Automated Coding of Counsellor and Client Behaviours in Motivational Interviewing Transcripts and Application to a Fully Generative Motivational Interviewing Chatbot

by

Soliman Tieu Wajid Ali

A thesis submitted in conformity with the requirements
for the degree of Master of Applied Science

Graduate Department of Electrical and Computer Engineering
University of Toronto

Automated Coding of Counsellor and Client Behaviours in Motivational Interviewing Transcripts and Application to a Fully Generative Motivational Interviewing Chatbot

Soliman Tieu Wajid Ali
Master of Applied Science
Graduate Department of Electrical and Computer Engineering
University of Toronto
2025

Abstract

Motivational Interviewing (MI) is a widely-used talk therapy approach employed by clinicians to guide clients toward healthy behaviour change. Evaluating MI sessions and training MI counsellors relies on *behavioural coding*, the classification of counsellor and client utterances into predefined categories. Recent advances in Large Language Models (LLMs) now make it possible to automate not only behavioural coding, but the delivery of MI itself. This dissertation introduces ***AutoMISC***, which performs utterance-level parsing and behavioural coding under the Motivational Interviewing Skill Code, the original annotation scheme for MI. ***AutoMISC*** achieves an overall accuracy of 70% and a macro F1 score of 0.42 for counsellor speech (19 categories) and 0.41 for client speech (17 categories) against expert-aligned annotations using GPT-4.1 ($n = 821$ utterances). Additional validation showed that the codes predict session-level counselling quality in a widely-used MI transcript dataset at 87% accuracy, and align with existing annotations in another dataset at 71% accuracy. We also demonstrate how the codes can visualize the trajectory of client motivation over a session alongside counsellor codes. We apply ***AutoMISC*** to the transcripts of a brief smoking cessation intervention experiment where tobacco smokers conversed with a fully generative MI counsellor chatbot evolved in collaboration with experienced MI clinician-scientists. Two versions were tested: (1) a single prompted LLM (106 participants), and (2) a two-stage approach which decouples technique selection from utterance generation (93 participants). Participant-reported confidence in quitting smoking was measured before the conversation and one week later. Both versions yielded an average increase in confidence of 1.7 on a 0-10 scale ($p < 0.001$ for both). The first version scored well on participant-reported perceived empathy, higher than typical human counsellors, while the second scored lower. ***AutoMISC***'s analysis of the transcripts provided deeper insights beyond participant-reported outcomes. Both versions showed adherence to MI standards in 99% of utterances. We found that the slope of the trajectory of the client's motivation correlates with the change in confidence (Spearman's $r = 0.28$, $p < 0.005$ for Version 1; $r = 0.20$, $p = 0.051$ for Version 2). This work demonstrates the potential synergy between automated MI delivery and automated MI behavioural coding.

To my parents

Acknowledgements

I would first like to thank my supervisor, Professor Jonathan Rose, for teaching me what it means to be a researcher and an engineer. The past two years have been an incredibly transformative experience for me, to say the least. I am deeply grateful for your patience, guidance, and dedication in mentoring me.

To my research collaborators, both at the lab and at CAMH, it has been a privilege to work with a group so rich in experience, diverse in discipline, and passionate to improve mental health. Every Digital Health Meeting opens my eyes to new ideas and perspectives.

Finally, I would like to thank my parents, whose hard work and sacrifices built the foundation of who I am today. I am eternally grateful for your love and support, and for your unwavering faith in me.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

The modern world faces an accelerating mental health crisis. Globally, rates of depression, anxiety, and addiction have risen sharply in recent years, with the COVID-19 pandemic further exacerbating these trends [20, 56]. Among these concerns, substance use disorders continue to pose a threat to public health, with tobacco use being one of the most pervasive and deadly. It is the leading cause of preventable death in Canada [18] and is responsible for over eight million deaths annually worldwide [26]. Notably, about 60% of tobacco smokers are ambivalent towards quitting [57], meaning that they hold motivations to quit that are impeded by conflicting reasons to maintain the status quo, a critical psychological barrier to successful cessation.

One way to address ambivalence directly is through counselling with human providers, who can apply successful treatments. To do so at scale requires solutions that are both effective and widely accessible. This motivates the development of two complementary technologies: automated counsellors capable of delivering high-quality mental health counselling, and automated systems for evaluating treatment effectiveness. This work focuses on the latter, which can serve a dual role: providing feedback to human counsellors for training and assessment, and operating in-the-loop with automated counsellors to guide in-session dynamics toward best practices and favourable outcomes. In this work, we both develop an automated assessment tool, and use it to evaluate an automated counsellor system.

1.1 Motivational Interviewing and Behavioural Coding

One counselling approach that directly targets the ambivalence underlying tobacco addiction is Motivational Interviewing (MI) [49], a method designed to guide individuals towards a target behaviour change. Originally introduced to treat alcohol use disorders, MI has become widely applied in addiction treatment, chronic disease management, coaching, nutrition, and more. Its core premise is that people are more likely to commit to change when they articulate their own motivations for it, rather than being directed by an external authority.

Training MI counsellors and evaluating MI sessions relies on the classification of each behaviour exhibited in an MI session according to an established taxonomy, called *behavioural coding*[8]. This work adopts the Motivational Interviewing Skill Code (MISC) [36], the original MI annotation framework, which provides comprehensive, mutually exclusive, utterance-level labels for both counsellor and client language. These codes are used to assess counsellor fidelity to MI principles and to classify

the amount of "change" indicated in client speech. Traditionally, this coding is performed manually by trained human annotators, a process that is costly, time-consuming, and unscalable.

1.2 Automated MI Evaluation Using LLMs

Recent advances in the field of Natural Language Processing (NLP) [79, 15] have produced Large Language Models (LLMs), which are capable of handling complex linguistic tasks with high performance. Their comprehensive capabilities, including detecting nuance and subtlety in natural language, make them promising tools for automating classification tasks like behavioural coding. This work presents *AutoMISC*, a transcript classification approach for MI based on the Motivational Interviewing Skill Code (MISC) [36]. The *AutoMISC* system uses pretrained LLMs to perform utterance-level behavioural code annotation of MI transcripts under the MISC 2.5 taxonomy. *AutoMISC* is validated in several ways: first, we compare its annotations (across both closed-source and open-source LLMs) with expert-aligned human annotations. Then, we show that its annotations align with those in the publicly available AnnoMI dataset [83]. We also show that the annotations can predict the binary counselling quality ratings at the session level in the High/Low Quality Counselling dataset [63]. The annotations can also be used to visualize the trajectory of client motivation (and counsellor techniques) over the course of a session, providing fine-grained insights into the session's internal dynamics.

1.3 Automated MI Talk Therapy Using LLMs

Perhaps even more impactful than their comprehensive capabilities, the generative capabilities of LLMs enable them to produce context-appropriate, human-like responses, positioning them as potential conversational agents for delivering MI itself. Early efforts in automated therapy began with pattern-matching engines like ELIZA [80], which provided scripted responses to client inputs based on a fixed set of input-output rules. Subsequent dialogue systems expanded these rule-based architectures, but remained fundamentally constrained for decades by rigid, pre-determined interaction patterns, which are perceived by clients as repetitive or overly generic [9]. Once this barrier was shattered by LLMs, automated therapy using LLMs became a significant area of study [32, 76].

The predecessor to the chatbots presented in this work, MIBot v5.2 [13, 12], combined scripted and generated responses to produce a chatbot that could significantly increase smokers' confidence to quit (measured prior to and one week after a brief intervention), a validated proxy of actual behaviour change [29, 1]. More recent work by others [71] confirmed that fully generative MI chatbots have potential for reducing alcohol use. This motivated the current generation of MIBot: MIBot v6.3, which uses a fully generative approach, developed as part of a large, ongoing collaboration with expert MI clinicians and computer engineers (author contributions listed in Section 1.5).

Two versions of MIBot 6.3 were tested on human participants: Version A ($n = 106$ participants) uses a single, prompted LLM to generate counsellor responses, while Version B ($n = 93$ participants) uses a two-stage process in which the specific immediate therapeutic technique is selected prior to generating the response. Clients reported measurements of confidence to quit smoking and perceived empathy. We apply *AutoMISC* to the resulting transcripts to (1) measure counsellor adherence to MI standards and (2) quantify client motivation over the course of the session. Finally, we propose

a novel metric based on the automated codes that correlates with the client-reported change in confidence, showcasing the synergy of automated MI delivery and automated MI behavioural coding.

1.4 Focus and Goals

The objective of this work is to explore methods for automating the evaluation of MI transcripts, and to demonstrate their utility by applying them to the transcripts of fully generative MI chatbot experiments for smoking cessation. More specifically, the objectives are:

1. **Automated MI Evaluation:** Develop and validate *AutoMISC*, a system for automating the classification of counsellor and client behaviours in MI transcripts under the MISC 2.5 taxonomy using LLMs. Evaluate how factors such as model, prompt structure, and context size affect classification performance.
2. **Application to MIBot v6.3A/B:** Apply *AutoMISC* to the resulting transcripts of MIBot v6.3A/B to evaluate counsellor MI-adherence in a fully generative MI chatbot, monitor client motivation throughout the session, and find relationships between client language and post-therapy outcomes.

1.5 Contributions

To achieve these goals, this dissertation makes the following contributions:

- *AutoMISC*, an automated system for utterance-level MISC 2.5 [36] behavioural coding of MI transcripts, released as an open-source software package.
- Validation of *AutoMISC* across open and closed-source LLMs by measuring (1) assessing performance against expert-aligned human annotations, and (2) performance on public annotated datasets.
- Design and development of portions of the MIBot v6.3 infrastructure, a multi-agent dialogue system to house a generative counsellor chatbot and manage MI sessions.
- Initial design of MIBot v6.3B, a fully generative, expert-informed MI counsellor chatbot for moving smokers towards the decision to quit.
- Deployment of MIBot v6.3B to human participants, giving client-reported measurements of intervention effectiveness and perceived empathy.
- Application of *AutoMISC* to the transcripts from MIBot v6.3A/B showing a significant correlation between client behavioural codes and the post-therapy treatment outcome from the study.
- Four datasets totalling 589 transcripts annotated using the MISC 2.5 to support future work in automated evaluation of MI transcripts.

1.6 Organization

This thesis is organized as follows: Chapter 2 reviews relevant background on Motivational Interviewing, the Motivational Interviewing Skill Code, smoking cessation, and related prior work on automated MI evaluation and MI chatbots. Chapter 3 describes the design of the *AutoMISC* system. Chapter 4 gives the results of the validation experiments for *AutoMISC*. Chapter 5 describes the design of MIBot v6.3B, its associated infrastructure, and the human studies. Chapter 6 gives the results of the MIBot v6.3A/B human experiments, including both client-reported metrics and *AutoMISC*'s outputs. Chapter 7 describes the visualization of conversational trajectories, and the analysis of the relationship between client codes and post-therapy outcomes in the MIBot v6.3A/B transcripts. Chapter 8 concludes the work and discusses limitations and suggestions for future work.

Chapter 2

Background and Related Work

This chapter provides relevant background and an overview of related work to the present work. The first section describes Motivational Interviewing, methods for evaluating MI, and its application in smoking cessation. The next section discusses recent advancements in Natural Language Processing (NLP) that have enabled robust natural language understanding and generation, leading to the development of automated tools for evaluating MI and conversational agents for delivering MI. Finally, we review prior work in the areas of automated psychotherapy evaluation and MI chatbots.

2.1 Motivational Interviewing

Motivational Interviewing is a talk therapy approach that a *counsellor* (often a medical provider) applies to help a *client* (a patient or subject) move towards and achieve a target behaviour change, typically related to health. The conversation is meant to be collaborative, rather than directive, exploring the client's motivations for change and connecting them to their underlying values. The key premise of MI is that a person can be guided to talk themselves into change. This is formalized in the causal chain of MI [51], which proposes that counsellor techniques influence client language, which in turn drives behavioural outcomes. MI counsellors use specific kinds of utterances, such as open-ended *questions* to evoke motivation and *reflections* (which are restatements of client's words, usually connected to relevant ideas and facts) to encourage further contemplation around the target behaviour.

As clients express themselves, counsellors listen carefully for two categories of motivational language: *change talk* [49], which indicates motivation, commitment or action towards change, and *sustain talk*, which reflects reasons to maintain the status quo. Most clients exhibit both, indicating an internal state of *ambivalence* in which they wish to change but are held back by conflicting factors preventing them from changing. A key goal in MI is to help resolve ambivalence by inviting and strengthening change talk, while acknowledging but not reinforcing sustain talk.

In successful MI, as the therapeutic alliance develops, there is a progression in client change talk from a *preparatory* stage (expressions of desire, ability, reasons, or need for change) to a *mobilizing* stage (expressions of commitment, activation, or taking steps towards change). This progression reflects increasing client readiness for change and is predictive of actual behavioural outcomes [49, 3]. However, both change talk and sustain talk may occur in preparatory or mobilizing forms.

2.2 Evaluation of Motivational Interviewing Sessions

The evaluation of Motivational Interviewing counselling may be approached from multiple angles. The most direct methods involve assessing treatment effectiveness through assessment of clinical outcomes, such as reduction in substance use or improvements in health-related metrics. These outcomes are specific to the target behaviour and are discussed in the context of smoking cessation in Section 2.3. However, outcome measures alone do not capture the therapeutic processes that unfold *during* an MI session. It is important to assess whether a counsellor adheres to best practices, and if the client's language indicates that they are progressing towards or away from the target behaviour. This fine-grained analysis is an established method for counsellor training and assessment, and is also relevant in the context of automated counsellors, as these must be vetted thoroughly before they can be deployed clinically.

The gold standard for modelling these in-session processes is by segmenting an MI session into a series of salient behaviours, then classifying each one according to an established taxonomy, called *behavioural coding* [8]. Behavioural coding schemes assign labels to conversational content at the *utterance* level – a single unit of thought. Within a given speaker turn, which we will refer to as a *volley*, a counsellor or client may express multiple utterances in sequence. Thus it is important to first parse volleys into a set of utterances prior to assigning behavioural codes.

2.2.1 Behavioural Coding Frameworks for MI

Several behavioural coding frameworks have been developed to support the evaluation of MI, which vary in scope, granularity, and intended application. Table 2.1 below presents a summary of widely used MI behavioural coding frameworks. Note that some behavioural coding schemes include "global scores", which are coarse, session-level assessments of certain aspects of an MI conversation, such as judging the level of partnership between counsellor and client, usually in the form of Likert scales completed by the annotator. Due to the focus on modelling the granular, internal dynamics of MI sessions, global score-type evaluations are left out of the scope of this work.

Framework	**Release (Latest)**	**Purpose**	**Couns. Codes**	**Client Codes**
Motivational Interviewing Skill Code (MISC) [36]	1997 (v2.5, 2010)	Comprehensive evaluation of counsellor and client behaviours in MI Sessions (research)	$19 + 6^*$	15
Motivational Interviewing Treatment Integrity (MITI) [52]	2005 (v4.2, 2015)	Practitioner proficiency assessment for MI-adherence (used in training)	$10 + 4^*$	—
Client Language Assessment in Motivational Interviewing (CLAMI) [50]	2008	Coding of client speech; originally an addendum to MISC 2.1	—	8
Client Evaluation of MI (CEMI) [43]	2013	Client-reported assessment of counsellor MI adherence	—	$0 + 11^*$
OnePass [45]	2014	Faster proficiency assessment than MITI	$0 + 23^*$	—

Table 2.1: Common behavioural coding frameworks used in Motivational Interviewing (* = global scores).

In this work, we adopt the Motivational Interviewing Skill Code (MISC) framework [36], the original annotation scheme for MI, because it was designed for research and provides a comprehensive, mutually exclusive, fine-grained taxonomy for *both* counsellor and client codes.

2.2.2 The Motivational Interviewing Skill Code (MISC) 2.5

The MISC 2.5 framework defines 19 counsellor codes and 17 client codes[1]. The basic counsellor strategies (questions and reflections), as well as client codes (change and sustain talk) described in Section 2.2 have several sub-types in MISC 2.5. For example, counsellor reflections are further subdivided into a *Simple Reflection* (SR) which simply mirrors a client's statement, and a *Complex Reflection* (CR) in which the counsellor both mirrors and adds meaning or insight. Client codes are further subdivided by type as follows: *preparatory* change/sustain talk includes the codes for different levels: *Desire* (D+/D-), *Ability* (AB+/AB-), *Reason* (R+/R-), and *Need* (N+/N-). Similarly *mobilizing* change/sustain talk has levels called *Commitment* (C+/C-), *Activation* (AC+/AC-), and *Taking Steps* (TS+/TS-). The full classification taxonomy is provided in **??**. Notably, client codes can be interpreted along a "motivational axis", with preparatory speech indicating emerging motivations and mobilizing speech indicating stronger commitment towards change (or maintaining the status quo). This axis is important in the present work as it permits the modeling of the slope of the client motivation as discussed in Chapter 7.

MISC also provides session-level *summary scores* computed from frequency counts and ratios of behavioural codes across the session, intended as heuristic indicators of session quality in research and training contexts. These include:

- **Percentage MI-Consistent Responses (%MIC)**: the proportion of counsellor behaviours classified as MI-Consistent i.e. directly prescribed in Miller and Rollnick's *Motivational Interviewing* [49]. Higher values indicate greater adherence to MI standards.
- **Reflection-to-Question Ratio (R:Q)**: the ratio of reflective statements to questions posed by the counsellor. Values between 1 and 2 are considered good [52].
- **Percentage Change Talk (%CT)**: the proportion of client utterances coded as Change Talk, with higher values associated with improved behavioural outcomes [4].

Behavioural coding is traditionally done manually, by trained human annotators, thereby rendering it a resource-intensive and unscalable process. To overcome these constraints, this work automates behavioural coding in MI using LLMs, as outlined in Chapter 3.

2.3 MI for Smoking Cessation

Motivational Interviewing has been shown to be effective for smoking cessation across a range of populations and settings [49, 33, 64]. A meta-analysis of 31 clinical trials involving over 9400 participants concluded that current MI smoking cessation approaches can be effective for adolescents and adults [31]. MI is particularly well-suited to this domain because of the large proportion (60%+) [57] of smokers who are ambivalent: most smokers are aware of the health risks and express a desire to quit, but simultaneously feel unready or unmotivated to quit. Due to the brevity of the chatbot conversations in our human studies, long-term outcome tracking was not feasible. Instead, we use

[1] Although not listed in the MISC 2.5, we include "Activation+/-" in the client code set based on definitions in [49].

two validated survey instruments as proxy measures of clinical effectiveness: the Readiness Ruler and the Consultational and Relational Empathy (CARE) survey, described below. The in-session dynamics were evaluated by the automated behavioural coding system described in Chapter 3.

2.3.1 The Readiness Ruler

Measurement	Question	Score Range
Importance	How important is it to you right now to stop smoking?	0-10
Confidence	How confident are you that you would succeed at stopping smoking if you started now?	0-10
Readiness	How ready are you to start making a change at stopping smoking right now?	0-10

Table 2.2: The Readiness Ruler questionnaire.

The Readiness Ruler [39] is a brief survey completed by the client designed to assess motivation to change the target behaviour. It captures three dimensions: importance (how important it is to them to change the target behaviour), confidence (how confident they are that they could succeed in changing), and readiness (how ready they are to start changing), measured at the time of assessment. Table 2.2 gives the specific questionnaire asked to the participants in our human studies. In practice, the Readiness Ruler is used as a litmus test of client motivation, asked by the counsellor at various times in treatment. Tracking the client's progress along the ruler provides a starting point for counsellors to address barriers or reinforce client motivation to change. In the human studies covered in this work, the change in the scores on the readiness rulers measured at different times serves as short-term proxy for therapeutic efficacy. For smoking cessation, the confidence ruler is a known and validated predictor of actual behaviour change [29, 1], so it is used as the primary measure of the chatbot's effectiveness.

2.3.2 Consultation and Relational Empathy (CARE) Survey

The Consultation and Relational Empathy (CARE) Survey [46] is a survey completed by the client that measures the client's perception of a healthcare provider's empathy during an interaction. It consists of 10 questions, each rated on a scale of 0-5, which are then summed to yield a total score out of 50. The full survey is included in Appendix D.4. It is widely used in psychotherapy research as a measure of therapeutic alliance, which is particularly important in MI. The CARE survey is to measure the perceived empathy of the automated counsellor in the human studies described in Chapter 5.

2.4 Natural Language Processing

Natural Language Processing (NLP) is a field at the intersection of linguistics and machine learning, concerned with the computational modelling of natural language. Recent advancements in NLP have made it possible to perform complex language understanding and generation tasks using Large Language Models (LLMs): neural networks with billions (sometimes trillions) of parameters trained

on extensive corpora of text. These models can be easily prompted or fine-tuned for domain-specific applications. LLMs are sequence models, meaning that they take in a sequence (natural language encoded as a series of tokens) and output a sequence (tokens then decoded back to natural language). Most LLMs are auto-regressive, meaning that they generate text by predicting the next token in a sequence based on all previous tokens, and continue until they reach a designated stop token or maximum length.

Modern LLMs are almost exclusively based on the transformer architecture, first introduced by Vaswani et al. in 2017 [79]. The key innovation of the transformer architecture is the self-attention mechanism, which is said to capture dependencies between any parts of the input sequence regardless of their positions (something that earlier architectures struggled with). This makes them particularly effective at tasks that require contextual understanding of a conversation, or detection of subtle nuances in speaker intent, motivating their use for applications like psychotherapy evaluation or conversational agents for counselling.

While fine-tuning can improve performance on specialized tasks, it is often impractical in clinical domains due to the sensitive nature of psychotherapy data and limited availability of annotated datasets. Fortunately, LLMs exhibit properties that make them highly useful without fine-tuning. One such property is *few-shot learning*, the ability to learn a new task only from a few examples provided at inference time. This is an emergent property from scaling up the size of models, first demonstrated by OpenAI with GPT-3 [15]. This is part of a broader phenomenon known as *in-context learning*, where LLMs generalize from examples in their input context rather than requiring parameter updates through training. The introduction of Reinforcement Learning with Human Feedback (RLHF) further enhanced LLM's practical utility. OpenAI's InstructGPT [58] demonstrated how RLHF could align LLM outputs with human preferences, effectively making them able to answer what was asked of them instead of completion-style inference, and producing responses that are more relevant, appropriate, and empathetic, all of which are important for therapeutic applications.

These breakthroughs have profound implications for automated behavioural coding and psychotherapy, as it became possible to build systems capable of performing these tasks through prompting alone, without large-scale data collection, extensive feature engineering, or supervised training. The following section describes prior work in these applications and how the advent of LLMs transformed them.

2.5 Related Work

2.5.1 Automated Behavioural Coding in Psychotherapy

Early approaches to automated behavioural coding in psychotherapy relied on linguistic features selected and engineered by experts [17, 62] or topic modeling [7, 6] to detect specific behaviours such as asking questions and providing reflections, occasionally combined with another modality such as accoustic features [5]. Later, neural network-based approaches emerged [73, 28, 84, 38, 19, 25], improving classification accuracies in behavioural coding tasks by offering a more expressive and implicit model of the dialogues. More recent work has used BERT-based transformer models [24, 42] to extract contextual embeddings from counsellor and client utterances [74, 13, 61, 85, 23], sometimes complemented by other features such as voice [75] and facial information [53], which are

then passed to downstream neural network-based classifiers. These approaches performed well when the behavioural task is sufficiently constrained, although extensive training is required on datasets annotated with high-quality labels. Among the strongest results is by [23], which achieved a macro F1 score of 0.42 with 70% accuracy on 10 counsellor codes under the MITI coding framework [52], and macro F1 of 0.72 with 72% accuracy on three client codes.

The adaptation of LLMs in this space initially explored fine-tuning approaches [34]. More recent efforts have demonstrated that LLMs can be effectively prompted for behavioural coding without fine-tuning, through either zero-shot prompting [14, 44], few-shot prompting [72], or in-context learning [21], achieving high accuracy when compared with human labels. Notably, with few-shot prompting, [72] achieved Macro F1 scores of 0.31 on 16 counsellor codes and 0.32 on 10 client codes under MISC 2.1 [48], an earlier version of the MISC framework.

Despite these advances, prior work still has limitations in their behaviour coding capabilities. Many approaches focus exclusively on either counsellor or client speech, and often target only a small subset of behaviours. For MI in particular, no existing work has attempted fully automated coding of both speakers under the complete MISC 2.5 framework [36]. Moreover, prior work rarely connects automated behaviour coding to treatment outcomes, and few projects release code or software to support reproducibility or real-world use.

2.5.2 MI Datasets

There are several public, anonymized datasets supporting the task of MI behavioural coding. These include the High/Low Quality Counseling dataset [63], Counsel-Chat [82], AnnoMI [83], MI-TAGS [23], and BiMISC [72]. The datasets vary in their sources, as well as the levels of granularity in the labels they provide. While these datasets have supported progress in behavioural coding, most lack full MISC 2.5 coverage, are not publicly accessible, or offer only coarse labeling. There remains a need for high-quality, fully annotated MI datasets aligned with an existing behavioural coding framework such as MISC 2.5, to support more complex tasks such as fine-grained modelling of MI transcripts and prediction of client behaviours.

