

StuBot: Learning by Teaching a Conversational Agent Through Machine Reading Comprehension

Nayoung Jin[†], Hana Lee[†], Minsam Ko^{*}

Department of Applied Artificial Intelligence, Hanyang University, South Korea
{jinfficial, hanalee, minsam}@hanyang.ac.kr

[†] All the authors contributed equally to this research, ^{*} Corresponding author

Abstract

This paper proposes StuBot, a text-based conversational agent that provides adaptive feedback for learning by teaching. StuBot first asks the users to teach the learning content by summarizing and explaining it in their own words. After the users inputted the explanation text for teaching, StuBot uses a machine reading comprehension (MRC) engine to provide adaptive feedback with further questions about the insufficient parts of the explanation text. We conducted a within-subject study to evaluate the effectiveness of adaptive feedback by StuBot. Both the quantitative and qualitative results showed that learning by teaching with adaptive feedback can improve learning performance, immersion, and overall experience.

1 Introduction

Learning by teaching is a teaching method that enables students to teach other students. Many studies have revealed the effectiveness of learning by teaching from diverse perspectives, such as long-term learning (Fiorella and Mayer, 2014), motivation and self-esteem (Wagner and Gansemer-Topf, 2005), communication skills (Stollhans, 2016), and abilities to gather and structure information (Grzega and Schöner, 2008).

Some studies proposed situated and interactive learning with virtual agents using a learning by teaching approach, such as Betty’s Brain (Leelawong and Biswas, 2008), Curiosity Notebook (Law et al., 2020), and SimStudent (Matsuda et al., 2010). However, these studies mainly depended on structured interactions for the presentation (e.g., typing words in a form or clicking items). For example, the student had to use a structured concept map to express the causal relationships of a river ecosystem (Leelawong and Biswas, 2008) or to drag/drop math icons to teach formulas (Donggil, 2017). In addition, choosing the following message or response among the given candidates was adopted

to build a situation involving dialoguing with a virtual agent (Iwase et al., 2021). Although such structured interaction is convenient and suitable for well-defined and simple learning tasks, it limits the user’s expression.

Few studies have addressed the earlier ideas of using conversational agents in learning by teaching to support users’ free expression by allowing them to input texts (Law et al., 2020; Park and Kim, 2015). However, in these studies, giving feedback for learning by teaching was limited and rarely covered owing to the difficulties in understanding the user’s input texts and in extracting what needs to be improved.

In this paper, we propose StuBot, a text-based *conversational* agent that promotes learning by teaching, as shown in Figure 1. StuBot builds a virtual teacher-student situation. The user plays the role of a teacher by explaining a learning material to prepare StuBot for an exam. Based on the user’s explanation texts, StuBot continues to question insufficient parts of the explanation. To do this, StuBot internally solves exercise problems through machine reading comprehension (MRC) that generates answer texts by analyzing the explanation. If the generated answer is incorrect, StuBot asks the user a question about the key topic of the corresponding question. Then, the user prepares another explanation for better teaching StuBot. After several iterations, StuBot finally takes an exam, and its exam results are shown to evaluate the user’s learning and teaching.

We conducted a within-subject study ($n = 20$) to evaluate the effectiveness of StuBot. The participants were asked to study the given learning materials by teaching StuBot with and without adaptive feedback. The study results showed that the participants exhibited significantly greater learning performance with adaptive feedback from StuBot. In addition, StuBot increased the participants’ immersion in learning by teaching. Furthermore, the qual-

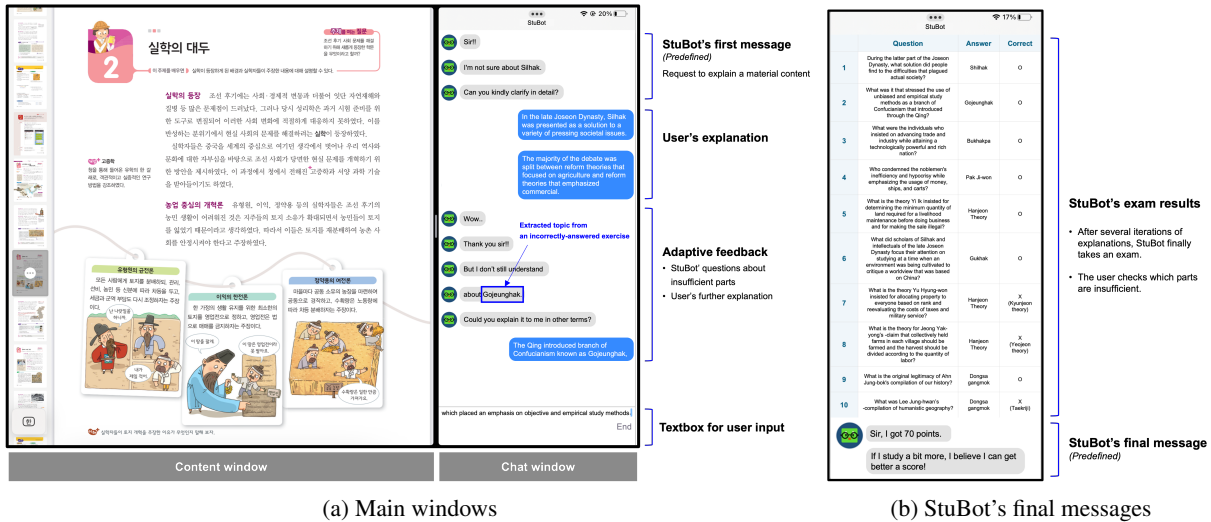


Figure 1: User interface of StuBot (with translated texts)

itative analysis results revealed the details about the participants' experiences with learning by teaching. Finally, on the basis of the findings, we discuss the implications of designing a conversational agent for learning by teaching.

2 Related Works

Many studies have proposed agent-based systems for learning, such as facilitating school administrative services (Pérez et al., 2020; Ranoliya et al., 2017) and improving knowledge of a specific subject (Hobert, 2019; Nguyen et al., 2019). Furthermore, MathBot explains mathematical concepts to students and guides them in solving exercise problems (Grossman et al., 2019). MAssistant helps users to trace the concepts they have learned and to build their own concept graphs (Jiang et al., 2019).

