*Bulletin of Shiseikan University,* Vol.11, March. 2024

論　文

# LLM-Empowered Essay Scoring Methods Integrated with ChatGPT and GPT API

○Reina NISHIHARA*1　Keisuke UEDA*2　Katsumi KIYONAGA*3

Keywords：automated essay scoring, AES, LLM, human scoring, improved prompt

## 1 Introduction

Since the release of ChatGPT by OpenAI in November 2022[1], several reports have highlighted the advantages and drawbacks of this Large Language Model (LLM). Users have begun to examine the capabilities of ChatGPT, a type of language model designed to predict the most likely succeeding word after a given word. However, it has been reported that ChatGPT struggles in solving mathematical problems in a step-by-step fashion.[2][3][4]. It has also been reported that ChatGPT can score English essays for IELTS through SNS and YouTube, and Mizumoto & Eguchi demonstrated that ChatGPT's scoring and the level of essays for TOEFL have a correlation[5]. This experiment indicated a subtle disparity between the initial and subsequent assessments conducted by ChatGPT across different levels of English proficiency.

In English courses in Japan, opportunities for English output are restricted due to feedback burdens. Additionally, English educators face the dilemma of objective essay-scoring difficulty. Undergraduates primarily focus on listening and reading for TOEIC, since a test score often required in job applications. Nonetheless, they shall need writing and speaking proficiency for business-related activities later on. Evidently, it would be advantageous to have more frequent essay-writing occasions with feedback. Scoring an essay at an advanced level is relatively easy due to the limited number and nature of mistakes. However, low-level essays display different patterns of errors. Some students submit essays with short, flawless but overly simple prose while others submit essays riddled with lexical and grammatical errors. While challenging themselves is necessary to push their boundaries, the severity of these errors impacts readability, and there is ample room for improvement. To encourage students to tackle sophisticated essay-writing, an impartial assessment criterion is crucial.

IELTS, administered by the British Council, is globally utilized to certify English proficiency. The examination consists of listening, reading, writing, and speaking tasks, with straightforward scoring criteria. Each section is rated on a scale of 1 to 9, corresponding to CEFR levels[6][7]. IELTS writing and speaking tests receive evaluation from proficient examiners, and the standards for each band score can be accessed[8]. These criteria are suitable for an impartial evaluation, regardless of the type of essay.

Although a correlation between the essay level and ChatGPT scoring was reported[5], the comparison between human scoring and ChatGPT scoring remains unclear.

This study was conducted to clarify the accuracy of the automated essay scoring (AES) competence of ChatGPT, leading to its calibration. Moreover, this effort to develop feedback prompts using ChatGPT would contribute to improving the fairness of essay evaluation by English educators. It would also be very helpful for learners to hone their writing skills by themselves.

---

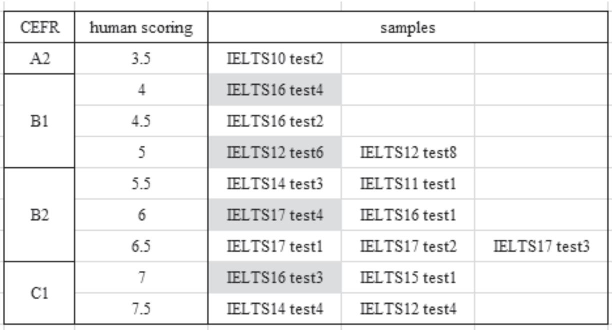
*1 中村学園大学　非常勤講師　　*2 無所属　　*3 至誠館大学　現代社会学部

## 2 Method

### *Material*

The sample answers provided with human scoring for the writing test Task 2 were selected from IELTS 10[9), 11[10), 12[11), 14[12), 15[13), 16[14), 17[15) published by Cambridge English (Table 1). Task 2 is an essay on a given topic that requires a minimum of 250 words. The selected essays below represent a wide range of scores, from 3.5 to 7.5, with one to three essays for each score.

**Table 1** Samples selected from a wide range of levels



| CEFR | human scoring | samples | | |
|---|---|---|---|---|
| A2 | 3.5 | IELTS10 test2 | | |
| B1 | 4 | IELTS16 test4 | | |
| | 4.5 | IELTS16 test2 | | |
| | 5 | IELTS12 test6 | IELTS12 test8 | |
| B2 | 5.5 | IELTS14 test3 | IELTS11 test1 | |
| | 6 | IELTS17 test4 | IELTS16 test1 | |
| | 6.5 | IELTS17 test1 | IELTS17 test2 | IELTS17 test3 |
| C1 | 7 | IELTS16 test3 | IELTS15 test1 | |
| | 7.5 | IELTS14 test4 | IELTS12 test4 | |

*Note: Colored four essays were used for the four-data analyses*

### *System Version*

Chat GPT-4 was utilized to check the output for Web UI prompts. The Python codes embedded GPT-3.5 Turbo API was used, and we utilized our personal API keys for this study. The version of Python utilized was 3.10.12, and the data were collected in late August 2023.

### *Parameter setting*

We utilized the GPT API (GPT-3.5 Turbo) through our Python programming to address the instability of the Web UI's output resulting from its parameter settings such as temperature[4)16). The term "temperature" originates from the notion of probability distribution. When the temperature value approaches zero, ChatGPT produces less random outputs[4)16). The temperature of the ChatGPT Web user interface has not been officially disclosed. However, discussions in the OpenAI Developer Forum suggest it may be between 0.7 and 1.0[17). We set temperature at 0.2 to program using the GPT API.

