An AI-Based Virtual Client for Educational Role-Playing in the Training of Online Counselors

Eric Rudolph[a], Natalie Engert[b] and Jens Albrecht[c]
Technische Hochschule Nürnberg Georg Simon Ohm, Nürnberg, Germany

Keywords: Counseling, Chatbot, Large Language Model, Persona, Role-Play.

Abstract: The paper presents the Virtual Client for Online Counseling (VirCo), a novel system for training online counselors through simulated client interactions. Addressing the rising need for digital communication skills in counseling, VirCo leverages large language models (LLMs) to create a chatbot that generates realistic responses from different client personas. The approach complements traditional role-playing methods in academic training, offering independent practice opportunities without direct supervision. To ensure privacy VirCo's chatbot interface uses an open-source LLM for response generation. The system's dataset comprises detailed persona descriptions and transcripts from role-play sessions, contributing to the authenticity of the training experience. The evaluation of the quality of the conversations utilized both human evaluators and LLMs. Results show a high degree of coherence and persona alignment in responses, highlighting VirCo's effectiveness as a training tool. The paper concludes by showcasing the features of the VirCo learning platform like compulsory assignments and multiple feedback mechanisms.

1 INTRODUCTION

Many people seek professional help in personal crisis situations. In this context, online counseling services play a vital role. At best, well-trained counselors can clarify the causes of problems in a dialogue and provide individual assistance. The American 988 crisis hotline, for example, is contacted over 400.000 times per month (Center for Behavioral Health Statistics and Quality, 2023). While telephone counseling is still dominating, many organizations also support the growing demand for text-based online counseling via email or chat.

The use of digital communication through online platforms presents unique challenges that require online counselors to receive specialized training before providing counseling to individuals seeking help in a digital environment. To ensure sufficient hands-on experience alongside theoretical learning, trainees should interact with clients in virtual environments as part of their education (DeMasi et al., 2020). Nonetheless, in academic settings, managing actual clients is uncommon. Thus, role-playing is frequently employed, because it offers several advantages over traditional case studies. In addition to immediate and immersive experiences that allow participants to embody and understand perspectives other than their own it also improves communication skills and empathy (Mianehsaz et al., 2023; Bharti, 2023; Sai Sailesh Kumar Goothy et al., 2019). Role-playing is also notably effective due to its ability to immerse individuals in another's perspective and it also provides the learner with a safe environment without harming potential clients in the real world (Kerr et al., 2021). Despite the several advantages role-playing can be difficult to implement due to the extensive supervision by trainers and the need to rely on role-play partners. As a result, course participants are restricted in their ability to practise independently and gain practical experience.

In this paper, we introduce VirCo, the **Vir**tual Client for Online **Co**unseling, a system to simulate clients for the training of online counselors (see figure 1 for a sample dialog). With VirCo, learners are able to engage autonomously with different clients and gain initial counseling experiences that are free from direct trainer involvement. Some effects, such as the emotional impact of recommended actions on the client, can only be observed in dialog with a real person. Thus, VirCo is not intended to replace role-

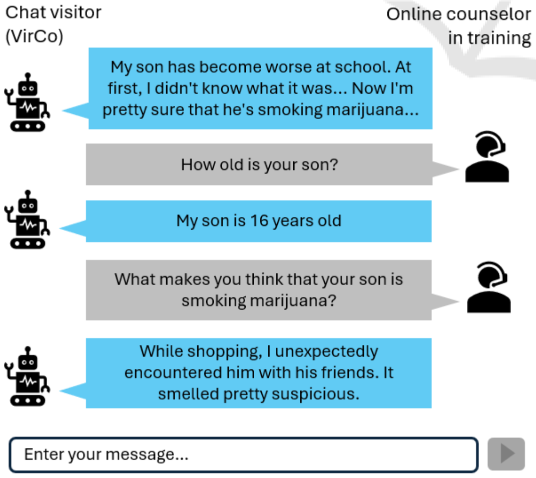
[a] https://orcid.org/0009-0003-0615-4780
[b] https://orcid.org/0009-0001-2493-0208
[c] https://orcid.org/0000-0003-4070-1787

Rudolph, E., Engert, N. and Albrecht, J.
An AI-Based Virtual Client for Educational Role-Playing in the Training of Online Counselors.
DOI: 10.5220/0012690700003693
Paper published under CC license (CC BY-NC-ND 4.0)
In *Proceedings of the 16th International Conference on Computer Supported Education (CSEDU 2024) - Volume 2*, pages 108-117
ISBN: 978-989-758-697-2; ISSN: 2184-5026
Proceedings Copyright © 2024 by SCITEPRESS – Science and Technology Publications, Lda.

playing but to complement it.

The idea of creating a chatbot for counselor training was explored before as discussed in section 2. Our solution is based on ideas from previous work, but makes use of the recent advances in natural language processing with large language models (LLMs). Thus, the core of the system is a chatbot which uses an LLM to generate answers. The chatbot is part of a comprehensive learning platform which includes functionality to practice with different personas (problem cases) and to provide automated and manual feedback to students. To represent the personas, we created transcripts of role-plays based on different problem scenarios like drug abuse of a child or quarrels between parents. The persona descriptions and the transcripts are used to shape the answers of the LLM. This paper gives an overview about the architecture of the whole learning platform and first evaluation results for the virtual client chatbot. Our contribution is as follows:

- We demonstrate, how large language models can be utilized for educational role-play.
- We present a methodology to simulate different personas with large language models.
- We introduce an architecture for a chatbot-based learning platform for educational role-play.
- We evaluate the course of conversation coherence and persona consistency of a persona-based client chatbot.