In particular, there have been virtual agents to motivate learners by situating them in a learning by teaching scenario. For example, Betty's Brain works in a scenario where the user prepares Betty, a virtual agent as a student, for an exam to qualify for a science club membership (Leelawong and Biswas, 2008). In Betty's Brain, the user explains a river ecosystem to Betty by filling in the blanks of a conceptual map (Leelawong and Biswas, 2008). Similarly, in SimStudent (Matsuda et al., 2010), the user helps a virtual agent pass a quiz. A virtual tutee system allows the user to build an intimate relationship with the conversation agent and teach it with responsibility (Park and Kim, 2015). Iwase et al. allowed the user to explain the theory of information technology to the virtual agent by selecting predefined dialogue sentences (Iwase et al., 2021). EnTAM teaches mathematics to virtual agents by

dragging and dropping mathematical icons to explain a formula (Donggil, 2017).

Such interactive systems for learning by teaching can be categorized into four subspaces according to two dimensions: (i) *presentation* and (ii) *feedback*. First, the interactive systems for learning by teaching can be distinguished by the existence of predefined structures for the presentation. Many prior systems were based on structured interactions that allow the user to explain the learning material to an agent by selecting the responses or typing the values corresponding to specific attributes (Leelawong and Biswas, 2008; Matsuda et al., 2010; Donggil, 2017; Iwase et al., 2021).

In contrast, non-structural interactions are also possible with conversational agents (Park and Kim, 2015; Law et al., 2020). The system based on conversational agents enables users to deliver their lessons in their own words. This can take advantage of verbal explanations during learning. Second, feedback types can also specify the systems for learning by teaching. Most systems support adaptive feedback by highlighting the insufficient parts of the users' knowledge according to their previous presentations. Such adaptive feedback to supplement previous understanding and explanations can improve the performance (Duran, 2017).

Prior systems mainly depended on structured presentations with adaptive feedback. The structured interaction can be convenient; however, allowing verbal expressions in the learner's own words can promote learning performance. Furthermore, it is difficult to construct more complex learning tasks, such as summarizing historical events or paraphrasing literature articles. Few studies have presented

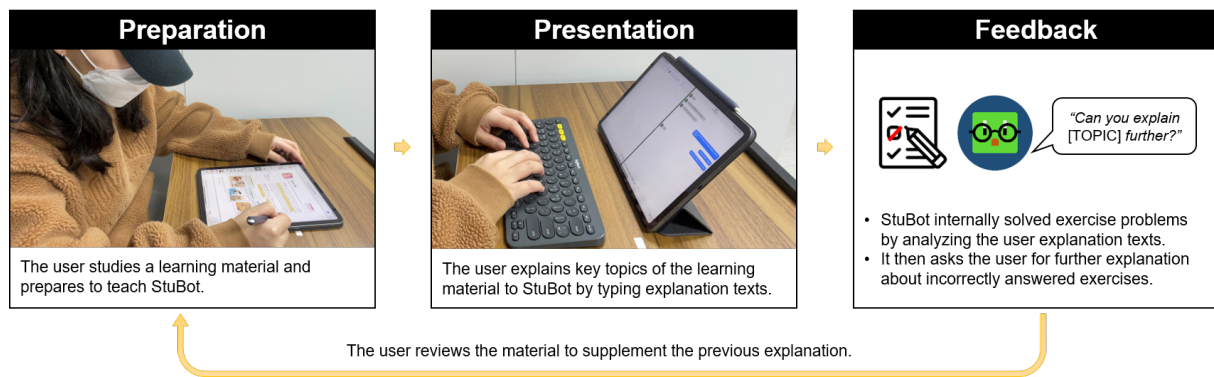


Figure 2: Learning by teaching StuBot

the initial idea and feasibility of using conversational agents for learning by teaching (Law et al., 2020; Park and Kim, 2015).

However, the feedback was limited in these studies owing to the difficulties in processing the user’s explanation texts. For example, providing feedback to users and correcting their previous teaching are not iterative (i.e., ending with the evaluation results rather than continuing the conversation) (Law et al., 2020), or feedback messages were typed by a human teacher (Park and Kim, 2015). This paper presents a new conversational agent that provides adaptive feedback to facilitate learning by teaching.

3 Workflow of StuBot

Bargh et al. (Bargh and Schul, 1980) explained learning by teaching in three steps (i.e., preparation, presentation, and feedback), and many methods to facilitate each step have been studied. For example, in the preparation step, motivating a learner to teach other students can increase the study time (Bargh and Schul, 1980; Duran, 2017) and learning performance (Hayes-Roth, 1977). In the presentation step, students can gain more benefits from the verbal presentation to the audience in terms of problem-solving skills (Gagne and Smith Jr, 1962) or long-term memory (Bargh and Schul, 1980). Furthermore, providing and receiving feedback can help learners concentrate on the content and supplement the current understanding (Wagster et al., 2007; Bargh and Schul, 1980; Duran, 2017).

As shown in Figure 2, StuBot is designed to support the three steps of learning by teaching. First, in the preparation step, the learner is asked to prepare StuBot for an exam. Similar to the agent for counselor training (Demasi et al., 2020), StuBot uses a student persona to increase immersion in learning. A learning session begins with StuBot’s message, “I have a test this week. Please teach me

about the *OOO*.” The learner then prepares a lesson that summarizes and explains the materials.

Second, the presentation step follows after the learner completes the preparation step. StuBot allows the learner to explain using a free-text form similar to that used in conventional messengers. The learner can send a message to StuBot by typing texts and pressing the Enter button. The user can input the explanation texts in multiple subsequent messages, and the sent messages can be modified or deleted while the End button to finish the explanations is not pressed.

Finally, in the feedback step, StuBot identifies insufficient parts by processing the user’s explanation texts. It asks the user for further explanation about the insufficiently explained topics using feedback messages such as “Can you explain *OOO* in another way?” The learner then prepares another explanation. StuBot continues to provide feedback until the learner covers all the important topics or the given learning time is over. After some iterations, StuBot finally takes an exam using all the previous explanations. Then, the test results are delivered by showing StuBot’s answer to each problem and its actual answer, and the learner checks which parts are insufficient.

4 Implementation

Figure 3 shows three modules for analyzing user’s explanation texts and generating feedback messages: answering, self-grading, and messaging.

