

# Beyond the Limits of AI Translation / Automated Evaluation

Kenji Kawauchi<sup>1)</sup>

Takae Saito<sup>2)</sup>

<sup>1)</sup>Department of Liberal Arts and Sciences, Faculty of Engineering, Maebashi Institute of Technology

<sup>2)</sup>Center for Liberal Arts, Utsunomiya Campus, Teikyo University

## Abstract

The aim of this paper is to elucidate, in a nutshell, what Generative AIs can and cannot do. As a starting point for an AI-based pedagogy in the field of English education, we verified the effectiveness of GPT-3.5, GPT-4, and Gemini concerning Discourse, Linguistic, and Formulaic Competences. The results revealed that while each tool is not a panacea for everything, the GPTs are very useful for all, except for one category of discourse competence.

## 1. Introduction

On November 30, 2022, the launch of ChatGPT-3.5 sent shockwaves around the world with enhanced conversational abilities, comprehension, contextual understanding and so on. It learns language patterns from enormous text data, enabling it to generate more sophisticated sentences and responses.

While AI-based pedagogy using AI translation (e.g. DeepL Translate) and AI automatic evaluation systems (e.g. QuillBot) available on Apps and websites are still in their embryonic stages, the appearance made a significant impact on the field of English language education in Japan, with Gally (2022) stating that it even heralds a “paradigm shift in English education.[1]”

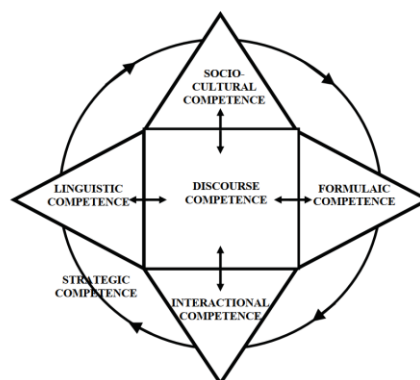
In the midst of a paradigm shift, attempts are being made to use AI translation in university English education (e.g., Yukishige et.al., 2022).[2] Regarding generative AI, Saito (2023) demonstrated the effective utilization of Generative AIs in English Language education.[3] Kawauchi & Saito (2023) elucidated what AI translation and automatic

evaluation systems can and cannot do, verifying the effectiveness of them. [4]

In this paper, we aim to clarify the “capabilities” and “limitations” using generative AIs (GPT-3.5, GPT-4, and Gemini) not employed in Kawauchi & Saito (2023)

## 2. Previous study

Kawauchi & Saito (2023) explored the potential for English sentence correction by AI translations and AI automatic evaluation systems, regarding three competencies (Discourse Competence, Linguistic Competence, Formulaic Competence) based on the model called “Communicative competence for language teachers” proposed by Celce-Murcia (2007) [5], as shown in Figure 1.



**Figure 1.** Communicative competence for language teachers (Celce-Murcia 2007: 45)

In their work, for AI translation, they used DeepL Translate and Google Translate, which were commonly used by Japanese university students learning English. Regarding AI evaluation systems, they verified using QuillBot and Grammarly. For convenience, they provide a summary of each verification item, verification content, and the systems used below:

**Table 1.** The summary of each verification item, verification content, and the systems

Verification item	Verification content	AI translation / evaluation systems
Discourse Competence	<ul style="list-style-type: none"> <li>• cohesion</li> <li>• coherence</li> <li>• discourse marker</li> </ul>	<u>Translation</u> <ul style="list-style-type: none"> <li>• DeepL Translate</li> <li>• Google Translate</li> </ul>
Linguistic Competence	<ul style="list-style-type: none"> <li>• vocabulary</li> <li>• grammatical accuracy</li> </ul>	<u>Automatic evaluation systems</u> <ul style="list-style-type: none"> <li>• QuillBot (Grammar Checker)</li> </ul>
Formulaic Competence	<ul style="list-style-type: none"> <li>• collocation</li> <li>• idiom</li> </ul>	<ul style="list-style-type: none"> <li>• Grammarly</li> </ul>

Text in (1) which is used in their study consists of 147 words (Henceforth *the text*). The first paragraph obtained from the web site is grammatically correct and forms meaningful sentences.<sup>2</sup> The remaining sentences are randomly collected from websites. Therefore, it is sentences that have no meaning and lack any context as a piece of writing. The Lexile measure of *the text* is 610L-800L. The CEFR vocabulary level using the CEFR-based Vocabulary Level Analyzer (ver. 2.0) is B1.1.

