"You tell me": A Dataset of GPT-4-Based Behaviour Change Support Conversations

Selina Meyer
selina.meyer@ur.de
Regensburg University
Regensburg, Germany

David Elsweiler
david.elsweiler@co.uk
Regensburg University
Regensburg, Germany

ABSTRACT

Conversational agents are increasingly used to address emotional needs on top of information needs. One use case of increasing interest are counselling-style mental health and behaviour change interventions, with large language model (LLM)-based approaches becoming more popular. Research in this context so far has been largely system-focused, foregoing the aspect of user behaviour and the impact this can have on LLM-generated texts. To address this issue, we share a dataset containing text-based user interactions related to behaviour change with two GPT-4-based conversational agents collected in a preregistered user study. This dataset includes conversation data, user language analysis, perception measures, and user feedback for LLM-generated turns, and can offer valuable insights to inform the design of such systems based on real interactions.

CCS CONCEPTS

• **Human-centered computing** → **Natural language interfaces**; **Human computer interaction (HCI)**; • **Computing methodologies** → **Language resources**.

KEYWORDS

conversational agents, behaviour change, large language models, dialogue, information behaviour

ACM Reference Format:
Selina Meyer and David Elsweiler. 2024. "You tell me": A Dataset of GPT-4-Based Behaviour Change Support Conversations. In *Proceedings of the 2024 ACM SIGIR Conference on Human Information Interaction and Retrieval (CHIIR '24), March 10–14, 2024, Sheffield, United Kingdom.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3627508.3638330

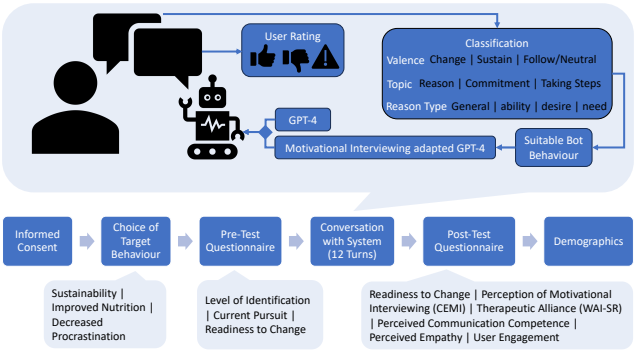

Figure 1: Each user interacts with one of two systems, where one system is prompted to adhere to Motivational Interviewing principles. Users interact with the systems for 12 turns. User turns are classified with respect to their implications regarding motivation for behaviour change. Each GPT-generated bot turn is rated as helpful, unhelpful, or harmful by the user, with an optional rating explanation.

1 INTRODUCTION

Chat interactions with conversational agents (CAs) are often studied in task-oriented domains, such as customer service [11], e-commerce [37, 38], or cooking [3, 15, 33]. In these scenarios, the CA and the user work together to solve a clear information need. However, information systems and other forms of online interaction are not only utilised to address information needs, but are also popular for leisure, or purely hedonistic purposes [12] or to address emotional or social needs [42]. Past research has highlighted the importance of meeting emotional needs when designing conversational agents, especially in sensitive contexts [22, 27, 41].

Consequently, there has been an increased interest in social influence dialogue systems, systems that automate behaviour and health interventions and are focused on addressing emotional needs more than information needs in recent years [7]. One conversational strategy that has been explored in this context is Motivational Interviewing (MI), a therapy approach aimed at increasing a person's motivation to change through supportive, non-confrontational conversation [31]. So far, conversational systems in this context have predominantly employed rule- or retrieval-based strategies [6, 16, 35, 39, 43, 44, 54]. However, rule-based conversational agents are often restrictive and fail to depict the same flexibility as conversations with human counsellors [1].

