Evaluation of feedback generated from agent-based social skills training systems

A qualitative analysis on the comprehensibility, usability, and improvement points of the generated feedback

Nikola Ntasi[1]

Supervisor(s): Willem-Paul Brinkman[1]**, Mohammed Al Owayyed**[1]

[1]EEMCS, Delft University of Technology, The Netherlands

A Thesis Submitted to EEMCS Faculty Delft University of Technology,
In Partial Fulfilment of the Requirements
For the Bachelor of Computer Science and Engineering
June 25, 2023

Name of the student: Nikola Ntasi
Final project course: CSE3000 Research Project
Thesis committee: Willem-Paul Brinkman, Mohammed Al Owayyed, Elmar Eisemann

Abstract

Chatbots are tools that can potentially be utilized in chat-based child helpline training. In this type of training, the quality of the feedback received is of vital importance. This paper aims to analyze the automated feedback generated by such a chatbot. The domains analyzed include user comprehension, usefulness, and potential improvement points. In a user study, a formative assessment and two interviews were conducted for each domain, respectively. For comprehensibility, all participants could easily understand the feedback report. They found that the bot could be easier to work around after reading the feedback, with the table being of much guidance. They found the transcript to be a welcome addition, but missing constructive feedback. Regarding improvement points, two of them were tightly related to the limitations of the chatbot, rather than the report itself. Extra guidance and instructions were deemed necessary by the participants, alongside an easier-to-read transcript interface.

1 Introduction

Chatbots are powerful tools that leverage machine learning to interact with users in a conversational manner [4]. They have gained popularity due to their ability to provide personalized assistance and enhance customer experiences. They are used in a wide range of domains, with child helplines potentially becoming one of them.

De Kindertelefoon [1] is the Dutch national child helpline, and children can opt to either chat via text with a trained counselor on their website or call the helpline. Based on their 2018 publications, they reported receiving more chat-based as opposed to call-based conversations [1] from children in need. Additionally, they report that chat-based conversations tend to last up to five times longer and children usually open up more about serious issues than they do on calls [18]. This evidence shows that the skills and training of helpline workers who engage in chat-based conversations are of paramount importance for successful child counseling.

To improve communication skills in this area, virtual child agents could be used to provide learners with scenarios to practice in. Training on an interactive virtual child in a conversational setting is particularly important, because this simulation provides a realistic and emotionally engaging scenario in which helpline workers can practice their skills safely. This means that they can train on a virtual child model with little to no risks compared to practicing on real individuals [12]. This type of role-playing, has demonstrated its effectiveness in assisting novice students who have limited access to clients in acquiring counseling skills, deepening cognitive comprehension, and enriching their emotional experiences [10][13].

The subject of this study is a training system developed by Sharon Grundmann [11], in which a virtual child agent (Lilobot) is used in a conversational setting. Lilobot simulates a child being bullied at school, and is intended to be used as a tool for learning how to structure counseling conversations in child helplines.

This chatbot is based on the BDI model [3]. This software model was developed for programming intelligent agents, where it simulates a real human's beliefs, desires, and intentions. By understanding the interplay between these three values, the BDI model can help psychologists and counselors gain insights into human decision-making, motivation, and behavior. Currently, after every interaction with Lilobot, a feedback report is generated at the end. This paper focuses on this feedback, which is of vital importance because it plays a critical role in skill acquisition and gaining insight about areas of improvement [14]. The aim here is to analyze to what extent this feedback is understandable by the participants but also analyze the usefulness and potential points of improvement of this feedback, with the hope of providing valuable insight for the potential next iteration. From the stated research direction, the following research questions are tied to the last three points, respectively:

1. To what extent is the feedback understandable?

2. How do participants feel better prepared from reading the feedback?

3. What other features or aspects would make this feedback even more useful?

The subsequent sections of the paper are structured as follows. Section 2 outlines the method employed in our user study. Section 3 presents the findings obtained from the study, alongside a discussion in Section 4. Section 5 elaborated on responsible research measures, and Section 6 concludes the paper by summarizing the findings and suggesting potential directions for future research.

2 Method

Here we will report relevant information for the reader to understand and potentially replicate this study, including detailed information on the sample, measures, and procedures used.

2.1 Participants

The recruitment of participants happened during a three-week course between May 2023 to June 2023 through peer referral sampling (i.e. snowball sampling), on-site at the Technical University of Delft. Since volunteering at De Kindertelefoon is accessible to the general public [2], we were not aiming for a specific target population, hence we considered this sampling method to be suitable for our research. It is worth acknowledging that such a method departs from probability-based sampling approaches [16], but for the scope of this experiment that shouldn't be a problem. The participants were only required to have a sufficient understanding of Dutch, as Lilobot's user interface was in Dutch. In our study, participants were invited to experiment as devised in the experiment procedure, followed by a set of interview questions.

