Interpretable Automatic Spoken Language Assessment

A case-study on using LLM-derived question-based interpretable features

Pavanpreet Singh Gandhi[1]

MSc Computational Statistics and Machine Learning

Dr Carlo Ciliberto & Paramdeep Khangura

Submission date: 22 September 2025

[1]**Disclaimer:** This report is submitted as part requirement for the MSc Computational Statistics and Machine Learning degree at UCL. It is substantially the result of my own work except where explicitly indicated in the text. The report may be freely copied and distributed provided the source is explicitly acknowledged.

Abstract

Automatic Spoken Language Assessment (SLA) has traditionally relied on either cascaded ASR-text pipelines or end-to-end speech models that operate as black boxes, limiting their pedagogical value in educational contexts where interpretable feedback is crucial. This work addresses the fundamental trade-off between predictive accuracy and interpretability by proposing a question-based approach that extracts interpretable features from speech Large Language Models (LLMs) for automatic SLA. To our knowledge, this is the first application of question-based interpretable feature extraction to speech LLMs in the spoken language assessment domain. We decompose holistic scoring into structured multiple-choice questions targeting specific speaking competencies, capturing the speech LLMs probabilistic responses as interpretable features for downstream regression. Our experiments on the Speak & Improve Corpus 2025 using Qwen2.5-Omni speech LLM demonstrate that this approach achieves competitive performance with dramatically fewer features—attaining a Pearson correlation coefficient of 0.8052 with just 42 interpretable features compared to 0.8282 with 3,584 black-box features from speech LLM hidden representations. We evaluate four distinct question sets and find that more comprehensive question sets generally improve performance, with our combined question set achieving 0.8132 correlation. Notably, our method shows exceptional data efficiency, with the rubric-based question set achieving 0.75 correlation using only 0.2% of training data (approximately 6 samples), comparable to BERT baselines trained on the full dataset (3000+ samples). The extracted features enable unsupervised decomposition of holistic scores into interpretable analytic scores across rubric dimensions, providing actionable feedback for learners while maintaining strong predictive performance. This work demonstrates the viability of interpretable speech LLM-based assessment systems that bridge the gap between accuracy and explainability in educational technology.

Code: https://github.com/pavanpreet-gandhi/interpretable-sla

Contents

Chapter 1

Introduction

Spoken Language Assessment (SLA) is the systematic evaluation of a speaker's oral proficiency, measuring various competencies such as pronunciation, fluency, grammar, vocabulary usage, and overall communicative effectiveness. This type of assessment is particularly valuable for second language (L2) learners, as it provides crucial feedback on their speaking abilities and guides their language development journey. With the widespread use of English, the number of L2 English learners worldwide continues to grow, creating a significant demand for scalable and accessible automatic SLA systems.

Traditionally, SLA is performed by human graders. These graders are typically trained to interpret a standardised rubric and score candidates across multiple dimensions or criteria within that rubric. For example, the Linguaskill Speaking Global Assessment Criteria [Cambridge English, 2020] based on the CEFR framework [Council of Europe, 2001] consists of three criteria: pronunciation and fluency, language resource, and discourse management. Each criterion is scored on a scale from A1 to C2, with A1 representing the lowest proficiency level and C2 the highest. While this approach ensures reasonably fair and nuanced evaluations, it is inherently limited by human availability, subjectivity, and scalability constraints.

Automatic SLA systems offer compelling advantages that address these traditional limitations. First, they provide immediate feedback, which is especially valuable for formative learning, enabling learners to engage in frequent practice without the logistical constraints of requiring human graders. Second, they offer scalability and cost-effectiveness by reducing dependence on human assessors, making language assessment more accessible to diverse populations. Third, they promise greater consistency by potentially reducing human bias and variability, assuming the availability of high-quality training data and robust model design.

However, automatic SLA presents unique challenges, particularly when compared to the more established field of Automatic Essay Scoring (AES). Speech is inherently a richer and more complex modality than text, containing paralinguistic information that is crucial for comprehensive assessment. Traditional approaches have relied on Automatic Speech Recognition (ASR) combined with AES pipelines [Wang et al., 2021, Raina et al., 2020], but these methods suffer from significant limitations. Transcriptions are lossy compressions of the original speech signal, where ASR inaccuracies can propagate through the assessment pipeline, and valuable paralinguistic information is often lost. These transcription inaccuracies are particularly problematic for L2 learners,

whose non-native pronunciation patterns often challenge ASR systems.

Recent advances in self-supervised speech representations [Bannò and Matassoni, 2022, Bannò et al., 2022, McKnight et al., 2023] and, more recently, speech LLMs [Ma et al., 2025] have demonstrated significant promise in overcoming the limitations of transcript-based approaches. These models can process raw audio directly, capturing the full richness of the speech signal and potentially providing more accurate assessments. However, these models operate as black boxes, lacking the interpretability that is crucial for educational applications.

Interpretability is especially critical in educational contexts (e.g., SLA and AES). From a pedagogical perspective, learners benefit greatly from understanding precisely where and how they need to improve their speaking skills. Unlike holistic scores that provide only aggregate performance measures, interpretable models can identify specific weaknesses amongst the various dimensions of speaking proficiency. This granular understanding enables learners to focus their practice efforts on areas most likely to yield improvement, potentially accelerating language acquisition and increasing motivation through clear, actionable guidance. Furthermore, teachers and tutors can leverage interpretable assessments to design personalized learning interventions, adapting instruction to individual learner profiles and monitoring progress across specific competency dimensions over time.

This work explores the novel application of interpretable features derived from speech LLMs for automatic SLA, drawing inspiration from recent advancements in interpretable AES [Eltanbouly et al., 2025]. We ask the central research question: **Can speech Large Language Models be used to develop interpretable systems for automatic spoken language assessment (SLA)?** To investigate this, we perform experiments on the Speak & Improve Corpus 2025 [Knill et al., 2024] using the Qwen2.5-Omni speech LLM [Xu et al., 2025]. Our contributions are three-fold. First, we demonstrate that a question-based methodology that extracts interpretable features using speech LLMs can achieve competitive performance with dramatically fewer features that are conceptually meaningful. Second, we evaluate different configurations of our approach to explore the trade-off between interpretability and predictive performance. Third, we explore the scalability of speech LLM-based SLA, demonstrating the advantage of interpretable features especially in extremely low-data regimes (less than 24 training examples).

The remainder of this report is structured as follows. Chapter 2 provides a comprehensive literature review covering spoken language assessment methodologies, self-supervised learning, speech LLMs and their applications in spoken language assessment, the parallel fields of mispronunciation detection and automatic essay scoring, and finally LLM-derived interpretable features. Chapter 3 details our proposed methodology for extracting interpretable features from speech LLMs and their application to automatic SLA. Chapter 4 describes our specific experimental design and implementation details. Chapter 5 presents and discusses our experimental results, including performance comparisons, interpretability analyses, and simulated low-data regimes. Finally, Chapter 6 concludes the report with a summary of findings, implications, limitations, and directions for future work.

Chapter 2

Literature Review

This chapter surveys prior work in automated spoken language assessment and related areas, organized into six thematic sections. We first review the evolution of automatic spoken language assessment in Section 2.1. We then examine self-supervised learning approaches that have become foundational to modern text and speech processing in Section 2.2. Next, we discuss the emergence of speech large language models in Section 2.3. We briefly cover the parallel field of mispronunciation detection in Section 2.4. We then examine developments in automatic essay scoring which offer useful parallels in Section 2.5. Finally, we review question-based approaches for deriving interpretable features from LLMs in Section 2.6, setting the stage for our proposed methodology.

2.1 Spoken Language Assessment

Automated spoken language assessment has a history dating back to the 1990s. Early systems were built on hidden Markov models (HMMs) with carefully engineered features. For example Bernstein et al. [1990] used an HMM-based speech recognizer to align a learner's speech to reference models and compute pronunciation scores. Such systems relied on expert-designed measures (e.g. segment durations, goodness-of-pronunciation scores) and provided the first proof-of-concept that computer evaluation of speaking proficiency could correlate well with human ratings. The next major leap came in the early 2010s with the advent of deep neural networks. Qian et al. [2012] replaced components of the HMM with a Deep Belief Network, one of the first instances of using deep learning models for pronunciation scoring. This marked the transition from purely expert-crafted features toward data-driven representations in SLA.

By the early 2020s, transformer-based encoders became widely adopted in a cascaded pipeline: an automatic speech recognizer (ASR) first generates a transcript, which is then processed by a text encoder to produce an embedding, followed by a regression model that predicts the proficiency score. For instance, Wang et al. [2021] proposed a system where ASR transcripts of learner speech are fed into transformer-based text encoders, such as BERT [Devlin et al., 2019] and XLNet [Yang et al., 2020], to extract semantic embeddings used for proficiency prediction. This approach enables the capture of content quality, vocabulary, and grammatical structures from transcribed speech. Similarly, Raina et al. [2020] developed a BERT-based SLA system and observed that appending a specific adversarial six-word phrase to the transcript could significantly increase the

predicted score. Both these works highlight the trend of leveraging large pre-trained language models for automated spoken language assessment.

A significant limitation of cascaded BERT-based graders is their reliance on ASR transcripts, which introduces two major problems: transcriptions are lossy representations that omit prosodic and acoustic cues essential for spoken language assessment, and they are prone to errors, particularly for non-native speakers with accents or disfluencies. To address these issues, researchers have explored end-to-end scoring methods that operate directly on audio using deep speech representations. Self-supervised models such as wav2vec 2.0 [Baevski et al., 2020] and HuBERT [Hsu et al., 2021] provide powerful audio encodings and can be fine-tuned for proficiency prediction without intermediate ASR, thereby implicitly capturing fluency and pronunciation cues. While studies such as Bannò and Matassoni [2022] and Bannò et al. [2022] showed that wav2vec 2.0–based systems can approach the performance of transcript-based systems, they still underperform BERT text graders on certain tasks, especially when assessing longer responses that require semantic understanding. Interestingly, both of these studies found that integrating audio and transcript-based text representations yields even stronger results, suggesting that audio embeddings capture acoustic characteristics (how something was said) but are less effective at capturing lexical and semantic content (what was said), and that combining both modalities leads to enhanced performance. This hypothesis is further supported by McKnight et al. [2023] which found similar results in conversational speaking tests.

Most recently, Ma et al. [2025] explored using speech LLMs for SLA, achieving state-of-the-art results on both private [Ludlow, 2020] and public [Knill et al., 2024] datasets, outperforming both cascaded BERT-based and end-to-end wav2vec 2.0-based systems. Initially, they found that zero-shot performance without any parameter fine-tuning was quite poor, which they attributed to the implicit positional bias of LLMs [Liusie et al., 2023] resulting in the LLM frequently predicting similar scores for all candidates. They then explored three different training schemes: classification with cross-entropy loss, classification with fair average loss (to allow the model to predict between the discrete levels), and regression. They found that classification with fair average loss performed the better than regression by a small margin (0.002 PCC improvement in initial tests). The authors hypothesised that the classification objective is more aligned with the pre-training objective of next token prediction which explains the slight performance edge. Interestingly, they also found that speech LLMs demonstrated strong generalization capabilities where models finetuned on one dataset were able to perform well on the other dataset.

2.2 Self Supervised Learning

Self-supervised learning (SSL) has emerged as a powerful paradigm for pre-training models on vast quantities of unlabeled data by having them solve a pretext task [Balestriero et al., 2023]. This process yields rich feature representations that are highly transferable to downstream applications. Two dominant approaches define the field: predictive and contrastive learning. Predictive, or generative, models learn by masking a portion of an input and training the model to predict the missing content, thereby learning local context and structure; a prime example is the masked language modeling used in BERT [Devlin et al., 2019]. In contrast, contrastive models learn a geometric embedding space by training an encoder to maximize the similarity between different

augmented views of the same data point (positive pairs) while simultaneously minimizing the similarity to other data points (negative pairs) [Chen et al., 2020]. This forces the model to learn discriminative features that are invariant to superficial transformations.

BERT (Bidirectional Encoder Representations from Transformers) is a foundational model in natural language processing that epitomizes the predictive SSL approach [Devlin et al., 2019]. Its pre-training is centered on Masked Language Modeling (MLM), where a percentage of input tokens in a sequence are randomly masked, and the model's objective is to predict the original identity of these masked tokens. By employing a Transformer encoder architecture, BERT can process the entire sequence at once, allowing it to fuse context from both the left and right sides of a masked token. This bidirectional context was a significant advancement over previous unidirectional models. The model is optimized using a cross-entropy loss function over the vocabulary, forcing it to learn deep representations of language syntax and semantics that have proven to be exceptionally effective for a wide array of language understanding tasks.

Wav2Vec 2.0 successfully adapted SSL principles to the challenging domain of continuous speech data [Baevski et al., 2020]. Its architecture consists of a convolutional feature encoder that generates latent representations of the raw audio, followed by a Transformer context network that builds contextualized representations. To handle the continuous nature of audio, Wav2Vec 2.0 introduced a hybrid objective centered on a contrastive task. After masking latent representations in the time domain, the model's contextualized output for a masked step must identify the correct quantized version of the original latent vector from a set of distractors sampled from the same utterance. This objective, optimized via a contrastive loss, forces the model to learn highly robust and discriminative speech features by not only predicting the masked content but also distinguishing it from other sounds.

HuBERT (Hidden-Unit BERT) sought to simplify the pre-training process for speech by more directly applying the BERT paradigm [Hsu et al., 2021]. The key innovation is a two-stage process that decouples target discovery from representation learning. In the first stage, an offline clustering algorithm (e.g., k-means) is applied to acoustic features to generate a finite set of discrete "hidden units," which serve as pseudo-labels. For the initial iteration, these features are 39-dimensional Mel-frequency cepstral coefficients (MFCCs); for subsequent iterations, features from the HuBERT model trained in the prior step are used. In the second stage, the HuBERT model is trained using a simple masked prediction task: it predicts the corresponding hidden unit ID for masked regions of the input audio. By converting the problem into a classification task over these discovered discrete units, the model is able to use the same cross-entropy loss as BERT, eliminating the need for a complex contrastive objective and demonstrating the power of the masked prediction framework for continuous data.

2.3 Speech Large Language Models

Speech large language models (speech LLMs) are foundation models that are able to generate text (and optionally audio) conditioned on text and audio input. They are usually constructed by connecting a speech encoder to a text-based LLM via a lightweight adapter that maps the speech encoders representations into the LLM's embedding space. The speech encoder processes raw audio into a sequence of vectors that represent the audio at a higher conceptual level. For

example, the Whisper model [Radford et al., 2022] is a transformer-based encoder-decoder model trained on large scale supervised automatic speech recognition (ASR) data. The encoder portion of Whisper takes in a spectrogram of the audio waveform (effectively a 2D image), converts the spectrogram into patches (effectively vertical slices of the spectrogram), and processes these patches with a series of transformer layers to produce a sequence of latent vectors that represent high-level features of the audio. This sequence of vectors is then passed through a lightweight adapter (e.g. a simple MLP) that maps the speech encoder's output into the LLMs embedding space. The text-based LLM (e.g., Qwen [Chu et al., 2023]) is typically a decoder-only transformer model trained on large scale text data with a next-token prediction objective. The LLM takes in the sequence of projected speech embeddings along with any text input and generates text output autoregressively, attending to both the speech and text context.

Typically, the speech encoder and the base LLM are pre-trained separately, the only parameters that are initialised from scratch come from the lightweight adapter that connects the two models. The entire stack is then finetuned end-to-end with audio-text input and text output pairs using standard next-token prediction loss. During finetuning, the audio encoder and the LLM can be kept frozen or optionally finetuned either fully or partially (e.g., using LoRA [Hu et al., 2021]). This end-to-end training allows the model to learn to align audio representations with text representations, enabling the LLM to effectively process and generate text conditioned on audio input. The result is a foundation model that integrates audio and language understanding, enabling it to perform complex tasks that require understanding and reasoning over both modalities.

Early work on speech LLMs fused pre-trained speech encoders with text-based LLMs. For example, AudioPaLM by Rubenstein et al. [2023] combined a speech representation model (AudioLM) with PaLM-2, showing that leveraging text-pretrained linguistic knowledge can significantly improve speech understanding and enable zero-shot speech translation. Building on this concept, Tang et al. [2024] introduced SALMONN, which integrated a text LLM with both speech and general audio encoders to handle spoken language, environmental sounds, and music within one model—achieving strong multi-task performance. Around the same time, Qwen-Audio [Chu et al., 2023] adopted a single Whisper-large-v2 encoder plugged into Qwen-7B and introduced multi-task pre-training with hierarchical tags to unify many audio tasks with a single audio encoder. Qwen2-Audio [Chu et al., 2024] then upgraded the encoder to Whisper-large-v3 and replaced hierarchical tags with natural language prompts while retaining Qwen-7B, scaling the same recipe for better instruction following. Subsequent works sought to enhance higher-level audio reasoning: Ghosh et al. [2024] proposed GAMA, which introduced a multi-layer audio feature aggregator (Audio Q-Former) and synthetic instruction tuning for complex audio question answering. Additionally, Microsoft's lightweight Phi-4-multimodal model [Microsoft et al., 2025] showed that even a 5.6B-parameter LLM can achieve robust speech understanding by connecting speech (and vision) encoders via efficient LoRA adapters. Most recently, Xu et al. [2025] presented Qwen2.5-Omni, a unified text–image–audio model that builds upon the architecture of Qwen2-Omni by upgrading to the Qwen2.5-7B base LLM while further training the speech encoder initialised from Qwen2-Audio. This model achieved state-of-the-art performance on audio evaluation benchmarks, notably topping the MMAU benchmark [Sakshi et al., 2024] for audio reasoning at the time of this dissertation.

2.4 Mispronunciation Detection

A closely related task to automated spoken language assessment is mispronunciation detection and diagnosis. Early systems relied on error pipelines based on recognition scores and hand-crafted rules, but recent work shifts to self-supervised speech models such as wav2vec 2.0 (and HuBERT), which are effective for both phone recognition and binary error detection [Peng et al., 2021, Wu et al., 2021, Xu et al., 2021]. Building on this, researchers have moved beyond local error flags toward holistic pronunciation assessment with self-supervised models [Kim et al., 2022], and most recently toward speech LLMs that take audio and reference text to predict utterance-level fluency and accuracy scores [Fu et al., 2024]. Empirically, speech LLMs are competitive for holistic scoring but do not consistently outperform strong self-supervised baselines for fine-grained error detection [Kim et al., 2022, Fu et al., 2024]. This is contrary to the trend in spoken language assessments where speech LLMs significantly outperform self-supervised models [Ma et al., 2025]. A possible explanation is that mispronunciation detection benefits less from the linguistic knowledge in LLMs, instead relying more on acoustic-phonetic cues that self-supervised models capture well. This suggests that speech LLMs may be most beneficial for higher-level spoken language assessment tasks that require both semantic and paralinguistic understanding and reasoning.

2.5 Automatic Essay Scoring (AES)

Automatic essay scoring (AES) has a longer history than spoken language assessment, dating back to Ellis Page's seminal 1966 article on "grading essays by computer" and the Project Essay Grade system [Page, 1966]. Because text-based large language models emerged first, AES was also quicker to adopt them compared to SLA. Early experiments with GPT-3 demonstrated the feasibility of using general-purpose LLMs for holistic scoring. For instance, Mizumoto and Eguchi [2023] applied GPT-3 [Brown et al., 2020] to the TOEFL11 (L2 writing) dataset, finding that it could produce scores with moderate agreement to human raters, especially when combined with linguistic features. Similarly Yancey et al. [2023] evaluated GPT-3.5 and GPT-4 [OpenAI et al., 2024] on short learner essays aligned to the CEFR scale, showing performance comparable to commercial AES engines when calibration examples were provided. In parallel, Naismith et al. [2023] investigated scoring discourse coherence, a notoriously difficult trait to score automatically, and demonstrated that GPT-4 could provide not only accurate trait scores but also interpretable rationales for its predicted scores. Notably, this work also highlighted the trend towards rationales and interpretability in AES as a means to enhance pedagogical value.

Following these initial results, prompting strategies quickly emerged as a means to improve LLM performance for AES. Mansour et al. [2024] showed that both GPT-3.5 and Llama 2 [Touvron et al., 2023] were highly sensitive to prompt design. Stahl et al. [2024] proposed prompting strategies that explicitly combined essay scoring with feedback generation, finding that requiring the model to justify its scores improved predictive accuracy. However, none of these prompt-based approaches achieved competitive performance on the ASAP dataset [Mathias and Bhattacharyya, 2018], indicating that purely prompt-based methods targeting holistic scoring may be inherently limited.

To address the limitations of prompt-based holistic scoring, researchers turned to analytic or

multi-trait scoring, which decomposes the overall assessment into multiple specific dimensions or traits. Do et al. [2024] framed multi-trait scoring as an autoregressive generation task with T5 [Raffel et al., 2023], demonstrating that pre-trained language models are effective at multi-trait scoring. Subsequently, Bannò et al. [2024] explored whether GPT-4 could perform analytic assessment without ground truth analytic scores by training a Longformer-based [Beltagy et al., 2020] holistic grader and prompting GPT-4 to provide analytic scores given the holistic score, obtaining significant correlations between predicted analytic scores and various features linked to the componential aspects of the CEFR levels. Building on this direction, Lee et al. [2024] proposed the MTS prompting framework to automatically decompose holistic scoring into trait-specific tasks through multiple conversational rounds, with this trait-decomposing strategy outperforming prior prompting approaches. Chu et al. [2025] further advanced this concept by introducing RMTS, which replaced trait-specific scores with rationales generated by fine-tuned small LLMs (e.g., T5) that are aggregated via text encoding. Finally, Eltanbouly et al. [2025] simplified this approach with TRATES, which uses a pre-trained LLM to extract trait-specific and task-specific features from essays by asking targeted questions and recording responses as "High", "Medium", or "Low" categories; these interpretable features are then fed to a regression model to predict final scores, achieving state-of-the-art performance on the ASAP and ASAP++ datasets, and demonstrating that LLM-derived interpretable features are highly effective for AES.

