

Development of a Mathematics Learning Support System through the Analysis of Handwritten Responses and Self-Explanations

Ryosuke Nakamoto

Doctoral Dissertation

Development of a Mathematics Learning
Support System through the Analysis of
Handwritten Responses and Self-Explanations

Ryosuke Nakamoto

Doctoral Dissertation

Development of a Mathematics Learning Support System through the Analysis of Handwritten Responses and Self-Explanations

Ryosuke Nakamoto
Supervisor: Professor Hiroaki OGATA

Abstract

The rapid expansion of educational technology, driven by initiatives like the GIGA School program, has made digital tools such as online quizzes and digital textbooks indispensable in modern classrooms. However, while these tools are useful for verifying correct answers, they often fail to capture students' thought processes or identify where they encounter difficulties. To effectively support students, it is crucial to understand their learning journeys and detect the points where they struggle. Research has shown that self-explanation—a process where students verbalize their reasoning—deepens understanding and enhances problem-solving abilities. Systems like i-START have successfully applied self-explanation to improve reading comprehension, fostering deep learning and providing insights into students' cognitive processes. Despite these benefits, accurately detecting where students face difficulties remains a challenge, particularly in mathematics, where the analysis of self-explanations is labor-intensive for teachers, limiting the practical use of such activities in classrooms.

To address these challenges, this doctoral thesis proposes a novel system that integrates self-explanation with pen-stroke data analysis, enabling the automatic detection of student difficulties and the provision of personalized feedback. Unlike traditional systems, this approach goes beyond simply evaluating correctness; it analyzes students' handwritten responses and self-explanations to uncover thought processes and identify specific points of cognitive difficulty. The system also leverages real-time data from classmates and historical datasets to improve the accuracy of difficulty detection.

This doctoral thesis investigates the development and evaluation of a self-explanation-based system designed to enhance mathematics education through personalized feedback and adaptive learning techniques. The research is structured around four interrelated studies, each tackling key challenges in detecting learning impasses, generating model answers, scoring self-explanations, and delivering real-time feedback to improve learning outcomes.

1. Impasse Detection with Self-Explanation: The first study established a framework for identifying learning impasses by analyzing students' problem-solving sequences using self-explanations, model answers, and pen-stroke data. The system achieved an 80% accuracy rate, laying the foundation for targeted feedback. However, relying solely on self-explanations limited accuracy due to variability in student responses. Incorporat-

ing additional features, such as textbook model answers, was identified as a potential enhancement to improve consistency and detection accuracy.

2. Automatic Generation of Model Self-Explanations: The second study developed a model to generate exemplary answers based on self-explanations collected in Study 1, achieving 72% coverage of essential knowledge components. However, only 64% of the generated answers adhered to the correct rubric step order, highlighting significant challenges in maintaining logical coherence and step-by-step reasoning. These results emphasize the inherent difficulty of processing open-ended self-explanations, where variability in student responses complicates the alignment of knowledge components with structured, sequential reasoning. This limitation underscores the need to explore hybrid approaches that integrate structured resources, such as textbook model answers or predefined templates, to guide the generation process while preserving the flexibility of self-explanations.

3. Data Augmentation and Scoring for Self-Explanation Analysis: The third study introduced data augmentation techniques using mathematical texts and Large Language Models (LLMs) to enhance scoring reliability. Models enriched with mathematical text datasets achieved the highest accuracy, outperforming those using LLM-generated data or mixed datasets. However, inconsistencies in LLM-generated data, particularly without fine-tuning, often compromised test performance. This suggests that LLM outputs should be selectively used and carefully curated to avoid introducing noise.

4. Real-Time Feedback for Enhancing Self-Explanation Skills: The fourth study presented the Self-Explanation-Based Automated Feedback (SEAF) system, providing immediate, personalized feedback to help students refine reasoning and sustain motivation. While initially effective for students with lower self-explanation skills, feedback quality diminished over time. This emphasizes the need for adaptive mechanisms that evolve with learners' needs to maintain engagement and effectiveness over the long term.

This thesis identifies critical limitations across the four studies: reliance on self-explanations and pen-stroke data alone reduced accuracy in detection and model generation; LLM-generated data introduced inconsistencies; and long-term feedback showed a decline in quality. Feedback systems must be carefully designed to suit specific use cases and timeframes while maintaining their relevance over extended periods. The findings underscore the need for feedback mechanisms that are both structured and adaptable,

supported by fine-tuned data generation and self-explanation approaches tailored to the learning context.

By addressing these limitations, this research highlights how educational technologies can be refined to become more accurate, reliable, and flexible, meeting the diverse needs of students in mathematics education. Future work will focus on enhancing these systems to better support a wide range of learners while ensuring scalability, adaptability, and sustained impact across varied educational contexts.

Contents

1	Introduction	1
1.1	Background of the Research	1
1.2	Issues and Proposal	2
1.3	Structure of This Thesis	3
2	Literature Review	6
2.1	Advancements in Educational Technology and Big Data	6
2.2	Self-Explanation in Math Learning	7
2.2.1	Theoretical Foundations	7
2.2.2	Research utilizing the Uniqueness of Self-explanations	7
2.2.3	The Difficulty of Processing Self-explanations	8
2.2.4	Self-Explanation and Handwriting Input Analysis in Mathematics Education	8
2.3	Challenges in Utilizing Self-Explanation for Impasse Detection	9
3	Preliminaries	11
3.1	Explanation of Data Collection and Analysis Methods	11
3.1.1	LEAF System	11
3.1.2	LEAF Framework	12
3.2	Data Collection and Labeling	12
3.2.1	Self-explanation and Penstroke Data Collection	12
3.3	Self-explanation Quality Scoring	15
3.3.1	Settings	15
3.3.2	Self-explanation Quality Scoring with Rubric and Index	16
3.4	Research Approach and Contribution	17

4	Experimental Study(1): Impasse Detection Methods with Self-explanation	19
4.1	Background and Motivation	19
4.1.1	Learning Impasse	20
4.1.2	Premise Conditions and the Definition of Impasse in Math Learning	21
4.2	Proposed Method	22
4.2.1	Rubric-Based Evaluative Framework	22
4.2.2	Labeling Impasses through Human Assessment	22
4.2.3	Feature Extraction from Handwriting Data Associated with Each Self-Explanation	23
4.3	Hierarchical Logistic Regression Analysis	24
4.3.1	Hierarchical Logistic Regression Model	24
4.3.2	Analysis of Self-Explanation Features (M_S , M_{S+G} , M_F)	26
4.3.3	Exploring General Learning Operations and The Impact of Memo and Quiz Actions (M_G , M_A)	28
4.3.4	Comprehensive Analysis in Final Model (M_F)	29
4.4	Discussion	29
4.4.1	Identifying Learning Impasses	29
4.4.2	Learning Difficulties in Online Mathematics Education in MF . . .	30
4.4.3	Limitations and Future Directions	31
5	Experimental Study(2): Automatic Generation of Exemplary Self-Explanations	32
5.1	Introduction	32
5.1.1	Motivation	32
5.1.2	Previous Research	33
5.2	Goals of This Study	34
5.2.1	Datasets	35
5.2.2	Preprocessing	38
5.3	Sample Answer Generation Model	38
5.3.1	Overall Architecture of the Proposed Model	38
5.3.2	Vectorizing Component	39
5.3.3	Clustering Component	40
5.3.4	Extracting Components	41
5.4	Experimental Setup	41

5.4.1	Grading Extracted Knowledge Components	41
5.4.2	Quality Evaluation with Metrics	42
5.4.3	Readability Analysis	43
5.5	Results	43
5.5.1	Knowledge Components Extraction Grading	43
5.5.2	Quality Evaluation with Metrics	44
5.6	Readability Analysis	44
5.7	Discussion and Limitations	49
5.7.1	Relationship between Knowledge Component Extraction and Similarity	50
5.7.2	Readability Analysis	51
5.7.3	Dataset Dependencies	51
5.7.4	Conclusion	52
6	Experimental Study(3): Preprocessing Proposal - Employing Data Augmentation Techniques for Self-Explanation Autoscoring	54
6.1	Introduction	54
6.2	Related Work	55
6.2.1	Automated Scoring of Self-Explanations: The Imperative for Rich Data	55
6.2.2	Augmenting Mathematical Self -Explanations Using Large Language Models	55
6.2.3	Proposal	56
6.3	Problem Setting: The Learning Task	57
6.3.1	Assessment of Self-Explanation Quality	57
6.3.2	The Text Regression Model Description	58
6.4	The Proposed Method	58
6.5	Pseudo-Labeling Training Algorithm	59
6.5.1	Dataset Categorization, Function Definitions, and Model Learning .	59
6.5.2	Pseudo-Data Preparation: LLM Usage and Mathematical Material	62
6.5.3	Comparative Analysis of Original and LLM-Generated Dataset . . .	63
6.6	Experiments and Evaluations	67

6.6.1	Exploring the Influence of Self-Explanation Augmentation on Model Efficiency	67
6.6.2	Evaluating Optimal Quantity of Pseudo-Self-Explanation Data . . .	68
6.7	Discussion	71
6.7.1	Detailed Analysis of Results (RQ1)	71
6.7.2	Findings and Observations (RQ2)	71
6.7.3	Limitations	72
6.7.4	Conclusion	73
7	Experimental Study(4): Automatic Feedback for Enhancing Self-Explanation Skills	74
7.1	Introduction	74
7.2	Previous Studies	76
7.2.1	Feedback System and Our Positioning	76
7.2.2	Consideration on Existing Intelligent Tutoring Systems	77
7.2.3	Conceptual Framework and Research Inspiration	78
7.2.4	Positioning of the Research and Its Novelty	78
7.3	The Architecture of the SEAF System	79
7.3.1	Feedback Delivery	79
7.3.2	Data Processing	80
7.3.3	Personalized Messages	80
7.3.4	Self-explanation Score	83
7.3.5	Classmates' Self-explanations Reference	85
7.4	Analytical Methods and Framework for Analysis	87
7.4.1	Experiment Methodology	88
7.4.2	Analysis Preparations	89
7.5	Results of SEAF Learning Effects Experiments	91
7.5.1	Results of Statistical Analysis Based on Self-Explanation Level . . .	91
7.5.2	Repeated Measures ANOVA	91
7.5.3	Perception Analysis	93
7.6	Discussion	94
7.6.1	RQ1:Can a classmate's self-explanations be presented and useful to the learner as feedback?	94

7.6.2	RQ2: Which students benefit most from using the SEAF system? . . .	94
7.6.3	Limitations	95
7.6.4	Main Findings	96
7.6.5	Future Directions	97
7.6.6	Conclusion	98
8	Conclusion and Future work	99
8.1	Main Findings of Each Study	99
8.1.1	Study 1: Impasse Detection	99
8.1.2	Study 2: Automated Model Answer Generation	100
8.1.3	Study 3: Data Augmentation with Math Texts and Large Language Models	100
8.1.4	Study 4: Self-Explanation-Based Automated Feedback (SEAF) . . .	101
8.2	Overall Discussion and Limitations of the Study	101
8.3	Implications and Future Work	103
8.4	Conclusion	104
	References	106
	Appendices	113
	Institutional Review Board (IRB) Approval and Ethical Considerations	114

List of Figures

1.1	Research Overview	4
1.2	Research Approach Cycle	4
3.1	Architecture of LEAF (Majumdar et al., 2019)	12
3.2	BookRoll user interface: handwritten answer (left), quiz and quiz answer report (right)	14
3.3	Handwritten answer review playback and self-explanation input user interface	14
4.1	Labeling Method	23
4.2	Self-Explanation and Handwriting Example with Impasses Highlighted . .	24
5.1	An overview of the process of our proposed model, and the inputs and outputs	35
5.2	Histogram of the distribution of self-explanation responses over the quizzes in the dataset	36
5.3	A boxplot of the self-explanation quality score distribution	37
5.4	Workflow for the self-explanation processing	39
5.5	Overall of a proposed model architecture	40
5.6	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)	46
5.7	The evaluations of Machine Generated Self-Explanations and Human Gen- erated Self-Explanations	47
6.1	Workflow for the BERT regression models	58
6.2	The overview of the pseudo-training process.	59
6.3	The Detail of Pseudo Training Process	60

6.4	Boxplots of self-explanation scores.	64
6.5	Validation MAE plot	69
6.6	Test MAE plot	70
7.1	The architecture of the SEAF system	79
7.2	Personalized messages.	82
7.3	Missing knowledge detection.	84
7.4	The scatter plot for the self-explanation analysis score.	85
7.5	Reference examples of classmates' self-explanation.	86
7.6	Distribution of self-explanation Quality means in each experimental period group.	92
7.7	Results Plots of the SEAF usage survey	93

List of Tables

3.1	Definition of words	16
3.2	Rubrics and a sample answer of self-explanation in a quiz	16
3.3	Self-explanation quality score grading definition	17
3.4	Descriptive statistics of the collected self-explanations(Samples)	17
3.5	Examples of Self-Explanations with High and Low quality scores	18
4.1	Hierarchical Logistic Regression Features	25
4.2	Hierarchical Logistic Regression Model Groups	26
4.3	Logistic Regression Results (Significant Indicators)	27
5.1	Distribution of human graded quality scores	36
5.2	Explanatory notes to evaluators assessing the readability of human and generated examples.	42
5.3	Missing knowledge components of each quiz by human evaluation.	44
5.4	Descriptive statistics of extracted self-explanations' quality score.	44
5.5	The similarity evaluation between human-generated and machine-generated sentences: F1 metrics of BERTScore, BLEU, Rouge-1, Rouge-2, Rouge-L	44
5.6	RMSE and correlations between human evaluation score and the similarity evaluation score of each metric	45
5.7	Readability of self-explanations	45
5.8	The average of ratings of human-generated or machine-generated	45
5.9	Comparison of human-created answers and machine-generated answers	48
6.1	Descriptive Statistics of Graded Self-Explanations	57
6.2	Statistics for Original and Generated Datasets	60
6.3	Algorithm 1: Semi-supervised learning with pseudo-labeling	61
6.4	Comparative metrics for original and generated datasets.	64

6.5	LLM-generated samples	65
6.6	Math Texts samples	66
6.7	Datasets Overview	67
6.8	Model Performance for Various Datasets (MAE)	68
6.9	Validation MAE with varying amounts of added pseudo-self-explanatory data	69
6.10	Test MAE with varying amounts of added pseudo-self-explanation data . .	70
7.1	Comparative analysis of self-explanation features in prominent intelligent tutoring systems	78
7.2	Descriptive statistics of self-explanation features by group and level.	91
7.3	The Friedman test results with repeated treatments at different experimen- tal periods	92
7.4	The result of Conover's post hoc test (P-Values)	92
7.5	Results of the SEAF usage survey	93

Chapter 1

Introduction

1.1 Background of the Research

The rapid growth of digital education tools, along with initiatives like the Global and Innovation Gateway for All (GIGA), has significantly changed classroom dynamics. The widespread distribution of individual devices to students and the increasing use of online quizzes and digital textbooks have made education more accessible, interactive, and data-driven. However, despite their benefits, many of these tools focus primarily on whether students' answers are correct, without delving deeply into their thought processes or identifying where they encounter difficulties during their learning process.

A promising area of research is learning analytics (LA), which leverages the data generated during students' learning activities to offer insights into their cognitive processes. By analyzing learning logs, predictive models can identify where students may need support and offer timely interventions to improve learning outcomes (Ferguson, 2012; Chen et al., 2021; Siemens, 2012). However, many current systems remain limited by focusing only on answer correctness, without capturing the reasoning or struggles behind students' responses.

Understanding students' cognitive processes is crucial for identifying learning *impasses*—moments when learners recognize gaps in their understanding, prompting deeper engagement and problem-solving (J. S. Brown and VanLehn, 1980; VanLehn et al., 1992; VanLehn et al., 2002). These impasses are critical for learning, as they trigger reflection and correction, leading to more significant learning gains (VanLehn and Jones, 1993). Identifying these moments of struggle can enable instructional interventions that effectively address learning gaps.

Many educational tools, however, fail to capture these moments of cognitive struggle. They primarily focus on verifying answers instead of understanding why students struggle. Prior research has explored methods to detect impasses through indicators like time intervals between actions or proficiency data from digital platforms (Iiyama et al., 2017; Ochoa et al., 2013), but these approaches do not fully reveal the underlying cognitive processes. A more sophisticated method is required to detect and address these impasses effectively.

1.2 Issues and Proposal

Incorporating self-explanation into digital learning environments offers a potential solution to this challenge. Self-explanation, where students articulate their thought processes while solving problems, has been shown to foster deeper reflection and understanding, especially in subjects like mathematics (Rittle-Johnson et al., 2017; Bisra, Liu, Nesbit, et al., 2018). For instance, systems like i-START have successfully employed self-explanation to enhance reading comprehension by providing insights into students' cognitive processes (McNamara et al., 2004). Despite its proven effectiveness, self-explanation remains challenging in subjects like mathematics, where the cognitive difficulties that arise during problem-solving are often complex and difficult to detect.

Additionally, guiding and assessing students' self-explanations requires considerable effort from educators, making it difficult to scale. Current systems are not equipped to process the rich data students produce nor to integrate impasse detection with self-explanation activities without imposing a significant burden on teachers. As a result, the widespread adoption of self-explanation strategies remains limited.

To address these limitations, this thesis proposes an automated system that integrates self-explanation with advanced learning analytics to better detect learning impasses and provide real-time feedback. By analyzing students' problem-solving processes and incorporating data such as pen-stroke logs, the system aims to offer personalized, timely interventions that improve learning outcomes. The goal is not only to enhance impasse detection but also to provide a scalable solution for the broader implementation of self-explanation techniques in digital education.

1.3 Structure of This Thesis

This thesis is structured around four interconnected studies, each contributing to the development of an AI-driven system that enhances the detection of learning impasses and provides real-time feedback. The four studies are logically sequenced, with Study 1 and Study 4 serving as the core objectives of the research, and Study 2 and Study 3 supporting these goals by automating and scaling key tasks.

Study 1: Impasse Detection

The first study focuses on the primary goal of detecting learning impasses. By analyzing both textual self-explanations and pen-stroke data, this study aims to identify steps when students struggle during problem-solving. Detecting impasses early allows for targeted interventions, laying the foundation for improving learning outcomes.

Study 2: Model Answer Generation for Self-Explanation

To support the implementation of Study 1, Study 2 addresses the challenge of generating model answers for self-explanation tasks. This study uses unsupervised learning techniques to develop methods for automatically producing exemplary self-explanations based on students' prior learning logs.

Study 3: Automatic Scoring Model for Self-Explanations

To support the data collection of Study 2, Study 3 introduces an automatic scoring model that evaluates self-explanations, leveraging Large Language Models (LLMs) to scale up the data volume. This model addresses the scalability challenge by automating the evaluation of self-explanations, ensuring that feedback can be provided more efficiently and at scale.

Study 4: Real-Time Feedback System

The fourth study returns to the main objective of providing real-time feedback. Building on the results of the previous studies, this system continuously monitors students' self-explanations and problem-solving behaviors to offer personalized, adaptive support. This system enhances the learning process by offering timely interventions to help them overcome cognitive barriers. Study 4 thus completes the research's core goal of providing automated, real-time feedback to improve learning outcomes.

As illustrated in Figure 1.1, this research tackles the core challenge of advancing beyond traditional digital tools, which often focus only on correctness, to a more comprehensive approach that captures and analyzes students' cognitive processes. Figure 1.2

Overall Issues	Many digital tools only check answers, ignoring the critical need to track students' thought processes. Identifying difficulties, even with self-explanation, takes time.			Students struggle with writing self-explanations without training.
Problems	A detection system is needed.	Automatic generation of model answers is needed for detection.	High-quality data extraction is required	Support is needed for writing better self-explanations.
Aim	Develop a method to detect impasses in mathematical reasoning accurately.	Generate example self-explanations from collected students' answers.	Enhance accuracy of automatic scoring for self-explanations.	Develop adaptive feedback that not only corrects but also aids in writing self-explanations, improving data quality.
Research	<i>Study 1:</i> Developed a method to detect impasses in students' explanations by comparing them to model answers and analyzing pen strokes.	<i>Study 2:</i> Created a model to generate answers, identifying key knowledge for detecting learning issues.	<i>Study 3:</i> Proposed an automatic scoring model using data augmentation with LLMs and handwriting data.	<i>Future Study:</i> Use adaptive feedback to correct and guide better self-explanations.
Remaining Problems	Need sample self-explanations for automatic generation.	Extract high-quality data for preprocessing.	Collect more high-quality self-explanations.	Improve feedback system to personalize based on data.
Publications	ICCE 2021, RPTL 2024	AIED 2022, RPTL 2023	ICALT2024a,b	ICALT2024a, ILE2024(under review)

Figure 1.1: Research Overview

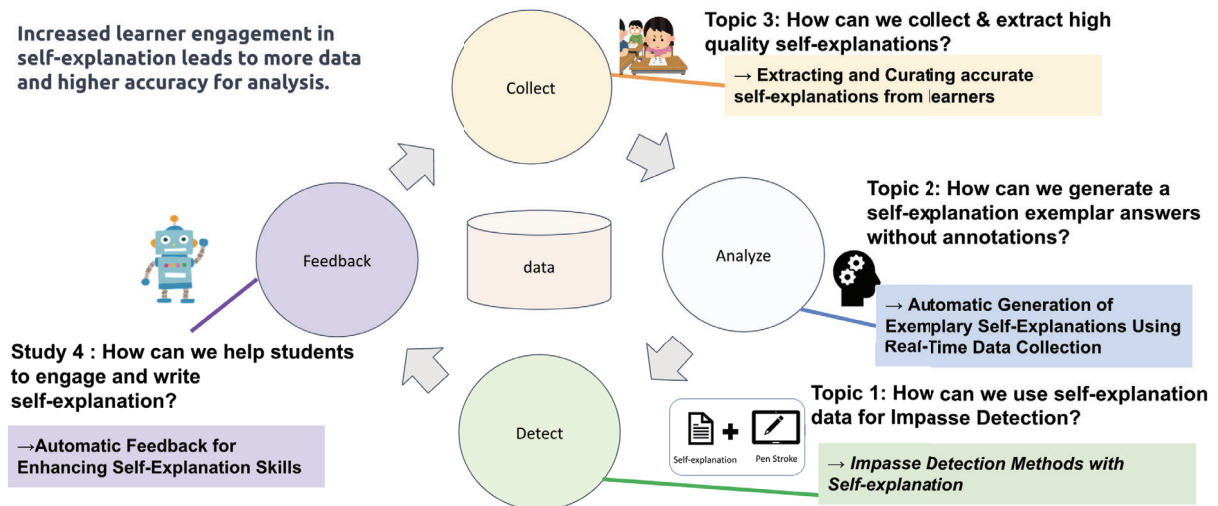


Figure 1.2: Research Approach Cycle

provides a visual representation of the continuous cycle of data collection, analysis, impasse detection, and feedback that powers the proposed system. This cycle is integral to the system's success: as students engage in deeper self-explanation, they produce richer data, which in turn leads to more accurate detection of cognitive struggles. This process enables the delivery of more personalized and timely interventions, ultimately leading to improved learning outcomes. Collectively, these studies form the basis of a comprehensive AI-driven system designed to detect learning impasses, generate model self-explanations, assess student responses, and deliver real-time feedback.

Chapter 2

Literature Review

2.1 Advancements in Educational Technology and Big Data

Educational big data has started to play a significant role in enhancing learning processes. This is particularly evident in web-based learning environments, where self-explanation strategies are integrated to assist learners. The emergence of digital learning platforms has significantly expanded the horizons of educational research, particularly in the realm of learning behaviors.

In parallel, the field of educational technology has focused on understanding and surmounting cognitive barriers known as "impasses" that obstruct the comprehension of mathematical concepts. Education today seeks not only to deliver content but also to foster deep cognitive engagement, holistic comprehension, and enduring curiosity among learners (Hattie et al., 1996). Central to this endeavor is the concept of "impasse," where learners encounter substantial barriers on their path to understanding. Extensive research into cognitive skill acquisition, without direct tutor intervention, consistently reveals the presence of impasses (J. S. Brown and VanLehn, 1980; Rubin, 1993; Siegler and Jenkins, 1989; VanLehn et al., 1992). An impasse arises when a student becomes acutely aware of a gap or deficiency in their comprehension, leading to feelings of uncertainty, error detection, or doubts about the accuracy of their actions. The significance of impasse detection lies in its direct relationship with the learning process. When students do not encounter impasses, their learning tends to be fragmented, regardless of the quality or depth of instructional explanations. Conversely, the emergence of an impasse serves as a catalyst for learning. In such instances, instructional interventions have demonstrated a profound

impact on learning outcomes. An impasse, by highlighting a knowledge deficit, motivates learners to actively bridge the gap through introspective reflection, consultation of resources, or seeking insights from peers and instructors. However, detecting impasses in unstructured tasks poses significant challenges, requiring substantial time and resources. This presents a burden for educators and limits widespread implementation.

2.2 Self-Explanation in Math Learning

2.2.1 Theoretical Foundations

The theoretical underpinnings of self-explanation in mathematics education emerge from the cognitive learning theories that emphasize the importance of metacognition and reflection in the learning process. Self-explanation has been consistently shown to facilitate mathematics learning, primarily through enhancing both conceptual and procedural knowledge (Renkl, 2017; Rittle-Johnson et al., 2017). Conceptual knowledge pertains to understanding mathematical principles such as equivalence, while procedural knowledge develops through engaging in problem-solving practices (Rittle-Johnson et al., 2001; Star, 2005). The act of self-explanation supports comprehension by aiding knowledge integration and directs learners' attention towards structural features of content, promoting deeper learning (M. T. H. Chi, 2000; McEldoon et al., 2013). Self-explanation has been shown to improve students' conceptual and procedural knowledge in math by helping them focus on important details (Bisra, Liu, Nesbit, et al., 2018; Renkl, 2017; Rittle-Johnson, 2017a). Conceptual knowledge refers to abstract ideas, while procedural knowledge is tied to specific problem-solving practices (Rittle-Johnson and Schneider, 2015; Rittle-Johnson et al., 2001, 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.

2.2.2 Research utilizing the Uniqueness of Self-explanations

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, devel-

oping models to assess understanding from self-explanation behavior, and using strategies to encourage further self-explanation for improved comprehension (Conati and 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).

2.2.3 The Difficulty of Processing 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.

2.2.4 Self-Explanation and Handwriting Input Analysis in Mathematics Education

Self-explanations and handwriting logs were utilized in a lot of research in mathematics education. Hodds et al. (2014) found that students who received the training generated higher quality explanations and performed better on a comprehension test in math quizzes (Hodds et al., 2014). Renkl (1997) stated that in mathematics quizzes solving, more successful learners tended to self-explain by predicting the next steps and identifying the overall goal structure and its subgoal objectives.

On the other hand, handwriting data analytics has been studied for detecting answer

stuck points (Iiyama et al., 2017). Kishi and Miura (2018) developed a system to detect a learner’s weak points using time intervals of pen stroke data in a mathematics quiz, and Ochoa et al. (2013) could discriminate between experts and non-experts in groups of students solving mathematical problems with multi-modal learning analytics, including pen stroke data (Kishi and Miura, 2018; Ochoa et al., 2013).

