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

Aditya Kamlesh Parikh *(aditya.parikh@ru.nl)*, Cristian Tejedor-García, Catia Cucchiarini, Helmer Strik
Center For Language Studies (CLS), Radboud University Nijmegen, The Netherlands

Challenge of L2 Speech Assessment

- **≋ Multi-dimensional task**
  Accuracy · Fluency · Prosody
- **◔ Expensive & slow**
  Expert scoring is costly, and hard to scale
- **◎ High L2 variability**
  Large amount of variation lowers agreement
- **↕ Aspect imbalance**
  Higher agreement on accuracy

SpeechLLM-Based L2 Speech Assessment

- **◍ Rubric-aligned assessment**
  Follow human scoring criteria for assessment.
- **◎ Multi-rater learning**
  Learn from multiple human judgments
- **◉ Uncertainty-aware**
  Predict confidence and variance
- **◌ Calibrated outputs**
  Produce interpretable confidence intervals

Research Question

To what extent can a SpeechLLM approximate human ratings in multi-aspect (accuracy, fluency, and prosody) performance assessment of L2 reading speech?

Experimental Setup

Input
Audio + Rubric Instructions

Base Model
Qwen2-Audio-7BInstruct (LoRA ~10M params)

Output
Sentence Level Scores or Labels for Accuracy, Fluency, Prosody

Dataset
SpeechOcean762
English L2 Speech
L1 Mandarin
5 Raters
Train/Test Split
2500/2500

Model Finetuning Strategies

DiCl: Discrete Classification

Predict **5 ordinal labels**

(Very Poor, Poor, Fair, Good and Very Good)

SRR.M: Single-Rubric Regression (Mean Squared Error)

Predict **continuous score (1–10)**

One model per aspect

MRR.M: Multi-Rubric Regression (MSE)

Jointly predict:
Accuracy Fluency Prosody

One model for all 3 aspects

MRR.G: Multi-Rubric Regression (Gaussian Negative Log-Likelihood)