2.5.3 MI Chatbots

The development of chatbots for Motivational Interviewing has evolved from its early roots in rule-based systems like ELIZA [80]. Early MI chatbots similarly relied on scripted, rule-based architectures [59, 54, 67], however, their lack of flexibility in providing customized responses (and sometimes user input) often hindered the development of the therapeutic alliance. In response, subsequent work incorporated Natural Language Understanding (NLU) modules using neural networks to classify client intent [2], marking the gradual shift away from rigid rule-based approaches. This shift continued as transformer models were adopted for both language understanding and generation of certain MI skills like *reflections* [70, 13] and *giving advice* [81]. One notable contribution was the Technology-Assisted Motivational Interviewing chatbot coach [66], which used a multiple classifiers and generative models to produce more contextually appropriate counsellor responses, to assist with training MI counsellors. Most recently, the advent of LLMs have enabled a new generation of MI chatbots that demonstrate fluency and contextual awareness with minimal engineering, either through prompting or fine-tuning. The author is a contributor to one such work, MIBot v6.3A [44], which uses OpenAI's GPT-4o to

deliver fully generative MI counselling to move smokers towards the decision to quit (see Chapter 5). Other notable work includes MIcha [47], which showed through a randomized control trial that even brief interactions with LLM-based chatbots can significantly increase client readiness to change.

In the following two chapters, we describe and then validate ***AutoMISC***, the automated behavioural coding system that we designed to annotate the transcripts from the MIBot v6.3A and 6.3B experiments under the MISC 2.5 framework.

Chapter 3

Design: Automated Coding System

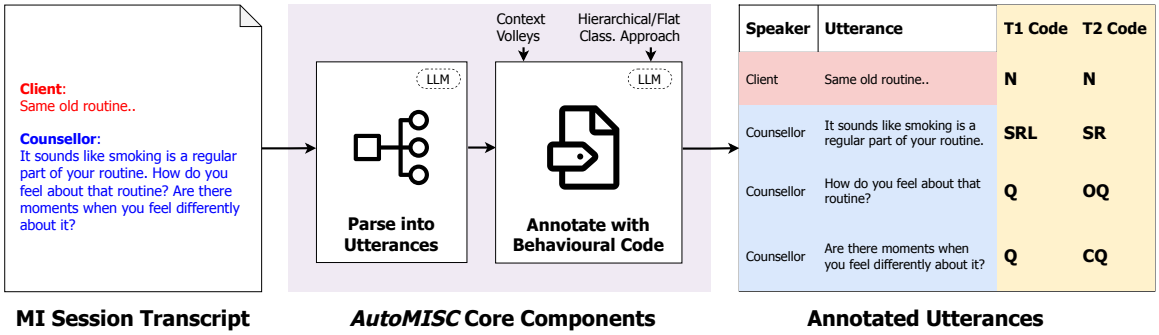


Figure 3.1: Overview of the *AutoMISC* system.

This chapter describes the design of *AutoMISC*, our automated behavioural coding system for Motivational Interviewing transcripts. As outlined in Chapter 1, one of the central goals of this work is to enable the evaluation of MI sessions at scale by automating the utterance-level classification of counsellor and client language under the Motivational Interviewing Skills Code (MISC). *AutoMISC* was designed to perform this multi-step process without human intervention, bypassing the traditional barriers of manual high-quality MI coding.

3.1 *AutoMISC* System Design

Figure 3.1 illustrates the architecture of the system. MI conversations are initially modelled as sequences of turns of speech from either the counsellor the counsellor or client, with each turn referred to as a *volley*. In the first step of the pipeline, each volley is parsed into utterances (individual units of thought, as described in detail below). Each utterance is then annotated with a behavioural code from the MISC framework (see Section 2.2.2). The input to *AutoMISC* is a transcript file that simply separates dialogue into volleys and identifies speaker roles as either counsellor or client. The outputs from the system is the parsed and annotated corpus, from which the MISC session-level summary scores can be computed, and the dynamics of the conversation can be visualized.

The remainder of this chapter describes the core components of *AutoMISC* in further detail, namely the utterance parser and the behavioural code annotator. The prompts for the modules are listed in

Appendix A.

3.1.1 Segmentation of Volleys into Utterances

The segmentation of each volley into one or more utterances is not simply a separation into sentences, because an utterance may be multiple sentences or a portion of a single sentence. This makes the task semantically complex, so we use a prompted pre-trained Large Language Model model to perform this task.

As an example, the following client volley is a single utterance: "I need help. I don't think I can do this on my own. " By contrast, the following counsellor volley would be split into two utterances at the comma: "You have a strong desire to quit, but you're not sure you have the ability."

The prompt for the utterance segmentation begins with definitions of *volley* and *utterance* from the MISC manual and then the general task of separation of utterances. It includes four few-shot example input-output pairs sourced from the MISC manual. The full parser module system prompt is provided in Appendix A.1.

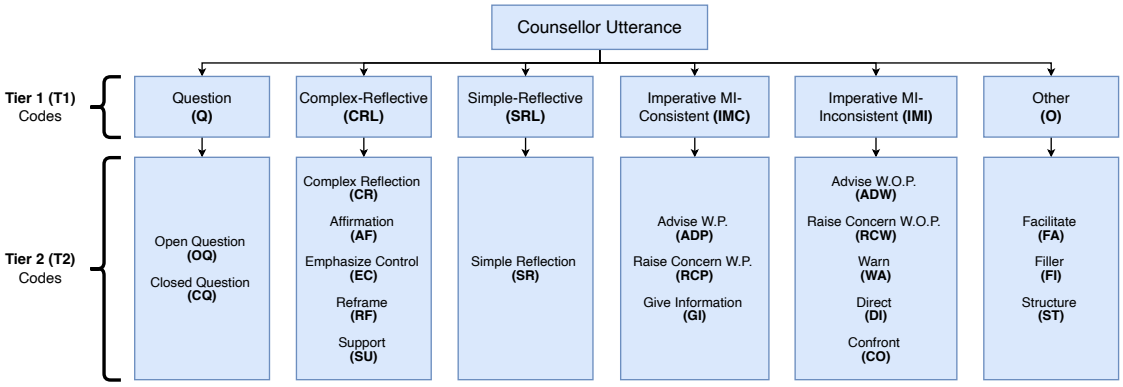


Figure 3.2: *AutoMISC* counsellor utterance classification taxonomy.

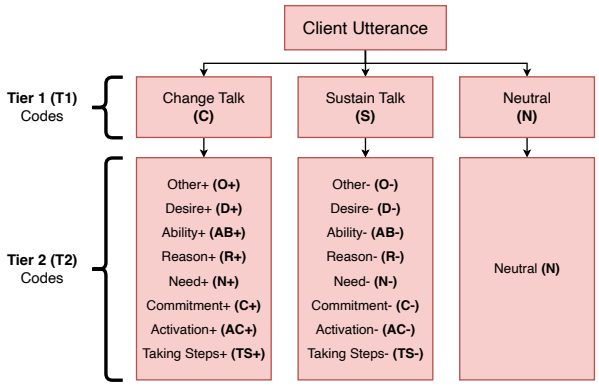


Figure 3.3: *AutoMISC* client utterance classification taxonomy.

3.1.2 Automated Coding

The classification of each utterance into a behavioural code is handled by the annotator module, which is also a prompted large language model.

A key decision in annotation is whether to use a **hierarchical** or a **flat** classification approach. This question was motivated by our manual coding work (described below in section 3.1.3) where we found it very helpful to decompose the task into two steps: in this hierarchical setup, the first step classifies an utterance into a higher-level grouping of similar MISC codes that we call *Tier 1* codes, then the second step refines the classification to the fine-grained MISC code (the *Tier 2* codes). This contrasts with the flat approach which classifies directly to the fine-grained MISC code. We hypothesized that a language model might see performance gains from this decomposition, at the cost of doubling the number of inference calls. For client utterances, the three Tier 1 categories are intuitively Change Talk (C), Sustain Talk (S), and Neutral Talk (N). For counsellor utterances, we grouped the 19 fine-grained codes into six groupings based on (human-perceived) semantic similarity and ease of disambiguation. The full sets of Tier 1 and Tier 2 codes for counsellor and client language are shown in Figure 3.2 and Figure 3.3 respectively. We compare this to the *flat* approach in which the model selects directly from the full set of Tier 2 codes in Section 4.1.3.

A second key parameter for the annotator module is to decide how much prior conversation context is needed for high classification accuracy. The module takes in a parameter called **number of context volleys** which sets how many volleys prior to the one under consideration to include in the prompt. We hypothesized that performance would improve with additional context up to a point of diminishing returns, discussed further in Section 4.1.2. Similar to the classification approach, this creates a tradeoff: more context may improve accuracy, but at the cost of increased token consumption.

Each prompt to the annotator module includes a task description, the available label set, the context window, and finally the target utterance for classification. In the hierarchical mode, the Tier 2 prompt is templated to include only the candidate codes associated with the selected Tier 1 label. The prompt templates are shown in Appendix A.2. Once annotation is complete, the summary scores described in Section 2.2.2 are computed.

3.1.3 Consensus Labels & Annotator Alignment with Experts

To evaluate and refine *AutoMISC*, we created a reference dataset of known-good human annotations, which we will refer to as the *consensus labels*. These were created using a combination of members of our research team which included both computer engineers and experienced MI clinicians specializing in smoking cessation. To produce reliable annotations, we first trained a team of three undergraduate research interns and one graduate student to annotate transcripts from the MIBot v6.3A dataset [**mahmood2025fully**] using the MISC 2.5 schema. We used an iterative process in which the goal was to achieve substantial inter-rater reliability, commonly quantified as Fleiss' Kappa $\kappa \geq 0.6$ [22]. The iterative process was as follows:

1. The four annotators independently label five transcripts.
2. The inter-rater reliability (IRR) is computed using Fleiss' κ across all codes, counsellor and client.

3. If $\kappa < 0.6$ for any category, an alignment meeting is held, together with expert MI clinicians to resolve discrepancies.

We completed two iterations: In the first round, annotators labelled the first five transcripts from the dataset (a total of $n = 367$ utterances) but did not meet the IRR threshold for all codes. A two-hour alignment meeting was held, during which consensus labels were produced for that sample. In the second round, annotators labelled a new set of five transcripts ($n = 454$ utterances), after which the IRR target was reached. Training was deemed complete, and consensus labels were consolidated across both sets, yielding a reference set of $n = 821$ utterances (580 from the counsellor, 241 from clients). Figure B.1.1 in Appendix B gives the pairwise Cohen's Kappa matrices between raters before and after training.

3.1.4 Classification Prompt Evolution

The initial classification prompts for the annotator module were derived directly from the definitions of behavioural codes in the MISC 2.5 manual and Miller and Rollnick's *Motivational Interviewing* [49]. These were evolved based on classification performance against the consensus labels of the reference dataset, using OpenAI's GPT-4o [1]. There were two key issues found with the prompts: The first concerned Open versus Closed Questions (OQ vs CQ): ***AutoMISC*** initially overused the OQ label. This was resolved by improving the prompt so that questions answerable with a "yes", "no", or short factual response should be coded as CQ in the Tier 2 counsellor classification prompt, as shown in Appendix A.2.

The second issue concerned Imperative-MI-Inconsistent vs Imperative-MI-Consistent (IMI vs IMC). Here the issue is that an imperative/directive statement is only MI-Consistent if permission was granted to do so, and that permission may be one or *more* volleys prior to the utterance being coded. It was observed that these permissions could be delivered in subtle ways, which were hard to detect. This was addressed by adding a Chain of Thought reasoning process around permission to the end of the T1 counsellor classification prompt, as shown in Appendix A.2.

Once these systemic issues were addressed, the next step was to validate that the system performed as expected. In the following chapter we describe the experiments conducted to validate ***AutoMISC*** across its input configurations, primarily by comparing its output to the consensus labels, but also via comparison to existing datasets.

[1] `gpt-4o-2024-08-06`

Chapter 4

Results: *AutoMISC* Validation Experiments

The *AutoMISC* system was validated primarily using macro F1 score and accuracy, measured on the first 10 conversations (a total of $n = 821$ utterances) from the MIBot v6.3A dataset [**mahmood2025fully**], using the consensus labels described in Section 3.1.3 as ground truth. We also validate against the labels of the AnnoMI dataset [83], and we show that the annotations can predict counselling quality in the HLQC dataset [63].

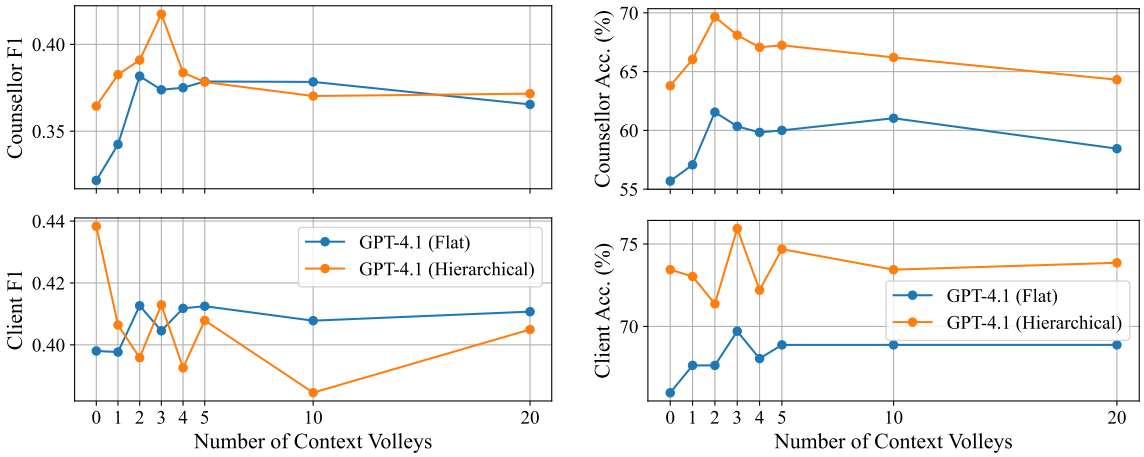
4.1 Experimental setup

AutoMISC is configured with three input parameters: (1) the language model used for annotation, (2) the classification structure (hierarchical vs. flat), and (3) the number of prior volleys provided as context to the model (the latter two introduced in Section 3.1.2). The models chosen were selected for diversity both in model provider, using both open- and closed-source, and a range of model sizes, as follows: OpenAI's GPT-4o [1] and GPT-4.1[2], Alibaba's Qwen3-30b-a3b[3], and Google's Gemma-3-12b[3]. The OpenAI models were accessed through the company's for-pay APIs, and the other models were run on an M3 Macbook Pro with 36GB of unified memory, and makes use of the native GPU acceleration. Wall-clock inference times per utterance were approximately 2 seconds for the OpenAI models, 7 seconds on the Qwen model and 16 seconds on the Gemma model. The utterance parsing step was done by GPT-4o in all cases, to enable direct comparison of classification/coding/annotation accuracy between the different models.

4.1.1 Parameter tuning

Figure 4.1 gives the classification performance (macro F1 score and accuracy) versus the number of context volleys for GPT-4.1, separated into different plots by speaker (counsellor/client) and classification approach (hierarchical vs. flat). Results for the other three models are given in

[1] `gpt-4o-2024-08-06`
[2] `gpt-4.1-2025-04-14`
[3] Quantized to 4-bit parameters



(a) Macro F1 Score vs. Context Volleys

(b) Accuracy vs. Context Volleys

Figure 4.1: Effect of context size and classification approach (hierarchical/flat) on counsellor and client classification performance (GPT-4.1, $n = 821$ utterances)

Appendix C. The accuracy is greater than F1 because the most common behavioural codes achieve good accuracy across the 19 counsellor codes and the 17 client codes.

4.1.2 Number of Context Volleys

For counsellor codes, Figure 4.1 (top row) shows that performances improves with additional context up until 2-3 volleys, after which it plateaus or declines. The initial increase is likely due to the fact that all the "IMC" codes require permission to be granted in a preceding volley. The degraded performance with longer contexts might be attributed to the model attending to less relevant context in the earlier volleys.

The client coding performance appears to simply plateau or degrade with added context. This is likely because change and sustain talk is self-evident within an utterance and may even shift rapidly between change talk and sustain talk within the same volley [49], making additional context less informative.

4.1.3 Hierarchical vs. Flat Classification Approach

Figure 4.1 shows that the hierarchical classification approach is almost uniformly better across all tested models and context window sizes, but the flat approach achieves similar or even higher macro F1 scores in a few configurations, mostly on the client codes.

4.2 Validation Results

Table 4.1 gives the F1 and accuracy scores for the model and parameter settings that achieved the highest macro F1 score. Complete numerical results across all configurations are in Appendix C.

The highest-performing model and configuration overall was GPT-4.1 using 3 prior volleys as context and the hierarchical classification structure. It achieves a macro F1 score of 0.42 and 68%

Model	Class. Struct.	Context Volleys	T1 Cou.		T1 Cli.		T2 Cou.		T2 Cli.		T2 Avg.	
			F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.
GPT-4.1	hier.	3	**.80**	**82**	**.87**	**88**	**.42**	**68**	**.41**	**76**	**.42**	**70**
GPT-4o	flat	2	–	–	–	–	.41	61	.41	65	.41	62
Qwen3-30b-a3b	hier.	0	.61	69	.77	78	.28	55	.35	63	.30	57
Gemma-3-12b	hier.	1	.60	70	.80	81	.30	54	.40	59	.33	56

Table 4.1: Best accuracy (%) and macro F1 scores with consensus labels across models, classification approach and context window sizes for each speaker and code tier ($n = 821$ utterances).

accuracy on the full set of 19 MISC counsellor codes. On the 17 client codes it achieves an F1 score of 0.41 and 76% accuracy. The smaller open-source models achieved competitive results on both counsellor and client coding. For instance, Gemma-3-12b reached 0.40 Macro F1 on client codes, outperforming the larger Qwen3-30b-a3b model.

Work	T2 Couns.	T1 Client	T2 Client
BiMISC	0.31 (16)	0.68 (3)	0.32 (10)
MI-TAGS	0.42 (10)	0.72 (3)	–
AutoMISC	**0.42** (19)	**0.87** (3)	**0.41** (17)

Table 4.2: Reported macro F1 scores from prior work compared to *AutoMISC*. Values in parentheses indicate the number of classes.

Table 4.2 compares *AutoMISC*'s classification performance to prior work reported in the original publications introducing the [72] and MI-TAGS [23] datasets. In spite of the larger label spaces covered, our results meet or exceed these results across both speaker roles.

Confusion matrices for the best performing configurations are included in Appendix C.

4.3 Supplementary Validation Experiments

4.3.1 Comparison to AnnoMI

As a secondary form of validation, we compare *AutoMISC*'s labels (using our best-performing configuration, GPT-4.1 with three context volleys and the hierarchical classification approach) against those from the AnnoMI dataset [83]. This dataset contains 133 MI conversations professionally transcribed and coded under a custom volley-level coding scheme by experienced MI practitioners. Each volley in the dataset has up to three counsellor codes (drawn from questions, reflections, and therapist input categories) and a single client code indicating Change Talk (C), Sustain Talk (S), or Neutral Talk (N). Although inspired by MITI/MISC, it differs significantly from the MISC coding used in this work. To make a direct comparison between *AutoMISC* and the AnnoMI codes, the AnnoMI codes were transformed in the following ways:

1. *AutoMISC* Tier 1 utterance-level **client codes** are aggregated across each volley through a majority vote. Ties are broken using the hierarchy C>S>N. The resulting aggregated labels are compared to AnnoMI's single client label per volley using Cohen's κ, accuracy, and a confusion matrix.

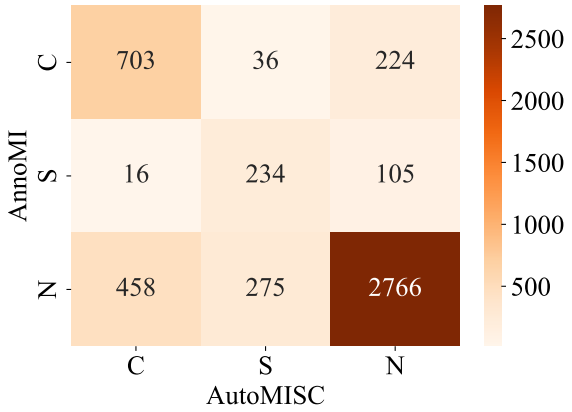


Figure 4.2: Confusion matrix comparing *AutoMISC* and AnnoMI client codes (aggregated to volley-level C/S/N).

2. For **counsellor codes**, an AnnoMI volley-level label is considered matched if for each counsellor code there exists at least one corresponding utterance-level code in *AutoMISC*'s annotations for that volley, according to the mapping shown in Table 4.3. A volley-level match occurs only if all AnnoMI codes are covered.

AnnoMI Code	**Mapped MISC 2.5 Codes**
Question: open	{OQ}
Question: closed	{CQ}
Reflection: simple	{SR}
Reflection: complex	{CR, RF, AF}
Therapist input: information	{GI}
Therapist input: advice	{ADP, ADW}
Therapist input: options	{ADP, ADW, EC, ST}
Therapist input: negotiation	{ADP, ADW, EC, ST, RCP, RCW, WA, CO, DI}
None of the above	{FA, FI, SU}

Table 4.3: Mapping from AnnoMI counsellor labels to MISC 2.5 codes used by *AutoMISC*.

With this mapping the *AutoMISC* client coding achieves a Cohen's $\kappa = 0.51$ (which is considered 'moderate' agreement) and an accuracy of 77% over $n = 4817$ volleys. Figure 4.2 gives the confusion matrix betwen the C, S, and N codes between *AutoMISC* and AnnoMI.

The counsellor code accuracy is 65% over $n = 4882$ volleys.

4.3.2 Predicting High/Low Quality on the HLQC Dataset

The High Low Quality Counselling (HLQC) dataset [63] is a publicly available[4] corpus of transcribed MI counselling demonstrations sourced from public websites. The HLQC dataset was designed to support the development of "data-driven methods for the automatic evaluation of counselling quality" and contains 258 transcribed MI sessions rated as either *high* or *low* quality by expert MI

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

practitioners. HLQC does not include fine-grained behavioural codes for a direct comparison with *AutoMISC*. However, the binary quality rating offers an opportunity to assess whether *AutoMISC*'s outputs align with expert judgments at the session level, using the following process: *AutoMISC* was run on the HLQC dataset using the best-performing configuration, and the three MISC summary scores described in subsection 2.2.2 were produced. These are used to predict binary counselling quality by training a logistic regression classifier using leave-one-out cross-validation (LOOCV).

Predictor(s)	**Acc. (%)**	**F1**	**AUC**
%MIC	**87**	**0.90**	0.933
R:Q	70	0.79	0.741
%CT	75	0.80	0.729
All Combined	86	0.89	**0.940**

Table 4.4: LOOCV classification performance for predicting binary session-level MI quality on HLQC using summary scores derived from *AutoMISC* ($n = 258$).

As shown in Table 4.4, the %MIC summary score is the most predictive individual feature, achieving 87% accuracy and an AUC of 0.93. Combining all three summary scores yields an overall accuracy of 86% accuracy and an AUC of 0.94. These results are consistent with those reported in the original HLQC study, where handcrafted MITI-derived features achieved 83–87% accuracy [63].

These results demonstrate that *AutoMISC*'s summary scores can serve as evaluators of counselling quality. This highlights the potential for applications of automated coding in MI quality assessment. In the next chapters, we employ *AutoMISC* to analyze transcripts from two different versions of an ongoing project on talk therapy counselling for moving towards the decision to quit smoking.

Chapter 5

Design: MIBot v6.3B

As described in Chapter 1, a secondary goal of this work is to apply *AutoMISC* to automated MI conversations, to extract insights from the in-session dynamics, beyond traditional effectiveness metrics. This chapter describes the design and deployment of the two chatbot versions tested and the associated human studies. The work in this chapter and the subsequent one is part of the larger, ongoing MIBot project, involving contributions from many team members. Section 6.5 provides attribution for the contributions.

The predecessor to this work, MIBot v5.2 [13], used a hybrid approach of five scripted responses and five generative reflections. The sixth generation of MIBot represented a major architectural shift to fully generative chatbots, taking advantage of the conversational capabilities of modern LLMs. This shift raised key design questions of how responses should be generated and how much control to exert over the process. The first experimental version (MIBot v6.3A) tested whether the knowledge encoded in the prompt and parameters of a single-prompted LLM could produce high-quality MI responses end-to-end without further control. The results from this experiment have been published in [44]. To explore a higher degree of control, MIBot v6.3B decomposes the task into two stages, described in the following section. The results for version 6.3B have not yet been published, so they are included in this dissertation for the first time in Chapter 6.