2.3 Challenges in Utilizing Self-Explanation for Impasse Detection

Impasses in learning, particularly in mathematics, are crucial moments that prompt deeper cognitive engagement and problem-solving (VanLehn et al., 1992). The detection of impasses is vital as it leads to the employment of various strategies by learners to overcome these hurdles. The integration of intelligent tutoring systems (ITS), Mathia, and I-START reveals a focus on self-explanation as a means to aid learners in overcoming these impasses, each system offering unique approaches to incorporating this strategy.

While self-explanation is a valuable educational tool, its application in detecting learning impasses presents several challenges. Firstly, there is a need for well-established methods and technologies to effectively detect learners’ impasses. The existing literature recognizes the utility of self-explanation but falls short in offering concrete, reliable methods for impasse detection, especially in complex subjects like mathematics. Also, the fragmentary nature of self-explanation writings and the challenge of integrating self-explanation prompts into web-based learning tools persist (Chamberland et al., 2015; Panaite et al., 2018).

Secondly, the traditional approach to analyzing self-explanation often adopts a top-down methodology, relying on predefined templates or model answers. This approach, while structured, limits the flexibility and adaptability of self-explanation in diverse educational settings. It often does not account for the unique thought processes and learning trajectories of individual students, thereby missing out on the nuanced understanding of their learning impasses.

Thirdly, the act of self-explaining, especially in subjects requiring complex mathematical reasoning, can be quite demanding for students. It requires them to articulate their thought processes and problem-solving steps in writing, a task that can be particularly challenging for those with varying levels of academic proficiency. The process of gener-

ating explanations to oneself demands a deep level of cognitive engagement. Students must not only recall information but also restructure it in their own words, a process that involves critical thinking and reflection. This cognitive effort, while beneficial for deep understanding, is time-consuming. The nature of self-explanation often requires students to integrate new knowledge with their existing understanding. Making these connections and constructing a coherent explanation is a complex task that cannot be rushed if it is to be effective. This complexity is particularly pronounced in subjects like mathematics, where concepts are often abstract and require detailed and thoughtful explanation.

Moreover, the implementation of self-explanation as a learning strategy also raises concerns regarding its practicality and adoption in regular classroom settings. Teachers are required to invest considerable effort in preparing and guiding students through the process of self-explanation. This labor-intensive nature of self-explanation, coupled with the need for specialized training and resources, poses significant challenges to its widespread adoption and effectiveness as an educational tool.

In summary, while self-explanation holds potential for enhancing learning and detecting impasses, its practical application in educational settings is hindered by methodological limitations, the complexity of the task for learners, and the resource-intensive nature of its implementation. Addressing these challenges is crucial for leveraging the full potential of self-explanation in educational technology and pedagogy.

Chapter 3

Preliminaries

3.1 Explanation of Data Collection and Analysis Methods

3.1.1 LEAF System

The Learning and Evidence Analytics Framework (LEAF) system (Flanagan and Ogata, 2018) stands at the forefront of this research, representing an innovative amalgamation of data collection and analysis methodologies within an educational context. At its core, LEAF utilizes the BookRoll digital reading system and the LAViEW learning analytics tool to gather and analyze detailed handwriting data and student interactions in a Japanese secondary school setting. This system is instrumental in capturing pen strokes and linking them with self-explanations, providing a detailed insight into the student's cognitive process during problem-solving. This research employs a diverse array of technologies to support Self-Directed Learning (SDL) in various contexts. These include a Learning Management System (LMS) such as Moodle, the BookRoll e-Book reading system, and a range of mobile and wearable health technologies. Each of these components plays a vital role in creating a comprehensive and integrated learning environment, as depicted in the figure 3.1.

In LEAF, Teaching-Learning Cases (TLCs) are formulated to generate evidence from specific teaching-learning activities. TLCs follow a structured format including Context, Problem, Indicator, Intervention, Control, and Results. This format provides a comprehensive framework for analyzing and understanding the nuances of various educational scenarios, enabling educators to draw valuable insights for effective teaching strategies.

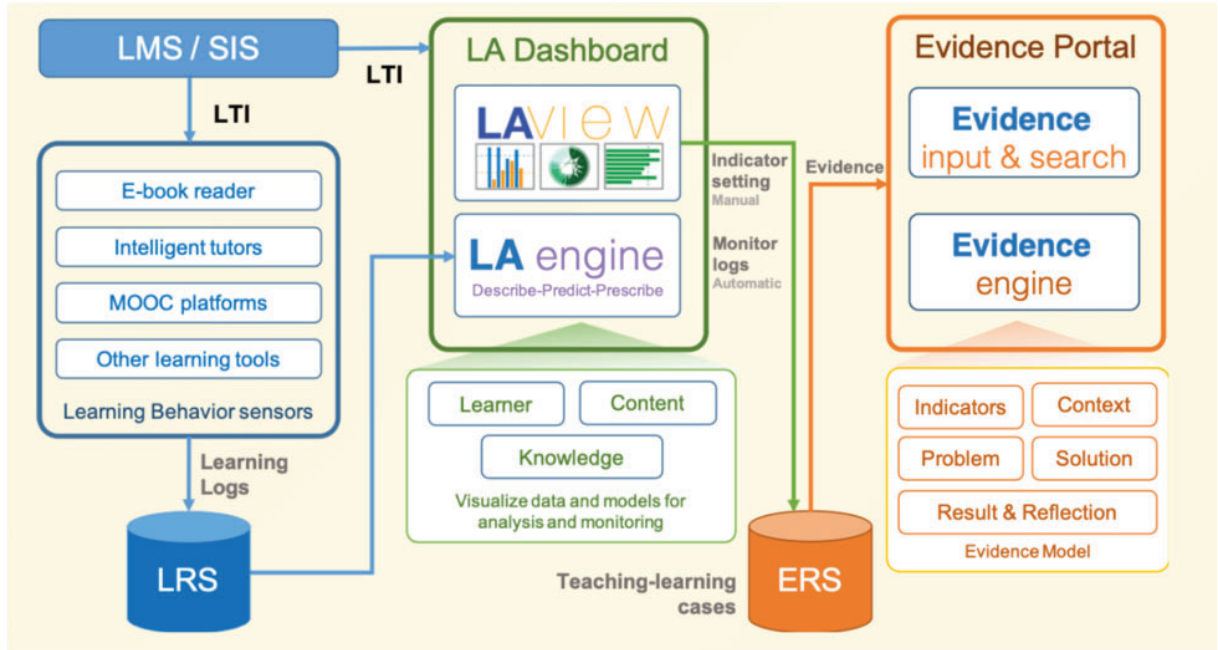


Figure 3.1: Architecture of LEAF (Majumdar et al., 2019)

3.1.2 LEAF Framework

Expanding on the LMS-independent platform concept by Flanagan and Ogata (2018), this research utilizes a framework that integrates a learning log database centered around BookRoll. This framework employs Experience API (xAPI) statements to import logs from various learning tools, encompassing a wide range of learning activities. The LMS serves as an access point to a plethora of learning support tools, while BookRoll provides in-depth log data crucial for analytics. The LEAF framework enhances this system by incorporating the Evidence Record Store (ERS) and the Evidence Analytics Tool. The ERS is a dedicated database for storing analytical findings, and the Evidence Analytics Tool is used for aggregating and presenting these findings in a user-friendly manner. This setup facilitates a structured approach to capturing and analyzing data from technology-enhanced teaching-learning practices.

3.2 Data Collection and Labeling

3.2.1 Self-explanation and Penstroke Data Collection

We collected the data over the period of Jan. 2021 to June. 2023 using the LEAF and a learning analytics tool LAVIEW where students and teachers can monitor and reflect

on their learning. The platform was deployed in a Japanese secondary school and has been in use for several years. Firstly, students were asked to view the quiz and write down their answers using a stylus and tablet computer with the handwriting input in BookRoll as shown in Figure 3.2. BookRoll captures the handwriting data as a series of vectors representing the coordinates and velocity of pen strokes, which allows realistic playback of the handwritten answers and fine-grained analysis of the students’ answering process. Preliminary analysis of such data has been examined by Yoshitake et al. (2020) to support group learning. Secondly, when the students had finished answering the quiz, they were asked to review their handwritten answers in LAViEW, and input explanations of how they solved the quiz. The students input a sentence of explanation everytime they think they have completed some step in their answers during the playback. Therefore, the self-explanations are temporally associated with the pen stroke data.

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. The self-explanation of the answer in Figure 3.3 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.”

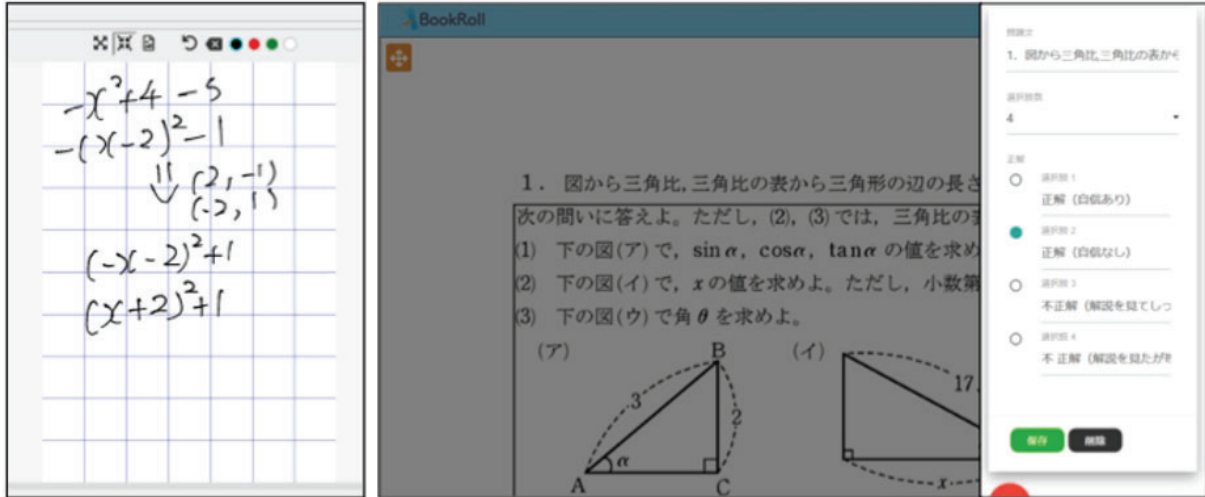


Figure 3.2: BookRoll user interface: handwritten answer (left), quiz and quiz answer report (right)

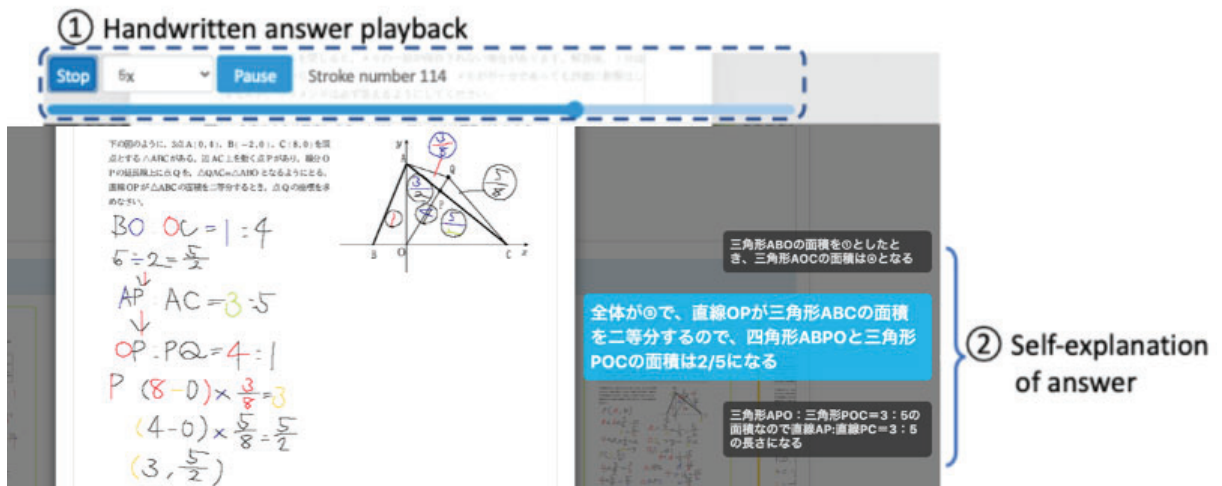


Figure 3.3: Handwritten answer review playback and self-explanation input user interface

3.3 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. Self-explanations in our study were assessed based on three main criteria: coherence, clarity, and relevance. Specifically, 'coherence' gauges the logical flow of the explanation, 'clarity' measures its understandability, and 'relevance' ensures the inclusion of all pertinent knowledge concepts and procedural elements. For consistent evaluation, we adapted the rubric and scoring definitions with varied solutions or strategies. Instead of a detailed sentence-by-sentence breakdown, our approach evaluates explanations on a holistic, quiz-by-quiz basis, offering a comprehensive insight into the learner's understanding of the topic. The three studies conducted in this research vary in terms of the timing of data collection. Therefore, in this chapter, we define the quality of self-explanation, propose scoring methods, and present scoring results for representative datasets, taking into account the differences in the datasets collected at different timings. Additionally, the datasets used in the experiments in the following chapters will be defined in each respective chapter.

3.3.1 Settings

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 and 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. These rubrics are designed to be system-judgeable, focusing on task-specific mathematical aspects. Table 3.1 delineates the definitions of key terms utilized in this chapter. Table 3.2 presents the rubrics and a sample answer of self-explanation in a quiz. Also, the evaluators created an objective scoring index, which focused on the inclusion of essential knowledge components and logical explanations that are shown in Table 3.3.

Table 3.1: 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 3.2: 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 $y=\frac{5}{6}x$, 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 $\triangle QAC$ was found from $\triangle AQO$ and $\triangle OQC$, and the coordinates of Q were found.

3.3.2 Self-explanation Quality Scoring with Rubric and Index

The evaluations were carried out by two scorers, who initially rated the self-explanations individually with using rubrics and scoring index. After the initial round of evaluations, the two scorers then compared their evaluations with each other, and if they 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 agreement through discussion.

Table 3.3: 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 3.4: Descriptive statistics of the collected self-explanations(Samples)

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

3.4 Research Approach and Contribution

To overcome the inherent challenges of self-explanation in learning, as mentioned in Chapter 2, this research introduces a novel approach that combines self-explanation with pen-stroke data for the impasse detection in LEAF Framework. This method significantly diverges from traditional systems like Mathia, which primarily focus on broader problem-solving contexts. Our system automatically detects impasses in learning and generates relevant feedback based on the collected data. This approach is predicated on the principle that increased learner engagement in self-explanation leads to more data availability, thereby deepening the understanding of the learning process.

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. Our exploration is supported by a rich tapestry of related works that bridge cognitive theories and data

Table 3.5: 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.

analytics, providing a theoretical bedrock for our approach to identifying learning impasses through pen stroke data.

To the best of our knowledge, our study is the first to combine self-explanations and pen stroke data in identifying stuck points, where the pen stroke data provides temporal and spatial hints on “where” the students were stuck and self-explanations help to infer specifically “what” knowledge the students were stuck on. Also, the integration of the feedback system into educational platforms has the potential to revolutionize teaching and student engagement. This advancement opens up opportunities for personalized learning experiences, ultimately leading to improved educational outcomes. The system, with its real-time analysis and bottom-up impasse detection, signifies a significant contribution to educational technology, setting a new standard for personalized learning support. Its positive impact on self-explanation in subjects like mathematics highlights the transformative potential of this technology in enhancing overall educational practices.

Chapter 4

Experimental Study(1): Impasse Detection Methods with Self-explanation

The focus of this chapter is analyzing pen stroke data and self-explanations collected from students while solving mathematical problems. The aim was to identify impasses in learning processes by closely observing patterns and anomalies in pen stroke behaviors and self-explanations.

4.1 Background and Motivation

The initial study of our study aims to uncover the underlying patterns within pen stroke data that signal learning impasses. By analyzing the nuanced ways in which students interact with digital writing surfaces, we intend to extrapolate meaningful insights into their understanding and problem-solving processes. Self-explanation has been widely recognized for its learning effects for a long time (Bisra, Liu, Nesbit, et al., 2018). It is a good indicator of the students' understanding towards solving a quiz because there are cognitive factors behind explanation (Lombrozo, 2006). Self-explanation is defined as explaining concepts, procedures, and solutions in order to deepen understanding of material and to make sense of relatively new information (M. Chi, Leeuw, Chiu, and LaVancher, 1994; Rittle-Johnson, 2006).

However, few research has been conducted on the use of self-explanations in learning analytics to identify impasses, to examine intervention methods, and to use them to recommend quizzes. Students who can well explain their solving processes possibly have clear solution strategies and good understanding of necessary knowledge and skills to

solve the quiz. On the other hand, students who cannot explain how they solved the quiz may be stuck in some solving steps. If a system can automatically identify a student’s impasses, the teacher can not only learn about the student’s level of understanding, but the system can also support the student to overcome his or her difficulties.

McNamara et al. (2004) created an interactive tutoring system called iSTART to support the development of self-explanation skills in reading comprehension . The system guides learners through the exercise to support active reading and thinking by providing automatic evaluation and scoring of self-explanations and appropriate scaffolding. Natural language processing techniques, such as Latent Dirichlet Allocation (LDA) Analysis topic modeling, were applied to extract the characteristics of student self-explanation artifacts, and their similarity with reading materials are used to analyze the self-explanations for automatic scoring. This system was modeled on an effective human-delivered intervention called self-explanation reading training (SERT).

4.1.1 Learning Impasse

Learning impasses represent pivotal moments in a student’s educational journey, particularly in the field of mathematics (VanLehn et al., 2003). Warli and Rahayu (2020) emphasize the cognitive challenges students face when understanding complex mathematical concepts, especially in group problem-solving contexts. These impasses often manifest as feelings of being stuck, detecting errors, or experiencing doubt, even when students are technically correct in their actions. In response to these difficulties, students may review previous knowledge, consult external resources, or seek help from peers or instructors. These moments of struggle are pivotal in fostering genuine learning, as they lead to a deeper grasp of the subject, the correction of misconceptions, and the enhancement of problem-solving skills. This understanding aligns with our study’s focus on impasse detection and resolution in mathematics education.

Extensive research on cognitive skill acquisition consistently underscores the prevalence of impasses in the learning process, particularly in the absence of tutors (VanLehn and Jones, 1993; VanLehn et al., 1992). Therefore, the automated detection of these mathematical impasses holds significant potential for advancing educational research. This study delves deeper into the concept of learning impasses in the context of self-explanation in mathematics learning. It aims to detect these impasses in automated learning environments, contributing to the field by defining and identifying impasses and

evaluating the role of self-explanations in overcoming these challenges.

4.1.2 Premise Conditions and the Definition of Impasse in Math Learning

Our approach to detecting impasses aligns with the definition proposed by VanLehn et al. (2003), where an impasse is seen as a point of cognitive struggle or uncertainty encountered by a student while trying to understand or apply a specific piece of knowledge in mathematics. Such impasses can manifest in various forms, including being stuck, detecting an error, or experiencing doubt, even when the action is technically correct. They prompt students to delve into their memory, consult external resources like textbooks, or seek assistance from others.

In line with this understanding, our research posits that the ability to produce detailed self-explanations is a crucial indicator of a student’s comprehension of mathematical concepts. We find that difficulties in providing coherent self-explanations often signal these underlying cognitive impasses. Therefore, we rigorously analyze the quality of self-explanations, understanding that clear and thorough explanations indicate a solid grasp of concepts, while vague or incomplete ones point to areas of weak comprehension.

We closely examine instances where self-explanations by students reveal significant impasses in mathematics learning. A high-quality self-explanation is not only logically structured but also elucidates the student’s pathway through the problem-solving process. This type of explanation demonstrates a deep understanding of the mathematical concept, showcasing the student’s ability to integrate and apply their knowledge effectively. Conversely, an ineffective self-explanation, characterized by a lack of coherence and clarity, points to potential learning impasses. These impasses might manifest as fragmented thoughts, unclear connections between ideas, or an inability to express the underlying mathematical reasoning. Such explanations often signal a superficial understanding or a disconnect in cognitive processing, where the student might be able to perform certain steps of the problem-solving process but fails to comprehend the overarching concepts or principles.

4.2 Proposed Method

4.2.1 Rubric-Based Evaluative Framework

This study introduces a rubric-based evaluative system for assessing students' reasoning in mathematics, particularly focusing on identifying learning impasses. Rubrics, as defined by Andrade (2000) and Arter & Chappuis (2007), provide a structured method to assess students' work, classifying performance across a range from outstanding to deficient. General rubrics give an overview of performance levels, while task-specific rubrics are tailored to specific mathematical elements (Thompson & Senk, 1998). The rubrics in this study evaluate problem-solving processes step by step, facilitating systematic judgment by the evaluative system.

Verschaffel et al. (1999) outlined five essential steps in mathematical problem-solving: drawing representations, listing elements, simplifying figures, executing calculations, and assessing solutions. Although self-explanation is not directly linked to any particular step, it provides insight into the underlying knowledge components involved in these steps. We hypothesize that students' self-explanations reveal their procedural knowledge during problem-solving.

As shown in Chapter 3, Table 3.1 delineates the definitions of key terms utilized in this chapter and Table 3.2 presents the rubrics and a sample answer of self-explanation in a quiz. These rubrics and samples serve as a foundation for detecting learning impasses.

4.2.2 Labeling Impasses through Human Assessment

The assessment process involves manual evaluation of students' handwriting and self-explanations, culminating in labeling problem-solving steps as either correct or incorrect. This approach allows for a more nuanced understanding of students' cognitive processes, especially for those who rely on rote memorization or struggle to articulate their thought processes. This method transcends simple correctness assessments, striving to uncover the reasons behind students' responses.

We focus on two outcomes:

- **Correct Step:** Self-explanations that are coherent and well-structured, indicating comprehension, receive a score of '1' (no detected impasse).
- **Incorrect Step:** Errors in solutions or weak self-explanations, especially in tackling

geometric concepts, receive a score of ‘0’, signaling difficulty.

Figure 4.1 presents a typical handwritten answer and its associated self-explanation, which evaluators use to classify parts as correct or incorrect.

- Stroke 34: Calculated the area of triangle ABC.
- Stroke 60: Calculated the area of triangle ABO.
- Stroke 182: Determined the coordinates of point P and derived the equation of line OP.
- Stroke 201: Expressed the coordinates of point Q in terms of variables.
- Stroke 291: Determine the value of k.
- Stroke 316: Substituted the value of k and obtained the answer.

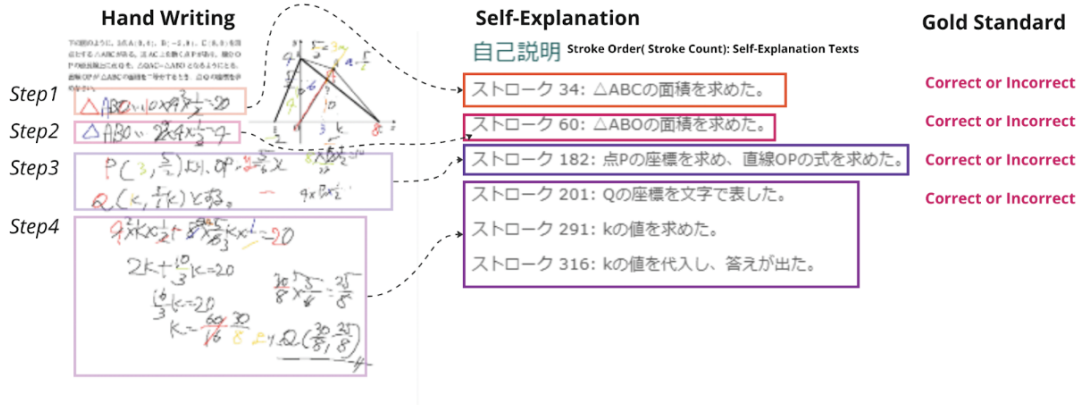


Figure 4.1: Labeling Method

4.2.3 Feature Extraction from Handwriting Data Associated with Each Self-Explanation

In our research, we employ a comprehensive approach that combines handwritten data with self-explanation texts to gain a profound understanding of students’ learning processes, building upon prior studies in handwriting detection (Iiyama et al., 2017; Kishi and Miura, 2018; Nakamoto et al., 2021). Figure 4.2 provides an overview of feature extraction associated with each self-explanation of a student. This method involves analyzing handwritten data, including pen strokes and inputs, to reveal the temporal and procedural aspects of problem-solving, which are crucial for identifying impasses by observing the sequence and methodology students employ in tackling mathematical problems.

Integrating these insights with self-explanation data allows us to form a holistic view of how students learn and pinpoint areas of difficulty. We adopt a methodology inspired by Kishi and Miura (2018), preparing model solutions for each quiz that encapsulate

essential steps and knowledge components for correct answers. We then compare students' self-explanations to these models and identify the most relevant self-explanation sentences. Finally, we generated the handwriting data associated with the most relevant self-explanation sentence. The detailed features for this key factor are outlined in Table 4.1.

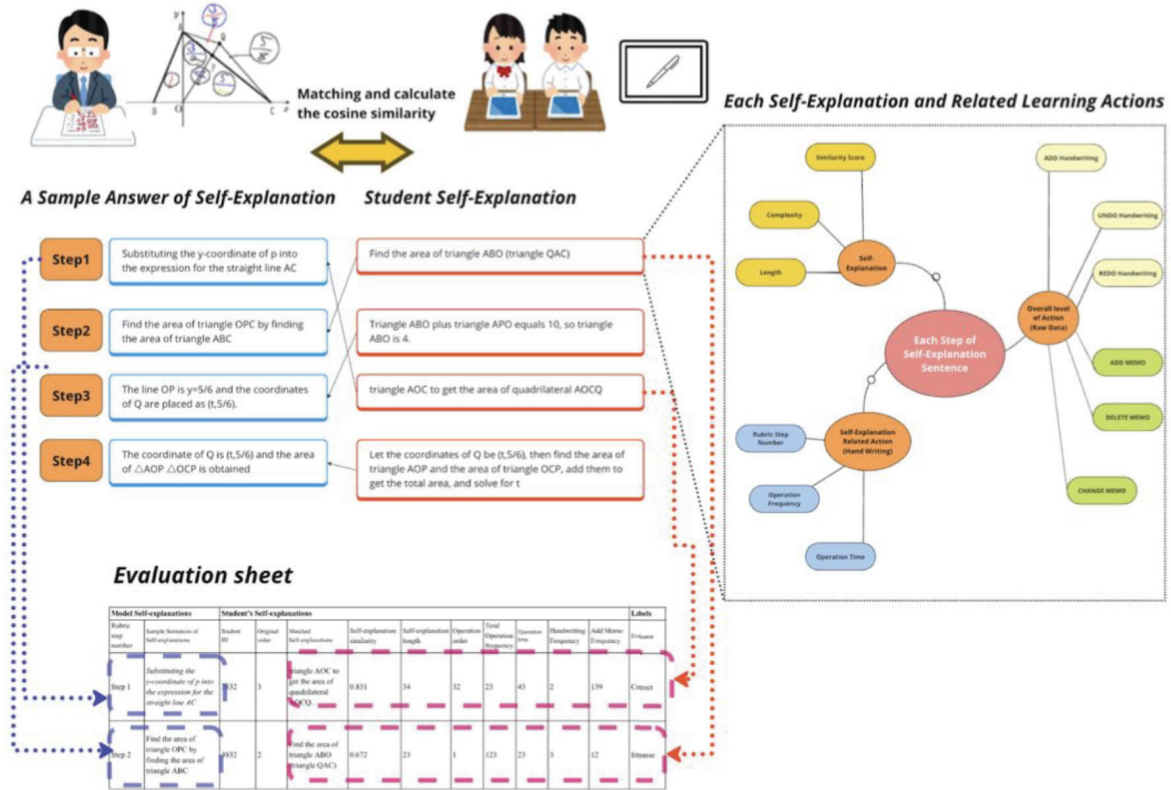


Figure 4.2: Self-Explanation and Handwriting Example with Impasses Highlighted

4.3 Hierarchical Logistic Regression Analysis

4.3.1 Hierarchical Logistic Regression Model

To identify significant predictors of learning impasses, we use a hierarchical logistic regression model, which sequentially introduces features related to self-explanation quality, handwriting data, and operational engagement. Table 4.2 summarizes the key feature groups.

