Transferapy - A therapeutical language app

Max Feucht 2742061
Ruby Uwadiegwu 2752363

NLP Assignment 3, 10th of May 2023

1 Introduction

Language models use probability distributions over sequences of words to analyze and generate human language, making them a powerful tool for a lot of applications. With recent advances in language models that use Deep Learning techniques (large language models - LLMs) to generate very realistic linguistical utterances, LLMs are increasingly used in the sensible context of psychotherapy. AI-powered therapeutic chatbots such Woebot, Wysa, Pyx Health, X2ai, or clare&me [2][3][1][4][5], are just a few examples of how language models are already used to improve the mental health of people. A severe limitation of current applications, however, is the missing link between traditional in-person psychotherapy and digital therapy assistants. We aim to bridge this gap by building *Transferapy*, an LLM-based therapy chatbot that works in cooperation with the human psychotherapist in order to augment the therapy provided both in-session, as well as in-between physical sessions.

As of today, there are platforms available for clients to share and discuss their feelings between two therapy sessions. Using such inter-sessional language model interventions is successful for other aspects of development, such as speech development in children [7]. However, clients may still feel a disconnection between their therapist and the inter-sessional intervention, possibly because the intervention may not fully replicate the interpersonal nature of therapy. If the intervention is not directly linked to the therapy sessions or if the therapist is not involved in the use of the intervention, clients may feel disconnected or disjointed, leading to less engagement with the intervention and a weaker therapeutic alliance with their therapist [8].

Our Web App idea to overcome this limitation is simple: to provide a platform on which the client can "speak" to a LLM-based chatbot, that is connected to the human therapist and has access to the content of the in-person therapy sessions. The chatbot thus serves as a communication instance for clients between in-person sessions. What makes our *Transferapy* stand out is that it has access to the same knowledge as the human therapist, and likewise provides the human therapist with the information of the in-between chat communication. The former is achieved by feeding summarized transcripts of the therapy session to the bot after each session to "update" it, and the latter is achieved by

providing the therapist with a summarized transcript of the chat content generated between sessions. Additionally, we want to make our chatbot feel more natural by allowing it to adopt the conversation style of the human therapist; each chatbot would thus be unique for each client-therapist pair. By combining these features, we aim to reduce the feeling of disconnection or disjointedness in clients and increase the therapeutic alliance with their therapist and ultimately the therapeutic outcome.

Importantly, this Web App does not aim to therapize or give critical advice to clients; its role is more thought of as an outlet for clients when they need support in-between sessions. Therefore, the model has to be waterproof to circumvent sensible utterances about diagnoses, recommendations for actions, interpretations, or trauma-inducing, insensitive, and, harmful or stigmatizing questions.

Regarding the linguistic meaningfulness of the app, the LLM must be able to perform various language-related tasks such as sentiment analysis, and must also be able to provide sound answers in the style of the therapist.

The LLM used to build the prototype for *Transferapy* in the assignment at hand is GPT-3.5. When building the "final" product, not the raw model would be used, but a fine-tuned version of it. However, fine-tuning the model was not within the scope of this assignment.

2 How it works

Transferapy offers clients a simple chat window, in which the user can type their message, which appears in blue on the right side. GPT-3.5 is then queried using this message to generate a response to the message, which is subsequently displayed on the left side (function `get_bot_response`, line $19 - 39$ in app.py). GPT-3.5 is queried with a temperature of 0.2 to provide relatively standardized and "non-experimental" answers, and without a frequency or presence penalty (using openai.Completion.create). A technical hurdle in a conversational setting was the fact that with every query to the OpenAI API, GPT-3.5 is queried "from scratch", without access to the prompt it has been queried with before. Thus, in order to provide a seamless conversation experience for the user, the model had to be provided with the conversation history for every new query. The first prompt could go without context, but the following queries consist of the conversation history and the present query. By doing so, the model has access to the contextual information of the conversation history, but can also be provided with other background information, e.g., by adding a session transcript summary to the conversation history. In the code at hand, this was achieved through defining the global variable `conversation` in the function `initialize_conversation` in lines $66 - 73$ in app.py. Thus, every query, every model output, but also a previous session's transcript is added to `conversation`, and not the query itself, but `conversation` is fed to GPT-3.5. An example of how `conversation` looks like under the hood is displayed in Appendix B. Using this technique makes the input queries longer, the longer the conversation is

held, and the more contextual information is provided (such as in-person session transcripts or medical information).