The shift to fully-generative counsellors introduced new challenges in managing the now free-flowing conversations, such as determining when a conversation should end, and ensuring that generated responses are appropriate and safe. We therefore developed an observer agent framework in which auxiliary LLM agents provide this oversight. This infrastructure was shared across both versions and is described in Section 5.2.

5.1 MIBot v6.3B Initial Design

The goal of the chatbot is to engage a client in an initial MI conversation about smoking cessation by generating a context-appropriate volley in response to the client's most volley, given the conversation history. Unlike v6.3A, v6.3B decomposes the volley generation into two steps, each handled by a separate prompted LLM: (1) The *selection* of the therapeutic technique, handled by a Behavioural Code Selector Agent and (2) The *generation* of the corresponding counsellor response, handled by the Response Generator Agent. Figure 5.1 illustrates an overview of the MIBot v6.3B system.

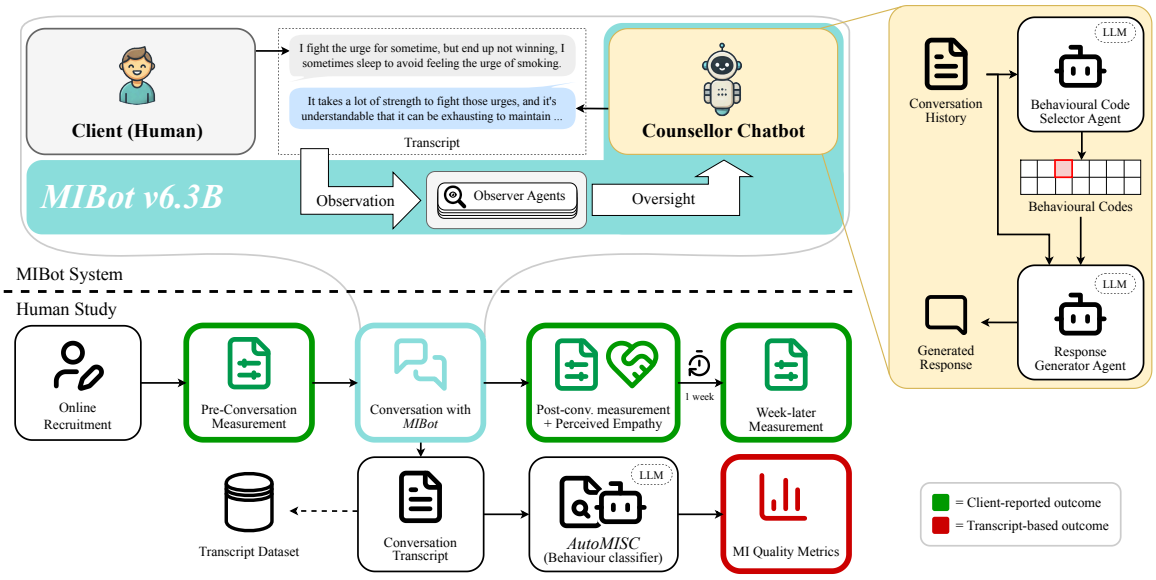


Figure 5.1: Overview of the MIBot v6.3B System.

5.1.1 Behavioural Code Selector Agent

The Behavioural Code Selector Agent gives the type of technique the counsellor should use to respond to the client's most recent volley, given the conversation history. The repertoire of possible techniques was initially based on the behavioural codes in the Motivational Interviewing Treatment Integrity manual (MITI) [52]. Although there is significant overlap between the MITI and MISC, MITI was chosen here because it was designed for efficiency in clinical trials and supervision, with a smaller label space (10 codes vs 19). This was later expanded to 14 codes through iterative prompt engineering in collaboration with expert MI clinicians. The codes are listed in Table 5.1.

Option	Behavioural Code
1	Giving Information
2	Persuade with Permission
3	Open Question
10	Closed Question
4	Simple Reflection
5	Complex Reflection
6	Double-Sided Reflection
7	Affirmation
8	Seeking Collaboration (Permission)
9	Emphasizing Autonomy
11	Seeking Collaboration (Feedback)
12	Structuring Statement
13	Summary
14	End of Conversation

Table 5.1: Available Options for the Behavioural Code Selector Agent

The first version of the prompt simply instructed the model to choose a contextually appropriate behavioural code. This later evolved to a Chain-of-Thought (CoT)-style prompt in which the model

was asked to provide the following information to inform its technique selection:

- Summarize the last client volley, conversation progress, counsellor's contributions to the session thus far, whether the counsellor has requested client permission, and if the client agreed.
- Analyze the last client volley for change/sustain talk, if trust/goals have been established, and the direction/progress of the conversation
- Given the available options, select one and provide reasoning for its use case
- Identify conversation stage (Establishing Trust, Establishing Goals, Exploring Options, Concluding Conversation)
- Re-state appropriate option to output.

To minimize the influence of the model's internal knowledge of MITI codes, the options were anonymized in the prompt ("Option 1", "Option 2", etc.) rather than described with their true names. The prompt is available in Appendix D.1.

5.1.2 Generation of Counsellor Response

The Response Generator Agent outputs a coherent, context-aware response to the client's most recent volley based on the selected technique and the conversation history. This prompt is the same as the system prompt from MIBot v6.3A [44], but with an additional templated instruction to generate an utterance of the the selected technique. The model used in the experiments was GPT-4o[1], the most powerful model available at the time. The prompt and templated descriptions for each code are available in Appendix D.1.

5.2 MIBot v6.3 Infrastructure

In both versions of MIBot v6.3, the counsellor bot is situated in a framework of autonomous observer agents, each one a prompted LLM (GPT-4o) responsible for monitoring a specific aspect of the ongoing conversation. Each turn, prior to the generation of the counsellor response, each observer examines the conversation, and some may intervene as necessary, as described below:

5.2.1 Moderator

The *moderator* reviews the counsellor's most recent utterance and determines whether it could potentially harm the client. While OpenAI's internal guardrails [55] 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 both versions of the chatbot, the re-generated counsellor utterance succeeded within four generation attempts and never failed to produce an acceptable utterance. The prompt for this observer agent is given in Appendix D.2.

[1] gpt-4o-2024-08-06

5.2.2 Off-Track Classifier

We were concerned that some of our paid participants (see Section 5.3) might intentionally steer the conversation far off from the topic of smoking, so we built a classifier to monitor conversations in real-time to detect this. 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. The prompt for this observer agent is given in Appendix D.2.

5.2.3 End Classifier

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. The prompt for this observer agent is given in Appendix D.2.

5.2.4 Usage Tracker

This observer tracks the number of turns, character length of conversation, number of tokens, and API cost. It raises a flag for the counsellor to end the conversation if it exceeds 900 characters in length within 15 minutes, with additional checks every five minutes thereafter up until 30 minutes, where the flag is raised unconditionally. While this approach unfortunately resulted in some conversations being cut short, it was a necessary design choice to align with our study compensation model, which paid participants for their time, described in the following section.

5.3 MIBot v6.3 Implementation Details

Both versions of MIBot v6.3 are implemented as Python backend web servers using the Sanic[68] framework, and hosted on Amazon Web Services (AWS). The conversations and observer agents are implemented as Python classes. For the user interface, we re-use the text-only chat frontend from the previous version, MIBot v5.2. Participants interact with the chatbot through the Prolific online behavioural research platform (www.prolific.com) [60], which allows researchers to recruit participants and manage studies. To enable concurrent usage while maintaining ability to monitor sessions in real time, the backend server runs with four worker processes, where each worker can handle any participant's incoming request, update the shared conversation state, and then become immediately available to serve the next request. Conversation state is made consistent across workers using a shared in-memory dictionary of active conversation objects, implemented using Sanic's shared context mechanism. To match backend capacity, we limit the number of concurrent participants to four on the Prolific recruitment platform. In a production-scale deployment, this architecture would be extended using an external state store such as Redis or a database to persist and share conversation state across multiple machines or containers.

5.4 Feasibility Study with Human Smokers

5.4.1 Participant Recruitment

A total of 93 English-speaking participants were recruited to evaluate the capability of MIBot v6.3B through the Prolific online recruitment platform [60]. The inclusion criteria for participants were: they must be adults (18+), fluent in English, have at least 90% approval rate on prior tasks performed on the Prolific platform, and must be current smokers of at least five cigarettes per day. Participants were compensated a total of £6.50 for completing two tasks one week apart. This group was also filtered from a larger group of 165 participants to select those who exhibited low confidence that they will succeed in quitting[2]. Finally, the recruitment was set to enrol equal numbers of male and female participants. Although the balance was affected by the above filter, the final sex proportion was evenly split (47 male, 46 female). Participant ages ranged from 18-73 years old, with a median of 40 years (mean=40, SD=12). The median time taken to complete the conversational part of the study was 23 minutes (mean=25, SD=9). Appendix D.3 provides more details on participant demographics.

5.4.2 Ethics Approval

The MIBot v6.3 studies were approved by the University of Toronto Research Ethics Board (REB) under protocol #49997 [65]. All participants completed an electronic consent form describing the study procedure, compensation, and data collection. All personally identifiable information was de-identified for research and release. Study data for version 6.3A is released on GitHub[3].

5.4.3 Study Design

The study design followed a pattern commonly employed in MI research (e.g., [77, 41, 27, 16, 40]) and therapeutic chatbot evaluations (e.g., [13, 30]). Participants in our study were taken through the following four steps (illustrated in the lower half of Figure 5.1):

1. In a **pre-conversation survey**, participants rated themselves on the **readiness ruler** survey (see Table 2.2).
2. Participants then engaged in a conversation with the counsellor chatbot described in Section 5.1, 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 [46, 10], 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 D.4).
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 [11, 27], and the most predictive part of the ruler is the self-reported confidence to succeed, which we used as our primary metric of

[2] 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.

[3] https://github.com/cimhasgithub/MIBOT_ACL2025

effectiveness [29, 1].

5.4.4 *AutoMISC*: Assessment of Counsellor and Client Language

We apply *AutoMISC* (described in Chapter 3) to the transcripts of both versions of the chatbot using the best-performing model and parameter configuration (GPT-4.1 with three context volleys using the hierarchical classification approach), yielding 199 MI conversations that are parsed from volleys into utterances and labelled at the utterance level according to the MISC 2.5 scheme. From the sequences of codes, we calculate the MISC summary scores: Percentage MI-Consistent (%MIC), Reflection-to-Question ratio (R:Q), and Percentage Change Talk (%CT) (described in Section 2.2.2). The first two metrics are provisional measures of MI quality, where higher %MIC and R:Q between 1-2 are indicative of counsellor proficiency. The third metric is a measurement of the client's motivation to change, with higher values associated with better behavioural outcomes. However, these static ratios provide only a coarse summary of each conversation and overlook the temporal dynamics of the session. In Chapter 7, we take a finer-granularity look at the relationship between the client codes and the post-session outcomes.

Chapter 6

Results: Human Study

In this chapter we give the results for the MIBot 6.3B experiment, including the Readiness Rulers, CARE Survey, and MISC summary scores. We compare it to its contemporary MIBot v6.3A as well as its predecessor, MIBot v5.2, and human healthcare professionals where applicable.

6.1 Effect on Participants' Readiness to Quit Smoking

Recall from Section 5.1 that the human participants in the MIBot studies completed the readiness ruler survey on three occasions: immediately prior to the conversation with the chatbot, immediately following 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 [29]. Table 6.1 presents data at those points in time for each of the three readiness rulers: importance, confidence, and readiness.

Ruler Attribute	**Ver.**	**Pre-Conv.**	**Post-Conv.**	**1Wk Later**	**Δ(1Wk –Pre)**	***p*-value (Δ)**
Importance	5.2	5.5 (2.9)	6.0 (2.8)	6.2 (2.8)	0.7 (2.0)	$<0.001^{\dagger}$
	6.3A	5.7 (2.6)	6.3 (2.9)	6.1 (2.7)	0.5 (1.7)	<0.005
	6.3B	6.1 (2.6)	6.4 (2.6)	6.1 (2.9)	0.0 (2.2)	0.68
Confidence	5.2	3.3 (2.3)	4.1 (2.5)	4.7 (2.7)	1.3 (2.0)	$<0.001^{\dagger}$
	6.3A	2.8 (2.0)	4.6 (2.6)	4.5 (2.7)	**1.66 (2.4)**	<0.001
	6.3B	2.7 (1.8)	4.1 (2.3)	4.5 (2.4)	**1.73 (2.2)**	<0.001
Readiness	5.2	4.9 (2.8)	5.3 (2.7)	5.4 (2.9)	0.4 (1.7)	$<0.01^{\dagger}$
	6.3A	5.2 (2.8)	5.9 (2.8)	5.5 (3.0)	0.3 (2.4)	0.22
	6.3B	5.0 (2.8)	5.6 (2.6)	5.3 (2.6)	0.3 (2.1)	0.21

Table 6.1: Average (SD) ratings on Readiness Ruler Survey on Importance, Confidence, and Readiness to quit smoking. Statistical significance using Wilcoxon signed-rank test († = two-tailed t-test).

Both MIBot v6.3A and v6.3B achieved a significant average increase in confidence of +1.7 on the ten-point scale, with v6.3B scoring slightly higher than v6.3A. For comparison, MIBot v5.2 (described in Chapter 5's introduction) reported an average change in confidence of +1.3. While none of the works are directly comparable to one another due to the differences in time of experiment and

starting average confidence, all works recruited a similar number of low-confidence participants. As another point of comparison, human counsellors in a prior study [69] achieved an average confidence increase of +2.5 points after five MI sessions delivered over a ten-week period.

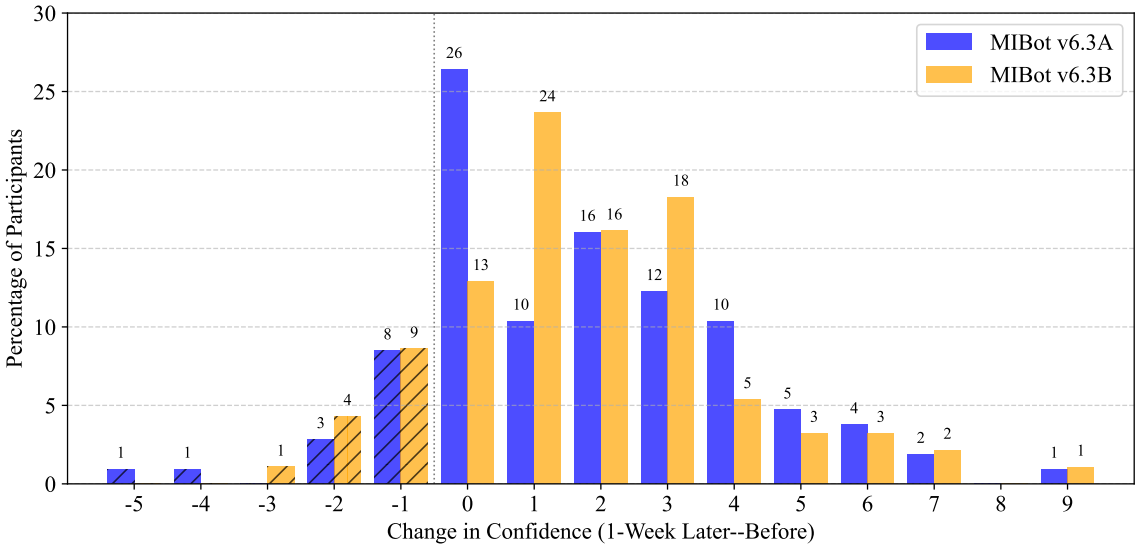


Figure 6.1: Distribution of Change in Confidence (1-Week Later – Before Conversation).

Figure 6.1 presents the distribution of week-later changes in confidence scores for v6.3A and v6.3B. Both distributions are right-skewed, but v6.3B shows a broader peak between +1 and +3, while version 6.3A is more sharply centered at 0, with 26% of participants reporting no change in confidence. Decreases in confidence were rare across both versions, with 13% in v6.3A and 14% in v6.3B, mostly falling by 1-2 points, with a few outliers.

In regards to the other readiness ruler attributes, version 6.3A showed a significant average increase of 0.5 in the participants' perceived importance of quitting, while 6.3B showed no statistically significant change. Version 5.2 showed modest but significant increases across all ruler attributes, slightly outperforming v6.3A on importance and readiness. The changes in the remainder of readiness ruler attributes across all versions were not statistically significant.

Table 6.2 shows that both the starting average confidence levels and week-later changes varied by demographic group across MIBots version 5.2, 6.3A, and 6.3B. While all subgroups experienced statistically significant gains, larger increases were observed for participants that are younger (under 30), non-white, or employed full-time. Also, female participants and older participants report lower starting confidence levels across all versions. Notably, female participants reported a greater increase in confidence with version 6.3B than with 6.3A ($\Delta = 2.0$ vs 1.7), while the opposite pattern occurred for male participants ($\Delta = 1.5$ vs 1.7). The largest subgroup effect size occurred for non-white participants, particularly for versions 5.2 and 6.3A. All participant demographics are given in Table D.3.1 in Appendix D.3.

Demo. Factor	Value	Ver.	Count, n (%)	Pre-conv.	Post-conv.	1Wk Later	Δ(1Wk – Pre)
Sex	Male	5.2	49 (49)	3.5 (2.6)	4.2 (2.9)	4.7 (2.8)	1.2 (2.5)
		6.3A	49 (46)	3.2 (1.7)	4.7 (2.2)	4.9 (2.5)	1.7 (2.3)
		6.3B	47 (51)	3.0 (2.1)	4.1 (2.4)	4.5 (2.4)	1.5 (2.0)
	Female	5.2	51 (51)	3.1 (1.9)	3.9 (2.1)	4.6 (2.6)	1.5 (2.2)
		6.3A	57 (54)	2.5 (2.1)	4.4 (2.8)	4.1 (2.9)	1.7 (2.5)
		6.3B	46 (49)	2.5 (1.5)	4.2 (2.1)	4.4 (2.4)	**2.0 (2.4)**
Age	<30	5.2	60 (60)	3.6 (2.5)	4.1 (2.6)	5.0 (2.7)	1.4 (2.5)
		6.3A	26 (25)	3.7 (2.1)	5.5 (2.5)	5.7 (2.7)	1.9 (3.1)*
		6.3B	21 (23)	3.4 (1.6)	4.5 (2.1)	5.3 (2.6)	1.9 (2.2)
	30+	5.2	40 (40)	3.0 (2.0)	4.1 (2.3)	4.1 (2.5)	1.2 (2.0)
		6.3A	80 (75)	2.5 (1.8)	4.3 (2.5)	4.1 (2.6)	1.6 (2.1)
		6.3B	72 (77)	2.5 (1.8)	4.0 (2.3)	4.2 (2.3)	1.7 (2.2)
Ethnicity	White	5.2	65 (65)	3.2 (2.3)	3.8 (2.4)	4.0 (2.5)	0.8 (1.9)
		6.3A	80 (75)	2.7 (1.9)	4.3 (2.6)	4.0 (2.6)	1.4 (2.2)
		6.3B	72 (77)	2.6 (1.8)	4.0 (2.2)	4.3 (2.3)	1.8 (2.3)
	Other	5.2	35 (35)	3.5 (2.3)	4.7 (2.6)	5.8 (2.6)	2.3 (2.7)
		6.3A	26 (25)	3.3 (2.0)	5.3 (2.4)	5.8 (2.8)	**2.5 (2.7)**
		6.3B	21 (23)	3.3 (1.9)	4.5 (2.6)	4.9 (2.7)	1.6 (2.0)*
Employment status	Full-Time	5.2	41 (41)	3.2 (1.9)	4.6 (2.5)	5.0 (2.7)	1.8 (2.2)
		6.3A	49 (46)	3.2 (1.9)	4.8 (2.3)	5.1 (2.6)	1.9 (2.3)
		6.3B	50 (54)	2.8 (1.5)	4.3 (2.3)	4.7 (2.5)	1.9 (2.5)
	Other	5.2	59 (59)	3.4 (2.6)	3.7 (2.4)	4.4 (2.6)	1.1 (2.4)
		6.3A	57 (54)	2.5 (2.0)	4.3 (2.8)	3.9 (2.8)	1.4 (2.4)
		6.3B	43 (46)	2.7 (2.1)	3.9 (2.2)	4.2 (2.3)	1.5 (1.9)

Table 6.2: Average (SD) self-reported confidence to quit smoking reported at different points in the studies (MIBot v5.2, 6.3A, 6.3B), segmented by demographic factor. Statistical significance calculated using one-sided Wilcoxon signed-rank test ($^* = p < 0.01$, all others $p < 0.001$).

Counsellor	CARE Score	% Perfect Score
MIBot v5.2	36	3
MIBot v6.3A	**42**	**11**
MIBot v6.3B	36	9
Humans*	40	N/A

Table 6.3: Average CARE scores and (%) perfect scores for MIBot v5.2, MIBot v6.3A/B and *human healthcare practitioners [37].

6.2 CARE Survey for Perceived Empathy

Each participant rated their perceived empathy of the chatbot via the CARE survey [46]. Table 6.3 presents the mean CARE scores for the three versions of MIBot discussed, alongside the average CARE score reported in a meta-analysis of 64 studies involving human healthcare practitioners[37]. Figure 6.2 provides the distribution of CARE scores among participants in each version, and Figure 6.3 shows the question-wise mean CARE score for each version.

MIBot v6.3A clearly outperforms the other versions in perceived empathy, achieving the highest

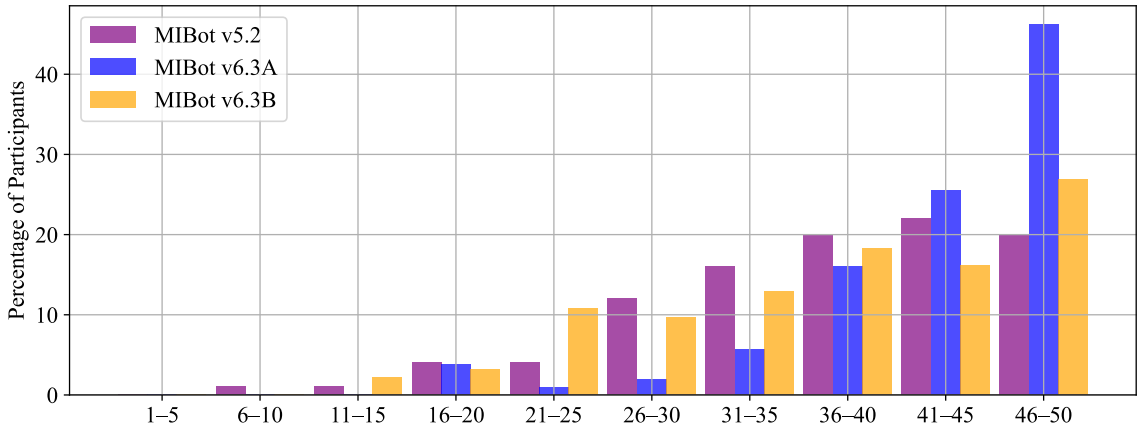


Figure 6.2: Distribution of CARE scores for MIBot v5.2, 6.3A, and 6.3B.

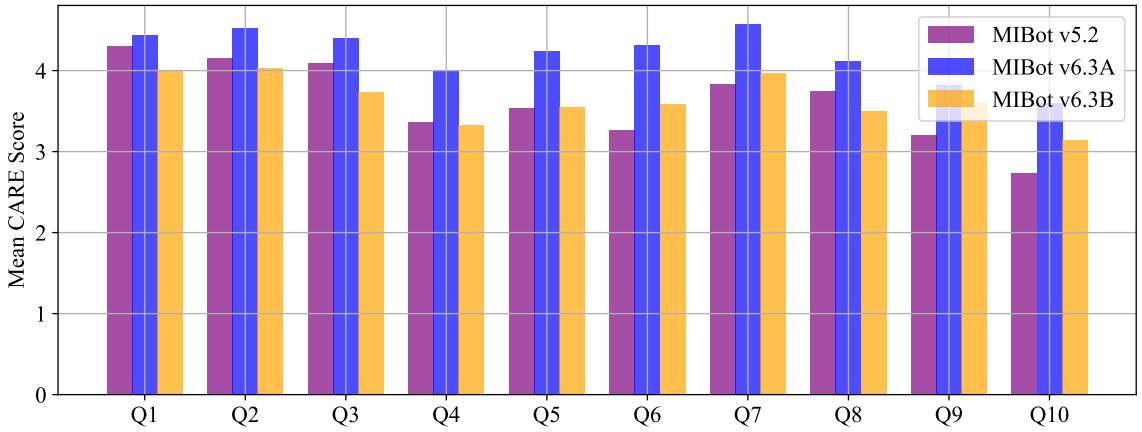


Figure 6.3: Question-wise mean CARE scores for MIBot v5.2, 6.3A, and 6.3B.

average CARE score of 42 and a higher mean score on all 10 questions, also exceeding the human benchmark of 40. The distribution is heavily left-skewed with the majority of participants assigning scores in the upper ranges, while the other versions have more broad distributions. Surprisingly, MIBot v6.3B performed poorly, receiving the same score (36) as MIBot v5.2, a rigid conversation of five scripted questions and generated reflections. Version 6.3B falls behind v5.2 on the first three questions, namely "making you feel at ease" (Q1), "letting you tell your story" (Q2), and "really listening" (Q3), suggesting that the decomposition of responses may interfere with conversational rhythm. All versions perform poorly on questions 9 and 10, which are about "helping you take control" and "making a plan of action with you", although v6.3B lags further behind v6.3A on these questions compared to others.

