RABIT: Rationale-Based Distillation Towards Interpretable Automatic Speaking Assessment via a Small Language Model

Bi-Cheng Yan, Hong-Yun Lin, Fu-An Chao, Jiun-Ting Li, Berlin Chen
National Taiwan Normal University, Taipei, Taiwan
{bicheng, fuanchao, berlin}@ntnu.edu.tw

Abstract

Automatic speaking assessment (ASA) manages to quantify the language competence of foreign language learners by providing a proficiency score based on their spoken response. Existing efforts in ASA typically employ a neural grader integrated with a set of handcrafted features to assess learners' oral proficiency from multiple facets. Despite decent performance, the black-box nature of these neural graders remains a significant barrier to providing interpretable explanations for the grading results. In light of this, we propose RABIT for ASA, a novel rationale-based knowledge distillation framework for interpretable grading decisions via a small language model. Specifically, RABIT first extracts multi-faceted grading rationales from a large language model (LLM) pertaining to the learner's response and the scoring guidelines. Subsequently, a compact yet efficient language model, equipped with distinct output heads, is jointly optimized to estimate a proficiency score while generating a sequence of grading rationales in an autoregressive manner. A series of experiments conducted on General English Proficiency Test (GEPT) dataset validates the feasibility and superiority of our method over several cutting-edge baselines.

1 Introduction

Spurred by the global demand for foreign language proficiency in both the workforce and academia, there is a growing need for the assessments of language competence (Davis and Norris, 2024). In response, the development of automatic speaking assessment (ASA) systems has garnered significant attention, figuring prominently in the fields of computer-assisted language learning (Zechner and Evanini, 2019) and large-scale language testing (Singla et al., 2021). Such systems offer a broad spectrum of applications to mitigate the disproportion between the limited number of instructors and the expanding population of foreign language learners, ranging from low-stakes contexts (e.g., providing informative feedback for both instructors and learners in course placement (Evanini and Wang, 2013)), to high-stakes scenarios (e.g., serving as a reliable reference for professionals in admission testing (Evanini et al., 2017; Chen and Li, 2016)).

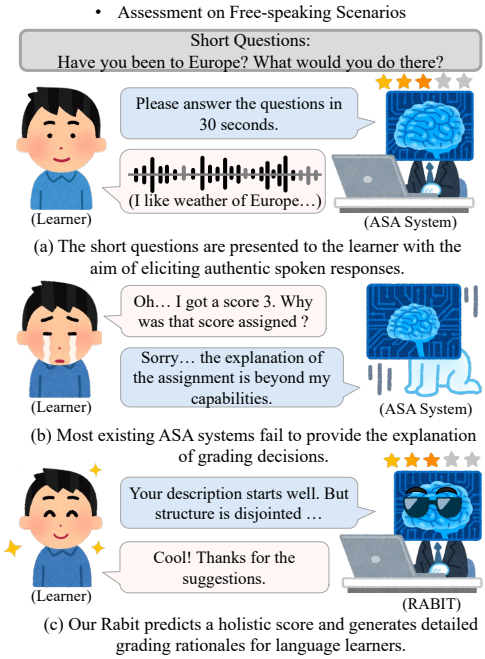

Figure 1: Motivations of our approach. (a) Assessment in the free-speaking scenario, where the ASA models aim to evaluate learner's language proficiency based on the spoken response. (b) Existing ASA systems often operate as a black-box process, lacking transparency in their grading decisions. (c) Our proposed method (RABIT) jointly optimizes proficiency estimation and rationale generation to achieve transparent and reliable assessment results.

ASA seeks to quantify the extent of language proficiency in foreign language learners across multiple dimensions, including delivery (fluency and pronunciation), language use (vocabulary and grammar), topic development (content and discourse) (Qian et al., 2019; Zechner and Evanini,

Proceedings of the 21st Workshop on Innovative Use of NLP for Building Educational Applications, pages 108–117

2019), and others. One of the de-facto archetypes of ASA is instantiated in free-speaking scenarios, as shown in Figure 1(a), where the language learner is presented with short, open-ended questions and instructed to respond based on their personal experiences or opinions. A leading strand of research in ASA derives a set of handcrafted features from the synergy of learners' spoken responses and the presented short questions. These features are then fed into a neural grader to predict either a holistic score (i.e., a categorical value of overall speaking proficiency) or analytic scores (i.e., continuous numerical scores for specific aspects). Specifically, to characterize learners' pronunciation clarity and spoken delivery, commonly used features include confidence scores of recognized linguistic units (e.g., words or phones), time-alignment information (e.g., speaking rate, pause frequency, and filled pauses), and statistical measures of fluency and pitch contours (Zechner et al., 2009; Chen et al., 2010). Grammatical accuracy and syntactic complexity are captured by the transcriptions of the learner response through part-of-speech tagging, syntactic dependency parsing, and morphological analysis (Moore et al., 2015; Qian et al., 2018). Lastly, content coherence and topical relevance are assessed by evaluating the learner response and the presented short questions, where the BERT-based semantic features combined with graph-based neural networks have been extensively investigated (Li et al., 2023; Bannò and Matassoni, 2023; Singla et al., 2021; Li et al., 2024).