The "client view" of the Transferapy web app is exemplary displayed in Figure 1. An important feature (which is not visible in the client view of *Transferapy*) is an automatic update of the model, as soon as a new session is initialized. If a first chat message is sent after a therapy session, *Transferapy* automatically a summary of the transcript of the previous session to GPT-3.5 (function `update_bot`, line 45 - 62 in app.py, called automatically in line $28-29$ in app.py). By doing so, the model is provided with an abstracted context of the client's situation and the therapist's knowledge thereof. This is done automatically by querying a database that contains therapy transcripts and querying a new instance of GPT-3.5 with the instruction "*Summarize this meeting transcript in bullet points: "* + the transcript string (function `get_transcript_summary`, line $98 - 128$ in app.py). The resulting summary is then fed to the instance of GPT-3.5 that is holding the conversation with the client, to provide it with context about the session. This enables to query *Transferapy* with questions about the session content, without having to repeat the content to the chatbot. An example of this behavior is shown in Figure 1 - the corresponging example transcript with its summary is given in Appendix A, and the question displayed in Figure 1 refers to the content of this summary. Note that we use a simple .csv file here as a replacement for a real database. The .csv file is included in the folder and the functionality can, thus, be tested.

The following screenshots demonstrate, how *Transferapy* is considering the context in it's answers and is able to accurately answer a question about the in-person session, without having "explicit" access to that information through the chat. The model is provided with that information through the automatic update described in the "How it works" section. Importantly, the question was not chosen based on the contents of the summary, but independently and without knowledge, of how the model summarized the in-person therapy session.

The web app is also available for the therapist. In this format, the web app offers more functionalities than in the client view. Specifically, the therapist can not only access the content of the chat message-by-message but can additionally update the LLM manually through the "Update" button on top of the chat window. They can retrieve the summarized content of the conversation between the client and the chatbot through the "Summarize" button - when clicking the button, the summary appears as a speech bubble in the chat and can be copied. The Update function works in the same way as described above (function `update_bot`), but can also be triggered manually. This can be useful if the model requires additional information outside of the therapy transcript, such as medication information or information from other practitioners. Currently, the "Update" button only updates the bot with the most recent transcript; in future updates to the app, this button should allow one to specifically select what information the model shall be updated with. Furthermore, the "Summarize" function is vital for the therapist, as it provides them with a quick overview of important information that was disclosed during the client-bot conversation. Using the "Update" and the "Summary" function, the chatbot and the therapist

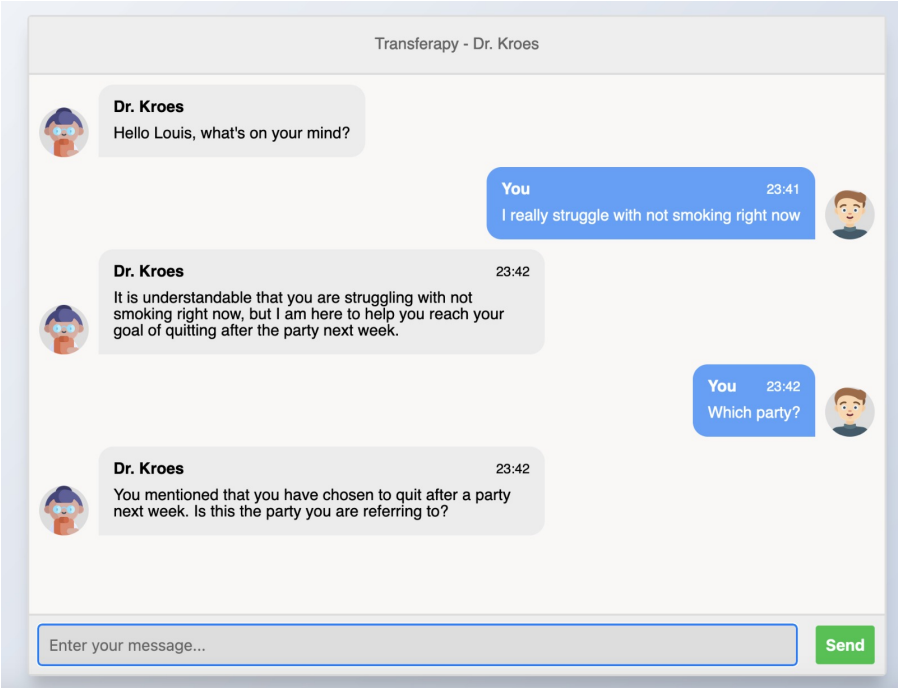

Figure 1: Screenshot of how *Transferapy* demonstrated context knowledge and answers accurately to a question about an in-person session, due to the automatic update performed when launching a new chat (represented in the client view)

can exchange information both ways. An example of the therapist view is given in 2.

