Josai International University Bulletin Vol. 33, No. 7, March 2025 37-59

〈Research Notes〉

An Exploratory Investigation into the Effectiveness of Fine-tuning GPT Models for Specialized Use in English Language Oral Assessments

Dunstan Henderson

Abstract

In the final quarter of 2022, OpenAI released its first generative pre-trained transformer large language model (LLM), featuring fine-tuning capabilities, a developer's suite, a playground environment, and an updated user interface for non-developers in the form of ChatGPT 3.0. This launch sparked widespread public interest, granting greater access to researchers and developers who applied the model across various sectors, including healthcare services and business analysis applications (Marquis et al., 2024). However, despite the increased interest, limited research has examined GPT 3's potential benefits in educational settings, particularly in language learning and assessment. This study explores the alignment of scores of a fine-tuned GPT model for grading oral English assessments with that of the true values scores of a human grader in university-level English as a Second Language (ESL) courses using few-shot learning. Two models were trained and compared against a human scoring standard: one specialized model and another designed for general multi-test applications. Both models were fine-tuned using a limited size training data set and language processing (NLP) techniques, employing a specifically designed grading matrix and output format. Following training, actual test data from students was input into the models, and the resulting grades were recorded in an Excel file. The models were evaluated using three analytic metrics for artificial intelligence models: Spearman's rank correlation (SRC), p-value, and mean absolute error (MAE)—metrics particularly suited for small training datasets accommodating the study's limitation of small datasets. The results of the specialized model produced an SRC of 0.706, and p-value of 0.022, indicating that the specialized model demonstrated a relationship that was moderately statistically significant in its predictions, with a reasonable degree of alignment with human scores indicated by the MAE value of 0.617. On the contrary, the generalized model failed to achieve statistical significance or alignment. However, it must be noted that the investigation was hampered by the very limited training data size that resulted in frequent misalignment and overfitting, leading to uncertainty in meaningful results.

Keywords: Artificial intelligence, GPT, oral assessment, NLP, prompt design

1. Introduction

OpenAI's ChatGPT model was the company's first publicly available AI chatbot, consolidating GPT models one through three. ChatGPT introduced the world to the capabilities of AI technology, and its accessibility to the general public led to a growing pool of practical applications. These innovations have trickled down to the education sector, with applications ranging from analyzing student academic writing to assisting students with language pronunciation (Roumeliotis & Tselikas, 2023).

Before GPT, early studies on AI-powered machines in the 1990s focused on datasets like the Switchboard-1 Telephone Speech Corpus, created by Godfrey and Holliman in 1993. This dataset included over 2,000 recorded phone conversations on 70 different topics intended to aid the development of Automatic Speech Recognition (ASR) systems by capturing human conversation dynamics and speech variability (Graff & Bird, 2000). The data set represented the largest attempt to gather data for ASR researchers.

As a result, ASR software began to emerge, such as Dragon Naturally Speaking, which debuted in the late 1990s (McCrocklin & Edalatishams, 2020). This software, primarily a language model transcription tool, was used by researchers and educators. Building on this technology, subsequent studies explored its application in second language classrooms, particularly for pronunciation development. ASR-powered applications scored learners' pronunciation styles, with researchers reviewing transcription accuracy. While these studies noted improvements in student pronunciation, they also revealed substantial drawbacks, including low adoption in classrooms due to ASR's mediocre reliability and transcription accuracy (McCrocklin et al., 2018). This highlighted the challenge of insufficient data for model improvement.

Three decades later, the feasibility of AI-based applications in language education regained traction. Studies began evaluating AI's potential, particularly in oral English education. For example, Zhou (2019) examined English language students' use of AI-powered apps like LAIX and Casually Speak to prepare for oral assessments. However, a survey of 200 students showed a steep decline in app usage—from 20% after one month to an additional 8% drop within a year. The study concluded that inadequate scoring mechanisms, limited spoken corpora, and unsupervised usage discouraged sustained engagement and raised concerns about students bypassing responsible-use guidelines, leading to increased cheating.

To address these issues, AI applications shifted focus from assisting students with oral skills to supporting educators with written assessments. One early example was the Duolingo English Test (DET), which emerged in 2016 as an evolution of Duolingo's 2012 language-learning platform (Settles et al., 2020). The DET used customized NLP techniques, integrating item response theory with multiple-choice assessment creation. This approach enabled a binary scoring system that minimized ambiguity in student responses and improved reliability. In practice, DET provided quantitative results for English proficiency

assessments, reducing grading time. However, it remained limited in scope, as it did not assess speaking or listening skills.

Renewed interest in AI-based oral test assessments emerged with Koizumi (2022), who highlighted the challenges of validity and reliability in oral assessments, often influenced by human graders' subjectivity, varying skill levels, and biases. Koizumi emphasized the need for standardization using rubrics that account for variability and provide actionable feedback with quantifiable results in both formative and summative testing frameworks.

In Japan, prioritizing oral skills assessment has become a national focus in recent years. Since the 2022 academic year, all third-year public junior high schools in Tokyo have been required to conduct oral examinations, with evaluations outsourced to private contractors. In the first wave of examinations, audio data was sent offshore to the Philippines for assessment by local graders (Honda & Tsuchidate, 2022). While this approach provided an alternative to in-house grading, it raised concerns about labor intensity and potential data security breaches under foreign legal frameworks.

To contribute to the field of AI in education, this study explores the potential of a fine-tuned GPT-3.5 turbo model to generate predictions closely aligned with actual scores and to determine whether a monotonic relationship exists between its predictions and human-assigned grades when trained with a small dataset using a process known as few-shot learning (IBM, 2024). The Spearman Rank Correlation (SRC), p-values and Mean Absolute Error (MAE) will provide valuable insights into the mechanics of the models' predictability operations and guide further fine-tuning of the model by targeted improvements to address specific weaknesses.

This study has two exploratory objectives. The primary goal is to create a highly specialized model that demonstrates statistically significant performance with acceptable alignment to human scoring standards. The secondary objective is to develop a more generalized model suitable for broader assessment purposes through prompt design and additional small size data sets.

2. Method

2.1 Processes – Process One

There were three major processes involved in this project, as shown in Figure 1. The study involved 50 consenting students enrolled in a CEFR A1-A2 level English course (Kenny & Woo, 2011) as a second language class focusing on oral fluency at a university in Japan. Of the 50 datasets gathered, 40 were used for creating training data, and the remaining 10 were reserved for evaluating model performance. This very small sample size would affect the statistical significance of the study's results but could still indicate whether further exploration of the models is warranted. Each student was required to complete a series of three oral tests, designed as part of their course outcome requirements. The tests

consisted of topic-specific scripted interviews contained in a rubric developed for use by both the fine-tuned model and human grader.

