

A FULLY GENERATIVE MOTIVATIONAL INTERVIEWING CHATBOT FOR MOVING SMOKERS
TOWARDS THE DECISION TO QUIT AND LLM-BASED SYNTHETIC SMOKERS

by

Zafarullah Mahmood

A thesis submitted in conformity with the requirements
for the degree of Master of Applied Science
Edward S. Rogers Sr. Department of Electrical & Computer Engineering
University of Toronto

A Fully Generative Motivational Interviewing Chatbot for Moving Smokers Towards the Decision to Quit and LLM-Based Synthetic Smokers

Zafarullah Mahmood
Master of Applied Science

Edward S. Rogers Sr. Department of Electrical & Computer Engineering
University of Toronto
2025

Abstract

Tobacco use is responsible for 7 million preventable deaths each year. For a large population of smokers, a therapeutic approach called Motivational Interviewing (MI) has shown effectiveness, yet many lack access to this support. This research presents an accessible alternative in the form of an expert-informed, large language model (LLM)-based **chatbot** that provides MI to smokers through unscripted, natural conversation. Additionally, this work contributes towards the development of **synthetic smokers**—LLM-based agents that simulate the demographic and behavioural characteristics of real human smokers, and can be used to test the chatbot during development.

A study with 106 smokers showed that a single conversation with the chatbot yielded a statistically significant average increase of +1.7 in participants’ confidence to quit at one-week follow-up. On a metric of perceived empathy, the chatbot scored higher than its predecessors, but significantly lower than typical human counsellors. An automated analysis of the transcripts showed the chatbot adhered to MI in 98% of the utterances (higher than human counsellors) and elicited a high proportion of motivational language (59%) from participants, comparable to high-quality human-delivered counselling sessions.

The fidelity of the synthetic smokers was validated by installing participants’ behavioural attributes into their “doppelgängers”. The automated analysis of transcripts showed that the level of motivation in doppelgängers’ language strongly correlated ($r_s = 0.57, p < 0.0001$) with their twin participants. In addition, synthetic smokers modulated their behaviour according to counselling quality, producing 54% less motivational language when interacting with a *confrontational* chatbot as compared to an *MI-adherent* one. This shows their potential as a tool for testing automated counselling systems and training MI counsellors.

*To the children who aged too soon,
and the futures that never were.*

Acknowledgements

I take this opportunity to extend my deepest gratitude to those who were instrumental in the completion of this work.

Dr. Jonathan Rose, C.M.,

I could not have asked for a better mentor. Our discussions and your honest critiques were instrumental in my development, not just as a researcher, but as an individual. You taught me the value of thinking and communicating clearly, and for that, I am forever grateful.

Colleagues and friends at UofT, CAMH, and beyond,

Your kindness towards me and devotion to our shared goal have been a source of my motivation. Thank you for teaching me things I didn't know existed.

Ammi,

Thank you for your endless encouragement and for being my biggest cheerleader from afar.

My brother, Aman,

Thank you for your constant support and belief in me.

To my partner in life, Rueshan,

This work would have been immeasurably less fulfilling and interesting without you. You made this a richer, more meaningful experience.

Lastly, I thank Allah, my Light and the Light of the Heavens and the Earth.

Contents

1	Introduction	1
1.1	Motivational Interviewing for Smoking Cessation	2
1.2	Large Language Models and the Automation of Talk Therapy	2
1.3	Focus and Goals	3
1.4	Contribution	3
1.5	Organization	3
2	Background and Related Work	5
2.1	Motivational Interviewing (MI)	5
2.1.1	Effectiveness of MI in Smoking Cessation	7
2.2	Foundational Language Models	8
2.3	Chatbots for Talk Therapy	10
2.4	LLM-Based Chatbots for Talk Therapy	11
2.5	Persona Creation Using LLMs	13
2.5.1	Prompt-Based Persona Conditioning	14
2.5.2	Style Transfer and Fine-Tuned Personas	14
2.5.3	Retrieval-Augmented Persona Memory	15
2.6	Evaluation of LLM-Based Persona Creation	15
2.7	LLM-Based Synthetic Subjects (Doppelgängers) in Behavioural Experiments	17
2.7.1	Current Validation Methods for LLM Doppelgängers	18
3	Development and Deployment of MIBot	20
3.1	Chatbot Design Process	20
3.1.1	Single-Prompt Architecture Rationale	20
3.1.2	Iterative Prompt Development	21
3.2	Observers	24
3.2.1	Moderator	24
3.2.2	Off-Track Conversation Classifier	24
3.2.3	End Classifier & Termination Process	25
3.3	MIBot System Design	25
3.3.1	Overview of the Application	25
3.3.2	Containerization	26
3.4	Deploying MIBot to AWS	27

3.4.1	Components of ECS	27
3.4.2	Other Components of the Deployment	28
3.4.3	Deployment Pipeline	29
4	Evaluation of MIBot: Design of the Human Feasibility Study	30
4.1	Ethics Approval and Consent	30
4.2	Participant Recruitment	30
4.3	Initial Screening Criteria	31
4.4	In-Study Screening	31
4.5	Participant Characteristics	31
4.6	Study Procedure	31
4.7	Survey Instruments	34
4.7.1	Readiness Ruler	34
4.7.2	CARE Measure	34
4.7.3	Qualitative Feedback	34
4.7.4	Follow-Up Quit Attempt Survey	34
4.8	Automated Conversation Analysis	35
5	Evaluation of MIBot: Results from the Human Feasibility Study	36
5.1	Primary Outcome: Readiness to Quit	36
5.2	Perceived Empathy: CARE Scale Assessment	39
5.3	Comparing Fully Generative MIBot v6.3 With Partially Scripted MIBot v5.2	42
5.4	AutoMISC Analysis	45
5.4.1	Contextualizing AutoMISC Metrics With the HLQC Dataset	45
5.4.2	Counsellor Behaviour Metrics	45
5.4.3	Client Change Talk Analysis	47
5.5	Behavioural Outcomes	47
5.6	Conversation Analysis	47
5.7	User Experience and Feedback	49
5.8	Discussion	50
5.9	Conclusion	51
6	Development of Synthetic Smokers: Preliminary Experiments	53
6.1	Objectives and Overview	53
6.2	Goals for a Synthetic Smoker	54
6.2.1	Defining Fidelity	55
6.2.2	Defining Representativeness	56
6.3	Development of Synthetic Smokers	57
6.3.1	The Baseline Synthetic Smoker	57
6.3.2	Controlling Verbosity	58
6.3.3	Installing Resistance	59
6.4	Conclusion	61
7	Methods for Attribute Installation in Synthetic Smokers	62
7.1	The Doppelgänger Methodology: A Data-Driven Approach	62

7.2	The Conversational Impact Modelling Experiment	63
7.3	Experiments with the Incremental Installation of Attributes	67
7.3.1	Methodology	67
7.3.2	Evaluation	71
7.3.3	Results	71
7.4	Synthetic Smokers' Sensitivity to Counselling Quality	71
7.4.1	Methodology	73
7.4.2	Evaluation	74
7.4.3	Results	74
7.5	Conclusion	74
8	Conclusion and Future Directions	75
8.1	Summary of Contributions	75
8.2	Future Directions	75
8.3	Concluding Remarks	76
A	MIBot Prompt Evolution	77
B	Observer Agent Prompts	79
C	Prompt for Virtual Smoker Client	81
D	History of the MIBot Project	83
E	Readiness Ruler Questions	85
F	CARE Survey Instrument	86
F.1	Results from the CARE survey	87
G	MIBot Feedback from Study Participants	89
H	AutoMISC	91
I	Deployment Artifacts	96
J	System Architecture Diagram	99
K	Overview of the Dataset	100
L	Consent from Participants	103
M	Summary of Literature Survey on Synthetic Doppelgänger	104
N	Example of Attribute Installation in a Doppelgänger	109
O	Full Transcript of a Doppelgänger Interaction	110
	Bibliography	121

List of Tables

2.1	Examples of change talk and sustain talk	6
2.2	Examples of MI skills	6
3.1	Initial MIBot Prompt	23
3.2	Final MIBot Prompt	23
4.1	Baseline Characteristics of Enrolled Participants	32
5.1	MIBot Readiness Ruler Summary	37
5.2	Confidence Changes by Baseline Confidence	38
5.3	Confidence Changes by Smoking Characteristics	38
5.4	Confidence Changes by Demographics	39
5.5	MIBot vs. Human CARE Scores	40
5.6	CARE Scores by Demographics and Behaviour	41
5.7	Mean CARE Scores per Question	42
5.8	AutoMISC Counsellor Metrics for MIBot vs. Human	45
5.9	AutoMISC Client Metrics for MIBot vs. Human	47
5.10	Quit Attempt Behaviour Before and After MIBot	47
5.11	Conversation Dynamics between Participants and MIBot: Quantitative Metrics	48
5.12	Segmentation of MIBot Participants based on their Experiences	50
6.1	Statistics on Volley Length for Low- and High-Verbosity Synthetic Smokers	58
6.2	% Change, sustain, and neutral talk for high- and low-resistance synthetic smokers	60
6.3	Readiness rulers reported by high- and low-resistance synthetic smokers	60
7.1	Ablation study on doppelgängers' self-reported post-conversation confidence	64
7.2	Effect of LLM type and reasoning on synthetic smokers' self-reported confidence	65
7.3	Multi-ruler conversational impact modelling results	66
7.4	Correlation between %CT and human smoker attributes in the MIBot dataset	68
7.5	Incremental attribute installation and its effect on the fidelity of doppelgängers	72
7.6	Fidelity of doppelgänger's motivational language stratified by sex	72
7.7	Effect of counselling quality on doppelgängers' motivational language	74
D.1	Summary of major MIBot versions	84

G.1	Top 10 most frequently mentioned positive words in participant feedback	89
G.2	Top 10 most frequently mentioned negative words in participant feedback	89
H.1	Demographic Information of Annotators	95
I.1	API Gateway Resource and Method Configuration	98
K.1	Description of the Columns in data.csv	100
K.2	Description of the Columns in conversations.csv	102
M.1	Validation Methods for LLM-based Synthetic Subjects	104

List of Figures

3.1	Overview of the MIBot system	24
3.2	MIBot Sanic Application Overview	26
3.3	Containerized MIBot Application	27
3.4	MIBot Deployment on AWS ECS	27
4.1	Feasibility Study Protocol Overview	33
5.1	Distribution of Confidence Changes	37
5.2	Confidence Change Distributions for MIBot v5.2 and v6.3A	43
5.3	CARE Score Distributions for MIBot v5.2 and v6.3A	43
5.4	Comparison of Mean CARE Scores per Question	44
5.5	Comparison of MISC summary score distributions for MIBot and the HLQC dataset	46
6.1	Distribution of volley length for default and fixed-verbosity synthetic smokers	59
7.1	Distribution of post-conversation confidence scores for humans compared to doppelgängers	65
7.2	Scatter plot of human-reported vs. doppelgänger-reported post-conversation confidence	66
7.3	Doppelgänger creation and validation process	69
7.4	Brief overview of doppelgänger creation and validation process	70
7.5	Scatterplot of change fraction for humans and doppelgängers	72
7.6	Distribution of change fraction (CF) for humans and doppelgängers	73
F.1	Distribution of CARE scores for <i>MIBot v5.2</i> and <i>MIBot v6.3A</i>	87
F.2	Question-wise mean CARE scores for <i>MIBot v5.2</i> and <i>MIBot v6.3A</i>	88
G.1	Word cloud representation of participant feedback	90
H.1	AutoMISC system diagram	91
H.2	Cohen’s κ between rater pairs on behaviour code annotations	92
J.1	MIBot system architecture	99

Chapter 1

Introduction

Tobacco use remains the leading cause of preventable disease and death in Canada, where more than 70% of lung cancer cases are attributable to smoking. Smoking is responsible for the deaths of approximately 45,000 Canadians annually ([Poirier et al., 2019](#)). Despite public health campaigns and cessation aids, 4.6 million Canadians continue to smoke, with higher prevalence among Indigenous communities, individuals in rural areas, and those with low incomes ([Canadian Partnership Against Cancer, 2020](#)). These groups also experience reduced access to smoking prevention, early diagnosis, and treatment services, which contributes to inequities in health outcomes.

One of the most effective evidence-based interventions for smoking cessation is talk therapy. Several therapeutic modalities have shown effectiveness in smoking cessation, including cognitive behavioural therapy (CBT) ([Beck et al., 1979](#)), acceptance and commitment therapy (ACT) ([Hayes et al., 1999](#)), and motivational interviewing (MI) ([Miller, William R. and Rollnick, Stephen, 2023](#)). The latter is the focus of this thesis. Access to such interventions, however, is hindered by several systemic challenges. First, traditional talk therapy can be prohibitively expensive for individuals without private insurance or those under financial constraints; although CBT is clinically effective, publicly funded CBT remains scarce in Canada ([Andrade, 2020](#); [Gratzer and Goldbloom, 2016](#)). Second, therapy services are limited in availability, especially in rural and remote regions: residents in rural Canada face long distances to providers, transportation difficulties, fewer clinicians, and lower socioeconomic resources, all constraining access ([Burns et al., 2007](#); [James et al., 2021](#)). Third, there is a substantial shortage of culturally safe mental health services tailored to the unique historical and social contexts of Indigenous populations ([Josewski, 2023](#); [Hart-Wasekeesikaw, 2009](#)). Indigenous communities experience systemic underfunding, long wait times, inadequate provider availability, and cultural mismatches between Indigenous clients and predominantly non-Indigenous therapists ([National Collaborating Centre for Indigenous Health, 2020](#)).

One way to address the lack of accessible mental health care is through chatbot-based counselling, which can serve users, especially when other options are limited. Chatbots are always available, require no appointments, protect anonymity, and can scale to thousands of users. Once developed, their marginal cost is nearly zero, making them especially suitable for underserved populations ([Torous and Roberts, 2017](#); [Miner et al., 2016](#)). Compared to traditional therapy, they offer a low-cost, scalable complement.

1.1 Motivational Interviewing for Smoking Cessation

Motivational interviewing (MI) has emerged as particularly effective for smoking cessation. MI is a client-centred counselling method developed in the early 1980s by William R. Miller and later expanded in collaboration with Stephen Rollnick (Miller and Rollnick, 1991; Miller, William R. and Rollnick, Stephen, 2023). It is designed to strengthen individuals’ intrinsic motivation to change their health-compromising behaviours, such as smoking. MI has been widely applied in addiction treatment settings and has shown consistent effectiveness in encouraging smoking cessation, particularly among individuals not yet ready to quit (Bischof et al., 2021; Lundahl et al., 2010).

The hallmark of MI is its collaborative, non-confrontational style, which relies on empathy, active listening, and evocation. Rather than telling clients what to do, MI counsellors aim to guide them in articulating their own reasons for change. The core principles of MI are: (1) expressing empathy through reflective listening; (2) developing discrepancy between current behaviour and personal goals; (3) rolling with resistance rather than confronting it; and (4) supporting self-efficacy (Rollnick et al., 2008). These principles are operationalized using skills such as open-ended questions, affirmations, reflections, and summaries—commonly referred to as the OARS model (Miller, 2023).

A critical insight from MI is that *ambivalence*—a state in which individuals simultaneously hold conflicting feelings about change—is often the first psychological barrier to change. Smokers may simultaneously recognize the harms of smoking and yet be unwilling to change due to perceived benefits or entrenched habits (Brown et al., 2023). Resolving this ambivalence is a necessary precursor to behaviour change. Numerous studies have demonstrated that MI increases the number of quit attempts and increases confidence to quit among ambivalent smokers (Abar et al., 2013; Gwaltney et al., 2009). MI promotes a shift from “sustain talk” (statements that favour the status quo) to “change talk” (statements that favour behavioural change), and the frequency and strength of change talk during a session have been shown to predict actual behavioural outcomes (Apodaca and Longabaugh, 2009).

1.2 Large Language Models and the Automation of Talk Therapy

The automation of talk therapy, once confined to rule-based systems and predefined scripts, has shifted since the advent of large language models (LLMs). The development of transformer-based architectures (Vaswani et al., 2017), particularly autoregressive generative models like GPT-3 and GPT-4o, has enabled machines to produce responses that are coherent, contextually appropriate, and human-like in natural dialogue (OpenAI, 2023). These capabilities have made fully generative counselling by LLMs not only possible but increasingly effective for therapeutic applications (Miner et al., 2020; Lee et al., 2021).