Although existing efforts have achieved commendable performance on various spoken assessment tasks by leveraging a set of handcrafted features to qualify learners' language competence, the opacity of scoring decisions remains a crucial challenge. The interpretability of scoring decisions is nearly sidelined, which might undermine the efficiency of language acquisition and diminish pedagogical utility, as depicted in Figure 1(b). In light of this, we propose a novel two-step knowledge distillation framework for ASA, termed rationale-based distillation for interpretable grading decisions (RABIT). By harnessing the reasoning capabilities of large language models (LLMs), RABIT first distills grading rationales into a lightweight yet effective small language model, which then in turn learns to generate interpretable, fine-grained scoring decisions while assessing the language proficiency of foreign language learners. To explicitly illustrate how the learner responses align with the scoring rubrics, RABIT first prompts an LLM (e.g., LLaMA3.1-8B and LLaMA3.2-3B (Touvron et al., 2023)) with systematic rubric guidelines to elicit multi-faceted grading rationales conditioned on the transcriptions of learner responses and the associated short questions. Afterwards, a small language model (e.g., SmolLM2-135M-Instruct (Allal et al., 2025)) augmented with two distinct output heads is jointly optimized to predict proficiency scores and generate corresponding grading rationales in an autoregressive manner. Extensive experiments on the General English Proficiency Test (GEPT) dataset demonstrate that our methods consistently outperform several competitive baselines. Beyond performance gains, exploring a small LLM to generate grading rationales for foreign language learners introduces an additional layer of transparency to the field of ASA research.

In summary, the main contributions of this work are at least three-fold: (1) As one step towards interpretable grading decisions, RABIT is among the first attempts to evaluate language proficiency of foreign language learners while offering grading rationales, opening up a new research avenue for ASA; (2) Furthermore, to adapt small language models for ASA, we introduce prompt vectors in conjunction with a score prediction head to stabilize proficiency score estimation while maintaining the computational efficiency; (3) Extensive experiments on a realistic spoken assessment dataset, comprising authentic oral responses and expert annotations, confirm the practical and pedagogical utilities of RABIT.

2 Rationale-based Distillation for Interpretable Grading Decisions (RABIT)

This section sets out with a problem definition of ASA task and then sheds light on the proposed knowledge distillation framework, rationale-based distillation for interpretable grading decisions (RABIT). Our approach aims to improve the transparency and the practical utility of ASA through two pivotal steps, namely LLM rationale probing and small language model training, as illustrated in Figure 2(a). In the first step of RABIT, we leverage an LLM to generate multi-faceted grading rationales by conditioning on the transcriptions of learner responses, the short questions, the rubric guidelines, and the expert-annotated proficiency score. Subsequently, in the second step, these dis-

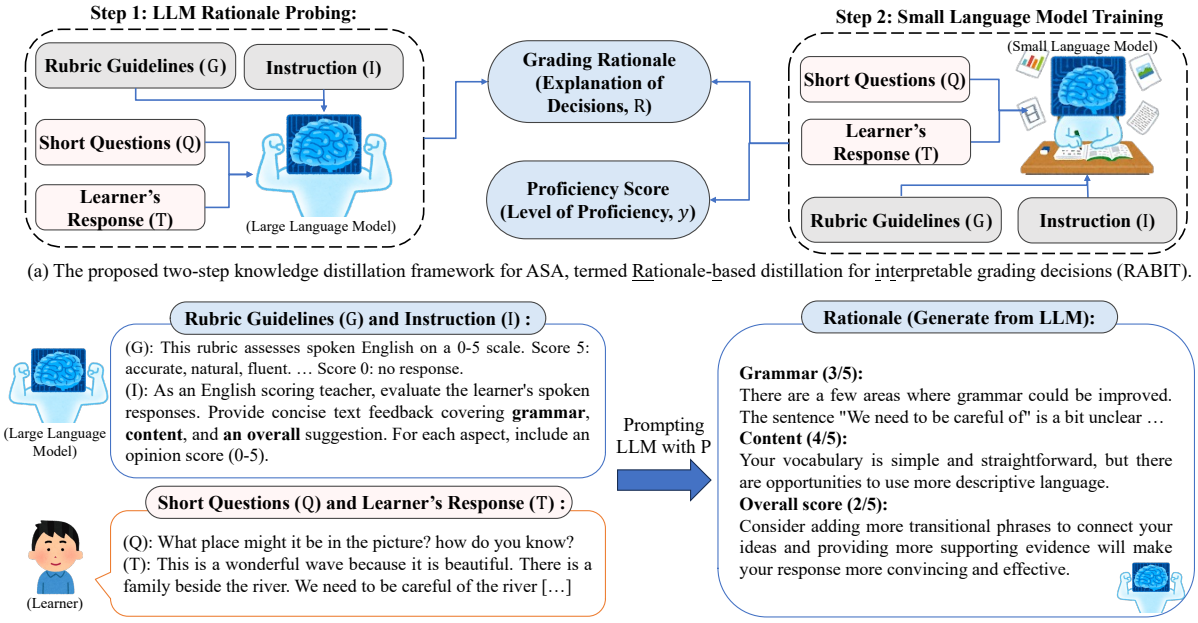

Figure 2: Overview of the proposed RABIT framework for interpretable knowledge distillation. (a) The two-stage training process, comprising LLM rationale probing and small language model (SLM) distillation. (b) A running example illustrating the rationale generation and distillation process in the first stage of RABIT.

tilled rationales serve as additional supervision for a small language model, which is jointly trained to estimate proficiency scores while producing the corresponding grading rationales.

