IIAI Open Conference Publication Series
IIAI Letters on Institutional Research
Vol.004, LIR294
DOI: https://doi.org/10.52731/lir.v004.294

# Experiments of Automatic Scoring Using Generative AI in a Summary Essay Learning System

Takahiro Yamasaki *, Ayako Hiramatsu †

## Abstract

We are developing an e-learning system aimed at improving Japanese language proficiency. This system focuses on the task of summarizing texts. An e-learning system designed for on-demand learning requires an automatic scoring function to provide feedback to users. This paper focuses on the scoring function using generative AI. Generative AI can be utilized in two ways: generating various patterns of correct answers and using it in the scoring process itself. When generating model answers, the user's responses are compared to these model answers for scoring, so the method of comparison is also considered. This paper reports experimental results on the differences between these applications.

*Keywords:* e-learning, Japanese essay summarization, natural language processing, automatic scoring, generative AI

## 1 Introduction

Currently, e-learning, which utilizes information and communication technology (ICT) for education and learning in various situations, is widely used. In on-demand e-learning, creating educational content and understanding the learners' status are important. In actual operation, the complexity of creating educational content poses a challenge for providers. For learning assessments in class assignments, multiple-choice questions and word input that can be automatically graded are often used to provide immediate feedback. In the case of essay-type assignments, scoring is often based on the presence of keywords, making it difficult to provide instant evaluation for highly flexible essay assignments. In universities, the ability to use Japanese effectively is emphasized as part of the initial year education. It is said that improving writing skills involves understanding basic rules and practicing writing many essays. E-learning provides a self-study environment where learners can use limited time effectively and solve problems appropriate for their level, enabling effective learning. With the recent spread of online classes, the demand for such systems has increased, and many universities are reporting their implementation. Pointing out problematic parts and considering correction methods can lead to more effective learning.

* Dept. of Electrical, Electronic and Information Engineering, Osaka Sangyo University, Osaka, Japan
† Dept. of Information Systems Engineering, Osaka Sangyo University, Osaka, Japan

However, for problems requiring long Japanese texts, after submitting their answers, learners receive feedback where graders actually read the essays, provide written advice, and reply to the learners. This process does not provide real-time advice, resulting in a significant time lag before learners can consider corrections, reducing the number of assignments they can tackle. Additionally, in e-learning, answers are input via PCs or tablets, assuming the use of proofreading and spell-check functions. Therefore, even for essay-type questions, learners do not practice the fundamental skill of writing by hand.

This study aims to develop an e-learning system that enables self-learning through the practice of Japanese essay writing by summarizing short essays [1]. The main features of this learning system include instant scoring and feedback on the summarized essays, a function to recognize handwritten answers, and a function to automatically generate a large number of practice problems. This paper first discusses the objectives and challenges related to summary-type essay questions and outlines the e-learning system we aim to develop. Furthermore, we report experimental results on how generative AI can be utilized as an indicator for automating the scoring of answers. The scoring of answers employs two methods: one where generative AI directly evaluates the answers, and another where the similarity between the answer text and the model answer is evaluated using sentence embedding models.

## 2 About Summary-type Essay Questions

It is said that the way to improve the ability to write essays and similar texts is by understanding the basic rules and then repeatedly practicing writing. There are two types of essay questions where a prompt is provided: one where you express your own opinions and another which is a summary test that requires you to condense a given text. This research focuses on summary problems which necessitate basic practice in understanding the provided text and the ability to concisely express its content. For implementing summary test learning in e-learning, it is necessary to provide a variety of texts to summarize and to immediately evaluate answers, providing comments on areas for improvement.

Various systems are used for the automatic scoring of essays, including those in operational use [2]. Particularly, the automatic evaluation and scoring of English essays and essays have been studied for a long time [3][4][5][6], with each method scoring based on keywords, grammatical correctness, phrasing, and structure. The assessment of logical structure might use cue words (connectives) or rules based on phrases within the text. These essay questions assume an opinion-based prompt, emphasizing logical development. Moreover, while a vast number of essays scored by experts may be necessary, this method cannot be used to provide a variety of topics because it does not allow for the use of past answer examples.