### *Flow*

First, we tried an original prompt to point out and correct errors as well as to assess the essay level on the ChatGPT Web UI and compared the scoring with human scoring provided in the books[8)9)10)11)12)13)14)15). Typical AES systems provide a grade, but we also offer improvement feedback that can address the issues overlooked by a human examiner. Since the inconsistency of the output has been reported[5), we conducted three trials for each test and compared the results with human evaluations. The original prompt we tested first was as follows:

*"You are a skilled IELTS examiner. Please evaluate the following essay written for task 2 in the 4 assessment factors, "Task Response", "Coherence and Cohesion", "Lexical Resource", and "Grammatical Range and Accuracy" from 0 to 9. Calculate the overall score as the average of the four criteria scores, rounded to the nearest 0.5. If the average ends in 0.25, round up to 0.5, and if it ends in 0.75, round up to the nearest whole number. Also, point out each grammatical error and lexical error exhaustively. Indicate inappropriate cohesive devices exhaustively. Could you give the candidate advice to improve as well?"*

Given the practical implications of the discrepancy observed between ChatGPT′s scoring of the original prompt and each human scorer (as shown in Figure 5), we pursued two methods: modifying the prompt according to reported ChatGPT behavior and analyzing the GPT API via Python codes (Figure 1). We selected four essays for this investigation to discern trends. The four essays are as follows: IELTS16 test3 (human scoring 7)[14), IELTS17 test4 (human scoring 6)[15], IELTS12 test6 (human scoring 5)[11], and IELTS16 test4 (human scoring 4)[14].

ChatGPT displays certain behaviors. It was reported that each method of adding "let's think step by step" at the end of the prompt and pre-step conversation for the main question improved the chain of thought (CoT)[2] [3]. Therefore, we conducted additional trials to improve the method. As a zero-shot approach, we added the prompt "let's think step by step" at the end of the original prompt. Additionally, we utilized a few-shot approach by consulting Chat GPT on the IELTS writing criteria for each band before presenting the original prompt.

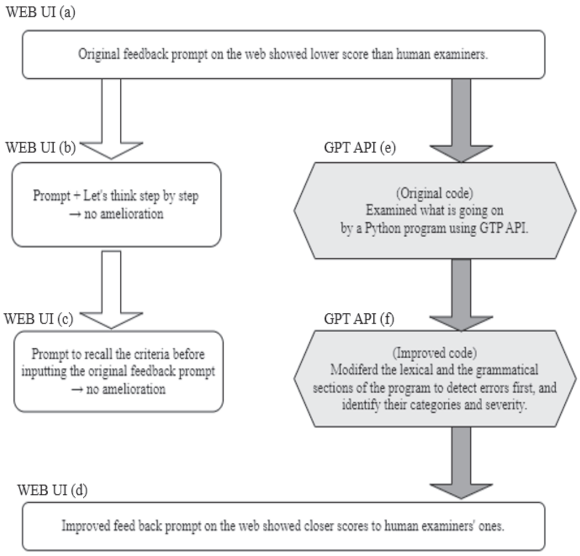
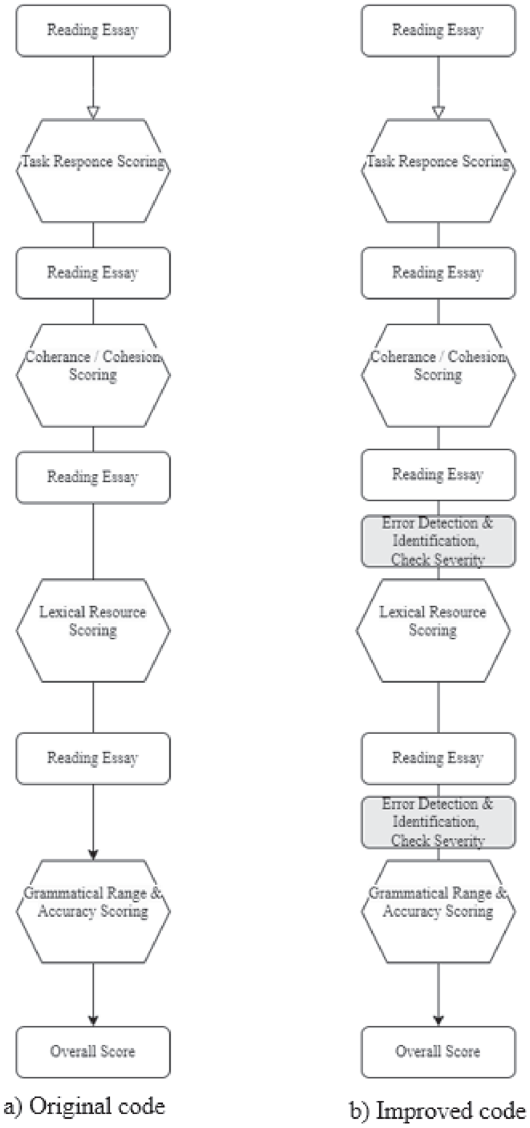

WEB UI (a)
Original feedback prompt on the web showed lower score than human examiners.
WEB UI (b)
Prompt + Let's think step by step → no amelioration
GPT API (e)
(Original code) Examined what is going on by a Python program using GTP API.
WEB UI (c)
Prompt to recall the criteria before inputting the original feedback prompt → no amelioration
GPT API (f)
(Improved code) Modiferd the lexical and the grammatical sections of the program to detect errors first, and identify their categories and severity.
WEB UI (d)
Improved feed back prompt on the web showed closer scores to human examiners' ones.

**Figure 1** Flow to improve a prompt for ChatGPT's Web UI

Subsequently, we analyzed the output of the GPT API using Python code as multiple parameters impact the Web UI output[4][16]. For instance, temperature seems to be set at 0.7-1.0[17]. Writing Python code is an effective means of eliminating the impact of parameters and encourages the linking of prompts to reach conclusions. Our original Python code analyzed four sections individually in a sequential manner and computed overall scores. Additionally, the code was designed to detect errors and provide exhaustive corrections. This version is referenced as the original Python code. After reviewing the results, we discovered that certain errors were being counted twice, which may have led to lower scores. To address this, we included the prompt outlined in Figure 3 as part of the lexical resource and grammatical range & accuracy criteria in our Python code (Figure 2, 3). This updated version is now referred to as the improved Python code. Our data was collected from analyzing 16 sample essays.


