Research and Practice in
Technology Enhanced Learning

RESEARCH Free and Open Access

# Unsupervised techniques for generating a standard sample self-explanation answer with knowledge components in a math quiz

Ryosuke Nakamoto [1]*, Brendan Flanagan [2], Yiling Dai [3], Kyosuke Takami [2,4] and Hiroaki Ogata [3]

*Correspondence:
s0527225@gmail.com
Graduate School of Social Informatics, Kyoto University, Kyoto, Japan
Full list of author information is available at the end of the article

**Abstract**

Self-explanation is a widely recognized and effective pedagogical method. Previous research has indicated that self-explanation can be used to evaluate students' comprehension and identify their areas of difficulty on mathematical quizzes. However, most analytical techniques necessitate pre-labeled materials, which limits the potential for large-scale study. Conversely, utilizing collected self-explanations without supervision is challenging because there is little research on this topic. Therefore, this study aims to investigate the feasibility of automatically generating a standardized self-explanation sample answer from unsupervised collected self-explanations. The proposed model involves preprocessing and three machine learning steps: vectorization, clustering, and extraction. Experiments involving 1,434 self-explanation answers from 25 quizzes indicate that 72% of the quizzes generate sample answers containing all the necessary knowledge components. The similarity between human-generated and machine-generated sentences was significant with moderate positive correlation, $r(23) = .48$, $p < .05$.The best-performing generative model also achieved a high BERTScore of 0.715. Regarding the readability of the generated sample answers, the average score of the human-generated sentences was superior to that of the machine-generated ones. These results suggest that the proposed model can generate sample answers that contain critical knowledge components and can be further improved with BERTScore. This study is expected to have numerous applications, including identifying students' areas of difficulty, scoring self-explanations, presenting students with reference materials for learning, and automatically generating scaffolding templates to train self-explanation skills.

**Keywords:** Self-explanation, Rubric, Knowledge components, Summarization, Natural language processing

## Introduction

Self-explanation is defined as generating explanations for oneself, explaining concepts, procedures, and solutions to deepen understanding of learning materials and make sense of relatively new information (Chi et al., 1994; Rittle-Johnson, 2006). While self-explanation has been found to be a potentially highly effective pedagogical method, it often relies on preparation by instructors which has been suggested as a limiting factor in its widespread implementation (Bisra et al., 2018). Several methods have been proposed, including the design of system interfaces that effectively monitor and support self-explanation, models that can assess case understanding from self-explanation behaviors, and strategies that effectively elicit further self-explanation to improve student case understanding (Arner et al., 2021; Boonthum et al., 2007; Conati & Vanlehn 2000; McNamara et al., 2004). However, an overview of past studies from the perspective of automating analysis with artificial intelligence reveals that analysis methods take a top-down approach. It is necessary to prepare the model answers for self-explanation in a fixed structure and capture these data in the form of an annotated dataset (Jackson et al., 2010; Nakamoto et al., 2021; Panaite et al., 2018; Panaite et al., 2019). This can limit the scalability in implementing these tools in practical environments. Therefore, in this study, we propose a method using unsupervised learning to automatically generate standard answers from collected self-explanations.

Generating model answers for self-explanations can be very difficult, especially when there is no prior information on how to solve a specific problem. To overcome this challenge, our study proposes a bottom-up approach that analyzes data collected from self-explanations. One of the main advantages of this approach is its ability to be applied to previously unseen problems that lack labeling, so long as there is input data available. While research in this area of self-explanation analysis is still limited, this study aims to investigate the potential for automatically generating standard self-explanation examples that incorporate the necessary conceptual knowledge components required to solve a given problem.

This paper explores the potential of self-explanations to generate standard sample answers in three different scenarios. In the first scenario, the study aims to identify knowledge components missing from students' self-explanations by comparing them to model answers. It will also investigate whether the system can extract sentences from high-quality self-explanations. In the second scenario, the standard sample answer will be provided as a reference for learners who struggled with a particular quiz question. Finally, in the third scenario, the study proposes a novel approach for automatically generating self-explanation scaffold templates based on sample sentences. This approach utilizes a glossary-based support system that encourages students to practice their self-explanation skills by looking up concepts from a curated list and filling in blank templates (Berthold &

Renkl, 2009; Berthold et al., 2009; Rittle-Johnson et al., 2017). Overall, the unique contribution of this paper is that it offers valuable insights into the potential of unsupervised learning methods as a tool for generating sample self-explanation answers in different learning scenarios.

The methods proposed in this study aim to promote learning of mathematical concepts and procedures through self-explanation. By understanding the learning process through self-explanation (Bisra et al., 2018), knowledge acquisition could be enhanced beyond what is possible with reference books alone. To achieve this goal, the paper presents a method for generating model responses based on the analysis of collected self-explanations. We conducted evaluations of our proposed approach using real data from three different perspectives to determine its appropriateness; (1) the ability of the model to generate self-explanations that contain relevant key component information for each quiz (Knowledge Components Extraction Grading), (2) the quality of the generated self-explanations compared to human-generated self-explanations as measured by established metrics (The Quality Evaluation with Metrics), and (3) the clarity of the model's responses for students (Readability Analysis). Through these evaluations, the paper aims to establish the fundamental methods for creating automatic model responses that can be used to analyze self-explanations and promote learning.

## Related work

### Effects of self-explanation in mathematics and its utilization

Self-explanation has been shown to improve students' conceptual and procedural knowledge in math by helping them focus on important details (Bisra et al., 2018; Renkl, 2017; Rittle-Johnson, 2017). Conceptual knowledge refers to abstract ideas, while procedural knowledge is tied to specific problem-solving practices (Rittle-Johnson & Schneider, 2015; Rittle-Johnson et al., 2001; Rittle-Johnson et al., 2015; Star, 2005). Verschaffel et al. (1999) identified five steps for solving math problems: drawing pictures, making lists, simplifying numbers, making calculations, and evaluating the solution. While self-explanation is not directly tied to any specific step in the problem-solving process, it focuses on the knowledge elements involved. Although these steps provide a framework for problem-solving, self-explanation goes beyond the procedural aspects and emphasizes the explanation of necessary knowledge components.

In our study, we examined the possibility of self-explanation by hypothesizing that students follow these problem-solving steps and that self-explanation occurs in relation to the knowledge elements, specifically tied to procedural knowledge.

The use of self-explanation in web-based learning has been explored through various methods. Some include designing interfaces that monitor and support self-explanation, developing models to assess understanding from self-explanation behavior, and using strategies to encourage further self-explanation for improved comprehension (Conati & Vanlehn, 2000). Crippen and Earl (2007) created a web-based learning tool that helps students with structured problem-solving. iSTART, an interactive tutoring system developed by McNamara et al. (2004), uses natural language processing techniques to evaluate and score self-explanation in reading. The system extracts features from learners' self-explanation artifacts and compares them to the reading material, providing appropriate scaffolding to improve comprehension. The results indicate that iSTART can effectively support self-explanation in various disciplines (Jackson et al., 2010). The present paper focuses on self-explanation in mathematics which differs from reading comprehension in that the source material often contains less text information, making it difficult to compare to learners' self-explanation artifacts. Therefore, we propose a method to generate examples for the purpose of scoring from collected self-explanations. To evaluate the effectiveness of this method we compared the similarity of the generated examples to human examples as measured by established metrics.

## Research utilizing the uniqueness of self-explanations

Self-explanation writing is often incomplete and fragmented, with learners focusing on what they don't understand rather than writing for others (Chamberland et al., 2015). As a result, unexpected answers are common in self-explanations (Panaite et al., 2018). Factors such as training in self-explanation writing and writing skills can also affect the quality of the writing (Hodds et al., 2014). Moreover, self-explanations do not typically provide clear answers or an effective and simple way to use them.

To address these challenges, various methods have been proposed for processing self-explanations. For example, Panaite et al. (2018) found that using the occurrence of fixed expressions and word counts in English reading comprehension as a filter can improve the quality of extracted sentences and the accuracy of subsequent automatic scoring. Panaite et al. (2019) used rule-based automatic and machine learning methods to score accumulated self-explanation data. Nakamoto et al. (2021) proposed a method for checking whether students can explain each step of their self-explanation by comparing the similarity between their writing and a human-created example solution. They determined whether the required information and vocabulary for the unit were included and identified knowledge gaps if elements were missing. The present paper incorporates the use of filtering methods proposed in previous research, such as: the identification of fixed expressions, and the appropriateness of self-explanation length to improve the quality of generated sample self-explanations written in Japanese for mathematics questions.