Recent years have seen proposals for methods incorporating machine learning [7][8], including studies that vectorize each essay using BERT and score them using Support Vector Regression (SVR), and research that integrates feature-based models, deep learning models, and hybrid models using item response theory to improve scoring accuracy. Another study employed deep learning models to estimate the logical structure of texts for automatic scoring, and another fine-tuned BERT to consider context in automatic scoring. Much of the related research targets essays written according to a prompt rather than summary-type essays. While there are fewer studies on scoring using generative AI, employing it could potentially achieve more accurate scoring than traditional methods.

In this research, the scoring criteria for summary problems are considered roughly in four levels, focusing on content rather than grammar or descriptive rules. However, the boundaries of evaluation are ambiguous and often vary by scorer. Therefore, evaluations are typically performed by multiple scorers. Since there is no single way to express ideas concisely, evaluations can differ even if similar words are used. Given this, it is necessary to determine evaluations by comparing with multiple correct examples. However, when providing comments, it is necessary to advise on how to narrow the gap to the closest correct example.

## 3 Short Essay Self-learning Systems

This study presents an overview of the e-learning system designed to improve Japanese language proficiency, as illustrated in Figure 1. This system provides scoring and advice for learners' answers while also automatically generating various types of questions, offering a wide range of problems.

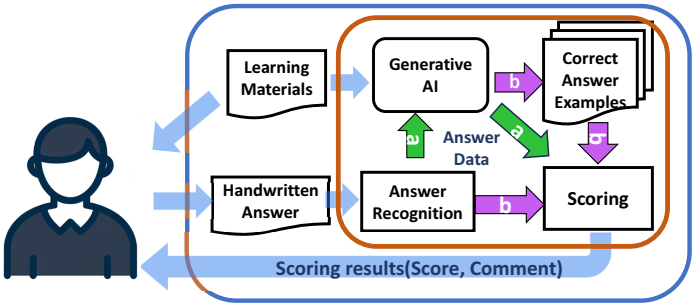


Figure 1: e-Learning System for Essay Summarization

The system presents learners with summary-type essay questions, to which they respond by writing summaries by hand. Next, the system captures these handwritten answers as images, converts them into character codes using Optical Character Recognition (OCR) technology, and digitizes them. When capturing images, the system corrects for noise and distortion to enhance the accuracy of character recognition. During OCR, it saves multiple characters with high probability, and if recognition is difficult due to learners' handwriting peculiarities, it swaps the acquired characters as needed to prevent misrecognition and estimates the appropriate text, thus digitizing the handwritten answers.

Two methods are used for scoring. The first method (a) involves providing the generative AI directly with the task text, response data, and scoring criteria, and scoring is conducted accordingly. An overview of this method is shown in Figure 2. The generative AI acts as a substitute for human scorers, performing the scoring of essays.

The second method (b) first uses generative AI to create model answers from the given summary-type essay prompts. When summarizing, it specifies conditions such as the length and style of the text, and the system creates the intended model answers. Then, using a sentence embedding model, it generates vectors for both the response data and the model answers, and evaluates their similarity by quantifying it numerically. The obtained scores are then quantified according to the scoring criteria. An overview of this method is shown in Figure 3. Finally, based on the scoring results, the system identifies issues in the learners' responses and provides advice tailored to their abilities, supporting the improvement of their writing skills.

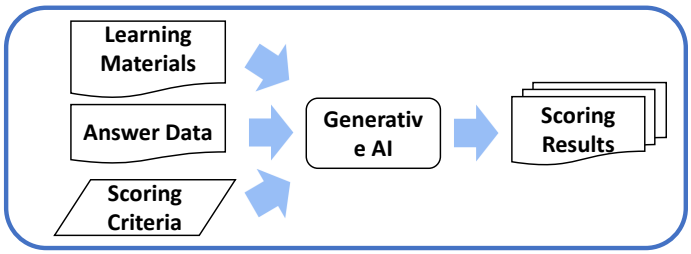


Figure 2: Automated Scoring method (a)

