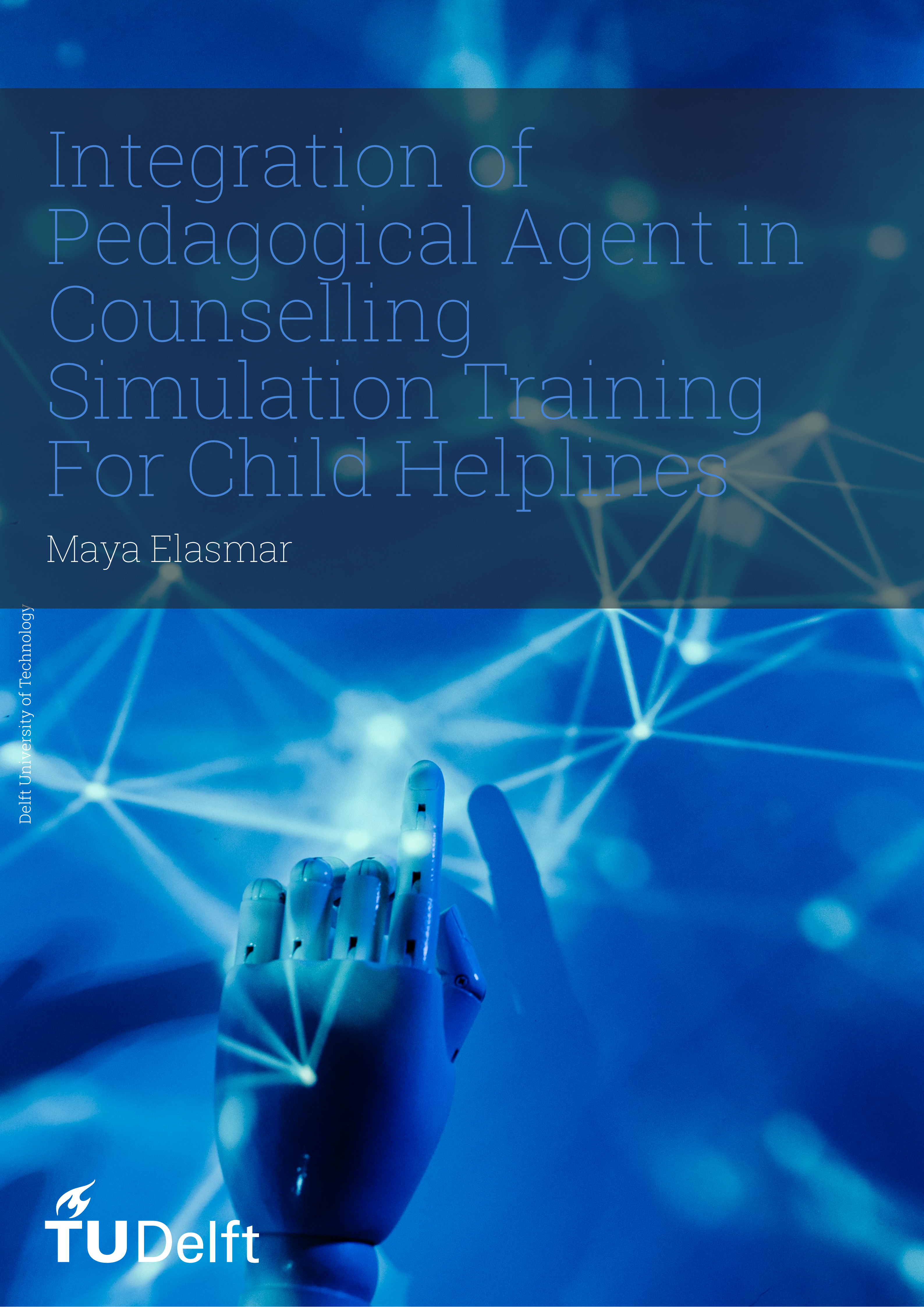


Integration of Pedagogical Agent in Counselling Simulation Training For Child Helplines

Maya Elasmar

Delft University of Technology

Integration of Pedagogical Agent in Counselling Simulation Training For Child Helplines

by

Maya Elasmar

Student Name
Maya Elasmar

Thesis Committee:	Willem-Paul Brinkman	TU Delft, Main Supervisor
	Mohammed Al Owayyed	TU Delft, Daily Supervisor
	Jie Yang	TU Delft, Committee Member
Project Duration:	Feb, 2024 - March, 2025	
Faculty:	Faculty of Electrical Engineering, Mathematics and Computer Science, Delft	
Cover:	Image by Tara Winstead from pexels.com	

TU Delft

Abstract

Child helplines provide a platform for children facing serious challenges, allowing them to share their stories and receive emotional support and guidance through counselling sessions. To ensure that volunteers are well prepared for these interactions, effective training programmes are essential. Recently, a BDI-based conversational agent was developed to train counsellors through role-play simulations, with the agent taking on the role of the child. However, this simulation training does not provide guidance to the trainees that caused a decrease in the trainees' self-efficacy. In this research, we aim to enhance the effectiveness of counselling training through role-play simulation by integrating a pedagogical agent. We integrated an adaptive pedagogical agent by applying the scaffolding technique, where we taught a set of skills divided into three modules. The pedagogical agent also provided feedback and hints as additional guidance methods. We evaluated this design through a mixed study involving 22 participants, comparing an intervention group trained with the pedagogical agent to a control group using a standard training approach. We measured the participants' performance, self-efficacy, and perceived usefulness of the system. While the intervention group showed higher mean scores across all measures compared to the control group, the differences were not statistically significant, indicating a possible underpowered experiment. This study contributes to the integration of pedagogical agents in simulation training systems for child helplines by proposing a framework that combines scaffolding, adaptivity, and structured learning.

Acknowledgments

First and foremost, I would like to thank my supervisors, Dr. Willem-Paul Brinkman and Mohammed Al Owayyed. I have learned so much from Dr. Brinkman—not only during this thesis but also through the courses he has taught, which inspired me to pursue this research within his group. I deeply appreciate his invaluable guidance and feedback throughout this process, without which this thesis would not have been possible. Similarly, I am grateful for Mo's guidance, valuable feedback, support and immediate assistance whenever I needed it. I would also like to extend a special thanks to Nele Albers for her help in finding participants for the focus group and for her invaluable guidance during the group meetings. Additionally, I am thankful to all the research group members for their guidance during the weekly meetings. A special thanks goes also to my fellow master's student, Adarsh Denga, with whom I have shared this journey.

On a personal note, I would like to thank my family for their endless support. My mother, also my best friend, has always been there to listen, motivate, and uplift me whenever I felt discouraged—not only throughout this thesis but throughout my entire life. She even went out of her way to help me recruit participants for my experiment. My father has always believed in me and worked tirelessly throughout his life to ensure that my siblings and I could receive the best education possible. My older brother, Abdelrahman, has always been kind and supportive. My ambitious brother, Abdelaziz, has inspired me to be independent, offered the best advice, encouraged me to pursue this master's degree, and has been a role model for me. Finally, I would like to extend my heartfelt thanks to the rest of my family and friends for their continuous encouragement and support throughout this journey.

Maya Elasmar
Amsterdam, 2025

Contents

1

Introduction

1.1. Motivation

Children might face serious circumstances such as domestic, emotional, sexual abuse, bullying, or mental health problems, often feeling powerless about their situation [1, 74]. Child helplines worldwide provide a platform where children can seek anonymous and confidential support, offering advice and assistance to those in need [38]. These services include crisis intervention, emotional support, information and advice, referrals to additional resources, preventive education, and follow-up support. In the Netherlands, De Kindertelefoon has served as the primary child helpline for 44 years [2], offering a safe space for children to freely and confidentially discuss sensitive topics [39]. The organisation provides both phone and chat services [2].

A new counsellor, especially one specialising in chat counselling, requires specific qualifications and competencies [74]. These include the ability to ask precise and effective questions [26], reflect on conversations [5], and summarise key points [32]. Additionally, child helplines require volunteers to possess certain skills such as empathy, sensitivity to others' needs, excellent communication skills, emotional resilience, and the ability to think on their feet [20, 66]. These required skills are depicted in figure 1.1.

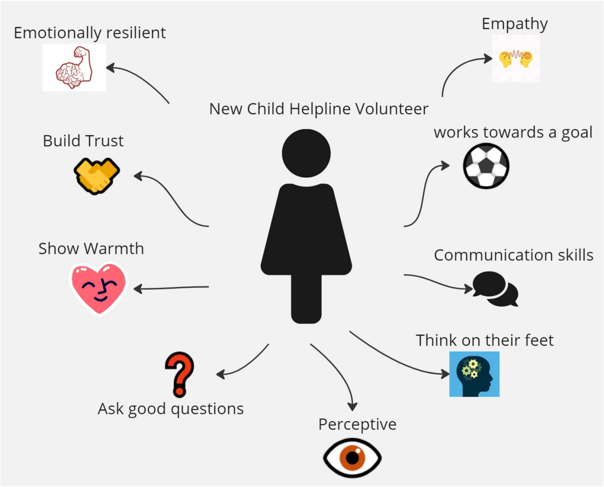


Figure 1.1: The required skills for a new volunteer at a child helpline requires.

The international child helpline offers e-courses to train volunteer counsellors in the required skills [19]. Similarly, child helplines such as De Kindertelefoon [3], the Scottish Child Helpline [20], and the UK Child Helpline provide training through online modules and in-person sessions [66]. Their programs include workshops, seminars, and role-play simulations, demanding significant time commitments from volunteers and supervisors who participate in role-playing simulations as children. These training programs also require costs associated with facilities and materials [75].

The use of virtual agents as an alternative to traditional training methods can potentially address some of the limitations [12], such as the significant costs and time investments required for organising training sessions. Additionally, they offer repetitive practising, which allows trainees to master the required skills [58]. To this end, a conversational agent has been created to train new volunteers for a child helpline [40].

1.2. Previous Research

The conversational agent, Lilobot, is built upon the Belief-Desire-Intention (BDI) framework. In this training simulation, the agent plays the role of a child with an issue, i.e. a bullied child [40]. Trainees can use this Lilobot to apply for example five-phase model, which is a structured counselling approach aimed at improving the effectiveness of counselling sessions [74]. However, initial evaluation with Lilobot showed that training with Lilobot results in a decrease in trainees' self-efficacy [40]. The researchers attributed this result to the absence of supervision, which is typically present in real training sessions, and the lack of actionable information tailored to the specific situations encountered during training.

To tackle these limitations, Martha et al. (2019) [53], Lane et al. (2013) [48] and Murray et al. (2010) [61] found that a pedagogical agent found to have positive impact on the learner's self-efficacy. Hence, our research proposes to integrate a virtual character designed to guide users through multimedia learning environments, a pedagogical agent. Pedagogical agents can serve as instructors, coaches, tutors, or learning companions, engaging in conversation while providing educational guidance, feedback, and support through autonomous actions and visual representations [41].

1.3. Research Questions

This research investigates the integration of a pedagogical agent into simulation training for child helpline counsellors in training. The study seeks to explore how integrating pedagogical agent can further enhance the training process. The study states the following research question:

How can the integration of the pedagogical agent enhance the effectiveness of counselling training through role-play simulation training ?

This research question is broken down into the following sub-questions:

1. *What are the key design considerations for the integration of a pedagogical agent?*
2. *How can a pedagogical agent be integrated in a simulation training environment?*
3. *To what extent does an integrated pedagogical agent in the simulation training environment contribute to the trainees' learning?*

1.4. Research Methods

To address the research question and its sub-questions, we followed the socio-cognitive engineering (SCE) approach [64]. To answer the first sub-question, we conducted a literature review, and organised a focus group session with an expert from De KinderTelefoon and two PhD students in psychology. These gave us insights about the operational demands of our system, human factors knowledge required to meet the operational demands and existing pedagogical agents. Using these insights, we put a design forward to address the second sub-question. Based on the design, we have conducted an experiment to evaluate the integration of a pedagogical agent into the simulation training. This aimed to answer the third sub-question. The results of the experiment were analysed using t-test with deltas. Finally, we drew conclusions to answer the research question and discussed some limitations and future work.

2 Foundation

In this chapter, we aim to address the first sub-question:

What are the key design considerations for the integration of a pedagogical agent?

To address the first sub-question, we followed the SCE approach for foundation, which consists of existing technology, operational demands, and human factors knowledge. This approach was useful in exploring existing pedagogical agents to identify features and strategies that could be integrated into our system. Furthermore, operational demands helped us look into the problem of the training simulation and what the stakeholders' needs are. As our system is made for humans, we needed to understand human factors knowledge. To address these sections, we conducted a literature study and organised a focus group session. Based on insights from the literature study and the focus group, we compiled a list of design considerations.

2.1. Literature Study

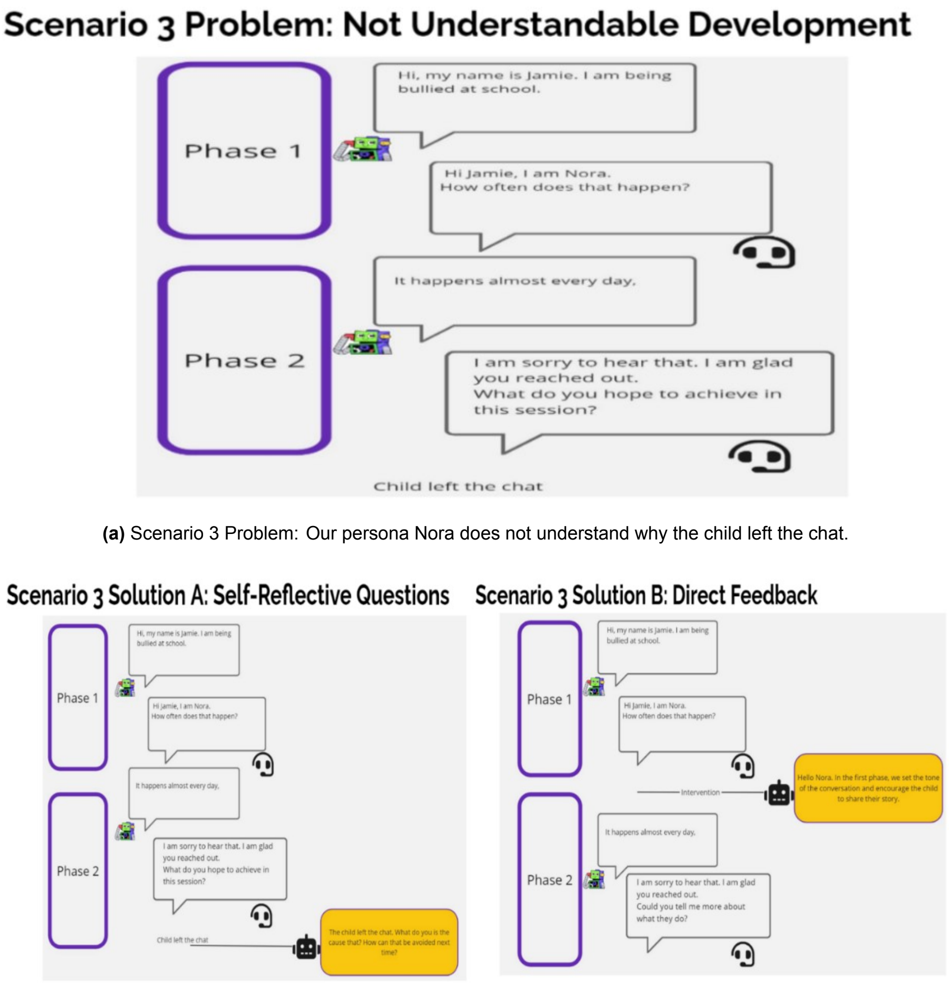
For our literature study, we used Google Scholar as the main search machine. We looked at a wide variety of papers, depending on the topic of the sections. For existing technologies, we mainly read review papers about pedagogical agents. These gave us insights about what a pedagogical agent is, what roles it has, and what strategies it applies. For operational demands, we looked at previous research with Lilobot: the BDI-based Lilobot [40], the emotional BDI-based Lilobot [52], and Lilobot with feedback [13]. This is to understand what challenges, problems, and findings were found when users interacted with the training simulation. Finally, for the human factors (physical, cognitive, social, cultural or emotional), we searched papers that gave us insights about human factors knowledge theories based on the stakeholders' needs and values, such as the cognitive load theory [76], social-cultural theory [80], and the constructivist learning theory [7].

2.2. Focus Group Setup

A focus group discussion is a qualitative method that is frequently used in research to gain a deeper understanding of social issues [68, 37]. Hence, we organised a focus group session via Microsoft Teams, inviting an expert from De Kindertelefoon and two PhD students in Psychology. These participants are key stakeholders, as the simulation is designed for child helplines like De Kindertelefoon and for volunteer counsellors.

Our aim was to understand their concerns about integrating a pedagogical agent into Lilobot and to see how these concerns align with the literature. During the session, we presented six scenarios based on potential problems that a trainee interacting with the virtual child might encounter, which were derived from previous research with Lilobot. The purpose was to demonstrate the consequences and trade-offs of different designs. For each scenario, we presented a concrete example of a persona named Nora, who represented a volunteer dealing with Lilobot, see for an example figure 2.1a.

For each problem, we proposed two potential, opposing solutions, see for an example figure 2.1b, based on teaching and guidance methods used in education. The purpose was to gain insights into their concerns regarding these solutions and prompt discussions. After each scenario, we initially asked which solution they preferred and why. We emphasised on getting their reasoning as we want to know their concerns more than what is the best solution. To gain further clarifications, we posed follow-up questions. The presented scenarios and the design concerns behind them are shown in table 2.1.



Scenario 3 Problem: Not Understandable Development

Hi, my name is Jamie. I am being bullied at school.

Hi Jamie, I am Nora.
How often does that happen?

It happens almost every day.

I am sorry to hear that. I am glad you reached out.
What do you hope to achieve in this session?

Child left the chat

(a) Scenario 3 Problem: Our persona Nora does not understand why the child left the chat.

Scenario 3 Solution A: Self-Reflective Questions

Hi, my name is Jamie. I am being bullied at school.

Hi Jamie, I am Nora.
How often does that happen?

It happens almost every day.

I am sorry to hear that. I am glad you reached out.
What do you hope to achieve in this session?

Child left the chat

The child left the chat. What do you is the cause that? How can that be avoided next time?

Scenario 3 Solution B: Direct Feedback

Hi, my name is Jamie. I am being bullied at school.

Hi Jamie, I am Nora.
How often does that happen?

Hello Nora. In the first phase, we set the tone of the conversation and encourage the child to share their story.

It happens almost every day.

I am sorry to hear that. I am glad you reached out.
Could you tell me more about what they do?

(b) Scenario 3 Solutions: (A) The agent asks self-reflective questions, (B) The agent gives direct feedback.

