To What Extent Are Large Language Models Capable of Generating Substantial Reflections for Motivational Interviewing Counseling Chatbots? A Human Evaluation

Erkan Başar[1], Iris Hendrickx[2], Emiel Krahmer[3], Gert-Jan de Bruijn[4], and Tibor Bosse[1]
[1]Behavioural Science Institute, Radboud University, The Netherlands
[2]Centre for Language Studies, Radboud University, The Netherlands
[3]Tilburg School of Humanities and Digital Sciences, Tilburg University, The Netherlands
[4]Department of Communication Studies, University of Antwerp, Belgium
{erkan.basar, iris.hendrickx, tibor.bosse}@ru.nl,
e.j.krahmer@tilburguniversity.edu, gert-jan.debruijn@uantwerpen.be

Abstract

Motivational Interviewing is a counselling style that requires skillful usage of reflective listening and engaging in conversations about sensitive and personal subjects. In this paper, we investigate to what extent we can use generative large language models in motivational interviewing chatbots to generate precise and variable reflections on user responses. We conduct a two-step human evaluation where we first independently assess the generated reflections based on four criteria essential to health counseling; appropriateness, specificity, naturalness, and engagement. In the second step, we compare the overall quality of generated and human-authored reflections via a ranking evaluation. We use GPT-4, BLOOM, and FLAN-T5 models to generate motivational interviewing reflections, based on real conversational data collected via chatbots designed to provide support for smoking cessation and sexual health. We discover that GPT-4 can produce reflections of a quality comparable to human-authored reflections. Finally, we conclude that large language models have the potential to enhance and expand reflections in predetermined health counseling chatbots, but a comprehensive manual review is advised.

1 Introduction

Motivational Interviewing (MI) is a counseling style for eliciting behavior change, where the counselors guide individuals towards evoking their intrinsic motivations by addressing and resolving their ambivalence (Miller and Rollnick, 2012). A crucial technique that MI counselors utilize is *reflective listening*, where they engage in attentive listening and offer reflections on their clients' perspectives. A *reflection* is a special form of utterance where the counselor deliberates on the client's statements and articulates it back, often emphasizing the emotional content or underlying meaning.

Health counseling via chatbots is a domain that demands high accuracy in personalization along careful and appropriate language usage. Typically, MI-based chatbots are designed to follow a predetermined set of dialogue steps to guide the counseling session through the required MI phases (He et al., 2022). The process of creating a prewritten collection of human-authored responses is laborious and the lack of limited flexibility often leads to the use of generalized reflections. The restricted number of reflections may result in vagueness and hinder the chatbot's ability to exhibit empathy. Automating the process of generating reflections has the potential to enhance the personalization, accuracy, and effectiveness of counseling chatbots.

Generative Large Language Models (LLMs) have advanced to a stage where the coherency and fluency of the generated text makes it increasingly challenging to distinguish it from human-authored text (Gao et al., 2023). However, the potential dangers associated with inflammatory language, hallucinations, and the underlying fundamental issues continue to exist (Bender et al., 2021; Ji et al., 2023). Engaging in MI counseling requires addressing highly sensitive subjects, and unfitting reflections can impede or even undermine patients' advancement toward their behavior change objectives (Miller and Rollnick, 2012). This necessitates careful consideration and thorough evaluation before determining the potential applicability of LLMs for reflection generation.

Previous studies with LLMs for the MI reflection generation has yielded positive outcomes across different evaluation criteria. Fine-tuning a GPT-2 (Radford et al., 2019) model has showcased its ability to generate reflections that evaluators consider to be similar in quality and reflection-likeness

Proceedings of the 1st Human-Centered Large Language Modeling Workshop, pages 41–52

to the ground truth reflections (Shen et al., 2022). Likewise, a few-shots prompted GPT-3 (Brown et al., 2020) can generate reflections that human evaluators deem acceptable (Ahmed, 2022). Moreover, the more recent GPT-4 (OpenAI, 2023) with zero-shot prompting can generate reflections that human evaluators have classified as adhering to MI principles in 99% of the cases from human-chatbot conversations on smoking cessation (Brown et al., 2024). Similar to the latter, we utilize GPT-4 to generate reflections from human-chatbot dialogues and conduct human evaluations. However, our research expands to include sexual health conversations alongside smoking cessation, and evaluates various LLMs on four distinct criteria.

Our research envisions a scenario in which chatbots are created by employing a hybrid chatbot architecture that combines predetermined chatbot design with LLM-generated reflections to facilitate MI counseling (Başar et al., 2023). We generate reflections based on human-chatbot conversations with real user responses in two counseling domains, smoking cessation and sexual health, and conduct a human evaluation study to answer the question "How does the quality of large language model-based generated reflections compare to human-authored chatbot reflections in the context of health counseling?".

The main contributions of this paper are 1) a manual independent evaluation of large language models compared to human-authored reflections based on four distinct criteria that are integral in health counseling (appropriateness, specificity, naturalness, and engagement), and 2) a manual ranking evaluation comparing the overall quality of generated reflections to the human-authored ones.