My son has become worse at school. At first, I didn't know what it was... Now I'm pretty sure that he's smoking marijuana...

How old is your son?

My son is 16 years old

What makes you think that your son is smoking marijuana?

While shopping, I unexpectedly encountered him with his friends. It smelled pretty suspicious.

Figure 1: Excerpt of an example conversation with the virtual client (VirCo). VirCo simulates a concerned mother who assumes that her son smokes marijuana.

2 RELATED WORK

The famous ELIZA, developed in 1966, is a pioneering chatbot that served as the model for many others. It simulates a therapist by posing and responding to particular queries from the individual interacting with it (Weizenbaum, 1966). Since then, many publications have been published in the area of chatbot for counseling, especially in the area of psychotherapy and mental health. See (Boucher et al., 2021) and (Xu and Zhuang, 2022) for an overview. These counseling chatbots are mainly intended to support diagnostics and behavior change, or simply to deliver supportive content. However, there are ongoing discussions about the feasibility of using AI-powered chatbots as a substitute for counselors or therapists.

Thus, several studies have sought to simulate the character of the client rather than the therapist. These chatbots are intended to be used for training aspiring professionals. (Tanana et al., 2019) introduced ClientBot, a patient-like chatbot imitating a visitor in a psychotherapy session. DeMasi et al. made a significant contribution to the research on chatbots in counseling contexts through the creation of the Crisisbot (DeMasi et al., 2019; DeMasi et al., 2020). Crisisbot simulates a caller to a suicide prevention hotline. In order to provide different problem scenarios of clients (personas), Crisisbot also introduced a multi-task training framework to construct persona-specific responses (DeMasi et al., 2020). An overview of persona-based conversational AI was given in (Liu et al., 2022).

One approach to provide consistent answers is to retrieve candidate responses from a corpus of prototype conversations as input for the generation of the actual response (Tanana et al., 2019; DeMasi et al., 2019; Liu et al., 2022). To further improve the retrieval, several works used an utterance classifier to predict the type of the next response (Cao et al., 2019; Park et al., 2019; DeMasi et al., 2020). Generative models like Seq2Seq were then used to generate responses based on the selected candidates (Tanana et al., 2019; DeMasi et al., 2020; Liu et al., 2022).

Generating responses which are both, consistent to a given persona and coherent to the course of conversation, is a crucial prerequisite for a realistic learning experience in such a setting. But this is a challenging requirement, because the conversations are basically open-domain and don't follow a clear structure. The state-of-the-art in 2019/20 did not produce satisfying results in this regard, because the models generated sometimes unrealistic, distracting or irrelevant responses (Tanana et al., 2019) or were not reliably consistent (DeMasi et al., 2020). In Crisisbot, the generated responses were also generally shorter than the real responses extracted from the corpus (DeMasi et al., 2020). This had a negative impact on the learning experience of aspiring counselors.

The situation changed with the rapid development of Large Language Models since the launch of ChatGPT. LLMs are very good at generating coherent dialogues. A typical drawback, hallucination, might even be beneficial in our setting. Lee et al. introduced a methodology to prompt LLMs for long open-domain conversations utilizing few-shot in-context learning and chain-of-thought (Lee et al., 2023) . Chen et al. showed that LLMs can effectively be used for counselor and client simulation without fine-tuning (Chen et al., 2023). Our approach combines the power of LLMs with ideas from the discussed previous work on client-simulating counseling chatbots.

3 DATASET

The basis for the Virtual Client evaluation is a dataset consisting of two parts: the persona descriptions and for each persona a set of simulated conversations. The persona descriptions were created by domain experts based on documented email counseling sessions and public forum posts[1]. To simulate a variety of counseling settings, we created so far seven persona descriptions. A key aspect of these personas is the definition of a main concern, which represents the client's motivation for using chat counseling. In particular, problems in the area of addiction counseling or educational counseling, such as the following problem description, were defined:

> *"Elke is worried about her 16-year-old son Lukas, who is in the 10th grade. She suspects that he is using drugs or smoking mariuhana because of his circle of friends. He was generally not a bad student, but unfortunately his grades have deteriorated recently. She has already tried to talk to Lukas about it, both about his school performance and his drug use, but he keeps blocking it and doesn't want to talk about it. Elke suspects that his son's changed behavior is due to his circle of friends, as they have a bad influence on him."*

Additionally, the linguistic characteristics of the personas are carefully considered to capture the communication preferences and styles of the clients.

Online counselor trainees were then asked to use these persona descriptions to role-play chat counseling conversations. The role-plays lasted approximately one hour, based on the typical length of chat counseling training sessions. Further dataset statistics can be found in Table 1.

Table 1: Dataset statistics.