A total of ten individuals participated in the study, with nine being male and one female. All participants fell within the age range of 18-24. Out of the ten participants, all except one had prior experience with chatbots. Three participants reported using chatbots frequently (more than 10 times

[1] https://www.kindertelefoon.nl/

a month), four participants mentioned occasional usage (2-10 times a month), and two participants reported rare usage (once a month or less).

2.2 Materials

Here we will lay out and explain all the materials that the participant will use. The first two train the participant on conversational skills, the third one is the bot with which the participants talk to, and finally, the last is the feedback report. These materials will be presented sequentially to the participant and all form a cohesive structure.

The Five Phase Model

Communication skills are essential for a counselor so that they can help the child, regardless of their knowledge about the issue that child might be facing. Hence, De Kindertelefoon developed The Five Phase Model with guidelines as to how you can structure such a conversation [18]. This model aims to help center the conversation around the child, so that the counselor can guide him/her to find a viable solution. This phase overview alongside conversation techniques associated with each phase will be handed out in printed form.

Pre-training transcript

Snippets of example conversations taken from Sindahl [17] will be used. The aim here is to show an example of how The Five-Phase Model works. We avoided including a transcript of a conversation with Lilobot, as one of the limitations of the chatbot is that it's very deterministic. This means that the responses that the participant might remember from the transcript will lead to the same outcome (Lilobot reaching out to the teacher regarding the bullying). This would completely skew our results and compromise the validity of the experiment. We also made sure that the theme of the conversation snippets was modified to avoid any triggers for the participants. This conversation transcript can be found in Appendix C.

Lilobot

Lilobot is a chatbot based on the BDI model. This is a software model developed for programming intelligent agents. The BDI model suggests that people's actions are driven by a combination of their beliefs, desires, and intentions. The purpose of this model lies in utilizing these concepts to address a specific problem in agent programming, in our case bullying. Beliefs refer to the thoughts and opinions that this agent holds about itself and the world around it. Desires represent the wants, needs, or goals it has, meanwhile, intentions are the conscious plans or decisions it makes to act in a certain way. Intentions are based on a combination of beliefs and desires. Here we want to change the agent's beliefs through certain prompts, which will lead to it forming new desires and intentions, respectively. With Lilobot we are giving the participant a model simulating a child who is bullied with a certain BDI model configuration, aiming to guide it to a certain intentional outcome. The participants will however not be told about this model before their interaction with Lilobot. This is because it would skew the experiment results of the other two researchers.

Feedback Report

This will be a Word document, consisting of navigation instructions, a table with the Lilobot's beliefs according to the BDI model, and a transcript of the conversation. The first page consisting of the instructions and the table can be found in Appendix B. It shows an overview of the chatbot's beliefs and which phase (from the Five-Phase Model) they're tied to. In addition, the intensity of each belief at the beginning and the end of the conversation for each belief is presented. Apart from the table, we also have the transcript. Below in Figure C, we can see a snippet from an example conversation with Lilobot. In this example, the Kindertelefoon (KT) asked how long Lilobot has been bullied, and we can see that the bot's belief value of "I think that the helpline worker is interested in my story" increases. We can also observe that Lilobot intends to talk more about its feelings, which follows considering the response it gives. By understanding the interplay between beliefs and intentions, the feedback helps the participant to gain insight into the child's decision-making and motivation.

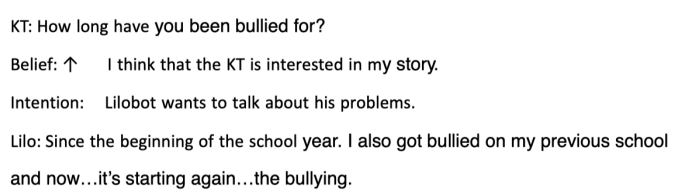

Figure 1: A snippet from the transcript found in the feedback report with Lilobot, translated from Dutch.

2.3 Measures

Formative Assessment

For usability testing, we will use an observational methodology called the "Thinking Aloud" method [9]. This method is used to understand participants' behaviors, thoughts, and motivations by having them narrate their thought process. The goal here is to understand how the users interact and navigate the feedback report, as well as analyze how easy or hard it is for them to digest the information presented.

To avoid priming [15] at this stage, we will avoid giving the user any specific tasks besides looking at the report itself. Their only task is to simply verbalize their thoughts as they move through the feedback report.

Alongside this process, we want to know to what extent the feedback is understandable. Hence, formative assessment [5] will be conducted, which involves asking the participant to provide detailed explanations of the feedback to the researchers. This will help evaluate their comprehension and understanding of the given information. We will ask the participant about different elements of the report, such as the table, the percentages shown, and why certain beliefs change in regard to certain prompts. All the primary qualitative data recorded at this stage will be transcribed be in the form of a descriptive text.