4.1 Answering module

StuBot creates answer texts for 10 predefined exercises based on a user’s explanation texts. The exercises are factoid questions that ask users to type short texts related to the key topics of the learning material. StuBot adopts a model of MRC that learns a predictive function f which generates an

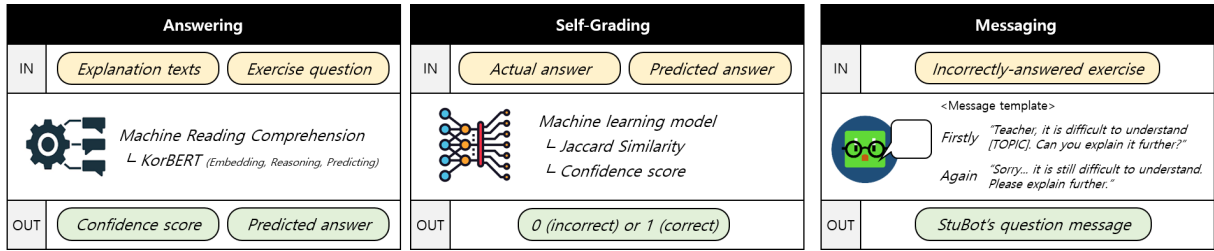


Figure 3: Internal workflow of StuBot for feedback messaging

appropriate answer A by receiving a context C and a related question Q : $f(C, Q) \rightarrow A$ (Baradaran et al., 2020). We used a ETRI’s trained MRC model (ETRI, 2021) that adopts KorBERT (Korean Bidirectional Encoder Representations from Transformers), an extension of BERT (Devlin et al., 2018). KorBERT performs morphological analysis to harness characteristics of the Korean language and processes texts at the morpheme level.

First, the model was trained on unlabeled data using a masked language model (MLM) in the pre-training step. KorBERT predicts the original vocabulary of some words randomly masked in the input according to the context. MLM trains a deep bidirectional transformer by fusing its left/right contexts to enable learning of the contextual relations between words in a text. For this pretraining, an extensive Korean natural language dataset was used, which contains 4.7 billion morphemes (ETRI, 2021), including news articles and encyclopedias.

Next, in the fine-tuning step, KorBERT begins with the pretrained weights, and all of the weights are fine-tuned by training with labeled data from downstream tasks, i.e., machine reading comprehension in this case. Finally, MRC predicts and extracts an answer text by analyzing relevances between the given question text and each explanation sentence through the trained KorBERT.

To evaluate the performance of the trained KorBERT on our requirement, we collected 30 passages from Korean Wikipedia and created 10 factoid questions per passage as positive cases, which can be answered according to the passage because it contains its answer text. We also made another 10 questions per passage as negative cases, which were not relevant to the corresponding passage. We inputted both the 300 positive and the 300 negative cases to the MRC model. These negative cases are to test whether StuBot can give wrong answers when the user input insufficient explanation.

Five human judges were recruited to determine whether each generated answer text by the MRC

model was correct, given its corresponding question and passage. In most cases, the five judges were unanimous. For the other cases, the final decision was made by the majority. The result showed that the MRC model correctly answered 290 questions (96.7%) among the 300 positive cases and incorrectly answered all the negative questions. These results show that StuBot with this model can solve exercises correctly as long as the user inputs a sufficient explanation.

4.2 Self-grading module

Next, StuBot self-grades the answer text using the MRC model by comparing it with its actual answer value corresponding to a question. This aims to identify the insufficient parts of the user’s explanation. If StuBot incorrectly answered many questions, the user’s explanation should be improved.

However, simply comparing the created answer text and the actual text can be too strict for grading. The answer text may consist of multiple words and be decorated with additional adjectives or adverbs. Therefore, the created answer texts might not be exactly the same as the actual texts. We built a simple machine learning model for grading to deal with this problem.

We trained a model of support vector machine by using two input features: (i) Jaccard similarity between the answer text and the actual text and (ii) the confidence value of the MRC model, which indicates how accurate the outputted answer text is. We used the above 600 answer texts created by the MRC engine and their actual values to train and test the models. Among the 600 pairs of MRC-created answers and actual ones, 290 pairs were regarded as correct by the five human judges and the others were regarded as negative cases. 80% of the pairs were used to train a model, with the rest being used for the testing. The positive and negative cases were equally distributed in both the training and test datasets. As the results, the trained model achieved 96.5% of accuracy.

4.3 Messaging module

Once the insufficient parts of the user explanation were identified in the previous module, StuBot prepares adaptive feedback messages for the user. To simulate learning by teaching, we designed StuBot as a human student who requests extra help from the user. StuBot seeks further explanation for the topics covered by the wrongly answered exercises. Note that StuBot's feedback is in the form of descriptive questions. We considered descriptive questions rather than other forms (e.g., multiple choice or fill in the blanks) because answering a descriptive question involves learning-related activities, such as analyzing contents and writing a summary in the user's own words.

The first feedback message is composed by inserting the actual answer into one of six templates, asking for further explanations about the actual answer. For example, StuBot sends a message such as *"Teacher, it is difficult to understand [TOPIC]. Can you explain it further?"* Next, if the user inputs another new explanation by responding to the first feedback message, StuBot performs the three modules again. If StuBot wrongly answered the same exercise again despite the user's supplement, a different template is used for the next message, such as *"Sorry... it is still difficult to understand. Please explain further."* If StuBot does not correctly answer the same exercise two consecutive times, it skips it and moves to the next problem.

5 Evaluation

5.1 Study design

We designed a within-subject experiment by considering the individual differences in learning performance, referring to the experiments in earlier studies (Fiorella and Mayer, 2014; Leelawong and Biswas, 2008; Wagster et al., 2007; Matsuda et al., 2010). We recruited 20 participants (10 males and 10 females) via a social networking service. They had a mean age of 23.4 years ($SD = 2.32$) and were college students who were mostly majoring in computers. The experiment took them approximately 1.5 h. They were compensated for about 10 USD. We denoted the participants as P1–P20.

For comparison with earlier conversational agents for learning by teaching (Law et al., 2020; Park and Kim, 2015), we prepared *Baseline*. *Baseline* operates similarly to StuBot. It provides graded results on the agent's answers on exercises according to the input explanation and encourages the

user to learn more by checking what is incorrectly answered. However, it does not interact with the user through feedback messages to explicitly ask for further explanations.