6.3 Assessing Counsellor Adherence to MI and Client Motivation

AutoMISC was applied to the 93 transcripts from the MIBot v6.3B study, as well as the 106 transcripts

Dataset	%MIC	R:Q	%CT
HLQC_LO	44 (26.9)	1.0 (2.1)	37 (33.8)
HLQC_HI	88 (14.6)	2.3 (3.9)	**68 (26.5)**
MIV6.3A	**99 (0.01)**	0.9 (0.3)	**68 (26.7)**
MIV6.3B	**99 (0.02)**	**1.7 (0.6)**	64 (24.6)

Table 6.4: Comparison of MISC summary metrics in MIBot v6.3 experiment transcripts and the HLQC dataset.

from MIBot v6.3A. As a point of comparison, we use the **HighLowQualityCounselling (HLQC)** dataset [63] introduced in Section 4.3.2, which consists of real MI counselling sessions labelled with a binary quality rating (154 high-quality, 104 low-quality). We compute the MISC summary scores separately for each session in each subset and then compared both versions of MIBot v6.3 against them. Table 6.4 summarizes the computed MISC metrics across datasets, and Figures 6.4-6.6 illustrate the distribution of the session-level metrics within each dataset via violin plots. We also show the distributions of the raw Tier 1 codes for both counsellor and client language in each dataset in Figure 6.7.

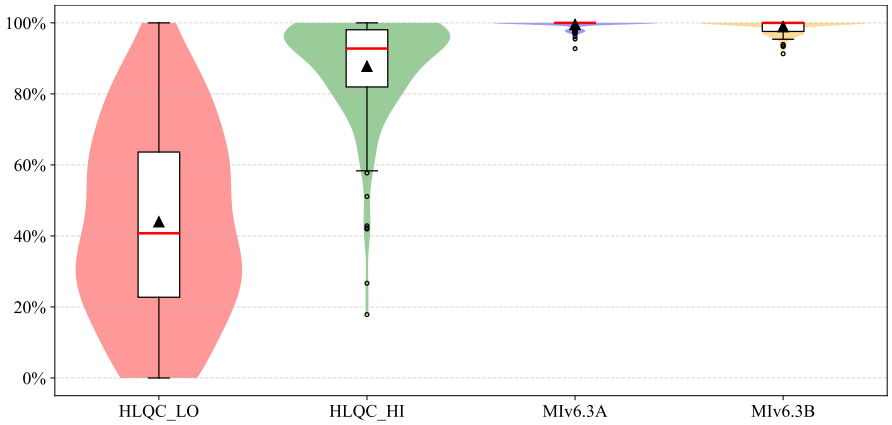


Figure 6.4: Distribution of Percent MI-Consistent Responses (%MIC) across sessions in each dataset.

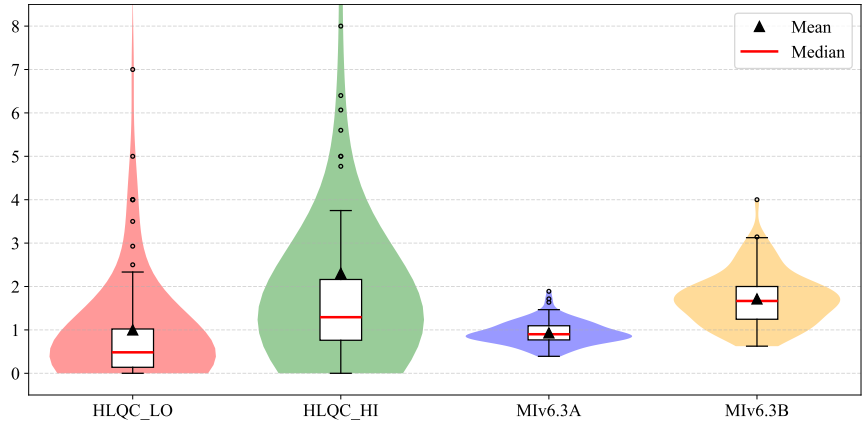


Figure 6.5: Distribution of Reflection-to-Question Ratio (R:Q) across sessions in each dataset.

In both versions of MIBot v6.3, a very high fraction of the chatbot counsellor utterances are

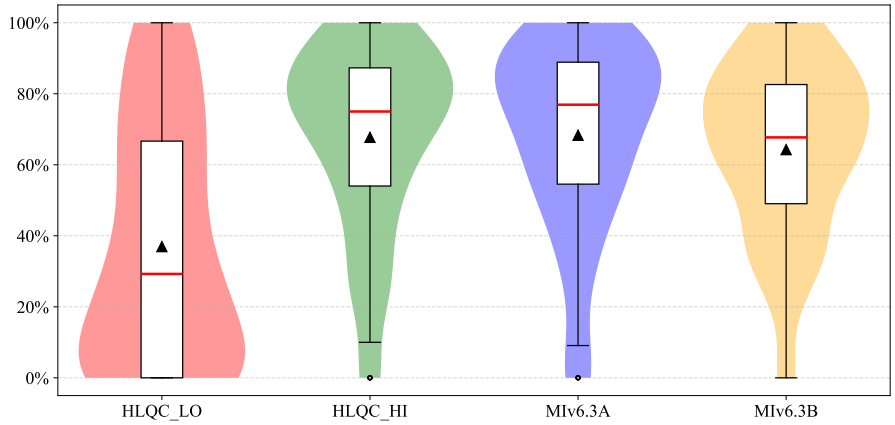


Figure 6.6: Distribution of Percent Client Change Talk (%CT) across sessions in each dataset.

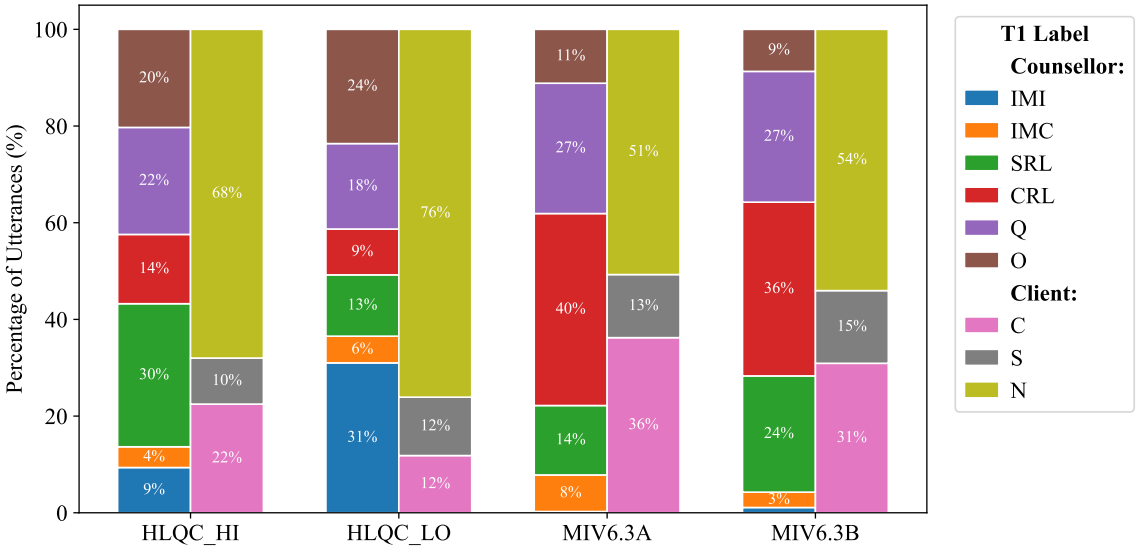


Figure 6.7: Distribution of Tier 1 behavioural codes for counsellor and client language across datasets.

MI-compliant (%MIC in the table), exceeding those in the high-quality dataset with far less variance. While the average R:Q ratio for version 6.3B falls within the recommended range of 1-2, version 6.3A falls just short of it, surprisingly lower than the low-quality subset, although with much lower variance. Finally, Version 6.3A achieves a 68% Client Change Talk, matching that of HLQC_HI subset in mean and variance, surpassing 6.3B at 64%, and sharply contrasting with the 37% in the HLQC_LO subset.

6.4 Discussion

While both versions of MIBot v6.3 achieved comparable improvements in participants' confidence to quit smoking, the differences in performance yield insights into automated counsellor response strategy. As shown in Table 6.1, version 6.3B had a more consistent effect across participants, with 73% of participants reporting an increase in confidence compared to 61% in 6.3A, and slightly lower standard deviation (2.2 vs 2.4). However, version 6.3A showed a small but statistically significant

increase in importance to quit ($\Delta = 0.5$, $p < 0.005$), whereas version B showed no statistically significant change on average ($\Delta = 0.0$, $p = 0.68$).

We hypothesize that this difference arises from the decomposition of response generation in version 6.3B, which may interfere with the model's ability to generate a natural and emotionally resonant response. In version 6.3B, the Response Generator Agent is constrained by the Behavioural Code Selector Agent's choice, even if a more nuanced or contextually appropriate strategy may be available, that may be more fluently expressed using the flexible single-prompted approach of 6.3A. This likely weakens the fluidity and perceived authenticity of responses in 6.3B, leading to dimished emotional engagement, and weaker impact on abstract motivational attributes like importance to change.

This is further directly supported by the poor performance of version 6.3B on the CARE survey, which is substantially lower than version 6.3A (36 vs 42; see Table 6.3) and matches the average CARE score of MIBot v5.2, where the conversation consisted of a rigid script of five questions and five reflections. Moreover, version 6.3B exhibits a higher R:Q ratio than 6.3A (1.7 vs 0.9; see Table 6.4) but a lower Percentage Client Change Talk (64% vs 68%), suggesting that while decomposition may help enforce adherence to MI best practices, it may simultaneously have negative implications on client engagement.

With these results are not directly comparable, they reveal a design tradeoff: exertion of control on response generation can enhance consistency and reliability, but may indirectly dampen the therapeutic alliance with the client.

6.5 Chatbot Contribution Attribution

The work presented in Chapters 5 and 6 is part of a large, long-term project with many collaborators. Below we describe the various contributions of each team member as well as the author.

- Zafar Mahmood: As first author on the MIBot v6.3A paper [**mahmood2025fully**], Mahmood led the design and deployment of version 6.3A. He was a key actor in the evolution the system prompt in consultation with expert MI clinicians on our research team and tested it using synthetic clients. Mahmood was also responsible for setting up the cloud infrastructure for the experiments, including deployment on AWS ECS and storage via S3. He also secured ethics approval and conducted the deployment and data collection for the MIBot v6.3A study.
- Michelle Collins: Collins contributed to the development of the Behavioural Code Selector Agent used in version 6.3B, by evolving its prompt in collaboration with expert MI clinician-scientist collaborators. She also contributed to the system prompt of version 6.3A.
- Sihan Chen: Chen implemented a backend function to upload conversation artifacts to AWS S3. He also contributed to the End Classifier observer's prompt.
- Yi Chen (Michael) Zhao: Zhao developed the Off-Track observer agent, which detects when the topic of the conversation diverges from smoking cessation.

The author of this work made the following contributions:

- Designed and implemented the modular observer agent framework underlying both versions of MIBot v6.3, which houses the counsellor bot and the associated observer agents.

- Contributed early versions of ***AutoMISC*** during the development of MIV6.3A, providing analysis that informed its iterative refinement.
- Designed an implemented MIBot v6.3B, introducing the decomposition of response generation into two stages.
- Deployed MIBot v6.3B to participants using the Prolific platform, managing recruitment, compensation, cloud deployment, and data collection.

Chapter 7

Motivational Trajectories and Correlation with Post-Therapy Outcome

A core assumption in MI is that client language influences and shapes downstream behavioural outcomes. In this chapter we explore how client language modeled as a sequence of MISC codes relates to post-therapy outcomes, by applying the *AutoMISC* system to the transcripts of the MIBot v6.3A/B human studies and looking for relationships or associations between the sequences of codes and the readiness ruler values reported by the experiment participants.

The MISC 2.5 summary scores, such as percent change talk (%CT), offer a coarse measure of client motivation, but they obscure the progression of motivation through a session. Prior work [3] showed that the change in strength of client *commitment* language (a subset of change talk) over a session is a good predictor of drug use outcomes at follow-up. This motivates the idea to visualize MI transcripts by plotting utterance behavioural codes over time, an idea common in talk therapy research [35].

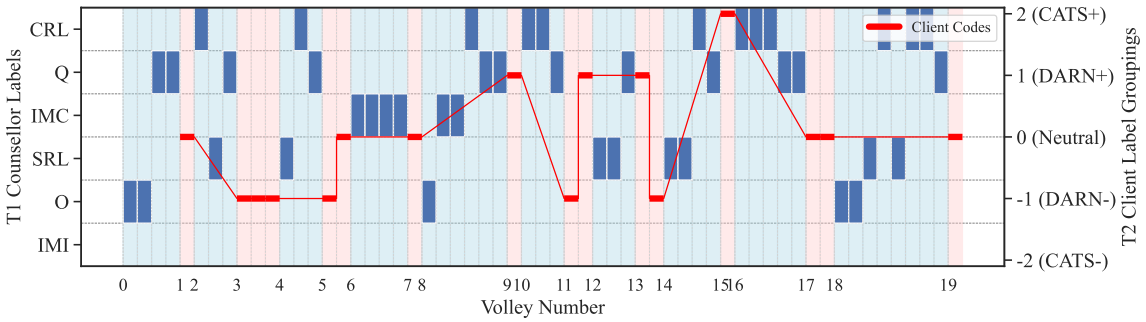


Figure 7.1: Example Visualization of MI session. Red: Client Speech codes. Blue bars: Counsellor Speech T1 codes

7.1 Visualization of Client Motivation Trajectories

Figure 7.1 shows an example *conversational trajectory* which is derived from ***AutoMISC*** codes of counsellor and client speech in a session from the MIBot v6.3A experiment. The x-axis shows progression along the session in two ways: the thin vertical lines delineate an utterance, while the solid blue or pink colour delineates a complete volley composed of one or more utterances. The left Y-axis shows the Tier 1 categories of the counsellor speech that were determined by ***AutoMISC***. The right Y-axis gives the Tier 2 categories of client speech ordered from the bottom as the strongest sustain talk, and at the top to be the strongest change talk, with neutral talk in the middle. Figure 7.1 shows a trajectory for a session in which the client's talk shows a somewhat upward trend from sustain talk to strong change talk. It also gives a sense of the kinds of MI skills that the counsellor was employing. We feel that this level of detail could play a useful role in the evaluation of the skills of the counsellor and the impact of the session on the client. In the next section we illustrate the latter with a metric computed from the client speech (red line) trajectory. As outlined in subsection 2.2.2, MISC 2.5 client codes can be interpreted along a "motivational axis": Change talk maps to positive values, sustain talk to negative values, and neutral talk to zero. By mapping the codes to numerical values, normalizing by session length, and plotting each client utterance across time, we derive a *motivational trajectory* for the client over the course of a session. Specifically, we map 17 T2 client codes to five numerical values based on the preparatory vs. mobilizing distinction to accommodate for variation in client expressiveness. The mapping can be found in Table 7.1.

Tier 2 Client Codes	**Score**
Commitment+, Activation+, Taking Steps+	**2**
Desire+, Ability+, Reason+, Need+	**1**
Other+, Neutral, Other−	**0**
Desire−, Ability−, Reason−, Need−	**-1**
Commitment−, Activation−, Taking Steps−	**-2**

Table 7.1: Mapping of Tier 2 client codes to numerical motivation scores.

7.2 Correlation of Client Code Sequence to Therapy Outcome

In this section we show how the sequence of client codes can be used to derive a quantitative metric that correlates with a therapy outcome. The metric, which is called the *motivation slope* is computed as the slope of a linear regression fit to the motivational trajectory of the client over the course of a session (the red line in Figure 7.1). We apply this analysis to the transcripts and outcome data from MIBot v6.3A ($n = 106$) and MIBot v6.3B ($n = 93$). The motivational trajectories are extracted by ***AutoMISC*** using its best-performing configuration, GPT-4.1 with three context volleys and the hierarchical classification approach. The outcome variable is the change in self-reported confidence to quit smoking [29, 1], measured one week after the conversation compared to the pre-conversation baseline.

Figure 7.2 shows two illustrative example trajectories from MIV6.3A: one in which the client confidence change was +5 a week later and trajectory is rising (orange), and one with a change of -2 and a falling trajectory (blue). Figure 7.3 shows scatterplots of motivational slope vs. week-later

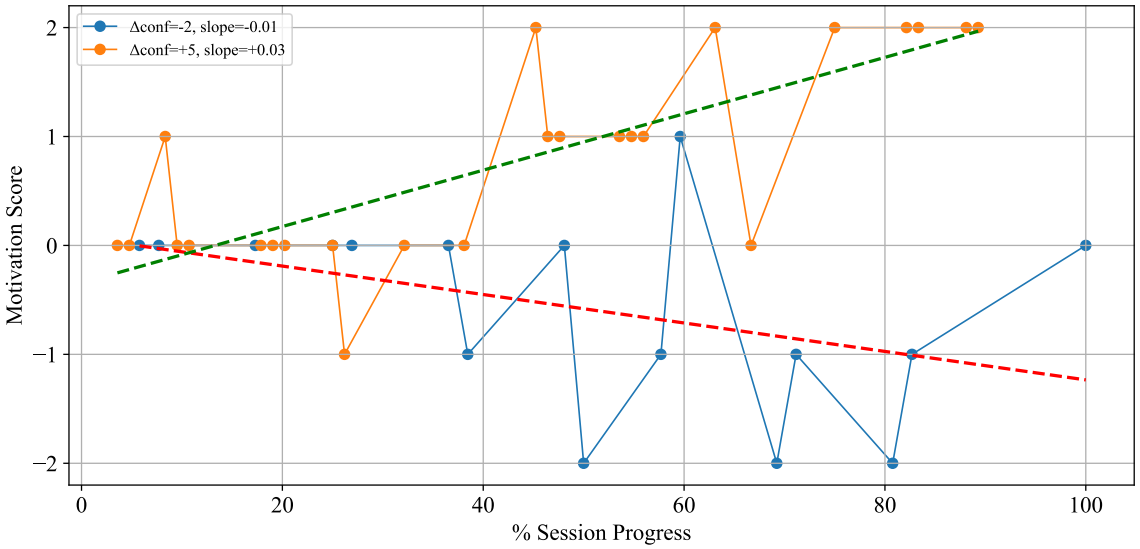
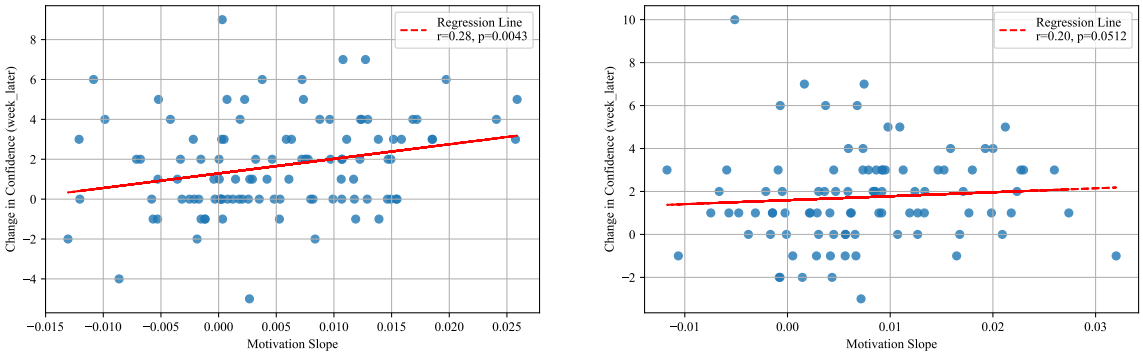


Figure 7.2: Two sample client motivation trajectories from the MIBot v6.3A study.

client-reported change in confidence for all the transcripts from each study using the different versions of MIBot v6.3. Figure 7.4 and Figure 7.5 display overlays of all motivational trajectories in each version and their associated regression lines. The green lines are from those clients with positive confidence change, and the red lines are for clients with negative change in confidence.

For each chatbot version, we compute Spearman's correlation between the change in confidence and five session-level features: pre-session confidence, motivation slope, and the three MISC summary scores, Percent MI-Consistent Responses (%MIC), Reflection-to-Question Ratio (R:Q), and Percent Client Change Talk (%CT). The results are in Table 7.2. To account for multiple comparisons, we apply a Bonferroni correction over 10 tests ($\alpha = 0.005$).



(a) MIBot v6.3A

(b) MIBot v6.3B

Figure 7.3: Motivational slope vs. week-later change in confidence scatterplot for MIBot v6.3A/B.

The aggregate pattern of progression of client motivation across both versions (see Figure 7.4 and 7.5) shows a general upward trend in motivational scores. The early part of sessions (0-40% progress) is denser with preparatory sustain talk ($y = -1$), while the latter half shows greater activity above $y = 1$, indicating a shift towards mobilizing change talk. This is complemented by the lower density

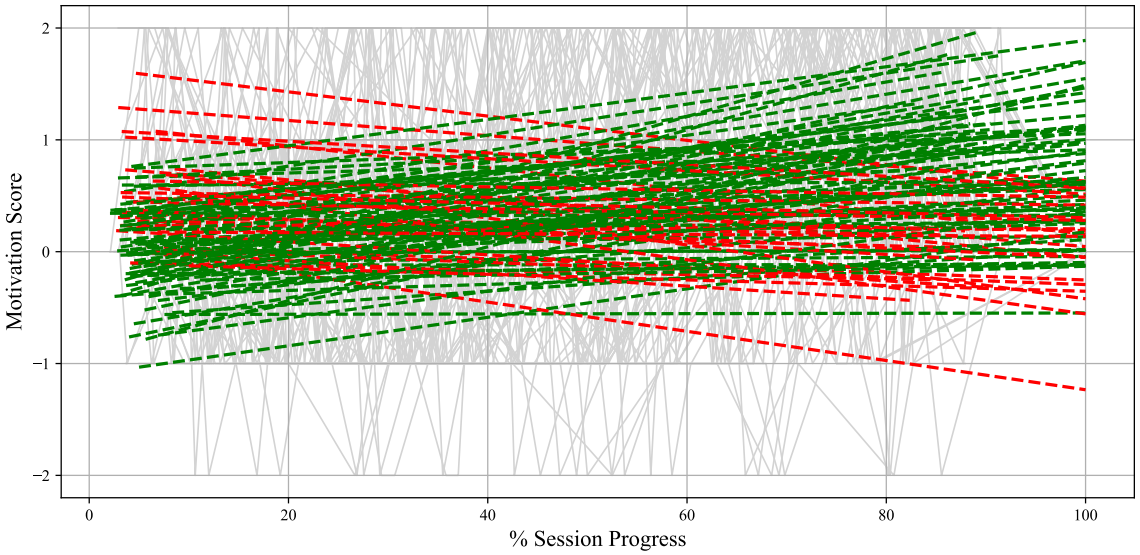


Figure 7.4: All client trajectory slopes for MIBot v6.3A.

Feature	**Ver.**	**Spearman** r	p**-value**
Pre-confidence	6.3A	-0.11	0.26
	6.3B	-0.25	<0.05
Motivation slope	6.3A	**0.28**	**<0.005**
	6.3B	0.20	0.051
% MIC	6.3A	0.01	0.07
	6.3B	0.05	0.64
R:Q	6.3A	0.10	0.32
	6.3B	-0.04	0.67
% CT	6.3A	0.17	0.08
	6.3B	0.28	<0.01

Table 7.2: Spearman correlations between session-level features and week-later change in confidence, split by chatbot version. Significant after Bonferroni correction at $\alpha = 0.005$ shown in bold.

below $y = 0$ in the latter half of the session.

Table 7.2 shows that for MIBot v6.3A, motivational slope was significantly correlated with week-later change in confidence ($r = 0.28$, $p < 0.005$), outperforming all other conversation features including pre-session confidence, %MIC, R:Q ratio, and %CT, none of which have statistically significant correlation. In v6.3B, the correlation was slightly weaker with marginal statistical significance ($r = 0.20$, $p = 0.051$), and did not survive Bonferroni correction. Interestingly, for v6.3B, the pre-confidence and %CT each had stronger correlations with the change in confidence and greater statistical significance than motivational slope, though they also did not survive Bonferroni correction. We note that the Bonferroni correction is considered to be conservative [78].

