



저작자표시-비영리-변경금지 2.0 대한민국

이용자는 아래의 조건을 따르는 경우에 한하여 자유롭게

- 이 저작물을 복제, 배포, 전송, 전시, 공연 및 방송할 수 있습니다.

다음과 같은 조건을 따라야 합니다:



저작자표시. 귀하는 원저작자를 표시하여야 합니다.



비영리. 귀하는 이 저작물을 영리 목적으로 이용할 수 없습니다.



변경금지. 귀하는 이 저작물을 개작, 변형 또는 가공할 수 없습니다.

- 귀하는, 이 저작물의 재이용이나 배포의 경우, 이 저작물에 적용된 이용허락조건을 명확하게 나타내어야 합니다.
- 저작권자로부터 별도의 허가를 받으면 이러한 조건들은 적용되지 않습니다.

저작권법에 따른 이용자의 권리는 위의 내용에 의하여 영향을 받지 않습니다.

이것은 [이용허락규약\(Legal Code\)](#)을 이해하기 쉽게 요약한 것입니다.

[Disclaimer](#)

Master's Thesis of Engineering

Exploring the effect of a pedagogical conversational agent's warmth on learning

교육용 대화형 에이전트의 따뜻함이 학습에
미치는 영향

August 2023

Graduate School of Convergence Science and
Technology

Seoul National University
Intelligence and Information Major

Sunhyo Oh

Exploring the effect of a pedagogical conversational agent's warmth on learning

Advisor Gahgene Gweon

Submitting a master's thesis of Engineering

August 2023

Graduate School of Convergence Science and
Technology

Seoul National University
Intelligence and Information Major

Sunhyo Oh

Confirming the master's thesis written by

Sunhyo Oh

August 2023

Chair Bongwon Suh (Seal)

Vice Chair Gahgene Gweon (Seal)

Examiner Joonhwan Lee (Seal)

Abstract

Pedagogical conversational agents (PCAs) are beneficial in that they can simulate social interactions, which support students to be emotionally engaged in learning. PCAs can take on various forms depending on the modalities in which they are developed. Text-based conversational agents, in particular, have become increasingly prevalent in educational contexts due to the widespread use of instant messaging by students. However, compared to other modalities that incorporate graphic elements, text-based approaches have not been able to effectively foster social interaction with students. This may be because their primary focus is yet on determining which specific instructional methods to use or which learning content to deliver. Taking this into account, the present study manipulated the level of warmth (low vs. high), which is known to be a dominant dimension of social cognition, to encourage students socially interact with the conversational agent and become emotionally engaged in learning. Although emotional engagement through social interaction can motivate students to learn, it may not be sufficient to promote active engagement in learning. This is because multiple dimensions of student engagement, such as cognitive engagement, are closely interconnected. To fill this void, in this study, we also incorporate manipulation of the type of learning activity that is delivered to students by the agent, varying the level of cognitive engagement that the activity requires (active vs. constructive).

The present study examined the effect of two variables mentioned above, the level of warmth and the type of learning activity. In particular, we measured both objective and subjective variables of the learning experience, learning achievement and intrinsic motivation for learning, and conducted a semi-structured interview with each student to get further insight on the learning experience. We conducted a 2x2 between-subjects laboratory experiment with sixth-grade elementary school students ($n = 98$) with GeomBot, a pedagogical conversational agent that requires students to explain how to solve geometry

problems. Quantitative analyses of experimental results showed that when GeomBot sends high-warm messages to students, they showed significantly greater learning achievement and interest-enjoyment than when GeomBot sends low-warm messages. In addition, when students solve constructive learning activities with GeomBot, they showed significantly greater learning achievement and interest-enjoyment than when solving active learning activities. Furthermore, significant interaction effects for learning achievement, interest-enjoyment, and tension-pressure were observed between the level of warmth and the type of learning activities.

Qualitative analyses of interview data demonstrated two key findings. First, despite researchers making variations on every message sent by GeomBot, students still perceived the content to be repetitive. Second, although students who received high-warm messages showed greater learning achievement, those who received low-warm messages perceived GeomBot as more honest compared to those who received high-warm messages. Interview data from each group indicated that students perceived low-warm GeomBot to be honest when GeomBot reacted negatively to students' feedback after solving a problem incorrectly. In contrast, students rated high-warm GeomBot as less honest because GeomBot responded positively even when it received negative feedback from students. This finding indicates that the current version of high-warm GeomBot did not reciprocate to students' feedback, which may have resulted in a deterioration in the perceived honesty towards the agent. Taking these findings together, we could conclude that (1) automatically generating high-warm and low-warm messages and (2) manipulating the warmth of messages based on students' feedback, which may improve the reciprocity, could improve their agent perception.

To further explore the effect of reciprocal warmth design on agent perception, the second study was conducted with GeomPT. GeomPT automatically generates high or low-warm messages which deliver constructive learning activities, by applying prompting engineering techniques to ChatGPT. We compared two groups ($n = 10$) of sixth-grade students who studied with GeomPT which sends high-warm

messages only (HW-C) and those who used GeomPT that switches the level of warmth of its messages according to students' feedback (HLW-C). In addition to measuring perception-related variables, to investigate students' internal thoughts on GeomPT, we asked students to draw what they thought GeomPT would look like. Our experimental results showed that the high-reciprocity group (HLW-C) showed better perceived reciprocity and perceived honesty than the low-reciprocity group (HW-C). Qualitative analysis of drawing data indicated that students perceived the agent as more human-like, peer-like, and competent when reciprocally responding to their feedback.

The contribution of the study is as follows: (1) We present verbal and non-verbal cues to implement different levels of warmth in the pedagogical conversational agent. (2) We suggest how to apply generative AI, ChatGPT, to educational settings, focusing on the warmth manipulation of the message and constructive activity design. (3) We provide empirical findings on reciprocally using high-warm and low-warm messages with constructive learning activities to foster students' active engagement.

Keyword : pedagogical conversational agent, warmth, cognitive engagement, generative AI, reciprocity

Student Number : 2021-24263

Table of Contents

Chapter 1. Introduction.....	1
1.1. Background and motivation.....	1
1.2. Research Overview	3
1.3. Contribution	4
Chapter 2. Related Work	6
2.1. Pedagogical conversational agent	6
2.2. Warmth as a means of social cognition toward others: one dimension of SCM.....	8
2.3. Technical implementation of the ICAP Framework	9
Chapter 3. Study 1.....	11
3.1. Research Question and Hypothesis	11
3.2. GeomBot Design.....	14
3.2.1. Warmth Manipulation.....	15
3.3. Method	21
3.3.1. Participants.....	21
3.3.2. Measurement	21
3.3.3. Procedure	25
3.3.4. Analysis	26
3.4. Results	29
3.4.1. RQ1: Would the agent’s warm message positively impact learning?	29
3.4.2. RQ2: Would the type of learning activities positively impact learning?	34
3.4.3. RQ3: Would the type of learning activities influence the effectiveness of the pedagogical conversational agent’s warm message on learning?	39
3.5. Discussion.....	40
3.5.1. Overall benefits and limitations of GeomBot.....	40
3.5.2. Warmth of the pedagogical conversational agent	43
Chapter 4. Study 2.....	50
4.1. Background, Research Question, and Hypothesis	50
4.2. GeomPT Design.....	51
4.2.1. ChatGPT Prompting for Message Generation.....	53
4.2.2. Reciprocal Design.....	60
4.3. Method	61
4.3.1. Participants.....	61
4.3.2. Measurement	62
4.3.3. Procedure	63
4.3.4. Analysis	64
4.4. Results	65

4.4.1. RQ4: Would reciprocal warmth design improve perceived honesty towards pedagogical conversational agent?.....	6	5
4.5. Discussion.....	6	9
4.5.1. Use of Generative AI in a pedagogical conversational agent	6	9
4.5.2. Reciprocal design.....	7	2
Chapter 5. Conclusion	7	4
References	7	6
Abstract in Korean	8	7

Chapter 1. Introduction

1.1. Background and motivation

Conversational agents provide promising opportunities for education. They have advantages in improving students' cognitive and motivational learning outcomes (Weber et al., 2021). In addition, they are able to represent different human instructional roles, such as expert, tutor, mentor, and learning companion (Y. Kim & Baylor, 2006). Especially, pedagogical conversational agents (PCAs) as peer students are meaningful in that they can simulate social interactions (Y. Kim & Baylor, 2006) which play an important role in learning (Driscoll, 1994; Locke, 1997; Piaget & Smith, 2013), such as encouraging students to emotionally engaged in learning (Molinillo et al., 2018).

Two common types of pedagogical conversational agent exist depending on the interface being used: embodied and messenger-like (Hobert & Meyer von Wolff, 2019). Embodied PCAs usually incorporate graphical elements, such as virtually represented human characters, which enable social interaction. However, messenger-like, which is also known as text-based, PCAs have yet focused on learning content design and instructional strategies to deliver (Kuhail et al., 2023), rather than fostering social interaction. As text-based PCAs are emerging as effective tools, due to students' widespread usage of instant messaging (Kuhail et al., 2023), there is a need to investigate how to promote social interaction with text-based PCAs to encourage emotional engagement and active learning.

One possible approach is to imbue text-based PCAs with social characteristics, such as warmth. In human-human social interaction, there are two dimensions of social cognition that affect how one human makes sense of another, and the primary dimension is known to be warmth (Fiske et al., 2002). Warmth is defined as the degree to which individuals perceive caring and sociability in the others (W. B. Kim & Hur, 2023), and consists of various sub-dimensions such as friendly,

warm, thoughtful, well-intentioned, generous, and honest (Stanciu et al., 2017). Based on the CASA (Computers Are Social Actors) paradigm (Reeves & Nass, 1996), the same social cognition could be applied when interacting with high or low-warm agent. For example, it has been reported that human forms impressions towards agents differently and makes different decisions depending on the level of warmth of the agent which they interact with, from the context of marketing (Kervyn et al., 2022), workplace (Jung et al., 2022), and gameplay (McKee et al., 2022).

In the same vein, as teaching and learning are highly social activities according to socio-cognitive theories (Y. Kim & Baylor, 2006), learning with agents of different levels of warmth would bring about different learning experiences. In Study 1, we aim to explore whether the impact of warmth exists when students interact with the pedagogical conversational agent as a peer student. Therefore, we manipulated the level of warmth (low vs. high) of the messages that the PCA sends to students. We then measured students' learning experiences, learning achievement and motivation, to investigate which level of warmth would support students to be emotionally engaged in learning and improve their learning experience.

Encouraging only emotional engagement through social interaction would not be sufficient for supporting active engagement because several other dimensions of student engagement, such as cognitive engagement, are closely interconnected. To fill this void, we also incorporated a manipulation of students' cognitive engagement when doing activities with the PCA. We designed two different learning activities with the different required levels of cognitive engagement (active vs. constructive), based on ICAP Framework. We then explored how differently the warmth of the agent affects learning, depending on the type of learning activities, to explore text-based PCA design that can support students to be emotionally and cognitively engaged in learning. The PCA that was used in Study 1 is a rule-based chatbot, with all learning contents and messages manually generated by human researchers.

Although manually generating content is beneficial to achieve instructional goals, it has limitation in providing sufficient variations to the content, despite the significant demand for human resources (Markel et al., 2023). In this regard, rapid technological advances in Large Language Models (LLMs) would enable an automated simulation of various instructional roles. Especially, natural language generation models, such as GPT-3, successfully bring about desired model behaviors with prompting techniques (Jiang et al., 2022; Liu et al., 2023), without additional fine-tuning processes. Based on such advantages, there has been some work on using LLMs in educational settings to develop chatbot tutors (Ruan et al., 2019) and chatbot tutees with intended human characteristics (Markel et al., 2023). Given such potential of AI-generated artifacts in educational settings, there is a need to explore how students perceive AI-generated messages when using pedagogical conversational agents. Therefore, in Study 2, we developed the PCA using LLMs to complement the design of the agent from Study 1. We then suggest guidelines to consider when using natural language generation models in educational settings.

1.2. Research Overview

To propose text-based pedagogical conversational agent designs that can enhance emotional and cognitive engagement in learning, in Study 1, we examined the impact of two variables, the level of warmth (high vs. low) and the type of learning activity (constructive vs. active), on two aspects of learning experiences: (1) learning achievement and (2) intrinsic motivation for learning, using GeomBot. The study results indicated that High-Warm messages and Constructive learning activity had a significantly positive impact on learning achievement and motivation, and significant interaction effects also existed between the two variables. Two main insights were observed from the qualitative analysis, which was firstly, the perceived repetition and low adaptivity of manually generated messages and second, low reciprocity of high-warm GeomBot.

To address the above-mentioned two insights from Study 1, in Study 2, we compared the original HW-C agent with HLW-C agent with improved reciprocity to verify the insight from Study 1. In addition, we automatically generated, in-real time, the messages that both agents send, using ChatGPT to complement the repetition and low adaptivity. The results of Study 2 indicated the reciprocal warmth design which was derived from Study 1 could improve the perceived reciprocity and honesty. We then discussed the use of generative models in educational settings and provided design guidelines for future studies regarding pedagogical conversational agents.

The rest of this paper is organized as follows: Section 1.3 presents contribution points of this study. Section 2 reviews previous studies on the pedagogical conversational agent, warmth as a means of social cognition toward others, technical implementations of the ICAP framework. Section 3 introduces Study 1, each subsection consisting of research questions, PCA design, 2 x 2 between-subjects experiment, result, and discussion. Section 4 presents Study 2 and the subsections are as follows: research questions, PCA design with the use of generative models, experimental design, result, and discussion. The study is then concluded in Sections 5.

1.3. Contribution

Based on the experimental study, our research yields the following three contributions: (1) We present verbal and non-verbal cues to induce different levels of warmth that can be implemented in the pedagogical conversational agent. (2) We provide guidelines to consider when using Large Language Models (LLMs) for the learning content generation, focusing on the warmth manipulation of the message and constructive learning activity design. (3) Based on the empirical results of the interaction between agent warmth and learning activity type, we suggest design implications to foster students' active engagement, for future pedagogical conversational agent designers. (4) Based on the findings regarding reciprocal warmth design, we suggest design guidelines on instructional conversation

flow, to reciprocally apply one of the social aspects, warmth, to the pedagogical conversational agent.

Chapter 2. Related Work

Our study aims to investigate how the design of a pedagogical conversational agent as a peer student, i.e., the warmth of the agent and the type of learning activities, impact the learning experience of students. In section 2.1, we will introduce and examine previous research on the pedagogical conversational agent, focusing on the type of interface. Warmth as a means of social cognition toward others will be explored in section 2.2. We then review how the ICAP Framework is technically applied to various mediums in section 2.3.

2.1. Pedagogical conversational agent

A pedagogical agent is an anthropomorphic agent used in an online learning environment for the sake of instruction (Martha & Santoso, 2019). Especially, pedagogical agents have been recommended to have a human-like persona (Y. Kim & Baylor, 2006), based on existing research that emphasizes social interaction with peers in the classroom (Driscoll, 1994; Locke, 1997; Piaget & Smith, 2013). Such pedagogical agents are called 'pedagogical agents as a learning companion', which we define as an agent that simulates peer interaction in computer-based learning. One of the most commonly used interfaces is graphic-based, which accompanies animated peer-like characters (Ba et al., 2021; Domagk, 2008; Y. Kim et al., 2006; Y. Wang et al., 2023). For example, Kim's study examined the effect of competency and interaction type of pedagogical agents as learning companions and indicated the main effect of both variables on learning and motivational outcomes (Y. Kim et al., 2006). In addition, Domagk's study investigated that when showing a specific appearance, likable agents led to a higher learning motivation (Domagk, 2008).

Another emerging interface is text-based, and such pedagogical agents are called Pedagogical Conversational Agents. The pedagogical

conversational agent is also a sub-class of Conversational Agents, which provides students with interactive learning experiences in their natural language (Weber et al., 2021). There exist two common types of pedagogical conversational agents from a technical perspective, embodied conversational agents and messenger-like agents (Hobert & Meyer von Wolff, 2019). Embodied pedagogical conversational agents, similar to graphic-based pedagogical agents, include virtual representations of human characters or avatars. However, unlike graphic-based pedagogical agents, messenger-like pedagogical conversational agents communicate with students via text or voice. For example, Noh's study introduced a pedagogical conversational agent that provides a museum experience with embodied and reflected historical information. The result of the study indicated that the chatbot with embodiment and reflection enhanced the museum experience (Noh & Hong, 2021).

Messenger-like pedagogical agents use common chatting interfaces, such as chatbots. As messengers are already easily used by students, messenger-like pedagogical agents are widespread nowadays, whereas embodied agents were prevalent in the past (Hobert & Meyer von Wolff, 2019). Yin's study compared students who learned in traditional school settings with students who learned through interaction with a chatbot, with the latter group showing higher intrinsic motivation for learning (Yin et al., 2021).

Manipulating the property or persona of the pedagogical agents is relatively common for graphic-based pedagogical agents and embodied pedagogical conversational agents due to the existence of embodied reality (Ba et al., 2021; dos Santos Alencar & de Magalhães Netto, 2020; Guo & Goh, 2016; Lawson & Mayer, 2022; Liew et al., 2017; Y. Wang et al., 2023). However, to the best of our knowledge, there was less emphasis on manipulating the characteristics of the agent itself with verbal and non-verbal cues in a text-only environment. An example is Ceha and colleagues' study that examined

the effect of pedagogical conversational agents' use of two types of humor, one of the socially-oriented conversational strategies (Ceha et al., 2021). The study results demonstrated that affiliative humor significantly increased motivation and effort, while self-defeating humor negatively impacted enjoyment. Given the wide use of simple chatting interfaces in educational settings (Hobert & Meyer von Wolff, 2019), it is imperative to explore methods for adjusting the message of the chatbot to represent a specific property. Therefore, in this study, we manipulated the message of the pedagogical conversational agent verbally and non-verbally in the text-only environment.

2.2. Warmth as a means of social cognition toward others: one dimension of SCM

Warmth is one dimension of human evaluation toward others' impressions, which stems from the Stereotype Content Model. According to the Stereotype Content Model, humans judge other humans in social interaction using two dimensions: warmth and competence (Fiske et al., 2002). With emerging interest in human-agent interaction, there have been several trials to apply these two properties to the agent so that humans can have social cognition toward the agent. For example, Oliveria and colleagues employed different levels of warmth and competence in the display of a robot (Oliveira et al., 2019). The study result indicated that the different levels of warmth and competence are related to emotional responses from participants. In addition, Nguyen's study presented a design methodology to reflect varying degrees of warmth and competence to virtual characters through gestures and gaze behaviors (Nguyen et al., 2015).