A participant was first asked to respond to a pre-survey. The pre-survey included a quiz to evaluate the participants' prior knowledge about Korean history and several quick questions about demographics and background. The participants then took a learning session. In a learning session, the participants studied a given material using a given agent, either StuBot or *Baseline*. At the end of each session, the participants responded to a post-survey that contained a quiz to evaluate the learning performance (Three question types: multiple-choice, descriptive, and comprehensive questions) and questions to measure immersion. In addition, the post-survey included open-ended questions about overall experiences in learning by teaching. The participants took such a learning session twice with different materials and agents. The order of both the agent and the material for each participant was controlled via randomization. The details of the learning session and measurements are described in Appendix A and B.

5.2 Materials

This study used Korean history because the participants in common would likely have learned related topics at least once. Also, historical materials have been widely used in studies on learning and teaching methods (Edmunds, 2006). We prepared two materials with about 300–350 words from a textbook of Korean history in a middle school. In addition, we prepared 10 factoid exercise problems that StuBot and *Baseline* must solve internally. The problems were selected from a teaching manual book for each material.

5.3 Analysis methods

We conducted both quantitative and qualitative analyses. We conducted a paired *t*-test (two-tailed) to analyze the learning performance and experiences according to the agent type. In addition, we performed an analysis of generalized estimating equation (GEE) (Diggle et al., 2002). Because our study used repeated measures for a participant, we considered GEE, which handles repetitive measurement data with multiple independent variables, like in (Amini et al., 2017; Nam et al., 2017). We used four dependent variables: the three scores about learning performance (i.e., the multiple-choice, de-

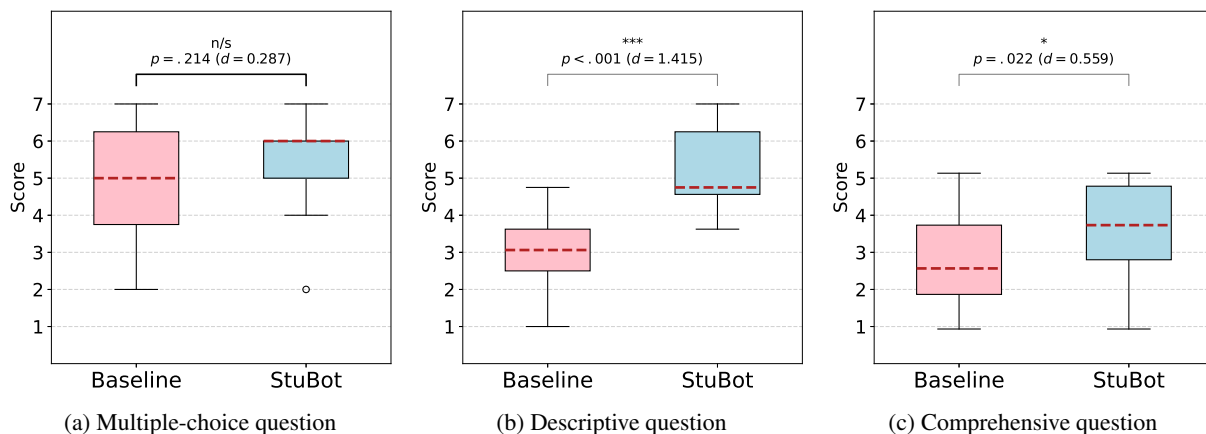


Figure 4: Paired t -test results on learning performance

scriptive, and comprehensive questions) and the immersion score. There were two independent variables: agent type and prior knowledge.

Furthermore, guided by (Braun and Clarke, 2012), we conducted a thematic analysis of the participants' responses to the open-ended questions. Two researchers first generated the initial codes by summarizing the responses to the related questions. Next, we attempted to find patterns from the results of our study objectives. Finally, we defined and named themes by clarifying the extracted patterns.

5.4 Quantitative results

5.4.1 Learning performance

Figure 4 presents the paired t -test results of the participants' quiz scores. We found that the difference in the scores was significant and large on the descriptive questions ($t = 6.329$, $p < .001$, $d = 1.415$). Similarly, the scores on the comprehensive question were significantly different and showed a medium level of the effect size ($t = 2.502$, $p = .022$, $d = 0.559$). The comprehensive question was the most difficult type, as its average score was the lowest, followed by the descriptive ones and the multiple-choice ones (comprehensive: $M = 5.2$, $SD = 1.49$ > descriptive: $M = 4.1$, $SD = 1.48$ > multiple-choice: $M = 3.22$, $SD = 1.25$).

The GEE analysis gave similar results, as shown in Table 1. We found a single independent variable that significantly affected the scores on the descriptive questions. Using StuBot was significantly and positively related to it ($p < .001$, $\beta = 2.063$). Moreover, using StuBot significantly and positively affected the scores on the comprehensive question ($p = .010$, $\beta = 2.100$). The prior knowledge was not significantly related to learning performance.

Variables	β	SE	z	p
(Intercept)	4.907	0.743	6.608	<.001
Agent type				
<i>StuBot</i>	0.600	0.455	1.319	.187
<i>Baseline</i>	reference category			
Prior knowledge	-0.002	0.183	-0.010	.992

(a) Multiple-choice question

Variables	β	SE	z	p
(Intercept)	2.685	0.350	7.669	<.001
Agent type				
<i>StuBot</i>	2.063	0.318	6.493	<.001
<i>Baseline</i>	reference category			
Prior knowledge	0.115	0.085	1.348	.178

(b) Descriptive question

Variables	β	SE	z	p
(Intercept)	6.032	1.534	3.933	<.001
Agent type				
<i>StuBot</i>	2.100	0.818	2.567	.010
<i>Baseline</i>	reference category			
Prior knowledge	-0.050	0.376	-0.134	.894

(c) Comprehensive question

Table 1: GEE analysis results on learning performance

5.4.2 Immersion

Figure 5 presents the paired t -test results on the immersion scores. The participants' immersion tended to be higher when they used StuBot (Baseline: $M = 3.61$, $SD = 0.73$; StuBot: $M = 4.27$, $SD = 0.48$). The total immersion score was significant and showed a large effect size ($t = 5.329$, $p < .001$, $d = 1.192$) shown in Table 2, the GEE analysis showed similar results, indicating that using StuBot was significantly and positively related to the total immersion score ($p < .001$, $\beta = 0.660$). Specifi-