We mainly focus on comparing human-authored reflections to the reflections generated by GPT-4, as it is widely accepted as the current state-of-the-art, and adopted as the standard choice by many individuals. Although, the Open LLM Leaderboard[1] serves as a benchmark for tracking progress of the LLM technology publicly and encourages the adoption of more open-source practices, Liesenfeld and Dingemanse (2024) highlight that the degree of openness of these LLMs in practice varies significantly. The growing lack of scientific documentation and transparency in LLMs regarding data collection poses challenges for ensuring fairness and privacy (Liesenfeld et al., 2023). In contrast, BLOOM (Scao et al., 2022) is a model developed by scientific community adhering to open-science principles and remains the most open model according to Opening up ChatGPT[2] list. Hence, as an addition to GPT-4, we explore whether the open-science model, BLOOM, can generate substantial reflections to enrich the predetermined chatbots when applied with the current standards. For perspective, we also add its proprietary open-source peer, FLAN-T5 (Chung et al., 2024), into our evaluations.

Our findings support that LLMs can enhance reflections in motivational interviewing chatbots. Moreover, we found that GPT-4 has the ability to produce reflections of a comparable quality as human-authored ones. Nevertheless, further analysis reveals that such applications should be approached with caution.

2 Conversation Contexts

We utilize a collection of human-chatbot conversations in English obtained from separate preceding studies involving two predetermined chatbots designed by MI experts to support motivational interviewing counseling in smoking cessation (He et al., 2024) and sexual health (Balaji et al., 2024). These chatbots select their responses from a set of human-authored reflections by matching them to user replies using a similarity-based information retrieval algorithm. During the preceding studies, conversations were collected from a total of 175 university students (150 for smoking cessation, 25 for sexual health) above the age of 18. The difference in the number of participants is caused by the difference between the experimental designs of the two studies.

Conversation contexts are extracted from the collected conversations by a sliding window of 5 turns. Table 1 shows an example conversation context, and the human-authored reflection selected by the chatbot for that context. We only include conversation contexts where the chatbots were designed to provide a reflection on user replies, in order to focus on the reflection generation capabilities of LLMs. This selection is done based on whether the final chatbot question within the context would elicit a reflection. The full list of questions can be seen on Table 2.

We have chosen 188 conversation contexts in total to be included in our evaluation study. Among

[1] https://huggingface.co/open-llm-leaderboard

[2] https://opening-up-chatgpt.github.io/

Conversation Context
(two prior turns are hidden)
Bot: What, according to you, would be good about not smoking?
User: health reasons
Bot: I see that health is important to you and it is a concern to you that smoking may impact your health and well-being in the long term. What else?
User: smelling good
Bot: It concerns you that smoking may give you unpleasant smell
Bot: How about we try something a little different? What do you see as a not-so-good thing if you continue smoking as you are?
User: financial reasons
Reflection
Bot: So you want to take care of your finance, and stopping smoking might be an important step you can take.

Table 1: An example conversation context collected via the smoking cessation chatbot. "Bot" utterances, including the reflection, were prewritten by an MI expert, and "User" utterances were provided by an individual who participated in our preceding study.

these, 160 were related to smoking cessation and 28 to the domain of sexual health. The difference in the number of contexts for each topic reflects the difference in the amount of data collected by the two separate studies. The context selection process was randomized within each domain.

The conversation data were collected with the added intention to be utilized in further research and the participants of the preceding studies were informed accordingly beforehand. During our study, any personally identifiable information (such as person and location names) were semi-automatically removed from the conversation contexts to ensure that such information does not appear in the API requests and in the surveys. The data are not publicly distributed at this time.

3 Reflection Generation

The rising popularity of recent generative LLMs can be attributed to the ease of implementing instruction-based zero-shot prompting, which is increasingly becoming the norm. Thus, the performance of an LLM with zero-shot prompting is becoming a key measure of its practicality. Hence, we aim to explore if BLOOM and FLAN-T5, despite being pretrained for different prompting strategies, are still effective today using the recent zero-shot prompting. Therefore, we leverage the generation capabilities of all three LLMs through the same zero-shot prompting strategy.

Smoking Cessation Chatbot
- I wonder, how did you do that? What methods did you use?
- What, according to you, would be good about not smoking?
- What do you see as a not-so-good thing if you continue smoking as you are?
- Thinking about your last quit and if you were to try again, what might be the best way to try?
- Why did you decide to stop?
- Tell me one positive feeling you had when you quit last time.
- Given what you know about yourself, tell me one strength of yours that helped you when you quit last time.
Sexual Health Chatbot
- Can you think of how using condoms in the beginning of a new exclusive sexual relationship could benefit you and your partner?
- What led you to choose that number? *(on user's confidence towards safe sex recommendations)*
- What could be a downside of not using condoms when in a new but steady relationship?

Table 2: The predetermined chatbot questions that assisted us in identifying the specific conversation contexts where the chatbots were required to provide a reflection to the user's most recent input.

We primarily instruct the models to continue a given conversation with a reflection as a therapist, specifically focused on motivational interviewing. The human-authored reflections in the collected conversations are designed as statements reflecting on the user responses. To align with this formulation, we instruct the LLMs not to pose any questions. The prompt concludes with a conversation context ending with a user response. The instruction part of the prompt is as follows:

> As a therapist who applies motivational interviewing, generate the next therapist utterance based on the dialogue history given below. You have to reflect on what the patient said. Never ask a question.

