The Development of an AI-Assistant to Therapists in a Chat-based Psychological Intervention: Gathering Users' First Impressions of the Experience*

Neha Deshpande[1,*], Mariam Fishere[2], Stefan Hillmann[1], Jorge P. Marqués[3], Catarina B. Ferreira[4], Sofia Silva[4], Ricardo Barroso[4] and Klaus M. Beier[2]

[1]Technische Universität Berlin, Straße des 17. Juni 135, Berlin, 10623, Berlin, Germany
[2]Charité Universitätsmedizin Berlin, Luisenstr. 57, Berlin, 10117, Germany
[3]International University of Catalonia, Josep Trueta, Sant Cugat del Vallès, 08195, Barcelona, Spain
[4]University of Trás-os-Montes and Alto Douro, Quinta dos Prados, Vila Real, 5000-801, Portugal

Abstract

The rapid growth of Artificial Intelligence (AI) presents significant opportunities to enhance mental healthcare, particularly in addressing the sensitive and complex issue of the consumption of Child Sexual Abuse Materials (CSAM). This study explores the design, development, and evaluation of an AI assistant aimed at supporting therapists during live, chat-based interventions for individuals who consume CSAM. The AI assistant provides real-time assistance by summarizing previous chats, offering message suggestions, and providing a semantic search tool. Using a participatory design approach, therapists tested the first prototype of the AI assistant in simulated therapy sessions, providing detailed feedback on its usability, effectiveness, and impact on the therapeutic process. The findings revealed that therapists generally had a positive experience using the AI assistant. Key factors contributing to this positive reception included the user-centered design approach, which ensured that the assistant was tailored to meet the therapists' needs. These outcomes suggest that AI-based support systems could play a valuable role in augmenting therapy for individuals with problematic sexual behaviors. As part of our ongoing research, further testing in real-world settings will be the next step to fully assess its potential.

Keywords

Human Computer Interaction (HCI), Usability, Online therapy intervention, AI-based assistant, User Experience Questionnaire (UEQ)

1. Introduction

Mental health disorders are a major global health issue, with over 150 million people in Europe affected. However, access to adequate care remains limited, with one in seven individuals receiving no support [1]. The COVID-19 pandemic has worsened this crisis, highlighting the need for scalable and cost-effective interventions [2]. Artificial intelligence (AI) has gained attention as a tool to enhance mental healthcare, with applications spanning diagnosis, treatment planning, early intervention, and support for both patients and clinicians [3, 4].

While AI has been extensively explored in mainstream mental health applications aimed at patients (e.g., diagnostic tools, chatbots), its use in specialized areas, such as therapeutic interventions for individuals consuming Child Sexual Abuse Material (CSAM), remains limited [5, 6]. CSAM consumption is a growing public health issue, with approximately 12.2% of offenders having a history of contact sexual offenses, requiring tailored interventions [7, 8]. However, specialized resources for such interventions are scarce due to a shortage of trained professionals, stigma, and the complexity of treatment strategies. AI-driven therapeutic solutions have the potential to address these challenges by offering scalable, evidence-based support. However, integrating AI into clinical practice raises concerns regarding efficacy,

Joint Proceedings of the ACM IUI Workshops 2025, March 24-27, 2025, Cagliari, Italy
*Corresponding author.
n.deshpande@tu-berlin.de (N. Deshpande); mariam.fishere@charite.de (M. Fishere); stefan.hillmann@tu-berlin.de (S. Hillmann); jpiqueras@uic.es (J. P. Marqués); acferreira@utad.pt (C. B. Ferreira); asofiasilva@utad.pt (S. Silva); rbarroso@utad.pt (R. Barroso); klaus.beier@charite.de (K. M. Beier)

ethics, and its impact on the therapeutic relationship [9]. AI should augment, not replace, human therapists, helping to alleviate resource constraints while preserving the essential human connection required for effective therapy.

1.1. Human-machine interaction: the case of AI-based assistants

The design and implementation of AI-driven mental health interventions require a deep understanding of human-computer interaction (HCI) principles, especially when supporting mental health professionals like therapists and peer counselors. AI systems are being used to scale training and offer automated feedback, enhancing clinical skills. For example, AI-powered tools can simulate patient interactions and provide real-time feedback, refining therapeutic techniques such as Motivational Interviewing (MI) before counselors work with real clients [10, 11].

Prior research on AI-assisted therapy has shown positive results in improving therapist efficiency, patient engagement, and recovery rates. [12] found that therapists could not distinguish AI-generated transcripts from human conversations and rated AI-assisted dialogues higher. While AI can support administrative tasks and enhance communication, integrating AI into therapeutic contexts requires balancing algorithmic decision-making with human expertise. Platforms like HAILEY demonstrate how AI can increase empathy in peer-support settings, with a 19.6% improvement in conversational empathy [13]. However, adoption is influenced by trustworthiness and perceived reliability [14], and user engagement improves when AI is socially competent, with natural language capabilities and empathetic responses [15, 16]. Moreover, human-in-the-loop mechanisms are essential for ethical oversight, particularly in sensitive areas like therapeutic interventions [17]. AI's integration into mental healthcare can offer scalable, effective interventions and continuous support for clinicians' development, improving both therapist skills and therapeutic systems.

1.2. Acceptance, usability, and experience of AI

The success of AI in therapeutic contexts depends heavily on its usability, acceptance, and integration into existing clinical workflows. AI-driven tools can assist mental health professionals by automating routine tasks like data recording, case classification, session summarization, and intervention recommendations [18]. These functionalities reduce cognitive load, allowing clinicians to focus on direct patient care. However, AI adoption is influenced not only by technological efficacy but also by social, psychological, and contextual factors [19, 20, 21].

Prior research indicates that users engage more effectively with AI when it is perceived as a socially competent actor. Therefore, incorporating anthropomorphic design elements (e.g., natural language capabilities, empathetic responses) can significantly enhance user acceptance [15, 16]. Furthermore, user-centered design principles emphasize that AI systems should be tailored to users' cognitive abilities, ethical concerns, and work practices to ensure seamless integration into therapeutic settings [22, 23]. Given the ethical complexities of AI-driven therapy for CSAM consumers, it is essential that AI systems are developed with strong privacy safeguards, ethical oversight, and human-in-the-loop mechanisms to prevent misuse and ensure alignment with therapeutic standards [17].

This paper includes six sections that are structured as follows: in the first section, we highlight AI's advances in the field of mental health with a focus on the concepts of human-machine interaction, acceptance, and usability of AI. In the second section, we present the current study. The methodology is explained in the third section, where we introduce a prototype of an AI-based assistant for the therapists. In the fourth section, we provide our evaluation of the users' acceptance and usability of the prototype. Then, in the fifth section, we discuss our findings. Finally, in the sixth section, we offer the conclusions of the study, the implications, as well as the limitations and future research perspectives.

2. The present study

As part of a larger research initiative, an online psychological intervention is being administered via chat by several mental health professionals, whom we refer to as ”therapists” in this study. The intervention is designed for individuals who consume Child Sexual Abuse Material (CSAM) (excluding offenders), whom we will refer to as ”clients” throughout this paper. The chat intervention is anonymous and spans a duration of four weeks, during which therapists conduct 50-minute sessions every week and monitor the clients' CSAM use. To primarily reduce therapists' cognitive load and to investigate additional ways AI can support therapeutic processes, we developed a prototype of an AI-based assistant. In the present study, we evaluate therapists' overall experience with this AI assistant. Specifically, we explore how therapists perceive and rate the assistant's usability and identify ways in which they believe AI could enhance the therapeutic process as part of the intervention. The prototype was tested in a simulated lab setting as a precursor to real-world implementation. The code used for the implementation is available at: https://git.tu-berlin.de/neha.deshpande/flask_ai_app

3. Methodology

Designing for AI comes with its challenges, [24] identified two sources of AI's distinctive design challenges: (1) uncertainty surrounding AI's capabilities, and (2) AI's output complexity, spanning from simple to adaptive complex. To address these challenges, some techniques that could facilitate a successful human-AI interaction have been mapped out. Accordingly, we adopted a typical User-Centered AI Design process [24] that involved the following steps: (1) Understanding the User, (2) Problem Definition and Ideation, (3) Prototyping the AI assistant followed by, (4) Testing and Evaluation as shown in figure 1. These steps are discussed in the next subsections in more detail.