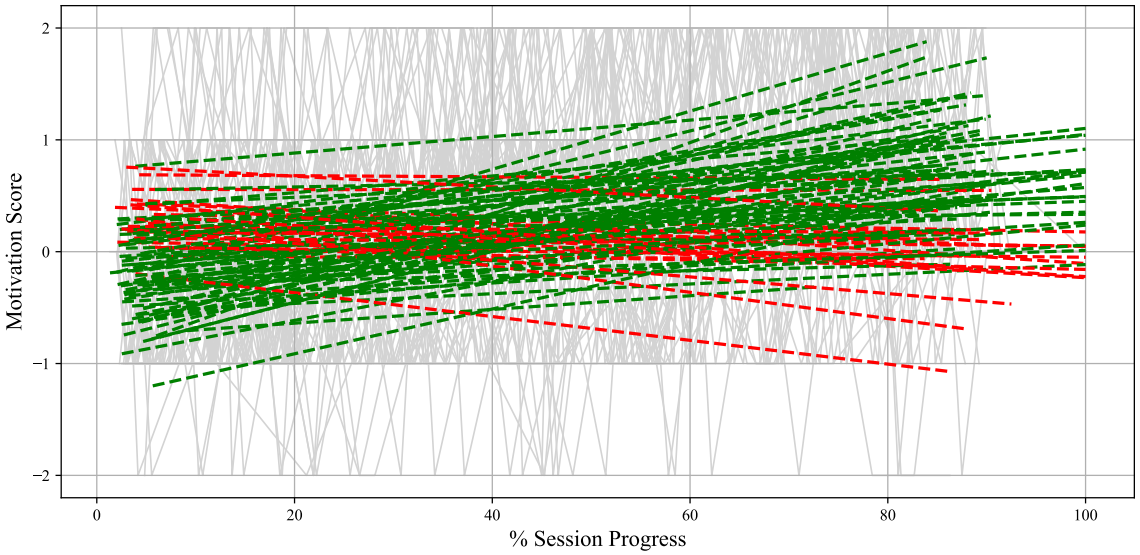


Figure 7.5: All client trajectory slopes for MIBot v6.3B.

7.3 Discussion

These results highlight the value of fine-grained sequential analysis of client language over the static session-level summary metrics. Unlike the MISC summary scores which collapse the session into a single aggregate value, trajectory-based analyses preserve the temporal dynamics of the session, an important consideration given that client motivation may flucutate rapidly even within a single volley [49]. MI counsellors are trained to recognize and respond to these moment-to-moment shifts, so there is utility in methods that capture such fluctuations.

For MIBot v6.3A, the motivational slope was more strongly correlated with the change in confidence than any other conversational feature tested. This serves as a form of validation for *AutoMISC*: it is not only capable of classifying utterances with reasonable reliability (see Section 4), but producing sequences of behavioural codes that reflect core assumptions in MI theory. However, as outlined in the preceding section, this effect was only found to be statistically significant for version 6.3A, as the results for 6.3B reflect an opposing narrative.

We hypothesize that the decomposition of response generation using two LLM agents in 6.3B may have inadvertently reduced the fluidity of the counsellor's responses. By constraining the model to first select a technique prior to generation, the system may have defaulted to a narrower, more set of techniques, as discussed in Section 6.4. This may have weakened the therapeutic alliance, thereby reducing client openness and increasing the prevalence of neutral talk or ambiguous motivational signals. In contrast, the single-prompted v6.3A offered greater expressive freedom to the model, which may have allowed the model to establish and maintain a deeper therapeutic alliance with the client, thereby eliciting clearer motivational progressions.

Chapter 8

Conclusions, Future Work and Limitations

8.1 Summary

This dissertation introduces *AutoMISC*, an LLM-based system for fully automated utterance-level annotation of counsellor and client speech in Motivational Interviewing (MI) transcripts under the MISC 2.5 framework. *AutoMISC* achieves classification performance equal to or exceeding prior approaches on expert-aligned annotations, and demonstrates alignment with annotations in existing datasets like AnnoMI. We use the annotations to predict the binary ratings of session quality in the High-Low Quality Counselling (HLQC) dataset. *AutoMISC* works both with state-of-the-art APIs and locally hosted models, making it suitable for use in privacy-sensitive settings such as talk therapy. We further apply *AutoMISC* to analyze the transcripts from a fully generative MI chatbot intervention for smoking cessation.

Two versions of the chatbot were evaluated: a single-prompt counsellor (v6.3A), and a two-agent design (v6.3B) which exerts greater control over generation by separating technique selection from response generation. These chatbots operate within a system of observer agents to manage the conversation. Both versions show potential in moving tobacco smokers towards the decision to quit, based on the measured increases in client-reported confidence to quit measured before and one week after the conversation. Applying *AutoMISC* to the transcripts yielded nuanced, language-based insights that complemented the readiness rulers and perceived empathy measurements. Comparing the MISC summary metrics to the HLQC dataset indicated strong adherence to MI in both versions.

To model relationships between client language and post-therapy outcomes, we propose a metric called *motivation slope* to quantify the progression of client motivation within an MI session. This metric correlates significantly with the week-later change in confidence in version 6.3A, with marginal significance for 6.3B. Collectively, these findings demonstrate the potential synergy of automated MI behavioural coding and automated MI talk therapy.

8.2 Limitations

8.2.1 *AutoMISC*

While ***AutoMISC*** delivers promising results in automating MI behaviour coding, several limitations should be noted. First, the consensus labels we used as ground truths were not directly labeled by MI experts, but instead by annotators aligned by experts. Despite our effort in iteratively refining the labels to meet the IRR threshold, one could argue that such indirect supervision may introduce discrepancies and limit the fidelity of our consensus labels. This concern is reinforced by the imbalance in the consensus label set. Second, while our system is grounded in the MISC 2.5 framework [36], it does not strictly follow all recommended coding procedures, such as doing a first pass and providing global scores before parsing and assigning behaviour codes, nor does it rely on modalities beyond text, such as vocal and visual cues that are essential for accurate interpretation and coding. Our proposed two-tiered coding flow was also designed heuristically and not grounded in MISC 2.5 or any other prior MI literature, whose validity and utility need to be confirmed by future research. Third, our validation experiments are imperfect due to limitations and constraints from the datasets used. For the AnnoMI [83] dataset, there might be inconsistencies in the mapping between the MISC labels and their custom volley-level coding scheme; for the HLQC [63] dataset, the high and low quality labels for transcripts are provided using their own custom criteria, thus may not always reflect the true quality of the transcripts. Finally, while we demonstrate ***AutoMISC*** 's ability to run on local models to address privacy concerns, our best results are still achieved using proprietary models such as GPT-4.1, leaving room for future work to improve open-source models further and provide better guarantees in regards to privacy.

8.2.2 MIBot v6.3 experiments

As with behavioural coding, the delivery of MI is influenced heavily by paralinguistic and visual cues such as intonation, timing, and facial expression. Our text-only interface may therefore have affected both behavioural coding and the in-session therapeutic dynamics themselves. The participants in our studies were paid, and may have been incentivized to report more favourable post-therapy outcomes. Although the primary outcome (change in confidence to quit) is a validated predictor of subsequent behaviour change, it remains an indirect measure of treatment effectiveness. Additionally, the results across different versions of MIBot are not directly comparable, as they did not follow protocol standards for randomization. Finally, only one model (GPT-4o) was tested in the MIBot v6.3 experiments, which limits generalizability across models.

8.3 Future Work

Classification tools like ***AutoMISC*** can be used within fully automated MI systems to track client state and counsellor adherence to MI, supplying reward signals that can steer counsellor generation towards appropriateness and safety. These signals can also support the training and evaluation of human MI counsellors through instantaneous feedback and simulated practice. Future studies should adopt stronger clinical designs. In particular, multi-session longitudinal study protocols and randomized control trials are needed to properly assess treatment effectiveness. With larger datasets, it will

be possible to directly predict treatment outcomes from the behavioural codes and conversational trajectories produced by *AutoMISC*. Future MI chatbots should expand beyond the textual interface to incorporate additional modalities. Deployment of systems with audio support is a concrete next step that may improve both coding fidelity and counsellor-client dynamics.

Appendix A

AutoMISC System Design Supplementary Material

Appendices A.1 and A.2 show the prompts for each of the core components of the *AutoMISC* system.

A.1 Parser Module Prompt

The Parser module is fed a system prompt, followed by several input-output pairs from the MISC manual ("few-shots"), and finally the target volley for parsing. It is constrained to return a list of strings using a structured output schema (defined using Pydantic). The prompt and few-shot examples are as follows:

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.

```
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.'']
```

A.2 Annotator Module Classification Prompts

The annotator module uses either a hierarchical or flat classification approach. In the hierarchical approach, the model first chooses a Tier 1 code, then selects a Tier 2 code from the subset associated with that Tier 1 category. Following the classification prompt, the annotator module is given a configurable number of volleys prior to the target utterances as context for classification, then the target utterance itself, templated in another prompt we call the "User Prompt". The model output is constrained using a structured output schema (Pydantic) to return only an explanation string and one code abbreviation from either the T1 or T2 grouping. Below we list out the Tier 1, Tier 2 and Flat classification prompts for both counsellor and client, as well as the user prompt.

Tier 1 Counsellor Prompt

You are an expert annotator of Motivational Interviewing (MI) counselling sessions. Your task is to classify the counsellor's final utterance in a given session excerpt into one of the following groupings of MISC 2.5 behavioural codes:

Classification Categories:

1. **C-Reflective (CRL)** - Deeply engages with or affirms the client's perspective.
    - *Behavioural Codes*: Affirm (AF), Support (SU), Complex Reflection (CR), Reframe (RF), Emphasize Control (EC)
    - **Affirm (AF)**: Communicates something positive or complimentary about the client's strengths or efforts.
    - **Support (SU)**: Sympathetic, compassionate, or understanding comments, which agree or side with the client.
    - **Complex Reflection (CR)**: A reflective listening statement that adds significant meaning or emphasis to what the client said, conveying a deeper or richer picture of the client's statement.
    - **Reframe (RF)**: Suggests a different meaning for an experience expressed by the client, usually changing the emotional valence of meaning but not the depth.
    - **Emphasize Control (EC)**: Acknowledges, honours, or emphasizes the client's autonomy and freedom of choice.
2. **S-Reflective (SRL)** - Mirrors or paraphrases the client's statement without adding extra insight (includes summarizing statements).
    - *Behavioural Codes*: Simple Reflection (SR)
    - **Simple Reflection (SR)**: A reflective listening statement which simply repeats or paraphrases the client's words or meaning, often with a slight change in wording or

emphasis.

3. **Imperative-MICO (IMC)** - **With client permission**, provides advice, raises a concern, or gives information.
    - *Behavioural Codes*: Advise with Permission (ADP), Raise Concern with Permission (RCP), Give Information (GI)
    - **Advise With Permission (ADP)**: After receiving permission, gives advice, makes a suggestion, or offers a solution or possible action.
    - **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.
    - **Giving Information (GI)**: Provides information to the client, explains something, educates or provides feedback, or discloses personal information.
4. **Imperative-MIIN (IMI)** - **Without client permission**, provides advice, raises a concern, warns, directs, or confronts the client.
    - *Behavioural Codes*: Advise Without Permission (ADW), Raise Concern Without Permission (RCW), Warn (WA), Direct (DI), Confront (CO)
    - **Advise Without Permission (ADW)**: Offers suggestions or guidance WITHOUT asking or receiving permission.
    - **Raise Concern Without Permission (RCW)**: 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.
    - **Direct (DI)**: Gives an order, command, or direction. The language is imperative.
    - **Confront (CO)**: Directly disagrees, argues, corrects, shames, blames, seeks to persuade, criticizes, judges, labels, moralizes, ridicules, or questions the client's honesty.
5. **Question (Q)** - Asks a question in order to gather information, understand, or elicit the client's story.
    - *Behavioural Codes*: Open Question (OQ), Closed Question (CQ)
    - **Open Question (OQ)**: A question is open only if it cannot be answered with "yes" or "no" in any grammatically valid or logically plausible way. The question must structurally require an elaboration, explanation, or descriptive narrative that goes beyond a binary or fixed-option response.
    - **Closed Question (CQ)**: A question is closed if it is about confirmation, factual information-seeking, curiosity about presence/absence of something, request for specific information or choices, or *it can be answered with "yes" or "no" under any grammatically valid interpretation, even if that answer is awkward, contextually unhelpful, or unlikely.*
6. **Other/Neutral (O)** - Structural or facilitative utterances that do not engage in MI techniques.
    - *Behavioural Codes*: Filler (FI), Facilitate (FA), Structure (ST)
    - **Filler (FI)**: Pleasantries such as "good morning", "nice weather we're having", etc.
    - **Facilitate (FA)**: Simple utterance that functions as a "keep-going" acknowledgement e.g. "Mm-hmm", "I see", "Go on"
    - **Structure (ST)**: Used to make a transition from one topic or part of a session to another. Also used to give information about will happen directly to the client throughout the course of treatment or within a study format, in this or subsequent sessions.

Category assignment instructions

1. **General instructions**
   (a) Analyze the given context and counsellor's final utterance.
   (b) Identify its primary function.
   (c) If the utterance involves **advice, suggestions, or information**, follow the **Permission Chain of Thought Guide** below before choosing between **IMC** and **IMI**.
   (d) For other types of utterances, assign the category directly.
   (e) Justify your choice in 1-2 sentences for category assignment except IMI and IMC.
2. **Permission Chain of Thought (Only when assigning IMC or IMI)**
   When the utterance involves **giving advice, suggestions, guidance, or information** (when deciding between **IMC** or **IMI**), you **must first apply this step-by-step reasoning** to determine if permission is given:
   (a) Check for Client Permission or Interest Expression: Examine the recent client utterances in the provided context. Has the client given permission—either explicitly by directly asking for advice, suggestions, or ideas or implicitly by showing openness, curiosity, or requesting information in a way that reasonably invites guidance. Both explicit and implicit permission are equally valid—there is no difference in weight between them. If either is present, permission is considered granted.
   (b) Check for Counsellor's Prior Permission-Seeking: Has the counsellor previously asked for permission to give advice, suggestions, or information and received agreement? If so, permission is also considered granted.
   (c) If Yes (to 1 or 2): Classify the utterance as IMC (permission has been granted).
   (d) If No: Classify the utterance as IMI (no permission has been granted).
   (e) **Carry Permission Forward:** Once permission—explicit or implicit—is granted, it remains **active** for all **topically related** suggestions, guidance, or information, **even if the counsellor's next utterance introduces a shift in topic or phrasing**. **Do NOT revoke permission just because the surface topic evolves naturally**, as long as the advice remains part of the **same overarching discussion or client goal**. **Permission only expires** if there is a **clear and substantive topic shift**, or if the client **disengages** or **withdraws interest**. In most cases, permission is granted in **recent client utterances**, but **prior permissions—especially implicit ones—can remain valid across multiple counsellor turns** if the conversation stays aligned with the client's intent or focus. You should **assume permission is still valid** unless there is strong evidence that the advice no longer relates to the client's earlier request, concern, or area of engagement. **Apply this permission reasoning chain ONLY when the utterance's function is to provide advice, suggestions, guidance, or information.**
   For all other categories (**CRL, SRL, Q, O**), permission is **not relevant**. Assign these categories based on their definitions without using this permission reasoning.

Output Format

- **explanation**: Use brief reasoning for all category assignments expect IMC and IMC. When the category is IMC or IMI, use the full chain of thought for determining permission as the justification.
- **label**: Provide only "CRL", "SRL", "IMC", "IMI", "Q", or "O".

Tier 2 Counsellor Prompt Template

You are an expert annotator of Motivational Interviewing (MI) counselling sessions. Your task is to assign a category label to the counsellor's final utterance in a given session excerpt.
Classification Categories
The utterance must be assigned one of the following labels:

{{spec}}

Output Format

- **explanation**: Briefly justify your choice in 1-2 sentences.
- **label**: Provide only the appropriate label.

Final instructions

1. Analyze the counsellor's final utterance.
2. Identify its primary function and intent.
3. Provide a brief explanation for your choice.
4. Assign the appropriate label based on the categories provided above.

The {{spec}} parameter is replaced by one of the following depending on what the Tier 1 code was:

```
CRL: |
  - **Complex Reflection (CR)**: A reflective listening statement that adds significant meaning or emphasis
      to what the client said, conveying a deeper or richer picture of the client's statement.
  - **Affirm (AF)**: Communicates something positive or complimentary about the client's strengths or
      efforts.
  - **Support (SU)**: Sympathetic, compassionate, or understanding comments, which agree or side with the
      client.
  - **Reframe (RF)**: Suggests a different meaning for an experience expressed by the client, usually
      changing the emotional valence of meaning but not the depth.
  - **Emphasize Control (EC)**: Acknowledges, honours, or emphasizes the client's autonomy and freedom of
      choice.
SRL: |
  - **Simple Reflection (SR)**: A reflective listening statement which simply repeats or paraphrases the
      client's words or meaning, often with a slight change in wording or emphasis.
IMC: |
  - **Advise With Permission (ADP)**: After receiving permission, gives advice, makes a suggestion, or
      offers a solution or possible action.
  - **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.
  - **Giving Information (GI)**: Provides information to the client, explains something, educates or
      provides feedback, or discloses personal information.
IMI: |
  - **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
Q: |
  - **Closed Question (CQ)**: A question is closed if it can be answered with ``yes'' or ``no'' under any
      grammatically valid interpretation, even if that answer is awkward, contextually unhelpful, or
      unlikely.
       To determine is a question is CQ, always check for its grammatical structure first. If the utterance
            can be interpreted in a way that permits a yes/no response, you must classify it as CQ.
       This includes any form of:
          - any utterance containing or beginning with grammatical constructions that use auxiliary or
              modal verbs, existence/presence checks, or binary/framed prompts must be labeled as CQ.
              These include, but are not limited to, questions that:
            - Begin with or contain modal/auxiliary verbs such as:
              Can, Could, Do, Does, Did, Are, Is, Was, Were, Will, Would, Have, Has, Had, Might, May,
                  Should, Shall, Must followed by a subject and verb/complement.
            - Ask about existence, availability, or presence using forms like:
              Is there, Are there, Do you have, Have you got, Would it be, Could it be, Might it be, Is it
                   possible that...
            - Implicitly or explicitly present binary choices or confirmatory framing, including
                structures like:
              Do you ever, Would you say, Are you thinking about, Would you like, Is this something you,
                  Have you thought about, Do you feel like, Do you think, Does it feel like, Do you
                  notice...
            If the utterance contains any clause that permits a grammatically valid yes/no or short
                factual response, even if additional elaboration is possible, it must be labeled CQ.
          - even if it appears to invite elaboration.
          - confirmation or factual information-seeking
          - curiosity about presence/absence of something
          - request for specific information or choices
       If there is any ambiguity between CQ and OQ, always label it as CQ.
  - **Open Question (OQ)**:
       A question is open only if it cannot be answered with ``yes'' or ``no'' in any grammatically valid
           or logically plausible way.
       The question must structurally require an elaboration, explanation, or descriptive narrative that
           goes beyond a binary or fixed-option response.
       Questions that seem to encourage elaboration but could be reduced to a yes/no response are still CQ,
            not OQ.
       Use this label only when there is no grammatical path to yes/no answers -- no exceptions.
O: |
  - **Facilitate (FA)**: Simple utterance that functions as a ``keep-going'' acknowledgement e.g. ``Mm-hmm
      '', ``I see'', ``Go on''
  - **Filler (FI)**: Pleasantries such as ``good morning'', ``nice weather we're having'', etc.
  - **Structure (ST)**: Gives information about will happen directly to the client throughout the course of
       treatment or within a study format, in this or subsequent sessions.
```

Tier 1 Client Prompt

You are an expert annotator of Motivational Interviewing (MI) counselling sessions. Your task is to assign a category label to the client's final utterance in a given session excerpt.
Classification Categories
The utterance must be assigned one of the following labels:

1. **Change Talk (C)** - The client expresses a stance toward **changing** the target behavior.
    - **Commitment** to change (e.g., stating/implying an intention to change, considering

alternatives, making plans to change).
- **Reasons** for change (including personal, health, or emotional factors).
- **Desire** to change (e.g., "I really want to quit.").
- **Optimism** about their ability to change (e.g., "I think I can do it.").
- **Need** to change (e.g., "I have to stop before it gets worse.").
- **Recent steps** toward change (e.g., "I cut back this week.").

2. **Sustain Talk (S)** - The client expresses a stance toward **maintaining** the target behavior.
   - **Commitment** to maintaining the target behaviour (e.g., stating/implying an intention to continue, dismissing alternatives, making plans to continue).
   - **Reasons** for maintaining the target behaviour (e.g., stress relief, social reasons).
   - **Desire** to continue the target behaviour (e.g., "I enjoy it too much to quit.").
   - **Pessimism** about their ability to change (e.g., "I don't think I can quit.").
   - **Need** to maintain the target behaviour (e.g., "I need cigarettes to cope.").
   - **Recent steps** reinforcing the target behaviour (e.g., "I bought another pack today.").
3. **Neutral (N)** - The utterance does not clearly support or oppose change.
   - Following along with the counsellor without expressing a stance.
   - Asking questions (e.g., "What are the benefits of quitting?").
   - Providing factual or general statements about the behaviour.

Output Format

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

Tier 2 Client Prompt Template

You are an expert annotator of Motivational Interviewing (MI) counselling sessions. Your task is to assign a category label to the client's final utterance in a given session excerpt.
Classification Categories
The utterance must be assigned one of the following labels:

{{spec}}

Output Format

- **explanation**: Briefly justify your choice in 1-2 sentences.
- **label**: Provide only the appropriate label.

Final instructions

1. Analyze the client's final utterance.
2. Identify its primary function and intent.
3. Provide a brief explanation for your choice.
4. Assign the appropriate label based on the categories provided above.

The {{spec}} parameter is replaced by one of the following depending on what the Tier 1 code was:

```
C: |
   - **Desire (D+)**: The client expresses a desire to change the target behaviour, e.g. ``I want to quit
```

```
      smoking''.
  - **Ability (AB+)**: The client expresses optimism about their ability to change, e.g. ``I think it's
      possible for me to quit''.
  - **Reasons (R+)**: The client provides reasons for changing the target behaviour, e.g. ``My children are
       begging me to quit''.
  - **Need (N+)**: The client expresses a need to change the target behaviour, e.g. ``I've got to quit
      before it gets worse''.
  - **Commitment (C+)**: The client expresses a commitment to change, e.g. ``I'm going to quit smoking''.
  - **Activation (AC+)**: The client leans towards action, e.g. ``I'm willing to give it another try''.
      This includes suggestions of alternatives to the target behaviour.
  - **Taking Steps (TS+)**: The client mentions recent steps towards change, e.g. ``I cut back on smoking
      this week''.
  - **Other (O+)**: The client makes a statement that supports change but does not fit into the other
      categories. This usually includes problem recognition or hypotheticals.
S: |
  - **Desire (D-)**: The client expresses a desire to maintain the target behaviour, e.g. ``I enjoy smoking
       too much to quit''.
  - **Ability (AB-)**: The client expresses pessimism about their ability to change, e.g. ``I don't think I
       can quit''.
  - **Reasons (R-)**: The client provides reasons for maintaining the target behaviour, e.g. ``Smoking is
      the only way I can relax''.
  - **Need (N-)**: The client expresses a need to maintain the target behaviour, e.g. ``I need to have my
      morning cigarettes''.
  - **Commitment (C-)**: The client expresses a commitment to maintain the target behaviour, e.g. ``I'm not
       going to quit smoking''.
  - **Activation (AC-)**: The client leans towards inaction, e.g. ``I'm not ready to quit yet''. This
      includes suggestions of maintaining the target behaviour.
  - **Taking Steps (TS-)**: The client mentions recent steps reinforcing the target behaviour, e.g. ``I
      bought two packs today''.
  - **Other (O-)**: The client makes a statement that supports maintaining the target behaviour but does
      not fit into the other categories. This usually includes problem recognition or hypotheticals.
N: |
  - The utterance does not clearly support or oppose change. There is no further categorization, so just
      use ``N''.
```

Flat Counsellor Prompt

You are an expert annotator of Motivational Interviewing (MI) counselling sessions. Your task is to classify the counsellor's final utterance in a given session excerpt into one of the following groupings of MISC 2.5 behavioural codes:

Classification Categories:
The utterance must be assigned one of the following labels:

- **Affirm (AF)**: Communicates something positive or complimentary about the client's strengths or efforts.
- **Support (SU)**: Sympathetic, compassionate, or understanding comments, which agree or side with the client.
- **Complex Reflection (CR)**: A reflective listening statement that adds significant meaning or emphasis to what the client said, conveying a deeper or richer picture of the client's statement.
- **Reframe (RF)**: Suggests a different meaning for an experience expressed by the client,

usually changing the emotional valence of meaning but not the depth.