For the third feature of adapting the conversation style of the chatbot to that of the therapist to work, we plan to fine-tune the model with historic transcripts of the therapist's utterances in a therapeutical context. As fine-tuning GPT-3.5 exceeds the scope of this assignment (next to other limitations outlined in the Limitations section), we have not implemented this feature yet.

For testing the script, please run `python app.py` in the terminal when in the correct folder, and then run `http://127.0.0.1:5000/` in the browser.

3 Limitations

Although *Transferapy* is a promising tool to augment traditional psychotherapy, there are a few limitations.

A major technical hurdle to achieving a well-performing therapy chatbot, that has access to large amounts of contextual data, is the current state of the

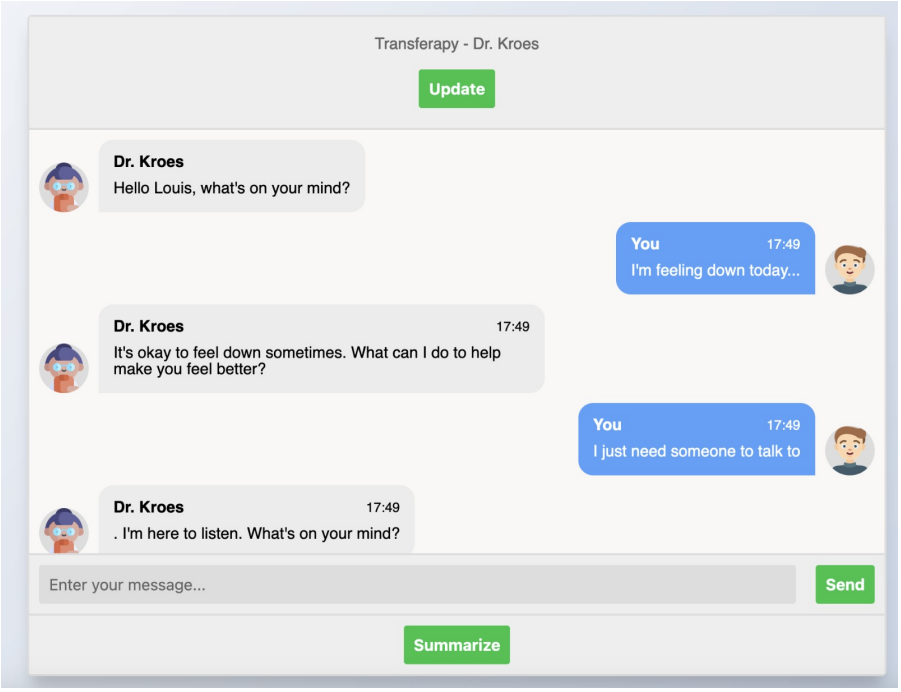

Figure 2: Screenshot of how *Transferapy* looks like in the therapist view

art of LLMs such as GPT-3.5, and the number of tokens they can process in a single input-output pair. Presently, the GPT-3.5 model has a maximum token limit of 4096, which can together be used for processing a single model input and generating the corresponding output. Although 4096 tokes equal a text amount of around 8000 words, this limit can be reached fast when considering how GPT-3.5 has to be queried in order to provide it with contextual information (described in "How it works", i.e., querying GPT-3.5 with all queries added to the `conversation` variable). As the input query grows the longer the conversation and the more contextual information provided, the 4096 token limit poses a challenge, especially in high-usage settings. If an input surpasses the token limit, it must be shortened or condensed, potentially impacting the context, content, and semantic meaning of the text. Overcoming this token limit is crucial for therapy chatbots to approach the same level of "knowledge" as human therapists, allowing language models to access comprehensive contextual information and deliver a more authentic conversational experience.