Given the high flexibility of large language models (LLMs) and their ability to generate human-like language, they are increasingly considered as tools to generate texts for sensitive use cases, such as mental health and counselling [1, 46–48]. This bears several advantages, but also many safety hazards [4]. Consequently, research on the topic is currently mainly limited to studying LLMs capabilities to perform specific behaviours in isolation – both of a longer conversation and of the user [1, 2, 32, 48]. In many such cases, existing datasets (i.e. EmpatheticDialogues) are used both to synthesise new conversational turns by the CA, and to evaluate the goodness

CHIIR '24, March 10–14, 2024, Sheffield, United Kingdom

ACM ISBN 979-8-4007-0434-5/24/03.
https://doi.org/10.1145/3627508.3638330

System Prompt:
You are a counsellor and help the user with the goal [target behaviour]. Never talk about yourself in the following, and concentrate fully on the user. Do not give active tips. If the user is talking about another topic, don't respond and lead them back to the [target behaviour]. Always speak in the second-person singular. You use the conversational strategy [action]. Definition of [action]: [description] Keep your answer as short as possible.
Description for action Reframe:
A reframe is a reflection that highlights a different perspective in the client's statement. It is particularly useful for defusing sustain talk. A therapist can listen to sustain talk and reframe the statement into a neutral statement, change talk or an affirmation.

Table 1: System prompt passed to GPT-4 in the MI-adapted condition, with description of the sample action *Reframe*

of generations through similarity-based metrics or human empathy judgements [21].

This system-oriented focus has led to a disregard of the user-side and a lack of resources and evaluations for studying real user interactions with LLM-based agents [52]. This is explored more frequently in the context of rule-based and retrieval-based conversational agents for counselling. For instance, He et al. [16] compared a rule-based MI-chatbot with a non-MI adherent conversational agent regarding user engagement, perceived empathy and therapeutic alliance. Given the new capabilities LLMs add to the picture of conversational AI, previous studies regarding user expectations for and behaviour towards CAs might not translate to these new technologies. Therefore, while researchers currently focus heavily on controlling specific LLM behaviours, one crucial aspect of CA-user interactions is overlooked: the unpredictability of user utterances.

To address this, we present a dataset of user interactions about three target behaviours (healthy nutrition, less procrastination, increased sustainability) with two GPT-4-based conversational agents, where one system is prompted to adhere to MI principles, and the other is not. Each conversation spans 12 turns, of which 5 turns are generated by GPT-4. The beginning and end of each conversation are rule-based and reflect different phases of an MI-session. This dataset can drive the exploration of user behaviour when interacting with LLM-based conversational agents in the context of social influence dialogue systems and help shed light on the controllability of such systems within the context of multi-turn conversations. The insights gained in the analysis of the dataset have the potential to inform the user-oriented design of such systems in the future.

2 DATASET

The dataset stems from a preregistered online user study aimed at evaluating the controllability and efficacy of GPT-4-based motivational behaviour change support [1]. Overall, the dataset consists of 2149 conversational turns, collected in 185 chats with 164 study participants. The complete dataset can be found on OSF. In this section, we describe the study setup and the conversational system and provide information about the study participants and post-processing steps.

2.1 Study Setup

At the beginning of the study, participants were provided with three personas to choose from, each focusing on a different target behaviour (healthier eating, sustainable living, less procrastination). These target behaviours were chosen, because they represent common nonmedical behaviour change goals [13, 14, 34, 40]. Participants were asked to choose the situation they identified with the most and put themselves in the persona's shoes during their interaction with the conversational system. This level of abstraction was added to avoid any potential harms interacting with the system might induce. However, in post-processing it became apparent that participants struggled to adhere to the level of abstraction, as most participants seemed to interact with the system from their own perspective.

Before the start of the conversation, participants were asked to what extent they identify with the chosen situation and are currently pursuing the target behaviour on two 10-point likert scales. Following He et al. [16], we also collect six measures pertaining to participants' experiences conversing with the system and their perceived relationship with the conversational agent:

