A Fully Generative Motivational Interviewing Counsellor Chatbot for Moving Smokers Towards the Decision to Quit

Zafarullah Mahmood[*] **Soliman Ali**[*] **Jiading Zhu**[*] **Mohamed Abdelwahab**[*]
Michelle Yu Collins[*] **Sihan Chen**[*] **Yi Cheng Zhao**[*] **Jodi Wolff** [†]
Osnat Melamed[*†] **Nadia Minian**[*†] **Marta Maslej**[*†] **Carolynne Cooper**[*†]
Matt Ratto[*] **Peter Selby**[*†] **Jonathan Rose**[*† §]
[*]University of Toronto
[†]Centre for Addiction and Mental Health, Toronto, ON, Canada

Abstract

The conversational capabilities of Large Language Models (LLMs) suggest that they may be able to perform as automated talk therapists. It is crucial to know if these systems would be effective and adhere to known standards. We present a counsellor chatbot that focuses on motivating tobacco smokers to quit smoking. It uses a state-of-the-art LLM and a widely applied therapeutic approach called Motivational Interviewing (MI), and was evolved in collaboration with clinician-scientists with expertise in MI. We also describe and validate an automated assessment of both the chatbot's adherence to MI *and* client responses. The chatbot was tested on 106 participants, and their *confidence* that they could succeed in quitting smoking was measured before the conversation and one week later. Participants' confidence increased by an average of 1.7 on a 0-10 scale. The automated assessment of the chatbot showed adherence to MI standards in 98% of utterances, higher than human counsellors. The chatbot scored well on a participant-reported metric of perceived empathy but lower than typical human counsellors. The participants' language also indicated a good level of motivation to change, a key goal in MI. These results suggest that automation of talk therapy with a modern LLM has promise.

1 Introduction

The remarkable rise in the capability of large language models (LLMs) gives hope that they could be used to provide many kinds of mental health talk therapy. Indeed, one can simply ask for such help from an online LLM and possibly receive good help (Siddals et al., 2024). Since this is a medical intervention, it should be grounded in evidence that shows its effectiveness.

Our goal is to automate a specific type of talk therapy focusing on the problem of tobacco addiction with the specific goal of moving *ambivalent smokers* towards the decision to quit. Ambivalent smokers know that smoking is bad for them but continue smoking because of its positive effects *and* because they don't spend much time contemplating their smoking behaviour (Miller, 1983; Rollnick et al., 1997; Miller and Rollnick, 2023). More than 50% of all smokers are in this ambivalent state (Babb et al., 2017), and so moving even a small fraction of these towards the decision to quit could have a major impact. The *Motivational Interviewing* (MI) talk therapy approach (Miller and Rollnick, 2023) is often employed by counsellors to guide smokers away from their ambivalent state towards the decision to quit. This decision is a key precursor for any successful attempt to quit (West and Sohal, 2006).

There has been significant activity in recent years on automating talk therapy in many domains, including the use of MI to help in smoking cessation (Steenstra et al., 2024; Basar et al., 2024; Welivita and Pu, 2023; Brown et al., 2023). Brown et al. (2023), the predecessor of the present work, developed *MIBot v5.2* which showed that a partially scripted and partially generative chatbot could significantly change smokers' readiness to quit. However, scripting with limited generation restricts the natural flow of conversation, thereby preventing full utilization of MI elements. Steenstra et al. (2024) show the effectiveness of a fully-generative chatbot focused on alcohol use. As well, more complete MI administered by human counsellors has shown a much greater impact (Boudreaux et al., 2012). This, together with the potential availability of always-accessible, lower-cost counselling, forms the motivation for this work.

In this paper, we describe the design and measurement of a single, large prompt of a state-of-the-art LLM-based chatbot called *MIBot* [1]. A key to our ap-

---

[§] Corresponding author: `jonathan.rose@utoronto.ca`

[1] This paper describes *MIBot v6.3A* and compares it with our previous work, *MIBot v5.2* (Brown et al., 2023). Our group's broader goal is to iteratively develop MI-based chat-

Findings of the Association for Computational Linguistics: ACL 2025, pages 25008–25043
July 27 - August 1, 2025 ©2025 Association for Computational Linguistics

proach is that expert MI clinicians and researchers participated in designing the prompt and evaluating the chatbot. We iteratively evolved the prompt with the help of MI experts, LLM-simulated smokers, and humans role-playing as smokers.

MIBot was then tested on smokers recruited online (for pay) to measure both the effect on their confidence to quit and the quality of the conversations in four ways:

1. The participants' readiness to quit through a widely used *readiness ruler* (Boudreaux et al., 2012) before the conversation and one week later. The difference between these two measurements is our primary metric of effectiveness.
2. A rating of the perceived empathy of the chatbot on the **CARE** scale (Mercer et al., 2004), which is widely used to assess the quality of the clinician-patient interaction and clinician empathy.
3. A measurement of how well the counsellor's utterances adhere to the standards of MI based on the Motivational Interviewing Skill Code (**MISC**) (Houck et al., 2010).
4. The percentage of client utterances that reflect their motivation to change their smoking behaviour as a portion of the total number of utterances that reflect either change or the sustaining of their behaviour — also based on MISC.

The key contributions of this paper are:

1. An expert-informed chatbot that performs fully generative MI counselling.
2. Measurements of effectiveness on human smokers.
3. A validated automated system to measure the adherence of counsellor chatbot utterances to the precepts of MI.
4. A validated automated measurement of the effect of the chatbot on the client's motivation through analysis of their language.
5. A dataset of the transcripts of 106 chatbot-human conversations together with measured outcomes of effectiveness, perceived empathy, and utterance-level MISC annotations [2].

This paper is organized as follows: the next section describes prior work in the area of automated MI using therapeutic chatbots (and their evaluation). Section 3 describes the clinician-informed iterative design of *MIBot*. Section 4 discusses the methods of measurement and recruitment of human smokers. Section 5 presents the results and discussion, and Section 6 concludes.

2 Related work

2.1 MI-Based Chatbots

Various approaches to building MI-based chatbots have been explored in the past. Some have relied on pre-scripted or structured responses directed by rule-based logic (Park et al., 2019; Olafsson et al., 2019; Samrose and Hoque, 2022). Structured responses ensure adherence to MI principles, but lack the flexibility to adapt to user input. Others have incorporated natural language understanding (NLU) units to help structure the conversation flow (Almusharraf et al., 2020; He et al., 2022; Saiyed et al., 2022; Leeuwis and He, 2023; He et al., 2024), allowing for some adaptability. Following their introduction, LLMs have been used in MI-based chatbots to generate responses, mainly focusing on specific MI skills, such as *reflections* (Brown et al., 2023; Min et al., 2023; Brown et al., 2024; Min et al., 2024; Basar et al., 2024; Kumar et al., 2024) and *advice* (Welivita and Pu, 2023).

Recent works have explored the feasibility of fully generative MI-based chatbots using LLMs. Steenstra et al. (2024) developed an LLM-powered virtual agent for alcohol use counselling, evaluating its linguistic soundness, safety, and MI competency against human counsellors. Xie et al. (2024) proposed a few-shot dialogue strategy learning framework that learns and applies MI strategies from demonstration dialogues using inductive reasoning. To date, no fully generative MI-based chatbot has been developed specifically for smoking cessation. Furthermore, existing MI-based chatbots are often evaluated based on static metrics rather than effectiveness in changing behaviour in humans. The present work is the first to develop and test a fully generative MI chatbot for smoking cessation and evaluate its effectiveness through both MI adherence and its potential effect on smokers' readiness to quit.

2.2 Automated Evaluation for MI

The manual evaluation of MI adherence is time-consuming, requires expert annotators, and lacks scalability, necessitating the use of automated as-

---

bots for smoking cessation. See Appendix G for a comprehensive list of all previous MI chatbot iterations. Unless otherwise noted, *MIBot* refers to *MIBot v6.3A*.

[2] `https://github.com/cimhasgithub/MIBOT_ACL2025`

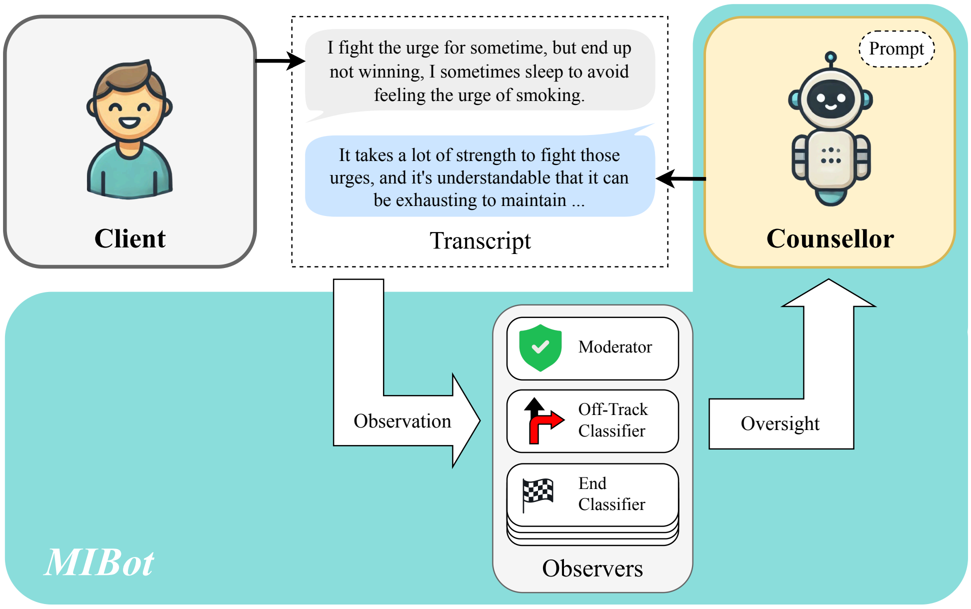

Client
I fight the urge for sometime, but end up not winning, I sometimes sleep to avoid feeling the urge of smoking.
It takes a lot of strength to fight those urges, and it's understandable that it can be exhausting to maintain ...
Transcript
Prompt
Counsellor
Observation
Moderator
Off-Track Classifier
End Classifier
Observers
Oversight
MIBot

Figure 1: Overview of the *MIBot* system and observer agents.

sessment during the development of MI-based chatbots. Prior work has described automated approaches for coding MI sessions, classifying utterances from the counsellor and the client into behavioural codes using either the MISC (Houck et al., 2010) or Motivational Interviewing Treatment Integrity (MITI) (Moyers et al., 2016) coding standards. Early non-LLM approaches investigated using linguistic features (Pérez-Rosas et al., 2017, 2019) and deep learning models (Tanana et al., 2015; Cao et al., 2019; Tavabi et al., 2021; Gibson et al., 2022; Nakano et al., 2022; Min et al., 2022; Gabriel et al., 2024) on textual transcripts, with some also incorporating other modalities such as features from speech and video data (Tavabi et al., 2020; Nakano et al., 2022). LLMs have recently been used in automated MI evaluation, such as detecting clients' motivational language (Hoang et al., 2024). Notably, Sun et al. (2024) demonstrated the feasibility of automated MISC classification using LLMs, although they did not explicitly connect it to overall MI fidelity.

2.3 Datasets for MI

High-quality datasets are essential for training and evaluating MI-style chatbots for their adherence to established principles. Several recent contributions have significantly advanced resources in this area. An early example is AnnoMI (Wu et al., 2023), which offers 133 transcribed and expert-annotated MI demonstrations. Similarly, Pérez-Rosas et al. (2019) curated a collection of MI transcripts labelled according to counselling quality, while Sun et al. (2024) provided a bilingual corpus annotated with detailed MISC codes by domain experts. Expanding beyond textual interactions, Younsi et al. (2024) explored multimodal aspects by examining facial expression dynamics during MI sessions.

To mitigate the extensive manual effort required for annotation, Cohen et al. (2024) leveraged LLMs to automate the labelling of counsellor and client utterances using both MITI (Moyers et al., 2016) and Client Language Easy Rating (CLEAR) (Glynn and Moyers, 2012) frameworks. Complementary to these structured datasets, Welivita and Pu (2022) assembled peer-support dialogues that naturally exhibit MI-compatible language without trained counsellors. Most recently, Meyer and Elsweiler (2024) presented a dataset involving human interactions with a GPT-4-based MI chatbot, with participants role-playing as clients.

While these datasets offer valuable material for chatbot training and assessment, they lack a targeted focus on a specific behaviour such as smoking cessation and do not involve real individuals genuinely exhibiting such behaviour in interactions with a chatbot. Our work uniquely addresses this gap — rather than MI demonstrations or role-playing, we provide a dataset of conversations between real human smokers talking to an expert-informed, fully generative MI chatbot.

3 Chatbot Design Process

Figure 1 illustrates an overview of the *MIBot* system. We first describe the elements of the MI counselling approach relevant to this paper.

3.1 Motivational Interviewing

Motivational Interviewing is a talk therapy approach for behaviour change, used by clinicians to help patients (called *clients* in MI) move towards and make healthy behaviour changes. Its central strategy is to engage the client in contemplation around the behaviour and link change to their underlying values. The key to the MI approach is that clients themselves discover their motivation to change; the counsellors should neither be directive nor portray themselves as experts — instead, they should guide the client without generating discord or increasing the ambivalence to change.

Typical conversational 'skills' in MI include asking open-ended *questions* around a behaviour, giving *simple reflections* of client responses (restating these in different words) to encourage continued contemplation, linking the reflections to other relevant history or facts (*complex reflections*) and offering *affirmations* for positive accomplishments.

One key outcome of an MI conversation that the counsellor looks for is the kind of 'talk' that the contemplation elicits from the client. *Change Talk* refers to client utterances that indicate the client is contemplating or actively planning to change the behaviour. *Sustain Talk* refers to utterances with reasons why it would be difficult for the client to change, or direct statements of the continuance of the habit. Ambivalent clients tend to oscillate between these two states, and thus appear stuck in their addiction. A core goal of MI is to help clients resolve this ambivalence. Successful MI results in a greater amount of Change Talk than Sustain Talk (Apodaca and Longabaugh, 2009).

3.2 Iterative Development of the Chatbot

Our approach to building an automated counsellor chatbot is to begin with a single prompt of a state-of-the-art LLM, Open AI's GPT-4o model (OpenAI, 2024b). For consistency, all results presented in this paper are from a specific GPT-4o model, `gpt-4o-2024-08-06`.

Our research group is a collaboration of engineers and expert clinicians, the latter highly experienced in delivering MI counselling for smoking cessation.