A further limitation pertains to the ability to adapt a therapist's conversation style, to increase the feeling of familiarity between the client and the chatbot. Achieving such behavior is quite computationally difficult due to a number of

reasons. Firstly, it requires robust training data, which means that earlier conversations between the therapist and client(s) must be transcribed and analyzed for the Web App purpose. This is quite challenging as acquiring transcribed or recorded client-therapist conversations poses legal and ethical concerns. Not only do the clients have to consent to these recordings, but also third-party transcribers should be chosen cautiously, and the model on which the data is trained must be very rigorously inspected for harmful output. Secondly, training models to accurately replicate a therapist's style can be computationally intensive and may require substantial computational resources, like fine-tuning the LLM generating the responses. Although OpenAI offers fine-tuning of models on their servers as a paid service, doing this for every therapist-patient pair can become a significant financial burden. Thirdly, real-time responsiveness in the style of the therapist can also be difficult to implement due to the complexity of the style of the therapist. And, because this is a platform for multiple clients and therapists, it can be difficult to store all of this information. Lastly, the client must consent to the conversations being recorded, stored, and used for further development of the model. Clients who do not consent, may not be able to use all the features of *Transferapy* or worse, not be able to use *Trasferapy* at all.

4 Future improvements

Further improvements can be made regarding our Web App.

Firstly, there should be a functionality that notifies the therapist as soon as worrisome or harmful content is detected in user input, e.g., mention of imminent danger or self-harm. This functionality is relevant due to a number of reasons:

Safety and risk management: By immediately informing the therapist about the possibility of immediate danger or self-harm, there is a direct focus on identifying potential risks to the client's safety and well-being. Appropriate measures can be made to mitigate those risks.

Immediate Support and Intervention: Similarly, this immediate action allows the therapist to provide support and intervention on time. The therapist can promptly respond to the client's needs.

Ethical and Legal Responsibilities: Therapists have an ethical and legal responsibility to ensure the safety and well-being of clients. This mechanism of alerting the therapist guarantees that appropriate measures are taken to protect the client.

Additionally, it is necessary to incorporate a filtering functionality into the model's output to ensure that *Transferapy* does not engage in actual therapizing. Again, the purpose of *Transferapy* is to provide an outlet for the client in between in-person sessions with a therapist, and not to provide therapizing advice itself. A filtering mechanism would assess whether the content includes discussions of diagnoses, inquiries about diagnoses, dubious or impractical action recommendations, as well as harmful or stigmatizing questions. We strongly

recommend collaborating with subject matter experts to develop such a filtering mechanism, although it was not feasible within the scope of this course.

Lastly, as already mentioned above, incorporating the feature of mimicking a client's therapist's conversation style is a vital improvement, in which we see great potential to increase the effectiveness and acceptance of the application.

In conclusion, developing *Transferapy* requires immediate content notification for the therapist, safety measures for the client, ethical considerations, client consent for recording, as well as mimicking therapists' conversation styles, all to create a supportive and safely functioning platform. Although there are still necessary developments needed to improve *Transferapy* for real-life use in the future, the present assignment provided a first outline of how such a service could work for both clients and therapists.

5 Contribution

The authors of this assignment divided the work equally and contributed equally. While Max did most of the coding, Ruby did most of the writing. However, both contributed to each other's work.

Appendix A

This Appendix demonstrates how knowledge transfer from an in-person therapy session to the chatbot can work. First, we provide an exemplary session transcript that contains all the "knowledge" from the in-person session. In the next step, this transcript is summarized and added to the `conversation` variable, to provide the model with the context of the session. Lastly, we demonstrate how this works in practice: the client asks something about the last session, and the model answers correctly, without being provided with that information explicitly in the chat.