- **Emphasize Control (EC)**: Acknowledges, honours, or emphasizes the client's autonomy and freedom of choice.
- **Simple Reflection (SR)**: A reflective listening statement which simply repeats or paraphrases the client's words or meaning, often with a slight change in wording or emphasis.
- **Advise With Permission (ADP)**: After receiving permission, gives advice, makes a suggestion, or offers a solution or possible action.
- **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.
- **Giving Information (GI)**: Provides information to the client, explains something, educates or provides feedback, or discloses personal information.
- **Advise Without Permission (ADW)**: Offers suggestions or guidance WITHOUT asking or receiving permission.
- **Raise Concern Without Permission (RCW)**: 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.
- **Direct (DI)**: Gives an order, command, or direction. The language is imperative.
- **Confront (CO)**: Directly disagrees, argues, corrects, shames, blames, seeks to persuade, criticizes, judges, labels, moralizes, ridicules, or questions the client's honesty.
- **Open Question (OQ)**: A question is open only if it cannot be answered with "yes" or "no" in any grammatically valid or logically plausible way. The question must structurally require an elaboration, explanation, or descriptive narrative that goes beyond a binary or fixed-option response.
- **Closed Question (CQ)**: A question is closed if it is about confirmation, factual information-seeking, curiosity about presence/absence of something, request for specific information or choices, or *it can be answered with "yes" or "no" under any grammatically valid interpretation, even if that answer is awkward, contextually unhelpful, or unlikely.*
- **Filler (FI)**: Pleasantries such as "good morning", "nice weather we're having", etc.
- **Facilitate (FA)**: Simple utterance that functions as a "keep-going" acknowledgement e.g. "Mm-hmm", "I see", "Go on"
- **Structure (ST)**: Used to make a transition from one topic or part of a session to another. Also used to give information about will happen directly to the client throughout the course of treatment or within a study format, in this or subsequent sessions.

Category assignment instructions

1. **General instructions**
   (a) Analyze the given context and counsellor's final utterance.
   (b) Identify its primary function.
   (c) If the utterance involves **advice, suggestions, or information**, follow the **Permission Chain of Thought Guide** below before choosing ADP, ADW, RCP, or RCW.
   (d) For other types of utterances, assign the category directly.
   (e) Justify your choice in 1-2 sentences for category assignment except ADP, ADW, RCP, or RCW.
2. **Permission Chain of Thought (Only when assigning IMC or IMI)**
   When the utterance involves **giving advice, suggestions, guidance, or information** , you

must first apply this step-by-step reasoning to determine if permission is given:

(a) Check for Client Permission or Interest Expression: Examine the recent client utterances in the provided context. Has the client given permission—either explicitly by directly asking for advice, suggestions, or ideas or implicitly by showing openness, curiosity, or requesting information in a way that reasonably invites guidance. Both explicit and implicit permission are equally valid—there is no difference in weight between them. If either is present, permission is considered granted.

(b) Check for Counsellor's Prior Permission-Seeking: Has the counsellor previously asked for permission to give advice, suggestions, or information and received agreement? If so, permission is also considered granted.

(c) If Yes (to 1 or 2): You may classify the utterance as ADP/RCP (permission has been granted).

(d) If No: Classify the utterance as ADW/RCW (no permission has been granted).

(e) **Carry Permission Forward:** Once permission—explicit or implicit—is granted, it remains **active** for all **topically related** suggestions, guidance, or information, **even if the counsellor's next utterance introduces a shift in topic or phrasing**. **Do NOT revoke permission just because the surface topic evolves naturally**, as long as the advice remains part of the **same overarching discussion or client goal**. **Permission only expires** if there is a **clear and substantive topic shift**, or if the client **disengages** or **withdraws interest**. In most cases, permission is granted in **recent client utterances**, but **prior permissions—especially implicit ones—can remain valid across multiple counsellor turns** if the conversation stays aligned with the client's intent or focus. You should **assume permission is still valid** unless there is strong evidence that the advice no longer relates to the client's earlier request, concern, or area of engagement. **Apply this permission reasoning chain ONLY when the utterance's function is to provide advice, suggestions, guidance, or information.**

Output Format

- **explanation**: Use brief reasoning for all category assignments except ADP/ADW/RCP/RCW. When the category is one of these, use the full chain of thought for determining permission as the justification.
- **label**: Provide only the appropriate label abbreviation.

Flat Client Prompt

You are an expert annotator of Motivational Interviewing (MI) counselling sessions. Your task is to classify the client's final utterance in a given session excerpt into one of the following groupings of MISC 2.5 behavioural codes:

Classification Categories:
The utterance must be assigned one of the following labels:

- **Desire+ (D+)**: The client expresses a desire to change the target behaviour, e.g. "I want to quit smoking".
- **Ability+ (AB+)**: The client expresses optimism about their ability to change, e.g. "I think it's possible for me to quit".

- **Reasons+ (R+)**: The client provides reasons for changing the target behaviour, e.g. "My children are begging me to quit".
- **Need+ (N+)**: The client expresses a need to change the target behaviour, e.g. "I've got to quit before it gets worse".
- **Commitment+ (C+)**: The client expresses a commitment to change, e.g. "I'm going to quit smoking".
- **Activation+ (AC+)**: The client leans towards action, e.g. "I'm willing to give it another try". This includes suggestions of alternatives to the target behaviour.
- **Taking Steps+ (TS+)**: The client mentions recent steps towards change, e.g. "I cut back on smoking this week".
- **Other+ (O+)**: The client makes a statement that supports change but does not fit into the other categories. This usually includes problem recognition or hypotheticals.
- **Desire- (D-)**: The client expresses a desire to maintain the target behaviour, e.g. "I enjoy smoking too much to quit".
- **Ability- (AB-)**: The client expresses pessimism about their ability to change, e.g. "I don't think I can quit".
- **Reasons- (R-)**: The client provides reasons for maintaining the target behaviour, e.g. "Smoking is the only way I can relax".
- **Need- (N-)**: The client expresses a need to maintain the target behaviour, e.g. "I need to have my morning cigarettes".
- **Commitment- (C-)**: The client expresses a commitment to maintain the target behaviour, e.g. "I'm not going to quit smoking".
- **Activation- (AC-)**: The client leans towards inaction, e.g. "I'm not ready to quit yet". This includes suggestions of maintaining the target behaviour.
- **Taking Steps- (TS-)**: The client mentions recent steps reinforcing the target behaviour, e.g. "I bought two packs today".
- **Other- (O-)**: The client makes a statement that supports maintaining the target behaviour but does not fit into the other categories. This usually includes problem recognition or hypotheticals.
- **Neutral (N)**: The utterance does not clearly support or oppose change. This can include following along with the counsellor without expressing a stance, asking questions (e.g., "What are the benefits of quitting?"), or providing factual or general statements about the behaviour.

Output Format

- **explanation**: Briefly justify your choice in 1-2 sentences.
- **label**: Provide only the appropriate label abbreviation.

Final Instructions

1. Analyze the counsellor's final utterance.
2. Identify its primary function and intent.
3. Provide a brief explanation for your choice.
4. Assign the appropriate label based on the categories provided above.

User Prompt

Session Transcript
The following is an excerpt of a MI counselling session transcript:

{{ transcript }}

Target Utterance for Classification
Below is the target {{ speaker }} utterance in the session excerpt:

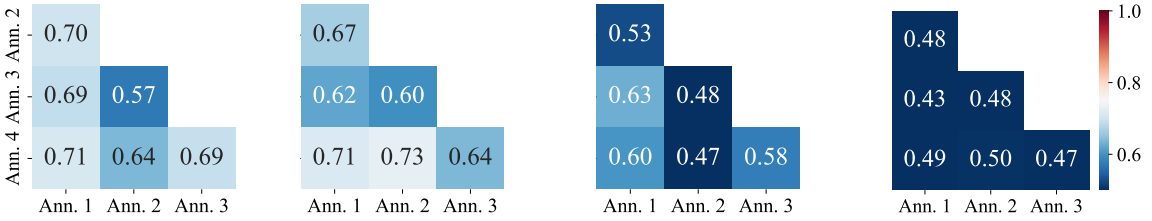
{{ utterance }}

Appendix B

Expert Alignment of Annotations

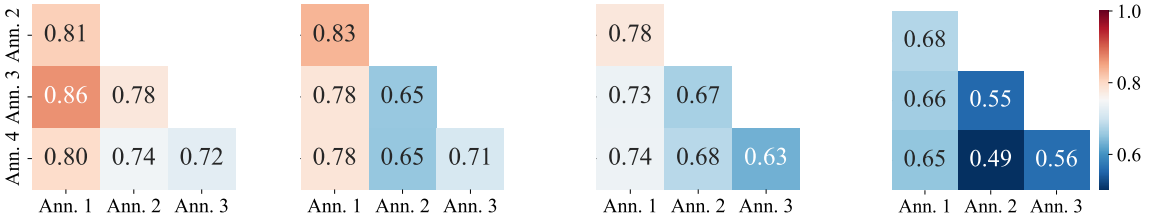
B.1 Inter-rater reliability before vs. after alignment

Figures B.1.1 and B.1.2 show the Cohen's Kappa between each pair of manual annotators before and after alignment, respectively. The process is described in full in Section 3.1.3.



T1 Counsellor (a) Fleiss' κ = **0.66**
T1 Client (b) Fleiss' κ = **0.66**
T2 Counsellor (c) Fleiss' κ = **0.54**
T2 Client (d) Fleiss' κ = **0.47**

Figure B.1.1: Pairwise Cohen's Kappa (and Fleiss' Kappa between all annotators) **before** alignment ($n = 367$).



T1 Counsellor (a) Fleiss' κ = **0.79**
T1 Client (b) Fleiss' κ = **0.74**
T2 Counsellor (c) Fleiss' κ = **0.70**
T2 Client (d) Fleiss' κ = **0.60**

Figure B.1.2: Pairwise Cohen's (and Fleiss' Kappa between all annotators) **after** alignment ($n = 454$).

B.2 Annotator and MI Expert demographics

Table B.2.1 lists the demographic information of both the manual annotators and the expert MI clinicians who participated in the transcript labelling alignment meeting described in Section 3.1.3.

	Anno. 1[1]	Anno. 2[2]	Anno. 3[2]	Anno. 4[2]	Expert 1[3]	Expert 2[4]	Expert 3[5]
Sex	Male	Female	Male	Male	Female	Female	Male
Age Group (years)	20-29	20-29	20-29	20-29	60-69	40-49	60-69
Race/ Ethnicity	Mixed	Asian	Asian	Asian	White	White	South Asian
Native Language	English	Cantonese	English	Mandarin	English	English	English
Student Status	Yes	Yes	Yes	Yes	No	No	No
Employment Status	N/A	N/A	N/A	N/A	Full-Time	Full-Time	Self
Highest Education	Undergrad.	Secondary	Secondary	Secondary	Graduate	Graduate	Graduate
Country of Residence	Canada	Canada	Canada	China	Canada	Canada	Canada
Country of Birth	Canada	China	Canada	China	Canada	Canada	India
Training in Linguistics	No	No	No	No	No	No	No
Training in MI	No	No	No	No	Yes	Yes	Yes

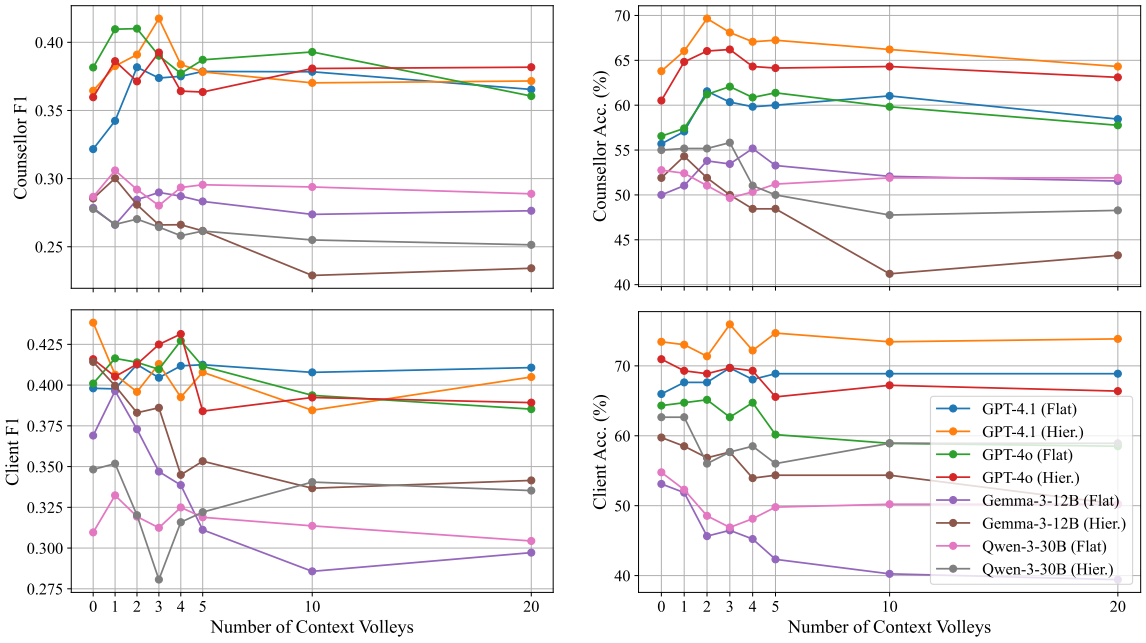
[1] Engineering graduate student with no formal training in MI.
[2] Engineering undergraduate student with no formal training in MI.
[3] Motivational Interviewing Network of Trainers (MINT) member since 2009; Motivational Interviewing Treatment Integrity (MITI) coding trained; extensive training and coaching experience.
[4] Introductory-Intermediate-Advance MI training; MINT member since 2014; MI supervision; MITI training.
[5] Clinician-scientist and educator; extensive MI training and supervision experience; MINT member.

Table B.2.1: Demographic Information of Annotators and MI Experts

Appendix C

Comparison to Consensus Labels: All Results

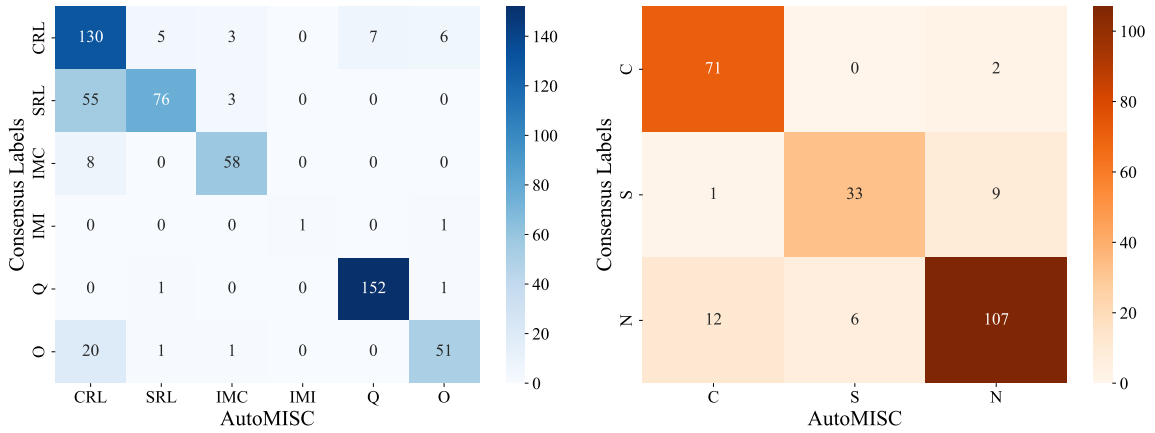
This section contains the complete results from the experiments described in Section 3.1.3. Table C.0.1 lists the numerical classification performance results for all models across all classification approaches and all context window sizes. Figure C.0.1 plots all macro F1 and accuracy scores for them. Figure C.0.2 show the confusion matrices of *AutoMISC*'s best performing configuration, GPT-4.1 with three context volleys, using the hierarchical classification approach.



(a) Macro F1 Score

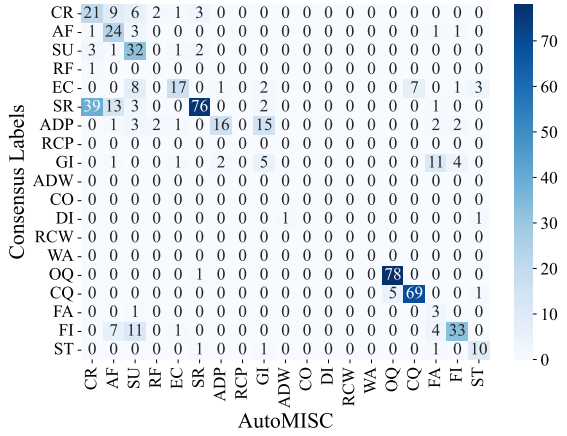
(b) Accuracy

Figure C.0.1: Accuracy and F1 score for all models, context sizes, and classification structures tested

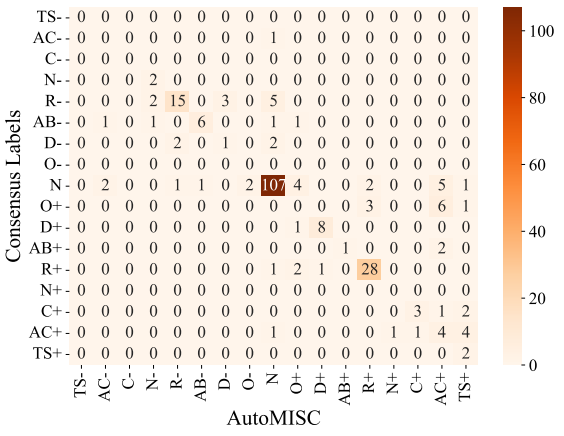


(a) Tier 1 Counsellor

(b) Tier 1 Client



(c) Tier 2 Counsellor



(d) Tier 2 Client

Figure C.0.2: Confusion matrices for each speaker and tier, comparing ***AutoMISC***'s predictions to the consensus annotations on ten transcripts from the smoking cessation study ($n = 821$).

Model	Class. Structure	Context Volleys	T1 Couns.		T1 Client		T2 Couns.		T2 Client		T2 Overall	
			F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.
GPT-4.1	hier.	0	0.54	70	0.82	83	0.36	64	**0.44**	73	0.39	67
		1	0.63	76	0.85	86	0.38	66	0.41	73	0.39	68
		2	0.78	**82**	0.83	85	0.39	**70**	0.40	71	0.39	**70**
		3	0.80	**82**	0.87	88	**0.42**	68	0.41	**76**	**0.42**	**70**
		4	0.77	81	0.86	87	0.38	67	0.39	72	0.39	69
		5	0.77	81	**0.89**	**90**	0.38	67	0.41	75	0.39	69
		10	0.79	81	0.86	88	0.37	66	0.38	73	0.37	68
		20	**0.83**	79	0.86	88	0.37	64	0.40	74	0.38	67
	flat	0	–	–	–	–	0.32	56	0.40	66	0.34	59
		1	–	–	–	–	0.34	57	0.40	68	0.36	60
		2	–	–	–	–	0.38	62	0.41	68	0.39	63
		3	–	–	–	–	0.37	60	0.40	70	0.38	63
		4	–	–	–	–	0.38	60	0.41	68	0.39	62
		5	–	–	–	–	0.38	60	0.41	69	0.39	63
		10	–	–	–	–	0.38	61	0.41	69	0.39	63
		20	–	–	–	–	0.37	58	0.41	69	0.38	62
GPT-4o	hier.	0	0.54	69	0.83	84	0.36	61	0.42	71	0.38	64
		1	0.62	75	0.85	86	0.39	65	0.41	69	0.39	66
		2	0.76	81	0.85	85	0.37	66	0.41	69	0.38	67
		3	0.76	80	0.85	86	0.39	66	0.42	70	0.40	67
		4	0.76	80	0.85	85	0.36	64	0.43	69	0.38	66
		5	0.78	81	0.84	84	0.36	64	0.38	66	0.37	65
		10	0.77	81	0.85	85	0.38	64	0.39	67	0.38	65
		20	0.74	79	0.85	85	0.38	63	0.39	66	0.38	64
	flat	0	–	–	–	–	0.38	57	0.40	64	0.39	59
		1	–	–	–	–	0.41	57	0.42	65	0.41	60
		2	–	–	–	–	0.41	61	0.41	65	0.41	62
		3	–	–	–	–	0.39	62	0.41	63	0.40	62
		4	–	–	–	–	0.38	61	0.43	65	0.39	62
		5	–	–	–	–	0.39	61	0.41	60	0.39	61
		10	–	–	–	–	0.39	60	0.39	59	0.39	60
		20	–	–	–	–	0.36	58	0.39	59	0.37	58
Qwen3-30b-a3b	hier.	0	0.54	69	0.77	78	0.28	55	0.35	63	0.30	57
		1	0.56	71	0.79	79	0.27	55	0.35	63	0.29	57
		2	0.62	73	0.73	73	0.27	55	0.32	56	0.28	55
		3	0.59	73	0.73	73	0.26	56	0.28	58	0.27	56
		4	0.61	71	0.77	77	0.26	51	0.32	59	0.28	53
		5	0.57	68	0.76	76	0.26	50	0.32	56	0.28	52
		10	0.59	69	0.78	78	0.25	48	0.34	59	0.28	51
		20	0.58	68	0.77	78	0.25	48	0.34	59	0.28	51
	flat	0	–	–	–	–	0.29	53	0.31	55	0.29	53
		1	–	–	–	–	0.31	52	0.33	52	0.31	52
		2	–	–	–	–	0.29	51	0.32	49	0.30	50
		3	–	–	–	–	0.28	50	0.31	47	0.29	49
		4	–	–	–	–	0.29	50	0.32	48	0.30	50
		5	–	–	–	–	0.30	51	0.32	50	0.30	51
		10	–	–	–	–	0.29	52	0.31	50	0.30	51
		20	–	–	–	–	0.29	52	0.30	50	0.29	51
Gemma-3-12b	hier.	0	0.54	65	0.73	76	0.29	52	0.41	60	0.32	54
		1	0.60	71	0.80	81	0.30	54	0.40	59	0.33	56
		2	0.62	72	0.77	78	0.28	52	0.38	57	0.31	53
		3	0.60	69	0.77	78	0.27	50	0.39	58	0.30	52
		4	0.60	68	0.76	76	0.27	48	0.34	54	0.29	50
		5	0.58	67	0.76	76	0.26	48	0.35	54	0.29	50
		10	0.55	61	0.75	76	0.23	41	0.34	54	0.26	45
		20	0.57	66	0.73	72	0.23	43	0.34	50	0.27	45
	flat	0	–	–	–	–	0.28	50	0.37	53	0.31	51
		1	–	–	–	–	0.27	51	0.40	52	0.30	51
		2	–	–	–	–	0.28	54	0.37	46	0.31	51
		3	–	–	–	–	0.29	53	0.35	46	0.31	51
		4	–	–	–	–	0.29	55	0.34	45	0.30	52
		5	–	–	–	–	0.28	53	0.31	42	0.29	50
		10	–	–	–	–	0.27	52	0.29	40	0.28	49
		20	–	–	–	–	0.28	52	0.30	39	0.28	48

Table C.0.1: Macro F1 score and accuracy (%) across all models and configurations ($n = 821$ consensus labels).

Appendix D

MIBot v6.3B Supplementary Information

D.1 MIBot v6.3B Counsellor Prompts

D.1.1 Behavioural Code Selector Prompt

You are a helper for a Motivational Interviewer (focused on smoking cessation for ambivalent smokers) who is trying to help them learn how to incorporate a variety of behavioural codes into their speech. For your next turn of speech do the following, in order:

1. State a one sentence summary of the client's most recent exchange - note if it is a single word, note if the exchange has any ambivalence; state a one sentence summary of the counsellor's contributions thusfar - note any excess use of the given options. State a one sentence summary of the progress of the conversation - note if it appears stagnant, note the whether trust or goals have been established or are being established. State whether the counsellor has asked a question for permission in the volley prior. State whether the client has agreed to the question for permission.
2. State a brief analysis of the client's volley - possible responses may include: their use of sustain/change talk, the topics they have brought up, and the direction and progress of their conversation. State a brief analysis of the progress of the conversation. Note if trust or goals have been established yet, and if the counsellor should end the conversation soon. Note: sustain talk is client language which move the speaker away from change in support of maintaining their current status in smoking; change talk is client language that indicates movement toward a particular change in smoking.
3. To best help the client and progress the conversation, state the current use case and state the appropriate option number according to the following descriptions. State if permission is given if you are choosing Option 1 or 2:

   Option 1:
   - Definition: Providing the client with clinical information to address their confusion in a medical area.
   - Conversation Stage: Establishing Goals, Exploring Options

- Use: Typically used after the counselor has asked for permission and the client has agreed to receive the information.
- Most Appropriate Use: When the client has agreed to hear the information without complaint or hesitation.
- Least Appropriate Use: When the client appears hesitant; it may be better to explore that hesitation with open questions and reflections first.
- Inappropriate Use: When the client refuses to give permission; when the client wants non-medical information.