Interview

Here we aim to gather primary qualitative data that will later be used for analysis. The first interview question will tackle

the usefulness aspect of the feedback report. We will give them a minute of time to reflect on how they felt better prepared from reading the feedback report, noting down important points they make. In case of ambiguity, we will further ask them about what they exactly meant. The same line of logic goes for the second interview question. For future iterations we wanted to gather data about the possible new features which would make the feedback report even more useful. Akin to the previous task, the user's responses will be transcribed in the form of descriptive text. The two interview questions can be found in Appendix A.

2.4 Study Procedure

This user study involved a collaborative effort with two other researchers who were also investigating various aspects of Lilobot. Specifically, they focused on examining the noticeability of behavioral and belief changes, alongside the believability of the bot. Together with this paper they contribute to a more holistic evaluation of this model. Careful consideration was given to the sequencing of the questionnaires and interviews for the participants. Lilobot was set up on only one laptop, due to the challenging environment configurations for the NLP model, the Java Spring Framework alongside the required processing power. As such, for all participants there was one moment of measurement with the same setup. Figure 2 gives an overview of the experiment order.

Consent Form and Demographics Survey

The experiment commenced after the participant signed the ethical consent form and afterward completing a demographics survey consisting of the age range, gender, and prior experience with chatbots.

Training

After collecting this data, the participant is presented with The Five-Phase model and shown a conversation snippet to familiarize him/herself with how the model works. In practice, De Kindertelefoon does not permit the counselor to intervene in the child's situation, for example by calling the school or their parents. Rather, the main goal is to foster an environment where the child can get emotional support so that they can devise strategies that can be followed to solve the problem at hand. This will be made clear to participants.

Interaction with Lilobot

Subsequently, the participants get familiarized with Lilobot. We first introduced them to some of the limitations, such as the absence of an emotional model, the inability to recognize multiple intents in the user prompts, and the restriction of using only one sentence as a response, among others [11]. The way that the conversation can go is twofold. Lilobot can either leave mid-conversation in case it loses trust in the participant (failed attempt), or it can decide to talk to the teacher about the bullying (successful attempt). In both cases, a feedback report was generated in the end. That is the focus area of this paper, evaluating how well the participants understand that feedback through usability testing, examining whether they find it useful in any way, and giving ideas for future improvement.

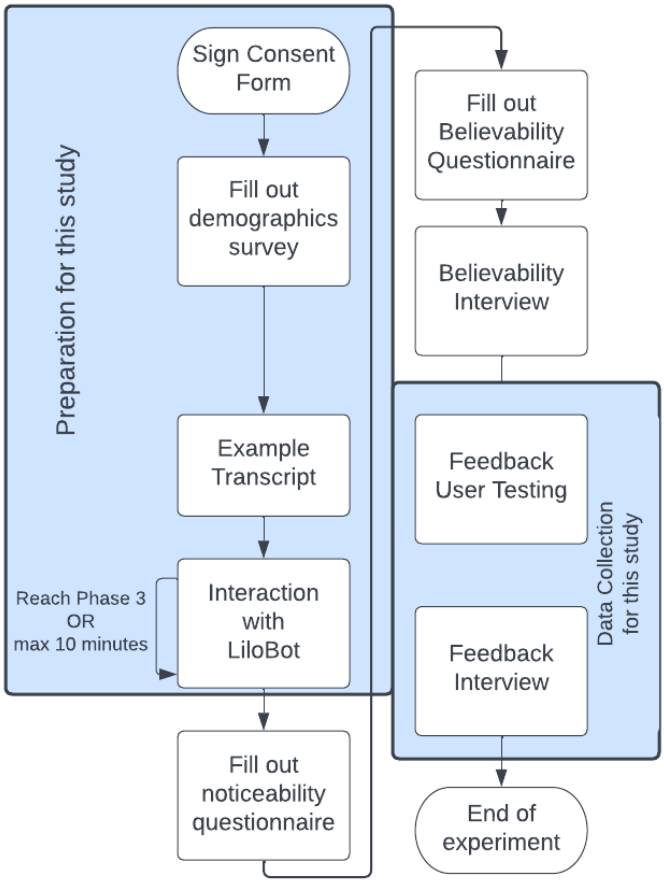

Figure 2: The sequence of the study organization.

During this stage we set up two constraints. Firstly, the participant needed to at least reach the third phase during their interaction. This is because jumping to the fourth phase was harder compared to other phases, since the participant lacked proper experience. Secondly, the other two researchers deemed 10 minutes to be enough time for the participant to get a sense of the bot's believability, noticeability, and a proper feedback report in the end.

Feedback Report

After the interaction with the bot, the two researchers carried out their questionnaires and interviews with the participant, leaving the feedback evaluation study to be done last. Here the participant will be presented with the auto-generated feedback, on which we will measure the participant's input. Following this, we reach the end of the experiment.