- **Therapeutic Alliance.** The Working Alliance Inventory–short revised (WAI-SR) measures the relationship between a counsellor and a client [53]. This questionnaire can be used to measure the degree to which interactions with the system are perceived as helpful and supportive by participants.
- **Perception of MI.** The Client Evaluation of Motivational Interviewing Scale (CEMI) is a measure designed to evaluate MI adherence of counsellors without the need of expert judgements. It is intended to be completed by clients directly after MI sessions, and serves as a proxy for the degree to which users feel that the spirit of MI was met during the interaction [23–25].
- **User Engagement.** Since we focus on the conversational aspect of interactions, we follow [16] and use the User Engagement Scale–Short Form (UES-SF) without the aesthetic appeal subscale [17, 36].
- **Perceived Empathy & Perceived Communication Competence.** We base the questions used for these measures on [16].
- **Readiness to Change.** Participants are asked to indicate their position on the *Contemplation Ladder* [5, 49], a tool to measure readiness for behaviour change on an 11-point likert scale, both before and after interaction with the conversational system. The values indicated on the contemplation ladder can be translated into stages of change as defined by the transtheoretical model [10, 51][2].

[1] Preregistering the study meant we publicly shared the study design, research questions, and analysis plan before data collection. Our time-stamped preregistration, along with extensive information about the study design, can be found here.

[2] the stages of change are precontemplation (no thoughts of changing), contemplation (beginning to consider change), preparation (preparing for change), action (taking steps to change), and maintenance (maintaining change).

2.2 Conversational System

Throughout the conversation, we used classifiers introduced and evaluated in previous work [28, 30], to evaluate participants' stance on behaviour change. Each user utterance was classified with a valence (change, sustain), a topic (Commitment, Taking Steps, Reason), and, if the topic is reason, a reason type (general, ability, need, desire). Each conversation consisted of three phases, based on the main MI processes and the structure of MI sessions [8, 31]:

Phase 1: Engaging and Focusing (4 turns). The system introduces itself, asks the participant which behaviour they would like to change and why, asks a "scaling question" to determine their level of readiness to change and follows up on the scaling question by asking why the participant did not choose a higher or lower value.

Phase 2: Evoking (5 turns). During this part of the conversation, all bot utterances are generated by GPT-4. Participants were randomly assigned to one of two conditions:
GPT-4: The complete conversation history is passed to GPT-4 to solicit a bot utterance.
MI-adapted GPT-4: The classification of the current user utterance is used to identify a suitable bot behaviour, as previously mapped through MI-literature [8, 31]. The complete conversation history is passed to GPT-4 with a system prompt setting the context of the conversation passing a definition of a suitable bot behaviour as given in MI-literature (see Table 1).

In both conditions, participants have to rate each bot utterance as either good/helpful, bad/unhelpful, or offensive/harmful and are given the option to explain their choice in free text format. We also set the same parameters for calls to the GPT-4 api in both conditions, setting the temperature to 0.7 and max_tokens to 100.

Phase 3: Conclusion (3 turns). The system summarises the conversation (generated by GPT) and invites the participant to add to the summary. The participant is then invited to define concrete next steps to take, and the system ends the conversation with a goodbye message. Except for the final bot utterance, participants are again prompted to rate the bot utterances and optionally explain their rating choice.

2.3 Participants

The study was run on Prolific. It was performed in German with German native speakers. Each participant was paid €4 as compensation for their efforts, and the median completion time was 20.9 minutes. Participants were between 19 and 72 years old (mean=32.28, sd=9.45). 41.4% were female. The majority of the participants were highly educated. 31.2% had completed a university entry-level high school diploma and 56% had at least a bachelor's degree. We provide demographic data for each participant in the dataset.

Most participants chose procrastination as their area to work on, with the least participants interested in increasing sustainability (see Table 2). Participants generally had a high level of identification with their chosen target behaviour (only 25% indicated a level of identification of 7 or less). Mean readiness to change at the beginning of the interaction was 6.25 (sd=2.27, min=0, max=10). This indicates that the interventions tested were relevant to the chosen participant pool and that the conversations collected can be expected to reflect realistic user-chatbot interactions. To ensure the validity of the post-test questionnaires, participants had to answer an attention check.

Target Behaviour	MI-adapted GPT-4	GPT-4
healthier eating	26	21
sustainable living	2	10
less procrastination	51	46
sum	79	78

Table 2: Number of chats per target behaviour and condition.

2.4 Post-Processing

We share the raw collected data in its original state. In addition, we apply a number of post-processing steps and share the processed data in separate files. In the following, we summarise all post-processing steps applied to the dataset.