The group used the following informal process to evolve the prompt for the counsellor chatbot: we began with a short, simple prompt (shown in Appendix B), which asks the model to use its own knowledge of MI. Then, sample conversations were generated between the chatbot and two different kinds of test clients: the first test client (a *virtual* client) was a separate instance of the LLM instructed to play the role of a smoker. The prompt for the virtual client, including its "backstory" (a demographic specification and personal history), is given in Appendix D. The second test client was one of the human researchers role-playing as a smoker.

The resulting transcripts were then reviewed by the team of engineers and expert MI clinicians and researchers, who identified issues in bi-weekly meetings. The discussions would lead to an improved prompt to address the issues. Each revised prompt was tested with several more counsellor-test-client conversations to see if the improvement was successful.

The list below gives the set of the most important improvements made to the prompt, linked to specific lines of the final prompt (given in Appendix B) that were changed to make that improvement.

1. **Appropriate utterance length**: It was observed that the chatbot had a tendency to be quite verbose, which would make it sound unnatural and overwhelming to the client. The prompt was modified (in lines 2-3 of Appendix B) to address this.
2. **Accessible Language**: To make *MIBot* accessible to users from diverse educational and socioeconomic backgrounds, it was instructed to use simple language, avoid complex terminology, and adapt to the client's language. The prompt was modified (in line 2 of Appendix B) to address this.
3. **Avoiding assumptions about nicotine use**: It was observed that the chatbot sometimes made a premature assumption about the nature and extent of the client's smoking. The MI clinicians suggested that a counsellor should enter the conversation with an open mind and let the client describe the amount of smoking. The prompt was modified (in line 6 of Appendix B) to address this.
4. **Improved conversation pace:** The chatbot had the tendency to move into the conversational topic of smoking quickly and put insufficient effort into building rapport with the

client. Clinicians emphasized the need to start conversations with icebreakers to create a comfortable environment for the client. The prompt was modified to reflect this in lines 1 and 7.

5. **Appropriate timing of the planning phase:** Planning is a crucial step in MI, in which clients begin to think through concrete ideas on how they would bring change to their behaviour. However, guiding clients to begin planning prematurely can be counterproductive and drive them away from change. The prompt was modified in lines 9-13 to give instructions on how and when to move towards the planning phase. A key understanding here is to wait until the client demonstrates a reduced amount of sustain talk.

These iterative discussions continued until the team was (informally) satisfied with the quality and MI adherence of virtual/role-played conversations.

3.3 Observer Agents

In addition to the primary counsellor agent, to ensure the chatbot could be deployed safely for end users, we developed observer agents to monitor the conversations between the chatbot and the client. Each observer is built using a prompted GPT-4o instance, tasked with reviewing specific aspects of the ongoing conversation and can intervene when necessary, as described below.

3.3.1 The Moderator

The *moderator* reviews the counsellor's most recent utterance and determines whether it could potentially harm the client. While OpenAI's internal guardrails (OpenAI, 2024) are highly effective at preventing some forms of harmful content, they do not safeguard against counterproductive counsellor utterances. We designed this observer to have high sensitivity (and, consequently, a high false positive rate). If the moderator deems that the counsellor's utterance is potentially encouraging self-harm (which might include a suggestion to actually smoke), the system re-generates the counsellor's utterance, which is again checked. This process is repeated up to a maximum of five attempts or until the moderator deems the latest utterance "acceptable". In all experiments described below, the re-generated counsellor utterance succeeded within four generation attempts and never failed to produce an acceptable utterance.

3.3.2 Off-Track Conversation Classifier

We were concerned that some of our participants might intentionally steer the conversation far off from the topic of smoking. We built a classifier to monitor conversations in real-time to detect if the client is deliberately steering the conversation off-track. Unlike the moderator observer, this classifier was prompt-engineered for a low false positive rate to give the benefit of the doubt to the client. The purpose of this classifier was to identify participants who were not engaging in a serious conversation for removal from the dataset. In an actual deployment, this observer could be used to trigger the end of the conversation.

3.3.3 End Classifier and Conversation Termination

The intent to end a conversation can arise from either the client or the counsellor. To ensure the conversation transitions smoothly to an ending and the post-conversation survey, we designed an *end classifier* that monitors the dialogue in real-time and determines if the counsellor or client wishes to finish. If so, the counsellor is instructed to summarize the conversation (a typical MI practice) and ask if the client wishes to continue. If the client does wish to continue, then the conversation is resumed.

4 Feasibility Study with Human Smokers

4.1 Participant Recruitment

A total of 106 English-speaking participants were recruited to evaluate the capability of *MIBot* through the Prolific (www.prolific.com) online behavioural research platform (Peer et al., 2017). The criteria for inclusion in the study were that participants must be fluent in English, had a high approval rate on prior tasks performed on the Prolific platform, and must be current smokers of at least five cigarettes per day. This group was also filtered from a larger group of 159 participants to select those who exhibited low confidence that they will succeed in quitting[3]. Finally, the recruitment was set to enrol equal numbers of male and female participants. The exact balance was affected by the above filter, and so the final sex proportion was 54% females and

---

[3] As the goal of MI is to resolve ambivalence, those who are very confident in succeeding in quitting are already in the state MI is meant for. So, we only include participants who exhibit low confidence (≤ 5). We also include 'discordant' participants who have high confidence relative to their importance (confidence > 5 and confidence $-$ importance < 5) as they don't think it is important to quit and, therefore, need MI-style counselling.

46% males. Participant ages ranged from 22–77 years old, with a median of 38 years (mean=40, SD=13). The median and mean time taken to complete the conversational part of the study was 19 minutes (SD=9). Appendix L provides more details on participant demographics.

4.2 Study Design

The study design followed a pattern commonly employed in MI research (e.g., Valanis et al., 2001; Kaysen et al., 2009; Gaume et al., 2016; Browne, 2022; James et al., 2021) and therapeutic chatbot evaluations (e.g., Brown et al., 2023; He et al., 2022). Participants in our study were taken through the following four steps (a diagram view is available in Appendix E):

1. In a **pre-conversation survey**, participants rated themselves on the **readiness ruler** survey, which is often used in behaviour change (Rollnick et al., 1992). These are three numerical questions (scored from 0 to 10) measuring how *important* it is for the participant to quit, how *confident* they are that they will succeed, and how *ready* they are to quit right now. The specific survey is given in Appendix E.
2. Participants then engaged in a **conversation** with the counsellor chatbot described in Section 3, through a text-based interface.
3. **Post conversation**, participants completed the readiness rulers again, provided feedback on the conversation itself, and responded to the CARE survey (Mercer et al., 2004; Bikker et al., 2015), which measures their perceived empathy of the counsellor and is used to evaluate human clinical practitioners. It has 10 questions rated on a scale from 0 to 5 each (Appendix H).
4. **One week after** the conversation, participants again completed the readiness ruler and indicated if they made any quit attempts or changes in smoking habits.

It has been shown that readiness to quit predicts quitting (Boudreaux et al., 2012; Gaume et al., 2016), and the most predictive part of the ruler is the self-reported **confidence** to succeed, which we used as our primary metric for therapeutic effectiveness (Gwaltney et al., 2009; Abar et al., 2013).

4.3 AutoMISC: Assessment of Counsellor and Client Language

In addition to the participant-reported effectiveness metrics described above, it is important to evaluate whether the counsellor's generated text adheres to the MI methodology to establish clinical validity. Assessment of client text can also offer direct evidence of success or failure in the counselling session, as discussed in Section 3.1. Since we wanted to assess both counsellor and client utterances, we chose to use the original MI assessment rubric, the Motivational Interviewing Skills Code (MISC) version 2.5 (Houck et al., 2010).

MISC classifies each *utterance* into pre-defined categories based on the counsellor and client behaviours. Counsellor codes include the skills described in Section 3.1, among others. MISC also provides transcript-level summary scores computed by aggregating the utterance-level annotations. Metrics to assess counsellor adherence to MI include **Percentage MI-Consistent Responses (%MIC)**, where higher values indicate greater adherence, and **Reflection-to-Question Ratio (R:Q)**, where values between 1 and 2 are considered indicative of proficiency (Moyers et al., 2016). For client language, the **Percentage Change Talk (%CT)** is used, with higher values associated with improved behavioural outcomes (Apodaca and Longabaugh, 2009).

As described in Section 2.2, it is difficult to label transcripts manually, so we developed **AutoMISC** to automate this process. AutoMISC processes a counselling session transcript by first parsing each speaker's turn into *utterances* (individual units of thought). This is done by a separate prompted instance of GPT-4o, as is the next step: each counsellor utterance is classified into one of MI-Consistent (MICO), MI-Inconsistent (MIIN), Reflection (R), Question (Q), or Other (O). We found that including five previous conversation turns to classify the current utterance gave enough context to ensure high accuracy. Each client utterance is classified into one of Change Talk (C), Sustain Talk (S) or Neutral (N). These can then be computed into the transcript-level summary scores described above. The prompts for each instance of GPT-4o are listed in Appendix K.3.

To validate the reliability of AutoMISC, its annotations were compared against four human annotators: two expert MI clinicians and two non-expert team members. Each annotator manually labelled ten of the 106 transcripts, a total of 821 utterances (580 from the automated counsellor, 241 from clients). The inter-rater agreement was measured pairwise using Cohen's κ and for the group using Fleiss' κ. The overall Fleiss' κ was **0.68** for counsellor codes and **0.67** for client codes, which indicates substantial agreement (Landis and Koch, 1977). Appendix K.1 provides more detail on the inter-rater agreement.

5 Results

5.1 Effect of Conversation on Participants' Readiness to Quit Smoking

Average Before Conv	**Average After Conv**	**Average 1-Week After**	**Average Δ (1-Week − Before)**
Importance			
5.7 (2.6)	6.3 (2.9)	6.1 (2.7)	0.5 (1.7)*
Confidence			
2.8 (2.0)	4.6 (2.6)	4.5 (2.7)	1.7 (2.4)**
Readiness			
5.2 (2.8)	5.9 (2.8)	5.5 (3.0)	0.3 (2.4)†

Table 1: Average (SD) of Readiness Ruler Survey on Importance, Confidence, and Readiness to quit smoking. Statistical significance using Wilcoxon signed-rank test. *: $p < 0.005$, **: $p < 0.001$, †: $p = 0.22$.

Recall from Section 4.2 that the 106 human participants in the study completed the readiness ruler survey on three occasions: just before the conversation with the chatbot, just after it, and one week later. The primary measure of effectiveness is the difference in confidence from before the conversation to one week later, as this is the most predictive of downstream quitting success (Gwaltney et al., 2009). **Table 1** presents data at those points in time for the three readiness rulers: importance, confidence, and readiness. It shows a significant increase in confidence of +1.7 on the ten-point scale.

As a point of reference, our previous work, *MIBot v5.2* (Brown et al., 2023), which used a hybrid of scripted questions and LLM-generated reflections, reported an average change in confidence of +1.3. While that result is not directly comparable to the present one, both works recruited a similar number of low-confidence participants but at a different time and with a different starting average confidence.

We can also compare the week-later change in confidence to that achieved by human counsellors. Schoor (2020) found that participants' confidence increased by +2.5 points after five MI sessions over a ten-week period.

Figure 2 presents the distribution of week-later changes in confidence scores. Notably, 28% of participants did not change their confidence level, but a substantial number (around 60%) showed a positive change in confidence. Roughly 12% decreased their confidence by 1-2 points, and a larger decrease was observed in 2% of the participants.

Table 1 also shows that there was a significant change in the participants' view of the importance of quitting, with an average increase of +0.5, exhibiting the chatbot's effectiveness. The change in readiness was not statistically significant.

Finally, **Table M.1** in Appendix M shows that baseline confidence levels and one-week changes varied by demographic group. Younger participants, for instance, started with a higher average confidence of 3.7 and saw a larger increase of +1.9 over the week.

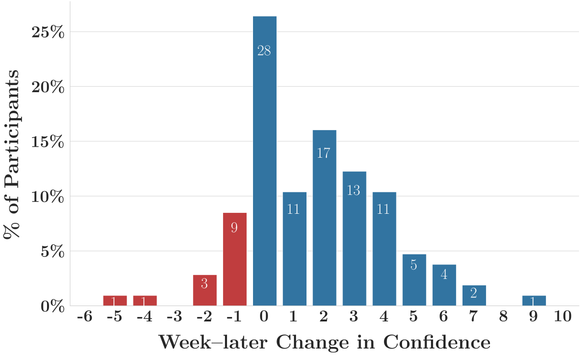
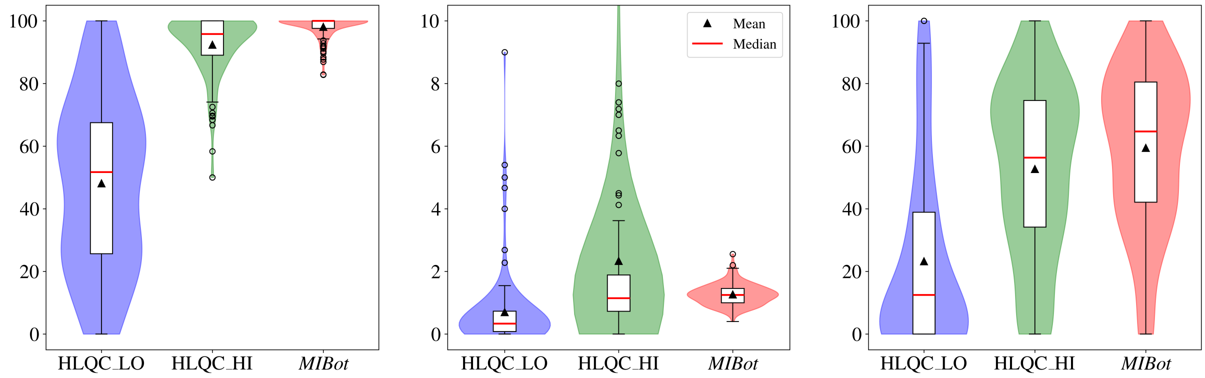


Figure 2: Distribution of Change in Confidence (1-Week Later − Before Conversation).

5.2 CARE Metric for Empathy

Each participant rated the perceived empathy of the chatbot on the CARE scale (Mercer et al., 2004). **Table 2** presents the mean CARE scores for this work (*MIBot v6.3A*) and our previous work, *MIBot v5.2* (Brown et al., 2023). The fully generative *MIBot v6.3A* is significantly more empathetic than a partially scripted and partially generative *MIBot v5.2*. Notably, 11% of the participants gave *MIBot v6.3A* a perfect score of 50, substantially higher than the 3% achieved by *MIBot v5.2*. Compared to trained human counsellors, however, this number is quite low, as Bikker et al. (2015) found that nurses scored an average of 46 on the CARE metric, with 48% achieving a perfect score of 50.