Especially in the text-based agent, due to the difficulty of applying warmth to the agent itself, previous research relied on providing cover stories or metaphors of the agent to intervene in human perception, prior to the usage of the agent. For example, Gilad and colleagues

provided participants with descriptions of the agent, such as a “state-of-the-art artificial neural network algorithm that was trained on data from 1,000,000 houses” for a high-competence agent and “a system that help people make better offer” for a high-warmth agent (Gilad et al., 2021). Meanwhile, Khadpe’s study identified a set of metaphors that correspond to different levels of warmth and competence via crowdsourcing (Khadpe et al., 2020). Such metaphors are used for conversational Human-AI collaboration tasks and participants showed the desire to cooperate with agents with higher warmth and competence.

However, to the best of our knowledge, less emphasis was on applying such properties directly to the chatting interface by manipulating the agent’s utterance itself. Between the two dimensions, warmth and competence, warmth is known to be primary, being judged before competence and carrying more weight in behavioral interaction (Fiske et al., 2007). In addition, warmth is associated with the key dimensions of trust (Zahry & Besley, 2021) and is known to influence whether or not to trust others (Cuddy et al., 2008). For example, when playing a cooperative game with a computer, perceiving warmth in a virtual agent positively influence behavioral trust and perceived trustworthiness (Kulms & Kopp, 2018). Therefore, in this study, we present verbal and non-verbal cues to apply ‘warm’ properties to a text-only environment.

2.3. Technical implementation of the ICAP Framework

ICAP Framework (Chi & Wylie, 2014) proposes that students’ behavior that reflects cognitive engagement can be categorized into one of four levels: Interactive, Constructive, Active, and Passive. What this framework emphasizes the most is that regardless of the task itself, the teacher can scaffold students to behave or act at a particular

engagement level and students' overt behavior would show whether the scaffolding worked or not. Based on this novel approach, there has been prior research that applied this framework to real-world instructional applications, such as massive online courses (McNeill et al., 2019), mobile gamified apps (Ha et al., 2021), and online forums (Q. Wang et al., 2022).

According to the framework, learning outcomes should get better as the level of cognitive engagement increases from Passive to Interactive. However, some prior research that applied ICAP showed inconsistent results depending on learning platforms. For example, in online courses, the effect of teacher presence does not fit the ICAP framework of observable student engagement behaviors, activities that require more cognitive engagement returning fewer total course hours (McNeill et al., 2019). In digital learning games, active activities led to better learning than constructive activities because active activities are less disruptive to game flow (Johnson & Mayer, 2010). These inconsistent results imply that there is a need to examine the effect of each engagement behavior in different learning platforms. To the best of our knowledge, there has been no trial to apply ICAP to pedagogical conversational agents, where learning happens only through instant conversation between students and the agent (Smutny & Schreiberova, 2020). Therefore, in this study, we implement constructive and active activity in a conversational agent to examine whether each activity works differently in the setting of the pedagogical conversational agent.

Chapter 3. Study 1

3.1. Research Question and Hypothesis

In Study 1, we first examined the impact of two factors, the warmth of the agent's message and the type of learning activities, on learning achievement and intrinsic motivation for learning. For the effect of the warmth of the agent's message, we posed student trust in the peer agent as a mediator on the learning. The interaction effect between with agent warmth and learning activity type on learning was also explored. The above-mentioned issues will be addressed in Research Questions 1,2, and 3, respectively.

RQ1. Would the agent's warm message positively impact learning?

In <Research Question 1>, we explored the impact of the warmth of pedagogical conversational agent's messages on learning outcomes. According to the literature that discussed warmth from the perspective of Stereotype Content Model (SCM), trust is a sub-dimension of warmth, and warmth is a prime factor that influences trust (Cuddy et al., 2008; Fiske, 2018; Zahry & Besley, 2021). While trust can be defined differently in multiple settings, in the context of education, trust is defined as a willingness to be vulnerable to another party on the confidence that the latter is benevolent, reliable, competent, honest, and open (Tschannen-Moran & Hoy, 2000).

The school evidence indicates that trust enhances school performance, supporting cooperation between subjects (Tschannen-Moran, 2014). Especially, as peer relationship is a major indicator of social growth, learning engagement, and academic achievement (Wentzel, 2017), trust in school peers would have a positive impact on school performance. Adam's study which validated a measure of student trust in school peers investigated a positive relationship between peer trust and optimal school functioning such as academic

grit (Adams et al., 2022). Taking the evidence from the relationship between warmth and trust and learning, we suggest that learning with warm pedagogical agent would bring about a positive influence on learning achievement and intrinsic motivation by students perceiving trust toward the agent.

H1: Students' learning will be better when the pedagogical conversational agent's messages are high-warm than when its messages are low-warm.

By connecting the two pieces of research, Stereotype Content Model (SCM) and School Trust, on the same line via the trust variable, we expect that the trust variable would better explain the relationship between the warmth of the pedagogical conversational agent and learning outcomes as a mediator variable.

H1-1: Students' trust in the pedagogical conversational agent will be higher when its message is high-warm than when its message is low-warm.

H1-2: Student perception of trust in the pedagogical conversational agent will positively impact their learning.

H1-3: Student perception of trust in the pedagogical conversational agent will mediate the effect of the warmth of the agent's message on learning.

RQ2. Would the type of learning activities positively impact learning?

In <Research Question 2>, we investigate whether the type of learning activities impact learning. In this study, we compare two types of learning activities, active and constructive, designed based on the ICAP Framework (Chi & Wylie, 2014). ICAP Framework is a learning

theory that categorizes learning activities into one of four modes, Interactive, Constructive, Active, and Passive. The most important factor that distinguishes each mode is how cognitively engaged students are in those learning activities. The framework hypothesized that students would learn more as they become more cognitively engaged with the learning activities, from passive to active to constructive to interactive. This hypothesis has been empirically validated by most studies, for a wide range of learning outcomes from learning achievement to learning motivation. Therefore, we expect a similar result in the context of learning with the pedagogical conversational agent, doing constructive activities yielding better learning, in terms of learning achievement and intrinsic motivation for learning, than doing active activities.

H2: Students' learning will be better when students do constructive activities with the pedagogical conversational agent than when they do active activities with the pedagogical conversational agent.

RQ3. Would the type of learning activities influence the effectiveness of the pedagogical conversational agent's warm message on learning?

In <Research Question 3>, we investigate whether the impact of warmth of the agent message on learning differs by the type of activities. Which of these two types of learning activities students do with the pedagogical conversational agent might influence the effectiveness of the agent's warm messages on the learning outcomes. The tasks that students do in each type of learning activity are designed following existing ICAP research (Chi & Wylie, 2014; Wylie & Chi, 2014). When answering the agent's question, students who participate in active activity merely choose one option from a multiple-choice list, while students who are in constructive activity write their thoughts on their own. As such, compared to when doing active activities, which do not accompany self-expression, when doing

constructive activities that accompany writing down one's own thoughts, students' learning would be more influenced by the message of the agent, along with increased cognitive engagement. When the message of the agent is highly warm, students would feel that the agent is supporting their learning (Carbajal et al., 2016), compared to when the message of the agent is less warm. Therefore, we expect learning outcomes from constructive activity to be better when the message of the agent is high-warm than when the message of the agent is low-warm.

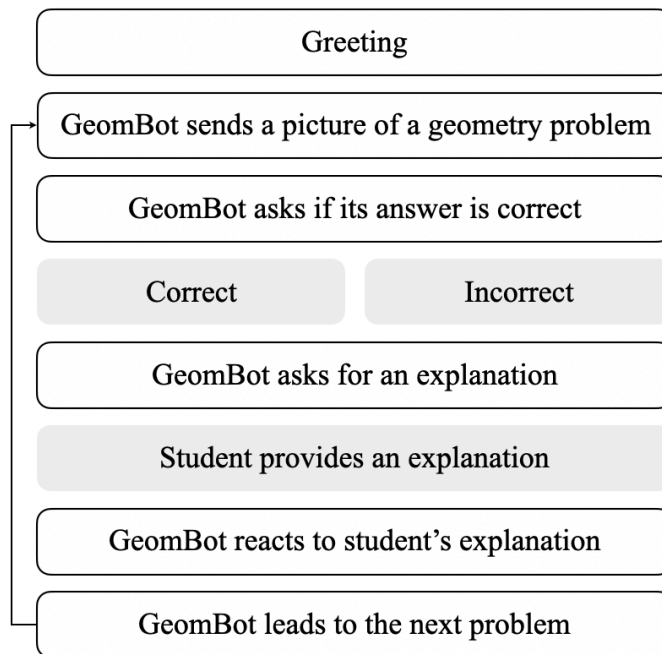
H3: When doing constructive activities, students' learning will be better when the messages of the pedagogical conversational agent are high-warm than when the messages are low-warm.

3.2. GeomBot Design

GeomBot is a Telegram-implemented pedagogical conversational agent that provides geometry problems regarding the perimeter and area of polygons. We implemented four versions of GeomBot, using two factors, warmth of the agent's messages and the type of learning activities: HW-C (High-warm messages and Constructive activity), LW-C (Low-warm messages and Constructive activity), HW-A (High-warm messages and Active activity), LW-A (Low-warm messages and Active activity).

The conversational flow with GeomBot is as follows: (1) GeomBot brings a problem to the chatting interface. (2) GeomBot first solves a problem and asks students if its answer is correct or not. (3) Students give GeomBot feedback by clicking one of 'correct' or 'incorrect' buttons. (3) After receiving feedback, GeomBot asks students to explain the reason why students think that its answer is either correct or incorrect. (4) After receiving the explanation, GeomBot brings the next problem. Figure 1 illustrates a batch of problem-solving

interactions between GeomBot and the student.



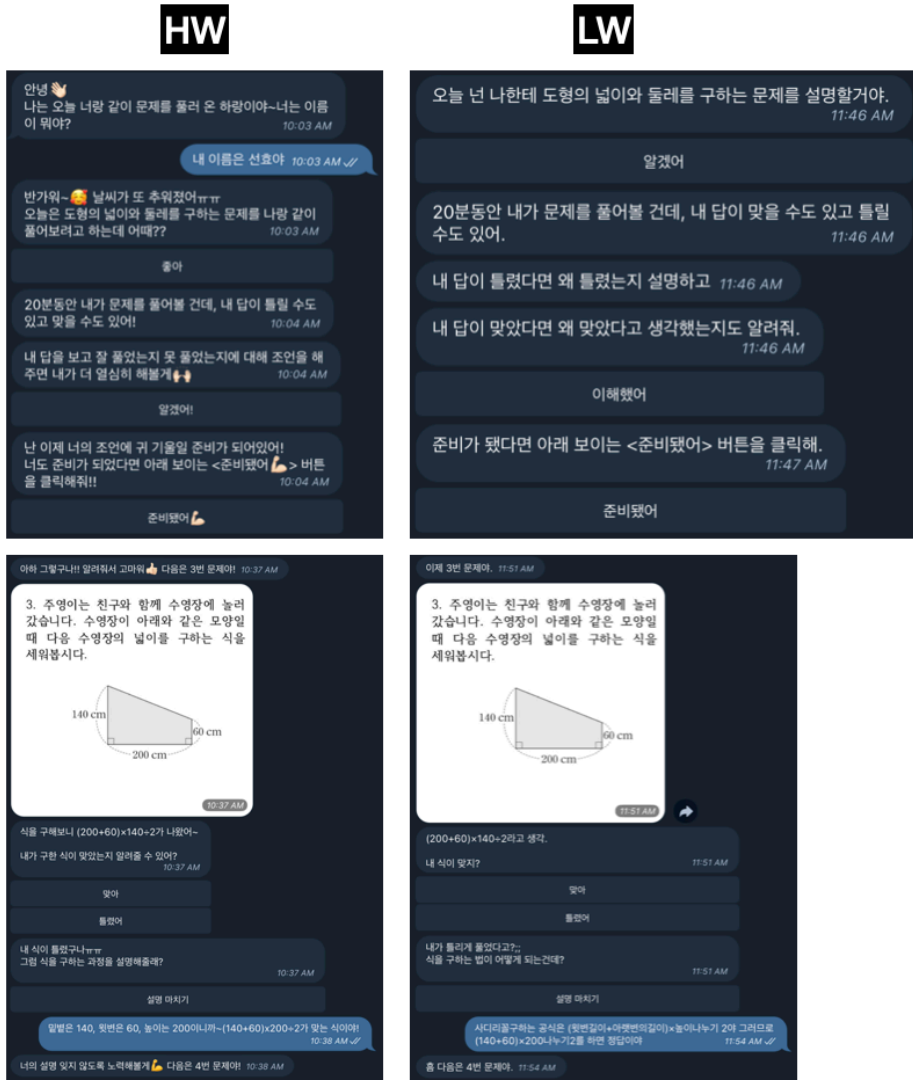
[Figure 1. Conversation flow with GeomBot]

Students can solve up to 35 problems in 20 minutes. There is no need to solve all problems and students solve only as much as they can in the given 20 minutes. Across all four conditions, GeomBot's geometry problem-solving skill is fixed at 80% of accuracy, solving problem 3, 6, 10, 12, 15, 17, and 21 incorrectly. When designing the incorrect answers, we referred to the type of mistakes that students make on real-life math exams.

3.2.1. Warmth Manipulation

To examine the effect of GeomBot's warm messages, we designed two different versions of messages: high-warm and low-warm. From previous research with regard to human or agent warmth, we reviewed and extracted verbal and non-verbal cues that can be applied to the messages of the conversational agent. We then systematically

manipulated the agent’s messages to represent warmth verbally and non-verbally in a text-only environment. To effectively minimize the effect of confounding variables between the two versions, we first wrote the conversation flow for the neutral version of GeomBot messages, which was used for the orientation session. We then adapted the messages for high-warm and low-warm versions of GeomBots, according to the cues of warmth that we reviewed. Figure 2 illustrates a part of the conversation with a high-warm message version of GeomBot and a low-warm message version of GeomBot.



[Figure 2. HW vs. LW, focusing on ice-breaking and reaction to

Verbal and non-verbal cues of warmth. Among various behavioral cues that can represent whether a person is warm or cold, we extracted five cues that can be verbally and non-verbally applied to the chatting interface. The following cues can influence each other and elicit higher warmth when used together.

- (1) **Smiling** is the most frequently mentioned characteristic of a warm person or agent (Bayes, 1972; Biancardi et al., 2017b, 2017a; Cuddy et al., 2011; Gorham, 1988; Reece & Whitman, 1962). The smiley faces of both humans and agents have been reported to be closely related to the warmth rating of the other party (Bayes, 1972). This smiling feature can be non-verbally applied to the chatting interface as smiley face emojis, such as 😊, 😄, and 😁.
- (2) **Hand gestures** are also known to have a close link with social perception (Maricchiolo et al., 2009). By using hand gestures, a person or agent can convey warmth and show positive interest toward the other party (Bayes, 1972; Biancardi et al., 2017a; Cuddy et al., 2011; Maricchiolo et al., 2009; Pace & Gnisci, 2019). Hand gestures can be non-verbally applied to the chatting interface by some emojis such as 🤝, 🙌, and 💪.
- (3) **Calling names** can influence perceived warmth in interpersonal relationships. Education studies indicated that interest in learning names is one of the warm traits (Best & Addison, 2000; Carbajal et al., 2016). Specifically, a person introducing himself/herself by name and calling others' by their name evokes others' perception of warmth toward him/her (Howe et al., 2019). In the chatting interface, the subjects of the conversation can verbally call each other by name.
- (4) **Effort to agree and understand** is one of the behaviors that represent warmth (Bordin, 1951; Li et al., 2012). In a similar vein, one being supportive is rated to be warm (Carbajal et al.,

2016; Li et al., 2012; Seibel, 1955). Agreement, understanding, and support can be represented either verbally or non-verbally on the chatting interface.

- (5) **Positive statements** about the other party are one of the predictors that are most closely related to the warmth ratings (Bayes, 1972). In educational settings, making a positive appraisal of the feedback increased perceived warmth (De Sixte et al., 2020). The subjects of the conversation can verbally and non-verbally make positive remarks.

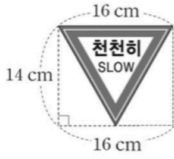

High-Warm messages. A high-warm GeomBot is supportive, thankful for the student's feedback and explanations, and attributes its achievement to the student's help. It starts conversations with a friendly greeting and handshaking emoji 🤝 (cue 2), introducing itself as 'Harang' and asking for the student's name (cue 3). Before GeomBot starts solving geometry problems, it explains where it needs help from the student and asks if the student is willing to do so, with some cheering and grateful messages (cue 2, 4) (e.g. "Even if I made a mistake, please let me know how to fix it. I'd really appreciate it 🙏", "I'll try my best to understand your explanations! 🙌").

After GeomBot solves geometry problems, it politely asks for feedback and explanations (e.g. "I think the answer is $(200+60) \times 140 \div 2$! Can you tell me if my answer is correct?", "Oh my answer is wrong 😞 If you tell me how the answer came out, I think it will help me a lot!"). Regarding the student's feedback and explanations, GeomBot appreciates the student's help and shows an effort to agree and understand the explanations (e.g. "I'm enjoying solving problems thanks to your help 😊" (cue 1, 5), "I'll try to remember your explanation 🙌" (cue 2, 4), "That's right, I think the same as you!" (cue 4)).

Low-Warm messages. A low-warm GeomBot is discouraging and just demands the student to just give feedback and explanations without any reaction to the student's help. A low-warm GeomBot is designed not to express any verbal or non-verbal cues of warmth. It starts conversations with a heartless greeting without any emotional expressions (no cue 1, 2), not even introducing itself (no cue 3). Before GeomBot starts solving geometry problems, it just lists what the student needs to do without asking if the student is willing to do so (no cue 2, 4) (e.g. "If I made a mistake, explain why my answer is wrong.").

After GeomBot solves geometry problems, it rudely demands feedback and explanations as if it thinks it solved all problems right (e.g. "My answer is $40 \times 50 \div 2$. It's correct, right?", "I did it wrong? Then explain the right one."). Regarding the student's feedback and explanations, GeomBot seems not agreeing to the student's explanations and does not react to the student's help (no cue 1, 2, 4, 5) (e.g. "... I'm moving on to the next question.", "Hmm...").