5 participants did not pass the attention check, and 16 participants led multiple (at least two) chats with the system. Since the chat interactions of these participants can nevertheless give valuable information, we decided to keep the data points in the dataset, even though the measures collected might not be valid in those cases. We provide a separate file with duplicate submission IDs and mark participants who did not pass the attention checks. 16 chats experienced technical issues. For example, these led to users having to send a message more than once before the system replied. Again, we marked these instances, but kept the associated data, as it can potentially be used to evaluate how users handle such technical errors, especially when the underlying conversational agent is highly capable as is the case for GPT-4.

We also calculate the overall results for each post-processing questionnaire (plus the values for each user engagement subscale), the $\Delta readinesstochange$ by subtracting pre-conversation from post-conversation values, and translate the readiness to change indicated by participants at the start of the conversation to the corresponding stage of change as defined in [5, 49].

Since we notice that in many instances, user utterances are not directly related to change, instead constituting reactions to bot utterances, we run a prefilter[3] trained on the same dataset as the valence, topic, and reason type classifiers over each user utterance. This prefilter classifies an utterance as change related or Follow/Neutral (not directly related to behaviour change). We then introduce follow/neutral as a third label for user utterance valences, resulting in valence distinctions of change talk, sustain talk, or follow/neutral.

Finally, to make the dataset appealing to a wider audience, we translate all conversational turns and rating explanations into English using a pre-trained machine translation model[4] [50]. We share both the German original and the English translations of the conversational data.

[3] https://huggingface.co/selmey/behaviour_change_prefilter_german
[4] https://huggingface.co/Helsinki-NLP/opus-mt-de-en – While some basic quality control of the translations was conducted, subtle nuances in language may not be fully captured. Readers should be aware of potential translation imperfections.

3 POTENTIAL APPLICATIONS

Due to the various variables explored in the dataset, its potential applications are vast, and range from behavioural analysis to the training of classifiers and generative models to improve LLM-outputs in the context. Thus, the dataset is relevant for research in human-computer interaction and information behaviour, as well as natural language processing. Below, we share two potential applications of the dataset, as well as a non-exhaustive list of research questions to be explored. Further relevant research questions related to the data set are described in our previous work [26, 27] and in the preregistration of the user study [29].

3.1 Exploration of User Expectations

The first message sent to participants in each conversation informed them that the goal of the conversation is not to give advice, but instead to help them reflect about behaviour change, in order to create their own plans for change. Although this message told participants not to expect advice and facts, a non-negligible share of participants voiced discontent with the lack of advice in the MI-adapted condition, either in their rating justifications, or in the chat itself, as can be seen below:

> **CA:** Understood! Finally, please tell me again what your next step will be.
> **Participant:** You tell me.

In contrast, other users took the questions posed by the conversational system as nudges to reflect and come up with solutions themselves. A qualitative analysis of the chats could shed light on the expectations different users pose to such systems. Many of the variables we collected (i.e. demographic data, readiness to change, chosen target behaviour) have the potential to influence user behaviour and can be explored as differentiating aspects.

3.2 Anticipating Information Needs

Throughout the conversations, the users shared information needs, both explicitly and implicitly. Note the following example:

> **CA:** [...] How will you make sure you follow your diet plan?
> **Participant:** That's the hard part, it has to be simple, not complicated and expensive food, and it has to be tasty too. I can never be sure, I have to try it out.

Even though the participant does not directly ask the conversational agent for help here, they still voice an information need that could be addressed by the system. Being able to anticipate these implicit information needs has the potential to greatly improve conversational search. This dataset can serve as a resource to analyse how users implicitly voice information needs in conversations with LLMs. Future research could then combine LLMs to provide flexible emotional support with retrieval to resolve implicitly expressed information needs identified through interactions with the LLM.

3.3 Relevant Research Questions

How does user interaction with the system differ based on conversational condition?
Variables to explore include utterance classifications, bot utterance ratings given by users, self-disclosure by the user, and how cooperative the user is in interaction with the system.

Can conversations with a chatbot be used to draw conclusions about a user's readiness to change their behaviour?
Another avenue to explore would be whether the conversational logs can be used to draw conclusions about chat success.