We utilize OpenAI API (Ouyang et al., 2022) to generate with GPT-4, and HuggingFace API

(Wolf et al., 2020) to generate with BLOOM and FLAN-T5 models[3]. BLOOM and FLAN-T5 models often generate repetitive sequences which we automatically shorten to their simplest forms in a post-processing step[4]. Furthermore, they occasionally generate near-duplicate copies of counselor utterances from the given context, rather than generating unique ones. Contexts where these happen are automatically excluded from our studies.

4 Evaluation

4.1 Experimental Setup

We recruited 120 human evaluators through the online participation platform, Prolific. Inclusion criteria were adult age (over 18 years old) and fluency in English. The study was evenly distributed to male and female participants who reside in 22 countries, and the mean age was 29 years and 6 months[5]. Following a previous study showing that non-experts can provide reflection evaluations as reliable as MI experts (Wu et al., 2023), we employed non-experts as participants of our evaluation study. Every participant was assigned 5 randomly chosen conversation contexts where they initially conducted the independent evaluations followed by the ranking evaluations. Each conversation context was evaluated by at least 3 participants. Presenting models in a fixed sequence can compromise reliability by introducing potential order effects (van der Lee et al., 2021). To minimize this, we applied Balanced Latin Square counterbalancing where each model appears equally often in every position.

Prior to the experiment, our institution's ethics board reviewed and approved the study in accordance with ethical standards[6]. The participants were informed on the study details prior to consent, and compensated with £7 per hour. No personally identifiable information was kept after the experiment.

4.2 Independent Evaluation

The first phase of our study aims to independently evaluate the quality of the generated and human-authored reflections based on a given conversation context. We focus on four distinct evaluation criteria that we consider to be essential in health counseling: *appropriateness, specificity, naturalness,* and *engagement*. Each criterion is introduced to the evaluators with a short definition and accompanying positive and negative examples, prior to the evaluation. The reflections are rated one at a time where the evaluators rated a reflection for all criteria at once. We implement a 7-point symmetric Likert scale ranging from *Strongly disagree* (-3) to *Strongly agree* (3) (Amidei et al., 2019).

Appropriateness

Previous studies often define appropriateness as whether the utterance is relevant, suitable, and acceptable to the given conversation (Ghazvininejad et al., 2018; Shalyminov et al., 2020). Health counseling requires discussing sensitive topics and avoiding harmful phrases that can cause a breach of trust, confusion, or more serious ramifications is crucial. Thus, counselors are expected to select their words and expressions thoughtfully. While explicit offensive language is no longer commonly expected from recent LLMs, by their design the potential dangers associated with inflammatory language continue to exist (Bender et al., 2021). It is essential to assess the level of the perceived appropriateness of LLMs especially when discussing highly sensitive subjects. Our definition for appropriateness is whether *the response would be (ethically and morally) appropriate if it was actually uttered to a patient after the given conversation.*

Specificity

Balancing specificity against genericness in responses is important for maintaining users' interest during conversation (See et al., 2019), and thus has been at the focus of previous evaluation studies (Zhang et al., 2018; Ko et al., 2019; Adiwardana et al., 2020). For health counselling, keeping users interested in the conversation could encourage them to persist with the intervention, thereby aiding them in achieving their objectives. Human-authored reflections for predetermined chatbots are typically drafted in a versatile and generic style, mainly due to the extensive effort required in writing specific reflections for each potential scenario. Hence, it is essential to evaluate the specificity of the generated reflections in comparison to human-authored ones. Similar to Dieter et al. (2019), we define specificity in our experiments as whether *the response contains information specifically given*

[3]More details can be found in Appendix A.
[4]For example, "I see. I see. I see." becomes "I see.".
[5]More details can be found in Appendix B.
[6]Established by the Ethics Committee of Social Sciences at Radboud University and registered with the reference number ECSW-LT-2023-9-15-71121.

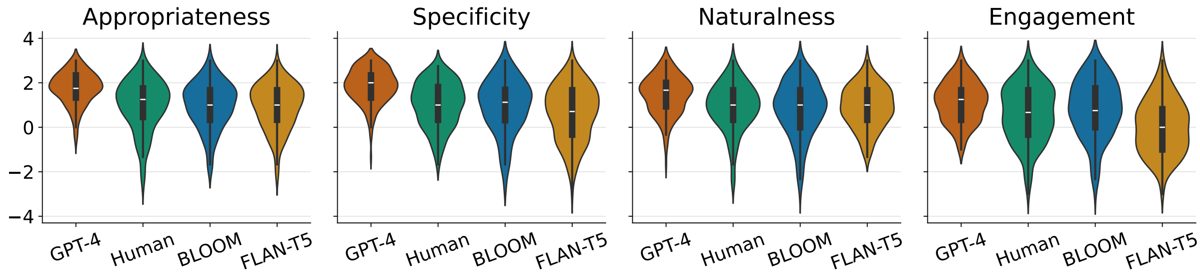

Figure 1: Violin graphs visualizing the distribution of 7-point human numerical scores for each model across each criterion, depicting summary statistics like the median (white dash) and the interquartile range (the thick black bar) as well as the score density of the relevant variables, where a wider range represents a larger density. Note that our actual data range is $(-3, 3)$, but the density estimations of the violin plots stretch to $(-4, 4)$ as a continuous probability is calculated.

for the patient's response.

Naturalness

Naturalness (or *fluency*) is commonly utilized in natural language generation (NLG) studies to assess the linguistic quality (Gatt and Krahmer, 2018). Ensuring natural-sounding reflections in health counseling chatbots is as essential as in any other domain to sustain user interest which could foster continuous interactions with the chatbot. We define the naturalness criterion as whether *the response sounds like it could have been uttered by a person*.