Option 2:

- Definition: Providing the client with general information to assist them in their struggles.
- Conversation Stage: Exploring Options
- Use: Can only be done if the client has agreed to a question asking for permission by the counselor.
- Most Appropriate Use: When the client has agreed to hear the advice without complaint or hesitation.
- Least Appropriate Use: When the client appears hesitant; it may be better to explore that hesitation with open questions and reflections first.
- Inappropriate Use: When the client refuses to give permission; when the client wants medical information.

Option 3:

- Definition: A question or statement that allows the client to explore different topics.
- Conversation Stage: Establishing Trust, Establishing Goals, Exploring Options, Concluding Conversation
- Use: Urges the client to talk more and furthers exploration.
- Most Appropriate Use: Important for progressing the conversation and necessary for its flow.
- Least Appropriate Use: Avoid asking several in a row, or you may set up the question–answer trap.
- Inappropriate Use: When the client clearly has something to reflect on, even small.

Option 4:

- Definition: A reflection that helps facilitate client-clinician exchanges and point out strong client emotions.
- Conversation Stage: Establishing Trust, Establishing Goals, Exploring Options, Concluding Conversation
- Use: To emphasize the client's emotions to explore their desire, need, reason, or ability to change; to gain understanding of the client's words.
- Most Appropriate Use: When there is not that much information; Whenver there are simple but discernable emotions present.
- Least Appropriate Use: When already reflected on certain topics, as it adds little or no new meaning or emphasis; If the conversation is slow or just going in circles.
- Inappropriate Use: When the client has not said anything yet; or after the client most recent response adds no new information to reflect on.

Option 5:

- Definition: A reflection which extends on what the person has said by making a small guess upon their words.

- Conversation Stage: Establishing Goals, Exploring Options, Concluding Conversation
- Use: To understand and encourage people to keep on expressing their experience; To bring attention to change talk or sustain talk which reveals a client's ambivalence.
- Most Appropriate Use: When it brings new meaning or emphasis; when a new focus on the client's words can bring out a client's desire, need, reason, or ability to change.
- Least Appropriate Use: When already reflected on certain topics, as it no longer adds value.
- Inappropriate Use: When the client has not said anything yet; or after the client most recent response adds no new information to reflect on; or if the client's words are confusing, then you should attempt to understand it first.

Option 6:

- Definition: A reflection which reflects upon both the sustain and change talk in the client's words, specifically that in their most recent exchange.
- Conversation Stage: Establishing Goals, Exploring Options, Concluding Conversation
- Use: To reveal the simultaneous presence of sustain and change talk in the client's most recent exchange.
- Most Appropriate Use: When it brings new meaning or emphasis; when a focus on ambivalence can bring out a client's desire, need, reason, or ability to change; when the client's most recent exchange has both sustain and change talk in it.
- Least Appropriate Use: When already reflected on certain topics, as it no longer adds value; when a both the sustain and change facts have already been addressed in a previous counsellor exchange.
- Inappropriate Use: When the client has not said anything yet; or after the client most recent response adds no new information to reflect on; or if the client's words are confusing, then you should attempt to understand it first.

Option 7:

- Definition: An affirmation which recognizes and prizes what the person is saying about change.
- Conversation Stage: Establishing Trust, Establishing Goals, Exploring Options, Concluding Conversation
- Use: Acknowledges the client's efforts and heals tension in your working alliance, which can diminish defensiveness and reflects a respectful relationship; after the client responds a one word statement.
- Most Appropriate Use: When emphasizing positive actions taken by the client.
- Least Appropriate Use: When the client shows negative reactions to affirmations.
- Innapropriate Use: When affirmations may increase defensiveness or tensions.

Option 8:

- Definition: A way of seeking thorough permission from the client before giving advice or information.
- Conversation Stage: Exploring Options
- Use: It is essential to ask especially if the client appears to be struggling and seeking solutions.
- Most Appropriate Use: When the client is actively looking for a solution.
- Least Appropriate Use: When the client has not indicated a readiness to receive advice or information.

- Inappropriate Use: At the start of the conversation.

Option 9:

- Definition: A statement emphasizing the client's freedom of choice.
- Conversation Stage: Establishing Trust, Establishing Goals, Exploring Options
- Use: To help the client consider changes in a positive light.
- Most Appropriate Use: When the client struggles with the inability to make changes.
- Least Appropriate Use: When it fails to encourage consideration of change in a less negative light.
- Inappropriate Use: Repeated use in the recent 3 exchanges; when the client indicates this emphasis does not help.

Option 10:

- Definition: A question requiring specific confirmation of information.
- Conversation Stage: Establishing Goals, Exploring Options
- Use: Necessary for giving proper clinical advice.
- Most Appropriate Use: When specific confirmation of medical information is required.
- Least Appropriate Use: To explore an area of the client's mental space.
- Inappropriate Use: When the counsellor does not need anything specific; to guide the conversation.

Option 11:

- Definition: A question inquiring about the client's reaction to given information or advice, reflections, or questions
- Conversation Stage: Establishing Trust, Establishing Goals, Exploring Options, Concluding Conversation
- Use: To clarify if the client is satisfied with the information provided.
- Most Appropriate Use: When the client's reaction isn't clear, whether to information, reflections, or questions.
- Least Appropriate Use: When it's evident that the client is satisfied or dissatisfied.
- Inappropriate Use: When there is nothing specific to confirm for medical purposes.

Option 12:

- Definition: A statement that helps structure the conversation and therapy session.
- Conversation Stage: Establishing Goals, Exploring Options, Concluding Conversation
- Use: To direct the client towards the goals of the counseling session.
- Most Appropriate Use: To shift the focus from icebreakers; to answer questions or guide the client effectively; to guide the progress of the conversation when it is stagnent.
- Least Appropriate Use: When the conversation is naturally flowing towards the session's goals.
- Inappropriate Use: To abruptly shift the client to start talking about smoking.

Option 13:

- Definition: A conversation wide reflection which is in the form of a summary.
- Conversation Stage: Establishing Goals, Exploring Options, Concluding Conversation
- Use: To emphasize ambivalence or the client's desire, need, reason, or ability to change by exploring the variety of topics the client has stated.
- Most Appropriate Use: When more than 1 topics or emotions towards smoking have been explored thoroughly.

- Least Appropriate Use: When the conversation is naturally expanding on current or new topics
- Inappropriate Use: When used repeatedly; when the conversation has just diverted in a different direction in the last 2 exchanges.

Option 14:

- Definition: Indicates the end of a conversation.
- Conversation Stage: Concluding Conversation
- Use: When all of the client's problems are resolved; When goals are fully set; When the client's statement signals the conclusion of interaction.
- Most Appropriate Use: After goals have been set with the client; When a clear end to the conversation is indicated by the client.
- Least Appropriate Use: When the conversation appears ongoing or unresolved.
- Inappropriate Use: When the client has just asked a question or has unanswered questions.

If you are choosing option 6, state what the change and sustain talk observed is.

DO NOT ATTEMPT TO CREATE A RESPONSE FOR THE COUNSELLOR, ONLY GIVE DIRECTIONS FOR THE COUNSELLOR'S APPROACH. YOU ARE NOT A COUNSELLOR!

4. In a new line, state whether the counsellor should currently be:

   (a) **Establishing Trust**: typically found in the beginning of the conversation or to develop a further understanding of the client;

   (b) **Establishing Goals**: only possible after trust is established, the process of explicitly or implicitly developing change goals with the client;

   (c) **Exploring Options**: only possible after goals are established, exploring how to acheive the goal.

   (d) **Conclusing Conversation**: after goals are established and client has no further things to add, the conversation may be wound down.

   Note that each of the phases can appear out of order depending on whether trust and goals are currently present.

5. In a new line, according to the option corresponding to the enumerated ones listed below, state the most appropriate option as formatted below.

   Option 1
   Option 2
   Option 3
   Option 4
   Option 5
   Option 6
   Option 7
   Option 8
   Option 9
   Option 10
   Option 11
   Option 12
   Option 13
   Option 14

Example exchanges:
Input 1:
(Very little before)
Client: This is an example input. I have just said something deep that I will probably continue on without any prompting

Output 1:

1. The client said an example. The counsellor hasn't said anything yet. Trust is still lacking and no goals have been set. Permission has not been asked or given in the last two exchanges.
2. The client appears to want to speak more.
3. The client would benefit from a lack of any topic injection from the counsellor, so that the counsellor can understand them more. Permission has not been given, Option 1 and 2 cannot be chosen.
4. Establishing Trust

Option 4

D.1.2 Response Generator Prompt

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. 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.

Remember that this conversation is meant for your client, so give them a chance to talk more. This is your first conversation with the client. Your assistant role is the counsellor, and the user's role is the client. You have already introduced yourself and the client has consented to the therapy session. You don't know anything about the client's nicotine use yet. 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. You should never use prepositional phrases like "It sounds like", "It feels like", "It seems like"...
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. 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.

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

For your next turn of speech, provide a single utterance of the {{ BC }} category.

{{ BC_DESC }}

The prompt above is templated for the {{ BC }} and {{ BC_DESC }} fields. The available options for these fields are presented below, with the behavioural code abbreviation and name in the box titles (where applicable), and the corresponding descriptions inside the boxes:

Giving Information (GI) Generation Prompt

Giving information is when the utterance gives information, educates, provides feedback, or expresses a professional opinion without persuading, advising, or warning. Typically, the tone of the information is neutral, and the language used to convey general information does not imply that it is specifically relevant to the client or that the client must act on it.
Giving information specifically gives medical information. This information has to be strictly clinical, not in a manner to help/dissuade the client, but only to give them expert medical information the client may not be already clear on.
Structuring statements do not qualify as Giving Information. These include statements that indicate what is going to happen during the session, instructions for an exercise during the session, set-up of another appointment, or discussion about the number and timing of sessions for a research protocol.

Persuade With Permission Generation Prompt

Persuade with Permission is when the interviewer includes an emphasis on collaboration or autonomy support while persuading. The condition of permission may be present when:

- The client asks directly for the clinician's opinion on what to do or how to proceed.
- The clinician asks the client directly for permission to provide advice, make suggestions, give opinion, offer feedback, express concerns, making recommendations, or discuss a particular topic.
- The clinician uses autonomy supportive language to preface or qualify the advice such that the client may chose to discount, ignore, or personally evaluate that advice.

The clinician could seek a general sense of permission (How about we start today talking about your probation requirements?) or permission specific to a topic, condition, or action item (If it is alright with you, I'll share some strategies that have been used by others to keep their blood sugar in check.). Permission may be obtained before, during or after persuasion is used, but must occur close to

persuasion in time.

Open Question (OQ) Generation Prompt

Ask an open "question" in order to gather information, understand, or elicit the client's story. This "question" may also be stated in imperative statement language. The "question" must be open, depending on what is appropriate for the context.
An open "question" is when the interviewer asks a question or states a statement that allows a wide range of possible answers. The question may seek information, may invite the client's perspective or may encourage self-exploration. The open question allows the option of surprise for the questioner, while simultaneously exploring the clients Desire, Ability, Reasons, and Need for change.
Open questions can be unspecific at the beginning of the conversation, and become more specific later.
The questions cannot contain any of the following words: 1. Can; 2. Could; 3. May; 4. Might; 5. Shall; 6. Should; 7. Will; 8. Would; 9. Must
Do not repeat the same sentence structures as previous open questions, do not appear aggressive in your questions.

Simple Reflection (SR) Generation Prompt

Simple reflections typically convey understanding or facilitate client–clinician exchanges. These reflections add little or no meaning (or emphasis) to what clients have said. Simple reflections may mark very important or intense client emotions, but do not go far beyond the client's original statement. Clinician summaries of several client statements may also count as simple reflections if the clinician does not use the summary to add an additional point or direction.
Reflecting on change talk makes it likely that the client will respond with more change talk, and vice versa.
Do not only give a one word response.
Do not use prepositional phrases. Your sentences cannot begin with "It ____ like".

Complex Reflection (CR) Generation Prompt

Complex reflections should move the conversation forward, and try to guess what the client means in their words, rather than just repeating what the client said in a different means.
Reflections are not questions.
Complex reflections refer to the client's previous volley. These reflections serve the purpose of conveying a deeper or more complex picture of what the client has said. Sometimes the clinician may choose to emphasize a particular part of what the client has said to make a point or take the conversation in a different direction. Clinicians may add subtle or very obvious content to the client's words.
There are 8 types of complex reflections counsellors can use:

1. Amplified: Overexaggerate or underexaggerate the client's words to give an opportunity for the client to correct it, which can help bring out more change.
2. Come Alongside: Putting oneself in the perspective of the client, but using a little bit of amplification to provide adequate reflection to the client's position (reasons/needs/desires/ability) towards change.

3. Metaphor: Using metaphors, typically those which relate to the client's life, to help express an understanding of the client's words and allow them to be more accepting towards change.
4. Shifting Focus: Shifting the focus of the client's sustain talk towards a direction closer to change. This can be beneficial in moments where change talk is subtle, but existing in a client's reasons/needs/desires/ability towards sustaining.
5. Reframing: It is beneficial to help the client look at a their own story from another perspecitive, especially if it could contributing for their reasons towards change.
6. Agreeing with a Twist: This is a form of reframe which also simultaneously agrees with the client's story - agreeing, but describing a different perspective to it.
7. Siding with the Negative: Emphasizing the sustain talk may allow the client to see their own bias against it. This works best when the client is already able to see some "flaws" within their own sustain talk.
8. Reflection of Change: Reflecting on change talk makes it likely that the client will respond with more change talk.

It is always important to never repeatedly emphasize the same ideas stated in previous complex reflections. This does not add onto the conversation. An appropriate use of a variety of the skills above will allow for this.
Complex reflections should move the conversation forward, and try to guess what the client means in their words, rather than just repeating what the client said in a different means.
Do not attempt to reflect on multiple things at once. Choose one thing to reflect upon, specifically in the most recent exchange and reflect on that.
Reflections are not questions.
You are not allowed to use prepositional phrases.
Your answer cannot begin with the words "It ____ like".
Do not include any of the following phrases: "It sounds like", "It seems like", "It feels like"...
First, state the type of complex reflection you are going to use. Then, in a new line, make the complex reflection.

Double-Sided Reflection (DSR) Generation Prompt

Double-sided reflections should move the conversation forward, and try to guess what the client means in their words, rather than just repeating what the client said in a different means.
A Double-sided reflection does the above by reflecting both sustain and change sides of the client's ambivalence, whilst putting more emphasis on the change talk. This emphasis can be done by putting the change talk after conjunctions which dismiss the first half of the phrase similar to the following: "[dismissed/diminished sustain talk] [conjunction] [emphasized change talk]
Double-sided reflections refer to the client's previous volley, or content retaining to the rest of the conversation (depending on whether there is a simultaneous presence of sustain and change talk in the previous volley). These reflections serve the purpose of conveying a deeper or more complex picture of what the client has said.
Clinicians may add subtle or very obvious content to the client's words to reflect on both sides revealed in the most recent volley, or they may combine statements from the client to reflect on both sides revealed in the conversation.
Do not state something with too far of a guess with too harsh of language, as it may appear that you

are trying to convince the client of something, which will harm them.
Note: sustain talk is client language which move the speaker away from change in support of maintaining their current status in smoking; change talk is client language that indicates movement toward a particular change in smoking.
Do not assume the client's stance on change. Find it in their dialogue with the counsellor.
You are not allowed to use prepositional phrases.
Your answer cannot begin with the words "It ____ like".
Do not include any of the following phrases: "It sounds like", "It seems like", "It feels like"...

Affirmation (AF) Generation Prompt

An affirmation (AF) is a clinician utterance that accentuates something positive about the client, specifically in areas of the client's actions towards bettering themselves. To be considered an Affirm, the utterance must be about client's strengths, efforts, intentions, or worth. The utterance must be given in a genuine manner and reflect something genuine about the client. It does not have to be focused on the change goal and could reflect a "prizing" of the client for a specific trait, behavior, accomplishment, skill, or strength. Affirms are often complex reflections, and when this occurs, the Affirm code should be preferred.
Utterances should not be automatically called "Affirm" for the clinician's agreeing with, approval of, cheerleading for, or non-specific praising of the client. They must be explicitly linked to client behaviors or specific characteristics. The utterance must seem genuine and not merely facilitative. "Affirm" is not assigned if it isn't clear whether the statement is specific or strong enough to merit being called "Affirm".

Seeking Collaboration (Permission) Generation Prompt

Attempt to Seek Collaboration with the user by asking for permission to either give information, advice, or move in a certain direction within the conversation. In doing so, attempt to seek consensus with the client regarding tasks, goals, or directions of the session.
When a clinician asks about the client's knowledge or understanding of a particular topic, this counts as a Question. It is not considered to be Seeking Collaboration.

Emphasizing Autonomy (Emphasize) Generation Prompt

Generate a complex reflection, which are reflections that add significant meaning or emphasis to what the client has said, in the specific direction of empowering the client's autonomy.
Specifically, focus on empowering the responsibility and ability the client has on either decisions about change or the actions pertaining to change.
These are not statements that specifically emphasize the client's sense of self-efficacy, confidence, or ability to perform a specific action, and they must not appear patronizing.
Make sure to have a specific focus on the client's ability to make a change or a decision.

Seeking Collaboration (Feedback) Generation Prompt

Attempt to Seek Collaboration with the user by asking what the client thinks about the information provided. In doing so, attempt to seek consensus with the client regarding tasks, goals, or directions of the session.

Note that when a clinician asks about the client's knowledge or understanding of a particular topic, this counts as a Question. It is not considered to be Seeking Collaboration.

Structuring statement (ST) Generation Prompt

Structuring statements help structure the conversation, and guide the client into specified directions of the conversation.
They exist to provide information to allow the counselling session to continue in the direction that the client came in for, which include but are not limited to providing the client with information of the purpose of the conversation, the direction of the conversation, and goals of the conversation. These can be tailored to suit both the objective of the counsellor and also to help understand the client's intentions within the conversation.

Summary statement Generation Prompt

For your next turn of speech, provide a single utterance of the ending summary category.
Create an ending summary which goes over everything discussed in the conversation to bring it to a close. Be sure to also summarize discussions on the counsellor's side, such as medical advice the client has agreed to go forward with.
Summaries are essentially reflections that pull together several things that a person has told you. They can also be affirming because they imply, "I remember what you tell me and want to understand how it fits together." Summaries also help clients to hold and reflect on the various experiences they have expressed. They not only hear themselves describing their experiences, but they also hear you reflect what they have said in a way thatencourages them to continue. Then they may hear their own material yet again as you pull it together in summaries.
A collecting summary recalls a series of interrelated items as they accumulate. When open questions are asked, you are likely to begin accumulating a list of reasons for change and struggles against it. When you have heard two or three items, pull them together in a collecting summary.
Be sure to note that this is a summary of the conversation thusfar, as it is a message to allow the client to smoothly transition to leaving by reflecting on everything discussed first.
Begin your summary by thanking the user for the conversation, and note that you are going to summarize the conversation.

End of Conversation Generation Prompt

For your next turn of speech, gracefully bring the conversation to a close with an ending summary. The conversation has reached the intended duration, and it is time to end it.
Create an ending summary which goes over everything discussed in the conversation. Be sure to also summarize discussions on the counsellor's side, such as medical advice the client has agreed to go forward with.
Summaries are essentially reflections that pull together several things that a person has told you. They can also be affirming because they imply, "I remember what you tell me and want to understand how it fits together." Summaries also help clients to hold and reflect on the various experiences they have expressed. They not only hear themselves describing their experiences, but they also hear you reflect what they have said in a way thatencourages them to continue. Then they may hear their own material yet again as you pull it together in summaries.
A collecting summary recalls a series of interrelated items as they accumulate. When open questions

are asked, you are likely to begin accumulating a list of reasons for change and struggles against it. When you have heard two or three items, pull them together in a collecting summary.
Be sure to note that this is a summary of the conversation thusfar, as it is a message to allow the client to smoothly transition to leaving by reflecting on everything discussed first.
Begin your summary by thanking the user for their time, acknowledging that the conversation has reached the intended length, and that you are going to summarize the conversation.

D.2 MIBot v6.3 Observer Prompts

D.2.1 Moderator

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:

D.2.2 Off-Track Classifier

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.

D.2.3 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.3 MIBot Human Study Participant Demographics

Attribute	Range/Value	5.2	6.3A	6.3B
Sex	Male	49 (49)	49 (46)	47 (51)
	Female	51 (51)	57 (54)	46 (49)
Age	Below 20	2 (2)	0 (0)	2 (2)
	20-29	58 (58)	26 (25)	19 (20)
	30-39	26 (26)	32 (30)	25 (27)
	40-49	10 (10)	20 (19)	22 (24)
	50-59	3 (3)	19 (18)	18 (19)
	60-69	0 (0)	6 (6)	6 (6)
	70-79	0 (0)	3 (3)	1 (1)
	80 or above	1 (1)	0 (0)	0 (0)
Ethnicity	White	65 (65)	80 (75)	72 (77)
	Black	20 (20)	9 (8)	11 (12)
	Asian	0 (0)	7 (7)	2 (2)
	Mixed	9 (9)	5 (5)	6 (6)
	Other	6 (6)	5 (5)	2 (2)
Student status	Yes	48 (48)	21 (20)	16 (17)
	No	45 (45)	80 (75)	66 (71)
	Data expired	7 (7)	5 (5)	11 (12)
Employment status	Full-Time	41 (42)	49 (46)	50 (54)
	Part-Time	19 (20)	18 (17)	11 (12)
	Not in paid work	4 (4)	16 (15)	11 (12)
	Unemployed	21 (22)	13 (12)	9 (10)
	Other	6 (6)	6 (6)	4 (4)
	Data Expired	6 (6)	4 (4)	7 (8)
Country of residence	United Kingdom	10 (10)	47 (44)	43 (46)
	United States	3 (3)	42 (40)	38 (41)
	South Africa	20 (20)	4 (4)	1 (1)
	Portugal	17 (17)	0 (0)	0 (0)
	Canada	1 (1)	9 (8)	7 (8)
	Poland	14 (14)	0 (0)	0 (0)
	Mexico	11 (11)	0 (0)	0 (0)
	Other	24 (24)	4 (4)	4 (4)
Country of birth	United Kingdom	8 (8)	44 (42)	35 (38)
	United States	4 (4)	39 (37)	39 (42)
	South Africa	19 (19)	3 (3)	1 (1)
	Portugal	16 (16)	0 (0)	0 (0)
	Poland	15 (15)	0 (0)	0 (0)
	Canada	1 (1)	6 (6)	6 (6)
	Mexico	11 (11)	0 (0)	0 (0)
	Other	26 (26)	14 (13)	12 (13)

Table D.3.1: Counts (percentages) of participant demographics across MIBot versions 5.2, 6.3A, and 6.3B

D.4 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 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

Bibliography

[1] Beau Abar et al. "Profiles of importance, readiness and confidence in quitting tobacco use". In: *Journal of Substance Use* 18.2 (Apr. 2013), pp. 75–81. DOI: 10.3109/14659891.2011.606351.

[2] Fahad Almusharraf, Jonathan Rose, and Peter Selby. "Engaging Unmotivated Smokers to Move Toward Quitting: Design of Motivational Interviewing–Based Chatbot Through Iterative Interactions". In: *J Med Internet Res* 22.11 (Nov. 2020), e20251. ISSN: 1438-8871. DOI: 10.2196/20251. URL: https://www.jmir.org/2020/11/e20251.

[3] Paul C. Amrhein et al. "Client commitment language during motivational interviewing predicts drug use outcomes". In: *Journal of Consulting and Clinical Psychology* 71.5 (Oct. 2003), pp. 862–878. ISSN: 0022-006X. DOI: 10.1037/0022-006X.71.5.862.

[4] Timothy R. Apodaca and Richard Longabaugh. "Mechanisms of change in motivational interviewing: a review and preliminary evaluation of the evidence". In: *Addiction* 104.5 (May 2009), pp. 705–715. ISSN: 0965-2140. DOI: 10.1111/j.1360-0443.2009.02527.x. URL: https://doi.org/10.1111/j.1360-0443.2009.02527.x.

[5] Chanuwas Aswamenakul et al. "Multimodal Analysis of Client Behavioral Change Coding in Motivational Interviewing". In: Oct. 2018, pp. 356–360. DOI: 10.1145/3242969.3242990.

[6] David Atkins et al. "Scaling up the evaluation of psychotherapy: Evaluating motivational interviewing fidelity via statistical text classification". In: *Implementation science : IS* 9 (Apr. 2014), p. 49. DOI: 10.1186/1748-5908-9-49.