Several technical innovations have enabled the adaptation of LLMs to therapeutic contexts. The self-attention mechanism within transformer models supports long-range dependency modelling and context tracking across multi-turn dialogues (Vaswani et al., 2017). Fine-tuning an LLM on domain-specific corpora (e.g., MI transcripts) can make the model outputs clinically grounded (Kong and Moon, 2025), while reinforcement learning with human feedback (RLHF) has been shown to improve the relevance, tone, and empathetic quality of a generated response (Ouyang et al., 2022). Together, these capabilities enable LLMs to generate novel therapeutic reflections, adhere to evidence-based counselling principles, and adaptively modulate tone.

1.3 Focus and Goals

The overarching goal of this project is to develop and evaluate a fully generative chatbot capable of providing MI counselling to help smokers move towards the decision to quit. This involves designing and implementing an MI counsellor that adheres to the core principles of MI, evaluating its effectiveness with human smokers using both human assessments and automated metrics, and creating *synthetic smokers*—LLM-based personas that mimic human smokers and consistently display realistic behavioural traits during MI conversations. The realism of these synthetic smokers is validated by comparing their conversational characteristics and behaviours with those observed in actual human smokers. Collectively, these efforts contribute to the broader vision of advancing AI-assisted mental health support through technologies that are safe, accessible, and grounded in evidence-based practices.

1.4 Contribution

This work makes the following contributions to the development and validation of a generative MI-based chatbot for smoking cessation:

1. Design and iterative refinement of a single-prompted, fully generative chatbot capable of conducting motivational interviews with human smokers.
2. Execution of an empirical study in which compensated human participants interacted with the chatbot, providing qualitative feedback and pre- and post-interaction self-reports on their readiness to quit smoking.
3. Development of a methodology to construct *synthetic smokers*—LLM-based agents embedded with the demographic and behavioural characteristics of human smokers.
4. Creation of evaluation techniques to assess the extent to which these synthetic agents approximate human smokers in controlled behavioural experiments.
5. Validation of synthetic smokers by quantifying their similarity to human smokers using the aforementioned evaluation techniques.

1.5 Organization

This thesis is structured as follows:

- Chapter 2 reviews relevant background on the Motivational Interviewing approach, foundational language models, chatbots for talk therapy, persona construction using LLMs, and prior work on the development and validation of LLM-based synthetic subjects in behavioural experiments.
- Chapter 3 presents the design of the fully generative MI counselling chatbot, along with details of its deployment.
- Chapter 4 describes the design of the human feasibility study.
- Chapter 5 reports the results of a study in which recruited smokers interacted with the chatbot.

- Chapter 6 describes a method for constructing synthetic smokers using large language models.
- Chapter 7 introduces a methodology for evaluating synthetic smokers.
- Chapter 8 concludes the work by discussing its limitations and outlining directions for future research.

Chapter 2

Background and Related Work

This chapter establishes the conceptual and empirical foundations for our contributions to automated talk therapy and the development of synthetic agents. We begin by surveying research on the clinical efficacy of motivational interviewing for smoking cessation. We then examine the emergence of transformer-based LLMs as conversational agents, with a focus on recent efforts to develop motivational interviewing counsellor chatbots using LLMs. We then survey methods for constructing synthetic agents, with a particular emphasis on LLM-based synthetic patients designed for behavioural research. We particularly focus on the techniques of persona *installation* via prompting, which seek to *embed* demographic and behavioural characteristics into synthetic agents. We also outline the inherent limitations of prompt-based persona installation, including issues of consistency, depth, and stereotype propagation. Finally, we critically review existing approaches for validating the fidelity of persona installation and compare them with our own validation methods.

2.1 Motivational Interviewing (MI)

Motivational Interviewing (MI) is a counselling technique originally developed in the 1980s to treat alcohol dependence and has since been applied across a wide range of health interventions, including smoking cessation (Miller, 1983; Miller, William R. and Rollnick, Stephen, 2023). MI is defined as a collaborative, client-centred conversational method designed to elicit intrinsic motivation for change by helping individuals resolve ambivalence. Instead of confronting or directing the client, the MI practitioner adopts a guiding stance: asking open-ended questions, listening reflectively, and echoing the client’s own change-relevant statements. This approach aims to elicit *change talk* (client statements in favour of change) while reducing *sustain talk* (arguments for maintaining the status quo) (Miller and Rose, 2009). By evoking a person’s own reasons for change in a non-judgmental and supportive way, MI strengthens their perceived autonomy and self-efficacy. Table 2.1 presents examples of change and sustain talk alongside MI-consistent responses by the counsellor.

MI Principles and Style

Underlying MI’s conversational approach are several core principles. First, the counsellor should express empathy and use reflective listening to understand the client’s perspective and build rapport. Second, MI

Speaker	Change Talk	Sustain Talk
Client	“I know I should quit smoking because my kids hate the smell, and I don’t want them to pick up the habit.”	“I’ve tried quitting before, but I always end up lighting one when I’m stressed out. It’s just who I am.”
Counsellor	“It sounds like you really care about setting a good example for your children.”	“So smoking feels like a part of how you manage difficult emotions.”

Table 2.1: Examples of change talk and sustain talk in motivational interviewing, with corresponding counsellor responses. The table shows client statements expressing motivation to change (change talk) and resistance to change (sustain talk), and how a counsellor might respond to each in an MI-consistent manner.

works to reveal any discrepancy between the client’s own goals or values and their current behaviour (Miller, 2023). Third, the counsellor acknowledges the client’s resistance rather than confronting it; resistant remarks are met with understanding and are used as opportunities to examine the client’s thoughts further, instead of provoking an argument. Finally, MI supports self-efficacy by emphasizing the client’s autonomy and capability in effecting change—the individual is encouraged that they have the strength and choice to quit if they so decide. These principles are often operationalized through specific conversational techniques summarized as OARS: Open-ended questions, Affirmations, Reflective listening, and Summaries (Rollnick and Miller, 1995). Examples of how MI counsellors use these skills are given in Table 2.2.

MI’s empathetic, autonomy-supportive atmosphere is particularly important for *ambivalent smokers*—those who may be defensive or unsure about quitting. It helps reduce resistance and increases engagement in the conversation about change (Miller, 1983; Miller, William R. and Rollnick, Stephen, 2023). Notably, MI’s strategy of guiding clients to articulate their own arguments for change is grounded in evidence that clients’ *change talk* during sessions predicts a greater likelihood of subsequent behaviour change (Miller and Rose, 2009). Thus, MI sessions explicitly aim to encourage change talk and soften sustain talk, steering the dialogue in a direction where the client’s language shifts towards change.

MI Skill	Example
Open-ended Question	“What are some things you’ve thought about when it comes to cutting back on drinking?”
Affirmation	“You’ve shown a lot of strength in coming here today and being open about what’s going on.”
Reflective Listening	“So you’re feeling stuck. You want to make a change, but you’re also worried you might fail again.”
Summary	“Let me see if I’ve got this right: you’ve been thinking more about quitting, especially since your health scare, but it’s been hard to imagine your daily routine without smoking. At the same time, you’ve started walking more and cutting back already.”

Table 2.2: Examples of core skills in motivational interviewing (OARS). The table provides examples for open-ended questions, affirmations, reflective listening, and summaries.

Measuring the Effectiveness of MI Counselling

As a structured therapeutic approach, MI uses well-defined success criteria. Researchers and clinicians use several strategies to evaluate the quality of MI conversations and their effect on client motivation. One

common approach is the Motivational Interviewing Skill Code (MISC), a coding system that categorizes counsellor utterances and client responses to quantify adherence to MI principles (Houck et al., 2010a). Using the MISC, independent annotators (or *coders*) can rate how well a counsellor’s statement aligns with MI techniques (for example, by counting reflections, questions, and advice) and determine the proportion of client change talk vs. sustain talk. High MI-consistent scores (e.g., a high ratio of reflections to questions, or a high percentage of client change talk) are associated with better outcomes, and such coding schemes are often used in training and research to ensure the fidelity of MI delivery.

Another practical tool is the “Readiness Ruler,” a simple self-reported measure of a client’s readiness to change (on a 0-10 scale for readiness, importance, or confidence) (Boudreaux et al., 2012). In the context of smoking cessation, a counsellor might ask, “On a scale from 0 to 10, how ready are you to quit smoking?” The Readiness Ruler is an effective way to track changes in motivation before and after an intervention. For example, an increase in a smoker’s readiness score after an MI session would indicate movement toward a decision to quit. Both the MISC coding of session transcripts and readiness scaling of clients are valuable evaluation methods: the former measures the level of motivational language by the client and ensures the conversational style remains true to MI, and the latter provides an outcome-oriented metric of the client’s motivational state.

While MI provides the therapeutic approach, the technological means to deliver it at scale in a conversational manner have only recently become available with the advent of large language models.

2.1.1 Effectiveness of MI in Smoking Cessation

MI has been widely adopted in smoking cessation efforts, particularly due to its relevance for smokers who experience ambivalence about quitting. For instance, a national survey found that over half of U.S. smokers express conflicting attitudes toward cessation (Babb et al., 2017), necessitating interventions that can address such ambivalence.

Over the past two decades, numerous clinical trials and meta-analyses have assessed the efficacy of MI counselling in helping tobacco users quit. A meta-analysis of 31 randomized trials involving over 9,000 smokers reported that MI markedly increased the likelihood of abstinence compared with control conditions, with a pooled odds ratio of approximately 1.45 (Heckman et al., 2010). Similarly, a Cochrane review of 28 studies found that MI-based counselling produced higher six-month quit rates than brief advice, with relative risks ranging from 1.2 to 1.3 (Lindson-Hawley et al., 2015). While these effect sizes are modest, the evidence consistently suggests that MI improves both quit attempts and abstinence, especially when delivered by trained practitioners in clinical or community settings.

The effectiveness of MI is consistent across various smoking populations and settings. Studies have shown positive outcomes with MI delivered by various professionals (physicians, nurses, trained counsellors) and in formats ranging from a single brief session to multiple sessions (Lindson-Hawley et al., 2015). Even a short, 15-20 minute MI-based conversation in a primary care visit can measurably boost a smoker’s likelihood of quitting relative to no counselling, especially when the practitioner adheres closely to MI principles (Zanjani et al., 2008).

One reason MI is particularly effective for smoking cessation is its alignment with the psychology of ambivalence common among smokers. Many smokers acknowledge the health risks of tobacco while

simultaneously relying on it for stress relief or as a habitual comfort, resulting in decisional conflict. MI directly engages this ambivalence by creating a non-confrontational space in which smokers can articulate and examine their mixed feelings, ultimately shifting the balance toward change. Empirical evidence suggests that MI not only improves cessation outcomes but also improves intermediate factors such as motivation, readiness to quit, and self-efficacy (Boudreaux et al., 2012; Hettema et al., 2005). These motivational gains are important, as a readiness to quit is a well-established precursor to cessation success (West and Sohal, 2006).

2.2 Foundational Language Models

The recent improvements in conversational AI have been advanced by foundational language models (Miller, 2021): extremely large neural networks pre-trained on vast corpora of text, which can be adapted to many tasks. These models serve as a foundation that can be specialized for specific applications via *fine-tuning*. Modern foundation models almost universally employ the **transformer** architecture (discussed in Section 2.2) to capture long-range context and dependencies in dialogue. Consequently, this allows LLM-based chatbots to understand and generate coherent multi-turn conversations and provide consistent responses to the client’s statements over long contexts (e.g., multiple sessions).

Foundation models are trained with self-supervised objectives on enormous text datasets, learning a broad range of linguistic patterns, factual knowledge, and even elusive interaction norms. The GPT series of models exemplifies how scaling up model size and data leads to *emergent capabilities*. GPT-3 (Brown et al., 2020) (175 billion parameters) demonstrated astonishing *few-shot learning*: it can perform a new language task given only a few examples. This few-shot ability is a direct consequence of training on various data at scale, which allows models to perform *in-context learning*, i.e., adapting to a new task described in the prompt. Such capabilities are invaluable for building a therapeutic chatbot. Instead of laboriously collecting and annotating thousands of counselling dialogues to train a model, it is possible to prompt a pre-trained LLM with instructions and examples of MI, and the model will generalize to produce appropriate counsellor responses (Xie et al., 2024). Later in this thesis, a foundation model is prompted with an expert-informed prompt. This method uses the model’s generative capabilities while allowing the customization of the output to be MI-consistent. This approach follows a broader trend in NLP: using large pretrained models as a base and conditioning them via prompts or lightweight fine-tuning to perform specialized dialogue tasks (Wei et al., 2022).

While OpenAI’s GPT models were pioneers, the field of foundational models has rapidly expanded, creating a competitive and varied environment. Other major technology firms have developed their own flagship models, often with distinct architectural philosophies and target applications. Google’s Gemini family, for instance, was designed from the ground up to be natively multimodal, processing text, images, audio, and video as interleaved sequences, a departure from models that handle modalities via separate components (Gemini Team, Google et al., 2023). Anthropic’s Claude series, particularly the Claude 3 models, has advanced benchmark performance while being developed with a strong emphasis on safety through methods like Constitutional AI; these models are also noted for their exceptionally large context windows (Anthropic, 2024).

A notable counter-trend to these proprietary, closed-weight models has been the proliferation of high-performance open-weight alternatives. Meta has been a leader in this domain, with its Llama series (Touvron et al., 2023) providing researchers and developers with high-performance base models that can be studied

and fine-tuned with transparency. This movement has catalyzed innovation across academia and industry. Similarly, organizations from other countries have made substantial contributions; for example, China’s DeepSeek-AI has released a series of proficient open-source models, including specialized variants for code generation and models built on Mixture-of-Experts (MoE) architectures (DeepSeekAI, 2024).

The diversification extends beyond just scale. Microsoft has investigated the development of smaller, more curated models with its Phi series and has demonstrated that models trained on high-quality, “textbook-like” data can achieve remarkable performance on reasoning tasks, challenging the “bigger is always better” assumption (Li et al., 2023b). Concurrently, specialized models targeting niche domains have emerged. For instance, xAI’s Grok is designed to use real-time information from the X social media platform (xAI, 2023), while enterprise-focused offerings like Amazon’s Titan (Amazon Web Services, 2023) and IBM’s Granite (IBM, 2023) prioritize data provenance, reliability, and integration within their respective cloud ecosystems. This rich array of foundational models, each with its own philosophy on openness, scale, and specialization, provides a broad array of tools that can be adapted for highly specific and sensitive applications.

This is a promising development for therapeutic chatbots: as these models advance, a carefully adapted version can exhibit even more natural dialogue and advanced motivational strategies. At the same time, research into controllability, i.e., giving developers and clinicians the tools to direct an LLM’s behaviour, is growing in importance (Fernandez et al., 2025). Techniques like system prompts, chain-of-thought prompting for reasoning (Wei et al., 2022), and lightweight policy modelling (Du et al., 2024) are some emerging trends to further improve an LLM’s use as a goal-oriented dialogue system.

The Transformer Neural Architecture

Transformers are the architectural backbone of virtually all modern large language models, and they play a central role in the chatbot developed in this work. The transformer architecture (Vaswani et al., 2017) departed from previous neural network designs by using self-attention as its sole mechanism to “compute representations of its input and output without using sequence-aligned RNNs or convolution” (Vaswani et al., 2017). This allowed parallel processing of each input token in the context to compute its deep context-dependent representation. In a transformer, each input token’s representation can be computed by attending to every other token in the context (the complete conversation up until now), which enables the model to capture long-range relationships and contexts. This is implemented through multi-head self-attention layers: the model computes attention weights that represent different types of relationships (e.g., semantic similarity, positional relevance) across the sequence. By stacking multiple self-attention layers, transformers can build very deep representations of text. Crucially, they scale efficiently on parallel hardware because each layer’s computations can be parallelized (unlike the sequential nature of RNNs). This scalability has allowed training of extremely large models with hundreds of billions of parameters on massive datasets.