2.5 Data Analysis

Observational Analysis for Comprehensibility

In formative assessment, researchers often assess the success and efficiency of users in completing specific tasks. In our case, there aren't really any specific tasks besides interpreting the table and the belief changes in the transcript correctly, so we will only try to assess to which extent the participants understand these elements.

Hence, here we will perform observational analysis [7]. This involves analyzing qualitative data gathered through direct observation of users interacting with the feedback report. We will closely analyze user behavior, interactions, and difficulties encountered during the testing session. We will extract and identify patterns, recurring themes, or issues that arise. This analysis helps uncover usability problems, user preferences, and areas for improvement.

Thematic Analysis for Usability and Points of Improvement

After gathering all primary qualitative data for usability and points of improvement, we will perform thematic analysis [8] on this data. This method is used to identify, analyze, and interpret patterns or themes within a dataset. It involves systematically organizing and categorizing the data to identify recurring topics, ideas, or concepts that are relevant to the research objectives. Thematic analysis aims to uncover the underlying meanings and patterns within the data, providing insights into the experiences, perspectives, or phenomena being studied.

Additionally, we will perform simultaneous coding [6], which offers advantages such as enhanced inter-coder reliability, reduced bias, comprehensive analysis, and opportunities for reflexivity and critical reflection.

3 Results

The thematic analysis regarding usefulness and points of improvement produced seven themes in total. It's important to consider that the some of meaning units from the two interview questions were mixed between the two. This happened because participants mentioned points regarding improvements during the usefulness interview, and vice versa. By mixing them, we got a more holistic overview of these two domains.

Figure 3 depicts a table with the theme names accompanied by their respective categories. For each category, the number of codes related to it is shown alongside the total number of quotes per theme.

Regarding simultaneous coding, Cohen's kappa [19] is a statistical measure used to assess the level of agreement between two raters or evaluators when dealing with categorical data. It takes into account both the observed agreement and the agreement that would be expected by chance. In our usefulness and points of improvement domains, Cohen's kappa statistics were 0.52 and 0.58, respectively. These were both moderate agreements.

3.1 Comprehensibility

It took on average five to six minutes for the participants to go through the feedback report. Surprisingly enough, all but one participant skipped the initial explanatory text of the paper. Despite this, all of them could comprehend the information in the feedback easily, with little to no further clarifications.

Table

When the participants were first shown the report, the element which caught their attention first was the table. They directly navigated the beliefs and the respective percentages that they were before and after the interaction. However, half of them reported not knowing whether a certain percentage entailed a good or bad value. There was no performance metric they could refer to, so they based their interpretation of these percentages on intuitive reasoning.

Theme	Category (with respective # of quotes)	Total # quotes
Table Guides on Phases	table useful (n = 7)	7
Tinkering with the Model	learning workaround (n = 6)	6
Mixed feelings about the transcript	transcript constructive (n = 1), belief change useful (n = 2), transcript constructive (n = 5)	8
A more realistic Five Phase Model	move between phases (n = 2), merging phases(n = 2), new phase approach (n = 3), making progress (n = 2)	9
Getting Suggestions	human guidance (n = 1), performance metric (n = 2), understanding belief change (n = 5), understanding intention change (n = 3)	11
Phrasing guidance	phrasing guidance (n = 5), bot doesn't understand (n = 2)	7
Better User Interface	transcript more readable (n = 5), graph over percentages (n = 1)	6

Figure 3: Theme overview. For each category, the number of codes related to it is shown alongside the total number of quotes per theme.

All of the participants made sense as to why the phases were tied to certain beliefs. They often glanced at the Five-Phase Model printout to see if the beliefs linked with their respective phase objective were logically coherent. For example, they examined why the belief "I think the helpline worker is interested in my story" was connected to the second phase, where the main objective there is to get a clear view of the child's story, perspective, personality, network, and competencies.

Transcript

Moving to the transcript, the first comment made by all participants was that there was a lot of text. This made the readability a bit slow, but the content presented was straightforward. The element that captivated their attention the most was the belief change. This is because it was presented to them earlier in the table, and they were familiar with it. In the majority of cases, they could derive meaning from the prompts they made and how the beliefs were affected. But there were cases where it was more challenging for them to interpret. For example, when some participants said "We're gonna find a solution together" the belief " I think the helpline worker can help me" increased, but the belief "I think the helpline worker can solve my problem" decreased. This scenario is elaborated on in the points of improvement domain.

Regarding the intention of the bot, most participants did not make any remarks about it. All of them quietly acknowledged it and moved on to the next prompt. When asked why, they responded that there was no direct link between the beliefs and the intention, even though the intention was understandable. They didn't see how the intention would be of any use to their interaction because it was lacking a structure that they could learn from, unlike the beliefs.