(1)

Despite speeds of up to 186 mph, the view through the window of the fast train is a natural match with these wide empty horizons where the landscape never seems blurred or cluttered.

We entered a richly decorated room. The doctor ordered him to take regular exercise. The Titanic sank on its maiden voyage. He was writhing on the ground in excruciating pain. Invading that country was an utterly stupid thing to do. I'd like to buy two bars of soap, please. I always try to do my homework in the morning,

<sup>1</sup> The evaluations of the AI translation and AI automated assessment systems presented here were conducted from March to early May 2023. It is noteworthy that the rapid development of these systems may result in different outputs at present, so readers are advised to bear this in mind.

<sup>2</sup> The first paragraph in *the text* is from 'Welcome to the train in Spain' (<https://www.iol.co.za/travel/welcome-to-the-train-in-spain-1414014>).

after making my bed. He has been asked to give a presentation about his work. At first her eyes filled with horror, and then she burst into tears. Their behavior was enough to drive anybody to crime. Mary whispered softly into John's ear. She placed her keys gently on the table and sat down.

Through their validation test, they summarized as shown in Table 2:

**Table 2.** Summary of the verification results

Verification item	Verification content	AI translation / evaluation systems			
		DeepL	Google	QuillBot	Grammarly
Discourse Competence	<ul style="list-style-type: none"> <li>• cohesion</li> <li>• coherence</li> <li>• discourse marker</li> </ul>	Bad	Bad	Bad	Bad
Linguistic Competence	<ul style="list-style-type: none"> <li>• vocabulary</li> <li>• grammatical accuracy</li> </ul>	Bad	Bad	Good	Good
Formulaic Competence	<ul style="list-style-type: none"> <li>• collocation</li> <li>• idiom</li> </ul>	Bad	Bad	Bad	Bad

No systems were at a practical level regarding discourse competence and formulaic competence. In terms of evaluating linguistic competence, the results are much better than those for others.

### 3. Methods

#### 3.1 Generative AIs

In this study, we used GPT-3.5, GPT-4, and Gemini to verify each verification item and content which are employed in the Kawauchi & Saito (2023)'s study as described in the previous section.

#### 3.2 Material

The same *the text* as in the section 2 was used to validate GPT-3.5, GPT-4, and Gemini.

### 4. Discourse Competence

The relation between two or more sentences can be divided into two categories: cohesion and coherence. Both relate to the structure of the sentences and are common in that they are concepts that guarantee the continuity of the sentences. Cohesion emphasizes the continuity of meaning between sentences, whereas coherence refers to the consistency of meaning throughout

the entire text (de Beaugrande & Dressler 1981:3-7).[6] Following Fox Tree (2010) [7], discourse marker is defined as below:

*Discourse markers* are often expressed in natural conversation and seem to refer to expressions that are not directly related to the content of the conversation but are necessary to facilitate the conversation.

Using GPT-3.5, GPT-4, and Gemini, we verify whether generative AIs can assess discourse competence such as cohesion, coherence, and discourse marker.<sup>3</sup> The text to be evaluated is *the text* in (1). This time, we gave the following instructions to the generative AIs with a sentence like a request to a person:<sup>4</sup>

- (2) a. Does the following text have cohesion?
- b. Does the following text have coherence?
- c. Are discourse markers used appropriately?

## 4.1 Validation I

### 4.1.1 Cohesion

The responses below are obtained from GPT-3.5, GPT-4, and Gemini for the question (2a).

#### (3) GPT-3.5

No, the text does not have cohesion. Cohesion refers to the logical connections between sentences or paragraphs that create a smooth flow of ideas. The provided text consists of disjointed sentences that do not relate to each other or form a coherent narrative. Each sentence is independent and does not build upon the previous one, resulting in a lack of cohesion.