Large generative models like GPT-3 (Brown et al., 2020), GPT-4o (OpenAI, 2023), and others use transformers that predict text autoregressively, i.e., one token at a time, conditioned upon the preceding context. This autoregressive setup is well-suited for dialogue generation, as the model always conditions on the conversation history when producing the next part of its response.

Empirically, the advent of transformer-based LLMs has yielded dramatic improvements in dialogue systems. Models like PaLM (Chowdhery et al., 2022) (540 billion parameters) and Google DeepMind’s Gemini (Gemini Team, Google et al., 2025) employ the same transformer building blocks, but at a greater scale and sometimes

with enhancements like sparsity or routing. The general finding is that larger transformers not only produce more fluent text but also exhibit emergent behaviours such as reasoning, abstraction, and skilful dialogue skills that smaller models lack (Zoph et al., 2022; Berti et al., 2025).

Before the advent of transformer-based LLMs, however, the development of therapeutic chatbots had a long and varied history.

2.3 Chatbots for Talk Therapy

Computer-based *chatbots* have long been examined as a means to deliver talk therapy through natural language. Early attempts date back to the 1960s, when systems like ELIZA simulated a Rogerian psychotherapist by pattern-matching user prompts and responding with scripted phrases (Weizenbaum, 1966). While ELIZA’s author intended it as a trivial demonstration, many users unexpectedly found the experience cathartic, mistaking the program for a genuine empathic listener. This serendipitous use of ELIZA foreshadowed both the potential and limitations of early therapeutic chatbots. In the 1970s, Colby and colleagues developed PARRY, a system that modelled paranoid thought patterns to mimic a patient with paranoid schizophrenia (Colby et al., 1971). PARRY’s ability to engage psychiatrists in text dialogue was striking for its time, but like ELIZA, it relied on manually created rules and keywords rather than any true language understanding. These early systems demonstrated that even simple keyword-driven dialogues could evoke an illusion of conversation, yet they lacked memory, contextual understanding, and flexibility. As such, they could not move beyond superficial interactions.

Through the 1980s and 1990s, progress in *talk therapy chatbots* stagnated. Some researchers, however, turned to expert systems and logic-based approaches for clinical use, but these were not true free-form chatbots. Several publications in the mid-1980s examined whether computers could mimic a psychotherapist’s reasoning or assist with brief psychotherapy techniques, often by guiding patients through structured question-answer routines (Hartman, 1986; Sampson, 1986; Servan-Schreiber, 1986). These systems remained largely rule-based: they followed predetermined scripts or decision trees derived from therapeutic principles, without the ability to truly “understand” natural language. A few projects showed modest success. For example, an early interactive program for cognitive-behavioural therapy (CBT) was tested for treating depression (Selmi et al., 1990). However, research on talk therapy chatbots in this era gained little attention. Nonetheless, the idea of computer-aided therapy persisted. By the late 1990s, researchers began developing digital self-help programs that delivered therapy exercises through a desktop computer. One of the first randomized trials of computerized CBT was an interactive software, Beating the Blues, that demonstrated that guided online CBT could markedly reduce anxiety and depression symptoms (Proudfoot et al., 2003). These efforts, while not “chatbots” in the modern sense, established an important proof of concept: computers could deliver legitimate mental health interventions following psychological approaches, even if early systems were highly scripted and simplistic.

The next generation of therapeutic chatbots emerged in the 2010s alongside advances in natural language processing. These systems often adopted a *hybrid approach*, combining scripted decision flows with modest machine learning components (e.g., classifiers to detect user sentiment or intent). Notable examples include conversational agents for mental health like Woebot (Fitzpatrick et al., 2017) and Wysa (Chang et al., 2024), which deliver principles of CBT or other interventions via a chat interface. Woebot engaged users with brief daily check-ins and mood tracking, intermixing scripted prompts and pre-written empathetic replies. Its

dialogues were structured as a branching tree, augmented by simple natural language processing at certain nodes. For example, when Woebot recognized a word indicating loneliness, it replied with a comforting phrase. In a randomized controlled trial with college students, Woebot markedly reduced self-reported depression symptoms over two weeks compared to an information-only control, illustrating the promise of such automated support (Fitzpatrick et al., 2017). Similarly, Wysa (Chang et al., 2024), a CBT-based mental health chatbot, has been evaluated in real-world and clinical settings, with some studies suggesting reductions in depression and anxiety with use. These *rule-based* or *hybrid chatbots* can deliver psycho-education and guide users through therapeutic exercises like breathing or re-framing thoughts. Users often report them as convenient and stigma-free support between or instead of human therapy sessions.

One of the earliest rule-based chatbots for MI-based smoking cessation was developed by Almusharraf (2019). The developers took a human-centred design approach and collected free-form responses from 121 smokers to train the chatbot’s natural language understanding (NLU) components (or intent classifiers). The chatbot was then tested on 100 additional smokers, who showed a marked increase in confidence to quit smoking—a mean increase of 0.8 points on a 0-10 scale ($p = 0.00005$) one week after the interaction. This demonstrated the system’s potential to help unmotivated smokers move toward quitting.

To this point, all the chatbots we have discussed are rule-based. The lack of free-form language generation in rule-based chatbots means these bots often repeat canned phrases, which patients can find formulaic or insincere. This contributes to low engagement and high attrition: many users try these chatbots only briefly and discontinue when the interaction feels stagnant or “bot-like.” Indeed, Limpanopparat et al. (2024) highlighted that many *rule-based* chatbot interventions suffered from users dropping out early, undermining their long-term effectiveness. In contrast, the exceptional conversational capabilities of LLM-based chatbots may be able to provide an engaging, almost human-like experience to users. As such, the section below focuses on the recent developments in LLM-based chatbots.

2.4 LLM-Based Chatbots for Talk Therapy

Recent breakthroughs in LLMs have started a new wave of development in AI chatbots for talk therapy. Unlike rule-based systems, these models can generate free-form and contextually relevant responses. They can also hold a far more natural and flexible dialogue with users. In the therapy domain, this means a chatbot can potentially *respond to anything* a client says, while maintaining a supportive and goal-oriented stance.

LLM-based chatbots have been shown to provide talk therapy using principles from popular therapy techniques such as MI or CBT (Mahmood et al., 2025; Kian et al., 2024; Ye et al., 2025). For example, Shen et al. (2020) fine-tuned a GPT-2 model to produce *reflections* in an MI counselling style. Human evaluators rated the model-generated reflective statements as slightly higher in quality than real practitioner reflections, suggesting that a well-trained language model can capture the essence of empathic, complex paraphrasing. Although the reflections produced by Shen et al. (2020) were intended as a training aid for human therapists (to provide examples of good reflective listening) rather than for direct use with patients, this work demonstrated the capacity of transformers to generate novel therapeutic responses that adhere to MI principles. Similarly, Brown et al. (2023) used LLM-generated reflections in an MI chatbot and showed that the system was more empathetic and led to a higher increase in confidence to quit smoking compared to a previous, non-generative version of the chatbot.

Several fully generative therapeutic chatbots have been developed in the last two years. One notable system is TAMI (Technology-Assisted Motivational Interviewing), a chatbot coach for smoking cessation that integrated transformer-based language understanding and generation (Saiyed et al., 2022). TAMI used intent classifiers to recognize user inputs and a transformer model to generate MI-consistent replies, including both simple and complex reflections. In a 2022 pilot trial with 34 smokers, users rated TAMI as highly competent in MI skills, though overall satisfaction with the bot was moderate (3 out of 5). These findings indicated that while the bot successfully employed MI techniques, there remained room to improve the user experience, highlighting the need for even more natural dialogue and empathy.

The latest therapeutic chatbots often use advanced foundational language models, either *fine-tuned* on therapy data or guided via careful prompting. For instance, **MIcha** is a GPT-4o-based chatbot designed to deliver motivational interviewing for behaviour change (Meyer and Elswailer, 2025). In a randomized controlled trial, MIcha’s brief conversations markedly increased users’ readiness to change unhealthy behaviours compared to a control (an unprompted GPT-4o instance). On a 0–10 *Readiness to Change* (Biener and Abrams, 1991) scale, interactions with the *MI-adapted* chatbot led to a mean increase of 0.87 (SD=2.02), compared to the control, which led to an increase of 0.73 (SD=2.05) (Meyer and Elswailer, 2025). The chatbot was prompted to use MI principles, and the study found that using MI-consistent language in generation not only improved outcomes but also helped mitigate harms like inappropriate advice or user distress. Interestingly, the researchers observed that users fell into distinct interaction styles—“cooperative” vs. “resistant”—which influenced conversation outcomes. Such insights hint that LLM chatbots might be able to adapt their approach in real-time to different client personalities, a level of personalization impossible with one-size-fits-all scripts.

Another important work is by Heinz et al. (2025), who developed **Therabot** and described it as the first generative AI therapy chatbot to undergo a full clinical trial. Therabot uses a fine-tuned LLaMA-2-70B model to deliver CBT for depression, anxiety, and eating disorders, and was evaluated in a four-week **randomized controlled trial**. Participants who engaged with Therabot showed substantial reductions in symptom severity: about a 51% drop in depression symptoms, with anxiety and eating disorder symptoms also markedly decreased relative to a waitlist control (30% and 18% drop, respectively). Strikingly, users of the AI agent rated their *therapeutic alliance* with the chatbot on par with the alliance they typically feel with humans. This suggests that a well-designed LLM chatbot, by virtue of highly responsive and understanding dialogue, can engender a sense of rapport and trust close to that of real therapy. The chatbot incorporated safety guardrails and human clinician oversight for crisis situations. For example, Therabot was programmed to recognize signs of a user in acute distress and provide gentle crisis intervention messages while simultaneously alerting human support (with an option to connect to a crisis line). These safeguards illustrate the hybrid approach often taken in practice: using LLMs for free-form therapeutic conversation but backing them with controlled protocols for high-risk scenarios. Overall, the success of Therabot’s trial is a strong proof-of-concept that LLM-driven chatbots can deliver clinical-level mental health benefits in practice.

LLMs also have a very long context window, which allows them to recall details from earlier sessions and respond with original reflective statements that make the client feel heard. In the case of multi-session interventions, techniques such as session summarization and retrieval-augmented generation (RAG) can be used to tailor the interaction style for the upcoming session. An approach to this has been presented by Corda et al. (2024), who used LLMs to generate personalized advice for sleep improvement.

Applications of LLMs in Improving Therapeutic Chatbots

In addition to **generating** coherent, context-aware responses for therapeutic chatbots, LLMs can be used in **modelling the clients** for training and evaluation purposes. We examine this idea in depth in Section 2.7. LLMs have also been employed in **evaluating** chatbot performance: rather than relying solely on labour-intensive human annotation of transcripts, researchers have begun using LLMs such as GPT-4o to automatically assess whether a conversation adheres to MI principles and even to rate the quality of reflections produced. Early studies show reasonably good agreement between GPT-based evaluators and human judgments (Scholich et al., 2025).

Limitations and Challenges in Using LLM-Based Therapeutic Chatbots

Despite the promising outlook of LLM-based chatbots as therapists, some major challenges are yet to be overcome. One such challenge is the LLM’s tendency for **hallucination**, i.e., to generate plausible-sounding but incorrect or ungrounded statements. In general usage, models like ChatGPT have been found to produce some factual error or fabrication in roughly 20% of their responses (Li et al., 2023a). In a mental health context, such hallucinations could translate to unhelpful or even harmful guidance. For example, making up an unfounded statistic about a treatment or misinterpreting a user’s story in a way that breaks rapport. Solutions being examined include grounding the model in verified psychoeducational content and implementing filters for medical advice (Amugongo et al., 2025).

There is also a complex issue relating to empathy and authenticity. While LLMs can be prompted to respond with empathetic phrases (and often do so convincingly), some claim that the interaction still ‘feels different’ from a human, especially over time. The empathy is a “simulation rather than an attuned human reaction” (Seitz, 2024), which can create a gap in how the support is perceived. For example, a study comparing GPT-3-based chatbots to human therapists found that the bots tended to overuse formulaic, reassuring, and affirming statements, yet did not probe deeply into clients’ feelings or ask many detailed questions. Therapists, in contrast, elicited more elaboration from clients and used complex reflections and occasional self-disclosures to build connection (Scholich et al., 2025). This suggests that current general-purpose LLMs, if used “out-of-the-box,” may lean towards a superficial counselling style—politely supportive but missing opportunities to examine the client’s experience in depth. In crisis situations, these gaps become even more pronounced: the same study noted that unspecialized chatbots handled scenarios of suicidal ideation or severe distress inadequately, often failing to ask about safety or to encourage seeking professional help.

2.5 Persona Creation Using LLMs

LLMs can be equipped with artificial personas to produce engaging and contextually appropriate behaviour in dialogue systems. Before the advent of LLMs, early work on persona-grounded dialogue demonstrated that conditioning responses on a predefined personal profile can address problems of blandness and inconsistency in chit-chat models. For example, Zhang et al. (2018) introduced the PersonaChat dataset where dialogue agents were given *persona profiles* (e.g., “I have a dog. I like camping.”) and showed that conditioning on such profiles yielded more specific and engaging conversations compared to profile-agnostic models. This persona-conditioning approach was found to improve next-utterance prediction and overall dialogue coherence, as the model maintains a consistent identity throughout an interaction. Subsequent research built on this idea by integrating persona information into both retrieval-based and generative chatbots, confirming that

persona grounding can improve user engagement and the realism of agent utterances (Roller et al., 2021; Li et al., 2016). At the same time, these studies revealed challenges: agents often still produced contradictions to their stated *persona profiles* or reverted to generic responses if the persona was not reinforced strongly (Kim et al., 2020; Song et al., 2020). Using pre-trained LLMs for persona creation mitigates this problem to a large extent as the *persona profiles* can be attended to on every token generation.

2.5.1 Prompt-Based Persona Conditioning

A common approach to creating an LLM-based persona is to inject *persona profiles* at inference time with prompt engineering. In this method, the model is steered by a carefully designed prompt that delineates the character’s identity, backstory, or speaking style. For instance, a system message might instruct: “You are a 45-year-old smoker who has tried quitting multiple times and is ambivalent about quitting again.” This kind of backstory conditioning provides an initial context that biases the LLM’s generation towards the persona’s perspective. Contemporary instruction-following models like GPT-4o accept system prompts that effectively establish such roles, enabling zero-shot persona adoption without additional training (Ouyang et al., 2022). Empirical work has shown that even concise persona descriptions in the prompt can sometimes markedly influence an LLM’s lexical choices, tone, and factual claims in ways consistent with that persona (Madotto et al., 2019; Liu et al., 2024).

In more elaborate setups, designers provide a series of in-context demonstrations, e.g., example dialogues or question-answer pairs that exemplify the persona’s behaviour (what might be called **behavioural exemplification**). By seeing a few turns of a persona in action, the LLM can learn to mimic the speaking style and attitudes illustrated by those examples (Joshi et al., 2024). This few-shot prompting strategy has been used to install personas ranging from cheerful customer service agents to sarcastic comedians, with qualitative improvements in staying “in character” (Gururangan et al., 2020). Prompt-based methods have the advantage of not requiring model fine-tuning, but they rely on the model’s context window and can degrade as a conversation progresses and new context pushes out the initial persona prompt.

2.5.2 Style Transfer and Fine-Tuned Personas

Another line of work treats persona adoption as a controllable text style problem. Rather than (or in addition to) conditioning the model on a persona description, one can post-process or constrain generation to match a target style associated with the persona. For example, a base LLM might first generate a candidate response based on conversational context, and then a style transfer model rewrites that response in the voice of a 45-year-old male smoker from a particular background.

Prior research on text style transfer provides tools for altering attributes like formality, sentiment, or dialect while preserving meaning (Niu and Bansal, 2018; Sudhakar et al., 2019). These techniques have been extended to dialogue personalization, such as Niu and Bansal (2018), which used a politeness classifier and style-specific language model to make a chatbot consistently polite or rude on demand. Also, Zhang et al. (2018) trained a transformer-based dialogue model with style embeddings for persona and emotion to generate stylized responses in one pass. Such style-controlled generation can intensify persona-specific linguistic quirks (choice of words, syntax, formality level) and has shown success in producing responses that human evaluators identify as having a distinct personality (Zhang et al., 2018).