In this section, we analyze the efficacy of various predictors in identifying learning impasses in middle school students' mathematical comprehension through hierarchical

Table 4.1: Hierarchical Logistic Regression Features

Feature Group	Feature Name	Description	Mean	SD
Self-Explanation Quality Indicators	Self-Explanation Engagement Level	The count of strokes or actions taken associated with each self-explanation, indicating the degree of user engagement.	0.06	0.19
	Self-Explanation Length	The character count of each self-explanation, serving as an indicator of the explanation's level of detail.	0.08	0.22
	Self-Explanation Complexity	The number of nodes (words) in each self-explanation, determined through morphological analysis using MeCab, indicating the complexity of the explanation.	0.09	0.22
	Self-Explanation Sequence	The order of strokes (operations) in each self-explanation, important for understanding user operation patterns.	0.25	0.24
	Self-Explanation Similarity Score	A score based on cosine similarity between self-explanations and reference sentences, indicating the similarity of the text.	0.08	0.22
Self-Explanation Related Action (Generated Handwriting Features)	Total Activity Duration	The total sum of time intervals between consecutive operations for each user, reflecting the overall duration of user activity(Include time intervals).	0.10	0.23
	Generated Handwriting Features Count	The total count of operations performed by each user, indicating the level of user activity.	0.30	0.25
	Total Activity Duration	The total sum of time intervals between consecutive operations for each user, reflecting the overall duration of user activity.	0.22	0.21
	Operation Segment Duration	A segment of time data associated with specific user operations, segmented based on user interactions, indicating the temporal aspects of user interactions.	0.23	0.23
	Accurate Activity Duration	The total of calculated times, reflecting a more accurate measure of user activity duration(Exclude time intervals. only the time for operations).	0.30	0.22
	Rubric Step Number	An ordinal scale (typically ranging from 1 to 4) that quantifies the number of rubric steps.	2.50	1.12
Overall Level of Action (Raw Operational Data)	ADD_HW_MEMO	Addition of hardware-related memo.	0.04	0.17
	UNDO_HW_MEMO	Undo hardware-related memo.	0.09	0.22
	CLOSE	Close an item or section.	0.21	0.38
	OPEN_RECOMMENDATION	Open a recommended item.	0.05	0.18
	TIMER_PAUSE	Pause a timer.	0.01	0.10
	CLOSE_RECOMMENDATION	Close a recommended item.	0.03	0.17
	ADD MEMO	Addition of a general memo.	0.02	0.13
	QUIZ_ANSWER_CORRECT	Provide a correct answer in a quiz.	0.04	0.19
	PAGE_JUMP	Jump to a specific page.	0.08	0.24
	QUIZ_ANSWER	Provide an answer in a quiz.	0.01	0.09
	CHANGE MEMO	Modify an existing memo.	0.00	0.00
	DELETE_MEMO	Delete a memo.	0.04	0.17
	CLEAR_HW_MEMO	Clear hardware-related memos.	0.04	0.17
	REDO_HW_MEMO	Redo a hardware-related memo.	0.09	0.22
	ADD_RECOMMENDATION	Add a recommendation.	0.21	0.38

Table 4.2: Hierarchical Logistic Regression Model Groups

Feature Group / Model Name	M_S	M_G	M_A	M_{S+G}	M_F (Final)
Self-Explanation Quality Indicators	○			○	○
Self-Explanation Related Action (Generated Handwriting Features)		○		○	○
Overall Level of Action (Raw Operational Data)			○		○
Total Features	5	6	15	11	26

logistic regression. The analysis is organized under three subheadings: (1) Analysis of Self-Explanation Features, (2) Exploring General Learning Operations and the Impact of Memo and Quiz Actions, and (3) Comprehensive Analysis in the Final Model. Each subheading represents a distinct model or combination of models to examine different aspects of student engagement and self-explanation features, offering a holistic understanding of the factors influencing learning impasses.

4.3.2 Analysis of Self-Explanation Features (M_S , M_{S+G} , M_F)

In our study, examining the efficacy of various predictors in identifying impasses in middle school students' mathematical comprehension, we employed distinct models that focused on self-explanation features and general learning activities. The analysis, segmented into different models— M_S , M_{S+G} , and M_F —revealed insightful findings on the predictors' influence on detecting comprehension difficulties, as indicated by their respective AUC values.

Model M_S (AUC = 0.740): This model focused on self-explanation features. Four self-explanation features demonstrated a statistically significant association with successful impasse detection:

- Self-Explanation Length: Odds Ratio: 15.69
- Self-Explanation Similarity Score: Odds Ratio: 8.75
- Self-Explanation Engagement Level: Odds Ratio: 1.63
- Self-Explanation Sequence: Odds Ratio: 0.12

Table 4.3: Logistic Regression Results (Significant Indicators)

Model		Predictors	Odds Ratio	Coeff.	Std Error	t Value	p-value
M _S	AUC = 0.740	Self-Explanation Length	15.69	2.75	0.21	13.09	<0.001
	LogL = -2133.1	Self-Explanation Similarity Score	8.75	2.17	0.19	11.50	<0.001
	AIC = 4278.1	Self-Explanation Engagement Level	1.63	0.49	0.18	2.72	0.010
	BIC = 4315.3	Self-Explanation Sequence	0.12	-2.10	0.15	-13.83	<0.001
M _G	AUC = 0.655	Generated Handwriting Features Count	3.43	1.23	0.15	8.38	<0.001
	LogL = -2360.8	Rubric_Step_Number	0.65	-0.42	0.03	-13.46	<0.001
	AIC = 4731.7	Accurate Activity Duration	0.49	-0.71	0.24	-2.88	0.004
	BIC = 4762.6						
M _A	AUC = 0.603	QUIZ_ANSWER_CORRECT	2.18	0.78	0.21	3.71	<0.001
	LogL = -2433.0	ADD_MEMO	0.56	-0.58	0.11	-5.29	<0.001
	AIC = 4898.0	ADD_HW_MEMO	0.53	-0.64	0.15	-4.13	<0.001
	BIC = 4997.0						
M _{S+G}	AUC = 0.777	Self-Explanation Length	18.44	2.91	0.22	13.43	<0.001
	LogL = -2027.7	Self-Explanation Similarity Score	10.11	2.31	0.20	11.79	<0.001
	AIC = 4075.4	Generated Handwriting Features Count	2.30	0.83	0.24	3.48	<0.001
	BIC = 4137.2	Rubric_Step_Number	0.62	-0.48	0.04	-13.58	<0.001
		Self-Explanation Sequence	0.13	-2.02	0.16	-12.86	<0.001
M _F	AUC = 0.802	Self-Explanation Length	24.76	3.21	0.23	13.94	<0.001
	LogL = -1941.8	Generated Handwriting Features Count	15.98	2.77	0.31	8.82	<0.001
	AIC = 3933.7	Self-Explanation Similarity Score	9.15	2.21	0.20	10.87	<0.001
	BIC = 4088.4	Accurate Activity Duration	2.03	0.71	0.31	2.31	0.021
		ADD_MEMO	0.75	-0.28	0.13	-2.15	0.031
		Rubric_Step_Number	0.60	-0.51	0.04	-14.10	<0.001
		OPEN	0.52	-0.65	0.28	-2.34	0.019
		Self-Explanation Sequence	0.26	-1.34	0.18	-7.34	<0.001
		ADD_HW_MEMO	0.06	-2.84	0.26	-10.88	<0.001

These positive coefficients corroborate the hypothesis that a higher level of detail in self-explanations aligns with better comprehension and course success.

Model M_{S+G} (AUC = 0.777): This model integrated generated handwriting features counts with self-explanation features, enhancing the predictive power of each feature group. The significant predictors included:

- Self-Explanation Length: Odds Ratio: 18.44
- Self-Explanation Similarity Score: Odds Ratio: 10.11
- Generated Handwriting Features Count: Odds Ratio: 2.30
- Rubric Step Number: Odds Ratio: 0.62

This integration suggests that both qualitative aspects of self-explanation and quantitative measures of operational engagement are pivotal in identifying learning impasses.

Model M_F (AUC = 0.802): This model achieved the highest AUC among the self-explanation focused models, underscoring the significance of combining various predictors:

- Self-Explanation Length: Odds Ratio: 24.76
- Self-Explanation Similarity Score: Odds Ratio: 9.15
- Generated Handwriting Features Count: Odds Ratio: 15.98

This model's high AUC value reinforced the importance of a multifaceted approach, combining both qualitative and quantitative elements of student engagement to predict learning impasses effectively.

4.3.3 Exploring General Learning Operations and The Impact of Memo and Quiz Actions (M_G , M_A)

The M_G model (AUC = 0.655) highlighted the role of operational factors such as:

- Generated Handwriting Features Count: Odds Ratio: 3.43
- Total Activity Duration: Odds Ratio: 1.37

Although this model exhibited a lower AUC compared to self-explanation focused models, it still underlined the relevance of general engagement levels in learning activities. The model’s findings align with the notion that higher operational engagement is positively associated with mathematical comprehension and course success.

The M_A model (AUC = 0.603) centered on memo-related actions and quiz answers, revealing a nuanced perspective on impasse detection:

- Quiz Answers Correct: Odds Ratio: 2.18
- Add Memo: Odds Ratio: 0.56
- Add HW Memo: Odds Ratio: 0.53

While quiz answers were significant predictors, memo actions had a lesser impact, indicating a more complex relationship with learning comprehension.

4.3.4 Comprehensive Analysis in Final Model (M_F)

The final model, M_F (AUC = 0.802), gathered significant predictors from the previous models, providing a holistic view of the factors influencing impasse detection. This model’s high AUC value reinforced the importance of a multifaceted approach, combining both qualitative and quantitative elements of student engagement to predict learning impasses effectively.

4.4 Discussion

4.4.1 Identifying Learning Impasses

This study employed hierarchical logistic regression to better understand learning challenges in online secondary school mathematics education. Our results demonstrate that combining various data types effectively identifies these challenges. Initially, a model focusing solely on self-explanation features, such as explanation length, achieved notable accuracy (AUC = 0.740), highlighting that self-explanation traits are important indicators of student understanding and potential learning barriers.

In contrast, models incorporating operational metrics, including the number of operations and total activity duration, achieved moderate accuracy (AUC = 0.655), underscoring the role of operational engagement in the learning process. While operational

metrics alone yielded moderate accuracy, combining them with self-explanatory data substantially improved model precision ($AUC = 0.777$), demonstrating the advantages of an integrated data approach. Most notably, our comprehensive model, combining self-explanation and operational factors, achieved the highest accuracy ($AUC = 0.802$), confirming the effectiveness of a holistic approach that incorporates detailed operational, engagement, and self-explanation data. These findings underscore the critical roles that both self-explanatory attributes and operational engagement play in identifying learning challenges in secondary school mathematics.

4.4.2 Learning Difficulties in Online Mathematics Education in MF

The 'Self-Explanation Length' emerged as a significant predictor, with an odds ratio of 24.76, indicating that students who provide longer, more detailed explanations tend to better understand complex mathematical concepts. Detailed self-explanations have been shown to promote deeper cognitive processing, reinforcing learning (M. Chi, 2000; Rittle-Johnson et al., 2017), thereby revealing thought processes and misunderstandings. Additionally, the 'Self-Explanation Similarity Score' had a substantial odds ratio of 9.15, indicating a strong link between alignment to reference material and understanding. High-quality self-explanations support knowledge transfer by focusing on content structure (Rittle-Johnson, 2006), while self-explanation complexity and sequence did not correlate strongly with accuracy, offering insights into less predictive aspects of learning difficulties.

The 'Rubric Step Number' suggests that later, more complex stages in problem-solving pose greater challenges, as indicated by a negative correlation coefficient of -0.2007. The 'Generated Handwriting Features Count,' with an odds ratio of 15.98, highlighted the role of active engagement. Frequent handwriting operations suggest a higher level of interaction with the learning material and emphasize the importance of active problem-solving and information manipulation in understanding complex concepts (McNamara et al., 2004). Handwriting features, such as stroke count and timing, also signify cognitive effort and problem-solving, supporting the concept of impasse-driven learning (VanLehn et al., 1992), while 'Accurate Activity Duration' showed a positive association, further underscoring sustained effort and engagement in overcoming learning impasses and improving comprehension.

4.4.3 Limitations and Future Directions

This study established a framework for identifying learning impasses using self-explanations, model answers, and pen-stroke data, achieving an 80% accuracy rate. However, several limitations warrant consideration.

First, the reliance on self-explanation quality introduces inherent variability, as well-written explanations are assumed to indicate an absence of learning barriers, while poorly articulated responses may suggest underlying issues. This assumption may lead to misjudgments, as students with strong understanding might omit detailed explanations, or those lacking training in self-explanation skills might produce subpar responses despite grasping the material. Such variability can introduce biases, undermining the system’s ability to accurately detect impasses.

Second, the model’s success is contingent on identifying consistent patterns in student behavior, relying on the assumption that learners share similar problem-solving methods and learning styles. This assumption may not hold universally, particularly in diverse educational contexts where students exhibit varied reasoning approaches. Poorly articulated or incomplete self-explanations further reduce the system’s accuracy, as the model struggles to identify clear patterns from inconsistent input.

Additionally, the scope of this research was limited to middle school mathematics, specifically linear equations and functions, within a structured four-step problem-solving process. While this provided a focused environment for testing the framework, it restricts the generalizability of the findings. The system’s applicability to other subjects, educational levels, and more complex or open-ended problem-solving scenarios remains unexplored.

Future research should address these limitations by incorporating additional features, such as textbook model answers or structured prompts, to provide a more consistent reference for evaluating self-explanations. Expanding the system’s scope to include diverse topics, grade levels, and problem-solving formats will enhance its adaptability and relevance. Moreover, training students in effective self-explanation techniques could mitigate the variability in input quality, improving the system’s accuracy. Longitudinal studies across varied educational contexts are also essential to validate the framework’s robustness and generalizability.

Chapter 5

Experimental Study(2): Automatic Generation of Exemplary Self-Explanations

5.1 Introduction

5.1.1 Motivation

The first study of this research underscored the importance of self-explanation in educational settings and introduced a novel approach for detecting impasses in mathematics learning. This approach integrates self-explanation texts, pen-stroke data, and model answers for quizzes. A key challenge identified in this study is the creation of rubrics and model answers, which are crucial for effectively detecting learning impasses. The manual development of these educational tools is both labor-intensive and may not fully address the varied needs of individual learners or the specific nuances of different mathematical concepts.

To overcome this challenge, the development of an automated system for the generation of rubrics and model answers is imperative. This next study of the research will focus on creating such a system, aiming to streamline the process of generating sample answers with accompanying rubrics. This advancement will not only enhance the efficiency of impasse detection but also cater to a broader range of learning requirements, thereby enriching the overall educational experience in mathematics learning.

5.1.2 Previous Research

While self-explanation is recognized as a highly effective pedagogical method, its implementation often hinges on the preparation by instructors. This reliance has been identified as a limiting factor in its broader application (Bisra, Liu, Nesbit, et al., 2018). Existing methods include designing system interfaces for monitoring and supporting self-explanation, developing models to assess understanding from self-explanation behaviors, and strategies to elicit further self-explanation for improved student case understanding (Arne et al., 2021; Boonthum et al., 2007; Conati and VanLehn, 2000; McNamara et al., 2004).

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. There are several studies to utilize and analyze the self-explanations. However, an overview of past studies shows a predominant top-down approach in AI-assisted analysis, necessitating fixed-structure model answers and annotated datasets (Jackson et al., 2010; Panaite et al., 2018, 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 in self-explanation analysis is still limited, the aim of this study is to investi-

gate the potential for automatically generating standard self-explanation examples that incorporate the necessary conceptual knowledge components required to solve a given problem. This study 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, Eysink, and Renkl, 2009, 2009; Rittle-Johnson, 2017a, 2017b). Overall, the unique contribution of this study 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.

To achieve this goal, the chapter 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 chapter aims to establish the fundamental methods for creating automatic model responses that can be used to analyze self-explanations and promote learning.

5.2 Goals of This Study

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 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, Liu, Nesbit, et al., 2018).

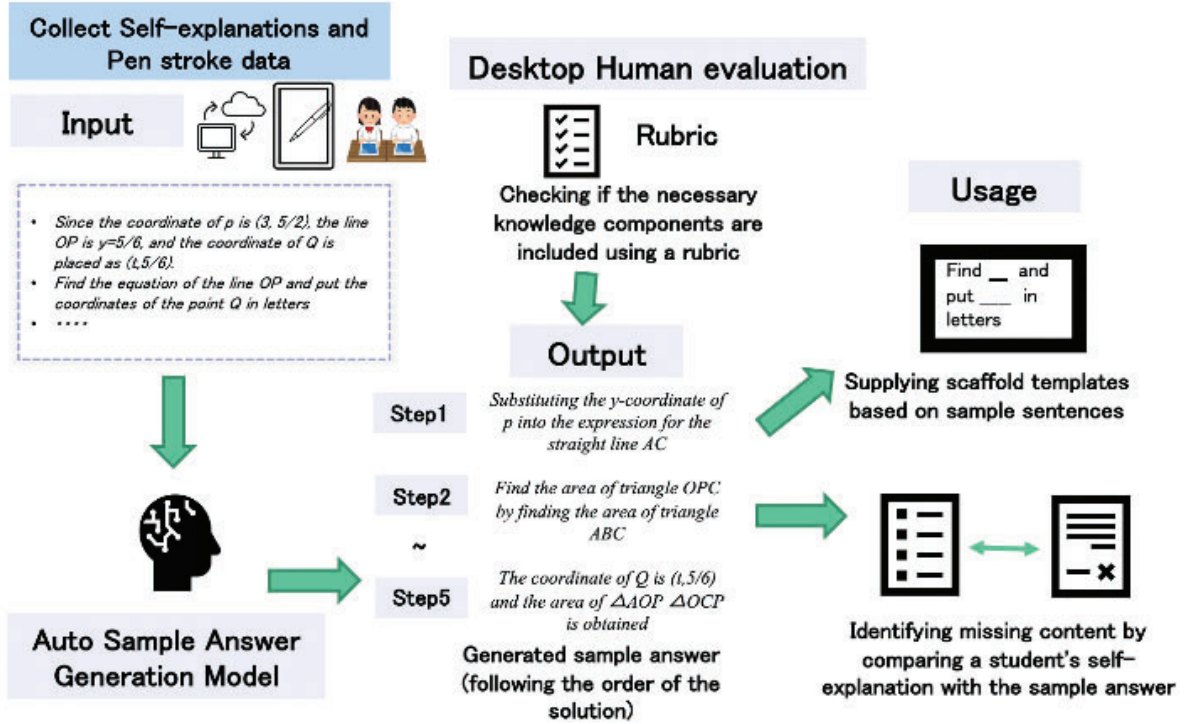


Figure 5.1: An overview of the process of our proposed model, and the inputs and outputs

5.2.1 Datasets

The data was collected between January 1, 2020, and December 31, 2021, using the LEAF platform. We conducted a study in 2021 using 1434 self-explanations from 25 quizzes to evaluate the machine’s ability to generate sample self-explanations automatically. The descriptive statistics of the self-explanations in the dataset and the evaluations given by the scorers. The scoring results were rated on a scale of 1 to 5 and sorted into categories that are shown in Table 3.4. The distribution of self-explanation quality scores was analyzed, revealing a correlation between the length of self-explanations and their assigned scores.

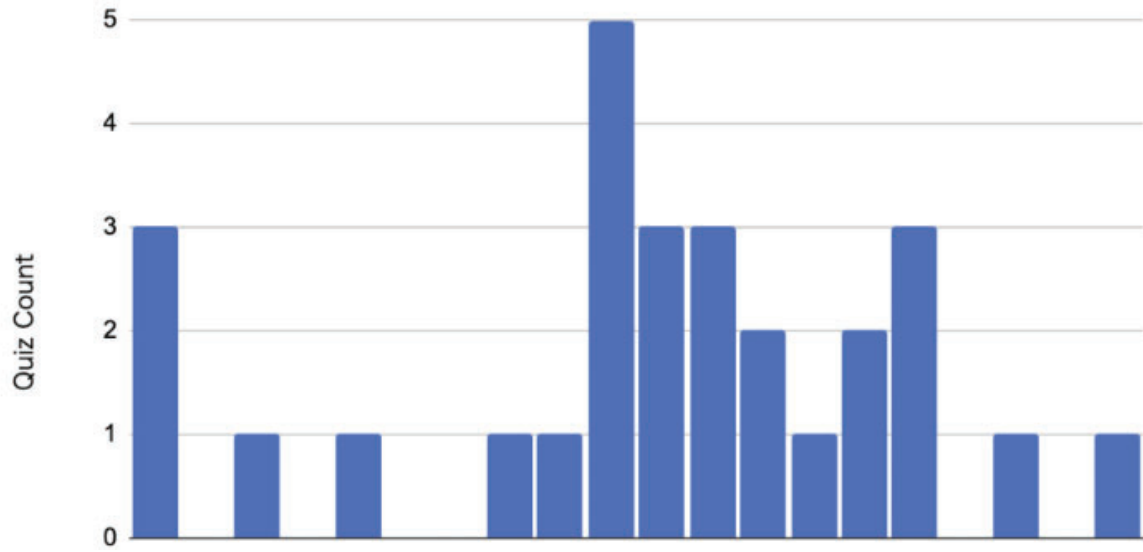


Figure 5.2: Histogram of the distribution of self-explanation responses over the quizzes in the dataset

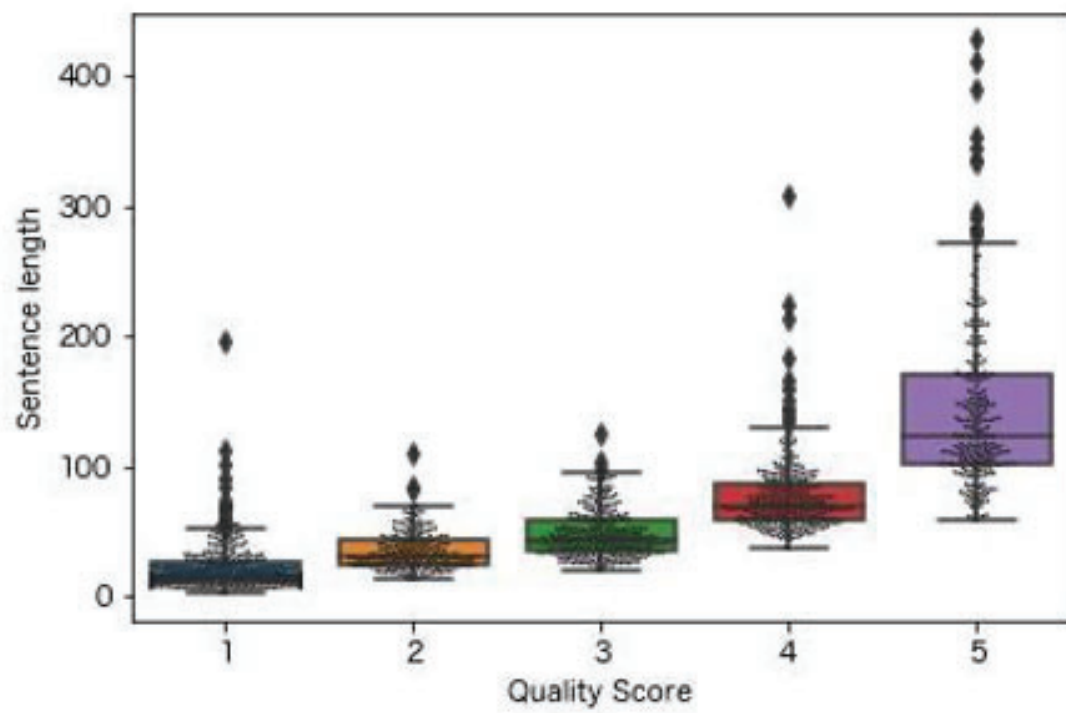
Longer self-explanations typically received higher scores, indicating a more comprehensive understanding and explanation of the problem-solving process. Examples of evaluated self-explanations are provided in Table 5.1 and Figure 5.3, illustrating the range of scores and the corresponding quality of explanations. These examples highlight the importance of logical structure and the inclusion of knowledge components in achieving higher scores.

A histogram of the distribution of self-explanation responses is shown in Figure 5.2, 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.

Table 5.1: 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

Figure 5.3: A boxplot of the self-explanation quality score distribution



5.2.2 Preprocessing

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 5.4. Firstly, we employed two characteristics to filter input data based on the quality and length distribution as shown in table 5.1. 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", "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.

5.3 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.

