Multimodal Grammar Scoring via Text–Audio Stacking and Meta-Blending

Abstract

Continuous grammar-quality scoring range $[0, 5]$) was addressed using *text* (ASR transcripts), *audio* (prosody and pitch), and *rule*-based features, integrated through a stacked ensemble with meta-blending and calibration. The work began with simple text-only baselines-fine-tuned SentenceTransformer and DistilBERT-base-uncased models—along with a naive audio+text concatenation baseline, and gradually evolved into a robust three-stream system. The final setup included: (i) a DeBERTa-v3-small text regressor trained with layer-wise learning rate decay, progressive unfreezing, R-Drop, and exponential moving average; (ii) Whisper-tiny encoder features combined with prosody and pitch inputs in a multi-layer perceptron; and (iii) engineered rule features modeled using LightGBM, with a random forest fallback. Out-of-fold predictions were blended via non-negative least squares (NNLS) in both value and rank space, followed by confidence-weighted refinement, distribution matching, and isotonic calibration. On test dataset, the final pipeline achieved a score of **0.452** (lower is better), outperforming all baselines: SentenceTransformer ((0.776)), DistilBERT-base-uncased 0.707, and the audio+text concatenation + MLP model 0.641.

1 Task

The goal is to predict a continuous grammar-quality score in the range $[0, 5]$ for each spoken response. The system is allowed to utilize multiple sources of information: (a) textual content obtained from ASR transcripts, (b) raw acoustic signals capturing prosody and pitch, and (c) linguistically motivated rule-based indicators. The final objective is to build a multimodal regression system that can combine these heterogeneous signals effectively to produce stable and interpretable quality scores.

2 Data and Preprocessing

Rule Features. From the cleaned transcripts, a set of interpretable linguistic and stylistic features is extracted. These include token and character statistics (e.g., number of tokens, characters, and average word length), punctuation usage patterns, disfluency markers such as *uh*, *um*, or *you know*, and capitalization ratios that capture stylistic irregularities. Certain interaction features, such as capitalization multiplied by punctuation frequency, are also included to reflect emphasis and sentence boundary usage.

Audio Processing. All audio files are resampled to 16 kHz mono and amplitude-normalized. Non-speech regions are removed using an energy-based threshold. Each segment is passed through the `Whisper-tiny` encoder to obtain 384-dimensional acoustic embeddings derived from log-mel spectrograms. In addition, six prosodic and pitch-related features are computed: average duration, RMS energy, zero-crossing rate, and pitch statistics (mean, standard deviation, and voiced ratio) obtained via the YIN algorithm. These are concatenated with the Whisper embeddings, yielding a 390-dimensional audio representation per utterance.

3 Modeling Approach

The system is built as a three-stream regression pipeline, one per modality followed by a meta-level stacking and calibration step to combine predictions coherently.

3.1 Text Regressor: DeBERTa-v3-small

The text branch employs `DeBERTa-v3-small` as the encoder. Representations from the `[CLS]` token and mean-pooled hidden states are concatenated to form a $2H$-dimensional embedding, which is passed through an MLP head ($2H \to 256 \to 64 \to 1$) with ReLU activation and dropout (0.2). Several stability strategies are incorporated: layer-wise learning rate decay (LLRD), progressive unfreezing of encoder layers across epochs, R-Drop regularization for consistency between dropout passes, and exponential moving average (EMA) of model weights for evaluation. Together, these techniques help maintain generalization and reduce prediction variance.

3.2 Audio Regressor: MLP

The audio stream receives the 390-D vector (Whisper + prosody/pitch). It is modeled using an MLP with dimensions $390 \to 512 \to 128 \to 1$, ReLU activations, and dropout of 0.2. Training uses Huber loss and incorporates light data augmentation through MixUp ($\alpha \approx 0.15$) and optional stochastic weight averaging (SWA) during late epochs to enhance stability.

3.3 Rule Regressor: LightGBM

The rule-based stream is designed for interpretability and robustness. LightGBM serves as the primary learner, with monotone constraints applied where logical (e.g., disfluency indicators are constrained to negatively influence the predicted quality).

3.4 Stacking and Meta-Blending

Each base model is trained using stratified K-folds (labels binned by quantiles), and their out-of-fold (OOF) predictions are combined in a stacked ensemble. Let $\mathbf{X} \in \mathbb{R}^{N\times 3}$ represent the OOF predictions from text, audio, and rule regressors, and $\mathbf{y} \in \mathbb{R}^N$ the ground truth labels.