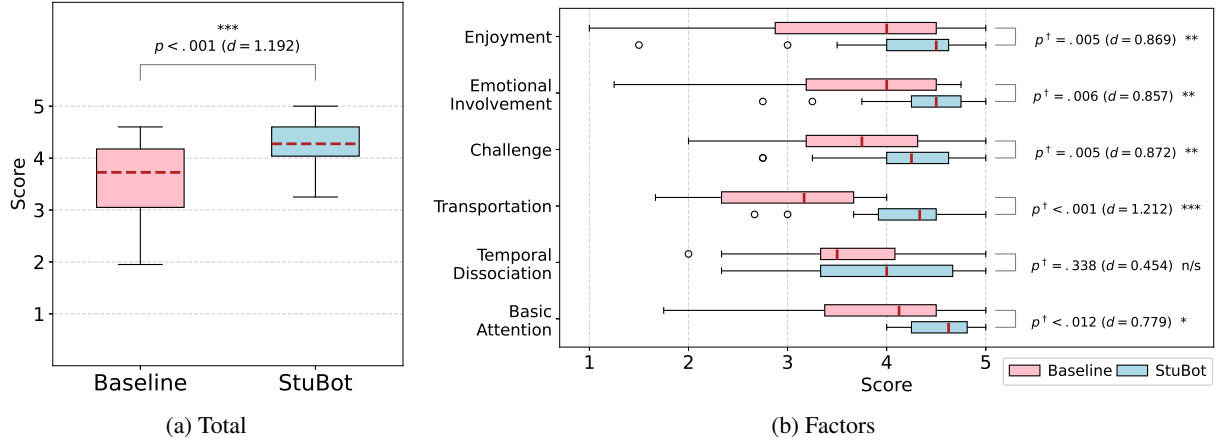


Figure 5: Paired t -test results on immersion (p^\dagger : adjusted p -value by Bonferroni correction)

Variables	β	SE	z	p
(Intercept)	3.711	0.264	14.053	<.001
Agent type				
<i>StuBot</i>	0.660	0.121	5.468	<.001
<i>Baseline</i>	reference category			
Prior knowledge	-0.028	0.066	-0.425	.671

Table 2: GEE analysis results on immersion

cally, the use of StuBot showed higher values for the subfactors than those of the use of Baseline, except for the temporal dissociation. In particular, the differences between StuBot and Baseline were the largest in the transportation factor ($p < .001$, $d = 1.212$), indicating that the participants interacted more with StuBot and were more willing to teach it in the learning by teaching situation.

5.5 Qualitative results

5.5.1 User experience with StuBot

First, our qualitative analysis revealed the benefits of learning by teaching. We found that it helped participants understand the overall storyline in detail. One participant (P5) commented, “*To teach StuBot, I had to study harder to understand its overall storyline entirely. So, I think that is why I could get a higher score.*” Although most responses were positive about learning by teaching, there were few opinions about its difficulties, mostly due to unfamiliarity. For example, one participant (P14) said, “*This is my first time to teach someone, so I found it difficult to explain what I learned.*”

Second, we found that StuBot’s feedback increased awareness about what the participants knew or did not know. One participant (P3) said, “*When I studied the material at first, I was not sure about*

how well I knew. However, with StuBot, I could clearly understand what I was insufficient in.” Similarly, P14 responded, “*By teaching StuBot, I could catch part of what I missed.*” Such increased awareness helped the participants learn more by supplementing their insufficient or incorrect understanding. One participant (P10) said, “*When using this [Baseline], I was not sure about my explanation. However, StuBot asked me about the difficult parts of my explanation. I could know which part is insufficient and study that part to better teach StuBot.*” Another participant (P13) responded, “*StuBot’s messages were helpful because they let me know what is lacking in my explanation.*”

Finally, there were some responses regarding the various impressions with the agent. For example, some participants reported feeling a sense of accomplishment when checking the agent’s final score on the exam. For example, one participant (P4) said, “*With StuBot, I felt like I had studied well in detail.*” Another participant (P13) said, “*I realized that I had taught StuBot very well after seeing that it finally achieved a high score.*” On the other hand, some participants were disappointed when their agents showed a low score. For example, one participant (P6) said, “*With Baseline, I thought I explained it to the agent properly, but I realized I missed a lot later.*” Another participant (P7) said, “*I felt like I didn’t teach the agent well.*” Moreover, some participants said that StuBot’s feedback gives the impression that it is like a real student. For example, one participant (P7) said, “*I felt that StuBot acted like a real student when it asked me a question.*” Another participant (P16) said, “*StuBot’s messages reminded me of the time when I taught some students at an academy.*”

Sender	Message
StuBot	<i>“Ah, I see now what it is! Then, please simplify how the Gukhak was conducted.”</i>
User	<i>“Gukhak is study that break away from the conventional Chinese-centered worldview and instead focus on gaining an in-depth understanding of the values that are held in Korea.”</i>
StuBot	<i>“Sir, can you clarify in more detail?”</i>
User	<i>“‘Hunminjeongeum Unhae’ was prepared by Shin Kyung-jun and ‘Daedong Yeojido’, a map describing the geography of Korea in detail was compiled by Kim Jeong-ho. Taekriji, a collection of papers on humanistic geography, was another project undertaken by Lee Jung-hwan. It indicates that we have begun to undertake study on Korean culture in a more active manner and move away from the ideals of Chinese culture.”</i>

(a) Participant #6

Sender	Message
StuBot	<i>“Thank you sir!! But I don’t know much about Gojeunghak yet, so would you please explain?”</i>
User	<i>“It is one of the areas of research that originated during the Qing Dynasty, and it is a field of research that seeks to provide objective solutions to the challenges that face our modern society.”</i>
StuBot	<i>“I’m sorry sir, but I still don’t understand.. If there is anything further you can tell me about this, I would really appreciate it.”</i>
User	<i>“For instance, Gojeunghak is a subject that our society has to address. Due to the fact that we were a rural community, the issues that developed the quickest were those pertaining to agriculture. Many different solutions, such as different ways of allocating land, were presented in order to resolve this problem.”</i>

(b) Participant #10

Table 3: Failure cases: the user’s first explanation seemed sufficient, but StuBot failed to output the correct answer internally and asked for further explanation (All messages are translated, and their original texts are presented in the Appendix C).

5.5.2 Failure Cases

During the experiment, StuBot showed a similar performance to the evaluation results for each model, described in Section 4. We found that 9.1% of the answers generated by MRC were wrong even though the participants’ explanations were sufficient. Moreover, the self-grading module showed a 0.9% error rate. For example, as shown in Table 3, P6 and P10 correctly answered StuBot’s first question, but StuBot asked the participants to explain more. Such errors mostly occurred when the user explanation terms differed from the exercise question and its actual answer, even though their semantics were identical.