5.1 Written-out Transcript[6]

Therapist: Good morning. It's really nice that you've come along to see me to talk about stopping smoking. Client: Good morning. Therapist: What have you got in mind, what do you want from us? Client: Well, I need some medication if you can offer me some. Therapist: Absolutely, we're in the right place for that. Client: And, uh, whatever else you can do just to get me to go smoke-free. Therapist: Okay. Well, just remind me again why it is you want to stop. Client: It's, uh, just these days– I think these days I'm-I'm finding that, um, I'm smoking 20, 30 a day, I'm getting sick of it. It's, you know, I've seen my friends have stopped smoking too and it's kind of they're looking at me weird these days like I'm the only one left who's still smoking. Therapist: Oh, good, and so feeling under pressure here. Client: Especially the other night I was at the restaurant, it was a bit– you know, it was a bit embarrassing I'm the only one walk walking out. Therapist: And so that-that can't feel a little uncomfortable being by yourself. So that sounds like one of the reasons you have to quit is it's a little getting smelly and it's getting embarrassing. Client: Mm-hmm. Therapist: Any other reasons? Client: Um, it's expensive these days. Therapist: How much does it actually cost you a week? Client: It's-it's about– Um, I'll buy and go through a pack a day, so it's almost £7. Therapist: £7 a week or a day? Client: £7 a day today, so- Therapist: So your pack a day— Client: -you're talking about £40, uh— Therapist: £50 a week. Client: £50 a week. Therapist: Even more mess, you see. Client: Yeah. Therapist: Okay. So we're talking about spending £50 a week- Client: Yeah. Therapist: -being one of the few left smoking in your crowd, and not liking the smell. That sounds like three reasons. Client: Mm-hmm. Therapist: Do you think that's enough to make you say goodbye to cigarettes? Client: Yeah. Therapist: Great, okay. So have you got a date in mind when you'd like to stop smoking? Client: Well, what you do– What do you suggest? Therapist: Generally, it works well if you find a time when you're not under a huge amount of other stresses and strains. Client: Mm-hmm. Therapist: What are the main reasons why you actually smoke? What do you enjoy most about it? Client: Um, it's just entertainment. Something, um, when I bored, it's got– I've got– I smoke- Therapist: Mm-hmm. Client: -or sometimes I go out, but as I said, I'm the only one kind of outside. Therapist: Okay. What do you think you'll miss most about it? Client: The– Um, it takes away a lot of my

stresses. Therapist: Mm-hmm, what sort of stresses? Client: Just, uh, when I- when I go down, uh, for-for a break at work. Therapist: Mm-hmm. Client: I never go down, I just need a break away from it all, just shut off. So, uh, I-I really enjoyed that. The cigarette just helps me calm down. Therapist: Yeah. And what is your favorite cigarette of the day? Client: It's definitely the one at the end of the evening. Therapist: Okay, so the end of the day- Client: Yeah. Therapist: -the thing that helps you just unwind, put an end to the work section and move into your own time. Client: Yeah. Therapist: Okay. So after you've had the quit date, what do you think you might do when it it's the first day at work and it's time to go home if you're not gonna have a cigarette? Client: I guess, um, I'll just- I'll just have, you know, some fresh air as well, but this time without a cigarette in my hand but to try and give myself regular breaks maybe, I don't know. Therapist: Okay. Client: It's one of those tough things. Therapist: Have you managed to stop smoking in the past? Client: I've managed for a couple of weeks. Therapist: Okay. And what got you back? Client: It was, um– To be honest, it was stress from an argument with my partner I remember. Therapist: So something just happened- Client: Yeah. Therapist: -so it took you back into it. All right. Let me just make sure I've understood what's going on here. You really want to stop smoking because you're kind of the last guy in your group still smoking, it costs a lot of money, and you don't really like the smell. Client: Mm-hmm. Therapist: The main reason that you smoke is because, at the end of the day, it gives you a kind of relaxation, and if there's quite a little stress that's going on which you're using cigarettes to help you cope with. Client: Mm-hmm. Therapist: What you'd like to do is stop and you'd like to sort of like choose a time to quit when it's going to be there's not as much stress around as possible. When do you think would be a good time for you to choose to be your quit date? Client: I guess, um, at the weekend I've got a party coming up. Therapist: Okay. Client: So, uh, maybe Monday morning. Therapist: So Monday morning after the party will be a good day, and that's next week. Okay. Now, one of the things that we do in the pharmacy is recommend that you use medication to help you. You know this increases your chance of success and it makes it more likely that can manage any cravings that come along. A lot of people like using nicotine replacement therapy which comes in a variety of formats like patches, gum, you've probably seen people using any of these products. And other options are tablets which you can get from the GP. Client: Mm-hmm. Therapist: So, yeah, again– Um, so have you something in mind that you might like to use to help you through managing the cravings? Client: Yeah. I'm thinking, um, one of these patches. Therapist: Okay. So a patch is something you've already seen, and- Client: Yeah. Therapist: -and, again, I'll get some out and I'll show you what they look like. This course of medication lasts for 10 to 12 weeks. And we strongly recommend that people stay on the whole course of medication and that that gives you the best chance of success. Stopping smoking is not easy but it's possible and with the type of support that we offer here in the pharmacy, you absolutely make your chance of success four times greater than trying on your own, so, that's why it's really good that you've come to see me. Now, what you said is you really want