Dataset component	Count
Number of conversations	56
Average messages per conversation	40.39
Counselor messages	1125
Client messages	1137

The data was collected in four phases, with the participation of different trainees. Although all conversations relate to the defined persona descriptions, the actual conversations include some variations. For instance, simulated clients contacted the wrong counseling centre and had to be referred to the correct centre by the counselor. The writing style also varies significantly. During counseling sessions, clients may exhibit varying levels of cooperation. Some may provide detailed answers to the counselor's questions, while others may be uncooperative, providing brief responses or struggling to articulate their problems clearly.

4 SYSTEM OVERVIEW

The Virtual Client can be accessed by online counselor trainees through a software portal which includes a learning management system and an interface for answer generation (Figure 2).

VirCo Learning Platform. The VirCo Learning Platform presents itself to trainees as a professional chat interface for online counseling, similar to what a trained counselor might use for counseling via a computer or mobile phone. The unique aspect of this platform is that trainees interact with a simulated client, programmed to respond to their statements using a large language model. The learning platform incorporates gamification elements to improve the learning experience for users. The platform itself is further described in section 6.

VirCo Bot. After a learner sends a message to the client on the learning platform, it is forwarded to the response generation interface. Due to privacy concerns, we prefer not to use cloud-based LLMs like ChatGPT. Instead, the open-source LLM Vicuna-

[1] All persona descriptions, conversations and prompts have actually been created in German, as the Virtual Client is intended for German-speaking students. We translated the examples to English for this publication.

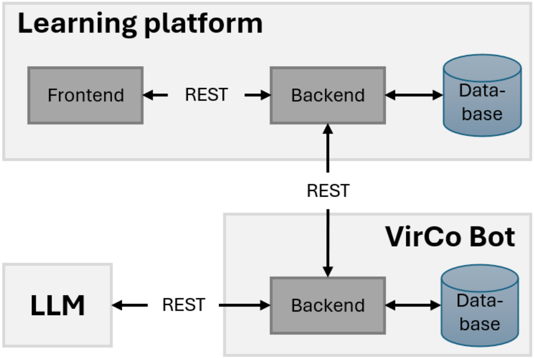


Figure 2: Overview of the VirCo architecture, which comprises two independent components: the learning platform and the chatbot interface. These components can function separately and communicate with each other via REST.

13B-1.5 (Chiang et al., 2023) is used to generate responses based on the previously described personas and dataset. The model is hosted on a private High Performance Computing (HPC) cluster. However, switching to OpenAI models for evaluation purposes is also possible.

The Vicuna 13B model was selected because a preliminary test showed that it fulfills the following requirements:

- It is able to carry out role-plays and respond to previous conversations.
- It is open source and can therefore be hosted on a private environment.
- With 13 billion parameters, it has a comparatively small size for the described capabilities.
- It has an understanding of the German language and is able to respond in German.

For prompting the model, a role-play prompt was created, which contains a role assignment, a placeholder for the description of the persona and a placeholder for a conversation history. Role-play prompts with an LLM often enhance its capabilities (Shanahan et al., 2023; Kong et al., 2023). The structure of such prompts often follows a pattern in which the persona of the character is described, followed by the inclusion of the conversation history, as shown in the following example in the context of online counseling:

> *“Pretend you are {name}. {name} is currently in counseling and is writing with her/his social counselor. Give statements that {name} would give. {problem_description} Chat history: {chat_history} {name} answers briefly and concisely”*

The placeholders in the prompt are then replaced by the persona (e.g. *Elke* from above) and conversation-specific data. The name and problem description are derived from the persona description (see section 3), while the conversation history is retrieved from the database. An example of a conversation history was already shown in Figure 1. For this purpose, the maximum token length was set to 256, as a message in the chat should not be too long, and a stop sequence was added, which interrupts the generation at the phrase “Counselor:”, as the language model would otherwise often unintentionally generate the next reply from the counselor. An example replacement for the placeholder *chat_history* from our data set looks like this:

> *Client: Hello, I’m Elke and my child Lukas is taking drugs ... Can you help me here? Counselor: Hello Elke. Great that you got in touch. I’m Marie and I’m a counselor. I’d be happy to try and help you, but first I’d like to clarify the technical and organizational framework with you. Is that okay? Client:*

Once the response has been generated by the LLM, it is sent back to the learning platform and displayed to the user. All conversations are logged to the database as input for the feedback module.

5 EVALUATION

For the purpose of evaluation, we segmented each of the 56 conversations into individual conversational turns. Initially, we analyzed the first two messages if initiated by the client, as illustrated in the concluding remarks of section 4, or the opening message if initiated by the counselor. Subsequently, we incrementally included additional turns of conversation. To illustrate, the initial sequence comprises alternating turns between the client and counselor, progressively building up to sequences involving multiple exchanges (e.g., Client - Counselor - Client - Counselor). This process resulted in a dataset comprising 1,125 entries for thorough evaluation. This means we generated an answer for each counselors message. The data was then rated by human evaluators as well as automatically by GPT-3.5 and GPT-4 on two tasks, dialogue coherence (task 1) and persona consistency (task 2). We used GPT-3.5 and GPT-4 because they often achieve competitive correlation with human judgment for natural language generation tasks (Wang et al., 2023). The task for dialogue coherence was to evaluate whether the generated answer fits to the conversation history. The task for consistency was to evaluate whether the generated answer fits to the persona. The integration of human and LLM evaluations serves distinct purposes:

LLM Empowered Evaluation: enables the analysis of large datasets with consistency and efficiency. LLMs can process extensive conversations, identifying patterns and metrics at scale. Since conventional metrics such as ROGUE or BLEU score do not provide reliable results for generated language tasks LLMs have become the standard for the automatic evaluation of dialogues (Chang et al., 2023)[p. 26] i. e. in (Lin and Chen, 2023; Liu et al., 2023).