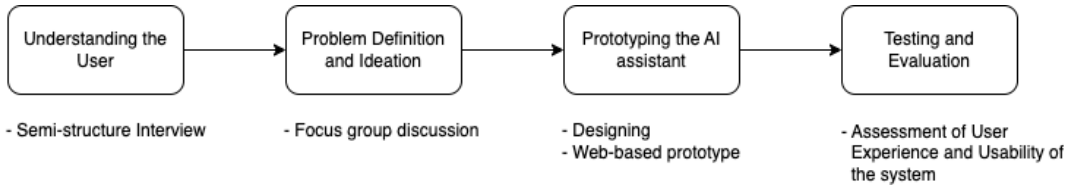

Figure 1: Design thinking process utilized for this study

3.1. Understanding the User

To understand therapists' needs and the domain of therapy, we conducted a 1-hour semi-structured interview with a therapist involved in online chat-based psychological interventions (see Section 2). The interview explored challenges faced during therapy sessions, the use of AI tools, and therapists' habits. Key points raised by the therapist, which informed our design process, included:

1. Difficulty remembering client details
2. Ensuring client engagement
3. Challenges in assessing client mood, especially on sensitive topics
4. Struggles with paraphrasing sentences to enhance empathy

3.2. Problem Definition and Ideation

To further define the problem, we conducted a virtual focus group [25] with six therapists involved in the project to explore their views on AI in therapy for CSAM users. Information about the therapists involved is mentioned in sub-section 4.1.1. The focus group discussions were informed by insights from the initial interviews. The key ideas generated from the focus group discussion are as follows:

1. AI can suggest 3-5 relevant responses based on past chat data
2. It adapts to the client's writing style, including vocabulary and emoticons
3. It tracks client mood during sessions
4. It provides a summary of pre-session questionnaires
5. It assists by completing messages of therapists
6. Therapists should maintain control over the AI assistant
7. The AI assistant should collaborate with the therapist, not the reverse

In response to the raised ideas, three main features were narrowed down to be incorporated into our prototype for further testing. The selected features were as follows, (1) response suggestions, (2) semantic search feature, and (3) chat summary, which were combined into an "AI assistant" that would work together with the therapist to collaboratively support them during live therapy sessions.

3.3. Prototyping of the AI assistant

The prototyping of the AI assistant included two stages namely (1) Design and (2) Development. As described by [26] and [24], it is important to design keeping in mind the existing work practices of the target users and letting the technology adapt to user needs. Following this logic, we created several mock-ups to fulfill the three main features planned to be developed. To provide more control and freedom of choice to the therapists, the AI assistant is displayed only when the therapist presses the "Ask AI" button which is placed within the chat interface as shown in Figure 4. Firstly, to implement these initial ideas, we used Streamlit [1], an open-source Python library to create basic mock-ups. These mock-ups helped shape both the user interface and the interaction flows for each feature in the prototype. Following their creation, active discussions with therapists refined these flows, which were then finalized for development into a functional prototype. An important step in this process involved data pre-processing as explained below.

3.3.1. Data Source and Pre-processing

Data used for the mock-ups mentioned in the previous sub-section came from the publicly available counsel-chat dataset [2], due to the unavailability of domain-specific data. This data was sufficient for testing feature design. Whereas, chat data for developing AI features was collected from simulated therapist training sessions with the same group of therapists who tested the prototype (see Section 4). In these sessions, one therapist acted as a client and another as a therapist. The data underwent initial cleaning to remove unnecessary messages, after which it was structured in a database with labels ("client" and "therapist") and metadata such as date, time, and chat room number (Figure 3). Although simulated, the dialogues were based on real-life use cases designed to train therapists in this specialized domain of therapy.

3.3.2. Interface and AI Implementation

The implementation of the AI assistant involved choosing the right models for the features mentioned above and integrating them smoothly within the chat interface. This chat interface was developed with Flask [3], a web-based Python framework. The interaction design for each feature was designed separately after considering the discussions from the focus group discussion and the user interview. The design and development for each feature is described below.

[1] https://streamlit.io/
[2] https://huggingface.co/datasets/nbertagnolli/counsel-chat
[3] https://flask-socketio.readthedocs.io/en/latest/deployment.html

Response suggestions

As established during the focus group discussion, the purpose of response suggestions was to suggest the next best responses to the therapist during an on-going therapy session. Suggestions were presented in descending order of their relevance given the last few messages as context. The following interaction flow was finalized for this feature which was visualized by creating mock-ups with Streamlit as shown in Figure 2.

- When there is a new message from the client, three possible responses should be suggested
- Therapists should be able to click on a suggestion so that it appears in the input box
- The responses suggested by the AI assistant should be able to be modified by the therapist before sending them through to the client to give control and responsibility to the therapists.

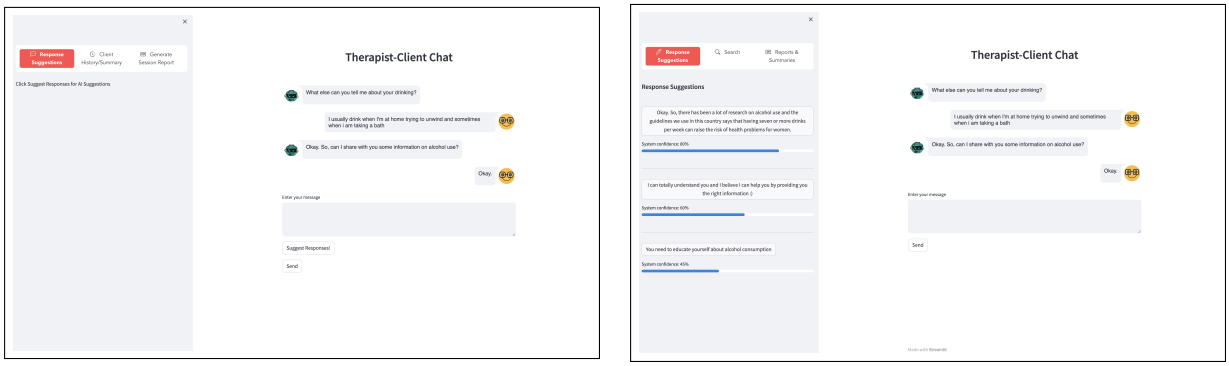

Figure 2: Streamlit mock-ups displaying the 'Response Suggestions' feature with the "Suggest Responses" button and relevance scores displayed next to the chat interface.

For implementing the functional prototype, we chose a model to match and retrieve therapist messages based on previous conversation context, using the available data as described in Sub-section 3.3.1. This decision was made due to the limited availability of domain-specific training data, which made developing a custom model challenging. Additionally, using LLMs via an API was not feasible due to content moderation restrictions, as these models cannot process sensitive topics common in CSAM-related therapy. This approach ensured both the efficiency and safety of the system in real-time interactions. To facilitate smooth operation, we implemented an efficient system for storing and retrieving message embeddings. An embedding is a dense numerical vector that represents the semantic meaning of a piece of text in a high-dimensional space and are computed by machine learning models like the transformers model [27]. The embeddings of all messages exchanged between therapists and clients were saved into a FAISS (Facebook AI Similarity Search)[4] database. FAISS is a highly efficient and scalable library designed for fast similarity-based search in large datasets of dense vector embeddings. Each message from the dialogue data was stored in this database, along with its sentence embedding and metadata, including the role (therapist or client), chat room identifier, date, and timestamp (Figure 3). To compute these embeddings, a sentence transformer model named "SBERT" was utilized. SBERT, short for Sentence-BERT, is a modification of the BERT (Bidirectional Encoder Representations from Transformers) [28] model designed specifically for encoding sentence-level semantics. SBERT incorporates siamese and triplet network architectures, enhancing BERT's capabilities to generate fixed-dimensional embeddings for sentences. This modification enables SBERT to capture semantic similarities between sentences, making it particularly effective for tasks like semantic textual similarity, sentence retrieval, etc. All sentence embeddings were pre-computed for each message using the "all-MiniLM-L6-v2" model [5] which was utilized via the Haystack framework [6], which provides seamless integration for retrieval and embedding tasks.