to do it and that next week sounds like a good day for you to start because the party is over and done with. Client: Mm-hmm. Therapist: You've got the time at the end of the day when you finish work, which might be the first time you've gone without a cigarette. So, what I'd like to recommend is that you use two products perhaps and I'm thinking of you're having a patch and then suggesting that you might use something that you can top up with. Often, a person who smokes knows exactly what it feels like when they're craving for a cigarette. What does it feel like to you, when you can't have a cigarette and you want one? Client: I just get irritable. Therapist: Okay. Client: Yeah. Therapist: And that's exactly one of the symptoms of nicotine withdrawal that I'm looking to stop happening to you. Client: Hmm. Therapist: So, I would strongly recommend that you have the patch. I'll talk to you about how to use that and that we have another product that I leave you to use as when any of those feelings of irritation are coming up, particularly in that first week. Client: Okay. Therapist: So, you're happy with the date for [unintelligible 00:06:16]? Client: Yeah, absolutely. Therapist: Okay and would you like to come back and see me again? Client: Well, definitely. Therapist: And what would you like me to do if you don't come back? Client: A nice phone call. Therapist: I'd be very happy to do that.

5.2 Summary of the same transcript, generated by Transferapy

The following text shows the summary the model is provided with, to learn the context of the in-person therapy session:

- Client wants to stop smoking and is looking for medication and other support
- Reasons for wanting to stop include feeling like the only one still smoking in their group, cost, and smell
- Smoking is used as a form of relaxation and to cope with stress
- Client has managed to stop for a couple of weeks in the past, but was taken back into it due to an argument with their partner
- Client has chosen to quit on Monday morning after a party
- Therapist recommends using a patch and another product to manage cravings
- Client is happy with the date and will come back for another visit or receive a phone call

5.3 Demonstration of knowledge transfer

The following screenshots demonstrate, how *Transferapy* is considering the context in it's answers and is able to accurately answer a question about the in-person session, without having "explicit" access to that information through the chat. The model is provided with that information through the automatic update described in the "How it works" section. Importantly, the question was not chosen based on the contents of the summary, but independently and without knowledge, of how the model summarized the in-person therapy session.

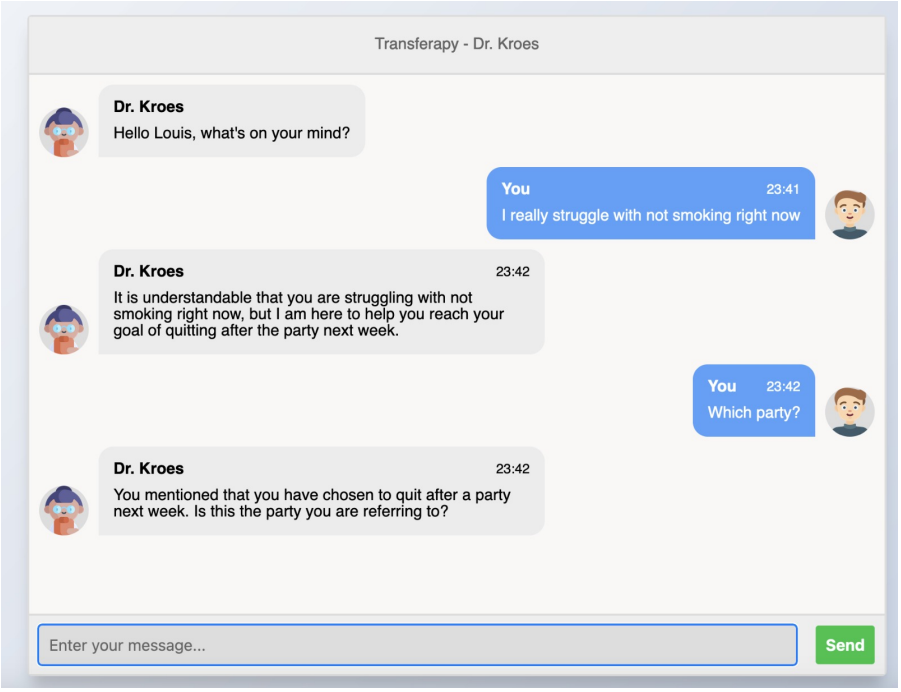

Figure 3: Screenshot of how *Transferapy* demonstrated context knowledge and answers accurately to a question about an in-person session, due to the automatic update performed when launching a new chat

Appendix B

Here is displayed, what the `conversation` variable looks like "under the hood". Again, not a single query, but the whole conversation up to this point is passed to GPT-3.5 in each additional query.