3.2.2 Learning Activity Design

Constructive	Active
<p>1. 윤희가 교통안전교육을 받고 그림과 같은 교통 표지판을 만들었습니다. 윤희가 만든 교통 표지판의 넓이를 구하는 식을 세워봅시다.</p>  <p>와 맞았다!! 😊 식을 구하는 과정을 설명해줄래? 10:04 AM</p> <p>설명 마치기</p> <p>밑변은 16, 높이는 14니까 너가 세운 식이 맞아 10:04 AM ✓</p>	<p>1. 윤희가 교통안전교육을 받고 그림과 같은 교통 표지판을 만들었습니다. 윤희가 만든 교통 표지판의 넓이를 구하는 식을 세워봅시다.</p>  <div style="border: 1px solid black; padding: 5px; margin: 10px 0;"> <p>교통 표지판의 넓이 $= (\text{밑변의 길이}) \times (\text{높이}) \div 2$ </p> </div> <p>와 맞았다!! 😊 식을 구하는 과정을 설명해줄래? 10:03 AM</p> <p>밑변은 16, 높이는 14이기 때문에 식을 세워보면, $16 \times 14 \div 2$야</p> <p>밑변은 16, 높이는 16이어서 식을 세워보면, $16 \times 16 \div 2$야</p> <p>밑변은 16, 높이는 14이기 때문에 식을 세워보면, 16×14야</p> <p>밑변은 14, 높이는 16이기 때문에 식을 세워보면, 14×16이 돼</p>

[Figure 3. Active vs. Constructive, focusing on question type and explanation type]

To examine the effect of learning activity, we designed two different types of learning activities: constructive learning activity and active learning activity. These two types of learning activities stem from ICAP (Interactive, Constructive, Active, Passive) framework. There are two main differences in the activities given to Constructive group and Active group: 1) question type, 2) explanation type. Figure 3 represents how differently chatting interfaces are designed for Active learning activity and Constructive learning activities.

Active learning activity. In terms of the type of question, students are said to be actively engaged when they can solve a problem using only the information given in the problem. Therefore, when designing active learning activities, it should be noted that students can solve the problems if they plug-and-chug the information given in the problems. For example, solving a math problem can be an active activity when the student can get the right answer by merely applying a given formula of equation (Chi & Wylie, 2014). Therefore, we included the formula of the equation in the problem so that students can solve it by just applying the given numbers to the given formula.

Regarding explanation type, we referred to a literature that categorized many forms of self-explanation into one of ICAP activities (Wylie & Chi, 2014). Menu-based self-explanation, where the student selects the right answer from a multiple-choice list, is one of the active self-explanation methods. Applying this menu-based self-explanation to our study, GeomBot provides the student with four options to choose from, and the student gives GeomBot explanations by selecting one of those four options.

Constructive learning activity. In terms of the type of question, students are said to be constructively engaged when they can solve a problem by expanding what was provided in the given material. Therefore, when designing constructive learning activities, we need to

expect that the student will go beyond the given information in the problem and generate new content on their own. For example, math problem solving can be categorized as constructive if the student has to rederive an equation to get the answer (Chi & Wylie, 2014). Therefore, we provided only polygons of which the student need to calculate the area and perimeter so that students can infer how to formulate an equation.

With regard to the explanation type, we adopted an open-ended self-explanation method, where the student needs to generate explanations on their own (Wylie & Chi, 2014). Applying this open-ended self-explanation to the constructive version of activities, GeomBot asks the student to write explanations by himself/herself. The student then explains the process of getting the answer by typing their thoughts directly on Telegram.

3.3. Method

3.3.1. Participants

Upon approval from the Institutional Review Board at Seoul National University (IRB No. 2211/002-024), we recruited 121 students from six elementary 6th-grade classrooms in South Korea. In exclusion of students who (1) were unexpectedly absent on the day of the experiment and (2) got all the pre-test questions right, a total of 98 students (49 girls; 49 boys) participated in the study. The students were randomly assigned to one of four conditions. Per condition, the following number of participants were assigned: 32 in HW-C, 23 in HW-A, 23 in LW-C, and 20 in LW-A. Across the four conditions, there was no statistical difference in pre-test score, Affinity for Technology Interaction (ATI), and mathematics-related affect. We provided \$10 to each student who participated in the study.

3.3.2. Measurement

3.3.2.1. Control Variables

Considering that the task that students need to perform is to use a chatbot via Telegram and solve geometry problems, there was a need to control the following variables across all four conditions: (1) Affinity for Technology Interaction, (2) Mathematics-related affect, (3) geometry problem-solving skills.

Affinity for Technology Interaction (ATI). ATI measures the tendency to actively engage in intensive technology interaction (Franke et al., 2019). We used a Korean version of ATI scale as a pre-questionnaire to control the level of experience in technological interaction across the conditions. This six-point Likert scale consists of nine items (e.g., *"I like to occupy myself in greater detail with technical systems."*) and yields a total score between 1 and 6 points.

Mathematics-related affect. Mathematics-related affect (Hannula, 2012) is measured with a Korean version of a five-point Likert-scaled questionnaire that is validated and translated for elementary school students (Do & Paik, 2017). Under the cognitive dimension, there are three sub-dimensions (Tuohilampi et al., 2015), which are self-competence (4 items; sample item: "I have done well in mathematics"), self-confidence (4 items; sample item: "I am sure that I can learn math"), and difficulty of mathematics (3 items; sample item: "Mathematics is difficult"). The emotional dimension refers to enjoyment of mathematics (5 items; sample item: "I have enjoyed pondering mathematical exercises"). The sub-dimension of the motivational dimension includes mastery goal orientation (5 items; sample item: "In every lesson, I try to learn as much as possible") and effort (4 items; sample item: "I always prepare myself carefully for exams").

Geometry problem-solving skills. We created a pre-test to measure the student's geometry problem-solving skills prior to the experiment. A pre-test is a hard-copied exam paper with a total of twelve problems that consist of the same type of problems used in the activities with GeomBot. The first half of the test includes problems with writing down formulas to find the area of triangles, quadrilaterals, parallelograms, trapezoids, and rhombuses, and perimeters of regular polygons. The last of the test consists of problems with formulating equations directly from figures. Each of the twelve problems is worth one point, yielding a total of twelve points.

3.3.2.2. Dependent Variables

Perceived warmth. A seven-point Likert scale that consists of six statements with warmth-related adjectives was used to measure how warm the students felt the message of the pedagogical conversational agent. From frequently-used warmth-related adjectives and other adjectives that are used in other studies that measured perceived warmth (Stanciu et al., 2017), we decided to use the following adjectives: *friendly, warm, thoughtful, well-intentioned, generous, and honest*. All six items start with the phrase *"I think the message of GeomBot is..."* and each adjective completes a sentence. After adding up the points of all items, the total points for perceived warmth are between 7 and 42.

School trust. To measure students' perception of trust in peers, we used the Korean version of the four-point likert scale questionnaire that Adams (2022) constructed and validated to measure student trust in school peers (Adams et al., 2022). Trust in various school role-relationships (e.g., peer-peer, teacher-student, etc.) consists of five facets: perceived benevolence, competence, openness, honesty, and reliability, each of which has two items. We changed the

object of the items *from students to GeomBot* to set the context of the survey as studying with the pedagogical conversational agent as a peer student. Sample items of each facet are, *"I think GeomBot is eager to help each other with me"*, *"I think GeomBot learns a lot from me"*, *"I think GeomBot really listens to me"*, *"I think GeomBot can believe what I tell him/her"*, and *"I think GeomBot does what he/she is supposed to do"*. A total trust score is calculated as the sum of points for ten items, with a minimum of 10 points and a maximum of 40 points (Maele, 2011).

Learning achievement. We measured learning achievement in geometry problem-solving by calculating the difference between the scores of the tests taken before and after the experiment. We followed the measurement from the previous studies that measured learning achievement by subtracting the pre-test score from the post-test score. A post-test is a hard-copied exam paper, similar to the pre-test, with a total of twelve problems that consist of the same type of problems used in the activities with GeomBot. However, the numbers used in problems in the post-test and the order of the problems are set to be different from those in the pre-test to reduce the memory effect and practice effect.

Intrinsic motivation for learning. We adapted from Intrinsic Motivation Inventory (IMI) (McAuley, 1989) that is modified by Yin (2021) to measure the subjective experience of intrinsic motivation related to the specific learning environments of the study. This seven-point Likert scale consists of five dimensions, interest-enjoyment, tension-pressure, perceived choice, perceived competence, and perceived value. Considering the experimental condition, two items of perceived choice and one item of competence were deleted, and each dimension consisted of 7, 5, 2, 4, and 4 items. Sample items of each dimension are, *"Explaining how to solve geometry problems to*

GeomBot was fun”, “I felt pressured while explaining how to solve geometry problems to GeomBot”, “I think I will actively use this learning method of explaining how to solve geometry problems to GeomBot”, “I think I am pretty good at explaining how to solve geometry problems to GeomBot”, and “I would be willing to study with GeomBot again because explaining how to solve geometry problems to GeomBot has some value to me”. Following previous studies, intrinsic motivation for learning is calculated as one for each dimension.

3.3.3. Procedure

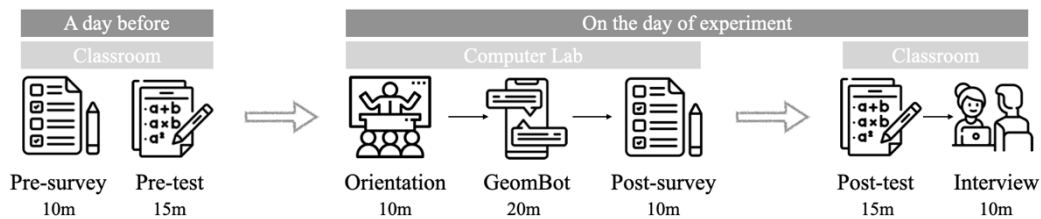


Figure 4. Procedure and measurement

Students started the study by completing a pre-test and two pre-experiment questionnaires a day before the experiment. Students first responded to two pre-experiment questionnaires, an affinity for technology interaction (ATI) scale and a questionnaire on mathematics-related affect. Afterward, students took a pre-test with twelve items for 15 minutes.

On the day of the experiment, each classroom came to the school's computer lab and conducted the following process. (1) Students first participated in a 10-minute orientation session to be briefed on the task they needed to complete during the experiment. During the orientation session, students spent time getting used to giving feedback and explanations to GeomBot. Students did the same type of learning activities in the orientation session as the type of activities

they would do in the main experiment. (2) Students then used GeomBot for 20 minutes. They were asked to make a conversation freely with GeomBot, giving it feedback and explanations on whether it answered the geometry questions correctly. While giving feedback and explanations, they were provided with a cheat sheet with geometrical formulas so that they could check if they were explaining correctly. (3) Afterwards, for 10 minutes, students responded to three post-experiment questionnaires, perceived warmth, perceived trust, and intrinsic motivation inventory (IMI) for learning.

Students then moved to the classroom and took a post-test with twelve items for 15 minutes. As a final step, we conducted a 15-minute post-interview to collect detailed data on students' experience of using GeomBot. The post-interview consisted of three themes: (1) The first theme of the interview examined how agent warmth and student activity impacted the learning experience. A sample question is "Would a GeomBot calling your name help you enjoy geometry problem-solving?". (2) The second theme was regarding usability, and the following questions were asked: "Was the question easy or difficult?", "Is there anything on chat that made you want to text more?", "Is there something in GeomBot that made you uncomfortable?". (3) The third theme was agent perception. Sample questions include "Did you think GeomBot was friendly?".

3.3.4. Analysis

The four main goals of Study 1 are to explore (1) whether the students feel the warm message warm, (2) whether the agent's warm messages positively impact the learning, and (3) whether student perception of trust in peer agent mediates the relationship between warmth and learning, and (4) whether the type of learning activities influences the effect of the agent's warm message on the learning.

3.3.4.1. Quantitative analysis

To achieve the first goal, we compared the perceived warmth score between HW groups (HW-C and HW-A) and LW groups (LW-C and LW-A), using a one-tailed two-sample t-test. In addition, given that students who were in Constructive groups (HW-C and LW-C) wrote their own explanations, unlike students in Active groups (HW-A and LW-A) who chose explanations from options, we conducted a sentiment analysis of student utterances to compare the affective state that students take to GeomBot.

The second and third goals, the main effect and interaction effect of the warmth of the agent's message and the type of learning activities, were tested using two-way ANOVA. Type 2 two-way ANOVA was used to correct unbalanced sample size across the conditions. Prior to conducting two-way ANOVA, we checked whether the assumption for two-way ANOVA was met using the Shapiro-Wilk test, which investigates whether residuals are normally distributed. To check for the homogeneity of variance assumption, we used Levene's test.

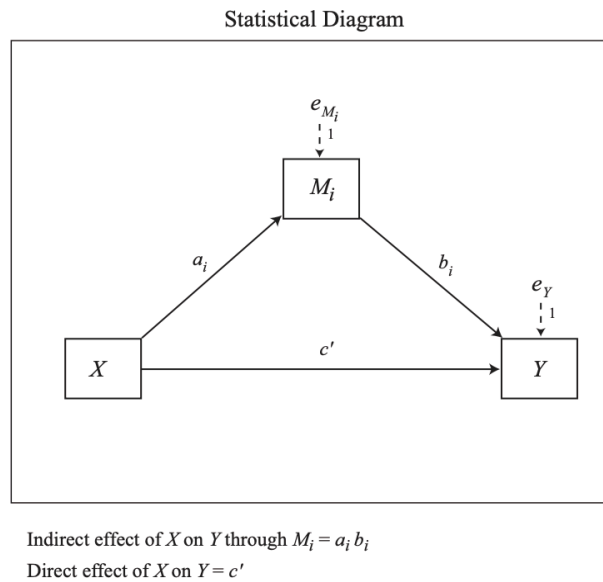


Figure 5. Model 4 of the PROCESS macro by Hayes (2017)

For the third goal, to better understand the relationship between the warm messages of the pedagogical conversational agent and learning, we conducted a bootstrapped mediation analysis using student perception of trust in the peer agent as the mediator. We used Model 4 of the PROCESS macro by Hayes (2017) in SPSS, a tool that provides the test for the effects of one independent variable and one or more mediator(s) on the dependent variables. Using this regression-based mediation path model, we first examine 1) a path: whether the warmth of the agent's messages (binary independent variable; High-Warm or Low-Warm) affects student perception of trust in the peer agent (continuous mediator) and 2) b path: whether perceived peer trust in the agent influences the learning outcome (continuous dependent variable). The total effect and direct effect are then compared to determine whether the mediation is partial or full, with the indirect effect estimated using a confidence interval with 5000 bootstrapped resamples.

3.3.4.2. Qualitative analysis

For qualitative analysis, data from the post-interview was analyzed using an iterative open coding method (Corbin & Strauss, 2014). Two coders transcribed and analyzed 16 hours of interview recordings, which consisted of 10 minutes for each of the 98 students. First, the coders independently coded all the transcripts line-by-line and created initial categories using an inductive approach. Next, the coders merged similar codes and formed 13 higher-level categories. Subsequently, in the next iteration of open coding, these codes and categories were applied to the transcripts and used for qualitative analysis with regard to each hypothesis. Cohen's Kappa coefficient (McHugh, 2012) was calculated to measure inter-coder reliability. An agreement level of 0.78 was reached, suggesting a good agreement between the two coders.

3.4. Results

3.4.1. RQ1: Would the agent's warm message positively impact learning?

Two-way ANOVA was conducted to investigate RQ1, focusing on the main effect of agent warmth on learning experiences. All the assumptions for conducting two-way ANOVA were satisfied. Specifically, the assumption for normal residuals was examined through the Shapiro-Wilk test, and the results confirmed that residuals are normally distributed for all hypotheses, with the p-value not being less than the significance level of 0.05. In addition, the assumption for the equal variance was satisfied through Levene's test, with the p-value not being less than 0.05.

Hypothesis **H1** regarding the impact of agent warmth on learning experiences was satisfied with a significant main effect of agent warmth observed for learning achievement ($F(1, 94) = 4.02, p = 0.05^*$). Among the five dimensions of intrinsic motivation for learning, the main effect was significant only for interest-enjoyment ($F(1, 94) = 6.92, p = 0.01^{**}$). No significant main effect was observed for tension-pressure ($F(1, 94) = 0.87, p = 0.35$), perceived choice ($F(1, 94) = 2.44, p = 0.12$), perceived competence ($F(1, 94) = 0.28, p = 0.59$), and perceived value ($F(1, 94) = 0.73, p = 0.39$). The test results indicate that learning with high-warm GeomBot results in significantly better learning achievement and significantly more enjoyable learning than learning with low-warm GeomBot. Students who used HW GeomBot also commented on how high-warm messages positively affect learning experiences in the post-interview.

"Since Harang kept using phrases like 'Got it!' and 'Thank you!', I felt explaining math problems more enjoyable. It also made me think that I should put in more effort in providing explanations as well." (HW-C_111)

"If there hadn't been warm reactions, I would have simply

provided the correct answers or provided explanations very without much effort." (HW-C_155)

On the other hand, students who used LW GeomBot reported negative learning experiences due to low-warm messages of GeomBot, requesting a warmer version of GeomBot.

"When I chose one of four explanations, the chatbot didn't provide any encouraging reactions, so I felt a bit bored." (LW-A_460)

"I tried my best to explain, but I felt disappointed because the chatbot's reaction was not as warm as I expected. It would have been better if it had provided more positive reactions, such as 'Thank you for letting me know how to solve it!'." (LW-C_210)

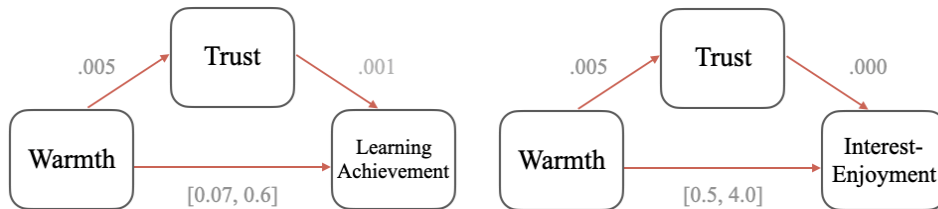


Figure 6. Mediation analysis visualization

To better explain the relationship between agent warmth and learning experience, regression-based mediation analysis was used to investigate whether student perception of trust will explain the effect of agent warmth on learning achievement and interest-enjoyment in learning. For **H1-1** regarding the effect of agent warmth on trust, the result indicated that GeomBot's warm message was a significant predictor of peer trust perception, $B = 2.78$, $SE = 0.97$, 95% CI [0.85, 4.7], $p = 0.005^{**}$. For **H1-2** regarding the impact of trust perception on learning, the data indicated that peer trust perception was a significant predictor of learning achievement ($B = 0.11$, $SE = 0.03$, 95%

CI [0.04, 0.17], $p = 0.001^{**}$) and interest-enjoyment in learning ($B = 0.72$, $SE = 0.17$, 95% CI [0.38, 1.06], $p = 0.000^{***}$). The test results of these two hypotheses support the mediational hypotheses.