# Methods

## Goals of our models

In this section, we introduce the proposed method and show how the system generates sample answers of self-explanations. An overview of the inputs and outputs of the model proposed in this research is shown in Figure 1 along with possible uses of the outputs. The input into the model of the proposed method includes the self-explanations of all students who explained the questions and their accompanying handwriting data that was collected when the student answered a question. The output is a single example response statement to the question. This example response is then compared to a student's self-explanation and can be used as a template to prompt self-explanation, which has been identified as an important task for supporting self-explanation (Bisra et al., 2018).

A rubric is a set of guidelines used to evaluate students' work. General rubrics provide an overview of performance levels, while task-specific rubrics specify the mathematical aspects of a task that determine each level of performance. Rubrics can be used to evaluate various tasks, and are well-suited for tasks with multiple solutions or strategies (Thompson & Senk, 1998), but in this case, simplified rubrics were used to assess whether or not students were able to solve math problems step-by-step, with the intention of the rubric being system judgeable. Table 1 shows the definitions of key terms used in this paper.

The output, the final goal of this study, is a sample answer as shown in Table 2. Ideally, this sample answer should demonstrate the knowledge components required to solve the quiz at each step.

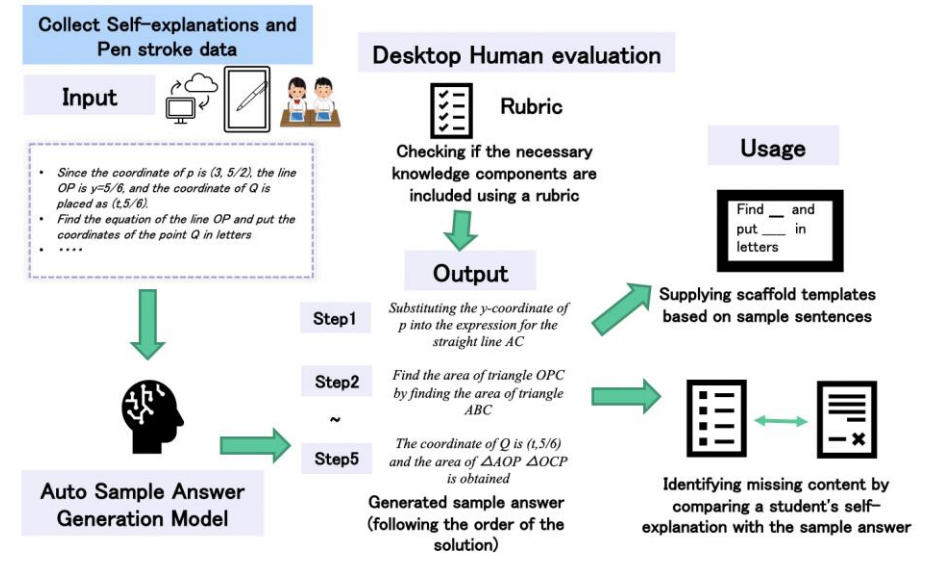

Collect Self–explanations and Pen stroke data
Input
Since the coordinate of p is (3, 5/2), the line OP is y=5/6, and the coordinate of Q is placed as (t,5/6).
Find the equation of the line OP and put the coordinates of the point Q in letters
••••
Auto Sample Answer Generation Model
Desktop Human evaluation
Rubric
Checking if the necessary knowledge components are included using a rubric
Output
Step1
Substituting the y-coordinate of p into the expression for the straight line AC
Step2
Find the area of triangle OPC by finding the area of triangle ABC
~
Step5
The coordinate of Q is (t,5/6) and the area of ΔAOP ΔOCP is obtained
Generated sample answer (following the order of the solution)
Usage
Find __ and put ___ in letters
Supplying scaffold templates based on sample sentences
Identifying missing content by comparing a student's self–explanation with the sample answer

**Fig. 1** An overview of the process of our proposed model, and the inputs and outputs

**Table 1** Definition of words

| Name | Definition |
|---|---|
| Knowledge Components | Specific conceptual knowledge or unique unit elements required to solve this quiz. |
| Step | The procedure or order in which the knowledge components appear in the quiz. |
| Rubric | Can-do descriptors that clearly describe all the knowledge components of the quiz by steps and are used to create labels and evaluations. |
| A Sample Answer (of self-explanations) | A single standard answer of self-explanations of a quiz with knowledge components, which are prepared according to the step rubric number. |

**Table 2** Rubrics and a sample answer of self-explanation in a quiz

| Number | Rubric | Sample answer of self-explanations |
|---|---|---|
| Step 1 | Be able to find the equation of a linear function from two points. | Substituting the y-coordinate of p into the equation of the line AC. |
| Step 2 | Be able to find the equation of the line that bisects the area of a triangle. | Find the area of triangle ABC, then find the area of triangle OPC. |
| Step 3 | Be able to represent a point on a straight-line using letters (P-coordinates). | With the line OC as the base, find the y-coordinate of p, which is the height. P’s coordinate is (t, -1/2t+4). |
| Step 4 | Be able to represent a point on a straight-line using letters (Q-coordinate). | Since the coordinates of P are (3,5/2), the line OP is y=⅚, and the coordinates of Q are (t,5/6). |
| Step 5 | Be able to formulate an equation for area based on relationships among figures. | Finally, the area of △QAC was found from △AQO and △OQC, and the coordinates of Q were found. |

## Dataset details

When constructing a data-based model, it is essential to consider the data set as the center of the model. This section defines what a good self-explanation is, which is crucial for generating the model, and provides an overview of the characteristics of a good self-explanation.

### *Dataset acquisition*

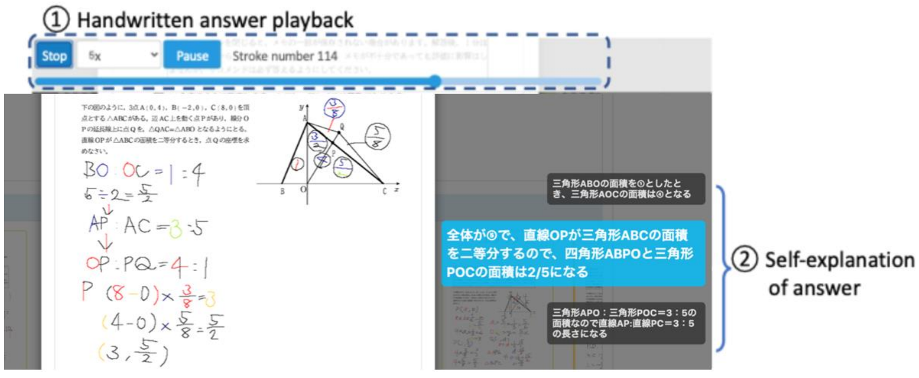
The data was collected between January 1, 2020, and December 31, 2021, using the LEAF platform (Flanagan & Ogata, 2018), which includes a digital reading system called BookRoll and a learning analytics tool called LAViEW. This platform has been used for several years in a Japanese secondary school. Students were asked to view the quiz and write their answers using a stylus and tablet computer with handwriting input in BookRoll. BookRoll captures the handwriting data as a series of vectors that represent the coordinates and velocity of pen strokes, enabling realistic playback of the handwritten answers and fine-grained analysis of the students’ answering process (Flanagan et al. 2021).

The utilization of pen stroke log data yields several advantages over character recognition. Firstly, pen stroke logs capture precise details such as movement and direction, providing a comprehensive understanding of the writing process. This granularity allows for a nuanced analysis of the data. Secondly, pen stroke data captures contextual information beyond the characters themselves. It enables the observation of stroke sequences and self-explanation timing, reconstructing the temporal aspect of writing. This time series data offers valuable insights into stroke order, pauses, and handwriting patterns, facilitating anomaly detection.

After completing the quiz, students were asked to use LAViEW to review their handwritten answers and explain how they arrived at their solutions as shown in Figure 2. Students would input a sentence of explanation each time they believed they had completed a step in their answer while the playback occurred. This ensured that their self-explanations were associated with the pen stroke data chronologically. The self-explanation of the answer in Figure 2 contained the following steps from top to bottom: “If the area of triangle ABO is 1, the area of triangle AOC is 4. Since the entire area is five and the line OP bisects the area of triangle ABC, the area of quadrilateral ABPO and triangle POC is 2/5. The area of triangle APO compared to triangle POC is 3:5, so the length of line AP compared to line PC is 3:5.”