[4]https://faiss.ai/index.html
[5]https://huggingface.co/nreimers/MiniLM-L6-H384-uncased
[6]https://docs.haystack.deepset.ai/docs/intro

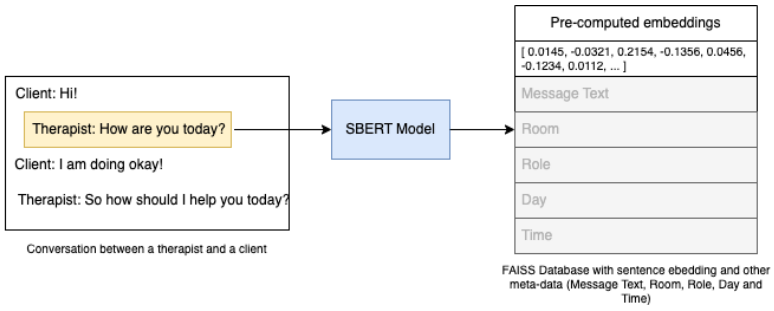

Figure 3: Process of pre-computing the embeddings and saving into a database for easier processing in real-time

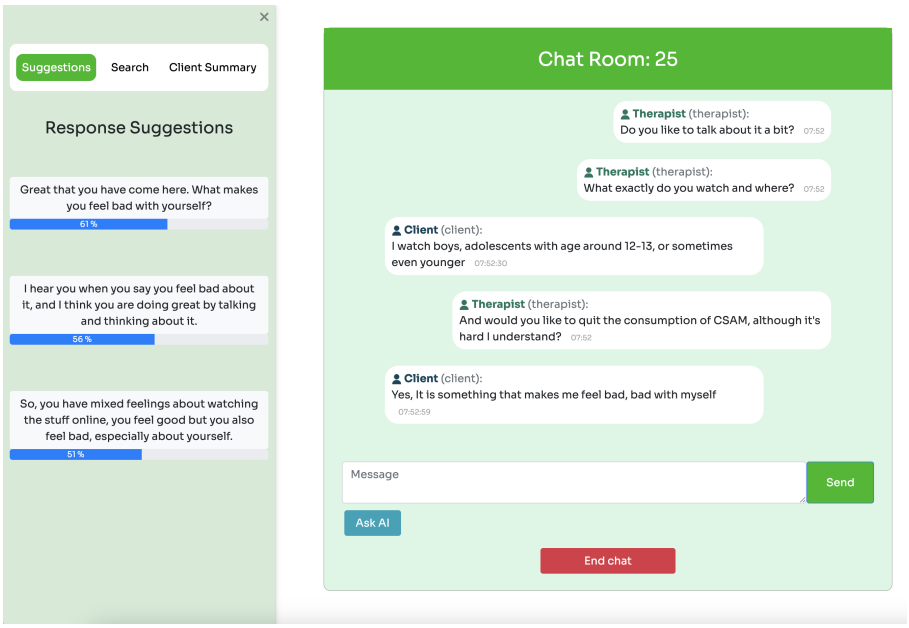

Figure 4: The 'Response Suggestions' feature integrated within the chat interface

When a new client message is received, the system compares its sentence embedding with pre-computed therapist message embeddings using cosine similarity [29]. Cosine similarity measures how similar two sentences are by comparing their embeddings, where a smaller angle indicates higher similarity [30]. The system ranks therapist messages based on similarity and presents the top 3 with their similarity scores in the interface, enabling quick access to the most relevant responses. This workflow ensures efficient, real-time interactions, as shown in Figure 4.

Semantic search

To seamlessly integrate the semantic search feature into the chat interface, the chosen interaction flow enabled the therapist to input a search term, which checked through all the prior messages within the chat conversation. Subsequently, the AI assistant retrieved and displayed the top 10 most semantically similar messages. The therapists had the flexibility to click on any of these ten search results, instantly prompting the chat interface to auto-scroll to the corresponding segment of the conversation displaying the selected message. Two variations of mock-ups were created as shown in Figure 5, with differences in the placement of the search bar.

Out of these two, the first mock-up (left in figure 5) was chosen to be developed into the Flask web app. As this feature also involves semantic matching similar to the response suggestions feature, we used the sentence transformers model "all-MiniLM-L6-v2" which worked similar to the previous feature described in sub-section 3.3.2 by computing an embedding for the search term/terms entered by the

therapist and cosine similarity measure to search through the previous messages in that particular chat room again using the pre-computed embeddings as explained in sub-section 3.3.2. The semantic search feature next to the chat interface is shown in Figure 6.

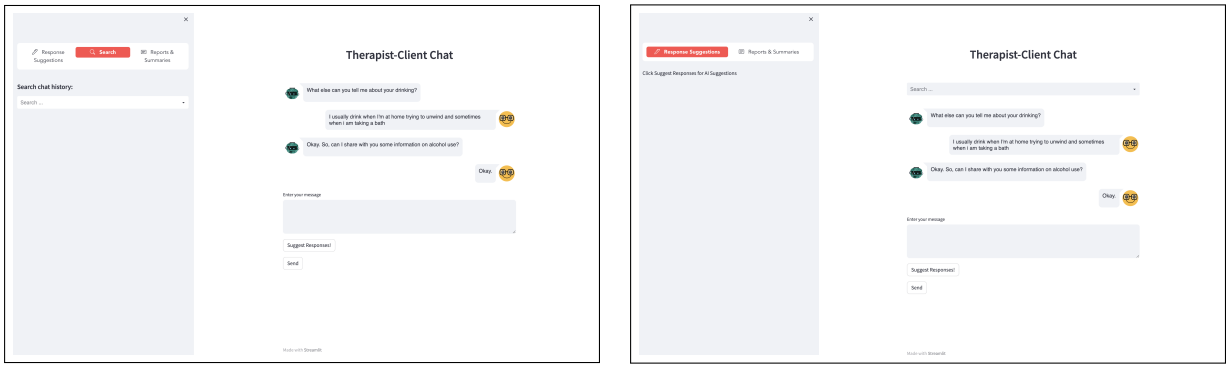

Figure 5: Streamlit mock-ups displaying the search bar next to the chat interface (left) and integrated into the chat interface (right)

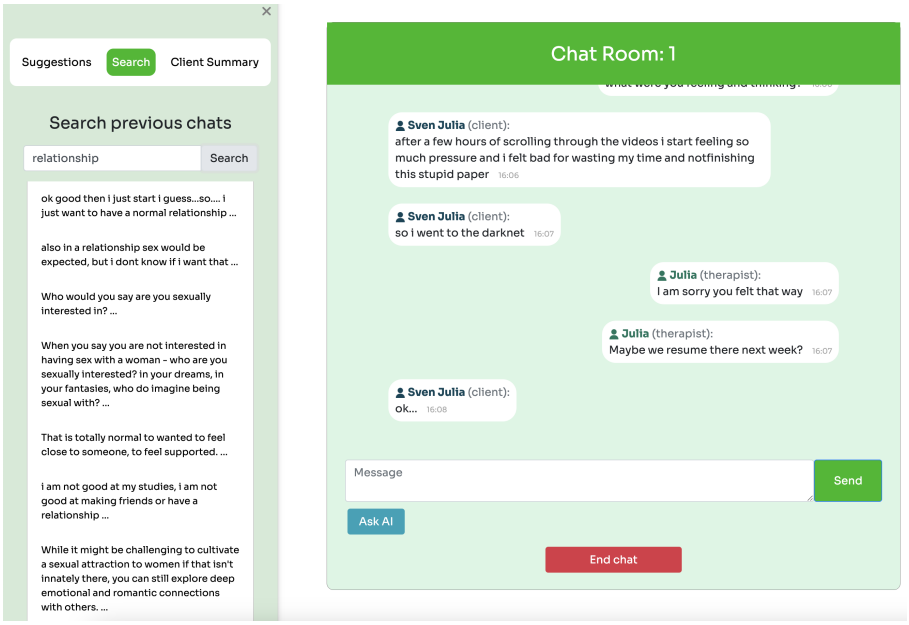

Figure 6: Semantic search feature

Chat Summary

This feature was implemented to help therapists learn about or recall their client's history. Therapists can generate a summary of the chat conversation at any point during the session, which can be modified to add any missing information deemed important for future use. To achieve these goals, two Streamlit mock-ups were created (Figure 7): one with a "Generate chat summary" button (on the left) for real-time summarization and recomputation, and another allowing therapists to copy the summary to their personal notes (on the right). Based on these mock-ups and therapist feedback, the final interaction flow was developed, including the "Generate chat summary" button and a copy icon for saving the summary to personal notes. These features were integrated into the web application, as shown in Figure 8.