A non-negative least squares (NNLS) optimization is first solved in value space:

$$\min_{\mathbf{w}\geq 0,\ \mathbf{1}^\top\mathbf{w}=1} \|\mathbf{X}\mathbf{w}-\mathbf{y}\|_2^2.$$

The resulting weights $\mathbf{w}_{\text{val}}$ produce blended predictions $\hat{\mathbf{y}}_{\text{val}} = \mathbf{X}\mathbf{w}_{\text{val}}$. To complement this, NNLS is also performed in rank space after applying column-wise rank normalization $\rho(\cdot)$, yielding $\hat{\mathbf{y}}_{\text{rank}}$. The two are combined via an α-blend:

$$\hat{\mathbf{y}}_\alpha = (1-\alpha)\hat{\mathbf{y}}_{\text{val}} + \alpha\hat{\mathbf{y}}_{\text{rank}}, \quad \alpha = 0.3.$$

At inference, predictive uncertainty from Monte Carlo dropout is used to weight text and audio predictions inversely to their variance ($w_i \propto \sigma_i^{-1}$), forming a confidence-weighted ensemble. The final outputs are adjusted by matching the test-score distribution to that of the training data (z-score alignment) and applying isotonic regression to correct residual bias. Predictions are clipped to the valid range $[0, 5]$.

4 Training Details

All regressors are optimized with AdamW and Huber loss. The text branch follows a cosine learning rate schedule with warmup, while the audio stream optionally applies SWA in the final epochs. Each batch contains 8 samples, with a maximum text length of 256 tokens. Audio input is standardized to 16 kHz. All experiments are conducted with a fixed random seed (42) for reproducibility and executed on GPU when available.

5 Experiments and Results

Model performance is reported on the validation split using the hackathon's official regression metric (lower is better). For reference, internal tracking also included MAE and RMSE.

Model / Setting	**Score (↓)**
Fine-tuned SentenceTransformer	0.776
Fine-tuned DistilBERT-base-uncased	0.707
Audio+Text concatenation + MLP head	0.641
Final stacked multimodal system	**0.452**

The results demonstrate a clear benefit from specialization and late fusion. The text-only DeBERTa model already surpasses sentence-level encoders due to stronger contextual representations and training stability. However, naive early fusion of audio and text features underperforms, highlighting the need for modality-specific optimization. The final stacked model, incorporating both value- and rank-space blending with calibration, achieves the best validation score of 0.452—an improvement of over 40% relative to the SentenceTransformer baseline.