### Self-explanation quality scoring

To build a good model, it’s important to have good-quality input data. Therefore, the first step was to evaluate the quality of the self-explanations. The evaluators created an objective scoring index, which focused on the inclusion of essential knowledge components and logical explanations that are shown in Table 3. The scoring results were rated on a scale of 1~5 and sorted into categories that are shown in Table 4. The evaluations were carried out by two scorers, who initially rated the self-explanations individually. After the initial round of evaluations, the two scorers then compared their evaluations with each other, and if they



下の図のように，3点A(0,4)，B(−2,0)，C(8,0)を頂点とする△ABCがある。辺AC上を動く点Pがあり，線分OPの延長線上に点Qを，△QAC=△ABOとなるようにとる。直線OPが△ABCの面積を二等分するとき，点Qの座標を求めなさい。

BO : OC = 1 : 4
$5 \div 2 = \frac{5}{2}$
AP : AC = 3 : 5
OP : PQ = 4 : 1
P $(8-0) \times \frac{3}{8} = 3$
$(4-0) \times \frac{5}{8} = \frac{5}{2}$
$(3, \frac{5}{2})$

三角形ABOの面積を①としたとき、三角形AOCの面積は④となる

全体が⑤で、直線OPが三角形ABCの面積を二等分するので、四角形ABPOと三角形POCの面積は2/5になる

三角形APO：三角形POC＝3：5の面積なので直線AP:直線PC＝3：5の長さになる

**Fig. 2** Handwritten answer review playback and self-explanation input user interface

**Table 3** Self-explanation quality score grading definition

| Graded score | Description |
|---|---|
| 1 (Unacceptable) | The number of steps for which self-explanations are filled in for the steps required for the solution is minimal, and/or there were fixed expressions in the students' self-explanations (e.g., mistaken patterns, boredom.) |
| 2 (Poor) | Self-explanations are mostly provided for the steps required for the solution. Although, they are more like bullet points than explanations. |
| 3 (Fair) | Self-explanations are mostly provided for the steps required for the answer. The average level of self-explanations among all respondents. |
| 4 (Very Good) | Self-explanations are provided for most of the steps required for the answer, but there is room for improvement as an explanation (Logic, expressions). |
| 5 (Excellent) | Self-explanations are mostly provided for the steps required for the answer, and the explanation is logical and well-written. |

**Table 4** Descriptive statistics of the collected self-explanations

| Number of quizzes | Cumulative total of answers | Number of individuals | Sentence length in Japanese characters | | Quality score | |
|---|---|---|---|---|---|---|
| | | | M | SD | M | SD |
| 25 | 1,434 | 117 | 62.0 | 53.6 | 2.91 | 1.42 |

**Table 5** Distribution of human graded quality scores

| Quality score | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| Number of answers | 350 | 231 | 303 | 292 | 258 |
| Sentence length in Japanese characters (Mean) | 21.2 | 34.6 | 47.8 | 76.5 | 142.6 |

differed, they discussed the reasons why and together came to a final value for the evaluation. In using this method there were no tiebreakers as the two scorers were able to come to an agreement through discussion. The descriptive statistics of the self-explanations in the dataset and the evaluations given by the scorers are shown in Table 5 and Figure 3.

Table 5 and Figure 3 show longer self-explanations tended to have higher scores, while shorter ones tended to have lower scores. Self-expressions with less than 50 characters tended to receive a score of 3 or lower, while those with 40 or more characters tended to receive a score of 4 or higher.

Examples of evaluated self-explanations are shown in Table 6, with the first example demonstrating a logically structured self-explanation that achieved a score of 5 despite its lengthy nature. This makes it easy to extract the characteristics of the problem-solving process from the content. It contains a total of five Knowledge Components, and was awarded the highest score. Conversely, the second example achieved a score of 2 and consists of fragmented content presented in a relatively short passage when compared to other self-explanations. While it includes two Knowledge Components, it was deemed insufficient as a self-explanation of the process as it did not fulfill the necessary problem-solving steps, and this resulted in it receiving a score of 2.

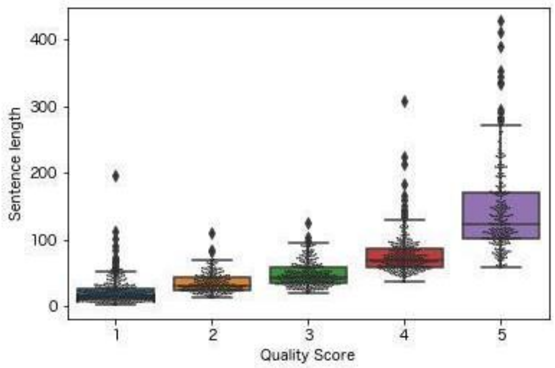


**Fig. 3** A boxplot of the self-explanation quality score distribution

**Table 6** Examples of self-explanations with High and Low quality scores

| Quality score | Sentence length | Self-explanation | Evaluation |
|---|---|---|---|
| 5 | 138 | The area of triangle AOC is ④ when triangle ABO has an area of ①, the sum of the areas is ⑤, line OP divides triangle ABC into two equal parts so the areas of quadrilateral ABPO and triangle POC are 2/5, the length ratio of AP to PC is 3:5 because the area ratio of triangle APO to POC is 3:5, the length ratio of OP to PQ is 4:1 because the area ratio of triangle ACO to QAC is 4:1, and we obtained the coordinates of point P from the coordinates of points A and C, and the coordinates of point Q from the coordinates of points O and P. | Clearly describing the relationships between different elements in the problem. Showing the use of specific mathematical concepts and formulas, such as the area of a triangle and the length ratio. |
| 2 | 39 | To find the equation of line OP and the coordinates of point Q, we first assign Q as a variable and then calculate the areas of triangles AOQ and QOC using the coordinates of Q. | Vague and lacks the necessary information to understand the process and reasoning behind the calculations. |

## The architecture of our approach

Generating sample self-explanations requires extracting good self-explanations from collected data, which forms the basis of our proposed model. An overview of the process that was used to filter and extract good self-explanations from the collected data and how it was preprocessed is shown in Figure 4. Firstly, we employed two characteristics to filter input data based on the quality and length distribution as shown in Figure 3 and Table 5. It was observed that the threshold for the sentence length of a self-explanation is around 40 characters and can be used to filter low quality self-explanations from high quality examples. Next, we excluded sentences that contained fixed expressions, such as: “make”,

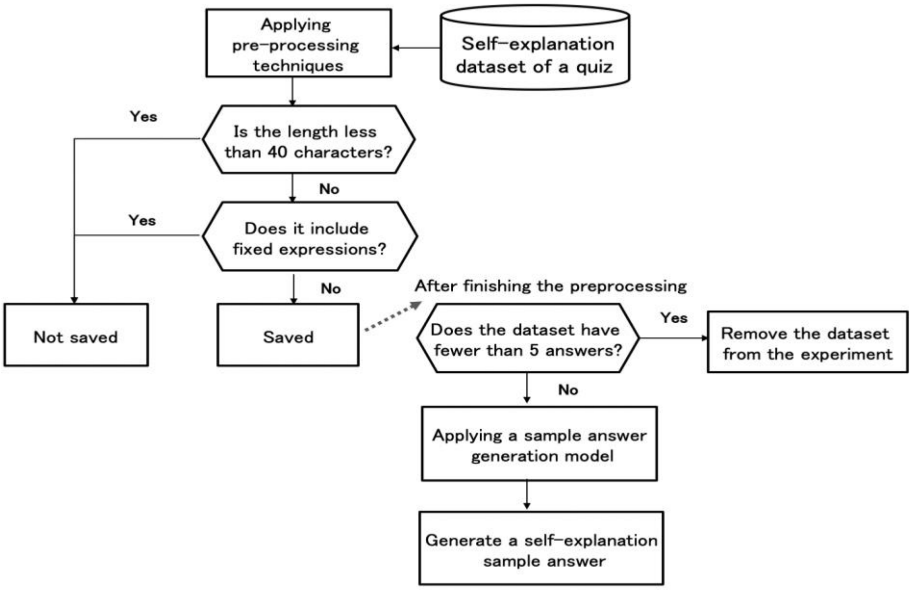

Applying pre-processing techniques
Self-explanation dataset of a quiz
Is the length less than 40 characters?
Yes
No
Does it include fixed expressions?
Yes
No
Not saved
Saved
After finishing the preprocessing
Does the dataset have fewer than 5 answers?
Yes
Remove the dataset from the experiment
No
Applying a sample answer generation model
Generate a self-explanation sample answer

**Fig. 4** Workflow for the self-explanation processing

"error", "impossible", "unreasonable", "to be annoyed", "irritated", "worried", "in difficulty", "strange", and "erased". Although these expressions are idiomatic, we removed them as they are not appropriate as model self-expressions. Finally, we omitted questions with less than five self-explanation responses as we determined that they did not have adequate data for analysis. A histogram of the distribution of self-explanation responses is shown in Figure 5, where the y-axis represents the number of quizzes that have the number self-explanation responses represented by in the same bin. For example, the bin on the left-hand side represents 3 quizzes have five or less self-explanation responses.