An alternative approach to style transfer is fine-tuning the LLM on persona-specific data. By training on

dialogues where an agent consistently speaks with a given persona (or on monologues/writings representative of that persona), the model internalizes the patterns of that character. Fine-tuning was the primary method in early persona-chat systems, often combined with latent variable models to encode persona traits (Li et al., 2016). For example, Roller et al. (2021) described fine-tuning a 9B-parameter pre-trained model on the PersonaChat dataset and other persona-annotated dialogues, which yielded a chatbot that humans found more consistent in personality and context coherence than a non-fine-tuned baseline. However, fine-tuning large LLMs for each persona is costly and inflexible; hence, prompt-based steering and modular style transfer are increasingly favoured for flexible persona switching.

2.5.3 Retrieval-Augmented Persona Memory

Maintaining a consistent persona over long interactions is a known difficulty. As conversations stray from the initial topic or exceed the model’s context length, the agent may “forget” its persona or drift in style. Modern systems address this with retrieval-augmented generation (RAG) techniques, wherein relevant persona information is fetched from an external store and fed into the model at each turn (Shuster et al., 2022; Xu et al., 2022). In a persona-aware RAG pipeline, the agent might have a specific memory of persona facts or dialogue history that is indexed (perhaps through a vector database). Prior to each response, the system retrieves the most pertinent persona snippets and prepends them to the LLM’s input. BlenderBot (Shuster et al., 2022) incorporates a long-term memory component that stores both the user’s persona and the bot’s own persona, and it learns to decide when to retrieve these memory entries to ground its responses (Shuster et al., 2022). This helps the agent avoid contradicting earlier statements about itself or repeating questions the user has answered. Similarly, Multi-Session Chat by Xu et al. (2022) uses retrieval to carry a persona across multiple dialogue sessions, ensuring that a chatbot remembers a user’s personal details and prior conversations even after many interactions.

Retrieval-based persona conditioning has been reported to improve consistency and factual alignment with the persona profile, though it requires reliable triggering mechanisms to decide when persona memory is needed. Recent research prototypes like **PersonaRAG** explicitly combine user profiling with RAG to create digital avatars that can “remember” and adapt with user interactions (Kimara et al., 2025).

Creating a persona is one thing; ensuring it is a faithful representation of the intended character is another. This has led to a variety of evaluation methods.

2.6 Evaluation of LLM-Based Persona Creation

To gauge how well a model represents a persona, researchers have employed various evaluation methods, ranging from automatic metrics to human judgment and psychometric tests. One important aspect is **logical consistency**: the agent should not produce utterances that conflict with the given persona. To measure this, Welleck et al. (2019) introduced the Dialogue Natural Language Inference (DNLI) corpus and associated metrics. DNLI consists of dialogue turns paired with persona sentences, labelled for entailment or contradiction (e.g., given persona: “I have a dog,” does the reply entail or contradict it?). By testing outputs against persona statements using NLI models, one can quantify contradiction rates. This approach revealed that vanilla persona-based models often ignore or contradict persona facts. This led to solutions including *unlikelihood training* and *controlled text generation* that penalize inconsistencies (Li et al., 2020; Kim et al.,

2020). Improved models show lower contradiction rates on DNLI and PersonaChat, indicating better persona fidelity (Kim et al., 2020).

Beyond logical consistency, evaluators examine whether the content and style of the agent’s responses align with the expected persona profile. **Linguistic style matching** is one tool: for example, does a model supposed to embody a high-extraversion persona use more social and positive-emotion words, as real extraverts do? Studies have used resources like the Linguistic Inquiry and Word Count (LIWC) lexicon to analyze generated language for personality markers (Jiang et al., 2024). Jiang et al. (2024) conducted a thorough experiment assigning GPT-4o various Big Five personality profiles and found that the model’s word choices and tone shifted in accordance with the target traits (e.g., “extroverted” personas produced more talkative, upbeat narratives than “introverted” ones). They also had the persona-infused model complete a standard 44-item Big Five Inventory questionnaire in prompt form, and the scores derived from the LLM’s answers correlated strongly with the intended trait levels (Jiang et al., 2024). This suggests that with the appropriate prompting, an LLM can consistently express designated personality traits to a degree measurable by psychological scales.

The work in Shu et al. (2024) cautions that slight variations in how questions are phrased or ordered can lead to inconsistencies in LLM persona questionnaire results, indicating that current prompting strategies may not always capture a model’s “true” persona in a reliable way (Shu et al., 2024). As a complementary approach, **human evaluation** remains crucial. Researchers often ask human annotators to judge if a conversation excerpt “sounds like” it was spoken by the intended persona (e.g., does this really feel like a 45-year-old smoker speaking?). In the PersonaLLM study, humans could correctly identify certain personality traits from a model’s stories at rates far above chance (more than 80% accuracy for clearly manifested traits) (Jiang et al., 2024). Similarly, the recent PersonaGym framework (Samuel et al., 2025) proposes a battery of scenarios where a persona-equipped agent is queried on various tasks (from factual questions to moral dilemmas) and rated (by LLM-based or human evaluators) on whether its responses align with the persona’s expected knowledge, behaviour, and preferences. Such multi-dimensional evaluations aim to go beyond surface traits, testing whether the persona remains consistently applied across different contexts and over extended dialogues.

Known Issues with LLM-Based Persona Creation

Current LLM persona construction techniques face major limitations. One concern is the reinforcement of stereotypes and “flattened” personas. The work in Liu et al. (2024) highlights that when an LLM is asked to embody a persona with incongruous or uncommon trait combinations (for example, a persona that is a political liberal but supports a traditionally conservative policy), the model often defaults to the stereotype—producing opinions more congruent with typical liberals or typical supporters of that policy, rather than faithfully holding the unusual combination of views. They found a nearly 10% drop in steerability for such incongruous personas, with the model sometimes slipping into the demographically expected stance instead of the target stance. This indicates that LLMs have difficulty representing complex, less common identities, tending instead to regress to more stereotypical patterns present in their training data.

Another limitation is maintaining coherence over long interactions. The context window of even modern LLMs (e.g., 4K to 32K tokens) is finite, which means lengthy conversation sessions may push the initial persona prompt out of scope. Without special handling, the model may start deviating from its role after many turns. Memory-augmented strategies as described above only partially mitigate this; errors can accumulate if the retrieval mechanism brings back irrelevant or incomplete persona details. Ensuring that a persona’s

voice and knowledge remain steady over hours of conversation is an open challenge.

There is also the issue that LLM-simulated personas lack a genuine internal life or the ability to experience emotions, which can lead to shallow or inconsistent modelling of complex human traits like empathy, remorse, or motivation. A persona might verbally claim to be “depressed” or “highly motivated,” but the model does not feel these states, potentially yielding dialogue that rings hollow or fails to adapt when tested in emotionally charged situations. [Gupta et al. \(2024\)](#) show that persona assignment can introduce hidden biases in reasoning. For instance, an LLM role-playing as an aggressive character might systematically favour combative responses in moral reasoning tasks, raising concerns that persona-conditioned LLMs could magnify certain biases under the guise of “staying in character.” Ethically, developers must be careful that personas do not become a vehicle for harmful or unfair stereotypes (e.g., a “mentally ill persona” that unintentionally produces stigmatizing language). Transparent documentation of how a persona is constructed and what its limits are is recommended when deploying persona-driven bots in sensitive domains ([Wu et al., 2025](#)).

The ability to create and evaluate personas is a stepping stone towards a more ambitious goal: creating synthetic subjects that can act as stand-ins for humans in behavioural experiments.

2.7 LLM-Based Synthetic Subjects (Doppelgängers) in Behavioural Experiments

The extent to which LLMs can emulate human behaviour has become a central question in computational social science and AI research. One promising approach to assess how closely LLMs behave compared to humans is to replicate human behavioural experiments using LLM-based synthetic agents, each configured with a persona derived from a human participant in a study. We refer to such a synthetic agent as the participant’s *doppelgänger*. In principle, a doppelgänger can substitute for the human subject, enabling faster, safer, and more scalable experimentation. In practice, however, creating effective doppelgängers presents several challenges: (1) persona profiles may lack sufficient richness to capture the full complexity of human behaviour; (2) LLMs may fail to consistently adhere to the installed persona; and (3) the model’s internal knowledge may dominate or distort the intended profile, leading to exaggeration or stereotype perpetuation. As a result, both the construction and rigorous validation of doppelgängers, i.e., scientifically measuring how closely they reproduce the responses of their human counterparts, are important. In the following sections, we examine a range of recent studies that have attempted to create LLM-based doppelgängers in disparate experimental domains (e.g., social, behavioural, clinical) and critique their approach to the creation and validation of such doppelgängers.

LLM Doppelgängers in Social Surveys

[Argyle et al. \(2023\)](#) introduced the concept of “silicon samples” as proxies for human survey respondents. They conditioned GPT-3 with thousands of real participants’ demographic backstories (e.g. age, gender, race, education, political affiliation) from U.S. survey data. They then generated GPT-3’s answers to the same questionnaires people had answered. The validation of these simulated respondents was quantitative: they measured how closely the distribution of answers from the LLM doppelgängers matched the actual human survey distributions across many items and correlations. Notably, the GPT-3 doppelgängers exhibited high

algorithmic fidelity—meaning they reproduced not only overall response proportions but also subgroup differences and attitude inter-correlations present in the human data. For example, “a conservative older male” persona in the model would respond to political questions in line with real conservatives of that demographic, while a “liberal young female” persona’s simulated responses aligned with that group’s patterns.

Argyle et al. (2023)’s validation method was statistical and aggregate. The authors had to correct some skewed marginals in the model’s raw outputs to better align with known sample proportions, indicating that the LLM did not automatically produce a perfectly representative sample without adjustment. More importantly, demonstrating that an LLM can match population-level distributions does not guarantee that any single doppelgänger behaves indistinguishably from its human counterpart in an interactive setting. The questions were mostly closed-ended survey items rather than free-flowing conversations, so it remains unclear whether the same fidelity would hold in open-ended behavioural or linguistic responses. In short, while Argyle et al. (2023) provided evidence that LLMs can emulate group-level attitudes with impressive granularity, their validation method (comparing survey response patterns) overlooks dynamic interaction traits and relies on the assumption that distributional similarity implies a faithful reproduction of individual human behaviour. This leaves a gap when considering applications like therapy dialogues where moment-by-moment language use matters.

2.7.1 Current Validation Methods for LLM Doppelgängers

Although prior works have made important strides in developing LLM-based doppelgängers and inventing ways to validate them, their methods often emphasize either macro-level correspondence (distributions and outcomes) or subjective appraisal. None of the existing studies in the behavioural or health context rigorously validated the *process* of the synthetic subject’s behaviour using domain-specific behavioural metrics. This is where our work seeks to contribute a novel approach. We install richly detailed smoking-related personas into an LLM (drawing on real smokers’ demographics, dependence levels, and motivation scores) and then validate these doppelgängers by analysing their conversational output using Motivational Interviewing (MI) metrics that have known relevance to behaviour change. Specifically, we evaluate the language of the synthetic smokers for its balance of *change talk* vs. *sustain talk* and we examine their pre- and post-conversation *readiness-to-change* ratings. These measures provide an objective, theory-driven way to assess whether the doppelgängers are behaving like real smokers at a given stage of change. For example, a genuine smoker ambivalent about quitting will often produce a low fraction of change talk (more sustain talk defending the status quo) and a low readiness score; our validation checks if the corresponding LLM doppelgänger exhibits a similar profile. By correlating the doppelgängers’ MI-consistent behaviours with those of their human counterparts and comparing distributions of metrics like $C/(C + S)$ talk ratios (change talk fraction) across groups, we can quantitatively gauge fidelity in the very domain of interest.

Unlike Argyle et al. (2023)’s demographic backstories or Aher et al. (2023)’s game outcomes, our validation focuses on conversational indicators of motivation, which are specialized and domain-specific. Similarly, in contrast to the general plausibility ratings used by Cook et al. (2025) and Haider et al. (2025), our approach uses validated psycholinguistic coding (e.g., the MISC scheme for MI dialogue) to measure the doppelgänger’s behaviour against established human benchmarks for “change talk” frequency. We also borrow the concept of a readiness ruler from clinical psychology and ask the synthetic smokers to self-report their readiness to quit before and after a counselling dialogue to see if the LLM’s self-assessment shifts in tandem with the human’s, and if not, where the discrepancies lie.

Table [M.1](#) in appendix [M](#)

Our review of recent literature on the validation of LLM-based doppelgängers (Table [M.1](#) in appendix [M](#)) reveals a lagging rigour in validation methodology. While LLMs demonstrate a remarkable ability to replicate the outcomes of human behaviour across diverse domains, this capability may not be synonymous with replicating the human experience or cognitive process. The recurring limitations identified in this analysis point to a clear agenda for future research. The primary challenge is to move beyond outcome-based validation and develop robust methods for assessing process fidelity. This will require interdisciplinary approaches that integrate techniques from cognitive science, psychology, and linguistics to probe the reasoning, emotional, and social dynamics of agent behaviour. Bridging the gap between aggregate statistical accuracy and individual believability remains a critical open problem, suggesting a need for hybrid validation frameworks. Ultimately, for LLM agents to become truly reliable proxies for human subjects, the field must prioritize the development of validation techniques that are as sophisticated as the models they seek to evaluate.

Conclusion and Future Directions

So far, we have established the foundations for this work, drawing from two distinct but converging areas: the clinical science of Motivational Interviewing and the engineering of Large Language Models. We reviewed MI as an evidence-based method for smoking cessation that guides individuals to articulate their own reasons for change. Its effectiveness is measurable through specific behavioural indicators like the balance of client “change talk” versus “sustain talk” and shifts in self-reported readiness scores. We also traced the development of therapeutic chatbots, from early rule-based systems to modern generative agents that use LLMs, which offer unprecedented conversational fluidity.

The creation of synthetic agents introduces a major challenge: how can we confidently test and validate these systems? The recent practice of using LLMs to create synthetic “doppelgängers” of human participants for behavioural research offers a path forward, but existing validation techniques often lack rigour. Methods that rely on aggregate statistical matching or subjective plausibility ratings do not confirm whether a synthetic subject behaves realistically within the specific interactions of a therapeutic conversation.

This thesis confronts this validation challenge directly. We propose and implement a novel methodology centred on creating MIBot, an MI-based chatbot for smoking cessation, and testing it against a cohort of synthetic smoker doppelgängers. Our primary contribution is a new validation method. Instead of using generic benchmarks, we validate our synthetic smokers using the core, theory-driven metrics of MI itself. This domain-specific validation creates a reliable experimental environment for developing automated therapeutic agents. Having laid this conceptual groundwork, the next chapter moves from theory to practice, detailing the design, development, and deployment of MIBot and the construction of the synthetic smoker personas used in our study.

Chapter 3

Development and Deployment of MIBot

The recent advances in Large Language Models (LLMs) present an opportunity to automate various forms of mental health talk therapy, including Motivational Interviewing (MI) for smoking cessation. This is an important area of need, as over half of all smokers are in an ambivalent state, where they are aware of the harms of smoking but have not yet decided on quitting ([Babb et al., 2017](#)). Guiding these individuals towards a decision to quit is a central precursor for any successful quit attempt ([West and Sohal, 2006](#)).

This chapter details the design and deployment of MIBot, an LLM-based MI chatbot created for this purpose. The work builds on a predecessor system, MIBot v5.2 ([Brown et al., 2023](#)), which, being partially scripted, had limitations in conversational naturalness. The development and evaluation of the current MIBot system are described in detail by [Mahmood et al. \(2025\)](#). This chapter complements that work by focusing on the technologies, design considerations, and implementation choices underlying the chatbot’s deployment for a human feasibility study.

The chapter is structured as follows. Section [3.1](#) outlines the clinician-informed, iterative process used to develop the chatbot’s core prompt. Section [3.2](#) details the system architecture, including the observer agents designed to ensure safety and conversational coherence. Finally, Section [3.3](#) describes the deployment of MIBot to a secured cloud infrastructure for a human feasibility study, including its technical implementation and the containerization of the MIBot service.

3.1 Chatbot Design Process

The design of MIBot was an iterative process that combined expertise in MI with prompt engineering for a modern LLM ([OpenAI, 2024](#)).