2.1 Problem Definition

Given a sequence of short questions Q, a foreign language learner is instructed to produce a spoken response X that reflects their personal experiences or opinions. An ASA system is then tasked with estimating a categorical proficiency score $y \in \{1, 2, 3, 4, 5\}$ for X, where y denotes a distinct level of language competence across multiple facets including delivery, language use, and topic development.

2.2 Step1 of RABIT: LLM Rationale Probing

For each instance of training data, consisting of a spoken response X paired with the short questions Q and the corresponding proficiency score y , we first transcribe X into a sequence of words T via a speech recognizer (e.g., Whisper-large-v3 (Radford et al., 2023)). Subsequently, a structured prompt template P is curated to elicit a sequence of grading rationales R from an LLM. The prompt template $\mathrm{P} = [\mathrm{G}; \mathrm{I}; \mathrm{Y}; \mathrm{Q}; \mathrm{T}]$ comprises the following components: the rubric guidelines G, which outline scoring rubrics used by human experts; the instruction I, which describes the role of the LLM in generating multi-faceted proficiency scores along with explanations; the score reference Y, which specifies y for the LLM to guide rationale elicitation; and the short questions Q coupled with the transcribed response T. The first step of RABIT is illustrated with a running example in Figure 2(b). Notably, since the rationale generation relies on the ASR transcriptions, the instruction I positions LLM as a language instructor to elicit multi-facted grading decisions from a textual perspective, focusing on the grammatical accuracy, content relevance, and overall proficiency score. To ensure the reliability of distilled rationales, we filter out instances where the LLM-generated scores conflict with y. In preliminary experiments, we evaluated the feasibility of large multimodal language models (e.g., Phi-4-multimodal (Abouelenin et al., 2025), and QwenAudio (Chu et al., 2023)) to generate grading rationales directly from the speech signal X and the task instruction I. However, these models are prone to generating hallucinated feedback. We attribute this to the scarcity of expert-annotated ASA data and its underrepresentation during the training of these multimodal foundation models.

2.3 Step2 of RABIT: Small Language Model Training

Although LLMs have demonstrated remarkable performance on various NLP benchmarks, the high computational overhead hinders practical deployment in real-world ASA applications, particularly on resource-limited edge devices such as mobile phones or personal computers. To this end, following the extraction of rationales in Step 1, we fine-tune a small language model (e.g., SmolLM2 (Allal et al., 2025)) in Step 2 to builds a computationally efficient and interpretable ASA model. As illustrated in Figure 3, we employ a small language model as the backbone, which jointly learns to predict the proficiency score and generate grading rationales based on a set of handcrafted features via parameter-efficient instruction tuning (Hu et al., 2022). Our interpretable ASA model consists of three components: proficiency-specific encoders, proficiency-specific projectors, and a small language model. In the following, we delve into the functional details of each component.

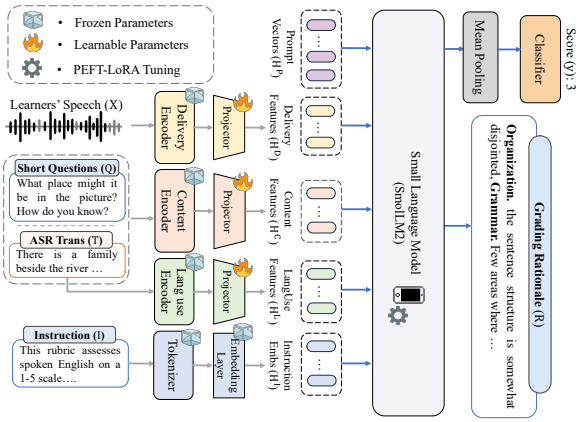

Figure 3: The proposed multi-modal ASA model in the second stage of RABIT, which jointly estimates proficiency scores and generates corresponding grading rationales.

Proficiency-specific Encoders. To capture the supra-segmental pronunciation cues in the learners' speech, we leverage a pre-trained audio encoder (e.g., Wav2vec2.0 (Baevski et al., 2020)) for delivery feature extraction (Lo et al., 2024; Chao et al., 2022; Yan et al., 2025), which takes the speech signal X as input and maps it to a sequence of high-level acoustic representations $\mathrm{H}_0^{\mathrm{D}}$. The language-use features $\mathrm{H}_0^{\mathrm{L}}$ are derived from one-hot encodings of part-of-speech tags, syntactic dependency labels, and morphological features, all of which are linearly projected and packed together via individual linear transformations to form a unified representation (Peng et al., 2024). To capture the content coherence between a learner's response and short questions, a pre-trained text encoder (e.g., BERT (Devlin et al., 2019)) is adopted to extract content features $\mathrm{H}_0^{\mathrm{C}}$ by formatting an input text sequence as the concatenation of short questions Q and the transcribed response T, delimited by the [SEP] token to model their inter-textual relationships.