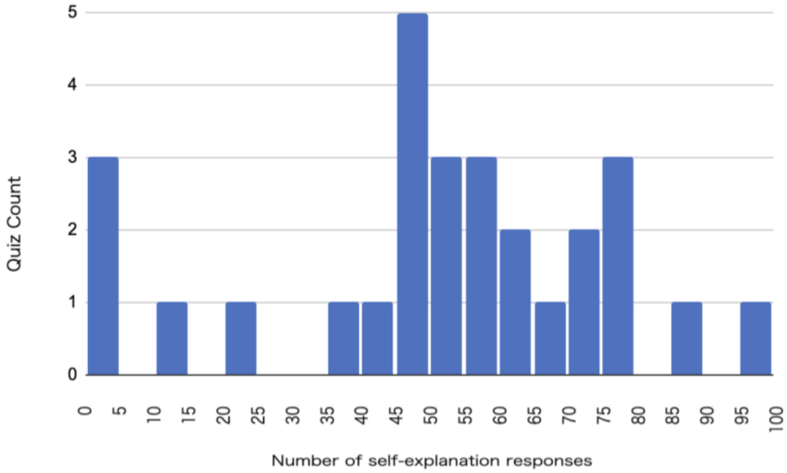


**Fig. 5** Histogram of the distribution of self-explanation responses over the quizzes in the dataset

## Sample answer generation model

Once a set of problems containing five or more responses had been collected, the proposed self-explanation model was constructed. In this section, we outline the underlying methodology used in constructing the proposed model and the output generated by this approach.

### *Overall architecture of the proposed model*

Text summarization is a crucial area in the design of deep learning models for natural language processing (NLP), and automated scoring of summaries is preferred over manual scoring (Crossley et al., 2019; Iqbal & Qureshi, 2020). Previous studies have used NLP tools like LSA and machine learning approaches (León et al., 2006; Ozsoy et al., 2011). There are two main types of summarizations: extractive, which extracts the essence from the entire text, and abstractive, which generates sentences. This study focuses on a sample self-explanation for students, requiring textual correctness. Extractive summarization was chosen due to limited data, and semantic chunks were extracted using unsupervised techniques.

The proposed model consists of three main components: vectorization, clustering, and extraction, as illustrated in Figure 6. The vectorization step involves transforming the textual data into numerical representations suitable for analysis. Clustering techniques are then applied to group the semantic chunks based on their similarities. Finally, the extraction component identifies the most representative sentences within each semantic cluster. In order to determine the most representative sentences for each semantic cluster, we implemented a method using the chronological order from the pen stroke data. By analyzing the pen stroke log, we were able to reorder the generated sentences based on

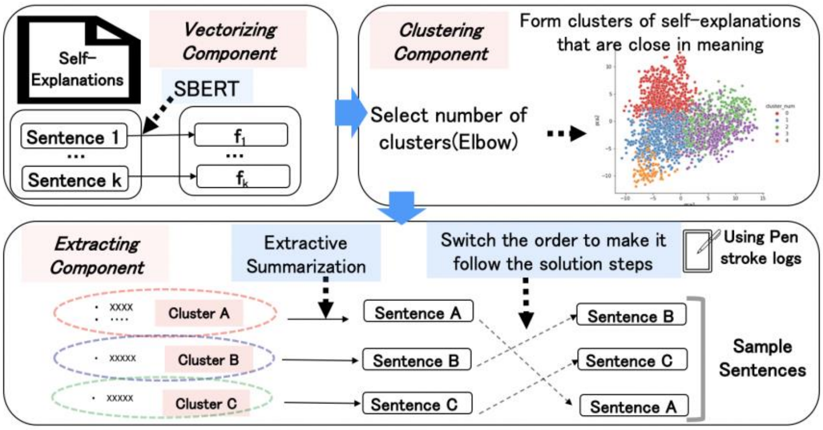


**Fig. 6** Overall of a proposed model architecture. The input is all the self-explanation sentences associated with the quiz, and the output is the summarization of knowledge components for the quiz

their position in the problem. This allowed us to identify and select the sentences that best captured the essence of each semantic cluster.

### *Vectorizing component*

We used Sentence BERT (SBERT; Reimers and Gurevych, 2019) and the BERT Japanese pre-trained model (Suzuki, 2019) as the vectorizing component for the following reasons. Firstly, BERT is a deep learning model based on the transformer architecture (Vaswani et al., 2017) that outperforms existing models in natural language processing on various tasks (Devlin et al., 2019). SBERT fine-tunes BERT and has significantly improved sentence embedding methods (Reimers & Gurevych, 2019). Secondly, since students are not well-trained in writing self-explanations, we expected a lack of uniformity in their descriptions, expressions, and content. BERT's versatility allows it to be applied to various tasks without changing the model's structure, which we thought was appropriate for this research given the lack of uniformity in the self-explanations.

### *Clustering component*

As the clustering component, we employed an unsupervised learning model, K-means. The purpose of creating meaning-intensive clusters through unsupervised learning is to reproduce the mathematical steps taken to produce the solution. Math problems have a method that often builds on previously learnt knowledge component units and often has to be solved by combining several basic knowledge component units. The authors assumed that a junior high school math problem would probably contain at least two steps and at most six steps of unit knowledge components, but this varies from problem to problem and requires a flexible model design. For example, when solving an equation, if the problem can be solved by simply shifting $x$, it is not easy to describe it further in the self-explanation. However, if it is a linear function, and students need to find the area of a triangle, find the formula of a straight line, or find the coordinates of a vertex, the steps become much more complicated. These variations need to be handled in a data-driven manner. In this study, the sum of squared errors was used to automatically determine the number of clusters using the elbow method, and it was determined in the range of 2 to 6 is optimal.

### *Extracting components*

To extract components for each semantic cluster, we identified the most representative sentences by sorting them according to their position in the problem, which was obtained from the chronological data of pen strokes. We used the LexRank algorithm (Erkan & Radev, 2004) to extract the most representative sentences from each cluster, based on the frequency of occurrence of each sentence. LexRank is a graph-based method that represents sentences in a graph structure and creates a summary by analyzing the relationships between the nodes that represent each sentence or word. Table 7 provides examples of self-explanations that were separated by the clustering model and extracted by LexRank.

**Table 7** Intermediate output examples for each process in the proposed model

| Cluster | Clustered sentences | Extracted sentence | Final order |
|---|---|---|---|
| A | [‘Find the equation of the straight line AC’, ‘Find the X coordinate of point P’, ‘Find the equation of the straight line OP’, ‘Let the X coordinate of point Q be t. The Y coordinate is 5/6t’, ‘Find the value of t’, ‘Find the value of 5/6t’, Considering the coordinates of points A and C, I found the coordinates of point P.] | Considering the coordinates of points A and C, I found the coordinates of point P. | 2 |
| B | [‘Find the area of triangle OPC’, ‘Find the height of triangle OPC when the base is OC’, ‘Find the area of triangle ABO (triangle QAC)’, ‘Find the area of triangle OAC’, ‘Add the area of triangle AOQ and the area of triangle QCO (quadrilateral AOQC)’, ‘Since the area of quadrilateral AOQC is 20, create an equation equal to 16/3t’, If the area of triangle ABO is 1, then the area of triangle AOC is 4.] | If the area of triangle ABO is 1, then the area of triangle AOC is 4. | 4 |
| C | [‘The area of triangle ABC is 20’, ‘The area of triangle OPC is 10 since it bisects triangle ABC’, ‘The Y coordinate of P is like this’, ‘This is the equation for line segment AC’, ‘Since the equation for line segment AC is known, the X coordinate of P can also be found’, ‘The equation for line segment OP is like this’, ‘Let the coordinates of Q be (t,5/6t)’, ‘First, find the area of quadrilateral OCQA’, ‘Quadrilateral OCQA - triangle OCA = triangle CQA, which is also 10’, ‘Substitute t and find the coordinates of Q’, Substituting the Q value (p,q) into the equation of OP, then q=5/6p, so Q(p,5/6p).] | Substituting the Q value (p,q) into the equation of OP, then q=5/6p, so Q(p,5/6p). | 3 |
| D | [‘First, find the straight line AC’, ‘Next, find the area of △ABC and find the area of △OPC. Also find the Y coordinate of point P’, ‘Then, find the X coordinate of point P and find the straight line OP’, ‘Next, find △AOB and find the area of △AQO. Also find △AOC and find the area of □AOCQ’, ‘Let the coordinates of Q be (t, 5/6) and find the area of △AOP and △OCP and add them to find t’, Considering the coordinates of point O and point P, I found the coordinates of point Q.] | Considering the coordinates of point O and point P, I found the coordinates of point Q. | 5 |
| E | [‘The area of triangle ABC is 10x4x1/2 = 20’, ‘Therefore, the area of triangle OPC is 10’, Since the whole is five and, the line OP bisects the area of triangle ABC, the area of quadrilateral ABPO and triangle POC is 2/5.’ The base of triangle OPC is 8, so the height is 5/2. The Y coordinate of P is also 5/2’, ‘Since △ABO = △QAC, both areas are equal’] | Since the whole is five and, the line OP bisects the area of triangle ABC, the area of quadrilateral ABPO and triangle POC is 2/5. | 1 |

## Experimental setup

Our approach was evaluated using real data from three perspectives to determine its suitability. Firstly, we tested the model's ability to extract relevant key information for each quiz (Grading Extracted Knowledge Components). Secondly, we compared the quality of the self-explanations generated by the model to human-generated explanations using established metrics (Quality Evaluation with Metrics). Finally, we assessed the readability of the model's responses for students (Readability Analysis). We assessed the structure and readability of the sentences generated by the model and analyzed two out of the 25 questions in detail.