3.1.1 Single-Prompt Architecture Rationale

The initial architectural decision for MIBot was to use a single, detailed prompt to define the chatbot’s behaviour. This “start simple” approach is a basic engineering principle, where the goal is to first build and test the most straightforward solution before considering more complex alternatives. In the context of an MI chatbot, a single prompt that encapsulates the counsellor’s entire persona, skills, and decision-making logic

provides a strong baseline for evaluation. If this simple architecture can be shown to be effective, it avoids the premature introduction of more complex systems, such as those involving multiple, adaptively selected prompts or a separate “Behaviour-Change” selector module. The iterative development process described in the following section was therefore focused on improving this single prompt to its maximum potential.

3.1.2 Iterative Prompt Development

The MIBot prompt was improved through structured feedback from both engineers and experienced MI clinicians who regularly met biweekly over the course of development. The process began with a minimal prompt that instructed the model to act as an MI counsellor. This baseline was tested through simulated counselling sessions with two types of test clients:

1. **Virtual smoker clients**, which were separate GPT-4o instances given detailed backstories and instructed to role-play smokers with varying attitudes toward quitting (see Appendix [O](#) for an example transcript). Chapter [6](#) describes in detail the creation of these basic LLM-based virtual smokers and subsequent research to improve their capabilities.
2. **Human role-playing as smokers**—members of our research team who adopted smoker personas and interacted with the chatbot to test how each version of the prompt behaved in different scenarios.

After each testing cycle, transcripts were reviewed in bi-weekly meetings to identify shortcomings in the appropriate use of MI skills, adherence to MI principles, tone, pacing, and client engagement. These findings informed successive prompt revisions. This process led to several central prompt revisions:

1. **Utterance Length Control**: Early versions tended toward long, paragraph-like responses, which risked dominating the conversation, a practice antithetical to MI, in which the client leads the process of contemplation. The prompt was amended to include explicit length constraints (“Keep your responses short. Do not talk more than your client.”) and to encourage brevity while maintaining reflective depth.
2. **Accessible Language**: To ensure inclusivity across educational and socioeconomic backgrounds, clinicians requested avoidance of jargon and adaptation to the client’s linguistic style. This was codified in the prompt as “Avoid using complex terminology... maintain simplicity in the conversation.”.
3. **Avoidance of Assumptions**: The model occasionally assumed the client’s nicotine consumption patterns. The prompt was revised to explicitly instruct the counsellor that “You don’t know anything about the client’s nicotine use yet” to preserve an open, inquiring stance.
4. **Rapport-Building Before Smoking Focus**: Initial versions of the prompt engaged with smoking behaviour too early, bypassing engagement and focusing stages. The revised prompt added guidance to “open the conversation with a general greeting and friendly interaction” before gradually steering toward smoking ambivalence.
5. **Guarding Against Premature Planning**: Planning is an MI process best introduced after sufficient evocation of Change Talk. The prompt included multi-step criteria for initiating planning, explicitly instructing the model to wait for reduced Sustain Talk and to confirm readiness before launching into planning and discussing concrete steps for quitting.

This iterative process continued until virtual and role-played conversations consistently met MI quality

expectations as determined by an informal consensus of the team.

Tables 3.1 and 3.2 show how the prompt developed from a simple instruction about the LLM's role to a detailed guideline addressing issues identified in the chatbot's performance. The full prompts are provided in Appendix A.

- 1 You are a skilled motivational interviewing counsellor.
- 2 Your job is to help smokers resolve their ambivalence towards smoking using motivational interviewing skills at your disposal.
- 3 Your next client is {client_name}. Start the conversation by greeting {client_name}.

Table 3.1: The initial system prompt used for MIBot. This version of the prompt is very simple and only instructs the model to act as an MI counsellor and greet the client.

- 1 You are a skilled motivational interviewing counsellor. Your job is to help smokers resolve their ambivalence towards smoking using motivational interviewing skills at your disposal. Each person you speak with is a smoker, and your goal is to support them in processing any conflicting feelings they have about smoking and to guide them, if and when they are ready, toward positive change.
- 2 Here are a few things to keep in mind:
 1. Try to provide complex reflections to your client.
 2. Do not try to provide advice without permission.
 3. Keep your responses short. Do not talk more than your client.
 4. Demonstrate empathy. When a client shares a major recent event, express genuine interest and support. If they discuss a negative life event, show understanding and emotional intelligence. Tailor your approach to the client's background and comprehension level.
 5. Avoid using complex terminology that might be difficult for them to understand, and maintain simplicity in the conversation.
- 3 Remember that this conversation is meant for your client, so give them a chance to talk more.
- 4 This is your first conversation with the client. Your assistant role is the counsellor, and the user's role is the client.
- 5 You have already introduced yourself and the client has consented to the therapy session.
- 6 You don't know anything about the client's nicotine use yet.
- 7 Open the conversation with a general greeting and friendly interaction, and gradually lead the conversation towards helping the client examine ambivalence around smoking, using your skills in Motivational Interviewing.
- 8 You should never use prepositional phrases like "It sounds like," "It feels like," "It seems like," etc.
- 9 Make sure the client has plenty of time to express their thoughts about change before moving to planning. Keep the pace slow and natural. Don't rush into planning too early.
- 10 When you think the client might be ready for planning:
 1. First, ask the client if there is anything else they want to talk about.
 2. Then, summarize what has been discussed so far, focusing on the important things the client has shared.
 3. Finally, ask the client's permission before starting to talk about planning.
- 11 Follow the guidance from Miller and Rollnick's "Motivational Interviewing: Helping People Change and Grow," which emphasizes that pushing into the planning stage too early can disrupt progress made during the engagement, focusing, and evoking stages.
- 12 If you notice signs of defensiveness or hesitation, return to evoking, or even re-engage the client to ensure comfort and readiness.
- 13 Look for signs that the client might be ready for planning, like:
 1. An increase in change talk.
 2. Discussions about taking concrete steps toward change.
 3. A reduction in sustain talk (arguments for maintaining the status quo).
 4. Envisioning statements where the client considers what making a change would look like.
 5. Questions from the client about the change process or next steps.

Table 3.2: The final system prompt for MIBot, developed through an iterative, clinician-informed process. This version includes detailed instructions on MI skills, tone, pacing, and specific guidelines for handling different conversational stages, such as when to initiate planning.

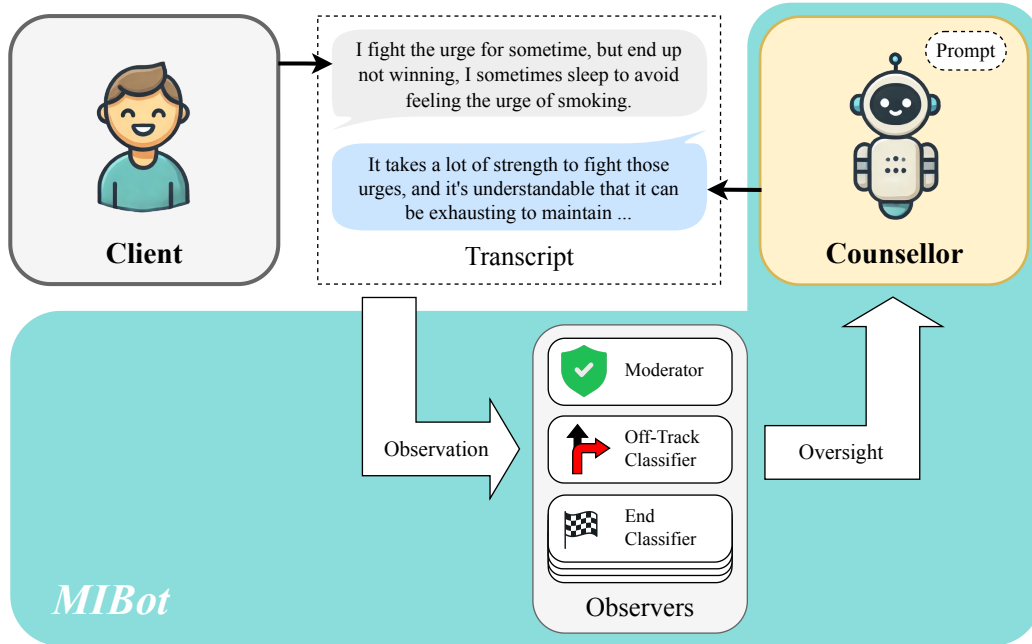


Figure 3.1: Overview of the MIBot system, which includes a core counsellor agent and a set of observer agents for safety and coherence. The system interacts with the user, and the observers monitor the conversation for harm, off-track dialogue, and conversation termination cues. Taken from [Mahmood et al. \(2025\)](#).

3.2 Observers

To improve safety in the deployment of MIBot, the core counsellor agent was supplemented with a set of *observer agents*, independent instances of GPT-4o prompted to monitor specific aspects of the conversation in real time. The output of these agents is used to intervene when necessary. Each observer was specialized through prompt engineering to perform a specific task in real time (see Appendix B for the full prompts).

3.2.1 Moderator

The *Moderator* evaluates the counsellor’s most recent utterance for potential harm before it is displayed to the client. While OpenAI’s internal safety systems reduce many risks, they do not address all possible counterproductive counselling behaviours, such as inadvertently reinforcing *sustain talk* or suggesting self-harm. The Moderator was intentionally configured for high sensitivity, accepting a higher false positive rate to reduce the risk of harmful or counterproductive content. If a counsellor’s utterance is flagged, it is regenerated and re-evaluated, with up to five regeneration attempts permitted. In all study conversations, an acceptable utterance was produced within four attempts, and no session failed to pass moderation.

3.2.2 Off-Track Conversation Classifier

The *Off-Track Classifier* detects when a client is steering the dialogue away from smoking cessation in a deliberate or sustained manner. Its prompt was tuned for low false positive rates to preserve conversational flexibility. In the feasibility study described in Chapter 4, this observer’s primary role was retrospective — identifying conversations for exclusion where the participant was not engaging seriously with the in-

tervention. In a live deployment, it can be used to trigger early termination or redirection to the main topic.

3.2.3 End Classifier & Termination Process

The *End Classifier* monitors both parties' dialogue to determine if the conversation is reaching a natural conclusion. It prioritizes the client's intent when making this determination, ensuring the conversation is not ended prematurely. Upon detecting an intent to close, it instructs the counsellor to deliver a concise summary of main discussion points — a standard MI practice — and to confirm with the client whether they wish to continue. If the client declines, the conversation is terminated and any post-session procedures, such as surveys, are initiated.

Design Rationale: All observers were implemented as separate, stateless LLM calls, each with prompts tailored to their decision criteria. This modular approach allowed independent improvement of their sensitivity–specificity balance without affecting the primary counsellor prompt. The Moderator favoured recall over precision to err on the side of client safety, whereas the Off-Track Classifier did the opposite, favouring conversational autonomy. The End Classifier's logic explicitly distinguished between topic changes and true conversation endings, reducing false terminations.

The prompted GPT-4o, together with observers, constitute the complete MIBot system, as illustrated in Figure 3.1.

3.3 MIBot System Design

3.3.1 Overview of the Application

MIBot is implemented as a containerized software system that can be run locally for development and deployed to cloud-based systems. In this section, we discuss the implementation details of MIBot, from the structure of the Python microservice and its integration with the OpenAI API to the containerization and the deployment of the service on Amazon Web Services.

The core of MIBot is a lightweight Python web application built with the Sanic framework ([Sanic Community Organization, 2025](#)). In MIBot, `app.py`, the main entry point, uses Sanic to configure routes for all external interactions and instantiates the conversation engine. An overview of the application's architecture is presented in Figure 3.2 (see Appendix J for more detailed diagrams). When a user sends a request to the `/chat` endpoint containing their message and `client_id`, the web server updates the state of the conversation and requests the next turn from it. The `Conversation` object relays this to the `Counsellor`, which in turn sends a request to the OpenAI API with the accumulated conversation history and the current client message, and receives a response containing the generated counsellor turn.

Each `Observer` attached to the `Conversation` inherits from a base class defining an asynchronous `observe()` method. As noted earlier, the `Moderator` observer screens counsellor utterances for safety and appropriateness; the `Off-Track Classifier` observer assesses whether the client is steering the conversation away from smoking cessation; and the `End Classifier` determines when a session should conclude. These observers are implemented as separate GPT-4o API calls with their own prompts. After each turn, the `Conversation` object iterates over all `Observer` objects, collects their observations, and makes real-time

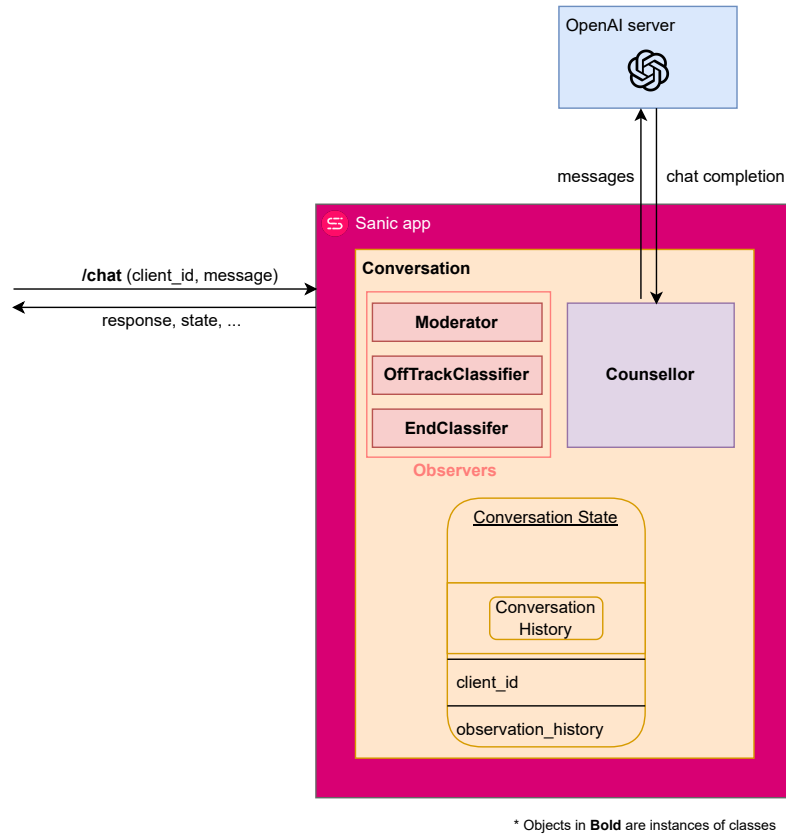


Figure 3.2: An overview of the Sanic web application that encapsulates the MIBot code. The application exposes REST APIs for external interactions, such as handling chat messages and providing conversation transcripts.

decisions (e.g., whether to end the conversation) before updating its state. If the generated output from the Counsellor is deemed suitable for the client, it is sent to the client along with relevant metadata.

A single Sanic app can handle multiple clients at once by creating replicas of the Conversation object, each uniquely identified by the `client_id`.¹ The microservice also exposes some additional endpoints: `/get_transcript` provides a downloadable transcript of the conversation for post-session analysis; `/health` returns a simple 200 OK response so that load balancers and orchestrators can perform health checks; and `/info` exposes build metadata such as the current version of the prompt.

3.3.2 Containerization

For reproducibility and ease of deployment, the microservice is packaged in a Docker container. Containerization provides several advantages for the development, testing, and deployment of MIBot, including environment consistency, portability, isolation, scalability, faster deployment and rollbacks, simplified CI/CD integration, and reproducibility, as discussed in detail by Sloane (Sloane, 2025). An overview of the containerized application is presented in Figure 3.3.

The MIBot container image is defined by a Dockerfile (see Appendix I) specifying the base Python runtime,

¹For the human feasibility study, to keep track of the participants and their conversations, we explicitly use `prolific_id` as `client_id`. Prolific (www.prolific.com) is the platform we use to recruit participants and conduct our feasibility study. See Chapter 4 for further details.