Proficiency-specific Projectors. To bridge the proficiency-specific encoders and the small language model, the projectors map the respective proficiency features into token representations, aligning the proficiency features with the latent space of the small language model. In this work, we employ a convolutional projector that encapsulates fine-grained linguistic cues into compact token representations with a Macaron-style neural structure, where a Transformer block is sandwiched between two separate depth-wise convolutional layers. Each convolution layer incorporates a GELU activation function ((Lee, 2023)) and a kernel size of 3. For down-sampling, the first layer utilizes a kernel stride of $k/2$, and the second employs a kernel stride of k. We compress the delivery, language-use, and content features along the temporal dimension at reduction rates of 64, 32, and 32, respectively (i.e., $k = 16$, 8, and 8).

Small Language Model. Conditioned on a set of handcrafted features extracted from proficiency-specific encoders, our language model aims to generate a sequence of grading rationales and estimate proficiency scores for language learners. Empirically, we found that training a small language model to output both the rationale sequence and a proficiency score within a single natural language sequence often leads to inconsistent output formats. In contrast to prior works (Chu et al., 2025; Hsieh et al., 2023), our approach introduces learnable prompt vectors H^{P} to the inputs, which are integrated with the handcrafted features for the small language model. Following forward propagation through the language model, the outputs of these prompt vectors are pooled by an averaging operation and subsequently passed to a classification head to predict the proficiency score y. A bit more terminology: The input sequence to our language model, denoted as $\mathrm{H} = [\mathrm{H}^{\mathrm{D}}; \mathrm{H}^{\mathrm{L}}; \mathrm{H}^{\mathrm{C}}; \mathrm{H}^{\mathrm{I}}; \mathrm{H}^{\mathrm{P}}]$, comprises prompt vectors H^{P}, instruction embeddings H^{I} and a set of handcrafted features, where $\mathrm{H}^{\mathrm{D}}, \mathrm{H}^{\mathrm{L}}, \mathrm{H}^{\mathrm{C}}$ are the projected features derived from

the delivery, language-use, and content coherence representations (i.e., $\mathrm{H}_0^{\mathrm{D}}$, $\mathrm{H}_0^{\mathrm{L}}$, and $\mathrm{H}_0^{\mathrm{C}}$) via their respective projectors. H^{I} denotes the textual representation of the instruction I, projected through the embedding layer of the small language model. Our ASA model employs a small language model parameterized by θ to generate the grading rationales $\mathrm{R} = (r_1, r_2, \cdots, r_N)$ by minimizing the negative log-likelihood:

$$\mathcal{L}_{rat} = -log\mathcal{P}_\theta(\mathrm{R}|\mathrm{H}), \quad (1)$$

$$= -\sum_{i=1}^{N} log\mathcal{P}_\theta(r_i|\mathrm{R}_{0:i-1}, \mathrm{H}), \quad (2)$$

where $\mathrm{R}_{0:i-1}$ denotes previously generated tokens. On a separate front, our model also predicts a proficiency score $\hat{y}$ via a classification head. To stabilize the score prediction, the classifier relies on the mean-pooled representation $\bar{\mathbf{h}}^{\mathrm{P}}$, derived by applying an average pooling operation to the output representations of prompt vectors H^{P} after processing through the language model. Proficiency score estimation is optimized by minimizing the cross-entropy loss between the distributions of predicted score $\hat{\mathbf{y}}$ and the ground-truth score y (represented as a one-hot vector $\mathbf{y}$):

$$\mathcal{L}_{score} = -\sum_{c=1}^{5} \mathbf{y}_c log\hat{\mathbf{y}}_c, \quad (3)$$

$$\hat{\mathbf{y}} = \mathcal{P}_\theta(\hat{\mathbf{y}}|\bar{\mathbf{h}}^{\mathrm{P}}) = softmax(\mathrm{W}\bar{\mathbf{h}}^{\mathrm{P}} + \mathbf{b}), \quad (4)$$

where W and $\mathbf{b}$ are learnable weights of the classification head. In the training phase, we keep the small language model frozen and integrate a LoRA module to minimize the following objective function:

$$\mathcal{L} = \mathcal{L}_{score} + \alpha\mathcal{L}_{rat}, \quad (5)$$

where $\alpha \in [0, 1]$ is a control parameter balancing the losses of rationale generation and proficiency score estimation.

3 Experimental Setups

GEPT Dataset. In this paper, a series of experiments were carried out on an in-house dataset derived from General English Proficiency Test (GEPT), a standardized assessment administered by the Language Training and Testing Center (LTTC) in Taiwan. The GEPT dataset targets English learners across various proficiency levels, comprising listening, reading, writing, and speaking.

Dataset		Human-rated Score				
		S1	S2	S3	S4	S5
Train		4	61	317	310	27
Dev		0	6	39	39	6
Test	Seen Prompt	0	11	37	39	3
	Unseen Prompt	0	4	142	137	17

Table 1: Statistics of Picture-description Section in GEPT Dataset.

For our ASA experiments, we specifically focus on the picture description section of the GEPT dataset, in which foreign language learners are presented with a picture and instructed to answer short questions. This subset consists of 1,199 recordings, each paired with an expert-annotated proficiency scores, the short questions, learner response and the corresponding pictures. Each spoken response was independently evaluated by two qualified experts adhering to the same standardized rubric. To derive a single categorical proficiency score, we averaged the two ratings and applied the floor function. The GEPT dataset was partitioned into training, development, and seen-prompt test sets in an 8:1:1 ratio. An isolated unseen-prompt test set comprising 300 recordings was reserved to evaluate the generalization capability of the proposed model to novel picture description learning scenario. The dataset statistics are provided in Table 1.