What impact does user behaviour have on chat success?
Research on early LLM-based generative chatbots in the context of general chit-chat has shown that unclear user utterances have a significant impact in the ability of such systems to perform well [45]. Other user behaviour aspects could have similar effects.

Which bot behaviours are most popular with users, and which bot behaviours are most effective in inducing an increase in readiness to change?
In this context, it is also worth exploring to what degree those groups intersect.

How do users interact with LLM-based systems in the context of behaviour change, compared to more restrictive (rule- or retrieval-based) systems?
The dataset can be used to expand on existing literature on user interactions with those traditional conversational agents, i.e. regarding user perceptions of chatbot mistakes [9] and self-disclosure in conversation with different types of chatbots [18–20].

4 ETHICAL CONSIDERATIONS

Participants were informed about the fact they are conversing with a GPT-based conversational agent in the informed consent. They were also explicitly told that their utterances are passed on to a third party, and advised not to share any information they would not feel comfortable disclosing in an anonymous forum. We urged participants to take on the role of one of three personas. Although we observed that the majority of participants did not consistently maintain these personas during their interactions with the system, they were repeatedly reminded of the experimental nature of the system through the solicitation of a judgement about response helpfulness and safety after each bot turn.

5 CONCLUSION

In this resource paper, we describe the collection process and potential applications of a dataset of interactions between 164 users and two GPT-4 based conversational systems. In addition to the conversations between users and systems, the dataset contains a variety of information that can be used to analyse user behaviour and conversation success: pre- and post-test measure results, utterance classifications regarding user's stance on behaviour change, ratings of LLM-generated bot-turns and rating explanations. Interactions with the MI-adapted condition also include information on the type of utterance prompted to GPT-4. The dataset can be used to analyse user and GPT-4 behaviour in the context of behaviour change, and serves as a valuable resource for the improvement of such systems in the future.

REFERENCES

[1] Imtihan Ahmed, Eric Keilty, Carolynne Cooper, Peter Selby, and Jonathan Rose. 2022. Generation and Classification of Motivational-Interviewing-Style Reflections for Smoking Behaviour Change Using Few-Shot Learning with Transformers. (2022).

[2] John W Ayers, Adam Poliak, Mark Dredze, Eric C Leas, Zechariah Zhu, Jessica B Kelley, Dennis J Faix, Aaron M Goodman, Christopher A Longhurst, Michael Hogarth, and Davey M Smith. 2023. Comparing Physician and Artificial Intelligence Chatbot Responses to Patient Questions Posted to a Public Social Media Forum. *JAMA internal medicine* 183, 6 (June 2023), 589—596. https://doi.org/10.1001/jamainternmed.2023.1838

[3] Sabrina Barko-Sherif, David Elsweiler, and Morgan Harvey. 2020. Conversational agents for recipe recommendation. In *Proceedings of the 2020 Conference on Human Information Interaction and Retrieval*. 73–82.

[4] Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (Virtual Event, Canada) *(FAccT '21)*. Association for Computing Machinery, New York, NY, USA, 610–623. https://doi.org/10.1145/3442188.3445922

[5] Lois Biener and David B Abrams. 1991. The Contemplation Ladder: validation of a measure of readiness to consider smoking cessation. *Health psychology* 10, 5 (1991), 360.

[6] Maya Boustani, Stephanie Lunn, Ubbo Visser, Christine Lisetti, et al. 2021. Development, Feasibility, Acceptability, and Utility of an Expressive Speech-Enabled Digital Health Agent to Deliver Online, Brief Motivational Interviewing for Alcohol Misuse: Descriptive Study. *Journal of medical Internet research* 23, 9 (2021), e25837.

[7] Kushal Chawla, Weiyan Shi, Jingwen Zhang, Gale Lucas, Zhou Yu, and Jonathan Gratch. 2023. Social Influence Dialogue Systems: A Survey of Datasets and Models For Social Influence Tasks. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, Dubrovnik, Croatia, 750–766. https://aclanthology.org/2023.eacl-main.53

[8] Dawn Clifford and Laura Curtis. 2016. *Motivational interviewing in nutrition and fitness*. Guilford Publications.