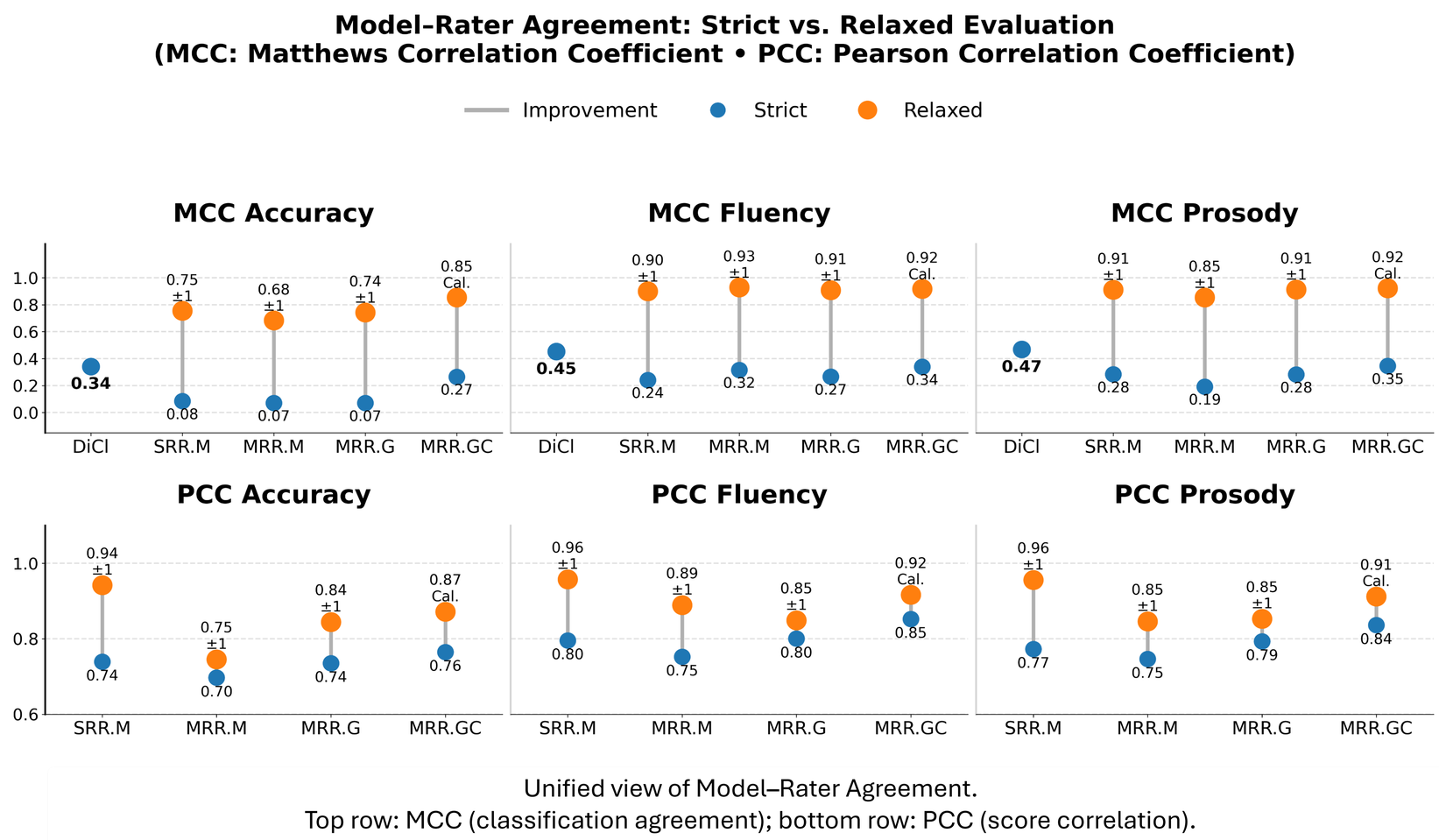
Predict **mean ± variance (μ, σ^2)**

Predict uncertainty and penalizes wrong predictions

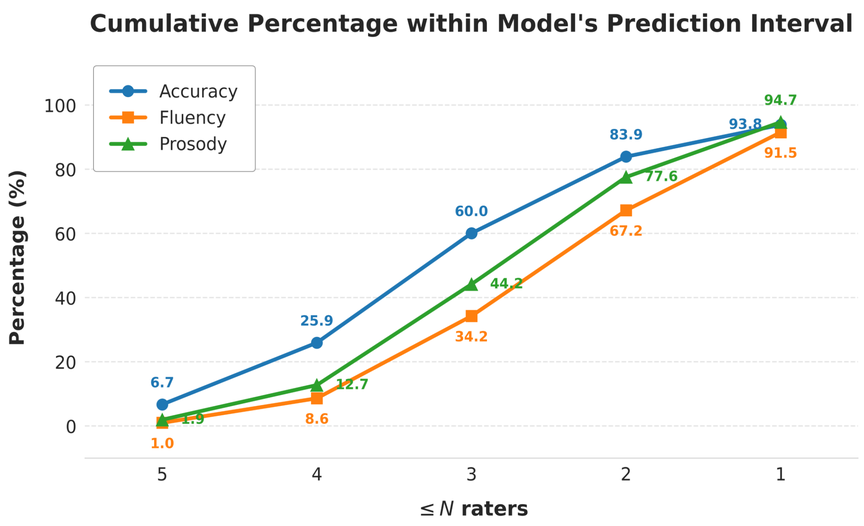
MRR.GC: Multi Rubric Multi Rater Regression with Gaussian Negative Log-likelihood and Conformer Prediction

Learn from **5 raters**
Output [low - high] calibrated interval

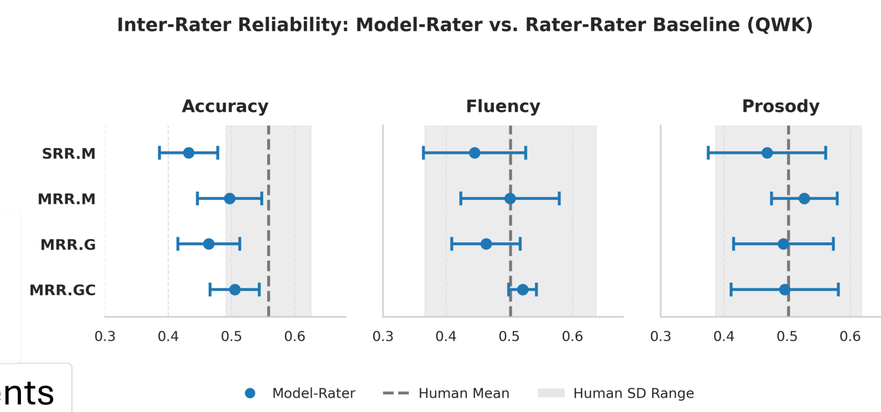
Results



Unified view of Model–Rater Agreement.
Top row: MCC (classification agreement); bottom row: PCC (score correlation).
Lower point = strict evaluation; upper point = relaxed (±1 tolerance or calibrated range).
The vertical gap reflects performance gains under relaxed evaluation.



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



- Multi-rater + calibrated modeling (MRR.GC) yields the strongest alignment with human judgments
- Large gains under ±1 tolerance highlight evaluation sensitivity, not true model capability

Conclusions and Insights

- SpeechLLMs can closely approximate human ratings, especially when modeling multi-rater variability and uncertainty.
- Humans agree more on accuracy, while the model performs better on fluency and prosody due to temporal pattern modeling.
- Conformal calibration provides an alternative to tolerance by learning adaptive confidence intervals.
- Model gains come from capturing human variability, not eliminating it.

NWO
Responsible AI for Voice Diagnostics
AiNed Fellowship Grants
NGF.1607.22.013
(RAIVD project)

Radboud Universiteit