Implementation Detailed. In the first step of RABIT, we employed LLaMA-3.1-8B-Instruct for rationale generation. In the second step, we used SmolLM2 model family as the backbone for our interpretable ASA model. To extract various handcrafted features, we adopted Wav2vec2.0[1] to extract delivery features. Content coherence was modeled using a BERT-based text encoder[2], while language-use features were derived from word-level linguistic units. Following Peng et al. (2024), language-use features include POS tags, dependency labels, and morphological features processed by the spaCy toolkit[3] based on the transcription of learner responses. As for the model training, we employed the Adam optimizer with a learning rate

[1]https://huggingface.co/facebook/wav2vec2-large-xlsr-53

[2]https://huggingface.co/google-bert/bert-base-uncased

[3]https://spacy.io/

Model	Seen-Prompt Test			Unseen-Prompt Test		
	Accuracy	Weighted-F1	Macro-F1	Accuracy	Weighted-F1	Macro-F1
$^{+}$Qian2019	55.56	54.80	–	64.33	62.80	–
$^{+}$SAMAD	65.56	64.80	–	69.67	68.40	–
Content Model	60.44 (±3.20)	59.41 (±3.30)	48.28 (±3.85)	67.00 (±1.39)	65.14 (±0.99)	38.56 (±3.20)
Delivery Model	52.00 (±2.88)	48.41 (±4.70)	31.72 (±4.85)	52.33 (±2.68)	49.42 (±2.81)	26.89 (±1.35)
Multi-ASA	63.56 (±2.88)	62.85 (±2.92)	44.60 (±5.15)	67.00 (±5.15)	65.37 (±4.65)	41.24 (±4.49)
Multi-ASA-Rat	66.89 (±1.64)	66.31 (±1.90)	45.07 (±3.70)	68.40 (±1.83)	66.67 (±2.31)	38.67 (±3.44)
RABIT	**70.00** (±1.11)	**69.09** (±1.10)	**56.54** (±6.97)	**71.66** (±1.52)	**69.71** (±1.70)	**47.39** (±1.47)

Table 2: Performance comparison with state-of-the-art methods on the GEPT dataset in terms of Accuracy (%), Weighted F1-score (%), and Macro F1-score (%). Standard deviations over 3 runs are reported in parentheses.$^{+}$Results for Qian2019 and SAMAD are reported in Peng et al. (2024).

of 0.001 and a batch size of 3. A warm-up strategy was applied, initiated with 1/100 of the maximum learning rate, and adjusted over steps with a cosine scheduler. All experiments were implemented with PyTorch and executed on a single NVIDIA 2080Ti GPU. To ensure reproducibility and mitigate the effects of randomness, we conducted three independent trials with different random seeds. Each model was trained for 100 epochs, and performance is reported as the average across the three runs. The optimal checkpoint for each trial was selected based on the minimum validation cross-entropy loss. We are committed to making all code, configurations, and logs available upon acceptance.

Evaluation Metric. We evaluate the performance of our model using the following metrics: 1) Classification accuracy (%), representing the percentage of correctly predicted proficiency levels ranging from 1 to 5; 2) Macro F1-score (%), which averages the F1-scores across all proficiency levels, treating each class with equal importance; and 3) Weighted F1-score (%), which weights the F1-score of each proficiency score by its support (number of true instances) to account for label imbalance. The weighted F1-score is particularly crucial for multi-class classification tasks involving imbalanced or skewed labels, as it offers a more representative evaluation of overall performance.

Comparative Methods. To evaluate the performance of our proposed framework, we compare RABIT with the following state-of-the-art (SOTA) methods. 1) Conventional methods: **Qian2019** is an iconic ASA model that extracts various handcrafted features from transcriptions and the corresponding time-alignment information derived from learner speech (Qian et al., 2019). Subsequently, **SAMAD** advances the neural architecture of Qian2019 by incorporating Transformer blocks to capture long-term dependencies of linguistic cues, while introducing soft-label optimization to mitigate the label imbalance problem in ASA model training (Peng et al., 2024). 2) Multi-faceted neural graders: **Multi-ASA** employs the same proficiency-specific encoders as RABIT to extract delivery, language-use, and content-related features. In contrast to the use of small language model for proficiency assessment, Multi-ASA adopts a simple concatenation for feature fusion. The fused features are subsequently passed through a linear projection layer followed by a mean pooling operation to predict the proficiency score. Moreover, to assess the efficacy of grading rationales in ASA, **Multi-ASA-Rat** integrates grading rationales as auxiliary input. We leverage a BERT encoder to extract semantic representations from the grading rationales, which are then fused with handcrafted features within the Multi-ASA to collaborative proficiency enhance proficiency score estimation. 3) Single-faceted neural graders: To analyze the contribution of each modality-specific encoder in RABIT, we report the assessment performance using only the content encoder (**Content Model**) and delivery encoder (**Delivery Model**). Both models extract the corresponding features and assess the proficiency score via simple mean pooling followed by a linear classifier.