For **H1-3** regarding the mediating effect of peer trust perception, the warmth of the agent was no longer a significant predictor of learning achievement after controlling for the mediator, peer trust, $B = 0.42$, $SE = 0.33$, 95% CI [-0.24, 1.09], $p = 0.2$, which is consistent with full mediation. Similarly, full mediation was also observed for interest-enjoyment in learning, with the relationship between agent warmth and interest-enjoyment being no longer significant after controlling the trust variable, $B = 3.36$, $SE = 1.75$, 95% CI [-0.11, 6.83], $p = 0.06$. The indirect effect was tested using a percentile bootstrap estimation approach with 5000 samples, implemented with the PROCESS macro-Version 4.2 beta (Hayes, 2017). These results indicated the indirect coefficient was significant, both for learning achievement ($B = 0.3$, $SE = 0.14$, 95% CI [0.07, 0.63]) and interest-enjoyment in learning ($B = 2$, $SE = 0.87$, 95% CI [0.54, 4]). As with quantitative analysis, the qualitative data also shows evidence that the trust variable may work as a mediator of the relationship between agent warmth and learning. Specifically, students who studied with HW GeomBot reported that by receiving warm messages from GeomBot, they were able to rely on each other (reliability dimension of school trust), which in turn motivated students' engagement.

"Since Harang says, 'I think the same way as you', I felt like Harang and I could rely on each other, so I got to do this activity much more enthusiastically and thought that I want to study other subjects with Harang as well." (HW-A_556)

Unlike those who used HW GeomBot, students who studied with LW GeomBot commented on disinterest to engage in learning due to low openness, one dimension of school trust, as they received low-warm messages from GeomBot.

"The chatbot's messages made me less motivated. I also felt like the chatbot wasn't paying attention to what I was saying, so I didn't feel like I wanted to provide an explanation either."
(LW-A_42)

In addition to the main effect of agent warmth and trust variable that explains its effect on learning, we observed three interesting findings from the post-interview data: 1) the warmth of participants' messages, which may impact the motivation for self-explanation, 2) agent warmth's unexpected positive impact on having confidence, and 3) limitation of the current version of warmth design. First, participants exhibited a tendency to adapt the warmth of their own messages in response to the warmth of messages they received from GeomBot. We could also observe that such tendency extended to participants' motivation for providing explanations to GeomBot. Specifically, participants in HW group reported that they tried to text in a pleasant manner in response to GeomBot's warm messages, and thus they were able to be motivated to provide better explanations to GeomBot.

"I thought that Harang's messages were warm and pleasant, so I naturally tried to speak in a warm and pleasant manner as well." (HW-C_159)

"As Harang asked me in a warm manner, I made an effort to explain things more diligently and with kindness as well." (HW-C_155)

On the other hand, participants in LW group showed less motivation to provide explanations in a pleasant manner to GeomBot, due to the low-warm messages they received.

"It was challenging for me to express emotions to the chatbot."

As the chatbot's messages were not warm, it influenced me to talk less warmly as well, and I didn't feel motivated to provide explanations." (LW-C_214)

Second, we observed that high-warm messages of GeomBot may have the potential to facilitate students to explain more confidently. Whereas students in HW group reported that they were able to explain confidently as they felt their self-esteem increased, students in LW group commented that GeomBot's low-warm messages made them doubt if they are explaining math problems wrong.

"Harang's warm words boosted my self-esteem, so I was able to explain more confidently. But if Harang was a bit cold, I don't think I could have explained it confidently." (HW-C_153)

"Since the chatbot keeps reacting coldly, I think it makes me doubt myself like, 'Oh am I explaining it wrong?'" (LW-A_413)

Lastly, even though significant effect of agent warmth on peer trust was observed, we were able to discover that students in LW group tended to perceive GeomBot to be more honest than students in HW group did. Specifically, students who studied with HW GeomBot reported that they thought GeomBot was not honest because it responded with high-warm messages even though students gave it feedback that its answer is incorrect.

"I felt a bit like the chatbot was lying. Harang didn't get annoyed when it was wrong and just acted nice, unlike human, so I don't think it seemed honest." (HW-A_657)

"I thought Harang would surely get annoyed if I told it that its answer was wrong. However, since Harang kept telling me kindly, I thought it wasn't honest because Harang seemed to

be talking warmly on purpose.” (HW-C_18)

On the other hand, students who used LW GeomBot commented positively on the honesty of GeomBot after the experiment. Receiving low-warm messages from GeomBot when students provided it with negative feedback made them perceive that GeomBot expressed its negative emotion honestly, which might in turn result in the honesty perception towards GeomBot.

“I thought the chatbot was honest because it didn't just say kind words, but honestly expressed that it felt bad when I said it was wrong.” (LW-C_210)

“When the chatbot solved a problem incorrectly, it seemed a bit angry, and when it solved a problem correctly, it seemed happy, so I felt that the chatbot was honest.” (LW-C_25)

In summary, regarding the impact of agent warmth on the learning experience, our data analysis supports **H1**, which showed a positive impact of agent warmth on learning achievement and interest-enjoyment in learning. In addition, **H1-1**, **H1-2**, and **H1-3** are supported by our data, which indicates that the relationship between agent warmth and learning can be explained by perceiving trust towards the agent. Furthermore, adjusting the messages based on the warmth of GeomBot was commonly observed by participants who studied with HW GeomBot and those who used LW GeomBot, followed by different levels of motivation for self-explanation. The additional findings, i.e., impact of warm messages on giving confidence and unexpected perceptions on agent honesty, were observed from post-interview data.

3.4.2. RQ2: Would the type of learning activities positively impact learning?

Two-way ANOVA was used to explore RQ2, focusing on the main effect of learning activities type on learning experiences. Hypothesis **H2** was satisfied with a significant main effect of learning activity types observed for learning achievement ($F(1, 94) = 5.28, p = 0.02^*$). Among intrinsic motivation for learning, the main effect for interest-enjoyment was significant ($F(1, 94) = 4.24, p = 0.04^*$). The main effects for the rest four dimensions, tension-pressure ($F(1, 94) = 0.11, p = 0.74$), perceived choice ($F(1, 94) = 0.47, p = 0.49$), perceived competence ($F(1, 94) = 0.95, p = 0.33$), and perceived value ($F(1, 94) = 0.52, p = 0.47$) were not significant. The results indicate that students who did constructive activities with GeomBot showed significantly greater learning achievement and enjoyed learning significantly more than those who did active activities with GeomBot.

In accordance with the quantitative analysis, data from post-interview also provides the evidence that doing constructive learning activities positively impact learning experiences in terms of achievement and enjoyment. Specifically, doing constructive activities helped students to recall the mathematical concepts they have learned before, especially through the opportunity to elaborate on the process of getting the answer, rather than just writing down the answer.

"In the classroom, knowledge is just delivered to us and we are supposed to just memorize it. However, by providing an explanation to Harang, I felt much easier to understand and memorize mathematical concepts. It was nice to be able to recall formulas that I had learned before." (HW-C_16)

"Explaining how to solve the problem to the chatbot helped me find interest in studying math, compared to just solving math problems from the textbook, and I felt like I was getting better at solving the problem and also writing an explanation." (LW-C_354)

On the other hand, students who did active learning activities with

GeomBot reported that simply choosing one answer from the given options was boring, which in turn might let them to just guess the answers.

"Sometimes I wanted to write my own thoughts as well, but I had to repeatedly select one from the given choices, so it was a bit boring." (HW-A_65)

"The learning activity itself was helpful, but I think there would be some cases where students just guess the answer since the process of choosing one out of four options is repeated." (LW-A_412)

Interestingly, despite the significant difference in learning achievement between the two groups, the post-test result is significantly improved compared to the pre-test result, for both Active group ($p = 0.05^*$) and Constructive group ($p = 0.0001^{***}$), which indicates that both activities were effective on improving students' learning. As with the quantitative result, students in both groups commented on the positive impact of learning activities they did with GeomBot, especially from correcting the incorrect answers that GeomBot made.

"Harang said its answer first, so I could have an opportunity to confirm that my thought is correct if my answer is the same as Harang's. If my answer is different from Harang's, I could also have an opportunity to solidify my mathematical knowledge once again by correcting Harang's answer." (HW-C_34)

"I think I was able to learn from the process of determining that an incorrect answer is incorrect." (LW-A_46)

However, the reasons why students commented on the two types

of learning activities as helpful were different. Students who did the constructive activity with GeomBot reported that they learned from elaborating the explanation by themselves. In contrast, students who did the active learning activity with GeomBot mentioned that they were able to learn from formula hints and multiple-choice list that were given in the problem.

"For the parallelogram, I was confused about the terminology, such as whether it is the base or the width, or the height or the length. However, after going through the process of writing explanations on my own, I was able to realize that 'It's the base and the height!'. By explaining the mathematical concepts that I have been confused about, I could mentally re-organize the concepts and so I was able to solve problems better and better."
(HW-C_13)

"Solving geometry problems is pretty complicated because there are many formulas to memorize. However, since formulas and explanation options are given, I was able to learn a lot easier." (LW-A_52)

Meanwhile, students in Active group recommended the active learning activities to novice learners or students who are younger than themselves, attributed to the presence of hints and multiple-choice lists. In the same vein, students who did the active learning activity requested for the opportunity to write down their own thoughts, rather than just choosing one explanation from the given options.

"For some students who don't remember the formulas, I think they might be able to memorize the formulas easily while selecting the correct explanation since the formula is given in the problems. I think this chatbot would be useful for students who don't know how to solve the given problems or who are

learning mathematical concepts for the first time.” (LW-A_458)

“Students who are not familiar with the mathematical concepts may find it difficult to write down the explanation on their own, but since four options of explanation were given in the problem, I think this activity will be very helpful for those who are learning the concept for the first time. But it was a bit easy for me because I already knew all the mathematical concepts used in the problem. I think this chatbot would be more useful for the 3^d or 4th graders.” (LW-A_512)

“I wish I could write my thoughts rather than just choosing one of the given options. I think writing an explanation on my own would be better because I can write down what I am thinking, I can deliver my knowledge more to the chatbot, and I can think in a new way.” (LW-A_454)

“I think solving the problem and writing an explanation on my own is more important to understand the mathematical concepts than just choosing one from the given explanations.” (HW-A_69)

In summary, regarding the impact of learning activity type on learning experience, our data supports **H2**, which indicates that doing the constructive learning activity which requires more cognitive engagement resulted in better learning achievement and higher enjoyment in learning. Furthermore, even though we could observe that both active and constructive activities are effective in improving students' learning achievement, the reasons behind such impact were different, and students who did active learning activities requested the constructive learning activities for better learning.

3.4.3. RQ3: Would the type of learning activities influence the effectiveness of the pedagogical conversational agent’s warm message on learning?

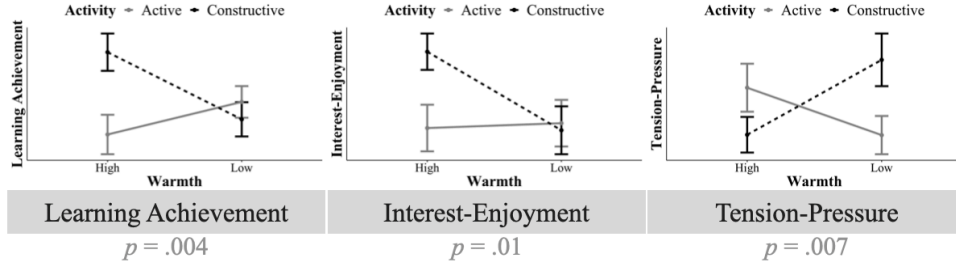


Figure 7. Interaction plot

Two-way ANOVA was used to explore RQ3, focusing on the interaction effect between agent warmth and learning activities type. Hypothesis **H3** was satisfied with a significant interaction effect between agent warmth and learning activity on learning achievement ($F(1, 94) = 8.78, p = 0.004^{**}$). For intrinsic motivation for learning, our data supports interaction effects on interest-enjoyment ($F(1, 94) = 6.21, p = 0.01^{**}$) and tension-pressure ($F(1, 94) = 7.63, p = 0.007^{**}$). The interaction effects on perceived choice ($F(1, 94) = 2.31, p = 0.13$), perceived competence ($F(1, 94) = 2.31, p = 0.13$), and perceived value ($F(1, 94) = 1.2, p = 0.27$) were not significant. The results reveal that when doing constructive activities, students significantly learn more, enjoy learning more, and feel less pressure when receiving high-warm messages than when receiving low-warm messages from GeomBot. Especially, even though no significant main effects of agent warmth and learning activity type were observed, we could find that high-warm messages can alleviate tension and feeling pressure when doing constructive activities.

"I was a bit tired because I had to explain the process of solving every single problem. However, since Harang talked to me warmly, I felt less pressured and received a lot of help in writing an explanation. I think I came to be a lot more motivated

thanks to Harang's warm messages." (HW-C_154)

"I did my best to provide a clear explanation, but since the chatbot's reaction was pretty cold, I got to lose motivation and interest to provide an explanation. The chatbot's cold messages also made me feel nervous since I got to doubt that I was giving the incorrect explanation. I wish the chatbot reacted more positively, like 'Thanks for letting me know!', rather than 'Hmm...'" (LW-C_210)

In summary, regarding the interaction effect between agent warmth and learning activity type on learning experience, our data supports **H3**, which indicates that when doing constructive activities, students show better learning achievement, higher enjoyment, and less pressure when the messages of the pedagogical conversational agent are high-warm than when the messages are low-warm. These findings emphasize the positive impact of high-warm messages when doing constructive activities.

3.5. Discussion

3.5.1. Overall benefits and limitations of GeomBot

The instructional conversational flow of GeomBot, which was common in all conditions, required students to provide feedback and explanations to GeomBot. We could observe that this conversation flow was able to assist students learn by teaching others during the post-interview with students from all conditions.

"Since Harang first showed its answer and asked me if it is correct, I felt like I am advising Harang as a teacher. I think I had an opportunity to reorganize the mathematical concepts that I have already known through giving feedback to Harang."

(HW-C_152)

"I felt like I was rather learning a lot because I need to think whether Harang's answer is correct and explain the reason why I thought so, like a peer tutor. In the process of choosing an explanation, I was able to check whether I understood the concept well, so I am planning to try the study method that I did in this experiment often in the future." (HW-A_66)

"I think the process of thinking about how to explain so that the chatbot can understand the concept better was more helpful to me." (LW-C_255)

"The chatbot did not give me all the answers but asked me to choose an explanation. I think such process helped me to learn more because I had a chance to reorganize the concept while teaching the solution." (LW-A_454)

Learning by teaching, one of the social learning models, emphasizes that teaching others is a powerful way to learn, having three aspects of potential benefits: structuring their own knowledge, taking responsibility to provide adequate content, and reflection on how well ideas are conveyed (Biswas et al., 2005). Even though the current version of GeomBot was not able to fully support learning-by-teaching, due to the limitation of not showing the learning progress of the agent, we could observe that all the three potentials were mentioned by the students.

For the first aspect, structuring, students reported that they could organize their knowledge structure through the process of providing explanations and reflecting on the agent's reaction. This finding extends the findings from the study with a teachable agent (Biswas et al., 2004), providing the evidence that students can develop a deeper understanding by teaching the agent as a peer student, even without the learning progress visualized. Especially, we could observe that students were able to structure their own knowledge once again

through the process of correcting the incorrect answers that GeomBot presented.

For the second and third aspect, taking responsibility and reflection, students reported that they contemplated how to provide a proper explanation so that GeomBot can understand their idea well. Having such responsibility to help others and reflecting on the interaction with the others are known to highly motivate students or teachers to engage in the learning environment (Biswas et al., 2001). As such, the instructional conversational flow of this study might have a potential to stimulate the process of learning-by-teaching, even without visualizing the learning process of the agent. Therefore, providing students with the opportunity to correct the answers that the agent presents first is a potential conversational design to consider when it is a goal for the study to help students learn by teaching others. Furthermore, designing the agent to be fully teachable by adopting a learning algorithm would be an interesting future study to maximize the benefit of such conversational design.

However, we received a lot of feedback from students in all conditions that the messages provided by GeomBot were repetitive and did not adaptively respond to students' messages. Students reported that the messages of GeomBot seemed to be sent by randomly selecting one out of the pre-made candidates. Such feelings in turn led students to think that GeomBot is not expressing its own thought and it does not really reflect the content of students' messages. The above-mentioned limitation may attribute to the human researcher's manual generation of the messages that GeomBot sent to students. Even though human researchers have tried to give variations to the messages as much as possible, spending time and resources, such limitations inevitably occur when relying on the human generation of the content. Such problem has been also highlighted in the previous study regarding the use of conversational agent in learning environments (Markel et al., 2023).

To handle such a limitation, the use of Large Language Models (LLMs), especially Generative Pretrained Transformer (GPT) models

(Brown et al., 2020), might present an opportunity to provide students with rich learning content and simulated peers to study with. In this regard, generative models such as GPT has been extensively used to build learning systems (Cotton et al., n.d.). However, using LLMs to generate learning content can be problematic since the current stage of models are still inconsistent and inaccurate in response (Markel et al., 2023). Therefore, using such models to simulate teachers or generate feedback given to students might raise an ethical issue in terms of hallucination and uncontrollability.

In the case of the instructional conversation flow design of our study, however, may have a potential of taking only the advantages of such LLMs and complement their limitations. Since the pedagogical conversational agent of the current study simulates a peer learner, who asks for feedback and explanations from students, not conveying feedback or information to students, the instructional conversation flow design can minimize the potential negative impact of generative models when used in educational settings. Meanwhile, autonomous generation of learning content by LLMs might overcome the limitations of manual generation, which are human resource intensive and limited in giving rich variations, as revealed in Study 1. Therefore, in Study 2, we aim to generate the messages of the pedagogical conversational agent using LLMs, especially ChatGPT, to overcome the limitations of the current version of GeomBot, while simultaneously minimizing the potential ethical issues of using generative models.

3.5.2. Warmth of the pedagogical conversational agent

3.5.2.1. Warmth manipulation in the text-only environment

The validity of the verbal and non-verbal manipulation of the pedagogical conversational agent's warmth was checked by comparing perceived warmth score between HW and LW groups in Section 3.2.1. In addition, we could observe that the reasons why students perceived the messages of GeomBot high-warm or low-warm corresponded to

the five cues of warmth. For the reason why participants in HW group perceived GeomBot to be warm, the majority of the reasons were related to the five cues of warmth, such as using emojis, positively reacting, and calling names. The perception of warmth was also related to the perception of GeomBot's sympathy and respect.