Figure 2.1: Scenario 3: Nora wants to understand how the scenario unfolded to help trainees reflect and learn from their experiences.

2.3. Technology

In this section, we studied the existing pedagogical agents to gain insight about available options and strategies that we could use in a training simulation integrating a pedagogical agent.

2.3.1. Roles & Architectural Design of Pedagogical Agents

Roles of Pedagogical Agent

Our aim is to integrate a pedagogical agent into the training simulation. A pedagogical agent acts as a tutor or instructor to enhance the student learning [72]. It is an intelligent agent developed based on pedagogical learning theory that can operate continuously and autonomously to support student activities [53]. A mentor agent, a pedagogical agent as a mentor, is supposed to work collaboratively with the learner to achieve goals [46]. Kim et al. (2016) [46] found that this type of agent improved both the learner's self-efficacy and overall learning outcomes. Additionally, they identified another type, the motivator agent, which demonstrates competence to the learner while simultaneously developing a

Scenario Problem	Solution A	Solution B	Design Concern
1: Nora gets stuck during simulation training.	The agent gives gradually decreasing guidance.	The agent gives gradually increasing guidance.	To understand their concern about guidance over time.
2: Nora struggles to apply the five phase model.	Explain the entire model at once.	Break the model phase by phase.	To understand their concern about the given amount and the presentation of training information.
3: Nora wants to understand how the scenario unfolded.	Self-reflection: the agent asks questions that lets them self-reflect on how the scenario unfolded	Direct feedback: the agent tells them how the scenario unfolded.	To understand their concerns about the reinforcement of learning.
4: Need for personalised guidance	User-driven guidance: request guidance by pressing button	Agent-driven guidance: agent recognises when support is needed.	To understand their concerns about trainees with different needs.
5: Trainee unsure how to handle a sensitive topic, like sexual abuse.	Agent takes over the conversation and demonstrates appropriate actions.	Provide step-by-step guidance to the trainee.	To understand their concerns about an agent helping in sensitive topics or unexpected situation.
6: Children might respond differently to similar solutions.	Allow the agent to lead trainees to negative outcomes to illustrate potential consequences.	Agent explains possible reactions and how to avoid negative outcomes.	To understand their concerns about interacting with children having different reactions.

Table 2.1: Scenarios, Problems, and Solutions shown to the focus group

social relationship to motivate the learner. However, their research also revealed that encouragement and support alone were insufficient for learners to reach their learning objectives.

Dai et al. (2022) [27] examined the impact of pedagogical agents on learning was examined. They found that the embodiment form (2D, 3D, recording of actual humans) of a pedagogical agent showed no clear relationship between the learning impact and the agent [27]. They also studied the different roles (expert, mentor, motivator, learner), which showed no improvement on the learning outcomes. However, an agent with multiple functions such as: demonstrating, coaching, information source, showed either no or positive effects on learning. They also identified this limitation in learning outcomes due to the usually small sample size in evaluation and that papers study short-retention of learning. Schroeder et al. (2015) [72] also found that the pedagogical effectiveness of pedagogical agents often stems from how they present information rather than their interactive elements, such as voice or embodiment [72]. Martha et al. (2019) [53] examined significant studies on the empirical evidence of pedagogical agents. They found that over 75% of these studies reported a positive impact on students' learning outcomes, and 50% showed a positive impact on student behaviour. Sikstrom et al. (2022) [73] did a study on the overall perception of the pedagogical agent. They found that the improvements of a pedagogical agent on the learning is inconsistent and sometimes contradictory. However, they also found that though the mixed reviews, the pedagogical agents are perceived more beneficial than distracting.

The reviewed papers in the previous paragraph were published in different years, 2022, 2015 and 2019, respectively. We compared them to see what the common findings are. They argued that the quality of research pedagogical agents was not high and could be better in terms of reporting, methodology and reliability of the results. They also mentioned that an agent's role of providing information and guiding is more effective. Furthermore, they even argue that there should be more studies about an agent taking multiple roles and functions. Finally, they find a limitation in research paper studying the student behaviour and their learning outcomes, as papers usually focus on one.

Pedagogical agent Software Architecture
Above, we discussed the roles of a pedagogical agent, but how does such as pedagogical agent work from a software perspective. To answer that, we looked at the architectural components of a pedagogical agent. One of these components is the knowledge base, which includes what the agent knows, its strategies and the student model [30]. The knowledge base manager helps users use and update the knowledge stored in the knowledge base [30]. Followed these components, Devedzic et al. (2002) [30] described the reasoning engine in their book. The reasoning engine serves a control module and decision-maker, determining the agent's actions, responses to stimuli, and other teaching-related activities. Johnson et al. (1998) [43] also described the reasoning engine as the engine that handles the decision-making of such an agent based on the student's model and updates the agent's mental state based on it. Finally, a pedagogical agent also has a communication component. This one is responsible for perceiving the dynamic learning environment and responding accordingly [30], and recognises situations where the pedagogical agent can intervene, such as specific student actions, co-learner progress, and the availability of desired information.

2.3.2. Cognitive Apprenticeship in Training with Pedagogical Agents

As mentioned above, a pedagogical agent can serve as a tutor to guide learners through tasks and activities [72]. According to Johnson et al. (2001) [44], an agent can teach both small actions and larger subplans, typically by first demonstrating the action and then allowing the learner to perform it. This approach has been shown to result in more robust interactions with learners.

Our focus group expressed concerns that trainees may struggle to handle complex situations if they have not seen a similar scenario at least once before encountering it themselves. This aligns with the cognitive apprenticeship model in education practice [29], which is a process where learners learn from a more experienced person by way of cognitive and metacognitive skills and processes [24]. It helps handle complex tasks and the cognitive processes involved in learning. The first method in this model is modelling, which encompasses demonstrating the process by an expert [23]. Johnson et al. (2001) [44] also discussed that, instead of first demonstrating and then allowing learners to perform, an agent can explain what to do and why. This also aligns with the second model of the cognitive apprenticeship, which is coaching [24]. It encompasses a mentor observing the student and providing guidance.

Next, we studied the use of the cognitive apprenticeship model in science. The findings showed that it is effective to practice scientists before entering the field, as well as improvement in content knowledge [71]. Finally, looking at its effect on the student behaviour, Sadler et al. (2010) [71] found that many studies showed improvements in the student's confidence and self-efficacy. However, they think that the evidence is mixed and relies on the self-reported data. Although, Sadler et al. (2010) [71] studied the model in scientific classes, we still believe it could be beneficial if applied as the pedagogical agent takes the role of a tutor or a mentor as this is what usually happens in educational settings [29]. It has also showed its positive effects on multiple aspects, such as: practising, learning knowledge and possibly student behaviour.

2.3.3. Adaptive learning

Pedagogical agents in virtual learning environments utilise personalised methods to guide students through course material, delivering tailored content, monitoring progress, and offering personalised support [84, 57]. Xu et al. (2005) [84] and Maryadi et al. (2017) [57] found personalised learning positively effective on the learning outcomes. Although, these studies showed positive effects, they did not study the long-term effects of personalised learning. Recent studies suggest that personalised learning experiences can be enhanced by integrating intelligent pedagogical agents with Intelligent Tutoring Systems (ITS) [87]. However, this suggestion was based on a discussion, and not evaluated in the paper.

Martin et al. (2020) [56] examined the integration of ITS with pedagogical agents. This integration enables the adaptation of teaching processes to individual learner characteristics, including knowledge levels, learning styles, and psychological traits. Such an approach closely aligns with the principles of adaptive learning, which involves dynamically adjusting instructional content based on students' comprehension and responses to embedded assessments, as well as their learning preferences. Martin et al. (2020) [56] also found that ITS is increasingly recognised not only for its technological innovations, but also for its ability to enhance the academic trajectory and satisfaction of learners in diverse educational contexts. Further evidence of ITS effectiveness is supported by a systematic review by Mousavinasab et al. (2021) [60], which found ITS to be highly effective in improving student performance, particularly through learner-based assessments. However, they also identify a significant gap in research on the broader educational outcomes of ITSs. Specifically, they found a notable gap in understanding how ITSs influence aspects such as critical thinking, problem-solving abilities, and decision-making skills.

The framework of ITS and adaptive learning align as they both consist of three models [69]: learner model, content model and instructional model.

The Learner Model (student module in ITS)
The learner model refers to the dynamic representation of the emerging knowledge and skill of the student [67]. It includes learner attributes [56], such as: knowledge, motivation and preference.

The Content Model (expert module in ITS)
The content model refers to the knowledge base to be taught to the learner [56]. It serves as the source of knowledge to be presented to the student, which includes generating questions, explanations and responses [67].

The Instructional Model (tutoring module in ITS)
The instructional model refers to the algorithm that assists in adapting the instruction based on the content and learner model [56]. It is also referred to as the adaptation model as it defines what, when, and how adaptation can occur. For example, if a learner has been evaluated as a beginner in a particular procedure, this model will show some step-by-step demonstrations of the procedure before asking the user to perform the procedure on his or her own [69]. When a learner gains expertise, this model might decide to present increasingly complex scenarios.

2.3.4. Pedagogical Agent with scaffolding

One of the strategies used by pedagogical agents in adaptive learning is scaffolded guidance [45]. Scaffolding can take several forms, such as hints, prompts, feedback, illustrations, or interactive features [31]. Duffy et al. (2015) [31] have found that scaffolding had a positive outcome for performance-approach students but caused negative feelings to mastery-approach students as they find it controlling. However, they had an unequal sample size across the conditions, which could have caused biased results.

Martha et al. (2020) [54] investigated the implementation of a pedagogical agent within ITS that employed metacognitive scaffolding, which demonstrated a positive impact on learning outcomes. Metacognitive scaffolding is structured into four phases: planning, monitoring, evaluation, and reflection. During the planning phase, students are guided to prepare for their assignments by setting goals and outlining strategies. In the monitoring phase, they receive support to comprehend the task at hand and track their progress. The evaluation phase involves the pedagogical agent identifying student mistakes and assisting with their correction. Lastly, in the reflection phase, students are provided with performance feedback along with suggestions for improvement. A few years after, they studied the effects of it that showed positive outcomes on the learning outcomes [55]. Cheng et al. (2009) [17] investigated the implementation of the pedagogical agent with horizontal scaffolding in HINTS, a health problem-solving system for clinical cases through simulation. By this, each section is taught separately from the other section [17]. They also applied vertical scaffolding, which considers how to build up the scaffolds to support students' learning over several sections. They do that by dividing the cases in three stages, each with a checkpoint to ensure that the student is on the correct path. This is called the regression model. When the students become better, they switch to the straight model, where the students continue to progress regardless of correctness. Cheng et al. (2009) [17] found this system to be stress-reducing and learning supportive for new students. However, this system did not apply customised hints to the student. Van Lehn et al. (2011) [78] also investigated the application of scaffolding within the adaptive learning system ITS. They found that scaffolding helps students self-repair and construct knowledge effectively, either applied within an adaptive learning system or with a human tutor. Finally, Sikstrom et al. (2022) [73] found that a pedagogical agent with scaffolding fosters self-regulative strategies. From this, we can get that scaffolding as a method has a positive outcome in whatever setting it is applied.

2.4. Operational Demands

In this section, we started by explaining the training simulation environment to understand how it works and what problem it has without our envisioned pedagogical agent. The purpose is to identify the stakeholders' needs and values that will eventually help to understand the design concerns.

2.4.1. Lilobot: a BDI-based Conversational Agent

Lilobot is a role-play simulation in which Lilobot plays the role of a child with an issue, e.g. a bullied child. In this simulation, the trainees can use Lilobot to apply for example the five phase model, which is explained in detail in appendix A. Lilobot is a BDI-based conversational agent. The BDI framework comprises three core concepts: Belief, Desire, and Intention [70, 33, 81]. Beliefs within this framework represent the agent's understanding of its environment, incorporating both external factors and its own internal state. For instance, in the case of Lilobot [40], beliefs include perceptions like being in control, trusting KinderTelefoon, or being asked about a confidant, continuously updated based on user input. Desires serve as the basis for the agent's intentions, guiding its actions and responses during interactions [33]. These desires reflect the agent's motivational state and include objectives such as Lilobot wanting to discuss its problem. Intentions represent the explicit sequence of actions needed to achieve these objectives, aligning with the agent's current state [81].

2.4.2. Problem scenario

In the SCE approach, the challenges faced without the envisioned system are depicted in a short story [63]. Hence, we provided a short story of a trainee interacting with Lilobot to show the challenges faced in this training simulation from the perspective of a trainee. The elements in this story are based on the feedback of those who interacted with Lilobot in previous research [13, 40]. Let's consider a scenario involving a persona called Hannah, see table 2.2.

Problem Scenario
Name: Hannah, **Age:** 21, **Major:** Business Administration
Scenario Description:
She received an explanation of the five-phase model before starting training with the simulation. Throughout the session, she noticed that Lilobot often failed to understand her statements and occasionally repeated its responses, based on participant feedback in evaluating Lilobot [40]. By phase 2 of the Five-Phase Model, Hannah felt stuck and struggled to progress, based on feedback in Grundmann's paper [40]. During the training, Hannah wished there were more interactive elements to keep her engaged. She received all the information before the start of the simulation training, but there were no reminders or additional interactions during the simulation, based on limitations identified in the evaluation of Lilobot with a feedback system [13]. As she advanced to phase 4, Lilobot consistently prompted Hannah to contact its teacher, which confused her, as she was unsure how to proceed. This was an issue participants encountered [40]. As a consequence, Hannah ended the conversation. When Hannah received feedback at the end, she struggled because Lilobot didn't understand her, not because she didn't apply the required skill well, reflecting participant feedback [40, 13]. Hannah found it not useful since she could not seek clarification and felt the feedback did not accurately reflect the interaction. After this training, Hannah felt less confident about becoming a good volunteer for a child helpline, based on previous evaluations [40, 13].

Table 2.2: Problem Scenario

Value Story

Although, our persona Hannah was fictional, the issues she encountered are based on the feedback and evaluation of real people interacting with Lilobot. From this story, we can learn the trainees' values and understand more about the design concerns. In table 2.3, we made a list of a few needs and values of the stakeholders that we identified and formulated based on the problem scenario.

User Need	Rationale
The trainees would want to overcome obstacles.	Participants got stuck during certain phases in the training simulation [40].
The trainees would want to be able to handle unexpected situations.	Trainees can be in an unexpected situation that causes the child to leave the conversation [40].
The trainees value an interactive element during the simulation that can be engaging and serve as reminders.	Trainees preferred having a paper explaining the five phase model during the simulation training over immediate feedback [13].
The trainees value a clear actionable feedback and one where they can also give their reflection on what happened	Trainees found the feedback not accurate to what really happened during the simulation [40].

Table 2.3: The user needs for a training simulation for child helplines based on previous evaluations of Lilobot [40, 13].

2.5. Human Factors

In this section, we focused on the human factors that should be addressed in the design to meet operational demands [63]. Based on the user needs identified in table 2.3, we addressed the following topics: cognitive load management, guidance, and active learning.

2.5.1. Cognitive Load Management

During scenario 2 regarding the frequency and the presentation of training information in Table 2.1, our focus group was concerned that giving too much information at a time might overload the learner. They were also concerned about being distracted by information that is currently not relevant to the task at hand. These concerns align with the cognitive load theory of John Sweller [76] that suggests that our working memory is only able to hold a small amount of information at once and that instructional methods should not overload it to maximise the learning [76]. This theory also focused on instructional strategies for decreasing extraneous cognitive load, which is load that is imposed by processes not directly relevant for learning [77]. One of the strategies to manage the cognitive load is fading guidance strategy [59]. This means replacing a uniform sequence of tasks with a varying sequence of tasks that initially offer sizeable learner guidance, gradually reducing this support until no guidance is provided, which is also known as scaffolding [18]. Another potential strategy is fragmentation, where complex skills are subdivided into subskills without considering their interactions and coordination demands [47]. These subskills are gradually integrated and taught as a unified set of skills over time.

2.5.2. Guidance

As seen in the problem story, our person got stuck at phase 2 of the five phase model. Also, like we identified in table 2.3, trainees would want to overcome these obstacles. Our focus group was also concerned at scenario 1, 3 and 5 about trainees not being able to proceed and for them to see what they did wrong and for that stick with them for a long time. Johnson et al. (1998) [43], Finch et al. (2020) [34] and Baylor et al. (2005) [8] also expressed concerns that the virtual agent, called Steve, should assist the students when they are in need of assistance or having a question.

Guidance is supported by Vygotsky's Socio-Cultural Theory and the concept of the Zone of Proximal Development (ZPD) [80, 82]. Vygotsky proposed that cognitive development occurs through social interactions, where learners engage in activities that are mediated and guided by more knowledgeable others, such as teachers or peers. One of the methods that teachers use to guide students is scaffolding [18]. Previously, we discussed in broad the effects of scaffolding as guidance method. Another method is problem-based learning, in which a student learns about a subject by solving an open-ended problem, either individually or in groups [21]. Yew et al. (2016) [86] also found that it is an effective teaching and learning approach, particularly when it is evaluated for long-term knowledge retention.

2.5.3. Active learning

As we identified in table 2.3, the stakeholders value an interactive element for engagement and reminders for knowledge. Our focus group was concerned in scenario 5 of an agent helping in sensitive topics, about the students not remembering their mistakes or learn from them if they are simply told by someone. They were also concerned about that these critics might lower their self-efficacy. This concern aligns with the strategy of self-assessment, in which learners assess their own performance, evaluate and reflect on the quality of their learning process and outcomes according to selected criteria to identify their own strengths and weaknesses [85]. Self-assessment is part of active learning [11], in which students actively engage with material through activities like discussions and problem-solving [14]. The constructionism learning theory supports the concept of active learning, which states that people construct knowledge through active engagement and experience with the material, followed by self-reflection on their learning process [7, 9]. Furthermore, active learning has a positive impact on the self-efficacy of students [36, 42]. Besides self-assessment, another technique used for active learning is quick quizzes, in which every period of time the teacher proposes a short quiz to the students about the things that been learned [25]. A similar technique to self-assessment is self-reflection. In this technique, the teacher asks the students a question that requires them to reflect on their learning or to engage in critical thinking [14]. Both techniques self-assessment and self-reflection are also parts of the cognitive apprenticeship model that we discussed earlier [23]. Besides, our focus group was also concerned about the accuracy and correctness of a student's own self-reflection. This is also a concern in self-regulated learning, in which individuals set their goals and monitor their development based on their personal and external constraints [65]. Personal and external constraints can be in the form of feedback [16], i.e. the learner regulates their development based on their personal feedback, but when they receive external feedback, they adjust their goals and performance.

2.6. Design Considerations

In this chapter, we gained insights from the literature study and our focus group. Based on these, we have set a list of design considerations that should be taken into account in our system. The list of design considerations and the underlying design concern are shown in table 2.4.