[9] Marianna A de Sá Siqueira, Barbara CN Müller, and Tibor Bosse. 2023. When do we accept mistakes from chatbots? The impact of human-like communication on user experience in chatbots that make mistakes. *International Journal of Human–Computer Interaction* (2023), 1–11.

[10] Carlo C DiClemente and James O Prochaska. 1998. Toward a comprehensive, transtheoretical model of change: Stages of change and addictive behaviors. (1998).

[11] Ela Elsholz, Jon Chamberlain, and Udo Kruschwitz. 2019. Exploring language style in chatbots to increase perceived product value and user engagement. In *Proceedings of the 2019 Conference on Human Information Interaction and Retrieval*. 301–305.

[12] David Elsweiler, Max L Wilson, and Brian Kirkegaard Lunn. 2011. Understanding casual-leisure information behaviour. In *New directions in information behaviour*. Vol. 1. Emerald Group Publishing Limited, 211–241.

[13] Dariush D Farhud. 2015. Impact of lifestyle on health. *Iranian journal of public health* 44, 11 (2015), 1442.

[14] Joseph R Ferrari, Jean O'Callaghan, and Ian Newbegin. 2005. Prevalence of procrastination in the United States, United Kingdom, and Australia: arousal and avoidance delays among adults. *North American Journal of Psychology* 7, 1 (2005).

[15] Alexander Frummet, David Elsweiler, and Bernd Ludwig. 2022. "What Can I Cook with these Ingredients?"-Understanding Cooking-Related Information Needs in Conversational Search. *ACM Transactions on Information Systems (TOIS)* 40, 4 (2022), 1–32.

[16] Linwei He, Erkan Basar, Reinout W Wiers, Marjolijn L Antheunis, and Emiel Krahmer. 2022. Can chatbots help to motivate smoking cessation? A study on the effectiveness of motivational interviewing on engagement and therapeutic alliance. *BMC Public Health* 22, 1 (2022), 726.

[17] Marianne Holdener, Alain Gut, Alfred Angerer, et al. 2020. Applicability of the user engagement scale to mobile health: a survey-based quantitative study. *JMIR mHealth and uHealth* 8, 1 (2020), e13244.

[18] Eunbin Kang and Youn Ah Kang. 2023. Counseling chatbot design: The effect of anthropomorphic chatbot characteristics on user self-disclosure and companionship. *International Journal of Human–Computer Interaction* (2023), 1–15.

[19] Jieon Lee, Daeho Lee, and Jae-gil Lee. 2022. Influence of Rapport and Social Presence with an AI Psychotherapy Chatbot on Users' Self-Disclosure. *International Journal of Human–Computer Interaction* (2022), 1–12.

[20] Yi-Chieh Lee, Naomi Yamashita, Yun Huang, and Wai Fu. 2020. " I hear you, I feel you": encouraging deep self-disclosure through a chatbot. In *Proceedings of the 2020 CHI conference on human factors in computing systems*. 1–12.

[21] Yanran Li, Ke Li, Hongke Ning, Xiaoqiang Xia, Yalong Guo, Chen Wei, Jianwei Cui, and Bin Wang. 2021. Towards an online empathetic chatbot with emotion causes. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2041–2045.

[22] Irene Lopatovska and Jessika Davis. 2023. Designing Supportive Conversational Agents With and For Teens. In *Proceedings of the 2023 Conference on Human Information Interaction and Retrieval*. 328–332.

[23] Michael B Madson, Richard S Mohn, Julie A Schumacher, and Alicia S Landry. 2015. Measuring client experiences of motivational interviewing during a lifestyle intervention. *Measurement and Evaluation in Counseling and Development* 48, 2 (2015), 140–151.

[24] Michael B Madson, Richard S Mohn, Allan Zuckoff, Julie A Schumacher, Jane Kogan, Shari Hutchison, Emily Magee, and Bradley Stein. 2013. Measuring client perceptions of motivational interviewing: factor analysis of the Client Evaluation of Motivational Interviewing scale. *Journal of Substance Abuse Treatment* 44, 3 (2013), 330–335.