2.6 Question-based LLM-derived Interpretable Features

Similar to TRATES [Eltanbouly et al., 2025], several works have explored using LLMs as generators of human-interpretable features for downstream prediction tasks. McInerney et al. [2023] introduced CHILL, which extracts clinically meaningful features from unstructured notes by prompting an LLM with expert-designed domain-specific questions. While their interpretable features did not outperform BERT or TF-IDF features in downstream prediction tasks, they achieved a much more parsimonious model ($\sim$ 100 vs $\sim$ 100k features) with interpretable features that aligned well with clinical priors, and notably found that using LLM confidence levels rather than binary representations significantly improved performance. Similarly, Benara et al. [2024] developed QA-EMB to generate interpretable features for predicting fMRI voxel responses, asking LLMs yes/no questions and using binary feature embeddings from output tokens to outperform both established interpretable baselines and black-box BERT-based models with only 29 questions. Balek et al. [2025] extended this approach across diverse text classification and regression datasets (scientometrics, banking, hate speech, and food hazard), achieving superior performance against TF-IDF and SciBERT-based embeddings while also using binary feature embeddings rather than probabilities. Finally, Sam et al. [2025] adopted a similar question-based approach to predict LLM task performance and detect adversarial prompting. These works collectively demonstrate the versatility of question-based LLM-derived interpretable features across domains. Interestingly, some approaches construct continuous embeddings whilst others use binary/discrete features, suggesting that the optimal feature representation may be task-dependent.

In the next chapter, we will combine these ideas to develop our approach for spoken language assessment. We intend to use a speech LLM as the backbone for scoring, but rather than relying on it as a black box, we will extract question-based interpretable features that correspond to

defined proficiency traits. By prompting the model with trait-specific questions and collecting its response probabilities, we aim to build a scoring system that not only achieves high accuracy but also produces interpretable evidence for its scores. The fusion of trait-guided prompts with question-based interpretable features derived from speech LLMs is a novel direction that, as our literature review has shown, holds great promise for creating transparent and robust automated assessment tools.

Chapter 3

Methodology

In this chapter, we detail our proposed methodology for developing an interpretable automatic spoken language assessment system using speech Large Language Models (LLMs). We begin by formally defining the problem setup in Section 3.1, then describe our question-based approach for extracting interpretable features using the speech LLM in Section 3.2. Within the question-based approach, we cover question set design in Section 3.2.1, feature extraction via speech LLMs in Section 3.2.2, and the final regression model in Section 3.2.3. The overall methodology is summarized in Figure 3.2, which appears later in this chapter.

3.1 Problem Setup

The problem is framed as a supervised learning task. We are provided with a dataset of N spoken language assessments, denoted as $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{N}$. Each data point $(\mathbf{x}_i, y_i)$ consists of two components. The input, $\mathbf{x}_i \in \mathbb{R}^{l_i}$, is the raw audio signal from a learner's spoken response, represented as a variable-length vector of waveform samples, where l_i is the length of the audio in seconds times the sample rate. The target, $y_i \in \mathbb{R}$, is a scalar value representing the holistic score or grade assigned to the assessment by a human expert rater.

The primary objective is to develop a model that can accurately predict the score y_i given the audio input $\mathbf{x}_i$. This is a regression problem where we aim to learn a function $f : \mathcal{X} \to \mathbb{R}$ that maps an audio input from the space of possible audio signals $\mathcal{X}$ to a predicted score $\hat{y}_i = f(\mathbf{x}_i)$. The goal is to minimize the discrepancy between the predicted scores and the true human-assigned scores across the dataset, typically measured by a loss function such as mean squared error.

Beyond predictive accuracy, a crucial requirement for our model is interpretability. In the context of educational assessment, a "black-box" model that provides a score without justification is of limited pedagogical value. Therefore, a key secondary objective is that the model's predictions should be explainable to some extent. This interpretability is essential for facilitating constructive feedback for the learner, highlighting specific areas of strength and weakness (e.g., in pronunciation, fluency, or grammar) that contributed to the final score.

3.2 Question-based Approach

To address the dual requirements of accuracy and interpretability, we propose a question-based approach. Instead of training an end-to-end model to predict a score directly from audio, we use a speech LLM as an intermediate feature extractor. The core idea is to decompose the complex task of holistic scoring into a series of simpler, more constrained sub-problems, framed as multiple-choice questions. The speech LLMs responses to these questions, captured as probabilities for each option, form a rich, interpretable feature set. A simple regression model is then trained on these features to produce the final score. This methodology, inspired by the recent trend of LLM-derived question-based interpretable features [Eltanbouly et al., 2025, McInerney et al., 2023, Benara et al., 2024, Balek et al., 2025] and applied to the audio domain, allows us to leverage the natural language and audio understanding of a large foundation model while maintaining features that are easier to interpret whilst still being predictive. The process can be broken down into three main steps: question set design (Section 3.2.1), feature extraction (Section 3.2.2), and final regression (Section 3.2.3). An overview of the entire approach is illustrated in Figure 3.2.

3.2.1 Question Set Design

The first step is to specify a set of multiple-choice questions whose answers are likely to be informative for determining the final score. This stage provides a crucial opportunity to inject domain knowledge into the system. The questions act as a structured guide, focusing the speech LLM on specific linguistic and paralinguistic traits relevant to different aspects of speaking proficiency.

Formally, we define a question set $\mathcal{Q} = \{q_1, q_2, \ldots, q_M\}$ consisting of M multiple-choice questions. Each question q_j is associated with a set of possible answer options $\mathcal{O}_j = \{o_{j,1}, o_{j,2}, \ldots, o_{j,K_j}\}$, where K_j denotes the number of options for question j. The complete question set can thus be represented as $\mathcal{Q} = \{(q_j, \mathcal{O}_j)\}_{j=1}^{M}$. For simplicity, we often assume that all questions have the same number of options K, so that $K_j = K$ for all $j \in \{1, 2, \ldots, M\}$, though this constraint can be relaxed in practice. An example of a single question along with its options is shown in figure 3.1.

Example Question

Question: Can an average listener understand without strain or repetition?

- **High:** Speech is always clear; no repetition or clarification needed.
- **Medium:** Mostly clear; occasional repetition or clarification needed.
- **Low:** Often unclear; listener must frequently ask for repetition or guesses meaning.

Figure 3.1: Example of a multiple-choice question with three options.

There is considerable flexibility in how these questions can be formulated. Questions can be expert-designed, where a subject matter expert, such as an experienced grader, designs questions that mirror their own internal evaluation process. This is especially effective since the expert is usually able to inject their domain specific knowledge into the model via natural language through

the questions that they design. Alternatively, a text-based LLM (e.g. GPT-5) can be prompted to act as a subject matter expert to automate the expert design process. This is particularly useful when expert time is limited or when scaling to multiple assessment contexts. A key consideration here is to make sure that the LLM is provided with sufficient context about the assessment task to allow it to generate relevant and meaningful questions.

If a detailed scoring rubric is available, its criteria can be directly converted into rubric-based questions, where each rubric dimension becomes a question with the corresponding grade levels as multiple-choice options. For example, suppose a rubric includes a "Pronunciation" criterion with levels such as: (A) Excellent pronunciation with minimal errors, (B) Good pronunciation with minor errors, and (C) Fair pronunciation with noticeable errors. This can be reformulated as the question: "How would you rate the speaker's pronunciation?" with options A, B, and C corresponding to the original rubric levels. This approach ensures tight alignment between the extracted features and the established rubric.

Furthermore, the question set can be dynamically refined rather than remaining fixed. Once an initial model has been trained, it is possible to examine instances where prediction errors are largest and identify underlying patterns in these failures. These insights can guide the development of additional, more focused questions that target the model's identified weaknesses, creating an iterative refinement process. This iterative approach can even be automated by presenting a speech LLM with challenging examples that exhibit large prediction errors and prompting it to generate new questions that would better differentiate between scores for such difficult cases. This process, which bears conceptual similarity to gradient boosting methods, was introduced by Benara et al. [2024]. We do not explore this iterative refinement process in this dissertation, but mention it here as a promising approach for practical applications.

Examples of specific question sets used in our experiments, including their rationale and design principles, are detailed in Section 4.3.

3.2.2 Feature Extraction via Speech LLMs

Once the question set is defined, we use a speech LLM to generate features for each audio sample. Specifically, we employ Qwen2.5-Omni [Xu et al., 2025] as our speech LLM. For every data point $(\mathbf{x}_i, y_i)$ in our dataset and for each question q_j in our question set, we prompt the speech LLM with both the audio $\mathbf{x}_i$ and the text of the question q_j. The model is tasked with selecting the most appropriate option for the multiple-choice question.

However, instead of recording the actual generated token (the argmax of the predicted probabilities), we capture the predicted probability associated with each valid option, allowing us to construct a continuous probabilistic feature representation. This richer representation captures the model's confidence in each option which can be more informative than a single categorical choice 5.1. Crucially, for the probability of each option to be well defined and comparable, each option must map to a single token in the speech LLMs vocabulary. This is generally feasible by specifying short options (e.g., "High" or "Low", "A" or "B", "1" or "2", etc.). Note that this does not mean that each short option cannot be associated with a longer descriptive text within the prompt, as shown in figure 3.1.

More formally, for each audio sample $\mathbf{x}_i$ and question q_j, we apply the speech LLM to obtain

a feature vector $\mathbf{z}_{i,j} \in \mathbb{R}^K$, where each element corresponds to the probability of selecting one of the K answer options. Specifically, if the options for question q_j are $\mathcal{O}_j = \{o_{j,1}, o_{j,2}, \ldots, o_{j,K}\}$, then:

$$\mathbf{z}_{i,j} = [P(o_{j,1}|\mathbf{x}_i, q_j), P(o_{j,2}|\mathbf{x}_i, q_j), \ldots, P(o_{j,K}|\mathbf{x}_i, q_j)]$$

where $P(o_{j,k}|\mathbf{x}_i, q_j)$ represents the probability that the speech LLM assigns to option k given the audio input $\mathbf{x}_i$ and question q_j. By repeating this process for all M questions in our set, we obtain M feature vectors $\mathbf{z}_{i,1}, \mathbf{z}_{i,2}, \ldots, \mathbf{z}_{i,M}$, each of dimension K. These vectors are then concatenated to form the final feature representation for audio sample $\mathbf{x}_i$:

$$\mathbf{z}_i = [\mathbf{z}_{i,1}; \mathbf{z}_{i,2}; \ldots; \mathbf{z}_{i,M}] \in \mathbb{R}^{M \times K}$$

This procedure, applied to all N audio samples in the dataset, yields a feature matrix $\mathbf{Z} \in \mathbb{R}^{N \times (M \times K)}$. Figure 3.2 illustrates this feature extraction process.

3.2.3 Final Regression Model

With the interpretable feature matrix $\mathbf{Z}$ constructed, the final step is to predict the assessment scores. We fit a regression model to learn the mapping from our question-based feature vectors to the ground-truth scores: $f : \mathbb{R}^{M \times K} \to \mathbb{R}$. We train this regression model using the dataset $\{(\mathbf{z}_i, y_i)\}_{i=1}^N$, where $\mathbf{z}_i$ is the feature vector for audio sample $\mathbf{x}_i$ and y_i is the corresponding human-assigned score.

While any regression algorithm could theoretically be applied at this stage, our empirical analysis reveals that simple linear regression achieves equivalent or superior performance to more sophisticated non-linear approaches (see Section 5.2). Although the effectiveness of linear regression might initially suggest enhanced interpretability through coefficient analysis, this advantage is largely negated by the high correlation observed among the extracted features, which complicates meaningful interpretation of individual regression weights. That being said, the features themselves remain interpretable on their own which is still useful. A detailed examination of these feature correlations and their implications for interpretability is presented in Section 5.5.

Formally, we define our linear regression model as:

$$\hat{\mathbf{y}} = \mathbf{Z}\mathbf{w} + w_0\mathbf{1}$$

where $\hat{\mathbf{y}} \in \mathbb{R}^N$ is the vector of predicted scores, $\mathbf{Z} \in \mathbb{R}^{N \times (M \times K)}$ is our feature matrix, $\mathbf{w} \in \mathbb{R}^{M \times K}$ is the vector of regression coefficients, $w_0 \in \mathbb{R}$ is the intercept term, and $\mathbf{1} \in \mathbb{R}^N$ is a vector of ones. To include the intercept term in matrix form, we augment our feature matrix as $\tilde{\mathbf{Z}} = [\mathbf{1}, \mathbf{Z}] \in \mathbb{R}^{N \times (M \times K + 1)}$ and define $\tilde{\mathbf{w}} = [w_0, \mathbf{w}^T]^T \in \mathbb{R}^{M \times K + 1}$, giving us $\hat{\mathbf{y}} = \tilde{\mathbf{Z}}\tilde{\mathbf{w}}$. The optimal coefficients are obtained via the normal equation:

$$\tilde{\mathbf{w}}^* = (\tilde{\mathbf{Z}}^T\tilde{\mathbf{Z}})^{-1}\tilde{\mathbf{Z}}^T\mathbf{y}$$

where $\mathbf{y} \in \mathbb{R}^N$ is the vector of ground-truth scores. In practice, we utilize the scikit-learn [Pedregosa et al., 2011] `LinearRegression` implementation.

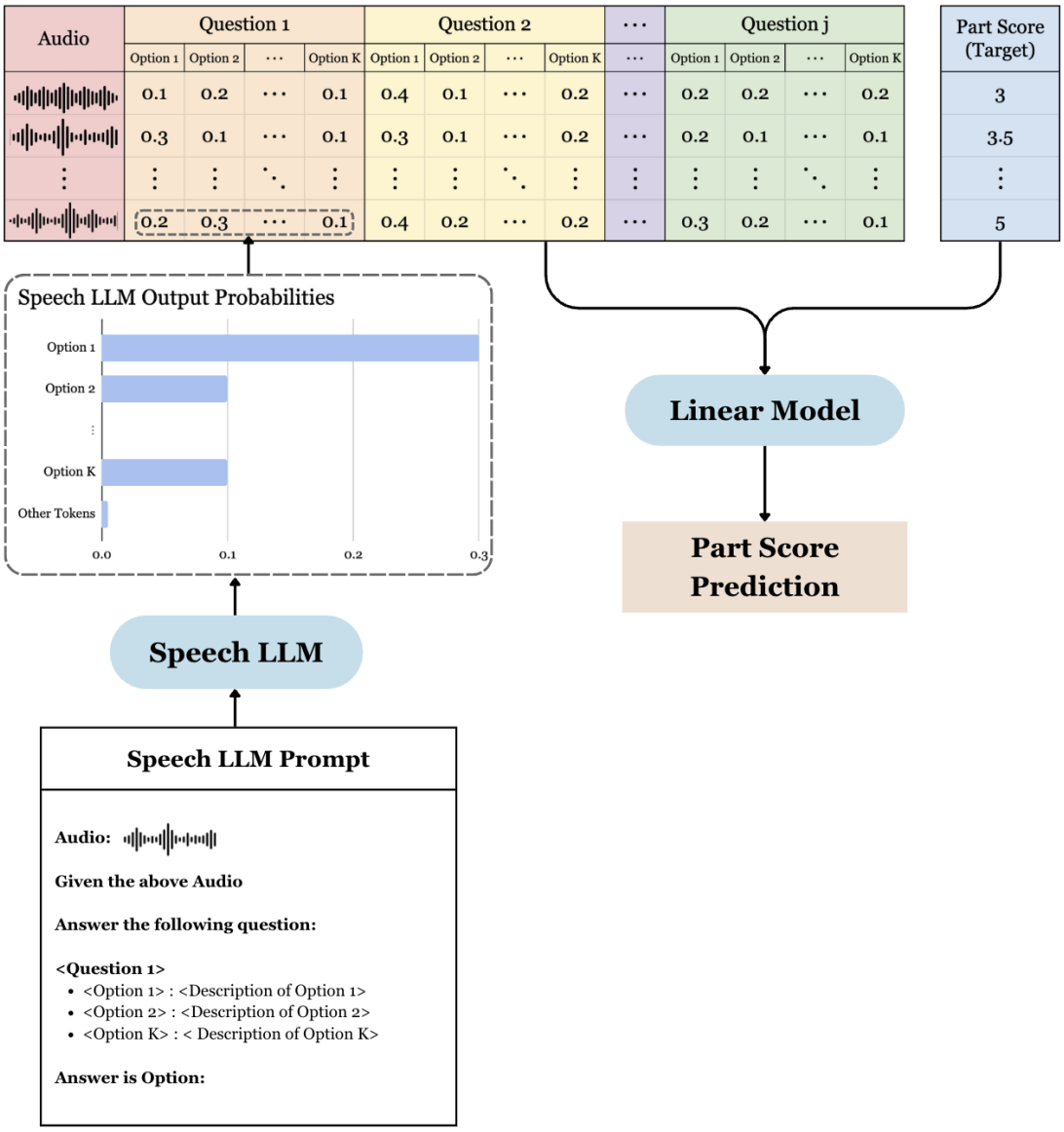

Figure 3.2: Schematic illustration of the question-based feature extraction process. The speech LLM is prompted to answer each question for every audio sample. The output probabilities for each option are collected into a table to form the interpretable feature matrix $\mathbf{Z}$, which is then used to train a regression model to predict the target variable (part scores).

Chapter 4

Experiments

In this chapter, we describe our experimental setup to evaluate our proposed method. We begin by introducing the Speak & Improve Corpus 2025 dataset used for our experiments in Section 4.1. Next, we outline the baseline methods against which we compare our approach in Section 4.2, including a cascaded BERT-based system and an end-to-end speech LLM representation-based system. We then detail the four specific question sets employed in our experiments in Section 4.3: the *Initial Question Set*, *Direct Scoring Question*, *Rubric-based Question Set*, and *Revised Question Set*. The complete question sets and prompts are provided in Appendix A. We discuss implementation details including our use of Qwen2.5-Omni and vLLM for inference in Section 4.4. Finally, we describe the two-stage calibration procedure and five evaluation metrics (RMSE, PCC, SRC, P@0.5, P@1.0) used to assess model performance in Section 4.5.

4.1 Dataset

We perform our experiments on the Speak & Improve Corpus 2025 [Knill et al., 2024], a publicly available dataset with 315 hours of L2 English learner speech recordings with human-assigned CEFR-aligned holistic scores. The data comes from the Speak & Improve web-app, a research project developed at the University of Cambridge, where L2 English learners from around the world practice their speaking skills by completing various speaking tasks and receiving automated feedback.

A single test consists of five parts, of which all except part 2 have been released. Part 1 is an interview where the learner is asked 8 questions about themselves. They are given 10 seconds of speaking time for the first 4 questions, and 20 seconds for the second 4. The first two questions, which may contain personal identity information, are not marked and are not included in this corpus. Part 2 consists of a read-aloud task where the learner must read aloud 8 sentences. Part 3 is a long turn task where the learner has 1 minute to give their opinion on a specific topic using 3 questions to guide them. Part 4 is another long turn task where the learner has 1 minute to describe a process depicted in a diagram. Finally, Part 5 is a communication activity where the learner has to respond to 5 questions relating to an overall topic, each for up to 20 seconds.

Each part receives a score from 1.0 (corresponding to A1 on the CEFR scale) to 6.0 (C2 on the CEFR scale) in increments of 0.5, with half scores allowed. The lowest score in the dataset is

2.0 (A2) and the highest is 5.5 (C1+), reflecting the L2 learner population. The overall test score is computed as the average of all parts except part 2, which is not included in the corpus. The dataset comes with 3 pre-defined splits: train (∼3000 submissions per part), dev (438 submissions per part), and eval (442 submissions per part). Audio files are provided as 16 kHz single-channel FLAC files.

For our approach, we employ the same question set for all parts, but train individual regression models for each part separately, with the overall predicted score calculated as the mean of the predicted part scores. It is important to note that the dataset lacks analytic scores for individual assessment sub-components, meaning our method must extract interpretable features representing these sub-components in an unsupervised manner.

4.2 Baselines

To evaluate the effectiveness of our question-based approach, we compare against two *black-box* baseline methods: a cascaded BERT-based system and an end-to-end speech LLM representation-based system.

4.2.1 BERT Baseline

Our first baseline follows the established cascaded pipeline approach used as the baseline in the Speak & Improve Challenge 2025 [Qian et al., 2024]. This baseline comprises of an ASR component to convert speech-to-text and a BERT-based text grader to predict scores from the transcripts.

First, OpenAI Whisper (small) [Radford et al., 2022] is used to generate transcripts from the audio recordings. These transcriptions are normalised by removing punctuation and converting to lowercase. Next, a BERT-based grader is finetuned to predict scores from the transcripts. These graders consist of a BERT-base-uncased [Devlin et al., 2019] text encoder followed by four multi-head self-attention layers whose outputs are concatenated and passed through two fully connected layers before finally mapping to a single scalar output. The weights for the encoder of the BERT-based grader are initialised from BERT-base-uncased, and a separate BERT-based grader is finetuned for each test part (Parts 1, 3, 4, and 5). The final submission-level score is computed as the average of the per-part predicted scores. We will refer to this baseline as the *BERT Baseline* throughout the rest of this dissertation. Further details regarding this baseline can be found in Qian et al. [2024].