Design Concern	**Design Consideration**
D1: The cognitive load can increase when giving a lot of information at once.	A pedagogical agent should teach the skills to the trainees gradually.
D2: Each trainee needs assistance at different phases in the training simulation.	A pedagogical agent should give personalised guidance to trainees based on their needs.
D3: Trainee should positively engage in the training simulation.	A pedagogical agent could motivate the trainees.
D4: Trainees' self-efficacy can decrease and forget important knowledge after a short time.	A pedagogical agent could incorporate active learning elements, such as quick quizzes and self-reflection exercises, to enhance self-efficacy and retention of information.
D5: Trainees can get stuck in unexpected situations during the training simulation.	A pedagogical agent could offer situation-based problem-solving support and guidance.

Table 2.4: The Design Considerations list defined through our literature study and focus group.

3 Design

In this chapter, we aim to address the following sub-question:

How can a pedagogical agent be integrated in a simulation training environment?

To answer this question, we propose a solution that integrates an adaptive pedagogical agent within the simulation training environment. This solution is designed to address the design considerations of table 2.4. In this chapter, we describe the adaptivity of our system. We also describe how we incorporated the scaffolding mechanism into the adaptive system. Finally, we described the remaining components of our design, feedback and hints.

3.1. Training Through Adaptivity and Scaffolding

In our design, we incorporated adaptivity and scaffolding to ensure that the pedagogical agent provides guidance to the trainees during their training with Lilobot. Adaptivity allows the system to adjust the training process based on each trainee's performance [56], while scaffolding offers structured support that is gradually removed as trainees gain competence [45].

To illustrate the overall flow of our design, shown in Figure 3.1, we consider the example of a trainee named Denis. During his first session, Denis begins with Module 1, which focuses on addressing a child's concern. The system introduces Denis to the purpose of the module and explains what is expected of him. When the session begins with Lilobot, routine tasks such as greetings and farewells, which are not part of Module 1, are automated by the system, as shown in Figure 3.3.

As the session progresses, at key points aligned with the module's objectives, the system identifies moments where Denis is required to engage actively and provide input. To signal these moments, an input box appears along with a notification prompting Denis to respond in accordance with the module's specific goals. Once Denis completes this part, the system resumes automation of the remaining conversation segments that are unrelated to Module 1. The system evaluates Denis's inputs and provides feedback based on it. Based on Denis's performance, he either passes the module and moves on with the training or he is required to repeat Module 1.

If Denis passes Module 1, he progresses to Module 2. We start then a new session, where he must now provide inputs for both Modules 1 and 2 skills. Meanwhile, the system continues to automate the inputs related to the skills of module 3 that have not yet been introduced to Denis. By handling these inputs automatically, the system reduces cognitive load [76] and allows Denis to focus on reinforcing previously learned skills and developing new ones at a manageable pace. When Denis passes Module 2, a new session starts where there is no automation, and Denis has to give inputs during the entire conversation. During each session, Denis can ask for a hint if he is stuck. As discussed in the foundation chapter, these hints act as a form of scaffolding [31], aligned with the principles of the Zone of Proximal Development (ZPD) [80, 82], ensuring Denis receives targeted guidance to overcome challenges while continuing to build his skills progressively.

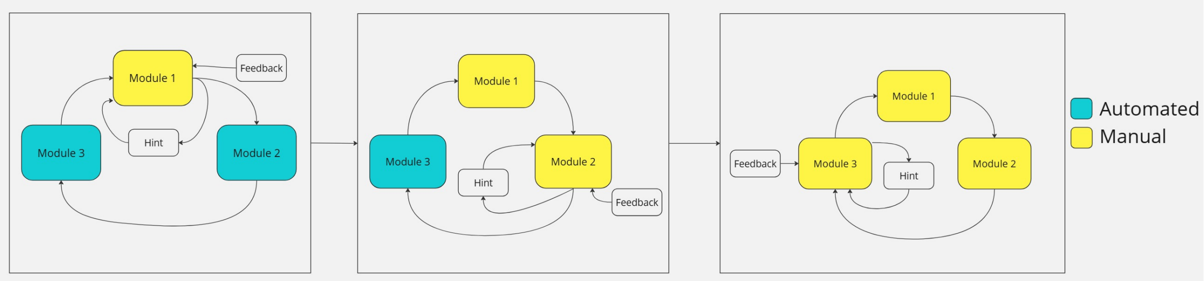


Figure 3.1: The design of Lilobot integrates adaptive guidance through scaffolding, ensuring trainees master skills in each module before progressing.

3.1.1. Adaptivity

Adaptivity is a core feature of our pedagogical agent, allowing it to tailor the training process to each trainee's individual performance. As mentioned in chapter 2, an adaptive framework exists of three models: the learner, content and instructional model. The three models representing our system are shown in 3.2. In our case, the trainees are often volunteers without formal training or degrees in counselling. Therefore, we assume that they begin with little to no counselling knowledge. However, they are expected to possess basic human skills, such as empathy and the ability to engage in natural conversation. These foundational skills form the basis of our learner model, allowing the system to build upon and refine these abilities throughout the training process. The content model in our design includes our modules. Module 1 focuses on the child's concerns, and module 2 focuses on the child's resources (abilities, relationships and goals). The final module is about reflecting, showing empathy and summarising some key points. The modules' structure follows the structure of Sindhal's handbook *Chat Counselling For Children And Youth* [74]. We followed their structure as this book explains what chat counselling is, the structure of a chat counselling conversation, and what is important for a chat counsellor to do. Our instructional model integrates scaffolding and adaptive guidance to progressively train trainees across three modules. Scaffolding is implemented by structuring the skills into three sequential modules that build upon each other, and by using an automated pilot to manage non-relevant parts of the conversation during initial sessions, allowing trainees to focus on core skills. Adaptivity is implemented in the design through performance-based progression to the next module, personalised feedback provided after each session, and hints offered during the session when the trainee demonstrates a need for additional support.

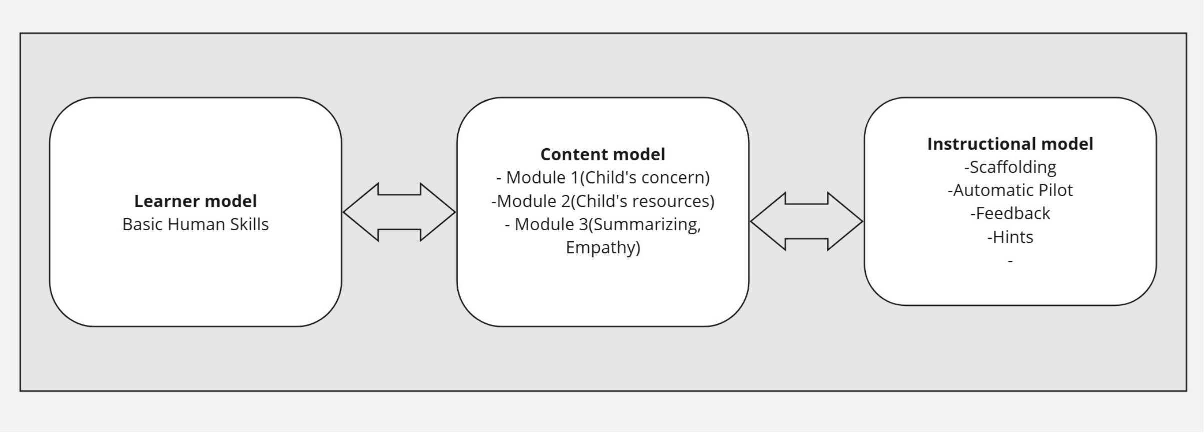


Figure 3.2: The representation of the ITS' models for the simulation training system with a pedagogical agent.

3.1.2. Adaptive With Scaffolding

Our adaptive design incorporates the scaffolding mechanism, where the guidance gradually decreases and each module scaffolds to the other. As the trainee advances, the session becomes more manual, with fewer explanations provided about the content of earlier modules. The modules are structured according to the importance of the content, with each module providing the foundational knowledge needed for the next.

Module 1: Understanding the Child's Concern
This module focuses on helping trainees understand the child's concerns and getting to know their story.

- **Scaffolding to Module 2:** The understanding gained in Module 1 enables trainees to explore the child's resources in the next module. These resources include identifying the child's response to their issues, who they trust, and what their goals are.

Module 2: Exploring the Child's Resources
In this module, trainees learn to delve deeper into the child's context by understanding their coping mechanisms, support systems, and goals.

- **Scaffolding to Module 3:** The insights gathered in Module 1 and 2 prepare trainees to identify key moments for showing empathy and summarising key points, such as their story or goal.

Module 3: Empathy & Summarising Key Points
This module incorporates the learnings from the previous stages, enabling trainees to demonstrate empathy effectively and summarise some points during the conversation.

3.1.3. Automatic Pilot

For the automated pilot, we predefined an optimal conversation flow where each message is labelled with its sender (either the virtual child or the pedagogical agent) and associated module, see line 5. The algorithm, shown in Algorithm 1, ensures that the automated pilot dynamically steps in to manage parts of the conversation based on the current module and the trainee's progress, see lines 6 through 30.

The conversation flow is divided into 3 modules. Each module has specific rules for when the automated pilot intervenes:

- **Module 1:** The pedagogical agent begins by delivering Segment 1. The algorithm allows the trainee to contribute with 4 inputs, meeting the passing criteria of asking 3-5 questions about the child's concerns. After this, the pedagogical agent completes Segments 2 and 3 automatically. See lines 8 through 18.
- **Module 2:** The pedagogical agent starts by delivering Segment 1. The trainee then provides 10 inputs, split into 4 for Module 1 and 6 for Module 2. Once these inputs are collected, the pedagogical agent resumes control to deliver Segment 3. See lines 19 through 26
- **Module 3:** The algorithm allows the trainee to control the entire conversation without intervention, ensuring full engagement and practice. See lines 27 through 29

The algorithm works by monitoring the trainee's input count and determining which segments have been completed. If the trainee's input count meets the predefined thresholds, the algorithm triggers the automated pilot to deliver the subsequent segments, ensuring smooth progress through the training session. This approach maintains a balance between user interaction and automated guidance, helping the trainee focus on their training objectives while covering all necessary conversation elements efficiently.

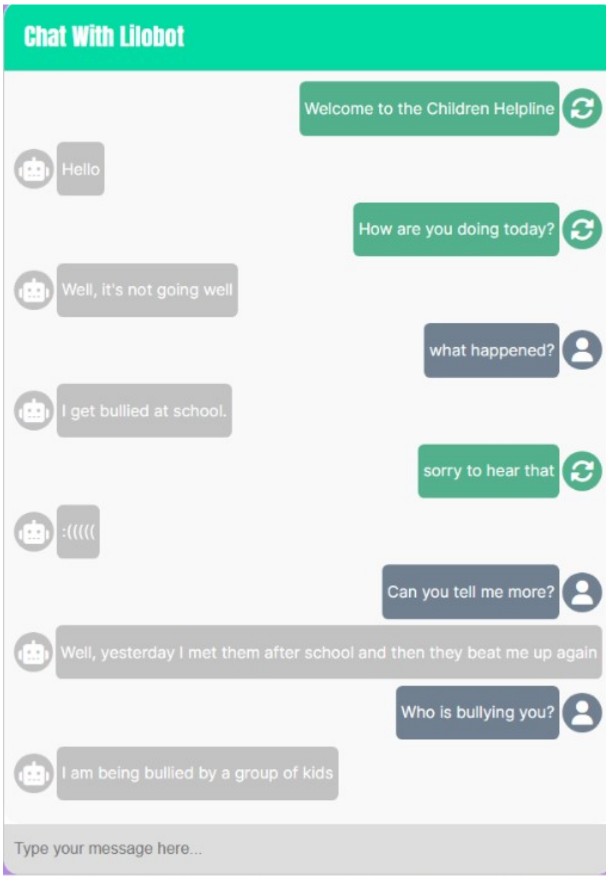
3.1.4. Adaptive Feedback & Hints

The system also provides personalised, adaptive feedback after each practice session. The feedback is tailored to the objective of the newly taught module. For example, if the trainee is in session 2, including modules 1 and 2, they will only get feedback about module 2 to keep the focus only on module 2. We do this to reduce the cognitive load [76]. The content of the feedback will be about whether they passed the session and a reflection about their performance for the currently taught module. If they fail, the feedback will also contain a suggestion about what they should do to pass the module.

The Keystroke-Level Model (KLM) [10] states that the average time a human needs when interacting with a computer is 10-15 seconds. We doubled that time as the trainees will likely not be familiar with the task and need more time to think about it. Hence, if the trainees get stuck by not giving input for half a minute, that will provoke a hint about their next step. Our hints will be based on their current session and the same for all trainees. Moreover, the trainees get only one hint per session.

Algorithm 1 Algorithm for the Automated Pilot

```
Initialise:
    TraineeInputCount $\leftarrow 0$
    CurrentModule $\leftarrow$ 1
    SegmentCompletion $\leftarrow$ {1: FALSE, 2: FALSE, 3: FALSE}
    OptimalFlow $\leftarrow$ Predefined conversation flow (Array of Messages)
function PredefinedConversationControl(CurrentModule, TraineeInput)
    Increment TraineeInputCount by 1
    if CurrentModule = 1 then
        if SegmentCompletion[1] = FALSE then
            ExecuteSegment(1, ”PedagogicalAgent”)
            SegmentCompletion[1] $\leftarrow$ TRUE
        else if TraineeInputCount $\geq 4$ and SegmentCompletion[2] = FALSE then
            ExecuteSegment(2, ”PedagogicalAgent”)
            SegmentCompletion[2] $\leftarrow$ TRUE
        else if SegmentCompletion[2] = TRUE and SegmentCompletion[3] = FALSE then
            ExecuteSegment(3, ”PedagogicalAgent”)
            SegmentCompletion[3] $\leftarrow$ TRUE
        end if
    else if CurrentModule = 2 then
        if SegmentCompletion[1] = FALSE then
            ExecuteSegment(1, ”PedagogicalAgent”)
            SegmentCompletion[1] $\leftarrow$ TRUE
        else if TraineeInputCount $\geq 10$ and SegmentCompletion[3] = FALSE then
            ExecuteSegment(3, ”PedagogicalAgent”)
            SegmentCompletion[3] $\leftarrow$ TRUE
        end if
    else if CurrentModule = 3 then
        Allow User to Control Entire Conversation
    end if
end function
function ExecuteSegment(SegmentID, Sender)
    Messages $\leftarrow$ Filter(OptimalFlow, segment = SegmentID and sender = Sender)
    for all Message in Messages do
        DeliverMessage(Message)
    end for
end function
function DeliverMessage(Message)
    Display(Message.content)
end function
```



Chat With Lilobot

Welcome to the Children Helpline

Hello

How are you doing today?

Well, it's not going well

what happened?

I get bullied at school.

sorry to hear that

:(((((

Can you tell me more?

Well, yesterday I met them after school and then they beat me up again

Who is bullying you?

I am being bullied by a group of kids

Type your message here...

Figure 3.3: An example of an automated part of the conversation. The automated messages are in a green textbox with a cycle icon. The light grey textbox with the robot icon represents the virtual child. The dark grey textbox with the user icon represents the trainees' input.

4 Evaluation

In this chapter, we aim to address the following sub-question:

To what extent does an integrated pedagogical agent in the simulation training environment contribute to the trainees' learning?

To evaluate the developed prototype of our design in chapter 3, we conducted an experiment to investigate the learning effect of our pedagogical agent integration in a role-play simulation training. In this chapter, we started by explaining our experiment set-up, followed by the results of our experiment. To answer our sub-question, we formulated hypotheses covering: the learning outcomes, self-efficacy and perceived usefulness. As mentioned in chapter 2, a pedagogical agent with scaffolding can improve the learning outcomes [73, 60]. We also discussed that a pedagogical agent can improve the learner's self-efficacy [46]. Davis et al.(1989) [28] discussed that a system with a high perceived usefulness is one in which a user believes that its usage will enhance their job performance.

Hence, our formulated hypotheses are:

- **H1:** Trainees using the pedagogical agent with Lilobot apply the taught skills better than those using Lilobot only.
- **H2:** Trainees using the pedagogical agent show higher self-efficacy levels than those using Lilobot only.
- **H3:** Trainees using the pedagogical agent perceive the training as more useful than those using Lilobot only.

4.1. Methods

In this section, we explained our study design, the participants, materials, measures, procedure and the data preparation & statistical analysis. We registered the design of this study with the Open Science Framework (OSF) [1]. This experiment was also approved by the TU Delft Human Ethics Research Committee (HREC reference number: 4695).

4.1.1. Study Design

We followed a mixed study design. Participants were divided into two groups, with each group experiencing only one of the study's conditions (between-subjects design). Additionally, all participants completed baseline and post-measurements to assess changes in performance and self-efficacy over time (within-subjects design). This combination allowed us to compare outcomes across groups while accounting for individual differences and initial performance levels.

[1] https://osf.io/9gtbm

Pilot Experiment

We conducted a pilot experiment with 4 participants to test the flow of the experiment and ensure there were no errors. No data was collected during this pilot, and these participants did not take part in the actual experiment.

4.1.2. Participants

We recruited 22 participants via email or personal network. The sample size was determined based on a power analysis implemented in R. The R script B.1 is shown in appendix B. This script calculates the necessary number of participants to achieve a power of approximately 0.72, a Cohen's d effect size of 0.8 and a correlation of 0.7 between baseline and post-measurements. The large effect size was expected due to the targeted design of the pedagogical agent and its scaffolding capabilities, which were specifically tailored to address the learning objectives.

The participant group consisted of 63.6% (14) females and 36.4% (8) males. The largest age group was 25–34 years old, comprising 45.5% (10) of the participants, followed by 31.8% (7) in the 18–24 age range. The remaining participants were aged 35 and older. Additionally, 77.2% (17) of the participants had a high level of education, holding at least a bachelor's degree or higher.