3.2 Usefulness

Tinkering with the model

Half of the users reported having learned a great deal from the feedback report, mainly in terms of the model that the bot was implemented. More precisely, they felt like they were better prepared to tackle the next conversation as they had more insight into how the model worked.

"It's useful to see what it thinks and what exactly I said to make it trust me more so I know how to work around it in the future as well."

However, they also found that this would make the bot deterministic, meaning that devising a sequence of hard-coded prompts would always yield a positive outcome. They said that in this scenario the feedback helped them learn how to work around the bot and pass to the next phase quickly, which doesn't give them a sense of having learned to tackle a delicate matter with a real child.

"I found the feedback useful, because I can learn how to manipulate it. It's not a particularly smart chatbot."

Table guides on phases

The majority of participants found the table to be the most useful element in the report. They felt that seeing beliefs tied to certain phases guided them towards which prompts to make when interacting with Lilobot.

"The first time around I was really thinking about what to say and now I know which beliefs I need to tackle first, so I can do that."

One participant reported that seeing the belief percentages also helped to tackle areas of improvement.

"So the scores with percentages are quite useful cause it tells you which areas you're lacking and which areas you can improve upon."

Mixed feelings about the transcript

The feedback regarding the transcript had mixed opinions. Most participants appreciated this element, as they could see a step-by-step walkthrough of how each prompt affected the model.

"I guess I know if an approach worked 'cause I see a bunch of upward arrows, so at least my approach wasn't terrible."

However, they didn't regard the transcript to be constructive. It only showed changes in the beliefs or intentions without explaining why it happened, meaning that in certain cases participants were left wondering what they did wrong.

"I didn't get a sense of where I could improve and what I could say next time."

3.3 Extra Features

Better transcript readability

A remark made by almost half of the participants was that the transcript could have been easier to read. The distinction between the messages from the child and the helpline worker alongside the belief change is not very prominent to the user and requires mental effort to distinguish.

"If stuff was highlighted in different colors it would have been easier to read. For example if value is 'I don't trust KT' and it's low then turn it red." Some participants even suggested turning the transcript into a table-like format.

"I think it would have been nicer as a table, with the first column being the helpline worker or child, second the message and third the belief. It would make it way more readable."

A more realistic Five Phase Model

In their accounts, more than half of the participants found that the discrete separation and flow of phases in the feedback report wouldn't resemble a conversation with a real child.

"The phases don't have to follow a sequential order. Thinking you can help it is only applicable to phase 3, but why not 1 or 2? This feels a bit unnatural, phases can be a bit more intertwined."

In their view, a real-life conversation will hardly ever follow such a structured approach, and having the bot jump from one phase to another would prepare them better.

"This model is focused on completing every phase, meanwhile the child might just want the bullying to stop and he can jump to any phase."

Better comprehension

Half of the users ran into phrasing problems, and for half of them it made the experience with the bot problematic. In most cases they needed more than one session with the bot because the bot wouldn't understand the message.

"Often I have the right intention of the message but then I didn't phrase it the right way so it leaves the conversation or it doesn't respond." These users suggested having a reference guide as to how you can formulate the sentences would greatly contribute to the usability of the bot.

"Well, it would help to have a dictionary and guidelines on how to use this system."

Getting suggestions

Lastly, the vast majority of participants stated that more guidance is needed with the feedback report. While it was clear how the elements were connected and how the model worked, there was a lot of room for interpretation.

"Would be useful having a model that revolves around problem solving. Like how well did you solve the problem or something." In many instances the participants found the bot to be incoherent with its behavior. For example, they couldn't understand why a certain belief was altered when the prompt they made was used for a different purpose.

"Why does the belief 'I think the helpline worker can solve my problem' go down when I say that we can find a solution together?" "I don't understand why at first when I ask if there is someone that you trust she leaves the conversation."

In our bullying scenario, the bot approaches the counselor thinking that they are going to fix the bullying situation and that it can trust them. And when the participant says otherwise it loses trust and may leave the conversation. Because of this, guidance as to why this happened was deemed useful if they were to be included in the report.

4 Discussion

In this section we reflect on the aforementioned themes and we attempt to give a possible explanation of the results.

Understanding the report

All the participants interpreted the report the right way. They could explain what every element meant and how they were interconnected to each other. This however doesn't go without reason. The experiment took place on-site at TU Delft,

through snowball sampling. Hence, we expected most participants to be students at this university. Having a technical background means that they took a very methodological and structured approach to the report. It's important to consider that comprehensibility results might not be the same with participants from other domains.

Double Coding

The Cohen kappa coefficient ranges from -1 to 1, with different interpretations for different values. A kappa value of 1 indicates perfect agreement beyond chance, while a value of 0 indicates agreement that is equivalent to chance alone. Negative values represent an agreement that is worse than chance. In our analysis, the kappa statistics were 0.52 and 0.58, for the first and second interview questions, respectively. A statistic of 0.67 or greater is considered a satisfactory to solid agreement [6]. From this we can infer that our coding could have been more reliable, resulting in an enhanced credibility of the analysis.