Since conventional summarization models struggled with capturing key dialogue information, two large language models, BART [31] and FLAN-T5-base [32], fine-tuned on the SAMSum dataset [33], were

used. The models, bart-large-cnn-samsum [7] and flan-t5-base-samsum [8], were accessed via Huggingface [9]. Due to token limits, input dialogues were split into segments for summarization and combined into a single summary.

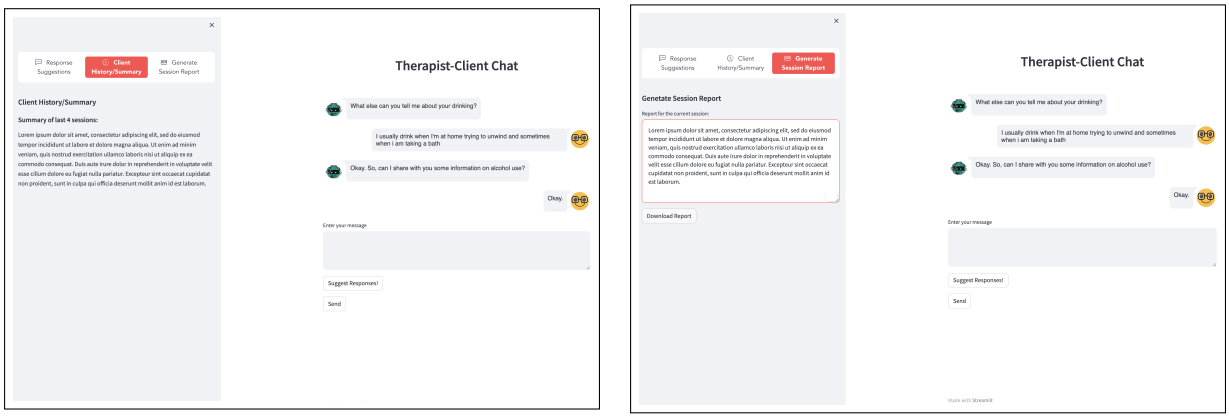

Figure 7: Mock-ups displaying a client summary: (left) a static summary of the client and (right) a summary inside a textbox for modification by therapists.

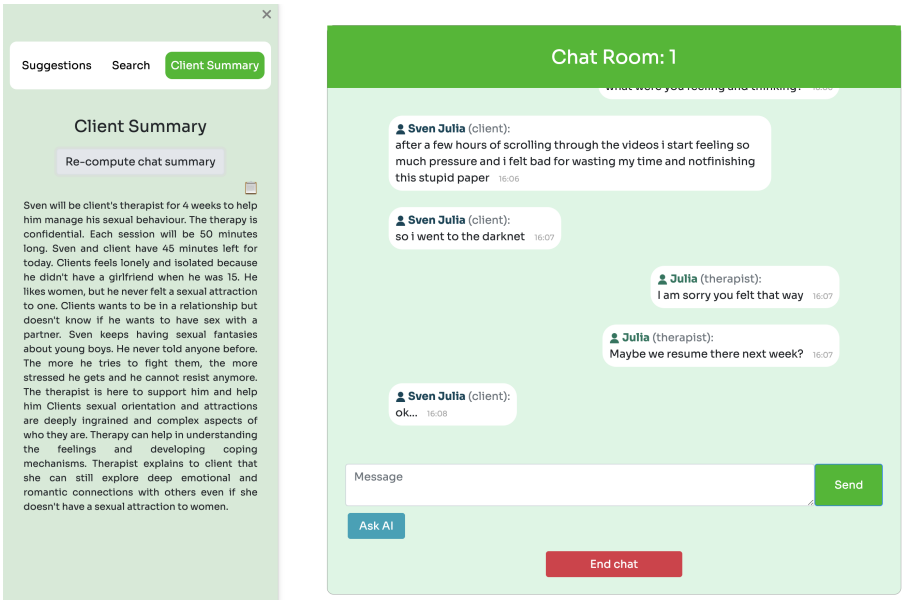

Figure 8: Chat summarization feature

3.3.3. Testing and Evaluation

After a functional prototype was ready and deployed, a procedure for testing it was designed. The next section discusses the information on participants, the exact procedure, and the measures used to test the prototype.

[7] https://huggingface.co/philschmid/bart-large-cnn-samsum
[8] https://huggingface.co/philschmid/flan-t5-base-samsum
[9] https://huggingface.co

4. Acceptance- and Usability-Evaluation

4.1. Methods

4.1.1. Participants

Six therapists (4 female, 2 male) were recruited for testing, with diverse language proficiencies (German, English, Czech, Spanish, and Portuguese). Participants varied in therapeutic experience, with only one having prior exposure to chat-based therapy and AI tools, while others had limited AI use for tasks like paraphrasing and research.

4.1.2. Measures

The user experience of conducting a chat session with a client, including access to the AI-based assistant, was evaluated using the User Experience Questionnaire (UEQ; Laugwitz et al., 2008) and two components from the UEQ+ (Schreep & Thomaschewski, 2019), focusing on the chat and response suggestion features (Section 4.2). The usability of the semantic search and summarization features was assessed with the System Usability Scale (SUS) and summary quality was evaluated through a brief questionnaire.

The UEQ comprises six scales (attractiveness, efficiency, perspicuity, dependability, stimulation, and novelty), each with four items rated on a 7-point Likert scale. For this study, four scales from the UEQ (attractiveness, efficiency, perspicuity, dependability) and two from the UEQ+ (usefulness, clarity) were selected. The SUS, a 10-item scale, measured the effectiveness, efficiency, and user satisfaction of the AI assistant's search and summarization features, with responses rated on a 7-point scale to align with the UEQ+.

Chat summarization was assessed through a brief questionnaire evaluating grammaticality, non-redundancy, referential clarity, focus, structure and additionally, two more questions were included to assess the content correctness of the summaries: content coverage and informativeness [34]

4.1.3. Statistical Methods

The statistical analysis corresponded to the description of the central tendency and dispersion measures for each item in the questionnaires for the group of respondents. Both the UEQ+ and the SUS contained items rated on a 7-point Likert scale (semantic differentials in the case of the UEQ+ items). The responses of the items were normalized to 0 -items ranging from -3 to 3- matching the usual reporting style of the UEQ. Firstly, the means and standard deviation (SD) of the individual item responses were calculated for the 6 therapists. Then, for each scale (composed of 4 items) the means and SD were obtained considering the responses of all items inside the scale. Secondly, the means and SD were calculated for each item on the SUS, and tfor each of the studied features: semantic search and summarization. Finally, the summarization evaluation questionnaire contained 7 questions rated on a 5-point Likert scale. The median and the Interquartile range were reported for each question. The results were stratified by two factors: the AI model used and summary size (See Sec. 4.2 for more details). All the summaries were evaluated by two separate therapists, but the results were calculated for all the responses independent of the evaluators. To test the Inter-rater reliability of the items, Krippendorff's alpha for ordinal data was calculated using implementation by [35]. Moreover, an automatic summarization evaluation metric was included: the Recall-Oriented Understudy for Gisting Evaluation (ROUGE), introduced by [36]. The ROUGE metric compares a generated summary to a reference text (in our case, the complete dialog) and computes the overlap of words (1-grams) between them. In practice, this metric can be interpreted as the recall score between the summary and a reference text.

4.2. Procedure

The testing procedure, conducted online by the first author, comprised three steps, each involving two participants (therapist and client). Figure 9 illustrates the session structure.