### Grading extracted knowledge components

The authors evaluated the machine-generated self-explanations using an established evaluation index for each quiz. Table 2 shows the criteria and results of the evaluation, which was based on the ability of evaluators to recall necessary knowledge components. To ensure consistency, two authors and one assistant evaluated the answers and any discrepancies were resolved through consultation, resulting in an improved Fleiss' kappa coefficient of 0.870 which indicates good reliability of agreement between the raters (Fleiss, 1971). Additionally, the authors evaluated the extracted sentences using a quality score to determine the appropriateness of the newly generated sentences.

### Quality evaluation with metrics

Next, we evaluated the similarity of human-created and machine-generated sentences using several metrics: BERTScore, BLEU (BiLingual Evaluation Understudy), and ROUGE (Recall-Oriented Understudy for Gisting Evaluation). Compared to the other metrics, BERTScore is expected to capture the meaning of the whole sentence better and be more robust for paraphrasing because it uses BERT embeddings which are generated from words and their context (Zhang et al., 2020). BLEU is a widely used metric for evaluating models such as machine translation. It evaluates how many N-grams in the generated text are included in the correct text (Papineni et al., 2002). The same is for ROUGE, a method based on an N-gram-based agreement (Lin, 2004).

In addition, we conducted a Spearman correlation analysis to investigate the relationship between the summary index and human evaluation. The aim of this analysis was to explore the possibility of using representative metrics as an alternative to the labor-intensive process of scoring Human Evaluation Scores as defined in the formula below. The Human Evaluation Score (HES) was scored according to how well machine-generated answers met the knowledge components against the evaluation index in the following form. Root Mean Squared Error (RMSE) was also calculated to check the error variance.

$$Human\ Evaluation\ Score = \frac{Num\ of\ Rubrics\ -\ Missing\ knowledge\ components}{Num\ of\ Rubrics}$$

### Readability analysis

Additionally, the authors conducted a survey to assess the readability of the self-explanation text generated by their system. Seven evaluators, including five students and two assistants, were asked to review the generated self-explanation and judge its readability. They were given written definitions and explanations as shown in Table 8 to guide their evaluation. The questionnaire items and analytical methods used in the survey were based on a previous study by Drori et al. (2021) on automatic question generation for mathematics.

## Results

### Knowledge components extraction grading

The human evaluation results of the generated self-explanation examples are shown in Table 9 and Table 10. In 72% of the quizzes, all five rubrics (knowledge components) were successfully generated, while 16% were missing one and 8% were missing two rubrics. Meanwhile, Table 10 shows that the quality score of the extracted sentences had an average of 4.49, with the bottom 25% scoring 4.15, indicating that the proposed method can effectively extract sentences. Most sentences were extracted from sentences with a score of 4 or 5, demonstrating the high quality of the extracted sentences.

**Table 8** Explanatory notes to evaluators assessing the readability of human and generated examples

| Items | Description |
|---|---|
| Explanations | Below are the model answers for each question self-explanation generated by a human or AI. The model answers essentially describe the required solution process and are based on the assumption that the answers can be understood by reading them. Please read the model answers to the following self-explanation sentences, answer whether the Japanese are easy or difficult to read, and answer whether AI or a human created it. There are 50 questions in total. |
| Definition of readability | Can the student define the steps necessary to answer the questions and explain them logically? Is it suitable for students to read as an example of self-explanation answers in class? |

**Table 9** Missing knowledge components of each quiz by human evaluation

| Missing knowledge components | 0 | 1 | ≥ 2 |
|---|---|---|---|
| Number of quizzes | 18 | 4 | 3 |
| Probability density | **0.72** | 0.16 | 0.12 |

**Table 10** Descriptive statistics of extracted self-explanations' quality score

| Number of quizzes | Mean of quality score | SD | Min | 25% | 50% | 75% | Max |
|---|---|---|---|---|---|---|---|
| 25 | **4.49** | 0.558 | 2.60 | 4.15 | 4.60 | 5.00 | 5.00 |

## Quality evaluation with metrics

Table 11 and Table 12 demonstrated the degree of similarity between human-generated and machine-generated sentences. The overall similarity was 0.715, with the best-performing generative model in BERTScore exhibiting a significant correlation of R=0.48. Table 11 provides F1 Metrics scores for various metrics, with BERTScore achieving the highest similarity metric of 0.719 and ROUGE-1 following closely with an average of 0.443. Correlations and RMSE were evaluated between HES and the different metrics, with tests of no correlation conducted. The results revealed a moderate correlation between HES and BERTScore $r(23) = .48$, $p < .05$, and between HES and BLEU $r(23) = .46$, $p < .05$. RMSE analysis showed a minor error of 0.273 for BERTScore, while the other metrics exhibited errors exceeding 0.5, indicating a significant difference.

## Readability analysis

Table 13 and Figures 7 and 8 illustrate the average difficulty ratings on a scale of 1 (most difficult) to 5 (easiest). The survey participants rated human-generated sentences as easier to read and slightly more appropriate than machine-generated sentences. The survey also asked the examinees to identify whether a sentence was created by a human or a machine, and the results showed that the opinions on the sentences created by machines were divided, while the human-created sentences were judged to be human in many cases. However, there were many cases where human-created sentences were misidentified as machine-created sentences, indicating that determining the creator of the self-explanation remains difficult. Moreover, 65% of the machine-generated questions were rated as human-generated, and 75% of the human-written questions were rated as human-generated in Table 14. However, in both cases, the human-generated model answers scored higher than the machine-generated ones.

Table 15 provides examples of instances where the machine was able to successfully extract all the knowledge components, as well as cases where it could not. For linear function questions, the scoring was based on the rubric presented in Table 2. For the equation, four rubrics were established: “matching all denominators”, “making appropriate transitions”, “calculating a binary equation as a linear equation”, and “deriving the answer by approximating”. The table compares the answers created by humans and those generated by the machine.

For the question involving a linear function, the human-created answer provided a step-by-step explanation and solution, including finding the area of triangle ABC, the coordinate of point P, and the coordinates of point Q. The machine-generated answer, on the other hand, made an incorrect assumption and provided an incomplete explanation.

**Table 11** The similarity evaluation between human-generated and machine-generated sentences: F1 metrics of BERTScore, BLEU, Rouge-1, Rouge-2, Rouge-L

| BERTScore | | BLEU | | ROUGE-1 | | ROUGE-2 | | ROUGE-L | |
|---|---|---|---|---|---|---|---|---|---|
| M | SD | M | SD | M | SD | M | SD | M | SD |
| **0.719** | **0.032** | 0.300 | 0.093 | **0.443** | **0.232** | 0.235 | 0.194 | 0.384 | 0.198 |

**Table 12** RMSE and correlations between human evaluation score and the similarity evaluation score of each metric

| | BERTScore | BLEU | ROUGE-1 | ROUGE-2 | ROUGE-L |
|---|---|---|---|---|---|
| Correlations | **0.48**** | **0.46**** | 0.11 | 0.34* | 0.23 |
| RMSE | **0.273** | **0.582** | 0.510 | 0.655 | 0.533 |

*Note. **p < 0.05, *p < 0.1*

**Table 13** Readability of self-explanations

| Label | Means of readability | SD |
|---|---|---|
| Machine-generated | 3.006 | 0.276 |
| Human-generated | 3.823 | 0.200 |
| Overall | 3.415 | 0.404 |

**Table 14** The average of ratings of human-generated or machine-generated

| Label | Rated as human | Rated as machine-generated |
|---|---|---|
| Machine-generated | 0.65 | 0.35 |
| Human-generated | 0.75 | 0.25 |