#### (4) GPT-4

The text you provided does not exhibit cohesion. It comprises a series of unrelated sentences that do not logically connect or flow from one to the next. Cohesion in text involves the use of linking words, references, and other mechanisms to ensure that the different parts of the text connect meaningfully. In this case, each sentence appears to address a completely different subject or scenario without any clear thematic or narrative linkage between them, disrupting the sense of a coherent and unified text.

#### (5) Gemini

No, the text lacks cohesion. Cohesion refers to how sentences are linked together to create a smooth flow of ideas. This text presents several unrelated sentences. Here's what's missing for cohesion:

**Topic:** There's no clear main topic connecting the sentences. They jump from trains to rooms to historical events and personal anecdotes.

**Transitions:** The sentences lack transitions that show how they relate to each other. Words like "however," "because," or "afterwards" could help create a logical flow.

**Reference:** Some pronouns ("him," "her") lack clear antecedents (the word they refer back to) making it unclear who or what they're referring to.

For the text to be cohesive, it would need to focus on a single topic or a sequence of related events. Transitions and clear references would be needed to tie the sentences together.

### 4.1.2 The Assessment of Cohesion

Cohesion is morphological or meaningful continuity that connects sentences. GPT-3.5 does not refer to cohesion but gives feedback on coherence. In other words, this shows a failure to distinguish between cohesion and coherence. GPT-4 as well as GPT-3.5 conflate them. Furthermore, it encourages the use of discourse marker without mentioning cohesion. However, it abstractly gives feedback about cohesion, as underlined in (4). "Reference" means, for example, an agreement between an antecedent and a pronoun over two sentences. Therefore, GPT-4's response is not explicit. Gemini, like other generative AIs, does not seem to be able to distinguish between cohesion and coherence. However, a significant difference is seen in Gemini. It differs from GPT-4 in that it provides specific examples and explanations of "Reference,"

<sup>3</sup> We tested without modifying the temperature setting or fine-tuning of GPT3.5 and GPT4 in subsequent verifications.

<sup>4</sup> The following verifications took place on March 4, 2024.

as underlined in (5). Considering these facts, it can be concluded that Gemini is somewhat superior in terms of cohesion.

## 4.2 Validation II

### 4.2.1 Coherence

Here are the responses obtained from GPT-3.5, GPT-4, and Gemini for the question (2b).

#### (6) GPT-3.5

The provided text lacks coherence. It appears to be a collection of unrelated sentences that do not flow logically from one to the next. Each sentence seems to stand alone without any clear connection to the preceding or following sentences. To improve coherence, the text should be revised to create a logical sequence of ideas or events that follow a clear narrative or thematic thread.

#### (7) GPT-4

The text provided lacks coherence. It is a series of unrelated sentences that do not build on each other to form a cohesive argument or narrative. Each sentence appears to address a different topic, without any apparent connection or logical progression from one to the next. For a text to be coherent, it should have a clear and logical flow of ideas where each part contributes to the development of a central theme or argument, which is not evident in this text.

#### (8) Gemini

No, the text lacks coherence. Here's why:

- Coherence refers to the ideas in a text being logically connected and making sense together. This text presents a series of unconnected sentences about various topics.
- It jumps from a train journey to an unrelated scene in a room, then mentions historical facts and personal actions, all without any clear connection.

For a text to be coherent, the sentences should flow logically from one another, building a unified idea or story.

### 4.2.2 The Assessment of Coherence

As readers are aware, *the text* has no meaningful connection and holds no significance as a whole. All generative AIs correctly point out the lack of coherence in *the text*.

Just to be sure, let's use a coherent text to verify the generative AIs' competence. We evaluate them using the beginning of Yasunari Kawabata's novel "Snow Country" translated by

Edward George Seidensticker.[8]

#### (9) *Snow Country*

The train came out of the long tunnel into the snow country. The earth lay white under the night sky. The train pulled up at a signal stop. A girl who had been sitting on the other side of the car came over and opened the window in front of Shimamura. The snowy cold poured in. Leaning far out the window, the girl called to the station master as though he were a great distance away. The station master walked slowly over the snow, a lantern in his hand. His face was buried to the nose in a muffler, and the flaps of his cap were turned down over his face.