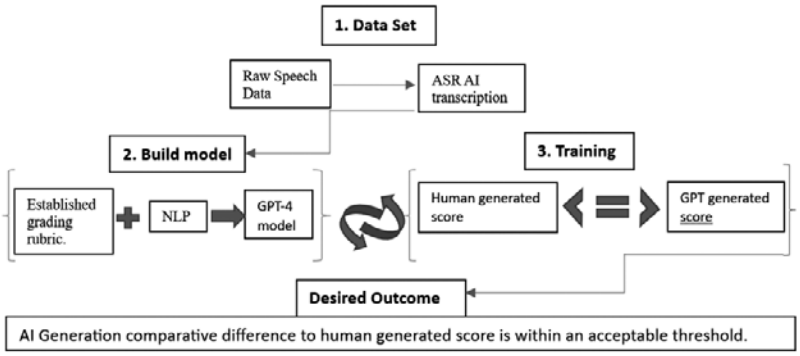

Figure 1　*Training Flow Diagram*

2.1.1 *Rubric Development for Fine-Tuning Conditionals*

Prior to data collection, a rubric was designed to account for entropic loss calculations and binary conditional classifications. Two key considerations guided the rubric's design: The first was whether the rubric reflected the students' ability to produce English conversational oral output aligned with course outcomes, and second, whether the rubric accommodated the cognitive processes of both human and AI graders.

The rubric was informed by Schema Theory (Piaget, 1952; Alexander, 2003; Ghosh & Gilboa, 2014) and Cognitive Load Theory (Sweller, 1988; van Merrienboer & Sweller, 2005). These frameworks emphasized the importance of managing cognitive load for both human and AI assessors. Research by Lake et al. (2017) demonstrated that the learning schema of an AI model mirrors that of a human assessor, as machine learning systems make inferences and representations based on small datasets, akin to human cognitive processes.

The grading rubric for student performance was influenced by Roever and Ikeda's research (2021), on interactional competence. Their rubric included six criteria for assessing role-play development:

"Facility with the language
Language use to deliver intended meanings
Language use for mitigation
Social actions
Engagement in interaction
Turn organization" (Roever & Ikeda, 2022, p11)

Of these, three criteria—language use for mitigation,

engagement in interaction, and turn organization—were selected as outcome markers for Tests 1 through 3. The exclusion of the other three criteria was justified by two factors: the low-level language proficiency of students in the CEFR A1-A2 course, and the need to control vocabulary and sentence structures in dialogues for ease of analysis during successive model training iterations.

The selected criteria were defined as follows:

- Language Use for Mitigation (LM): Encompassed opening and closing conversations and the use of politeness strategies.
- Engagement in Interaction (EI): Included responses to questions with reasoning. The explicit use of "because" served as a reasoning marker for the GPT grader. Indicators of interactive listening, such as repeating statements, signified comprehension checks.
- Turn Organization (TO): Addressed the student's ability to know when to speak in a dialogue.

In the final rubric, these three criteria were mapped to a binary grading mechanism supplemented by target vocabulary and grammar requirements, which were constants across all prompt fields. These elements were incorporated into the training data for Test 1 seen in Appendix A. To score a point, students needed to meet all prompt conditions. If one or more conditions were unmet, no point was awarded. However, the oversimplification of the rubric for ease of grading may have inadvertently impacted the efficacy in measuring the student's English oral language ability.

The cognitive load for both AI and human graders was managed by designing a rubric with simple, binary scoring conditions. This approach reduced cognitive strain and minimized error rates. For the fine-tuned model, its limited dataset inherently constrained its cognitive capacity, which aligned well with the rubric's straightforward structure. The binary design also facilitated descriptive and metric analysis during training loop iterations, contributing to model improvement.

Using binary-choice assessment Söderbom (2009) emphasized the transition of fine-tuned models from random guessing to informed decision-making. This process underscored the importance of structured training loops incorporating clear NLP prompts, detailed feedback, and incremental adjustments to the rubric. These measures guided the model toward achieving alignment patterns with that of the human grader.

2.1.2 *Data Collection*

The tests were administered in person by the researcher, who graded the students' performances live while simultaneously recording their responses. Audio data was captured in the Waveform Audio (.wav) file format using a custom-built Python application. This was developed from the Pydub library for audio recording and Tkinter for the graphical user interface (GUI). The recorded responses were

systematically categorized and securely stored on a solid-state drive for subsequent processing and analysis.

2.1.3 *Data processing*

Data processing was performed using a bespoke application developed by the researcher. The initial step involved transcription using OpenAI's Whisper Multilingual (default) and Whisper-en, an English ASR model. Following transcription, a total of 1,997 transcription and audio files were meticulously reviewed and cleaned to correct phrasing errors, preserve meaning and context, and ensure that all transcriptions were in English when using the default Whisper ASR model.

After the cleaning process, the transcription files were combined and reformatted into a JavaScript Object Notation Line (jsonl) compatible file format. This format was prepared specifically for use with the fine-tuned GPT model, developed in Process Three, to facilitate its analysis and training loops.

2.2 Process Two and Three

In process two the engineered prompts were combined with the cleaned transcription files and added to training data for both models. A software application was developed to assist with the creation and formatting of training data. The jsonl file data was formatted in accordance with Open AI's formatting guidelines, and classified into three specific categories: System, User and Assistant as seen in the code snippet seen in Figure 2.0. Briefly, the system contains context and environmental parameters for the GPT model to be aware of. The user contains the communication between the user and AI that serves as a space for placing the analysis object, and assistant contains output format and grading rubric criteria.

The system role was used to state that the AI was grading the dialogue and placing emphasis on points of rubric interest. The user role was populated with a two-person dialogue of single statement and causal response. Finally, the *assistant* role managed the AI's response following a predefined output based on the oral test rubric. These were allocated a binary true false switch to facilitate grading and analysis. While creating the training data, considerations were made in terms of token usage and fine-tuning entropic loss values while in training loops.

2.2.1 *Token Usage*

Tokens refer to the individual packets of text that the model processes. There is no fixed standard for what constitutes a token; it may include integers, individual words, or character strings (OpenAI, 2022). As an example, the phrase, including the brackets and quotation marks *("the cat sat on the mat")*, would have a token count of 10. Each word is allocated a token, and punctuation marks, such as brackets and quotation marks, are also allotted as separate tokens. Token size and usage are critical for fine-tuning as they influence how the model interprets and learns from the data. Minimizing

unnecessary tokens, such as line breaks, spaces, punctuation, and positional markers (e.g., brackets and square brackets), is essential to enhance fine-tuning efficacy and cost effectiveness.