Limitations of Lilobot

A significant number of participants felt that they could "workaround" with Lilobot after reviewing the feedback. The rationale behind this will be elaborated as follows. Lilobot is an initial iteration chatbot, and as such it has significant limitations. One of them is that it doesn't have an emotional model, so whenever participants show sympathy it doesn't recognize it. The Natural Language Processing (NLP) implementation is also very limited. This means that there were very specific sentence structures that participants had to adhere to for the NLP model to parse their input. In addition, the bot's responses were hard-coded and sparse, and they were classified into response domains (like the length of getting bullied or how it felt because of it).

Learning workaround

We propose two sequentially linked reasons as to why participants thought they could "work around" the bot to jump to the next phase quicker. It's important to consider that the participants only tackled the bullying domain, but also the limitations of the virtual child.

First, interacting with the bot conditioned them to use particular words and sentences while avoiding others. The phrasing problem means that the participant had to conduct more than one session with Lilobot, and by doing so they recognized the same responses from it. This became repetitive and it gave them a sense of control and predictability. Subsequently, after looking at the table (deemed very helpful by most participants) they saw which beliefs were tied to which phase. We already know that the conversation with Lilobot only supports a sequential flow of phases. Hence, knowing how to formulate proper prompts and having a sense of the responses they might receive, they formulated their input in such a way as to tackle the belief tied to the current phase to jump to the next one.

Following this line of reasoning, this is why we suspect that a considerable number of participants perceived the feedback as a "workaround" rather than constructive feedback.

The need for extra guidance

Half of the participants suggested adding extra instructions and guidance to the feedback, especially regarding the transcript. In contrast to the brief introduction to only the Five Phase Model that they received, volunteers of the helpline usually complete a 30-hour face-to-face group training including a wide range of domains. Hence, their training was very limited and they lacked the experience and knowledge that a volunteer would have gained regarding different child counseling topics.

Regarding other points of improvement

In terms of the more realistic Five Phase Model and the phrasing problem, it is important to consider that this all depends on the implementation of the virtual child. The feedback report only gives feedback regarding the current version of Lilobot. If this virtual child gets updated in a future iteration, then the feedback report will reflect that.

5 Responsible Research

Careful consideration was put into how we could conduct this experiment ethically. Below are some of the main points which contribute to the responsibility of this research.

The experiment commenced after obtaining ethical approval from TU Delft, with approval number 2960. When signing the consent form, it was made clear to the participants that they could withdraw from the study at any time. During the data collection process for demographics, we used Qualtrics [2] which is GDPR (General Data Protection Regulation) [3] compliant and provides technology that enables the strongest privacy and security law worldwide. As for the transcription of participant responses during the interview, we implemented measures for data anonymization. These measures were implemented to ensure that participants could not be identified through personal data in the event of a data leak. In addition to this, we also randomized the participant numbers.

The code necessary to implement the chatbot is openly accessible in the TU Delft repository. This implies that once the environments are set up and the installations are completed, anyone can readily execute the chatbot. Upon concluding the conversation, a feedback report will be generated, which can be utilized to reproduce this study.

The dataset for this study is available in the 4TU database [4].

6 Conclusions and Future Work

The feedback was understandable to all the participants, with some of them skipping the initial instructions on how to navigate it. However, this should be approached with a degree of skepticism given the expected technical background.

In terms of usefulness, most participants found the table to be the primary element that would guide their future interactions. Seeing which beliefs were tied to which phase gave

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

[3] https://gdpr-info.eu/

[4] https://doi.org/10.4121/85110f4b-40e1-4567-9f6e-e97c6337ad92

them a sense of direction. However, half of the users perceived the table as being a workaround for the virtual child. This is because of the repetitive responses they encountered and other limitations of Lilobot, as discussed earlier.

In terms of proposed features and improvements, more than half the participants stated that improved comprehension from the bot and a more realistic Five Phase Model would resemble a real-life scenario with the child. This is however dependent on the implementation of the virtual agent. Regarding extra guidance and suggestions, we deemed it to be due to the lack of proper training from the participants. One last noteworthy point of improvement for the future iteration is a better user interface for transcripts, using tables and colors.

In summary, the feedback report is constructive for most, but the limited implementation of Lilobot made some participants regard it as a workaround. One intriguing field to explore further involves enhancing Lilobot's capabilities by potentially incorporating an emotional model and possibly a more advanced NLP model. This would naturally raise the question: *'How would the feedback report be perceived any differently in case the of a perfect virtual child model?'*

References

[1] De kindertelefoon | Jaarverslag 2019 , May 2023. [Online; accessed 30. May 2023].