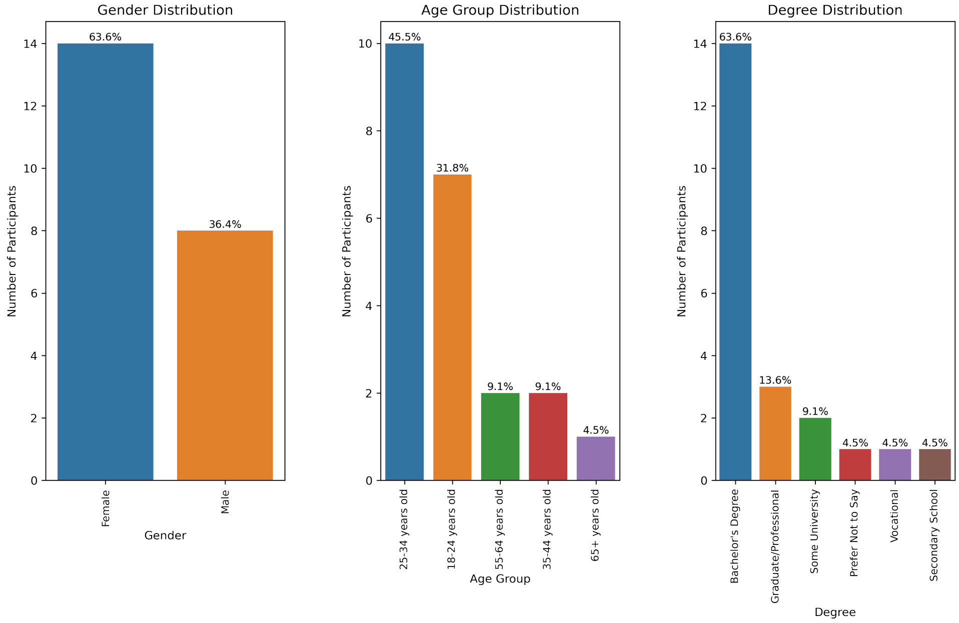


Figure 4.1: Distribution of Participants by Gender, Age Group, and Educational Level: The left plot represents the gender distribution, the middle plot illustrates the count of participants across different age groups, and the right plot highlights the educational levels of the participants.

4.1.3. Materials

The experiment was made available to participants through an online survey hosted on Qualtrics [2]. Qualtrics hosted the informed consent form, and the questionnaires, and directed our participants to our system.

Prototype

The prototype layout was adapted from a previous thesis on Lilobot [40]. The chat functionality utilised basic input detection techniques, including exact matching, regex matching with prioritisation, and loose matching with overlap scoring. The matching utterance examples were predefined by Grundmann et al. (2025) and Al Owayyed et al. (2024) [40, 4]. Additionally, the automatic pilot was implemented using a predefined conversation flow, based on the algorithm outlined in Figure 1. The scenarios for Lilobot were adapted from Al Owayyed et al. (2024)[4].

Both the intervention and control conditions involved a pre-session, a training session, and a post-session. The primary difference between the conditions lay in the training session. The intervention

[2] https://www.qualtrics.com/nl/

group interacted with the prototype featuring our design, incorporating the three modules and the automatic pilot. In contrast, the control group interacted solely with the standard Lilobot interface.

For both groups, we provided an overview of the required skills where we explained how they can apply them. The intervention group's training content was structured across the three modules, as illustrated in Figure 4.2, while the control group received all training content at once, as shown in Figure 4.3.

Whenever a participant did not give input for 30 seconds, a hint appeared initially in a form pop-up, see Figure B.1a in appendix B.2.1, once the trainees read it, they clicked on the button 'Got it' and the pop-up disappeared, and the hint appeared on the right side of the screen above the explanation, see Figure B.1b in appendix B.2.1. Figure B.1a also shows an example of a hint. Similarly, at the end of each conversation, feedback was displayed first as a pop-up and then moved to the left side of the screen. This is illustrated in Figures B.2a and B.2b in appendix B.2.2. Both hints and feedback were available to participants in both groups.

Finally, a timer was implemented for each session to ensure participants spent the required amount of time in the session and did not attempt to bypass the process by having only one conversation before moving to the next session. Both the pre- and post-sessions were scheduled for 5 minutes to maintain comparability. Similarly, the training sessions for both the intervention and control groups were designed to last 18 minutes for consistency. For the intervention group, the training was divided into three modules with durations of 3 minutes, 5 minutes, and 10 minutes, respectively. These durations were determined based on the results of the pilot experiment.

The prototype was implemented using HTML, JavaScript, and CSS within IntelliJ IDEA. The code can be found online on GitHub [3].

Other Material

The participants were required to sign an informed consent form that provided information about the purpose of the study, the procedure of the experiment, the data collection and the risks of participating. Additionally, it included tick boxes highlighting key points that participants had to agree on to take part in the experiment. It can be found in the appendix B.3. The Qualtrics survey included instructions outlining the next steps in the experiment. Similarly, in the prototype before each session, we prepared a page that included instructions about the next session and a welcome and end page that had instructions about interacting with our system in general. You can see these instruction pages in appendix B.7, Figure B.3.

4.1.4. Measures

To test our hypotheses, we used two types of measures. To measure their performance (H1), we analysed the trainees' inputs during their interactions with Lilobot. For self-efficacy (H2) and perceived usefulness (H3), we utilised point-scale questionnaires. In addition to these measures, we also collected demographic data from the participants.

Performance

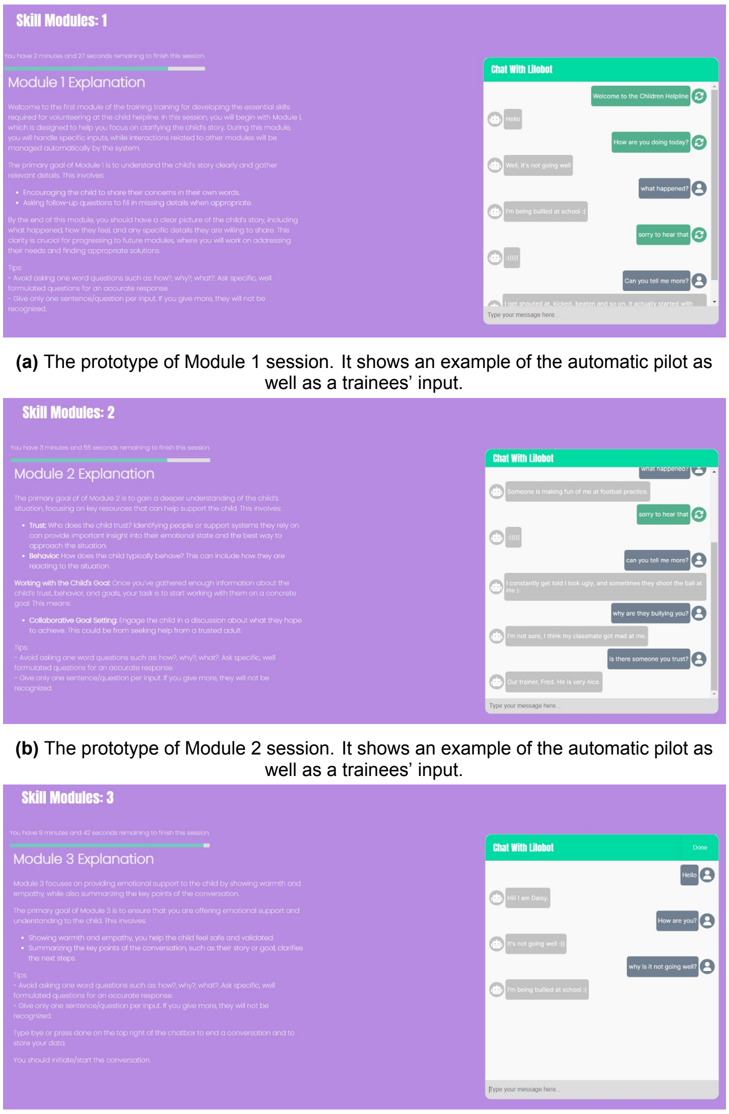
Our system categorised each user input to a predefined utterance, for example, when the trainee types ”hello”, it is categorised to the utterance ”request_chitchat_greeting”. These utterances were defined by Grundmann et al. (2025) [40], who introduced Lilobot. To evaluate trainees' performance, the utterances were classified into three distinct categories: good, neutral, and poor.

A good classification was assigned when the trainee's input was specific, or relevant, and appropriate for interacting with the child. For example, ”How many kids are bullying you?” was classified as good because it is a focused and meaningful question. Inputs were classified as neutral if they were not particularly relevant, such as ”What is your age?”. Finally, the poor category included inputs that were inappropriate or counterproductive at the current stage of the conversation, such as ”You are overreacting.” which is something a trainee should not say to a child. The categorisation list is shown in appendix B.5.

For each utterance, we assigned a performance value:

- **Good utterances** were assigned either 0.5 or 1, depending on the quality of the trainee's inputs, see Tables B.2, B.5, B.8 in appendix B.5.
- **Neutral utterances** were assigned 0, see Tables B.3, B.6, B.9 in appendix B.5.

[3] https://github.com/MayaElasmar/Pedagogical-Agent-Integration-Into-Simulation-Training.git



You have 2 minutes and 27 seconds remaining to finish this session.

Module 1 Explanation

Welcome to the first module of the training training for developing the essential skills required for volunteering at the child helpline. In this session, you will begin with Module 1, which is designed to help you focus on clarifying the child's story. During this module, you will handle specific inputs, while interactions related to other modules will be managed automatically by the system.

The primary goal of Module 1 is to understand the child's story clearly and gather relevant details. This involves:

- Encouraging the child to share their concerns in their own words.
- Asking follow-up questions to fill in missing details when appropriate.

By the end of this module, you should have a clear picture of the child's story, including what happened, how they feel, and any specific details they are willing to share. This clarity is crucial for progressing to future modules, where you will work on addressing their needs and finding appropriate solutions.

Tips:
- Avoid asking one word questions such as: how?, why?, what?. Ask specific, well formulated questions for an accurate response.
- Give only one sentence/question per input. If you give more, they will not be recognized.

Welcome to the Children Helpline

Hello

How are you doing today?

Well, it's not going well

what happened?

I'm being bullied at school :(

sorry to hear that

:(((((

Can you tell me more?

I get shouted at, kicked, beaten and so on. It actually started with

(a) The prototype of Module 1 session. It shows an example of the automatic pilot as well as a trainees' input.

You have 3 minutes and 55 seconds remaining to finish this session.

Module 2 Explanation

The primary goal of of Module 2 is to gain a deeper understanding of the child's situation, focusing on key resources that can help support the child. This involves:

- **Trust:** Who does the child trust? Identifying people or support systems they rely on can provide important insight into their emotional state and the best way to approach the situation.
- **Behavior:** How does the child typically behave? This can include how they are reacting to the situation.

Working with the Child's Goal: Once you've gathered enough information about the child's trust, behavior, and goals, your task is to start working with them on a concrete goal. This means:

- **Collaborative Goal Setting:** Engage the child in a discussion about what they hope to achieve. This could be from seeking help from a trusted adult.

Tips:
- Avoid asking one word questions such as: how?, why?, what?. Ask specific, well formulated questions for an accurate response.
- Give only one sentence/question per input. If you give more, they will not be recognized.

what happened?

Someone is making fun of me at football practice.

sorry to hear that

:(((((

can you tell me more?

I constantly get told I look ugly, and sometimes they shoot the ball at me):

why are they bullying you?

I'm not sure, I think my classmate got mad at me.

is there someone you trust?

Our trainer, Fred. He is very nice.

(b) The prototype of Module 2 session. It shows an example of the automatic pilot as well as a trainees' input.

You have 9 minutes and 42 seconds remaining to finish this session.

Module 3 Explanation

Module 3 focuses on providing emotional support to the child by showing warmth and empathy, while also summarizing the key points of the conversation.

The primary goal of Module 3 is to ensure that you are offering emotional support and understanding to the child. This involves:

- Showing warmth and empathy, you help the child feel safe and validated.
- Summarizing the key points of the conversation, such as their story or goal, clarifies the next steps.

Tips:
- Avoid asking one word questions such as: how?, why?, what?. Ask specific, well formulated questions for an accurate response.
- Give only one sentence/question per input. If you give more, they will not be recognized.

Type bye or press done on the top right of the chatbox to end a conversation and to store your data.

You should initiate/start the conversation.

Hello

Hiii I am Daisy.

How are you?

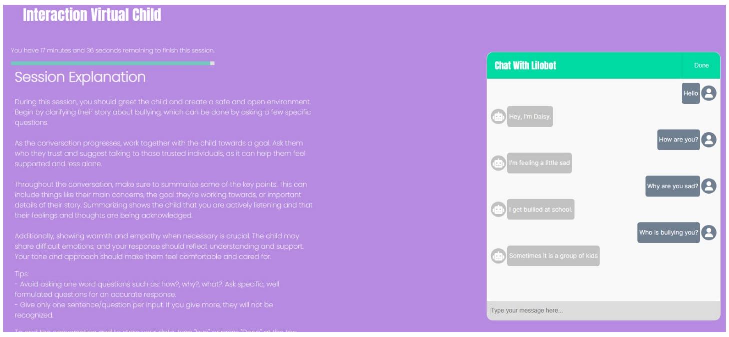
It's not going well :((

why is it not going well?

I'm being bullied at school :(

(c) The prototype of Module 3 session. It shows also that the automatic pilot is not available in this session.

Figure 4.2: The training prototype with modules for the intervention group.



You have 17 minutes and 36 seconds remaining to finish this session.

Session Explanation

During this session, you should greet the child and create a safe and open environment. Begin by clarifying their story about bullying, which can be done by asking a few specific questions.

As the conversation progresses, work together with the child towards a goal. Ask them who they trust and suggest talking to those trusted individuals, as it can help them feel supported and less alone.

Throughout the conversation, make sure to summarize some of the key points. This can include things like their main concerns, the goal they're working towards, or important details of their story. Summarizing shows the child that you are actively listening and that their feelings and thoughts are being acknowledged.

Additionally, showing warmth and empathy when necessary is crucial. The child may share difficult emotions, and your response should reflect understanding and support. Your tone and approach should make them feel comfortable and cared for.

Tips:
- Avoid asking one word questions such as: how?, why?, what?. Ask specific, well formulated questions for an accurate response.
- Give only one sentence/question per input. If you give more, they will not be recognized.

To end the conversation and to store your data, type "bye" or press "Done" at the top

Hello

Hey, I'm Daisy.

How are you?

I'm feeling a little sad

Why are you sad?

I get bullied at school.

Who is bullying you?

Sometimes it is a group of kids

Figure 4.3: The training prototype for the control group. There are no modules or an automatic pilot in this training session.

- **Poor utterances** were assigned a value of -1, see Tables B.4, B.7, B.10 in appendix B.5.

For each session, we calculated the performance across all conversations, using the weighted average. Inputs from the automatic pilot were excluded entirely from all calculations. We calculated the maximum points the trainee can obtain in a conversation and the points they obtained. Then, we divided the obtained points by the maximum points to get their performance. If the trainee had more than one conversation in the session, we calculated the average of these performances. If the last conversation was not finished because they ran out of time, then it is not included in the calculation. If the trainee could only manage one conversation but did not manage to finish it because of the time, the performance is also 0.

Self-efficacy

To measure self-efficacy, we developed a questionnaire inspired by three existing measures for self-efficacy including the Counseling Self-Estimate Inventory (COSE) by Larson et al.(1992) [49], and the SE-12 questionnaire by Axboe et al. (2016) [6] and Grundmann et al. (2025) [40]. We checked each of these questionnaires and selected the relevant items to our topic. The selected items and their sources are explained in more detail in Appendix B.6. The questionnaire consists of 9 statements that the participants rated on a 10- point scale with values from -5 ‘strongly disagree’, 0 ‘neutral’ to +5 ‘strongly agree’.

Perceived Usefulness

To measure the perceived usefulness, we developed a questionnaire based on an existing measure, which is the performance expectancy in the Unified Theory of Acceptance and Use of Technology (UTAUT) [35, 79]. We reviewed the performance expectancy items from the paper by Fitranie et al. (2021) [35], selected those relevant to our system, and modified certain terms to ensure their applicability. The questionnaire consists of 6 statements. Each statement was rated on a 7-point scale from 1 ’strongly disagree’, 4 ’neutral, to 7 ’strongly agree’.

Qualitative Measurement

In addition to collecting quantitative measures, we sought to understand participants’ subjective experiences with our system, particularly their opinions on the guidance provided by the pedagogical agent. To explore this, we included two open-ended questions: ”What did you like most about the system?” and ”What did you like least about the system?”

4.1.5. Procedure

The participants in the experiment followed a structured procedure, as illustrated in Figure 4.4. First, they accessed a link to Qualtrics on a laptop and selected a unique participant ID from a predefined list of 30 unique IDs. After selecting their ID, they signed an informed consent form and proceeded to fill out a demographics questionnaire. Subsequently, they completed a pre-questionnaire designed to assess their self-efficacy before being directed to the experimental system. Using the chosen participant ID, they signed up for the experiment and engaged in a 5-minute pre-session with the conversational agent, during which the pedagogical agent was absent, to measure their pre-training performance.

We created a list of participant ID’s: P1- P30. The participants randomly chose one of these. If they chose an even ID number, then they were placed in the intervention condition, while those with odd ID numbers were placed in the control condition. In the intervention condition, participants did three training sessions with a conversational agent guided by an adaptive pedagogical agent. The training consisted of a 3-minute session with Module 1, a 5-minute session with Modules 1 and 2, and a 10-minute session with all three modules. The duration of each session increased because participants had to complete more tasks and provide inputs for additional modules as they progressed. Participants in the control condition interacted with the system for 18 minutes without the pedagogical agent. Instead of the pedagogical agent. To ensure comparability with the intervention condition, the same virtual child was used, and the session duration was matched to the intervention’s 18 minutes.

After completing their assigned training sessions, participants engaged in a 5-minute post-session to assess their post-training performance. The same virtual child was used in all sessions to maintain consistency across participants. In total, there were three scenarios, and each module session in the intervention group was assigned to one of these scenarios. For the control group, as well as the pre- and post-sessions, each conversation was randomly assigned one of these scenarios. After completing the post-session, participants were redirected to Qualtrics, where they re-entered their Participant ID and completed a post-questionnaire on self-efficacy, followed by a questionnaire on perceived usefulness. Finally, they answered an open-ended questionnaire containing two additional questions to conclude the experiment.

4.1.6. Data Preparation & Statistical Analysis

The data was collected in two formats. First, the questionnaires on Qualtrics captured answers on Self-Efficacy and Perceived Usefulness. Second, participants’ interactions with Lilobot and their performance

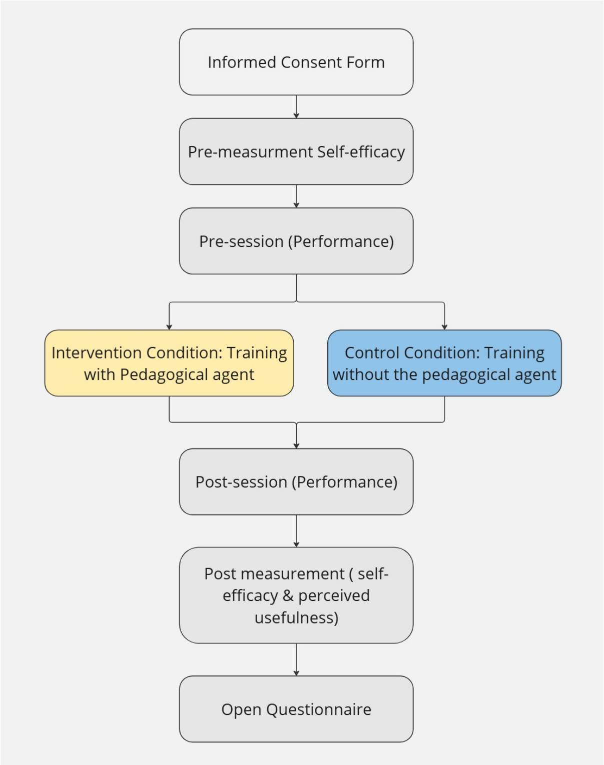

Informed Consent Form
Pre-measurment Self-efficacy
Pre-session (Performance)
Intervention Condition: Training with Pedagogical agent
Control Condition: Training without the pedagogical agent
Post-session (Performance)
Post measurement (self-efficacy & perceived usefulness)
Open Questionnaire

Figure 4.4: The procedure of our experiment.

during these interactions were saved as Excel files, which they uploaded to Qualtrics. The data cleaning and preprocessing were conducted in Python 3 using Jupyter Notebook. We cleaned the data by deleting incomplete surveys or participants who did not the conversation seriously. Further, we preprocessed the data by splitting the pre-survey in two: one for intervention group and one for control group. Then, we made sure these correspond with the post-survey based on the participant ID. Additionally, we calculated the average performance for each participant in Python and stored these in separate data frames. For the Self-Efficacy and Perceived Usefulness items, we calculated the average score of all items per participant and stored them in separate data frames. All items in both questionnaires were positively worded, so no score-reversing was required.