We think that the failure cases were posed by the limitations of our MRC approach. Given an answer, our MRC uses the selective (extractive) method, which extracts terms from the explanation. Even if the user input text is semantically correct, our answering and self-grading modules can output wrong results if it does not contain terms in the question or actual answer.

Interestingly, the negative impacts of the errors on the user experience seemed not severe. Despite

the internal errors, the participants just continued the conversation with StuBot. When the errors occurred, the participants inputted another explanation text by using different expressions or completing the sentence. The participants tried explaining it better because they were a teacher.

The generative method of MRC can help reduce internal errors, but we believe that the selective method is more suitable for StuBot. If MRC is too great, it can predict the correct answer even if the user explanation is insufficient; StuBot likely skips adaptive feedback. We think there is an appropriate design point between the MRC performance and system effectiveness (i.e., learning supports).

6 Discussion

6.1 Awareness

The results show that StuBot can improve the self-awareness of the learner’s understanding level. These results are in line with earlier studies that addressed the benefits of questioning (Dillon, 2006) and self-awareness of the learning status (Bargh and Schul, 1980). In particular, the explicitness of

asking for a further explanation can be essential for efficiently increasing the learner’s awareness. In Baseline, the agent provides exam test results to the users but does not explicitly ask them to input another explanation. Even though the time lengths of the learning sessions with a different agent were not different, the learning performance was greater when using StuBot. In addition, some participants mentioned that they liked to receive feedback messages because StuBot pointed out where to focus.

Earlier systems with conversational agents for learning by teaching suffered from computationally understanding the user description, and, therefore, their feedback was limited (Law et al., 2020) or was done by a human teacher (Park and Kim, 2015). However, we believe that other advanced natural language processing, such as the generation of automatic question expression (Qu et al., 2021), can also improve the explicitness of adaptive feedback.

6.2 Interactivity

Interactivity is known to be essential for task performance in a virtual environment (Coulter et al., 2007; Sowndararajan et al., 2008). Similarly, we found that StuBot made the most significant difference in the transportation of the immersion score, which is related to interactivity. This may be because StuBot plays a proactive role by giving questions. Similarly, Liu et al. (Liu et al., 2020) showed that the proactive conversational agent can help increase interactivity. Further studies can be performed to design a conversational agent’s roles and effective moments of intervention during learning.

6.3 Intimacy

The conversational agent’s adaptive feedback can motivate learners to focus on a learning by teaching situation by increasing the intimacy. Many participants in our study mentioned the different impressions they had with the agent during and after learning. For example, some participants expressed pride in the outcome that the agent finally made. Other participants were disappointed by the agent’s lower score on the final exam than expected. Furthermore, a participant commented that StuBot reminded him of teaching students in real-world settings. Such findings can extend prior studies on agent-based learning by teaching without textual conversation (Jenkins et al., 2007; Kim et al., 2020) by addressing another important design consideration. Further studies are needed to understand the learner’s impressions with the student agent and

harness them to increase motivation and accomplishment in learning.

6.4 Inconvenient Interaction

Finally, designing conversational agents for learning by teaching may be different from designing conventional chatbot agents in education that mainly aim to offer convenience. With a conventional chatbot, the user mostly takes a proactive role by requesting something from the agent. Then, the agent interprets the user intent according to the input texts and responds to it appropriately.

However, the interaction with StuBot is different. StuBot requests first, and then the user responds. StuBot bothers the users for learning effects. This can be seen as an *inconvenient interaction* that is uncomfortable for the users but benefits them by requiring them to take action or perform an explicit activity (Rekimoto and Tsujita, 2014). In the post-survey, all the participants commented that they preferred to use StuBot over Baseline despite its annoying requests. For example, one participant commented that it was difficult to explain to StuBot, but it made him ponder and study more in detail. Therefore, it is necessary to address the issue of inconvenience with an agent because the user may give up using the system if it is too inconvenient. Further studies are needed to determine the appropriate task loads from a demanding agent.

7 Conclusion

This paper presented StuBot that asks the learner to teach a learning material and provides adaptive feedback to further explain the insufficient parts. To do this, StuBot overcomes the limitations of existing conversational agents for learning by teaching by using advanced natural language processing. We conducted a within-subject study using StuBot. The results reveal the effectiveness of StuBot and provide practical implications for designing conversational agents for learning by teaching.

This work can bring new opportunities for using MRC. In this study, MRC does not need to be perfect because StuBot has a student persona. Moreover, asking a question by StuBot after the user’s good explanation (i.e., error) was empirically acceptable because the user was a teacher who tries to explain better. We believe that our approach can extend application areas that MRC with practically technical limitations can apply.

Limitations

Our study still has several limitations and the results should be carefully understood in this context. First, the sample size of this study may not be sufficient to generalize the results; therefore, further experiments should be conducted on a large number of participants. Second, the experiments were conducted in a controlled laboratory setting, and, therefore, there could be different observations in practical situations. Third, the failure to consider all the possible variables in a single study owing to concerns about the complexity of the experiments and variable confounding can be extended by subsequent studies with other variables (e.g., chatbot's tone or personality). Finally, the learning materials and contents were limited to a specific type, and, therefore, further studies need to be conducted to analyze the effects of conversational agent-based learning by teaching in other learning domains.

Ethics Statement

All subjects participated voluntarily and provided their written informed consent to participate in this study. The Study was approved by the institutional review board of Hanyang University (HYUIRB-202204-013).

Acknowledgments

This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (Ministry of Science and ICT) (No. RS-2022-00155885, Artificial Intelligence Convergence Innovation Human Resources Development (Hanyang University ERICA)).