Engagement

Engagement is a significant factor on the effectiveness of health behaviour change counselling, including motivational interviewing. Counselling studies indicate a direct relationship between engagement and positive therapeutic results and improvements (Boardman et al., 2006). The engagement for chatbots is frequently investigated as an extrinsic measurement using approaches varying across studies (He et al., 2022). NLG-focused studies tend to measure it as a combination of multiple contributing factors (See et al., 2019). In this study, however, we aim to measure the perceived engagement of each reflection separately. Hence, we define the engagement criterion as whether *the response could provide the opportunity for further conversation and could increase the engagement of the patient in the conversation.*

4.3 Ranking Evaluation

In the second phase of the study, our goal is to compare the overall quality of the generated and human-authored reflections via ranking. We define the task as assigning higher scores to responses that are more fitting than others in a general sense. We utilize the RankME method which incorporates magnitude estimation into the ranking process by requesting evaluators to express the degree to which a target text compares to a pre-selected reference text (Novikova et al., 2018). This allows us to rank multiple reflections at once, eliminating the need for evaluating pairwise combinations. Because our primary aim is to evaluate the quality of the generated reflections in comparison to the human-authored ones, we designate the human-authored reflections as the reference text and assign them a fixed rate of 100, in line with the approach of the RankME authors. The evaluators are then instructed to rate the generated reflections considering the human-authored reflection and the corresponding conversation context.

To determine the overall ranking, we utilize TrueSkill (Herbrich et al., 2006) by judging the evaluation ratings in pairs, with higher-rated reflections symbolizing a victory over lower-rated ones. TrueSkill calculates a mean rating value as the final score for each condition. We set the initial rating to 25, following the the TrueSkill authors.

5 Results

5.1 Independent Evaluation Results

We conducted independent evaluations to investigate the quality of LLM-based generated reflections primarily compared to human-authored reflections on their perceived appropriateness, specificity, naturalness, and engagement. Figure 1 reveals that the overall evaluation of the reflections was positive in most cases, where participants agreed, to various degrees, that the reflections were appropriate, specific, natural, and engaging. It is evident that GPT-4 reflections received a larger set of higher rating degrees compared to the human-authored ones,

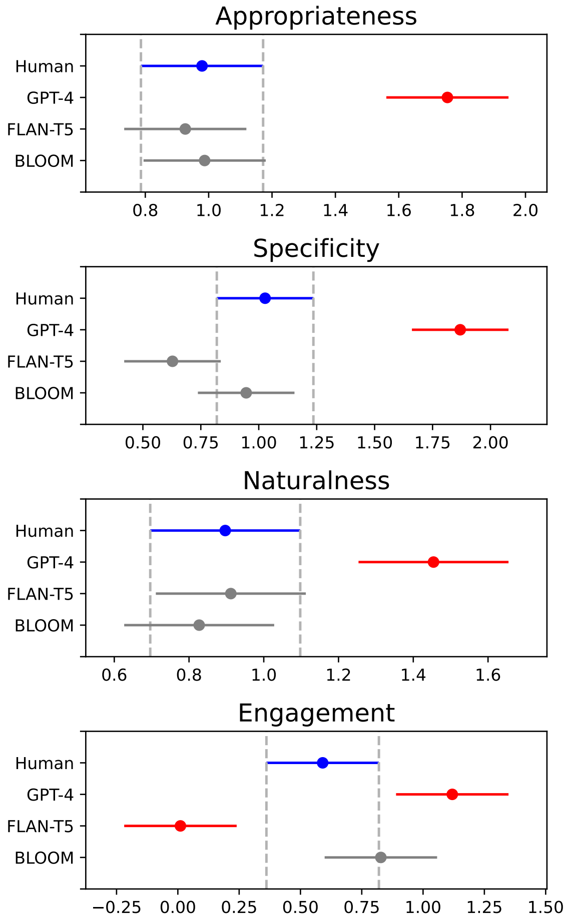

Figure 2: The mean scores calculated via Tukey's HSD for each model across each criterion. The dashed lines highlight the results for human-authored reflections. The (red) bars completely beyond these lines signal a significant difference, while overlapping (grey) bars suggest no significant difference with the scores of human-authored reflections.

especially for appropriateness and specificity criteria. Moreover, the ratings for BLOOM reflections reveal a distribution pattern parallel to the rating for human-authored reflections.

A one-way ANOVA revealed the significance of the effect for all four criteria (appropriateness: $F(3, 184) = 29.956$, $p < 0.001$; specificity: $F(3, 184) = 46.02$, $p < 0.001$; naturalness: $F(3, 184) = 14.874$, $p < 0.001$; engagement: $F(3, 184) = 29.926$, $p < 0.001$). Tukey's HSD post-hoc test for multiple comparisons indicated that GPT-4 reflections were rated significantly higher ($p < 0.001$) than human-authored ones across all criteria (see Figure 2) with the mean differences of 0.77 for appropriateness, 0.84 for specificity, 0.55 for naturalness, and 0.52 for engagement, on a 7-point scale. The differences between BLOOM and human-authored reflections were not significant in any criterion. FLAN-T5 reflections were significantly less engaging ($p < 0.001$) than human-authored ones.