4.2.2 Speech LLM Representations

Our second baseline utilizes the multimodal capabilities of speech LLMs for direct end-to-end scoring. We prompt Qwen2.5-Omni [Xu et al., 2025] to predict scores directly from audio inputs and extract the final layer hidden state representation corresponding to the predicted options token. This hidden state, with 3584 dimensions, is conditioned on the task prompt and the audio input, allowing it to capture rich semantic and acoustic information relevant to the scoring task. A regression model is then trained on these representations to predict final scores. Given that the feature dimensionality (3584) exceeds the available training samples per test part, we employ ridge regression rather than ordinary least squares to address the resulting rank deficiency of the

Gram matrix. Following the same structure as the BERT baseline and our proposed approach, we train separate models for each test part and compute the final submission-level score by averaging the predicted part scores.

This baseline slightly differs from the regression approach used in Ma et al. [2025], who fine-tune LoRA adaptors along with a final linear layer. However, our approach of training a regression model on fixed final layer representations is computationally cheaper and performs comparably well in practice (see Section 5.2 and Ma et al. [2025]). Therefore, this baseline is near state-of-the-art making it significantly challenging to outperform. However, unlike our question-based approach, this method does not provide any interpretable features or insights into the scoring process. We refer to this baseline as the *Speech LLM Representations* throughout the rest of this dissertation.

4.3 Question Sets

Our experiments evaluate four distinct question sets: *Initial Question Set*, *Direct Scoring Question*, *Rubric-based Question Set*, and *Revised Question Set*. These question sets differ in the number of questions, the number of options per question, and their design methodology. Table 4.1 summarizes the key characteristics of each question set.

Question Set	**Questions**	**Options**	**Design Methodology**
Initial Question Set	14	"High", "Medium", "Low"	LLM generated
Direct Scoring Question	1	"A", "B", "C", "D", "E", "F"	One question
Rubric-based Question Set	3	"A", "B", "C", "D", "E", "F"	Rubric based
Revised Question Set	30	"High", "Low"	Expert enhanced

Table 4.1: Overview of question sets used in experiments.

The *Initial Question Set* consists of 14 questions generated using ChatGPT o3. The LLM was prompted with the Linguaskill Speaking Global Assessment Criteria rubric [Cambridge English, 2020] and instructed to generate questions that would be both necessary and sufficient for determining final scores. Following the TRATES methodology [Eltanbouly et al., 2025], each question has three options: "High", "Medium", and "Low". The complete question set is provided in Section A.1 in Appendix A.

The *Direct Scoring Question* exists within a question set with a single question that asks the speech LLM to predict the final score directly. The speech LLM is presented with the Linguaskill Speaking Global Assessment Criteria rubric [Cambridge English, 2020] verbatim (with options A1 to C2 mapped to F to A) and asked to select a single option representing the final score across all three dimensions of the rubric (Pronunciation and Fluency, Language Resource, and Discourse Management). The goal of this question set is to assess the value of having multiple questions versus a single direct scoring question. The prompt used for this question is shown in Section A.2 in Appendix A.

The *Rubric-based Question Set* consists of three questions, each corresponding to one of the three criteria in the Linguaskill Speaking Global Assessment Criteria rubric [Cambridge English, 2020]. Each question has six options (A, B, C, D, E, F) corresponding to the six levels in the rubric. This question set is designed to directly reflect the rubric structure while still decomposing

the scoring task into multiple dimensions. The advantage of this approach is that it ensures tight alignment between the extracted interpretable features and the rubric criteria. The complete *Rubric-based Question Set* is provided in Section A.3 in Appendix A.

The *Revised Question Set* is a more comprehensive version of the *Initial Question Set*, expanded to 30 questions that incorporate feedback from subject matter experts—specifically, experienced oral assessment graders who regularly evaluate spoken language proficiency. Starting from the initial 14 questions, these experts modified existing questions and contributed additional ones based on their practical grading experience, resulting in a comprehensive set of 30 questions. Notably, this question set also contains generic questions about the audio quality (e.g., background noise, volume) to help the model account for extraneous factors that may affect intelligibility. A key modification in this question set is the reduction from three response options to two: "High" and "Low". The Medium option was deliberately removed to encourage the speech LLM to make more decisive judgements rather than defaulting to neutral responses. This binary choice format forces the model to "pick a side" for each assessed trait, potentially leading to more discriminative features while also reducing the dimensionality of the feature space. The complete *Revised Question Set* is provided in Section A.5 in Appendix A.

4.4 Implementation Details

Our implementation leverages vLLM [Kwon et al., 2023] for speech LLM inference, providing high-throughput serving with optimized memory management through PagedAttention. The system uses a generic system prompt: *"You are a helpful assistant."* for simplicity. Note that typical generation parameters such as temperature, top-k, and top-p are irrelevant for our methodology, as we only extract the probability distribution over the vocabulary for the immediate next token following the prompt, rather than performing autoregressive text generation.

Given the Speak & Improve datasets restrictions prohibiting the use of model APIs and requiring only self-hosted open-source models [Knill et al., 2024], we deployed the 7B parameter Qwen2.5-Omni model on cloud GPU infrastructure using vLLM. This self-hosting approach presented significant engineering challenges, including configuring audio-specific tokenization schemes, managing multi-modal data processors, and handling audio encoding requirements that distinguish speech LLMs from standard text-based models.

A crucial optimization involves consecutive inference on the same audio sample to preserve the common key-value (KV) cache. When processing multiple questions for a single audio recording, all questions for that sample are processed sequentially before moving to the next audio file. This allows the model to maintain the computationally expensive audio representation in its KV cache across questions, dramatically reducing inference overhead compared to re-encoding the audio for each question. The scale of inference required was substantial, with approximately 48,000 audio files processed for each experiment experiment, and the multi-question approach multiplying the total number of inference calls by the number of questions in each set. To address I/O bottlenecks arising from reading audio files, we implemented batching strategies that process audio files in groups of 10,000 before passing them to vLLM, optimizing throughput by ensuring vLLM has sufficient data for its internal batching mechanisms.

For embedding extraction, we utilized Hugging Face's `transformers` library [Wolf et al., 2020]

and adopted the Apache Parquet file format [apa, 2013] for efficient storage and retrieval of the embedding data generated throughout our experiments.

In our experiments, we explore two strategies for posing questions to the speech LLM: sequential question inference and batch question inference. With **sequential question inference**, each question is posed to the speech LLM in separate inference calls. For a question-set with M questions, this requires M distinct inference requests. This allows the model to focus exclusively on one assessment dimension at a time, potentially leading to more focused and accurate evaluations. The prompt used for sequential inference is shown in Figure 4.1.

Sequential Inference Prompt Template

You are an examiner grading a spoken English test for non-native (L2) speakers.
You will be given an audio file of a candidate's speech, and you must answer the following question based on the audio content.
Question: {question}
Options: {options}
Please select **ONLY one** of the options ({option_keys}) that best describes the candidate's performance. Respond with the option only.
Answer:

Figure 4.1: Prompt template used for sequential question inference with the speech LLM.

With **batch question inference**, all questions are presented simultaneously in a single inference call, with the speech LLM producing a structured JSON response containing scores for all criteria. This strategy significantly reduces computational costs by requiring only one inference call per audio sample. However, the simultaneous consideration of multiple assessment criteria may lead to cognitive overload [Cheng et al., 2023], where the model struggles to maintain equal attention across all questions.

For the *Rubric-based Question Set* and the *Revised Question Set*, we evaluate both inference strategies to assess their impact on performance. The prompts used for batch inference are shown in Section A.4 and Section A.6 in Appendix A.

4.5 Calibration and Evaluation Metrics

Due to the distribution of scores in the dataset being dense in the middle and sparse at the extremes, any fitted model tends to heavily regress towards the mean. Following standard practice in the field [Qian et al., 2024, Ma et al., 2025], we compute calibration parameters to correct for this effect.

The calibration procedure consists of two stages, as illustrated in Figure 4.2. First, for each assessment part, we fit a linear transformation to predict the ground-truth part scores from the predicted part scores on the dev set. This linear transformation is then applied to all predicted part scores. The predicted overall score is computed as the average of these calibrated predicted part scores. Second, an additional linear calibration is applied to the predicted overall score by fitting a linear transformation on the dev set to predict the ground-truth overall scores from the predicted overall scores (which are themselves the average of the calibrated part scores). This two-stage calibration procedure represents standard practice in the field [Qian et al., 2024, Ma

et al., 2025] and is necessary to ensure that predicted scores are well-aligned with ground-truth scores. This calibration procedure is applied consistently to all models, including the baselines, to ensure fair comparison.

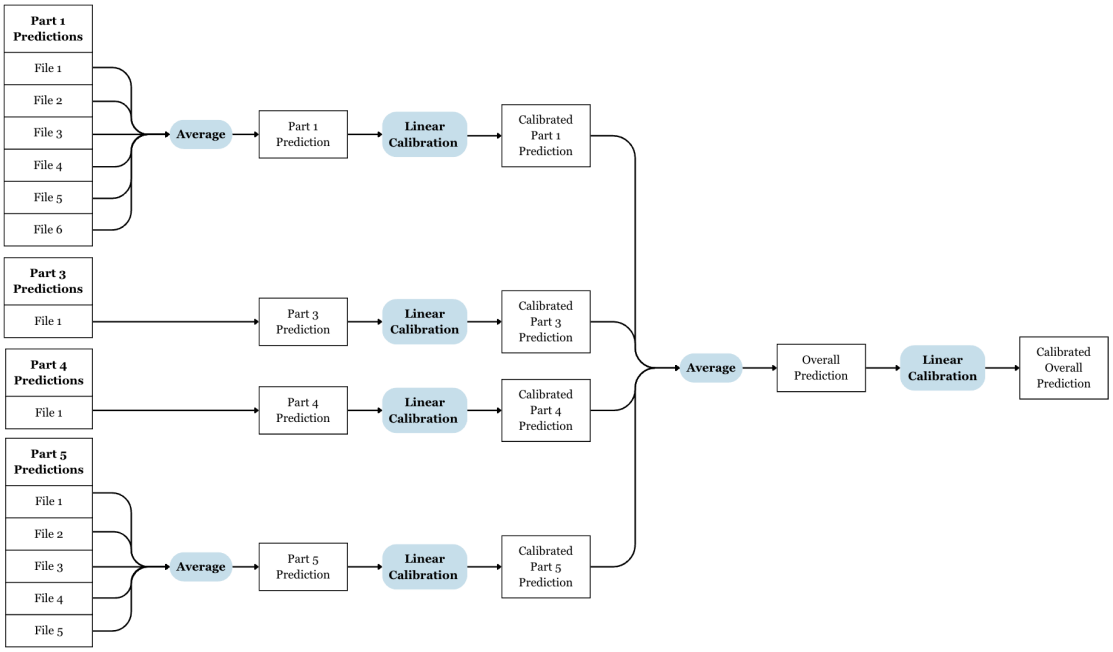

Figure 4.2: Illustration of the two-stage calibration procedure used to align predicted scores with ground-truth scores. First, predictions from all files of a given speaker are averaged to obtain speaker-level scores (this is relevant for parts 1 and 5 where multiple files exist per speaker). Next, a linear calibration is applied to each part's predicted scores. The overall predicted score is computed as the average of the calibrated part scores, followed by a second linear calibration to align with ground-truth overall scores. Note that each linear calibration has its own slope and intercept parameters that are fitted on the dev set.

For evaluation, we employ five standard metrics commonly used in spoken language assessment. Root Mean Square Error (RMSE) measures the average magnitude of prediction errors, with lower values indicating better performance. Pearson Correlation Coefficient (PCC) quantifies the linear relationship between predicted and ground-truth scores, ranging from -1 to 1, where values closer to 1 indicate stronger positive correlation. Spearman Rank Correlation (SRC) is essentially PCC applied to the ranks of the data rather than the raw values. It assesses the monotonic relationship between predictions and true scores, providing insight into how well the model preserves the relative ordering of submissions. Precision at 0.5 (P@0.5) represents the percentage of predictions within 0.5 points of the ground truth, while Precision at 1.0 (P@1.0) indicates the percentage within 1.0 point. Both precision metrics serve as measures of practical accuracy for educational assessment applications. It is important to note that the calibration procedure does not affect PCC or SRC, as these correlation metrics are invariant to linear transformations of the predictions.

Chapter 5

Results and Discussion

In this chapter, we present and discuss the results from various experiments with our proposed question-based approach. We begin by comparing different feature transformations in Section 5.1, followed by an evaluation of different regression models in Section 5.2. We then analyze the impact of different question sets on performance in Section 5.3 and provide visual inspection of these results in Section 5.4. Next, we analyze and discuss the interpretability of the extracted features in Section 5.5. Finally, we investigate the scalability of our approach in low data regimes in Section 5.6.

For consistency with the Speak & Improve Challenge 2025 [Qian et al., 2024] and Ma et al. [2025], all results reported in this chapter are evaluated on the dev set unless otherwise specified. Hyperparameters for all models were tuned using cross-validation on the training set, ensuring the dev set remained completely unseen during model development. Calibration parameters, as described in Section 4.5, were still computed on the dev set following standard practice. Results are reported as overall submission-level scores, computed as the average of predicted part scores across test components. All metrics are presented as mean ± standard deviation over 1000 bootstrapped samples of the data.

5.1 Comparison of Transformations

In this section, we compare different transformations of the speech LLM outputs to construct the feature matrix. We evaluate four different feature transformations of the *Initial Question Set* features to understand how the representation of the speech LLMs responses affects downstream performance. The **original feature matrix** (represented as "Original" in table 5.1) contains the raw log-probabilities of each option as output by the speech LLM. The **discrete feature matrix** (represented as "Discrete" in table 5.1) contains one-hot encoded representations of the selected option for each question (i.e., the argmax of the log-probabilities), effectively converting the continuous probability distribution into a categorical choice. The **exponentiated feature matrix** (represented as "Exp." in table 5.1) contains the exponentiated log-probabilities, which correspond to the actual probabilities of each option. Finally, the **softmax feature matrix** (represented as "Softmax" in table 5.1) contains the softmax-normalised probabilities of each option, ensuring that the probabilities for each question sum to 1, unlike the exponentiated feature

matrix where probabilities may sum to less than 1 due to the non-zero probabilities assigned to invalid tokens.

The results are shown in Table 5.1. Several key observations emerge from this comparison. First, the discrete feature matrix performs significantly worse than all continuous representations, indicating that the richer information contained in the probability distributions is crucial for performance. These findings can be explained by two theories: (1) the log-probabilities capture the model's uncertainty about each option, which provides valuable information for the regression model, and (2) the continuous probabilities help overcome the LLMs' positional bias [Liusie et al., 2023] of selecting the same options regardless of the question or audio content. Among the continuous representations, the exponentiated feature matrix performs best, followed closely by the original log-probability feature matrix, and then the softmax-normalised feature matrix. The poorer performance of the softmax-normalised features may be attributed to the fact that normalising only across valid options discards probability mass assigned to other tokens in the vocabulary, potentially reducing signal if the model's certainty over the full vocabulary is meaningful for assessment.

Feature Transformation	**RMSE**	**PCC**	**SRC**	**P@0.5**	**P@1.0**
Original Features (logprobs)	0.4278 ± 0.0086	0.7529 ± 0.0111	0.7525 ± 0.0121	77.67 ± 1.01	97.62 ± 0.35
Discrete Features	0.4964 ± 0.0097	0.6628 ± 0.0140	0.6661 ± 0.0148	69.94 ± 1.11	95.56 ± 0.48
Exponentiated Features (probs)	**0.4200 ± 0.0081**	**0.7640 ± 0.0103**	**0.7630 ± 0.0116**	**78.75 ± 1.00**	**97.75 ± 0.35**
Softmax Normalized Features (probs)	0.4401 ± 0.0088	0.7355 ± 0.0116	0.7390 ± 0.0125	76.33 ± 1.04	97.52 ± 0.37

Table 5.1: Performance comparison of linear models using different feature transformations applied to the *Initial Question Set*.

5.2 Comparison of Models

After establishing the best feature transformation, we compare different regression models trained on the exponentiated feature matrix from the *Initial Question Set*. Specifically, we compare linear regression, ridge regression, elastic net regression, KNN regression with various distance metrics (euclidean, manhattan), support vector regression (SVR) with linear and RBF kernels, XGBoost regression, and multi-layer perceptron regression networks. We performed comprehensive hyperparameter tuning for each model using cross-validation on the training set to ensure optimal performance. The results are shown in Table 5.2. The results indicate that linear regression performs comparably to all other models, indicating that the relationship between the extracted features and the final scores is largely linear.

Model	**RMSE**	**PCC**	**SRC**	**P@0.5**	**P@1.0**
Linear Regression	0.3945 ± 0.0150	0.8052 ± 0.0153	0.8150 ± 0.0165	81.76 ± 1.82	**97.46 ± 0.74**
Ridge Regression	0.3944 ± 0.0150	0.8053 ± 0.0153	0.8153 ± 0.0165	**81.99 ± 1.82**	**97.46 ± 0.74**
ElasticNet Regression	**0.3942 ± 0.0150**	**0.8055 ± 0.0153**	0.8155 ± 0.0165	81.76 ± 1.83	**97.46 ± 0.74**
KNN Regression	0.4180 ± 0.0162	0.7779 ± 0.0176	0.7878 ± 0.0187	79.43 ± 1.99	96.77 ± 0.83
Support Vector Regression	0.3944 ± 0.0151	0.8053 ± 0.0153	**0.8163 ± 0.0163**	81.73 ± 1.87	**97.46 ± 0.74**
XGBoost Regression	0.4068 ± 0.0158	0.7911 ± 0.0167	0.8009 ± 0.0178	80.63 ± 1.87	97.24 ± 0.77
MLP Regression	0.4049 ± 0.0161	0.7934 ± 0.0166	0.8103 ± 0.0165	79.90 ± 1.96	**97.46 ± 0.73**

Table 5.2: Performance comparison of various regression models using features from the *Initial Question Set*.

5.3 Comparison of Feature Sets

We now compare the performance of linear models trained on exponentiated feature matrices across different question sets. As detailed in Section 4.3, we evaluate four distinct question sets: the *Initial Question Set* with 14 questions using "High"/"Medium"/"Low" options, the *Direct Scoring Question* with 1 question using 6 CEFR levels, the *Rubric-based Question Set* with 3 questions using 6 CEFR levels, and the *Revised Question Set* with 30 questions using "High"/"Low" options. For the rubric-based and revised question sets, we compare both individual inference where each question is posed separately and batch inference where all questions are presented simultaneously in a single prompt, as described in Section 4.4.

Our comparison also includes two baseline methods: the *BERT Baseline*, which uses Whisper ASR followed by a BERT text grader, and the *Speech LLM Representations*, which extracts 3584-dimensional hidden state representations from the final layer of Qwen2.5-Omni conditioned on a direct scoring prompt, as detailed in Section 4.2.

To further analyze our approach, we evaluate a *Combined Question Set* that concatenates features from all individual question sets (including both batch and individual inference where applicable). Additionally, we implement random projections of the *Initial Question Set* to match the dimensionality of the *Combined Question Set* (204). This involves multiplying the initial 42-dimensional feature vector by a random matrix $\mathbf{R} \in \mathbb{R}^{42\times 204}$ where each entry is drawn from a uniform distribution over $[0, 1]$, allowing us to assess whether performance gains are simply due to higher dimensionality enabling the linear model to capture more complex relationships, or if the specific information contained in different question sets is beneficial.

Finally, we concatenate the *Combined Question Set* features with the speech LLM representations to determine whether our question-based approach provides complementary information beyond the black-box speech LLM features. The model trained on this feature set uses ridge regression, consistent with the *Speech LLM Representations*, to handle the high-dimensional feature space (204 combined question features plus 3584 speech LLM features).

Feature Set	**RMSE**	**PCC**	**SRC**	**P@0.5**	**P@1.0**
BERT Baseline	0.445	0.746	0.750	73.3	96.6
Initial Question Set	0.3945 ± 0.0150	0.8052 ± 0.0153	0.8150 ± 0.0165	81.76 ± 1.82	97.46 ± 0.74
Direct Scoring Question	0.4454 ± 0.0171	0.7427 ± 0.0206	0.7484 ± 0.0220	75.35 ± 2.14	96.76 ± 0.84
Rubric-based Question Set	0.4203 ± 0.0151	0.7751 ± 0.0170	0.7847 ± 0.0186	79.46 ± 1.92	97.26 ± 0.78
Rubric-based Question Set (Batch Inference)	0.4247 ± 0.0165	0.7696 ± 0.0185	0.7717 ± 0.0204	78.59 ± 2.02	96.55 ± 0.87
Revised Question Set	0.4163 ± 0.0168	0.7799 ± 0.0183	0.7905 ± 0.0184	78.94 ± 1.95	97.48 ± 0.73
Revised Question Set (Batch Inference)	0.4188 ± 0.0156	0.7769 ± 0.0176	0.7868 ± 0.0186	79.95 ± 1.87	97.24 ± 0.77
Combined Question Set	0.3872 ± 0.0153	0.8132 ± 0.0151	0.8245 ± 0.0153	82.87 ± 1.83	97.93 ± 0.67
Random Projections of *Initial Question Set*	0.3945 ± 0.0150	0.8052 ± 0.0153	0.8150 ± 0.0165	81.76 ± 1.82	97.46 ± 0.74
Speech LLM Representations	**0.3727 ± 0.0146**	**0.8282 ± 0.0144**	**0.8397 ± 0.0144**	85.84 ± 1.73	**98.17 ± 0.65**
Combined Question Set + Speech LLM Representations	**0.3727 ± 0.0147**	**0.8282 ± 0.0146**	0.8378 ± 0.0148	**86.28 ± 1.68**	**98.17 ± 0.65**

Table 5.3: Performance comparison of linear regression models across different feature representations (evaluated on dev set for overall scores).