References

- Payam Amini, Roya Najafi-Vosough, Ali Ghaleiha, and Hossein Mahjub. 2017. Evaluation of bipolar disorder in several relapses over time using generalized estimating equations. *Iranian journal of psychiatry*, 12(3):182.
- Razieh Baradaran, Razieh Ghiasi, and Hossein Amirkhani. 2020. A survey on machine reading comprehension systems. *Natural Language Engineering*, pages 1–50.
- John A Bargh and Yaacov Schul. 1980. On the cognitive benefits of teaching. *Journal of Educational Psychology*, 72(5):593.
- Janell R Blunt and Jeffrey D Karpicke. 2014. Learning with retrieval-based concept mapping. *Journal of Educational Psychology*, 106(3):849.
- Virginia Braun and Victoria Clarke. 2012. Thematic analysis.
- Robert Coulter, Linda Saland, T Caudell, Timothy E Goldsmith, and D Alverson. 2007. The effect of degree of immersion upon learning performance in virtual reality simulations for medical education. *In-Medicine Meets Virtual Reality*, 15:155.
- Orianna Demasi, Yu Li, and Zhou Yu. 2020. A multi-persona chatbot for hotline counselor training. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3623–3636.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.
- Peter Diggle, Peter J Diggle, Patrick Heagerty, Kung-Yee Liang, Scott Zeger, et al. 2002. *Analysis of longitudinal data*. Oxford university press.
- James T Dillon. 2006. Effect of questions in education and other enterprises. *Rethinking Schooling*, pages 145–174.
- Song Donggil. 2017. Designing a teachable agent system for mathematics learning. *Contemporary Educational Technology*, 8(2):176–190.
- David Duran. 2017. Learning-by-teaching. evidence and implications as a pedagogical mechanism. *Innovations in Education and Teaching International*, 54(5):476–484.
- Julie Edmunds. 2006. How to assess student performance in history: Going beyond multiple-choice tests. *SERVE Center at University of North Carolina at Greensboro*.
- ETRI. 2021. KorBERT. https://aiopen.etri.re.kr/service_dataset.php.
- Logan Fiorella and Richard E Mayer. 2014. Role of expectations and explanations in learning by teaching. *Contemporary Educational Psychology*, 39(2):75–85.
- Robert M Gagne and Ernest C Smith Jr. 1962. A study of the effects of verbalization on problem solving. *Journal of experimental psychology*, 63(1):12.
- Joshua Grossman, Zhiyuan Lin, Hao Sheng, Johnny Tian-Zheng Wei, Joseph J Williams, and Sharad Goel. 2019. Mathbot: Transforming online resources for learning math into conversational interactions. *AAAI 2019 Story-Enabled Intelligence*.
- Joachim Grzega and Marion Schöner. 2008. The didactic model ldl (lernen durch lehren) as a way of preparing students for communication in a knowledge society. *Journal of Education for Teaching*, 34(3):167–175.
- Barbara Hayes-Roth. 1977. Evolution of cognitive structures and processes. *Psychological review*, 84(3):260.

- Sebastian Hobert. 2019. Say hello to 'coding tutor'! design and evaluation of a chatbot-based learning system supporting students to learn to program. In *ICIS*.
- Koya Iwase, Kota Gushima, and Tatsuo Nakajima. 2021. "relationship between learning by teaching with teachable chatbots and the big 5. In *2021 IEEE 3rd Global Conference on Life Sciences and Technologies (LifeTech)*, pages 191–194. IEEE.
- Marie-Claire Jenkins, Richard Churchill, Stephen Cox, and Dan Smith. 2007. Analysis of user interaction with service oriented chatbot systems. In *International Conference on Human-Computer Interaction*, pages 76–83. Springer.
- Charlene Jennett, Anna L Cox, Paul Cairns, Samira Dhoparee, Andrew Epps, Tim Tijs, and Alison Walton. 2008. Measuring and defining the experience of immersion in games. *International journal of human-computer studies*, 66(9):641–661.
- Lan Jiang, Shuhan Hu, Mingyu Huang, Zhichun Wang, Jinjian Yang, Xiaojun Ye, and Wei Zheng. 2019. Mas-sistant: a personal knowledge assistant for mooc learners. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): System Demonstrations*, pages 133–138.
- Soomin Kim, Jinsu Eun, Changhoon Oh, Bongwon Suh, and Joonhwan Lee. 2020. Bot in the bunch: Facilitating group chat discussion by improving efficiency and participation with a chatbot. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–13.
- Edith Law, Parastoo Baghaei Ravari, Nalin Chhibber, Dana Kulic, Stephanie Lin, Kevin D Pantasdo, Jessy Ceha, Sangho Suh, and Nicole Dillen. 2020. Curiosity notebook: A platform for learning by teaching conversational agents. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–9.
- Krittaya Leelawong and Gautam Biswas. 2008. Designing learning by teaching agents: The betty's brain system. *International Journal of Artificial Intelligence in Education*, 18(3):181–208.
- Yafei Liu, Hongjin Qian, Hengpeng Xu, and Jinmao Wei. 2020. Speaker or listener? the role of a dialog agent. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4861–4869.
- Noboru Matsuda, Victoria Keiser, Rohan Raizada, Arthur Tu, Gabriel Stylianides, William W Cohen, and Kenneth R Koedinger. 2010. Learning by teaching simstudent: Technical accomplishments and an initial use with students. In *International conference on intelligent tutoring systems*, pages 317–326. Springer.
- Ju-Hyun Nam, Myeong-Seob Lim, Hyun-Kyeong Choi, Jae-Yeop Kim, Sung-Kyeong Kim, Sung-Soo Oh, Sang-Baek Koh, and Hee-Tae Kang. 2017. Factors increasing the risk for psychosocial stress among korean adults living in rural areas: using generalized estimating equations and mixed models. *Annals of occupational and environmental medicine*, 29(1):1–12.
- Hien D Nguyen, Vuong T Pham, Dung A Tran, and Trung T Le. 2019. Intelligent tutoring chatbot for solving mathematical problems in high-school. In *2019 11th International Conference on Knowledge and Systems Engineering (KSE)*, pages 1–6. IEEE.
- SeungWon Park and ChanMin Kim. 2015. Boosting learning-by-teaching in virtual tutoring. *Computers & Education*, 82:129–140.
- José Quiroga Pérez, Thanasis Daradoumis, and Joan Manuel Marquès Puig. 2020. Rediscovering the use of chatbots in education: A systematic literature review. *Computer Applications in Engineering Education*, 28(6):1549–1565.
- Fanyi Qu, Xin Jia, and Yunfang Wu. 2021. Asking questions like educational experts: Automatically generating question-answer pairs on real-world examination data. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2583–2593, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Bhavika R Ranoliya, Nidhi Raghuwanshi, and Sanjay Singh. 2017. Chatbot for university related faqs. In *2017 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, pages 1525–1530. IEEE.
- Jun Rekimoto and Hitomi Tsujita. 2014. Inconvenient interactions: an alternative interaction design approach to enrich our daily activities. In *Proceedings of the 2014 International Working Conference on Advanced Visual Interfaces*, pages 225–228.
- Ajith Sowndararajan, Rongrong Wang, and Doug A Bowman. 2008. Quantifying the benefits of immersion for procedural training. In *Proceedings of the 2008 workshop on Immersive projection technologies/Emerging display technologies*, pages 1–4.
- Sascha Stollhans. 2016. *Learning by Teaching. Developing Transferable Skills*. Research-publishing. net.
- Mimi Wagner and Ann Gansemer-Topf. 2005. Learning by teaching others: A qualitative study exploring the benefits of peer teaching. *Landscape journal*, 24(2):198–208.
- John Wagster, Jason Tan, Yanna Wu, Gautam Biwas, and Dan Schwartz. 2007. Do learning by teaching environments with metacognitive support help students develop better learning behaviors? In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 29.