Reading Essay
Task Responce Scoring
Reading Essay
Coherance / Cohesion Scoring
Reading Essay
Error Detection & Identification, Check Severity
Lexical Resource Scoring
Reading Essay
Error Detection & Identification, Check Severity
Grammatical Range & Accuracy Scoring
Overall Score

a) Original code b) Improved code

**Figure 2** Flow chart of the original Python code and the improved Python code

After receiving the improved Python code output, we implemented the modifications to the UI prompt. To prevent any confusion, we prioritized command of error detection, classification, and severity checking. We then proceeded to evaluate the criteria for lexical resource, grammatical range & accuracy, task response, and coherence & cohesion. Additionally, we rewrote the calculation for the overall score. Refer to Figure 4 for the revised UI prompt details. The data was collected from the aforementioned sample essays.

_**Error Identification:** To assess "Lexical Resource" and "Grammatical Range and accuracy", identify all lexical and grammatical errors thoroughly._

_**Classify Errors:** Categorize the errors as either:_

_-**lexical** related to word choice, usage, and vocabulary range._

_-**Grammatical (related to sentence structure, verb tenses, punctuation, etc._

_Please illustrate and correct (A) errors exhaustively except for (B) errors._

_**Severity and Frequency:** Take into the account the type and severity of the (A) mistakes. Mistakes that impede comprehension should be given more weight than minor ones. Also, note the frequency of the errors._

_For the lexical resource check, (A) is lexical, and (B) is grammatical. For the grammatical check, (A) is grammatical, and (B) is lexical._

**Figure 3** The incorporated prompt to the improved Python code

_You are a skilled IELTS examiner. Assess the provided essay for task 2 using the following criteria:_

_1. **Error Identification:** To assess "Lexical Resource" and "Grammatical Range and Accuracy", identify all lexical and grammatical errors thoroughly._

_2. **Classify Errors:** Categorize the errors as either:_

_- **Lexical:** Related to word choice, usage, and vocabulary range._

_- **Grammatical:** Related to sentence structure, verb tenses, punctuation, etc._

_3. **Severity and Frequency:** Take into account the type and severity of the errors. Mistakes that impede comprehension should be given more weight than minor ones. Also, note the frequency of these errors._

_4. **Lexical Resource Considerations:** Examine the range and appropriateness of the vocabulary:_

_- Is there a balanced use of both common and uncommon words?_

_- Are words used appropriately in context?_

_- Are there moments where vocabulary enhances the clarity or depth of the message?_

_Rate the "Lexical Resource" on a scale from 1 to 9, based on the IELTS task 2 criteria._

_5. **Grammatical Range Considerations:** Analyze the variety and complexity of sentence structures:_

_- Are different sentence structures (simple, compound, complex) employed?_

_- Is there proficiency in various grammatical constructs?_

_Rate the "Grammatical Range and Accuracy" on a scale from 1 to 9, in line with the IELTS task 2 criteria._

_6. Rate the "Task Response" on a scale from 1 to 9, in accordance with the IELTS task 2 criteria._

_7. Rate the "Coherence & Cohesion" on a scale from 1 to 9, in accordance with the IELTS task 2 criteria._

_Indicate inappropriate cohesive devices._

_8. Indicate an exhaustive error list with the judge if it is a lexical error or a grammatical error and their corrections._

_*Calculation*_

_1. Calculate the sum of the four criteria scores._

_2. Divide the sum by 2._

_3. If the decimal is 0.5, round to the bigger whole number._

*If the decimal is not 0.5, round to the nearest whole number.*

*4. Divide the whole number by 2 again. Don't round up this time. This number is the overall score.*

**Figure 4** Improved prompt for Web UI of ChatGPT

After calibrating through trial and modification, we created a prompt to aid us in giving feedback on the content without taking into account English proficiency. The prompt is presented below:

*After reading the following essay, as an ordinary reader, please give a comment on the impression of the content, including the used examples. An advocate position on the writer is desirable unless the message is immoral. Please call the writer "you."*

## 4 Results

### *ChatGPT scoring using original prompt*

At this analysis comprising 16 essays, it was found that the higher a human scorer rated an essay, the higher the ChatGPT score using the original prompt, as demonstrated in Figure 5. As reported earlier[5)], a slight variation in output was observed among 3 trials for each essay.

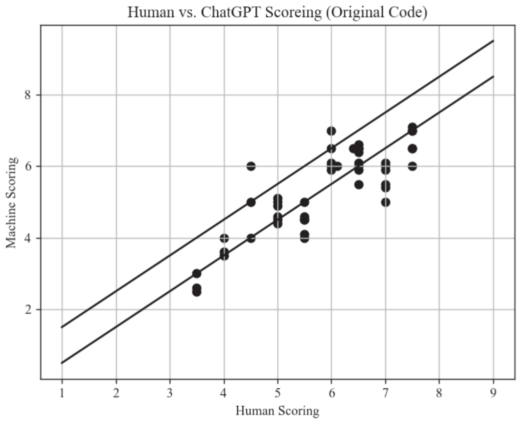


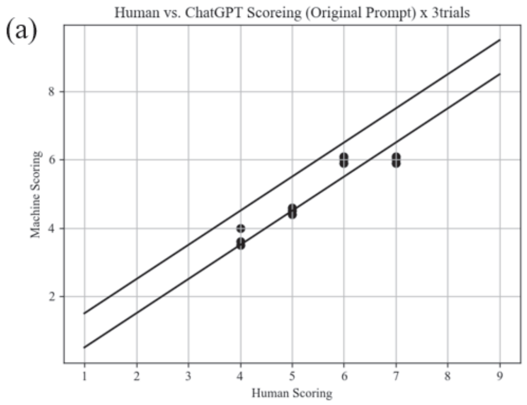
**Figure 5** Human vs. ChatGPT Scoring (Original Code) (16 samples x 3 output)