The results are presented in Table 5.3. Additional results for individual parts of the overall assessment and eval set performance are provided in Appendix B. Several key insights emerge from this comparison:

BERT Baseline: The cascaded *BERT Baseline* achieves an RMSE of 0.445 and PCC of 0.746, which is significantly worse than all our question-based approaches, despite having far more features (768 dimensions from BERT). This indicates that our interpretable question-based features are able capture more relevant information for scoring than the text-based black-box

features extracted by BERT from ASR transcripts.

Question Set Complexity and Performance: There is a general trend showing that more comprehensive question sets lead to better performance. The *Direct Scoring Question*, using only a single question, achieves the weakest performance with an RMSE of 0.4454 ± 0.0171 and PCC of 0.7427 ± 0.0206. The *Rubric-based Question Set* (3 questions) improves substantially to an RMSE of 0.4203 ± 0.0151 and PCC of 0.7751 ± 0.0170, while the *Initial Question Set* (14 questions) performs even better with an RMSE of 0.3945 ± 0.0150 and PCC of 0.8052 ± 0.0153. However, this trend has a notable exception: the *Revised Question Set* (30 questions) slightly underperforms the *Initial Question Set* (RMSE of 0.4163 ± 0.0168 vs 0.3945 ± 0.0150), likely due to the reduction from three response options ("High"/"Medium"/"Low") to two ("High"/"Low"), which indicates that the granularity of response options is also a critical factor in capturing predictive information.

Inference Strategy Effects: Batch inference only slightly underperforms individual inference across both question sets tested. For the *Rubric-based Question Set*, batch inference shows an RMSE increase of 0.0044 and PCC decrease of 0.0055 compared to individual inference. For the *Revised Question Set*, batch inference results in an RMSE increase of 0.0025 and PCC decrease of 0.0030. This performance drop is relatively small especially considering the standard deviations derived from bootstrapping, suggesting that batch inference may be a viable option when computational efficiency is a priority.

Combining Question Sets: The *Combined Question Set*, which concatenates features from all individual question sets, achieves dramatically superior performance with an RMSE of 0.3872 ± 0.0153 and PCC of 0.8132 ± 0.0151, which is substantially better than any individual question set. Crucially, the random projection experiment maintains identical performance to the initial question set, demonstrating that the performance gains from combining question sets are not merely due to increased dimensionality. Instead, this finding indicates that different question sets capture complementary information about speaking proficiency, further supporting the general trend that more questions lead to better performance.

Speech LLM Representations: The *Speech LLM Representations* achieve near state-of-the-art performance with a RMSE of 0.3727 ± 0.0146 and PCC of 0.8282 ± 0.0144. None of our interpretable feature sets managed to surpass this performance. This finding suggests that our question-based method essentially queries different aspects of the same underlying audio representation rather than extracting fundamentally new information. One way to conceptualize this is that each question provides a different semantic and conceptually meaningful "view" of the same underlying audio representation that can be interpreted more naturally than the black-box representation itself. Furthermore, the fact that combining question-based features with speech LLM representations yields no significant improvements (RMSE of 0.3727 ± 0.0147 vs 0.3727 ± 0.0146) supports this hypothesis, indicating that the question-based features do not provide substantial complementary information beyond what is already captured by the speech LLMs embeddings.

Overall, this reflects the fundamental trade-off in machine learning between interpretability and predictive accuracy. Our question-based approach achieves competitive performance with dramatically fewer features, ranging from 6 to 60, compared to the 3584-dimensional speech LLM representations. More importantly, each of these features corresponds to a human-interpretable assessment criterion with clear conceptual meaning, enabling educators and learners to potentially understand which aspects of speaking proficiency contribute to the final score.

5.4 Visual Inspection

To better understand the comparative behaviour of our different feature sets, we provide a visual analysis of predictions versus ground truth scores.

Figure 5.1 shows the predictions from the *Initial Question Set* model versus ground truth scores. As with most predictive models, the regression line has a slope less than 1, reflecting the expected regression to the mean phenomenon [Galton, 1886], which is particularly pronounced given the datasets score distribution with most samples in the middle range (3.0–4.0) and sparse tails.

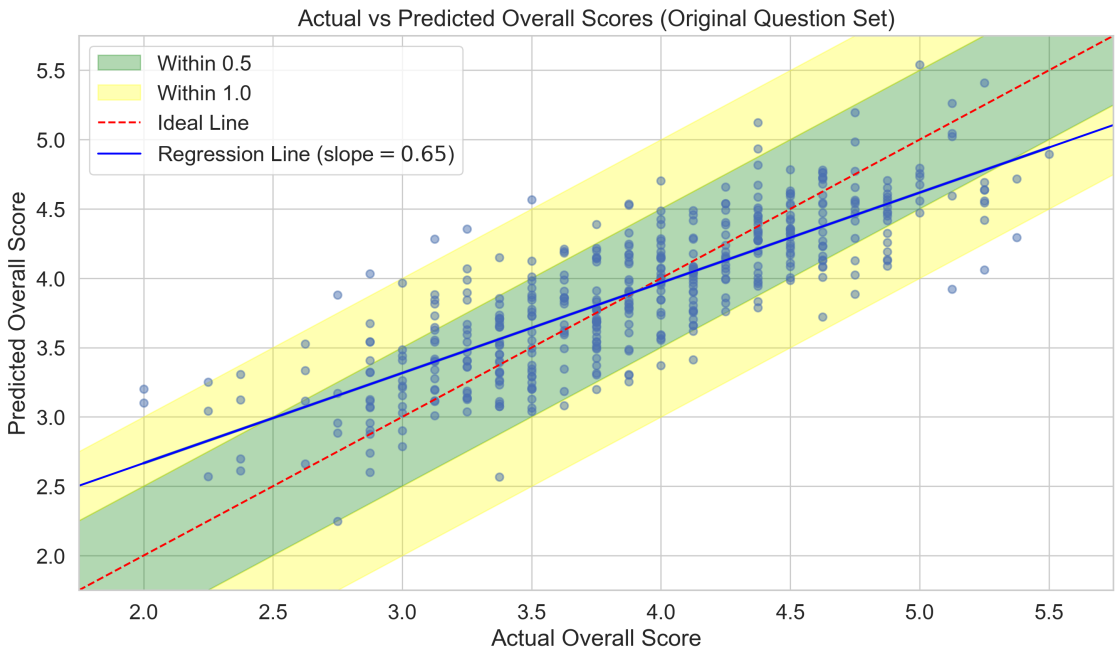

Figure 5.1: Predictions versus ground truth scores for the *Initial Question Set*. The diagonal line represents perfect predictions, while the fitted line shows the actual model behaviour.

More importantly, comparing across all feature sets in Figure 5.2 reveals differences in how well different feature representations capture the full range of scores. Feature sets with higher predictive performance (higher PCC) demonstrate slopes closer to 1 and better capture extreme values, with the *Speech LLM Representations* showing the steepest slope, followed by the *Combined Question Set*, and then individual question sets.

5.5 Interpretable Feature Analysis

In this section, we analyze the interpretable features extracted by our question-based approach to understand their relationships with each other and the target variable to extract useful insights.

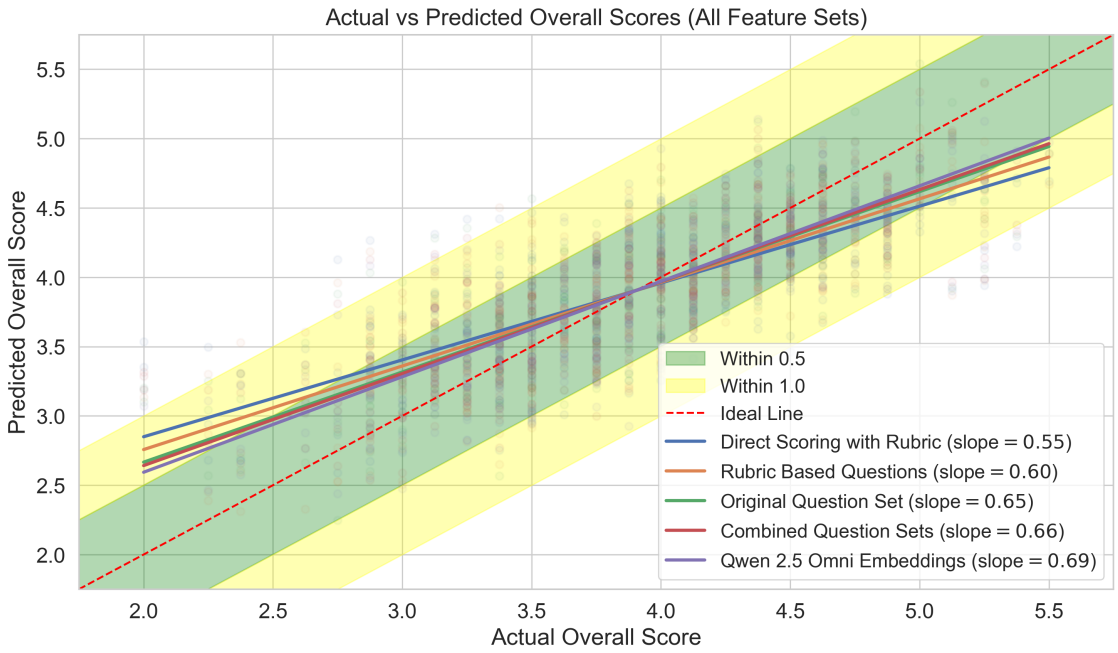

Figure 5.2: Regression lines of predictions versus ground truth scores across all feature sets. The plots show how different feature representations affect the regression line, with higher-dimensional features with higher PCC generally producing larger slopes that are able to better capture extreme target values.

5.5.1 Correlation Analysis

We begin by examining the correlation structure between features and the target variable - part score.

Figure 5.3 presents the correlation matrix for the *Initial Question Set*, revealing several important patterns. As expected, features exhibit intuitive correlations with part scores: "High" features correlate positively with the scores, "Low" features correlate negatively, and "Medium" features show mixed correlations depending on the specific question.

Additionally, we observe that "High" features are generally positively correlated with other "High" features and negatively correlated with Low features. Interestingly, these correlations seem to transcend rubric dimensions; for example, pronunciation and fluency questions show strong correlations with language resource questions. This could suggest that the speech LLM struggles to effectively distinguish between different dimensions of speaking proficiency, or it may reflect inherent correlations in the data itself, for example that candidates who speak more fluently also tend to use a wider range of vocabulary and grammar. That being said, the cross-dimension correlations seem to be slightly lower than within-dimension correlations, indicating that the speech LLM retains some ability to differentiate between different aspects of speaking proficiency.

Furthermore, some questions, such as pronunciation and fluency question 5, appear less correlated with others and also show weaker correlations with part scores, indicating they may contribute less predictive value to the overall assessment. Overall, the correlation analysis reveals high levels of multicollinearity across the feature set.

Figure 5.4 shows the correlation matrix for the *Rubric-based Question Set*.

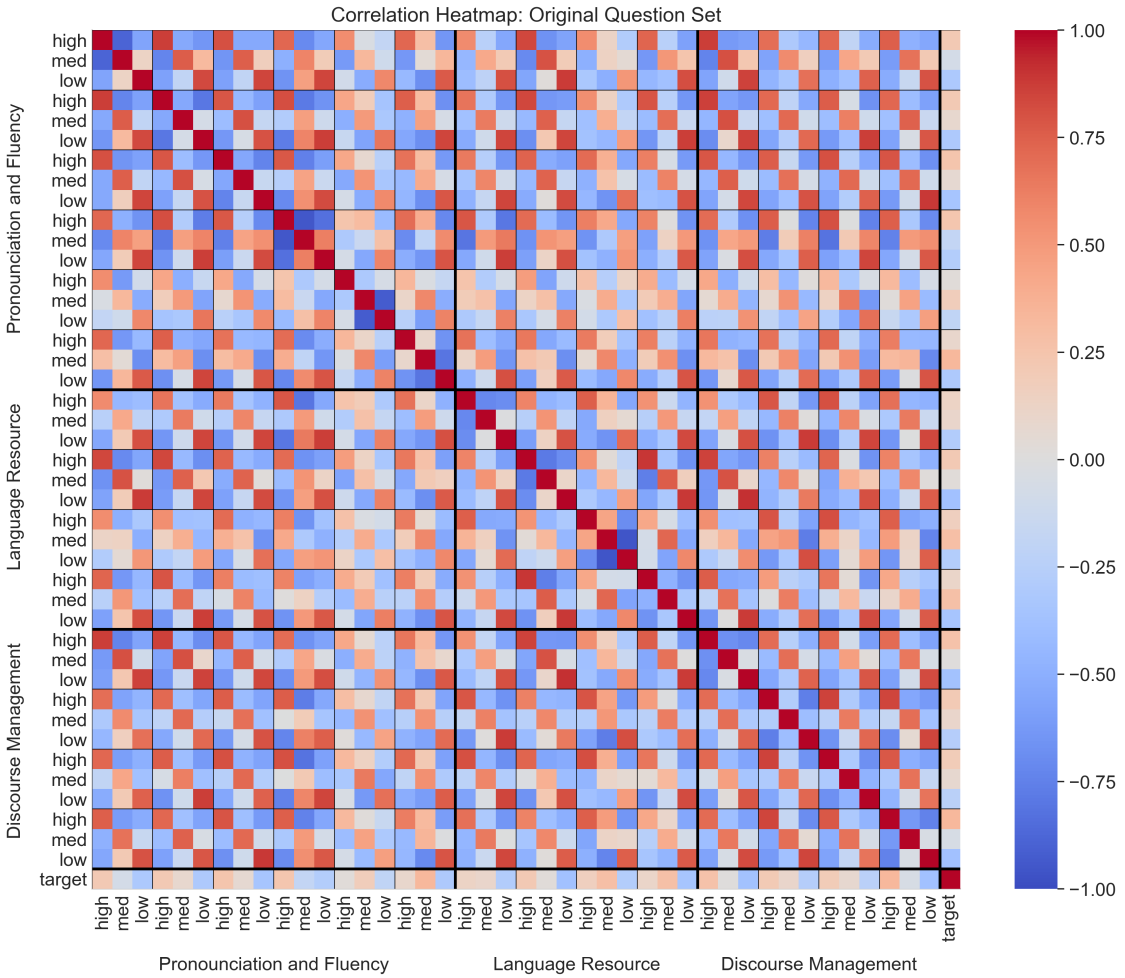

Figure 5.3: Correlation matrix between features and the target variable from the *Initial Question Set*. Each question has three features corresponding to the probabilities of the "High", "Medium", and "Low" options.

Grade A features exhibit very weak correlations with part scores, indicating that the speech LLM rarely assigns high probability to this option. Grade B features demonstrate the strongest positive correlations, while Grades C, D, and E show progressively more negative correlations. Grade F shows slightly less negative correlation than Grade E, likely due to similarly low assignment probabilities. Overall, the weaker alignment between feature correlations and prior expectations may stem from the increased number of response options (6 vs 3), which can lead to small and uninformative probability values assigned to extreme grades. Importantly, while grades show correlation across rubric dimensions, these cross-dimensional correlations are again weaker than within-dimension correlations, indicating that the speech LLM retains some ability to differentiate between different aspects of speaking proficiency.

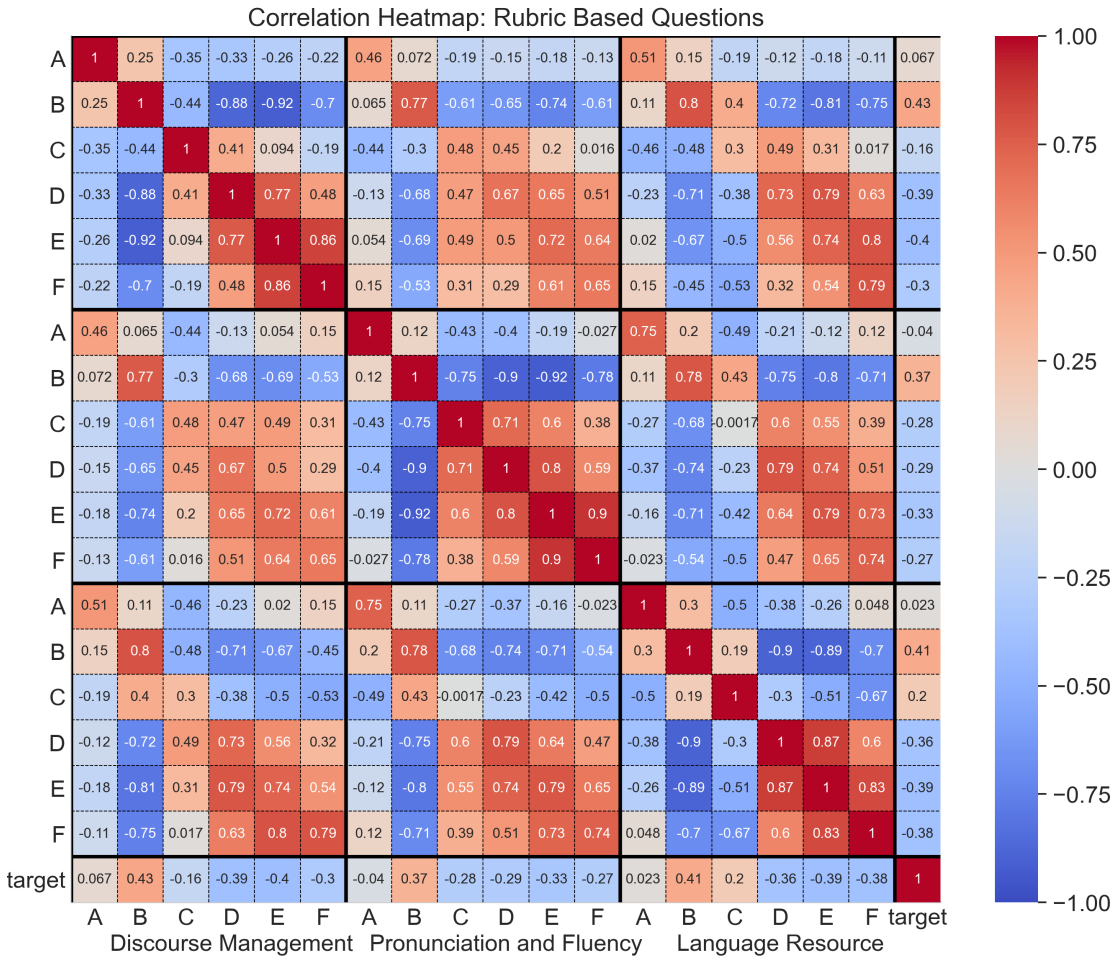

Figure 5.4: Correlation matrix between features and the target variable from the *Rubric-based Question Set*. Each rubric dimension has six features corresponding to the probabilities of grades A through F.

5.5.2 Regression Coefficients

Analysis of the regression coefficients for the *Rubric-based Question Set* reveals coefficients whose signs contradict intuitive expectations derived from correlation analysis, as illustrated in Figure 5.5. This counter-intuitive behaviour stems from the high degree of multicollinearity present among the extracted features. For example, Discourse Management Grade B has a strongly negative coefficient while Language Resource Grade B has a strongly positive coefficient, despite both showing similar correlation patterns with the target scores. Increasing the regularization parameter in ridge regression to mitigate multicollinearity makes coefficients more aligned with intuitive expectations, but this comes at the cost of significantly degraded model performance as excessive regularization effectively reduces the model to a constant predictor.

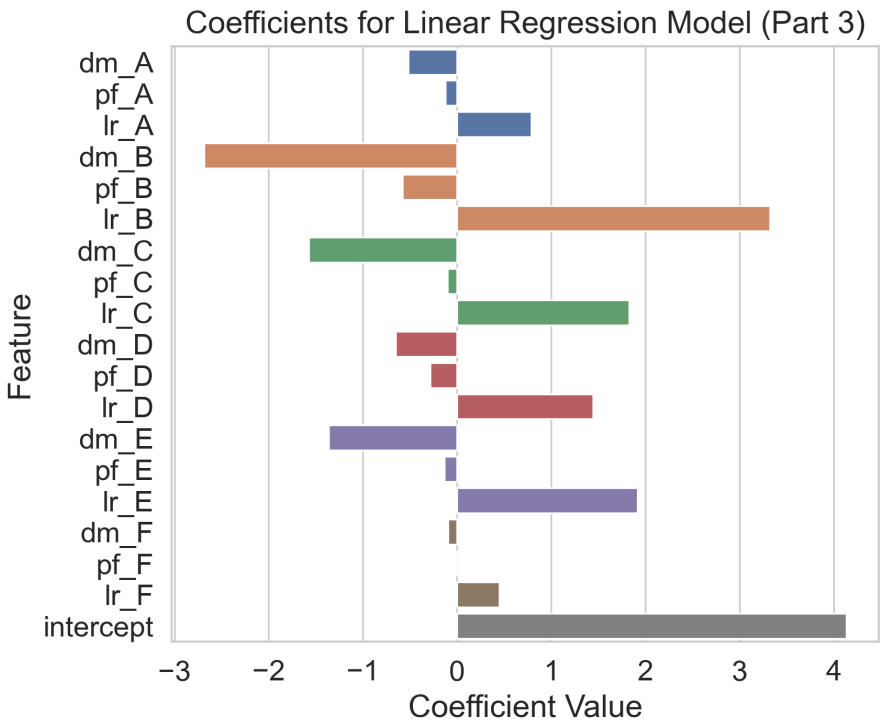

Figure 5.5: Regression coefficients for the *Rubric-based Question Set*. Each rubric dimension (pronunciation and fluency: pf, language resource: lr, discourse management: dm) has six grade options (A-F). The counter-intuitive coefficient signs demonstrate that high-grade features (A) don't consistently receive positive coefficients, and low-grade features (F) don't consistently receive negative coefficients, illustrating the impact of multicollinearity among features.