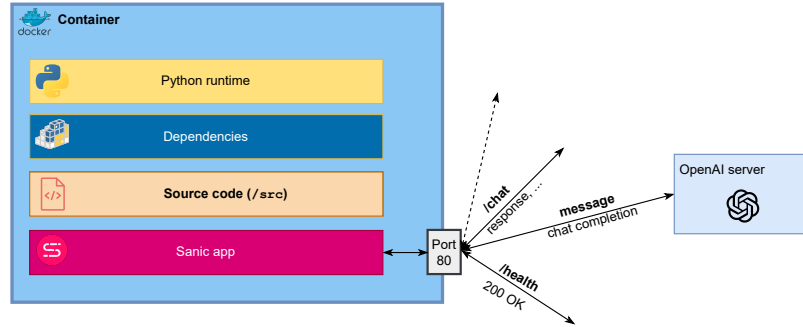


Figure 3.3: The containerized Sanic application for MIBot. The application is packaged in a Docker container for reproducibility and ease of deployment, with the container image stored in Amazon Elastic Container Registry (ECR).

required dependencies, the source code, and main entry point (viz. `app.py`). This image is stored in Amazon Elastic Container Registry (ECR). The container registry stores all the images built by the CI/CD pipeline, but only the image with the production tag is used for deployment.

3.4 Deploying MIBot to AWS

MIBot is deployed as a service on Amazon **Elastic Container Service (ECS)**. ECS is a fully managed service provided by Amazon Web Services (AWS) that “makes it easier to deploy, manage, and scale applications using containers” (Amazon Web Services, 2025a). We now discuss each component of ECS.

3.4.1 Components of ECS

An overview of the deployment architecture is presented in Figure 3.4.

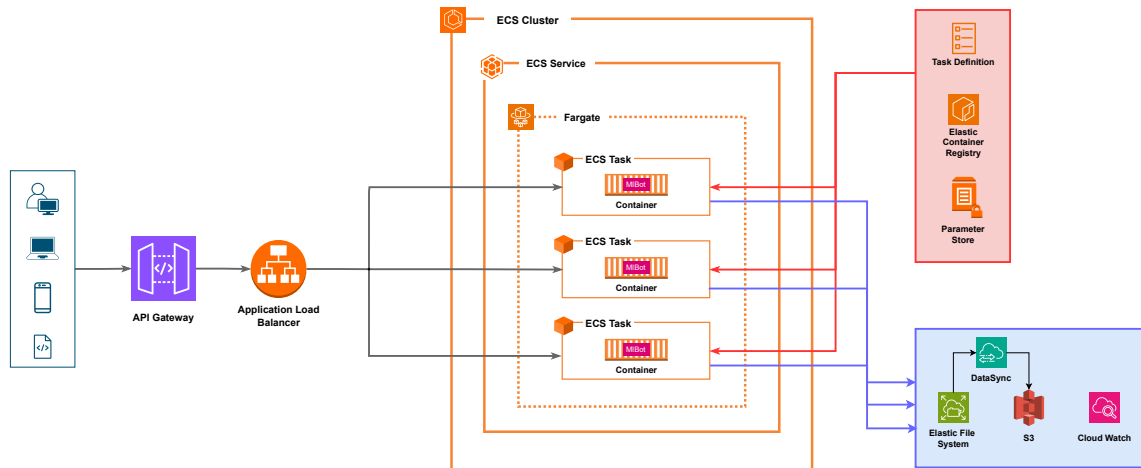


Figure 3.4: An overview of the deployment of the containerized MIBot application on Amazon Elastic Container Service (ECS). The diagram shows the key components, including the ECS cluster, the ECS service, and the ECS task definition, as well as the interaction with other AWS services like Elastic Load Balancer (ELB) and API Gateway.

1. ECS Cluster to deploy MIBot to ECS, we first provisioned an **ECS cluster** (`mibot-v6-cluster`). An ECS cluster is a *logical* grouping of heterogeneous compute resources. In the context of AWS, a **compute resource**

is any AWS-managed infrastructure component that provides processing power for running applications or workloads. These resources can be **user-managed**, like Elastic Compute Cloud (EC2), where users rent virtual machines, or **fully managed**, like AWS Fargate, which is used to run containerized applications without direct server management. A cluster is therefore a logical grouping of such compute resources.

In our ECS cluster, however, we only used **AWS Fargate** as the computing resource. In addition to allowing for the deployment of the containerized MIBot application without server management, AWS Fargate also offers *spot runs* for cost optimization. In the FARGATE_SPOT mode, tasks run on spare compute capacity. If a container receives no traffic in the last two hours and AWS reclaims the capacity, the task will be terminated. ECS will detect this event and will almost immediately instantiate a new task for the application.

2. ECS Service Inside the ECS cluster, we created an **ECS Service** (`mibot-v6-service`). The ECS Service contains the deployment configuration of the application, for example, the number of desired replicas of the application (also called *tasks*) that should run at any given time. For our study, we set this to use two tasks. If one of the tasks fails, the ECS Service replaces it automatically. It can also be configured to increase the number of tasks when it detects higher-than-normal traffic. The Service is connected to an Elastic Load Balancer (ELB) to distribute incoming traffic evenly among tasks. It also defines deployment (rolling update, blue/green deployment) and rollback strategies, and can use the AWS circuit breaker to roll back failed deployments automatically.

3. ECS Task Definition The final component in the deployment of MIBot is defining a **Task**. A Task runs a specific container after downloading it from the Elastic Container Registry (ECR). The Task definition specifies environment variables and API credentials that are securely stored in AWS Systems Manager Parameter Store and are injected into the container (e.g., `OPENAI_API_KEY`) when the task is started. It further defines a *health check* for the container. The health check sends a request to the `/health` endpoint on the container's port 80 every five minutes. If the response is anything other than `200 OK` or it does not get a response within one minute, it deems the container unhealthy. The Service terminates the Task and replaces it with a new one. The task definition further specifies the required CPU and memory (1024 CPU units (1 vCPU) and 4 GB, respectively, in the case of MIBot). The containers also mount a persistent Amazon Elastic File System (EFS) volume to store conversation transcripts and evaluation metrics. Furthermore, all container logs are written to AWS CloudWatch for retrospective analysis of the system's behaviour.

3.4.2 Other Components of the Deployment

Load Balancer We configured an Elastic Load Balancer (`mibot-elb`) with three subnets for high availability, with three instances of load balancers in three different Availability Zones. The load balancers are *application* load balancers (ALB) ([Amazon Web Services, 2025b](#)), which are internet-facing and associated with a security group permitting inbound traffic on port 443. The DNS names allow external users to connect to the service through a custom domain name ([Shopify, 2025](#)). The ALB routes incoming HTTP requests to the ECS service's target group and internal health check requests to the `/health` endpoint.

API Gateway We also provisioned an AWS API Gateway, which acts as a reverse proxy and enables secure TLS termination and request throttling. The gateway exposes HTTPS endpoints for `/chat`, `/get_transcript`, `/health`, `/info`, and `/s3_upload`. Each path includes `OPTIONS` methods to enable cross-origin requests and

defines the expected response headers. This prevents client browsers from blocking requests due to Cross-Origin Resource Sharing (CORS) policies or displaying security warnings.

DataSync Conversation transcripts and metadata are stored on both an encrypted EFS volume and AWS S3 only when the participant clicks on the final Submit button at the end of the study session. EFS acts as a redundant data layer in case the upload to S3 fails. For eventual consistency, we periodically copy files from EFS to S3 using AWS DataSync, which runs a daily CRON job.

3.4.3 Deployment Pipeline

We adopted DevOps practices for automated deployment. Every time we push a special production Git tag to the remote main branch, a GitHub workflow builds the Docker image, runs unit tests, and, if tests pass, pushes the image to Amazon ECR. Another workflow triggers a CloudFormation deployment that updates the ECS task definition with the new image tag and performs a rolling deployment of the ECS Service. Deployment uses a circuit breaker configuration: if the service fails health checks, the rollout is automatically rolled back to the previous stable revision. By integrating the deployment pipeline into version control, we ensured automated deployments tied to code changes.

Chapter 4

Evaluation of MIBot: Design of the Human Feasibility Study

The deployed MIBot system was used in a feasibility study with human smokers. The goal of this study was to determine the effect of a single-session interaction with MIBot and assess its safety for delivery to ambivalent smokers in a real-world setting.

4.1 Ethics Approval and Consent

The protocol was approved by the University of Toronto Research Ethics Board (Protocol Number 49997; approved August 3, 2024) (Rose et al., 2025) and adhered to all institutional guidelines. Before participating in the study, prospective participants reviewed an online consent form that outlined study aims, procedures, compensation, potential risks, and data handling practices (excerpts from the consent form are available in Appendix L). Participation required explicit electronic consent. Risks were described as minimal but included the possibility that discussing smoking could cause stress or temporarily increase cravings. No personally identifying information was collected, and all data were de-identified prior to release.

4.2 Participant Recruitment

Participants were recruited via Prolific,¹ an online behavioural research platform with pre-screened participant pools and built-in demographic filters. Prolific was chosen for its ability to target recruitment to specific smoking status, demographic characteristics, and quality-control thresholds, and for its established use in prior MI chatbot studies (Brown et al., 2023; Almusharraf et al., 2020). Participants received £5.50 for the main session and £1.00 for the follow-up survey, exceeding Prolific’s recommended hourly rates.

¹<https://www.prolific.com>.

4.3 Initial Screening Criteria

Eligibility screening was implemented at two levels: (1) Prolific pre-screen filters, applied before invitation to the study; and (2) an in-study screening step prior to chatbot interaction. The first set of filters required that all invitees be 18 years or older, be fluent in English, have an approval rate of at least 90% on prior Prolific studies, and self-identify as a *current smoker* of at least five cigarettes per day, with a history of smoking at this rate for at least one year.

In addition, recruitment was set up to aim for a nearly equal sex balance. Although the final sample reflected slight deviations due to subsequent exclusion filtering, this pre-allocation ensured coverage across male and female participants.

4.4 In-Study Screening

Upon accessing the study website, participants completed a smoking status confirmation question identical to Prolific’s pre-screening question. From an initial pool of 159 participants, we screened for ambivalence using the Readiness Ruler, as described in Section 4.7.1. To ensure participants were in a state where MI could be beneficial, we included those with a pre-conversation *confidence-to-quit* score of 5 or lower on a 10-point scale. We also included *discordant* participants: those who, despite having high confidence (a score greater than 5), rated the importance of quitting at least five points lower than their confidence. This screening process resulted in a final sample of 106 participants.

Participants who met all criteria and provided informed consent proceeded to the survey phase. Those who did not meet the eligibility criteria were redirected to the Prolific platform without completing the study.

4.5 Participant Characteristics

Baseline demographics of enrolled participants are summarized in Table 4.1. All were English-speaking adults who self-identified as current daily smokers and passed pre-screening and in-study eligibility checks on Prolific. The sample was approximately balanced by sex and covered a broad range of ages. Unless otherwise specified, values are presented as counts (n) and percentages (%). Age is reported as median, mean (SD), and range.

4.6 Study Procedure

The full study procedure comprised four major phases: (1) pre-conversation surveys; (2) a single text-based conversation with MIBot; (3) immediate post-conversation surveys; and (4) a one-week follow-up survey. Figure 4.1 illustrates the progression.

Phase 1: Pre-Conversation Surveys Before interaction with MIBot, participants completed:

1. **Heaviness of Smoking Index (HSI)** (Heatherton et al., 1989) survey, which assesses nicotine dependence with two questions:
 - (a) number of cigarettes smoked per day (scored 0–3), and

Characteristic	n (%)
Total participants	106
Sex	
Female	57 (53.8)
Male	49 (46.2)
Language	
English-speaking	106
Age summary	Range 22–77; median 38; mean 40 (SD 13)
Age groups (years)	
Below 20	0 (0.0)
20 to 29	26 (24.5)
30 to 39	32 (30.2)
40 to 49	20 (18.9)
50 to 59	19 (17.9)
60 to 69	6 (5.7)
70 to 79	3 (2.8)
Above 79	0 (0.0)
Ethnicity	
White	80 (75.5)
Black	9 (8.5)
Asian	7 (6.6)
Mixed	5 (4.7)
Other	5 (4.7)
Student status	
No	80 (75.5)
Yes	21 (19.8)
Data expired	5 (4.7)
Employment status	
Full-time	49 (46.2)
Part-time	18 (17.0)
Not in paid work	16 (15.1)
Unemployed	13 (12.3)
Other	10 (9.4)
Country of residence	
United Kingdom	47 (44.3)
United States	42 (39.6)
Canada	9 (8.5)
South Africa	4 (3.8)
Other	4 (3.8)
Country of birth	
United Kingdom	44 (41.5)
United States	39 (36.8)
Canada	6 (5.7)
Kenya	3 (2.8)
South Africa	3 (2.8)
Germany	2 (1.9)
Other	9 (8.5)

Table 4.1: Baseline characteristics of the 106 enrolled participants in the feasibility study. The table includes demographic information such as sex, age, ethnicity, student and employment status, and country of residence and birth.

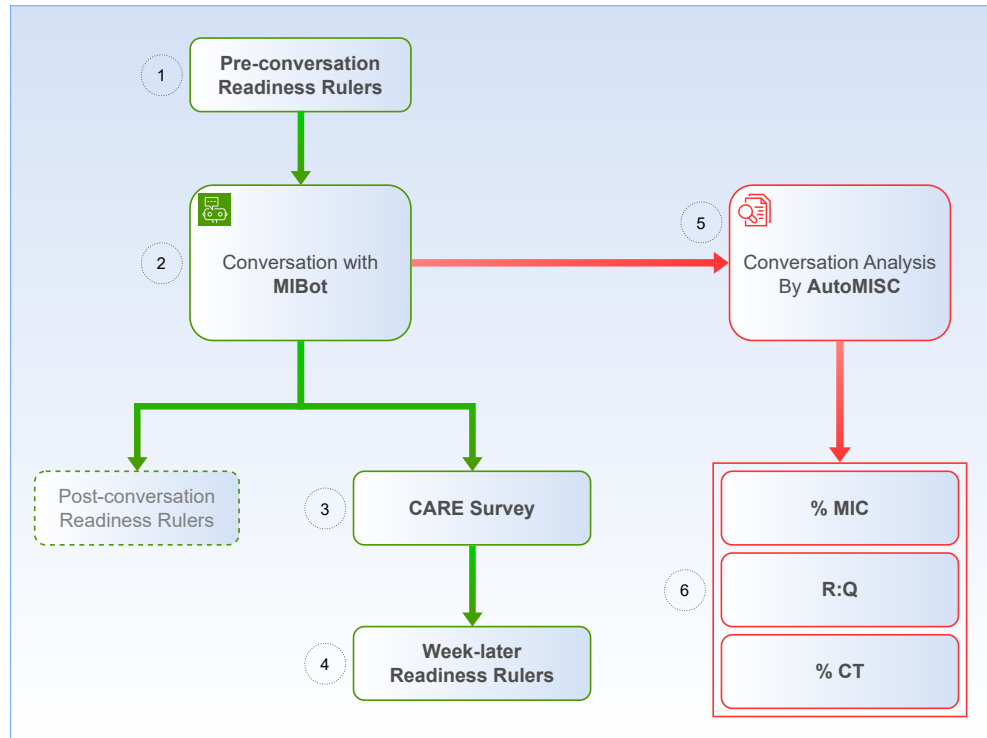


Figure 4.1: Overview of the feasibility study protocol, illustrating the four main phases: pre-conversation surveys, a single text-based conversation with MIBot, immediate post-conversation surveys, and a one-week follow-up survey.

(b) time to first cigarette after waking (scored 0–3).

The HSI score is the sum of these two items, ranging from 0 to 6, with higher scores indicating greater nicotine dependence.

2. **Quit Attempts in the Past Week:** the number of conscious attempts to abstain from smoking for at least 24 hours during the preceding seven days.
3. **Readiness Ruler:** three questions measuring self-reported ratings of importance, confidence, and readiness to quit on 0–10 scales (Section 4.7.1).

Phase 2: Conversation with MIBot Participants engaged in a single MI-style conversation via a web-based text interface.

Phase 3: Post-Conversation Surveys Immediately after the conversation, participants repeated the Readiness Ruler, completed the CARE empathy scale (Section 4.7.2), and provided qualitative feedback on the chatbot’s performance.

Phase 4: One-Week Follow-Up Seven days later, participants were invited via Prolific to complete a follow-up survey. This included a third administration of the Readiness Ruler and questions about quit attempts and changes in smoking behaviour over the past week.