The graph illustrates the range of scores within 0.5 points of the human score on the IELTS test. Candidates may opt for re-scoring by paying a fee and may achieve slightly higher scores[18)]. If a score is changed through re-scoring, the fee for re-scoring is refunded. That means, 0.5 points, a minimum unit, can be an allowable margin of error through human scoring. Out of 48 data points (16 essays x 3 times), 30 scores fell within the 0.5 point range, while 18 scores fell outside of it. The furthest deviation was a score 2 points lower than its corresponding human evaluation.

### *Four-data analyses*

To investigate effective means of improvement, we conducted four data analyses. We selected essays with human scores of 7, 6, 5, and 4 points to determine trends across wide levels. Figure 6 displays the scores given by ChatGPT scoring using the original prompt for these four essays, which demonstrate a similar scoring tendency to that observed in the 16 samples (Figure 5).

As indicated in Figure 6, the trials involving the addition of "let's think step by step" at the end of the original prompt and asking for scoring criteria before inputting the original prompt did not result in improvement. The scores for both band score 7 essays were over half a point lower than human scoring, and their scoring trends remained similar to those of the original prompt (Figure 6a, b, c).



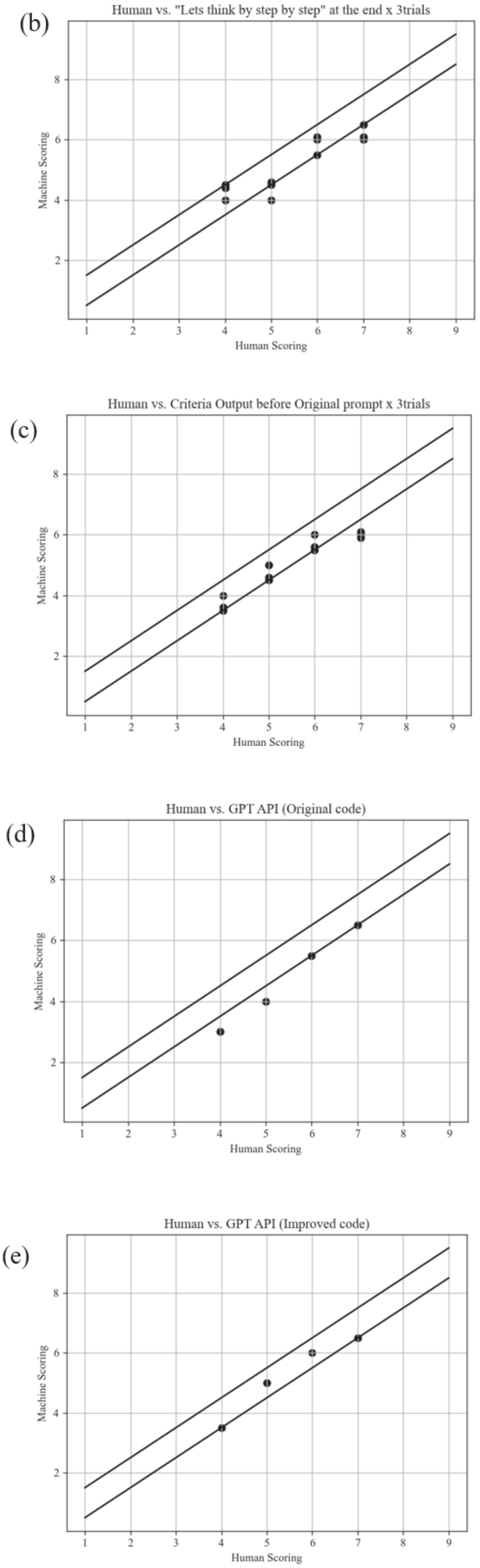
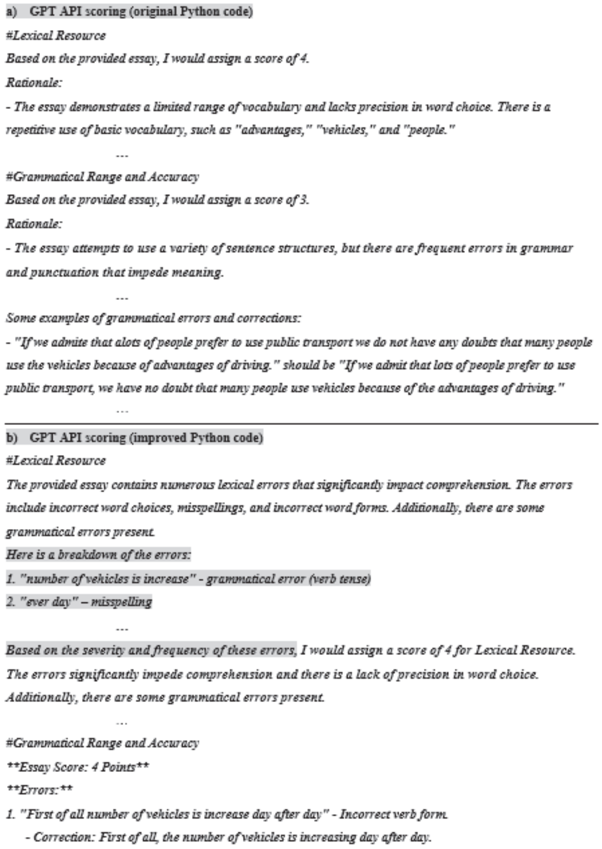


**Figure 6** Four-data Analyses

*Note: (a) Human scoring and ChatGPT scoring using the original prompt (4 samples), (b) Human scoring and the prompt with "let's think step by step" at the end, (c) Human scoring and the prompt after the preparation of criteria output, (d) Human scoring and GPT API Scoring with the original Python code, and (e) Human scoring and GPT API Scoring with the improved Python code.*

The criteria that ChatGPT answered were paraphrased from the publicly available ones[8], and they were accurate, except for the word limit, one of criteria for score 1. The result shows another approach was necessary.