Additionally, token usage serves as a metric for OpenAI to calculate the cost of training a model and its associated API usage, particularly when embedded into applications like grading systems. In this research, tokens were employed during every training session and API call for grading oral tests, and usage was billed accordingly by OpenAI.

2.2.2 *Training Analytics – Entropic Loss*

In the training phase the Entropic loss metric, also known as cross-entropy loss, is an essential metric for measuring and improving the training of classification models (OpenAI, 2024; Gunel et al., 2021; Mao et al., 2023). When a model makes probabilistic predictions, it assigns confidence scores to each possible answer. Put simply as an example, the model might estimate it is 80% confident the answer is "A" and 20% confident it is "B" when compared to the true score.

The entropic loss metric compares the model's confidence scores to the actual answers, quantifying how far the model's predictions deviate from perfect alignment. A smaller entropic loss value indicates better predictions and improved model performance. The entropic loss is mathematically defined as follows:

$$L = -\frac{1}{\mathrm{N}} \sum_{i=1}^{N} \sum_{c=1}^{C} yi, c \log (pi, c)$$

In the formula, L denotes the total loss, N is the number of data samples while C is the total number of classes the model is attempting to predict. y_i, c is a binary indicator showing if a specific sample i belongs to a particular class c. If true, the value is 1; otherwise, 0 for the remaining classes. To quantify demerit allocations for incorrect estimates, a logarithmic penalty, $log(p_i, c)$ is applied. For example, in a prediction of 80% for the correct answer, the penalty is low, whereas 20% certainty would warrant a high penalty and force the models to learn and improve.

Two summations are performed. The first, adds all error for likely answers on all class c, and i samples, and the second adds up the total of all examples in the batch, a data processing area calculated from the total training data size. Lastly, $-\frac{1}{\mathrm{N}}$ is used to convert the negative previous summation result, and then averaged across the number (N) of samples equating the loss of the batch.

Once the training was complete, the remaining ten students' oral test transcriptions were processed through the secondary program that incorporated the fine-tuned model's API and job reference number from OpenAI.

2.2.3 *Model Performance Metrics*

Each fine-tuned model underwent a series of analyses, which included a Spearman Rank Correlation

(SRC), p-value, and Mean Absolute Error (MAE) metrics, explained in the following sections, alongside scatterplots to guide alterations in the training data and improve model performance. The SRC, p-value, and MAE analyses were implemented in a custom application using Python's SciPy library (scipy.stats) for the statistical calculations and Matplotlib (matplotlib.pyplot) for generating scatterplot graphs.

2.3 Spearman Rank Correlation.

The Spearman Rank Correlation (SRC) is a valuable metric for measuring the rankings of two sets of variables, commonly applied in AI research involving small datasets and few-shot learning (Wang & Brown, 2008; Wu, Zhang, & Bao, 2015; Mumtaz & Giese, 2022). SRC is a non-parametric measurement tool, in that it does not impose strict assumptions about variable distributions. Instead, it evaluates the order or rank of variables rather than their exact values, making it suitable for monotonic data where one variable consistently increases or decreases with the other, without changes in the direction of data points. The metric is mathematically defined by the following formula (Spearman, 1904; Janse et al., 2021; The Knowledge Academy, 2023):

$$r_s = 1 - \frac{6\Sigma\, d_i^2}{n(n^2 - 1)}$$

In this equation $d_i = Rank\ (X_i) - Rank(Y_i)$ means that $Rank\ (X_i)$ represents the model score prediction against the actual scores ($Rank(Y_i)$), or in the case of this investigation,

$$d_i = Human\ grader\ (X_i) - FineTuned\ GPT\ (Y_i)$$

The values are then calculated to produce the difference between the ranks of the variables and subsequently squared. The calculation is then completed by assigning n with the number of test runs. If the values of rs = +1, then the variables have a prefect positive monotonic relationship. If rs = -1, then there is a perfect negative monotonic relationship and finally an rs = 0 indicates no monotonic relationship exists.

2.4 Probability value (p-value).

The p-value is often associated with the Spearman Rank Correlation (SRC) metric to serve as a binary indicator for null hypothesis agreement or rejection, thereby determining the statistical significance of monotonic relationships. This makes it particularly suited for evaluating AI models (Beaujean et al., 2011).

After the SCR is calculated the r_s value is inserted into the z-statistic for a sample size greater than 30 or a t-statistic of less than 30, ideal for this investigation's dataset size. This assessment determines whether the observed correlation is statistically significant or the result of random chance (Wasserstein & Lazar, 2016). This is illustrated in the following formula.

$$p = P\left(Z \geq |z| \middle| H_0\right)$$

The SCR, the r_s value replaces the Z variable seen below.

$$p = P\left(R_s \geq |r_s| \middle| H_0\right)$$

The p-value results are indicative of the following. If p is less than 0.05 then the observed result is unlikely to have occurred by random chance, rejecting the null hypothesis (H_0). However, If the p-value is greater or equal to 0.05 then the results are not statistically significant, and the null hypothesis (H_0) stands. This threshold ensures that any observed correlations were judged for their likelihood of being due to either random variation or genuine monotonic relationships.

2.5 Mean Absolute Error.

Mean absolute error (MAE) metric measures the average absolute difference between predicted values and actual values and has seen reliable use in AI regression model development using few shot-learning low data sizes (Qi et al., 2020a; 2023; Willmott & Matsuura, 2005). MAE provides insights into the relationship between predicted data and actual data, offering a reliable measure of model performance (Unal et al., 2023). The metric is useful for guiding further training development and project feasibility. MAE is well-suited for quantifying the alignment of predicted scores with human-assigned grades, providing a foundational evidence base for refining NLP prompt design and data size balances.

The MAE is calculated using the following equation.

$$MAE = \frac{1}{n}\sum_{\{i=1\}}^{n} \left| y_i^{\{\{pred\}\}} - y_i^{\{\{true\}\}} \right|$$

In this formula, n represents the total number of observations, $y_i^{\{\{pred\}\}}$ is the predicted value of an observation compared to the actual value $y_i^{\{\{true\}\}}$. The absolute value operation ($|.|$)

ensures that the true values are maintained numerically positive to prevent positive and negative prediction cancelations that result in a misleading low average error.

In relation to this research, the equation is populated with model specific variables.

$$MAE = \frac{1}{n}\sum_{\{i=1\}}^{n} \left| y_i^{\{\{model\}\}} - y_i^{\{\{human\ scores\}\}} \right|$$

In the equation, calculates the sum of the absolute difference of the Fine-tuned model score ($y_i^{\{\{model\}\}}$) and the human score ($y_i^{\{\{human\ score\}\}}$) and divide it by the total number of tests (n). In this case, 10.