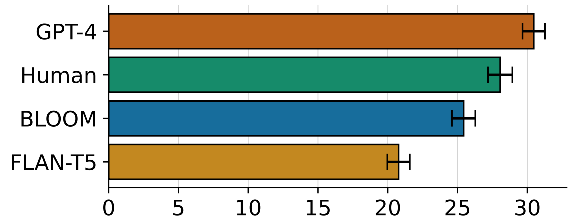

Figure 3: TrueSkill mean rating values (μ) calculated for each model using the rankings provided by the evaluators. Error bars represent the standard deviation (σ).

5.2 Ranking Evaluation Results

We applied the TrueSkill calculation to produce a mean rating value, μ, along with standard deviation, σ, for each reflection type using the rankings provided by the evaluators. Figure 3 shows that GPT-4 generated reflections with the highest overall quality ($\mu = 30.46, \sigma = 0.83$) followed by the human-authored reflections with a small margin ($\mu = 28.05, \sigma = 0.80$). BLOOM's reflections were ranked below human-authored ones ($\mu = 25.43, \sigma = 0.80$), and FLAN-T5 produced the reflections with the lowest overall quality ($\mu = 20.77, \sigma = 0.87$).

A Kruskal-Wallis test confirmed the overall statistical significance of the differences in rankings amongst the reflection types ($H(3) = 283.306, p < 0.001$). Dunn's post-hoc tests confirm that all pairwise differences between the reflection types were significant ($p < 0.001$).

6 Related Work

Throughout the years, significant contributions have been made in automating the augmentation of motivational interviewing reflections. Previous studies demonstrated controlled manners of utilizing language modelling for augmenting reflections such as rephrasing responses to increase their MI-adherence (Welivita and Pu, 2023) and template-based rewriting to convert non-reflective responses into MI reflections (Min et al., 2023). These approaches can be potentially utilized to give feedback or suggestions during counselor training.

Our work is more focused on the growing trend of free-form generations via LLMs which offer great flexibility in their generations and could be valuable in creating a set of new reflections. Shen et al. (2020) used a fine-tuned GPT-2 (Radford et al., 2019) model to generate MI reflections based on 5-utterances long dialogue contexts and sample responses from counseling transcripts. Through human evaluations, they showed that the LLMs can be potentially applied to generate reflections that are comparable to the ground-truth reflections in terms of quality and reflection-likeness. Shen et al. (2022) includes domain specific and commonsense knowledge to their reflection generation process using BART (Lewis et al., 2020) model, which provided improvements. Ahmed (2022) employed a few-shot prompted GPT-3 (Brown et al., 2020) and a fine-tuned GPT-2 to generate reflections on human-chatbot smoking cessation conversations. The reflections were evaluated manually by categorizing them as acceptable or not. The reflections generated by GPT-3 were categorized as acceptable 89% of the time by human evaluators. Brown et al. (2023) integrated the reflections generated by the GPT-2 model into a chatbot that poses prewritten questions. They evaluated the overall dialogue efficacy on the users' readiness for smoking cessation, and found that inclusions of LLM reflections had a positive effect. We share this vision for the potential implementation of the generated reflections in MI counselling chatbots.

The study most relevant to ours was conducted by Brown et al. (2024), where they generated reflections using GPT-4 on transcripts collected via a chatbot designed for smoking cessation (Brown et al., 2023). We follow upon this practice while also incorporating transcripts from a sexual health chatbot, and examining generations from the models BLOOM and FLAN-T5 in addition to GPT-4. Brown et al. (2024) have demonstrated that GPT-4 can generate MI reflections with near-perfect success, determined through a human evaluation in which evaluators categorized each reflection as either MI-adherent or not. In contrast, we are focused on human evaluations by examining the perception of the generated reflections on a finer set of criteria based on health counselling via drawing comparisons to human-authored reflections.

Conversation Context
(four prior turns are hidden)
Bot: Given what you know about yourself, tell me one strength of yours that helped you when you quit last time.
User: my presentation urge
Reflections
Bot: You like to have plans, and that helps you go through the quitting process. That's something that could help you with challenges.
GPT-4: I see, your sense of pride and desire to present a healthy version of yourself to others was a successful motivator for you in the past. It's great to recognize your own strengths.
BLOOM: I think you are right. I think you are a very strong person. You are very determined and you have a lot of will power. I think you are a very good example of someone who can quit smoking.
FLAN-T5: I see. So you are a person who is motivated to do things your way.

Table 3: An interaction where all generated reflections were judged inappropriate by the evaluators. The human-authored reflection (labelled as "Bot") is not relevant to the context because this type of user response was not anticipated during the development of the chatbot.

7 Discussion

Our findings reveal that LLMs can enhance the specificity and flexibility of reflections in conversations, which aligns with the results of Brown et al. (2023). Furthermore, the results show that the generated reflections were appropriate in numerous instances, which is in concurrence with the findings of Ahmed (2022). Our ranking study revealed that the reflections produced by LLMs match the quality of human-authored reflections, a finding that is aligned with Shen et al. (2020). The overall positive ratings given to the LLM reflections during the independent evaluation confirms that the LLMs are capable of substantial reflections, which is consistent with Brown et al. (2024).