We executed the original Python code that combined four criteria to gauge ChatGPT's ability without any parameters (Figure 6d). The results indicated lower scores for essays marked with a score of 4 and 5. Nonetheless, this experiment was advantageous as it detected a greater number of mistakes by systematically identifying errors based on the essay. Double-counting of errors was reduced in this experiment, yet some were still identified (Table 2). Additionally, based on our experiences as English teachers, it seems that careless mistakes have a greater impact on ChatGPT's scoring than on human scoring. Despite some reading difficulties, we are able to comprehend the primary message. While examining an essay, human graders typically identify errors as they read, classify them, and take their severity into account when determining the number of errors. Therefore, the code was enhanced to detect errors, categorize them, and assess their severities (Figure 6e). The improved Python code scores were within a range of 0.5 points for each essay. Examples of the output are depicted in Figure 7. The output of the improved Python code was clearer than the original code. Although double counting errors persisted, it occurred less frequently than in the original code. For instance, in the IELTS16 test3 (human scoring 7)[14], the inclusion of "cavy" instead of “cavity,” and “loads of bread” instead of “loaves of bread” are examples of words that were double counted as both lexical and grammatical errors.



**a) GPT API scoring (original Python code)**

*#Lexical Resource*

*Based on the provided essay, I would assign a score of 4.*

*Rationale:*

*- The essay demonstrates a limited range of vocabulary and lacks precision in word choice. There is a repetitive use of basic vocabulary, such as "advantages," "vehicles," and "people."*

...

*#Grammatical Range and Accuracy*

*Based on the provided essay, I would assign a score of 3.*

*Rationale:*

*- The essay attempts to use a variety of sentence structures, but there are frequent errors in grammar and punctuation that impede meaning.*

...

*Some examples of grammatical errors and corrections:*

*- "If we admite that alots of people prefer to use public transport we do not have any doubts that many people use the vehicles because of advantages of driving." should be "If we admit that lots of people prefer to use public transport, we have no doubt that many people use vehicles because of the advantages of driving."*

...

**b) GPT API scoring (improved Python code)**

*#Lexical Resource*

*The provided essay contains numerous lexical errors that significantly impact comprehension. The errors include incorrect word choices, misspellings, and incorrect word forms. Additionally, there are some grammatical errors present.*

*Here is a breakdown of the errors:*

*1. "number of vehicles is increase" - grammatical error (verb tense)*

*2. "ever day" – misspelling*

...

*Based on the severity and frequency of these errors, I would assign a score of 4 for Lexical Resource. The errors significantly impede comprehension and there is a lack of precision in word choice. Additionally, there are some grammatical errors present.*

...

*#Grammatical Range and Accuracy*

***Essay Score: 4 Points***

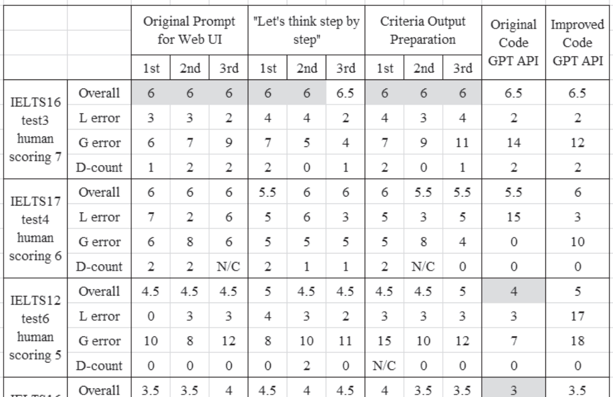
***Errors:***

*1. "First of all number of vehicles is increase day after day" - Incorrect verb form.*

*- Correction: First of all, the number of vehicles is increasing day after day.*