Human Evaluation: provides often a more robust result with nuanced understanding of conversational context and emotional undertones. Reliable human evaluators are essential for identifying complex conversational dynamics in the context of counseling sessions that AI might overlook.

Conducting both evaluation approaches enables a more comprehensive insight. In addition, both approaches are compared to assess their degree of overlap and to analyze the differences. We start by explaining the automatic evaluation procedure, as human evaluators basically got the same instructions as the LLM.

5.1 Automatic Evaluation by LLMs

The automatic evaluation was carried out with the OpenAI models "GPT-3.5 1106 Turbo" (hereafter abbreviated as GPT-3.5) and "GPT-4 1106 Preview" (hereafter abbreviated as GPT-4). The evaluation prompts are inspired by G-Eval, a framework for the evaluation of Natural Language Generation tasks (Liu et al., 2023). Both prompts begin with a task description of the evaluation. Afterwards there are task specific evaluation criteria and a scoring description. This is followed by conversation-specific information.

5.1.1 Task 1: Conversation History Rating

The following prompt is used for the automated generation of ratings and evaluations with GPT. The first section is about familiarizing the model with the instruction. Since GPT tends to evaluate the course of the conversation as a whole instead of just the generated response based on the course of the conversation, this is explicitly described here:

> *"The following conversation shows a chat counseling session between a client and a counselor. The context of the chat between these two people is online social counseling. Your task is to evaluate to what extent the generated message matches the previous conversation. Please evaluate only whether the generated message matches the previous conversation."*

The next part outlines the criteria for evaluation. The main focus of the evaluation criteria lies on the coherence to the conversation history, ensuring the chat maintains logical consistency; Content Accuracy, requiring that the answers are correct and relevant to the advisor's message; and the Flow of the Conversation, emphasizing the need for a natural and uninterrupted progression of the chat without sudden topic shifts or confusing elements.

> *"Evaluation criteria:*
> *Coherence: The chat should be coherent.*
> *Content accuracy: The content of the answers must be correct and appropriate to the counselors message.*
> *Flow of the conversation: The continuation of the chat should have a natural and smooth flow, without abrupt changes of topic or confusing contexts. Please keep in mind that this is a chat. Short, precise and direct answers can occur here."*

The third section provides a scoring system for the evaluation. A score of 0 signals a fully coherent response that is contextually accurate and flows well with the previous conversation. A score of 1 indicates basic coherence and content alignment with minor grammatical or spelling errors. A score of 2 denotes a lack of coherence, confusion, or repetitive words or sentences, indicating a mismatch with the ongoing conversation. The human evaluators rate those scores on a website with a green (0: fits good), yellow (1: fits mediocre) and red (2: doesn't fit) radio button.

> *"Score rating:*
> *0: The generated answer is coherent with the previous conversation. The content of the generated answer is correct and the flow of the conversation is appropriate to the previous course.*
> *1: The generated answer is basically coherent with the course of the conversation so far and the content matches the course of the conversation, but the generated answer contains grammatical errors or spelling mistakes*
> *2: The generated answer does not match the course of the conversation so far. It is confused or there are many repetitions of words or sentences"*

The following section outlines the structure for conducting the evaluation. It highlights that the evaluation will be based on the previously mentioned criteria. The format includes placeholders for the course of the conversation (history) and the generated response (answer) from the LLM. Note that the text in curly brackets is also a placeholder for the conversation history, the generated response.

> *"The evaluation is based on the evaluation criteria.*
> *Course of the conversation:*
> *{history}*
> *Generated response:*
> *{answer}"*

The last block specifies the format in which the evaluation results are to be displayed. This is added for post-processing purposes.

> *"Please structure your answer as JSON with the attributes rating and reason. Output only the JSON format: "*

5.1.2 Task 2: Persona Consistency Rating

The second task is centered on evaluating the consistency of a generated answer with a predefined character persona. The prompt, akin to the first one, shifts focus towards assessing persona consistency. The differences between this prompt and the prompt before were marked in bold.