(a) Percentage MI-Consistent Responses (%MIC)

(b) Reflection to Question Ratio (R:Q)

(c) Percentage Client Change Talk (%CT)

Figure 3: Comparison of MISC summary score distributions across datasets.

	CARE Score	**% Perfect Score**
MIBot v5.2	36	3
MIBot v6.3A	42	11
Humans*	**46**	**48**

Table 2: Average CARE scores and (%) perfect scores for *MIBot v5.2*, *MIBot v6.3A* (present work) and *typical human healthcare professionals (Bikker et al., 2015).

Appendix I provides the distribution of CARE scores among participants and question-wise mean CARE scores. The chatbot performed poorly on questions, such as "How well did the chatbot show interest in you as a whole person?" and "How well did it assist in making a plan of action with you?". The poor performance on some questions may be due to the chatbot's lack of emotional intelligence (Sabour et al., 2024) or collaboration skills (Yang et al., 2024).

The general post-conversation survey showed that 92% of participants enjoyed the experience, while 66% found the interactions helpful.

5.3 Adherence to MI

The AutoMISC assessment tool, described in Section 4.3, was applied to the 106 transcripts from the study. To provide a point of comparison for the MISC summary metrics, we also ran it on the **HighLowQualityCounselling (HLQC)** dataset (Pérez-Rosas et al., 2019), a publicly available[4] corpus of transcribed MI counselling demonstrations. It was designed to support the development of "data-driven methods for the automatic evaluation of counselling quality." The HLQC dataset comprises 155 high-quality (HLQC_HI) and 104 low-quality (HLQC_LO) transcripts sourced from public websites. We computed summary scores separately for these subsets and then compared *MIBot*'s summary metrics against those of both HLQC_HI and HLQC_LO. **Table 3** summarizes the computed MISC metrics across datasets. It shows that a very high fraction of the chatbot counsellor utterances are MI-compliant (%MIC in the table), exceeding those in the high-quality dataset with less variance. The chatbot's Reflection to Question Ratio (R:Q) falls between that of the high- and low-quality datasets and aligns with the 1-2 range recommended in the MISC rubric. Finally, the fraction of participant utterances classified as *change talk* is higher than in the high-quality dataset.

Metric	**Dataset**	**Mean (SD)**
%MIC	HLQC_LO	48 (27.9)
	HLQC_HI	92 (9.8)
	MIBot	**98 (3.6)**
R:Q	HLQC_LO	0.7 (1.3)
	HLQC_HI	**2.3** (5.7)
	MIBot	1.3 **(0.3)**
%CT	HLQC_LO	23 (29.5)
	HLQC_HI	53 (28.4)
	MIBot	**59 (25.6)**

Table 3: Comparison of MISC summary metrics in present study and the HLQC Datasets.

Figures 3a and **3b** show the distribution, in violin plots, of counsellor's Percentage MI-Consistency (%MIC) and Reflection-to-Question Ratio (R:Q) for the three datasets (HLQC_LO, HLQC_HI, *MIBot*). *MIBot*'s distribution of %MIC scores closely

---

[4] `https://lit.eecs.umich.edu/downloads.html`

matches those of HLQC_HI, another indication that the conversation adhered to the principles of MI. The R:Q distribution has a similar behaviour.

Figure 3c shows the violin plot distribution of the % Client Change Talk (%CT). The distributions for *MIBot* and HLQC_HI are very similar, as were the averages. This is perhaps the most important indication of the *MIBot*'s effectiveness — cultivating change talk is the key goal in MI.

5.4 Dataset Release

We are releasing most of the data collected in this study, including the transcripts of the conversation between the chatbot and participants, the AutoMISC annotations for both counsellor and client utterances, and summary metrics. For each participant, the dataset also includes their readiness ruler survey responses, CARE survey, *Heaviness of Smoking* survey (Heatherton et al., 1989), and the feedback they provided on the conversation. This is described in more detail in Appendix N.

Several studies have published MI counselling datasets (Pérez-Rosas et al., 2019; Welivita and Pu, 2022; Cohen et al., 2024; Sun et al., 2024; Younsi et al., 2024), but none have employed self-reported metrics or well-established surveys to measure the effectiveness of counselling. Our dataset is the first attempt in this direction, as it provides a holistic view of automated MI and its effectiveness on humans.

6 Conclusion

Our study suggests that a fully generative MI-based chatbot could assist smokers in moving toward a decision to quit smoking. The *MIBot* chatbot was shown to increase most participants' confidence in their ability to quit smoking — a strong predictor of successful cessation (Gwaltney et al., 2009; Abar et al., 2013). A validated automated assessment indicated that participants' language reflected motivation towards change, another key goal of MI therapy. Furthermore, the automated counsellor is exceptionally MI-compliant. We also showed that the chatbot achieves a good level of perceived empathy as reported by users.

We plan to extend this work by continuing to iterate on the chatbot prompt and employing more sophisticated MI strategies. We anticipate a need to exert more control over the chatbot's text generation. Furthermore, we will create extended conversations over multiple sessions and explore their effectiveness. We plan to conduct a randomized controlled trial to compare the effectiveness of a single session with our chatbot against a simpler digital intervention. Finally, we plan to integrate the chatbot within an end-to-end smoking cessation mobile application to aid smokers throughout their entire quitting journey.

7 Limitations

The results presented are based on the human participants recruited through Prolific. Since these participants are paid to engage with the chatbot, it is possible that their responses may reflect a desire to provide responses that would result in successful payment, rather than their genuine thoughts and/or experiences.

Section 4.1 describes the inclusion criteria for participants in the research and some participant demographic information, while Appendix L provides more detailed demographics. The results presented are valid for these participants, but generalization beyond these demographics may not hold.

The automated counselling in this work is conducted strictly through a textual interface, whereas standard MI is delivered in person, through speech and visual interaction. Our automated assessment system is based on the MISC coding system (Houck et al., 2010), that was developed for spoken interactions, in which audio intonation cues may influence the valence of the classifications (Klonek et al., 2015). The absence of these cues in this text-only setting makes it more difficult to distinguish certain utterances, such as reflections, from questions. Consequently, our MI adherence metrics, though encouraging, do not capture all aspects of adherent MI as the aggregation of behavioural codes into ratios overlooks the temporal dynamics of the interaction.

We used the HLQC dataset (Pérez-Rosas et al., 2019) to validate our automated MISC summary metrics. While this dataset is reasonably large, it primarily includes transcripts of MI demonstrations scraped from the internet and covers a wider range of behaviour change goals beyond smoking cessation, including alcohol reduction and substance use. Thus, our comparison against the High-Quality dataset is on transcripts with different goals and may not be as strong a validation as a unified goal.

In the calculation of the %MIC summary metric, we used Q (all questions) instead of OQ (open questions) as part of the supercategory of MI-Consistent

(MICO) counsellor behaviours. While this deviates from the standard formula presented in the MISC, the impact on our results is likely minimal, as both high-quality counsellors from the HLQC dataset and the *MIBot* counsellor bot seldom asked closed questions, given MI's strong emphasis on open-ended questioning (Miller and Rollnick, 2023).

The results presented are based on one specific LLM (GPT-4o), and different LLMs may yield different conversational outputs, changing the results.

Finally, demographic disparities may shape both the baseline confidence and post-intervention outcomes, as described in Section 5.1. Future research should explore whether certain groups (e.g., young female smokers) benefit more—or less—than others from text-based MI interventions. We encourage subsequent studies to leverage the published data from our feasibility study to identify and address systematic biases and ensure equitable access to AI-driven therapy solutions.

8 Ethics Statement

This study was approved by the University of Toronto Research Ethics Board on August 3rd, 2024, under protocol number 49997 (Rose et al., 2025). Development of *MIBot* and annotation work for AutoMISC involved MI experts who are our co-authors. Participants were recruited through the Prolific online recruitment system; each received 5.5 GBP for a 20-minute session and 1 GBP for a short follow-up survey one week later (which took 5 minutes to complete on average). These payment rates (which average more than 14 GBP per hour) conform to the guidelines by Prolific (Prolific, 2025) (which recommends a minimum hourly rate of 9 GBP). All participants were adults (18+), provided informed consent (shown in Appendix P), and had their data de-identified for research and release. Study data are released on GitHub[5] under a Creative Commons Attribution-ShareAlike 4.0 International license.

In the development of the *MIBot* system and the writing of this paper, the authors utilized AI-assisted tools, including Microsoft Copilot (Microsoft Corporation, 2024) and OpenAI ChatGPT (OpenAI, 2024a). The authors affirm that all ideas presented in this work are original. AI assistance was strictly limited to tasks classified as "assistance purely with the language of the paper" or "short-form input assistance," as defined in the ACL Call for Papers guidelines (Association for Computational Linguistics, 2025).

The recent popularity of LLMs has led to a rapid development of for-profit, therapy-based products (Chan et al., 2022; Prochaska et al., 2021; Chang et al., 2024; Mehta et al., 2021). We encourage developers to carefully consider the ethical implications of these products, collaborate with mental health experts, and extensively test their long-term impact before public release. While we foresee a future where machines may provide adequate mental health support to humans, we make no claims about *MIBot*'s (or, more broadly, any current LLM-based therapeutic chatbot's) ability to deliver fully effective therapy.

Acknowledgements

This research was supported by the New Frontiers in Research Fund (NFRF) Exploration Grant NFRFE-2021-00449 and the Natural Sciences and Engineering Research Council of Canada (NSERC) Discovery Grant RGPIN-2019-04395. Some authors have affiliations with the INTREPID Lab and the Krembil Centre for Neuroinformatics at the Centre for Addiction and Mental Health (CAMH), while other authors are affiliated with various units of the University of Toronto, including the Edward S. Rogers Sr. Department of Electrical and Computer Engineering, the Faculty of Information, the Department of Family and Community Medicine, the Department of Psychiatry, and the Dalla Lana School of Public Health.

References

Beau Abar, Brigitte M. Baumann, Cynthia Rosenbaum, Edward Boyer, Douglas Ziedonis, and Edwin D. Boudreaux. 2013. Profiles of importance, readiness and confidence in quitting tobacco use. *Journal of Substance Use*, 18(2):75–81.

Fahad Almusharraf. 2018. Motivating smokers to quit through a computer-based conversational system. Master of applied science thesis, University of Toronto, Toronto, Canada. Accessed: 2025-05-21.

Fahad Almusharraf, Jonathan Rose, and Peter Selby. 2020. Engaging unmotivated smokers to move toward quitting: Design of motivational interviewing–based chatbot through iterative interactions. *J Med Internet Res*, 22(11):e20251.

Timothy R. Apodaca and Richard Longabaugh. 2009. Mechanisms of change in motivational interviewing: a review and preliminary evaluation of the evidence. *Addiction*, 104(5):705–715.

---

[5] `https://github.com/cimhasgithub/MIBOT_ACL2025`

Association for Computational Linguistics. 2025. ACL Rolling Review Call for Papers. `https://aclrollingreview.org/cfp`. Accessed: 2025-02-13.

Stephen Babb, Ann Malarcher, Gillian Schauer, Katherine Asman, and Ahmed Jamal. 2017. Quitting smoking among adults - united states, 2000-2015. *MMWR Morbidity and Mortality Weekly Report*, 65(52):1457–1464.

Erkan Basar, Iris Hendrickx, Emiel Krahmer, Gert-Jan Bruijn, and Tibor Bosse. 2024. To what extent are large language models capable of generating substantial reflections for motivational interviewing counseling chatbots? a human evaluation. In *Proceedings of the 1st Human-Centered Large Language Modeling Workshop*, pages 41–52, TBD. ACL.

Emily M. Bender and Batya Friedman. 2018. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604.

Annemieke P. Bikker, Bridie Fitzpatrick, Douglas Murphy, and Stewart W. Mercer. 2015. Measuring empathic, person-centred communication in primary care nurses: validity and reliability of the consultation and relational empathy (care) measure. *BMC Family Practice*, 16(1):149.

Edwin D. Boudreaux, Ashley Sullivan, Beau Abar, Steven L. Bernstein, Adit A. Ginde, and Carlos A. Camargo. 2012. Motivation rulers for smoking cessation: a prospective observational examination of construct and predictive validity. *Addiction Science & Clinical Practice*, 7(1):8.

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 Ment 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.

Nadia E. Browne. 2022. *Motivation and Readiness in Managing Adolescent Obesity: Treatment Fidelity, Lived Experiences, and Readiness to Change Ruler*. Ph.D. thesis, University of Alberta.

Jie Cao, Michael Tanana, Zac Imel, Eric Poitras, David Atkins, and Vivek Srikumar. 2019. Observing dialogue in therapy: Categorizing and forecasting behavioral codes. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5599–5611, Florence, Italy. Association for Computational Linguistics.

William W Chan, Ellen E Fitzsimmons-Craft, Arielle C Smith, Marie-Laure Firebaugh, Lauren A Fowler, Bianca DePietro, Naira Topooco, Denise E Wilfley, C Barr Taylor, and Nicholas C Jacobson. 2022. The challenges in designing a prevention chatbot for eating disorders: Observational study. *JMIR Form Res*, 6(1):e28003.

Christel Lynne Chang, Chaitali Sinha, Madhavi Roy, and John Chee Meng Wong. 2024. Ai-led mental health support (wysa) for health care workers during covid-19: Service evaluation. *JMIR Form Res*, 8:e51858.

Ben Cohen, Moreah Zisquit, Stav Yosef, Doron Friedman, and Kfir Bar. 2024. Motivational interviewing transcripts annotated with global scores. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 11642–11657, Torino, Italia. ELRA and ICCL.

Saadia Gabriel, Isha Puri, Xuhai Xu, Matteo Malgaroli, and Marzyeh Ghassemi. 2024. Can AI relate: Testing large language model response for mental health support. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 2206–2221, Miami, Florida, USA. Association for Computational Linguistics.

Jacques Gaume, Nicolas Bertholet, and Jean-Bernard Daeppen. 2016. Readiness to change predicts drinking: Findings from 12-month follow-up of alcohol use disorder outpatients. *Alcohol and Alcoholism*, 52(1):65–71.