#### (10) GPT-3.5

Yes, the text demonstrates coherence. Coherence refers to the smooth flow and logical connection of ideas within a text. In this passage, the events progress logically from the train emerging from a tunnel into snow country, to a girl opening a window, to her interaction with the station master. The descriptions of the setting and characters contribute to a unified scene, enhancing the overall coherence of the passage.

#### (11) GPT-4

The text you provided does have coherence. It presents a series of events that are logically connected and create a clear narrative scene. Each sentence follows from the one before in a way that makes sense, contributing to a unified description of a scene at a train station in the snow country. The actions and descriptions are logically connected, painting a vivid picture that the reader can follow and understand. The characters' actions and the setting details are consistent and contribute to the overall mood and imagery, aiding the text's coherence.

#### (12) Gemini

Yes, the text has coherence. Here's why:

- Logical Flow: The passage starts with the train entering a snowy landscape, setting the scene. It then focuses on a specific action – a girl opening the window. This action leads to the interaction between the girl and the station master, which is the main focus of the excerpt.
- Transition Words: Though not abundant, the text uses "The" effectively to connect sentences and maintain a clear flow.
- Cause and Effect: The girl opening the window allows her to call the station master, who then responds by walking over. Overall, the passage provides a clear picture of a train arriving at a snowy station and the interaction between a passenger and the station master.

As can be seen from the underlined part, the generative AIs can appropriately evaluate a text

with coherence.

### 4.3 Validation III

#### 4.3.1 Discourse marker

Now we show the outputs for the discourse marker.

##### (13) GPT-3.5

The text lacks appropriate discourse markers to connect the sentences and create coherence. Discourse markers are words or phrases that help guide the reader through the text, indicating relationships between ideas or signaling transitions between different parts of the discourse.

In the provided text, there are no explicit discourse markers such as "however," "therefore," "meanwhile," etc., which would help to link the sentences more effectively and enhance the overall flow of the text. Incorporating appropriate discourse markers would improve the coherence and readability of the passage.

##### (14) GPT-4

The passage you provided contains a series of sentences that seem unrelated to each other, suggesting that discourse markers are either missing or not used effectively to connect the ideas and events described. Discourse markers are words or phrases used to manage the flow of discourse, such as indicating sequence (e.g., "first," "then"), contrast (e.g., "however," "on the other hand"), or cause and effect (e.g., "because," "therefore").

In the provided text, the sentences are presented without clear logical or thematic connections, making the passage feel disjointed: There is no apparent logical or thematic connection between sentences like "Despite speeds of up to 186 mph..." and "We entered a richly decorated room."

The use of "Despite" at the beginning suggests a contrast that is never completed within the context of that sentence or linked to subsequent information.

The sentences appear as isolated statements without transitions or markers that would indicate their relationship to one another, whether in terms of time, cause and effect, contrast, or elaboration.

Effective use of discourse markers would involve integrating these sentences in a way that clarifies their relationships, providing a coherent narrative or argument. As it stands, the passage lacks this coherence, indicating inappropriate or

##### (15) Gemini

The passage doesn't use many discourse markers, and the ones used aren't strictly necessary. Here's a breakdown: Despite: This one functions well. It indicates a contrast between the train's speed and the clarity of the view. However, the other sentences could stand alone without the following markers:

- We entered a richly decorated room.  
(This is a simple statement)
- The doctor ordered him to take regular exercise.