In R, we used the cleaned data frames containing these scores for the analyses. To calculate the deltas by subtracting the pre-scores from the post-scores for each participant, representing the improvement in performance or Self-Efficacy. These deltas were computed for both the intervention and control groups for performance and Self-Efficacy. An unpaired t-test was then conducted to compare the deltas across the two groups. For Perceived Usefulness, since there were no pre-training scores, we conducted an unpaired t-test to compare the mean post-training scores between the two groups.

Finally, the qualitative responses to the open-ended questions were analysed using thematic analysis, following the methodology proposed by Braun and Clarke [15]. To ensure the validity of this analysis, a person with a bachelor's degree in computer science conducted double-coding on the qualitative data. Initially, a coding scheme was developed and agreed upon for each question. Subsequently, each coder independently assigned all responses at once to one of the predefined themes. In R, we calculated a confusion matrix on the assignment of both coders and used this confusion matrix to calculate the Cohen's Kappa [22]. After completing the independent coding, the coders reviewed and discussed their differences in assignments. Through this discussion, they jointly decided on the final coding for all responses, ensuring consensus on the assigned themes.

All the data, Jupyter notebooks of the data cleaning and R Scripts of the analysis are available online for download through the 4TU.ResearchData repository [4].

[4] https://doi.org/10.4121/2e329e76-5f48-4d7f-bd70-4bd6d3238846

4.2. Results

4.2.1. Performance

The comparison of performance scores between pre- and post-sessions reveals that the intervention group demonstrated a higher average performance difference compared to the control group, as shown in Figure 4.5. To evaluate hypothesis H1, we conducted a Welch two-sample *t*-test to compare the performance deltas between the intervention group ($M = 0.13, SD = 0.19$) and the control group ($M = 0.01, SD = 0.19$). The results showed no statistically significant difference, $t(19.99) = 1.46, p = .160$, with a 95% confidence interval of $[-0.05, 0.29]$. This suggests that the intervention did not lead to significantly greater improvements in performance compared to the control group.

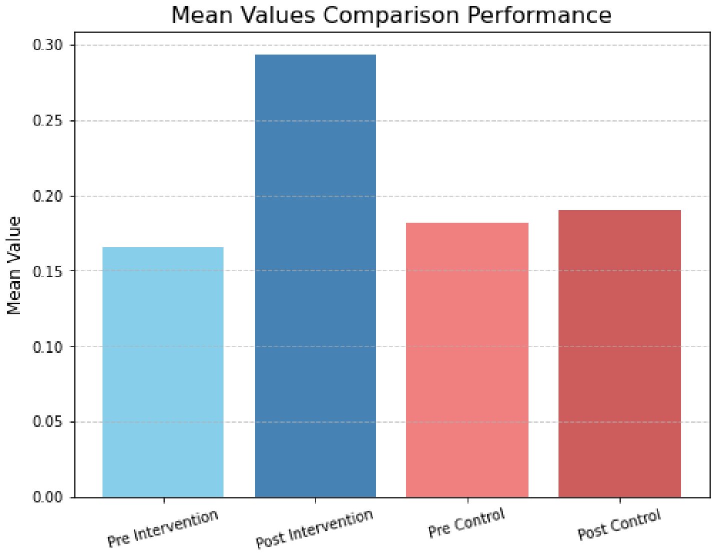


Figure 4.5: The means of the performance of pre- and post-session for the intervention(Blue) and the control(Red) group.

4.2.2. Self-Efficacy

First, we examined self-efficacy deltas between the intervention and control groups. As shown in Figure 4.5, the intervention group demonstrated a slight increase in self-efficacy from the pre- to post-session, whereas the control group showed a decrease. To evaluate hypothesis H2, we conducted a Welch two-sample *t*-test to compare self-efficacy deltas between the intervention group ($M = 0.04, SD = 0.89$) and the control group ($M = -0.33, SD = 1.44$). The results indicated no statistically significant difference, $t(16.7) = 0.73, p = .473$, with a 95% confidence interval of $[-0.70, 1.45]$. These findings do not provide sufficient evidence to support a difference in self-efficacy improvements between the intervention and control groups.

4.2.3. Perceived Usefulness

To evaluate hypothesis H3, we conducted a Welch two-sample *t*-test to compare perceived usefulness scores between the intervention group ($M = 5.47, SD = 0.83$) and the control group ($M = 5.29, SD = 0.89$). The results were not statistically significant, $t(19.92) = 0.49, p = .629$, with a 95% confidence interval of $[-0.58, 0.94]$. These results indicate no significant difference in perceived usefulness between the two groups. Additionally, we visualised the means of both groups in Figure 4.7. The average of the intervention is higher by only 0.2.

4.2.4. Explorative Results

To see whether there were negative effects of the training sessions, we analysed the percentage of participants whose performance and self-efficacy decreased in the post-session compared to the pre-session. As shown in Table 4.1, 27% of participants in the intervention group experienced a decline in performance, whereas this percentage was higher for the control group at 36%. Regarding self-efficacy, the majority of participants in the control group showed a decrease in the post-session (55%), while in the intervention group, the percentage was lower, with 36% experiencing a decline.

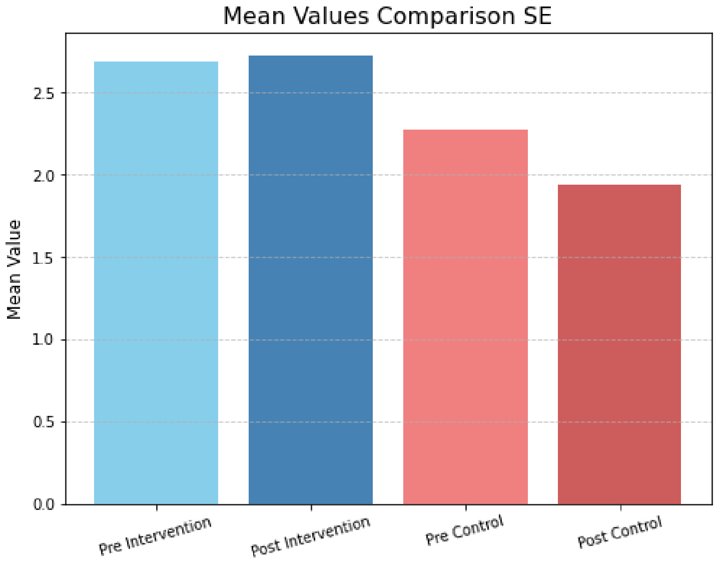


Figure 4.6: The means of the self-efficacy of pre- and post-session for the intervention(Blue) and the control(Red) group.

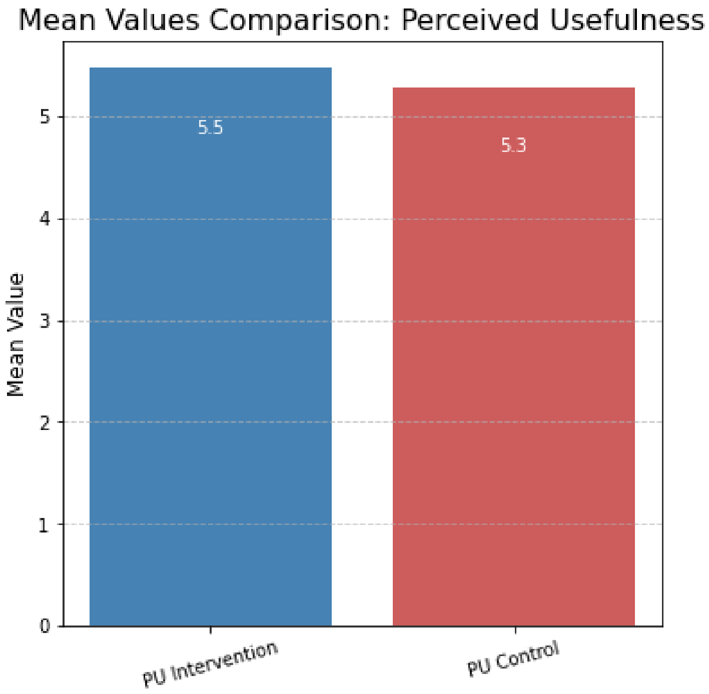


Figure 4.7: The means the perceived usefulness of the intervention(blue) and control(red) groups.

Modules Training

Besides our hypotheses, we also analysed how participants performed during the training session, specifically in the intervention group, which consisted of three modules. First, we analysed the performance data across the three modules, visualised in Figure 4.8. The boxplots reveal that the median performance improved from Module 1 to Module 3, indicating that participants developed their skills over time.

Control Training

For the control group, participants engaged in an 18-minute session during which they could have one or more conversations with Lilobot. We examined whether their performance improved with an increasing number of conversations. As illustrated in Figure 4.9, participants who engaged in multiple conversations generally had a higher average performance compared to those with fewer conversations. Upon

Table 4.1: Summary of Negative Percentages in Performance and SE Data

Dataset	Percentage Negative	Interpretation
Intervention Performance	27.27%	Decrease in performance after training
Control Performance	36.36%	Decrease in performance after training
Intervention Self-Efficacy	36.36%	Decrease in SE after training
Control Self-Efficacy	54.55%	Decrease in SE after training

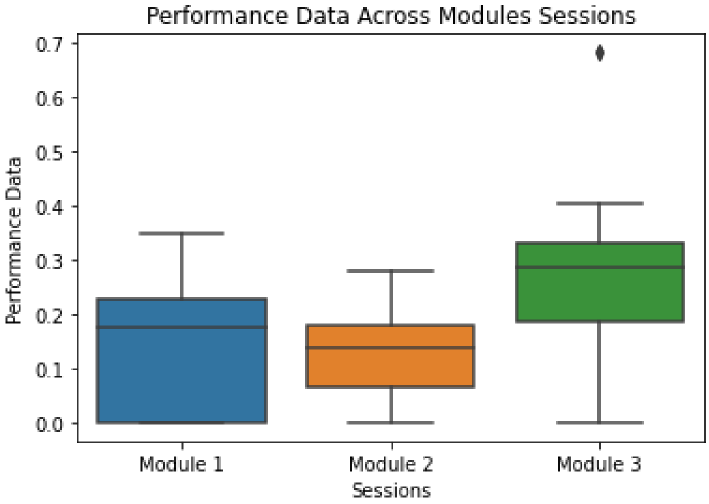


Figure 4.8: Boxplots showing the performance of participants across the three modules. The median performance improves from Module 1 to Module 3, but variability increases in Module 3.

inspecting the conversations, we observed that participants with fewer conversations tended to have longer individual conversations with Lilobot, while those with multiple conversations had shorter interactions.

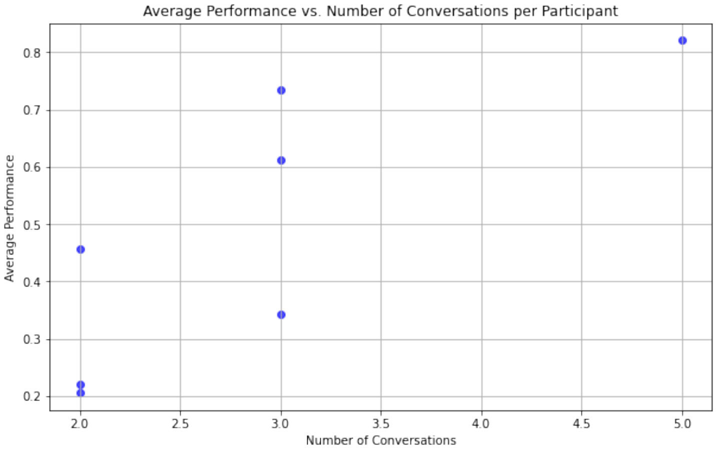


Figure 4.9: Scatter plot showing the relationship between the number of conversations and the average performance for participants in the control group. Each point represents a participant, with performance calculated as the average score across all conversations.

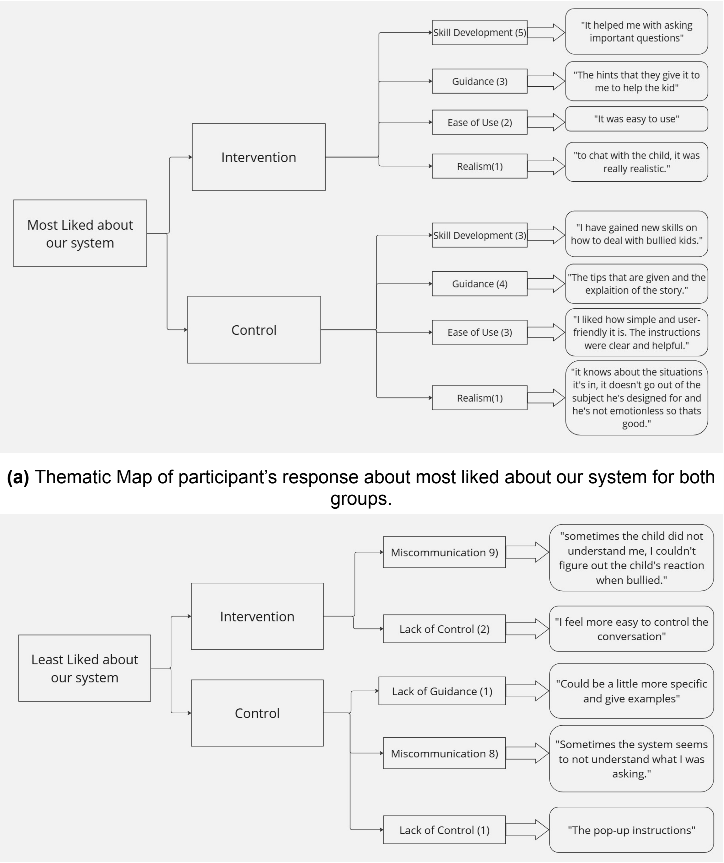
Open Questions

We conducted a qualitative measurement using two open-ended questions to gather participants' opinions about the system—specifically, what they liked most and least. The Intercoder reliability for the double coding was assessed using Cohen's Kappa [22], which indicated substantial agreement ($\kappa = 0.72$, 95% CI [0.57, 0.87]).

For the question, *“What did you like most about our system?”*, four themes were identified: Skill Development, Guidance, Ease of Use, and Realism. These themes with an example response are illustrated in Figure 4.10a. Skill Development describes responses where participants mentioned that they learned something from interacting with the system (nIntervention=5, nControl=3), for example: ”It helped me with asking important questions”. Guidance refers to responses highlighting the system’s guidance methods (nIntervention=4, nControl=4), such as hints, explanations, or the automatic pilot. As an example, there was a response ”The hints that they give it to me to help the kid.” Ease of Use encompasses responses indicating that the system was user-friendly (nIntervention=3, nControl=3), for example: ”it was easy to use”. Finally, Realism includes responses describing the chatbot as realistic (nIntervention=1, nControl=1), for example: ”to chat with the child, it was very realistic”.

For the question, *“What did you like least about our system?”*, three themes were identified: Miscommunication, Lack of Control, and Lack of Guidance, as illustrated in Figure 4.10b. Miscommunication refers to responses where participants mentioned that the chatbot did not always understand their input (nIntervention=9, nControl=8), for example: ”Sometime the child did not understand me”. Lack of Control refers to responses expressing a preference for control over the conversation (nIntervention=2, nControl=1), for example: ”I feel more easy to control the conversation”. Lack of Guidance relates to responses criticising the system’s guidance methods(nControl=1).

The majority of both groups mentioned that Lilobot did not always respond correctly to their question or did not understand the input. Both groups mentioned that they learned some new skill, or about being a counsellor from the system. Moreover, both groups mentioned that they liked the hints and the provided skills explanation. Unfortunately, the intervention group did not mention anything regarding the modules or automatic pilot. On the contrary, one mentioned that they liked being in control of the conversation.


Most Liked about our system
Intervention
Skill Development (5)
"It helped me with asking important questions"
Guidance (3)
"The hints that they give it to me to help the kid"
Ease of Use (2)
"It was easy to use"
Realism(1)
"to chat with the child, it was really realistic."
Control
Skill Development (3)
"I have gained new skills on how to deal with bullied kids."
Guidance (4)
"The tips that are given and the explaition of the story."
Ease of Use (3)
"I liked how simple and user-friendly it is. The instructions were clear and helpful."
Realism(1)
"it knows about the situations it's in, it doesn't go out of the subject he's designed for and he's not emotionless so thats good."
Least Liked about our system
Intervention
Miscommunication 9)
"sometimes the child did not understand me, I couldn't figure out the child's reaction when bullied."
Lack of Control (2)
"I feel more easy to control the conversation"
Control
Lack of Guidance (1)
"Could be a little more specific and give examples"
Miscommunication 8)
"Sometimes the system seems to not understand what I was asking."
Lack of Control (1)
"The pop-up instructions"

(a) Thematic Map of participant’s response about most liked about our system for both groups.

(b) Thematic Map of participant’s response about least liked about our system for both groups.

Figure 4.10: Thematic Map of participant’s response about most liked (Top Figure) and least liked (Bottom Figure) about our system.

4.3. Discussions

For our experiment, we hypothesised that incorporating the pedagogical agent would lead to higher learning outcomes. Despite higher means in performance differences between the pre- and post-sessions, our results showed no significant difference. Hence, we fail to reject the null hypothesis and there is no sufficient evidence to support H1. Although one of our design considerations was reducing cognitive load through the scaffolding mechanism, both the intervention and control groups received the same amount of information within the 18-minute training session. While the intervention group received the information in smaller chunks, they were still able to progress to the next module without necessarily mastering the skills from the previous one. This may have limited the effectiveness of the scaffolding approach in reinforcing learning. In contrast, real-life training sessions at a UK child helpline span over 11 weeks [66], allowing for a more gradual learning process and better retention of skills.