5.3.1 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 and Qureshi, 2020). Previous studies have used NLP tools like LSA and machine learning approaches (León et al., 2006; Özsoy 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 5.5. The vectorization

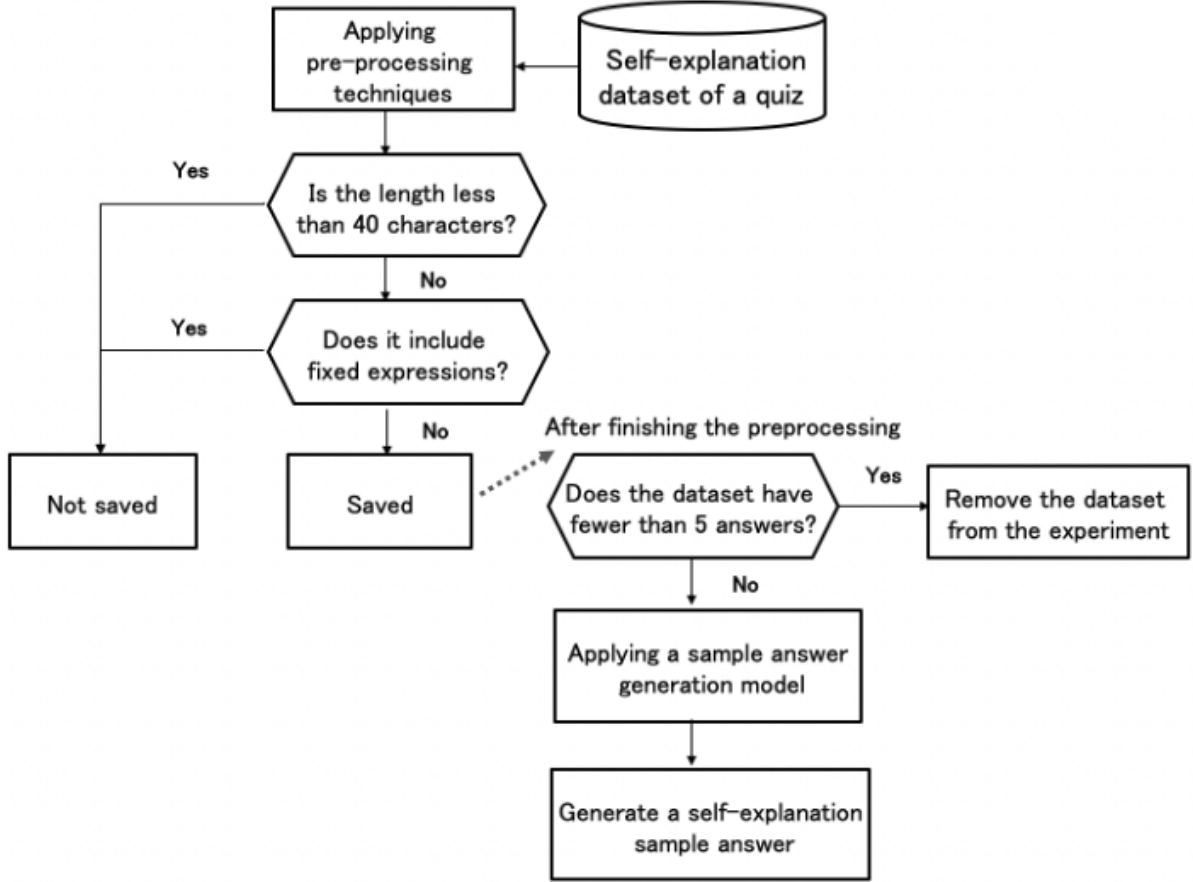


Figure 5.4: Workflow for the self-explanation processing

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

5.3.2 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

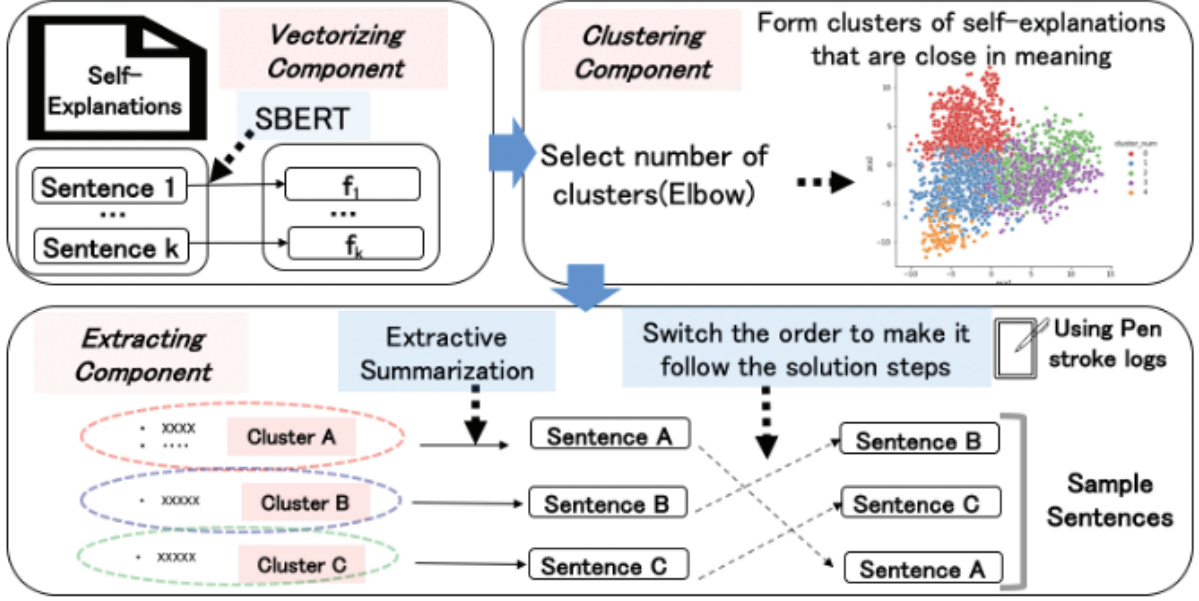


Figure 5.5: Overall of a proposed model architecture

et al., 2017) that outperforms existing models in natural language processing on various tasks (Devlin et al., 2018). SBERT fine-tunes BERT and has significantly improved sentence embedding methods (Reimers and 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.

5.3.3 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.

5.3.4 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 and 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 5.2 provides examples of self-explanations that were separated by the clustering model and extracted by LexRank.

5.4 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.

5.4.1 Grading Extracted Knowledge Components

The authors evaluated the machine-generated self-explanations using an established evaluation index for each quiz. Table 3.2 shows the criteria and results of the evaluation, which was based on the ability of evaluators to recall necessary knowledge components (Fleiss, 1971). To ensure consistency, two authors and one assistant evaluated the answers

Table 5.2: 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?

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.

5.4.2 Quality Evaluation with Metrics

Next, we evaluated the similarity of human-created and machine-generated sentences using several metrics: BERTScore (Zhang et al., 2020), BLEU (BiLingual Evaluation Understudy) (Papineni et al., 2002), and ROUGE (Recall-Oriented Understudy for Gisting Evaluation)(M. Chi, Leeuw, Chiu, and Lavancher, 1994). 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 (M. Chi, Leeuw, Chiu, and Lavancher, 1994).

In addition, we conducted a Spearman correlation analysis to investigate the relation-

ship 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 evaluation index in the following form. Root Mean Squared Error (RMSE) was also calculated to check the error variance.

$$\text{Human Evaluation Score} = \frac{\text{Num of Rubrics} - \text{Missing knowledge components}}{\text{Num of Rubrics}} \quad (5.1)$$

5.4.3 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. 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.

5.5 Results

5.5.1 Knowledge Components Extraction Grading

The human evaluation results of the generated self-explanation examples are shown in Table 5.3 and Table 5.4. In 72% of the quizzes, all five rubrics (knowledge components) were successfully generated, while 16% were missing one and 8% were missing two rubrics. However, achieving alignment between Knowledge Components and Rubric Step Order proved challenging, with only 64% of the quizzes successfully maintaining coherence in both aspects. This indicates that ensuring both accurate knowledge components and logical step order requires further improvement.

Meanwhile, Table 5.4 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 5.3: Missing knowledge components of each quiz by human evaluation.

Missing knowledge components	0	1	≥ 2
Num of quizzes	18	4	3
Probability Density	0.72	0.16	0.12

Table 5.4: Descriptive statistics of extracted self-explanations’ quality score.

Num 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

5.5.2 Quality Evaluation with Metrics

Table 5.5 and Table 5.6 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 5.5 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.

Table 5.5: The similarity evaluation between human-generated and machine-generated sentences: F1 metrics of BERTScore, BLEU, Rouge-1, Rouge-2, Rouge-L

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

5.6 Readability Analysis

Table and figures 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

Table 5.6: RMSE and correlations between human evaluation score and the similarity evaluation score of each metric

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

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 5.7: 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 5.8: 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

Table 5.9 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.

The level of Readability of human-generated and machine-generated

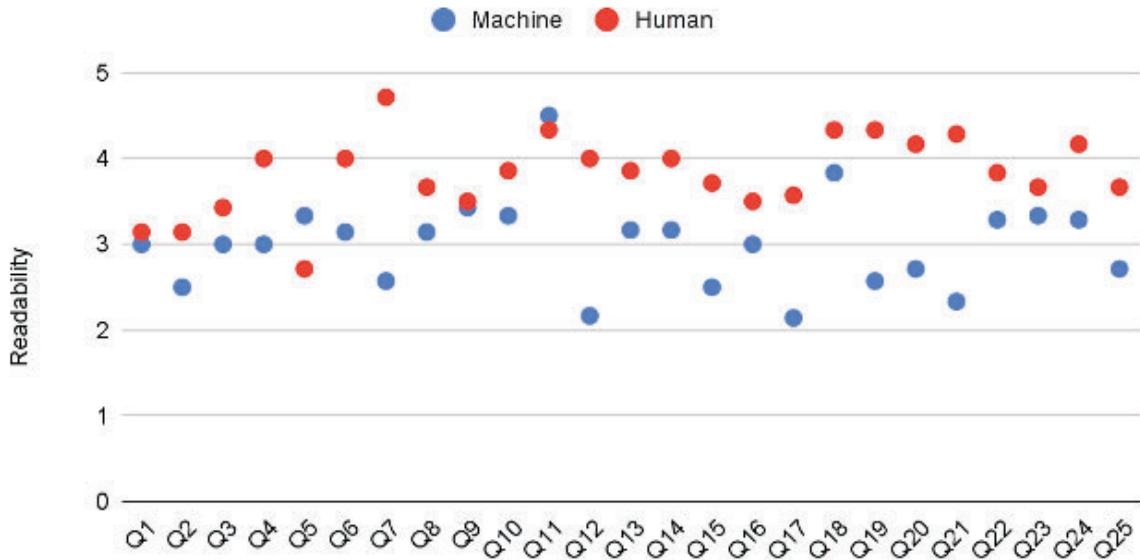


Figure 5.6: Student survey question: For each of 50 questions students are asked to rate if the question is (i) Human or machine-generated and (ii) the Difficulty level of each self-explanation on a scale between 5(easiest) and 1 (hardest)

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.

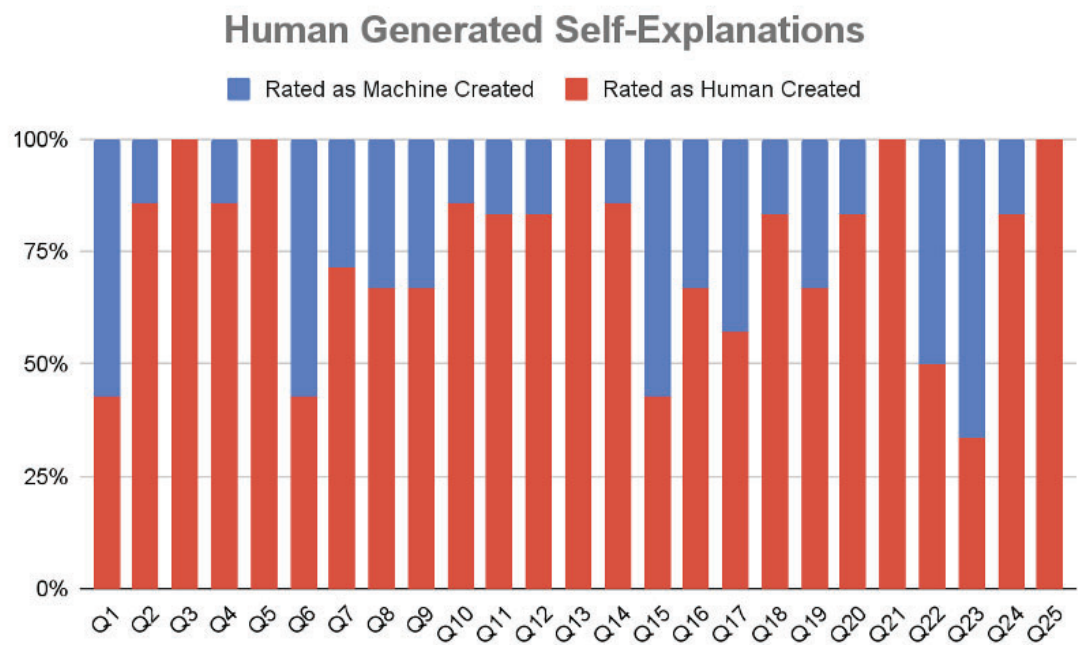
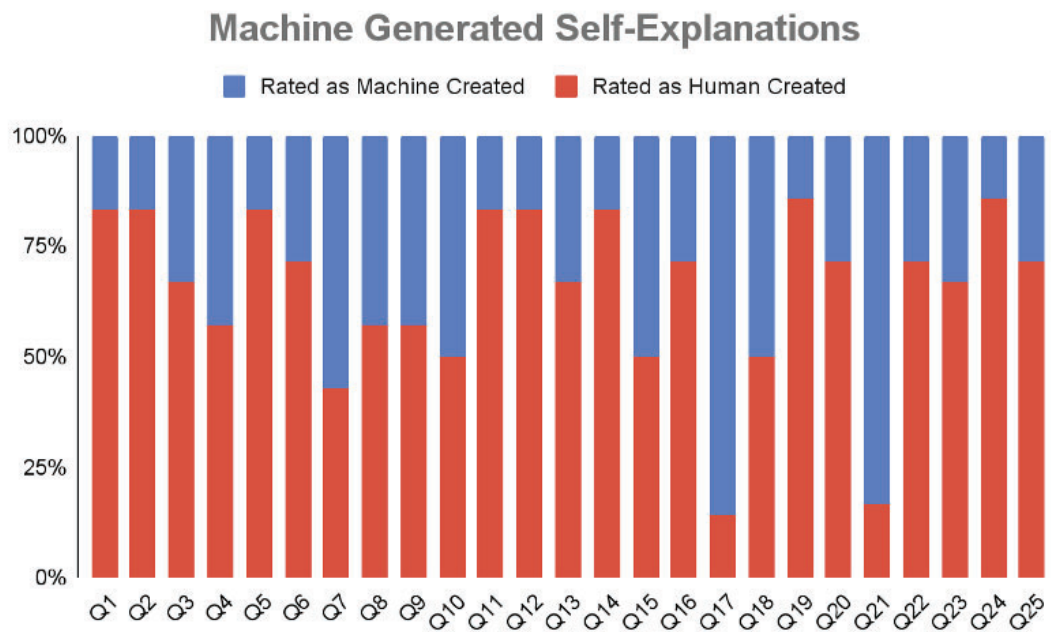


Figure 5.7: The evaluations of Machine Generated Self-Explanations and Human Generated Self-Explanations

Table 5.9: 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 BERTScore: 0.74	Missing knowledge components: 2 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 $\triangle QAC$ using t , subtract the area of $\triangle OAC$ from the sum of the areas of $\triangle AQO$ and $\triangle OQC$. 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 a 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.

5.7 Discussion and Limitations

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 establishes a valuable benchmark for future research, particularly given the lack of prior benchmarks for direct comparison. While accuracy is undoubtedly important, the primary aim of this study was to facilitate learners' cognitive processes rather than to maximize accuracy alone. In this light, even partially accurate or incorrect self-explanations can serve as catalysts for deeper understanding by encouraging learners to engage in critical reflection and actively evaluate their own knowledge.

However, a notable challenge emerged in aligning Knowledge Components with Rubric Step Order, with only 64% of cases achieving accurate alignment. This underscores the inherent complexity of simultaneously ensuring the correctness of extracted knowledge components and the logical coherence of their sequential order. Achieving consistency in both dimensions is crucial, as misalignment may detract from the clarity and usefulness of machine-generated self-explanations. Addressing this challenge will be a key focus for future refinement and research, as improving this alignment has the potential to significantly enhance the overall effectiveness of these explanations in supporting structured learning. 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.

5.7.1 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.

5.7.2 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.

5.7.3 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.

5.7.4 Conclusion

In this chapter we 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, Eysink, and Renkl (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.

As outlined in the Dataset Dependencies section, our methodology is heavily reliant on the collected datasets, with a particular emphasis on the pivotal role of data cleansing in achieving success. Obtaining higher-quality datasets necessitates a nuanced approach, involving not only threshold-based processing but also the implementation of more adaptable scoring methods for Deep Learning models. The upcoming chapter will delve into the process of constructing a learning model from a limited pool of self-explanation data.

Additionally, it will scrutinize strategies for developing a more precise self-scoring classifier specifically designed for evaluating the quality of self-explanations.

Chapter 6

Experimental Study(3): Preprocessing Proposal - Employing Data Augmentation Techniques for Self-Explanation Autoscoring

6.1 Introduction

This chapter focuses on the critical dependencies on datasets, as highlighted in the previous chapter. Data quality plays a pivotal role in the success of our methodology, particularly in the context of educational technology. Motivated by this, we aim to explore strategies to enhance the quality of collected datasets. Moving beyond conventional threshold-based processing methods, this research delves into the implementation of more flexible scoring techniques tailored for Deep Learning models.

The limited availability of self-explanation data poses a significant challenge in this domain. This chapter outlines the process of constructing a learning model designed to work effectively within these constraints. A key focus is the development of a more accurate self-scoring classifier capable of evaluating the nuanced quality of self-explanations. By addressing this challenge, we aim to contribute methodologies and insights that can improve the efficacy of educational technologies and support personalized learning experiences.

Our goal is to generate enriched self-explanation content and improve the accuracy of automated grading systems for Japanese mathematics education. This research emphasizes careful alignment with the intricacies of the Japanese language and the specific challenges of mathematical reasoning. Through this work, we strive to advance the ca-

pabilities of automated systems in providing precise and meaningful feedback, fostering deeper engagement and learning outcomes for students.

6.2 Related Work

6.2.1 Automated Scoring of Self-Explanations: The Imperative for Rich Data

Self-explanation, widely recognized for amplifying learning outcomes in various fields, notably mathematics, has found its stride in the digital learning environment (M. Chi, Leeuw, Chiu, and LaVancher, 1994; Rittle-Johnson, 2006, 2017b).

The endeavor to automate the scoring of self-explanation quality has seen the integration of NLP tools and cutting-edge neural network architectures (Funayama et al., 2023). Techniques like latent semantic analysis (LSA) and recurrent neural network (RNN) interfaced with machine learning underscore the capabilities of automated systems, often outshining traditional manual evaluation in both effectiveness and efficiency (Crossley et al., 2019; Funayama et al., 2023; León et al., 2006; Özsoy et al., 2011; Panaite et al., 2018, 2019). Furthermore, semi-supervised learning techniques, which capitalize on abundant unlabeled data, have exhibited the potential to refine scoring accuracy (Cascante-Bonilla et al., 2020). Yet, the quest for more representative samples of self-explanations, especially in languages other than English, remains a prevailing challenge.

6.2.2 Augmenting Mathematical Self -Explanations Using Large Language Models

Consistent data shortages and imbalances have long impeded automated classification. Techniques like the synthetic minority oversampling technique (SMOTE), adaptive synthetic (ADASYN), and its derivatives aim to counteract these with synthetic data generation (Chawla et al., 2002; Han et al., 2005; He et al., 2008; Salazar et al., 2021), but their effectiveness wanes for scant datasets. Although a variety of synthetic data tools and research exist (Antulov-Fantulin et al., 2012; Berg et al., 2016; Dahmen and Cook, 2019; El Emam, 2020; Ping et al., 2017; Rubin, 1993), their applicability is often curtailed when facing complex domain-specific data, particularly in predicting student outcomes (Berg et al., 2016; Flanagan et al., 2022; Peña-Ayala, 2018). Navigating this landscape, LLM have made a compelling case as tools for text data augmentation. For instance, Dai

et al.’s AugLLM (Dai et al., 2023) capitalizes on ChatLLM (T. Brown et al., 2020) to generate supplemental text, emphasizing the rising importance of LLMs in the realm of mathematical content (Lightman et al., 2023). However, in the context of mathematical education, the accurate auto-scoring of self-explanations presents stark challenges. Crafting these self-explanations is not just labor-intensive but demands expertise in both the subject and linguistic expression. Coupled with the lack of diverse and representative self-explanation samples, particularly in languages other than English, the task becomes even more daunting.

6.2.3 Proposal

Given these challenges, amassing a vast and diverse collection of self-explanation samples is demanding. This predicament further complicates the development of systems designed to aid learning through extensive sets of self-explanation examples. To address these challenges and enhance automated scoring of self-explanations, we propose a semi-supervised approach that leverages the LLM. While popular LLMs such as OpenAI’s GPT-3 (T. Brown et al., 2020) are commonly employed in English language settings, due to the nature of the target problem we specifically focus on the Japanese variant of the model developed by CyberAgent (CyberAgent, 2023), based on GPT-NeoX (Andonian et al., 2021). This approach aims to explore the model’s potential for generating self-explanation sentences, which will serve as the foundation for our regression models designed to predict self-explanation scores. By incorporating the semi-supervised methodology and leveraging advanced language models, we aim to improve the accuracy and effectiveness of auto-scoring in the self-explanation learning domain. Our research is anchored by two pivotal questions:

RQ1: To what extent can the integration of self-explanations generated by the LLM Japanese model and mathematical material be used to enhance the accuracy of the predictive regression model for self-explanation scores?

RQ2: What is the optimal quantity of artificially generated pseudo-self-explanation data required to effectively improve the predictive performance of the model?

These research questions provide insights into maximizing the utility of the LLM Japanese model and refining data augmentation techniques. The core findings from our research are twofold. First, we propose a strategy for advancing automated scoring in math education by synergizing LLM-generated content and mathematical material. Second, we

highlight the ideal quantity of artificial self-explanation data for peak pre-dictive accuracy.

6.3 Problem Setting: The Learning Task

In this section, we primarily introduce the original human-labeled data, which serves as a foundation for the subsequent pseudo-labeling of the unlabeled samples, thereby bolstering the training process. Prior to delving into methodological details, it is pivotal to define the distinct learning task under examination, which underpins our methodological foundation.

6.3.1 Assessment of Self-Explanation Quality

For the evaluation process, two independent evaluators employed these rubrics to rate the collected 2205 self-explanations, scoring them on a scale ranging from 1 to 5. A quadratic weighted Cohen’s kappa coefficient (Cohen, 1960) of 0.749 between the evaluators indicated a significant level of agreement. The subsequent analysis used the mean score derived from both evaluators, which categorized the self-explanation scores for a roughly uniform distribution. Descriptive statistics of the collected self-explanations are presented in Table 6.1. In anticipation of the machine learning methodologies outlined in the subsequent chapters, our dataset was segmented into three distinct categories. The train dataset, which incorporates 1420 self-explanations, forms the fundamental basis for both training our models and for LLM data augmentation. Meanwhile, the valid dataset, comprising 355 self-explanations, is earmarked for the crucial tasks of fine-tuning our models’ parameters. It also plays a significant role in the evaluation of model accuracy and in ensuring model robustness. Lastly, the test dataset, which consists of 431 self-explanations, is designated to provide a measure of the performance of our finalized models.

Table 6.1: Descriptive Statistics of Graded Self-Explanations

Data Type	Num of Quiz	Variations in Math Units	Total Answers	Sentence Length (Character Count)		Quality Score	
				Mean	SD	Mean	SD
Train	40	8	1420	67.8	56.8	2.94	1.34
Valid	37	8	355	67.3	59.3	2.92	1.31
Test	8	3	431	63.7	53.2	2.81	1.25

6.3.2 The Text Regression Model Description

Inspired by the work of Wang et al. (Wang et al., 2019), we employ BERT (Vaswani et al., 2017) and a pre-trained BERT Japanese model (Suzuki, 2019) as the backbone for our regression models, which are intended to predict the quality scores of self-explanations. Wang et al.’s methodology of injecting rubrics into the system influenced the architecture of our model, making it specifically attuned to the grading of short responses. BERT’s deep learning model, grounded on a transformer architecture, has been recognized for surpassing most preceding models in diverse natural language processing tasks (Devlin et al., 2018). Given its robust performance and compatibility with the Japanese language, BERT is an ideal choice for our study. Our model takes as the input the preprocessed self-explanation text and the corresponding quiz title and yields as the output the predicted quality score for each self-explanation.

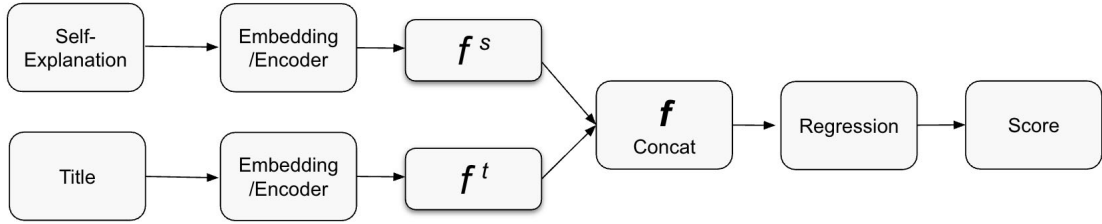


Figure 6.1: Workflow for the BERT regression models

6.4 The Proposed Method

In this section, we delve into our proposed method, building on the foundation. Our approach, illustrated in Figure 6.2, seamlessly blends human-labeled mathematical text data and LLM-generated data to enhance our machine-learning model. Drawing from Cascante-Bonilla’s semi-supervised framework (Cascante-Bonilla et al., 2020), we utilize pseudo-labeling as our primary technique. For human-labeled data, we lean on mathematical self-explanations, while the LLM and mathematical content texts help in producing pseudo-labeled samples to complement them. We gathered over 1420 self-explanation samples for the training model from undergraduate math students, which were further uti-

lized in the Japanese LLM (Step 1). Figure 6.2 provides a comprehensive visual overview of this pseudo-training mechanism and its integrated studies.

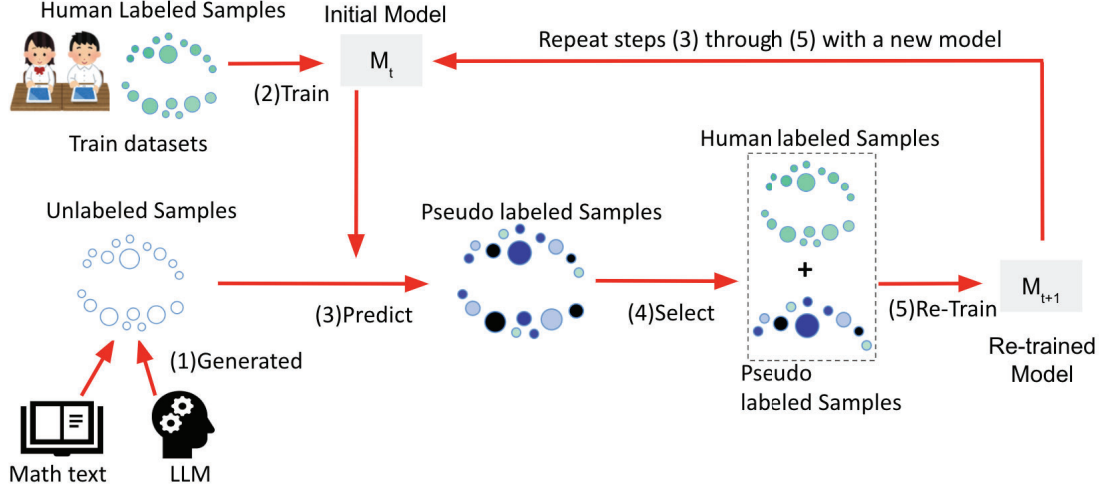


Figure 6.2: The overview of the pseudo-training process.

The pseudo-labeling technique commences with the training of an initial model using the labeled dataset (Step 2). This model then assigns labels to the unlabeled data, producing what we term ‘pseudo’ labels (Step 3). These newly formed pseudo-labels are then amalgamated with the original labeled dataset, initiating a cycle of continuous model enhancement (Steps 4–5). As the model’s predictive prowess escalates, the caliber of the pseudo-labels also elevates.

6.5 Pseudo-Labeling Training Algorithm

6.5.1 Dataset Categorization, Function Definitions, and Model Learning

Using the pseudo-labeling training algorithm, we integrate human-labeled datasets, LLM, and math texts and functions. The human-labeled datasets act as the primary foundation for our machine learning training, and the pseudo-labeled datasets further refine and expand it. A comprehensive breakdown and flow are described in detail in the following sections as definitions and pseudo-code. Figure 6.3 offers an illustrative representation of the training process, harmonizing with the methodology laid out in Figure 6.2.