- The Titanic sank on its maiden voyage.  
(This is a historical fact)
- He was writhing on the ground in excruciating pain.  
(This is a vivid description)
- Invading that country was an utterly stupid thing to do.  
(This is a strong opinion)
- I'd like to buy two bars of soap please.  
(This is a polite request)
- I always try to do my homework in the morning, after making my bed. (This is a routine)
- He has been asked to give a presentation about his work. (This is a passive voice statement)
- At first her eyes filled with horror, and then she burst into tears. (This describes a sequence of events, but a comma could work here too)
- Their behavior was enough to drive anybody to crime.  
(This is a cause-and-effect statement)
- Mary whispered softly in John's ear.  
(This is a simple action)
- She placed her keys gently on the table and sat down.  
(This describes two consecutive actions, a comma could work here too)

Discourse markers can be helpful for emphasizing transitions, adding nuance, or structuring complex ideas. However, in this case, the sentences are clear and concise on their own. Here are some examples of where discourse markers could be effective:

In contrast to the chaos outside, the room was peaceful. (This emphasizes the difference between two situations)

Furthermore, research suggests regular exercise can improve mental health. (This adds information to a previous statement)

However, some historians believe the sinking of the Titanic could have been prevented. (This introduces a counterpoint)

Overall, the passage doesn't misuse discourse markers, but they aren't necessary for clarity in these sentences.

#### 4.3.2 The Assessment

All generative AIs properly pointed out the lack of discourse markers in *the text*. Furthermore, they carefully explain examples of discourse markers, and the appropriate feedback is provided. Based on these results, it can be concluded that each generative AI has the capability to distinguish discourse markers.

## 5. Linguistic Competence

Linguistic competence comprises vocabulary and grammatical knowledge. They are subcategorized as syntax, morphology, lexicon, phonology, and orthography. The text used in this section intentionally employs sentences from *the text* that have been altered to be ungrammatical. It contains

a total of ten grammatical and morphological mistakes. The bolded words and phrases represent that they are changed from *the text*, and strikethroughs indicate letters or words that are deleted from *the text*.

(16) *The text* with wrong words and phrases

Despite speeds of up to 186 mph, the view through the window **off** the fast train **are** a natural match with these wide empty horizons where the landscape never seems blurred or cluttered.

We entered **a** richly decorated room. The doctor ordered him to take regular exercise. The Titanic sank on its maiden voyage. He was writhing on the ground in excruciating pain. Invading that country was an utterly stupid thing to do. I'd like to buy two bars of soap, please. I always try to do my homework in **the** morning, after **making** my bed. He has been asked **about** give a presentation about his work. At first her eyes filled with horror, and then she burst into tears. Their behavior was enough to drive anybody to crime. Mary whispered softly **with** John's ear. She placed **on the table her keys gently** and sat down.

5.1 Validation

Here, we verify how many mistakes each generative AI can detect and whether they can correct them accurately. The prompt we used is “How many errors are there in the following sentences? Also, please correct and itemize the mistakes.” We obtained following the responses.

(17) GPT-3.5

1. Despite the speed of up to 186 mph, the view through the window in the fast train is a natural match with these wide empty horizons where the landscape never seems blurred or cluttered. ((a)t-verb agreement error: “view” should be singular, and (b)“clutter” should be in the past tense)
2. We entered a richly decorated room. ((c)missing indefinite article)
3. I always try to do my homework in the morning, after making my bed. ((d)Missing definite article and (e)gerund form of the verb “make”)
4. He has been asked (f)to give a presentation about his work. (Missing preposition)
5. Mary whispered softly into John's ear. ((g)Preposition error: “with” should be “into”)
6. She gently placed her keys on the table and sat down. ((h)Word order correction)

(18) GPT-4

1. Despite speeds of up to 186 mph, the view through the window of the fast train is a natural match with these wide empty horizons where the landscape never seems blurred or cluttered.  
**Mistakes:** ①“speed” should be “speeds.”  
 ②“off” should be “of.” ③“are” should be “is.”  
 ④“clutter” should be “cluttered.”
2. We entered a richly decorated room.  
**Mistakes:** ⑤“a” is missing before “richly decorated room.”
3. I always try to do my homework in the morning, after making my bed.  
**Mistakes:** ⑥“the” is missing before “morning.”  
 ⑦“make” should be “making.”
4. He has been asked to give a presentation about his work.  
**Mistakes:** ⑧“about” should be “to.”
5. Mary whispered softly into John's ear.  
**Mistakes:** ⑨“with” should be “into.”
6. She gently placed her keys on the table and sat down.  
**Mistakes:** The correct placement and order should be ⑩“She gently placed her keys on the table.”