4 Experimental Results

Main Results and Performance Benchmarking. At the outset, we compare the proposed approach (viz. RABIT) with existing state-of-the-art ASA methods. From the results shown in the Table 2, we have the following observations. 1) The proposed RABIT consistently outperforms all baseline methods on both seen- and unseen-prompt test sets in terms of accuracy and F1-scores, demonstrating that RABIT surpasses conventional baselines as well as single- and multi-faceted neural graders, while highlighting the strong generalization capability of our framework across diverse evaluation scenarios. 2) Regarding multi-faceted neural graders, Multi-ASA-Rat achieves notable performance gains over its base model (Multi-ASA). By incorporating LLM-generated grading rationales as auxiliary inputs, the results demonstrate that semantic encodings of rationales are highly complementary to other handcrafted features, leading to enhanced assessment accuracy in ASA. In contrast, RABIT formulates rationale generation as a primary learning objective and predicts language proficiency through a small language model, which not only achieves superior performance compared to Multi-ASA-Rat but also significantly enhances the transparency of the assessment results. Furthermore, in a comparison within the single-faceted neural graders, it is evident that the Content Model achieves better results than the Delivery Model. This finding suggests that content coherence exerts a greater influence than spoken delivery in ASA, aligning with observations reported in previous studies (Qian et al., 2019; Bannò and Matassoni, 2023). Moving beyond individual model components, Multi-ASA integrates handcrafted features from content and delivery, alongside language-use features, into a unified neural architecture, resulting in superior performance over its single-faceted counterparts (viz. Content and Delivery models). 3) As to the conventional methods, SAMAD exhibits strong performance on both test sets, significantly outperforming Qian2019 and achieving competitive results to Multi-ASA-Rat. Moreover, SAMAD achieves superior results on the test set with unseen-prompts compared to other baseline methods. These performance gains primarily stem from its delivery modeling, which leverages ASR transcriptions coupled with the spoken signals to capture prosody and fluency of learner speech, effectively boosting the generalization capability across evaluation scenarios.

Setting	Seen Prompt		Unseen Prompt	
	Acc	wF1	Acc	wF1
Number of Prompts				
1	69.63	68.87	70.66	69.06
3*	**70.00**	**69.09**	**71.66**	**69.71**
5	68.51	67.64	71.44	69.49
Combination Weight for Rationale Generation (In a 3-prompt vector setting)				
0.0	66.67	66.18	70.33	68.85
0.3	69.63	68.71	71.22	69.28
0.5*	**70.00**	**69.09**	**71.66**	**69.71**
0.7	69.63	68.71	70.89	68.70
Impact of Base LM Choice (SmolLM2) (In a 3-prompt vector and weight setting of 0.5)				
135M-Instruct*	**70.00**	**69.09**	**71.66**	**69.71**
360M-Instruct	68.15	67.45	71.56	69.47

Table 3: Ablations of parameters in RABIT, reporting accuracy (Acc, %) and weighted F1-score (wF1, %). *Denotes the setting reported in Table 2.

Ablation Study on RABIT Parameters. In this section, we perform an ablation study to systematically examine the assessment performance of RABIT under various parameter configurations on both seen- and unseen-prompt test sets. We first investigate the impact of the number of prompt vectors (i.e., $|\mathrm{H}^{\mathrm{P}}| \in \{1, 3, 5\}$) employed in RABIT. The corresponding results are summarized in the first part of Table 3. Notably, RABIT facilitates proficiency score prediction through the integration of prompt vectors into the model input, rather than relying solely on natural language generation. As the number of prompt vectors increases, we observe performance gains on the unseen-prompt test set; however, these improvements remain marginal. Conversely, a noticeable decline in performance is evident on the seen-prompt test set. To strike an optimal balance between generalization across test sets and computational efficiency, we configure the RABIT with $|\mathbf{H}^{\mathrm{P}}| = 3$.

Subsequently, we investigate the combination weight α that balances the two learning objectives (i.e., proficiency estimation $\mathcal{L}_{score}$ and grading rationale generation $\mathcal{L}_{rat}$), as defined in Eq. (5). The weight parameter α is selected from

$\{0.3, 0.5, 0.7\}$. Furthermore, $\alpha = 0.0$ is included as a reference baseline variant, representing the performance without rationale guidance. As demonstrated in the second part of Table 3, our results suggest that the rationale generation task is crucial for the performance of ASA; when this task is omitted (i.e., $\alpha = 0.0$), the corresponding performance is significantly inferior to all other settings with rationale guidance ($\alpha > 0.0$). Furthermore, we find that the choice of combination weight has a relatively modest impact on overall performance. Increasing α from 0.3 to 0.7 produces only marginal gains on the seen-prompt test set, while inducing a slight performance degradation in unseen-prompt test set.

Finally, the third part of Table 3 examines how the scale of the backbone language model impacts ASA performance. Notably, our results demonstrate that RABIT, when configured with SmolLM-135M-Instruct, consistently outperforms its larger 360M counterpart across all evaluation metrics. This finding suggests that for RABIT, a compact model offers a more balanced model capacity, which is sufficient to capture grading nuances on seen prompts while effectively generalizing to the unseen-prompt test set. Considering both predictive power and computational efficiency, RABIT adopts SmolLM2-135M-Instruct as its default backbone model.