5.5.3 Question-level Scores

To address the interpretability challenges posed by multicollinearity while maintaining the pedagogical value of our approach, we implement an alternative interpretation strategy. We train separate linear regression models to predict part scores using only features corresponding to each individual rubric dimension. The predictions from these dimension-specific models can be interpreted as analytic scores for each rubric criterion, effectively decomposing the holistic scores into predicted analytic scores. We call these predictions **question-level scores**, as they represent the model's assessment of a candidate's performance specific to each question. The variability amongst these question-level scores provides insights into the candidate's strengths and weaknesses across different dimensions of speaking proficiency. This methodology is illustrated in Figure 5.6.

Figure 5.7 visualizes these dimension-specific predictions for various candidates, showing how each rubric criterion contributes to the overall assessment. This visualization proves especially valuable for educational applications, as it enables candidates to receive targeted feedback on their performance across different dimensions of speaking proficiency. Rather than receiving only a holistic score, learners can identify specific areas for improvement, such as pronunciation and fluency, language resource usage, or discourse management, facilitating more effective and focused practice strategies.

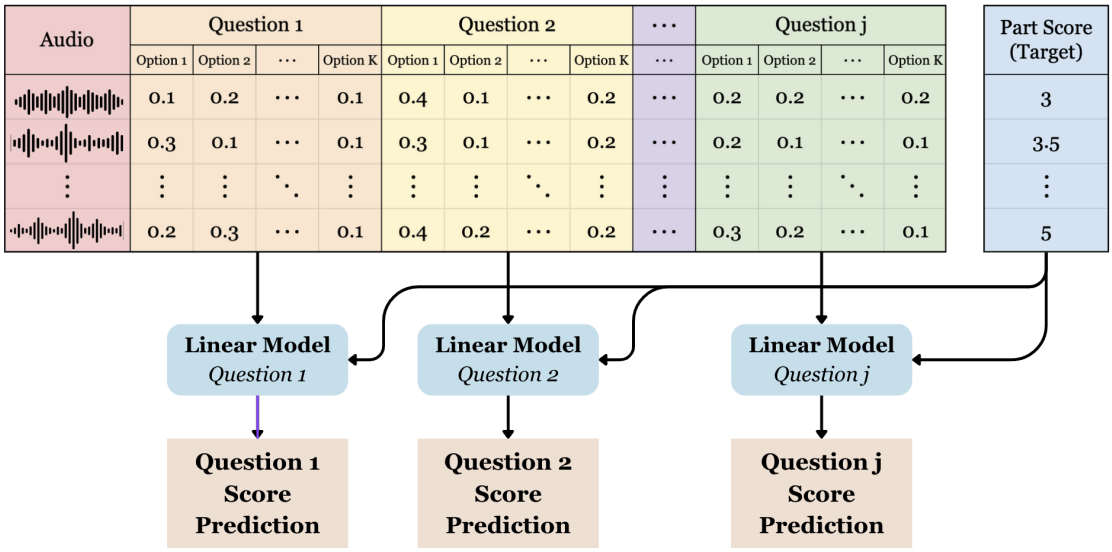

Figure 5.6: Illustration of the question-level scores methodology for obtaining analytic scores for each question. Separate linear models are trained to predict the same target variable (part score) for each question using only features corresponding to that question, and the predictions from these models are interpreted as question-level scores.

5.6 Low Data Regime

In this section, we investigate the scalability of speech LLM-based SLA, something that has not been studied in prior work. We evaluate the performance of the *Initial Question Set*, *Direct Scoring Question*, *Rubric-based Question Set*, and the *Speech LLM Representations* in simulated low data regimes by training models on various fractions of the training data and evaluating performance on the dev set. Specifically, we train models on 0.2% to 100% of the training data (approximately 6 to 3000 samples) and evaluate performance on the full dev set. Each experiment is repeated 100 times with different random seeds to account for variability in training data selection. We employ ridge regression with regularization strength selected via cross-validation on the training set for all models to ensure robustness even in low data regimes where the number of features may approach or greatly exceed the number of training samples. We report performance in terms of mean PCC and RMSE over the 100 trials for each training data fraction with standard deviations obtained from the performance variability within the 100 trials.

Figures 5.8 and 5.9 present the scaling curves for PCC and RMSE, respectively, revealing several important insights about the data efficiency of different feature representations.

Most remarkably, the *Rubric-based Question Set* demonstrates exceptional performance in extremely low data regimes. With only 0.2% of training data (approximately 6 samples), it achieves a PCC around 0.75 and RMSE around 0.44, which is comparable to the *BERT Baseline* trained on the full dataset. This suggests that speech LLMs can extract highly relevant information for scoring without requiring large amounts of training data, and that a well-chosen question set can enable effective SLA by guiding the model to focus on the most relevant aspects of the underlying audio representation. The superior performance of the *Rubric-based Question Set* compared to

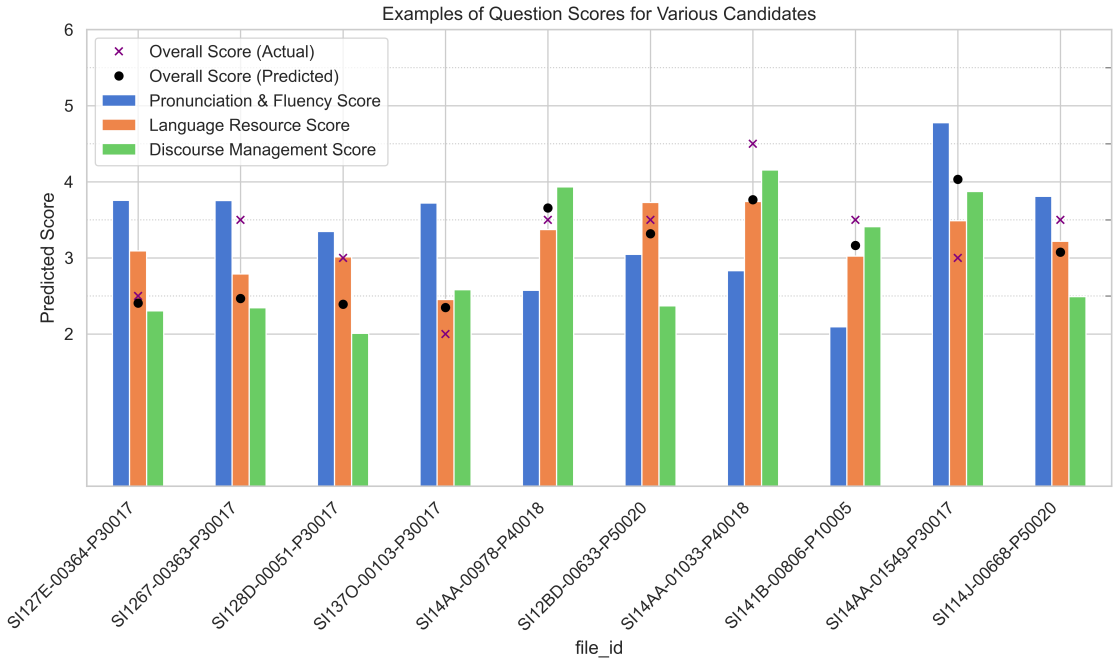

Figure 5.7: Question-level scores for different rubric dimensions across various sample candidates, illustrating how each criterion contributes to the overall assessment. Each file ID corresponds to a candidate's response for a specific test component.

other question-based approaches in this regime indicates that specific question design is critical, and that simply having more questions or options does not necessarily improve performance when training data is scarce.

Beyond approximately 0.8% of training data, the relative performance rankings stabilize, with more comprehensive feature sets generally achieving better performance. The *Speech LLM Representations* consistently outperform all question-based feature sets across all data regimes, while among the question-based approaches, the *Initial Question Set* (42 features) surpasses the *Rubric-based Question Set* (18 features) once sufficient training data becomes available. Interestingly, the question-based approaches appear to reach performance plateaus around 30% of the training data, while the *Speech LLM Representations* continue to improve with additional data, suggesting that question-based methods may have inherent asymptotic performance limitations compared to the full representational capacity of speech LLM embeddings.

The scaling curves also reveal substantially higher performance uncertainty in low data regimes, as evidenced by the larger standard deviations across trials. This increased variability indicates that the specific selection of training samples has a more pronounced impact on model performance when training data is limited, highlighting the importance of careful data curation in resource-constrained scenarios.

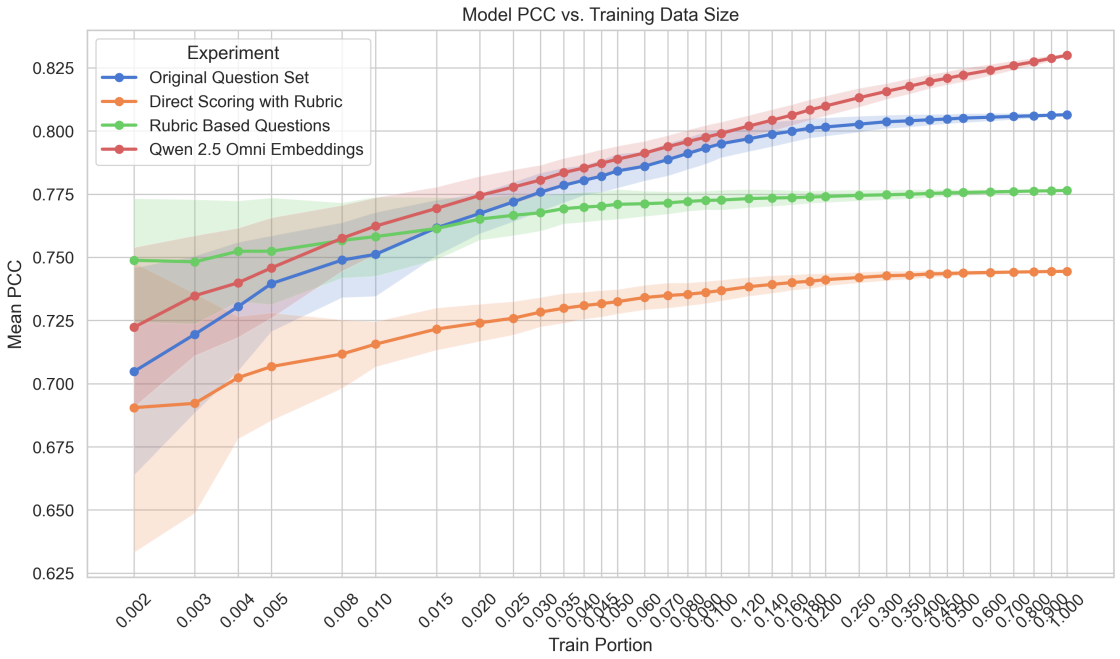

Figure 5.8: Scaling curves showing PCC as a function of training data size (log scale) across different feature sets. Each point represents the mean performance over 100 random sub-samples of the training data, with the shaded region indicating the standard deviation across the 100 sub-samples.

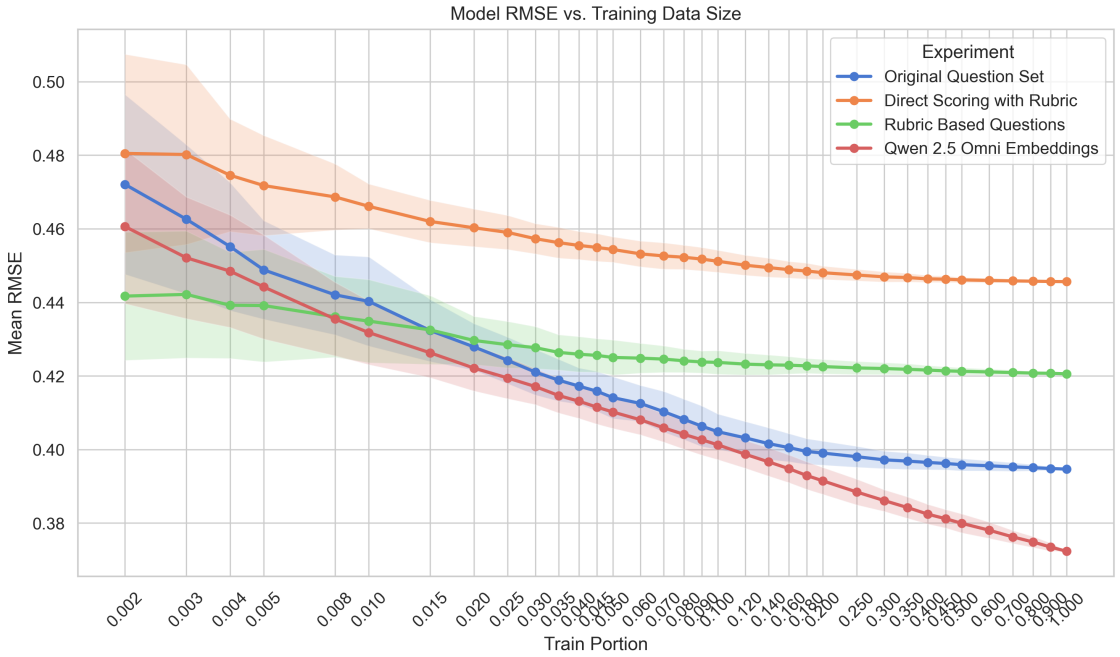

Figure 5.9: Scaling curves showing RMSE as a function of training data size (log-scale) across different feature sets. Each point represents the mean performance over 100 random sub-samples of the training data, with the shaded region indicating the standard deviation across the 100 sub-samples.

Chapter 6

Conclusion

This dissertation sought to answer the research question: **Can speech Large Language Models be used to develop interpretable systems for automatic spoken language assessment while maintaining competitive predictive performance?** Through multiple experiments on the Speak & Improve Corpus 2025, we investigated the effectiveness of question-based interpretable features derived from speech LLMs for automatic spoken language assessment (SLA).

6.1 Summary of Findings

Our question-based interpretable SLA approach achieves competitive performance with dramatically fewer features, attaining a PCC of 0.8052 with just 42 interpretable features compared to the black-box *Speech LLM Representation*'s PCC of 0.8282 with 3584 features. However, the direct use of speech LLM embeddings as features still outperforms our question-based approach across all configurations. Notably, combining question-based features with speech LLM embeddings yields no significant performance improvements, suggesting that our question-based method doesn't extract fundamentally new information but rather picks out relevant aspects from the same underlying audio representation in an interpretable manner.

Our experiments with feature transformations reveal that exponentiated features (probabilities) perform best, while discrete features perform significantly worse, indicating that the richer information contained in probability distributions is crucial for performance. The strong performance of simple linear regression models, comparable to more complex approaches like SVR and neural networks, demonstrates that the relationship between extracted features and final scores is largely linear. We also find that more questions generally lead to better performance, with PCC improving from 0.7427 for the *Direct Scoring Question* to 0.8052 for 14 questions in the *Initial Question Set*. Similarly, more options per question appear to improve performance, but with an important caveat: too many options can lead to extreme options consistently receiving low probabilities and becoming uninformative, as observed in the *Rubric-based Question Set*. Regarding inference strategy, batch inference only slightly underperforms individual inference (PCC decrease of 0.003-0.005), making it a viable option when computational efficiency is a priority.

Our correlation analysis reveals high multicollinearity among question-based features, even across different rubric dimensions, complicating the interpretation of regression coefficients which

often display counter-intuitive signs. This multicollinearity could indicate that the speech LLM struggles to effectively distinguish between different dimensions of speaking proficiency, or it may reflect inherent correlations in the data itself, such as candidates who speak more fluently also tending to use more sophisticated vocabulary. To address these interpretability challenges, we propose fitting separate regression models on features corresponding to each question to obtain question-level scores. These question-specific predictions effectively decompose holistic scores into analytic scores, providing valuable insights into candidates' strengths and weaknesses across different aspects of speaking proficiency.

In terms of scalability, the *Rubric-based Question Set* demonstrates exceptional performance in extremely low data regimes, achieving a PCC around 0.75 with just 0.2% of training data (approximately 6 samples), comparable to the *BERT Baseline* trained on the full dataset, and outperforming *Speech LLM Representations* in this regime in this regime. This suggests that speech LLMs can extract highly relevant information for scoring without requiring large amounts of training data, and that a well-chosen question set can enable effective SLA by guiding the model to focus on the most relevant aspects of the underlying audio representation. However, other question-based approaches do not perform as well in this regime, indicating that specific question design is critical.

Overall, these results collectively highlight the trade-off between interpretability and predictive accuracy in automated spoken language assessment, while demonstrating the viability of question-based interpretable features even in resource-constrained scenarios.

6.2 Implications for SLA and Educational Technology

Our question-based approach enables targeted feedback by decomposing holistic scores into interpretable dimensions such as pronunciation, fluency, and language resource usage. This allows learners to identify specific weaknesses and focus practice efforts on areas most likely to yield improvement, while educators can design personalized interventions based on individual learner profiles. The transparency of assessment criteria makes evaluation more understandable and actionable compared to black-box systems that provide only unexplained holistic scores.

The data efficiency demonstrated in our experiments—competitive performance with just 10% of training data—significantly reduces barriers to developing SLA systems. This makes automated assessment feasible for less-resourced languages and specialized domains where extensive annotated datasets are unavailable, potentially democratizing access to high-quality spoken language assessment tools worldwide.

6.3 Limitations and Future Work

A fundamental limitation of our approach is the lack of guarantees that the extracted interpretable features truly capture the intended linguistic concepts they are designed to assess. While our correlation analysis and the success of similar methodologies in other domains provide some evidence for conceptual alignment, we cannot definitively establish that high probabilities for "pronunciation quality" features genuinely reflect superior pronunciation rather than other correlated factors.

Future work should conduct comprehensive validation studies comparing our automatically generated question-level scores with expert assessments on the same dimensions, potentially through expert annotation of the specific traits our questions target.

The design of effective question sets remains a significant challenge. Our questions were developed through a combination of expert knowledge and LLM assistance, but creating question sets that are both comprehensive and yield disentangled features represents an open research problem. The high multicollinearity observed in our extracted features demonstrates the difficulty of designing questions that capture distinct aspects of proficiency while maintaining predictive power. Future work could explore automatic question generation through prompt optimization techniques, potentially using iterative refinement based on feature independence criteria or predictive performance on held-out data.

The question-based methodology we propose demonstrates considerable generality and could be readily adapted to other modalities and assessment tasks. Beyond spoken language assessment, this approach could be applied to any regression or classification task on unstructured data where interpretability is desired. Furthermore, similar approaches could be employed in other modalities such as pure text, images, or videos. The core principle of decomposing complex evaluation tasks into structured, interpretable sub-questions represents a promising direction for developing explainable AI systems across domains where human-interpretable feedback is crucial.

An important limitation of our analysis is that we lack the domain expertise to conduct meaningful error analysis at the individual candidate level. While we attempted to examine specific examples where different models achieved varying levels of accuracy—identifying cases where our question-based approach succeeded or failed relative to baseline methods—we were unable to discern meaningful patterns in these differences. Without deep expertise in spoken language assessment and the nuanced factors that influence human grader decisions, we could not determine why certain candidates received lower scores than others or why specific models performed better on particular types of responses. This limitation highlights the importance of involving domain experts in future research, particularly for conducting qualitative error analysis that could reveal insights into model behaviour and guide improvements to question design and feature extraction strategies.

Finally, bias and fairness considerations, while not explicitly studied in this work, represent critical concerns for any deployed assessment system. Future research should systematically investigate whether speech LLM-based features exhibit differential performance across demographic groups, including but not limited to native language background, gender, age, and socioeconomic status. The interpretable nature of our question-based features may actually provide advantages in this regard, as biased model behaviour could be more easily identified and addressed through targeted interventions on specific questions rather than requiring modifications to opaque black-box systems. Developing bias detection and mitigation strategies specifically tailored to question-based interpretable features represents an important avenue for ensuring equitable deployment of automated assessment technologies.

Bibliography

Apache parquet: A columnar storage file format. https://parquet.apache.org, 2013. Accessed: YYYY-MM-DD.

Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations, 2020. URL https://arxiv.org/abs/2006.11477.

Vojtěch Balek, Lukáš Sýkora, Vilém Sklenák, and Tomáš Kliegr. Llm-based feature generation from text for interpretable machine learning, 2025. URL https://arxiv.org/abs/2409.07132.

Randall Balestriero, Mark Ibrahim, Vlad Sobal, Ari Morcos, Shashank Shekhar, Tom Goldstein, Florian Bordes, Adrien Bardes, Gregoire Mialon, Yuandong Tian, Avi Schwarzschild, Andrew Gordon Wilson, Jonas Geiping, Quentin Garrido, Pierre Fernandez, Amir Bar, Hamed Pirsiavash, Yann LeCun, and Micah Goldblum. A cookbook of self-supervised learning, 2023. URL https://arxiv.org/abs/2304.12210.

Stefano Bannò, Hari K. Vydana, Kate M. Knill, and Mark J. F. Gales. Can GPT-4 do L2 analytic assessment? In Ekaterina Kochmar, Marie Bexte, Jill Burstein, Andrea Horbach, Ronja Laarmann-Quante, Anaïs Tack, Victoria Yaneva, and Zheng Yuan, editors, *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 149–164, Mexico City, Mexico, June 2024. Association for Computational Linguistics. URL https://aclanthology.org/2024.bea-1.14/.