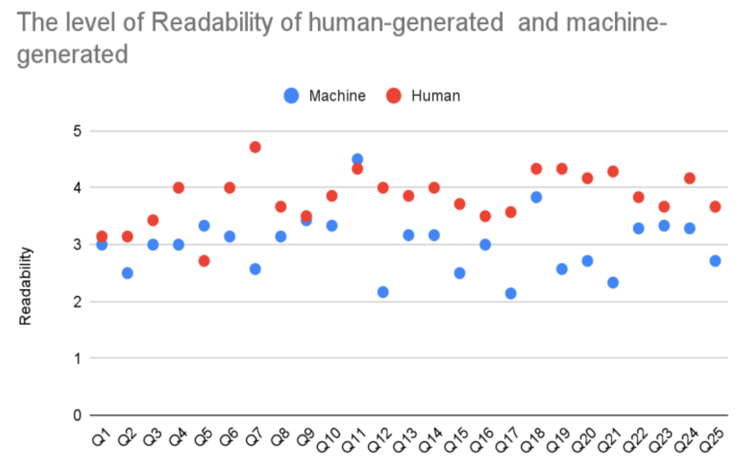


**Fig. 7** Student survey question: For each of 50 questions students are asked to rate if the question is (i) Human or machine-generated and (ii) the Difficulty level of each self-explanation on a scale between 5(easiest) and 1 (hardest)

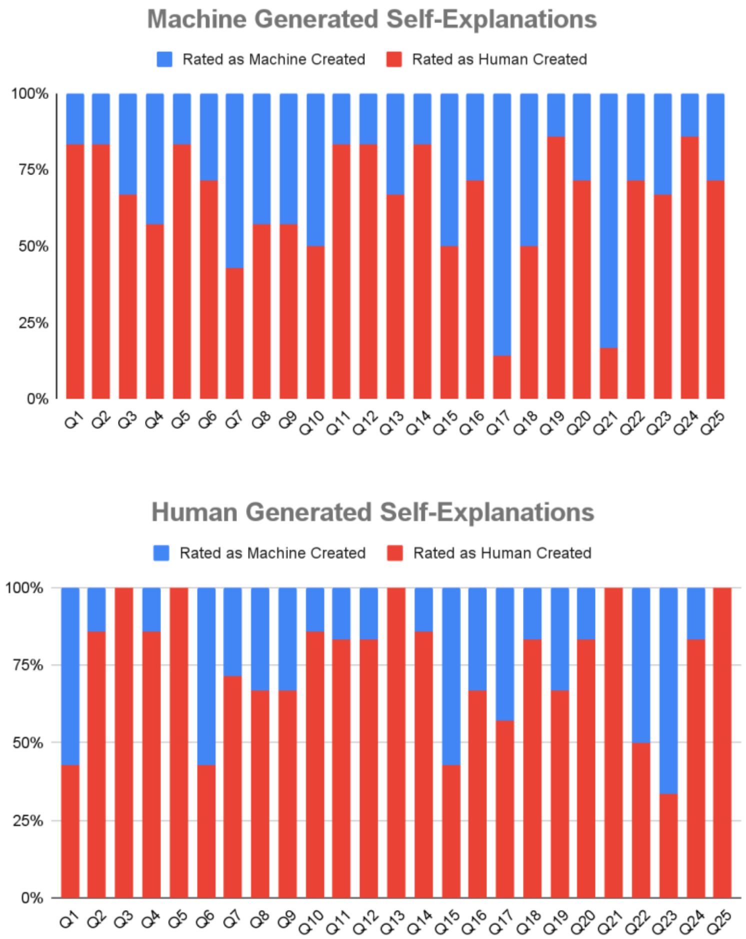


**Fig. 8** The evaluations of machine generated self-explanations and human generated self-explanations

For the equation question, the human-created answer showed a clear understanding of the problem and applied the correct mathematical concepts to derive the solution. In contrast, the machine-generated answer made a calculation error and provided an incomplete explanation.

In summary, the machine’s performance varied depending on the complexity of the question and the level of knowledge components required for the solution. While the machine was able to successfully extract all the knowledge components for some questions, it struggled with others. Further improvements are needed to enhance the machine’s performance in solving complex math problems.

**Table 15** Comparison of human-created answers and machine-generated answers

| Type | Linear function | Equation |
|---|---|---|
| Question | Find the coordinates of the point Q when the line OP bisects the triangle ABC. | If (-a+b)/3 - (a+2b)/4 + (2a+b)/5 = 0, find the value of (4a-3b)/(a-b). |
| Result in | Missing knowledge components: 0<br>BERTScore: 0.74 | Missing knowledge components: 2<br>BERTScore: 0.706 |
| Human-created | Line OP bisects triangle ABC, so the area of triangle OPC is 20/2, which is 10. Find the area of triangle ABC to get the area of triangle OPC. Let the coordinate of point P be (a, -1/2a+4). To express the area of △QAC using t, subtract the area of △OAC from the sum of the areas of △AQO and △OQC. If the x-coordinate of Q is t, the y-coordinate can be put as 5/6t | To get rid of the denominator, I multiplied both terms by 60, the least common multiple of the denominator. Since we paid the denominator, we are calculating integers. The value of a is expressed in terms of b so that the two-way equation can be calculated as a one-way equation. a = 2/11b is substituted for the value of an in the calculation. Substitute a = 2/11b for the value of a and calculate |
| Machine-generated | Since the whole is five and, the line OP bisects the area of triangle ABC, the area of quadrilateral ABPO and triangle POC is 2/5. Considering the coordinates of points A and C, I found the coordinates of point P. Substituting the Q value (p,q) into the equation of OP, then q=5/6p, so Q(p,5/6p). If the area of triangle ABO is 1, then the area of triangle AOC is 4. Considering the coordinates of point O and point P, I found the coordinates of point Q. | Do calculations. First, multiply both sides by 60. I rewrote the equation with all the denominators in place and moved 2b to the right of = to represent a by b |

*Note. BERTScore represents the similarity between human-generated and machine-generated sentences.*

## Discussion and limitations

We conducted a study in 2021 using 1,434 self-explanations from 25 quizzes to evaluate the machine's ability to generate sample self-explanations automatically. Our study aimed to answer three research questions. Firstly, we found that the machine generated sample answers that were 72% accurate in approximating human-generated knowledge components. Secondly, we evaluated the similarity between human-generated and machine-generated sentences using various metrics and found that BERTScore exhibited the highest similarity metric of 0.719, followed by ROUGE-1 with an average of 0.443. Furthermore, we found a moderate correlation between HES and BERTScore, as well as between HES and BLEU. Thirdly, we found that there was a difference in the readability of text, with the average human score being 3.8 and 3.0 for the machine. Interestingly, 65% of the respondents thought that the sentences created by the machine were human, while 25% of the human-created answers were mistaken as machine-generated. This experiment has implications for various information system developments, such as estimating user comprehension, automatically scoring self-explanations, and developing tools to promote awareness through self-explanation on the web.

The 72% accuracy rate achieved in this study can serve as a benchmark for future research, although it is challenging to compare with previous benchmarks since none have been established to date. While accuracy is critical, the objective of this study was to facilitate learners' cognitive processes rather than to achieve high accuracy rates. Thus, even incorrect self-explanations can be valuable in providing learners with insights, and the impact of low accuracy rates is minimal.

Our findings suggest that BERTScore may be a more accurate metric for evaluating the similarity between human-generated and machine-generated sentences compared to other commonly used metrics. These findings are in line with previous research (Zhang et al., 2020). BERTScore's use of contextual embedding provides a more effective means of capturing sentence meaning than relying on n-gram matching with BLEU. Furthermore, we believe that BERTScore's context-aware vectors mitigate the risk of unfairly inflating scores for candidates with many overlapping words, which is a drawback of ROUGE-1.

## Relationship between knowledge component extraction and similarity

The BERTScore showed that machine-generated sentences were similar to human-generated sentences semantically. However, the score for knowledge component extraction was still low despite the high similarity. This paragraph explores why this is the case and considers two factors that affect knowledge component extraction: mathematical units and clustering components.

In the example of linear functions, mathematical units such as "triangle" and "line segment" were frequently used and recognized as nodes, making it relatively easy to extract step-by-step solutions. However, in the case of equations, the thought process of solving was not extracted successfully. Many symbols such as "X" and "=" were used, making it difficult to recognize them as chunks of meaning.

While 20% of the dataset was from linear functions and triangle formulas, 80% were from factorization, square roots, and other equations. This affected the evaluation index, and it may be necessary to examine the suitability of the target problems for "knowledge component extraction" in the future. The second factor is some cases where extraction does not work well when generating semantic clusters in the model. The clustering component calculated the sum of squared errors and automatically set the number of clusters at 2 to 6. However, in some quizzes, the sum of squared errors exceeded 3000, and the semantic coherence was not well established. This is because the number of dimensions of the sentences encoded by BERT is 768, so the number would inevitably be large.