A high MAE indicates high deviation between the two scores, suggesting the model requires further training whereas a low would mean closer alignment reflective of a model with better performance.

3. Results

Two models were being trained with results separated into three fine-tuned models and their subsequent four runs. The primary model (hereby known as Fine-tune 1) was tuned to assess only Oral Test One. The two secondary models (hereby known as Fine-tune 2 and Fine-tune 3) were developed for a general assessment role trained with Oral Tests One through Three.

The results for all three fine-tuned models are as presented.

3.1 Fine-tune 1 Model

Fine-tune 1 contained 40 participant transcriptions as contributors to training sets created only from Oral Test One. Included was a very general system instruction prompt design establishing the model's working environment. A second language grader without extra procedural criteria.

```
"messages": [
        {
            "role": "system",
            "content": "You are grading second
language learning students conversational dialogue."
        } ]
```

In this system prompt design, the NLP instructions were limited to positioning the GPT model within the context of the task and did not contribute to the finer aspects of grading. This approach evaluated the impact that system instructions had on the entropic loss value during training, either increasing or decreasing it, providing a starting point for measurement of prompt design.

Seen in Figure 2, during training, the entropic loss value on the y-axis initially quickly declined between epochs 1-100 on the x-axis, indicating relatively fast adaptations to patterns in the training data sets. After this decline, the entropic loss stabilized, reaching its minimum value of 0.3767. This signified that the model had learnt the base structure of the data sets but was uncertain on optimal generalization or potential overfitting. That is to say, the core elements of the training data were learnt as well as some

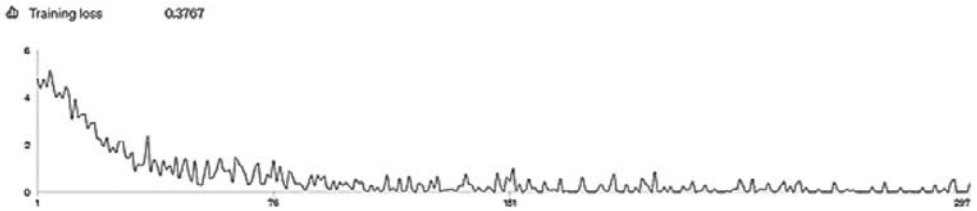

Figure 2 *Fine-tune 1 Model Training: Entropic Loss*

unnecessary noise in the data that created a model too specific and thus lost its generalizability for other test data.

In the four runs, the fine-tuned GPT model's scoring was consistent, each having a similar trend and minimal deviation in scores. The model produced accuracy rates of between 40% to 60% as seen in Table 1. The SRC analysis of run four came in at 0.706, indicating that the model had reached grading pattern alignment stability with the human grader after being trained on Oral Test One data sets only, and thus hyper-specializing the model. Additionally, the p-value of 0.022, confirmed positive for moderate statistical significance in that the observed correlation in ranking was unlikely to be due by chance. However, it must be noted that this was an exploratory study aimed at investigating if the model has a monotonic relationship and is worth the investment in additional training data and prompt designs for later robustness and reliability evaluations.

Table 1 *Fine-tune 1 Run Scores*

Test Number	1	2	3	4	5	6	7	8	9	10
Human Score	5	4	4	1	5	4	4	5	5	5
Fine-tune 1 Run 1	4	5	5	4	4	5	4	5	5	5
Fine-tune 1 Run 2	4	5	5	5	4	4	5	4	5	5
Fine-tune 1 Run 3	5	4.75	4.65	4	5	3.5	5	5	4.75	5
Fine-tune 1 Run 4	5	4.75	4.65	4	5	3.5	5	5	4.75	5

The MAE score was 0.617, indicating a moderately low average error and a good balance between rank and accuracy. This result warranted further fine-tuning as a test specific specialized fine-tuned grader model for future use.

The scatter plot for the first model, as shown in Figure 3, illustrates that the fine-tuned points generally clustered near perfect alignment, producing a moderate to strong distribution in score correlation supporting the SRC and MAE metric results. Seen in Table 2.

Table 2 *Fine-tune 1 Model Results*

Loop number	Entropic loss	SRC	P-value	MAE
Fine-tune 1	0.3767	0.706	0.022	0.617

Note. SRC = Spearman Rank Correlation; P-value = Probability value; MAE = Mean Absolute Error.

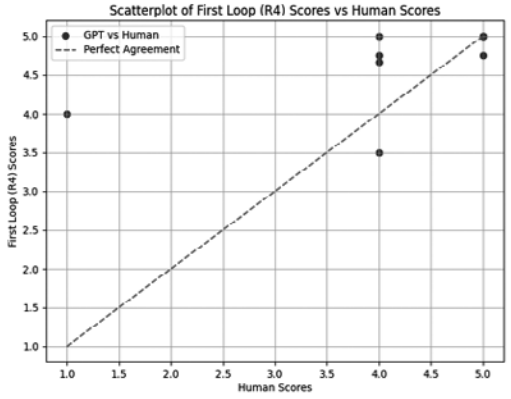

Figure 3 *Fine-tune 1 Resuit Distribution*

3.2 Fine-tune 2 Model

Creation of a more generalized model that could score Oral Tests Two and Three involved an additional 20 training data sets with an extra line of specific NLP instructions in its prompt design, namely the requirement of person B to answer person A's question. This was done with the intention to maintain the SRC and p-values to significant thresholds and decrease entropic loss and was a modification across all data sets. The modified system role is shown below.

```
"messages": [
        {
            "role": "system",
            "content": "You are grading second
language learning students conversational dialogue.
person_b must answer person_a question."
```

The training data was reformatted and compiled to create Fine-tune 2 model of which the results are seen in Table 3.

The new training data was used to create the Fine-tune 2 model and had an entropic loss of 0.2845, decreasing from the previous 0.33767 seen in Figure 4 and Table 3. This showed that the model only slightly benefited from the 20 extra training datasets, half from Oral Test Two and the other half from Oral Test Three along with the addition of minimal NLP instructions.

As in Fine-tune 1, the training loss showed a steep fall at the same epochs, suggesting a fast-learning rate. This behavior is representative of a model learning sphere that not only learnt core patterns but included the consumption of noise and extraneous details, which lead to the phenomenon known as overfitting.