Dataset categorization: Hereafter D_{type} represents an unlabeled dataset pertaining to a particular type, and D_{type}^* does a labeled dataset of the same category as D_{type} .

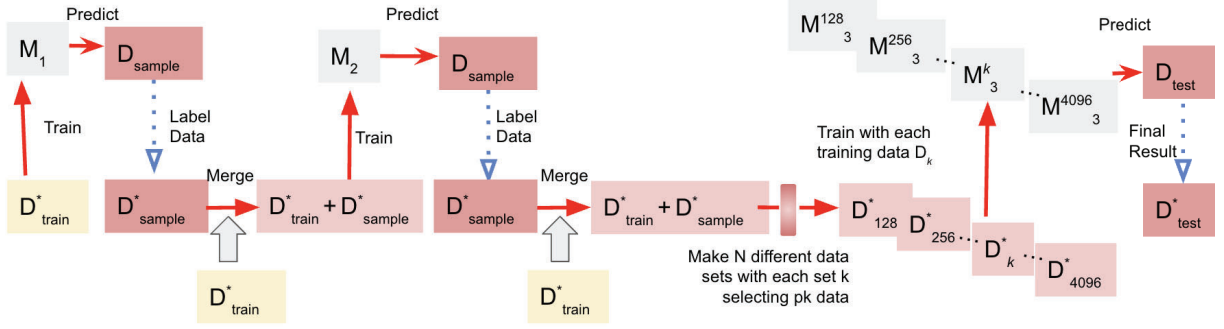


Figure 6.3: The Detail of Pseudo Training Process

Table 6.2: Statistics for Original and Generated Datasets

Data Type	Counts	Mean Score	Std
Original	2205	2.91	1.33
LLM	4096	3.39	1.44
Math	4096	3.87	1.77

D_{provided} is the composite dataset given to the model for training and evaluation, which includes the labeled training set D_{train}^* , test dataset D_{test} , and the generated unlabeled sample dataset D_{sample} .

$$D_{\text{provided}} = D_{\text{train}}^* + D_{\text{test}} + D_{\text{sample}} \quad (1)$$

Function definitions:

$$\text{Model}(\theta, D^*) = M \quad (2-1)$$

$$\text{Test}(M, D_{\text{test}}) = D_{\text{test}}^* \quad (2-2)$$

$$\text{Select}(D, k) = D_k \quad (\text{where } \text{Select}(D, n(D)) = D) \quad (2-3)$$

Model learning and final test:

$$M_1 = \text{Model}(\theta, D_{\text{train}}^*) \quad (3-1)$$

$$M_{t+1} = \text{Model}(\theta, D_{\text{train}}^* + \text{Select}(\text{Test}(M_t, D_{\text{sample}}), k_t)) \quad (3-2)$$

$$D_{\text{test}}^* = \text{Test}(M_T, D_{\text{test}}) \quad (\text{where } T \text{ is sufficient number of } t) \quad (3-3)$$

Table 6.3: Algorithm 1: Semi-supervised learning with pseudo-labeling

Algorithm 1: Semi-supervised learning with pseudo-labeling

Input:

- Labeled dataset: $\mathbb{D}_{\text{train}}^*$
- Unlabeled dataset: $\mathbb{D}_{\text{sample}}$
- Test dataset: $\mathbb{D}_{\text{test}}$
- Maximum number of iterations: T
- Confidence threshold for pseudo-labeling: *confidence*
- Convergence threshold for change in consecutive scores: ϵ

Output:

- Trained model: *NLP model*
- Final score of the model on $\mathbb{D}_{\text{test}}$: s

Procedure:

Initialize model:
NLP model $\leftarrow$ trained NLP model with $\mathbb{D}_{\text{train}}^*$

Initialize values:
previous score $\leftarrow 0$
iteration $\leftarrow 0$
convergence achieved $\leftarrow \text{False}$

While *iteration* $< T$ **and not** *convergence achieved* **do**

 a. **Predict pseudo-labels:**
 predictions $\leftarrow$ predicted labels of $\mathbb{D}_{\text{sample}}$ using *NLP model*
 confidence $\leftarrow$ threshold when labeling of $\mathbb{D}_{\text{sample}}$ using *NLP model*

 b. **Filter high confidence predictions:**
 $\mathbb{D}_{\text{sample}}^* \leftarrow \{ \text{sample} \mid \text{sample in } \mathbb{D}_{\text{sample}} \text{ and } \text{confidence of the sample} > \text{confidence} \}$

 c. **Merge labeled and pseudo-labeled data:**
 $\mathbb{D}_{\text{train}}^* \leftarrow \mathbb{D}_{\text{train}}^* + \mathbb{D}_{\text{sample}}^*$

 d. **Retrain the model:**
 NLP model $\leftarrow$ trained NLP model with $\mathbb{D}_{\text{train}}^*$

 e. **Evaluate current model performance:**
 current score $\leftarrow$ Evaluation metrics when labeling $\mathbb{D}_{\text{test}}$ using *NLP model*

 f. **Check for convergence:**
 If $|\text{current score} - \text{previous score}| < \epsilon$:
 convergence achieved $\leftarrow \text{True}$
 previous score $\leftarrow$ *current score*

 g. **Update iteration count:**
 iteration $\leftarrow$ *iteration* + 1

End Procedure

Parameter Setting in Our Study: Figure 6.3 provides a comprehensive outline of our experimental approach. Our dataset consists of both human-annotated and unlabeled samples. For the training process, we amassed 2205 self-explanation samples from student

contributors. In our setting, θ stands for a model built using logistic regression with text representation acquired from BERT, a state-of-the-art transformer-based model renowned for its superior performance on numerous NLP tasks. The iterative process continues for 3 timesteps; in other words, T is set to 3. The selection size at each timestep t , denoted by k_t , varies as follows: k_1 is equal to the total number of data points in D_{sample} , such as $n(D_{\text{sample}})$, whereas for the second time step, k_2 could be any one of the following: 128, 256, 512, 1024, 2048, or 4096. To distinguish them, we defined the $(t+1)$ th model, which is the same as M_{t+1} in Formula (3-2), which was learned with selected k_t training data as $M_t^{(k_t)}$. The following formula represents the concrete model learning method in the study:

$$M_1 = \text{Model}(\theta, D_{\text{train}}^*) \quad (4-1)$$

$$M_2 = \text{Model}(\theta, D_{\text{train}}^* + \text{Select}(\text{Test}(M_1, D_{\text{sample}}), n(D_{\text{sample}}))) \quad (4-2)$$

$$M_3^{2^{(i+7)}} = \text{Model}(\theta, D_{\text{train}}^* + \text{Select}(\text{Test}(M_2, D_{\text{sample}}), 2^{(i+7)}))$$

6.5.2 Pseudo-Data Preparation: LLM Usage and Mathematical Material

We employed a pseudo-labeling technique to enrich our dataset, sourcing additional self-explanation data via the Japanese LLM and mathematical material. Given alternatives such as OpenAI’s GPT-3 (T. Brown et al., 2020), our preference leaned towards Cyber-agent’s LLM (CyberAgent, 2023) due to its open-source availability and its adeptness in the Japanese language, perfectly complementing our dataset. To gather data, the LLM tackled mathematical contexts and formulated pertinent explanations. Our methodology was as follows:

- Random data selection: We began our process by randomly selecting 30% from our human-labeled training dataset to capitalize on the rich diversity of student-generated self-explanations.
- Keyword extraction: Ten keywords were extracted from each self-explanation, encapsulating its essence, guiding LLM to produce contextually relevant data.
- LLM generation: Armed with the extracted keywords, we then proceeded to prompt the LLM. Specifically, each set of 10 keywords was used as seed input, directing the LLM to generate contextually coherent pseudo-self-explanation data. The model

was given a directive to ‘elaborate based on the provided keywords,’ ensuring the generated content maintained relevance to the original self-explanation context.

Approximately 19,000 entries were generated, with a random subset of 4096 used for experiments. This combination of pseudo- and human-labeled data broadened our training set, enhancing the automated scoring system’s performance without compromising quality.

We also leveraged the math quiz texts dataset, populated with standard mathematical solutions. Its rich mathematical material and contextual problem-solving methods made it invaluable for generating mathematical self-explanations.

6.5.3 Comparative Analysis of Original and LLM-Generated Dataset

In Table 6.4 and Figure 6.4, a detailed comparative analysis between the original and the synthetically generated datasets is elucidated. Upon examination, it becomes evident that the augmented datasets exhibit a modest augmentation in their average quality metrics relative to the foundational dataset.

Table 6.5 showcases a range of these samples with their respective translations in English, alongside the scores predicted by our automated system. It is evident from the samples that as the complexity of the self-explanation increases, so does the predicted score. For instance, the first sample, which delves into the relationship between solutions and coefficients, is accorded a high score of 5.00 due to its comprehensive self-explanation. Conversely, the final example, which is a brief reflection on an error and an attempt to rectify it, receives a lower score of 1.23, reflecting its succinct nature.

Table 6.6 offers various math texts that were considered during our study. These texts, extracted from educational resources, shed light on various mathematical principles and theorems. The predicted self-explanation score gives an indication of the depth and comprehensiveness of the self-explanation provided in each math text. For instance, the math text delving into the ‘angle bisector and ratio, using Ceva’s theorem’ provides a nuanced exploration of the topic, earning it a top score of 5.00. In contrast, the examination of the sizes of the three angles of a triangle is more straightforward, reflected in its score of 2.66. The varied scores underline the diverse complexity levels present in math texts.

The inherent variability in the complexity of mathematical texts and self-explanations, as illustrated by Tables 5 and 6, underscores the importance of our automated scoring system. It demonstrates the system’s capability to adaptively discern and quantify the

nuances in self-explanation quality.

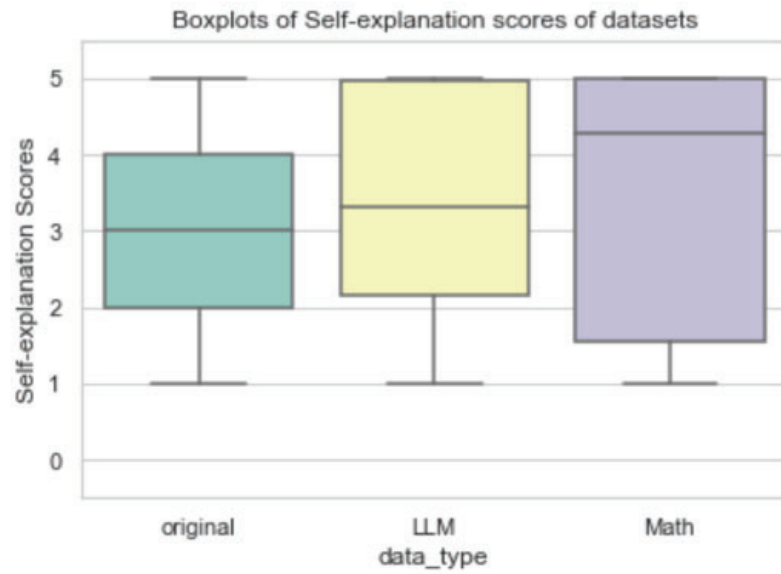


Figure 6.4: Boxplots of self-explanation scores.

Table 6.4: Comparative metrics for original and generated datasets.

Data Type	Counts	Mean Score	Std
Original	2205	2.91	1.33
LLM	4096	3.39	1.44
Math	4096	3.87	1.77

Table 6.5: LLM-generated samples

LLM-Generated Samples (Original)	LLM-Generated Samples (English Translation)	Predicted Self-Explanation Score
その後、$\alpha + \beta = -a - 2$と$\alpha\beta = 2a$の関係から解と係数の関係が分かる。 次に、問題で言及されたαとβを用いて式を展開し整理し、右辺を0にする。さらに式を工夫して代入が可能な形にする。そして、関係式$\alpha + \beta = -a - 2$と$\alpha\beta = 2a$を式に代入して簡略化し、$a^2 + a - 6 = 0$となる。これを因数分解してaの解を求めると、$a = -3$と$a = 2$が得られる。その後、$a = 2$を代入すると解が重解になり、解が一つしかなくなるため問題に合わない。したがって、正しい答えは$a = -3$である。その後、重積算や累乗などいろいろな計算が解に出てきて混乱するだろうから丁寧に説明していく。	After that, $\alpha + \beta = -a - 2$ and $\alpha\beta = 2a$ can be understood from the relationship between the solutions and coefficients. Next, expand and organize the expressions using the α and β mentioned in the problem, and set the right side to 0. Further manipulate the equation to transform it into a form that allows for substitution. Then, substitute the relationship $\alpha + \beta = -a - 2$ and $\alpha\beta = 2a$ into the equation and simplify it, resulting in $a^2 + a - 6 = 0$. Factorize this to find the solutions for a, which are $a = -3$ and $a = 2$. Afterward, when $a = 2$ is substituted, the answer becomes a repeated root, which means there is only one solution, making it inconsistent with the problem. Therefore, the correct answer is $a = -3$. Afterward, various calculations such as product of roots and exponentiation may appear in the solution, and it may be confusing, so I will explain it carefully.	5.00
代入してみたら、平方根の中が計算できた。αとβを求めた。αに-2を代入できた。解を吟味して、答えは-3だった。その後、$a = -4.7 \times 5 + 7$として計算したが誤差があった。そこで右側の5を左側の4に変形すると正しい値が求められた（ただし小数点以下を切り捨て）。 この問題は高校数学の範囲外であるが、中学 高校で習った知識で解くことができたので紹介したい。	Substituted it into the formula for solutions. Calculated the value inside the square root. Found α and β. Substituted $\alpha = -2$. Examined the solution, and the answer was -3. Afterward, calculated it as $a = -4.7 \times 5 + 7$, but there was an error. So I rearranged the 5 on the right side to 4, and the correct value was obtained (with the decimal part truncated). This problem is beyond the scope of high school mathematics, but it could be solved using knowledge learned in middle and high school, so I want to introduce it.	4.08
グラフを書いた傾きが-4の$y = x^2$の接線を求めた。その後、接点を原点とし両辺に点a, bを代入すると「sin-cos」と出たので、「COS(1)」で求めた。これで式の意味が分かったと思う。-5は角度が変化しないからゼロでないといけないのだ！	Used the slope obtained from graphing to find the tangent of $y = x^2$ with slope -4. Took the point of tangency as the origin and substituted points a and b on both sides. 'Sin-cos' was obtained, so I calculated it using 'COS(1)'. I think I understood the meaning of the equation now. -5 does not change the angle, so it must be nonzero!	3.02
$x = 3$を代入する。方程式を解き進める。解の吟味を書く。間違いに気づき、修正する。最終的な答えを書く。その後、再帰的に解く。	Substituted $x = 3$. Proceeded to solve the equation. Wrote the examination of the solutions. Noticed the mistake and corrected it. Wrote the final answer. Afterward, solve it recursively.	2.18
前のは間違えたため、全部消した。その後、通分してみた。	Since the previous one was incorrect, I deleted everything and then performed the common denominator. Afterwards, something like this.	1.23

Table 6.6: Math Texts samples

Math Texts Samples	Predicted Self-Explanation Score
Angle bisector and ratio, using Ceva's theorem: Revised version succeed math A problem 349 Δ (Triangle), let Δ have the angle bisector of $\angle$ and the point where it intersects the side, and the point that divides the side in the ratio. When the line intersects at point, find the length of the side.	5.00
Using Menelaus's theorem: Segment ratio and area ratio, revised version succeed math A problem 350 Δ , let be the point where it divides the side in the ratio, and the point where the segment is divided in the ratio, and the point where the extension of the segment intersects the side. Find the following segment ratios and area ratios $\Delta:\Delta$	4.93
Using the relationship between sides and angles: Range of values for side length in a triangle, revised version succeed math A problem 355, determine the range of values for so that a triangle with the following side lengths exists.	3.84
Using the relationship between the sizes of three sides: Proving inequalities related to segment lengths, revised version succeed math A important example 66, take point inside Δ , and join and prove that . Abbreviated.	3.13
Examining the sizes of the three angles of a triangle, revised version succeed math A important example 64, examine the sizes of the three interior angles of Δ .	2.66

6.6 Experiments and Evaluations

6.6.1 Exploring the Influence of Self-Explanation Augmentation on Model Efficiency

We embarked on an exploration to discern the influence of self-explanation augmentation on the efficiency of an automated self-explanation scoring model across diverse datasets. We used mean absolute error (MAE) metrics to evaluate model performance, giving insights into the extent of error deviation and the efficacy for individual items. Tables 7 and 8 lay out the results of our experiments, contrasting performances across different dataset permutations.

When we introduced augmented datasets into the mix, distinct variations in performance emerged. Remarkably, our model, when nurtured with a blend of the ‘math’ and ‘original dataset’, consistently delivered the most desirable MAE results. This underlines its superior predictive precision in assessing self-explanation quality. Such results lend credence to the efficacy of the model when trained with this specific data amalgamation.

On another note, the ‘mixed’ model—which weaves together human-graded samples, LLM-crafted pseudo-sentences, and mathematical content—also demonstrated notable improvements. This outcome underscores the model’s robustness and flexibility when fed with diverse data sources. Yet, the model named ‘only_LLM_math’, which exclusively depended on LLM-created sentences, trailed behind the foundational model in terms of performance. This observation underscores the criticality of harmonizing human-judged and machine-produced data to achieve optimal results.

Table 6.7: Datasets Overview

Dataset	Base_Line	LLM	Math	Mixed	Only_LLM_Math
Original (N = 1420)	•	•	•	•	
LLM- generated (N = 4096)		•		•	•
Math texts (N = 4096)			•	•	•
Total Number of Data	1420	5516	5516	9612	8192

Table 6.8: Model Performance for Various Datasets (MAE)

Data Type	Base_Line	LLM	Math	Mixed	Only_LLM_Math
Test	0.749	0.699	0.646	0.692	1.135
Val	0.602	0.341	0.358	0.336	1.033

6.6.2 Evaluating Optimal Quantity of Pseudo-Self-Explanation Data

The data from Tables, as well as the visual cues from Figures , underline the importance of calibrating the amount of added pseudo-self-explanation data. While adding more datasets often improves performance, there is a saturation point beyond which the returns diminish. This trend underscores the need for a balanced approach, optimizing the amount of data incorporated to ensure robust and generalizable results. It also emphasizes the intricate interplay between different datasets, with ‘LLM’ and ‘mixed’ data offering particularly promising outcomes. The ‘baseline’ row signifies the MAE when the model is trained only on the original dataset, devoid of any pseudo-self-explanation data. Each subsequent row shows the MAE when the model is trained with an increasing volume of pseudo-self-explanation data, ranging from 128 to 4096 datasets.

Upon examining the LLM model, we note an enhancement in performance when the added datasets increase from 128 to 256. Beyond this, the further addition of generated data does not significantly reduce the MAE, suggesting an optimal balance between data augmentation and model efficacy with an addition of 256 datasets. The `math` model displays a similar trend, with the lowest MAE achieved when 1024 datasets are added. Beyond this point, no substantial performance enhancement is observed with extra data. For the `mixed` model, we see a consistent improvement in performance with increased data, but this plateaus beyond 2048 datasets, where the MAE slightly increases. Conversely, the `only_LLM_math` model shows erratic trends. Its performance varies noticeably with the quantity of added data and consistently exceeds the baseline model’s MAE, regardless of the added data volume. This reveals potential difficulties when exclusively relying on generated pseudo-self-explanation data.

Table 6.9: Validation MAE with varying amounts of added pseudo-self-explanatory data

Dataset	Number of Datasets Added					
	128	256	512	1024	2048	4096
base_line			0.60			
LLM	0.57	<u>0.35</u>	0.51	0.49	0.40	0.34
math	0.40	0.50	0.43	<u>0.35</u>	0.40	0.36
mixed	0.59	<u>0.32</u>	0.52	0.44	0.40	0.34
only_LLM_math	1.19	0.90	0.96	<u>0.81</u>	1.02	1.03

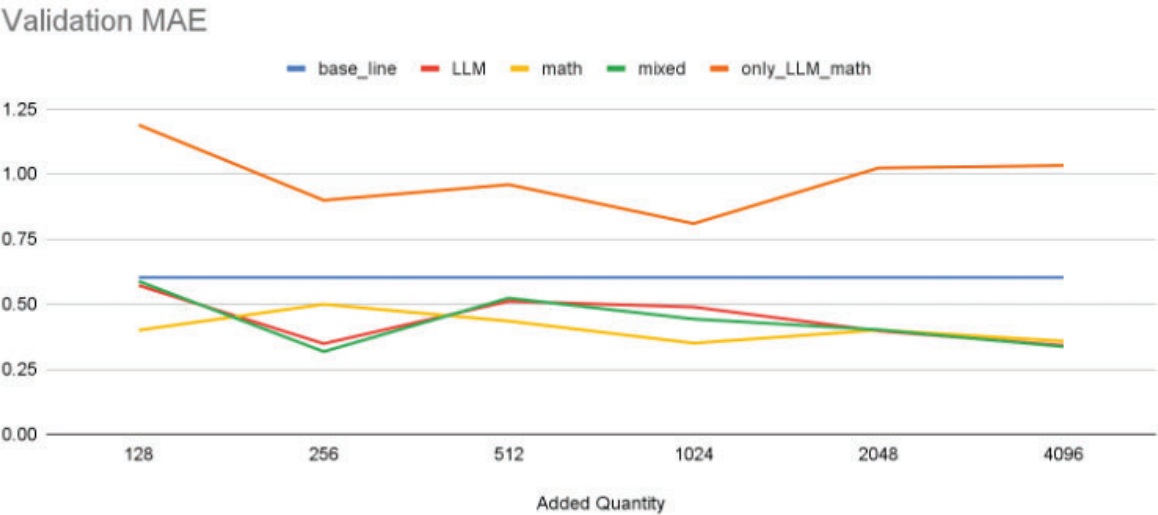


Figure 6.5: Validation MAE plot

Table 6.10: Test MAE with varying amounts of added pseudo-self-explanation data

Dataset	Number of Datasets Added					
	128	256	512	1024	2048	4096
base_line			0.75			
LLM	0.67	0.63	0.72	0.72	0.71	0.7
math	0.64	0.66	0.67	0.64	0.65	0.65
mixed	0.68	0.66	0.71	0.68	0.73	0.69
only_LLM_math	1.19	0.96	1.02	0.89	1.15	1.14



Figure 6.6: Test MAE plot

6.7 Discussion

6.7.1 Detailed Analysis of Results (RQ1)

Regarding Research Question 1, an in-depth analysis of the results reveals several noteworthy observations regarding the influence of self-explanation augmentation on the model’s performance. In the test category, we observe an improvement in the model’s performance when transitioning from the baseline to the LLM and `math` models. Notably, the `math` model achieves the lowest MAE at 0.646, which aligns with Dai et al. (2023) proposition that data augmentation at the semantic level improves robustness and consistency. However, the performance slightly deteriorates in the `mixed` model and substantially plummets in the `only_LLM_math` model. This suggests that an excessive concentration of LLM-generated self-explanations could impair the model’s predictive proficiency.

A similar pattern emerges when examining individual topics within the ‘test’ category. For instance, the model delivers optimal performance for ‘quadratic equations’ with the LLM-generated model, but the performance markedly deteriorates when solely relying on LLM-generated self-explanations. The validation category follows a similar trajectory, with the LLM, `math`, and `mixed` models outshining the baseline model. Once again, the `mixed` model achieves the smallest error. However, the ‘`only_LLM_math`’ model experiences a decline in performance, further highlighting the advantages of using a diverse dataset that encompasses both human-evaluated and machine-generated explanations.

6.7.2 Findings and Observations (RQ2)

Regarding Research Question 2, the results presented in Tables 6.10, and 6.9 provide valuable insights into determining the optimal quantity of generated pseudo-self-explanation data that can enhance the model’s performance. For the ‘LLM’ model, an initial improvement in model performance is observed as the number of added datasets increases from 128 to 256. Beyond this point, further augmentation of the generated data does not lead to a significant reduction in MAE, suggesting that adding 256 datasets strikes an optimal balance between data augmentation and model performance. The ‘`math`’ model exhibits a similar pattern, with the lowest MAE observed when 1024 datasets are added, and no significant performance improvements resulting from further data augmentation. The ‘`mixed`’ model, on the other hand, shows a general trend of performance enhancement with increased data augmentation, up to a threshold of 2048 datasets, beyond which the

MAE slightly increases. In contrast, the ‘only_LLM_math’ model does not present a consistent trend. Its performance fluctuates significantly as the volume of added data increases, and its MAE consistently surpasses that of the baseline model, regardless of the amount of added data. This underscores the challenges of solely leveraging generated pseudo-self-explanation data for augmentation, particularly when the model might lack domain-specific expertise, echoing concerns raised by (Dai et al. (2023)).

Our study underscores that while weaving in generated pseudo-self-explanation data offers potential advantages, there is a point of diminishing returns. Over-sampling methods, such as SMOTE and ADASYN, can expand small datasets and provide a broader learning spectrum for models (Salazar et al., 2021). Although our focus has been on LLM pseudo-labeling, merging this with oversampling could be a promising direction for future research, aiming to further refine our models. Our study underscores the importance of careful and context-aware adjustments when adopting data augmentation strategies in developing self-explanation auto-scoring models.

6.7.3 Limitations

This study highlights several limitations that warrant careful consideration. The dataset used for this research was restricted to mathematics, which may limit the generalizability of the findings to other subjects. Mathematics, with its unique linguistic and conceptual structures, may not fully represent the challenges present in other disciplines, where self-explanations might differ in complexity and focus. Furthermore, the methodology relied heavily on the ability of the LLM to generate pseudo-self-explanation data. While this approach enhanced the dataset’s size and diversity, it also introduced noise and errors, which impacted the overall system performance.

The quality and representativeness of the labeled data significantly influenced the model’s effectiveness. Poorly labeled or biased datasets compromised the system’s efficacy, highlighting the need for more precise and diverse training datasets. Additionally, performance variability across mathematical topics was evident. For instance, the system demonstrated higher prediction accuracy in topics such as the "property of a circle" (0.242) compared to "quadratic functions" (0.419) within the validation dataset. These disparities suggest that the effectiveness of self-explanation augmentation is closely linked to the inherent complexity of the subject and the linguistic nuances of the explanations.

Another challenge was the limited size of the evaluation dataset in certain mathemat-

ical categories. Small datasets for specific topics hindered the ability to draw definitive conclusions, as the ease of inference varied significantly across topics like linear functions, shapes, equations, and square roots. Some mathematical concepts are inherently more challenging for machine learning due to their linguistic or conceptual intricacies, further complicating system training and evaluation.

6.7.4 Conclusion

This study addressed the challenge of limited self-explanation data variety and quantity for preprocessing in automated scoring systems. By employing a semi-supervised approach with a Japanese LLM, we enhanced the dataset used for scoring mathematical self-explanations. The findings demonstrate that combining LLM-generated explanations with mathematical content improves system effectiveness. Notably, the results revealed an optimal threshold for the quantity of synthetic self-explanation data that benefits the system; exceeding this threshold yielded no further improvement, underscoring the need for balanced data augmentation strategies.