**Figure 7** Comparison of Output: Original Code vs. Improved Code (Excerpts)

**Table 2** Error Classification and double counted errors



| | | Original Prompt for Web UI | | | "Let's think step by step" | | | Criteria Output Preparation | | | Original Code GPT API | Improved Code GPT API |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | 1st | 2nd | 3rd | 1st | 2nd | 3rd | 1st | 2nd | 3rd | | |
| IELTS16 test3 human scoring 7 | Overall | 6 | 6 | 6 | 6 | 6 | 6.5 | 6 | 6 | 6 | 6.5 | 6.5 |
| | L error | 3 | 3 | 2 | 4 | 4 | 2 | 4 | 3 | 4 | 2 | 2 |
| | G error | 6 | 7 | 9 | 7 | 5 | 4 | 7 | 9 | 11 | 14 | 12 |
| | D-count | 1 | 2 | 2 | 2 | 0 | 1 | 2 | 0 | 1 | 2 | 2 |
| IELTS17 test4 human scoring 6 | Overall | 6 | 6 | 6 | 5.5 | 6 | 6 | 6 | 5.5 | 5.5 | 5.5 | 6 |
| | L error | 7 | 2 | 6 | 5 | 6 | 3 | 5 | 3 | 5 | 15 | 3 |
| | G error | 6 | 8 | 6 | 5 | 5 | 5 | 5 | 8 | 4 | 0 | 10 |
| | D-count | 2 | 2 | N/C | 2 | 1 | 1 | 2 | N/C | 0 | 0 | 0 |
| IELTS12 test6 human scoring 5 | Overall | 4.5 | 4.5 | 4.5 | 5 | 4.5 | 4.5 | 4.5 | 4.5 | 5 | 4 | 5 |
| | L error | 0 | 3 | 3 | 4 | 3 | 2 | 3 | 3 | 3 | 3 | 17 |
| | G error | 10 | 8 | 12 | 8 | 10 | 11 | 15 | 10 | 12 | 7 | 18 |
| | D-count | 0 | 0 | 0 | 0 | 2 | 0 | N/C | 0 | 0 | 0 | 0 |
| IELTS16 test4 human scoring 4 | Overall | 3.5 | 3.5 | 4 | 4.5 | 4 | 4.5 | 4 | 3.5 | 3.5 | 3 | 3.5 |
| | L error | 11 | 12 | 12 | 14 | 6 | 10 | 11 | 8 | 10 | 6 | 15 |
| | G error | 8 | 7 | 13 | 11 | 7 | 10 | 7 | 10 | 10 | 8 | 7 |
| | D-count | N/C | N/C | 1 | N/C | 0 | N/C | N/C | 0 | 1 | 2 | 0 |

*Note: Overall scores different from human scoring more than 0.5 points are highlighted. D-Count means “double count as both lexical and grammatical error”. N/C means “not classified”. In that case, we classified errors.*

***Expanding the improved Python code’s target to 16 Essays***

To evaluate the efficiency of the Python code previously discussed, an analysis was conducted on 16 essays (refer to Figure 8). Among the 16 essays, 14 were scored within 0.5 points of their human grading. The remaining two essays were scored one point higher than their human grading.

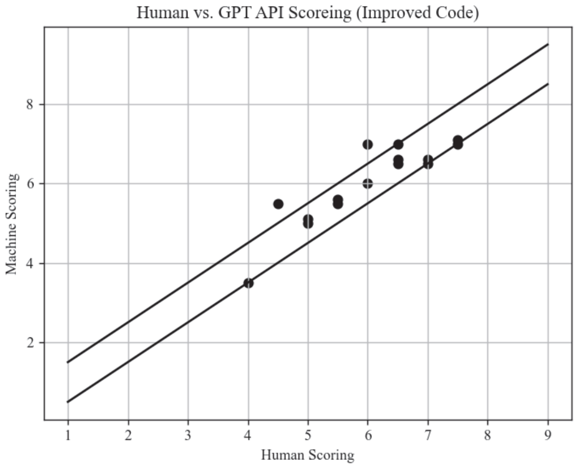


**Figure 8** Human vs. GPT API Scoring (Improved Code) (16 essays x 1 output)

***The improved prompt based on the improved Python code***

Considering the factors that enhanced the Python code’s output, we developed an improved prompt following our method. The environment for running a Python code embedded with GPT API could pose a challenge to English educators. Figure 9 displays a comparison between human scoring and ChatGPT scoring using the improved prompt on 16 sample essays. The plots in Figure 9 reveal a more robust correlation between human and machine scoring when compared to Figure 5. Of the 48 sets of data (16 essays x 3 times), 44 data points fell within the 0.5-point range, indicating that only four data points were outside that range.

The furthest point deviated 1.5 points from its human scoring. The process of four-data analysis served as a calibration to attain near human-like machine scoring by utilizing an improved prompt to match with human scoring. Nevertheless, the variance rate was higher than that of the improved Python code utilizing GPT API.

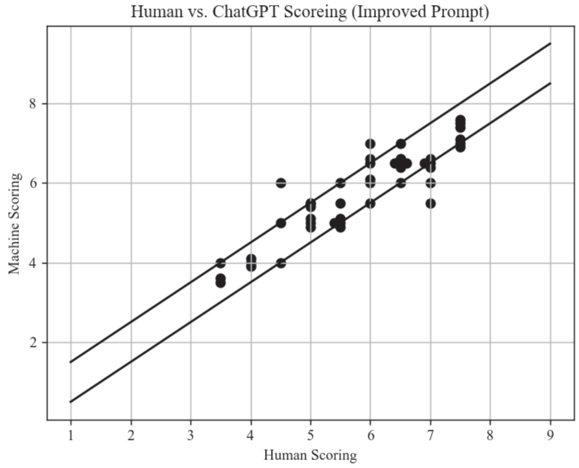


**Figure 9** Human vs. ChatGPT Scoring (Improved Prompt) (16 samples x 3 output)

***Calculation of overall score***

As mentioned in the introduction, ChatGPT does not work as a conventional calculator2)3)4). In the Web UI prompts, ChatGPT was assigned to compute the IELTS score. Nonetheless, we discovered that the outcome was sometimes inaccurate owing to the distinctive 0.5 decimal in the IELTS scoring structure. While ChatGPT demonstrated reliability in several instances, inaccuracies arose during the rounding up process. We rectified these errors as they were encountered, but certain miscalculations persisted despite a refinement in the Web UI prompt. Our preference is to execute computations with Python code to guarantee accuracy.

***The prompt to assist comments on the content***

This prompt described in the methods section was created to offer the writer encouragement as well as valuable feedback. Providing feedback on essay content can enhance learner motivation since students often dedicate ample time to construct effective plots and convey their experiences and thoughts through their essays. A Web UI prompt is appropriate for this purpose. ChatGPT finishes with generating its response while human examiners are reading the essay by themselves. The following is a sample response generated by ChatGPT (GPT-4) for IELTS17 test4 15).