4.7 Survey Instruments

The goal of the work is to move ambivalent smokers towards the decision to quit. We employed the following metrics to measure outcomes from the interaction. In addition to the primary outcome measures, we also used other survey instruments to obtain a thorough view of the chatbot's effectiveness.

4.7.1 Readiness Ruler

The Readiness Ruler ([Rollnick et al., 1992](#)) is a validated tool for assessing motivational state across three dimensions:

- **Importance:** "How important is it to you right now to stop smoking?"
- **Confidence:** "How confident are you that you would succeed at stopping smoking if you started now?"
- **Readiness:** "How ready are you to start making a change at stopping smoking right now?"

Responses were recorded on an 11-point scale (0 = "not at all", 10 = "extremely"). The week-later change in *confidence* from the pre-conversation value was used as the primary metric for the chatbot's effectiveness, as it is the most predictive of downstream quitting success ([Gwaltney et al., 2009](#); [Abar et al., 2013](#)).

4.7.2 CARE Measure

The Consultation and Relational Empathy (CARE) measure ([Mercer et al., 2004](#); [Bikker et al., 2015](#)) assesses perceived empathy in clinical encounters. Ten items evaluate the counsellor's ability to make the participant feel at ease, listen actively, appreciate the participant as a whole person, and collaborate on problem-solving. Each question is rated on a 1–5 scale (from "poor" to "excellent"), with a total score range of 10–50.

4.7.3 Qualitative Feedback

Three open-ended questions solicited subjective impressions:

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

These responses can inform future prompt improvements and provide contextual data for interpreting quantitative outcomes.

4.7.4 Follow-Up Quit Attempt Survey

One week later, apart from self-reporting their readiness rulers, participants also reported whether they had made any quit attempts in the preceding seven days, the number of attempts, and whether any changes in smoking habits had occurred. This included partial changes, such as a reduction in cigarettes per day.

4.8 Automated Conversation Analysis

In addition to participant-reported outcomes, conversations were analyzed for MI adherence and elicitation of motivational language using AutoMISC, an automated implementation of the Motivational Interviewing Skill Code v2.5 (Houck et al., 2010a). The AutoMISC system, primarily developed by Ali (2025) and described in that thesis, was further validated in (Ali et al., 2025). For this project, we used the existing AutoMISC system and contributed to the work described in (Mahmood et al., 2025). The analysis pipeline first segments each conversational volley into individual utterances. Then, it classifies counsellor utterances as MI-Consistent, MI-Inconsistent, Reflection, Question, or Other. Similarly, it classifies each client utterance as exhibiting Change Talk, Sustain Talk, or Neutral. Finally, it computes the following: %MI-Consistent Responses (%MIC), Reflection-to-Question Ratio (R:Q), %Client Change Talk (%CT). AutoMISC was validated against human coders, including two expert MI clinicians (Mahmood et al., 2025).

Readiness rulers (especially the week-later change in *confidence*), CARE, and AutoMISC summary metrics together provide an overall understanding of the MIBot intervention, assessing its effectiveness, perceived empathy, safety, and adherence to MI principles. In the next chapter, we report the results of our human feasibility study on these metrics.

Chapter 5

Evaluation of MIBot: Results from the Human Feasibility Study

This chapter presents a thorough evaluation of MIBot’s effectiveness through a feasibility study with 106 smokers. Building on the system design and implementation described in previous chapters, we assess MIBot’s performance across four critical dimensions established in the smoking cessation and motivational interviewing literature: behavioural change readiness, perceived therapeutic empathy, adherence to MI principles, and elicitation of client change talk.

The chapter is organized to progress from primary outcomes through measurements of therapeutic process, and finally, to behavioural changes and user experiences. First, we report the primary outcome of changes in readiness to quit smoking (Section 5.1). We then analyse how the chatbot performed on perceived empathy—measured using the CARE scale—and compare its performance to that of human healthcare professionals (Section 5.2). Next, we examine MIBot’s adherence to MI principles through AutoMISC analysis and examine whether MIBot could maintain fidelity to therapeutic standards while successfully eliciting change talk from clients (Section 5.4).

Following these core metrics, we investigate behavioural outcomes, including quit attempts and self-reported changes; examine conversation flow, both quantitatively and qualitatively (Section 5.6); and examine illustrative case studies and sample outliers. Finally, we analyse participant feedback (Section 5.7) and discuss broader implications of our findings (Section 5.8).

5.1 Primary Outcome: Readiness to Quit

Overall Changes in Readiness Rulers

Table 5.1 summarizes the mean (and standard deviation) of each readiness ruler before the conversation, immediately afterwards, and one week later. The table includes the change between the pre-conversation metric and one week later (Δ). The Wilcoxon signed-rank test was applied to assess the significance of the change. Participants’ confidence increased markedly from a baseline mean of 2.8 to 4.5 one week later

($\Delta = 1.66, p < 10^{-9}$). This represents an approximately 59% relative improvement and constitutes our primary outcome measure, aligning with MI theory that confidence (self-efficacy) is a major predictor of behaviour change (Gwaltney et al., 2009; Abar et al., 2013). Importance also increased, albeit more modestly ($\Delta = 0.5, p < 0.005$), while readiness exhibited a small, non-significant change ($\Delta = 0.23, p = 0.22$).

Ruler	Before mean (SD)	After mean (SD)	One Week mean (SD)	Δ mean (SD)
Importance	5.7 (2.6)	6.3 (2.9)	6.1 (2.7)	0.5 (1.7)**
Confidence	2.8 (2.0)	4.6 (2.6)	4.5 (2.7)	1.7 (2.4)***
Readiness	5.2 (2.8)	5.9 (2.8)	5.5 (3.0)	0.3 (2.4)

Table 5.1: Means and standard deviations (SD) of readiness rulers (0–10 scale) for importance, confidence, and readiness, measured at three time points: before the conversation, immediately after, and one week later. The table also shows the mean change (Δ) between the pre-conversation measurement and the one-week follow-up. Statistical significance of the change is assessed using the Wilcoxon signed-rank test (*** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$).

Figure 5.1 illustrates the distribution of one-week changes in confidence. Of the 106 participants, 64 (60.4%) showed improvement, 21 (19.8%) remained unchanged, and 21 (19.8%) experienced a decrease. The median change was +1.0 point, with the interquartile range spanning 0 to +3 points. Notably, 15 participants (14.2%) achieved gains of 5 or more points, representing substantial movement toward greater quitting confidence.

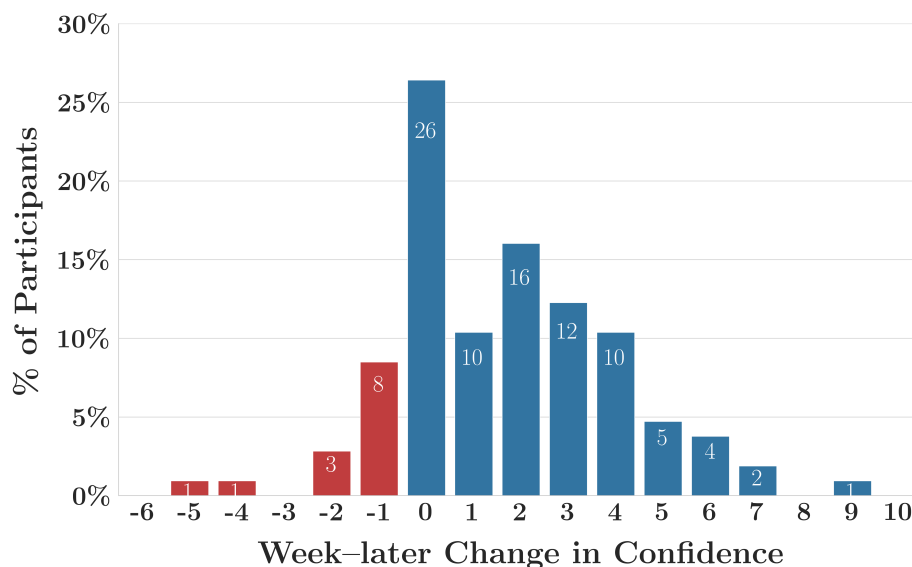


Figure 5.1: Distribution of one-week changes in confidence scores (post-conversation minus pre-conversation). The x-axis represents the change in confidence on a 0–10 scale, and the y-axis represents the number of participants. The majority of participants (60.4%) showed an improvement in confidence.

Stratified Analysis by Baseline Characteristics

To understand which participants benefited most from MIBot, we stratified outcomes by baseline characteristics. Table 5.2 shows that those starting with the lowest self-confidence ($n = 31$, confidence ≤ 1) experienced the largest improvement (+2.2 points), whereas participants with moderate or higher confidence gained approximately one and a half points. The three participants who began with very high confidence showed a decline, likely reflecting regression to the mean. Baseline confidence correlated negatively with the change

(Spearman $r = -0.21$, $p < 0.05$), indicating that MIBot is most beneficial for participants who are least confident in their ability to quit.

Baseline confidence range	Sample size	Baseline confidence mean (SD)	Post-conversation mean	1-week follow-up mean (SD)	Change from baseline mean (SD)
0–1	31	0.5 (0.5)	2.7	2.7 (2.5)	+2.2 (2.4)***
2–3	32	2.5 (0.5)	2.0	4.2 (2.6)	+1.7 (2.5)**
4–5	40	4.3 (0.5)	1.4	5.8 (2.0)	+1.5 (2.1)***
≥6	3	9.3 (0.6)	0.6	7.7 (3.2)	–1.7 (2.9)

Table 5.2: Longitudinal changes in self-reported confidence scores (0–10 scale), stratified by baseline confidence level. The table shows the mean confidence scores at baseline, immediately post-conversation, and at 1-week follow-up, for different subgroups of participants based on their initial confidence. The change from baseline to 1-week follow-up is also presented. Statistical significance is determined by the Wilcoxon signed-rank test (*** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$).

Confidence changes varied across smoking characteristics and quit history profiles. Participants were grouped by HSI (low 0–1, moderate 2–3, high 4–5, very high ≥6), daily cigarette consumption (<5, 5–9, 10–19, ≥20), whether they had made a quit attempt in the week before the study, and the number of prior attempts (0, 1–2, ≥3). The mean change in confidence for each subgroup is summarized in Table 5.3.

Participant characteristics	Sample size	Baseline confidence mean (SD)	Post-conversation mean (SD)	1-week follow-up mean (SD)	Change from baseline mean (SD)
<i>Heaviness of Smoking Index</i>					
Low (0–1)	28	3.5 (1.9)	5.5 (2.3)	4.9 (2.5)	+1.5 (2.7)**
Moderate (2–3)	55	2.7 (2.0)	4.1 (2.6)	4.7 (2.9)	+2.0 (2.3)***
High (4–5)	21	2.5 (1.8)	4.5 (2.3)	3.3 (2.4)	+0.8 (2.1)
Very high (≥6)	2	1.5 (2.1)	4.0 (5.7)	5.0 (2.8)	+3.5 (0.7)
<i>Daily cigarette consumption</i>					
<5	5	3.2 (1.3)	5.2 (2.0)	4.2 (2.6)	+1.0 (1.6)
5–9	32	3.5 (2.5)	5.2 (2.7)	5.2 (3.0)	+1.7 (2.9)**
10–19	38	2.8 (1.5)	4.2 (2.5)	4.5 (2.8)	+1.7 (2.1)***
≥20	31	2.1 (1.6)	4.2 (2.5)	3.7 (2.3)	+1.7 (2.3)***
<i>Pre-conversation quit attempt</i>					
Yes	34	3.1 (1.6)	5.1 (2.7)	5.7 (2.6)	+2.6 (2.2)***
No	72	2.7 (2.1)	4.3 (2.5)	3.9 (2.6)	+1.2 (2.3)***
<i>Number of prior attempts</i>					
0	72	2.7 (2.1)	4.3 (2.5)	3.9 (2.6)	+1.2 (2.3)***
1–2	16	3.3 (1.6)	6.1 (2.5)	6.9 (2.4)	+3.6 (2.0)***
≥3	18	2.8 (1.7)	4.3 (2.6)	4.7 (2.4)	+1.8 (2.0)**

Table 5.3: Longitudinal changes in quit confidence scores (0–10 scale), stratified by baseline smoking characteristics and quit history. The table shows mean confidence scores at baseline, post-conversation, and 1-week follow-up for subgroups based on Heaviness of Smoking Index (HSI), daily cigarette consumption, pre-conversation quit attempts, and number of prior attempts. The change from baseline is also shown. Statistical significance is determined by the Wilcoxon signed-rank test.

Participants with moderate nicotine dependence (HSI 2–3) showed the greatest gains (+2.0), compared with those with high dependence (+0.8). Daily consumption showed little systematic difference across groups.

Larger improvements were observed among participants reporting a conscious quit attempt in the week before the conversation ($n = 34$), who showed confidence increases of +2.6, compared with +1.2 among those with no recent attempts. Among those with one or two previous attempts ($n = 16$), the improvement was +3.6. These patterns suggest that participants who were already contemplating change benefited the most.

Demographic Patterns

Demographic characteristics	Sample size	Baseline confidence mean (SD)	Post-conversation mean (SD)	1-week follow-up mean (SD)	Change from baseline mean (SD)
<i>Sex</i>					
Female	57	2.5 (2.1)	4.4 (2.8)	4.1 (2.9)	+1.7 (2.5)***
Male	49	3.2 (1.7)	4.7 (2.2)	4.9 (2.5)	+1.7 (2.3)***
<i>Age</i>					
< 30 years	26	3.7 (2.1)	5.5 (2.5)	5.7 (2.7)	+1.9 (3.1)*
≥ 30 years	80	2.5 (1.8)	4.3 (2.5)	4.1 (2.6)	+1.6 (2.1)***
<i>Ethnicity</i>					
White	80	2.7 (1.9)	4.3 (2.6)	4.0 (2.6)	+1.4 (2.2)***
Other	26	3.3 (2.0)	5.3 (2.4)	5.8 (2.8)	+2.5 (2.7)***
<i>Employment status</i>					
Full-time	49	3.2 (1.9)	4.8 (2.3)	5.1 (2.6)	+1.9 (2.3)***
Other	57	2.5 (2.0)	4.3 (2.8)	3.9 (2.8)	+1.4 (2.4)***

Table 5.4: Longitudinal changes in quit confidence scores (0–10 scale), stratified by demographic characteristics. The table shows mean confidence scores at baseline, post-conversation, and 1-week follow-up for subgroups based on sex, age, ethnicity, and employment status. The change from baseline is also shown. Statistical significance is determined by the Wilcoxon signed-rank test.

Demographic stratification reveals several patterns that warrant careful interpretation. Younger participants (< 30 years) had higher baseline confidence (3.7, SD 2.1) than older groups (2.5, SD 1.8) and showed numerically larger improvements (mean +1.9, SD 3.1) compared to older participants (mean +1.6, SD 2.1). While baseline confidence differed by sex (2.5 for females versus 3.2 for males), week-later changes were identical (1.7). Participants identifying as non-White ethnicities had higher baseline confidence than White participants (3.3 versus 2.7) and showed larger gains (2.5 versus 1.4). These preliminary findings must be interpreted cautiously, as the study was not powered for subgroup analyses and baseline demographic differences were not statistically controlled. See Table 5.4 for a summary of confidence changes by demographic characteristics.

5.2 Perceived Empathy: CARE Scale Assessment

The Consultation and Relational Empathy (CARE) scale (Mercer et al., 2004) measures patients’ perceptions of their healthcare provider’s empathy through 10 questions rated from 1 (poor) to 5 (excellent) (see Appendix F for the full instrument). MIBot achieved a mean total score of 42 out of 50. To contextualize this performance, Table 5.5 compares MIBot’s scores with human healthcare professionals from Bikker et al. (2015).

While MIBot’s mean score approaches that of human providers, the percentage of MIBot interactions achieving perfect scores (11%) remains well below human benchmarks (48%).

Provider	Mean CARE	% Perfect scores
MIBot	42	11
Human healthcare professionals*	46	48

Table 5.5: Comparison of average CARE (Consultation and Relational Empathy) scores and the percentage of perfect scores between MIBot and typical human healthcare professionals. The data for human professionals is from ([Bikker et al., 2015](#)).