Despite these advancements, several challenges remain. The system’s effectiveness depends heavily on the users’ writing style, academic level, and motivation. The time-intensive nature of self-explanation poses a significant hurdle for large-scale data collection, complicating the practical application of feedback systems (McEldoon et al., 2013). Additionally, producing high-quality self-explanations requires proficiency in both subject matter and writing skills, which not all learners possess (Hodds et al., 2014; Panaite et al., 2018). These challenges highlight the need for continued exploration of methods to enhance feedback systems. Future research should focus on refining data augmentation strategies, improving the adaptability of systems to diverse learner profiles, and addressing the scalability of self-explanation-based methodologies. These efforts will be critical in developing systems that are not only effective but also practical for widespread educational use.

Chapter 7

Experimental Study(4): Automatic Feedback for Enhancing Self-Explanation Skills

7.1 Introduction

The ability of students to articulate complex mathematical thought processes through writing poses a significant challenge, particularly given the wide range of learning levels in a typical classroom. Self-explanation, a critical component of effective learning, encourages students to reflect on their reasoning, identify gaps in understanding, and solidify their knowledge. However, previous studies have highlighted substantial barriers to the widespread adoption of self-explanation as a learning tool. Two primary challenges emerge: the time-intensive nature of self-explanation and the need for robust automated feedback systems that can effectively support learners.

One of the primary challenges in promoting self-explanation is the significant time and effort required for students to articulate their reasoning in writing. This time-intensive process often discourages learners, limiting their engagement and reducing the broader applicability of self-explanation methods. Addressing this challenge necessitates innovative strategies that streamline the process and motivate students to actively participate. By developing and implementing feedback systems that are both efficient and engaging, this research seeks to lower the cognitive and temporal barriers associated with self-explanation, making it a more accessible and effective learning strategy.

Crafting high-quality self-explanations poses an additional challenge, as it demands a combination of subject matter expertise and writing proficiency—skills that many learners

may lack. Automated feedback systems hold considerable potential to bridge this gap by assessing the quality of self-explanations and providing personalized guidance. However, existing systems often fail to deliver feedback that is constructive, actionable, and tailored to individual learner needs. This research focuses on optimizing automated feedback systems to support students more effectively, particularly those who struggle to articulate their thought processes. The objective is to develop systems capable of evaluating self-explanations with precision while offering tailored suggestions to help learners refine their reasoning and communication skills.

To maximize effectiveness, it is crucial for feedback systems to deliver personalized and specific guidance. Without targeted support, students may disengage from self-explanation activities before they fully develop the skills needed to succeed (Arner et al., 2021). However, generating such detailed feedback for unstructured tasks—particularly in new or less conventional subject areas—requires significant time and resources. This preparation, setup, and customization place a substantial burden on educators, posing a challenge for widespread adoption in classrooms. Teachers aiming to incorporate self-explanation prompts into their instruction must navigate the difficulty of crafting feedback that is both accurate and actionable, which requires a comprehensive understanding of the subject matter and significant investment of time.

By addressing these interconnected challenges, this research aims to advance the field of educational technology by offering practical solutions that enhance student engagement, streamline the self-explanation process, and optimize feedback delivery. These efforts are designed to empower learners, foster deeper understanding, and promote the effective integration of self-explanation methodologies across diverse educational contexts, including mathematics education and beyond.

We introduce the self-explanation-based automated feedback (SEAF) system—a unique synthesis of established feedback principles and implementations. SEAF offers instantaneous feedback by harnessing real-time data analytics and machine learning techniques. This allows for a feedback system that is not just tailored but is adaptively attuned to learners’ evolving self-explanation patterns. Our study was designed to address two primary research questions, as follows:

RQ1: Can a classmate’s self-explanations be presented and useful to the learner as feedback?

RQ2: Which students benefit most from using the SEAF system?

7.2 Previous Studies

7.2.1 Feedback System and Our Positioning

Learning interventions are implemented to augment normal teaching to enhance the cognitive, metacognitive and affective aspects of learning, and improve motivation (Hattie et al., 1996). One area of particular interest that has been found to be an effective learning intervention is self-explanation, a cognitive process that involves generating explanations to deepen one’s understanding of a concept or solution (Bisra, Liu, Nesbit, et al., 2018; Rittle-Johnson, 2017b). Research has found that self-explanation can promote both conceptual and procedural knowledge, directing attention to the structural features of intellectual content (M. Chi, Leeuw, Chiu, and LaVancher, 1994; Rittle-Johnson, 2006). While the self-explanation strategy has been found to be effective, it requires scaffolding to encourage students to use it in effective ways. Reflection mechanisms demand higher cognitive skills that students might not have yet acquired, also some students could be reluctant or not be motivated to use their skills in the course of study, and as a result, only a small percentage of students engage in generating explanations in the spontaneous self-explanation condition (M. T. H. Chi et al., 1989; Renkl, 1997). In previous research, instructors and researchers have often employed scripts to facilitate the self-explanation task by manually creating prompts, chapter-based exercises and templates, and feedback (Bisra, Liu, Nesbit, et al., 2018), including scaffold templates based on sample sentences (Berthold, Eysink, and Renkl, 2009; Berthold and Renkl, 2009; Rittle-Johnson, 2017b). However, it has also been suggested that research into computer-generated prompts and feedback could improve the scalability of self-explanation while ensuring consistency in the quality of the support that is required (Arner et al., 2021). Feedback is pivotal in academic improvement (Hattie, 2009) and aligns with SDG 4’s focus on sustainable, quality education (Colglazier, 2015). Emerging educational systems, especially in STEM fields, offer scalable, real-time feedback that enhances learning (Hwang et al., 2020; Lu et al., 2020; Su et al., 2023). Prior studies on computer-aided feedback systems like iSTART (Boonthum et al., 2007; Levinstein et al., 2007; McNamara et al., 2004) and StairStepper (Arner et al., 2021) have focused on enhancing self-explanation skills in reading comprehension as well as in writing tasks like summarizing and paraphrasing. These systems provide automated feedback to support learners. However, as pointed out by O’Neil et al. (2014), the design of self-explanation prompts should strike a balance. They should en-

courage cognitive processes that wouldn't otherwise take place, without being so intricate that they distract or hinder the learner. To achieve maximum effectiveness, personalized and specific feedback for self-explanation is crucial. Lack of such targeted support may cause students to disengage from self-explanation activities before they can fully develop their skills in this area (Arne et al., 2021). However, generating feedback for unstructured tasks in new or less common areas demands significant time and resources. This level of preparation, setup, and customization is a burden for educators, making it difficult to widely implement these strategies in classrooms. This presents a challenge for teachers who are keen on incorporating self-explanation prompts into their instruction. Additionally, crafting feedback that is both accurate and actionable necessitates a comprehensive understanding of the subject matter.

7.2.2 Consideration on Existing Intelligent Tutoring Systems

In the realm of ITS, several platforms have established a significant presence by catering to various facets of educational needs. Table 7.1 offers a comprehensive overview of the features and functionalities related to self-explanation across four distinguished ITS platforms: SEAF, Mathia (Learning, 2023; Ritter et al., 2007), Assessment's ("Assessment's," 2023; Heffernan and Heffernan, 2014) and I-START (McNamara et al., 2004). The table elucidates the varying degrees to which each system emphasizes self-explanation, shedding light on the nuances of their approaches. At a glance, SEAF emerges as a leader in integrating peer-based explanations, a feature relatively unexplored in other platforms. While Mathia and Assessment's subtly incorporate self-explanation as part of their broader framework, I-START and SEAF prioritize it as a core instructional strategy.

While all platforms, including SEAF, exhibit personalization to varying extents, the depth and nature of this personalization differ. For instance, while Mathia focuses on real-time adjustments to student needs, I-START tailors its challenges based on text complexity. Notably, SEAF and Assessment's stand out for their adaptive nature, ensuring the content and feedback are in sync with the learner's evolving proficiency. The central focus of our analysis, self-explanation, showcases varying levels of integration across platforms. Both SEAF and I-START treat self-explanation as a core feature, with SEAF uniquely offering peer-based explanations. In contrast, Mathia and Assessment's incorporate self-explanation more subtly, using it as a tool within a broader problem-solving context. Another distinguishing facet is the integration of peer-based learning.

SEAF’s emphasis on sharing ‘pen strokes’ and peer explanations sets it apart from its contemporaries, offering learners a unique, collaborative perspective.

Table 7.1: Comparative analysis of self-explanation features in prominent intelligent tutoring systems

Feature/Aspect	SEAF (Ours)	Mathia	Assessment’s	I-START
Main focus area	Mathematics	Mathematics	Mathematics	Reading comprehension
Personalization	Yes, personalized message	Real-time adjustments	Adaptive learning	Adaptive to text complexity
Feedback mechanism	Peer-based explanations	Immediate, step-by-step	Hints and immediate feedback	Immediate on self-explanations
Self-explanation emphasis	High (core feature)	Moderate (within problem-solving context)	Moderate (user responses to hints/questions)	High (core training)
Peer-based learning	Yes (pen strokes, explanations)	No	Collaborative problem-solving (in some cases)	No
Real-world application	Not specified	Contextual word problems	Real-world problem sets	Complex real-world texts
Customizability	High (based on users’ data)	Not specified	Yes (teacher-customizable content)	Not specified

7.2.3 Conceptual Framework and Research Inspiration

The underpinning conceptual framework of this study intertwines the dynamic role of self-explanation in learning with the challenges encountered at critical learning impasses. Rooted in cognitive science, self-explanation is perceived as a catalyst that bridges the gap between a learner’s current comprehension and their aspirational knowledge base. This is particularly evident when learners actively articulate their thought processes (Renkl, 2017). The inception of this research stems from recognizing self-explanation’s diagnostic potential in identifying and addressing learning impasses. This insight is coupled with the ambition to harness the rich vein of educational big data, aiming to elevate the customization of learning experiences in digital environments.

7.2.4 Positioning of the Research and Its Novelty

This research is strategically positioned at the confluence of cognitive science, educational technology, and data analytics, marking a novel intersection in academic inquiry. Its unique contribution lies in elucidating the role of self-explanation within the context of mathematics education, especially in online learning platforms. Employing an innovative methodological approach, this study compares learner-generated explanations against those of experts, thereby deepening our comprehension of effective self-explanation strategies in educational settings. Moreover, it addresses a notable void in existing academic discourse by exploring the processing and enhancement of the quality of self-explanations.

This novel perspective not only enriches the literature but also provides practical insights into the application of ITS in mathematics education, paving the way for more informed and effective pedagogical strategies.

7.3 The Architecture of the SEAF System

The proposed system architecture consists of three main components: data collection, data processing, and feedback delivery as shown in Figure 7.1. In the data collection component, self-explanations are collected into the main database using online platforms. Secondly, the feedback delivery component monitors the database and instructs the data processing component to analyze the data. The data processing component is responsible for processing the raw data collected and generating insights to inform the feedback generation process. The feedback was delivered to the learner if requested through an online platform and learners can then review and reflect on the feedback in real-time as a result.

7.3.1 Feedback Delivery

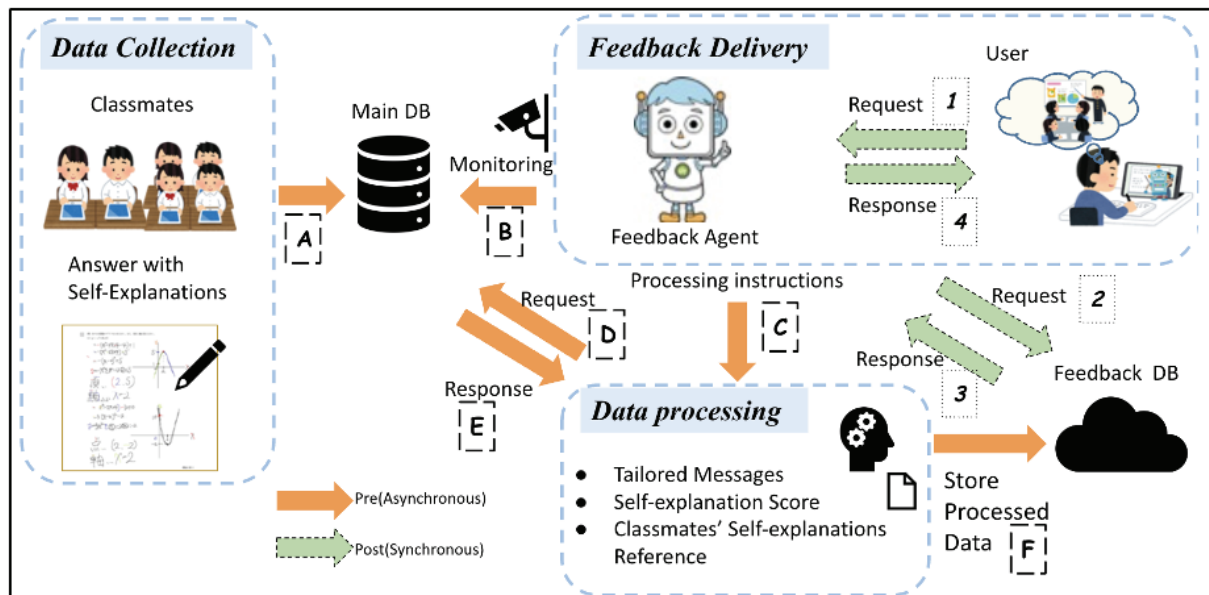


Figure 7.1: The architecture of the SEAF system

The feedback delivery component is a critical component, as it provides personalized and real-time feedback to each individual learner. The proposed architecture (Figure 7.1) comprises two modules: a synchronous module for sequential processing and an asynchronous module for database monitoring. The asynchronous module (Node B) monitors

the database for updates and changes, triggering the preprocessing component to process the updated data (Node C). It then retrieves the updated data from the main database (Node D and E), processes it, and saves the trained weight to the feedback database (Node F). The synchronous module (Node 1) processes each user request in the order received and requests the newest trained weight from the feedback database (Nodes 2 and 3). This allows for the efficient handling of multiple requests while maintaining their order. Finally, the architecture adapts quickly to database changes and provides up-to-date feedback to learners (Node 4).

7.3.2 Data Processing

The SEAF system’s data processing component involves analyzing the self-explanations of learners, identifying key concepts, evaluating writing quality, and understanding learners’ thought processes (Node F in Figure 7.1). Its personalized feedback features use classmates’ learning processes as a reference for comparison, helping learners understand their progress and identify improvement areas (Bisra, Liu, Nesbit, et al., 2018). Additionally, communicating strategic messages of learning in an individualized and optimized way may prevent mistakes and provide a deeper understanding (Dodeen, 2015; Hong et al., 2006). Social cognitive theory (Bandura and of Mental Health, 1986; Schunk and Pajares, 2005) served as a common background for the development of the three effective learning support tools. The system’s comparative analysis of self-explanation writing promotes healthy competition and collaboration among learners (Hall and Vance, 2010), potentially leading to better learning outcomes.

7.3.3 Personalized Messages

Personalized Messages Descriptions

The personalized messages feature is a tool that offers personalized feedback to each learner based on their self-explanation quality and learning progress. The system generates feedback messages automatically, using predefined templates that are partly customized to each learner’s learning logs. The feature encourages students to continue learning and improving their self-explanation writing skills by providing individualized feedback. Previous research (Dodeen (2015) and Takallou et al. (2015)) suggests that writing a good self-explanation will deepen students’ understanding, and taking time to

learn will reduce errors and deepen learning.

The feature also includes a missing knowledge detection module that compares the user’s self-explanation with the model answer to identify missing information, as illustrated in Figure 7.2. The system uses a summary model to derive the solution steps of the problem (Nakamoto et al., 2024) and extracts missing knowledge elements from the cosine similarity (Nakamoto et al., 2021). It presents the missing knowledge elements to the student and advises them to solve related problems to enhance their learning and understanding of self-explanation. Figures 7.3 show an example of the feedback message that the SEAF system generates for the student after analyzing their self-explanation writing. The message congratulates the student for answering the question correctly and provides positive reinforcement to encourage them to continue their learning journey. The system recommends that the student refers to others’ self-explanation and relevant sections of the textbook to try related problems and develop a habit of reflecting on their mistakes. Additionally, the system highlights areas where the student struggled and offers guidance on how to improve.

Generating Personalized Messages

To generate the personalized messages feature, we use a combination of NLP techniques and pre-defined templates, which are customized based on each learner’s learning logs. We begin by creating a model answer text from the collected data, which is then vectorized using the BERT Japanese pre-trained model which has been shown to be effective in many NLP tasks (Suzuki, 2019). These summarization sentences form the overall summary of the answer texts. BERT(a) represents the distributed representation obtained by applying the BERT algorithm to the text input a . We created a set of vectors $\mathbf{V}$ from a set of student answer texts $A = \{a_1, a_2, \dots, a_n\}$, where n represents the total number of students’ answer texts.

$$\mathbf{V} = \{\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n\}, \mathbf{v}_i = \text{BERT}(a_i) \quad (1) \quad (7.1)$$

We then perform k-means clustering on the vector set $\mathbf{V}$ to obtain k clusters. Each cluster $c_i = \{\mathbf{v}_{i1}, \mathbf{v}_{i2}, \dots, \mathbf{v}_{im_i}\}$ consists of m_i vector representations of answer texts, where m_i represents the total number of answer texts in the i -th cluster, and $1 \leq i \leq k$.

$$\mathbf{C} = \{c_1, c_2, \dots, c_k\}, c_i = \{\mathbf{v}_{i1}, \mathbf{v}_{i2}, \dots, \mathbf{v}_{im_i}\} \quad (2) \quad (7.2)$$

PDFダウンロード 色付け切替 下書き機能をオフにする 自己説明表示オフ

再生 1倍速(通常) 一時停止 再生ストローク0 AI先生の採点を始める ①feedback Button

現在22人がこの問題を回答しました。

AI先生の個別アドバイス

何度か試行錯誤をされていましたが、最終的によく正解できましたね。あなたの筆跡や自己説明から、クラスみんなの回答などから総合して、問題の途中まではよくできていましたが、「問題の後半におけるたすきや式に係る単元（例：角の大きさと弧の長さ、円周角と弧など）」に関する箇所間違いがあったようです。あと一息ですね。他の人の自己説明や教科書の該当箇所を参考に間違える問題にもチャレンジしてみてください。数学はとにかくたくさん問題を解くだけではなく、振り返りを行うことが重要です。どうして間違ったのか、解き方などを振り返り自分はどういう点で賢くなったのかを考える癖をつけましょう。どんな勉強でも最初は成果があまり出ません。しかし、基礎をしっかりと身につけて継続して続けることで、必ず成長することができます。この調子で頑張ってください。

AI先生 いつも皆さんのことを見守ってますよ

皆さんの能力は必ず伸びます！

学力

みんなが思う学力の向上曲線

現実の学力向上イメージ

大きなギャップ

不安

時間軸

成長を実感するまで時間がかかる→

ここに到達するまでが大変

AI先生の分析からあなたに推薦する問題を用意しています。次のボタンを押して推薦問題にチャレンジしてみてください。AI先生による分析結果を確認するには下のタブをクリックしてください。

AI問題推薦ページへJump ③Jump to Recommendation Page

② Individual feedback on self-explanations

Figure 7.2: Personalized messages.

For each cluster c_i , we used the LexRank algorithm (Erkan and Radev, 2004) to extract the most representative sentences from each cluster. 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. The set of summarization sentences $\mathbf{S}$ constitutes the overall summary of the answer texts. LexRank(c) represents the extracted sentences obtained by applying the LexRank algorithm to the cluster c .

$$\mathbf{S} = \{s_1, s_2, \dots, s_k\}, s_i = \text{LexRank}(c_i) \quad (3) \quad (7.3)$$

For step 2, we compare the answer summary with the user’s response, taking into account the stroke data associated with each sentence vector. We use this stroke data to determine which self-explanation moments take longer and where the pen stops, and then add this information to the comparison. We denote the set of model answer texts as $\mathbf{A} = \{a_1, a_2, \dots, a_n\}$, and the user’s response text as u . We first obtain their vector representations using BERT:

$$\mathbf{V}_{\mathbf{A}} = \{\mathbf{v}_{\mathbf{a}1}, \mathbf{v}_{\mathbf{a}2}, \dots, \mathbf{v}_{\mathbf{a}n}\} \quad (4) \quad (7.4)$$

Where $\mathbf{v}_{\mathbf{a}1}$ describes the vector representation of each answer text in the set $\mathbf{A}$, and $\mathbf{v}_{\mathbf{u}} = \text{BERT}(u)$. In addition, $\mathbf{V}_{\mathbf{A}}$ represents the set of vector representations of the model answer texts. We then create a feature vector for each sentence in $\mathbf{A}$ and u by incorporating stroke data $\text{stroke}_{A,i}$, stroke_u and timing information $\text{timing}_{A,i}$, timing_u , where $\text{stroke}_{A,i}$ and stroke_u represent the stroke data associated with the sentence in $\mathbf{A}$ and u , respectively. And $\text{timing}_{A,i}$ and timing_u represent the timing information associated with the i -th sentence in $\mathbf{A}$ and u , respectively.

For step 3, after extracting the self-explanations, stroke, and timing feature vectors, we utilize them to train a random forest classification model. This model is used to predict the overall feature, which indicates whether the user’s response is correct or incorrect.

7.3.4 Self-explanation Score

Self-explanation Score Descriptions

The Self-explanation Scoring feature in SEAF provides learners with an objective way to measure the quality of their self-explanation writing. SEAF evaluates the quality of self-explanation using a set of predefined criteria, based on the self-explanation summary model developed by Nakamoto et al. (2024). The SEAF system also provides a way for users to compare their scores with their classmates, which can be seen in Figure 7.4. In the first section of the feedback message, the system presents the average amount of writing done by classmates and advises users to write more if the number of words or vocabulary used is low. It also informs the user of their Self-explanation Analysis Score and how it compares to the class average. For example, if the user’s score was higher than

the class average, the SEAF message congratulates the user on their well-written self-explanation. Furthermore, the SEAF message also provides suggestions for improvement. For instance, if the user’s self-explanation sentence was shorter than the class average, the message suggests that writing more text could result in a higher score and a better understanding of the unit.

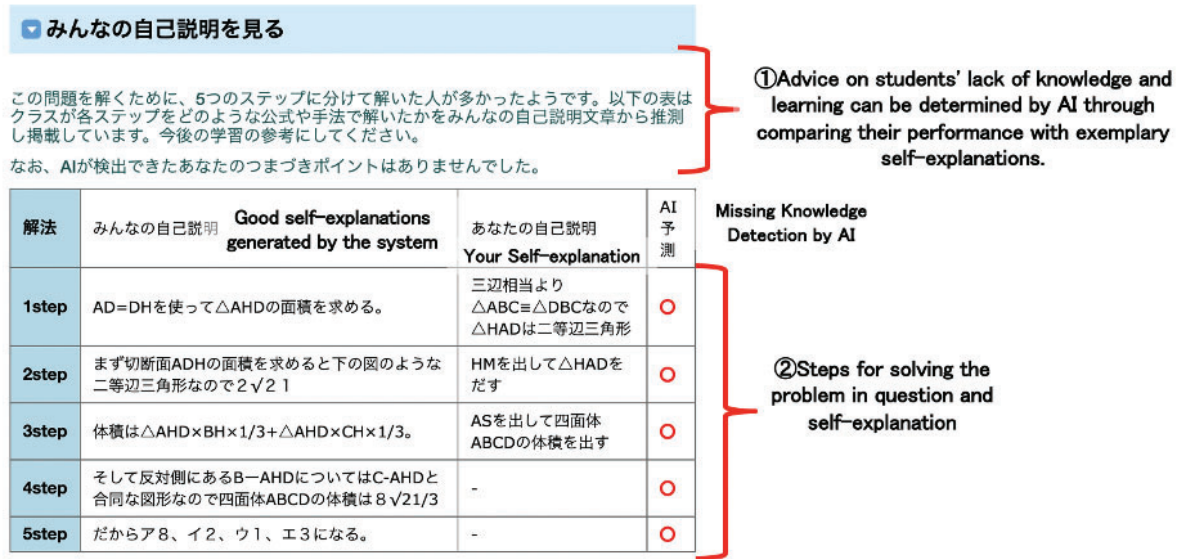


Figure 7.3: Missing knowledge detection.

Analyzing Self-explanation Score

The SEAF system compares the student’s self-explanation with the model answer and assigns a score based on a set of predefined criteria. Let u represent the textual input provided by the user and s represent a summarization sentence. The score is computed using the ensuing formula, wherein erf denotes the error function, and μ and σ are the mean and standard deviation, respectively, of a normal distribution having a mean of 50:

$$\text{score}(u, s) = 50 + 50 \times \text{erf} \left(\frac{\text{sim}(u, s) - \mu}{2\sigma} \right) \quad (5) \quad (7.5)$$

where $\text{sim}(x, y) = \frac{\mathbf{x} \cdot \mathbf{y}}{\|\mathbf{x}\| \|\mathbf{y}\|}$ represents the cosine similarity between vectors $\mathbf{x}$ and $\mathbf{y}$. To provide a more accurate assessment of the quality of the writing, the SEAF system adjusts the score based on the length of the student’s self-explanation. The adjusted score (AS) is calculated using the following formula:

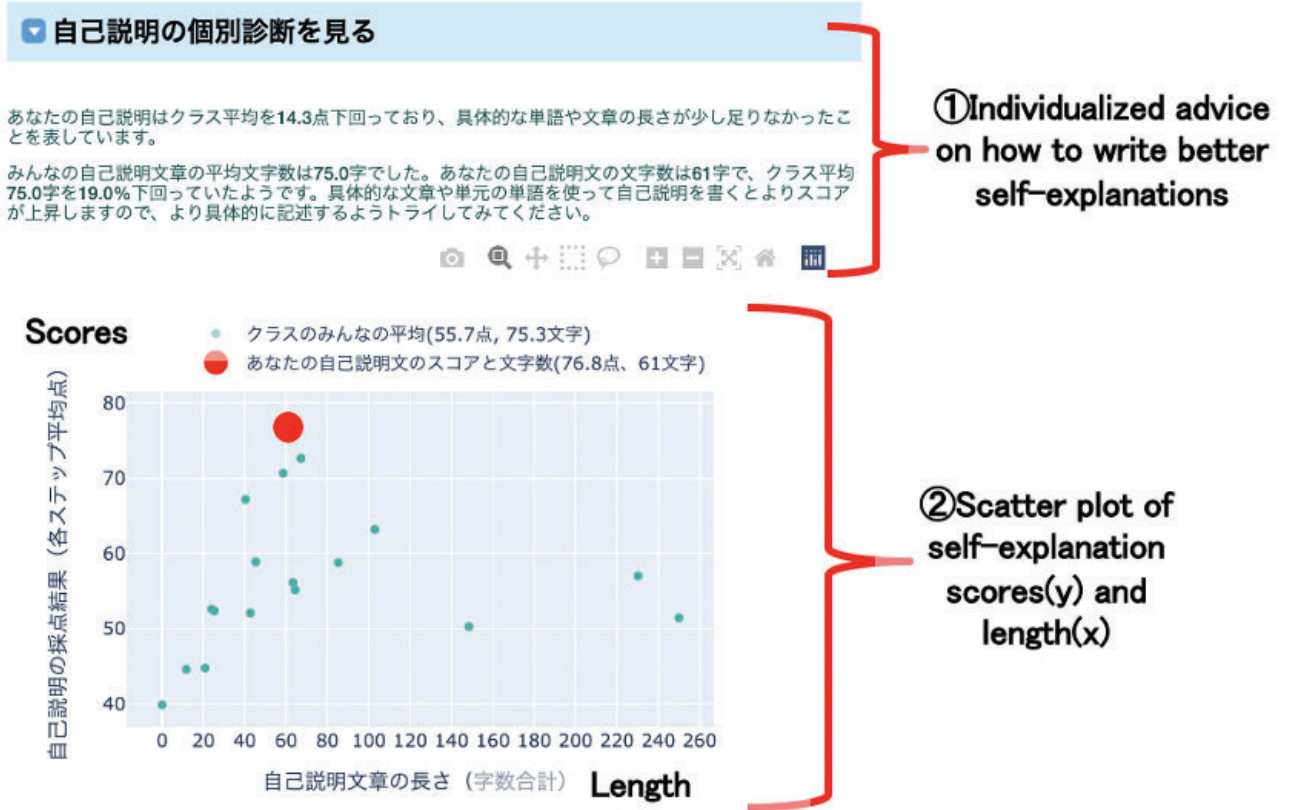


Figure 7.4: The scatter plot for the self-explanation analysis score.

$$AS = S \times \left(1 + \frac{L}{\text{mean length of class self-explanations}} \right) \quad (6) \quad (7.6)$$

where S is the similarity score between the student's self-explanation and the model answer, and L is the length of the self-explanation. This adjustment is made to encourage students to provide more specific explanations by requiring longer self-explanations and because individuals who provide longer self-explanations tend to receive higher scores according to the distributions of collected dataset.