Stefano Bannò and Marco Matassoni. Proficiency assessment of l2 spoken english using wav2vec 2.0, 2022. URL https://arxiv.org/abs/2210.13168.

Stefano Bannò, Kate M. Knill, Marco Matassoni, Vyas Raina, and Mark J. F. Gales. L2 proficiency assessment using self-supervised speech representations, 2022. URL https://arxiv.org/abs/2211.08849.

Iz Beltagy, Matthew E. Peters, and Arman Cohan. Longformer: The long-document transformer, 2020. URL https://arxiv.org/abs/2004.05150.

Vinamra Benara, Chandan Singh, John X. Morris, Richard Antonello, Ion Stoica, Alexander G. Huth, and Jianfeng Gao. Crafting interpretable embeddings by asking llms questions, 2024. URL https://arxiv.org/abs/2405.16714.

Jared Bernstein, Michael Cohen, Hy Murveit, Dimitry Rtischev, and Mitchel Weintraub. Automatic evaluation and training in english pronunciation. In *First International Conference on Spoken Language Processing (ICSLP 1990)*, pages 1185–1188, 1990. doi: 10.21437/ICSLP.1990-313.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners, 2020. URL https://arxiv.org/abs/2005.14165.

Cambridge English. Linguaskill speaking global assessment criteria. https://www.cambridgeenglish.org/Images/605504-linguaskill-speaking-assessment-criteria.pdf, 2020. Accessed: 2025-09-10.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations, 2020. URL https://arxiv.org/abs/2002.05709.

Zhoujun Cheng, Jungo Kasai, and Tao Yu. Batch prompting: Efficient inference with large language model apis, 2023. URL https://arxiv.org/abs/2301.08721.

SeongYeub Chu, JongWoo Kim, Bryan Wong, and MunYong Yi. Rationale behind essay scores: Enhancing s-llm's multi-trait essay scoring with rationale generated by llms, 2025. URL https://arxiv.org/abs/2410.14202.

Yunfei Chu, Jin Xu, Xiaohuan Zhou, Qian Yang, Shiliang Zhang, Zhijie Yan, Chang Zhou, and Jingren Zhou. Qwen-audio: Advancing universal audio understanding via unified large-scale audio-language models, 2023. URL https://arxiv.org/abs/2311.07919.

Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhifang Guo, Yichong Leng, Yuanjun Lv, Jinzheng He, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen2-audio technical report, 2024. URL https://arxiv.org/abs/2407.10759.

Council of Europe. *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*. Cambridge University Press, Cambridge, 2001.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019. URL https://arxiv.org/abs/1810.04805.

Heejin Do, Yunsu Kim, and Gary Geunbae Lee. Autoregressive score generation for multi-trait essay scoring, 2024. URL https://arxiv.org/abs/2403.08332.

Sohaila Eltanbouly, Salam Albatarni, and Tamer Elsayed. Trates: Trait-specific rubric-assisted cross-prompt essay scoring, 2025. URL https://arxiv.org/abs/2505.14577.

Kaiqi Fu, Linkai Peng, Nan Yang, and Shuran Zhou. Pronunciation assessment with multi-modal large language models, 2024. URL https://arxiv.org/abs/2407.09209.

Francis Galton. Regression towards mediocrity in hereditary stature., January 1886. URL https://doi.org/10.2307/2841583.

Sreyan Ghosh, Sonal Kumar, Ashish Seth, Chandra Kiran Reddy Evuru, Utkarsh Tyagi, S Sakshi, Oriol Nieto, Ramani Duraiswami, and Dinesh Manocha. Gama: A large audio-language model with advanced audio understanding and complex reasoning abilities, 2024. URL https://arxiv.org/abs/2406.11768.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. Hubert: Self-supervised speech representation learning by masked prediction of hidden units, 2021. URL https://arxiv.org/abs/2106.07447.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models, 2021. URL https://arxiv.org/abs/2106.09685.

Eesung Kim, Jae-Jin Jeon, Hyeji Seo, and Hoon Kim. Automatic pronunciation assessment using self-supervised speech representation learning, 2022. URL https://arxiv.org/abs/2204.03863.

Kate Knill, Diane Nicholls, Mark J.F. Gales, Mengjie Qian, and Pawel Stroinski. Speak & improve corpus 2025: an l2 english speech corpus for language assessment and feedback. *arXiv preprint arXiv:2412.11986*, 2024. URL https://arxiv.org/abs/2412.11986.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention, 2023. URL https://arxiv.org/abs/2309.06180.

Sanwoo Lee, Yida Cai, Desong Meng, Ziyang Wang, and Yunfang Wu. Unleashing large language models' proficiency in zero-shot essay scoring, 2024. URL https://arxiv.org/abs/2404.04941.

Adian Liusie, Potsawee Manakul, and Mark J. F. Gales. Mitigating word bias in zero-shot prompt-based classifiers, 2023. URL https://arxiv.org/abs/2309.04992.

Karen Ludlow. *Official Quick Guide to Linguaskill*. Cambridge University Press, 2020. ISBN 978-1108885256.

Rao Ma, Mengjie Qian, Siyuan Tang, Stefano Bannò, Kate M. Knill, and Mark J.F. Gales. Assessment of l2 oral proficiency using speech large language models. *arXiv preprint arXiv:2505.21148*, 2025. URL https://arxiv.org/abs/2505.21148.

Watheq Mansour, Salam Albatarni, Sohaila Eltanbouly, and Tamer Elsayed. Can large language models automatically score proficiency of written essays?, 2024. URL https://arxiv.org/abs/2403.06149.

Sandeep Mathias and Pushpak Bhattacharyya. ASAP++: Enriching the ASAP automated essay grading dataset with essay attribute scores. In Nicoletta Calzolari, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Koiti Hasida, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, Stelios Piperidis, and Takenobu Tokunaga, editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May 2018. European Language Resources Association (ELRA). URL https://aclanthology.org/L18-1187/.

Denis Jered McInerney, Geoffrey Young, Jan-Willem van de Meent, and Byron C. Wallace. Chill: Zero-shot custom interpretable feature extraction from clinical notes with large language models, 2023. URL https://arxiv.org/abs/2302.12343.

Simon Webster McKnight, Arda Civelekoglu, Mark JF Gales, Stefano Banno, Adian Liusie, and Kate M Knill. Automatic assessment of conversational speaking tests. 2023. doi: 10.17863/CAM.99725. URL https://www.repository.cam.ac.uk/handle/1810/353637.

Microsoft, :, Abdelrahman Abouelenin, Atabak Ashfaq, Adam Atkinson, Hany Awadalla, Nguyen Bach, Jianmin Bao, Alon Benhaim, Martin Cai, Vishrav Chaudhary, Congcong Chen, Dong Chen, Dongdong Chen, Junkun Chen, Weizhu Chen, Yen-Chun Chen, Yi ling Chen, Qi Dai, Xiyang Dai, Ruchao Fan, Mei Gao, Min Gao, Amit Garg, Abhishek Goswami, Junheng Hao, Amr Hendy, Yuxuan Hu, Xin Jin, Mahmoud Khademi, Dongwoo Kim, Young Jin Kim, Gina Lee, Jinyu Li, Yunsheng Li, Chen Liang, Xihui Lin, Zeqi Lin, Mengchen Liu, Yang Liu, Gilsinia Lopez, Chong Luo, Piyush Madan, Vadim Mazalov, Arindam Mitra, Ali Mousavi, Anh Nguyen, Jing Pan, Daniel Perez-Becker, Jacob Platin, Thomas Portet, Kai Qiu, Bo Ren, Liliang Ren, Sambuddha Roy, Ning Shang, Yelong Shen, Saksham Singhal, Subhojit Som, Xia Song, Tetyana Sych, Praneetha Vaddamanu, Shuohang Wang, Yiming Wang, Zhenghao Wang, Haibin Wu, Haoran Xu, Weijian Xu, Yifan Yang, Ziyi Yang, Donghan Yu, Ishmam Zabir, Jianwen Zhang, Li Lyna Zhang, Yunan Zhang, and Xiren Zhou. Phi-4-mini technical report: Compact yet powerful multimodal language models via mixture-of-loras, 2025. URL https://arxiv.org/abs/2503.01743.

Atsushi Mizumoto and Masaki Eguchi. Exploring the potential of using an ai language model for automated essay scoring. *Research Methods in Applied Linguistics*, 2(2), August 2023. ISSN 2772-7661. doi: 10.1016/j.rmal.2023.100050. Publisher Copyright: © 2023 The Author(s).

Ben Naismith, Phoebe Mulcaire, and Jill Burstein. Automated evaluation of written discourse coherence using GPT-4. In Ekaterina Kochmar, Jill Burstein, Andrea Horbach, Ronja Laarmann-Quante, Nitin Madnani, Anaïs Tack, Victoria Yaneva, Zheng Yuan, and Torsten Zesch, editors, *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 394–403, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.bea-1.32. URL https://aclanthology.org/2023.bea-1.32/.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red

Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo,

Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. Gpt-4 technical report, 2024. URL https://arxiv.org/abs/2303.08774.

Ellis B. Page. The imminence of... grading essays by computer. *The Phi Delta Kappan*, 47(5): 238–243, 1966. ISSN 00317217. URL http://www.jstor.org/stable/20371545.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Linkai Peng, Kaiqi Fu, Binghuai Lin, Dengfeng Ke, and Jinsong Zhan. A study on fine-tuning wav2vec2.0 model for the task of mispronunciation detection and diagnosis. In *Interspeech 2021*, pages 4448–4452, 2021. doi: 10.21437/Interspeech.2021-1344.

Mengjie Qian, Kate Knill, Stefano Bannò, Siyuan Tang, Penny Karanasou, Mark J.F. Gales, and Diane Nicholls. Speak & improve challenge 2025: Tasks and baseline systems. *arXiv preprint arXiv:2412.11985*, 2024. URL https://arxiv.org/abs/2412.11985.

Xiaojun Qian, Helen Meng, and Frank K. Soong. The use of dbn-hmms for mispronunciation detection and diagnosis in l2 english to support computer-aided pronunciation training. In *Interspeech 2012*, pages 775–778, 2012. doi: 10.21437/Interspeech.2012-238.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision, 2022. URL https://arxiv.org/abs/2212.04356.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer, 2023. URL https://arxiv.org/abs/1910.10683.

Vyas Raina, Mark J.F. Gales, and Kate M. Knill. Universal adversarial attacks on spoken language assessment systems. In *Interspeech 2020*, pages 3855–3859, 2020. doi: 10.21437/Interspeech.2020-1890.

Paul K. Rubenstein, Chulayuth Asawaroengchai, Duc Dung Nguyen, Ankur Bapna, Zalán Borsos, Félix de Chaumont Quitry, Peter Chen, Dalia El Badawy, Wei Han, Eugene Kharitonov, Hannah Muckenhirn, Dirk Padfield, James Qin, Danny Rozenberg, Tara Sainath, Johan Schalkwyk, Matt Sharifi, Michelle Tadmor Ramanovich, Marco Tagliasacchi, Alexandru Tudor, Mihajlo Velimirović, Damien Vincent, Jiahui Yu, Yongqiang Wang, Vicky Zayats, Neil Zeghidour, Yu Zhang, Zhishuai Zhang, Lukas Zilka, and Christian Frank. Audiopalm: A large language model that can speak and listen, 2023. URL https://arxiv.org/abs/2306.12925.

S Sakshi, Utkarsh Tyagi, Sonal Kumar, Ashish Seth, Ramaneswaran Selvakumar, Oriol Nieto, Ramani Duraiswami, Sreyan Ghosh, and Dinesh Manocha. Mmau: A massive multi-task audio understanding and reasoning benchmark, 2024. URL https://arxiv.org/abs/2410.19168.

Dylan Sam, Marc Finzi, and J. Zico Kolter. Predicting the performance of black-box llms through self-queries, 2025. URL https://arxiv.org/abs/2501.01558.

Maja Stahl, Leon Biermann, Andreas Nehring, and Henning Wachsmuth. Exploring llm prompting strategies for joint essay scoring and feedback generation, 2024. URL https://arxiv.org/abs/2404.15845.

Changli Tang, Wenyi Yu, Guangzhi Sun, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun Ma, and Chao Zhang. Salmonn: Towards generic hearing abilities for large language models, 2024. URL https://arxiv.org/abs/2310.13289.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open foundation and fine-tuned chat models, 2023. URL https://arxiv.org/abs/2307.09288.

Xinhao Wang, Keelan Evanini, Yao Qian, and Matthew Mulholland. Automated scoring of spontaneous speech from young learners of english using transformers. In *2021 IEEE Spoken Language Technology Workshop (SLT)*, pages 705–712, 2021. doi: 10.1109/SLT48900.2021.9383553.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi 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 M. Rush. Huggingface's transformers: State-of-the-art natural language processing, 2020. URL https://arxiv.org/abs/1910.03771.

Minglin Wu, Kun Li, Wai-Kim Leung, and Helen Meng. Transformer based end-to-end mispronunciation detection and diagnosis. pages 3954–3958, 08 2021. doi: 10.21437/Interspeech.2021-1467.

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, Bin Zhang, Xiong Wang, Yunfei Chu, and Junyang Lin. Qwen2.5-omni technical report, 2025. URL https://arxiv.org/abs/2503.20215.

Xiaoshuo Xu, Yueteng Kang, Songjun Cao, Binghuai Lin, and Long Ma. Explore wav2vec 2.0 for mispronunciation detection. In *Interspeech 2021*, pages 4428–4432, 2021. doi: 10.21437/Interspeech.2021-777.

Kevin P. Yancey, Geoffrey Laflair, Anthony Verardi, and Jill Burstein. Rating short L2 essays on the CEFR scale with GPT-4. In Ekaterina Kochmar, Jill Burstein, Andrea Horbach, Ronja Laarmann-Quante, Nitin Madnani, Anaïs Tack, Victoria Yaneva, Zheng Yuan, and Torsten Zesch, editors, *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 576–584, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.bea-1.49. URL https://aclanthology.org/2023.bea-1.49/.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. Xlnet: Generalized autoregressive pretraining for language understanding, 2020. URL https://arxiv.org/abs/1906.08237.

Appendix A

Full Question Sets and Prompts

In this chapter, we provide the full question sets and text prompts used in our experiments.

A.1 Initial Question Set

In this section, we list all the questions in the *Initial Question Set*. The questions are divided into three categories: Pronunciation and Fluency, Language Resource, and Discourse Management. Each question has three response options: "High", "Medium", and "Low".

Pronunciation and Fluency

1. **Intelligibility — Can an average listener understand without strain or repetition?**
    - **High:** Speech is always clear; no repetition or clarification needed.
    - **Medium:** Mostly clear; occasional repetition or clarification needed.
    - **Low:** Often unclear; listener must frequently ask for repetition or guesses meaning.
2. **Flow & Pausing — Is speech smooth and continuous?**
    - **High:** Smooth, continuous flow with only natural pauses.
    - **Medium:** Generally smooth but noticeable hesitations, false starts, or self-corrections.
    - **Low:** Choppy; long or frequent pauses and restarts disrupt comprehension.
3. **Stress & Rhythm — Are English stress patterns followed?**
    - **High:** Correct word and sentence stress; rhythm sounds natural and aids meaning.
    - **Medium:** Mostly correct but some misplaced or missing stresses.
    - **Low:** Stress patterns frequently wrong or monotone, obscuring meaning.
4. **Intonation — Does pitch vary appropriately to signal meaning?**
    - **High:** Varied, natural intonation that supports message and attitude.
    - **Medium:** Limited pitch range but some appropriate rises and falls.

- **Low:** Flat or erratic pitch; intonation confuses or fails to signal intent.

5. **L1 (first language) Phonological Influence — Do first-language sounds interfere?**
   - **High:** L1 influence minimal; never impedes understanding.
   - **Medium:** Noticeable L1 features but rarely impede comprehension.
   - **Low:** Strong L1 interference regularly causes misunderstanding.
6. **Automaticity of Retrieval — Does the speaker search for basic words or forms?**
   - **High:** Lexis and grammar retrieved instantly; no searching pauses.
   - **Medium:** Occasional brief pauses to find words or forms.
   - **Low:** Frequent pauses or reformulations to recall basic language.

Language Resource

1. **Vocabulary Range — How wide is the speaker's word choice?**
   - **High:** Broad, topic-specific lexis; idioms and collocations used flexibly.
   - **Medium:** Adequate range for the task; limited idiomatic use.
   - **Low:** Basic, repetitive vocabulary; gaps limit expression.
2. **Vocabulary Accuracy & Appropriacy — Are words precise and well-chosen?**
   - **High:** Words fit context and collocate naturally; register appropriate.
   - **Medium:** Occasional awkward or imprecise choices, but meaning clear.
   - **Low:** Frequent inappropriate, mistranslated, or vague words that blur meaning.
3. **Grammatical Range & Complexity — Are complex structures attempted?**
   - **High:** Frequent, accurate use of subordinate clauses and varied sentence types.
   - **Medium:** Mix of simple and some complex forms with errors.
   - **Low:** Mostly simple clauses; complexity avoided or unsuccessful.
4. **Grammatical Accuracy — How error-free is the grammar?**
   - **High:** Only minor slips; errors never impede meaning.
   - **Medium:** Regular minor errors; occasional meaning disturbance.
   - **Low:** Frequent errors that confuse or obscure meaning.

Discourse Management

1. **Overall Logic & Coherence — Is the main line of thought easy to follow?**
   - **High:** Argument/thread immediately clear and logical throughout.
   - **Medium:** Generally coherent but some jumps or unclear links.

- **Low:** Hard to follow; ideas seem random or disjointed.

2. **Progression & Development of Ideas — Are ideas expanded and linked?**
   - **High:** Points are developed with explanations, details, or examples.
   - **Medium:** Some expansion but parts remain thin or under-developed.
   - **Low:** Ideas presented as isolated statements with little or no development.
3. **Cohesive Devices — Are linking words used accurately and variedly?**
   - **High:** Wide range of cohesive devices used accurately and naturally.
   - **Medium:** Limited repertoire; some repetition or minor misuse.
   - **Low:** Few, incorrect, or overused linkers; connections unclear.
4. **Ability to Produce Extended Discourse — Can the speaker sustain longer turns?**
   - **High:** Maintains extended stretches of speech comfortably.
   - **Medium:** Produces connected sentences but turns remain brief.
   - **Low:** Restricted to very short or incomplete utterances; discourse breaks down.

A.2 Direct Scoring Prompt

In this section, we provide the full text of the *Direct Scoring Prompt* used in our experiments. The entire prompt effectively acts as a single question with six options: "A", "B", "C", "D", "E", and "F".

Direct Scoring Prompt

You are an examiner grading a spoken English test for non-native (L2) speakers.

You will be given an audio file of a candidate's speech.
Using ONLY the rubric below, evaluate the performance and assign a score for each criterion.
The scores range from A to F, with A being the highest and F the lowest.

Rubric:

PRONUNCIATION AND FLUENCY
A – Pronunciation is highly intelligible; stress, rhythm, intonation and connected speech are used effectively to express meaning. Flow of speech is effortless with only natural hesitation and pauses.
B – Pronunciation is intelligible; stress, rhythm, intonation and connected speech are used to express meaning well. Flow of speech is generally effortless with mostly natural hesitation and pauses.
C – Pronunciation is generally intelligible but L1 features may occasionally interfere; stress, rhythm and intonation are used to express meaning adequately. Some hesitation may be present while searching for language.
D – Pronunciation can generally be understood but L1 features may cause strain; attempts to

use stress, rhythm and intonation to express meaning are not always successful. Flow of speech is uneven, with some signs of false starts, self-correction, repetition and/or unnatural hesitation.
E – Pronunciation of single words and phrases may be intelligible but L1 features may make understanding difficult; attempts to use stress, rhythm and intonation to express meaning are unsuccessful. Utterances are short, with frequent hesitations and pauses.
F – Pronunciation of individual words may be intelligible but L1 features may cause excessive strain to a listener; little attempt is made to use aspects of stress, rhythm or intonation to express meaning. Utterances are limited to single words or phrases, with excessive hesitations and pauses making speech difficult to follow.

LANGUAGE RESOURCE
A – Displays full control of complex language, including a wide range of vocabulary (e.g. idiomatic expressions and collocations) and sophisticated syntactic structures. Lexical and/or grammatical errors, if present, are not noticeable.
B – Displays good control of complex language, including a range of vocabulary (e.g. attempts to use idiomatic expressions and collocations) and sophisticated syntactic structures. Lexical and/or grammatical errors, if present, are not intrusive.
C – There is an adequate range of grammar and vocabulary which is sufficiently accurate. Lexical and/or grammatical errors are present but generally do not impede meaning.
D – The range of grammar and vocabulary used is limited. Utterances using simple language are accurate but basic inaccuracies when attempting to use more complex language may impede communication of ideas.
E – The range of language is insufficient. Some utterances (e.g. single words or short phrases) may be accurate but inaccuracies in grammar and vocabulary restrict communication of ideas.
F – The range of language is very limited. Some accurate language (e.g. pre-packaged utterances) may occur but frequent inaccuracies mean the message is not communicated.