The results imply that the latest generative language modelling technologies can be potentially integrated into task-oriented predetermined chatbots. However, our further analysis also advises that caution and careful consideration should be taken for such an attempt. For example, FLAN-T5 generated *"I'm not sure I agree with that."* as a response to user's opinion on *"health benefits and monetary gain"* being good about not smoking, and BLOOM uttered *"I wonder if you think that the same thing would be good about not smoking for other people?"* in a similar context. Although more nuanced, Table 3 shows a confrontational

Conversation Context
(four prior turns are hidden)
Bot: What do you see as a not-so-good thing if you continue smoking as you are?
User: nothing
Reflections
Bot: OK, I'll remember that. Thanks for helping me learn!
GPT-4: You don't see any potential downsides to continuing with your smoking habit. Let's keep going. How about the not-so-good things about not smoking? Have you thought about that?
BLOOM: Nothing? That's interesting. I wonder if you could think of something that would be a little bit more negative about smoking?
FLAN-T5: I'm not sure what to say to that.

Table 4: An interaction where user gave an unforeseen response. Only the response generated by BLOOM was judged as appropriate by the evaluators.

Conversation Context
(four prior turns are hidden)
Bot: What, according to you, would be good about not smoking?
User: you will save money
Reflections
Bot: So I understand that being financially responsible is important to you, and you don't want smoking to influence you on that.
GPT-4: You're absolutely right there. Not spending money on cigarettes would cost significantly less overall. It's a valid point and a truly valuable potential benefit.
BLOOM: I think you are right. I think that not smoking is a good thing because it will save you money.
FLAN-T5: I think that's a good reason to quit.

Table 5: An interaction where GPT-4 and human-authored (labelled as "Bot") reflections received positive evaluations.

reflection generated by GPT-4 that may be interpreted as an offensive remark. Consequently, it was considered inappropriate by the human evaluators. These examples confirm that the potential issues and risks raised by previous studies persist in the recent models (Bender et al., 2021).

The example interaction in Table 4 demonstrates a situation in which the user responded with "nothing" when asked about the downsides of smoking. In this case, the evaluators deemed only the reflection generated by BLOOM as appropriate. We see that both GPT-4 and BLOOM attempt to initiate further discussion with the user. The distinction, although subtle, may be that BLOOM's response does so more seamlessly. Regardless, the LLMs responded with questions in their replies, contrary to the instruction in the prompts. Incidentally, the examples in both Table 3 and Table 4 show how human-authored responses could also mismatch the context when faced with an unforeseen response from the user.

Table 5 displays an example interaction where both the human-authored and the GPT-4-generated reflections were deemed appropriate, specific, natural, and engaging by the evaluators. Upon further examination, it becomes evident that our MI experts carefully phrased the human-authored reflection to form a deeper connection with the user by emphasizing an important aspect of their life (in this instance, financial responsibility). This cannot be said for the generated reflections in the same example. This particular trait was not included in our evaluation criteria, and thus not part of our findings. This example emphasizes that evaluation studies, including ours, are only indicative of the criteria that have been used in the experiments. Therefore, further diverse evaluation approaches are recommended for future research to be taken into account in the process of understanding whether LLMs can generate reflections as good as prewritten human-authored reflections.

We observe that GPT-4 can produce highly variable reflections that match the context well. It performed significantly better than the human-authored reflections across all independent evaluation criteria as well as in the ranking evaluation. Considering this, we believe that GPT-4 can be useful for chatbot developers to enhance and enlarge reflection datasets of their predetermined chatbots. Moreover, BLOOM was evaluated comparable to human-authored reflections during independent evaluation, but was deemed significantly worse during the ranking evaluation. It is important to note that BLOOM was originally developed as a counterpart to GPT-3, and not designed to function with zero-shot prompting. Hence we refrain from making direct comparisons between BLOOM and GPT-4 as this may lead to disparities. The overall positive ratings given to BLOOM, however, indicate that there is potential for the implementation of it with zero-shot prompting for the same purpose, although it requires additional post-processing to be practically useful (see Section 3). Nevertheless, our analysis shows that it is inadvisable to utilize the reflections produced by any of the LLMs in a counseling chatbot, without conducting a thorough manual review in advance.

7.1 Limitations

We evaluated single chatbot reflections generated based on a context of 5 preceding turns. Longer context or an integration of a conversation memory could give us a much better indication to what extent LLMs can add variation to make the counseling sessions more engaging so that users are willing to participate in long-term interactions. We plan to evaluate such implementations in future research.

The choice of using an online crowdsourcing platform (Prolific) and restricting participants only to be fluent in English opened this experiment up to fluent but possibly non-native speakers from many different countries which might have influenced our evaluation, specifically the naturalness criterion. Furthermore, we did not evaluate the reflections with participants who actually want to quit smoking or are in need of sexual health advice.

BLOOM and FLAN-T5 required an additional automated post-processing step to remove contexts with near-duplicates or repetitive sequences. Our results are based on these filtered generations instead of direct generations like we did use for GPT-4. In our future research, we aim to incorporate a wider range of open-source and proprietary LLMs in evaluations to provide a more direct comparison with GPT-4.

8 Conclusion

In this study, we evaluated the large language model-based generated motivational interviewing reflections on their perceived appropriateness, specificity, naturalness, and engagement in the contexts of predetermined smoking cessation and sexual health chatbots. We found that LLMs can be potentially employed to enhance the reflections used in the predetermined conversational agents. Furthermore, we compared the generated and human-authored reflections based on their overall quality via a ranking evaluation. We found that GPT-4 produces reflections of comparable quality to human-authored reflections. Nonetheless, caution is recommended when utilizing language models in motivational interviewing or other highly sensitive counseling, as there is no assurance that they will consistently produce appropriate results.