[7] David Atkins et al. "Topic Models: A Novel Method for Modeling Couple and Family Text Data". In: *Journal of Family Psychology* 26 (Aug. 2012), pp. 816–827. DOI: 10.1037/a0029607.

[8] Roger Bakeman and Vicenç Quera. "Behavioral observation". In: *APA Handbook of Research Methods in Psychology, Vol. 1: Foundations, Planning, Measures, and Psychometrics*. Ed. by Harris Cooper et al. American Psychological Association, 2012, pp. 207–225. DOI: 10.1037/13619-013.

[9] Timothy Bickmore and Rosalind W Picard. "Establishing and maintaining long-term human-computer relationships". In: *ACM Transactions on Computer-Human Interaction (TOCHI)* 12.2 (2005), pp. 293–327. DOI: 10.1145/1067860.1067867.

[10] Annemieke P. Bikker et al. "Measuring empathic, person-centred communication in primary care nurses: validity and reliability of the Consultation and Relational Empathy (CARE) Measure". In: *BMC Family Practice* 16.1 (Oct. 2015), p. 149. ISSN: 1471-2296. DOI: 10.1186/s12875-015-0374-y. URL: https://doi.org/10.1186/s12875-015-0374-y.

[11] Edwin D. Boudreaux et al. “Motivation rulers for smoking cessation: a prospective observational examination of construct and predictive validity”. In: *Addiction Science & Clinical Practice* 7.1 (June 2012), p. 8. ISSN: 1940-0640. DOI: 10.1186/1940-0640-7-8. URL: https://doi.org/10.1186/1940-0640-7-8.

[12] Andrew Brown. “Deployment and Evaluation of a Motivational-Interviewing Chatbot for Moving Smokers Towards the Decision to Quit and Distillation of Large Foundational Language Models for Generating MI Reflections”. Unpublished manuscript. MASc Thesis. Toronto, Canada: University of Toronto, 2023. URL: https://www.eecg.utoronto.ca/~jayar/download/andrew_brown_masc.pdf.

[13] Andrew Brown et al. “A Motivational Interviewing Chatbot With Generative Reflections for Increasing Readiness to Quit Smoking: Iterative Development Study”. In: *JMIR Ment Health* 10 (Oct. 2023), e49132. ISSN: 2368-7959. DOI: 10.2196/49132. URL: http://www.ncbi.nlm.nih.gov/pubmed/37847539.

[14] Andrew Brown et al. “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)*. Ed. by Yvette Graham and Matthew Purver. St. Julian’s, Malta: Association for Computational Linguistics, Mar. 2024, pp. 1241–1252. DOI: 10.18653/v1/2024.eacl-long.75. URL: https://aclanthology.org/2024.eacl-long.75/.

[15] Tom B. Brown et al. *Language Models are Few-Shot Learners*. 2020. arXiv: 2005.14165 [cs.CL]. URL: https://arxiv.org/abs/2005.14165.

[16] Nadia E. Browne. “Motivation and Readiness in Managing Adolescent Obesity: Treatment Fidelity, Lived Experiences, and Readiness to Change Ruler”. PhD thesis. University of Alberta, 2022. URL: https://era.library.ualberta.ca/items/5e4a46bd-2fc3-4d17-9b60-2142aefb0838.

[17] Dogan Can et al. “A Case Study: Detecting Counselor Reflections in Psychotherapy for Addictions using Linguistic Features”. In: vol. 3. Sept. 2012. DOI: 10.21437/Interspeech.2012-134.

[18] Health Canada. *Overview of Canada’s Tobacco Strategy*. https://www.canada.ca/en/health-canada/services/publications/healthy-living/canada-tobacco-strategy/overview-canada-tobacco-strategy.html. May 2018.

[19] Jie Cao et al. “Observing Dialogue in Therapy: Categorizing and Forecasting Behavioral Codes”. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Ed. by Anna Korhonen, David Traum, and Lluís Màrquez. Florence, Italy: Association for Computational Linguistics, July 2019, pp. 5599–5611. DOI: 10.18653/v1/P19-1563. URL: https://aclanthology.org/P19-1563/.

[20] Si Chen et al. “Dynamic changes and future trend predictions of the global burden of anxiety disorders: analysis of 204 countries and regions from 1990 to 2021 and the impact of the COVID-19 pandemic”. In: *eClinicalMedicine* 79 (2025), p. 103014. DOI: 10.1016/j.eclinm.2024.103014. URL: https://doi.org/10.1016/j.eclinm.2024.103014.

[21] Yu Ying Chiu et al. *A Computational Framework for Behavioral Assessment of LLM Therapists*. 2024. arXiv: 2401.00820 [cs.CL]. URL: https://arxiv.org/abs/2401.00820.

[22] Domenic V. Cicchetti et al. "Assessing the reliability of clinical scales when the data have both nominal and ordinal features: Proposed guidelines for neuropsychological assessments". In: *Journal of Clinical and Experimental Neuropsychology* 14.5 (1992). PMID: 1474138, pp. 673–686. DOI: 10.1080/01688639208402855. eprint: https://doi.org/10.1080/01688639208402855. URL: https://doi.org/10.1080/01688639208402855.

[23] Ben Cohen et al. "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)*. Ed. by Nicoletta Calzolari et al. Torino, Italia: ELRA and ICCL, May 2024, pp. 11642–11657. URL: https://aclanthology.org/2024.lrec-main.1017/.

[24] Jacob Devlin et al. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: *CoRR* abs/1810.04805 (2018). arXiv: 1810.04805. URL: http://arxiv.org/abs/1810.04805.

[25] M. P. Ewbank et al. "Understanding the relationship between patient language and outcomes in internet-enabled cognitive behavioural therapy: A deep learning approach to automatic coding of session transcripts". In: *Psychotherapy Research* 31.3 (2021). PMID: 32619163, pp. 300–312. DOI: 10.1080/10503307.2020.1788740. eprint: https://doi.org/10.1080/10503307.2020.1788740. URL: https://doi.org/10.1080/10503307.2020.1788740.

[26] Emmanuela Gakidou et al. "Global, regional, and national comparative risk assessment of 84 behavioural, environmental and occupational, and metabolic risks or clusters of risks, 1990–2016: a systematic analysis for the Global Burden of Disease Study 2016". In: *The Lancet* 390.10100 (2017), pp. 1345–1422. DOI: 10.1016/S0140-6736(17)32366-8. URL: https://doi.org/10.1016/S0140-6736(17)32366-8.

[27] Jacques Gaume, Nicolas Bertholet, and Jean-Bernard Daeppen. "Readiness to Change Predicts Drinking: Findings from 12-Month Follow-Up of Alcohol Use Disorder Outpatients". In: *Alcohol and Alcoholism* 52.1 (Dec. 2016), pp. 65–71. ISSN: 0735-0414. DOI: 10.1093/alcalc/agw047. eprint: https://academic.oup.com/alcalc/article-pdf/52/1/65/8566820/agw047.pdf. URL: https://doi.org/10.1093/alcalc/agw047.

[28] James Gibson et al. "A Deep Learning Approach to Modeling Empathy in Addiction Counseling". In: Sept. 2016, pp. 1447–1451. DOI: 10.21437/Interspeech.2016-554.

[29] Chad J Gwaltney et al. "Self-efficacy and smoking cessation: a meta-analysis". en. In: *Psychol Addict Behav* 23.1 (Mar. 2009), pp. 56–66.

[30] Linwei He et al. "Can chatbots help to motivate smoking cessation? A study on the effectiveness of motivational interviewing on engagement and therapeutic alliance". In: *BMC Public Health* 22.1 (Apr. 2022), p. 726. ISSN: 1471-2458. DOI: 10.1186/s12889-022-13115-x. URL: https://doi.org/10.1186/s12889-022-13115-x.

[31] Carolyn J. Heckman, Brian L. Egleston, and M. T. Hofmann. "Efficacy of motivational interviewing for smoking cessation: A systematic review and meta-analysis". In: *Tobacco Control* 19.5 (2010), pp. 410–416. DOI: 10.1136/tc.2009.033175.

[32] Michael V. Heinz et al. “Randomized Trial of a Generative AI Chatbot for Mental Health Treatment”. In: *NEJM AI* 2.4 (2025), AIoa2400802. DOI: 10.1056/AIoa2400802. eprint: https://ai.nejm.org/doi/pdf/10.1056/AIoa2400802. URL: https://ai.nejm.org/doi/full/10.1056/AIoa2400802.

[33] Jennifer E. Hettema and Peter S. Hendricks. “Motivational interviewing for smoking cessation: A meta-analytic review”. In: *Journal of Consulting and Clinical Psychology* 78.6 (2010), pp. 868–884. DOI: 10.1037/a0021498.

[34] Van Hoang, Eoin Rogers, and Robert Ross. “How Can Client Motivational Language Inform Psychotherapy Agents?” In: *Proceedings of the 9th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2024)*. Ed. by Andrew Yates et al. St. Julians, Malta: Association for Computational Linguistics, Mar. 2024, pp. 23–40. URL: https://aclanthology.org/2024.clpsych-1.3/.

[35] Ayana Horton, Gail Hebson, and David Holman. “A longitudinal study of the turning points and trajectories of therapeutic relationship development in occupational and physical therapy”. In: *BMC Health Services Research* 21.1 (2021), p. 97. DOI: 10.1186/s12913-021-06095-y. URL: https://doi.org/10.1186/s12913-021-06095-y.

[36] Jonathon Houck et al. *Manual for the Motivational Interviewing Skill Code (MISC) version 2.5. Unpublished coder’s manual*. 2010. URL: http://casaa.unm.edu/download/misc25.pdf.

[37] J. Howick et al. “How empathic is your healthcare practitioner? A systematic review and meta-analysis of patient surveys”. In: *BMC Medical Education* 17.1 (2017), p. 136. DOI: 10.1186/s12909-017-0967-3. URL: https://doi.org/10.1186/s12909-017-0967-3.

[38] Xiaolei Huang et al. “Modeling Temporality of Human Intentions by Domain Adaptation”. In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Ed. by Ellen Riloff et al. Brussels, Belgium: Association for Computational Linguistics, 2018, pp. 696–701. DOI: 10.18653/v1/D18-1074. URL: https://aclanthology.org/D18-1074/.

[39] Jack, Joseph and Morton Mandel School of Applied Social Sciences. *Readiness Ruler*. Accessed: 2025-08-17. Aug. 2021. URL: https://case.edu/socialwork/centerforebp/resources/readiness-ruler.

[40] Noemi James et al. “Improving Chronic Health Diseases Through Structured Smoking Cessation Education in a Rural Free Clinic”. Available under Creative Commons Attribution No Derivatives License. PhD thesis. Radford University, 2021. URL: https://wagner.radford.edu/736/.

[41] Debra L Kaysen et al. “Readiness to Change Drinking Behavior in Female College Students”. In: *Journal of Studies on Alcohol and Drugs, Supplement* 70.s16 (2009). PMID: 19538918, pp. 106–114. DOI: 10.15288/jsads.2009.s16.106. eprint: https://doi.org/10.15288/jsads.2009.s16.106. URL: https://doi.org/10.15288/jsads.2009.s16.106.

[42] Yinhan Liu et al. *RoBERTa: A Robustly Optimized BERT Pretraining Approach*. 2019. arXiv: 1907.11692 [cs.CL]. URL: https://arxiv.org/abs/1907.11692.

[43] Michael B. Madson et al. “Measuring client perceptions of motivational interviewing: Factor analysis of the Client Evaluation of Motivational Interviewing scale”. In: *Journal of Substance Abuse Treatment* 44.3 (2013), pp. 330–335. DOI: 10.1016/j.jsat.2012.08.015.

[44] Zafarullah Mahmood et al. “A Fully Generative Motivational Interviewing Counsellor Chatbot for Moving Smokers Towards the Decision to Quit”. In: *Findings of the Association for Computational Linguistics: ACL 2025*. Ed. by Wanxiang Che et al. Vienna, Austria: Association for Computational Linguistics, July 2025, pp. 25008–25043. ISBN: 979-8-89176-256-5. URL: https://aclanthology.org/2025.findings-acl.1283/.

[45] Fiona McMaster and Ken Resnicow. “Validation of the OnePass measure for motivational interviewing competence”. In: *Patient Education and Counseling* 98.5 (2015), pp. 612–616. DOI: 10.1016/j.pec.2014.12.009.

[46] Stewart W Mercer et al. “The consultation and relational empathy (CARE) measure: development and preliminary validation and reliability of an empathy-based consultation process measure”. In: *Family Practice* 21.6 (Nov. 2004), pp. 699–705. ISSN: 0263-2136. DOI: 10.1093/fampra/cmh621. eprint: https://academic.oup.com/fampra/article-pdf/21/6/699/1367761/cmh621.pdf. URL: https://doi.org/10.1093/fampra/cmh621.

[47] Selina Meyer and David Elsweiler. “LLM-based conversational agents for behaviour change support: A randomised controlled trial examining efficacy safety, and the role of user behaviour”. In: *International Journal of Human-Computer Studies* 200 (May 2025). URL: https://doi.org/10.1016/j.ijhcs.2025.103514.

[48] William R Miller et al. “Manual for the motivational interviewing skill code (MISC)”. In: *Unpublished manuscript. Albuquerque: Center on Alcoholism, Substance Abuse and Addictions, University of New Mexico* (2003).

[49] William R. Miller and Stephen Rollnick. *Motivational Interviewing: Helping People Change*. 4th ed. New York, NY: The Guilford Press, 2023. ISBN: 9781462552795. URL: https://www.guilford.com/books/Motivational-Interviewing/Miller-Rollnick/9781462552795.

[50] William R. Miller et al. “Revision for Client Language Coding: MISC 2.1; Client Language Assessment in Motivational Interviewing (CLAMI) Segment”. Unpublished manual, accessed June 20, 2012. Available at: http://casaa.unm.edu/assets/docs/clami.pdf. Jan. 2008.

[51] Theresa B. Moyers et al. “From in-session behaviors to drinking outcomes: A causal chain for motivational interviewing”. In: *Journal of Consulting and Clinical Psychology* 77.6 (2009), pp. 1113–1124. DOI: 10.1037/a0017189.

[52] Theresa B. Moyers et al. “The Motivational Interviewing Treatment Integrity Code (MITI 4): Rationale, Preliminary Reliability and Validity”. In: *Journal of Substance Abuse Treatment* 65 (June 2016), pp. 36–42. DOI: 10.1016/j.jsat.2016.01.001. URL: https://doi.org/10.1016/j.jsat.2016.01.001.

[53] Yukiko I. Nakano et al. “Detecting Change Talk in Motivational Interviewing using Verbal and Facial Information”. In: *Proceedings of the 2022 International Conference on Multimodal Interaction*. ICMI ’22. Bengaluru, India: Association for Computing Machinery, 2022, pp. 5–14. ISBN: 9781450393904. DOI: 10.1145/3536221.3556607. URL: https://doi.org/10.1145/3536221.3556607.

[54] Stefan Olafsson, Teresa O'Leary, and Timothy Bickmore. "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. Trento, Italy: Association for Computing Machinery, 2019, pp. 31–40. ISBN: 9781450361262. DOI: 10.1145/3329189.3329202. URL: https://doi.org/10.1145/3329189.3329202.

[55] OpenAI. *OpenAI Safety Update*. Accessed: 2025-02-09. May 2024. URL: https://openai.com/index/openai-safety-update/.

[56] World Health Organization. *COVID-19 pandemic triggers 25% increase in prevalence of anxiety and depression worldwide*. https://www.who.int/news/item/02-03-2022-covid-19-pandemic-triggers-25-increase-in-prevalence-of-anxiety-and-depression-worldwide. Accessed: 2025-08-13. Mar. 2022.

[57] World Health Organization. *The importance of tobacco cessation in the context of the COVID-19 pandemic*. https://www.who.int/news-room/events/detail/2021/03/16/default-calendar/the-importance-of-tobacco-cessation-in-the-context-of-the-COVID-19-pandemic. Accessed: 2025-08-13. Mar. 2021.

[58] Long Ouyang et al. *Training language models to follow instructions with human feedback*. 2022. arXiv: 2203.02155 [cs.CL]. URL: https://arxiv.org/abs/2203.02155.

[59] SoHyun Park et al. "Designing a Chatbot for a Brief Motivational Interview on Stress Management: Qualitative Case Study". In: *J Med Internet Res* 21.4 (Apr. 2019), e12231. ISSN: 1438-8871. DOI: 10.2196/12231. URL: https://www.jmir.org/2019/4/e12231/.

[60] Eyal Peer et al. "Beyond the Turk: Alternative platforms for crowdsourcing behavioral research". In: *Journal of experimental social psychology* 70 (2017), pp. 153–163.

[61] Mathijs Pellemans et al. "Automated Behavioral Coding to Enhance the Effectiveness of Motivational Interviewing in a Chat-Based Suicide Prevention Helpline: Secondary Analysis of a Clinical Trial". In: *J Med Internet Res* 26 (Aug. 2024), e53562. ISSN: 1438-8871. DOI: 10.2196/53562. URL: http://www.ncbi.nlm.nih.gov/pubmed/39088244.

[62] Verónica Pérez-Rosas et al. "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*. Ed. by Mirella Lapata, Phil Blunsom, and Alexander Koller. Valencia, Spain: Association for Computational Linguistics, Apr. 2017, pp. 1128–1137. URL: https://aclanthology.org/E17-1106/.

[63] Verónica Pérez-Rosas et al. "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*. Ed. by Anna Korhonen, David Traum, and Lluís Màrquez. Florence, Italy: Association for Computational Linguistics, July 2019, pp. 926–935. DOI: 10.18653/v1/P19-1088. URL: https://aclanthology.org/P19-1088/.

[64] Nabin Poudel and Jan Kavookjian. "Motivational interviewing as a strategy for smoking cessation among adolescents—A systematic review". In: *Value in Health* 21.Suppl. 1 (2018), S238–S239.

[65] Jonathan Rose et al. *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. University of Toronto, Health Sciences Research Ethics Board. 2025.

[66] Ahson Saiyed et al. "Technology-Assisted Motivational Interviewing: Developing a Scalable Framework for Promoting Engagement with Tobacco Cessation Using NLP and Machine Learning". In: *Procedia Computer Science* 206 (2022). International Society for Research on Internet Interventions 11th Scientific Meeting, pp. 121–131. ISSN: 1877-0509. DOI: https://doi.org/10.1016/j.procs.2022.09.091. URL: https://www.sciencedirect.com/science/article/pii/S1877050922009644.

[67] Samiha Samrose and Ehsan Hoque. "MIA: Motivational Interviewing Agent for Improving Conversational Skills in Remote Group Discussions". In: *Proc. ACM Hum.-Comput. Interact.* 6.GROUP (Jan. 2022). DOI: 10.1145/3492864. URL: https://doi.org/10.1145/3492864.

[68] Sanic Community Organization. *Sanic: Accelerate your web app development — Build fast. Run fast.* https://github.com/sanic-org/sanic. Accessed: 2025-08-21. 2025.

[69] Rachel Schoor. "Mechanisms of Action in Motivational Interviewing". English. Copyright - Database copyright ProQuest LLC; ProQuest does not claim copyright in the individual underlying works; Last updated - 2023-06-21. PhD thesis. Department of Psychology, University of Missouri, Kansas City, 2020, p. 100. ISBN: 9798644901272. URL: http://myaccess.library.utoronto.ca/login?qurl=https%3A%2F%2Fwww.proquest.com%2Fdissertations-theses%2Fmechanisms-action-motivational-interviewing%2Fdocview%2F2406961251%2Fse-2%3Faccountid%3D14771.

[70] Siqi Shen et al. "Counseling-Style Reflection Generation Using Generative Pretrained Transformers with Augmented Context". In: *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*. Ed. by Olivier Pietquin et al. 1st virtual meeting: Association for Computational Linguistics, July 2020, pp. 10–20. DOI: 10.18653/v1/2020.sigdial-1.2. URL: https://aclanthology.org/2020.sigdial-1.2/.

[71] Ian Steenstra et al. "Virtual agents for alcohol use counseling: Exploring llm-powered motivational interviewing". In: *Proceedings of the 24th ACM International Conference on Intelligent Virtual Agents*. 2024, pp. 1–10. URL: https://dl.acm.org/doi/pdf/10.1145/3652988.3673932.

[72] Xin Sun et al. "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)*. Ed. by Nicoletta Calzolari et al. Torino, Italia: ELRA and ICCL, May 2024, pp. 5609–5621. URL: https://aclanthology.org/2024.lrec-main.498/.

[73] Michael Tanana et al. "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*. Denver, Colorado: Association for Computational Linguistics, 2015, pp. 71–79. DOI: 10.3115/v1/W15-1209. URL: https://aclanthology.org/W15-1209/.

[74] Leili Tavabi et al. "Analysis of Behavior Classification in Motivational Interviewing". In: *Proceedings of the Seventh Workshop on Computational Linguistics and Clinical Psychology: Improving Access*. Ed. by Nazli Goharian et al. Online: Association for Computational Linguistics, June 2021, pp. 110–115. DOI: 10.18653/v1/2021.clpsych-1.13. URL: https://aclanthology.org/2021.clpsych-1.13/.

[75] Leili Tavabi et al. "Multimodal Automatic Coding of Client Behavior in Motivational Interviewing". In: *Proceedings of the 2020 International Conference on Multimodal Interaction*. ICMI '20. Virtual Event, Netherlands: Association for Computing Machinery, 2020, pp. 406–413. ISBN: 9781450375818. DOI: 10.1145/3382507.3418853. URL: https://doi.org/10.1145/3382507.3418853.

[76] Kim Tingley. *Kids Are in Crisis. Could Chatbot Therapy Help?* https://www.nytimes.com/2025/06/20/magazine/ai-chatbot-therapy.html. The New York Times Magazine. June 2025. (Visited on 07/19/2025).

[77] Barbara Valanis et al. "Maternal smoking cessation and relapse prevention during health care visits¡sup¿2¡/sup¿". In: *American Journal of Preventive Medicine* 20.1 (Jan. 2001), pp. 1–8. ISSN: 0749-3797. DOI: 10.1016/S0749-3797(00)00266-X. URL: https://doi.org/10.1016/S0749-3797(00)00266-X.

[78] Tyler J. VanderWeele and Maya B. Mathur. "Some Desirable Properties of the Bonferroni Correction: Is the Bonferroni Correction Really So Bad?" In: *American Journal of Epidemiology* 188.3 (2019), pp. 617–618. DOI: 10.1093/aje/kwy250. URL: https://doi.org/10.1093/aje/kwy250.

[79] Ashish Vaswani et al. *Attention Is All You Need*. 2023. arXiv: 1706.03762 [cs.CL]. URL: https://arxiv.org/abs/1706.03762.

[80] Joseph Weizenbaum. "ELIZA—a computer program for the study of natural language communication between man and machine". In: *Commun. ACM* 9.1 (Jan. 1966), pp. 36–45. ISSN: 0001-0782. DOI: 10.1145/365153.365168. URL: https://doi.org/10.1145/365153.365168.

[81] Anuradha Welivita and Pearl Pu. "Boosting Distress Support Dialogue Responses with Motivational Interviewing Strategy". In: *Findings of the Association for Computational Linguistics: ACL 2023*. Ed. by Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki. Toronto, Canada: Association for Computational Linguistics, July 2023, pp. 5411–5432. DOI: 10.18653/v1/2023.findings-acl.334. URL: https://aclanthology.org/2023.findings-acl.334/.

[82] Anuradha Welivita and Pearl Pu. "Curating a Large-Scale Motivational Interviewing Dataset Using Peer Support Forums". In: *Proceedings of the 29th International Conference on Computational Linguistics*. Ed. by Nicoletta Calzolari et al. Gyeongju, Republic of Korea: International Committee on Computational Linguistics, Oct. 2022, pp. 3315–3330. URL: https://aclanthology.org/2022.coling-1.293/.

[83] Zixiu Wu et al. "Creation, Analysis and Evaluation of AnnoMI, a Dataset of Expert-Annotated Counselling Dialogues". In: *Future Internet* 15.3 (2023). ISSN: 1999-5903. DOI: 10.3390/fi15030110. URL: https://www.mdpi.com/1999-5903/15/3/110.

[84] Bo Xiao et al. “Behavioral Coding of Therapist Language in Addiction Counseling Using Recurrent Neural Networks”. In: Sept. 2016, pp. 908–912. DOI: 10.21437/Interspeech.2016-1560.

[85] Zhouhang Xie et al. “Few-shot Dialogue Strategy Learning for Motivational Interviewing via Inductive Reasoning”. In: *Findings of the Association for Computational Linguistics: ACL 2024*. Ed. by Lun-Wei Ku, Andre Martins, and Vivek Srikumar. Bangkok, Thailand: Association for Computational Linguistics, Aug. 2024, pp. 13207–13219. DOI: 10.18653/v1/2024.findings-acl.782. URL: https://aclanthology.org/2024.findings-acl.782/.