[25] Michael B Madson, Margo C Villarosa, Julie A Schumacher, and Richard S Mohn. 2016. Evaluating the validity of the client evaluation of motivational interviewing scale in a brief motivational intervention for college student drinkers. *Journal of substance abuse treatment* 65 (2016), 51–57.

[26] Selina Meyer. 2021. Natural Language Stage of Change Modelling for "Motivationally-driven" Weight Loss Support. In *Proceedings of the 2021 International Conference on Multimodal Interaction*. 807–811.

[27] Selina Meyer. 2022. "I'm at my wits' end"-Anticipating Information Needs and Appropriate Support Strategies in Behaviour Change. In *Proceedings of the 2022 Conference on Human Information Interaction and Retrieval*. 396–399.

[28] Selina Meyer and David Elsweiler. 2022. GLoHBCD: A Naturalistic German Dataset for Language of Health Behaviour Change on Online Support Forums. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*. 2226–2235.

[29] Selina Meyer and David Elsweiler. 2023. Evaluating the Efficacy, Controllability, and Safety of LLM-driven Conversational Agents to Support Behaviour Change. (2023).

[30] Selina Meyer and David Elsweiler. 2023. Towards Cross-Content Conversational Agents for Behaviour Change: Investigating Domain Independence and the Role of Lexical Features in Written Language Around Change. *researchgate preprint 10.13140/RG.2.2.10419.30242* (2023).

[31] William R Miller and Stephen Rollnick. 2012. *Motivational interviewing: Helping people change*. Guilford press.

[32] Harsha Nori, Nicholas King, Scott Mayer McKinney, Dean Carignan, and Eric Horvitz. 2023. Capabilities of gpt-4 on medical challenge problems. *arXiv preprint arXiv:2303.13375* (2023).

[33] Elnaz Nouri, Robert Sim, Adam Fourney, and Ryen W White. 2020. Step-wise recommendation for complex task support. In *Proceedings of the 2020 Conference on Human Information Interaction and Retrieval*. 203–212.

[34] Office for National Statistics. [n. d.]. Most adults report making some changes to their lifestyle for environmental reasons. ([n. d.]). https://www.ons.gov.uk/peoplepopulationandcommunity/wellbeing/articles/mostadultsreportmakingsomechangestotheirlifestyleforenvironmentalreasons/2023-07-05

[35] Stefan Olafsson, Teresa O'Leary, and Timothy Bickmore. 2019. Coerced change-talk with conversational agents promotes confidence in behavior change. In *Proceedings of the 13th EAI International Conference on Pervasive Computing Technologies for Healthcare*. 31–40.

[36] Heather L O'Brien, Paul Cairns, and Mark Hall. 2018. A practical approach to measuring user engagement with the refined user engagement scale (UES) and new UES short form. *International Journal of Human-Computer Studies* 112 (2018), 28–39.

[37] Andrea Papenmeier, Alexander Frummet, and Dagmar Kern. 2022. "Mhm..."–Conversational Strategies For Product Search Assistants. In *Proceedings of the 2022 Conference on Human Information Interaction and Retrieval*. 36–46.

[38] Andrea Papenmeier, Dagmar Kern, Daniel Hienert, Alfred Sliwa, Ahmet Aker, and Norbert Fuhr. 2021. Dataset of Natural Language Queries for E-Commerce. In *Proceedings of the 2021 Conference on Human Information Interaction and Retrieval*. 307–311.

[39] SoHyun Park, Jeewon Choi, Sungwoo Lee, Changhoon Oh, Changdai Kim, Soohyun La, Joonhwan Lee, Bongwon Suh, et al. 2019. Designing a chatbot for a brief motivational interview on stress management: Qualitative case study. *Journal of medical Internet research* 21, 4 (2019), e12231.

[40] MGM Pinho, JD Mackenbach, Hélène Charreire, J-M Oppert, H Bárdos, K Glonti, H Rutter, Sofie Compernolle, Ilse De Bourdeaudhuij, JWJ Beulens, et al. 2018. Exploring the relationship between perceived barriers to healthy eating and dietary behaviours in European adults. *European journal of nutrition* 57 (2018), 1761–1770.