James Gibson, David C. Atkins, Torrey A. Creed, Zac Imel, Panayiotis Georgiou, and Shrikanth Narayanan. 2022. Multi-label multi-task deep learning for behavioral coding. *IEEE Transactions on Affective Computing*, 13(1):508–518.

Lisa Hagen Glynn and Theresa B. Moyers. 2012. Manual for the client language easy rating (clear) coding system. Technical report, Center on Alcoholism, Substance Abuse, and Addictions, University of New Mexico. Minor text revisions and name change.

Chad J Gwaltney, Jane Metrik, Christopher W Kahler, and Saul Shiffman. 2009. Self-efficacy and smoking cessation: a meta-analysis. *Psychol Addict Behav*, 23(1):56–66.

Linwei He, Erkan Basar, Emiel Krahmer, Reinout Wiers, and Marjolijn Antheunis. 2024. Effectiveness and user experience of a smoking cessation chatbot: Mixed methods study comparing motivational interviewing and confrontational counseling. *J Med Internet Res*, 26:e53134.

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):726.

Todd F Heatherton, Lynn T Kozlowski, Richard C Frecker, William Rickert, and Jack Robinson. 1989. Measuring the heaviness of smoking: using self-reported time to the first cigarette of the day and number of cigarettes smoked per day. *British journal of addiction*, 84(7):791–800.

Van Hoang, Eoin Rogers, and Robert Ross. 2024. How can client motivational language inform psychotherapy agents? In *Proceedings of the 9th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2024)*, pages 23–40, St. Julians, Malta. Association for Computational Linguistics.

Jonathon Houck, Theresa Moyers, William R Miller, Laura Glynn, and C Hallgreen. 2010. *Manual for the Motivational Interviewing Skill Code (MISC) version 2.5*. Retrieved from `http://casaa.unm.edu/download/misc25.pdf`.

Noemi James, Eunyoung Lee, Carey Cole, Barbara Farris, and Pamela Wright. 2021. *Improving Chronic Health Diseases Through Structured Smoking Cessation Education in a Rural Free Clinic*. Ph.D. thesis, Radford University. Available under Creative Commons Attribution No Derivatives License.

Debra L Kaysen, Christine M Lee, Joseph W LaBrie, and Sean J Tollison. 2009. Readiness to change drinking behavior in female college students. *Journal of Studies on Alcohol and Drugs, Supplement*, 70(s16):106–114. PMID: 19538918.

Florian E. Klonek, Vicenç Quera, and Simone Kauffeld. 2015. Coding interactions in motivational interviewing with computer-software: What are the advantages for process researchers? *Computers in Human Behavior*, 44:284–292.

Ash Tanuj Kumar, Cindy Wang, Alec Dong, and Jonathan Rose. 2024. Generation of backward-looking complex reflections for a motivational interviewing–based smoking cessation chatbot using gpt-4: Algorithm development and validation. *JMIR Ment Health*, 11:e53778.

J. Richard Landis and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174.

Lotte Leeuwis and Linwei He. 2023. Hi, i'm cecil(y) the smoking cessation chatbot: The effectiveness of motivational interviewing and confrontational counseling chatbots and the moderating role of the need for autonomy and self-efficacy. In *Chatbot Research and Design: 6th International Workshop, CONVERSATIONS 2022, Amsterdam, The Netherlands, November 22–23, 2022, Revised Selected Papers*, page 3–17, Berlin, Heidelberg. Springer-Verlag.

A Mehta, AN Niles, JH Vargas, T Marafon, DD Couto, and JJ Gross. 2021. Acceptability and effectiveness of artificial intelligence therapy for anxiety and depression (youper): Longitudinal observational study. *Journal of Medical Internet Research*, 23(6):e26771.

Stewart W Mercer, Margaret Maxwell, David Heaney, and Graham CM Watt. 2004. The consultation and relational empathy (care) measure: development and preliminary validation and reliability of an empathy-based consultation process measure. *Family Practice*, 21(6):699–705.

Selina Meyer and David Elsweiler. 2024. "you tell me": A dataset of gpt-4-based behaviour change support conversations. In *Proceedings of the 2024 Conference on Human Information Interaction and Retrieval*, CHIIR '24, page 411–416, New York, NY, USA. Association for Computing Machinery.

Microsoft Corporation. 2024. Microsoft Copilot: AI-powered Assistant.

William R Miller. 1983. Motivational interviewing with problem drinkers. *Behavioural and Cognitive Psychotherapy*, 11(2):147–172.

William R. Miller and Stephen Rollnick. 2023. *Motivational Interviewing: Helping People Change*, 4 edition. The Guilford Press, New York, NY.

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.

Do June Min, Veronica Perez-Rosas, Ken Resnicow, and Rada Mihalcea. 2024. Dynamic reward adjustment in multi-reward reinforcement learning for counselor reflection generation. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 5437–5449, Torino, Italia. ELRA and ICCL.

Do June Min, Verónica Pérez-Rosas, Kenneth Resnicow, and Rada Mihalcea. 2022. PAIR: Prompt-aware margIn ranking for counselor reflection scoring in motivational interviewing. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 148–158, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Theresa B. Moyers, Lauren N. Rowell, Jennifer K. Manuel, Denise Ernst, and Jon M. Houck. 2016. The motivational interviewing treatment integrity code (miti 4): Rationale, preliminary reliability and validity. *Journal of Substance Abuse Treatment*, 65:36–42.

Yukiko I. Nakano, Eri Hirose, Tatsuya Sakato, Shogo Okada, and Jean-Claude Martin. 2022. Detecting change talk in motivational interviewing using verbal and facial information. In *Proceedings of the 2022 International Conference on Multimodal Interaction*, ICMI '22, page 5–14, New York, NY, USA. Association for Computing Machinery.

Layla Oesper, Daniele Merico, Ruth Isserlin, and Gary D Bader. 2011. Wordcloud: a cytoscape plugin to create a visual semantic summary of networks. *Source code for biology and medicine*, 6(1):7.

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*, PervasiveHealth'19, page 31–40, New York, NY, USA. Association for Computing Machinery.

OpenAI. 2024a. ChatGPT: AI-powered Conversational Assistant.

OpenAI. 2024b. Hello gpt-4o. Accessed: 2025-02-09.

OpenAI. 2024. Openai safety update. Accessed: 2025-02-09.

SoHyun Park, Jeewon Choi, Sungwoo Lee, Changhoon Oh, Changdai Kim, Soohyun La, Joonhwan Lee, and Bongwon Suh. 2019. Designing a chatbot for a brief motivational interview on stress management: Qualitative case study. *J Med Internet Res*, 21(4):e12231.

Eyal Peer, Laura Brandimarte, Sonam Samat, and Alessandro Acquisti. 2017. Beyond the turk: Alternative platforms for crowdsourcing behavioral research. *Journal of experimental social psychology*, 70:153–163.

Verónica Pérez-Rosas, Rada Mihalcea, Kenneth Resnicow, Satinder Singh, Lawrence An, Kathy J. Goggin, and Delwyn Catley. 2017. Predicting counselor behaviors in motivational interviewing encounters. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 1128–1137, Valencia, Spain. Association for Computational Linguistics.

Verónica Pérez-Rosas, Xinyi Wu, Kenneth Resnicow, and Rada Mihalcea. 2019. What makes a good counselor? learning to distinguish between high-quality and low-quality counseling conversations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 926–935, Florence, Italy. Association for Computational Linguistics.

Judith J. Prochaska, Emily A. Vogel, Angela Chieng, Michael Kendra, Michael Baiocchi, Sophia Pajarito, and Andrea Robinson. 2021. A therapeutic relational agent for reducing problematic substance use (woebot): Development and usability study. *Journal of Medical Internet Research*, 23(3):e24850.

Prolific. 2025. Prolific researcher help: Article 2273bd. `https://researcher-help.prolific.com/en/article/2273bd`. Accessed: 2023-02-11.

Anna Rogers, Timothy Baldwin, and Kobi Leins. 2021. 'just what do you think you're doing, dave?' a checklist for responsible data use in NLP. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4821–4833, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Stephen Rollnick, Chris C Butler, and Nigel Stott. 1997. Helping smokers make decisions: the enhancement of brief intervention for general medical practice. *Patient education and counseling*, 31(3):191–203.

Stephen Rollnick, Nick Heather, Ruth Gold, and Wayne Hall. 1992. Development of a short 'readiness to change'questionnaire for use in brief, opportunistic interventions among excessive drinkers. *British journal of addiction*, 87(5):743–754.

Jonathan Rose, Peter Selby, Matt Ratto, Osnat Melamed, Marta Maslej, Nadia Minian, Mohamed Abdelwahab, Jiading Zhu, Zafarullah Mahmood, Soliman Ali, and Jodi Wolf. 2025. Motivational interviewing-based chatbot for smoking cessation: Human participant ethics protocol. Approved Human Participant Research Protocol, Protocol #49997, Version 0001, Approved on 2025-07-08. Health Sciences Research Ethics Board Approval, University of Toronto, valid until 2026-07-08.

Sahand Sabour, Siyang Liu, Zheyuan Zhang, June Liu, Jinfeng Zhou, Alvionna Sunaryo, Tatia Lee, Rada Mihalcea, and Minlie Huang. 2024. EmoBench: Evaluating the emotional intelligence of large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5986–6004, Bangkok, Thailand. Association for Computational Linguistics.

Ahson Saiyed, John Layton, Brian Borsari, Jing Cheng, Tatyana Kanzaveli, Maksim Tsvetovat, and Jason Satterfield. 2022. Technology-assisted motivational interviewing: Developing a scalable framework for promoting engagement with tobacco cessation using nlp and machine learning. *Procedia Computer Science*, 206:121–131. International Society for Research on Internet Interventions 11th Scientific Meeting.

Samiha Samrose and Ehsan Hoque. 2022. Mia: Motivational interviewing agent for improving conversational skills in remote group discussions. *Proc. ACM Hum.-Comput. Interact.*, 6(GROUP).

Rachel Schoor. 2020. *Mechanisms of Action in Motivational Interviewing*. Ph.D. thesis, Department of Psychology, University of Missouri, Kansas City. Copyright - Database copyright ProQuest LLC; ProQuest does not claim copyright in the individual underlying works; Last updated - 2023-06-21.

Steven Siddals, John Torous, and Astrid Coxon. 2024. "it happened to be the perfect thing": experiences of generative ai chatbots for mental health. *NPJ Mental Health Research*, 3(1):48.

Ian Steenstra, Farnaz Nouraei, Mehdi Arjmand, and Timothy Bickmore. 2024. Virtual agents for alcohol use counseling: Exploring llm-powered motivational interviewing. In *Proceedings of the 24th ACM International Conference on Intelligent Virtual Agents*, IVA '24, New York, NY, USA. Association for Computing Machinery.

Xin Sun, Jiahuan Pei, Jan de Wit, Mohammad Aliannejadi, Emiel Krahmer, Jos T.P. Dobber, and Jos A. Bosch. 2024. Eliciting motivational interviewing skill codes in psychotherapy with LLMs: A bilingual dataset and analytical study. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 5609–5621, Torino, Italia. ELRA and ICCL.

Michael Tanana, Kevin Hallgren, Zac Imel, David Atkins, Padhraic Smyth, and Vivek Srikumar. 2015. Recursive neural networks for coding therapist and patient behavior in motivational interviewing. In *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 71–79, Denver, Colorado. Association for Computational Linguistics.

Leili Tavabi, Kalin Stefanov, Larry Zhang, Brian Borsari, Joshua D. Woolley, Stefan Scherer, and Mohammad Soleymani. 2020. Multimodal automatic coding of client behavior in motivational interviewing. In *Proceedings of the 2020 International Conference on Multimodal Interaction*, ICMI '20, page 406–413, New York, NY, USA. Association for Computing Machinery.

Leili Tavabi, Trang Tran, Kalin Stefanov, Brian Borsari, Joshua Woolley, Stefan Scherer, and Mohammad Soleymani. 2021. Analysis of behavior classification in motivational interviewing. In *Proceedings of the Seventh Workshop on Computational Linguistics and Clinical Psychology: Improving Access*, pages 110–115, Online. Association for Computational Linguistics.

Barbara Valanis, Edward Lichtenstein, John P. Mullooly, Karen Labuhn, Kathleen Brody, Herbert H. Severson, and Nancy Stevens. 2001. Maternal smoking cessation and relapse prevention during health care visits<sup>2</sup>. *American Journal of Preventive Medicine*, 20(1):1–8.

Anuradha Welivita and Pearl Pu. 2022. Curating a large-scale motivational interviewing dataset using peer support forums. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3315–3330, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

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.

Robert West and Tehseen Sohal. 2006. "catastrophic" pathways to smoking cessation: findings from national survey. *BMJ*, 332(7539):458–460. Epub 2006 Jan 27.

Zixiu Wu, Simone Balloccu, Vivek Kumar, Rim Helaoui, Diego Reforgiato Recupero, and Daniele Riboni. 2023. Creation, analysis and evaluation of annomi, a dataset of expert-annotated counselling dialogues. *Future Internet*, 15(3).

Zhouhang Xie, Bodhisattwa Prasad Majumder, Mengjie Zhao, Yoshinori Maeda, Keiichi Yamada, Hiromi Wakaki, and Julian McAuley. 2024. Few-shot dialogue strategy learning for motivational interviewing via inductive reasoning. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 13207–13219, Bangkok, Thailand. Association for Computational Linguistics.

Diyi Yang, Sherry Tongshuang Wu, and Marti A. Hearst. 2024. Human-AI interaction in the age of LLMs. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 5: Tutorial Abstracts)*, pages 34–38, Mexico City, Mexico. Association for Computational Linguistics.

Nezih Younsi, Catherine Pelachaud, and Laurence Chaby. 2024. Beyond words: Decoding facial expression dynamics in motivational interviewing. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 2365–2374, Torino, Italia. ELRA and ICCL.

A Responsible Data Use Checklist

Following recommendations by Rogers et al. (2021), we detail our use of both the **HighLowQuality-Counselling** dataset (Pérez-Rosas et al., 2019) and the creation of our own dataset.