DISCOURSE MANAGEMENT
A – The logic behind the message is immediately apparent. There is a clear progression within the development of ideas.
B – The logic behind the message is easy to follow. There is a coherent progression within the development of ideas.
C – The logic behind the message is comprehensible but may require effort to identify. The relationship between ideas is generally clear.
D – There may be some relationship between ideas, but they appear generally disconnected.
E – Not applicable due to lack of extended discourse. (utterances are limited to short and incomplete sentences)
F – Not applicable due to lack of extended discourse. (utterances are limited to isolated words and memorised phrases)

INSTRUCTIONS
1. Evaluate the candidate's speech using the three criteria above.
2. Output a single overall grade based on the candidate's performance across all criteria.
3. Do not output anything else except a single letter grade for the overall performance.

Overall grade:

A.3 Rubric-based Question Set

This section lists all the questions in the *Rubric-based Question Set*. There are three questions corresponding to each rubric dimension: Pronunciation and Fluency, Language Resource, and Discourse Management. Each question has six response options: "A", "B", "C", "D", "E", and "F".

Pronunciation and Fluency

1. **Assign a mark (A–F) for the candidate's pronunciation and fluency.**

    - **A:** Pronunciation is highly intelligible; stress, rhythm, intonation and connected speech are used effectively to express meaning. Flow of speech is effortless with only natural hesitation and pauses.
    - **B:** Pronunciation is intelligible; stress, rhythm, intonation and connected speech are used to express meaning well. Flow of speech is generally effortless with mostly natural hesitation and pauses.
    - **C:** Pronunciation is generally intelligible but L1 features may occasionally interfere; stress, rhythm and intonation are used to express meaning adequately. Some hesitation may be present while searching for language.
    - **D:** Pronunciation can generally be understood but L1 features may cause strain; attempts to use stress, rhythm and intonation to express meaning are not always successful. Flow of speech is uneven, with some signs of false starts, self-correction, repetition and/or unnatural hesitation.
    - **E:** Pronunciation of single words and phrases may be intelligible but L1 features may make understanding difficult; attempts to use stress, rhythm and intonation to express meaning are unsuccessful. Utterances are short, with frequent hesitations and pauses.
    - **F:** Pronunciation of individual words may be intelligible but L1 features may cause excessive strain to a listener; little attempt is made to use aspects of stress, rhythm or intonation to express meaning. Utterances are limited to single words or phrases, with excessive hesitations and pauses making speech difficult to follow.

Language Resource

1. **Assign a mark (A–F) for the candidate's control of grammar and vocabulary.**

    - **A:** Displays full control of complex language, including a wide range of vocabulary (e.g. idiomatic expressions and collocations) and sophisticated syntactic structures. Lexical and/or grammatical errors, if present, are not noticeable.

- **B:** Displays good control of complex language, including a range of vocabulary (e.g. attempts to use idiomatic expressions and collocations) and sophisticated syntactic structures. Lexical and/or grammatical errors, if present, are not intrusive.
- **C:** There is an adequate range of grammar and vocabulary which is sufficiently accurate. Lexical and/or grammatical errors are present but generally do not impede meaning.
- **D:** The range of grammar and vocabulary used is limited. Utterances using simple language are accurate but basic inaccuracies when attempting to use more complex language may impede communication of ideas.
- **E:** The range of language is insufficient. Some utterances (e.g. single words or short phrases) may be accurate but inaccuracies in grammar and vocabulary restrict communication of ideas.
- **F:** The range of language is very limited. Some accurate language (e.g. pre-packaged utterances) may occur but frequent inaccuracies mean the message is not communicated.

Discourse Management

1. **Assign a mark (A–F) for the candidate's discourse management (organisation and coherence).**

    - **A:** The logic behind the message is immediately apparent. There is a clear progression within the development of ideas.
    - **B:** The logic behind the message is easy to follow. There is a coherent progression within the development of ideas.
    - **C:** The logic behind the message is comprehensible but may require effort to identify. The relationship between ideas is generally clear.
    - **D:** There may be some relationship between ideas, but they appear generally disconnected.
    - **E:** Not applicable due to lack of extended discourse. Utterances are limited to short and incomplete sentences.
    - **F:** Not applicable due to lack of extended discourse. Utterances are limited to isolated words and memorised phrases.

A.4 Rubric-based Question Set Batch Prompt

In this section, we provide the full text of the *Rubric-based Question Set Batch Prompt* used in our experiments. This prompt is used to extract features for all 18 questions in the *Rubric-based Question Set* in a single speech LLM call.

Rubric-based Question Set Batch Prompt

You are an examiner grading a spoken English test for non-native (L2) speakers.

You will be given an audio file of a candidate's speech.
Using ONLY the rubric below, evaluate the performance and assign a score for each criterion.
The scores range from A to F, with A being the highest and F the lowest.

Rubric:

PRONUNCIATION AND FLUENCY
A – Pronunciation is highly intelligible; stress, rhythm, intonation and connected speech are used effectively to express meaning. Flow of speech is effortless with only natural hesitation and pauses.
B – Pronunciation is intelligible; stress, rhythm, intonation and connected speech are used to express meaning well. Flow of speech is generally effortless with mostly natural hesitation and pauses.
C – Pronunciation is generally intelligible but L1 features may occasionally interfere; stress, rhythm and intonation are used to express meaning adequately. Some hesitation may be present while searching for language.
D – Pronunciation can generally be understood but L1 features may cause strain; attempts to use stress, rhythm and intonation to express meaning are not always successful. Flow of speech is uneven, with some signs of false starts, self-correction, repetition and/or unnatural hesitation.
E – Pronunciation of single words and phrases may be intelligible but L1 features may make understanding difficult; attempts to use stress, rhythm and intonation to express meaning are unsuccessful. Utterances are short, with frequent hesitations and pauses.
F – Pronunciation of individual words may be intelligible but L1 features may cause excessive strain to a listener; little attempt is made to use aspects of stress, rhythm or intonation to express meaning. Utterances are limited to single words or phrases, with excessive hesitations and pauses making speech difficult to follow.

LANGUAGE RESOURCE
A – Displays full control of complex language, including a wide range of vocabulary (e.g. idiomatic expressions and collocations) and sophisticated syntactic structures. Lexical and/or grammatical errors, if present, are not noticeable.
B – Displays good control of complex language, including a range of vocabulary (e.g. attempts to use idiomatic expressions and collocations) and sophisticated syntactic structures. Lexical and/or grammatical errors, if present, are not intrusive.
C – There is an adequate range of grammar and vocabulary which is sufficiently accurate. Lexical and/or grammatical errors are present but generally do not impede meaning.
D – The range of grammar and vocabulary used is limited. Utterances using simple language are accurate but basic inaccuracies when attempting to use more complex language may impede communication of ideas.
E – The range of language is insufficient. Some utterances (e.g. single words or short phrases) may be accurate but inaccuracies in grammar and vocabulary restrict communication of ideas.
F – The range of language is very limited. Some accurate language (e.g. pre-packaged utterances)

may occur but frequent inaccuracies mean the message is not communicated.

DISCOURSE MANAGEMENT
A – The logic behind the message is immediately apparent. There is a clear progression within the development of ideas.
B – The logic behind the message is easy to follow. There is a coherent progression within the development of ideas.
C – The logic behind the message is comprehensible but may require effort to identify. The relationship between ideas is generally clear.
D – There may be some relationship between ideas, but they appear generally disconnected.
E – Not applicable due to lack of extended discourse. (utterances are limited to short and incomplete sentences)
F – Not applicable due to lack of extended discourse. (utterances are limited to isolated words and memorised phrases)

INSTRUCTIONS
1. Evaluate the candidate's speech using the three criteria above.
2. Assign a single-letter mark for each criterion (A–F).
3. Do not output anything else except the marks for each criterion in the specified format.

OUTPUT FORMAT (only):

```
{
  "Pronunciation_Fluency": {
    "mark": "<letter>"
  },
  "Language_Resource": {
    "mark": "<letter>"
  },
  "Discourse_Management": {
    "mark": "<letter>"
  }
}
```

A.5 Revised Question Set

This section lists all the questions in the *Revised Question Set*. The questions are divided into four categories: Pronunciation and Fluency, Language Resource, Discourse Management, and Audio Quality. Each question has two response options: "High" and "Low".

Pronunciation and Fluency

1. **Intelligibility — Is the speech easy to understand?**
   - **High:** Speech is consistently clear and easy to follow.
   - **Low:** Speech is frequently unclear; the listener struggles to understand.

2. **Flow & Pausing — Does the speech move smoothly or is it disrupted by hesitations, fillers, or restarts?**
    - **High:** Continuous flow with only natural pauses; fillers and restarts are negligible.
    - **Low:** Frequent long pauses, fillers, or restarts break comprehension.
3. **Stress & Rhythm — Are English stress patterns and rhythm used correctly to aid meaning?**
    - **High:** Accurate word and sentence stress with natural rhythm that supports meaning.
    - **Low:** Misplaced or missing stress and uneven rhythm obscure meaning.
4. **Intonation — Does pitch vary appropriately to convey meaning and attitude?**
    - **High:** Varied, natural intonation enhances the message and stance.
    - **Low:** Flat, erratic, or unnatural pitch confuses meaning.
5. **Word Linking — Does the speaker connect words smoothly, so speech sounds joined rather than word-by-word?**
    - **High:** Words are smoothly linked; speech sounds connected and natural.
    - **Low:** Words are produced separately; lack of linking makes speech sound choppy.
6. **L1 Phonological Influence — Do sounds from the first language interfere with English pronunciation?**
    - **High:** L1 influence is minimal and never hinders understanding.
    - **Low:** Strong L1 interference regularly causes misunderstanding.
7. **Automaticity of Retrieval — Does the speaker pause to search for basic words or forms?**
    - **High:** Words and structures are retrieved instantly; no searching pauses.
    - **Low:** Frequent pauses or reformulations occur while recalling basic language.
8. **Disfluencies & Self-Correction — How often do false starts or mid-sentence corrections occur?**
    - **High:** Disfluencies are rare; self-corrections are subtle and non-disruptive.
    - **Low:** Numerous false starts or overt corrections disrupt speech flow.

Language Resource

1. **Vocabulary Range — How varied is the speaker's word choice?**
    - **High:** Broad, topic-specific vocabulary; idioms and collocations used flexibly.
    - **Low:** Basic, repetitive vocabulary; lexical gaps limit expression.

2. **Context-Appropriate Vocabulary — Are words suitable for the situation (formal or informal)?**
    - **High:** Words are precise and fit the situation naturally.
    - **Low:** Frequent imprecise or unsuitable word choices blur meaning.
3. **Grammatical Range & Complexity — Are complex structures attempted successfully?**
    - **High:** Frequent, accurate use of subordinate clauses and varied sentence types.
    - **Low:** Mostly simple clauses; attempts at complexity are unsuccessful or avoided.
4. **Grammatical Accuracy — How free from errors is the grammar?**
    - **High:** Only minor slips occur; errors never impede meaning.
    - **Low:** Frequent errors confuse or obscure meaning.
5. **Idiomaticity & Collocation — Are idioms and collocations used naturally?**
    - **High:** Idiomatic expressions and collocations are used appropriately and accurately.
    - **Low:** Idioms are rare or incorrect; collocation errors distract or confuse.
6. **Lexical Flexibility — Can the speaker paraphrase effectively when lacking a word?**
    - **High:** Paraphrases or circumlocutes skillfully without loss of meaning.
    - **Low:** Unable to paraphrase; communication breaks when a word is missing.
7. **Word Formation & Derivation — Are derived forms (e.g., adjectives from nouns) used correctly?**
    - **High:** Derived words are formed and used accurately, improving precision.
    - **Low:** Incorrect or limited word formation leads to ambiguity.

Discourse Management

1. **Overall Logic & Coherence — Is the main line of thought easy to follow?**
    - **High:** Argument or thread is immediately clear and logical throughout.
    - **Low:** Ideas seem random or disjointed; difficult to follow.
2. **Progression & Development — Are ideas expanded with explanations, details, or examples?**
    - **High:** Points are fully developed with supporting detail and clear links.
    - **Low:** Ideas are presented as isolated statements with little development.
3. **Cohesive Devices Accuracy — Are linking words (e.g., 'however', 'because') used correctly?**

- **High:** Linking words are used accurately to show logical relationships.
- **Low:** Linking words are often incorrect or missing, confusing relationships.

4. **Cohesive Devices Variety — Does the speaker use a range of linking words and phrases?**
   - **High:** A wide variety of cohesive devices is used naturally.
   - **Low:** Little variety; the same device is overused or devices are absent.
5. **Topic Maintenance — Does the speaker stay on the same topic throughout the response?**
   - **High:** Content remains consistently on topic.
   - **Low:** Frequent digressions or off-topic remarks confuse the main point.
6. **Organizational Signals — Does the speaker use signposting language (e.g., 'first', 'on the other hand') to guide the listener?**
   - **High:** Clear signposting helps the listener navigate through ideas.
   - **Low:** Absence or misuse of signposting leaves structure inferred.
7. **Economy & Redundancy — Is information conveyed concisely without unnecessary repetition?**
   - **High:** Delivery is concise with minimal redundancy.
   - **Low:** Frequent repetition or tangential content clouds the message.

Audio Quality

1. **Microphone Clarity — Is the recording free from muffling or distortion?**
   - **High:** Microphone captures speech clearly without noticeable distortion.
   - **Low:** Muffling, crackling, or distortion makes speech hard to analyse.
2. **Volume Level — Is the speaker loud enough to hear comfortably?**
   - **High:** Volume is comfortably audible throughout the recording.
   - **Low:** Volume is too low or varies so much that parts are hard to hear.
3. **Background Noise — Is there distracting noise in the environment?**
   - **High:** Background noise is minimal and does not interfere with speech.
   - **Low:** Noticeable noise (traffic, chatter, hum) competes with the speaker.
4. **Initial Lag — Is there a long silence or delay before the speaker starts talking?**
   - **High:** Speech starts promptly with only brief natural silence.
   - **Low:** Long silence or delay precedes speech, wasting processing time.

5. **Additional Speakers — Are other voices audible that might confuse transcription?**
   - **High:** No other speakers are audible or they are clearly separate.
   - **Low:** Other voices overlap or interrupt, risking transcription errors.
6. **Clipping or Distortion — Does the audio peak or crackle due to high input levels?**
   - **High:** No audible clipping; audio levels are well-balanced.
   - **Low:** Clipping or crackle distorts words and hampers analysis.
7. **Echo or Reverb — Is there echo that blurs the speech signal?**
   - **High:** Room acoustics are dry; no echo interferes with speech.
   - **Low:** Echo or strong reverb makes words blend together.
8. **Recording Completeness — Is any part of the speech cut off at the start or end?**
   - **High:** Recording captures the entire utterance without cuts.
   - **Low:** Speech is truncated, missing initial or final words.

A.6 Revised Question Set Batch Prompt

In this section, we provide the full text of the *Revised Question Set Batch Prompt* used in our experiments. This prompt is used to extract features for all 30 questions in the *Revised Question Set* in a single speech LLM call.

Revised Question Set Batch Prompt

You are an examiner grading a spoken English test for non-native (L2) speakers.

Given an audio file and evaluation questions, assess the candidate's spoken English performance.
Your task is to listen to the audio and answer the questions based on the candidate's speech.

QUESTIONS:

- pf_1. Intelligibility — Is the speech easy to understand?
  High: Speech is consistently clear and easy to follow.
  Low: Speech is frequently unclear; the listener struggles to understand.
- pf_2. Flow & Pausing — Does the speech move smoothly or is it disrupted by hesitations, fillers, or restarts?
  High: Continuous flow with only natural pauses; fillers and restarts are negligible.
  Low: Frequent long pauses, fillers, or restarts break comprehension.
- pf_3. Stress & Rhythm — Are English stress patterns and rhythm used correctly to

aid meaning?
High: Accurate word and sentence stress with natural rhythm that supports meaning.
Low: Misplaced or missing stress and uneven rhythm obscure meaning.

- pf_4. Intonation — Does pitch vary appropriately to convey meaning and attitude?
  High: Varied, natural intonation enhances the message and stance.
  Low: Flat, erratic, or unnatural pitch confuses meaning.

- pf_5. Word Linking — Does the speaker connect words smoothly, so speech sounds joined rather than word-by-word?
  High: Words are smoothly linked; speech sounds connected and natural.
  Low: Words are produced separately; lack of linking makes speech sound choppy.

- pf_6. L1 Phonological Influence — Do sounds from the first language interfere with English pronunciation?
  High: L1 influence is minimal and never hinders understanding.
  Low: Strong L1 interference regularly causes misunderstanding.

- pf_7. Automaticity of Retrieval — Does the speaker pause to search for basic words or forms?
  High: Words and structures are retrieved instantly; no searching pauses.
  Low: Frequent pauses or reformulations occur while recalling basic language.

- pf_8. Disfluencies & Self-Correction — How often do false starts or mid-sentence corrections occur?
  High: Disfluencies are rare; self-corrections are subtle and non-disruptive.
  Low: Numerous false starts or overt corrections disrupt speech flow.

- lr_1. Vocabulary Range — How varied is the speaker's word choice?
  High: Broad, topic-specific vocabulary; idioms and collocations used flexibly.
  Low: Basic, repetitive vocabulary; lexical gaps limit expression.

- lr_2. Context-Appropriate Vocabulary — Are words suitable for the situation (formal or informal)?
  High: Words are precise and fit the situation naturally.
  Low: Frequent imprecise or unsuitable word choices blur meaning.

- lr_3. Grammatical Range & Complexity — Are complex structures attempted successfully?
  High: Frequent, accurate use of subordinate clauses and varied sentence types.
  Low: Mostly simple clauses; attempts at complexity are unsuccessful or avoided.

- lr_4. Grammatical Accuracy — How free from errors is the grammar?
  High: Only minor slips occur; errors never impede meaning.
  Low: Frequent errors confuse or obscure meaning.

- lr_5. Idiomaticity & Collocation — Are idioms and collocations used naturally?
  High: Idiomatic expressions and collocations are used appropriately and accurately.
  Low: Idioms are rare or incorrect; collocation errors distract or confuse.
- lr_6. Lexical Flexibility — Can the speaker paraphrase effectively when lacking a word?
  High: Paraphrases or circumlocutes skillfully without loss of meaning.
  Low: Unable to paraphrase; communication breaks when a word is missing.
- lr_7. Word Formation & Derivation — Are derived forms (e.g., adjectives from nouns) used correctly?
  High: Derived words are formed and used accurately, improving precision.
  Low: Incorrect or limited word formation leads to ambiguity.
- dm_1. Overall Logic & Coherence — Is the main line of thought easy to follow?
  High: Argument or thread is immediately clear and logical throughout.
  Low: Ideas seem random or disjointed; difficult to follow.
- dm_2. Progression & Development — Are ideas expanded with explanations, details, or examples?
  High: Points are fully developed with supporting detail and clear links.
  Low: Ideas are presented as isolated statements with little development.
- dm_3. Cohesive Devices Accuracy — Are linking words (e.g., 'however', 'because') used correctly?
  High: Linking words are used accurately to show logical relationships.
  Low: Linking words are often incorrect or missing, confusing relationships.
- dm_4. Cohesive Devices Variety — Does the speaker use a range of linking words and phrases?
  High: A wide variety of cohesive devices is used naturally.
  Low: Little variety; the same device is overused or devices are absent.
- dm_5. Topic Maintenance — Does the speaker stay on the same topic throughout the response?
  High: Content remains consistently on topic.
  Low: Frequent digressions or off-topic remarks confuse the main point.
- dm_6. Organizational Signals — Does the speaker use signposting language (e.g., 'first', 'on the other hand') to guide the listener?
  High: Clear signposting helps the listener navigate through ideas.
  Low: Absence or misuse of signposting leaves structure inferred.
- dm_7. Economy & Redundancy — Is information conveyed concisely without unnecessary repetition?
  High: Delivery is concise with minimal redundancy.
  Low: Frequent repetition or tangential content clouds the message.

- aq_1. Microphone Clarity — Is the recording free from muffling or distortion?
  High: Microphone captures speech clearly without noticeable distortion.
  Low: Muffling, crackling, or distortion makes speech hard to analyse.
- aq_2. Volume Level — Is the speaker loud enough to hear comfortably?
  High: Volume is comfortably audible throughout the recording.
  Low: Volume is too low or varies so much that parts are hard to hear.
- aq_3. Background Noise — Is there distracting noise in the environment?
  High: Background noise is minimal and does not interfere with speech.
  Low: Noticeable noise (traffic, chatter, hum) competes with the speaker.
- aq_4. Initial Lag — Is there a long silence or delay before the speaker starts talking?
  High: Speech starts promptly with only brief natural silence.
  Low: Long silence or delay precedes speech, wasting processing time.
- aq_5. Additional Speakers — Are other voices audible that might confuse transcription?
  High: No other speakers are audible or they are clearly separate.
  Low: Other voices overlap or interrupt, risking transcription errors.
- aq_6. Clipping or Distortion — Does the audio peak or crackle due to high input levels?
  High: No audible clipping; audio levels are well-balanced.
  Low: Clipping or crackle distorts words and hampers analysis.
- aq_7. Echo or Reverb — Is there echo that blurs the speech signal?
  High: Room acoustics are dry; no echo interferes with speech.
  Low: Echo or strong reverb makes words blend together.
- aq_8. Recording Completeness — Is any part of the speech cut off at the start or end?
  High: Recording captures the entire utterance without cuts.
  Low: Speech is truncated, missing initial or final words.

INSTRUCTIONS:

1. Listen to the audio file carefully.
2. Read each question and its options carefully.
3. For each question, select the option that best describes the candidate's performance.
4. Do not output anything else except for the selected option (either "High" or "Low") for each question in the specified format.