7.3.5 Classmates' Self-explanations Reference

Classmates' Self-explanations Reference Descriptions

The third feature of the SEAF system is the provision of sample self-explanations that are relevant to the learning process. These examples can serve as a reference for learners to write their own self-explanations, and help them understand what makes a good self-explanation. Bisra, Liu, Nesbit, et al. (2018) found that self-explanation was slightly

less effective than reciprocal teaching, feedback, and interval learning (review effect), but implementing both mutual teaching and feedback in self-explanation could enhance learning. By viewing and analyzing other students' high-quality self-explanation writing, learners can gain a better understanding of what constitutes good self-explanation and how to improve their own writing.

Additionally, the SEAF system selects these examples based on their quality and relevance to the self-explanation writing task, which promotes collaboration and social learning. Fyfe and Rittle-Johnson (2016) reported that feedback was effective in promoting learning, particularly when it was immediate. Therefore, this feature not only serves as a reference for learners to follow but also provides immediate feedback that can enhance their learning experience.

This feature provides a visual representation of similar classmates' self-explanations and their corresponding pen strokes, as illustrated in Figure 7.5. Users can access two key pieces of information: (1) similar self-explanation sentences related to the student's responses, and (2) details about their pen strokes. By utilizing this feature, users can review and compare their own self-explanations and pen strokes with those of their classmates, allowing for a better understanding of their performance and potentially identifying areas for improvement.

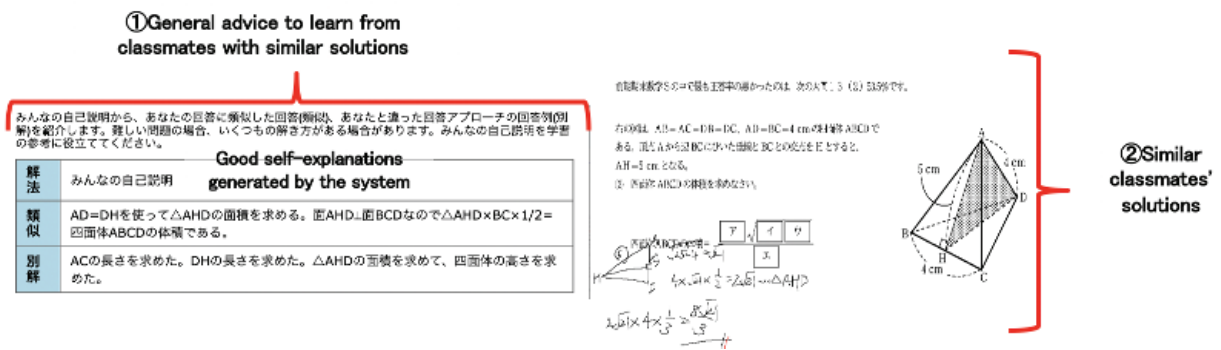


Figure 7.5: Reference examples of classmates' self-explanation.

Extracting Example Sentences

The system uses a cosine similarity algorithm to calculate the similarity score between the student's self-explanation and the model answer. Firstly, as a preprocessing step, build a text regression model that predicts 1 to 5 points, infer with the Text regression model with collected data, and use only data above a threshold value. Secondly, for a given user

input in the form of a Japanese sentence, the sentence is vectorized using BERT to obtain a d -dimensional vector representation. The cosine similarity between this vector and all vectors in each cluster is then calculated. The response closest to the user's sentence is chosen as the sentence corresponding to the cluster with the highest cosine similarity value. Let $\mathbf{S} = \{s_1, s_2, \dots, s_i, \dots, s_n\}$ be the set of all Japanese sentences, where s_i represents the i -th sentence and n is the total number of sentences. Using BERT, each sentence s_i is vectorized into a d -dimensional vector $\mathbf{x}_i$. The set of all vectors is represented as $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_i, \dots, \mathbf{x}_n\}$, where $\mathbf{x}_i$ represents the vector representation of s_i .

The system then uses a clustering algorithm to suggest high-quality self-explanation examples to each student based on their previous writing and learning progress. Using k-means clustering, the vectors in $\mathbf{X}$ are partitioned into k clusters, where k is a predetermined number of clusters. Let the resulting clusters be represented as $\mathbf{C}_1, \mathbf{C}_2, \dots, \mathbf{C}_k$. For a given user's Japanese sentence u , which is also vectorized using BERT to form vector $\mathbf{x}_u$, we calculate the cosine similarity between $\mathbf{x}_u$ and each vector $\mathbf{v}_{ij} \in \mathbf{C}_i$ as follows:

$$\text{sim}(u, \mathbf{C}_i) = \max_{1 \leq j \leq m} \text{sim}(\mathbf{x}_u, \mathbf{v}_{ij}) \quad (7)$$

Finally, the response closest to the user's sentence is chosen as the sentence corresponding to the cluster with the highest cosine similarity value:

$$\text{response}(u) = \text{argmax}_{i=1, 2, \dots, k} \text{sim}(u, \mathbf{C}_i) \quad (8)$$

where argmax represents $\mathbf{C}_i$ with the index i of the maximum value in the set. To ensure the privacy and anonymity of the student users, the system employs data masking techniques.

7.4 Analytical Methods and Framework for Analysis

The framework for analyzing the effectiveness of SEAF in evaluating the quality of writing self-explanations was based on a quasi-experimental design with purposeful sampling. As junior high school is a part of compulsory education in Japan, it is not ethically feasible to conduct a randomized experiment, and instead, participants were selected purposefully. In the experiment, both groups received the treatment at the same conditions, and in the analysis, they were divided based on their usage of the SEAF system. Additionally, we divided participants into a low and a high self-explanation level group to identify those who could benefit the most from the feedback. Lastly, we administered a questionnaire

to the high engagement group at the conclusion of the investigation to gather additional insights.

7.4.1 Experiment Methodology

To comprehensively investigate the influence of SEAF on students' self-explanation writing, the study was structured in three distinct studies: Before Experiment, Preliminary Experiment, and Actual Experiment. This structure allowed for capturing a holistic view of participants' evolution over time, factoring in both their natural state without the aid of SEAF and their progression with consistent SEAF usage. All 50 participants, drawn from the 3rd grade of junior high school, consistently partook in these three studies. Their participation in each study is emphasized to ensure a continuous data trajectory for each student and facilitate a clear understanding of SEAF's impact over time.

Before Experiment: This served as the baseline data acquisition period, where students wrote self-explanations without the intervention of SEAF, thereby capturing their innate self-explanation abilities.

Preliminary Experiment: Here, students were introduced to the SEAF tool for the first time. This study was conducted from November 2021 to March 2022. They were directed to pen their self-explanations, checking their pen strokes, and upon completion, press a button to garner feedback on their self-explanations. This study allowed for initial observations on SEAF's immediate effects on self-explanation quality.

Actual Experiment: Extending from April 2022 to June 2022, this study saw students' continued use of SEAF, reinforcing the habits developed during the Preliminary Experiment and providing insights into the tool's longer-term impact on students' self-explanation skills. During this study, the 49 questions were administered on a weekly basis over the course of three months to gauge the students' understanding and evolution in self-explanation abilities.

The frequency of feedback content usage during the Actual Experiment, along with the mean frequency of writings self-explanations, acted as pivotal metrics in gauging student engagement and their dedication to the self-explanation tasks assigned weekly. By organizing our study in these consecutive studies, we aim to capture not only the immediate effects of SEAF on self-explanation but also its sustained impact over an extended period. This structure, coupled with the consistent participation of all 50 students across studies, offers robust insights into SEAF's potential as a transformative educational tool.

7.4.2 Analysis Preparations

Evaluation of Self-explanations Based on Three Principal Criteria

In our study, we assessed self-explanations with rigor, focusing on three central tenets: coherence, clarity, and relevance. The 'coherence' evaluates the logical sequence of the explanation. 'Clarity' measures its comprehensibility, and 'relevance' ensures the capture of all vital knowledge constructs and procedural details. For a consistent evaluation, we adapted the rubric and scoring methodologies suggested as suggest in study 1. Thompson and Senk (1998) have endorsed these rubrics, making them suitable for tasks with multiple solution strategies. We crafted task-specific rubrics to ensure their precision and applicability. Methodology for Self-explanations Evaluation are following, and the result is shown in Table 4:

Scoring System: Two independent raters employed rubrics developed for this study to assess the self-explanations. A scale ranging from 1 to 5 was utilized for scoring.

Rater Consistency: Quadratic Weighted Cohen's kappa coefficient (Cohen, 1960) was employed to determine the level of agreement between the raters. A coefficient value of 0.749 indicates substantial agreement.

Data Processing: To maintain consistency in the evaluation, the mean score from both raters was used for further analysis.

Categorization of Scores: For a clear distribution, the self-explanation scores were categorized, aiming for uniformity. Participants were divided into two groups based on average self-explanation quality scores:

Participant Categorization Based on Engagement

To examine the influence of SEAF usage on self-explanation quality, we divided participants into low and high engagement groups based on feedback utilization frequency and the depth of self-explanation. Given educational constraints, a randomized experiment wasn't viable. Instead, we chose purposeful sampling to establish experimental groups. Engagement is pivotal to cultivating conceptual understanding, as it bridges motivation and learning (Blumenfeld et al., 2005; Sinha et al., 2015). This strategy, therefore, can be an alternative to randomized experiments when exploring the ramifications of computer-aided engagement on learning outcomes.

Out of 50 participants in the study, the provided SEAF usage metrics (mean, standard

deviation, minimum, and maximum) aggregate the data from both "Preliminary" and "Actual" experiment studies. Notably, participants recorded an average SEAF usage of 14.23, a standard deviation of 13.03, and a usage range between 1 and 52 times across the two studies. Participants were bifurcated based on SEAF usage into two categories: the high-engagement group, who accessed the feedback button at least four times during the study, and the low-engagement group, who did so less often. This division was driven by the imperative for a substantial statistical analysis.

SEAF Usage Metrics Analysis The study analyzed SEAF usage metrics to discern patterns and trends in student engagement. This analysis included:

- **Average Usage:** The mean number of times students accessed the SEAF system.
- **Engagement Range:** The minimum and maximum usage frequencies, providing insights into the range of student interaction with the system.
- **Standard Deviation:** To understand the variability in SEAF usage among students.
- **Group Classification:** Participants were classified into high and low engagement groups based on their interaction frequency with SEAF. This classification helped in drawing comparative analyses between different levels of user engagement.

7.5 Results of SEAF Learning Effects Experiments

7.5.1 Results of Statistical Analysis Based on Self-Explanation Level

The goal of this statistical analysis was to discern significant differences in self-explanation quality between different engagement groups during the actual experiment period. We aimed to ascertain if there was a relationship between engagement level and self-explanation quality, and if SEAF’s presence made a tangible difference within these groups. Table 7.2 presents the descriptive statistics concerning self-explanation characteristics, segmented by group and level. The ‘N’ in the table denotes the total count of self-explanation descriptions for each respective group.

Table 7.2: Descriptive statistics of self-explanation features by group and level.

Self-Explanation Level	Low-Engagement Group (N = 18)				High-Engagement Group N = 28				Welch’s <i>t</i> -Test	
	N	Self-Explanation Answers	M	SD	N	Self-Explanation Answers	M	SD	DF	t
High	9	29	3.034	0.973	16	198	3.76	1.097	38.9	3.631 ***
Low	9	69	2.05	1.05	12	131	2.62	1.11	144.6	3.520 ***

Note: *** $p < 0.01$.

7.5.2 Repeated Measures ANOVA

Since the experiment was repeated in this study, a repeated measures analysis was conducted to determine whether the SEAF was effective for each experimental treatment group. The subjects for the study were selected from those who had responded to the three-time period group, and Table 7.3 shows the number of subjects and their corresponding scores. N represents the number of unique users within the group. To test the null hypothesis that the means of the groups were equal during the experimental period, a non-parametric (rank-based) one-way repeated measures ANOVA, specifically the Friedman test, was performed. The results of the Friedman test with the chi-square test showed that the high self-explanation level group did not differ significantly from the students’ self-explanation Quality based on the experimental period, $\chi^2(2, N = 25) = 0.375, p = 0.829$. However, it was found that the low self-explanation level group was significantly different, $\chi^2(2, N = 21) = 8.77, p = 0.012$.

Table 7.3: The Friedman test results with repeated treatments at different experimental periods

Self-Explanation Level	Engagement	M(SD)				Statistics		
		N	T1	T2	T3	DF	W	χ^2
High	High	16	3.79 (0.51)	3.74 (0.60)	3.66 (0.73)	2	0.0152	0.375
Low	High	12	2.50 (0.73)	2.93 (0.82)	2.64 (0.45)	2	0.337	8.77 **

Note. T1: Before Experiment, T2: Preliminary Experiments, T3: Actual Experiment. ** $p < 0.05$.

Table 7.4: The result of Conover's post hoc test (P-Values)

Experimental Period	T1	T2	T3
T1	-	0.021	0.251
T2	-	-	0.135

Note. T1: Before Experiment, T2: Preliminary Experiments, T3: Actual Experiment.

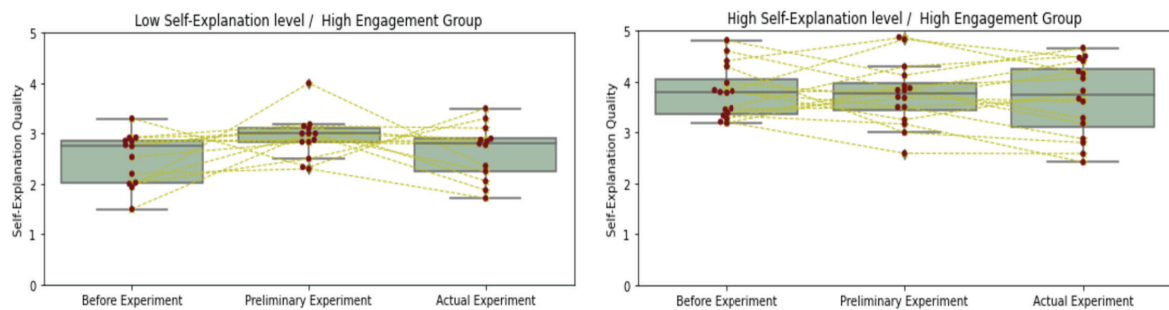
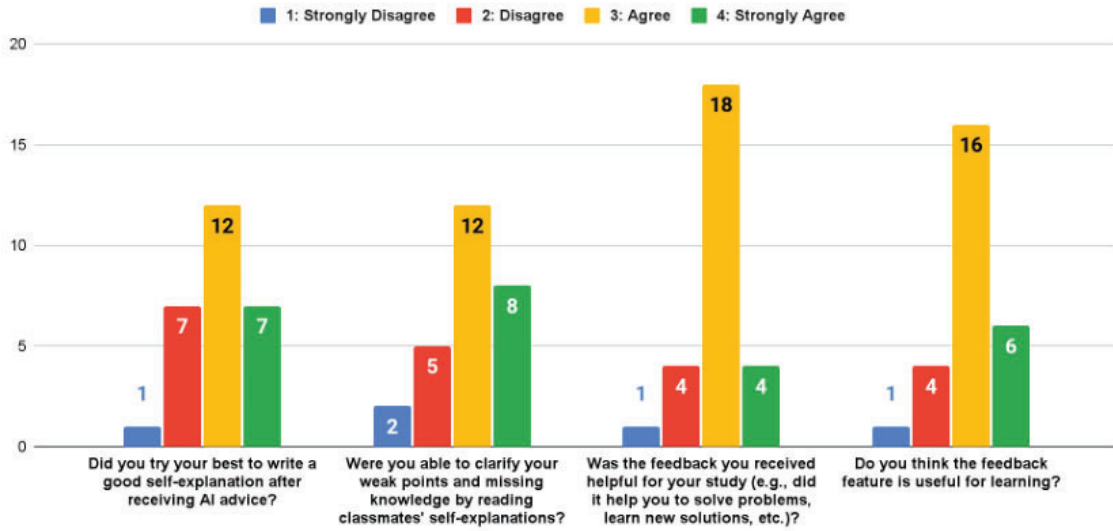


Figure 7.6: Distribution of self-explanation Quality means in each experimental period group.

Table 7.5: Results of the SEAF usage survey

No.	Question	Self-Explanation Level	N	M(SD)
1	Did you try your best to write a good self-explanation after receiving AI advice?	Low	12	2.92 (0.76)
		High	15	2.93 (0.85)
2	Were you able to clarify your weak points and missing knowledge by reading classmates' self-explanations?	Low	12	2.83 (0.90)
		High	15	3.07 (0.85)
3	Was the feedback you received helpful for your study (e.g., did it help you to solve problems, learn new solutions, etc.)?	Low	12	3.00 (0.41)
		High	15	2.87 (0.81)
4	Do you think the feedback feature is useful for learning?	Low	12	3.08 (0.49)
		High	15	2.93 (0.85)

Figure 7.7: Results Plots of the SEAF usage survey



7.5.3 Perception Analysis

A survey was conducted after the actual experiments to test user satisfaction and the learning effect of the SEAF. The decision to use a four-point scale was deliberate, considering the Japanese cultural tendency towards central tendency and previous research on scale accuracy and reliability. The survey responses predominantly came from the high engagement group, suggesting a connection between active SEAF interaction and the willingness to provide post-experiment insights. Figure 7.7 shows how learners rated the SEAF in terms of its interface, usefulness for learning, and future use.

7.6 Discussion

7.6.1 RQ1: Can a classmate’s self-explanations be presented and useful to the learner as feedback?

In line with Research Question 1 (RQ1), the study explored the utility of classmates’ self-explanations as feedback within the SEAF system. Grounded in social cognitive theory, SEAF aims to create a sustainable learning environment through the use of reciprocal teaching and feedback (Bisra, Liu, Nesbit, et al., 2018; Hall and Vance, 2010). Our analysis particularly focused on how this feedback model impacted learners with varying levels of self-explanation skills.

The results demonstrated a significant impact for the low self-explanation group, validating that peer-generated self-explanations can be a potent form of feedback for this demographic. Notably, Conover’s post hoc test showed significant improvements for these learners between the Before Experiment and Preliminary Experiment periods. This suggests that the SEAF system is both impactful and valuable for enhancing the quality of self-explanations among students who initially struggled.

However, the high self-explanation group did not show similar significant improvements, indicating that the system may have limitations in its applicability across all skill levels. This highlights the need for the careful curation of peer-generated self-explanations to avoid misleading or confusing learners, emphasizing that quality over quantity is crucial when selecting reference self-explanations. To sum up, the study supports SEAF’s hypothesis that classmates’ well-crafted self-explanations can serve as effective feedback, particularly for learners who are less adept at self-explanation. Future iterations of the SEAF system should focus on fine-tuning the reference selection process to make this form of peer feedback universally effective.

7.6.2 RQ2: Which students benefit most from using the SEAF system?

The results showed that the SEAF was more effective for students with low levels of self-explanation, suggesting that learners with low self-explanation skills may need more support than learners with high self-explanation skills. One possible explanation for this result is that the SEAF helped these students because they may have had difficulty writing self-explanations without prior strategies or training. This finding is consistent

with previous research suggesting that learners with low prior knowledge may need more guidance and support than learners with high prior knowledge (Arner et al., 2021).

Furthermore, many respondents indicated that the Low self-explanation level group was more useful the perceptual analysis, suggesting that the SEAF may have been a useful learning tool for learners who could not self-explain. Conversely, the question No. 2 suggests that self-explanation is also an effective strategy for students who already possess high self-explanation skills. Students who are accustomed to problem-solving through self-explanation may compare their self-explanations with others and effectively articulate the benefits they derive from the SEAF system.

Additionally, the results of the repeated measures analysis demonstrated a significant difference in self-explanation quality between the Before Experiment period and the Preliminary Experiment period for the low self-explanation group. The study suggests that the initial experience with the SEAF system may have a significant impact on its effectiveness. Participants exhibited positive responses and improvements in their self-explanation quality during the preliminary experiment, likely due to the novelty of the system. However, over time, as users became more familiar with the system, the impact diminished in the actual experiment. The Preliminary and Actual Experiments were conducted over a period of 4 months and 3 months, respectively. To enhance the system's effectiveness, it may be necessary to determine an appropriate duration for the experiment and introduce new stimuli when participants become accustomed to the intervention.

Finally, since this was Purposeful Sampling, more detailed experiments are needed even if the results are significant. In particular, given previous research (Blumenfeld et al., 2005) that has linked Engagement to learning, it is possible that students who are more motivated to learn and have a more positive attitude actually embraced the SEAF technology and used it more frequently, resulting in higher quality self-explanations. Further research is needed on the mechanism of frequent use by all, and on the display and messaging methods of Feedback according to learning motivation and personality.

7.6.3 Limitations

The implementation and evaluation of the SEAF system revealed several limitations that must be addressed to enhance its effectiveness and applicability in diverse educational contexts. While SEAF demonstrates significant potential in improving self-explanation skills and providing personalized feedback, key challenges emerged during its evaluation.

First, the system struggled to sustain feedback quality over time, leading to diminished engagement and reduced impact in later stages. This decline suggests that the feedback mechanisms lacked the adaptability needed to evolve alongside learners' progress and changing needs. Without dynamic updates to feedback strategies, the system's effectiveness may wane, particularly in long-term use.

Second, variability in the quality of self-explanations provided by learners posed a significant challenge. Low-quality or inconsistent inputs limited the system's ability to generate precise and meaningful feedback. Addressing this issue requires strategies such as scaffolding, guided prompts, or incentive systems to encourage learners to produce more thoughtful and detailed self-explanations.

Third, the evaluation of SEAF was constrained by the sample size of 50 participants, which, while sufficient for initial insights, restricts the generalizability of the findings. A larger and more diverse sample would provide a more robust assessment of SEAF's effectiveness across varied learner demographics and educational settings.

Finally, disparities in learner engagement revealed challenges in ensuring equitable participation. Most meaningful feedback originated from highly engaged learners, while less-engaged participants contributed minimally, highlighting the intricate relationship between motivation and effective self-explanation. Future efforts must explore personalized interventions and alternative feedback mechanisms to involve less-engaged learners and ensure equitable benefits across all student groups.

7.6.4 Main Findings

Despite these limitations, this study demonstrates the significant potential of the SEAF system to transform self-explanation practices in mathematics education. By providing immediate, personalized feedback, SEAF helps learners identify gaps in their understanding and refine their reasoning processes, contributing to deeper learning and more effective problem-solving.

A key strength of SEAF lies in its ability to balance personalization with scalability. The system adapts feedback to individual needs, illustrating the promise of adaptive learning technologies in educational environments. However, the decline in feedback quality over time highlights the need for mechanisms that sustain engagement and ensure relevance throughout prolonged use. Additionally, its reliance on high-quality learner inputs and reduced effectiveness for less-engaged students suggest areas requiring further

refinement. These findings suggest that future iterations of SEAF should integrate mechanisms to dynamically adjust feedback, maintain long-term engagement, and address the diverse needs of learners.

7.6.5 Future Directions

To enhance the SEAF system’s long-term effectiveness and applicability, several future directions are proposed. First, feedback mechanisms must be developed to dynamically adjust to learners’ progress and task complexity. Sustaining feedback quality over time is critical to maintaining engagement and relevance. Incorporating adaptive features, such as gamified elements or scaffolded interventions, could further motivate learners, particularly those who are less engaged.

Second, improving the quality of learner inputs is essential for maximizing SEAF’s potential. Guided prompts, scaffolding, and incentive mechanisms, such as rewards or peer-based collaboration, can encourage learners to produce higher-quality self-explanations. Additionally, fine-tuning feedback algorithms to align more closely with domain-specific challenges and cognitive demands will enhance the system’s overall precision and effectiveness.

Third, the sample size and diversity should be expanded in future studies to validate the system’s scalability and generalizability across different educational contexts. Rigorous evaluations, including randomized control trials (RCTs) and longitudinal studies, will be critical for assessing SEAF’s long-term impact and refining its adaptive features. These studies should explore how SEAF performs across various learning environments and levels of complexity, ensuring that the system remains relevant and effective for a broad range of learners.

Finally, SEAF’s design must address the issue of engagement over prolonged use. Feedback systems should evolve with learners’ needs and task difficulties, ensuring they remain targeted and effective. Ensuring that learners stay motivated to engage with the feedback system is essential for long-term success, particularly for complex and extended learning processes.

7.6.6 Conclusion

The SEAF system offers a promising approach to improving self-explanation skills and enhancing learning outcomes in mathematics education. By identifying gaps in understanding and providing personalized, real-time feedback, SEAF has demonstrated its potential to foster deeper engagement and more effective problem-solving. However, its success depends on addressing limitations related to adaptability, long-term feedback quality, and learner input variability.

Through iterative refinement and the integration of adaptive, context-aware mechanisms, SEAF can evolve into a robust and transformative educational tool. Its ability to balance personalization with scalability positions it as a valuable resource for diverse learners, paving the way for innovative applications in educational technology. Future research will focus on overcoming current challenges to ensure that SEAF meets the evolving needs of students and educators in varied educational settings.

Chapter 8

Conclusion and Future work

To conclude, we first present the main findings and contribution of this research and end by proposing directions for future research.

8.1 Main Findings of Each Study

This doctoral thesis presents a novel approach to enhancing mathematics education by integrating self-explanation methodologies with pen-stroke data and computational tools. The research focuses on building adaptive and personalized learning systems through impasse detection, automated model answer generation, data augmentation, and self-explanation-based automated feedback. These studies collectively highlight the potential of self-explanation to transform traditional educational tools while addressing key challenges in their design and implementation.