Qualitative Examination of Grading Rationales. Building on the evidence that rationale generation enhances the ASA performance of RABIT, we further conduct a qualitative comparison to evaluate the interpretability of the grading rationales generated by RABIT against those from the teacher LLM (viz. the source LLM used for distillation). As shown in Figure 4, the majority of the grading rationales produced by RABIT and the source LLM share similar semantic content (sentences highlighted in green). However, as illustrated by the sentences marked in purple, RABIT provides grammatical feedback by identifying errors rather than generating full corrected sentences. Notably, RABIT retains the emergent capabilities inherited from its base model, SmolLM2, allowing it to offer nuanced grading opinions or instructive feedback on the spoken response (sentences highlighted in orange). Moreover, as the rationale generation process is conditioned on the expert-annotated proficiency score, the generated rationales and proficiency scores exhibit a high degree of consistency. These qualitative results confirm that RABIT not only effectively captures the grading rationales of the teacher model but also preserves the intrinsic reasoning capabilities of the underlying small language model.

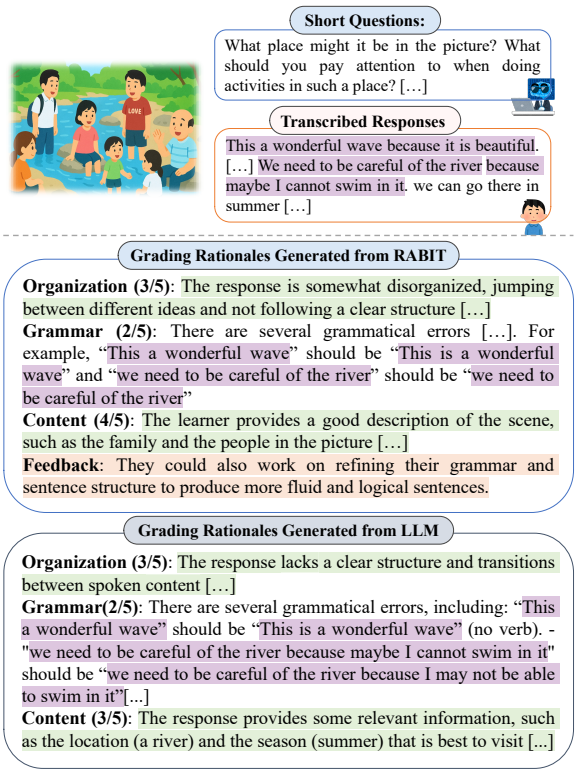

Figure 4: Qualitative comparisons of rationales generated by RABIT and the source LLM. Sentences highlighted in green indicate semantic consistency with the rationales generated by the LLM, while those in purple reflect semantic divergence. Sentences in orange highlight emergent capabilities exhibited by RABIT.

5 Conclusion and Future Work

In this paper, we presented an ASA model designed to provide a transparent grading process without compromising computational efficiency. This advancement highly broadens the scope of ASA system in the real-world applications. As a pivotal step toward interpretable assessment, we proposed RABIT, a rationale-based distillation framework that produces grading decisions while simultaneously estimating the proficiency scores of language learners via a small language model. To enhance the stability of proficiency score estimation, we introduced prompt vectors for the small language model, making a pioneering effort within the field of ASA. Extensive experiments on the picture-description section of the GEPT dataset validate the superiority of RABIT, revealing that integrating rationale generation during the learning process further elevates ASA performance.

Limitations

The proposed method is limited by the quality of the rationales generated from LLMs. Additionally, the visual information contained in the prompts (i.e., pictures) is nearly sidelined in current work. Despite this minimal integration of visual data, RABIT remains consistent with the standard ASA rubrics of the GEPT. These scoring rubrics prioritize spoken delivery and language use, while treating topic development as only a secondary facet of the overall proficiency score. To address these limitations, future work will investigate integrating visual features from the picture prompts and improving the reliability of rationale generation through multi-agent prompting or human feedback mechanisms. Furthermore, we plan to enhance the diversity of the generated rationales of RABIT by leveraging back-translation techniques (Shen et al., 2025).

Ethics Statement

This study adheres to the ACL Code of Ethics. We utilized anonymized datasets provided by LTTC under authorized access, ensuring that data curation aligns with ACL standards. While RABIT provides automated proficiency assessments, it is designed to supplement human raters and provide informative feedback for language learners and instructors in low-stakes learning scenarios. We also acknowledge potential biases inherent in teacher LLMs and encourage cautious interpretation of automated feedback.

References

Abdelrahman Abouelenin, Atabak Ashfaq, Adam Atkinson, Hany Awadalla, Nguyen Bach, Jianmin Bao, Alon Benhaim, Martin Cai, Vishrav Chaudhary, Congcong Chen, and 1 others. 2025. Phi-4-mini technical report: Compact yet powerful multimodal language models via mixture-of-loras. *arXiv preprint arXiv:2503.01743*.

Loubna Ben Allal, Anton Lozhkov, Elie Bakouch, Gabriel Martín Blázquez, Guilherme Penedo, Lewis Tunstall, Andrés Marafioti, Hynek Kydlíček, Agustín Piqueres Lajarín, Vaibhav Srivastav, and 1 others. 2025. Smollm2: When smol goes big–data-centric training of a small language model. *arXiv preprint arXiv:2502.02737*.

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 (NIPS)*, 33:12449–12460.

Stefano Bannò and Marco Matassoni. 2023. Proficiency assessment of l2 spoken english using wav2vec 2.0. In *IEEE Spoken Language Technology Workshop (SLT)*, pages 1088–1095. IEEE.