[2] Vrijwilliger Informatie, June 2023. [Online; accessed 4. Jun. 2023].

[3] Carole Adam and Benoit Gaudou. *BDI agents in social simulations: a survey*. PhD thesis, 2017.

[4] Eleni Adamopoulou and Lefteris Moussiades. Chatbots: History, technology, and applications. *Machine Learning with Applications*, 2:100006, December 2020.

[5] Randy Elliot Bennett. Formative assessment: A critical review. *Assessment in education: principles, policy & practice*, 18(1):5–25, 2011.

[6] Laila Burla, Birte Knierim, Jurgen Barth, Katharina Liewald, Margreet Duetz, and Thomas Abel. From text to codings: intercoder reliability assessment in qualitative content analysis. *Nursing research*, 57(2):113–117, 2008.

[7] Malgorzata Ciesielska, Katarzyna W Boström, and Magnus Öhlander. Observation methods. *Qualitative methodologies in organization studies: Volume II: Methods and possibilities*, pages 33–52, 2018.

[8] Victoria Clarke, Virginia Braun, and Nikki Hayfield. Thematic analysis. *Qualitative psychology: A practical guide to research methods*, 3:222–248, 2015.

[9] David W Eccles and Güler Arsal. The think aloud method: what is it and how do i use it? *Qualitative Research in Sport, Exercise and Health*, 9(4):514–531, 2017.

[10] Thomas C Froehle, Sharon E Robinson, and WAYNE J DE KURPIUS. Enhancing the effects of modeling through role-play practice. *Counselor Education and Supervision*, 22(3):197–206, 1983.

[11] Sharon Grundmann. A bdi-based virtual agent for training child helpline counsellors. 2022.

[12] Andrzej A Kononowicz, Luke A Woodham, Samuel Edelbring, Natalia Stathakarou, David Davies, Nakul Saxena, Lorainne Tudor Car, Jan Carlstedt-Duke, Josip Car, and Nabil Zary. Virtual patient simulations in health professions education: systematic review and meta-analysis by the digital health education collaboration. *Journal of medical Internet research*, 21(7):e14676, 2019.

[13] Lisa M Larson and Jeffrey A Daniels. Review of the counseling self-efficacy literature. *The Counseling Psychologist*, 26(2):179–218, 1998.

[14] Henry W Maier. Role playing: Structures and educational objectives. *The International Child and Youth Care Network*, 36, 2002.

[15] Daniel C Molden. Understanding priming effects in social psychology: An overview and integration. *Social Cognition*, 32(Supplement):243–249, 2014.

[16] Charlie Parker, Sam Scott, and Alistair Geddes. Snowball sampling. *SAGE research methods foundations*, 2019.

[17] T.N. Sindahl and Børns Vilkår. *Chat Counselling for Children and Youth: A Handbook*. Børns Vilkår, 2011.

[18] Trine Natasja Sindahl. *Chat Counselling for Children and Youth: A Handbook*. Børns Vilkår, 2011.

[19] Shuyan Sun. Meta-analysis of cohen's kappa. *Health Services and Outcomes Research Methodology*, 11:145–163, 2011.

A Interview Questions

1. How do you feel better prepared from reading this feedback?
2. What other feature or aspect do you think would make this feedback even more useful?

B Feedback Report

FEEDBACK GESPREK

Hier is een transcriptie van je gesprek met Lilobot met zijn gedachten tijdens het gesprek. Lilobot heeft een reeks overtuigingen en intenties die tijdens het gesprek constant worden bijgewerkt op basis van wat je tegen hem zegt. In de onderstaande tabel kun je zien wat Lilobot's overtuigingen waren aan het begin van het gesprek en aan het einde. Het transcript van het gesprek laat zien welke overtuigingen veranderen op basis van jouw berichten. Het symbool ↑ betekent dat de overtuiging toeneemt, terwijl ↓ betekent dat de overtuiging afneemt. Het transcript laat ook zien welke intenties Lilobot had op het moment in het gesprek. Al deze notaties zijn cursief weergegeven tussen jullie gesprek.

