Chen, Long Zhou, Yimeng Zhu, Tianrui Wang, Jinyu Li, Shujie Liu, Bo Ren, Linquan Liu, et al. 2023. On decoder-only architecture for speech-to-text and large language model integration. In *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 1–8. IEEE.

Xiaoming Xi, Derrick Higgins, Klaus Zechner, and David Williamson. 2012. A comparison of two scoring methods for an automated speech scoring system. *Language Testing*, 29(3):371–394.

Klaus Zechner, Lei Chen, Larry Davis, Keelan Evanini, Chong Min Lee, Chee Wee Leong, Xinhao Wang, and Su-Youn Yoon. 2015. Automated scoring of speaking tasks in the test of english-for-teaching (teft™). *ETS Research Report Series*, 2015(2):1–17.

8. Language Resource References

Junbo Zhang and Zhiwen Zhang and Yongqing Wang and Zhiyong Yan and Qiong Song and Yukai Huang and Ke Li and Daniel Povey and Yujun Wang. 2021. *speechocean762: An Open-Source Non-Native English Speech Corpus for Pronunciation Assessment*. ISCA.