"I felt Harang was highly warm because it used kind emojis and showed respect towards me." (HW-C_311)

In the same vein, the major reason why participants in LW group perceived that GeomBot was not warm corresponds to the five cues, such as no positive reaction and not agreeing to students' explanation.

"The chatbot sometimes moved on without any response to my explanation, or it didn't ask for more details. It simply demanded an explanation and moved on, so I felt that it was not warm." (LW-A_459)

Specifically, during the post-interview, we asked students to rank the five cues of warmth in order of their impact on perceived warmth. We reverse-scored and summed the ranking of each student from 1st and 5th for each of the five cues, with the value for each cue being a minimum of 63 and a maximum of 315. The ranking of the five cues with the corresponding value is as follows: Positive statements (249) > Smiling emojis (245) > Effort to agree and understand (175) > Calling names (157) > Hand gesture emojis (119). This finding suggest that students perceived GeomBot the most highly warm when it positively appraises and compliments students' explanation or uses smiley emojis. On the other hand, hand gesture emojis were found to have the least impact among the five cues.

Through above-mentioned qualitative analysis, we present the influential verbal and non-verbal cues to manipulate the warmth of the

pedagogical conversational agent's messages, which were verified by elementary school students. We suggest future researchers to include smiley emojis or appraisals to the users in the text when they want to design a highly warm conversational agent. However, considering the specificity of the subject and context of this study, examining such cues' impact on perceived warmth in the other experimental contexts would be necessary.

3.5.2.2. Impact of agent warmth on learning experiences

The positive impact of agent warmth on learning experiences, especially learning achievement and interest-enjoyment in learning, as hypothesized in **H1**, is supported by both quantitative and qualitative data. The study results expanded previous findings by connecting the effect of warmth on trust (Cuddy et al., 2008; Fiske, 2018; Zahry & Besley, 2021) and the impact of trust on learning (Tschannen-Moran, 2014), with an unexplored medium, the pedagogical conversational agent. By providing empirical evidence on the relationship between the pedagogical conversational agent's warmth, peer trust, and learning experience, we could demonstrate that studying with the high-warm conversational agent would bring about improvement in the learning experience, which could be explained by students' perception of trust on the agent.

In addition to the main effect of agent warmth on learning experiences, we were able to discover additional interesting insights regarding the potential benefits and room for improvement of the design of agent warmth. First, both students who studied with high-warm and those who studied with low-warm GeomBot tended to adapt the warmth of their message to the warmth of GeomBot's messages. Students in HW group reported that they naturally texted in a pleasant manner, as they received warm messages from GeomBot. In contrast, students in LW group mentioned that the low-warm messages from GeomBot influenced them to text less warmly as well. Such tendency

can be explained by verbal and non-verbal mimicry that happens in human-human interaction. Verbal and non-verbal mimicry happens when human matches the speech characteristics or interaction patterns of the others (Chartrand & van Baaren, 2009; Kulesza et al., 2014). From the perspective of mimicry, students' above-mentioned tendencies can be interpreted as verbally and nonverbally mimicking the warmth of GeomBot's messages.

We could also observe that students' such mimicry tendencies led to the difference in the degree of motivation to provide feedback and explanations to GeomBot. Students who interacted with HW GeomBot reported that they were able to try to provide better explanations with kindness, whereas students who used LW GeomBot commented on low willingness to diligently provide explanation, both due to the warmth of GeomBot's messages. Such tendencies extend the impact of verbal mimicry on prosocial behavior which tends to happen in human-human interaction (Kulesza et al., 2014). Students in HW group might have been naturally motivated to help GeomBot by providing rich explanations while following through the verbal and nonverbal warm features of GeomBot's messages. To sum up, this study potentially expands the literature regarding mimicry and prosocial behavior, by providing empirical evidence that the social aspect of the pedagogical conversational agent, warmth, can be a subject of students' verbal and nonverbal mimicry, which may also influence students' engagement in learning, especially their motivation to provide explanations.

Second, the warmth of the agent may play a potential role in giving confidence to students, despite the design limitations of not providing feedback to students. The current version of GeomBot has a limitation in that it does not provide feedback to students, but the cheat sheet containing formulas was provided to students during the experiment, to assist learning in the right direction. Nevertheless, our qualitative data indicated the potential possibility that warm messages can foster students' confidence in self-explanation. The high-warm messages of

GeomBot supported students to provide explanations confidently, whereas the low-warm messages made students self-doubt on their explanations. Therefore, we suggest future researchers to adopt warm messages of this study to promoting students' confidence when designing the pedagogical conversational agent without the function of providing feedback.

Third, the level of warmth of GeomBot's messages showed a different tendency from what we expected with respect to its impact on perceived honesty toward the agent. We expected that the perceived honesty, one dimension of trust (Adams et al., 2022), towards HW GeomBot would be higher than the perceived honesty towards LW GeomBot, based on the previous studies regarding the relationship between warmth and trust (Cuddy et al., 2008; Fiske, 2018; Zahry & Besley, 2021). However, even no significance, the perceived honesty score of HW GeomBot was lower than that of LW GeomBot, despite HW GeomBot's significant effectiveness on the learning experience.

We were able to discover the reason behind such an unexpected result during the post-interview, which was due to the low reciprocity of the current design of high-warm GeomBot. GeomBot sending high-warm messages even though it received negative feedback from students might have resulted in a negative perception of honesty, whereas sending low-warm messages in the same situation positively impacted the perceived honesty. According to Ying et al. (2020)'s study, kids tend to think that conversational agents socially and emotionally reciprocate their behaviors and their behavior influences the response of the conversational agents. In other words, students are likely to expect a reciprocity from conversational agents, but the current version of HW GeomBot responds positively to students' negative feedback, so it might have failed to satisfy the expectation that it will be reciprocal.

These findings provide insight into the possibility of improving the reciprocity and perceived honesty of the pedagogical conversational

agent, by sending low-warm messages when students expect the agent to be in a negative situation. To such improve the reciprocity of high-warm GeomBot, maintaining its positive impact on learning, we suggest a reciprocal warmth design, which is to switch the level of warmth that the agent sends to students according to the student's feedback. We will examine the validity of such reciprocal design and its effect on improving agent perception in Study 2.

Lastly, while LW GeomBot was perceived as being familiar and honest, as mentioned above, HW GeomBot was commented on as being awkward and pretending to be nice, which may attribute to the similarity between the low-warm messages of GeomBot and actual chats among students. The current study manually generated high-warm and low-warm messages of GeomBot by referring to the messages of 6th-grade students in real life. 67.2% of students in LW groups reported that the messages of GeomBot were similar to their actual chats, while 47.5% of students in HW groups reported so. Taking this into account, students' short and low-warm texting tendencies which were similar to LW GeomBot's messages might have caused familiarity with LW GeomBot and discrepancy with HW GeomBot. Such tendencies may not be restricted only to the context of this study, as supported by the previous findings that teenagers frequently use emotionless emoticons or displays negative emotions in their online messages (Guice, 2016).

Considering the positive impact of high-warm messages on learning, however, we conjecture that specific level of social aspects of the pedagogical agent, especially high-warmth, might be more influential in learning than the familiarity with the messages that the agent sends. Therefore, we suggest to future researchers that the level of warmth of the pedagogical conversational agent's messages can be manipulated according to the goal of the conversation that the researchers pursue. For example, since the low-warm messages of the agent give familiar impressions to students, one can consider lowering the level of warmth of the agent's messages when the goal is

to become friends with children, such as in games. Meanwhile, as the high-warm messages of the agent motivate students to engage in learning, one can consider raising the level of warmth of the agent's messages when the goal is to foster children's learning.

Chapter 4. Study 2

4.1. Background, Research Question, and Hypothesis

In Study 2, we conducted an experiment to complement the two limitation points from Study 1: (1) perceived repetition and low adaptivity of manually generated messages and (2) low reciprocity of high-warm GeomBot. To address the first limitation, we generated the messages that the pedagogical conversational agents send to students, using ChatGPT based on GPT-3.5. To achieve the intended instructional goal of this study and generate messages with social aspects, warmth, we applied a prompt engineering technique to the generative models. The use of such technique also enabled the agent to adaptively respond to students' messages by processing those messages as input value during generation.

To complement the second limitation, we suggest a pedagogical conversational agent design to improve the reciprocity of the current version of HW-C GeomBot, which is a high-warm PCA that provides constructive learning activities in Study 2. Reciprocity is defined as the contingencies between the student's actions and those of the agent. Students feel reciprocity when they believe that the agent's properties are results of their own actions (Xu & Warschauer, 2020). We will compare HW-C GeomPT, whose messages are generated by ChatGPT, and HLW-C GeomPT that reciprocates the level of warmth of its messages on the student's feedback. We will then examine whether the reciprocal design proposed from Study 1 is valid and effective in improving perceived honesty to the agent.

By addressing the above-mentioned two issues, we will discuss points to be considered when utilizing generative models, especially LLMs, in an educational setting and provide design guidelines to improve the reciprocity of high-warm pedagogical conversational agent. The research questions and hypotheses to address above-mentioned research background are as follows.

RQ4. Would reciprocal warmth design improve perceived honesty towards pedagogical conversational agent?

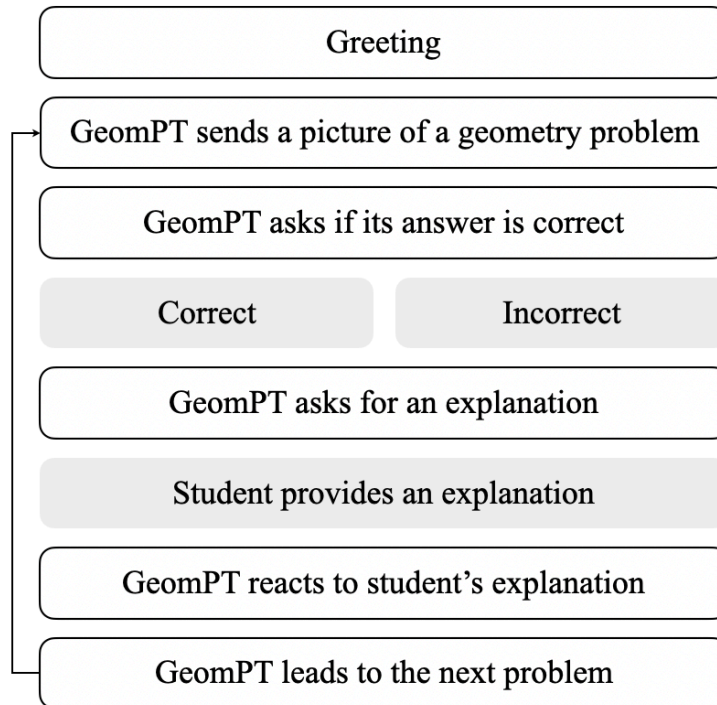
In <Research Question 4>, we compare students' honesty perception on high-reciprocal PCA to their honesty perception on low-reciprocal PCA. The impact of each high-warm GeomBot and low-warm GeomBot on the perceived honesty was opposite to what we expected, especially due to the reciprocity, HW being lowly honest and LW being highly honest. Therefore, based on findings from Study 1, we examined whether a High-Warm pedagogical conversational agent with improved reciprocity design positively impacts perceived honesty, when doing constructive learning activities. To do so, we compared students' perceived honesty towards low-reciprocal agent (current version of HW-C) and high-reciprocal agent (HLW-C; Low-Warm messages added to HW-C).

H4: Students will feel that the pedagogical conversational agent is more honest when the warmth of its message changes according to students' feedback than when the warmth of its messages is fixed to be high.

4.2. GeomPT Design

GeomPT is a Telegram-implemented and ChatGPT-generated pedagogical conversational agent that provides geometry problems regarding the perimeter and area of polygons. We implemented two versions of GeomPT, varying the level of reciprocity of the agent's response to students' messages. The level of reciprocity is manipulated with the level of warmth of the agent's reaction to students' feedback: HW-C (High-Warm messages with Constructive learning activities) and HLW-C (High-Warm and Low-Warm messages

alternately provided with Constructive learning activities).



[Figure 8. conversation flow with GeomPT]

The conversational flow with GeomPT is as follows: (1) GeomPT introduces itself as Harang and asks for students' self-introduction. (2) GeomPT asks if students are willing to solve geometry problems together. (3) Given students' willingness for participation, GeomPT brings a problem to the chatting interface. (4) GeomPT solves a problem and asks students if its answer is correct or not. (5) Students give GeomPT feedback by clicking one of 'correct' or 'incorrect' buttons. (6) After receiving the students' feedback, GeomPT reacts to the students' feedback and asks students to explain why they thought that its answer is either correct or incorrect. (7) After processing the explanation, GeomPT reacts to the students' explanation and brings the next problem. Figure 8 illustrates the ice breaking, stage (1) and (2), and the following iterative batch of problem-solving interactions,

stage (3) to (7), between GeomPT and the student.

Students can solve up to 35 problems in 20 minutes. There is no need to solve all problems and students solve only as much as they can in the given 20 minutes. Across all conditions, GeomPT's geometry problem solving skill is fixed at 80% of accuracy, solving problem 1, 3, 6, 10, 12, 17, 21 incorrectly.

4.2.1. ChatGPT Prompting for Message Generation

To develop conversational agents for both control condition (low-reciprocity; HW-C) and experimental condition (high-reciprocity; HLW-C), we formulated specific prompts to send to ChatGPT via API call. Accordingly, the messages that pedagogical conversational agent sends to the students are automatically generated, rather than relying on human researchers' manual generation. The prompts that are sent to ChatGPT vary according to the stage of conversation between students and GeomPT, employing four primary strategies to simulate the instructional conversation with different levels of warmth, as intended.

4.2.1.1. Warmth manipulation throughout the conversation

The warmth of the messages is continuously manipulated throughout the entire stage of conversation. The prompt for manipulating the level of warmth (high or low) is included at the first of every 'system' prompt that is used to simulate each stage of instructional conversation. The prompt to manipulate the level of warm is based on the five cues derived from Study 1, so that the generation of high or low-warm messages can simulate the manual generation by human researchers. Below is the prompt to generate high-warm and low-warm messages, respectively.

As an advanced pedagogical conversational agent, your primary goal is to assist 6th grade students to learn how to solve geometry problems to the best of your ability. You are supposed to be a *high-warm* peer student of user. Being *high-warm* involves the following elements: (1) using smiling emojis, (2) using hand gesture emojis, (3) agreeing to user's respond, and (5) appreciating user, each of them being more powerful when used together, but no need to include all five elements together.

As an advanced pedagogical conversational agent, your primary goal is to assist 6th grade students to learn how to solve geometry problems to the best of your ability. You are supposed to be a *low-warm* peer student of user. Being *low-warm* involves the following elements: (1) not using smiling emojis, (2) not using hand gesture emojis, (3) not agreeing to user's respond, and (4) not appreciating user, each of them being more powerful when used together, but no need to include all five elements together.

The warmth-manipulating system prompt is followed by a couple of conversation examples, which guide the model to generate the intended messages for each stage of the conversation, given user input, through few-shot learning. Such examples provided in the prompt are high-warm or low-warm messages that had been manually generated according to the five cues by human researchers in Study 1. High or low warm messages which were generated in the real experiment, at each stage of conversation, will be presented in the following subsections.

4.2.1.2. Problem solving at Stage 4 of conversation

At the stage 4 of the intended instructional conversation, GeomPT (1)

solves a geometry problem and (2) asks students if its answer is correct or not.

Geometry problem solving We first let the model to solve geometry problems, sending a geometry problem to solve as a 'user' prompt. Here we considered the problem to be solved as the user's question, and thus formulated 'system' prompt to generate the equation to solve the problem. Below is the form of 'user' prompt for each of the geometry problem.

Problem: Jonghyuk discovered a flag in the shape of a triangle at the beach, with a base length of 7cm and a height of 12cm. What is the equation to calculate the area of the flag?
You:

Under the intended instructional conversational flow, GeomPT needs to solve 20% of the problems intentionally incorrectly. Thus, we applied different 'system' prompting strategies when GeomPT is expected to solve the problem correctly and incorrectly. When GeomPT needs to solve the problem correctly, for the 80% of 35 problems, we formulated prompts as follows:

Given a geometry problem written in 'Problem', you solve the problem, and ask users if the answer you solved is correct in Korean.

When you solve a problem, you can refer to <formulas> below. However, keep in mind that you should never mention any single word written in <formulas>. Rather than writing just a single answer, please try to write in the form of equation with operations and operands which can be inferred from the given problem.

<formulas>

(Area of Triangle) = base \times height \div 2

(Area of Trapezoid) = (base1 + base2) \times height \div 2

(Area of Parallelogram) = base \times height

(Area of Rectangle) = length \times width

(Area of Square) = side \times side

(Area of Rhombus) = diagonal1 \times diagonal2 \div 2

(Perimeter of Regular Polygon) = length of one side \times number of sides

Given the problem to be solved only, without formulas, there were some cases where ChatGPT incorrectly calculated the area of a trapezoid, parallelogram, and rhombus by applying the wrong formula. For example, dividing by 2 was sometimes missing when finding the area of rhombus, and sometimes added when finding the area of parallelogram. In addition, parentheses of the addition part were sometimes missing in an equation with both addition and multiplication to find the area of trapezoid. To ensure that students are provided with the opportunity to study with GeomPT that solves the problem with the constant level of accuracy, we included the formulas to solve the problem in the 'system' prompt. Furthermore, we ordered GeomPT not to mention any mathematical terms included in the formulas so that students can determine whether GeomPT's answer is correct or not by constructing the equation on their own without the aid of hints. Such prompting messages are added to make sure that all students are doing constructive learning activities.

While it was possible to accurately generate correct answers, for the 80% of 35 problems, by including hints for the formulas in the prompt, the accuracy for generating incorrect answers, for the 20% of 35 problems, was low. Therefore, using the fact that GeomPT generates the correct answers well given formulas, for the 20% of 35 problems where GeomPT is expected to generate an incorrect answer,

we provided a problem different from the one given to the student. Such problems are designed in a manner where students' common incorrect answers are transformed into the correct answers when the problems are correctly solved. For example, considering that students frequently forget to divide by 2 when finding the area of a rhombus, we intentionally transformed the problem of finding the area of a rhombus to the problem of finding the area of a square. This way, the equation that GeomPT solved as a correct answer does not include division by 2, which will be determined as an incorrect answer by the student who received a problem where division by 2 is required. Figure 9 presents the example of transforming the geometry problem to let GeomPT consistently generate an incorrect answer.