(19) Gemini

1. Despite speed of up to 186 mph, the view through the window off the fast train are a natural match...  
**Mistake:** Subject-verb agreement. “Speed” is singular, so (a)the verb should be “is” instead of “are.”
2. We entered richly decorated room.  
**Mistake:** Missing article. (b)“Room” needs an article (a/an/the) in most cases.
3. He has been asked about give a presentation about his work.  
**Mistake:** (c)Missing infinitive marker “to” before “give.”
4. Mary whispered softly with John's ear.  
**Mistake:** (d)Preposition error. You whisper “to” someone, not “with” their body part.

5.2 The Assessment

GPT-3.5 judged that the wrong text has six errors, but it detected eight errors indeed. The detected errors are correctly modified. It is a high degree of error detection. GPT-4 detected all mistakes and corrected all mistakes correctly. On the other hand, Gemini only pointed out four out of ten errors and offered correct alternative suggestions for them. Therefore, considering the detection and correction capability of linguistic competence, we can conclude that the more practical system is GPT-4.

## 6. Formulaic Competence

Formulaic competence refers, roughly say, to fixed phrases, collocations, and idioms. In the next subsection, we will verify them using following text:

(20)

Despite speeds of up to 186 mph, the view through the window of the **quick** train is a natural match with these wide empty horizons where the landscape never seems blurred or cluttered. We entered a **magnificently** decorated room. The doctor ordered him to take **scheduled** exercise. The Titanic sank on its maiden voyage. He was writhing on the ground in **oppressive** pain. Invading that country was a **fully** stupid thing to do. I'd like to buy two **sticks** of soap please. I always try to study my homework in the morning, after **fixing** my bed. He has been asked to **play** a presentation about his work. At first her eyes filled with horror, and then she burst **with** tears. Their behavior was enough to **run** anybody to crime. Mary whispered sweetly in John's ear. She placed her keys softly on the table and sat down.

The text (20), based on (1), includes ten errors concerning idioms and collocations. The bolded words are replaced with words that are closer in meaning to the correct words but are unnatural.

### 6.1 Validation

The prompt we used here is “How many errors are there in the following sentences, concerning idioms and collocations? Also, please correct and itemize the mistakes.” We obtained following the responses.

(21) GPT-3.5

There are several errors concerning idioms and collocations in the sentences provided:

1. “quick train” should be “fast train” to maintain correct collocation.
2. “scheduled exercise” should be “regular exercise” to use the appropriate idiom.
3. “fully stupid” is not a common collocation; it should be “completely stupid.”
4. “sticks of soap” should be “bars of soap” for correct idiom usage.
5. “fixing my bed” should be “making my bed” for idiomatic accuracy.
6. “play a presentation” should be “give a presentation” to use the correct idiom.
7. “burst with tears” should be “burst into tears” for proper collocation.
8. “run anybody to crime” should be “drive anybody to crime” for idiomatic accuracy.

9. “whispered sweetly” is not a common collocation; it should be “whispered softly” or simply “whispered.”
10. (a) “placed her keys softly” should be “set her keys softly” for correct collocation.

(22) GPT-4

Here are the idiomatic and collocational errors in the sentences provided, along with the corrections:

1. (a) “a natural match with” should be “perfectly complements”
2. “take scheduled exercise” should be “take regular exercise”
3. “writhing on the ground in oppressive pain” should be “writhing on the ground in excruciating pain” -
4. “fully stupid” should be “completely stupid” or “utterly stupid”
5. “two sticks of soap” should be “two bars of soap”
6. “study my homework” should be “do my homework”
7. “fixing my bed” should be “making my bed”
8. “play a presentation” should be “give a presentation”
9. “burst with tears” should be “burst into tears”
10. “run anybody to crime” should be “drive anybody to crime”

(23) Gemini

There are 4 errors concerning idioms and collocations in the sentences:

1. “Quick train”: This is not a common way to refer to a fast train. A better option would be “high-speed train” or “express train.”
2. “Natural match”: While grammatically correct, “natural match” doesn't quite capture the intended meaning in this context. A better option would be “perfectly complements” or (a) “harmonizes beautifully with.”
3. “Fully stupid”: “Fully” is redundant here. “Stupid” is already a strong word. Simply saying “stupid thing to do” is sufficient.
4. “Play a presentation”: You don't “play” a presentation. You could say “give a presentation” or “deliver a presentation.”