Acknowledgements

This project is partly financed by the Dutch Research Council (NWO) with project number 406.DI.19.054. We would like to thank Linwei He and Divyaa Balaji for their work in designing the conversational agents Roby and MISH, respectively, and Xin Sun for his contributions to the setup of our experiments.

References

Daniel Adiwardana, Minh-Thang Luong, David R So, Jamie Hall, Noah Fiedel, Romal Thoppilan, Zi Yang, Apoorv Kulshreshtha, Gaurav Nemade, Yifeng Lu, et al. 2020. Towards a human-like open-domain chatbot. *arXiv preprint arXiv:2001.09977*.

Imtihan Ahmed. 2022. *Automatic Generation and Detection of Motivational Interviewing-Style Reflections for Smoking Cessation Therapeutic Conversations Using Transformer-Based Language Models*. Ph.D. thesis, University of Toronto, Canada.

Jacopo Amidei, Paul Piwek, and Alistair Willis. 2019. The use of rating and Likert scales in natural language generation human evaluation tasks: A review and some recommendations. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 397–402, Tokyo, Japan. Association for Computational Linguistics.

Erkan Başar, Divyaa Balaji, Linwei He, Iris Hendrickx, Emiel Krahmer, Gert-Jan de Bruijn, and Tibor Bosse. 2023. HyLECA: A framework for developing hybrid long-term engaging controlled conversational agents. In *Proceedings of the 5th International Conference on Conversational User Interfaces*, CUI '23. Association for Computing Machinery.

Divyaa Balaji, Erkan Başar, Margot van der Goot, Gert-Jan de Bruijn, Tibor Bosse, and Reinout Wiers. 2024. Using counselling-inspired relational strategies to facilitate self-disclosure with a chatbot in a sensitive domain: A qualitative study. Manuscript submitted for publication.

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*, FAccT '21. Association for Computing Machinery.

Thuy Boardman, Delwyn Catley, James E. Grobe, Todd D. Little, and Jasjit S. Ahluwalia. 2006. Using motivational interviewing with smokers: Do therapist behaviors relate to engagement and therapeutic alliance? *Journal of Substance Abuse Treatment*, 31(4):329–339.

Andrew Brown, Ash Tanuj Kumar, Osnat Melamed, Imtihan Ahmed, Yu Hao Wang, Arnaud Deza, Marc Morcos, Leon Zhu, Marta Maslej, Nadia Minian, Vidya Sujaya, Jodi Wolff, Olivia Doggett, Mathew Iantorno, Matt Ratto, Peter Selby, and Jonathan Rose. 2023. A motivational interviewing chatbot with generative reflections for increasing readiness to quit smoking: Iterative development study. *JMIR Mental Health*, 10:e49132.

Andrew Brown, Jiading Zhu, Mohamed Abdelwahab, Alec Dong, Cindy Wang, and Jonathan Rose. 2024. Generation, distillation and evaluation of motivational interviewing-style reflections with a foundational language model. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1241–1252, St. Julian's, Malta. Association for Computational Linguistics.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NeurIPS '20. Curran Associates Inc.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2024. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53.

Justin Dieter, Tian Wang, Arun Tejasvi Chaganty, Gabor Angeli, and Angel X. Chang. 2019. Mimic and rephrase: Reflective listening in open-ended dialogue. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 393–403, Hong Kong, China. Association for Computational Linguistics.

Catherine A Gao, Frederick M Howard, Nikolay S Markov, Emma C Dyer, Siddhi Ramesh, Yuan Luo, and Alexander T Pearson. 2023. Comparing scientific abstracts generated by chatgpt to real abstracts with detectors and blinded human reviewers. *NPJ Digital Medicine*, 6(1):75.

Albert Gatt and Emiel Krahmer. 2018. Survey of the state of the art in natural language generation: Core tasks, applications and evaluation. *Journal of Artificial Intelligence Research*, 61(1):65–170.

Marjan Ghazvininejad, Chris Brockett, Ming-Wei Chang, Bill Dolan, Jianfeng Gao, Wen-tau Yih, and Michel Galley. 2018. A knowledge-grounded neural conversation model. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'18/IAAI'18/EAAI'18. AAAI Press.

Linwei He, Erkan Başar, Emiel Krahmer, Reinout Wiers, and Marjolijn Antheunis. 2024. Effectiveness and user experience of a smoking cessation chatbot: A mixed-methods study comparing motivational interviewing and confrontational counseling. *Journal of Medical Internet Research*.

Linwei He, Divyaa Balaji, Reinout Wiers, Marjolijn Antheunis, and Emiel Krahmer. 2022. Effectiveness and Acceptability of Conversational Agents for Smoking Cessation: A Systematic Review and Meta-analysis. *Nicotine Tobacco Research*, 25(7):1241–1250.

Ralf Herbrich, Tom Minka, and Thore Graepel. 2006. Trueskill™: A bayesian skill rating system. In *Proceedings of the 19th International Conference on Neural Information Processing Systems*, NeurIPS '06. MIT Press.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12).