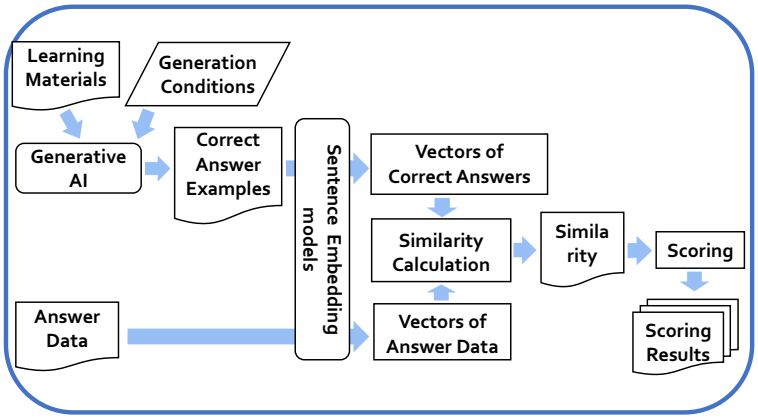


Figure 3: Automated Scoring method (b)

# 4 Automated Scoring Methods

## 4.1 Model Answers for Correct Examples

To ensure fair scoring of answer texts, this experiment establishes a four-level scoring criterion. Each level is defined as follows:

A: The main theme is concisely summarized (correct answer).

B: Not deviating from the main theme, but with some excess or deficiency.

C: Excerpt from the task text, but deviating from the main theme.

D: Contains content outside the task text, no longer a summary.

In this experiment, the model answers for the summary task were created using generative AI. A text of approximately 6000 characters was prepared as the task, and it was summarized to a length of 360 to 400 characters. To allow for diverse expressions, 10 model answers were generated. Although the generated model answers include some variations in expression, they capture the important points, and all received an A grade in manual evaluation.

## 4.2 Scoring Method (a)

ChatGPT is used for scoring by generative AI. The model used is GPT-4 Turbo (gpt-4-1106-preview), with parameters set to their default values. The scoring criteria are as shown in

the previous section, with A and D grades given to more than 10% of the answer data. After recognizing handwritten responses using the proposed method [1], scoring was conducted on 50 pieces of answer data written by 50 students. As an example of the scoring results, a portion (10 answers) of the results from five rounds of scoring is shown in Table 1. In this table, the evaluation scores of A, B, C, and D are represented numerically as 4, 3, 2, and 1, respectively. Despite providing the same prompt to ChatGPT, there is some variability in the scores. The average of these numerical scores is taken as the final scoring result for each answer.

Table 1: Scoring of 10 Answers by ChatGPT

| | No.1 | No.2 | No.3 | No.4 | No.5 | No.6 | No.7 | No.8 | No.9 | No.10 |
|---|---|---|---|---|---|---|---|---|---|---|
| 1st | 3 | 4 | 3 | 4 | 3 | 3 | 4 | 2 | 2 | 1 |
| 2nd | 3 | 4 | 3 | 3 | 3 | 3 | 3 | 2 | 2 | 1 |
| 3rd | 3 | 4 | 3 | 4 | 3 | 3 | 4 | 2 | 2 | 1 |
| 4th | 3 | 4 | 3 | 4 | 3 | 3 | 4 | 2 | 2 | 1 |
| 5th | 3 | 4 | 3 | 4 | 3 | 3 | 3 | 2 | 3 | 1 |

## 4.3 Construction of Embedding Model for Method (b)

In this research, we use three sentence embedding models: ”Sentence BERT,” ”Supervised SimCSE,” and ”Unsupervised SimCSE.” Sentence BERT utilizes a model that has been pre-trained on Japanese data made available on HuggingFace. For both Supervised and Unsupervised SimCSE, we adopt a pre-trained BERT provided by Tohoku University’s Kurohashi Laboratory as the base model. The training for Unsupervised SimCSE uses a dataset of 1 million sentences extracted from the Japanese Wikipedia, while the training for Supervised SimCSE utilizes the ”JSNLI” dataset published by the Kyoto University Language Media Laboratory. JSNLI is a dataset comprising about 500,000 pairs of premise and hypothesis sentences.

## 4.4 Scoring Method (b)

In this study, we utilize three sentence embedding models to evaluate based on the cosine similarity between the ten created correct examples and the answer data. Specifically, the cosine similarity between one correct example and ten answers for each model is presented in Table 2, and this similarity is divided into quarters for scoring. The closer the cosine similarity is to 1, the higher the relevance of the text is considered, and such answer data are rated higher.