""" [Input: SUMMARY OF PREVIOUS SESSION]

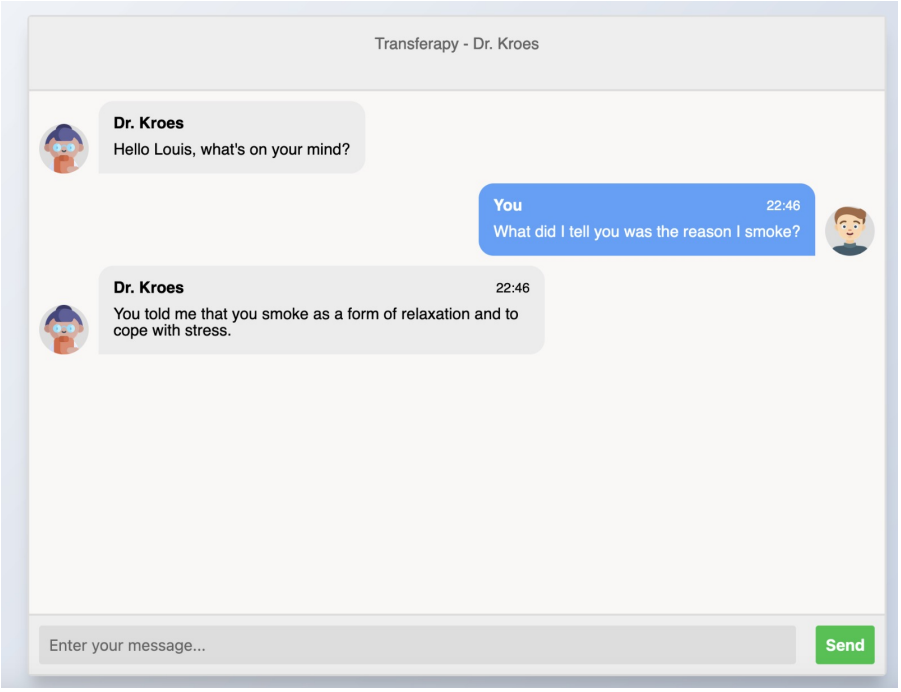

Figure 4: Screenshot of how *Transferapy* answers accurately to a question about an in-person session, due to the automatic update performed when launching a new chat

Input: I'm feeling down today...
Output: It sounds like you're having a tough day. Is there anything I can do to help?
Input:I just need someone to talk to
Output: I'm here for you. What's on your mind?
Input: I feel like I'm in need of a real friend
Output: That's understandable. It can be hard to find true friends. Is there someone you can reach out to? """

For the next message, GPT-3.5 thus not only receives the next query, but the following string + the new query, as in:

""" [Input: SUMMARY OF PREVIOUS SESSION]
Input: I'm feeling down today...
Output: It sounds like you're having a tough day. Is there anything I can do to help?
Input:I just need someone to talk to
Output: I'm here for you. What's on your mind?

Input: I feel like I'm in need of a real friend
Output: That's understandable. It can be hard to find true friends. Is there someone you can reach out to?
[Input: NEXT QUERY] """

References

[1] May 2023. URL: https://www.pyxhealth.com/.

[2] URL: https://play.google.com/store/apps/details?id=com.woebot&pli=1.

[3] URL: https://www.wysa.com/.

[4] URL: https://www.x2ai.com/individuals.

[5] URL: https://www.clareandme.com/.

[6] URL: https://github.com/uccollab/annomi.

[7] Oscar Saz et al. "Tools and Technologies for Computer-Aided Speech and Language Therapy". In: *Speech Communication* 51.10 (2009). Spoken Language Technology for Education, pp. 948–967. ISSN: 0167-6393. DOI: https://doi.org/10.1016/j.specom.2009.04.006. URL: https://www.sciencedirect.com/science/article/pii/S0167639309000661.

[8] Madalina Sucala et al. "The therapeutic relationship in e-therapy for mental health: a systematic review". In: *Journal of medical Internet research* 14.4 (2012), e110.