A Learning Session

A brief orientation followed to explain a learning tool (i.e., StuBot or Baseline) and to let the participants practice it at the beginning. After the orientation, the participants received a material and began the study. We set the learning time of each session to the same length of 15 min, including 7 min to type the first explanation text, by referring to other study settings (Fiorella and Mayer, 2014; Blunt and Karpicke, 2014).

After the participant's first explanation, the following procedure differed slightly depending on the learning tool of the session. When the participants used Baseline, they could use all the remaining session times to review the content according to the agent's exam test results. However, with StuBot, additional explanations may be required according to StuBot's feedback. Once all the feedback messages by StuBot had been resolved, the participants were allowed to use the remaining time for a review in the same way as with Baseline.

B Measurements

B.1 Prior knowledge

The prior knowledge was measured by a quiz, consisting of 10 multiple-choice questions about general Korean history (e.g., "*Which of the following best describes the period in which this genre painting was drawn?*"), with a time limit of 5 min. The questions were selected from the same textbook we used for the learning content. All the answers were scored on a 7-point scale and averaged.

B.2 Learning performance

First, seven multiple-choice questions asked the participants to choose one of the five answers that best fit a given statement (e.g., "*Which is not an appropriate explanation about the Silhak?*"). Second, four descriptive questions asked the participants to explain specific historical concepts or events (e.g., "*Describe at least two roles of the Confucian academy.*"). The participants were required to finish the multiple-choice questions and descriptive questions in 10 min. Both questions were selected from the same textbook that we used as the learning material. Finally, a comprehensive question asked the participants to summarize the learning content in a paragraph (e.g., "*Please summarize what you have learned in this chapter.*").

The participants' answers to the quiz questions were scored for each question type on a 7-point

scale. First, the multiple-choice questions were graded according to the answer sheet in the textbook, and the number of correct answers was used as a score.

Next, referring to (Edmunds, 2006), we recruited external human judges to grade the descriptive and comprehensive questions. The judges individually scored the participants' answers to the descriptive questions on a 7-point scale by referring to the learning materials and answer sheets. The individual scores were averaged to obtain the final score for the descriptive question. Grading a comprehensive question was performed more thoroughly because it did not have an answer sheet, unlike the multiple-choice and descriptive questions from the textbook. Two judges collaborated to create a guideline for grading according to the learning material. The guideline contained 15 key topics and possible variants that should be included in a summary. The participants earned 7 points for the key topic if the submitted summary included it, and all the points were averaged as the final score for a comprehensive question.

B.3 Immersion

The participants' immersion in learning was measured by slightly customizing the immersion questionnaire about the game (Jennett et al., 2008). The original questionnaire consisted of 31 five-point Likert scale questions, including six factors: basic attention, temporal dissociation, transportation, challenge, emotional involvement, and enjoyment. We excluded the game-specific items (e.g., "*To what extent was your sense of being in the game environment stronger than your sense of being in the real world?*").

As a result, our immersion questionnaire contained four questions for each of basic attention (e.g., "*To what extent did the StuBot hold your attention?*"), challenge (e.g., "*Were there any times during the StuBot in which you just wanted to stop?*"), and emotional involvement (e.g., "*To what extent did you feel emotionally attached to the StuBot?*"); three for each of temporal dissociation (e.g., "*To what extent did you notice events taking place around you?*") and transportation (e.g., "*To what extent did you feel that you were interacting with the StuBot?*"); and two for enjoyment (e.g., "*How much would you say you enjoyed using the StuBot?*"). The total score was calculated by averaging the participants' responses to 20 questions.

C Failure Cases

Sender	Message
StuBot	“오 그런 내용이구나! 그러면, 국학에 대해서 좀 더 쉽게 풀어서 설명해주세요 ㅠ ㅠ”
User	“국학은 그동안의 중국 중심의 세계관에서 벗어나 우리나라의 가치에 대해 더욱 깊이 연구하는 학문을 말해.”
StuBot	“ 쌤 ㅠ ㅠ 좀 더 자세하게 설명해주실 수 있나요??”
User	“신경준은 ‘훈민정음운해’를 편찬하기도 했고 김정호는 우리나라 지리에 대해 자세히 묘사한 지도인 ‘대동여지도’를 편찬하기도 했어. 이종환이 인문학적 지리서인 ‘택리지’를 편찬하기도 했지. 그만큼 중국적 가치관에서 벗어나 우리나라에 대한 연구를 더 활발히 진행하기 시작했다는 거야.”

(a) Participant #6

Sender	Message
StuBot	“ 쌤 감사합니다!! 근데 제가 고증학에 대해서 아직 잘 모르겠는데 설명해주세요 ㅠ ㅠ”
User	“청에서 들어온 학문의 갈래중에 하나인데 우리 사회의 문제를 객관적으로 해결하자라는 학문이야.”
StuBot	“ 쌤 죄송한데요 ㅠ ㅠ 아직 이해가 안돼요.. 이것에 대해 알고 있는 내용이 또 있으시면 알려주세요..”
User	“예를 들어서 고증학은 우리 사회가 직면한 문제 우리는 농촌사회 였기 때문에 농업에 관해 가장 빨리 문제가 생겼고 이것을 해결하기 위해 토지 분배 방법등 다양한 해결 방안이 제시된것처럼 우리의 문제를 객관적으로 바라보고 해결하는 학문이 고증학이야.”

(b) Participant #10

Table 4: Original texts of the dialogues presented in Table 3