The results also showed no significant difference between the deltas of the pre- and post-questionnaires for self-efficacy, although the intervention group had higher means than the control group. Hence, we fail to reject the null hypothesis and there is no sufficient evidence to support H2. One possible explanation is that both groups received adaptive feedback and static hints, which may have provided similar levels of guidance and support. To see more of the effect of the adaptive guidance on the self-efficacy, the feedback and hints should have been static for the control group and adaptive for the intervention group. Wu et al. (2023) [83] showed that adaptive support enhances learner's self-efficacy with a clear distinction between conditions that might have led to more conclusive effects.

The results of the t-test on the perceived usefulness showed no significant difference between the intervention group and the control group. Hence, we fail to reject the null hypothesis and there is no sufficient evidence to support H3. Both groups received hints when they encountered difficulties in progressing through the conversation. Although the intervention group benefited from a scaffolding mechanism and an automatic pilot designed to facilitate skill learning, both groups received the same amount of information within the same time frame. As a result, participants in the control group may have perceived that they had received sufficient information, especially since they had no reference point to compare their condition with. This lack of comparison might have influenced their perception, and it is possible that clearer differences would have emerged if participants had experienced both implementations. Furthermore, participants in both groups highlighted that they appreciated the hints and explanations provided during the training. Interestingly, none of the participants in the intervention group mentioned liking the modules training or the automatic pilot. On the contrary, some intervention group participants noted that they preferred having full control over the conversation, suggesting that this aspect of the design might not have been liked by everyone.

There are some interesting takeaways from the explorative results. In both intervention and control training sessions, the performance gets higher when they train more or have more conversations, see figure 4.8, 4.9. This observation aligns with the findings of Lee et al. (1991) [50], who emphasised that repetition and practice enhance skill acquisition, particularly when practice conditions promote active engagement and cognitive processing. The main feedback we got about the system in terms of guidance is that they liked the hints and the explanations of the skills. Some did not even like the automatic pilot as they liked to have control over the conversation.

4.4. Limitations

For the initial power calculation, we assumed a correlation between pre and post measurements of 0.7 and an effect size of 0.8 for the difference in improvement between the intervention and control groups. After analysing the actual results, we calculated the correlation and effect size for each measure. For performance, the correlations was -0.36 (intervention) and 0.26 (control), with an effect size of 0.62. According to Cohen's guidelines [22], an effect size of 0.62 falls within the medium-to-large range, suggesting a meaningful difference between groups. For self-efficacy, the correlations were 0.54 (intervention) and 0.30 (control), with effect size of 0.31. Using the performance values (as they represent the highest effect size), we recalculated the required sample size. The new calculation indicates that a sample size of 108 would achieve a power of 0.72. Hence, we believe that if H1 is still valid, the smaller correlation and effect size observed suggest that the experiment was underpowered. This could explain why, despite the fact that the sample means showed improvement, the t-test was unable to confirm this with statistical significance. A larger sample size than 22 would likely have provided sufficient power to detect meaningful differences.

Our qualitative analysis revealed frequent negative responses from participants that the chatbot sometimes failed to understand their input or repeated responses. This is probably a consequence of the relatively simple NLU approach we have applied in our prototype. This issue may have negatively impacted participants' performance. Morover, our original idea was to prevent trainees from progressing to the next session until they successfully passed the corresponding module. However, to accommodate the time limitations of this experiment, participants were allowed to advance to subsequent modules even if they had not successfully mastered the current one, potentially limiting their ability to fully learn the required skills.

5 Discussion & Conclusion

5.1. Conclusion

This research aimed to answer the following research question:

How can the integration of the pedagogical agent enhance the effectiveness of counselling training through role-play simulation training ?

We broke down this research question into the following three subquestions:

1. What are the key design considerations for the integration of a pedagogical agent?

We addressed this question through a literature review and a focus group discussion. The investigation focused on two main aspects: the technical and conceptual integration of a pedagogical agent. Based on our findings, we concluded that the pedagogical agent should teach skills to trainees gradually, in alignment with John Sweller's cognitive load theory [76], to avoid overwhelming their cognitive capacity.

Additionally, the pedagogical agent should provide personalised guidance tailored to the trainee's individual needs, ensuring a supportive learning experience. It is essential for the trainees to actively and positively engage with the training simulation; thus, the pedagogical agent should incorporate active learning elements and offer motivational feedback when trainees perform well. Finally, when trainees encounter difficulties or fail to make progress, the pedagogical agent should proactively provide them with the necessary guidance to help them move forward.

2. How can a pedagogical agent be integrated in a simulation training environment? We integrated the pedagogical agent into the simulation training environment in several ways. We developed an adaptive pedagogical agent that incorporates a scaffolding mechanism. Our adaptive framework adapted with the three models of Intelligent Tutoring Systems (ITS), shown in Figure 3.2. Our learner model is the trainee's prior knowledge about counselling skills. Our content model is the training content which was divided into three modules, each focusing on specific skills. Finally, our instructional model included the scaffolding mechanism, an automatic pilot, feedback and hints.

The training was structured across three sessions to align with these modules. All session include all three modules. In Session 1, the trainee works exclusively on Module 1, while an automatic pilot takes over the trainees' inputs relevant to Modules 2 and 3. If the trainee successfully completes Module 1, they progress to Session 2; otherwise, they repeat Session 1. In Session 2, the trainee trains with Modules 1 and 2, with the automatic pilot managing inputs for Module 3. Upon passing both Modules 1 and 2, they move to Session 3. Finally, in Session 3, the trainee applies all three modules without assistance from the automatic pilot.

Additionally, the adaptive pedagogical agent provided personalised feedback that adjusted dynamically to the trainee's performance. In contrast, the hints offered by the pedagogical agent were static, providing general suggestions.

3. To what extent does an integrated pedagogical agent in the simulation training environment contribute to the trainees' learning?

To address this question, we conducted an experiment in a mixed study design that combined between-subjects and within-subjects elements. The experiment involved 22 participants. The measures used included performance during the pre- and post-sessions, pre- and post-self-efficacy questionnaires, and a perceived usefulness questionnaire.

For all three measures, the t-tests on the deltas revealed no significant differences between the intervention and control groups. However, the intervention group consistently showed higher means across all measures. These findings shows a potential that while statistical significance was not achieved, there was a promising direction towards our hypotheses, especially the performance measure. Hence, there is a likelihood that this happened, because our experiment was underpowered.

From the qualitative analysis, we observed that many participants from both groups reported developing new skills through our system and appreciated the guidance provided by the pedagogical agent.

5.2. Limitations & Future Work

One major limitation of this study was the relatively small sample size (22 participants), which may have resulted in an underpowered experiment, reducing the likelihood of detecting statistically significant differences despite promising trends in the results. Future research should consider replicating this study with a larger sample size—108 participants, as projected in our power calculation based on the observed correlation and effect size.

Additionally, our pedagogical agent was limited in its ability to provide dynamic hints or an enhanced automatic pilot that could intervene based on real-time trainee input. Instead, it relied on predefined conversations. Another potential limitation was the training duration. The 18-minute sessions may not have provided sufficient time for participants to internalise and apply the skills effectively. Future studies could extend training over multiple sessions to allow for gradual learning and reinforcement. Furthermore, incorporating long-term assessments of skill retention and application—such as follow-up evaluations in real-world child helpline environments—could offer deeper insights into the system's practical impact.

Finally, the natural language understanding (NLU) component relied on basic intent classification methods rather than advanced deep learning-based techniques. Prior research suggests that large language models (LLMs) can significantly enhance conversational AI by improving response accuracy and contextual awareness [51, 62]. Future work could explore integrating LLMs to improve user interactions and adaptability in the training system.

5.3. Contributions

At an academic level, this study contributes to the broader field of incorporating pedagogical agents into learning systems. While the results were inconclusive, the results show a promising trend for further research and suggested the potential for finding more conclusive insights with a larger sample size or refined experimental design. We also propose a design framework for integrating pedagogical agents into simulation-based training systems. Our framework leverages scaffolding to introduce skills progressively, adaptivity to tailor guidance to individual needs, and structured learning models to organise training into modular steps. This approach provides a possible foundation for applying pedagogical agents across various domains requiring interactive skill development. Hence, we contributed with a research of the integration of pedagogical agent into simulation training for volunteer counsellors of child helplines.

5.4. Final Remarks

In this thesis, we integrated a pedagogical agent into a role-play simulation-based counselling training system designed for volunteers at child helplines. While results were inconclusive, they highlight a promising direction for future research, with the potential for stronger insights through larger sample sizes or refined designs. This study contributes to the integration of pedagogical agents in training systems by proposing a framework that combines scaffolding, adaptivity, and structured learning.

References

[1] Oct. 2023. url: https://childhelplineinternational.org/voices-of-children-young-people-around-the-world-2022-data/.

[2] url: https://www.kindertelefoon.nl/informatie-voor-volwassenen/over-ons.

[3] url: https://www.kindertelefoon.nl/vrijwilliger/vertel-me-meer.

[4] Mohammed Al Owayyed et al. "A Cognitive Conversational Agent for Training Child Helpline Volunteers". In: *Proceedings of the 24th ACM International Conference on Intelligent Virtual Agents*. IVA '24. GLASGOW, United Kingdom: Association for Computing Machinery, 2024. isbn: 9798400706257. doi: 10.1145/3652988.3696197. url: https://doi-org.tudelft.idm.oclc.org/10.1145/3652988.3696197.

[5] Norzaliza Alis, Wan Marzuki Wan Jaafar, and Ahmad Fauzi Mohd Ayub. "The Influence of Self-Reflection Towards Counselor Trainee Self-Development". In: *Middle-East Journal of Scientific Research* 19.Innovation Challenges in Multidisciplinary Research & Practice (2014). Corresponding Author: Wan Marzuki Wan Jaafar, Faculty of Educational Studies, Universiti Putra Malaysia, 43400, UPM Serdang, Selangor, Malaysia. Tel: +603-89468125., pp. 85–88. issn: 1990-9233. doi: 10.5829/idosi.mejsr.2014.19.icmrp.13.

[6] Mette K Axboe et al. "Development and validation of a self-efficacy questionnaire (SE-12) measuring the clinical communication skills of health care professionals". In: *BMC medical education* 16 (2016), pp. 1–10.

[7] Steve Olusegun BADA. "Constructivism Learning Theory: A Paradigm for Teaching and Learning". In: *IOSR Journal of Research & Method in Education* 5.6 (Ver. I Nov. 2015), pp. 66–70. doi: 10.9790/7388-05616670. url: https://www.iosrjournals.org.

[8] Amy L Baylor and Yanghee Kim. "Simulating instructional roles through pedagogical agents". In: *International Journal of Artificial Intelligence in Education* 15.2 (2005), pp. 95–115.

[9] Carl Bereiter. "Constructivism, socioculturalism, and Popper's World 3". In: *Educational Researcher* 23.7 (1994), pp. 21–23.

[10] Alan Blackwell. "Human computer interaction notes". In: *Advanced Graphics & HCI* 26 (2001), p. 2001.

[11] Charles C Bonwell and James A Eison. *Active learning: Creating excitement in the classroom. 1991 ASHE-ERIC higher education reports.* ERIC, 1991.

[12] Kim Bosman, Tibor Bosse, and Daniel Formolo. "Virtual Agents for Professional Social Skills Training: An Overview of the State-of-the-Art". In: *Intelligent Virtual Agents*. Ed. by Shweta Kapoor et al. Cham: Springer International Publishing, 2019, pp. 81–83. isbn: 978-3-030-16447-8. url: https://eudl.eu/pdf/10.1007/978-3-030-16447-8_8.

[13] Ayrton Braam. "A feedback system for a children's helpline training-chatbot". Contributor: Brinkman, W.P. (graduation committee) ORCID 0000-0001-8485-7092; Al Owayyed, M. (mentor); Gadiraju, Ujwal (graduation committee) ORCID 0000-0002-6189-6539. MA thesis. Delft University of Technology, Dec. 2023. url: http://resolver.tudelft.nl/uuid:d4efa0df-85ad-40de-88d7-ef21d1ff3d4f.

[14] Cynthia J. Brame. *Active Learning*. CFT Assistant Director. Accessed: 2024-06-26. url: https://cft.vanderbilt.edu/wp-content/uploads/sites/59/Active-Learning.pdf.

[15] Virginia Braun and Victoria Clarke. "Using thematic analysis in psychology". In: *Qualitative research in psychology* 3.2 (2006), pp. 77–101.

[16] Deborah L. Butler and Philip H. Winne. "Feedback and Self-Regulated Learning: A Theoretical Synthesis". In: *Review of Educational Research* 65.3 (1995), pp. 245–281. url: https://www.jstor.org/stable/1170684.

[17] Yuh-Ming Cheng et al. “Building a General Purpose Pedagogical Agent in a Web-Based Multimedia Clinical Simulation System for Medical Education”. In: *IEEE Transactions on Learning Technologies* 2.3 (July 2009), pp. 216–225. doi: 10.1109/TLT.2009.31.

[18] Morgane Chevalier et al. “The role of feedback and guidance as intervention methods to foster computational thinking in educational robotics learning activities for primary school”. In: *Computers & Education* 182 (2022), p. 104431. doi: 10.1016/j.compedu.2022.104431. url: https://doi.org/10.1016/j.compedu.2022.104431.

[19] Child Helpline International. *eLearning Platform: Our Courses*. Dec. 2022. url: https://childhelplineinternational.org/elearning-platform-our-courses/.

[20] Children 1st. *Children 1st Helpline Volunteer Role Description*. https://www.children1st.org.uk/media/2j5b3px4/children-1st-helpline-volunteer-role-description.pdf. Published: 19/07/2017. Children 1st, 2017.

[21] Richard E. Clark. *How Much and What Type of Guidance is Optimal for Learning from Instruction?* University of Southern California. n.d. url: https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=c3e72429596530bd8add73a28793c1ea57ace249.

[22] Jacob Cohen. “A coefficient of agreement for nominal scales”. In: *Educational and psychological measurement* 20.1 (1960), pp. 37–46.

[23] Allan Collins and Manu Kapur. *Cognitive Apprenticeship*. https://www.mcw.edu/-/media/MCW/Education/Academic-Affairs/OEI/Faculty-Quick-Guides/Cognitive-Apprenticeship-Theory.pdf. May 2022.

[24] Allan Collins and Manu Kapur. “Cognitive Apprenticeship”. In: pp. 109–127. doi: 10.1017/CBO9781139519526.008.

[25] Brian Robert Cook and Andrea Babon. “Active learning through online quizzes: better learning and less (busy) work”. In: *Journal of Geography in Higher Education* 41.1 (2017). Published online: 14 May 2016, pp. 24–38. doi: 10.1080/03098265.2016.1185772. url: https://doi.org/10.1080/03098265.2016.1185772.

[26] Mick Cooper and John McLeod. “Client Helpfulness Interview Studies: A Guide to Exploring Client Perceptions of Change in Counselling and Psychotherapy”. In: (Nov. 2015). doi: 10.13140/RG.2.1.2086.0885.

[27] Laduona Dai et al. “A systematic review of pedagogical agent research: Similarities, differences and unexplored aspects”. In: *Computers & Education* 190 (2022), p. 104607.

[28] Fred D. Davis. “Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology”. In: *MIS Quarterly* 13.3 (Sept. 1989), pp. 319–339. doi: 10.2307/249008.

[29] Vanessa P Dennen and Kerry J Burner. “The cognitive apprenticeship model in educational practice”. In: *Handbook of research on educational communications and technology*. Routledge, 2008, pp. 425–439.

[30] Vladan Devedzic and Andreas Harrer. “Architectural patterns in pedagogical agents”. In: *International Conference on Intelligent Tutoring Systems*. Springer. 2002, pp. 81–90.

[31] Melissa C. Duffy and Roger Azevedo. “Motivation matters: Interactions between achievement goals and agent scaffolding for self-regulated learning within an intelligent tutoring system”. In: *Computers in Human Behavior* 52 (2015), pp. 338–348. doi: 10.1016/j.chb.2015.05.041. url: https://doi.org/10.1016/j.chb.2015.05.041.

[32] Bernd-Joachim Ertelt, William E. Schulz, and Andreas Frey. *Counsellor Competencies: Developing Counselling Skills for Education, Career and Occupation*. Springer, 2022, pp. 33–38. url: https://link.springer.com/content/pdf/10.1007/978-3-030-87413-1.pdf.

[33] Loris Fichera et al. “Flexible Robot Strategy Design Using Belief-Desire-Intention Model”. In: *Communications in Computer and Information Science ((CCIS, volume 156))*. 2011. url: https://link.springer.com/chapter/10.1007/978-3-642-27272-1_5.

[34] Dylan Keifer Finch and Stephen H Edwards. “Using a pedagogical agent to support students learning to program”. In: *2020 ASEE Virtual Annual Conference Content Access*. 2020.

[35] Siska Fitrianie et al. “Factors affecting user’s behavioral intention and use of a mobile-phone-delivered cognitive behavioral therapy for insomnia: A small-scale UTAUT analysis”. In: *Journal of Medical Systems* 45 (2021), pp. 1–18.

[36] Chan Yuen Fook et al. “Relationship Between Active Learning and Self Efficacy Among Students in Higher Education”. In: *International Academic Research Journal of Social Science* 1.2 (2015), pp. 139–149.

[37] Tim Freeman. “’Best practice’ in focus group research: making sense of different views”. In: *Journal of Advanced Nursing* 56.5 (2006), pp. 491–497. doi: `10.1111/j.1365-2648.2006.04043.x`.

[38] Ruben G Fukkink, Rudy Ligtvoet, and Suzan Bruns. 2016. url: `https://onlinelibrary.wiley.com/doi/full/10.1111/chso.12150?casa_token=Nqa2FYY5sSoAAAAA%3AtCWHuENfgUT6fH4BLx5jSOT9niJhrm_l5Mq-aK7jIBbFgo-3BXC4VYm_Q5O_V2jcrVDHQou3OnENhw`.

[39] Ruben G. Fukkink and Jo M.A. Hermanns. “Children’s experiences with chat support and telephone support”. In: *Journal of Child Psychology and Psychiatry* (2009). doi: `10.1111/j.1469-7610.2008.02024.x`. url: `https://doi.org/10.1111/j.1469-7610.2008.02024.x`.

[40] S. Grundmann, M. Al Owayyed, M. Bruijnes, et al. “Lilobot: A Cognitive Conversational Agent to Train Counsellors at Children’s Helplines”. In: *Journal of Medical Systems* 49.5 (2025), p. 5. doi: `10.1007/s10916-024-02121-8`. url: `https://doi.org/10.1007/s10916-024-02121-8`.