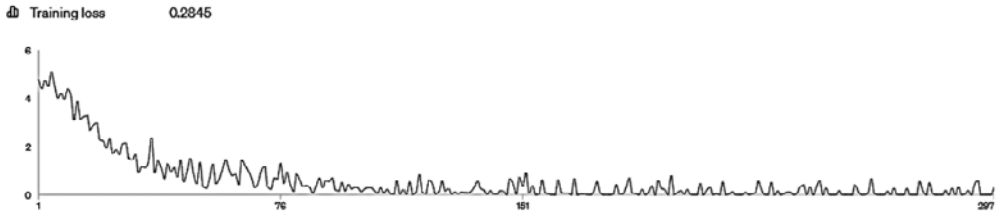

Figure 4 *Fine-tune 2 : Entropic Loss*

Table 3 *Fine-tune 2 Run Scores*

Test Number	1	2	3	4	5	6	7	8	9	10
Human Score	5	4	4	1	5	4	4	5	5	5
Fine-tune 2 Run 1	5	3	3	4	3	5	3	5	5	5
Fine-tune 2 Run 2	5	5	3	3	4	3	5	3	5	5
Fine-tune 2 Run 3	5	5	4	4	4	2	5	4	5	5
Fine-tune 2 Run 4	5	5	4	2	4	3	5	5	4	5

Fine-tune 2 model demonstrated moderate alignment with the human scores, achieving an SRC score of 0.457, particularly in the high score range of the human grader. However, it also overestimated low scores as indicated in the Fine-tune 2 run scores in Table 3. Noticeable deviations were observed, especially in the midrange scores. The p-value of 0.57 seen in Table 4 was too high to be statistically significant indicating the correlation ranking was likely due to chance. The MAE score of 0.6 was also an indicator of low average error range and reasonable balance that required further fine-tuning adjustments similar to that in Fine-tune 1.

Table 4 *Fine-tune 2 Model Results*

Loop number	Entropic loss	SRC	P-value	MAE
Second loop	0.2845	0.457	0.57	0.600

Note. SRC = Spearman Rank Correlation; P-value = Probability value; MAE = Mean Absolute Error.

In Figure 5 the scatter plot graph signified some agreement of the model with the human greater, however deviations seen in scores four and five showed that the model on occasions over and underestimated scores.

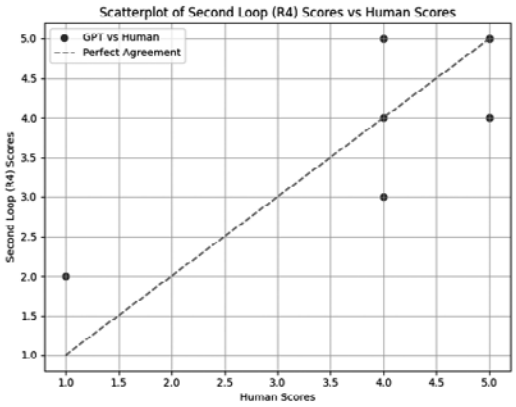

Figure 5 *Fine-tune 2 Model Scatter plot*

3.3 Fine-tune 3 Model

The focus of the Fine-tune 3 was on the exploration of prompt design on the model's prediction confidence scores and subsequent entropic loss value during training. A further two more learning parameters were added to the system role NLP instructions to increase its learning complexity.

```
{
    "messages": [
        {
            "role": "system",
            "content": "You are grading second
language learning students conversational dialogue.
person_b must answer person_a question. first greet
person a then describe how person_b is feeling\"
        ]}
```

Fine-tune 3 was trained and produced a significantly lower entropic loss measure of 0.0057 as seen in Figure 6. The loss was a sharp drop that stabilized in the 80th epoch with a very small amount of variation and much sooner than the previous two loops. It indicated that the additional NLP prompts provided a more task specific instruction set and suggested an increased detail in understanding of the tasks.

However, the improvement in entropic loss came at the expense of grading alignment scores across all four runs, seen in Table 5. The SRC score of 0.416, signified a depreciating trend of weaker alignment correlation to the human grader than in the Fine-tuned 2 model. This coefficient, being closer to zero, suggests an insignificant monotonic relationship between the model's predictions and the human grader's scores. The p-value of 0.688 was much higher above the threshold of $p<0.05$, further confirming its weakness and statistical irrelevance as seen in Table 6.

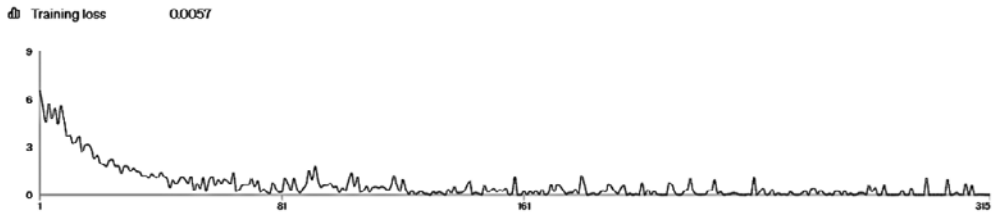

Figure 6 *Fine-tune 3 Entropic Loss*

Table 5 *Fine-tune 3 Run Scores*

Test Number	1	2	3	4	5	6	7	8	9	10
Human Score	5	4	4	1	5	4	4	5	5	5
Fine-tune 3 Run 1	5	5	4	2	3	4	5	5	5	5
Fine-tune 3 Run 2	4	3	5	5	4	3	5	3	5	4
Fine-tune 3 Run 3	5	5	4	3	5	4	5	5	2	4
Fine-tune 3 Run 4	5	4	3	4	3	5	4	5	5	3

The MAE score was 0.9, significantly higher than previous iterations, indicating unacceptable grading deviations. This result added to the weak SRC score and the statistical insignificance of the Fine-tune 3 Model, further highlighting its limited reliability and performance.

Table 6 *Fine-tune 3 Model Results*

Loop number	Entropic loss	SRC	P-value	MAE
Fine-tune 3	0.0057	0.146	0.688	0.900

Note. SRC = Spearman Rank Correlation; P-value = Probability value; MAE = Mean Absolute Error.

The scatterplot observations further revealed significant deviations from the human scores, with larger distances indicating substantial prediction errors. These outliers suggested that the model was struggling to achieve accurate predictions, highlighting its limited performance in aligning with human grading seen in Figure 7.