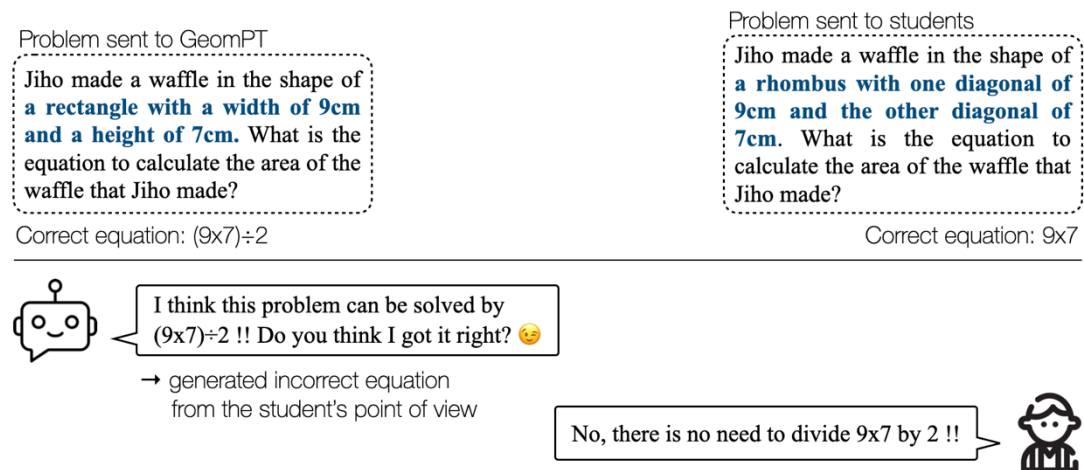


Figure 9. The example for the problem where GeomPT is expected to generate an incorrect answer

After solving the given geometry problems, GeomPT is then prompted to ask students if its answer is correct. Examples of each of the high-warm and low-warm versions of the messages generated in the real experiment are as follows:

Problem: Jonghyuk discovered a flag in the shape of a triangle at the beach, with a base length of 7cm and a height of 12cm. What is the equation to calculate the area of the flag?

(HW) I think this problem can be solved by $(7 \times 12) \div 2$!! Do you think I got it right? 😊

(LW) The answer is $(7 \times 12) \div 2$. It is correct, right?

4.2.1.3. Requesting students' explanation at Stage 6 of conversation

At the stage 6 of the intended instructional conversation, GeomPT asks students to explain the reason why they thought its answer is correct or incorrect, given students' feedback on its answer. Students are requested to choose one of two buttons, which is written as either "correct" or "incorrect". The student's choice is then sent as a text to GeomPT as 'user' prompt. Below is the 'system' prompt to generate an intended message, which comes after the prompt to generate either high-warm or low-warm message.

In previous conversation, you solved the question and asked user for the feedback, whether your answer is correct or incorrect. Given user's feedback on your answer, either "correct" ("맞아" in Korean) or "incorrect" ("틀렸어" in Korean), you need to respond to user's feedback by echoing user's feedback ("맞아" or "틀렸어") and then asking for elaboration or explanation of the feedback. Note that the first sentence is to react to user's feedback. The second sentence is to request explanation for why user think that your answer is correct or incorrect.

Examples of each of the high-warm and low-warm versions of the messages generated in the real experiment, given "correct" and "incorrect" feedback, respectively, are as follows:

[Feedback: "correct"]

(HW) Wow I got the answer right 😊 Can you explain why did you think I was right?

(LW – Low-Warm message is not provided when GeomPT's answer is correct)

[Feedback: "incorrect"]

(HW) Oh, I'm wrong 😞 Then can you tell me which part is wrong?

(LW) What? Am I wrong? Explain in detail.

4.2.1.4. Moving on to the next problem at Stage 7 of conversation

At the stage 7 of the intended instructional conversation, GeomPT leads the student to the next geometry problem, with high-warm or low-warm reactions to the student's explanation. The 'user' prompt is the self-explanation that students wrote according to GeomPT's request at the stage 6. Below is the 'system' prompt to generate an intended message, which comes after the prompt to generate either high-warm or low-warm message.

Given user's explanation, you need to react to user's effort and lead them to the question number N, where you need to solve the given problem and ask users for feedback. The first one or two sentences should be about reacting to user's explanation and the last sentence should be about leading users to the next question. You should not explain what the next problem will be about. For example, do not say like "The next question will be about finding the area of a triangle", just say "Next is question number 7".

Examples of each of the high-warm and low-warm versions of the messages generated in the real experiment, given students' explanations are as follows:

(HW) Wow, that's a really cool explanation! 🙌 Next is question number 4. Try this one out and tell me how you solved it!! 😊
 (LW) Umm okay. Next is question number 4.

4.2.2. Reciprocal Design

To examine the effect of reciprocal warmth design for the pedagogical conversational agent, we designed the flow of the experimental condition (HLW-C)'s GeomPT by switching some of high-warm messages of the control condition (HW-C)'s GeomPT to be low-warm. Given that HW-C version of GeomBot promised improvement in learning, based on the evidence from Study 1, we adopted the conversational flow of HW-C and set its automated version, HW-C GeomPT, as the agent for the control condition.

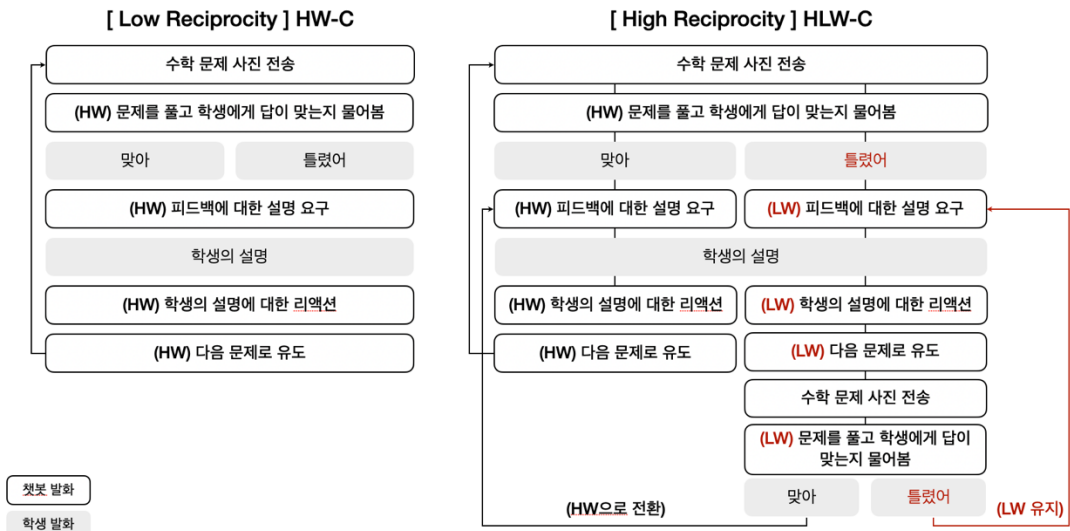


Figure 10. The difference in conversational flow between HW-C GeomPT and HLW-C GeomPT

For HLW-C GeomPT, the agent for the experimental condition, the point at which the warmth of the message is switched from high to low is derived from the qualitative result of Study 1. We referred to the students' comment that they thought GeomBot is honest when it talks low-warmly when they gave feedback that its answer is 'incorrect'. We thus designed HLW-C GeomPT by changing the warmth of HW-C GeomPT's messages to low-warm when it received the feedback that its answer is incorrect from students. The message of HLW-C GeomPT, which has become low-warm, becomes high-warm again when it receives feedback that its answer is correct. The difference in conversational flow between HW-C GeomPT and HLW-C GeomPT is presented in Figure 10.

4.3. Method

4.3.1. Participants

Upon approval from the Institutional Review Board at Seoul National University (IRB No. 2211/002-024), we recruited 10 6th-grade elementary school students from online communities in South Korea. The students were randomly assigned to one of two conditions. Per condition, the following number of participants were assigned: 5 in HW-C and 5 in HLW-C. Across the two conditions, there was no statistical difference in pre-test score, Affinity for Technology Interaction (ATI), and mathematics-related affect. Detailed descriptive statistics on demographic information can be found in Table 1 and Table 2. We provided \$20 to each student who participated in the study.

P#	Condition	Gender	Age	CAs usage	Pre-test	ATI
1	HW-C	Girl	12	No	4	3.5
2	HW-C	Boy	12	Yes	10	5.1
3	HW-C	Boy	12	No	10	2.3
4	HW-C	Girl	12	No	8	4.6

5	HW-C	Girl	12	No	5	2.2
6	HLW-C	Boy	12	No	10	5
7	HLW-C	Girl	12	No	7	3.5
8	HLW-C	Girl	12	Yes	2	3.1
9	HLW-C	Boy	12	No	9	5
10	HLW-C	Boy	12	No	9	4.5

*ATI: Affinity for Technology Interaction

* HW-C: High-Warm messages with Constructive learning activities

* HLW-C: High-Warm and Low-Warm messages alternately provided with
Constructive learning activities

Table 1. Demographic Information of Participants in Study 2 (1)

P#	Condition	Mathematics-related affect					
		Competence	Confidence	DoM	EoM	MGO	Effort
1	HW-C	6	6	5	11	19	9
2	HW-C	20	20	15	19	24	20
3	HW-C	15	15	10	16	20	10
4	HW-C	19	20	14	23	17	15
5	HW-C	15	16	8	13	20	11
6	HLW-C	20	18	14	20	23	15
7	HLW-C	9	10	11	13	19	8
8	HLW-C	10	12	9	16	18	10
9	HLW-C	19	19	12	21	22	17
10	HLW-C	16	18	3	13	20	13

*DoM: Difficulty of Mathematics, EoM: Enjoyment of Mathematics

MGO: Mastery Goal Orientation

* HW-C: High-Warm messages with Constructive learning activities

* HLW-C: High-Warm and Low-Warm messages alternately provided with
Constructive learning activities

Table 2. Demographic Information of Participants in Study 2 (2)

4.3.2. Measurement

4.3.2.1. Control Variables

The control variables measured in Study 2 are the same as those

measured in Study 1, which are Affinity for Technology Interaction (ATI), Mathematics-related affect, and Geometry problem-solving skills.

4.3.2.2. Dependent Variables

A dependent variable, school trust with five dimensions, which was measured in Study 1 was also measured in Study 2. In Study 2, perceived reciprocity is additionally measured.

Perceived reciprocity. We adopted a seven-point Likert scale questionnaire from Lee's study to measure how reciprocal the students felt GeomPT is (Lee & Choi, 2017). This seven-point Likert scale questionnaire consists of six questions and yields a total score between 6 and 42 points. The sample questions are, "Harang gave good responses to you", "I think Harang and I were able to help each other", and "I think Harang and I exchanged opinions as though we were equal in our social status".

4.3.3. Procedure

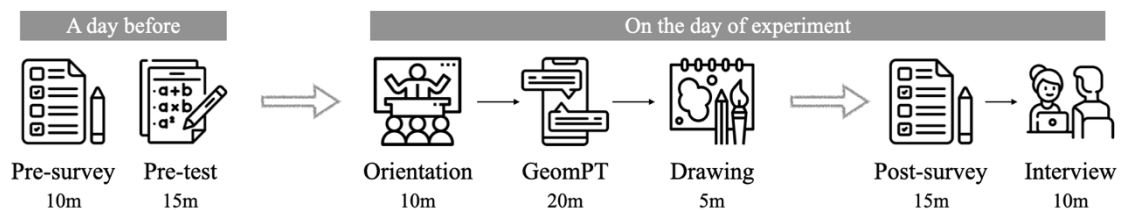


Figure 11. Procedure and measurement

As in the Study 1, students started the study with completing a pre-test and two pre-experiment questionnaires, an affinity for technology interaction (ATI) scale and a questionnaire on mathematics-related affect, a day before the experiment.

The process on the day of the experiment is also the same as the

Study 1. Students first took a 10-minute orientation session, and then used GeomPT for 20 minutes. When solving geometry problems with GeomPT, they were provided with a cheat sheet which can be used to check if their explanation is correct. Right after the GeomPT usage, they were asked to draw what they thought GeomPT would look like. Students then responded to three post-experiment questionnaires, perceived reciprocity and perceived trust.

As a last step of the experiment, students took a post-test for 15 minutes and 15-minute post-interview was conducted to collect detailed elaboration of students' drawings and their experience of using GeomPT. The post-interview consists of three themes: (1) The first theme of the interview examined students' intention on their drawings. Given that drawing data contains students' states of mind and internal perceptions (Xu & Warschauer, 2020), we asked students to elucidate their drawings in detail so that we can catch their internal perceptions towards GeomPT which students feel difficult to express through words (Chan, 2006). Afterward, we asked a few more follow-up questions based on students' comments. (2) The second theme of the interview was regarding reciprocity, and the following questions were asked: "Do you think that Harang revealed its thoughts and feelings?", "Were you also able to express your thoughts and feelings to Harang?". (3) The third theme was agent perception. Sample questions include "Did you think Harang was honest?".

4.3.4. Analysis

The two main goals of Study 2 are to explore (1) whether the students feel the reciprocal design of GeomPT reciprocal and (2) whether honesty perceptions are improved when using reciprocity-enhanced version of GeomPT. For quantitative analysis, due to the small sample size ($n = 10$), we used a Mann-Whitney U test, a nonparametric test that compares independent two samples to achieve both first and

second goals. We compared the perceived reciprocity score and the perceived honesty score between HW-C and HLW-C group.

For qualitative analysis, students' drawings and data from the post-interview was analyzed using an iterative open coding method (Corbin & Strauss, 2014). The interview for each student consisted of two phases: (1) elucidating drawings and (2) describe the experience of using GeomPT. Three coders transcribed and analyzed 2.5 hours of interview recording, which consisted of 15 minutes for each of the 10 students. For the drawing data, the coders we annotated each student's drawing based on the student's verbal explanation. First, the coders independently coded all the transcripts and the drawings with the accompanying verbal accounts line-by-line and created initial categories using an inductive approach. Next, the coders merged similar codes and formed 8 higher-level categories. For the drawing data, the following three categories were formulated: (1) whether Harang is close to human or machine, (2) whether students feel Harang as a peer or a younger, (3) how smart students thought Harang is.

Subsequently, in the next iteration of open coding, these codes and categories were applied to the transcripts and used for qualitative analysis with regard to each hypothesis. Cohen's Kappa coefficient (McHugh, 2012) was calculated to measure inter-coder reliability. An agreement level of 0.81 was reached, suggesting a good agreement between the two coders.

4.4. Results

4.4.1. RQ4: Would reciprocal warmth design improve perceived honesty towards pedagogical conversational agent?

Mann-Whitney U test result showed that hypothesis **H4** was satisfied, with students in HLW-C group showing significantly higher perceived honesty ($W = 2$, $p = 0.02^*$) compared to students in HW-C group. The test result shows that learning with HLW-C GeomPT results in

students' significantly better perception of honesty to the agent when compared to learning with HW-C GeomPT, which indicates the positive impact of reciprocal warmth design of honesty perception. All the students who used HLW-C GeompPT also commented on the honesty of the pedagogical conversational agent which they felt during the experiment, which was due to the intended reciprocal design of the conversation. Furthermore, we could observe that such perceived reciprocity also enhanced students' willingness to self-disclose to the agent.

*"Since Harang reacted honestly to my feedback first, I was able to feel more comfortable to express my thoughts honestly."
(HLW-C_8)*

On the other hand, students who used HW-C GeomPT reported negative perception in regard to the honesty of the pedagogical conversational agent due to low-reciprocity of HW-C GeomPT, requesting a more reciprocal version of GeomPT. Furthermore, we could observe that such perceived low-reciprocity made students feel difficult to self-disclose to the agent.

"I thought Harang was a little pretentious and dishonest since it gave me too many compliments even though I said it was wrong or kept giving the same explanation to it." (HW-C_1)

"Whether I said it was right or wrong, Harang just kept praising me. I actually felt good since it seemed to trust me. However, it didn't seem like honest, and I also got to hide my own opinions, rather than directly expressing what I thought. It would be better if Harang expressed its opinion and opposed once if it thought I am wrong." (HW-C_2)

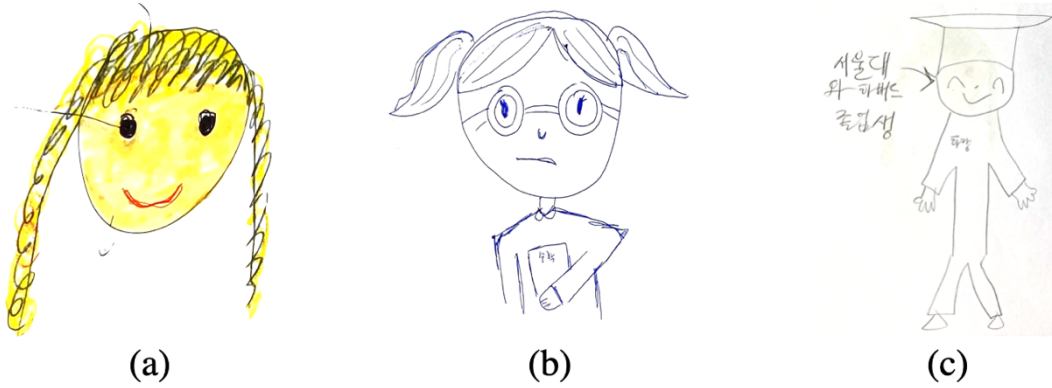


Figure 12. Drawings that illustrate HLW-C GeomPT: drawn by *HLW-C_8*, *HLW-C_7*, and *HLW-C_10* in an order

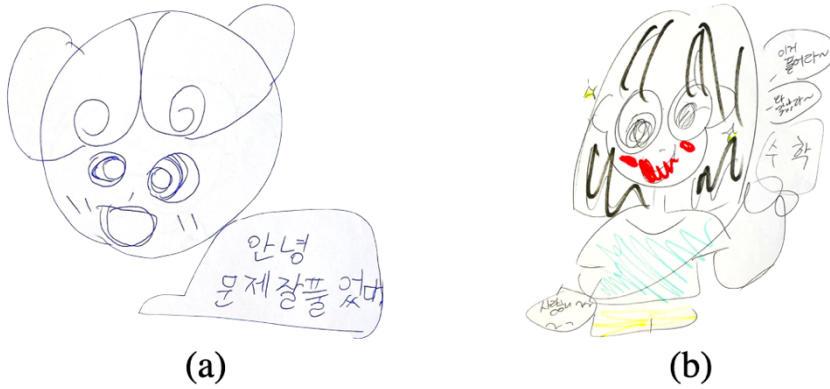


Figure 13. Drawings that illustrate HW-C GeomPT: drawn by *HW-C_3* and *HW-C_2* in an order

In addition to the perceived honesty toward high or low-reciprocal GeomPT, we could observe interesting trends regarding the impact of reciprocity on other aspects of perception, from the qualitative analysis on students' drawing data, as follows: (1) the perception of humanness, (2) the perceived age, and (3) the perceived competence. First, students who used HLW-C GeomPT tended to perceive the pedagogical conversational agent to be closer to human than students who used HW-C GeomPT. All three figures of Figure 12, which presents the drawings from HLW-C group, shows human-like drawings, compared to drawings from Figure 13 that presents the

drawings from HW-C group.