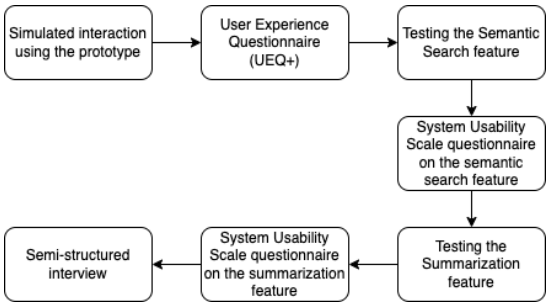

Figure 9: Flowchart illustrating the structure of the testing session.

- In the first step, participants engaged in a 20-minute simulated chat session. The therapist's use of the AI was observed by the first author, who then had the therapist complete the User Experience Questionnaire (UEQ+) regarding their chat interaction with the AI assistant.
- The second step focused on evaluating the semantic search and summarization features. The therapist assessed the semantic search by entering a search term and filling out the System Usability Scale (SUS). They then generated a chat summary and completed a second SUS questionnaire about the summarization feature. All responses were anonymous.
- In the third step, a semi-structured interview was conducted with the participants. Key questions included their experience with the chat interaction, any issues encountered, perceived helpfulness of the AI assistant, useful and non-useful features, and recommended changes.
- After the three-step procedure, two therapists independently assessed the quality of AI-generated summaries from four therapy session dialogues, evaluating linguistic and content quality. Summaries were generated with two models (BART and FLAN-T5) and three summary sizes: large, medium, and small. A total of 16 summaries were assessed by two therapists. Finally, the therapists selected the most suitable summary for real-life therapy sessions.

4.3. Results

In this section, we present and describe the results we have achieved with the methods and the procedure described above (cp. Sec. 4.1 and 4.2). The results are presented based on the measures used, as not all measures apply to every feature evaluated. First, we describe the results of the overall user experience (i.e. UEQ+) which mainly reflects the User Experience (UX) of chatting with support of the *response suggestion* as described in the procedure in Sec. 4.2. This is followed by the results on the usability ratings for the *semantic search* and the *chat summarization*. Finally, we present the ratings about the linguistic quality of the generated *chat summaries* and discuss our interpretation of the results.

4.3.1. Overall User Experience

Figure 10 shows the mean ratings for the single items in the UEQ questionnaire and the two components selected from the UEQ+. Most items have been rated by 6 therapists and all scales are in the range of -3 to 3. For scales Perspicuity, Dependability, and Usefulness, some items were responded to by 5 or 4 therapists (cp. column N in Table 1). Additionally, Table 1 shows the related numerical values and the standard deviations for all items as well as for the item means of the components. In total, all ratings are on the positive side of the scale.

The component Perspicuity (product is easy to understand and easy to learn) has very high ratings with a mean of 2.8. On the other hand, the component's Efficiency (mean 0.71) and Usefulness (mean 0.96) have low ratings. Especially, the item rewarding/rewarding has a very low rating (0.17). Also, the items inefficient/efficient and impractical/practical, both with a rating of 0.33, are close to the neutral point (i.e. 0) of their scales.

For completeness, we also show how our results are related to the benchmark data provided by [37] in Figure 11. That benchmark is based on the results of over 100 studies that have used the UEQ for

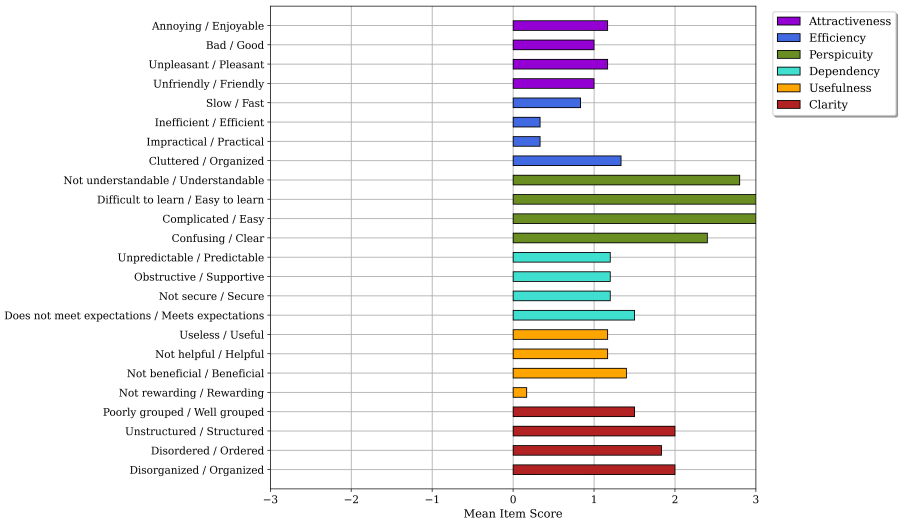

Figure 10: Mean scores of the single items of the User Experience Questionnaire as shown in Table 1.

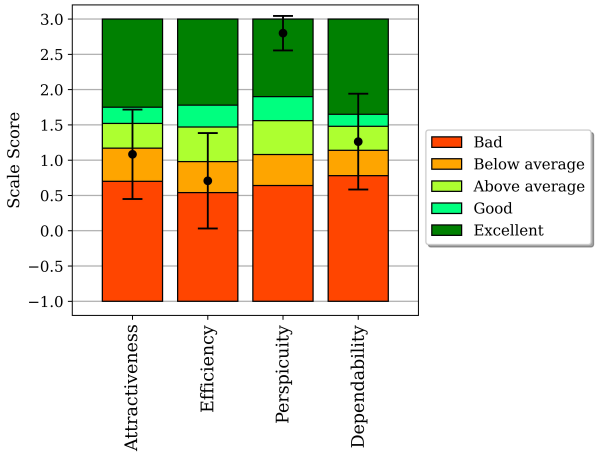

Figure 11: Location of our UEQ results in the UEQ core components in comparison to other applications in the UEQ benchmark (cp. [37]).

evaluation. In that benchmark, the ratings on Perspicuity are in the upper area of the Excellent range. For Dependability (product is predictable and controllable) our results are still above the average of studies in the benchmark, while Attractiveness and Efficiency are below the average.

4.3.2. Usability Ratings for Semantic Search and Chat Summarisation

Table 2 shows the mean ratings and the standard deviations for the items of the System Usability Scale (SUS) that have been used in our study. As for the UEQ, the scales are normalized to a range of -3 to 3 for the result presentation. This does not change the shape of the data, as we used a 7-point Likert scale (see Section 4.1).

SUS ratings have been collected for the *semantic search* (left in Table 2) and the *chat summarisation* feature (right in Table 2). SUS ratings could be collected from all 6 therapists for both features.

For the *semantic search*, the mean ratings of all items are 1 or above. Especially, for SUS-3 and SUS-4 the ratings are high (2.5 and 2.76 respectively). This is in accordance with the results on the overall user experience ratings for Perspicuity (see Figure 10.

For the *chat summarisation*, SUS-3 and SUS-4 are also high. Regarding the frequent usage of the features (SUS-1), the mean rating for summarisation (1.67) is higher than for the search (1.17). However, the standard deviation for SUS-1 is also high in relation to the scale (1.21 and 1.33 respectively). The

	Item Score		
Item	**N**	**Mean**	**SD**
Attractiveness	24	1.08	1.50
Annoying - Enjoyable	6	1.17	1.72
Bad - Good	6	1.00	1.41
Unpleasant - Pleasant	6	1.17	1.60
Unfriendly - Friendly	6	1.00	1.67
Efficiency	24	0.71	1.60
Slow - Fast	6	0.83	1.83
Inefficient - Efficient	6	0.33	1.86
Impractical - Practical	6	0.33	1.63
Cluttered - Organized	6	1.33	1.21
Perspicuity	20	2.80	0.52
Not understandable - Understandable	5	2.80	0.45
Difficult to learn - Easy to learn	5	3.00	0.00
Complicated - Easy	5	3.00	0.00
Confusing - Clear	5	2.40	0.89
Dependability	19	1.26	1.41
Unpredictable - Predictable	5	1.20	1.10
Obstructive - Supportive	5	1.20	1.92
Not secure - Secure	5	1.20	1.64
Does not meet expectations - Meets expectations	4	1.50	1.29
Usefulness	23	0.96	1.58
Useless - Useful	6	1.17	1.72
Not helpful - Helpful	6	1.17	1.33
Not beneficial - Beneficial	5	1.40	1.67
Not rewarding - Rewarding	6	0.17	1.72
Clarity	24	1.83	1.20
Poorly grouped - Well grouped	6	1.50	1.22
Unstructured - Structured	6	2.00	1.10
Disordered - Ordered	6	1.83	1.47
Disorganized - Organized	6	2.00	1.26

Table 1
Number of valid replies (N), mean value, and standard deviation (SD) for the User Experience Questionnaire (UEQ) and the the two additional components (Usefulness and Clarity) of the UEQ+.
Note: The item scores were normalised to 0 and range from −3 to 3. The scores for each of the six scales were calculated as the arithmetic mean of the 4 corresponding items inside the scale.