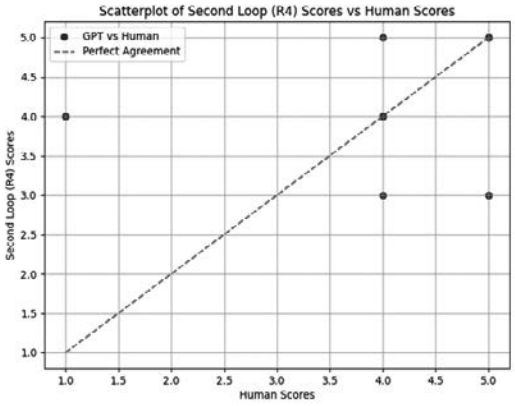

Figure 7 *Fine-tune 3 Model Scatter Plot*

4. Discussion

The results of this exploratory study identified the potentials and limitations of fine-tuning GPT-3.5 turbo models for assessing oral language proficiency in small dataset environments. The results from the Fine-tune 1 model provided encouraging initial outcomes, with the entropic loss in training stabilizing quickly and achieving a minimum value. This reflected the model's ability to learn foundational considerations from the training data, such as basic sentence structure and turn-taking dynamics in conversational exchanges. However, while the MAE and SRC metrics supported the close alignment to the true scores of the specialized Fine-tune 1 model, its hyper-specialization raised concerns about its adaptability to diverse datasets. For example, the model performed well within the specific parameters of Test One but showed limited versatility when faced with variations in language use or unseen variables used in conversations such as fillers, and word repetitions. This highlights a major trade-off in fine-tuning AI models, that is, achieving high performance on narrowly defined tasks often sacrifices generalizability needed for practical applications in language education.

Fine-tune 2 and Fine-tune 3 models showed further challenges in finding a balance in precision and flexibility. Although additional data and refined NLP instruction sets were incorporated to improve model performance, the results were inconsistent. Fine-tune 2 demonstrated a moderate SRC score, suggesting reasonable alignment with high-score ranges, but there were overestimations in low-score ranges and deviations in midrange scores. This signified deep irregularities in the model's grading logic. Qualitative observations revealed that the model may have misinterpret hesitation markers or filler words in student speech as indicators of poor performance, whereas the human grader accounted for these as natural elements of oral language at the beginner level. Moreover, the model's consistent overfitting could be attributed to the layout design of the training sets. A review of the training data further supported this, revealing a complete absence of hesitation markers and filler words as examples for the AI to

learn. Additionally, large spaces between the system, user, and assistant roles, may have contributed to substantial noise that could have resulted in overfitting.

Fine-tune 3 was designed to improve entropic loss through additional parameters in the prompt design. Although the entropic loss improvements implied better data learning at the training level, the model's SRC and MAE scores showed weaker alignment with human grading, and the high p-values confirmed the statistical insignificance of its predictions. These findings suggested that quantitative improvements in entropic loss do not necessarily make gains in grading accuracy.

The scatterplots and statistical analyses revealed important qualitative gaps in the model's understanding of human conversational dynamics. Fine-tune 3 appeared to learn surface-level patterns, but it struggled to generalize to unseen data where conversational structures were absent from the training data. These results emphasize the need to carefully compile training data and design NLP prompts to accommodate interactional conversation dynamics and turn-taking. The overfitting observed in Fine-tune 3 also underlined the need for iterative experimentation with rubric designs and training loops to achieve a balance between task-specific precision and adaptability across different datasets. Further studies with larger data sets and robustness metrics like bootstrapping, may also provide insights into specific divergences between model predictions and human grading.

5. Conclusion

The explorative investigation was moderately successful in creating a statistically significant model. Fine-tune 1 demonstrated a moderate statistical significance and alignment with human grading within a specially defined task, indicating further investment in development. However, the wider scope of Fine-tune 2 and Fine-tune 3 fell short of generalizability and reliability in measuring the monotonic relationship found between Fine-tune 1 and the human grader. These results highlighted the challenges of balancing specialization and adaptation in fine-tuning AI models for complex tasks like oral language assessment. Fine-fine 1 warranted future further development of the training process. An increase in data set size with succinct and efficient prompt designs will be critical to enhance the applicability of AI-driven assessment tools for language education.

6. Limitations

This project faced limitations in time and human resources on prominent tasks, such as programming applications, creating training datasets, and transcribing audio data all of which were performed by a single researcher. Managing these tasks alone made it difficult to progress beyond the multi-training

loop stage within the one-year period. A larger team with at least two researchers could have helped the project move forward faster and produced more robust results. However, assembling such a team proved difficult due to the specific technical expertise needed to work with AI training systems currently unavailable at the home institution.

As mentioned throughout this report, the project was also constrained by the use of a small dataset, which significantly impacted the training and evaluation of the fine-tuned GPT-3.5 turbo models. The ability of the models to generalize was limited, as it lacked sufficient exposure to a diverse range of speech patterns, hesitation markers, and filler words commonly found in beginner-level oral assessments. This constraint likely contributed to the models' tendency to overfit to the training data and their struggle to adapt to unseen datasets, resulting in weaker alignment with human grading in broader tests. The small dataset size further limited the reliability and robustness of statistical analyses, such as SRC and p-values, raising questions about the validity of the findings for larger-scale applications.

Regarding transcriptions of participant oral recordings, OpenAI's Whisper AI was used as the main ASR tool. Audio data was processed through its API and returned as text files. However, Whisper AI struggled with beginner-level Japanese pronunciations of English, causing frequent errors distinguishing pronunciation variations and local pronouns, causing incorrect word use in transcriptions and sometimes transcribing in the wrong language entirely. These issues increased the time required to manually correct the transcriptions as recorded in the audio file. The high error rate of Whisper AI ASR, raised concerns about its impact on future AI-powered applications and brought forth the issue of interdependency of the two, prompting a proposed future investigation into ASR ground up development of non-native English ASRs.

Notes:

[1] This research was a part of a presidential research grant issued by Josai International University in 2023.

References

Aldoseri, A., Al-Khalifa, K. N., & Hamouda, A. M. (2023). Re-Thinking Data Strategy and Integration for Artificial Intelligence: Concepts, Opportunities, and Challenges. *Applied Sciences*, *13*(12), 7082. https://doi.org/10.3390/app13127082

Beaujean, F., Caldwell, A., Kollar, D., & Kroeninger, K. (2011). P-Values for Model Evaluation. *Physical Review D*, *83*(1). https://doi.org/10.1103/PhysRevD.83.012004

Davila, A., Colan, J., & Hasegawa, Y. (2024). Comparison of Fine-tuning strategies for Transfer Learning Medical Image Classification. arXiv preprint arXiv:2406.10050.