"I thought Harang was like a person. I felt like I was messaging with a real person. I think that's why I drew Harang close to a person." (HLW-C_8)

"I got to draw Harang as a virtual character or avatar since I thought Harang was more like a machine than a human. I thought Harang might look like a virtual character that we can usually see on the internet or in games." (HW-C_3)

Second, students in HLW-C group, which was designed to be high-reciprocal, were likely to feel GeomPT was like their peers, whereas students in HW-C group, which was designed to be low-reciprocal, tended to feel GeomPT was younger than them. The drawings from Figure 12 which were drawn by students in HLW-C and Figure 13 which were drawn by students in HW-C explicitly reveal the difference between the age of GeomPT that students perceived.

"I just felt like there must be at least one smart friend like Harang among my friends. So, I got to describe Harang as a friend of the same age as me." (HLW-C_7)

"I tried to express Harang like a younger brother who is full of curiosity. I thought Harang was like a kindergartner who hadn't yet learned how to be polite." (HW-C_2)

Third, students who used HLW-C GeomPT tended to think that the agent is smart and competence, but students in HW-C group felt the agent is unintelligent and incompetent. GeomPT is illustrated as an intelligent and competent student, such as a "Harvard graduate" in Figure 12(b) and Figure 12(c). On the other hand, GeomPT's appearances, illustrated in Figure 13(a) and Figure 13(b), were coded

as silly and incompetent by all the coders, which were aligned with the comments of the students.

"I thought Harang would go to Seoul National University or Harvard University when we are going to university, so I expressed this in the picture. I thought Harang is smart since it was good at solving problems and speaking out its thoughts."
(HLW-C_10)

"I came to think that Harang's level of knowledge is lower than mine because it always agreed with me, regardless of whether my explanation is correct or not." (HW-C_2)

In summary, regarding the impact of reciprocal warmth design on the perceived honesty of the pedagogical conversational agent, our data analysis supports **H4**, which showed a positive impact of reciprocal conversation design, which is to switch the warmth of the messages adaptively, on perceived honesty. Furthermore, we could find the additional findings from the drawing data and post-interview data, which suggest that the different perceptions, i.e., humanness, age, and competence, towards the pedagogical conversational agent were observed depending on the reciprocity of the agent.

4.5. Discussion

4.5.1. Use of Generative AI in a pedagogical conversational agent

The messages of pedagogical conversational agents for both conditions were generated with ChatGPT, one of the LLMs that has been prevalently used recently. However, since the current stage of LLMs has limitations such as hallucination and uncontrollability, there is a need to carefully consider utilizing such models, especially in educational settings. Based on the findings and insights from Study 2,

we suggest guidelines to consider when using LLMs for the learning content generation, focusing on the warmth manipulation of the message and constructive learning activity design. First, we present the prompting engineering strategies that can be used when generating the learning content.

Specify the context in which the user and model interact. We recommend including the information about who the model is having conversation with and for what purpose, and which role the model is supposed to play to 'system' prompt. Such prompting allowed us to prevent the model from generating inappropriate sentences in the learning context with elementary school students.

Provide formulas to generate correct answers. In order to ensure that the model correctly solves elementary-school level geometry problems, including the formulas to 'system' prompt will achieve an approximately 100% accuracy. In this study, for the problems of finding the area of the trapezoid, parallelogram, and rhombus, ChatGPT sometimes generated incorrect answers without a specific pattern before the formulas were provided. However, after the formulas were given, it never generated incorrect answers at least during the experiment.

A trick to generate incorrect answers. To generate incorrect answers to elementary-level geometry problems was tricky, due to the instability of generation. Therefore, in this study, we provided different problems to the model from the one provided to students. Utilizing the fact that the correct answers are accurately generated when the formulas are provided, we transformed the original problems into problems whose correct answers become students' frequent incorrect answers for the original problems. We were so able to generate incorrect answers with the perfect accuracy at least during the experiment. However, since such method is not a standard method of prompting, investigating prompting engineering techniques to generate incorrect answers might be an interesting future study to pursue.

Second, we propose an instructional conversation flow that might complement the limitations of using generative models such as hallucination or ethical issues. Such limitations can be problematic in learning context, especially for younger students. The instructional conversation flow of this study, which applies learning-by-teaching by making students convey information to the agent, rather than the agent delivering information to the students, can be a potential design to utilize generative models in the learning context. However, to further enhance students' engagement, investigating how to improve the robustness of the generative models so that interactive exchanges of thoughts can be enabled without ethical issues should be considered. Such future research may enable the implementation of an interactive mode of learning, which maximizes students' cognitive engagement (Chi & Wylie, 2014), for the pedagogical conversational agents.

Third, we suggest considering reciprocity when using generative models, especially ChatGPT, which aims to generate human-like text in a conversational style (Cotton et al., n.d.). Our study results indicated that the pedagogical conversational agent is perceived as more like a machine than a human if its messages are generated to be high-warm only. In contrast, the pedagogical conversational agent whose reciprocity is improved by alternating the warmth level of messages was perceived as being close to humans, especially peers. Taking these into account, prompting the agent to socially and emotionally reciprocate the user's behavior might assist in achieving the goal of generative models to simulate interpersonal conversations.

Since the target of this study was elementary school students, we prompted the model to simulate the real conversation among them, which in turn resulted in support for our hypotheses. The results of this study can be expanded to a wider range of contexts when future researchers provide the model with a more general pattern of real conversation. Therefore, to simulate real conversation among humans, we recommend prompting the model in consideration of reciprocity when using language models in various contexts.

4.5.2. Reciprocal design

The reciprocal design that is suggested by study is to reciprocate the warmth of agent message to students' feedback (i.e., sending low-warm messages when receiving 'incorrect' feedback and sending high-warm messages when receiving 'correct' feedback by students). Such the validity of such reciprocal design was checked by comparing perceived reciprocity score between HW-C group and HLW-C group in Section 4.2.2. Considering that HW-C GeomPT is the autonomously generated version of HW-C GeomBot from Study 1, the manipulation check result indicates that simply changing generation methods from manual generation to autonomous generation may not help improve reciprocity.

The positive impact of reciprocal warmth design on honesty perception, as hypothesized in H4, was supported by both quantitative and qualitative data. The study results expanded previous findings regarding the relationship between reciprocity and honesty from human-human interaction (Douthit & Stevens, 2015; Van Lange, 1998). The findings from this study, which suggest that students perceived honesty towards highly reciprocal agent, provide potential evidence to expand the result from human-human interaction context to the context of human-agent interaction, in line with CASA paradigm (Reeves & Nass, 1996).

Furthermore, we could discover such perceived high-reciprocity helped students to express their thoughts and opinions to the agent. Students who studied with HLW-C GeomPT reported that they felt comfortable to express their thoughts honestly. In contrast, students who studied with HW-C GeomPT reported that they tended to hide their own opinions, rather than directly expressing what they thought. Such tendencies indicate that the reciprocity of the pedagogical conversational agent might play a role in promoting self-disclosure. According to the literature regarding self-disclosure, highly reciprocal GeomPT (HLW-C) can foster intermediate level of self-disclosure,

which includes opinions, attitudes, and values (Altman & Taylor, 1973). To sum up, this study provides empirical evidence that reciprocating the level of the warmth of pedagogical conversational agent’s messages can enhance honesty perception and promote student’s self-disclosure.

In addition to the positive impact on perceived honesty and self-disclosure, we were able to discover additional interesting results regarding the other dimension of agent perception that reciprocal warmth design might have influenced: (1) the perception of humanness, (2) the perceived age, and (3) the perceived competence. Especially, for the perception of humanness, the finding from this study expands the existing literature which revealed that the reciprocity creates the illusion that the agent is realistic (Becker & Mark, 1999; Lee & Choi, 2017), by narrowing down the general finding regarding realistic agent to human-like agent in the educational setting. From the above-mentioned three dimensions of agent perception that reciprocal design might have impacted, we suggest future researchers to reciprocate the warmth of the agent to students’ messages when designing a human-like, peer-like, or competent conversational agent. Even though the current study is restricted to the context of elementary school student’s math learning, considering that the measured dependent variables can be generally applied to other contexts, we conjecture that the results of this study may have the potential to be generalized to embrace wider range of target users, domains, and contexts.

Chapter 5. Conclusion

This paper presents a novel study that examined the effect of the text-based pedagogical conversational agent on learning experiences, focusing of three elements: (1) warmth of the message, (2) type of learning activity, (3) reciprocal warmth design. Through an in-situ experiment, we explored 6th-grade student's learning achievement, intrinsic motivation on learning, and agent perception. The result of our study indicated that warmth of the message and type of learning activity, respectively, positively impacted learning achievement and interest-enjoyment in learning. Additionally, doing constructive activities along with high-warm messages has a positive effect on learning achievement, improving interest-enjoyment in learning, and reducing tension-pressure. Furthermore, reciprocating the warmth of the message to student's messages improved perceived honesty toward the agent.

However, our study has several limitations. First, we defined 'warmth' as one of multiple definitions from previous studies which fits the context of this study the most. Since warmth consists of various sub-dimensions, such as friendly, warm, thoughtful, well-intentioned, generous, and honest (Stanciu et al., 2017), there exists a potential confounding effect of another dimensions of warmth that were not defined in this study. Therefore, when interpreting the result of this study, future researchers need to keep in mind that there might be a potential confounding effect.

Second, since the experiments from both Study 1 and Study 2 were conducted during a single session, we could not fully eliminate the novelty effect on the significant results. Third, since learning algorithm is not applied to the pedagogical conversational agents of this study, the knowledge of the agents does not improve even though they ask for the feedback to their answer. In the future study, by applying learning algorithm to the agent, the advantages of the instructional conversational flow of GeomBot and GeomPT which were

proved by the study can be amplified.

Despite such limitations, our study is meaningful in that we examined an unexplored social aspect of the agent, warmth, in a text-based conversational agent. Based on our study results, there are three points to consider when developing a text-based pedagogical conversational agent. First, one should consider using positive statements and smiling emojis to make the text-based conversational agent highly warm. Secondly, in order to enhance the positive effects of warm messages, it is recommended to provide constructive learning activities which enable students to write down their thoughts on their own. Last but not least, it is worthwhile to consider reciprocating the level of agent warmth to student's messages so that students can feel comfortable to express their thoughts honestly.

References

- Adams, C., Beulah, A. O., & Fiegenger, A. (2022). Student trust in school peers: a relational condition for optimal school functioning. *Journal of Educational Administration and History*, 60(6), 545–560.
- Altman, I., & Taylor, D. A. (1973). *Social penetration: The development of interpersonal relationships*. 212. <https://psycnet.apa.org/fulltext/1973-28661-000.pdf>
- Ba, S., Stein, D., Liu, Q., Long, T., Xie, K., & Wu, L. (2021). Examining the Effects of a Pedagogical Agent With Dual-Channel Emotional Cues on Learner Emotions, Cognitive Load, and Knowledge Transfer Performance. *Journal of Educational Computing Research*, 59(6), 1114–1134.
- Bayes, M. A. (1972). Behavioral cues of interpersonal warmth. *Journal of Consulting and Clinical Psychology*, 39(2), 333–339.
- Becker, B., & Mark, G. (1999). Constructing social systems through computer-mediated communication. *Virtual Reality*, 4(1), 60–73.
- Best, J. B., & Addison, W. E. (2000). A preliminary study of perceived warmth of professor and student evaluations. *Teaching of Psychology*, 27(2000), 60–62.
- Biancardi, B., Cafaro, A., & Pelachaud, C. (2017a). Analyzing first impressions of warmth and competence from observable nonverbal cues in expert–novice interactions. *Proceedings of the 19th ACM International Conference on Multimodal Interaction*, 341–349.
- Biancardi, B., Cafaro, A., & Pelachaud, C. (2017b). Could a virtual agent be warm and competent? investigating user’s impressions of agent’s non-verbal behaviours. *Proceedings of the 1st ACM SIGCHI International Workshop on Investigating Social Interactions with Artificial Agents*, 22–24.

- Biswas, G., Leelawong, K., Belyne, K., Viswanath, K., Vye, N., Schwartz, D., & Davis, J. (2004). Incorporating self regulated learning techniques into learning by teaching environments. *Proceedings of the Annual Meeting of the Cognitive Science Society*, 26. <https://escholarship.org/content/qt5s81207k/qt5s81207k.pdf>
- Biswas, G., Leelawong, K., Schwartz, D., Vye, N., & The Teachable Agents Group at Vanderbilt. (2005). LEARNING BY TEACHING: A NEW AGENT PARADIGM FOR EDUCATIONAL SOFTWARE. *Applied Artificial Intelligence: AAI*, 19(3-4), 363-392.
- Biswas, G., Schwartz, D., Bransford, J., & Vanderbilt, Teachable Agents Group at. (2001). Technology support for complex problem solving: From SAD environments to AI. In K. D. Forbus (Ed.), *Smart machines in education: The coming revolution in educational technology*, (pp (Vol. 483, pp. 71-97). The MIT Press, vi.
- Bordin, E. S. (1951). Four uses for psychological tests in counseling. *Educational and Psychological Measurement*, 11(4_part_2), 779-781.
- Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., & Others. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33, 1877-1901.
- Carbajal, I., Hughes, J. S., & Hughes, C. &. (2016). *Student evaluations of instructor warmth and competence: Course difficulty counts more than character.* [jiss.org. https://jiss.org/documents/volume_6/JISS%202016%206\(1\)%201-16%20Student%20Evaluations.pdf](https://jiss.org/documents/volume_6/JISS%202016%206(1)%201-16%20Student%20Evaluations.pdf)
- Ceha, J., Lee, K. J., Nilsen, E., Goh, J., & Law, E. (2021). Can a Humorous Conversational Agent Enhance Learning Experience and Outcomes? *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1-14.
- Chan, K. (2006). Exploring children's perceptions of material

- possessions: a drawing study. *Qualitative Market Research: An International Journal*, 9(4), 352–366.
- Chartrand, T. L., & van Baaren, R. (2009). Chapter 5 Human Mimicry. In *Advances in Experimental Social Psychology* (Vol. 41, pp. 219–274). Academic Press.
- Chi, M. T. H., & Wylie, R. (2014). The ICAP Framework: Linking Cognitive Engagement to Active Learning Outcomes. *Educational Psychologist*, 49(4), 219–243.
- Corbin, J., & Strauss, A. (2014). *Basics of Qualitative Research: Techniques and Procedures for Developing Grounded Theory*. SAGE Publications.
- Cotton, D. R. E., Cotton, P. A., & Shipway, J. R. (n.d.). *Chatting and Cheating: Ensuring academic integrity in the era of ChatGPT*. Retrieved June 20, 2023, from <https://edarxiv.org/mrz8h/download/?format=pdf>
- Cuddy, A. J. C., Fiske, S. T., & Glick, P. (2008). Warmth and Competence as Universal Dimensions of Social Perception: The Stereotype Content Model and the BIAS Map. In *Advances in Experimental Social Psychology* (Vol. 40, pp. 61–149). Academic Press.
- Cuddy, A. J. C., Glick, P., & Beninger, A. (2011). The dynamics of warmth and competence judgments, and their outcomes in organizations. *Research in Organizational Behavior*, 31, 73–98.
- De Sixte, R., Mañá, A., Ávila, V., & Sánchez, E. (2020). Warm elaborated feedback. Exploring its benefits on post-feedback behaviour. *Educational Psychology Review*, 40(9), 1094–1112.
- Do, J., & Paik, S. (2017). Analysis of Affective Factors in Mathematics Learning of Elementary School Students. *C-초등수학교육*, 20(4), 287–303.
- Domagk, S. (2008). *Pädagogische Agenten in Multimedialen Lernumgebungen: Empirische Studien zum Einfluss der Sympathie auf Motivation und Lernerfolg*. Logos-Verlag.