> *"The following conversation shows a chat counseling session between a client and a counselor. The context of the chat between these two people is online social counseling.* ***Your task is to evaluate whether the character described would write the generated response."***

The persona consistency is a direct evaluation criterion. An indirect way to evaluate this is to assess the flow of conversation, which may extend the persona description but still should follow a consistent problem case:

> *"Evaluation criteria:*
> ***Character consistency: The content of the answers must be consistent with the character description or expand on it in a meaningful and realistic way.***
> ***Flow of conversation: The continuation of the chat should be natural and appropriate to the character described.*** *Please keep in mind that this is a chat. Short, precise and direct answers can occur here."*

The score descriptions were also adjusted to task two:

> *"Score:*
> *0:* ***The generated answer is very realistic and is consistent with the character's description or expands on it in a meaningful way***
> *1:* ***The generated answer fits the character, but the person would probably express themselves differently based on the character description and previous history***
> *2:* ***The generated answer does not match the character. The person would definitely not express themselves in this way"***

After the description of the scores a character description is added. In order to prevent the model from rating the generated answer only on the course of the conversation, a sentence was added with a request to rate only the relation to the character description. The rest remains the same:

> *The evaluation is based on the evaluation criteria.*
> ***Character description:***
> ***{personality_condition}***
> *Course of the conversation:*
> *{history}*
> *Generated answer:*
> *{answer}*
> *Please structure your answer as JSON with the attributes Rating and Reason.* ***Please only evaluate the generated answer in relation to the character description and the course of the conversation.*** *Only output the JSON format:"*

5.2 Manual Evaluation by Humans

The manual evaluation process was carried out with five raters with an academic background in social sciences (hereinafter referred to as raters). It relies on the two tasks outlined in the previous section. A specialized web application was created for the rating which enabled evaluators to rate dialogues based on the specific criteria outlined in section 5.1. Evaluators can view dialogue histories, read task descriptions and generated answers, and score each task using a set of radio buttons. Additionally, for the second task, a persona description was provided on the right side for reference. To ensure comparability, all raters were provided with the same conversations.

5.3 Results

5.3.1 Results of Task 1

Table 2 provides a comparison of evaluation scores assigned by GPT-3.5 and GPT-4, alongside five individual raters. These scores are presented as percentages, reflecting the frequency with which each rater deemed the generated responses as coherent with the

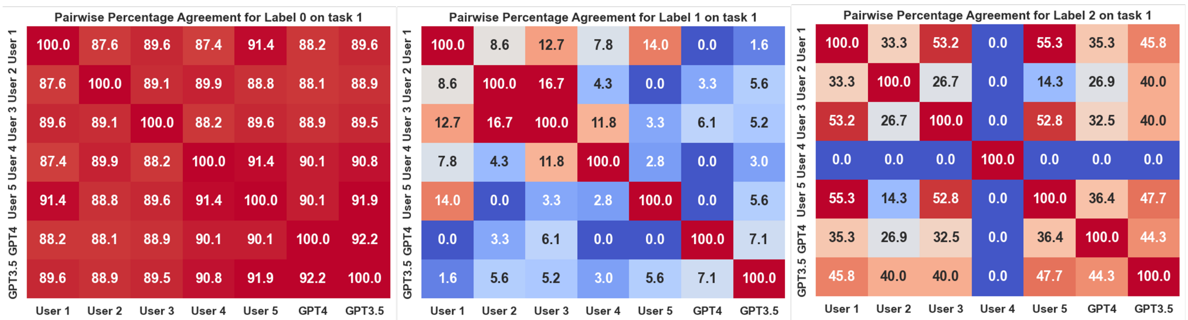


Figure 3: Pairwise percentage of agreement between all raters for each label on task 1 (conversation coherence).

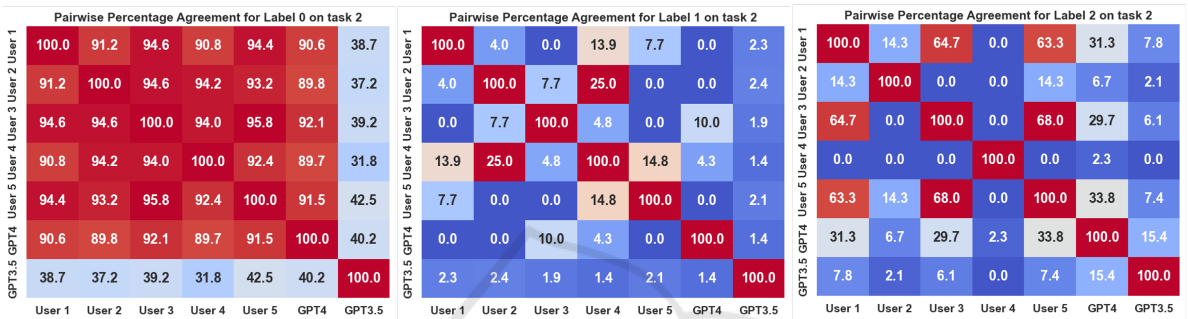


Figure 4: Pairwise percentage of agreement between all raters for each label on task 2 (persona consistency).

preceding discourse. Additionally, the table encapsulates the consensus assessment derived through a majority vote among the raters (1 through 5). The data vividly illustrates a general agreement among all parties — both human raters and AI — that at least 89% of the generated answers maintained coherence (Label 0) with the interview's prior context.

Table 2: Comparative Rating Percentages by GPT Versions and Raters for task 1.