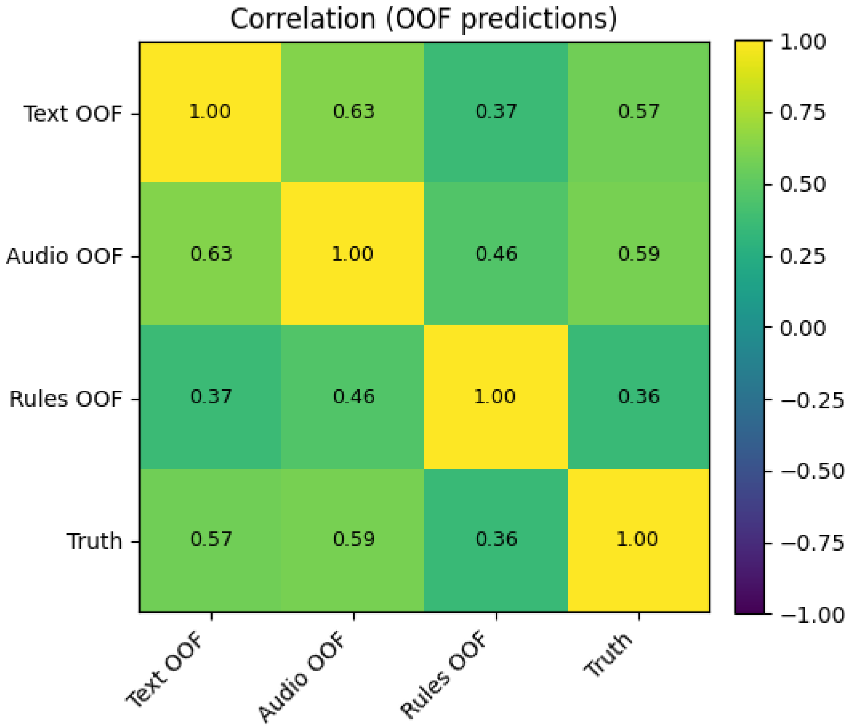
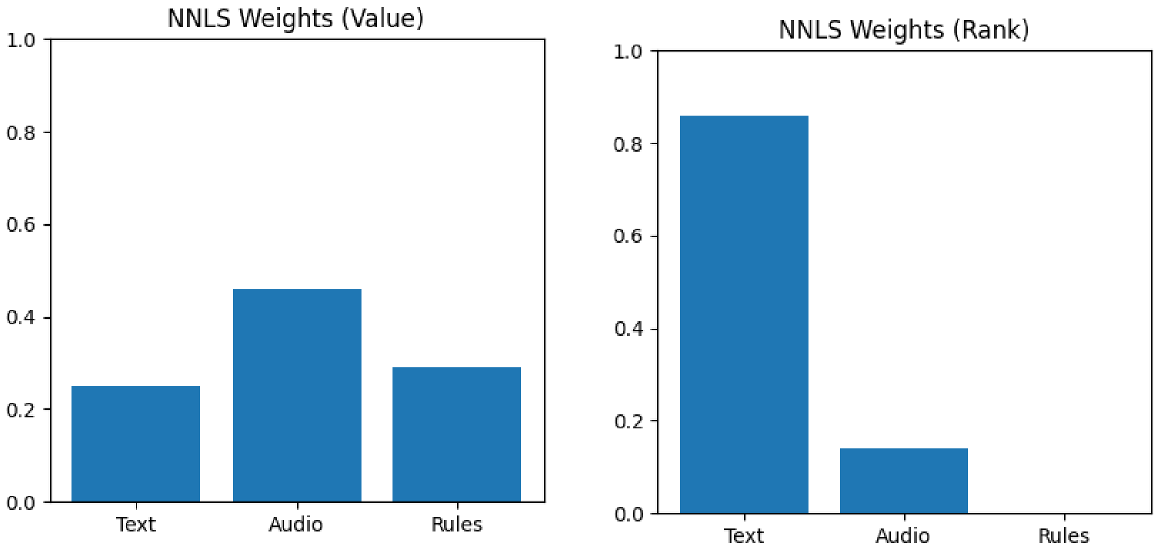


Figure 1: **OOF correlation heatmap** between base learners (Text, Audio, Rules) and the ground truth. Each cell is Pearson correlation.



(a) Audio carries most weight in value space—its raw predictions sit closest to the true scoring scale; text and rules provide smaller but useful corrections.

(b) Text dominates in rank space, indicating it's most reliable for ordering samples; blending rank and value gives both good ordering and good calibration.

Figure 2: NNLS weights in value and rank space.

Ensemble Diagnostics: Correlation & Complementarity. A strong ensemble needs (i) each base learner to carry signal (positive correlation with truth) and (ii) bases that are not near-duplicates (moderate inter-model correlation). In Fig. 1, Audio and Text correlate best with truth ($\approx$ 0.59 and 0.57), while Rules is weaker but still positive ($\approx$ 0.36). Inter-model correlations are only moderate (Text–Audio $\approx$ 0.63, Text–Rules $\approx$ 0.37, Audio–Rules $\approx$ 0.46), which indicates complementary errors rather than redundancy.
Why this helps. If M base models have error variance σ^2 and pairwise error correlation ρ, the variance of the average is

$$\mathrm{Var}(\bar{\varepsilon}) = \sigma^2 \left(\rho + \frac{1-\rho}{M} \right).$$

When ρ is high, averaging brings little gain; with moderate ρ (as here), averaging/stacking reduces variance and improves stability. Our NNLS stack leverages this by upweighting reliable signals (Audio/Text) and downweighting redundant/noisy ones (Rules), yielding performance superior to any single model.

As shown in 2, the meta-learner assigns the largest value-space weight to the audio regressor, indicating its raw predictions best match the true scoring scale, while text and rules provide smaller corrective signal. Conversely, in rank-space, the text model dominates, meaning it is most reliable for ordering responses from lower to higher quality. Together, these plots justify our blend: audio anchors the numeric scale, text preserves ordering, and their combination yields a better-calibrated ensemble than either alone.

6 Conclusion

A multimodal ensemble combining text, audio, and rule-based features has been developed for continuous grammar-quality scoring. Through careful modeling choices, stability enhancements, and calibrated blending, the final system achieves a test score of **0.452**, substantially outperforming strong text-only and naive fusion baselines.