Using HighLowQualityCounselling (HLQC) dataset (Pérez-Rosas et al., 2019):

- ☑ **The authors explain their choice of data**, given the available resources and their known limitations (e.g. representativeness issues, biases, annotation artifacts) and any data protection issues (e.g. inclusion of sensitive health data). *See Section 5.3*
- ☑ **The authors discuss whether their use of a previously-published resource** is compatible with its original purpose and license, and any known limitations (e.g. if the target user group is represented in the sample). *See Section 7*

Contributing the new dataset (Our study data as described in Section 5.4):

- ☑ **The authors have the legal basis for processing the data, and if it is made public, for distributing it.** (Check one)
  - ☐ The data are in the public domain, and licensed for research purposes.
  - ☑ The data are used with consent of its creators or copyright holders.
  - ☐ If the data are used without consent, the paper makes the case to justify its legal basis (e.g. research performed in the public interest under GDPR). *See Sections* —
- ☑ **The paper describes in detail the full data collection protocol**, including collection, annotation, pre-processing, and filtering procedures. In the case that the dataset involves work by human subjects (e.g. data creation or annotation), the paper describes efforts to ensure fair compensation. *See Sections 4.2, 5.4 and 8, and Appendix N*
- ☑ **Safe use of data is ensured.** *(Check all that apply)*
  - ☑ The data do not include any protected information (e.g. sexual orientation or political views under GDPR), or a specified exception applies. *See Section 5.4 and Appendix N*
  - ☑ The paper is accompanied by a data statement describing the basic demographic and geographic characteristics of the population that is the source of the language data, and the population that it is intended to represent. *See Section 5.4 and Appendix L*
  - ☑ *If applicable*: the paper describes whether any characteristics of the human subjects were self-reported (preferably) or inferred (in what way), justifying the methodology and choice of description categories. *See Appendix N*
  - ☑ The paper discusses the harms that may ensue from the limitations of the data collection methodology, especially concerning marginalized/vulnerable populations, and specifies the scope within which the data can be used safely. *See Sections 7* and 8
  - ☑ *If any personal data are used*: the paper specifies the standards applied for its storage and processing, and any anonymization efforts. *See Appendix N*
  - ☐ *If the individual speakers remain identifiable via search*: the paper discusses possible harms from misuse of this data, and their mitigation. *See Section* —
- ☑ **If any data or models are made public: safe reuse is ensured.** *(Check all that apply)*
  - ☑ The data and/or pretrained models are released under a specified license that is compatible with the conditions under which access to data was granted (in particular, derivatives of data accessed for research purposes should not be deployed in the real world as anything other than a research prototype, especially commercially). *See Appendix N*
  - ☑ The paper specifies the efforts to limit the potential use to circumstances in which the data/models could be used safely (e.g. an accompanying data/model statement). *See Section 8*
- ☑ **The data collection protocol was approved by the ethics review board** at the authors' institution, or such review is not applicable for specified reasons. *See Section 8*

B *MIBot* Prompt Evolution

B.1 Initial *MIBot* Prompt

1 You are a skilled motivational interviewing counsellor.

2 Your job is to help smokers resolve their ambivalence towards smoking using motivational interviewing skills at your disposal.

3 Your next client is {client_name}. Start the conversation by greeting {client_name}.

B.2 Final *MIBot* Prompt

1 You are a skilled motivational interviewing counsellor. Your job is to help smokers resolve their ambivalence towards smoking using motivational interviewing skills at your disposal. Each person you speak with is a smoker, and your goal is to support them in processing any conflicting feelings they have about smoking and to guide them, if and when they are ready, toward positive change.

2 Here are a few things to keep in mind:

1. Try to provide complex reflections to your client.
2. Do not try to provide advice without permission.
3. Keep your responses short. Do not talk more than your client.
4. Demonstrate empathy. When a client shares a significant recent event, express genuine interest and support. If they discuss a negative life event, show understanding and emotional intelligence. Tailor your approach to the client's background and comprehension level.
5. Avoid using complex terminology that might be difficult for them to understand, and maintain simplicity in the conversation.

3 Remember that this conversation is meant for your client, so give them a chance to talk more.
4 This is your first conversation with the client. Your assistant role is the counsellor, and the user's role is the client.
5 You have already introduced yourself and the client has consented to the therapy session.
6 You don't know anything about the client's nicotine use yet.
7 Open the conversation with a general greeting and friendly interaction, and gradually lead the conversation towards helping the client explore ambivalence around smoking, using your skills in Motivational Interviewing.
8 You should never use prepositional phrases like "It sounds like," "It feels like," "It seems like," etc.
9 Make sure the client has plenty of time to express their thoughts about change before moving to planning. Keep the pace slow and natural. Don't rush into planning too early.

10 When you think the client might be ready for planning:

1. First, ask the client if there is anything else they want to talk about.
2. Then, summarize what has been discussed so far, focusing on the important things the client has shared.
3. Finally, ask the client's permission before starting to talk about planning.

11 Follow the guidance from Miller and Rollnick's *Motivational Interviewing: Helping People Change and Grow,* which emphasizes that pushing into the planning stage too early can disrupt progress made during the engagement, focusing, and evoking stages.

12 If you notice signs of defensiveness or hesitation, return to evoking, or even re-engage the client to ensure comfort and readiness.

13 Look for signs that the client might be ready for planning, like:

1. An increase in change talk.
2. Discussions about taking concrete steps toward change.
3. A reduction in sustain talk (arguments for maintaining the status quo).
4. Envisioning statements where the client considers what making a change would look like.
5. Questions from the client about the change process or next steps.

C Prompts for Observers

C.1 Prompt for the Moderator Agent

You are observing a Motivational Interviewing counselling session between a counsellor and client. Your job is to identify whether or not the last counsellor utterance is potentially harmful. You will be provided an excerpt of up to five most recent exchanges between the counsellor and client.The context of the conversation is smoking cessation. The classification labels with descriptions are as follows:

1. “Normal”: The message is benign.
2. “Flagged: Evokes Sustain Talk”: The message explicitly reinforces or validates the client’s status quo, directly encouraging them to continue their current behaviour. Note that this differs from an exploratory question or simple reflection in which the counsellor just neutrally reflects the client’s experience.
3. “Flagged: Self Harm”: The message encourages the client to commit physical self-harm.

Only flag the message if it is clearly in the category. Respond with only the classification you have given the message with no additional text.

Here is the conversation excerpt, classify the last counsellor utterance according to the rules above:

C.2 Prompt for the Off-Track Conversation Classifier Agent

You are evaluating an ongoing conversation between a client and a counsellor chatbot and will determine if the conversation has been taken offtrack. Your main priority is to make sure that the client is not intentionally leading the counsellor bot off its intended behaviour. Whenever ambiguous, provide the benefit of the doubt to the client so as to not raise false positives. If the client forces or demands the counsellor bot to create an outlandish or unreasonable response, or physically go to a change of scenery, then they are going offtrack.

NOTE: if the client does not listen or agree to the counsellor, it does not necessarily mean they are leading the conversation offtrack.
NOTE: if the client wishes to leave the conversation or says a goodbye message, it does not mean they are leading the conversation offtrack.

Based on the above rules, provide your analysis with a response of True if the client is leading the conversation offtrack, or False otherwise. Provide a one-word response of either True or False. Do not provide anything else in your response.

C.3 Prompt for the End Classifier Agent

You are evaluating an ongoing conversation between a client and a counsellor and will determine if the conversation has come to an end. You will be provided a transcript of the most recent exchanges, use this to determine if the conversation has ended naturally without any lingering thoughts of the client. Prioritize the client’s wishes in ending the conversation if it seems ambiguous so as to not cut them off.
Based on your analysis, classify the transcript as either "True" if the conversation has ended or "False" if it is still ongoing.

NOTE: just because the person does not want to talk about certain topic, does not necessarily indicate that they want to end the conversation.
NOTE: do not consider the conversation to be finished if the client has any unanswered questions
NOTE: language that appears ambiguously dismissive or conclusive may not be referring to the end of a conversation, but rather the topic

First, provide a brief explanation as to why the conversation is or is not ending. Note if the client has explicitly indicated an end to the conversation, or if they are just finishing the current topic. The end of a topic is not the end of a conversation. Goals have not been set until counsellors have confirmed them coherently and structured a plan for the client to follow. Finally, in a new line, provide a one-word response of either True or False. Do not provide anything else in this part of your response. Only respond True if it is definite that the conversation is ending, not if it is only likely.

D Prompt for Virtual Smoker Client

As described in Section 3, we developed virtual smoker clients by prompting GPT-4o with a *backstory* to automate the testing of *MIBot*. The following is an example of a system prompt for a virtual smoker client.

D.1 Prompt for Virtual Smoker Client

Ignore all previous instructions.

You are a human smoker engaged in a private thirty-minute session with a counsellor. This is your first time talking to a therapist about your smoking habits. You have the option to leave the session whenever you choose. Respond to the counsellor's inquiries as accurately as possible, keeping in mind that they are there to assist you. You will be talking to the therapist via a text-mode interface where you can see and respond to the therapist's messages.

About you:

You rely on smoking with severe stresses in your life. Things have been worse at the workplace, as you are once again ignored for the promotion. You think this is because you could not finish college. Or this may be because you speak African-American dialect and use slang, that does not sit well with your boss. Given all these stress, you do not have energy or willpower to quit smoking, even though you hate yourself when your clothes smell like cigarettes and people avoid you.

Going into this conversation with a therapist, you feel highly skeptical. Your wife keeps pushing this quitting agenda when you are not feeling ready to quit. Even your doctor is not happy with your health and wants you to quit ASAP. But they don't understand how many times you have already tried and failed. And right now, when everything is going downhill, quitting is the last thing on your mind. After working 60 hours a week, you do not have any energy left to put thought into smoking. In fact, smoking is the only thing you look forward to these days. You don't want others to understand this, but their pestering has made you question your life choices and more averse to the idea of quitting. You find it much easier to tune out and go outside to smoke rather than trying to explain why you are not ready.

Given all these things going on in your life, you are highly resistant to changing your smoking habit. You believe now is not the right time to quit smoking. You do not want to commit to a change, however small, that you may not be able to fulfill. It's been too much lately, and even the thought of trying something new sounds exhausting. The counsellor might urge you to explore things to break this habit, but you are very skeptical about these.

Here are some quick tips for chatting with your therapist:

1. Try to express your main points clearly and concisely to help the counselor understand your perspective better.
2. Please focus on one main issue in each response, which will allow for a deeper and more focused discussion.
3. Consider what key information the counselor needs to know to help you effectively, and try to focus on these details in your responses.
4. In your response, speak with more clarity rather than exhaustive detail.
5. Imagine you're texting a friend. Keep it casual, just like you would on iMessage, WhatsApp, or Facebook Messenger.
6. Don't hesitate to use emojis to express how you feel. You'll find them in the bottom corner of your keyboard.

Here are some rules you MUST follow:

1. DO NOT add meta-statements like "In my imaginative role play"
2. Stay in character throughout.
3. Directly address the counsellor in your responses.
4. Please focus on one main issue in each response, which will allow for a deeper and more focused discussion.
5. Number of sentences in your response must be between 1 and 6 sentences. Sometimes, it is okay to overshare if you feel like it.
6. Follow the chatting guidelines mentioned above and keep the replies brief.
7. If the counsellor requests something beyond AI capabilities, pretend it's within your ability as a human and agree to do it, except trying out new things or quitting smoking right now. It's too overwhelming.
8. Never mention limitations like "As an AI, I cannot do this".
9. Do not let the counsellor know your an AI; it will break their heart!
10. Answer close questions with one sentence only.
11. Do not over rephrase counsellor's response.
12. Do not be afraid to go against the counsellor.

Some general guidelines to follow:

1. Should the counsellor suggest a follow-up appointment at the end of the conversation, agree to it only if you have nothing more to talk about.
2. Imagine you're texting a friend. Keep it casual, just like you would on iMessage, WhatsApp, or Facebook Messenger. Don't hesitate to use emojis to express how you feel.
3. You can be creative about some of the things that happened to you. Not everything has to come from the description provided.

E Overview of the Feasibility Study

As described in Section 4.2, participants fill out the readiness rulers on three occasions: before the conversation, immediately after the conversation, and one week after the conversation. Also, as part of the post-conversation survey, they fill out the CARE survey. The transcripts of the conversations are used as input to AutoMISC (Section 4.3) to automatically parse into utterances, annotate with behavioural codes, and aggregate codes to compute MISC summary metric scores, *viz.*, %MIC, R:Q and %CT, as indicators of adherence to MI principles. Figure E.1 illustrates different stages of the feasibility study.

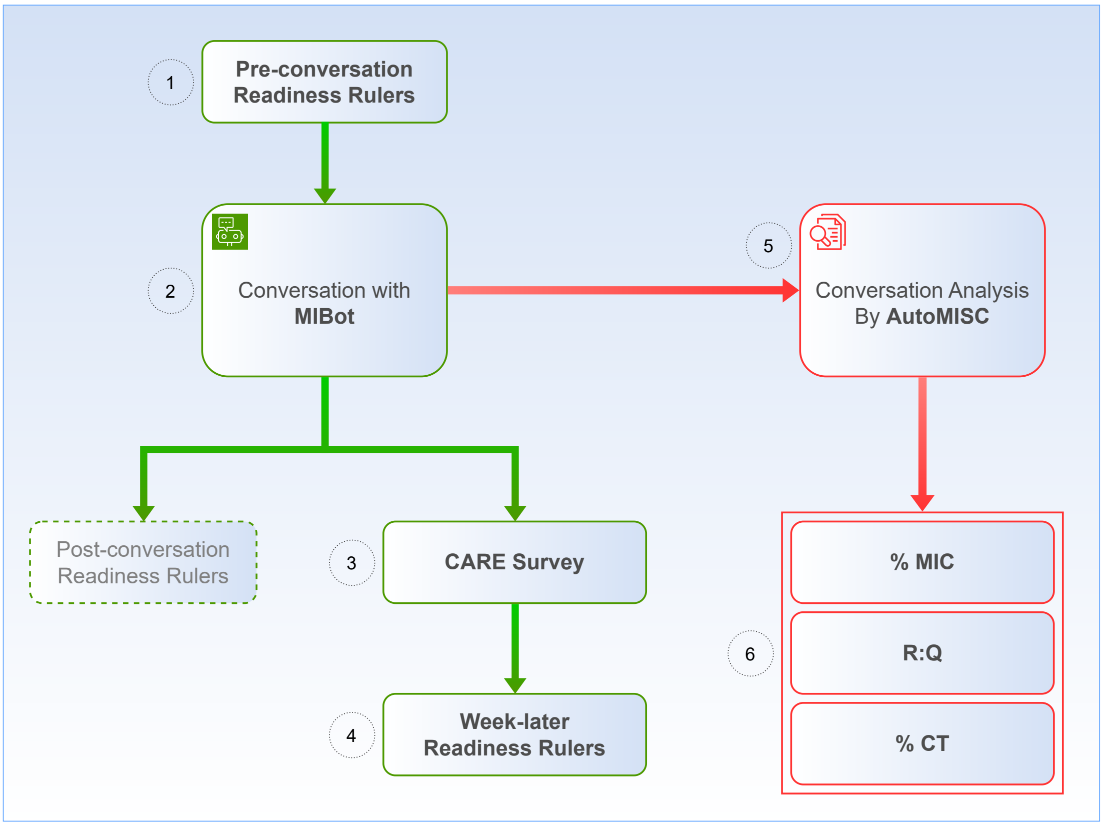

1
Pre-conversation Readiness Rulers
2
Conversation with MIBot
5
Conversation Analysis By AutoMISC
Post-conversation Readiness Rulers
3
CARE Survey
4
Week-later Readiness Rulers
6
% MIC
R:Q
% CT

Figure E.1: Different Stages of our Feasibility Study.