Overtuiging	Vijffasemodel	Begin	Eind	Verschil
Ik voel me in controle in het gesprek	alle fasen	50%	50%	0%
Ik denk dat KT mij begrijpt	alle fasen	10%	10%	0%
Ik voel me verbonden met KT	alle fasen	40%	65%	25%
Ik voel me veilig in het gesprek	alle fasen	100%	100%	0%
Ik denk dat KT geïnteresseerd is in mijn verhaal	fase 2	10%	100%	90%
Ik denk dat KT te vertrouwen is	fase 2	60%	70%	10%
Ik denk dat ik mijn verhaal heb verteld	fase 2	0%	100%	100%
Ik denk dat KT vraagt naar een wens	fase 3	0%	100%	100%
Ik denk dat KT vraagt naar een positieve wens	fase 3	0%	0%	0%
Ik denk dat KT het probleem kan oplossen	fase 3	100%	0%	-100%
Ik denk dat KT mij kan helpen	fase 3	70%	80%	10%
Ik denk dat ik competent ben om het probleem op te lossen	fase 4	30%	40%	10%
Ik denk dat KT het probleem voor mij gaat oplossen	fase 4	0%	0%	0%
Ik denk dat Juf Ellie mij kan helpen	fase 4	50%	100%	50%
Ik denk dat KT en ik samen tot een oplossing zullen komen	fase 4	0%	100%	100%
Ik denk dat KT vraagt naar een vertrouwenspersoon	fase 4	0%	100%	100%
Ik denk dat KT wil het gesprek beëindigen	fase 5	0%	100%	100%

TRANSCRIPT

KT: Hi

Overtuiging: ↑ Ik denk dat KT te vertrouwen is

Intentie: Lilobot wil het over zijn probleem hebben

Lilo: Hoi! Ik ben Lilobot

Overtuiging: ↑ Ik denk dat ik mijn verhaal heb verteld

Lilo: Ik word gepest op school

KT: Wie pest jou?

Overtuiging: ↑ Ik denk dat KT geïnteresseerd is in mijn verhaal

C Example Conversation Transcript

THE FIVE PHASE MODEL

1. BUILDING RAPPORT

OBJECTIVE: CREATE A WELCOMING ATMOSPHERE AND BUILD TRUST
METHOD: EMPATHY, RESPECT, SINCERE INTEREST, ACTIVE LEARNING

i. *Hallo Ik ben [naam]. Ik ben hier om te luisteren en te helpen!*
ii. *Wat is er aan de hand?*
iii. *Wil je dat ik help*

e.g.

COUNSELLOR: Hi. Welcome to the chat
CHILD: Hi there
COUNSELLOR: Before we start please tell me how old you are and if you are a boy or a girl?
CHILD: Girl 13 years old
COUNSELLOR: Thanks - then I can better adapt to what you tell. What would you like to talk about?

2. CLARIFY THE CHILD'S STORY

OBJECTIVE: GET A CLEAR VIEW OF THE CHILD'S STORY, PERSPECTIVE, PERSONALITY AND COMPETENCIES.
METHOD: ASK DETAILED QUESTIONS ABOUT THE CHILD'S STORY, ITS SUBTLETIES, ITS DEPTH AND CONRETE MANIFESTATIONS

i. *Hoe voel je je daarbij?*
ii. *Waarom kan je niet concentreren?*
iii. *Dus je weet niet hoe je beter kan worden in wiskunde?*

e.g.

COUNSELLOR: okay. So you have now told me that you have a problem with biting nails. And that you have moved to a children's home about 2 months ago, because you have ocd. And your father and sister also have ocd. And that you don't go to school at the moment.

3. SETTING GOAL FOR THE SESSION

OBJECTIVE: THAT BOTH PARTIES ARE AWARE OF WHAT THE CHILD MAY USE THE CONVERSATION FOR.

METHOD: CLARIFICATION

i. *Zoek je iemand om mee te praten?*

ii. *Waar wil je over praten?*

4. WORKING TOWARDS THE SESSION GOAL

OBJECTIVE: TO ENSURE, THAT THE CHILD MAY BENEFIT FROM THE CONVERSATION

METHOD: STIMULATING THE CHILD'S OWN PROBLEM-SOLVING SKILLS

i. *Wil je dat we samen een strategie opzoeken?*

ii. *Heb je al met de pesters gesproken?*

iii. *Hoe zou je dit kunnen oplossen?*

e.g.

COUNSELLOR: is there anything you have considered doing which might help?

CHILD: no not really

COUNSELLOR: ok. Then let us look at it together. If I asked you to find a solution, what would be the first thing you think about?

CHILD: Spik to the staff again - maybe

COUNSELLOR: Yes. I think this sound as a good idea. Is there one of them you trust?

CHILD: yes I think so. Thanks bye bye

5. ROUNDING OFF THE CONVERSATION

OBJECTIVE: THAT THE CHILD IS LEFT WITH AS FEW QUESTIONS AS POSSIBLE

METHOD: SUMMING UP AND CLARIFYING

i. *Bedankt dat je je verhaal met mij hebt gedeeld!*

ii. *Ik hoop dat we samen een oplossing hebben kunnen vinden.*

iii. *Onthoud dat er mensen zijn die om je geven en je willen helpen.*

e.g.

CHILD: yes thanks

COUNSELLOR: You are welcome. It was nice talking to you. It is great that you do something about it, and you are always welcome to write to us again - also if you need to find other solutions

CHILD: thanks bye bye

COUNSELLOR: bye bye