Fu-An Chao, Tien-Hong Lo, Tzu-I Wu, Yao-Ting Sung, and Berlin Chen. 2022. 3m: An effective multi-view, multi-granularity, and multi-aspect modeling approach to english pronunciation assessment. In *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pages 575–582. IEEE.

Lei Chen, Keelan Evanini, and Xie Sun. 2010. Assessment of non-native speech using vowel space characteristics. In *IEEE Spoken Language Technology Workshop*, pages 139–144. IEEE.

Nancy F Chen and Haizhou Li. 2016. Computer-assisted pronunciation training: From pronunciation scoring towards spoken language learning. In *Proceedings of Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)*, pages 1–7. IEEE.

SeongYeub Chu, Jong Woo Kim, Bryan Wong, and Mun Yong Yi. 2025. Rationale behind essay scores: Enhancing s-llm's multi-trait essay scoring with rationale generated by llms. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 5796–5814.

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*.

Larry Davis and John M Norris. 2024. *Challenges and innovations in speaking assessment*. Routledge.

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 conference of the North American chapter of the association for computational linguistics: human language technologies (NAACL)*, pages 4171–4186.

Keelan Evanini, Maurice Cogan Hauck, and Kenji Hakuta. 2017. Approaches to automated scoring of speaking for k–12 english language proficiency assessments. *ETS Research Report Series*, 2017(1):1–11.

Keelan Evanini and Xinhao Wang. 2013. Automated speech scoring for non-native middle school students with multiple task types. In *Proceedings of the Interspeech (INTERSPEECH)*, pages 2435–2439.

Cheng-Yu Hsieh, Chun-Liang Li, Chih-Kuan Yeh, Hootan Nakhost, Yasuhisa Fujii, Alex Ratner, Ranjay

Krishna, Chen-Yu Lee, and Tomas Pfister. 2023. Distilling step-by-step! outperforming larger language models with less training data and smaller model sizes. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8003–8017.

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

Minhyeok Lee. 2023. Gelu activation function in deep learning: a comprehensive mathematical analysis and performance. *arXiv preprint arXiv:2305.12073*.

Jiun-Ting Li, Tien-Hong Lo, Bi-Cheng Yan, Yung-Chang Hsu, and Berlin Chen. 2023. Graph-enhanced transformer architecture with novel use of cefr vocabulary profile and filled pauses in automated speaking assessment. In *SLaTE*, pages 109–113.

Jiun-Ting Li, Bi-Cheng Yan, Tien-Hong Lo, Yi-Cheng Wang, Yung-Chang Hsu, and Berlin Chen. 2024. Automated speaking assessment of conversation tests with novel graph-based modeling on spoken response coherence. In *2024 IEEE Spoken Language Technology Workshop (SLT)*, pages 825–832. IEEE.

Tien-Hong Lo, Fu-An Chao, Tzu-I Wu, Yao-Ting Sung, and Berlin Chen. 2024. An effective automated speaking assessment approach to mitigating data scarcity and imbalanced distribution. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 1352–1362.

Russell Moore, Andrew Caines, Calbert Graham, and Paula Buttery. 2015. Incremental dependency parsing and disfluency detection in spoken learner english. In *International Conference on Text, Speech, and Dialogue (TSD)*, pages 470–479. Springer.

Wen-Hsuan Peng, Sally Chen, and Berlin Chen. 2024. Enhancing automatic speech assessment leveraging heterogeneous features and soft labels for ordinal classification. In *IEEE Spoken Language Technology Workshop (SLT)*, pages 945–952. IEEE.

Yao Qian, Patrick Lange, Keelan Evanini, Robert Pugh, Rutuja Ubale, Matthew Mulholland, and Xinhao Wang. 2019. Neural approaches to automated speech scoring of monologue and dialogue responses. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8112–8116. IEEE.

Yao Qian, Rutuja Ubale, Matthew Mulholland, Keelan Evanini, and Xinhao Wang. 2018. A prompt-aware neural network approach to content-based scoring of non-native spontaneous speech. In *IEEE Spoken Language Technology Workshop (SLT)*, pages 979–986. IEEE.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning (ICML)*, pages 28492–28518. PMLR.

Zhanming Shen, Hao Chen, Yulei Tang, Shaolin Zhu, Wentao Ye, Xiaomeng Hu, Haobo Wang, Gang Chen, and Junbo Zhao. 2025. Cycle-instruct: Fully seed-free instruction tuning via dual self-training and cycle consistency. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5123–5137.

Yaman Kumar Singla, Avyakt Gupta, Shaurya Bagga, Changyou Chen, Balaji Krishnamurthy, and Rajiv Ratn Shah. 2021. Speaker-conditioned hierarchical modeling for automated speech scoring. In *Proceedings of the ACM international conference on information & knowledge management (CIKM)*, pages 1681–1691.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, and 1 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Bi-Cheng Yan, Ming-Kang Tsai, and Berlin Chen. 2025. Muffin: Multifaceted pronunciation feedback model with interactive hierarchical neural modeling. *IEEE Transactions on Audio, Speech and Language Processing*.

Klaus Zechner and Keelan Evanini. 2019. *Automated speaking assessment: Using language technologies to score spontaneous speech*. Routledge.

Klaus Zechner, Derrick Higgins, Xiaoming Xi, and David M Williamson. 2009. Automatic scoring of non-native spontaneous speech in tests of spoken english. *Speech communication*, 51(10):883–895.