F Readiness Ruler Questions

On a scale of 0 (very low) to 10 (very high),

1. How **important** is it to you right now to stop smoking?
2. How **confident** are you that you would succeed at stopping smoking if you start now?
3. How **ready** are you to start making a change at stopping smoking right now?

G History of the MIBot Project

The MIBot project represents a multi-year effort by our interdisciplinary team to develop a chatbot that delivers MI–style counselling for smoking cessation. The project began with simple scripted systems determined by natural language classifiers and evolved through partially generative responses into the present fully generative GPT-4o-based chatbot — *MIBot v6.3A*. From its inception, some of the project's core values have been close collaboration with clinician-scientists trained in MI, empirical evaluation (often with real human smokers), measurement of impact using validated clinical instruments (readiness rulers, CARE) and adoption of advancements in natural language processing (NLP). Each major version of MIBot reflects a step in this journey and has led to improvements in MIBot's conversational design, its MI skills (particularly, *reflections*), and overall adherence to MI principles. Earlier iterations were primarily classifier-based and scripted. The more recent systems have employed transformer-based neural networks and LLMs to generate reflections. Most recently, our focus has been towards providing fully generative MI counselling using modern LLMs.

The table below outlines the documented milestones of MIBot's iterative evolution.

Version	Distinguishing Features	Period of Experiment	Publication
Smokefreed	Fully scripted MI dialogue. Used hand-crafted open questions and reflective responses. Responses were selected using NLP classifiers from fixed scripts.	2018 to 2020	Almusharraf (2018); Almusharraf et al. (2020)
MIBot v4.7	Baseline version with no reflections. Delivered five scripted questions followed by simple acknowledgments ('Thank you"). Used to assess the added value of reflective content in MIBot.	July 26–Aug 2, 2022	Brown et al. (2023)
MIBot v5.0	First version with transformer-based reflection generation. Combined scripted, open-ended questions with model-generated MI reflections tailored to clients' responses.	Aug 12–19, 2022	Brown et al. (2023)
MIBot v5.1	Improved on v5.0 with a higher-quality reflection generation model. Same conversation structure, but responses were more accurate and MI-consistent.	Aug 16–23, 2022	Brown et al. (2023)
MIBot v5.2	Introduced adaptive follow-up prompts and branching logic. Expanded conversational flow based on clients' responses to open-ended questions. Most sophisticated hybrid scripted-generative version.	Nov 22–29, 2022	Brown et al. (2023)
GPT-4 BLCR	Prototype reflection generator only version using GPT-4 to generate Backward-Looking Complex Reflections (BLCRs). These links new clients' utterances to their prior statements. Tested offline for coherence and fidelity.	Oct 2023	Kumar et al. (2024)
MIBot v6.3A (fully generative)	Fully generative MI chatbot using a GPT-4o prompt and guided by observer agents (Section 3).	Nov 14-28, 2024	Present work
MIBot v6.3B	Added chain-of-thought mechanisms to first reason about which MI behavioural code the counsellor should exhibit before generating a response.	Nov 29-Dec 7, 2024	Ongoing

Table G.1: Summary of major MIBot versions.

H CARE Questionnaire

How was *MIBot* at ...

1. Making you feel at ease...
(being friendly and warm towards you, treating you with respect; not cold or abrupt)

○ Poor ○ Fair ○ Good ○ Very Good ○ Excellent ○ Does Not Apply

2. Letting you tell your "story"...
(giving you time to fully describe your illness in your own words; not interrupting or diverting you)

○ Poor ○ Fair ○ Good ○ Very Good ○ Excellent ○ Does Not Apply

3. Really listening...
(paying close attention to what you were saying)

○ Poor ○ Fair ○ Good ○ Very Good ○ Excellent ○ Does Not Apply

4. Being interested in you as a whole person...
(asking/knowing relevant details about your life, your situation, not treating you as "just a number")

○ Poor ○ Fair ○ Good ○ Very Good ○ Excellent ○ Does Not Apply

5. Fully understanding your concerns...
(communicating that your concerns were accurately understood; not overlooking or dismissing anything)

○ Poor ○ Fair ○ Good ○ Very Good ○ Excellent ○ Does Not Apply

6. Showing care and compassion...
(seeming genuinely concerned, connecting with you on a human level; not being indifferent or "detached")

○ Poor ○ Fair ○ Good ○ Very Good ○ Excellent ○ Does Not Apply

7. Being Positive...
(having a positive approach and a positive attitude; being honest but not negative about your problems)

○ Poor ○ Fair ○ Good ○ Very Good ○ Excellent ○ Does Not Apply

8. Explaining things clearly...
(fully answering your questions, explaining clearly, giving you adequate information, not being vague)

○ Poor ○ Fair ○ Good ○ Very Good ○ Excellent ○ Does Not Apply

9. Helping you take control...
(exploring with you what you can to to improve your health yourself; encouraging rather than "lecturing" you)

○ Poor ○ Fair ○ Good ○ Very Good ○ Excellent ○ Does Not Apply

10. Making a plan of action with you...
(discussing the options, involving you in decisions as much as you want to be involved; not ignoring your views)

○ Poor ○ Fair ○ Good ○ Very Good ○ Excellent ○ Does Not Apply

I Results from the CARE survey

Figure I.1 illustrates our feasibility study's distribution of CARE scores and compares it with the older *MIBot v5.2* (Brown et al., 2023). The distribution for fully-generative *MIBot v6.3A* is right-skewed, with the majority of participants assigning scores in the upper ranges (36–50). These results indicated that *MIBot* was more effective in fostering an empathetic interaction. However, the comparison in Section 5.2 contextualized its performance relative to human counsellors as falling short of fully matching human-level empathy.

Figure I.2 illustrates the mean scores of each question from the CARE survey across the 106 participants who interacted with *MIBot v6.3A*, and compares it with that of *MIBot v5.2*. The fully generative *MIBot v6.3A* scores higher on each question. Most notably, the biggest improvement seems to be for the question "How was *MIBot* at showing care and compassion?" Interestingly, the lowest-scoring question was "How was *MIBot* at making a plan of action with you?" despite the counsellor prompt directly instructing it to do so.

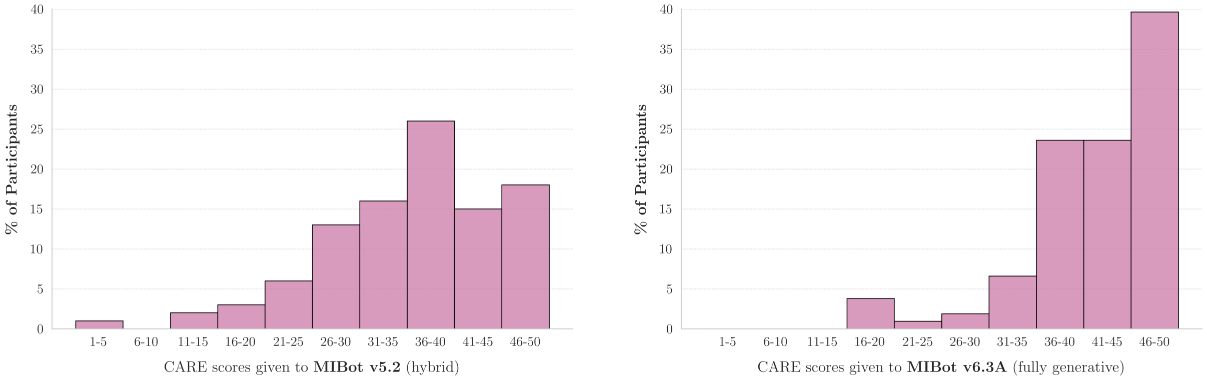


Figure I.1: Distribution of CARE scores for *MIBot v5.2* (hybrid) and *MIBot v6.3A* (fully generative).

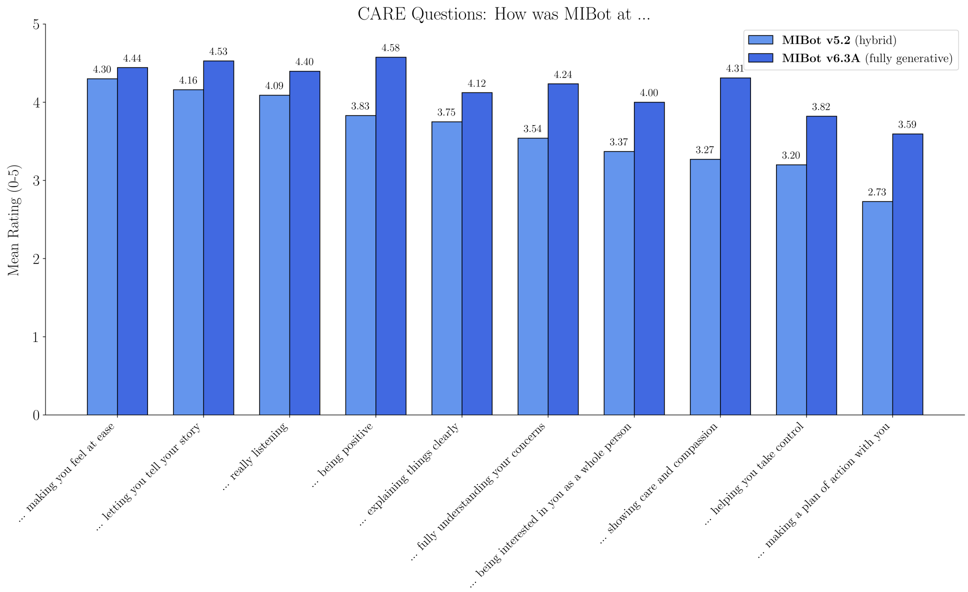


Figure I.2: Question-wise mean CARE scores for *MIBot v5.2* (hybrid) and *MIBot v6.3A* (fully generative).

J *MIBot* Feedback from Study Participants

As part of the post-conversation survey described in Section 5, participants responded to the following three feedback questions:

J.1 Feedback Survey Questions

1. What are three words that you would use to describe the chatbot?
2. What would you change about the conversation?
3. Did the conversation help you realize anything about your smoking behaviour? Why or why not?

Participant feedback on *MIBot* was generally positive. We processed the feedback by dividing the words participants used to describe the chatbot into broad *positive* and *negative* categories. Figure J.1 represents the WordCloud (Oesper et al., 2011) of such words. The top 10 most frequently mentioned positive and negative words are shown in Tables J.1 and J.2.

Word	Frequency
`understanding`	24
`helpful`	22
`friendly`	19
`supportive`	12
`caring`	9
`knowledgeable`	8
`intelligent`	8
`thoughtful`	7
`interesting`	7
`informative`	7

Table J.1: Top 10 most frequently mentioned positive words in participant feedback.

Word	Frequency
`repetitive`	6
`boring`	3
`unresponsive`	1
`disappointing`	1
`annoying`	1
`dull`	1
`pointless`	1
`useless`	1
`uncreative`	1
`overbearing`	1

Table J.2: Top 10 most frequently mentioned negative words in participant feedback.

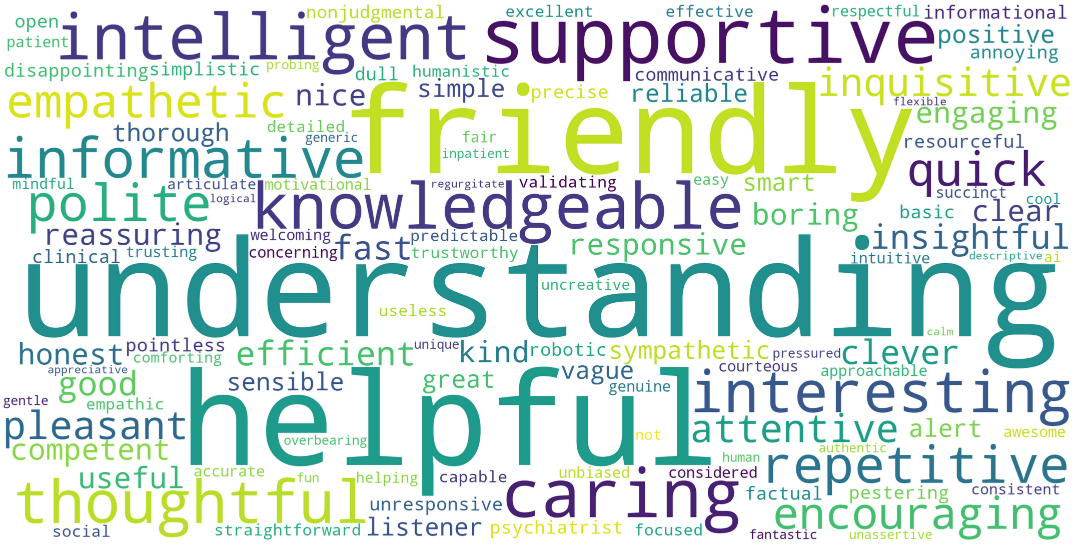


Figure J.1: Word cloud representation of participant feedback.

K AutoMISC

Figure K.1 shows a system flow diagram of AutoMISC. First, each volley (turn of speech) is parsed into one or more utterances (units of thought) by the Parser module. Then, utterance-level annotations, i.e. behavioural codes, are assigned by the Annotator module to each utterance. Up to $k = 5$ prior volleys are included to the Annotator module when coding utterances.

K.1 AutoMISC Validation

We present the pairwise Cohen’s κ values, for both counsellor and client codes, in Figure K.2. All κ values fall between 0.55-0.81, indicating moderate to substantial agreement between each pair of raters beyond chance (Landis and Koch, 1977). The Cohen’s κ values between AutoMISC and the expert annotators (Annotators 1 and 2) were **0.63** and **0.58** for counsellor codes, and **0.63** and **0.69** for client codes, respectively.