While the score for knowledge component extraction is moderate, it is not enough to replace human evaluation completely. The results suggest that the dependence on the dataset may also affect the evaluation index. Therefore, it is essential to consider different factors such as mathematical units, clustering components, and dataset dependence when using BERTScore for evaluation.

## Readability analysis

The study found that determining whether a text was human or machine-generated was not straightforward, and some machine-generated texts were judged to be human. However, interpretation of the content was still necessary, as self-explanation sentences were difficult to read and understand, regardless of whether they were made by humans or machines. The study also suggested that some puzzling points in machine-made texts may hinder comprehension compared to human-made texts that were generally easy to read.

These results are not extremely low compared to previous research. While there is no research on automatic generation in the field of self-explanatory mathematics, there are studies on text generation in mathematics. Drori et al. (2021) found that when measuring whether automatically generated math problems were created by humans or machines, students consistently rated human-created questions as more readable. In other words, at this stage of research, further investigation is needed to generate more human-like questions and texts in the complex domain of mathematics.

The use of the extraction model may have made the sentences difficult to read due to the prefixes connecting them to the original sentences. When combined with the element extraction results, it is possible that the necessary elements for the problem were extracted, but the sentences were not connected well, resulting in machine-like sentences that were difficult to read. To improve the quality of the machine-generated sentences, the study suggests post-processing the extracted sentences more smoothly or adding an element of abstraction to the model.

## Dataset dependencies

The proposed model is designed to generate sentences from collected data using a bottom-up approach. However, the model’s effectiveness is heavily reliant on the quality and quantity of the data used in the process. The study found that missing components in the data can significantly impact the model’s sentence generation quality. Additionally, the type of problem being addressed can also influence the model’s performance. While the model met the minimum requirements, it is evident that further improvements are necessary to enhance its overall effectiveness.

Despite these challenges, the model’s performance in extracting information from data with a quality score of 4 or higher was impressive. Although lower quality data may have been omitted during preprocessing, the model's ability to extract valuable sentences from good data was evident. Moving forward, the focus should be on developing strategies for collecting high-quality data and refining the preprocessing model to differentiate between good and bad data. Ultimately, this will lead to more accurate and effective sentence generation.

## Future work

This study investigated the feasibility of generating standardized sample answers. This method could be applied to three main tasks to support the use of self-explanation in learning: (A) identifying knowledge gaps in students' self-explanations, (B) providing support to learners who struggle with a specific problem, and (C) creating self-explanation scaffold templates using sample sentences. We proposed a new bottom-up approach to analyzing self-explanation data to generate standard solution examples, which aims to enhance learning through active thinking about mathematical concepts and procedures. We ensured that the generated standard sample answers are readable and effective, and further experiments are needed to assess its quality from a student's perspective.

To expand on this research, future studies should consider increasing the sample size and utilizing a variety of quizzes to facilitate deeper analysis and generalize the findings. Additionally, it is recommended to investigate the suitability of the target problem for knowledge component extraction and explore the generation of multiple solutions instead of relying on a single model answer. A comprehensive understanding of the effectiveness of this method can be gained by comparing it with template-based self-explanations, as proposed by Berthold et al. (2009). Template-based self-explanations involve providing students with predefined templates to assist them in constructing their self-explanations. Conducting such a comparison would enable a thorough evaluation of the advantages and limitations of each approach, thus identifying the most effective instructional strategies.

Furthermore, it would be valuable to explore the transferability of this method to different domains, such as English, in order to gain insights into its broader applicability. Considering alternative data sources, such as keystrokes instead of pen strokes, could also yield more accurate and relevant results. Exploring these possibilities would contribute to the overall advancement of the field of educational technology.

**Abbreviations**

BERT: Bidirectional Encoder Representations from Transformers; BERTScore: BERT based text generation evaluation Score; BLEU: BiLingual Evaluation Understudy; HES: Human Evaluation Score; LSA: Latent Semantic Analysis; NLP: Natural Language Processing; RMSE: Root Mean Squared Error; ROUGE: Recall-Oriented Understudy for Gisting Evaluation; SBERT: Sentence BERT.

**Acknowledgments**

Not applicable

### Authors' contributions

RN conducted the created machine learning models and data analysis and drafted the initial manuscript. BF, YD and KT provided insight, and editing of the manuscript. HO provided supervision for this research. All authors read and approved the final manuscript.

### Authors' information

Ryosuke Nakamoto is currently a doctoral student in the Graduate School of Informatics, Kyoto University. His research focuses on using the collected self-explanations and to examine how they can be used to support learning, including identifying students' stuck points and feedback to students in mathematics.

Brendan Flanagan is an Associate Professor at the Center for Innovative Research and Education in Data Science, Institute for Liberal Arts and Sciences, and the Graduate School of Informatics at Kyoto University. He received a bachelor's degree from RMIT University and master's and Ph.D. degrees from the Graduate School of Information Science and Electrical Engineering, Kyushu University. His research interests include: Learning Analytics, Educational Data Science, Educational Data Mining, NLP/Text Mining, Machine Learning, Computer Assisted Language Learning, and the Application of Blockchain in Education.

Yiling Dai is a Program-Specific Researcher at the Academic Center for Computing and Media Studies, Kyoto University. She received a Bachelor's degree from Zhejiang University, a Master's degree from the Graduate School of Business, Rikkyo University, and a PhD degree from the Graduate School of Informatics, Kyoto University. Her research interests include: Information Retrieval, Knowledge Discovery, Educational Data Mining and Learning Analytics.

Kyosuke Takami is a senior researcher at Education Data Science Center, National Institute for Educational Policy Research (NIER) in Japan. He has prior expertise in the application of neuroscience in education and received a PhD degree from Osaka university. He has been a high school teacher for more than 9 years. His research interests include: Learning Analytics, Educational Data Mining and Educational Technologies.

Hiroaki Ogata is a full Professor at the Academic Center for Computing and Media Studies, Kyoto University. His research interests include: Learning Analytics, Evidence-Based Education, Educational Data Mining, Educational Data Science, Computer Supported Ubiquitous and Mobile Learning, and CSCL.

**Funding**

This work was partly supported by JSPS Grant-in-Aid for Scientific Research (B) JP20H01722 and JP23H01001, (Exploratory) JP21K19824, (Early Career) JP23K17012, (A) JP23H00505, and NEDO JPNP20006.

### Availability of data and materials

Not applicable.

## Declarations

### Competing interests

The author declares no competing interests.

**Author details**

[1] Graduate School of Social Informatics, Kyoto University, Kyoto, Japan

[2] Center for Innovative Research and Education in Data Science, Institute for Liberal Arts and Sciences, Kyoto University, Kyoto, Japan

[3] Academic Center for Computing and Media Studies, Kyoto University, Kyoto, Japan

[4] Education Data Science Center, National Institute for Educational Policy Research (NIER), Tokyo, Japan

Received: 8 October 2022 Accepted: 21 July 2023
Published online: 1 January 2024 (Online First: 16 August 2023)

### References

Arner, T., McCarthy, K., & McNamara, D. (2021). iSTART StairStepper—Using comprehension strategy training to game the test. *Computers*, *10*(4), 48. https://doi.org/10.3390/computers10040048

Berthold, K., & Renkl, A. (2009). Instructional aids to support a conceptual understanding of multiple representations. *Journal of Educational Psychology*, *101*(1), 70–87. https://doi.org/10.1037/a0013247

Berthold, K., Eysink, T. H., & Renkl, A. (2009). Assisting self-explanation prompts are more effective than open prompts when learning with multiple representations. *Instructional Science*, *37*(4), 345–363. https://doi.org/10.1007/s11251-008-9051-z

Bisra, K., Liu, Q., Nesbit, J. C., Salimi, F., & Winne, P. H. (2018). Inducing self-explanation: A meta-analysis. *Educational Psychology Review*, *30*(3), 703–725. https://doi.org/10.1007/s10648-018-9434-x

Boonthum, C., Levinstein, I. B., & McNamara, D. S. (2007). Evaluating self-explanations in iSTART: Word matching, latent semantic analysis, and topic models. In A. Kao & S. R. Poteet (Eds.), *Natural language processing and text mining* (pp. 91–106). Springer, London. https://doi.org/10.1007/978-1-84628-754-1_6

Chamberland, M., Mamede, S., St-Onge, C., Setrakian, J., Bergeron, L., & Schmidt, H. (2015). Self-explanation in learning clinical reasoning: The added value of examples and prompts. *Medical Education*, *49*(2), 193–202. https://doi.org/10.1111/medu.12623

Chi, M., Leeuw, N., Chiu, M., & Lavancher, C. (1994). Eliciting self-explanations improves understanding. *Cognitive Science*, *18*(3), 439–477. https://doi.org/10.1207/s15516709cog1803_3

Conati, C., & Vanlehn, K. (2000). Toward computer-based support of meta-cognitive skills: A computational framework to coach self-explanation. *International Journal of Artificial Intelligence in Education*, *11*(4), 389–415.