Rater	Label 0	Label 1	Label 2
GPT-3.5	91.91	2.22	5.86
GPT-4	89.59	0.81	9.61
Rater 1	89.26	5.8	4.94
Rater 2	89.51	5.57	4.93
Rater 3	89.81	5.81	4.39
Rater 4	95.92	4.08	0
Rater 5	92.38	3.52	4.11
Majority (1-5)	92.01	3.54	4.45

Figure 3 shows the pairwise percentage agreement between all raters including the GPT models for each label on task 1 in form of a heatmap. Each cell in the heatmap represents the pairwise agreement between two raters (e.g., User 1 and User 2, or User 1 and GPT-3.5) for the respective label. We used a Jaccard-like metric to evaluate the agreement: Let $Resp_l(r)$ set of responses that rater r has rated with label l. The agreement $A_l(r_1,r_2)$ between two raters r_1 and r_2 with regard to label l is then calculated by:

$$A_l(r_1,r_2) = \frac{|Resp_l(r_1) \cap Resp_l(r_2)|}{|Resp_l(r_1) \cup Resp_l(r_2)|}$$

The diagonal cells, which are all 100%, represent the agreement of each rating entity with itself. It shows that when it comes to label 0 ("coherent") there are high levels of agreement among all entities, mostly above 87%. It indicates that the users and AI models tend to agree on the classification of label 0 on this task. The second heatmap for label 1 ("moderately coherent") shows very low agreement levels, indicated by cooler colors like blue and light red. Most values are below 20%, suggesting a high level of disagreement or variability in how label 1 is being classified. The agreement level on label 2 ("not coherent") ranges from 14.3 % to 55.3 %.

All in all, it can be said that there is little difference between the human raters and the GPT models. While raters often agree when a generated answer is coherent with the respective course of the conversation, they disagree about which course of the conversation is only moderately coherent and occasionally agree when the generated answer does not fit the course of the conversation at all.

5.3.2 Results of Task 2

Table 3 summarizes the percentage of each label assigned by GPT-3.5, GPT-4, the five human raters, and the majority decision among human raters. Similar to task 1, raters 1 to 5 and GPT-4 predominantly assigned label 0 (”consistent with the persona”). The exception here is GPT-3.5. This is because GPT-3.5 often takes the persona description very literally and argues that an upset persona responds directly with an emotional answer about the problem, which is not necessarily the case in a chat counseling session. For example the generated answer *“yes that’s okey”* on the conversation depicted at the end of section 4 was rated by GPT-3.5 with label 2 with the following reason:

> *“The generated answer does not match the character. Elke is very worried about her son and suspects that he is taking drugs. She would therefore probably not simply answer ’Yes, that’s okay’, but would rather talk about her worries or ask for specific help.”*

All other raters, including GPT-4, considered the generated answer to be consistent with the persona. GPT-4’s reasoning was as follows:

> *“The generated answer is coherent with the previous course of the conversation. The client agrees with the technical and organizational framework, which is a logical and appropriate next step in the conversation. There are no grammatical or content errors, and the flow of the conversation remains natural and coherent.”*

Table 3: Comparative Rating Percentages by GPT Versions and Raters for task 2.

Rater	Label 0	Label 1	Label 2
GPT-3.5	38.27	13.4	48.32
GPT-4	89.59	0.81	9.61
Rater 1	92.22	3.58	4.2
Rater 2	93.15	5.78	1.07
Rater 3	96.39	0.52	3.1
Rater 4	95.92	3.86	0.21
Rater 5	93.99	2.35	3.67
Majority (1-5)	94.34	2.53	3.13

Figure 4 shows the heat maps for task 2. The map on the left-side shows again a high percentage of agreement (mostly in the red shades indicating percentages from the high 80s to 100%) among raters 1 to 5 and GPT-4 for label 0, which means there is a general consensus that the answer fits the persona description well. The agreement with GPT-3.5 is significantly lower with percentages ranging from around 31.8% to 42.5%. This is in line with table 3.

In summary, it can be stated that the responses of the Vicuna model in the virtual client were assessed as both coherent and consistent in approx. 90% of cases. In the cases where this was not the case, the raters often disagreed.

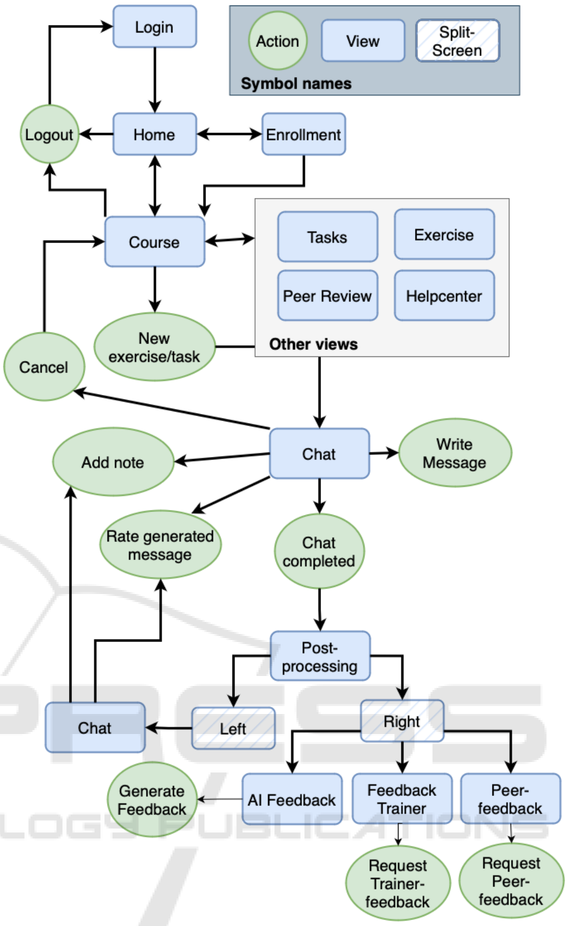


Figure 5: Userflow diagram of the learning platform.