*“Firstly, I'd like to commend you for addressing a very timely and relevant topic. The rising interest in alternative medicines and treatments is certainly a phenomenon seen globally, and you've made a valid attempt to capture both sides of the debate.*

*Your introduction provides a clear context by highlighting the ancient origins of herbal and natural treatments. However, it might be beneficial to delve a bit deeper into why people have historically turned to these remedies. This could provide a richer backdrop for your main points.*

*The examples you've used to illustrate the cons of alternative medicines, such as allergic reactions, were effective. You've rightly highlighted the potential dangers of self-medication without proper knowledge. However, while your point about environmental concerns regarding pharmaceutical production is interesting and valid, it seems a bit disconnected from the direct health implications you primarily focus on. You might consider adding an example or two about the therapeutic benefits of alternative medicines to give a fuller picture of their advantages.*

*Additionally, I noticed a few grammatical errors, such as "some people dead" (which should probably be "some people died") and "medicine which develops rocketly" (perhaps you meant "rapidly"). Proofreading can help make your argument more compelling.*

*Lastly, your conclusion is concise and clearly restates your stance on the issue. Overall, your essay paints a vivid picture of the debate on alternative vs. conventional medicine. It prompts readers to think about the choices they make and the potential repercussions of those decisions. Well done!"*

It seems enough to encourage the writer to try the next one. It is necessary for a teacher to check both the original essay and a generated comment with certainty. However, this system enables mass-scoring.

## 5 Discussion

### *Effect of modification*

According to Table 2, the improved Python code produced higher scores for low-level essays despite identifying more errors. This improvement may be attributed to accounting for severity. It is reasonable because the sample essays remained comprehensible to us, English instructors, despite having a number of errors.

### *Simplification of the modification of Python codes*

This time, the identification of errors and their classification was repeated in two criteria. However, it is possible to program to detect them at first, and refer to the classification later to simplify. For this trial, the Python codes were used to identify the causes of the disparity between human scoring and ChatGPT scoring. When it is applied to mass-scoring, it is desirable to simplify.

### *Limit of Web UI*

The Web interface of ChatGPT is user-friendly for all. Nevertheless, a token threshold is in place for both input and output[19]. Actually, a user can request ChatGPT to generate the following response. However, it appears that ChatGPT tries to summarize the answer within a one-time token, and its character count depends on the language.

### *Current Developments*

The prompt engineering is under development now. After collecting data, we noticed that the marking with ### or ``` is recommended for the target text by OpenAI[20] and the Prompt Engineer guide on GitHub[21]. Even though the target text lacked clear marking in our prompts, which was detected precisely by ChatGPT. It was clear because the detected errors were specific to the essay. Having said that, new findings may provide clues for further improving our prompts, leaving room for continued development.

### *Limit of this study*

Conducting this study on a substantial number of essays would be desirable; however, the availability of human-scored essays is restricted. Each book comprises four essays, some of which are sample essays without human scoring. To claim that the sample size is sufficient would be an overstatement. Nonetheless, IELTS score and CEFR are among the most dependable standards. Thus, this study can hold significance.

### *Application to English course at schools*

Correction of essays presents challenges for educators as it is a time-consuming task that must be done outside of class. Therefore, requiring students to submit essays every lesson is difficult. Many instructors opt for assigning a final essay or report instead of a final exam, and provide only a score without additional feedback. Objectively scoring content while reading it can also prove to be challenging. For the public exam, scoring criteria were pre-determined, allowing for ease in rating essays. Without objective criteria, it takes time to adjust criteria to student levels as assessment preparation. In these cases, ChatGPT’s objective scoring system is especially helpful. Scores are based on CEFR, providing relevant milestones for students and helping examiners to identify oversights. Additionally, the system enables frequent essay scoring. In fact, AES is used with human scoring in the Graduate Record Examination (GRE)5)22) in the US and Test of English as a Foreign Language (TOEFL)5)23).

In a classroom setting, students typically compose essays on a particular subject. To eliminate redundancy, instructors can employ ChatGPT's Web UI custom instructions to avoid repeated input of the topic.

***Elevating motivation***

Students describe their experiences and thoughts in their essays, presenting their viewpoints with supporting evidence. It is ideal for teachers to develop a closer relationship with each student by offering comments on the content in a classroom setting. We need to read their essays directly, but it is possible to enlist the assistance of ChatGPT to generate personalized feedback. This will decrease our workload.

This prompt encourages students to aspire to excel in their work. It is suggested to respond to the content in a conversational manner to establish a positive rapport. Acknowledging the quality of the content may motivate students to write even better essays in the future, creating a positive cycle.

## 6 Conclusion

It is factual that ChatGPT's scores were somewhat correlated with those given by humans. However, ChatGPT's IELTS sample essay scores were largely lower than those provided by humans. Moreover, occasionally, the disparity from human scores exceeded the acceptable margin of error of plus or minus 0.5 points. Attempts to integrate "let's think step by step" and requesting scoring criteria prior to scoring were unproductive in aligning scores. To remove parameter effects, we developed a Python program utilizing the GPT API (GPT-3.5 Turbo). We found that inaccuracies in lexical resources and grammatical accuracy were penalized more heavily than by human graders. To improve the program, we added error detection and classification while taking into account their severity. This led to greater alignment with human-based ratings. For the Web-based user interface, we developed an improved prompt that exhibits a stronger correlation with human scoring compared to the original one. Notably, the output of the improved Python code was superior. In addition, we designed a prompt to provide individualized feedback aimed at motivating learners to compensate for the lack of human impression.

## Acknowledgement

We would like to express our gratitude to Stewart Laird for his guidance on the IELTS examination.