Godfrey, J. J., Holliman, E. C., & McDaniel, J. (1992). Switchboard: Telephone Speech Corpus for Research and Development. *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, *1*, 517-520. https://doi.org/10.1109/ICASSP.1992.225858

Graff, D., & Bird, S. (2000). *Many Uses, Many Annotations for Large Speech Corpora: Switchboard and TDT as Case Studies* (No. arXiv:cs/0007024). arXiv. https://doi.org/10.48550/arXiv.cs/0007024

Gunel, B., Du, J., Conneau, A., & Stoyanov, V. (2021). *Supervised Contrastive Learning for Pre-trained Language Model Fine-tuning* (No. arXiv:2011.01403). arXiv. https://doi.org/10.48550/arXiv.2011.01403

Honda, Y., & Tsuchidate., S. (2022, November 28). Tokyo Includes 1st Speaking Test for English in Entrance Exam Yuka Honda and Soichi Tsuchidate. *The Asahi Shinbun*.

Hirai, A., & Kovalyova, A. (2023). Using Speech-to-Text Applications for Assessing English Language Learners' Pronunciation: A Comparison with Human Raters. In: Suárez, Md. M., El-Henawy, W.M. (eds) Optimizing Online English Language Learning and Teaching. *English Language Education*, vol 31. Springer, https://doi.org/10.1007/978-3-031-27825-9_17

Ikeda, N. (2017). Measuring L2 oral pragmatic abilities for use in social contexts: Development and validation of an assessment instrument for L2 pragmatic performance in university settings (Doctoral dissertation). University of Melbourne.

IBM. (2024). *What is Few-shot Learning?* IBM. Retrieved July 8, 2024, from https://www.ibm.com/topics/few-shot-learning

Janse, R. J., Hoekstra, T., Jager, K. J., Zoccali, C., Tripepi, G., Dekker, F. W., & Van Diepen, M. (2021). Conducting Correlation Analysis: Important Limitations and Pitfalls. *Clinical Kidney Journal*, *14*(11), 2332-2337. https://doi.org/10.1093/ckj/sfab085

Kenny, T., & Woo, L. (2011). *Nice Talking with You: Level 1*. Cambridge University Press.

Koizumi, R. (2022). L2 speaking assessment in secondary school classrooms in Japan. *Language Assessment Quarterly, 19*(2), 142-161.

Lake, B. M., Ullman, T. D., Tenenbaum, J. B., & Gershman, S. J. (2017). Building machines that learn and think like people. *Behavioral and Brain Sciences*, *40*, e253. https://doi.org/10.1017/S0140525X16001837

Loeb, S., Dynarski, S., McFarland, D., Morris, P., Reardon, S., & Reber, S. (2017). Descriptive analysis in education: A guide for researchers. *U.S Department of Eduction, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance.*

Mao, A., Mohri, M., & Zhong, Y. (2023). *Cross-Entropy Loss Functions: Theoretical Analysis and Applications* (No. arXiv:2304.07288). arXiv. https://doi.org/10.48550/arXiv.2304.07288

Marquis, Y. A., Oladoyinbo, T. O., Olabanji, S. O., Olaniyi, O. O., & Ajayi, S. A. (2024). Proliferation of AI Tools: A Multifaceted Evaluation of User Perceptions and Emerging Trend. *Asian Journal of Advanced Research and Reports*, *18*(1), 30-35. https://doi.org/10.9734/ajarr/2024/v18i1596

McCrocklin, S., & Edalatishams, I. (2020). Revisiting Popular Speech Recognition Software for ESL Speech.

TESOL Quarterly, *54*(4), 1086-1097. https://doi.org/10.1002/tesq.3006

McCrocklin, S., Humaidan, A., & Edalatishams, I. (2018). ASR Dictation Program Accuracy: Have Current Programs Improved? *Pronunciation in Second Language Learning and Teaching Proceedings*, *10*. https://doi.org/10.31274/psllt.15376

Mumtaz, S., & Giese, M. (2022). Hierarchy-based Semantic Embeddings for Single-valued & Multi-valued Categorical Variables. *Journal of Intelligent Information Systems*, *58*(3), 613-640. https://doi.org/10.1007/s10844-021-00693-2

OpenAI. (2022). *API Reference Documentation*. OpenAI. Retrieved September 9, 2024. from https://platform.openai.com/docs/api-reference

Piaget, J. (1952). The origins of intelligence in children. International Universities Press.

Qi, J., Du, J., Siniscalchi, S. M., Ma, X., & Lee, C.-H. (2020). Analyzing Upper Bounds on Mean Absolute Errors for Deep Neural Network Based Vector-to-Vector Regression. *IEEE Transactions on Signal Processing*, *68*, 3411-3422. https://doi.org/10.1109/TSP.2020.2993164

Roever, C., & Ikeda, N. (2021). What scores from monologic speaking tests can(not) tell us about interactional competence. *Language Testing, 39*(1), 7-29. https://doi.org/10.1177/02655322211003332

Roever, C., & Ikeda, N. (2021). What Scores from Monologic Speaking Tests Can(not) Tell Us About Interactional Competence. *Language Testing*, *39*(1), 7-29. https://doi.org/10.1177/02655322211003332

Roumeliotis, K. I., & Tselikas, N. D. (2023). ChatGPT and Open-AI Models: A Preliminary Review. *Future Internet*, *15*(6), 192. https://doi.org/10.3390/fi15060192

Söderbom, M. (2009). Applied econometrics lecture 10: Binary choice models. University of Gothenburg. Retrieved from https://www.soderbom.net/lecture10notes.pdf

Settles, B., T. LaFlair, G., & Hagiwara, M. (2020). Machine Learning–Driven Language Assessment. *Transactions of the Association for Computational Linguistics*, *8*, 247-263. https://doi.org/10.1162/tacl_a_00310

Songsiengchai, S., Sereerat, B., & Watananimitgul, W. (2023). Leveraging Artificial Intelligence (AI): ChatGPT for Effective English Language Learning among Thai Students. *English Language Teaching*, *16*(11), 68. https://doi.org/10.5539/elt.v16n11p68

Spearman, C. (1904). The Proof and Measurement of Association between Two Things. *The American Journal of Psychology, 15*(1), 72-101.

Sweller, J. (1988). Cognitive Load During Problem Solving: Effects on Learning. *Cognitive Science*, *12*(2), 257-285. https://doi.org/10.1207/s15516709cog1202_4

Thayyib, P. V., Mamilla, R., Khan, M., Fatima, H., Asim, M., Anwar, I., Shamsudheen, M. K., & Khan, M. A. (2023). State-of-the-Art of Artificial Intelligence and Big Data Analytics Reviews in Five Different Domains: A Bibliometric Summary. *Sustainability*, *15*(5), 4026. https://doi.org/10.3390/su15054026

The Knowledge Academy. (2023, October). *Spearman's Rank Correlation: A Comprehensive Guide*. Retrieved

from https://www.theknowledgeacademy.com/blog/spearmans-rank/

Unal, A., Asan, B., Sezen, I., Yesilkaynak, B., Aydin, Y., Ilicak, M., & Unal, G. (2023). *Climate Model Driven Seasonal Forecasting Approach with Deep Learning* (No. arXiv:2302.10480). arXiv. https://doi.org/10.48550/arXiv.2302.10480

Wang, J., & Brown, M. S. (2008). Automated Essay Scoring Versus Human Scoring: A Correlational Study. *Contemporary Issues in Technology and Teacher Education*, *8*(45), 310-325.