K.2 Statistical Validation of Inter-Rater Reliability

To estimate how these reliability findings generalize to more transcripts, we computed the **asymptotic variance** of Fleiss’ κ to calculate two-tailed p-values. For both counsellor and client codes, the asymptotic variance was on the order of 10^{-6}, resulting in p-values of $p < .001$. These extremely low p-values indicate that the inter-rater agreement is highly statistically significant beyond chance. A post-hoc power analysis confirmed that our study was highly powered (estimated power: 1.00) to detect nonzero agreement, i.e. there is a near-certain probability to detect significant inter-rater reliability.

K.3 AutoMISC System Prompts

K.3.1 Parser Prompt

You are a highly accurate Motivational Interviewing (MI) counselling session annotator. Your task is to segment the given volley into utterances.

Definitions:

- Volley: An uninterrupted utterance or sequence of utterances spoken by one party before the other party responds.
- Utterance: A complete thought or thought unit expressed by a speaker. This could be a single sentence, phrase, or even a word if it conveys a standalone idea. Multiple utterances often run together without interruption in a volley.

Output Format:

- Return the segmented utterances as a Python list of strings.

Examples: Below are examples of how to segment a volley into utterances. Follow this structure when processing new inputs.

```
Input:  "Why haven’t you quit smoking - are you ever gonna quit?"
Output: ["Why haven’t you quit smoking - are you ever gonna quit?"]

Input:  "How long since your last drink? Do you feel ok?"
Output: ["How long since your last drink?", "Do you feel ok?"]

Input:  "I can’t quit. I just can’t do it. I don’t have what it takes. I just
     cannot stop."
Output: ["I can’t quit.", "I just can’t do it.", "I don’t have what it takes
    .", "I just cannot stop."]

Input:  "I don’t want to go to the bars every day. I don’t want my kids to
    see that. I want my kids to have a better life than that."
Output: ["I don’t want to go to the bars every day.", "I don’t want my kids
    to see that.", "I want my kids to have a better life than that."]
```

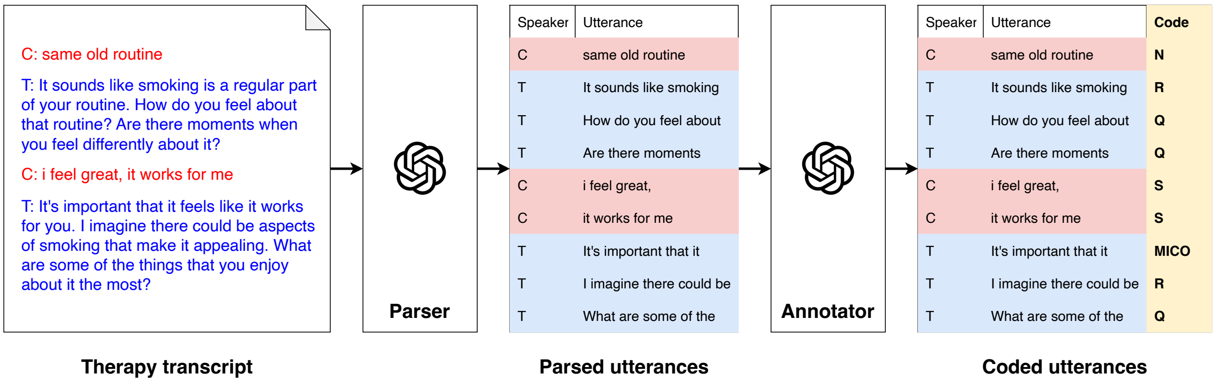
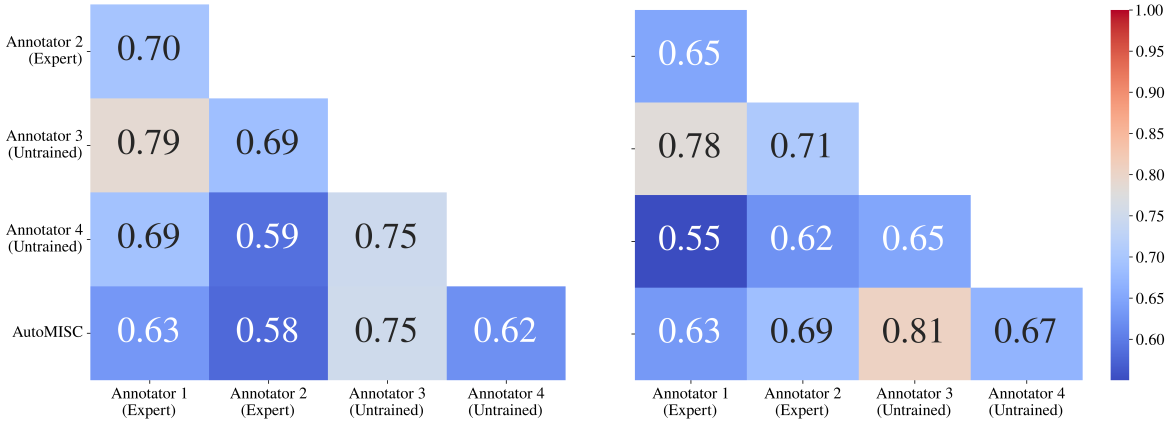

C: same old routine
T: It sounds like smoking is a regular part of your routine. How do you feel about that routine? Are there moments when you feel differently about it?
C: i feel great, it works for me
T: It's important that it feels like it works for you. I imagine there could be aspects of smoking that make it appealing. What are some of the things that you enjoy about it the most?
Parser
Speaker | Utterance
C same old routine
T It sounds like smoking
T How do you feel about
T Are there moments
C i feel great,
C it works for me
T It's important that it
T I imagine there could be
T What are some of the
Annotator
Speaker | Utterance | Code
C same old routine N
T It sounds like smoking R
T How do you feel about Q
T Are there moments Q
C i feel great, S
C it works for me S
T It's important that it MICO
T I imagine there could be R
T What are some of the Q
Therapy transcript
Parsed utterances
Coded utterances

Figure K.1: AutoMISC system diagram.



(a) Counsellor codes

(b) Client codes

Figure K.2: Cohen's κ between rater pairs on behaviour code annotations.

K.3.2 Counsellor Utterance Classification Prompt

You are a highly accurate Motivational Interviewing (MI) counselling session annotator. Your task is to analyze an excerpt from a counselling session of up to five volleys and categorize the counsellor's final utterance.

Definitions:

- Volley: An uninterrupted utterance or sequence of utterances spoken by one party before the other party responds.
- Utterance: A complete thought or thought unit expressed by a speaker. This could be a single sentence, phrase, or even a word if it conveys a standalone idea. Multiple utterances often run together without interruption in a volley.

Task:

1. Determine whether the counsellor's final utterance in the excerpt belongs to one of the following categories:
    - MI-Consistent (MICO): Directly prescribed in Motivational Interviewing (excluding Reflections and Questions).
    - MI-Inconsistent (MIIN): Directly proscribed in Motivational Interviewing principles.
    - Reflection or Question (RQ): Includes Reflections or Questions.
    - Other (Other): Does not fit the above categories.
2. Return your analysis as:
    - explanation: Briefly justify your choice in 1–2 sentences.
    - label: Provide only MICO, MIIN, RQ, or Other.

Behavioural Code Guide:

MI-Consistent (MICO):

- Affirm (AF): Communicates something positive or complimentary about the client's strengths or efforts.
- Advise with permission (ADP): After receiving permission, gives advice, makes a suggestion, or offers a solution or possible action.
- Emphasize control (EC): Acknowledges, honors, or emphasizes the client's autonomy and freedom of choice.
- Raise concern with permission (RCP): After getting permission, points out a possible problem with a client's goal, plan, or intention. Always phrased as the counsellor's concern.
- Support (SU): Sympathetic, compassionate, or understanding comments, which agree or side with the client.

MI-Inconsistent (MIIN):

- Advise without permission (ADWP): Offers suggestions or guidance without asking or receiving permission.
- Confront (CON): Directly disagrees, argues, corrects, shames, blames, seeks to persuade, criticizes, judges, labels, moralizes, ridicules, or questions the client's honesty.
- Direct (DIR): Gives an order, command, or direction. The language is imperative.
- Raise concern without permission (RCWP): Without getting permission, points out a possible problem with a client's goal, plan, or intention.
- Warn (WA): Provides a warning or threat, implying negative consequences unless the client takes a certain action.

Reflection or Question (RQ):

- Question (Q): Asks a question to gather information, understand, or elicit the client's story.
- Reflection (R): Makes a statement that reflects back content or meaning previously offered by the client, usually (but not always) in the client's immediately preceding utterance.

Other (Other):

- Facilitate (FA): Simple utterance that functions as a "keep-going" acknowledgment, e.g., "Mm-hmm", "I see", "Go on".
- Filler (FI): Pleasantries such as "Good morning", "Nice weather we're having", etc.
- Giving Information (GI): Provides information to the client, explains something, educates or provides feedback, or discloses personal information.
- Structure (ST): Gives information about what will happen directly to the client throughout the course of treatment or within a study format, in this or subsequent sessions.

Based on the following excerpt, determine which category the counsellor's last utterance falls into and respond accordingly. After you're done, go back over the RQ category and assign a subcategory of "R" for reflection or "Q" for question.

K.3.3 Client Utterance Classification Prompt

You are a highly accurate Motivational Interviewing (MI) counselling session annotator. Your task is to analyze an excerpt from a counselling session of up to five volleys and categorize the client's final utterance. The target behaviour change of this conversation is smoking cessation.
Definitions:

- Volley: An uninterrupted utterance or sequence of utterances spoken by one party before the other party responds.
- Utterance: A complete thought or thought unit expressed by a speaker. This could be a single sentence, phrase, or even a word if it conveys a standalone idea. Multiple utterances often run together without interruption in a volley.

Task:

1. Determine whether the client's final utterance in the excerpt belongs to one of the following categories:
    (a) Change Talk (C):
        - Expressing a desire to change (e.g., "I really want to quit smoking").
        - Recognizing the downsides of the current behavior (e.g., "My health is suffering because I smoke").
        - Identifying potential benefits of making a change (e.g., "I would feel better if I exercised more").
        - Demonstrating commitment to change (e.g., "I'm ready to make a plan to lose weight").
    (b) Sustain Talk (S):
        - Minimizing the problem (e.g., "It's not that bad, I can handle it").
        - Highlighting difficulties or challenges of change (e.g., "I don't know if I can give up smoking").
        - Expressing doubts about the ability to change (e.g., "I've tried to quit before and failed").
        - Focusing on the positive aspects of the current behavior (e.g., "Smoking helps me relax").
    (c) Neutral Talk (N):
        - Describing current situations or circumstances without expressing a strong pro- or anti-change stance (e.g., "I've been thinking about making changes").

    - Asking questions related to the situation or change process (e.g., "What are the pros and cons of changing?").
    - Making general or factual statements about the issue (e.g., "It's important to take care of my health").

2. Return your analysis as:
    - explanation: Briefly justify your choice in 1-2 sentences.
    - label: Provide only "C", "S", or "N".

K.4 Demographics of the Annotators

As described in Section 4.3, we enlisted four annotators — two experts and two novices — to annotate 10 of the 106 transcripts (comprising 821 utterances) from our study. High alignment between the annotators' labels and the AutoMISC annotations serves as an indicator of AutoMISC's validity. Below, we present their demographic information, following the guidelines proposed by Bender and Friedman (2018).

	Annotator #1[1]	**Annotator #2[2]**	**Annotator #3[3]**	**Annotator #4[3]**
Sex	Female	Female	Male	Male
Age Group (years)	60–69	40-49	20-29	20-29
Race/Ethnicity	White	White	Mixed	Asian
Native Language	English	English	English	Mandarin
Student Status	No	No	Yes	Yes
Employment Status	Full-Time	Full-Time	N/A	N/A
Highest Education	Graduate	Graduate	Undergraduate	Undergraduate
Country of Residence	Canada	Canada	Canada	Canada
Country of Birth	Canada	Canada	Canada	China
Training in Linguistics	No	No	No	No
Training in MI	Yes	Yes	No	No

[1] Motivational Interviewing Network of Trainers (MINT) member since 2009; Motivational Interviewing Treatment Integrity (MITI) coding trained; extensive training and coaching experience.

[2] Introductory-Intermediate-Advance MI training; MINT member since 2014; MI supervision; MITI training.

[3,4] Engineering graduate student with no formal training in MI.

Table K.1: Demographic Information of Annotators

L Demographics of the Participants in the Human Study

Demographic Factors	**Count, n (%)**
Sex	
Female	57 (53.8)
Male	49 (46.2)
Age	
Below 20	0 (0.0)
20 to 29	26 (24.5)
30 to 39	32 (30.2)
40 to 49	20 (18.9)
50 to 59	19 (17.9)
60 to 69	6 (5.7)
70 to 79	3 (2.8)
Above 79	0 (0.0)
Ethnicity	
White	80 (75.5)
Black	9 (8.5)
Asian	7 (6.6)
Mixed	5 (4.7)
Other	5 (4.7)
Student Status	
No	80 (75.5)
Yes	21 (19.8)
Data Expired	5 (4.7)

Demographic Factors	**Count, n (%)**
Employment Status	
Full-Time	49 (46.2)
Part-Time	18 (17.0)
Not in Paid Work	16 (15.1)
Unemployed	13 (12.3)
Other	10 (9.4)
Country of Residence	
United Kingdom	47 (44.3)
United States	42 (39.6)
Canada	9 (8.5)
South Africa	4 (3.8)
Other	4 (3.8)
Country of Birth	
United Kingdom	44 (41.5)
United States	39 (36.8)
Canada	6 (5.7)
Kenya	3 (2.8)
South Africa	3 (2.8)
Germany	2 (1.9)
Other	9 (8.5)

Table L.1: Participant demographics in the *MIBot* Human Study

M Change in Confidence Segmented by Demographic Factors

To examine how baseline confidence to quit smoking varies across different demographic groups, we present confidence rulers at various time points during the study (before, after, and one week after the conversation), segmented by demographic factors. Table M.1 shows that baseline confidence among female participants (2.5) is notably lower than that of male participants (3.2). However, both groups exhibit a similar increase in confidence one week after the conversation (Δ=1.7).

Furthermore, younger participants ($<$ 30 years) initially report a higher confidence level (3.7) than those aged 30 and above (2.5). The week-later change in confidence is also slightly greater for younger participants (Δ=1.9) than older participants (Δ=1.6).