[41] Agneta Gulz et al. “Building a Social Conversational Pedagogical Agent: Design Challenges and Methodological Approaches”. In: *Conversational Agents and Natural Language Interaction: Techniques and Effective Practices*. Ed. by Diana Perez-Marin and Ismael Pascual-Nieto. Hershey, PA: IGI Global, 2011, pp. 128–155.

[42] Petra Hendrickson. “Effect of active learning techniques on student excitement, interest, and self-efficacy”. In: *Journal of Political Science Education* 17.2 (2021), pp. 311–325.

[43] W Lewis Johnson, Erin Shaw, and Rajaram Ganeshan. “Pedagogical agents on the web”. In: *ITS*. Vol. 98. Citeseer. 1998, pp. 2–7.

[44] W. Lewis Johnson. “Pedagogical agent research at CARTE”. In: *AI Magazine* 22.4 (2001), pp. 85–97.

[45] W. Lewis Johnson and James C. Lester. “Pedagogical Agents: Back to the Future”. In: *International Journal of Artificial Intelligence in Education* 15.4 (2005), pp. 353–358. doi: `10.1007/s40593-005-0001-3`.

[46] Yanghee Kim and Amy L. Baylor. “Research-Based Design of Pedagogical Agent Roles: a Review, Progress, and Recommendations”. In: *International Journal of Artificial Intelligence in Education* 26 (2016). Published online: 7 July 2015, pp. 160–169. doi: `10.1007/s40593-015-0055-y`. url: `https://doi.org/10.1007/s40593-015-0055-y`.

[47] Paul Kirschner and Jeroen J. G. van Merriënboer. *Ten Steps to Complex Learning: A New Approach to Instruction and Instructional Design*. New York: Routledge, 2013. isbn: 978-0-415-89484-0.

[48] H Chad Lane et al. “The effects of a pedagogical agent for informal science education on learner behaviors and self-efficacy”. In: *Artificial Intelligence in Education: 16th International Conference, AIED 2013, Memphis, TN, USA, July 9-13, 2013. Proceedings 16*. Springer. 2013, pp. 309–318.

[49] Lisa M Larson et al. “Development and validation of the counseling self-estimate inventory.” In: *Journal of counseling Psychology* 39.1 (1992), p. 105.

[50] Timothy D Lee, Laurie R Swanson, and Anne L Hall. “What Is Repeated in a Repetition? Effects of Practice Conditions on Motor Skill Acquisition”. In: *Physical Therapy* 71.2 (Feb. 1991), pp. 150–156. doi: `10.1093/ptj/71.2.150`. url: `https://doi.org/10.1093/ptj/71.2.150`.

[51] Na Liu et al. “From llm to conversational agent: A memory enhanced architecture with fine-tuning of large language models”. In: *arXiv preprint arXiv:2401.02777* (2024).

[52] Dongxu Lu. “Emotion Model for Child Helpline Training Tool”. Master thesis. MA thesis. Delft University of Technology, 2023. url: `http://resolver.tudelft.nl/uuid:d07f81de-35de-4a2c-a363-9169d53c90b5`.

[53] Ati Suci Dian Martha and Harry B. Santoso. “The Design and Impact of the Pedagogical Agent: A Systematic Literature Review”. In: *International Journal of Information and Education Technology* 9.4 (2019), pp. 1–5. url: `https://files.eric.ed.gov/fulltext/EJ1204376.pdf`.

[54] Ati Suci Dian Martha et al. “A Scaffolding Design for Pedagogical Agents within the Higher-Education Context”. In: *Proceedings of the Conference on Educational Technologies (ICEduTech)*. Universitas Indonesia. Depok, Indonesia, 2020.

[55] Ati Suci Dian Martha et al. “The effect of the integration of metacognitive and motivation scaffolding through a pedagogical agent on self-and co-regulation learning”. In: *IEEE Transactions on Learning Technologies* 16.4 (2023), pp. 573–584.

[56] Florence Martin et al. “Systematic Review of Adaptive Learning Research Designs, Context, Strategies, and Technologies From 2009 to 2018”. In: *STEMPS Faculty Publications, STEM Education & Professional Studies* (2020). url: https://digitalcommons.odu.edu/stemps_fac_pubs/2020.

[57] Joseph Adrianus Maryadi, Harry Budi Santoso, and Yugo Kartono Isa. “Development of Personalized Pedagogical Agent for Student-Centered e-Learning Environment”. In: *2017 7th World Engineering Education Forum (WEEF)*. IEEE, Nov. 2017. doi: 10.1109/WEEF.2017.8467028.

[58] Kimberly H. McManama O’Brien et al. “Suicide risk assessment training using an online virtual patient simulation”. In: *Mhealth* 5 (2019), p. 31. doi: 10.21037/mhealth.2019.08.03. url: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6737388/.

[59] Jeroen J G van Merriënboer and John Sweller. “Cognitive load theory in health professional education: design principles and strategies”. In: *Medical Education* 44 (1 2010), pp. 85–93. doi: 10.1111/j.1365-2923.2009.03498.x.

[60] Elham Mousavinasab et al. “Intelligent tutoring systems: a systematic review of characteristics, applications, and evaluation methods”. In: *Interactive Learning Environments* 29.1 (2021), pp. 142–163.

[61] Melissa Murray and Gerson Tenenbaum. “Computerized pedagogical agents as an educational means for developing physical self-efficacy and encouraging activity in youth”. In: *Journal of Educational Computing Research* 42.3 (2010), pp. 267–283.

[62] Humza Naveed et al. “A comprehensive overview of large language models”. In: *arXiv preprint arXiv:2307.06435* (2023).

[63] Mark A. Neerincx. *Socio-Cognitive Engineering (SCE) Tool Manual*. Available at https://scetool.ewi.tudelft.nl/sites/default/files/sce_manual_v1.0.pdf. Apr. 2020.

[64] Mark A. Neerincx et al. “Socio-Cognitive Engineering of a Robotic Partner for Child’s Diabetes Self-Management”. In: *Frontiers in Robotics and AI* 6.118 (Nov. 2019), pp. 1–18. issn: 2296-9144. doi: 10.3389/frobt.2019.00118. url: https://doi.org/10.3389/frobt.2019.00118.

[65] Barbara M. Newman and Philip R. Newman. “Self-regulation theories”. In: *Theories of Adolescent Development*. Ed. by Richard M. Lerner and Laurence Steinberg. Routledge, 2020. url: https://www.sciencedirect.com/topics/psychology/self-regulation-theory.

[66] NSPCC. *Childline Volunteers*. 2024. url: https://join-us.nspcc.org.uk/volunteers/volunteers/childline/.

[67] Hyacinth S. Nwana. “Intelligent Tutoring Systems: an overview”. In: *Artificial Intelligence Review* 4 (1990), pp. 251–277.

[68] Tobias O. Nyumba et al. “The use of focus group discussion methodology: Insights from two decades of application in conservation”. In: *Methods in Ecology and Evolution* 8.1 (2018), pp. 20–32. doi: 10.1111/2041-210X.12860.

[69] Pipatsarun Phobun and Jiracha Vicheanpanya. “Adaptive intelligent tutoring systems for e-learning systems”. In: *Procedia - Social and Behavioral Sciences* 2 (2010). Received November 2, 2009; revised December 10, 2009; accepted January 18, 2010, pp. 4064–4069. url: https://www.sciencedirect.com.

[70] Anand Rao and Michael Georgeff. “BDI agents: From theory to practice”. In: *Proceedings of the First International Conference on Multi-Agent Systems (ICMAS 1995)*. San Francisco, CA, 1995, pp. 312–319.

[71] Troy D Sadler et al. “Learning science through research apprenticeships: A critical review of the literature”. In: *Journal of Research in Science Teaching: The Official Journal of the National Association for Research in Science Teaching* 47.3 (2010), pp. 235–256.

[72] Noah L. Schroeder and Chad M. Gotch. “Persisting Issues in Pedagogical Agent Research”. In: *Journal of Educational Computing Research* 53.2 (2015), pp. 183–204. doi: 10.1177/0735633115597625. url: https://jec.sagepub.com.

[73] Pieta Sikström et al. “How pedagogical agents communicate with students: A two-phase systematic review”. In: *Computers & Education* 188 (2022), p. 104564.

[74] TRINE N SINDAHL. “THE EFFECTS OF CHAT COUNSELLING”. In: *CHAT COUNSELLING FOR CHILDREN AND YOUTH A HANDBOOK*.

[75] Frederick Sundram et al. “Motivations, Expectations and Experiences in Being a Mental Health Helplines Volunteer”. In: *International Journal of Environmental Research and Public Health* 15.10 (2018). This article belongs to the Section Global Health, p. 2123. doi: `10.3390/ijerph15102123`.

[76] J. Sweller. “Cognitive Load during Problem Solving: Effects on Learning”. In: *Cognitive Science* 12 (1988), pp. 257–285.

[77] Jeroen JG Van Merrienboer et al. “Redirecting learners’ attention during training: Effects on cognitive load, transfer test performance and training efficiency”. In: *Learning and instruction* 12.1 (2002), pp. 11–37.

[78] Kurt VanLehn. “The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems”. In: *Educational psychologist* 46.4 (2011), pp. 197–221.

[79] Viswanath Venkatesh, James YL Thong, and Xin Xu. “Unified theory of acceptance and use of technology: A synthesis and the road ahead”. In: *Journal of the association for Information Systems* 17.5 (2016), pp. 328–376.

[80] I. M. Verenikina. “Vygotsky’s Socio-Cultural Theory and the Zone of Proximal Development”. In: *Expanding the Horizon. Information Systems and Activity Theory*. Ed. by H. M. Hasan, I. M. Verenikina, and E. L. Gould. Wollongong: University of Wollongong Press, 2003, pp. 4–14.

[81] Piek Vossen, Selene Baez, and Bram Kraaijeveld. “Leolani: A reference machine with a theory of mind for Social Communication”. In: Jan. 1970. url: `https://link.springer.com/chapter/10.1007/978-3-030-00794-2_2`.

[82] L. C. Webster, T. M. Hastings, and K. Garrett. “Effective Writing Strategies and Feedback in Counselor Education”. In: *Journal of Counselor Preparation and Supervision* 17.2 (2023). url: `https://digitalcommons.sacredheart.edu/jcps/vol17/iss2/9`.

[83] Ting-Ting Wu et al. “Leveraging computer vision for adaptive learning in STEM education: Effect of engagement and self-efficacy”. In: *International Journal of Educational Technology in Higher Education* 20.1 (2023), p. 53.

[84] Dongming Xu and Huaiqing Wang. “Intelligent agent supported personalization for virtual learning environments”. In: *Decision Support Systems* (Aug. 2005). doi: `10.1016/j.dss.2005.05.033`. url: `https://doi.org/10.1016/j.dss.2005.05.033`.

[85] Zi Yan. “Self-assessment in the process of self-regulated learning and its relationship with academic achievement”. In: *Assessment & Evaluation in Higher Education* 45.2 (2020), pp. 224–238.

[86] Elaine HJ Yew and Karen Goh. “Problem-based learning: An overview of its process and impact on learning”. In: *Health professions education* 2.2 (2016), pp. 75–79.

[87] SUN Yu and LI Zhiping. “Intelligent Pedagogical Agents for Intelligent Tutoring Systems”. In: *2008 International Conference on Computer Science and Software Engineering*. IEEE. Los Alamitos, CA, USA, Dec. 2008, pp. 420–423. doi: `10.1109/CSSE.2008.1272`. url: `https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=4721800`.

A

Appendix Foundation

A.1. The Five Phase Model

The five phase model is a structured counselling approach aimed at improving the effectiveness of counselling sessions [74], which was later adapted by the Danish helpline Børns Vilkår. The five phase model starts with setting a tone for the conversation, then encouraging the child to share their issue. Following this, the counsellor collaborates with the child to clarify their expectations and intentions for the session. Subsequently, the counsellor works towards maximising the child's potential benefits by enhancing problem-solving skills, thereby enabling effective addressing of challenges. Finally, the counsellor rounds the conversation by addressing any remaining concerns the child has.

B
Appendix Evaluation

B.1. Power Analysis

We made a power analysis in R script, B.1 to calculate the sample size of our experiment.

```

# The following code is based on a similar code  by Dr. Willem-Paul Brinkman.

# Load necessary libraries
set.seed(123)  # For reproducibility

# Constants
n_participants <- 11  # Number of participants
n_simulations <- 1000  # Number of simulations
correlation <- 0.71  # Correlation between baseline and post
R_squared <- correlation^2  # Calculate R-square
effect_size <- 0.8  # Effect size for Condition B

# Helper function to generate results
generate_condition <- function(n_participants, n_simulations, R_squared, effect_size = 0) {
  # Baseline
  baseline <- matrix(rnorm(n_participants * n_simulations, mean = 0, sd = 1),
                     nrow = n_simulations, ncol = n_participants)

  # Post
  post <- baseline * R_squared +
    matrix(rnorm(n_participants * n_simulations, mean = 0, sd = 1),
           nrow = n_simulations, ncol = n_participants) * (1 - R_squared)

  # Apply effect size adjustment for post if provided
  post <- post + effect_size

  # Delta (Post - Baseline)
  delta <- post - baseline

  return(list(baseline = baseline, post = post, delta = delta))
}

# Generate results for Condition A
condition_A <- generate_condition(n_participants, n_simulations, R_squared)

# Generate results for Condition B (with effect size adjustment)
condition_B <- generate_condition(n_participants, n_simulations, R_squared, effect_size)

# Example Outputs
cat("Condition␣A␣Delta:\n")
print(condition_A$delta[1:5, ])  # Show the first 5 deltas for Condition A

cat("\nCondition␣B␣Delta:\n")
print(condition_B$delta[1:5, ])  # Show the first 5 deltas for Condition B

mean_delta_A <- apply(condition_A$delta, 1, mean)
mean_delta_B <- apply(condition_B$delta, 1, mean)
```

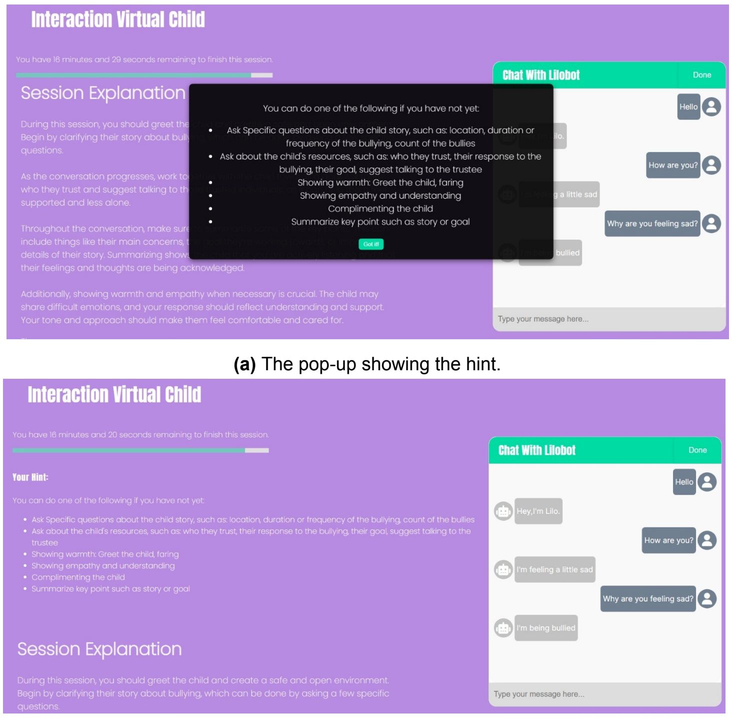
```

cat("Mean␣Delta␣for␣Condition␣A␣(first␣10␣simulations):\n", head(mean_delta_A, 10), "\n")
cat("Mean␣Delta␣for␣Condition␣B␣(first␣10␣simulations):\n", head(mean_delta_B, 10), "\n")

# Constants
alpha <- 0.05  # Significance level

# Perform t-tests for each simulation
p_values <- sapply(1:n_simulations, function(i) {
  t_test <- t.test(condition_A$delta[i, ], condition_B$delta[i, ])
  return(t_test$p.value)
})

# Count p-values less than alpha
significant_tests <- sum(p_values < alpha)

# Calculate power
power <- significant_tests / n_simulations

# Output results
cat("Number␣of␣significant␣tests␣(p␣<␣0.05):", significant_tests, "\n")
cat("Statistical␣Power:", power, "\n")
```

Listing B.1: Power Analysis R Script

B.2. Prototype

B.2.1. Hints

The hint was initially shown as a pop-up, see Figure B.1a. When the participant clicked 'Got it', the pop-up disappeared and the hint stayed at the left side of the screen, see Figure B.1b.



You have 16 minutes and 29 seconds remaining to finish this session.

Session Explanation

During this session, you should greet the

Begin by clarifying their story about bull

questions.

As the conversation progresses, work to

who they trust and suggest talking to th

supported and less alone.

Throughout the conversation, make sure

include things like their main concerns, t

details of their story. Summarizing shows

their feelings and thoughts are being acknowledged.

Additionally, showing warmth and empathy when necessary is crucial. The child may share difficult emotions, and your response should reflect understanding and support. Your tone and approach should make them feel comfortable and cared for.

You can do one of the following if you have not yet:

- Ask Specific questions about the child story, such as: location, duration or frequency of the bullying, count of the bullies
- Ask about the child's resources, such as: who they trust, their response to the bullying, their goal, suggest talking to the trustee
- Showing warmth: Greet the child, faring
- Showing empathy and understanding
- Complimenting the child
- Summarize key point such as story or goal

Hello

How are you?

Why are you feeling sad?

(a) The pop-up showing the hint.

You have 16 minutes and 20 seconds remaining to finish this session.

Your Hint:

You can do one of the following if you have not yet:

- Ask Specific questions about the child story, such as: location, duration or frequency of the bullying, count of the bullies
- Ask about the child's resources, such as: who they trust, their response to the bullying, their goal, suggest talking to the trustee
- Showing warmth: Greet the child, faring
- Showing empathy and understanding
- Complimenting the child
- Summarize key point such as story or goal

Session Explanation

During this session, you should greet the child and create a safe and open environment. Begin by clarifying their story about bullying, which can be done by asking a few specific questions.

Hello

Hey,I'm Lilo.

How are you?

I'm feeling a little sad

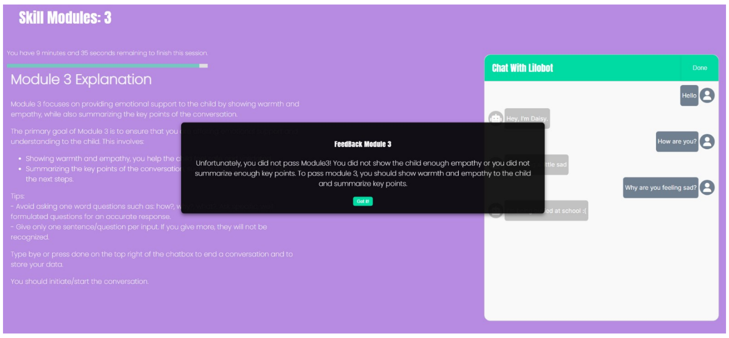
Why are you feeling sad?