Table 2: Cosine Similarity of 10 Answers

| | No.1 | No.2 | No.3 | No.4 | No.5 | No.6 | No.7 | No.8 | No.9 | No.10 |
|---|---|---|---|---|---|---|---|---|---|---|
| BERT | 0.88 | 0.89 | 0.89 | 0.84 | 0.86 | 0.89 | 0.89 | 0.88 | 0.85 | 0.71 |
| Supervised | 0.95 | 0.97 | 0.92 | 0.93 | 0.93 | 0.96 | 0.97 | 0.95 | 0.96 | 0.85 |
| Unsupervised | 0.93 | 0.94 | 0.94 | 0.93 | 0.92 | 0.95 | 0.95 | 0.93 | 0.90 | 0.93 |

In this experiment, the cosine similarities are categorized into four levels using quartiles. Specifically, the cosine similarity for each piece of answer data is calculated for each of the ten correct examples, and these results are then categorized into four levels based on the quartiles, with their average value used as the final scoring result. As an example, the results of scoring the cosine similarities of ten pieces of answer data using quartiles are shown in Table 3.

Table 3: Scoring of Answer by Method (b)

| | No.1 | No.2 | No.3 | No.4 | No.5 | No.6 | No.7 | No.8 | No.9 | No.10 |
|---|---|---|---|---|---|---|---|---|---|---|
| BERT | 4 | 4 | 4 | 3 | 3 | 4 | 4 | 3 | 3 | 1 |
| Supervised | 4 | 4 | 3 | 3 | 3 | 4 | 4 | 4 | 4 | 1 |
| Unsupervised | 4 | 4 | 4 | 3 | 3 | 4 | 4 | 3 | 3 | 1 |

## 5 Experiment Results

Table 4 shows examples of scoring results. Since scoring by a single evaluator may be biased, we adopted the average score of four evaluators as the ground truth. This table presents the scoring results of ChatGPT used in method (a) and each sentence embedding model used in method (b) (5 out of 50 answers). It can be observed that the answer to No.1 received high scores by any method, while No.2 received low scores.

Table 4: Scoring Results by Each Method

| | Average score (Correct) | Method (a) | Method (b) | | |
|---|---|---|---|---|---|
| | | ChatGPT | Sentence BERT | Supervised SimCSE | Unsupervised SimCSE |
| No.1 | 3.25 | 4.00 | 3.54 | 3.95 | 3.54 |
| No.2 | 1.25 | 1.00 | 1.00 | 1.00 | 1.00 |
| No.3 | 1.50 | 3.00 | 3.73 | 2.78 | 2.93 |
| No.4 | 2.50 | 3.80 | 3.18 | 2.90 | 2.47 |
| No.5 | 3.00 | 3.00 | 3.68 | 3.50 | 3.36 |

To verify the scoring accuracy, we calculated the correlation matrix and the root mean square error (RMSE) for the 50 data samples. The correlation matrix is shown above Table 6, and the RMSE values are shown below Table 6. For method (b), the results when the maximum value is taken are shown. Regarding the correlation coefficient with the ground truth scores, the Supervised SimCSE used in method (b) showed the highest value, which was almost equivalent to the ChatGPT used in method (a). On the other hand, the RMSE shows that both ChatGPT and Supervised SimCSE had the smallest errors. These results indicate that both ChatGPT and Supervised SimCSE have a moderate positive correlation with the ground truth, while both contain some errors. Based on these results, it can be concluded that none of the scoring methods are entirely accurate, and there is room for improvement to enhance the scoring accuracy.

To verify the scoring accuracy, we calculated the correlation matrix and the root mean square error (RMSE) for the 50 data samples. The correlation matrix is shown in Table 5, and the RMSE values are shown in Table 6. For method (b), the results when the maximum value is taken are shown. Regarding the correlation coefficient with the ground truth scores, the Supervised SimCSE used in method (b) showed the highest value, which was almost equivalent to the ChatGPT used in method (a). On the other hand, the RMSE shows that both ChatGPT and Supervised SimCSE had the smallest errors. These results indicate that both ChatGPT and Supervised SimCSE have a moderate positive correlation with the ground truth, while both contain some errors. Based on these results, it can be concluded that none of the scoring methods are entirely accurate, and there is room for improvement to enhance the scoring accuracy.