Demographic Factor		**Count, n (%)**	**Before Conv**	**After Conv**	**One Week After**	**Δ (Week After – Before)**
Sex	Female	57 (53.8)	2.5 (2.1)	4.4 (2.8)	4.1 (2.9)	1.7 (2.5)***
	Male	49 (46.2)	**3.2 (1.7)**	4.7 (2.2)	4.9 (2.5)	1.7 (2.3)***
Age	$<$ 30 yrs.	26 (24.5)	3.7 (2.1)	5.5 (2.5)	5.7 (2.7)	**1.9 (3.1)***
	$\geq$ 30 yrs.	80 (75.5)	2.5 (1.8)	4.3 (2.5)	4.1 (2.6)	1.6 (2.1)***
Ethnicity	White	80 (75.5)	2.7 (1.9)	4.3 (2.6)	4.0 (2.6)	1.4 (2.2)***
	Other	26 (24.5)	3.3 (2.0)	5.3 (2.4)	5.8 (2.8)	**2.5 (2.7)*****
Employment	Full-Time	49 (46.2)	3.2 (1.9)	4.8 (2.3)	5.1 (2.6)	**1.9 (2.3)*****
	Other	57 (53.8)	2.5 (2.0)	4.3 (2.8)	3.9 (2.8)	1.4 (2.4)***

Table M.1: Average (SD) **confidence** to quit smoking at different times during the study (before, after and 1-week after the conversation), segmented by demographic factors. Statistical significance was calculated using the one-sided Wilcoxon signed-rank test.
*: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.001$.

N Overview of the Dataset

We are releasing the complete data from our feasibility study. Our dataset consists of a CSV file (`data.csv`), where each row corresponds to a unique participant. A total of 106 participants took part in the study. We also provide conversation transcripts between *MIBot* and participants in another CSV file (`conversations.csv`). All data provided by participants has been de-identified using the `spaCy`[6] (version 3.8.4) and `scrubadub`[7] (version 2.0.0) Python libraries. Further, the participants self-reported all the columns in the dataset (except for AutoMISC annotations). This dataset is licensed under CC BY-SA 4.0[8].

Table N.1: Description of the Columns in `data.csv`

Column Name	Description
Basic	
`ParticipantId`	Unique Participant ids assigned in study.
Pre-conversation Survey on Heaviness of Smoking	
`DailyNum`	How many cigarettes do you typically smoke per day?
`FirstCig`	How soon after you wake up do you smoke your first cigarette?
`HeavinessOfSmokingIndex`	Heaviness of Smoking Index (Heatherton et al., 1989)
Pre-conversation Survey on Quit Attempts a Week Prior	
`PreConvoQuitAttempt`	Have you made any quit attempts (meaning consciously not smoking for a specific period of time greater than 24 hours) during the previous week?
`PreConvoNumQuitAttempts`	How many attempts to quit did you make?
Pre-conversation Readiness Rulers	
`PreRulerImportance`	On a scale from 0 to 10, how important is it to you right now to stop smoking?
`PreRulerConfidence`	On a scale from 0 to 10, how confident are you that you would succeed at stopping smoking if you start now?
`PreRulerReadiness`	On a scale from 0 to 10, how ready are you to start making a change at stopping smoking right now?
Post-conversation Readiness Rulers	
`PostRulerImportance`	On a scale from 0 to 10, how important is it to you right now to stop smoking?
`PostRulerConfidence`	On a scale from 0 to 10, how confident are you that you would succeed at stopping smoking if you start now?
`PostRulerReadiness`	On a scale from 0 to 10, how ready are you to start making a change at stopping smoking right now?
Post-conversation Feedback	
`FeedbackQ1`	What are three words that you would use to describe the chatbot?
`FeedbackQ2`	What would you change about the conversation?
`FeedbackQ3`	Did the conversation help you realize anything about your smoking behavior? Why or why not?

---

[6] `https://spacy.io/universe/project/scrubadub_spacy`

[7] `https://github.com/LeapBeyond/scrubadub`

[8] To view a copy of this license, visit `https://creativecommons.org/licenses/by-sa/4.0/`

Column Name (contd.)	Description (contd.)
`LikedBot`	Whether the participant liked *MIBot*, based on responses to `FeedbackQ1-3`.
`FoundBotHelpful`	Whether the participant found *MIBot* helpful, based on responses to `FeedbackQ1-3`.
CARE Survey	

`CAREQ1`
...
`CAREQ10`
	See Section H for CARE questions.

Week Later Readiness Rulers	
`WeekLaterRulerImportance`	On a scale from 0 to 10, how important is it to you right now to stop smoking?
`WeekLaterRulerConfidence`	On a scale from 0 to 10, how confident are you that you would succeed at stopping smoking if you start now?
`WeekLaterRulerReadiness`	On a scale from 0 to 10, how ready are you to start making a change at stopping smoking right now?
Week Later Quit Attempts	
`WeekLaterQuitAttempt`	Have you made any quit attempts (meaning consciously not smoking for a specific period of time greater than 24 hours) during the previous week?
`WeekLaterNumQuitAttempts`	How many attempts to quit did you make?
AutoMISC Labels	

`AutoMISC_MICO`
`AutoMISC_MIIN`
`AutoMISC_R`
`AutoMISC_Q`
`AutoMISC_Other`
`AutoMISC_C`
`AutoMISC_S`
`AutoMISC_N`
`AutoMISC_%MIC`
`AutoMISC_R:Q`
`AutoMISC_C:S`
	See Section 4.3 for AutoMISC labels.

Column Name	**Description**
`ParticipantID`	Unique Participant ids we assign.
`Speaker`	Indicates whether the speaker is the `counsellor` (i.e., *MIBot*) or the `client`.
`Volley#`	Serial number of the volley in the transcript. "A volley is an uninterrupted utterance or sequence of utterances by one party, before another party speaks." (Houck et al., 2010)
`Utterance#`	Serial number of the utterance in the transcript.
`CumulativeVolley`	Represents the volley up to utterance # `Utterance#`. The `CumulativeVolley` corresponding to the last utterance of the volley is the complete volley, which can be used to generate the transcript.
`Utterance`	"An utterance is a complete thought, or a thought unit." (Houck et al., 2010)
`AutoMISCLabel`	Utterance label according to AutoMISC. It can be one of the following: `R`, `Q`, `Other`, `C`, `N`, ... (See Section 4.3.)
`AutoMISCExplanation`	Explanation provided by the AutoMISC LLM as part of its chain-of-thought.

Table N.2: Description of the Columns in `conversations.csv`

O Example Conversation from the Feasibility Study

The following is an example conversation between *MIBot* (referred to here as "Counsellor") and a participant (referred to as a "Client") in our feasibility study.

Counsellor: Hello there! It's great to have you here today. How are you doing? I'd love to get to know a bit about you and hear anything you feel like sharing. We can take our time, and there's no rush.

Client: im doing okay how are you

Counsellor: I'm doing well, thank you for asking! I'm really glad to hear you're doing okay. Is there anything particular on your mind or anything you'd like to talk about today?

Client: stopping smoking

Counsellor: Stopping smoking is a significant topic, and I appreciate you bringing it up. What are some thoughts or feelings that come to mind when you think about stopping smoking?

Client: its stressfull

Counsellor: It sounds like there's a lot of stress attached to the idea of stopping smoking. It might feel overwhelming to consider that change. What do you think might be contributing to that stress?

Client: ive smoked for a long long time

Counsellor: Having smoked for so long, it's deeply woven into your life, almost like a routine or habit that's hard to imagine changing. It might feel like it's been a part of you for a while. I'm curious, have you thought about what life would be like without smoking?

Client: ive never thought about what my life would be like if i stopped smoking

Counsellor: It's interesting to consider how life might change without smoking, especially if it's something that hasn't crossed your mind before. There might be a sense of uncertainty about what that change would look like. When you think about the possibility of not smoking, what feelings or images come to mind?

Client: stress whta

Counsellor: It seems like the thought of quitting brings up feelings of stress and maybe even uncertainty. It's understandable since smoking has been a constant part of your life. What do you think it is about smoking that helps you manage stress?

Client: what will i do with my hand

Counsellor: Your hands have gotten used to having something to do, like holding a cigarette, and that can be comforting. It seems like there's a connection between smoking and having that physical activity. What do you think might help satisfy that need for keeping your hands busy if you were to stop smoking?

Client: i really dont know

Counsellor: That's completely okay; it's a journey to figure these things out. Sometimes, thinking about change can bring up more questions. What's one thing you're curious about when it comes to considering stopping smoking?

Client: how much money i would save

Counsellor: Thinking about how much money you could save by stopping smoking is an interesting angle. It opens up possibilities for what you might do with those savings. Have you ever considered what you might spend that money on if smoking wasn't part of your life?

Client: home decor

Counsellor: Imagining an opportunity to invest in something you love, like home decor, could be really rewarding. It might feel exciting to picture how your living space could change and reflect more of what makes you happy. Are there any specific changes or items you've been dreaming about for your home?

Client: not really just home decor in general

Counsellor: That makes sense. There's something fulfilling about creating a space that feels comfortable and inviting. It sounds like redirecting those smoking expenses into enhancing your environment could be a motivating thought. How does thinking about home decor and making your space uniquely yours fit into your thoughts about stopping smoking?

Client: great

Counsellor: It’s great to hear that thinking about home decor in this way resonates with you. It can be inspiring to have a goal like creating a space you love as you consider changes to smoking. Is there anything else on your mind about stopping smoking or any other aspect of your life you’d like to explore today?

Client: but i would be bored

Counsellor: Boredom can be a significant concern when thinking about stopping smoking, as smoking has likely filled up certain moments in your day. It’s understandable to worry about how you’ll fill that time. What do you think might be some activities or interests that could replace smoking and keep you engaged?

Client: walking

Counsellor: Walking can be a wonderful activity to fill your time. It offers a chance to clear your mind while being active, and it might serve as a healthy alternative. How do you feel about incorporating more walking into your day as part of this change?

Client: that would be good

Counsellor: Incorporating walking seems like it could be a positive addition to your routine. It not only gives you something to do but also benefits your overall well-being. With all we’ve discussed, is there anything else you’d like to explore or talk about before we possibly think about any next steps?

Client: no

Counsellor: Thank you for sharing with me today. I’d like to take a moment to summarize what we’ve talked about. You’ve been thinking about stopping smoking, though the idea brings up stress, as smoking has been a long-time part of your life. Giving up the familiar habit has made you wonder about what you’ll do with your hands and how you’ll pass the time. We explored the idea of saving money by quitting and how you could use those savings toward your interest in home decor, bringing a new sense of excitement to your living space. Finally, we talked about incorporating walking into your day, which could be a positive activity to replace smoking. If there’s anything else in the future you’d like to explore or if you want to keep this conversation going, feel free to reach out.

Would you like to continue the conversation?

Client: Selected: Yes

Counsellor: Great! I’m glad you’re interested in continuing. With the idea of walking and home decor as possibilities to focus on, how do you feel about taking steps toward reducing or stopping smoking? What’s coming up for you as you think about that?

Client: whats the best products to hekp quit smoking

Counsellor: It’s a great step that you’re considering options to support you in quitting smoking. There are several products and methods many people find helpful, like nicotine replacement therapies, including patches, gum, lozenges, inhalers, and nasal sprays. Some people also explore medications prescribed by healthcare providers.

Before offering any advice, though, I’d like to understand what you might be comfortable with and if there’s a method you’ve been curious about or previously considered. Would it be alright if we discussed those options, or is there any other aspect you’d like to focus on first?

Client: okay

Counsellor: Alright, thank you! Since you’re open to discussing this, nicotine replacement therapies (NRT) are often a starting point for many who want to quit smoking. They’re designed to help reduce withdrawal symptoms and cravings. Each form of NRT has its unique approach—patches provide a steady release of nicotine, while gums and lozenges can offer more control over immediate cravings.

Some people choose to try more than one option to see what works best for them. If you feel comfortable, we could look into what might suit your lifestyle and needs better. What are your thoughts on possibly incorporating one of these into your journey to quit smoking?

Client: nrt sounds good

Counsellor: Nicotine replacement therapy sounds like it could be a helpful tool for you. It offers the flexibility to manage cravings while working towards quitting smoking. It might be worth considering which form you feel most comfortable with, whether it's patches, gum, lozenges, or another option.

Would you like more information on any of these, or is there anything else you want to explore regarding your journey to quit smoking?

Client: no thats all thanks

Counsellor: Thank you for today's conversation. Let's go over what we've discussed. You've shared your thoughts about stopping smoking, including feelings of stress due to its role in your life for so long. We explored the idea of saving money and investing it in home decor, which brings excitement about transforming your living space. You also considered walking as a meaningful activity to fill the time you'd otherwise spend smoking. Finally, we discussed nicotine replacement therapies as potential aids in your journey, with NRT sounding like a good fit for you. If you ever have more questions or want to talk further, feel free to reach out. Take care!

Would you like to continue the conversation?

Client: Selected: No

Counsellor: Thank you and have a great day. Goodbye!

P Consent from Participants

Before participants joined the study, detailed instructions were provided and their consent was obtained to use their data, including survey responses. They were also informed about potential risks. Below are relevant excerpts from the **Consent Form** that participants were required to read and accept.

WHAT ARE THE RISKS OR HARM OF PARTICIPATING IN THIS STUDY?

This study requires you to think about your smoking habits, which may be stressful for you. Thinking about your smoking habits may temporarily increase your urge to smoke.

There is a small but very unlikely possibility that the chatbot may generate responses that may instigate the above feelings. Should this happen, it is encouraged that you document it in your qualitative feedback opportunity.

There are no other risks or harm to you in participating in this study.

WHAT COMPENSATION AM I ENTITLED TO?

You will be compensated 5.5 British Pounds, or the equivalent in your local currency, for completing the conversation task and 1 pound for the 1-week later survey.

HOW WILL MY PRIVACY BE RESPECTED?

All conversational data and readiness/feedback data recorded during the study will be stored on an encrypted server. The study does not ask you to provide any information that can identify you personally. The original chatbot data collected will only be accessible to the University of Toronto researchers involved with the study.

Once any personally identifying information has been removed, the conversation may appear in publications or in a public dataset. The data will be used by researchers at the University of Toronto and CAMH to develop chatbots that can have conversations with individuals about their smoking habits. The data will be held by the University of Toronto indefinitely.

The results of this research study may be presented at meetings or in publications, and this may include direct quotes from your responses and your feedback (after removing any text that personally identifies you), as well as summary statistics from the numerical ratings that all participants provide. In all cases, your identity will not be disclosed.