### 6.2 The Assessment

GPT-3.5 detected ten errors, but a suggestion like (21a) is a markedly unnatural collocation. Therefore, the error detection and correction capability of GPT-3.5 is nine out of ten.<sup>5</sup> GPT-4 also detected ten errors, but, unfortunately, it was not able to detect the unnatural phrase “quick train.” Additionally, the proposed phrase (22a) is unacceptable. This is because, if replaced with that phrase, the sentence would be ungrammatical.

<sup>5</sup> The proposed phrase “completely stupid” is different from the original sentence but does not count as an error since it is collocatable.

Therefore, the error detection and correction capability of GPT-4 is eight out of ten. Gemini had a notably low detection capability, detecting only four errors, but it corrected them accurately. It is noteworthy that a collocation (23a), which is not suggested by GPT-3.5 and GPT-4, is proposed and is an acceptable expression. These results indicate that GPT-4 can assess linguistic competence better than GPT-3.5, although Gemini is unreliable.

## 7. Conclusion and Remarks

We have explored the potential for detection and correction by generative AIs, regarding three competencies (Discourse Competence, Linguistic Competence, Formulaic Competence). The results can be summarized as follows:

**Table 3.** Summary of the verification results

Verification item	Verification content	Generative AI		
		GPT3.5	GPT4	Gemini
Discourse Competence	● cohesion	Bad	Bad	Neutral
	● coherence	Good	Good	Good
	● discourse marker	Good	Good	Good
Linguistic Competence	● vocabulary	Neutral	Good	Bad
	● grammatical accuracy			
Formulaic Competence	● collocation	Good	Good	Bad
	● idiom			

The results of the validation show that, except for cohesion, While Gemini is hardly at the practical level, GPT-3 and GPT-4 showed very good results in the verification of other competences. In particular, GPT-4 can be said to be at a practical level. In this study, the instructions were given to the generative AIs in sentences that were similar to requests to a person. However, it does not take into account prompt engineering, such as giving definitions of terms. Improving the prompts for better evaluation will be a challenge for the future.

In the field of English education, the challenge going forward is how to utilize AI translation and AI automatic evaluation systems,

as well as generative AI, in English language education. Moreover, it requires the rapid establishment of AI-based pedagogy considering prompt engineering. This is not a matter of the future; it is a real issue that we are facing now.

## Acknowledgment

This work was supported by JSPS KAKENHI Grant Number 23K00724.

## References

- [1] Gally, T. "Educational Practice of AI Translation," Talk presented at *Educational Practice of AI Translation*, LET, December 10, 2022.
- [2] Yukishige, M., Tsutada, K., Nishiyama, M., & Tom Gally. *Let's Work with AI*, Tokyo: Sanshusha, 2022
- [3] Saito, T. "Applications of Generative AIs in English Language Education," Talk presented at *Generative AI: Applications in University Education and Business Operations*, Joint FD/SD Workshop for National and Public Universities in Gunma Prefecture 2023, Maebashi Institute of Technology, December 20, 2023.
- [4] Kawachi, K. & Saito, T. "On the Limits of AI Translation / Automated Evaluation," *Utsunomiya Campus, Teikyo University, Annual Reports of the Humanities*, 29: 187-202, 2023.
- [5] Celce-Murcia, M. "Rethinking the Role of Communicative Competence in Language Teaching." In A. A. Soler., & M. P. Safont Jorda (eds.), *Intercultural Language Use and Language Learning*, 41-57, Dordrecht: Springer, 2007.
- [6] de Beaugrande, R., & Dressler, W. *Introduction to Text Linguistics*. London: Longman. 1981.
- [7] Fox Tree, J. E. "Discourse markers across speakers and settings." *Language and Linguistics Compass*, 3(1), 1-13. 2010.
- [8] Kawabata, Y. *Snow Country*. (Trans. E Seidensticker). New York: Tuttle Publishing. 1957.