## Reference

1) Open AI (2022) *Introducing ChatGPT.* https://openai.com/blog/chatgpt　（アクセス日　2023.9.23）
2) Wei J., et al. (2022) Chain of thought prompting elicits reasoning in large language models. https://arxiv.org/aba/2201.11903　（アクセス日　2023.9.25）
3) Kojima T. et al. (2022) Large language models are zero-shot reasoners. *36th Conference on Neral Information Proceeding Systems (NeurIPS 2022).* https://proceedings.neurips.cc/paper_files/paper/2022/file/8bb0d291acd4acf06ef112099c16f326-Paper-Conference.pdf（アクセス日　2023.9.28）
4) Wolfram S. (2023) *What is ChatGPT Doing…and Why Does It Work?* Wolfram Media Inc.
5) Mizumoto A. & Eguchi M. (2023) Exploring the potentials of using an AI language model for automated essay scoring. *Research Methods in Applied Linguistics*, *2* (2023) 100050
6) British Council. (2023) *Understanding and explaining*

*IELTS scores*. https://takeielts.britishcouncil.org/teach-ielts/test-information/ielts-scores-explained （アクセス日 2023.10.1）

7) Council of Europe. (2001) *Common European Framework of Reference for Languages: Learning, Teaching and Assessment*. Council of Europe.

8) British Council. (2023) *IELTS Writing Band Descriptors.* https://takeielts.britishcouncil.org/sites/default/files/ielts_writing_band_descriptors.pdf （アクセス日 2023.10.8）

9) Cambridge English. (2015) *IELTS 10 Academic authentic examination papers with answers*. Cambridge University Press and UCLES.

10) Cambridge English. (2016) *IELTS 11 Academic authentic examination papers with answers*. Cambridge University Press and UCLES.

11) Cambridge English. (2017) *IELTS 12 Academic authentic examination papers with answers*. Cambridge University Press and UCLES.

12) Cambridge English. (2019) *IELTS 14 Academic authentic examination papers with answers*. Cambridge University Press and UCLES.

13) Cambridge English. (2020) *IELTS 15 Academic authentic examination papers with answers*. Cambridge University Press and Cambridge Assessment 2020.

14) Cambridge English. (2021) *IELTS 16 Academic authentic examination papers with answers*. Cambridge University Press and Cambridge Assessment 2021.

15) Cambridge English. (2022) *IELTS 17 Academic authentic examination papers with answers*. Cambridge University Press and Cambridge Assessment 2022.

16) Open AI. (2023) *API reference.* https://platform.openai.com/docs/api-reference/introduction （アクセス日 2023.9.23）

17) jazzg (2023, April 19). *Web chat default temperature for GPT 3.5 and 4* [Online forum post]. Open AI. Developer Forum. https://community.openai.com/t/web-chat-default-temperature-for-gpt-3-5-and-4/167356 （アクセス日 2023.10.4）

18) IELTS. (2023) *Getting and sharing your results.* https://www.ielts.org/for-test-takers/results （アクセス日 2023.10.1）

19) Makvana M. (2023, July 10) *What is the character limit for ChatGPT? - Let's find out together.* How -To Geek https://www.howtogeek.com/895929/what-is-the-character-limit-for-chatgpt/（アクセス日 2023.10.5）

20) OpenAI. (2023) *Prompt engineering.* https://help.openai.com/en/collections/3675942-prompt-engineering （アクセス日 2023.10.3）

21) DAIR.AI (2023) *Prompt Engineering Guide.* https://www.promptingguide.ai/ （アクセス日 2023.10.4）

22) Educational Testing Service (ETS). (2023) *The GRE Test is the world's most widely used admissions test for graduate & professional school.* https://www.ets. org/gre.html （アクセス日 2023.10.8）

23) Educational Testing Service (ETS). (2023) *The TOEFL tests prepare students for university study, immigration and more.* https://www.ets.org/toefl.html （アクセス日 2023.10.8）

# 大規模言語モデルChatGPTとGPT APIを活用した英文エッセー機械採点

西原玲奈　上田圭介　清永克己

大規模言語モデルは英語学習に活用できると話題になっている。学習者に注目されがちであるが、教員にとっても然りである。これまでにOpenAI社のChatGPTは英文エッセー採点ができると報告されている[1)] 。本研究では、IELTS公式問題集巻末のサンプルエッセーを用いて掲載されている人間採点と比較し、ChatGPT(GPT4.0)の英文エッセー採点能力を評価した。Web ユーザーインターフェース(GPT4.0)の出力は、すでに報告されているように[1)] 相関はあるものの、ばらつきがあった。そこで、パラメーターの影響を排除するため、GPT APT(GPT3.5-Turbo)を組み込んだPythonプログラムを開発した。このプログラムではIELTSの4項目それぞれを判定する4ループを繋げ、最後に総合点を算出する。その結果、ChatGPTが人間採点より低く採点する傾向が明らかになった。筆者らはChatGPTのフィードバックで採点だけでなく誤りも指摘するように指示していたため、一部の誤りを語彙力と文法でダブルカウントしていること、また、他の英語試験の試験官を務めた経験を元に人間採点者に比べ個々の誤りを一様に深刻に捉えられていることに気づいた。そこで、語彙力ループと文法ループでは最初にミスを検出し、誤りの分類と程度を判定した後に採点基準をチェックするように、プログラムに組み込むプロンプトを改良した。この改良Pythonプログラムでは人間採点にかなり近い採点出力を得た。このことに基づき、使い勝手を考慮し、Web ユーザーインターフェース用の改良プロンプトを考案した。また、そのほかに採点だけではなく、学習者と良好な関係を築くため、英語能力とは関係なくエッセーの内容に対するフィードバックを出力するプロンプトも考案した。このようにChatGPTの助けを借りれば、従来、英語教員の負担であったエッセーフィードバックを軽減し、授業での実施頻度を高めることができ、また学習者にとっても自習環境が整うであろう。