OUTPUT FORMAT (only):

```
{
    "pf_1": "<Option>",
    "pf_2": "<Option>",
    "pf_3": "<Option>",
    "pf_4": "<Option>",
    "pf_5": "<Option>",
    "pf_6": "<Option>",
    "pf_7": "<Option>",
    "pf_8": "<Option>",
    "lr_1": "<Option>",
    "lr_2": "<Option>",
    "lr_3": "<Option>",
    "lr_4": "<Option>",
    "lr_5": "<Option>",
    "lr_6": "<Option>",
    "lr_7": "<Option>",
    "dm_1": "<Option>",
    "dm_2": "<Option>",
    "dm_3": "<Option>",
    "dm_4": "<Option>",
    "dm_5": "<Option>",
    "dm_6": "<Option>",
    "dm_7": "<Option>",
    "aq_1": "<Option>",
    "aq_2": "<Option>",
    "aq_3": "<Option>",
    "aq_4": "<Option>",
    "aq_5": "<Option>",
    "aq_6": "<Option>",
    "aq_7": "<Option>",
    "aq_8": "<Option>"
}
```

Appendix B

Additional Results

In this chapter, we list the results of different question sets for each individual part of the assessment and the overall assessment for both the development and the evaluation subset. For all of these results, the linear models were trained on the train subset and calibration coefficients were computed on the development subset.

B.1 Development Subset

B.1.1 Part 1

Feature Set	**RMSE**	**PCC**	**SRC**	**P@0.5**	**P@1.0**
Initial Question Set	0.5337 ± 0.0181	0.7296 ± 0.0207	0.7277 ± 0.0230	65.14 ± 2.31	93.34 ± 1.23
Direct Scoring Question	0.6366 ± 0.0206	0.5785 ± 0.0297	0.5815 ± 0.0313	54.48 ± 2.41	88.55 ± 1.53
Rubric-based Question Set	0.5428 ± 0.0174	0.7184 ± 0.0213	0.7114 ± 0.0242	61.99 ± 2.41	93.15 ± 1.19
Rubric-based Question Set (Batch Inference)	0.5580 ± 0.0183	0.6989 ± 0.0238	0.6804 ± 0.0275	62.22 ± 2.27	92.63 ± 1.29
Revised Question Set	0.5495 ± 0.0191	0.7099 ± 0.0223	0.7133 ± 0.0236	64.95 ± 2.31	92.44 ± 1.30
Revised Question Set (Batch Inference)	0.5483 ± 0.0181	0.7115 ± 0.0217	0.7110 ± 0.0234	61.26 ± 2.36	94.75 ± 1.07
Combined Question Set	0.5182 ± 0.0174	0.7477 ± 0.0193	0.7506 ± 0.0207	64.48 ± 2.41	94.96 ± 1.04
Random Projections of *Initial Question Set*	0.5337 ± 0.0181	0.7296 ± 0.0207	0.7277 ± 0.0230	65.14 ± 2.31	93.34 ± 1.23
Speech LLM Representations	0.4920 ± 0.0156	0.7761 ± 0.0178	0.7740 ± 0.0187	66.99 ± 2.28	96.12 ± 0.90
Combined Question Set + Speech LLM Representations	0.4940 ± 0.0159	0.7741 ± 0.0180	0.7711 ± 0.0191	68.17 ± 2.26	95.43 ± 0.99

Table B.1: Performance comparison of linear regression models across different feature representations (evaluated on dev set for part 1 only).

B.1.2 Part 3

Feature Set	RMSE	PCC	SRC	P@0.5	P@1.0
Initial Question Set	0.5451 ± 0.0237	0.6641 ± 0.0323	0.6623 ± 0.0316	65.46 ± 2.31	94.33 ± 1.13
Direct Scoring Question	0.5945 ± 0.0237	0.5793 ± 0.0357	0.5865 ± 0.0333	58.51 ± 2.31	92.52 ± 1.26
Rubric-based Question Set	0.5795 ± 0.0233	0.6072 ± 0.0357	0.6041 ± 0.0339	62.04 ± 2.36	92.09 ± 1.28
Rubric-based Question Set (Batch Inference)	0.5842 ± 0.0232	0.5986 ± 0.0354	0.6005 ± 0.0331	61.07 ± 2.33	92.52 ± 1.24
Revised Question Set	0.5661 ± 0.0229	0.6305 ± 0.0317	0.6340 ± 0.0308	64.94 ± 2.30	94.09 ± 1.16
Revised Question Set (Batch Inference)	0.5766 ± 0.0218	0.6124 ± 0.0319	0.6267 ± 0.0319	61.28 ± 2.27	93.66 ± 1.20
Combined Question Set	0.5326 ± 0.0225	0.6829 ± 0.0289	0.6925 ± 0.0281	67.74 ± 2.08	95.47 ± 1.03
Random Projections of *Initial Question Set*	0.5451 ± 0.0237	0.6641 ± 0.0323	0.6623 ± 0.0316	65.46 ± 2.31	94.33 ± 1.13
Speech LLM Representations	0.5115 ± 0.0207	0.7129 ± 0.0259	0.7254 ± 0.0250	68.39 ± 2.14	95.70 ± 0.97
Combined Question Set + Speech LLM Representations	0.5111 ± 0.0212	0.7133 ± 0.0266	0.7236 ± 0.0252	68.44 ± 2.10	95.48 ± 0.98

Table B.2: Performance comparison of linear regression models across different feature representations (evaluated on dev set for part 3 only).

B.1.3 Part 4

Feature Set	RMSE	PCC	SRC	P@0.5	P@1.0
Initial Question Set	0.5518 ± 0.0174	0.6496 ± 0.0252	0.6500 ± 0.0273	63.52 ± 2.27	93.85 ± 1.14
Direct Scoring Question	0.5887 ± 0.0193	0.5847 ± 0.0285	0.5767 ± 0.0311	59.85 ± 2.39	92.24 ± 1.24
Rubric-based Question Set	0.5887 ± 0.0191	0.5849 ± 0.0294	0.5792 ± 0.0308	58.09 ± 2.40	91.74 ± 1.32
Rubric-based Question Set (Batch Inference)	0.5681 ± 0.0191	0.6222 ± 0.0277	0.6172 ± 0.0298	61.05 ± 2.36	93.81 ± 1.13
Revised Question Set	0.5494 ± 0.0181	0.6534 ± 0.0251	0.6512 ± 0.0275	62.07 ± 2.31	93.83 ± 1.13
Revised Question Set (Batch Inference)	0.5810 ± 0.0196	0.5993 ± 0.0279	0.5979 ± 0.0304	60.09 ± 2.33	91.56 ± 1.33
Combined Question Set	0.5406 ± 0.0177	0.6673 ± 0.0247	0.6692 ± 0.0261	63.26 ± 2.29	95.47 ± 0.97
Random Projections of *Initial Question Set*	0.5518 ± 0.0174	0.6496 ± 0.0252	0.6500 ± 0.0273	63.52 ± 2.27	93.85 ± 1.14
Speech LLM Representations	0.5262 ± 0.0189	0.6887 ± 0.0241	0.6935 ± 0.0255	65.34 ± 2.36	94.49 ± 1.11
Combined Question Set + Speech LLM Representations	0.5265 ± 0.0189	0.6884 ± 0.0242	0.6918 ± 0.0255	65.74 ± 2.35	94.28 ± 1.12

Table B.3: Performance comparison of linear regression models across different feature representations (evaluated on dev set for part 4 only).

B.1.4 Part 5

Feature Set	RMSE	PCC	SRC	P@0.5	P@1.0
Initial Question Set	0.4811 ± 0.0167	0.7649 ± 0.0181	0.7725 ± 0.0195	70.68 ± 2.21	94.75 ± 1.08
Direct Scoring Question	0.5824 ± 0.0189	0.6260 ± 0.0273	0.6287 ± 0.0294	62.37 ± 2.35	90.86 ± 1.36
Rubric-based Question Set	0.5070 ± 0.0177	0.7343 ± 0.0208	0.7467 ± 0.0224	68.32 ± 2.23	93.87 ± 1.17
Rubric-based Question Set (Batch Inference)	0.5442 ± 0.0191	0.6848 ± 0.0246	0.6951 ± 0.0250	67.20 ± 2.18	91.33 ± 1.33
Revised Question Set	0.5251 ± 0.0201	0.7110 ± 0.0240	0.7244 ± 0.0239	68.31 ± 2.20	93.19 ± 1.18
Revised Question Set (Batch Inference)	0.5032 ± 0.0190	0.7388 ± 0.0217	0.7437 ± 0.0216	70.41 ± 2.19	95.01 ± 1.03
Combined Question Set	0.4668 ± 0.0167	0.7806 ± 0.0176	0.7868 ± 0.0179	72.93 ± 2.10	95.90 ± 0.95
Random Projections of *Initial Question Set*	0.4811 ± 0.0167	0.7649 ± 0.0181	0.7725 ± 0.0195	70.68 ± 2.21	94.75 ± 1.08
Speech LLM Representations	0.4581 ± 0.0164	0.7897 ± 0.0166	0.7982 ± 0.0166	73.57 ± 2.14	96.37 ± 0.90
Combined Question Set + Speech LLM Representations	0.4580 ± 0.0164	0.7898 ± 0.0166	0.7963 ± 0.0169	73.38 ± 2.14	96.37 ± 0.90

Table B.4: Performance comparison of linear regression models across different feature representations (evaluated on dev set for part 5 only).

B.1.5 Overall

Note this is the same as the data reported in Table 5.3.

Feature Set	RMSE	PCC	SRC	P@0.5	P@1.0
Initial Question Set	0.3945 ± 0.0150	0.8052 ± 0.0153	0.8150 ± 0.0165	81.76 ± 1.82	97.46 ± 0.74
Direct Scoring Question	0.4454 ± 0.0171	0.7427 ± 0.0206	0.7484 ± 0.0220	75.35 ± 2.14	96.76 ± 0.84
Rubric-based Question Set	0.4203 ± 0.0151	0.7751 ± 0.0170	0.7847 ± 0.0186	79.46 ± 1.92	97.26 ± 0.78
Rubric-based Question Set (Batch Inference)	0.4247 ± 0.0165	0.7696 ± 0.0185	0.7717 ± 0.0204	78.59 ± 2.02	96.55 ± 0.87
Revised Question Set	0.4163 ± 0.0168	0.7799 ± 0.0183	0.7905 ± 0.0184	78.94 ± 1.95	97.48 ± 0.73
Revised Question Set (Batch Inference)	0.4188 ± 0.0156	0.7769 ± 0.0176	0.7868 ± 0.0186	79.95 ± 1.87	97.24 ± 0.77
Combined Question Set	0.3872 ± 0.0153	0.8132 ± 0.0151	0.8245 ± 0.0153	82.87 ± 1.83	97.93 ± 0.67
Random Projections of *Initial Question Set*	0.3945 ± 0.0150	0.8052 ± 0.0153	0.8150 ± 0.0165	81.76 ± 1.82	97.46 ± 0.74
Speech LLM Representations	**0.3727 ± 0.0146**	**0.8282 ± 0.0144**	**0.8397 ± 0.0144**	85.84 ± 1.73	**98.17 ± 0.65**
Combined Question Set + Speech LLM Representations	**0.3727 ± 0.0147**	**0.8282 ± 0.0146**	0.8378 ± 0.0148	**86.28 ± 1.68**	**98.17 ± 0.65**

Table B.5: Performance comparison of linear regression models across different feature representations (evaluated on dev set for overall scores).

B.2 Evaluation Subset

Note that performance on the evaluation subset is noticeably lower than on the development subset. While this difference partially stems from calibration coefficients being computed on the development subset, the evaluation set appears to be inherently more challenging than the development set. This is evidenced by the observable performance gap between the two sets even before calibration is applied (analysis not shown here for brevity).

B.2.1 Part 1

Feature Set	RMSE	PCC	SRC	P@0.5	P@1.0
Initial Question Set	0.5596 ± 0.0236	0.7055 ± 0.0268	0.7120 ± 0.0290	64.12 ± 2.74	92.39 ± 1.56
Direct Scoring Question	0.6564 ± 0.0285	0.5383 ± 0.0387	0.5339 ± 0.0417	53.10 ± 2.89	87.67 ± 1.90
Rubric-based Question Set	0.6028 ± 0.0259	0.6577 ± 0.0303	0.6433 ± 0.0341	61.52 ± 2.78	89.36 ± 1.76
Rubric-based Question Set (Batch Inference)	0.6340 ± 0.0256	0.6292 ± 0.0350	0.6177 ± 0.0385	58.66 ± 2.90	87.02 ± 1.92
Revised Question Set	0.5660 ± 0.0255	0.6832 ± 0.0299	0.6810 ± 0.0333	63.30 ± 2.82	92.31 ± 1.57
Revised Question Set (Batch Inference)	0.5876 ± 0.0245	0.6591 ± 0.0300	0.6593 ± 0.0328	61.00 ± 2.77	91.37 ± 1.59
Combined Question Set	0.5126 ± 0.0229	0.7433 ± 0.0250	0.7488 ± 0.0265	69.36 ± 2.70	94.03 ± 1.39
Random Projections of *Initial Question Set*	0.5596 ± 0.0236	0.7055 ± 0.0268	0.7120 ± 0.0290	64.12 ± 2.74	92.39 ± 1.56
Speech LLM Representations	0.4862 ± 0.0224	0.7705 ± 0.0237	0.7775 ± 0.0246	72.34 ± 2.62	95.67 ± 1.19
Combined Question Set + Speech LLM Representations	0.4831 ± 0.0228	0.7734 ± 0.0239	0.7801 ± 0.0250	73.01 ± 2.54	95.67 ± 1.21

Table B.6: Performance comparison of linear regression models across different feature representations (evaluated on eval set for part 1 only).

B.2.2 Part 3

Feature Set	RMSE	PCC	SRC	P@0.5	P@1.0
Initial Question Set	0.4887 ± 0.0187	0.6916 ± 0.0275	0.6946 ± 0.0316	68.24 ± 2.66	95.36 ± 1.22
Direct Scoring Question	0.5521 ± 0.0207	0.5790 ± 0.0375	0.5832 ± 0.0403	57.54 ± 2.94	93.97 ± 1.34
Rubric-based Question Set	0.5533 ± 0.0202	0.5744 ± 0.0360	0.5572 ± 0.0410	63.29 ± 2.86	93.98 ± 1.38
Rubric-based Question Set (Batch Inference)	0.5417 ± 0.0202	0.5972 ± 0.0332	0.6162 ± 0.0359	63.26 ± 2.92	94.29 ± 1.31
Revised Question Set	0.5320 ± 0.0207	0.6166 ± 0.0323	0.6168 ± 0.0357	65.75 ± 2.79	93.63 ± 1.40
Revised Question Set (Batch Inference)	0.5463 ± 0.0237	0.5906 ± 0.0416	0.6048 ± 0.0377	63.02 ± 2.81	92.63 ± 1.51
Combined Question Set	0.5041 ± 0.0181	0.6656 ± 0.0286	0.6649 ± 0.0324	64.57 ± 2.81	96.34 ± 1.10
Random Projections of *Initial Question Set*	0.4887 ± 0.0187	0.6916 ± 0.0275	0.6946 ± 0.0316	68.24 ± 2.66	95.36 ± 1.22
Speech LLM Representations	0.4822 ± 0.0192	0.6997 ± 0.0268	0.7015 ± 0.0300	66.57 ± 2.67	96.64 ± 1.02
Combined Question Set + Speech LLM Representations	0.4853 ± 0.0190	0.6951 ± 0.0266	0.7001 ± 0.0295	65.89 ± 2.73	96.64 ± 1.03

Table B.7: Performance comparison of linear regression models across different feature representations (evaluated on eval set for part 3 only).

B.2.3 Part 4

Feature Set	RMSE	PCC	SRC	P@0.5	P@1.0
Initial Question Set	0.5667 ± 0.0232	0.6128 ± 0.0327	0.6004 ± 0.0369	59.03 ± 2.84	92.63 ± 1.50
Direct Scoring Question	0.6049 ± 0.0252	0.5362 ± 0.0443	0.5274 ± 0.0449	60.19 ± 2.86	89.63 ± 1.75
Rubric-based Question Set	0.6163 ± 0.0238	0.5137 ± 0.0442	0.4849 ± 0.0490	56.93 ± 2.90	88.95 ± 1.80
Rubric-based Question Set (Batch Inference)	0.5911 ± 0.0229	0.5666 ± 0.0388	0.5519 ± 0.0410	58.21 ± 2.93	92.30 ± 1.51
Revised Question Set	0.5836 ± 0.0218	0.5822 ± 0.0361	0.5574 ± 0.0406	58.91 ± 2.84	93.35 ± 1.45
Revised Question Set (Batch Inference)	0.6264 ± 0.0236	0.4972 ± 0.0414	0.4815 ± 0.0433	56.80 ± 2.85	89.93 ± 1.70
Combined Question Set	0.5643 ± 0.0223	0.6172 ± 0.0321	0.6082 ± 0.0371	62.29 ± 2.80	92.65 ± 1.52
Random Projections of *Initial Question Set*	0.5667 ± 0.0232	0.6128 ± 0.0327	0.6004 ± 0.0369	59.03 ± 2.84	92.63 ± 1.50
Speech LLM Representations	0.5344 ± 0.0215	0.6670 ± 0.0313	0.6610 ± 0.0354	66.50 ± 2.80	93.27 ± 1.47
Combined Question Set + Speech LLM Representations	0.5334 ± 0.0212	0.6684 ± 0.0306	0.6632 ± 0.0347	63.14 ± 2.80	92.94 ± 1.49

Table B.8: Performance comparison of linear regression models across different feature representations (evaluated on eval set for part 4 only).

B.2.4 Part 5

Feature Set	RMSE	PCC	SRC	P@0.5	P@1.0
Initial Question Set	0.5279 ± 0.0235	0.6908 ± 0.0305	0.6819 ± 0.0344	67.63 ± 2.78	93.64 ± 1.37
Direct Scoring Question	0.5938 ± 0.0274	0.5770 ± 0.0422	0.5657 ± 0.0411	58.29 ± 2.73	92.32 ± 1.55
Rubric-based Question Set	0.5539 ± 0.0253	0.6500 ± 0.0342	0.6392 ± 0.0382	65.31 ± 2.74	93.71 ± 1.42
Rubric-based Question Set (Batch Inference)	0.5606 ± 0.0291	0.6371 ± 0.0420	0.6151 ± 0.0402	65.01 ± 2.76	94.30 ± 1.34
Revised Question Set	0.5517 ± 0.0250	0.6521 ± 0.0347	0.6404 ± 0.0370	64.97 ± 2.69	93.66 ± 1.35
Revised Question Set (Batch Inference)	0.5581 ± 0.0255	0.6482 ± 0.0332	0.6524 ± 0.0348	63.61 ± 2.78	93.35 ± 1.41
Combined Question Set	0.5126 ± 0.0232	0.7127 ± 0.0295	0.7085 ± 0.0314	70.27 ± 2.59	95.02 ± 1.25
Random Projections of *Initial Question Set*	0.5279 ± 0.0235	0.6908 ± 0.0305	0.6819 ± 0.0344	67.63 ± 2.78	93.64 ± 1.37
Speech LLM Representations	0.4927 ± 0.0234	0.7420 ± 0.0278	0.7366 ± 0.0296	74.85 ± 2.58	94.24 ± 1.38
Combined Question Set + Speech LLM Representations	0.4887 ± 0.0232	0.7457 ± 0.0274	0.7409 ± 0.0292	74.55 ± 2.56	94.24 ± 1.38

Table B.9: Performance comparison of linear regression models across different feature representations (evaluated on eval set for part 5 only).

B.2.5 Overall

Feature Set	RMSE	PCC	SRC	P@0.5	P@1.0
Initial Question Set	0.4037 ± 0.0156	0.7773 ± 0.0186	0.7828 ± 0.0224	77.88 ± 2.30	98.96 ± 0.57
Direct Scoring Question	0.4536 ± 0.0187	0.7086 ± 0.0274	0.7093 ± 0.0303	70.60 ± 2.62	98.01 ± 0.79
Rubric-based Question Set	0.4381 ± 0.0170	0.7300 ± 0.0245	0.7224 ± 0.0295	75.21 ± 2.45	97.66 ± 0.87
Rubric-based Question Set (Batch Inference)	0.4257 ± 0.0175	0.7498 ± 0.0246	0.7418 ± 0.0283	75.30 ± 2.42	98.33 ± 0.69
Revised Question Set	0.4237 ± 0.0165	0.7503 ± 0.0214	0.7464 ± 0.0263	75.83 ± 2.41	98.61 ± 0.66
Revised Question Set (Batch Inference)	0.4461 ± 0.0174	0.7203 ± 0.0242	0.7206 ± 0.0283	73.79 ± 2.45	97.98 ± 0.79
Combined Question Set	0.3989 ± 0.0158	0.7832 ± 0.0186	0.7870 ± 0.0219	78.77 ± 2.31	98.98 ± 0.57
Random Projections of *Initial Question Set*	0.4037 ± 0.0156	0.7773 ± 0.0186	0.7828 ± 0.0224	77.88 ± 2.30	98.96 ± 0.57
Speech LLM Representations	**0.3845 ± 0.0151**	**0.8019 ± 0.0184**	**0.8007 ± 0.0227**	80.55 ± 2.20	**99.31 ± 0.48**
Combined Question Set + Speech LLM Representations	**0.3833 ± 0.0150**	**0.8028 ± 0.0182**	**0.8032 ± 0.0221**	**80.54 ± 2.22**	**99.31 ± 0.48**

Table B.10: Performance comparison of linear regression models across different feature representations (evaluated on eval set for overall scores).