Item		Semantic Search		Summarisation	
ID	**Text**	**Mean**	**SD**	**Mean**	**SD**
SUS-1	I think that I would like to use this feature frequently	1.17	1.33	1.67	1.21
SUS-2	I found this feature unnecessarily complex*	2.17	1.17	2.23	0.52
SUS-3	I thought the feature was easy to use	2.50	0.55	2.67	0.52
SUS-4	I found that the feature was well integrated in the interface	2.67	0.52	2.83	0.41
SUS-5	I felt very confident using the system	1.83	0.98	1.83	1.60
SUS-6	I felt confident about the results provided by the feature	1.00	1.10	0.50	1.05

Table 2
Mean and standard deviation (SD) for the System Usability Scale items answered by 6 therapists. The item scores were normalised to 0 and range from −3 to 3 from most negative to most positive evaluation.
* The sign of the normalised score for question "I found this feature unnecessarily complex" was inverted to allow an easier interpretation of the results.

lowest rating is SUS-6 (confidence in the results) for both features. The mean of 0.5 for SUS-6 for the summarisation is also the lowest in the entire SUS results.

The rating for SUS-2 (I found this feature unnecessarily complex) is high for both features, which is in line with SUS-3. The ratings for this item have been inverted for easier interpretation – higher values above 0 reflect a positive impact on usability. Thus, the participants do *not* agree with the statement that the usage is unnecessarily complex.

Model	Size	N	Linguistic Quality Median (IQR)				
			Q1	Q2	Q3	Q4	Q5
BART	Large	6	3 (2.3, 3.0)	3.5 (2.3, 4.0)	3.5 (3.0, 4.0)	3.0 (2.3, 3.8)	3.5 (2.3, 4.0)
	Medium	8	4 (3.0, 5.0)	4.5 (3.8, 5.0)	4.5 (3.8, 5.0)	4.0 (3.5, 5.0)	4.0 (3.8, 5.0)
	Small	2	3.5 (3.3, 3.8)	4.0 (3.5, 4.5)	4.5 (4.3, 4.8)	3.0 (3.0, 3.0)	3.5 (3.3, 3.8)
	All sizes	16	3.0 (3.0, 4.0)	4.0 (3.0, 5.0)	4.0 (3.0, 5.0)	3.5 (2.8, 4.3)	4.0 (3.0, 4.0)
FLAN-T5	Large	6	4.0 (3.3, 4.0)	3.5 (2.3, 4.0)	4.0 (2.5, 4.0)	4.0 (2.5, 4.0)	4.0 (2.5, 4.0)
	Medium	8	4.0 (3.0, 4.0)	3.0 (2.8, 4.3)	3.5 (2.0, 4.3)	3.0 (3.0, 4.3)	3.5 (2.0, 4.3)
	Small	2	3.0 (3.0, 3.0)	3.0 (3.0, 3.0)	3.5 (3.3, 3.8)	3.0 (3.0, 3.0)	3.5 (3.3, 3.8)
	All sizes	16	4.0 (3.0, 4.0)	3.0 (2.8, 4.0)	4.0 (2.0, 4.0)	3.0 (3.0, 4.0)	4.0 (2.8, 4.0)
All summaries		32	3.5 (3.0, 4.0)	3.5 (3.0, 4.3)	4.0 (3.0, 5.0)	3.0 (3.0, 4.0)	4.0 (2.8, 4.0)

Table 3
Results of the ratings on the linguistic quality of the generated chat summaries.
Note: The sizes of the summaries correspond to the fraction of summary to dialog, and are categorized in Large (>0.16), Medium (0.12–0.15) and Small (<0.11). For the Linguistic Quality items, the median and the interquartile range are reported. Column N corresponds to the number of evaluations for each row, considering there are two evaluators.

4.3.3. Linguistic and Content Quality of Chat Summaries

Table 3 includes the median and interquartile range (IQR) of each Linguistic Quality Item (see Section 4.1.2 for descriptions) stratified by model and summary size. The median and IQR for each cell are calculated with all the summaries corresponding to the given model and size, for each question separately, and for every evaluator. In addition, the totals for each model (considering all sizes) and the total for all summaries are also reported.

Similarly, Table 4 shows the same metrics for questions 6 and 7, corresponding to the Content Quality of the summaries. Additionally, the mean ROUGE metric for each subgroup and its SD are included next to this question, provided that the interpretation of this metric corresponds to the recall score between the summary and the complete dialog, that is, the percentage of overlapping words among all words in the reference text. In particular, Q6 can be seen as the recall of the information in the dialog that is considered relevant by the evaluator.

For all items (Q1 to Q7) in Table 3 and 4 higher values mean a better rating on the Likert-scale. With the exception of Q1, the ratings for summaries with BART (compared by all sizes) are equal to or higher than with FLAN-T5. Furthermore, there is a trend, especially in Q1 to Q5, that the highest ratings are given for medium-sized summaries. Due to the low number of study participants, no statistical tests were performed to support the observed trends.

When asked to choose the most adequate summary among the 4 produced for each dialog, a BART-generated summary was chosen 6 times, whereas a FLAN-T5-generated summary only 2 times. Also, summaries with Medium Size were chosen 6 times, and summaries with Large and Small sizes only 1 time each. These results are in line with the quantitative results on the questionnaire.

As for the inter-rater reliability, Krippendorff's alpha for ordinal data was calculated separately for each question and produced values ranging from -0.01 (Question 2) to -0.63 (Question 5). These values denote very low inter-rater reliability, to the point where the concordance between the two raters is lower than that expected by chance.

			Content, Median (IQR)		
Model	**Size**	**N**	**Q6**	**Q7**	**ROUGE, Mean (std)**
BART	Large	6	3.5 (2.3, 4.8)	3.0 (2.3, 4.5)	0.29 (0.08)
	Medium	8	4.0 (3.5, 5.0)	4.5 (3.5, 5.0)	0.20 (0.01)
	Small	2	2.5 (2.3, 2.8)	3.0 (2.5, 3.5)	0.13
	All sizes	16	4.0 (2.0, 5.0)	4.0 (2.0, 5.0)	0.23 (0.07)
FLAN-T5	Large	6	4.0 (3.3, 4.0)	4.0 (3.3, 4.8)	0.30 (0.07)
	Medium	8	3.0 (3.0, 4.3)	3.0 (2.8, 4.3)	0.23 (0.02)
	Small	2	3.0 (3.0, 3.0)	2.5 (2.3, 2.8)	0.16
	All sizes	16	3.0 (3.0, 4.3)	3.0 (2.8, 4.3)	0.25 (0.05)
All summaries		32	3.5 (3.0, 5.0)	3.5 (2.0, 5.0)	0.24 (0.07)

Table 4
Results of the ratings on the linguistic quality of the generated chat summaries.
Note: The sizes of the summaries correspond to the fraction of summary to dialog, and are categorized in Large (>0.16), Medium (0.12-0.15) and Small (<0.11). For the Content items, the Median and the Interquartile range are reported, and for the ROUGE score, the mean and standard deviation. Column N corresponds to the number of evaluations for each row, considering there are two evaluators. The ROUGE means and standard deviations are calculated with the available summaries on each subgroup, which corresponds to N/2.

4.3.4. Insights from the Interviews

The semi-structured interview provided valuable insights into the overall experience of interacting with the prototype. This section provides valuable feedback received from the therapists about the AI assistant and the prototype as a whole.