Wasserstein, R. L., & Lazar, N. A. (2016). The ASA's Statement on p-Values: Context, Process, and Purpose. *The American Statistician, 70*(2), 129-133.

Willmott, C., & Matsuura, K. (2005). Advantages of the Mean Absolute Error (MAE) Over the Root Mean Square Error (RMSE) in Assessing Average Model Performance. *Climate Research*, *30*, 79-82. https://doi.org/10.3354/cr030079

Wu, Y., Zhang, Z., & Bao, F. (2015). Secure Two-party Rank Correlation Computations for Recommender Systems. 2015 IEEE Trustcom/BigDataSE/ISPA, 732-739. https://doi.org/10.1109/Trustcom.2015.435

Yu, B., Kaku, A., Liu, K., Parnandi, A., Fokas, E., Venkatesan, A., Pandit, N., Ranganath, R.,Schambra, H., & Fernandez-Granda, C. (2024). Quantifying Impairment and Disease Severity Using AI Models Trained on Healthy Subjects. Proceedings of the National Science Foundation. Retrieved from https://par.nsf.gov/biblio/10521299-quantifying-impairment-disease-severity-using-ai-models-trained-healthy-subjects

van Merriënboer, J. J. G., & Sweller, J. (2005). Cognitive Load Theory and Complex Learning: Recent Developments and Future Directions. *Educational Psychology Review, 17*(2), 147-177. https://doi.org/10.1007/s10648-005-3951-0

Wu, Y., Zhang, Z., & Bao, F. (2015). Secure Two-party Rank Correlation Computations for Recommender Systems. 2015 IEEE Trustcom/BigDataSE/ISPA, 732-739. https://doi.org/10.1109/Trustcom.2015.435

Zhou, J. (2019). Construction of Artificial Intelligence-Based Interactive Oral English Teaching Platform Based on Application Problems of Present Intelligent Products. *IOP Conference Series: Materials Science and Engineering*, *569*(5), 052-055. https://doi.org/10.1088/1757-899X/569/5/052055

Appendix A

Oral Test One grading rubric properties.

Prompt 1

- *Opening of conversations* (LM)
- Probe - Target vocabulary/ appropriate grammar. [1pt]
- Topic management, reaction: Answers interviewer's questions (TO) [1pt]

Prompt 2

- Probe - Target vocabulary/ appropriate grammar. [1pt]
- Topic management, reaction: Answers interviewer's questions and provides reasoning using because (TO) (EI) [1pt]

Prompt 3

- Probe - Target vocabulary/ appropriate grammar. [1pt]
- Topic management, reaction: Answers interviewer's (TO) [1pt]

Prompt 4

- Exiting conversation (LM) [1pt]
- Probe - Target vocabulary/ appropriate grammar. [1pt]
- Topic management, reaction: Presents new topic to interviewer provides reasoning using because (EI) [1pt]

Prompt 5

- Closing conversation (LM) [1pt]
- Probe - Target vocabulary/ appropriate grammar [1pt]

Oral Test One grading rubric and assistant NLP for Fine-tuned models and human grader.

Dialogue Segment One

Segment	All correct	Not correct	Not correct	Not correct
A: Hi. How are you? B: Hi D I'm good.	1.Grammar: 1, 2.Opening question answered: 1, 3.Explanation: Answered Person A's Questions correctly., Total score: 1	1.Grammar: 1, 2.Opening question answered: 0, 3.Explanation: did not Answer Person A's Questions correctly., Total score: 0	1.Grammar: 0, 2.Opening question answered: 0, 3.Explanation: Did not use I am before the adjective., Total _score: 0	1.Grammar: 0, 2.Opening question answered: 0, 3.Explanation: Did not use I am before the adjective., Total score: 0

Dialogue Segment Two

Segment	All correct	Not correct	Not correct	Not correct
A: Why is that? B: Because I slept well last night. How about you?	1.Grammar: 1, 2.Opening explanation: 1 3.Includes the word (because): 1 4.includes the phrase (How about you?): 1 5.Explanation: Answered Person A's Questions correctly used 'because' and 'how about you'. Total score: 1	1.Grammar: 1, Answers person A questions: 1 3.Includes the word (because): 0 4.includes the phrase (How about you?): 1 5.Explanation: Answered person A's questions correctly. Total score: 0	1.Grammar: 1 Answers person A questions: 1 Includes the word (because): 0 4.includes the phrase (How about you?): 0 5.Explanation: Did not use the phrase (How about you?) or (because). Total score: 0	

Dialogue Segment Three

Segment	All Correct	Not Correct 1	Not Correct 2	Not Correct 3
A: I'm great B: I'm sorry I have to go now.	1.Grammar: 1, 2.States Intention to leave: 1, 3.Explanation: Informs speaker B that they have to leave the conversation. Total score: 1	1.Grammar: 1, 2.States Intention to leave: 0, 3.Explanation: does not Inform speaker A that they have to leave the conversation. Total score: 0	1.Grammar: 0, 2.States Intention to leave: 1, 3.Explanation: does not use correct sentence structure. Total score: 0	

Dialogue Segment Four

Segment	All Correct	Not Correct 1	Not Correct 2	Not correct 3
A: Why do you have to go? B: Because I have to go running practice this afternoon	1.Grammar: 1, 2.uses (because): 1, 3.Gives reason for leaving: 1, 4.Explanation: Used 'because' and gave speaker reason for leaving the conversation. Total score: 1	1.Grammar: 1, 2.Uses (because): 0, 3.Gives reason for leaving: 1, 4.Explanation: Did not use because at the beginning of the statement. Total score: 0	1.Grammar: 0, 2.uses (because): 1, 3.Gives reason for leaving: 1, 4.Explanation: Word order was not correct. Total score: 0	

Dialogue Segment Five

Segment	All Correct	Not Correct 1	Not Correct 2	Not correct 3
A: Ok be safe. B: Will do, bye.	1.Grammar: 1, 2.Answers person A questions: 1, 3.Explanation: answers person as request. 4.Total score: 1	1.Grammar: 1, 2.Answers person A questions: 0, 3.Explanation: does not answer person a request. Total score: 0	1.Grammar: 0, 2.Answers person A questions: 0, 3.Explanation: incorrect grammar. Total score: 0	