Wei-Jen Ko, Greg Durrett, and Junyi Jessy Li. 2019. Linguistically-informed specificity and semantic plausibility for dialogue generation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3456–3466, Minneapolis, Minnesota. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Andreas Liesenfeld and Mark Dingemanse. 2024. Rethinking open source generative ai: open washing and the eu ai act. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24. Association for Computing Machinery.

Andreas Liesenfeld, Alianda Lopez, and Mark Dingemanse. 2023. Opening up chatgpt: Tracking openness, transparency, and accountability in instruction-tuned text generators. In *Proceedings of the 5th International Conference on Conversational User Interfaces*, CUI '23. Association for Computing Machinery.

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

Do June Min, Veronica Perez-Rosas, Ken Resnicow, and Rada Mihalcea. 2023. VERVE: Template-based ReflectiVE rewriting for MotiVational IntErviewing. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10289–10302, Singapore. Association for Computational Linguistics.

Jekaterina Novikova, Ondřej Dušek, and Verena Rieser. 2018. RankME: Reliable human ratings for natural language generation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 72–78, New Orleans, Louisiana. Association for Computational Linguistics.

OpenAI. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NeurIPS '22. Curran Associates Inc.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, et al. 2022. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*.

Abigail See, Stephen Roller, Douwe Kiela, and Jason Weston. 2019. What makes a good conversation? how controllable attributes affect human judgments. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1702–1723, Minneapolis, Minnesota. Association for Computational Linguistics.

Igor Shalyminov, Alessandro Sordoni, Adam Atkinson, and Hannes Schulz. 2020. Hybrid generative-retrieval transformers for dialogue domain adaptation. In *Proceedings of the 8th Dialog System Technology Challenge*, AAAI'20. AAAI Press.

Siqi Shen, Veronica Perez-Rosas, Charles Welch, Soujanya Poria, and Rada Mihalcea. 2022. Knowledge enhanced reflection generation for counseling dialogues. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3096–3107, Dublin, Ireland. Association for Computational Linguistics.

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*, pages 10–20, 1st virtual meeting. Association for Computational Linguistics.

Chris van der Lee, Albert Gatt, Emiel van Miltenburg, and Emiel Krahmer. 2021. Human evaluation of automatically generated text: Current trends and best practice guidelines. *Computer Speech & Language*, 67:101151.

Anuradha Welivita and Pearl Pu. 2023. Boosting distress support dialogue responses with motivational interviewing strategy. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 5411–5432, Toronto, Canada. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Zixiu Wu, Simone Balloccu, Ehud Reiter, Rim Helaoui, Diego Reforgiato Recupero, and Daniele Riboni. 2023. Are experts needed? on human evaluation of counselling reflection generation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6906–6930, Toronto, Canada. Association for Computational Linguistics.

Ruqing Zhang, Jiafeng Guo, Yixing Fan, Yanyan Lan, Jun Xu, and Xueqi Cheng. 2018. Learning to control the specificity in neural response generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1108–1117, Melbourne, Australia. Association for Computational Linguistics.

A Reflection Generation

We employed the June 2023 edition of GPT-4, named as `gpt-4-0613`, with its default settings, including the `temperature` parameter set to 1. We employed the BLOOM version 176B parameters, coded as `bloom-176b`, and the FLAN-T5 version with 11.3B parameters, named as `flan-t5-xxl`. HuggingFace API interface was utilized to generate with these models with slight modifications to the default configurations: `return_full_text` was set to False, `no_repeat_ngram_size` was adjusted to 4, and `max_new_tokens` was limited to 100.

API calls were made in September 2023. The `openai` Python library was utilized to generate with GPT-4 while the `requests` Python library was facilitated to make calls to the HuggingFace API[7]. The models were utilized in accordance with their corresponding licenses and terms at the time of this study. OpenAI provides a Terms of Use[8]. BLOOM is authorized under BigScience RAIL License v1.0[9]. And FLAN-T5 authorized under Apache 2.0 license.

B Participant Demographic

While recruiting our participants, we have not placed any restrictions other than fluency in English and being older than 18 years old. As a result, we have attracted a wide range of participants in terms of demographic. The study involved individuals of various age groups, ranging from 18 to 58, including participants in their 20s, 30s, 40s, and 50s. The mean age of the participants was 29 years and 6 months, with the majority (14 individuals) falling into the 25-year-old category. Participants residing in 22 countries joined in our study including Austria, Canada, Czech Republic, Denmark, Estonia, Finland, France, Germany, Greece, Hungary, Iceland, Italy, Latvia, Mexico, Netherlands, New Zealand, Poland, Portugal, South Africa, Spain, United Kingdom, United States of America. However, half of the participants (60 individuals) were residing in South Africa.

C Correlation Analysis

We computed Pearson correlation coefficients to examine the linear relationships between each pair of four criteria. There was a positive correlation for all combinations; appropriateness and specificity ($r(186) = 0.63, p < 0.001$), appropriateness and naturalness ($r(186) = 0.41, p < 0.001$), appropriateness and engagement ($r(186) = 0.51, p < 0.001$), specificity and naturalness ($r(186) = 0.31, p < 0.001$), specificity and engagement ($r(186) = 0.51, p < 0.001$), naturalness and engagement ($r(186) = 0.41, p < 0.001$).

[7] https://api-inference.huggingface.co
[8] https://openai.com/policies/terms-of-use
[9] https://huggingface.co/spaces/bigscience/license