Interaction experience

Therapists appreciated the user-friendliness of the interface and expressed satisfaction over the adaptable nature of the AI assistant as it could be toggled on and off based on their needs. They found its integration within the system to be smooth.

Features provided by the AI-assistant

The therapist more acquainted with chat-based therapy seemed more comfortable with the AI assistant. The usage frequency of AI-suggested responses correlated with their quality; when effective, therapists actively monitored and utilized the suggestions. The displayed relevance scores alongside suggestions did not influence the therapist's choices. Despite not always being helpful, they didn't disrupt the therapists' interactions with the clients significantly.

Perceived Usefulness of Features

The response suggestions were not found to be helpful in most cases and were rarely used. While two therapists found them useful and utilized them frequently, others didn't use them, despite actively examining the suggestions in the sidebar. Some therapists felt the semantic search feature was less necessary during interactions.

Potential improvements

The semantic search feature demonstrated effectiveness with single-word search terms but struggled with multiple-word queries, displaying both relevant and irrelevant results that could have been filtered further by the AI model. The summarization feature was liked the most, with most therapists finding it highly beneficial, and summarizing the crucial points from the chat conversation.

4.4. Results Discussion

Overall, the ratings for the usability (SUS) of *semantic search* and *char summaries*, overall user experience (UEQ and UEQ+ components) when chatting with AI support and especially *response suggestions*, and the linguistic quality of the summaries are all on the positive part of the scale. Predominantly, they are not optimal and some are close to neutral ratings.

The most important issue is the neutral rating of the perceived efficiency with the UEQ (cp. Figure 10 in Section 4.3). Here, we see a great need to try to improve the interaction. From our interviews after the interaction, we see that not only the (limited) accuracy of the response suggestions or their integration into the user interface is the source of perceived issues. It seems that the general attitude of users towards AI applications has a major influence on their use and evaluation. In upcoming user studies, we will gather the corresponding attitude of the therapists as well as try better to explain the AI features of the chat assistant.

We see another interesting result in the ratings of the content of the chat summaries. The mean rating of SUS-6 (confidence about the content, cp. Table 2) is close to neutral with a mean of 0.5. On the other hand, the ratings on the content-related items of the linguistic quality questionnaire (Q6 and Q7, see Table 4 are quite positive for BART. Probably, the interference of the two different models (BART and FLAN-T5) as well as some grammatical issues (reflected in Q1 for BART in Table 3 led to the comparatively low ratings for SUS-6.

5. Discussion

In this section, we discuss the overall study results, focusing first on the usability of the different features as evaluated by the therapists, and then on their general experience with the system.

Two out of the six therapists found the response suggestions particularly useful, especially given the repetitive nature of the writing-based intervention. This feature allowed therapists to avoid redundancy in their responses while maintaining the flexibility to decide on the appropriateness of the suggestions. However, since the underlying model used for matching responses was based on simple sentence matching, it occasionally failed to capture the full context of the conversation or align with a therapist's unique conversational style, leading to less accurate suggestions. Regarding the chat summary, most therapists felt confident using the tool, though one expressed some reservations. Overall, it was considered valuable for tracking clients and assisting with report writing.

Regarding chat summaries, their quality ranged from acceptable to good, depending on the length. Medium-length summaries were rated as good, while longer summaries were deemed only acceptable. This preference likely varies by therapist's working style and the needs of each case. However, conclusions should be made cautiously, as inter-rater agreement was low, indicating high subjectivity in the evaluation. This highlights the need for larger studies with more evaluators, though we remain confident in the AI assistant's ability to meet therapists' needs.

Therapists rated the semantic search feature lower than the other two, but most participants expressed interest in using it and felt confident in its utility. However, one participant had reservations about its reliability. This finding should be interpreted cautiously, as the tool is likely used occasionally, for retrieving specific information or refreshing memory on previous topics. Thus, its effectiveness is context-dependent. Overall, semantic search proved useful for finding session-related keywords.

Most therapists found the prototype 'pleasant,' 'enjoyable,' and 'friendly,' but preliminary results show below-average ratings on the attractiveness dimension of the UEQ scale, similar to the efficiency dimension. These findings may suggest that some participants' needs were not fully met by the AI prototype. However, therapists generally had a positive view of the organization of the tool. Practicality ratings were mixed, with efficiency ratings mostly in the reasonable-to-poor range, and the speed of the tool perceived differently among participants. In terms of handling the AI assistant (perspicuity), therapists rated it highly for comprehension, ease of learning, and clarity. On the dependability dimension, the AI assistant was seen as consistent, with positive feedback on its predictability and fulfillment of expectations. Most participants viewed the tool as helpful, though some expressed concerns

about its security. Therapists held positive beliefs about the AI assistant's potential contribution to the therapeutic process, with favorable evaluations on both usefulness and clarity. The tool's structure, grouping, order, and organization were rated reasonably well, while most therapists considered it useful, though some had reservations about its level of help and benefit. As the AI's responses improve with more data from real therapy sessions, trust in the tool is expected to increase, potentially leading to higher ratings in future evaluations. However, due to the limited number of participants, meaningful statistical tests were impractical. Similar to previous studies [38, 39, 40], our findings highlight the potential of AI in therapy, while also acknowledging the need for further refinement to better align AI tools with therapists' unique styles, case-specific needs, and specific conversational requirements.

5.1. Limitations and future research

This study has several limitations that future research should address. First, the small sample size of six therapists and the focus on a specific intervention limit the generalizability of the findings to other mental health professionals. Second, the reliance on self-report measures may introduce biases, and some participants did not complete all questionnaires, potentially affecting the results. Despite anonymous administration, incomplete responses may still influence the interpretation of the findings. A third limitation involves the participants' prior exposure to AI, which was not considered in the analysis. Future studies should account for prior AI experience as a variable. Additionally, the messages used in this study were derived from a therapist training session, which may not fully reflect actual therapist-client conversations. Key factors such as message length, response time, and the client's exact mental state were not replicated, potentially affecting the burden on therapists. These aspects should be assessed in real-world client settings. Despite these limitations, this study represents a valuable first step in developing AI assistants for therapists. Future research should explore the use of AI tools in training new therapists and investigate the application of Large Language Models (LLMs), like Mistral, Llama, and Gemini, for tasks such as chat summarization, leveraging larger technical infrastructures. Furthermore, incorporating advanced therapy frameworks—such as classifying messages into categories like open-ended or reflection-based—could enhance response suggestions and improve the overall experience. Finally, utilizing larger models requires robust anonymization protocols and careful attention to ethical considerations.

6. Conclusion

This study evaluated therapists' experiences with a prototype AI assistant in a therapeutic intervention. Therapists assessed the prototype in a simulated setting and generally found it valuable for potential integration into practice. The findings suggest AI assistants could support interventions for clients with CSAM-related issues or problematic sexual behaviors by reducing cognitive load and providing personalized assistance in remote therapy. However, further empirical research is needed to assess their effectiveness in real-world settings and how their impact compares to simulated environments. Future development should focus on strengthening therapist-client relationships, refining AI-assisted therapy experiences, and aligning AI tools with therapists' needs to maximize clinical utility. Technically, future research should explore fine-tuned models trained on real-world therapeutic interactions and use LLMs' language capabilities to enhance adaptability and contextual understanding in AI-assisted therapy.

7. Acknowledgments

The described work was done in the frame of the STOP-CSAM project, funded by the European Commission in the activity ISF-2021-TF1-AG-CYBER with project ID 101084719. The development of the system as well as the the related study have been approved by the ethics committee of Charité – Universitätsmedizin Berlin, under the reference number EA4/173/23. The ethics approval covers the participation of the involved therapists.

References

[1] World Health Organization, Artificial intelligence in mental health research: New who study on applications and challenges, https://www.who.int/europe/news/item/06-02-2023-artificial-intelligence-in-mental-health-research--new-who-study-on-applications-and-challenges, 2023. Accessed: 13 November 2023.

[2] G. S. Duden, S. Gersdorf, K. Stengler, Global impact of the covid-19 pandemic on mental health services: A systematic review, Journal of Psychiatric Research 154 (2022) 354–377. doi:10.1016/j.jpsychires.2022.08.013.