Crippen, K. J., & Earl, B. L. (2007). The impact of web-based worked examples and self-explanation on performance, problem solving, and self-efficacy. *Computers & Education*, *49*(3), 809–821. https://doi.org/10.1016/j.compedu.2005.11.018

Crossley, S. A., Kim, M., Allen, L., & McNamara, D. (2019). Automated summarization evaluation (ASE) using natural language processing tools. In S. Isotani, E. Millán, A. Ogan, P. Hastings, B. McLaren & R. Luckin (Eds.), *Artificial Intelligence in Education. AIED 2019. Lecture Notes in Computer Science, vol 11625* (pp. 84–95). Springer, Cham. https://doi.org/10.1007/978-3-030-23204-7_8

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In J. Burstein, C. Doran & T. Solorio (Eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* (pp. 4171–4186). Association for Computational Linguistics. https://aclanthology.org/N19-1423.pdf

Drori, I., Tran, S., Wang, R., Cheng, N., Liu, K., Tang, L., Ke, E., Singh, N., Patti, T., Lynch, J., Shporer, A., Verma, N., Wu, E., & Strang, G. (2021). A neural network solves and generates mathematics problems by program synthesis: Calculus, differential equations, linear algebra, and more. *Proceedings of the National Academy of Sciences*. National Academy of Sciences. https://www.cs.columbia.edu/~idrori/drori2021math.pdf

Erkan, G., & Radev, D. (2004). LexRank: Graph-based lexical centrality as salience in text summarization. *Journal of Artificial Intelligence Research*, *22*(1), 457–479.

Flanagan, B., & Ogata, H. (2018). Learning analytics platform in higher education in Japan. *Knowledge Management & E-Learning*, *10*(4), 469–484. https://doi.org/10.34105/j.kmel.2018.10.029

Flanagan, B., Takami, K., Takii, K., Dai, Y., Majumdar, R., & Ogata, H. (2021). EXAIT: A symbiotic explanation education system. In M. M. T. Rodrigo et al. (Eds.), *Proceedings of the 29th International Conference on Computers in Education* (pp. 404–409). Asia-Pacific Society for Computers in Education.

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, *76*(5), 378–382. https://doi.org/10.1037/h0031619

Hodds, M., Alcock, L., & Inglis, M. (2014). Self-explanation training improves proof comprehension. *Journal for Research in Mathematics Education*, *45*(1), 62–101. https://doi.org/10.5951/jresematheduc.45.1.0062

Iqbal, T., & Qureshi, S. (2020). The survey: Text generation models in deep learning. *Journal of King Saud University - Computer and Information Sciences*, *34*(6), Part A, June 2022, 2515–2528. https://doi.org/10.1016/j.jksuci.2020.04.001

Jackson, G. T., Guess, R. H., & McNamara, D. S. (2010). Assessing cognitively complex strategy use in an untrained domain. *Topics in Cognitive Science*, *2*, 127–137.

León, J. A., Olmos, R., Escudero, I., Cañas, J. J., & Salmerón, L. (2006). Assessing short summaries with human judgments procedure and latent semantic analysis in narrative and expository texts. *Behavior Research Methods*, *38*(4), 616–627. https://doi.org/10.3758/bf03193894

Lin, C.-Y. (2004). ROUGE: A package for automatic evaluation of summaries. In M.-F. Moens & S. Szpakowicz (Eds.), *Text summarization branches out* (pp. 74–81). Association for Computational Linguistics. https://aclanthology.org/W04-1013.pdf

McNamara, D. S., Levinstein, I. B., & Boonthum, C. (2004). iSTART: Interactive strategy training for active reading and thinking. *Behavior Research Methods, Instruments, & Computers*, *36*(2), 222–233.

Nakamoto, R., Flanagan, B., Takami, K., Dai, Y., & Ogata, H. (2021). Identifying students' stuck points using self-explanations and pen stroke data in a mathematics quiz. In M. M. T. Rodrigo et al. (Eds.), *Proceedings of the 29th International Conference on Computers in Education* (pp. 521–530). Asia-Pacific Society for Computers in Education.

Ozsoy, M. G., Alpaslan, F. N., & Cicekli, I. (2011). Text summarization using Latent Semantic Analysis. *Journal of Information Science*, *37*(4), 405–417. https://doi.org/10.1177/0165551511408848

Panaite, M., Dascalu, M., Johnson, A., Balyan, R., Dai, J., McNamara, D., & Trausan-Matu, S. (2018). Bring it on! Challenges encountered while building a comprehensive tutoring system using ReaderBench. In C. P. Rosé, R. Martínez-Maldonado, H. U. Hoppe, R. Luckin, M. Mavrikis, K. Porayska-Pomsta, B. McLaren, B. du Boulay (Eds.), *Artificial Intelligence in Education. AIED 2018. Lecture Notes in Computer Science, vol 10947* (pp. 409–419). Springer, Cham. https://doi.org/10.1007/978-3-319-93843-1_30

Panaite, M., Ruseti, S., Dascalu, M., Balyan, R., McNamara, D. S., & Trausan-Matu, S. (2019). Automated scoring of self-explanations using recurrent neural networks. In M. Scheffel, J. Broisin, V. Pammer-Schindler, A. Ioannou & J.

Schneider (Eds.), *Transforming Learning with Meaningful Technologies. EC-TEL 2019. Lecture Notes in Computer Science, vol 11722* (pp. 659–663). Springer, Cham. https://doi.org/10.1007/978-3-030-29736-7_61

Papineni, K., Roukos, S., Ward, T., & Zhu, W. (2002). BLEU: A method for automatic evaluation of machine translation. In P. Isabelle (Ed.), *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics* (pp. 311–318). Association for Computational Linguistics. https://doi.org/10.3115/1073083.1073135

Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing* (pp. 3982–3992). Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-1410

Renkl, A. (2017). Learning from worked-examples in mathematics: Students relate procedures to principles. *ZDM Mathematics Education*, *49*(4), 571–584. https://doi.org/10.1007/s11858-017-0859-3

Rittle-Johnson, B. (2006). Promoting transfer: Effects of self-explanation and direct instruction. *Child Development*, *77*(1), 1–15.

Rittle-Johnson, B. (2017), Developing mathematics knowledge. *Child Development Perspectives*, *11*(3), 184–190. https://doi.org/10.1111/cdep.12229

Rittle-Johnson, B., & Schneider, M. (2015). Developing conceptual and procedural knowledge of mathematics. In R. C. Kadosh & A. Dowker (Eds.), *Oxford handbook of numerical cognition* (pp. 1118–1134). Oxford University Press.

Rittle-Johnson, B., Loehr, A. M., & Durkin, K. (2017). Promoting self-explanation to improve mathematics learning: A meta-analysis and instructional design principles. *ZDM Mathematics Education*, *49*(4), 599–611.

Rittle-Johnson, B., Schneider, M., & Star, J. R. (2015). Not a one-way street: Bidirectional relations between procedural and conceptual knowledge of mathematics. *Educational Psychology Review*, *27*, 587–597. https://doi.org/10.1007/s10648-015-9302-x

Rittle-Johnson, B., Siegler, R. S., & Alibali, M. W. (2001). Developing conceptual understanding and procedural skill in mathematics: An iterative process. *Journal of Educational Psychology*, *93*, 346–362. https://doi.org/10.1037//0022-0663.93.2.346

Star, J. R. (2005). Reconceptualizing procedural knowledge. *Journal for Research in Mathematics Education*, *36*(5), 404–411.

Suzuki, M. (2019). *Pretrained Japanese BERT models*. https://github.com/cl-tohoku/bert-japanese

Thompson, D. R., & Senk, S. L. (1998). Using rubrics in high school mathematics courses. *Mathematics Teacher*, *91*(9), 786–793.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, *30*, 5998–6008.

Verschaffel, L., De Corte, E., & Vierstraete, H. (1999). Upper elementary school pupils' difficulties in modeling and solving nonstandard additive word problems involving ordinal numbers. *Journal for Research in Mathematics Education*, *30*(3), 265–285. https://doi.org/10.2307/749836

Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., & Artzi, Y. (2020). BERTScore: Evaluating text generation with BERT. In Proceedings of International Conference on Learning Representations (pp. 1–15). https://openreview.net/pdf?id=SkeHuCVFDr

## Publisher's Note

The Asia-Pacific Society for Computers in Education (APSCE) remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.