Table 5: Verification of Scoring Accuracy (Correlation Matrix)

| | Correct score | ChatGPT | Sentence BERT | Supervised SimCSE | Unsupervised SimCSE |
|---|---|---|---|---|---|
| Correct score | 1.000 | 0.466 | 0.366 | 0.461 | 0.244 |
| ChatGPT | 0.466 | 1.000 | 0.311 | 0.394 | 0.301 |
| Sentence BERT | 0.366 | 0.311 | 1.000 | 0.799 | 0.495 |
| Supervised SimCSE | 0.461 | 0.394 | 0.799 | 1.000 | 0.584 |
| Unsupervised SimCSE | 0.244 | 0.301 | 0.495 | 0.584 | 1.000 |

Table 6: Verification of Scoring Accuracy (RMSE)

| | ChatGPT | Sentence BERT | Supervised SimCSE | Unsupervised SimCSE |
|---|---|---|---|---|
| Correct | 0.78 | 0.88 | 0.78 | 0.86 |

# 6 Conclusion

This paper focuses on the creation of model answers and the automated scoring of answer texts in the development of an e-learning system aimed at improving Japanese language proficiency. Model answers were created using generative AI, and two methods were adopted for automated scoring: direct scoring by generative AI and similarity evaluation using sentence embedding models. The sentence embedding models used for scoring included Sentence BERT, Supervised SimCSE, and Unsupervised SimCSE, demonstrating the production of high-quality model answers with the use of ChatGPT.

In scoring the answer texts, the correlation coefficients and RMSE were calculated against manual scoring, which served as the standard for correct answers. The results showed that both methods exhibited a certain level of positive correlation; however, the accuracy of the scoring is still not high enough, indicating the need for improvements in the accuracy of automated scoring. Enhancing scoring accuracy could be possible through revising the scoring criteria and improving the methods of quantifying similarities, and these aspects will be addressed in future work. Additionally, increasing the amount of answer

data and conducting experiments with problem texts of different genres and lengths will enhance the diversity of the data.

Furthermore, to identify issues, which is the ultimate goal of the scoring support system, the use of ChatGPT and specifying clear prompts for scoring rationales will be tested. The feasibility of displaying issues will also be further explored, with the aim of developing a system capable of automated scoring and presenting issues.

## Acknowledgments

This work was supported by JSPS KAKENHI Grant Number JP23K02645.

## References

[1] T. Yamasaki and A. Hiramatsu, "A Study of Correcting Handwritten Answers for Short Essay Self-learning Systems," Proc. 14th International Conf. on Learning Technologies and Learning Environments (LTLE 2023), 2023.

[2] T. Ishioka, "Latest Trends in Automated Essay Scoring and Evaluatio," Japanese Society for Artificial Intelligence, Vol.23, No.1, 2008, pp.17–24.

[3] J. Burstein and M. Wolska, "Toward evaluation of writing style: Finding overly repetitive word use in.student essays, " Proc. 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL '03), 2013, pp. 35–42.

[4] E. B. Page, "New computer grading of student prose, using modern concepts and software, " Experimental Education, Vol.62, No.2, 1994, pp.127–142.

[5] T. K. Landauer, D. Laham, and P. Foltz, "Automated scoring and annotation of essays with the intelligent essay assessor," Automated Essay Scoring: A Crossdisciplinary Perspective, 2003, pp.87–112.

[6] S. Elliot, "IntelliMetric: From Here to Validity," Automated Essay Scoring: A Crossdisciplinary Perspective, 2003. pp.71–86.

[7] V. S. Kumar and D. Boulanger, "Automated Essay Scoring and the Deep Learning Black Box: How Are Rubric Scores Determined ?," Artificial Intelligence in Education, Vol.31, 2021, pp.538–584.

[8] D. Ramesh and S. K. Sanampudi, "An automated essay scoring systems: a systematic literature review," Artificial Intelligence Review, Vol.55, 2022, pp.2495–2527.