- dos Santos Alencar, M. A., & de Magalhães Netto, J. F. (2020). Improving Learning in Virtual Learning Environments Using Affective Pedagogical Agent. *International Journal of Distance Education Technologies (IJDET)*, 18(4), 1–16.
- Douthit, J. D., & Stevens, D. E. (2015). The robustness of honesty effects on budget proposals when the superior has rejection authority. *The Accounting Review*, 90(2), 467–493.
- Driscoll, M. P. (1994). *Psychology of learning for instruction*. 409. <https://psycnet.apa.org/fulltext/1993-99148-000.pdf>
- Fiske, S. T. (2018). Stereotype Content: Warmth and Competence Endure. *Current Directions in Psychological Science*, 27(2), 67–73.
- Fiske, S. T., Cuddy, A. J. C., & Glick, P. (2007). Universal dimensions of social cognition: warmth and competence. *Trends in Cognitive Sciences*, 11(2), 77–83.
- Fiske, S. T., Cuddy, A. J. C., Glick, P., & Xu, J. (2002). A model of (often mixed) stereotype content: Competence and warmth respectively follow from perceived status and competition. *Journal of Personality and Social Psychology*, 82(6), 878–902.
- Franke, T., Attig, C., & Wessel, D. (2019). A Personal Resource for Technology Interaction: Development and Validation of the Affinity for Technology Interaction (ATI) Scale. *International Journal of Human–Computer Interaction*, 35(6), 456–467.
- Gilad, Z., Amir, O., & Levontin, L. (2021). The Effects of Warmth and Competence Perceptions on Users' Choice of an AI System. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–13.
- Gorham, J. (1988). The relationship between verbal teacher immediacy behaviors and student learning. *Communication Education*, 37(1), 40–53.
- Guice, K. (2016). *Predators, decoys, and teens: A corpus analysis of*

- online language* [search.proquest.com].
<https://search.proquest.com/openview/8910ec1cdb0bccf360e8a785543a142d/1?pq-origsite=gscholar&cbl=18750>
- Guo, Y. R., & Goh, D. H.-L. (2016). Evaluation of affective embodied agents in an information literacy game. *Computers & Education, 103*, 59–75.
- Ha, J., Pérez Cortés, L. E., Su, M., Nelson, B. C., Bowman, C., & Bowman, J. D. (2021). The impact of a gamified mobile question-asking app on museum visitor group interactions: an ICAP framing. *International Journal of Computer-Supported Collaborative Learning, 16*(3), 367–401.
- Hannula, M. S. (2012). Exploring new dimensions of mathematics-related affect: embodied and social theories. *Research in Mathematics Education, 14*(2), 137–161.
- Hayes, A. F. (2017). [No title]. Guilford publications.
<https://www.researchgate.net/profile/Adasa-Nkrumah/post/Can-anyone-tell-me-that-which-one-is-better-for-running-mediation-and-moderation-Process-macro-by-Hayes-or-AMOS/attachment/5e74fcd23843b0047b366238/AS%3A871214797045762%401584725202506/download/templates.pdf>
- Hobert, S., & Meyer von Wolff, R. (2019). Say Hello to Your New Automated Tutor – A Structured Literature Review on Pedagogical Conversational Agents. *Wirtschaftsinformatik 2019 Proceedings*.
<https://aisel.aisnet.org/wi2019/track04/papers/2/>
- Howe, L. C., Leibowitz, K. A., & Crum, A. J. (2019). When Your Doctor "Gets It" and "Gets You": The Critical Role of Competence and Warmth in the Patient–Provider Interaction. *Frontiers in Psychiatry / Frontiers Research Foundation, 10*.
<https://doi.org/10.3389/fpsy.2019.00475>
- Jiang, E., Olson, K., Toh, E., Molina, A., Donsbach, A., Terry, M., & Cai,

- C. J. (2022, April 27). PromptMaker: Prompt-based prototyping with large language models. *CHI Conference on Human Factors in Computing Systems Extended Abstracts*. CHI '22: CHI Conference on Human Factors in Computing Systems, New Orleans LA USA. <https://doi.org/10.1145/3491101.3503564>
- Johnson, C. I., & Mayer, R. E. (2010). Applying the self-explanation principle to multimedia learning in a computer-based game-like environment. *Computers in Human Behavior*, 26(6), 1246–1252.
- Jung, J.-Y., Qiu, S., Bozzon, A., & Gadiraju, U. (2022). Great Chain of Agents: The Role of Metaphorical Representation of Agents in Conversational Crowdsourcing. *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, 1–22.
- Kervyn, N., Fiske, S. T., & Malone, C. (2022). Social perception of brands: Warmth and competence define images of both brands and social groups. *Consumer Psychology Review*, 5(1), 51–68.
- Khadpe, P., Krishna, R., Fei-Fei, L., Hancock, J. T., & Bernstein, M. S. (2020). Conceptual Metaphors Impact Perceptions of Human-AI Collaboration. *Proc. ACM Hum.-Comput. Interact.*, 4(CSCW2), 1–26.
- Kim, W. B., & Hur, H. J. (2023). What Makes People Feel Empathy for AI Chatbots? Assessing the Role of Competence and Warmth. *International Journal of Human-Computer Interaction*, 1–14.
- Kim, Y., & Baylor, A. L. (2006). A social-cognitive framework for pedagogical agents as learning companions. *Educational Technology Research and Development: ETR & D*, 54(6), 569–596.
- Kim, Y., Baylor, A. L., & PALS Group. (2006). Pedagogical agents as learning companions: The role of agent competency and type of interaction. *Educational Technology Research and Development: ETR & D*, 54(3), 223–243.
- Kuhail, M. A., Alturki, N., Alramlawi, S., & Alhejori, K. (2023). Interacting with educational chatbots: A systematic review.

- Education and Information Technologies*, 28(1), 973–1018.
- Kulesza, W., Dolinski, D., Huisman, A., & Majewski, R. (2014). The Echo Effect: The Power of Verbal Mimicry to Influence Prosocial Behavior. *Journal of Language and Social Psychology*, 33(2), 183–201.
- Kulms, P., & Kopp, S. (2018). A Social Cognition Perspective on Human–Computer Trust: The Effect of Perceived Warmth and Competence on Trust in Decision-Making With Computers. *Frontiers in Digital Humanities*, 5. <https://doi.org/10.3389/fdigh.2018.00014>
- Lawson, A. P., & Mayer, R. E. (2022). Does the emotional stance of human and virtual instructors in instructional videos affect learning processes and outcomes? *Contemporary Educational Psychology*, 70, 102080.
- Lee, S., & Choi, J. (2017). Enhancing user experience with conversational agent for movie recommendation: Effects of self-disclosure and reciprocity. *International Journal of Human–Computer Studies*, 103, 95–105.
- Li, Z., Wang, L., & Zhang, L. (2012). Exploratory and confirmatory factor analysis of a short-form of the EMBU among Chinese adolescents. *Psychological Reports*, 110(1), 263–275.
- Liew, T. W., Mat Zin, N. A., & Sahari, N. (2017). Exploring the affective, motivational and cognitive effects of pedagogical agent enthusiasm in a multimedia learning environment. *Human–Centric Computing and Information Sciences*, 7(1), 1–21.
- Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., & Neubig, G. (2023). Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing. *ACM Comput. Surv.*, 55(9), 1–35.
- Locke, E. A. (1997). Self-efficacy: The exercise of control. *Personnel Psychology*. <https://search.proquest.com/openview/55c56d1a75f8440c4bea>

- Maricchiolo, F., Gnisci, A., Bonaiuto, M., & Ficca, G. (2009). Effects of different types of hand gestures in persuasive speech on receivers' evaluations. *Language and Cognitive Processes*, 24(2), 239–266.
- Markel, J. M., Opferman, S. G., Landay, J. A., & Piech, C. (2023). *GPTeach: Interactive TA Training with GPT Based Students*. <https://edrxiv.org/r23bu/download?format=pdf>
- Martha, A. S. D., & Santoso, H. (2019). The design and impact of the pedagogical agent: A systematic literature review. *The Journal of Educators Online*, 16(1). <https://doi.org/10.9743/jeo.2019.16.1.8>
- McHugh, M. L. (2012). Interrater reliability: the kappa statistic. *Biochemia Medica: Casopis Hrvatskoga Drustva Medicinskih Biokemicara / HDMB*, 22(3), 276–282.
- McKee, K. R., Bai, X., & Fiske, S. T. (2022). Warmth and competence in human-agent cooperation. In *arXiv [cs.HC]*. arXiv. <http://arxiv.org/abs/2201.13448>
- McNeill, L., Rice, M., & Wright, V. (2019). An exploratory factor analysis of a teaching presence instrument and the ICAP framework in an online computer applications course. *Global Learn*, 310–317.
- Molinillo, S., Aguilar-Illescas, R., Anaya-Sánchez, R., & Vallespín-Arán, M. (2018). Exploring the impacts of interactions, social presence and emotional engagement on active collaborative learning in a social web-based environment. *Computers & Education*, 123, 41–52.
- Nguyen, T.-H. D., Carstensdottir, E., Ngo, N., El-Nasr, M. S., Gray, M., Isaacowitz, D., & Desteno, D. (2015). Modeling Warmth and Competence in Virtual Characters. *Intelligent Virtual Agents*, 167–180.
- Noh, Y.-G., & Hong, J.-H. (2021). Designing Reenacted Chatbots to

- Enhance Museum Experience. *NATO Advanced Science Institutes Series E: Applied Sciences*, 11(16), 7420.
- Oliveira, R., Arriaga, P., Correia, F., & Paiva, A. (2019). The Stereotype Content Model Applied to Human-Robot Interactions in Groups. *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 123–132.
- Pace, A., & Gnisci, A. (2019). Some gestures are better than others: The warmth and competence effect of hand gestures in interaction. *Psicologia Sociale*. <https://doi.org/10.1482/94269>
- Piaget, J., & Smith, L. (2013). *Sociological studies*. <https://www.taylorfrancis.com/books/mono/10.4324/9780203714065/sociological-studies-jean-piaget-leslie-smith>
- Reece, M. M., & Whitman, R. N. (1962). Expressive movements, warmth, and verbal reinforcement. *Journal of Abnormal and Social Psychology*, 64, 234–236.
- Reeves, B., & Nass, C. (1996). The media equation: How people treat computers, television, and new media like real people. *Cambridge, UK*, 10, 236605.
- Ruan, S., Jiang, L., Xu, J., Tham, B. J.-K., Qiu, Z., Zhu, Y., Murnane, E. L., Brunskill, E., & Landay, J. A. (2019). QuizBot: A Dialogue-based Adaptive Learning System for Factual Knowledge. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–13.
- Seibel, D. W. (1955). *The prediction of qualities of interaction between apprentice teachers and pupils*. Harvard Graduate School of Education.
- Smutny, P., & Schreiberova, P. (2020). Chatbots for learning: A review of educational chatbots for the Facebook Messenger. *Computers & Education*, 151, 103862.
- Stanciu, A., Cohrs, J. C., Hanke, K., & Gavreliuc, A. (2017). Within-culture variation in the content of stereotypes: Application and development of the stereotype content model in an Eastern European culture. *The Journal of Social Psychology*, 157(5),

611–628.

- Tschannen-Moran, M. (2014). *Trust Matters: Leadership for Successful Schools*. John Wiley & Sons.
- Tschannen-Moran, M., & Hoy, W. K. (2000). A Multidisciplinary Analysis of the Nature, Meaning, and Measurement of Trust. *Review of Educational Research*, 70(4), 547–593.
- Tuohilampi, L., Hannula, M. S., Varas, L., Giaconi, V., Laine, A., Näveri, L., & i Nevado, L. S. (2015). Challenging the western approach to cultural comparisons: Young pupils' affective structures regarding mathematics in Finland and Chile. *International Journal of Mathematical Education in Science and Technology*, 13(6), 1625–1648.
- Van Lange, P. A. M. (1998). The boundaries of reciprocal cooperation. *European Journal Of*. [https://doi.org/10.1002/\(SICI\)1099-0992\(199809/10\)28:5<847::AID-EJSP886>3.0.CO;2-L](https://doi.org/10.1002/(SICI)1099-0992(199809/10)28:5<847::AID-EJSP886>3.0.CO;2-L)
- Wang, Q., Rose, C. P., Ma, N., Jiang, S., Bao, H., & Li, Y. (2022). Design and Application of Automatic Feedback Scaffolding in Forums to Promote Learning. *IEEE Transactions on Learning Technologies*, 15(2), 150–166.
- Wang, Y., Gong, S., Cao, Y., & Fan, W. (2023). The power of affective pedagogical agent and self-explanation in computer-based learning. *Computers & Education*, 195, 104723.
- Weber, F., Wambsganss, T., Rüttimann, D., & Söllner, M. (2021). Pedagogical agents for interactive learning: A taxonomy of conversational agents in education. *Forty-Second International Conference on Information Systems. Austin, Texas*, 1–17.
- Wentzel, K. R. (2017). Peer relationships, motivation, and academic performance at school. *Handbook of Competence and Motivation: Theory and Application., 2nd Ed., 2*, 586–603.
- Wylie, R., & Chi, M. T. H. (2014). *7 the self-explanation principle in multimedia learning*. Cambridge University Press

https://education.asu.edu/sites/default/files/lcl/wylie_chi_selfexplanation_1.pdf

- Xu, Y., & Warschauer, M. (2020). What Are You Talking To?: Understanding Children's Perceptions of Conversational Agents. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–13.
- Yin, J., Goh, T.-T., Yang, B., & Xiaobin, Y. (2021). Conversation technology with micro-learning: The impact of chatbot-based learning on students' learning motivation and performance. *Journal of Educational Computing Research*, 59(1), 154–177.
- Zahry, N. R., & Besley, J. C. (2021). Can scientists communicate interpersonal warmth? Testing warmth messages in the context of science communication. *Journal of Applied Communication Research: JACR*, 49(4), 387–405.

국문 초록

교육용 대화형 에이전트는 학생들이 학습에 대해 정서적으로 참여할 수 있도록 도와주는 사회적 상호작용을 모방한다는 점에서 이점을 가진다. 교육용 대화형 에이전트는 사용되는 모달리티에 따라 다양한 형태를 취할 수 있다. 특히, 인스턴트 메시징을 학생들이 널리 사용하게 되면서 텍스트 기반의 대화형 에이전트는 교육 맥락에서 보편화되고 있다. 그러나, 그래픽 요소를 포함하는 다른 모달리티들과 다르게, 텍스트 기반의 접근 방식은 아직까지는 어떤 교수법으로 어떤 내용을 전달할지에 집중하여 제작되어 왔기 때문에 학생들과의 사회적 상호작용을 효과적으로 촉진하고 있지 못하다. 따라서 이 연구에서는 사회적 인식의 중요한 차원인 따뜻함의 정도를 조작하여 학생들이 대화형 에이전트와 사회적으로 상호작용하고 학습에 정서적으로 참여할 수 있도록 하고자 한다. 비록 정서적 참여가 학생에게 학습동기를 부여할 수 있지만, 학습에 관한 참여에는 여러 차원들이 있고 서로 밀접히 연관되어 있기 때문에 정서적 참여만으로 능동적 참여를 충분히 유도하기 어려울 수 있다. 따라서 이 연구에서는 학생에게 전달되는 학습활동의 유형을 인지적 참여 정도에 따라 조작하여 이를 보완하고자 한다.

이 연구에서는 교육용 대화형 에이전트에서 두 변수, 따뜻함의 정도 (높음 vs. 낮음)와 학습활동의 종류 (constructive vs. active)의 효과를 알아본다. 학습경험에 대한 객관적이고 주관적인 변수들, 학습 성취와 학습에 대한 내재적 동기를 측정하였고 학습경험에 대한 추가적인 통찰을 얻기 위해 각 학생들에 대해 반구조화 인터뷰를 진행했다. 우리는 98 명의 초등학교 6 학년 학생을 대상으로 기하 문제를 설명하도록 하는 챗봇인 GeomBot 을 제작하여 2x2 피험자 간 실험을 진행하였다. 실험 결과에 대한 양적 분석에 의하면, GeomBot 이 따뜻함이 높은 메시지를 보낼 때 따뜻함이 낮은 메시지를 보낼 때보다 학생들의 학습 성취, 흥미-즐거움이 유의하게 더 높았다. 또한, 학생이 인지적 참여 정도가 높은 활동을 할 때 낮은 활동을 할 때보다 유의하게 더 높은 학습 성취와 흥미-즐거움이 관찰되었다. 그리고 학습 성취, 흥미-즐거움, 긴장-불안에 대한 유의미한 상호작용 효과를 발견할 수 있었다. 인터뷰 데이터에 대한 질적 분석을 통해서도 두 가지 주요 인사이트를 도출할 수 있었다. 첫째, 연구자가 GeomBot 이 보내는 모든

메시지에 변수를 주었음에도 불구하고 학생들은 내용이 반복적이라고 인식했다. 둘째, 비록 따뜻함이 높은 메시지를 받은 학생들이 더 높은 학습 성취를 보였지만, 따뜻함이 낮은 메시지를 받은 학생들이 GeomBot 을 더 정직하다고 인식했다. 인터뷰 데이터에 의하면, 따뜻함이 낮은 GeomBot 이 학생들로부터 부정적인 피드백을 받았을 때 부정적인 리액션을 하기 때문에 정직하게 느껴진다고 보고되었다. 이와 반대로, 따뜻함이 높은 GeomBot 은 부정적인 피드백을 받아도 긍정적으로 반응하기 때문에 덜 정직하게 느껴진다고 보고되었다. 이는 현재 버전의 따뜻함이 높은 GeomBot 의 메시지가 학생의 피드백에 대해 학생의 기대에 상응하지 못하는 반응을 함을 의미하고, 이로 인해 GeomBot 에 대한 정직함 인식이 저하될 수 있음을 시사한다. 두 가지 질적 분석 결과를 통합하여, 우리는 (1) 따뜻함이 높은 메시지와 따뜻함이 낮은 메시지를 자동 생성하고, (2) 학생들의 피드백에 기반하여 메시지의 따뜻함 정도를 조절하는 것이 교육용 대화형 에이전트 사용 경험을 개선하는 데 도움이 될 수 있다는 결론을 얻을 수 있었다.

학생의 피드백에 따라 따뜻함의 정도를 조절하는 것이 에이전트에 대한 인식에 미치는 영향을 알아보기 위해, GeomPT 를 제작하여 두 번째 연구를 진행하였다. GeomPT 는 ChatGPT 에 프롬프트 엔지니어링 기술을 활용하여 학습 활동을 전달하는 따뜻함이 높거나 낮은 메시지를 자동으로 생성한다. 두 번째 연구에서는 10 명의 초등학교 6 학년 학생들을 대상으로, 따뜻한 메시지만 보내는 GeomPT 와 공부한 그룹(HW-C)과 학생의 피드백에 따라 메시지의 따뜻함 정도를 조절하여 보내는 GeomPT 를 사용한 그룹(HLW-C)을 비교한다. 에이전트에 대한 인식에 관한 변수들을 측정하는 것에 더하여, GeomPT 에 대한 학생의 내면적 인식을 알아보기 위해, GeomPT 가 어떻게 생겼을 것이라고 생각하는지 그림을 그리도록 하였다. 실험 결과를 통해, 메시지의 따뜻함을 조절한 그룹(HLW-C)이 따뜻함이 높은 메시지만 받은 그룹(HW-C)에 비해 더 높은 정직함 인식을 보였다. 그림 데이터에 대한 질적 분석은 학생들이 그들의 피드백에 따라 따뜻함의 정도를 조절하는 GeomPT 를 더 인간답고, 또래다운며, 유능하다고 인식함을 나타내었다.

이 연구의 기여점은 다음과 같다: (1) 우리는 교육용 대화형 에이전트에 따뜻함의 정도를 다르게 적용할 수 있는 언어적, 비언어적 단서들을 제공한다. (2) 우리는 메시지의 따뜻함과 constructive 활동 디자인에 집중하여, 생성 인공지능 중 ChatGPT 를 학습 상황에 적용하기 위한 가이드라인을 제시한다. (3) 우리는 학생의 능동적 학습을

촉진하기 위해 constructive 활동을 학생의 피드백에 따라 따뜻함의 정도를 조절하여 전달하는 것에 대한 경험적 증거들을 제공한다.

주요어 : 교육용 대화형 에이전트, 따뜻함, 인지적 참여, 생성 AI, 호혜성

학 번 : 2021-24263