[3] S. Graham, C. Depp, E. E. Lee, C. Nebeker, X. Tu, H. C. Kim, D. V. Jeste, Artificial intelligence for mental health and mental illnesses: an overview, Current psychiatry reports 21 (2019) 1–18.

[4] F. Minerva, A. Giubilini, Is ai the future of mental healthcare?, Topoi (2023) 1–9.

[5] S. Swaminath, R. M. Simons, M. L. Hatwan, Understanding pedophilia: a theoretical framework on the development of sexual penchants, Journal of child sexual abuse 32 (2023) 732–748.

[6] H.-E. Lee, T. Ermakova, V. Ververis, B. Fabian, Detecting child sexual abuse material: A comprehensive survey, Forensic Science International: Digital Investigation 34 (2020) 301022.

[7] M. Stoltenborgh, M. H. Van Ijzendoorn, E. M. Euser, M. J. Bakermans-Kranenburg, A global perspective on child sexual abuse: Meta-analysis of prevalence around the world, Child maltreatment 16 (2011) 79–101.

[8] M. Henshaw, C. Arnold, R. Darjee, J. R. Ogloff, J. A. Clough, Enhancing evidence-based treatment of child sexual abuse material offenders: The development of the cem-cope program, Trends and Issues in Crime and Criminal Justice (2020) 1–14.

[9] N. Aoki, The importance of the assurance that "humans are still in the decision loop" for public trust in artificial intelligence: Evidence from an online experiment, Computers in Human Behavior 114 (2021) 106572.

[10] R. Louie, A. Nandi, W. Fang, C. Chang, E. Brunskill, D. Yang, Roleplay-doh: Enabling domain-experts to create llm-simulated patients via eliciting and adhering to principles (arxiv: 2407.00870). arxiv, 2024.

[11] A. Chaszczewicz, R. S. Shah, R. Louie, B. A. Arnow, R. Kraut, D. Yang, Multi-level feedback generation with large language models for empowering novice peer counselors, arXiv preprint arXiv:2403.15482 (2024).

[12] M. A. Kuhail, N. Alturki, J. Thomas, A. K. Alkhalifa, A. Alshardan, Human-human vs human-ai therapy: An empirical study, International Journal of Human–Computer Interaction (2024) 1–12.

[13] A. Sharma, I. W. Lin, A. S. Miner, D. C. Atkins, T. Althoff, Human–ai collaboration enables more empathic conversations in text-based peer-to-peer mental health support, Nature Machine Intelligence 5 (2023) 46–57.

[14] R. Larasati, A. De Liddo, E. Motta, Ai healthcare system interface: Explanation design for non-expert user trust, in: ACMIUI-WS 2021: Joint Proceedings of the ACM IUI 2021 Workshops, volume 2903, CEUR Workshop Proceedings, 2021.

[15] C. Canning, T. J. Donahue, M. Scheutz, Investigating human perceptions of robot capabilities in remote human-robot team tasks based on first-person robot video feeds, in: 2014 IEEE/RSJ International Conference on Intelligent Robots and Systems, IEEE, 2014, pp. 4354–4361.

[16] C. Pelau, D. C. Dabija, I. Ene, What makes an ai device human-like? the role of interaction quality, empathy and perceived psychological anthropomorphic characteristics in the acceptance of artificial intelligence in the service industry, Computers in Human Behavior 122 (2021) 106855.

[17] O. Ozmen Garibay, B. Winslow, S. Andolina, M. Antona, A. Bodenschatz, C. Coursaris, G. Falco, S. M. Fiore, I. Garibay, K. Grieman, et al., Six human-centered artificial intelligence grand challenges, International Journal of Human–Computer Interaction 39 (2023) 391–437.

[18] P. M. Doraiswamy, C. Blease, K. Bodner, Artificial intelligence and the future of psychiatry: Insights from a global physician survey, Artificial intelligence in medicine 102 (2020) 101753.

[19] S. Baek, Y. Kim, The study of behavioral intention of delivery application by applying the extended technology acceptance mode, The Korean Journal of Food And Nutrition 31 (2018) 185–194.

[20] R. de Kervenoael, R. Hasan, A. Schwob, E. Goh, Leveraging human-robot interaction in hospitality services: Incorporating the role of perceived value, empathy, and information sharing into visitors' intentions to use social robots, Tourism Management 78 (2020) 104042.
[21] C. Wang, Consumer acceptance of self-service technologies: An ability–willingness model, International Journal of Market Research 59 (2017) 787–802.
[22] D. Norman, User centered system design, New perspectives on human-computer interaction (1986).
[23] J. Preece, Y. Rogers, H. Sharp, Interaction Design: Beyond Human-Computer Interaction, John Wiley & Sons, New York, NY, 2002. Chapters 9–13.
[24] Q. Yang, A. Steinfeld, C. Rosé, J. Zimmerman, Re-examining whether, why, and how human-ai interaction is uniquely difficult to design, in: Proceedings of the 2020 chi conference on human factors in computing systems, 2020, pp. 1–13.
[25] R. A. Powell, H. M. Single, Focus groups, International journal for quality in health care 8 (1996) 499–504.
[26] H. Lieberman, User interface goals, ai opportunities, AI Magazine 30 (2009) 16–16.
[27] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need.(nips), 2017, arXiv preprint arXiv:1706.03762 10 (2017) S0140525X16001837.
[28] N. Reimers, I. Gurevych, Sentence-bert: Sentence embeddings using siamese bert-networks, arXiv preprint arXiv:1908.10084 (2019).
[29] G. Salton, C. Buckley, Term-weighting approaches in automatic text retrieval, Information processing & management 24 (1988) 513–523.
[30] F. Rahutomo, T. Kitasuka, M. Aritsugi, et al., Semantic cosine similarity, in: The 7th international student conference on advanced science and technology ICAST, volume 4, University of Seoul South Korea, 2012, p. 1.
[31] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. rahman Mohamed, O. Levy, V. Stoyanov, L. Zettlemoyer, Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension, in: Annual Meeting of the Association for Computational Linguistics, 2019. URL: https://api.semanticscholar.org/CorpusID:204960716.
[32] H. W. Chung, L. Hou, S. Longpre, B. Zoph, Y. Tay, W. Fedus, Y. Li, X. Wang, M. Dehghani, S. Brahma, et al., Scaling instruction-finetuned language models, arXiv preprint arXiv:2210.11416 (2022).
[33] B. Gliwa, I. Mochol, M. Biesek, A. Wawer, Samsum corpus: A human-annotated dialogue dataset for abstractive summarization, ArXiv abs/1911.12237 (2019). URL: https://api.semanticscholar.org/CorpusID:208010268.
[34] H. Hardy, S. Narayan, A. Vlachos, Highres: Highlight-based reference-less evaluation of summarization, 2019, pp. 3381–3392. doi:10.18653/v1/P19-1330.
[35] S. Castro, Fast Krippendorff: Fast computation of Krippendorff's alpha agreement measure, https://github.com/pln-fing-udelar/fast-krippendorff, 2017.
[36] C.-Y. Lin, ROUGE: A package for automatic evaluation of summaries, in: Text Summarization Branches Out, Association for Computational Linguistics, Barcelona, Spain, 2004, pp. 74–81. URL: https://aclanthology.org/W04-1013.
[37] M. Schrepp, A. Hinderks, J. Thomaschewski, Construction of a benchmark for the user experience questionnaire (ueq), International Journal of Interactive Multimedia and Artificial Intelligence 4 (2017) 40–44.
[38] M. Eshghie, M. Eshghie, Chatgpt as a therapist assistant: a suitability study, arXiv preprint arXiv:2304.09873 (2023).
[39] R. AlMakinah, A. Norcini-Pala, L. Disney, M. A. Canbaz, Enhancing mental health support through human-ai collaboration: Toward secure and empathetic ai-enabled chatbots, arXiv preprint arXiv:2410.02783 (2024).
[40] Z. Iftikhar, S. Ransom, A. Xiao, J. Huang, Therapy as an nlp task: Psychologists' comparison of llms and human peers in cbt, arXiv preprint arXiv:2409.02244 (2024).