8.1.1 Study 1: Impasse Detection

The first study developed a methodology for detecting learning impasses by comparing learner-generated self-explanations with system-generated model answers, supplemented by pen-stroke data analysis. This approach allowed the system to identify specific points where learners struggled, providing a foundation for personalized feedback and targeted interventions. However, the reliance on learner-generated self-explanations alone revealed limitations due to the variability and open-ended nature of these responses. Unlike structured and standardized textbook model answers, self-explanations often lacked precision and consistency, making accurate impasse detection more challenging. Integrating textbook solutions as an additional reference could enhance the system's accuracy and scala-

bility across diverse learning contexts.

Future research should focus on hybrid approaches that combine learner-generated self-explanations with curated resources, such as textbook solutions, to maximize the strengths of both. Expanding datasets to include diverse sources, such as teacher-created materials or peer-reviewed examples, could further improve the robustness and adaptability of the system.

8.1.2 Study 2: Automated Model Answer Generation

The second study shifted its focus to automating the generation of self-explanation model answers, reducing the reliance on manually curated solutions. By employing an unsupervised learning approach, the system was able to generate accurate answers that preserved essential knowledge components in mathematical solutions. This represents a significant step toward scalability and efficiency in self-explanation-based learning systems.

Despite these advancements, a critical limitation emerged in aligning *Knowledge Components* with the *Rubric Step Order*. Only 64% of the generated answers achieved accurate alignment, reducing their clarity and pedagogical value. Misalignment between components and logical sequences can confuse learners and hinder the effectiveness of automated explanations. Addressing this issue requires the development of more sophisticated reasoning algorithms capable of capturing the causal and procedural relationships within mathematical problem-solving. Additionally, integrating structured hints or rubrics could serve as scaffolding to improve the coherence of generated answers.

8.1.3 Study 3: Data Augmentation with Math Texts and Large Language Models

The third study explored data augmentation techniques by combining structured mathematical texts with LLMs to improve the quality of generated self-explanations. Mathematical texts provided a reliable and consistent source of structured knowledge, while LLMs contributed context-aware and nuanced explanations to enrich the dataset.

The results revealed that while the mixed dataset (math texts + LLMs) achieved strong validation performance, test accuracy was higher when using only mathematical texts. This discrepancy suggests that LLM-generated data, introduced noise and inconsistencies that negatively impacted generalizability. While LLMs offer the potential to enrich datasets with diverse perspectives, their outputs require careful curation and

post-processing to align with educational standards. Refining LLM-generated content and employing selective data augmentation strategies could improve the effectiveness of combining these resources.

8.1.4 Study 4: Self-Explanation-Based Automated Feedback (SEAF)

The fourth study introduced the Self-Explanation-Based Automated Feedback (SEAF) system, a tool designed to deliver real-time, personalized guidance based on learners' self-explanations. Building on insights from previous studies, SEAF aimed to enhance self-explanation skills and motivation by providing immediate feedback tailored to individual needs. The system demonstrated its ability to engage learners, address misunderstandings in real-time, and improve comprehension of complex problems.

However, SEAF faced challenges in maintaining its effectiveness over extended periods. Feedback quality diminished as the system struggled to adapt to learners' evolving needs, leading to reduced engagement and a decline in long-term impact. Additionally, the system did not fully address the diverse cognitive and motivational demands of different learners, limiting its generalizability. These limitations were compounded by the absence of rigorous evaluations, such as randomized control trials (RCTs), which restricted the ability to comprehensively assess SEAF's strengths and weaknesses.

Future research should focus on developing adaptive feedback mechanisms that dynamically adjust to learners' progress and changing needs. Incorporating gamified elements or scaffolded feedback could help sustain motivation and engagement over extended periods. Furthermore, conducting RCTs and longitudinal studies will be essential for evaluating the system's long-term efficacy and applicability across diverse educational contexts.

8.2 Overall Discussion and Limitations of the Study

This doctoral thesis investigates the integration of self-explanation methodologies into adaptive learning systems, offering significant advancements to the field of educational technology while identifying key areas for refinement. By synthesizing the findings across four interrelated studies, the research evaluates the effectiveness of self-explanation-based approaches, highlights their limitations, and outlines strategies to address these challenges.

Self-explanation fosters critical reflection, enabling learners to articulate their thought processes, identify gaps in understanding, and deepen their engagement with learning ma-

terials. These benefits were evident across the studies, which collectively demonstrated the potential of self-explanation systems to detect areas of difficulty, generate relevant model answers, and provide personalized feedback. For example, the impasse detection system (Study 1) effectively identified learning struggles by comparing self-explanations with model answers and analyzing pen-stroke data. This capability enabled the delivery of targeted and personalized interventions. Similarly, the automated generation of model answers (Study 2) preserved essential knowledge components and achieved commendable accuracy, demonstrating the feasibility of scaling self-explanation-based systems for broader educational applications.

Despite these successes, several limitations were identified. The open-ended nature of self-explanations, while promoting creativity and critical thinking, introduces variability and inconsistency that challenge automated systems. For instance, Study 2 revealed that only 64% of generated answers aligned with rubric step order, indicating difficulties in maintaining logical coherence and pedagogical clarity. In Study 3, the integration of mathematical texts and LLMs showcased the benefits of combining structured knowledge with advanced language models. However, inconsistencies in LLM-generated data, particularly without fine-tuning, compromised test accuracy. Study 4 further highlighted the need for adaptive feedback mechanisms, as SEAF’s effectiveness diminished over time due to its inability to evolve alongside learners’ changing needs and engagement levels.

To address these limitations, this thesis proposes several strategies. First, scaffolding mechanisms, such as guided prompts or structured templates, can improve the quality and consistency of learner-generated self-explanations, enabling systems to process them more effectively. Second, incorporating curated external resources, such as textbook solutions or domain-specific datasets, can enhance the reliability and scalability of these systems, particularly in structured educational domains. Third, adaptive feedback mechanisms are crucial to sustain engagement and relevance over extended periods. Incorporating gamified or scaffolded approaches can help maintain learners’ motivation and ensure that feedback evolves with their progress and needs.

Future research should prioritize hybrid approaches that balance the flexibility of self-explanations with structured supports. Integrating causal reasoning algorithms, fine-tuned LLMs, and dynamic feedback loops will be critical for addressing current limitations and improving system robustness. Additionally, conducting randomized control trials and longitudinal studies will be essential to validate the generalizability and long-term impact

of self-explanation systems across diverse educational contexts.

By addressing these challenges, future iterations of self-explanation-based adaptive systems can offer more robust, scalable, and inclusive solutions. These systems have the potential to transform educational practices by providing deeper insights into learning processes, fostering critical thinking, and delivering personalized support tailored to individual learner needs.

8.3 Implications and Future Work

This research offers valuable contributions to educational practice by enhancing adaptive learning systems and improving classroom efficiency. A key application lies in developing advanced adaptive learning environments. By detecting learning impasses and generating personalized feedback, educational platforms can adjust instructional content in real-time, ensuring students follow tailored learning pathways. For example, in mathematics education, the system could dynamically modify problem sets or instructional materials based on a student's demonstrated understanding or struggles, fostering a more individualized and effective learning experience.

Self-explanation itself represents a rich learning process, offering insights into how students think and where they encounter difficulties. This study demonstrated that by focusing on self-explanations alone, it is possible to capture students' learning processes and detect impasses in mathematics learning, albeit with limited accuracy. While this bottom-up approach provides a foundation for understanding learners' reasoning, achieving higher accuracy requires leveraging more diverse and comprehensive datasets. Expanding beyond self-explanations to include additional datasets, such as curated model answers or other structured resources, can significantly enhance detection accuracy. However, for the highly structured domain of education, where reliability and precision are paramount, integrating control models or external datasets is essential to ensure the system meets the accuracy standards required for practical use.

Additionally, this research has significant implications for reducing teacher workload. The system's capacity for automatic scoring and feedback generation allows teachers to shift focus from repetitive tasks to interactive and personalized instruction. For instance, the system could automatically assess lab reports or problem-solving exercises, providing immediate, detailed feedback, enabling teachers to concentrate on mentoring and address-

ing individual student needs.

Moreover, these methodologies can support specialized educational programs, such as those for students with learning disabilities. Adaptive feedback systems are particularly beneficial in special education contexts, where tailored interventions are crucial. For example, a reading comprehension program for students with dyslexia could leverage this system to offer immediate, context-aware corrective feedback, promoting sustained engagement and learning.

Future work will focus on integrating Generative AI and Retrieval-Augmented Generation (RAG) technologies to develop domain-specific LLMs for educational contexts. These advancements will enable systems to provide even more precise, personalized support by dynamically adjusting to evolving learner needs. However, challenges such as dependency on external APIs and data privacy concerns must be addressed. Strategies such as developing in-house AI models tailored to educational needs will mitigate risks, ensuring the system’s reliability and security.

In summary, this research underscores the importance of balancing bottom-up approaches that explore diverse datasets with the need for highly accurate and structured models to meet the rigorous demands of educational contexts. By refining these systems and expanding their datasets and algorithms, future research can develop more robust, reliable, and scalable educational technologies that cater to diverse learners and instructional needs.

8.4 Conclusion

This doctoral thesis makes significant contributions to the field of educational technology by integrating self-explanation methodologies into adaptive learning systems. The research addresses the intricate task of detecting learning impasses and providing personalized feedback, fostering deeper engagement and improved learning outcomes. The systems developed in this thesis offer innovative solutions to long-standing challenges in mathematics education and beyond.

Despite its limitations, this research lays a strong foundation for further exploration and refinement. By addressing challenges such as variability in self-explanations, long-term feedback sustainability, and data inconsistencies, future systems can become more robust, scalable, and inclusive. This work paves the way for continued innovation, en-

suring that educational technologies meet the diverse and evolving needs of learners and educators in various contexts.

References

- Andonian, A., Anthony, Q., Biderman, S., Black, S., Gali, P., Gao, L., Hallahan, E., Levy-Kramer, J., Leahy, C., Nestler, L., et al. (2021). *Gpt-neox: Large scale autoregressive language modeling in pytorch (version 0.0.1)*. <https://zenodo.org/record/7714278>
- Antulov-Fantulin, N., Bosnjak, M., Zlatic, V., Grcar, M., & Šmuc, T. (2012). Synthetic sequence generator for recommender systems-memory biased random walk on a sequence multilayer network. *Proceedings of the IFIP Working Conference on Database Semantics*.
- Arne, T., McCarthy, K., & McNamara, D. (2021). Start stair stepper—using comprehension strategy training to game the test. *Computers*, 10(4), 48. <https://doi.org/10.3390/computers10040048>
- 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>
- Assessment's. (2023). <https://new.assistments.org/>
- Bandura, A., & of Mental Health, N. I. (1986). *Social foundations of thought and action: A social cognitive theory*. Prentice-Hall, Inc.
- Berg, A., Mol, S., Kismihók, G., & Sclater, N. (2016). The role of a reference synthetic data generator within the field of learning analytics. *Journal of Learning Analytics*, 3, 107–128.
- 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.
- Berthold, K., Eysink, T., & Renkl, A. (2009). Assisting self-explanation prompts are more effective than open prompts when learning with multiple representations. *Instr. Sci.*, 37, 345–363.
- Berthold, K., & Renkl, A. (2009). Instructional aids to support a conceptual understanding of multiple representations. *J. Educ. Psychol.*, 101, 70–87.
- 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.
- Bisra, K., Liu, Q., Nesbit, J., Salimi, F., & Winne, P. (2018). Inducing self-explanation: A meta-analysis. *Educ. Psychol. Rev.*, 30, 703–725.
- Blumenfeld, P., Kempler, T., & Krajcik, J. (2005). *The cambridge handbook of the learning sciences: Motivation and cognitive engagement in learning environments*.

- Boonthum, C., Levinstein, I. B., & McNamara, D. S. (2007). Evaluating self-explanations in start: Word matching, latent semantic analysis, and topic models. https://doi.org/10.1007/978-1-84628-754-1_6
- Brown, J. S., & VanLehn, K. (1980). Repair theory: A generative theory of bugs in procedural skills. *Cognitive Science*, 4, 379–426.
- Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020). Language models are few-shot learners [abs/2005.14165]. *arXiv*.
- Cascante-Bonilla, P., Tan, F., Qi, Y., & Ordonez, V. (2020). Curriculum labeling: Revisiting pseudo-labeling for semi-supervised learning. *AAAI Conference on Artificial Intelligence*, 7–12.
- 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>
- Chawla, N., Bowyer, K., Hall, L., & Kegelmeyer, W. (2002). Smote: Synthetic minority over-sampling technique [abs/1106.1813]. *arXiv*.
- Chen, X., et al. (2021). Predictive analysis in educational settings: Improving performance through learning logs. *Educational Data Science Journal*, 7(2), 123–134.
- Chi, M., Leeuw, N., Chiu, M., & Lavancher, C. (1994). Eliciting self-explanations improves understanding. *Cognitive Science*, 18(3), 439–477.
- Chi, M. T. H. (2000). Self-explaining: The dual processes of generating inference and repairing mental models. 5, 161–238.
- Chi, M. T. H., Bassok, M., Lewis, M. W., Reimann, P., & Glaser, R. (1989). Self-explanations: How students study and use examples in learning to solve problems. *Cognitive Science*, 13(2), 145–182.
- Chi, M., Leeuw, N., Chiu, M., & LaVancher, C. (1994). Eliciting self-explanations improves understanding. *Cogn. Sci.*, 18, 439–477.
- Chi, M. (2000). Self-explaining: The dual processes of generating inference and repairing mental models. In R. Glaser (Ed.), *Advances in instructional psychology: Educational design and cognitive science* (pp. 161–238). Erlbaum.
- Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20, 37–46. <https://doi.org/10.1177/001316446002000104>
- Colglazier, W. (2015). Sustainable development agenda: 2030. *Science*, 349, 1048–1050.
- Conati, C., & VanLehn, K. (2000). Toward computer-based support of meta-cognitive skills: A computational framework to coach self-explanation. *Int. J. Artif. Intell. Educ.*, 11, 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.
- Crossley, S., Kim, M., Allen, L., & McNamara, D. (2019). Automated summarization evaluation (ase) using natural language processing tools. *International Conference on Artificial Intelligence in Education*, 25–29.
- CyberAgent. (2023). *Open-calm-7b*. <https://huggingface.co/cyberagent/open-calm-7b>

- Dahmen, J., & Cook, D. (2019). Synsys: A synthetic data generation system for healthcare applications. *Sensors*, 19, 1181.
- Dai, H., Liu, Z., Liao, W., Huang, X., Cao, Y., Wu, Z., Zhao, L., Xu, S., Liu, W., Liu, N., et al. (2023). Auggpt: Leveraging chatgpt for text data augmentation. *arXiv preprint arXiv:2302.13007*.
- Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.
- Dodeen, H. (2015). Teaching test-taking strategies: Importance and techniques. *Psychology Research*, 5(2), 108–113.
- 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*. <https://www.cs.columbia.edu/~idrori/drori2021math.pdf>
- El Emam, K. (2020). Seven ways to evaluate the utility of synthetic data. *IEEE Security & Privacy*, 18, 56–59.
- Erkan, G., & Radev, D. (2004). Lexrank: Graph-based lexical centrality as salience in text summarization. *Journal of Artificial Intelligence Research*, 22(1), 457–479.
- Ferguson, R. (2012). Learning analytics: Drivers, developments and challenges. *International Journal of Technology Enhanced Learning*, 4(5/6), 304–317. <https://doi.org/10.1504/IJTEL.2012.051816>
- Flanagan, B., Majumdar, R., & Ogata, H. (2022). Fine grain synthetic educational data: Challenges and limitations of collaborative learning analytics. *IEEE Access*, 10, 26230–26241.
- 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>
- Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5), 378–382. <https://doi.org/10.1037/h0031619>
- Funayama, H., Asazuma, Y., Matsubayashi, Y., Mizumoto, T., & Inui, K. (2023). Reducing the cost: Cross-prompt pre-finetuning for short answer scoring. *International Conference on Artificial Intelligence in Education*, 3–7.
- Fyfe, E., & Rittle-Johnson, B. (2016). The benefits of computer-generated feedback for mathematics problem solving. *Journal of experimental child psychology*, 147, 140–151.
- Hall, S., & Vance, E. (2010). Improving self-efficacy in statistics: Role of self-explanation & feedback. *Journal of Statistics Education*, 18.
- Han, H., Wang, W., & Mao, B. (2005). Borderline-smote: A new over-sampling method in imbalanced data sets learning. *Proceedings of the International Conference on Intelligent Computing*, 23–26.
- Hattie, J. (2009). *Visible learning: A synthesis of 800+ meta-analyses on achievement*. Routledge.
- Hattie, J., Biggs, J., & Purdie, N. (1996). Effects of learning skills interventions on student learning: A meta-analysis. *Review of educational research*, 66(2), 99–136.

- He, H., Bai, Y., Garcia, E., & Li, S. (2008). Adasyn: Adaptive synthetic sampling approach for imbalanced learning. *Proceedings of the 2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*, 1322–1328.
- Heffernan, N. T., & Heffernan, C. (2014). The assessment’s ecosystem: Building a platform that brings scientists and teachers together for minimally invasive research on human learning and teaching. *International Journal of Artificial Intelligence in Education*, 24, 470–497.
- 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/jresmetheduc.45.1.0062>
- Hong, E., Sas, M., & Sas, J. C. (2006). Test-taking strategies of high and low mathematics achievers. *The Journal of Educational Research*, 99(3), 144–155.
- Hwang, G. J., Xie, H., Wah, B. W., & Gašević, D. (2020). Vision, challenges, roles and research issues of artificial intelligence in education. *Comput. Educ. Artif. Intell.*, 1(1), 100001.
- Iiyama, M., Nakatsuka, C., Morimura, Y., Hashimoto, A., Murakami, M., & Minoh, M. (2017). Detecting answer stuck point using time intervals of pen strokes. *Transactions of Japanese Society for Information and Systems in Education*, 34(2), 166–171.
- Iqbal, T., & Qureshi, S. (2020). The survey: Text generation models in deep learning. *Journal of King Saud University - Computer and Information Sciences*, 34(6), 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.
- Kishi, K., & Miura, M. (2018). Detecting learners’ weak points utilizing time intervals of pen strokes. *International Journal of Learning Technologies and Learning Environments*, 1(1), 61–77.
- Learning, C. (2023). *Why cl: Research*. <https://www.carnegielearning.com/why-cl/research/>
- León, J., Olmos, R., Escudero, I., Cañas, J., & Salmerón, L. (2006). Assessing short summaries with human judgments procedure and latent semantic analysis in narrative and expository texts. *Behav. Res. Methods*, 38, 616–627.
- Levinstein, I. B., Boonthum, C., Pillarisetti, S. P., Bell, C., & McNamara, D. S. (2007). Start 2: Improvements for efficiency and effectiveness. *Behavior Research Methods*, 39, 224–232. <https://doi.org/10.3758/BF03193151>
- Lightman, H., Kosaraju, V., Burda, Y., Edwards, H., Baker, B., Lee, T., Leike, J., Schulman, J., Sutskever, I., & Cobbe, K. (2023). Let’s verify step by step. *arXiv preprint abs/2305.20050*.
- Lombrozo, T. (2006). The structure and function of explanations. *Trends in cognitive sciences*, 10(10), 464–470.
- Lu, V. N., Wirtz, J., Kunz, W. H., Paluch, S., Gruber, T., Martins, A., & Patterson, P. G. (2020). Service robots, customers, and service employees: What can we learn from the academic literature and where are the gaps? *J. Serv. Theory Pract.*, 30, 361–391.

- McEldoon, K. L., Durkin, K. L., & Rittle-Johnson, B. (2013). Is self-explanation worth the time? a comparison to additional practice. *British Journal of Educational Psychology*, 83(4), 615–632. <https://doi.org/10.1111/j.2044-8279.2012.02083.x>
- 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.
- McNamara, D., Levinstein, I., & Boonthum, C. (2004). Istart: Interactive strategy training for active reading and thinking. *Behavior Research Methods, Instruments, & Computers*, 36, 222–233.
- Nakamoto, R., Flanagan, B., Dai, Y., Takami, K., & Ogata, H. (2024). Unsupervised techniques for generating a standard sample self-explanation answer with knowledge components in a math quiz. *Res. Pract. Technol. Enhanc. Learn.*, 19, 016. <https://doi.org/10.58459/rptel.2024.19016>
- 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. *ICCE 2021*, 22–26.
- Ochoa, X., Chiluiza, K., Méndez, G., Luzardo, G., Guamán, B., & Castells, J. (2013). Expertise estimation based on simple multimodal features. *Proceedings of the 15th ACM on International conference on multimodal interaction*, 583–590.
- O’Neil, H. F., Chung, G. K. W. K., Kerr, D., Vendlinski, T. P., Buschang, R. E., & Mayer, R. E. (2014). Adding self-explanation prompts to an educational computer game. *Computers in Human Behavior*, 30, 23–28. <https://doi.org/10.1016/J.CHB.2013.07.025>
- Özsoy, M., Alpaslan, F., & Çiçekli, I. (2011). Text summarization using latent semantic analysis. *J. Inf. Sci.*, 37, 405–417.
- 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. *International Conference on Artificial Intelligence in Education*, 27–30.
- Panaite, M., Ruseti, S., Dascalu, M., Balyan, R., McNamara, D., & Trausan-Matu, S. (2019). Automated scoring of self-explanations using recurrent neural networks. *European Conference on Technology Enhanced Learning*, 16–19.
- Papineni, K., Roukos, S., Ward, T., & Zhu, W. (2002). Bleu: A method for automatic evaluation of machine translation. *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, 311–318. <https://doi.org/10.3115/1073083.1073135>
- Peña-Ayala, A. (2018). Learning analytics: A glance of evolution, status, and trends according to a proposed taxonomy. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 8, e1243.
- Ping, H., Stoyanovich, J., & Howe, B. (2017). Datasynthesizer: Privacy-preserving synthetic datasets. *Proceedings of the 29th International Conference on Scientific and Statistical Database Management*, 27–29.
- Reimers, N., & Gurevych, I. (2019). Sentence-bert: Sentence embeddings using siamese bert-networks. *Proceedings of the 2019 Conference on Empirical Methods in Nat-*

- ural Language Processing and the 9th International Joint Conference on Natural Language Processing*, 3982–3992. <https://doi.org/10.18653/v1/D19-1410>
- Renkl, A. (1997). Learning from worked-out examples: A study on individual differences. *Cognitive Science*, 21(1), 1–29.
- Renkl, A. (2017). Learning from worked-examples in mathematics: Students relate procedures to principles. *ZDM*, 49(4), 571–584.
- Ritter, S., Anderson, J. R., Koedinger, K. R., & Corbett, A. (2007). Cognitive tutor: Applied research in mathematics education. *Psychonomic Bulletin & Review*, 14(2), 249–255.
- Rittle-Johnson, B. (2006). Promoting transfer: Effects of self-explanation and direct instruction. *Child Dev.*, 77, 1–15.
- Rittle-Johnson, B. (2017a). Developing mathematics knowledge. *Child Dev Perspect*, 11, 184–190. <https://doi.org/10.1111/cdep.12229>
- Rittle-Johnson, B. (2017b). Developing mathematics knowledge. *Child Dev. Perspect.*, 11, 184–190.
- Rittle-Johnson, B., Loehr, A. M., & Durkin, K. (2017). Promoting self-explanation to improve mathematics learning: A meta-analysis and instructional design principles. *ZDM*, 49(4), 599–611.
- Rittle-Johnson, B., & Schneider, M. (2015). Developing conceptual and procedural knowledge of mathematics. In R. C. Kadosh & A.
- 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>
- Rubin, D. (1993). Statistical disclosure limitation. *Journal of Official Statistics*, 9, 461–468.
- Salazar, A., Vergara, L., & Safont, G. (2021). Generative adversarial networks and markov random fields for oversampling very small training sets. *Expert Systems with Applications*, 163, 113819.
- Schunk, D. H., & Pajares, F. (2005). Competence perceptions and academic functioning. In *Handbook of competence and motivation* (pp. 85–104). Guilford Publications.
- Siegler, R. S., & Jenkins, E. (1989). *How children discover new strategies*. Lawrence Erlbaum Associates, Inc.
- Siemens, G. (2012). Learning analytics: Envisioning a research discipline and a domain of practice. *Proceedings of the 2nd International Conference on Learning Analytics and Knowledge*, 4–8.
- Sinha, S., Rogat, T., Adams-Wiggins, K., & Hmelo-Silver, C. (2015). Collaborative group engagement in a computer-supported inquiry learning environment. *International Journal of Computer-Supported Collaborative Learning*, 10, 273–307.
- Star, J. R. (2005). Reconceptualizing procedural knowledge. *Journal for Research in Mathematics Education*, 36, 404–411.

- Su, P. Y., Zhao, Z. Y., Shao, Q. G., Lin, P. Y., & Li, Z. (2023). The construction of an evaluation index system for assistive teaching robots aimed at sustainable learning. *Sustainability*, 15(17), 13196. <https://doi.org/10.3390/su151713196>
- Suzuki, M. (2019). *Pretrained japanese bert models*. <https://github.com/cl-tohoku/bert-japanese>
- Takallou, F., Vahdany, F., Araghi, S. M., & Tabrizi, A. R. N. (2015). The effect of test taking strategy instruction on iranian high school students' performance on english section of the university entrance examination and their attitude towards using these strategies. *International Journal of Applied Linguistics and English Literature*, 4(6), 119–129.
- Thompson, D., & Senk, S. (1998). Using rubrics in high school mathematics courses. *Mathematics Teacher: Learning and Teaching PK-12*, 91, 786–793.
- VanLehn, K., & Jones, R. M. (1993). Learning by explaining examples to oneself: A computational model. In *Cognitive models of complex learning* (pp. 25–82). Kluwer Academic.
- VanLehn, K., Jones, R. M., & Chi, M. T. H. (1992). A model of the self-explanation effect. *The Journal of the Learning Sciences*, 2(1), 1–59.
- VanLehn, K., Jordan, P. W., Rose, C. P., Bhembé, D., Bottner, M., Gaydos, A., et al. (2002). The architecture of why2-atlas: A coach for qualitative physics essay writing. *Intelligent tutoring systems, 2002, 6th international conference*, 158–167.
- VanLehn, K., Siler, S., Murray, C., Yamauchi, T., & Baggett, W. (2003). Why do only some events cause learning during human tutoring? *Cognition and Instruction*, 21, 209–249.
- 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>
- Wang, T., Inoue, N., Ouchi, H., Mizumoto, T., & Inui, K. (2019). Inject rubrics into short answer grading system. *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 3–7.
- Warli, I., Cintamulya, & Rahayu, P. (2020). Scaffolding process based on students diagnostic difficulties in proving group problems by using mathematics mapping. *Journal of Physics: Conference Series*, 1422.
- Yoshitake, D., Flanagan, B., & Ogata, H. (2020). Supporting group learning using pen stroke data analytics. *28th International Conference on Computers in Education Conference Proceedings*, 634–639.
- Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., & Artzi, Y. (2020). Bertscore: Evaluating text generation with bert.

Appendices

Institutional Review Board (IRB) Approval and Ethical Considerations

As this study involves the use of student data, we acknowledge the importance of obtaining approval from the Institutional Review Board (IRB). We have taken the necessary steps to ensure compliance with ethical guidelines, and the study has been submitted to and approved by the IRB. Consent for using the students' data in our research is obtained from their guardians at the beginning of each academic year. We provide detailed information about the purpose of data collection, how it will be used, and the measures taken to ensure confidentiality and privacy. The guardians have the right to decline consent or withdraw their consent at any time without any negative consequences for the students.