[41] Amon Rapp, Lorenzo Curti, and Arianna Boldi. 2021. The human side of human-chatbot interaction: A systematic literature review of ten years of research on text-based chatbots. *International Journal of Human-Computer Studies* 151 (2021), 102630.

[42] Ian Ruthven. 2019. Making meaning: A focus for information interactions research. In *Proceedings of the 2019 conference on human information interaction and retrieval*. 163–171.

[43] Samiha Samrose and Ehsan Hoque. 2022. MIA: Motivational Interviewing Agent for Improving Conversational Skills in Remote Group Discussions. *Proceedings of the ACM on Human-Computer Interaction* 6, GROUP (2022), 1–24.

[44] Daniel Schulman, Timothy W Bickmore, and Candace L Sidner. 2011. An Intelligent Conversational Agent for Promoting Long-Term Health Behavior Change Using Motivational Interviewing.. In *AAAI Spring Symposium: AI and Health Communication*. 61–64.

[45] Abigail See and Christopher D Manning. 2021. Understanding and predicting user dissatisfaction in a neural generative chatbot. In *Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue*. 1–12.

[46] Ashish Sharma, Inna W Lin, Adam S Miner, David C Atkins, and Tim Althoff. 2023. Human–AI collaboration enables more empathic conversations in text-based peer-to-peer mental health support. *Nature Machine Intelligence* 5, 1 (2023), 46–57.

[47] Ashish Sharma, Kevin Rushton, Inna Wanyin Lin, David Wadden, Khendra G Lucas, Adam S Miner, Theresa Nguyen, and Tim Althoff. 2023. Cognitive Reframing of Negative Thoughts through Human-Language Model Interaction. *arXiv preprint arXiv:2305.02466* (2023).

[48] Siqi Shen, Charles Welch, Rada Mihalcea, and Verónica Pérez-Rosas. 2020. Counseling-Style Reflection Generation Using Generative Pretrained Transformers with Augmented Context. In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*. Association for Computational Linguistics, 1st virtual meeting, 10–20. https://aclanthology.org/2020.sigdial-1.2

[49] James D Slavet, LAR Stein, Suzanne M Colby, Nancy P Barnett, Peter M Monti, Charles Golembeske Jr, and Rebecca Lebeau-Craven. 2006. The Marijuana Ladder: Measuring motivation to change marijuana use in incarcerated adolescents. *Drug and Alcohol Dependence* 83, 1 (2006), 42–48.

[50] Jörg Tiedemann and Santhosh Thottingal. 2020. OPUS-MT — Building open translation services for the World. In *Proceedings of the 22nd Annual Conferenec of the European Association for Machine Translation (EAMT)*. Lisbon, Portugal.

[51] WF Velicer, JO Prochaska, JL Fava, GJ Norman, and CA Redding. 1998. Detailed overview of the transtheoretical model. *Homeostasis* 38 (1998), 216–33.

[52] Laura Weidinger, Maribeth Rauh, Nahema Marchal, Arianna Manzini, Lisa Anne Hendricks, Juan Mateos-Garcia, Stevie Bergman, Jackie Kay, Conor Griffin, Ben Bariach, Iason Gabriel, Verena Rieser, and William Isaac. 2023. Sociotechnical Safety Evaluation of Generative AI Systems. arXiv:2310.11986 [cs.AI]

[53] Fabian Wilmers, Thomas Munder, Rainer Leonhart, Thomas Herzog, Reinhard Plassmann, Jürgen Barth, and Hans Wolfgang Linster. 2008. Die deutschsprachige Version des Working Alliance Inventory-short revised (WAI-SR)-Ein schulenübergreifendes, ökonomisches und empirisch validiertes Instrument zur Erfassung der therapeutischen Allianz. *Klinische Diagnostik und Evaluation* 1, 3 (2008), 343–358.

[54] Bei Xu and Ziyuan Zhuang. 2022. Survey on psychotherapy chatbots. *Concurrency and Computation: Practice and Experience* 34, 7 (2022), e6170.