I'm being bullied

(b) The hint on the left side of the screen.

B.2.2. Feedback

The hint was initially shown as a pop-up, see Figure B.2a. When the participant clicked 'Got it', the pop-up disappeared and the feedback stayed at the left side of the screen, see Figure B.2b.



Skill Modules: 3

You have 9 minutes and 35 seconds remaining to finish this session.

Module 3 Explanation

Module 3 focuses on providing emotional support to the child by showing warmth and empathy, while also summarizing the key points of the conversation.

The primary goal of Module 3 is to ensure that you are offering emotional support and understanding to the child. This involves:

- Showing warmth and empathy, you help the child feel safe and validated.
- Summarizing the key points of the conversation, such as their story or goal, clarifies the next steps.

Tips:
- Avoid asking one word questions such as: how?, why?, what?. Ask specific, well formulated questions for an accurate response.
- Give only one sentence/question per input. If you give more, they will not be recognized.

Type bye or press done on the top right of the chatbox to end a conversation and to store your data.

You should initiate/start the conversation.

FeedBack Module 3

Unfortunately, you did not pass Module3! You did not show the child enough empathy or you did not summarize enough key points. To pass module 3, you should show warmth and empathy to the child and summarize key points.

Chat With Lilobot

Hello

Hey, I'm Daisy.

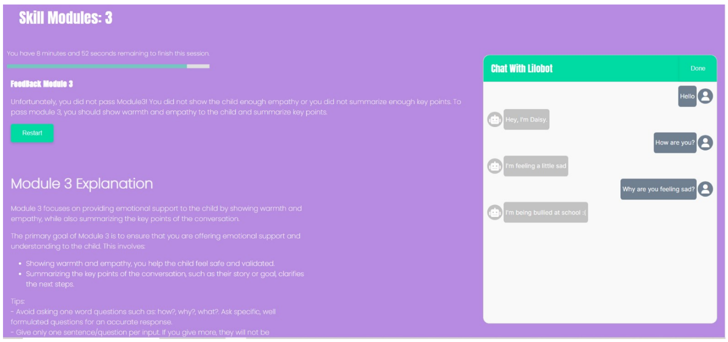
How are you?

I'm feeling a little sad

Why are you feeling sad?

I'm being bullied at school :(

(a) The pop-up showing the feedback.



Skill Modules: 3

You have 8 minutes and 52 seconds remaining to finish this session.

FeedBack Module 3

Unfortunately, you did not pass Module3! You did not show the child enough empathy or you did not summarize enough key points. To pass module 3, you should show warmth and empathy to the child and summarize key points.

Module 3 Explanation

Module 3 focuses on providing emotional support to the child by showing warmth and empathy, while also summarizing the key points of the conversation.

The primary goal of Module 3 is to ensure that you are offering emotional support and understanding to the child. This involves:

- Showing warmth and empathy, you help the child feel safe and validated.
- Summarizing the key points of the conversation, such as their story or goal, clarifies the next steps.

Tips:
- Avoid asking one word questions such as: how?, why?, what?. Ask specific, well formulated questions for an accurate response.
- Give only one sentence/question per input. If you give more, they will not be

Chat With Lilobot

Hello

Hey, I'm Daisy.

How are you?

I'm feeling a little sad

Why are you feeling sad?

I'm being bullied at school :(

(b) The feedback on the left side of the screen.

B.3. Informed Consent Form Simulation Training For Counselling

You are being invited to participate in an experiment titled integrating a pedagogical agent into a BDI-based training simulation. This study is being done by Maya Elasmar, and supervised by Willem-Paul Brinkman, and Mohammed Al Owayyed; All of which are affiliated with the TU Delft.

Research Purpose:
The purpose of this research study is to investigate the effectiveness of our system on the simulation training of volunteer counsellors in child helplines. In the simulation training, the trainees interact with a virtual chatbot that takes the role of a child with an issue. Hence, the trainees are dealing with a virtual child and not a real one. Specifically, we aim to explore how the incorporation of guidance can enhance the training process for counsellors, particularly focusing the required skills for a volunteer counsellor at a child helpline.

Experiment Procedure:
During the experiment, you will conduct one (or more) session with a virtual chatbot that acts as a child that is being bullied. You will be taking the role of a counsellor, using the required counselling skills to assist the virtual child. Throughout the session, you will receive guidance to help you apply these skills. The session lasts around 45 minutes.

Data Collection:
You will be asked to give your personal data (e.g. age group, degree and gender), which will be collected as categories for data analysis. We will ask you to fill out questionnaires before or after the session using the survey platform Qualtrics. They will be mainly about your opinion of the system, your knowledge, your self-efficacy and perceived usefulness. We will also ask you about your experience. Both the questionnaires and the data we extrapolate from them will be deleted from Qualtrics after collection and will be anonymised to be shared for scientific purposes.

Risks:
The virtual child is not based on a real story, but it is a realistic story. We advise people who are sensitive to the topics of bullying, violence, and emotional distress not to participate. If needed, you can contact a helpline in your country: https://findahelpline.com. In general, data can be leaked. We will minimise any risks by not storing any identifying information. The data will be stored privately and only accessible

by the researchers. The anonymised data will be stored after the research has been concluded, it will be published in a public repository (e.g., 4TU.ResearchData).

Compensation:
Your participation in this experiment is entirely voluntary, and you can withdraw at any time up until 3 days after the experiment. After that, the data cannot be removed. For more information, please contact:

If you agree and consent to this Opening Statement, you can now fill in the consent form below.

Consent
Please fill out the consent form below by answering either "Yes" or "No" for each question. Please note that a "No" for any single question renders you ineligible to participate in this study.

Statement	**Yes**	**No**
I have read and understood the experiment information above. I have been able to ask questions about the study and my questions have been answered to my satisfaction.	□	□
I consent voluntarily to be a participant in the experiment and understand that I can refuse to answer questions and withdraw from participation at any point during the experiment.	□	□
I understand that I cannot withdraw from participation once the experiment ends after three days.	□	□
I understand that taking part in the experiment involves interacting with a conversational agent that simulates a child.	□	□
I understand that taking part in the experiment involves filling in questionnaires about my knowledge of the topic, self-efficacy regarding counselling, and perceived usability of the system.	□	□
I understand that when I take part in the experiment, I will deal with a virtual agent that simulates a child that is being bullied.	□	□
I understand that the child is a virtual child who is suffering from bullying.	□	□
I understand that taking part in the experiment involves the risk of possible data leakage. The researcher does everything to mitigate this risk by storing the collected data safely, and publishing the anonymised data in a public repository.	□	□
I understand that the information retrieved during participating in this study will be used for research and can be published in a scientific paper.	□	□
I understand that personally identifiable information, such as (prolific ID, age group, gender, degree) will be collected.	□	□
I agree that my anonymised data will be accessible for all purposes, including for example educational, research, and commercial purposes.	□	□
I give permission for the data collected during the experiment that I provide to be archived in a public repository (e.g. 4TU Center for Research Data) so it can be used for future research and learning.	□	□
By ticking this box, I agree to participate in this study.	□	□

Table B.1: Consent Form Statements

Name: ____

Signature: ____

B.4. Demographics Questionnaire

How old are you?

o Under 18

o 18-24 years old

o 25-34 years old

o 35-44 years old

o 45-54 years old

o 55-64 years old

o 65+ years old

How do you describe yourself?

o Male

o Female

o Non-binary / third gender

o Prefer to self-describe ________________________

o Prefer not to say

What is the highest level of education you have completed?

o Some primary school

o Completed primary

o Some Secondary school

o Completed secondary school

o Vocational or Similar

o Some university but no degree

o University Bachelors Degree

o Graduate or professional degree (MA, MS, MBA, PhD, JD, MD, DDS etc.)

o Prefer not to say

B.5. Predefined Utterances

The predefined utterances were categorised for each module as Good, Neutral, or Bad, with each utterance assigned a value of -1, 0, 0.5, or 1. Table B.2 presents the good utterances for Module 1, while Tables B.3 and B.4 show the neutral and poor utterances, respectively. Similarly, Tables B.5, B.6, and B.7 present the good, neutral, and poor utterances for Module 2. Likewise, Tables B.8, B.9, and B.10 show the good, neutral, and poor utterances for Module 3.

B.6. Self-Efficacy Questionnaire

The self-efficacy questionnaire included items adapted from several existing questionnaires. Item 1, which assessed participants' confidence in their knowledge of counselling, was taken from the Counseling Self-Estimate Inventory (COSE) by Larson et al. (1992) [49]. Items 2, 3, and 5, which focused on encouraging the child to expand on their experiences and express emotions, were drawn from the SE-12 questionnaire by Axboe et al. (2016) [6]. Additionally, Item 5, which related to demonstrating empathy, was included in both the SE-12 questionnaire and Grundmann et al. (2025) [40]. Item 7, concerning collaborative goal setting with the child, was also sourced from Grundmann et al. (2025) [40].

To ensure alignment with the training content, we developed Items 4, 6, and 8, following a similar sentence structure to the existing items. These items addressed key skills taught in the modules, including summarising key points, understanding a child's reaction, and clarifying relationships.

Utterance	Score
trigger_unknown_what	0.5
request_bullying_who	1
request_unknown_who	0.5
request_bullying_details	1
request_unknown_details	0.5
request_bullying_count	1
request_bullying_location	1
request_unknown_location	0.5
request_bullying_duration	1
request_unknown_duration	0.5
request_bullying_frequency	1
request_unknown_frequency	0.5
request_unknown_when	0.5
request_bullying_why	1
request_unknown_why	0.5

Table B.2: Module 1 Good Utterances and their corresponding scores.

Utterance	Score
request_bullying_age	0
request_bullying_bullyage	0

Table B.3: Module 1 Neutral Utterances and their corresponding scores.

B.7. Instructions

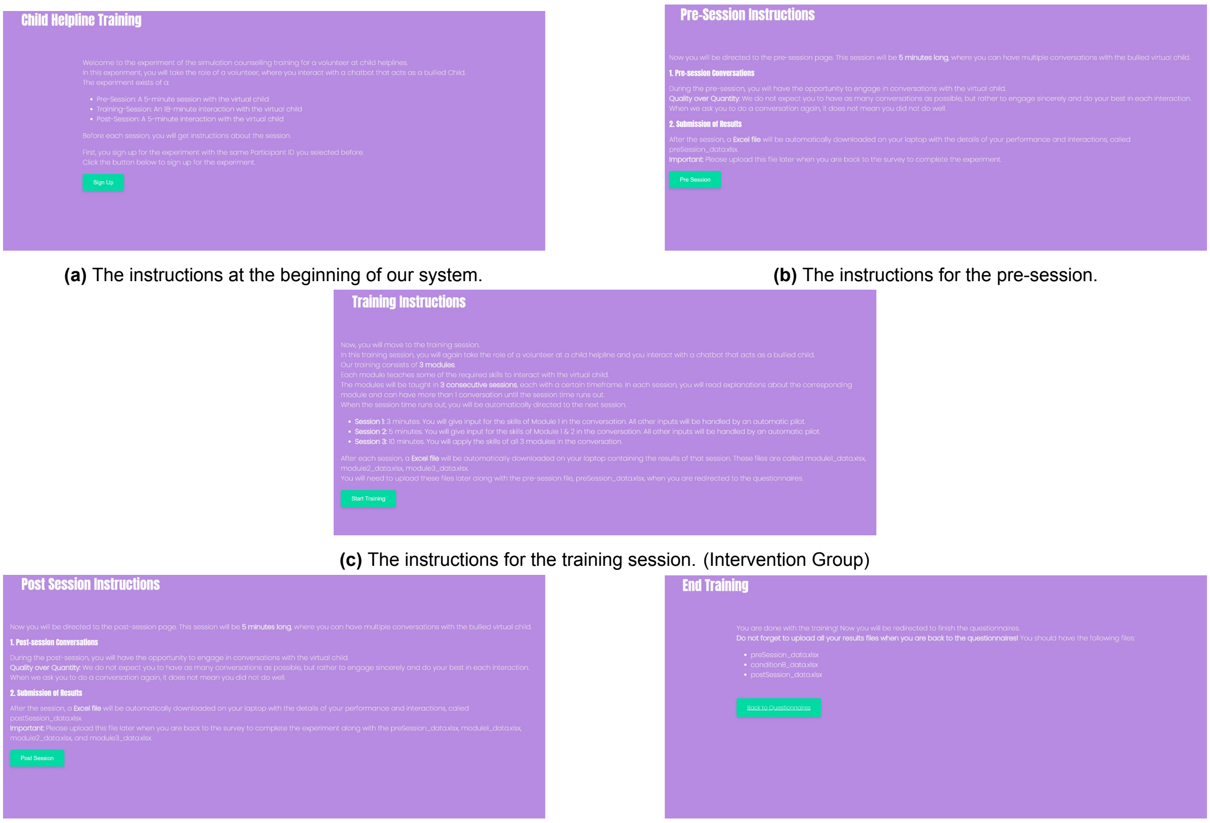
Figure B.3 shows the different instructions given to the participants of our experiment during their interaction with our system.

Utterance	Score
request_school_start	-1

Table B.4: Module 1 Poor Utterances and their corresponding scores.

Utterance	Score
request_bullying_response	1
request_unknown_response	0.5
request_bullying_confidant	1
request_unknown_confidant	0.5
request_bullying_parent	1
request_goal_what	1
request_goal_dream	1
confirm_goal_collaborate	1
request_goal_howchild	1
request_unknown_how	0.5
request_confidant_who	1
confirm_confidant_teacher	1
request_confidant_why	1
request_confidant_how	1
inform_confidant_help	1
inform_confidant_say	1

Table B.5: Module 2 Good Utterances and their corresponding scores.



Child Helpline Training

Welcome to the experiment of the simulation counselling training for a volunteer at child helplines.
In this experiment, you will take the role of a volunteer, where you interact with a chatbot that acts as a bullied Child.
The experiment exists of a:

- Pre-Session: A 5-minute session with the virtual child
- Training-Session: An 18-minute interaction with the virtual child
- Post-Session: A 5-minute interaction with the virtual child

Before each session, you will get instructions about the session.

First, you sign up for the experiment with the same Participant ID you selected before.
Click the button below to sign up for the experiment.

(a) The instructions at the beginning of our system.

Pre-Session Instructions

Now you will be directed to the pre-session page. This session will be **5 minutes long**, where you can have multiple conversations with the bullied virtual child.

1. Pre-session Conversations

During the pre-session, you will have the opportunity to engage in conversations with the virtual child.
Quality over Quantity: We do not expect you to have as many conversations as possible, but rather to engage sincerely and do your best in each interaction.
When we ask you to do a conversation again, it does not mean you did not do well.

2. Submission of Results

After the session, a **Excel file** will be automatically downloaded on your laptop with the details of your performance and interactions, called preSession_data.xlsx.
Important: Please upload this file later when you are back to the survey to complete the experiment.

(b) The instructions for the pre-session.

Training Instructions

Now, you will move to the training session.
In this training session, you will again take the role of a volunteer at a child helpline and you interact with a chatbot that acts as a bullied child.
Our training consists of **3 modules**.
Each module teaches some of the required skills to interact with the virtual child.
The modules will be taught in **3 consecutive sessions**, each with a certain timeframe. In each session, you will read explanations about the corresponding module and can have more than 1 conversation until the session time runs out.
When the session time runs out, you will be automatically directed to the next session.

- **Session 1:** 3 minutes. You will give input for the skills of Module 1 in the conversation. All other inputs will be handled by an automatic pilot.
- **Session 2:** 5 minutes. You will give input for the skills of Module 1 & 2 in the conversation. All other inputs will be handled by an automatic pilot.
- **Session 3:** 10 minutes. You will apply the skills of all 3 modules in the conversation.

After each session, a **Excel file** will be automatically downloaded on your laptop containing the results of that session. These files are called module1_data.xlsx, module2_data.xlsx, module3_data.xlsx.
You will need to upload these files later along with the pre-session file, preSession_data.xlsx, when you are redirected to the questionnaires.

(c) The instructions for the training session. (Intervention Group)

Post Session Instructions

Now you will be directed to the post-session page. This session will be **5 minutes long**, where you can have multiple conversations with the bullied virtual child.

1. Post-session Conversations

During the post-session, you will have the opportunity to engage in conversations with the virtual child.
Quality over Quantity: We do not expect you to have as many conversations as possible, but rather to engage sincerely and do your best in each interaction.
When we ask you to do a conversation again, it does not mean you did not do well.

2. Submission of Results

After the session, a **Excel file** will be automatically downloaded on your laptop with the details of your performance and interactions, called postSession_data.xlsx.
Important: Please upload this file later when you are back to the survey to complete the experiment along with the preSession_data.xlsx, module1_data.xlsx, module2_data.xlsx, and module3_data.xlsx.

(d) The instructions for the post-session. (Intervention Group)

End Training

You are done with the training! Now you will be redirected to finish the questionnaires.
Do not forget to upload all your results files when you are back to the questionnaires! You should have the following files:

- preSession_data.xlsx
- conditionB_data.xlsx
- postSession_data.xlsx

(e) The instructions for the rest of the experiment.(Control Group)

Figure B.3: The instructions given in our prototype to explain the participants of our experiment each step during the interaction with our system.

Utterance	Score
request_unknown_feeling	0
request_chitchat_greeting	0
request_chitchat_faring	0
inform_goal_help	0
request_goal_feeling	0
request_confidant_feeling	0
request_confidant_when	0
request_confidant_where	0
request_confidant_say	0
confirm_confidant_summary	0

Table B.6: Module 2 Neutral Utterances and their corresponding scores.

Utterance	Score
inform_goalhitstop_positive	-1
inform_unknown_positive	-1
inform_unknown_negative	-1
inform_goal_negative	-1
inform_goalhitstop_negative	-1
request_goal_howkt	-1

Table B.7: Module 2 Poor Utterances and their corresponding scores.

Utterance	Score
request_chitchat_greeting	1
request_chitchat_faring	1
confirm_bullying_summary	1
ack_unknown_empathize	1
ack_bullying_empathize	1
ack_unknown_compliment	0.5
ack_contactingkt_compliment	1
confirm_goal_summary	1
request_chitchat_goodbye	1

Table B.8: Module 3 Good Utterances and their corresponding scores.

Utterance	Score
ack_unknown_neutral	0
request_chitchat_end	0
confirm_chitchat_satisfaction	0
confirm_chitchat_questions	0

Table B.9: Module 3 Neutral Utterances and their corresponding scores.

Utterance	Score
ack_unknown_guilt	-1
ack_unknown_taunt	-1

Table B.10: Module 3 Poor Utterances and their corresponding scores.