6 LEARNING PLATFORM

The evaluated virtual client chatbot is the core of a comprehensive learning platform. In section 4 we already introduced the overall architecture and the interaction between the various components. This section focuses on the learning platform in the upper part of the figure 2. To get an overview of its functionality, figure 5 shows a userflow diagram. After login, course enrollment and course selection, which is common practice in a learning platform, an overview of the compulsory tasks and voluntary exercises completed so far with the virtual clients is displayed.

While a compulsory task involves a chat with a persona selected by the trainer and serves as an examination, the exercise serves as a voluntary opportunity to improve one’s chat counseling skills and as preparation for the compulsory task. In the latter case, the student can choose the persona from a number of personas pre-selected by the trainer. Optionally, a technical difficulty can also be configured, for example that the persona has internet problems.

During the chat session, a note field offers the opportunity to make notes for each message. This is intended to improve the student's ability to reflect. On the left-hand side, users have to possibility to reflect on the chat and rate the AI-generated messages (thumbs up/down). This helps to further improve the VirCo architecture in the future. Various options for requesting/generating feedback are displayed on the right-hand side:

- **Request Trainer Feedback.** To receive feedback from the trainer, a student can actively request feedback by clicking on a button. The trainer can configure for the course how much feedback he/she would like to give per student.
- **Request Peer Feedback.** To encourage student participation in providing feedback, our approach employs a coin-based incentive system. Initially, students are allocated a specific number of coins. Submitting feedback requires the expenditure of one coin, which is replenished upon providing feedback. This cycle aims to foster continuous engagement and contribution.
- **AI-Generated Feedback.** AI-generated feedback is another promising attempt to improve counseling skills. The advantage here is that the feedback is provided directly after the counseling session.

7 SUMMARY AND FUTURE WORK

In this study, we explore the application of large language models (LLMs) in the realm of education, specifically focusing on their potential for facilitating educational role-play to improve the training of online counselors. We introduce a comprehensive methodology designed to enable LLMs to simulate diverse personas. Our work includes the development of a novel architecture for a chatbot-based platform specifically tailored for educational role-play. This platform leverages the capabilities of LLMs to deliver interactive learning experiences and personalized feedback opportunities.

We also evaluated the persona consistency and the course of conversation coherence of the client chatbot. The input data for the natural language generation task was created by chat counseling role-plays where one person played the client and one person played the counselor. The evaluation showed that the answers generated by the model are realistic, generally correspond to the course of the interview and are consistent with the persona. However, it also shows that the human raters disagree about when a generated answer does not match the course of the conversation. It is therefore also not surprising that GPT-4 and GPT-3.5 do not agree with the human raters here either. Similar results were obtained when measuring persona consistency, with the difference that GPT-3.5 shows strong deviations from other raters and GPT-4.

Future improvements could concentrate on expanding the diversity and complexity of the personas and scenarios included in the dataset to cover a broader range of counseling situations. This expansion would demonstrate the system's adaptability and improve its utility as a practical training tool. Additionally, exploring how the virtual client responds to non-serious counselor messages could unveil limitations and inform necessary adjustments to its architecture, ensuring effective responses across a wider range of interaction types.

Another way to improve the virtual client is to compare different LLM models in this scenario. For this, a ranking task could be used instead of a rating task, as comparing different answers is often easier for both humans and language models than evaluating a single answer. The ranking can then be used as a preference dataset to further improve the virtual client through Direct Preference Optimization or Reinforcement Learning from Human Feedback.

We also described the feedback functionality of the learning platform, but this has not yet been evaluated. In addition, we want to delve deeper into the automatic generation of feedback and compare different feedback methods like the “Situation, Behavior, Impact”- or the “sandwich”-method with different LLM architectures. It is currently not possible to add further personas in the learning platform front end. We plan to add such functionality, whereby the course trainer is asked specific questions and a persona is then created on this basis. This will make it very easy to use the learning platform in areas other than online counseling.

In-depth research on the long-term impact of training with the platform will provide valuable insights into its efficacy, benefits, and limitations. An additional area of interest is the exploration of user interaction with the system, with a particular focus on user experience and interface design. Optimizing these aspects could significantly enhance engagement and the overall effectiveness of training sessions. Finally, given the rapid development of LLMs, ongoing updates and comparisons with new models will ensure that the platform remains at the forefront of technology, continuously improving its realism and effectiveness as a training tool.

REFERENCES

Bharti, R. K. (2023). Contribution of Medical Education through Role Playing in Community Health Promotion: A Review. *Iranian Journal of Public Health*.

Boucher, E. M., Harake, N. R., Ward, H. E., Stoeckl, S. E., Vargas, J., Minkel, J., Parks, A. C., and Zilca, R. (2021). Artificially intelligent chatbots in digital mental health interventions: A review. *Expert Review of Medical Devices*, 18(sup1):37–49.

Cao, J., Tanana, M., Imel, Z. E., Poitras, E., Atkins, D. C., and Srikumar, V. (2019). Observing Dialogue in Therapy: Categorizing and Forecasting Behavioral Codes. arXiv:1907.00326.

Center for Behavioral Health Statistics and Quality (2023). 2023 National Survey on Drug Use and Health (NSDUH): Prescription drug images for the 2023 questionnaires. Technical report, Substance Abuse and Mental Health Services Administration.

Chang, Y., Wang, X., Wang, J., Wu, Y., Yang, L., Zhu, K., Chen, H., Yi, X., Wang, C., Wang, Y., Ye, W., Zhang, Y., Chang, Y., Yu, P. S., Yang, Q., and Xie, X. (2023). A survey on evaluation of large language models. arXiv:2307.03109.

Chen, S., Wu, M., Zhu, K. Q., Lan, K., Zhang, Z., and Cui, L. (2023). LLM-empowered Chatbots for Psychiatrist and Patient Simulation: Application and Evaluation. arXiv:2305.13614.

Chiang, W.-L., Li, Z., Lin, Z., Sheng, Y., Wu, Z., Zhang, H., Zheng, L., Zhuang, S., Zhuang, Y., Gonzalez, J. E., Stoica, I., and Xing, E. P. (2023). Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. https://lmsys.org/blog/2023-03-30-vicuna/.

DeMasi, O., Hearst, M., and Recht, B. (2019). Towards augmenting crisis counselor training by improving message retrieval. In *Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology*, pages 1—11, Minneapolis, Minnesota. Association for Computational Linguistics.

DeMasi, O., Li, Y., and Yu, Z. (2020). A multi-persona chatbot for hotline counselor training. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3623—3636, Online. Association for Computational Linguistics.

Kerr, A., Strawbridge, J., Kelleher, C., Barlow, J., Sullivan, C., and Pawlikowska, T. (2021). A realist evaluation exploring simulated patient role-play in pharmacist undergraduate communication training. *BMC Medical Education*, 21(1):325.

Kong, A., Zhao, S., Chen, H., Li, Q., Qin, Y., Sun, R., and Zhou, X. (2023). Better zero-shot reasoning with role-play prompting. arXiv:2308.07702.

Lee, G., Hartmann, V., Park, J., Papailiopoulos, D., and Lee, K. (2023). Prompted LLMs as Chatbot Modules for Long Open-domain Conversation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 4536–4554, Toronto, Canada. Association for Computational Linguistics.

Lin, Y.-T. and Chen, Y.-N. (2023). Llm-eval: Unified multi-dimensional automatic evaluation for open-domain conversations with large language models. arXiv:2305.13711.

Liu, J., Symons, C., and Vatsavai, R. R. (2022). Persona-Based Conversational AI: State of the Art and Challenges. In *2022 IEEE International Conference on Data Mining Workshops (ICDMW)*, pages 993–1001. arXiv.

Liu, Y., Iter, D., Xu, Y., Wang, S., Xu, R., and Zhu, C. (2023). G-eval: Nlg evaluation using gpt-4 with better human alignment. arXiv:2303.16634.

Mianehsaz, E., Saber, A., Tabatabaee, S. M., and Faghihi, A. (2023). Teaching Medical Professionalism with a Scenario-based Approach Using Role-Playing and Reflection: A Step towards Promoting Integration of Theory and Practice. *Journal of Advances in Medical Education and Professionalism*, 11(1).

Park, S., Kim, D., and Oh, A. (2019). Conversation model fine-tuning for classifying client utterances in counseling dialogues. In *Proceedings of the 2019 Conference of the North*, pages 1448–1459. Association for Computational Linguistics.

Sai Sailesh Kumar Goothy, Sirisha D, and Movva Swathi (2019). Effectiveness of Academic Role-play in Understanding the Clinical Concepts in Medical Education. *International Journal of Research in Pharmaceutical Sciences*, 10(2):1205–1208.

Shanahan, M., McDonell, K., and Reynolds, L. (2023). Role-Play with Large Language Models. arXiv:2305.16367.

Tanana, M., Soma, C., Srikumar, V., Atkins, D., and Imel, Z. (2019). Development and evaluation of clientbot: Patient-like conversational agent to train basic counseling skills. In *Journal of Medical Internet Research*, volume 21(7), Online. Journal of medical Internet research.

Wang, J., Liang, Y., Meng, F., Sun, Z., Shi, H., Li, Z., Xu, J., Qu, J., and Zhou, J. (2023). Is chatgpt a good nlg evaluator? a preliminary study. arXiv:2303.04048.

Weizenbaum, J. (1966). Eliza—a computer program for the study of natural language communication between man and machine. In *Communications of the ACM*, volume 9(1), pages 36–45.

Xu, B. and Zhuang, Z. (2022). Survey on psychotherapy chatbots. *Concurrency and Computation: Practice and Experience*, 34(7):e6170.