Characteristic	making you feel at ease	letting you tell your story	really listening	being interested in you as a whole person	fully understanding your concerns	showing care and compassion	being positive	explaining things clearly	helping you take control	making a plan of action with you	CARE
DEMOGRAPHIC CHARACTERISTICS											
<i>Sex</i>											
Female (n=57)	4.6 (0.7)	4.6 (0.7)	4.5 (0.8)	4.2 (0.9) [†]	4.4 (0.8)	4.4 (0.9)	4.6 (0.7)	4.1 (1.1)	3.8 (1.3)	3.7 (1.5)	42.8 (6.4)
Male (n=49)	4.3 (1.0)	4.5 (0.9)	4.3 (1.0)	3.7 (1.3)	4.1 (1.0)	4.2 (1.0)	4.6 (0.8)	4.1 (0.9)	3.9 (1.3)	3.4 (1.6)	41.1 (8.6)
p ^a	.174	.844	.409	.026*	.078	.470	.738	.580	.438	.376	.489
<i>Age</i>											
<30 (n=26)	4.4 (1.0)	4.6 (0.7)	4.2 (1.2)	3.9 (1.2)	4.0 (1.3)	4.1 (1.0)	4.5 (0.7)	4.2 (1.0)	4.1 (1.1)	4.0 (1.3)	42.0 (8.3)
≥30 (n=80)	4.5 (0.8)	4.5 (0.8)	4.5 (0.8)	4.0 (1.1)	4.3 (0.8)	4.4 (0.9)	4.6 (0.7)	4.1 (1.0)	3.7 (1.4)	3.5 (1.6)	42.0 (7.2)
p ^a	.936	.549	.267	.536	.454	.164	.334	.869	.277	.079	.788
<i>Ethnicity</i>											
White (n=80)	4.5 (0.9)	4.5 (0.8)	4.4 (0.9)	4.0 (1.2)	4.2 (0.9)	4.3 (1.0)	4.6 (0.7)	4.1 (1.0)	3.7 (1.4)	3.5 (1.5)	41.9 (7.6)
Other (n=26)	4.3 (0.8)	4.5 (0.7)	4.4 (0.9)	4.0 (1.0)	4.2 (1.1)	4.2 (0.8)	4.4 (0.8)	4.1 (1.1)	4.2 (1.1)	4.0 (1.4)	42.3 (7.1)
p ^a	.134	.498	.738	.770	.949	.254	.122	.795	.099	.071	.933
<i>Employment</i>											
Full-Time (n=49)	4.3 (1.0)	4.3 (0.9)	4.2 (1.1)	3.8 (1.3)	4.1 (1.0)	4.1 (1.1)	4.5 (0.8)	4.1 (1.0)	3.8 (1.3)	3.6 (1.5)	40.7 (8.8)
Other (n=57)	4.6 (0.7)	4.7 (0.7) [†]	4.6 (0.8) [†]	4.2 (0.8)	4.4 (0.9)	4.5 (0.8) [†]	4.7 (0.6)	4.2 (1.1)	3.8 (1.4)	3.6 (1.5)	43.1 (6.0)
p ^a	.174	.013*	.028*	.144	.152	.012*	.296	.521	.850	.924	.263
BEHAVIOURAL CHARACTERISTICS											
<i>Confidence</i>											
0-1 (n=31)	4.5 (0.9)	4.5 (0.8)	4.4 (0.9)	3.9 (1.3)	4.4 (0.8)	4.4 (1.0)	4.5 (0.8)	4.1 (1.1)	3.7 (1.4)	3.3 (1.6)	41.5 (7.7)
2-3 (n=32)	4.5 (0.7)	4.6 (0.8)	4.6 (0.8)	4.0 (1.0)	4.2 (0.9)	4.4 (0.9)	4.7 (0.7)	4.0 (1.1)	3.8 (1.4)	3.2 (1.5)	42.1 (6.9)
4-5 (n=40)	4.3 (1.0)	4.5 (0.8)	4.3 (1.1)	4.0 (1.2)	4.2 (1.1)	4.2 (1.0)	4.5 (0.7)	4.2 (1.0)	4.0 (1.3)	4.0 (1.4)	42.1 (8.2)
≥6 (n=3)	4.7 (0.6)	5.0 (0.0)	4.7 (0.6)	4.7 (0.6)	3.7 (0.6)	4.7 (0.6)	5.0 (0.0)	4.3 (0.6)	3.7 (0.6)	4.3 (1.2) [†]	44.7 (0.6)
p ^b	.534	.471	.691	.621	.363	.623	.248	.962	.798	.032*	.922
<i>HSI</i>											
Low (0-1) (n=28)	4.1 (1.1)	4.3 (1.0)	4.1 (1.2)	4.0 (1.2)	4.1 (1.2)	4.0 (1.2)	4.4 (0.9)	3.9 (1.2)	3.7 (1.3)	3.8 (1.4)	40.4 (9.1)
Med. (2-3) (n=55)	4.5 (0.8)	4.6 (0.7)	4.5 (0.9)	4.0 (1.1)	4.3 (0.9)	4.3 (0.9)	4.6 (0.7)	4.2 (0.9)	4.0 (1.1)	3.6 (1.5)	42.5 (7.5)
High (4-5) (n=21)	4.7 (0.5)	4.7 (0.5)	4.5 (0.7)	4.0 (1.0)	4.4 (0.8)	4.6 (0.7)	4.7 (0.5)	4.4 (0.9)	3.5 (1.7)	3.3 (1.7)	42.8 (4.8)
V.high (≥6) (n=2)	5.0 (0.0)	5.0 (0.0)	5.0 (0.0)	5.0 (0.0)	4.5 (0.7)	4.5 (0.7)	5.0 (0.0)	3.0 (1.4)	3.5 (2.1)	3.0 (2.8)	43.5 (6.4)
p ^b	.126	.281	.169	.421	.890	.351	.423	.167	.672	.809	.678
<i>Cigarettes/Day</i>											
<5 (n=5)	4.2 (0.4)	4.2 (0.8)	4.2 (0.4)	4.0 (0.7)	5.0 (0.0)	4.2 (0.8)	4.8 (0.4)	4.2 (0.8)	4.2 (0.8)	4.4 (0.9)	43.4 (4.4)
5-9 (n=32)	4.4 (0.9)	4.5 (0.7)	4.4 (0.9)	4.1 (1.2)	4.2 (0.9)	4.3 (0.8)	4.5 (0.7)	4.2 (0.7)	3.9 (1.0)	3.9 (1.4)	42.3 (6.8)
10-19 (n=38)	4.3 (1.0)	4.5 (0.9)	4.4 (1.0)	3.9 (1.1)	4.2 (1.0)	4.2 (1.2)	4.6 (0.8)	4.1 (1.1)	3.8 (1.5)	3.5 (1.5)	41.6 (9.1)
≥20 (n=31)	4.6 (0.7)	4.6 (0.7)	4.3 (1.0)	4.0 (1.2)	4.2 (1.0)	4.5 (0.8)	4.7 (0.5)	4.1 (1.2)	3.8 (1.5)	3.3 (1.7)	42.0 (6.6)
p ^b	.199	.550	.453	.932	.176	.499	.521	.959	.916	.327	.960
<i>Quit Attempts</i>											
0 (n=72)	4.4 (0.9)	4.5 (0.8)	4.4 (1.0)	3.9 (1.2)	4.2 (1.0)	4.2 (1.0)	4.5 (0.7)	4.1 (1.1)	3.7 (1.4)	3.5 (1.5)	41.4 (7.9)
1-2 (n=16)	4.6 (0.6)	4.6 (0.8)	4.6 (0.6)	4.1 (0.9)	4.6 (0.5)	4.4 (1.0)	4.6 (0.6)	4.4 (0.9)	3.9 (1.1)	4.2 (1.2)	44.1 (5.4)
≥3 (n=18)	4.4 (1.0)	4.7 (0.8)	4.3 (1.0)	4.2 (1.1)	4.2 (0.9)	4.4 (0.8)	4.7 (0.8)	4.1 (0.8)	4.1 (1.2)	3.6 (1.7)	42.7 (7.5)
p ^b	.878	.279	.709	.583	.265	.564	.569	.523	.696	.220	.500

Mean(SD). [†] Higher mean(sig). ^a Mann-Whitney. ^b Kruskal-Wallis. *** $p < .001$, ** $p < .01$, * $p < .05$

Table 5.6: CARE (Consultation and Relational Empathy) scale scores, broken down by various demographic and behavioural characteristics of the participants. The table shows the mean and standard deviation of scores for each of the 10 CARE items, as well as the total CARE score.

Question-by-Question Analysis of CARE Survey

How was MIBot at...	Mean (SD)
Being positive	4.6 (0.7)
Letting you tell your “story”	4.5 (0.8)
Making you feel at ease	4.4 (0.9)
Really listening	4.4 (0.9)
Showing care and compassion	4.3 (0.9)
Fully understanding your concerns	4.2 (1.0)
Explaining things clearly	4.1 (1.0)
Being interested in you as a whole person	4.0 (1.1)
Helping you take control	3.8 (1.3)
Making a plan of action with you	3.6 (1.5)
Total Score	42.2 (7.5)

Table 5.7: Mean scores for each of the 10 questions on the CARE (Consultation and Relational Empathy) scale, rated on a 1-5 scale. The table is sorted from the highest-scoring to the lowest-scoring question.

Table 5.7 presents the mean scores for each CARE question, revealing specific strengths and weaknesses in MIBot’s empathic performance. The chatbot performed best on ‘being positive’ (4.6, SD 0.7) and ‘letting you tell your “story”’ (4.5, SD 0.8), while it scored lowest on ‘helping you take control’ (3.8, SD 1.3) and ‘making a plan of action with you’ (3.6, SD 1.5).

The weaker performance on some questions may be due to the chatbot’s limited emotional intelligence (Sabour et al., 2024) or collaboration skills (Yang et al., 2024). Table 5.6 presents CARE scale scores across demographic and behavioural characteristics. Female participants rated MIBot markedly higher than male participants on “being interested in you as a whole person” (4.2 versus 3.7, $p = .026$). Likewise, participants in non-full-time employment gave markedly higher scores than full-time workers on three dimensions: ‘letting you tell your story’ (4.7 versus 4.3, $p = .013$), ‘really listening’ (4.6 versus 4.2, $p = .028$), and ‘showing care and compassion’ (4.5 versus 4.1, $p = .012$). No significant differences were found across age, ethnicity, smoking intensity, or quit attempt categories.

5.3 Comparing Fully Generative MIBot v6.3 With Partially Scripted MIBot v5.2

To contextualize the performance of our fully generative approach, we compare MIBot v6.3A with its predecessor, MIBot v5.2, a hybrid system that combined scripted questions with LLM-generated reflections (Brown et al., 2023). MIBot v5.2 employed a hybrid approach using scripted open-ended questions followed by GPT-2 XL-generated MI-style reflections based on participants’ responses to the questions.

Readiness Ruler Comparisons

MIBot v5.2 achieved a mean confidence increase of 1.3 (SD 2.3, $p < 0.001$) from baseline to one week later among 100 participants, and MIBot v6.3A achieved an increase of 1.7 (SD 2.4, $p < 0.001$) among 106 participants. This difference was not statistically significant ($t(203.60) = 0.98$, $p = 0.165$, one-tailed, Cohen’s $d = 0.14$). As shown in Figure 5.2, the distribution of confidence changes reveals several patterns: MIBot

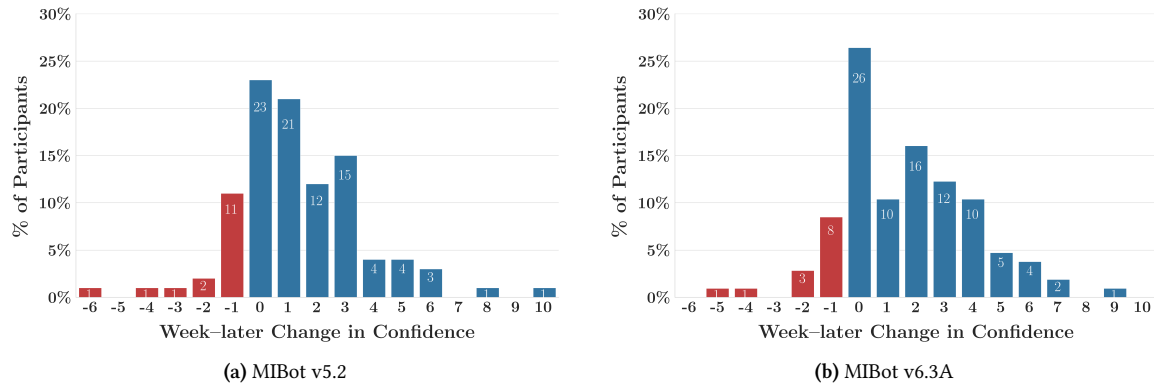


Figure 5.2: Distribution of week-later confidence changes from baseline for MIBot v5.2 and MIBot v6.3A. The x-axis shows the change in confidence score, and the y-axis shows the number of participants. Red bars indicate a decrease in confidence, while blue bars indicate an increase or no change.

v6.3A resulted in more participants maintaining their baseline confidence (28% vs. 23% with no change) and fewer experiencing decreased confidence (13% vs. 17%).

For importance to quit, MIBot v5.2 achieved a significant increase of 0.7 (SD 2.0, $p < 0.001$), while MIBot v6.3A showed a more modest but still significant gain of 0.5 (SD 1.7, $p < 0.005$). Readiness changes were minimal and non-significant for both versions (v5.2: 0.4, SD 1.7, $p = 0.01$; v6.3A: 0.3, SD 2.4, $p = 0.22$).

Perceived Empathy Comparisons

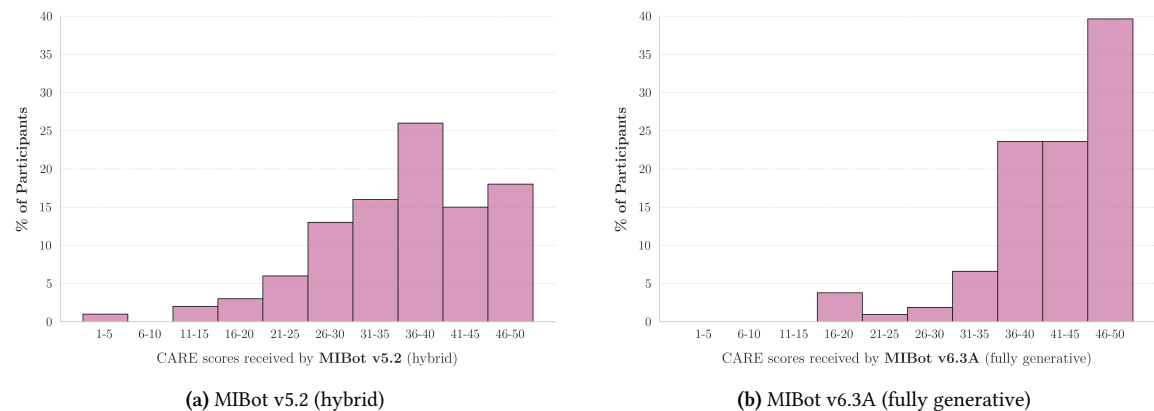


Figure 5.3: Distribution of CARE (Consultation and Relational Empathy) empathy scores for MIBot v5.2 and MIBot v6.3A. The x-axis represents the total CARE score, and the y-axis represents the number of participants. The distribution for the fully generative v6.3A is shifted to the right, indicating higher perceived empathy.

MIBot v5.2 achieved a mean CARE score of 36 (SD = 9.1), with only 3% of participants awarding a perfect score, while MIBot v6.3A achieved a mean score of 42 (SD = 7.5), with 11% receiving perfect scores. This difference was statistically significant ($t(191.67) = 4.96$, $p < .001$, Cohen's $d = 0.70$), representing a medium-to-large effect size, and suggests that the fully generative responses of v6.3A markedly improved perceived empathy. Figure 5.3 illustrates the distributional shift between versions. The hybrid v5.2 shows a relatively normal distribution centred in the mid-30s, with considerable spread across all score ranges, whereas v6.3A demonstrates a clear rightward skew, with 40% of participants rating the chatbot in the highest range (46–50).

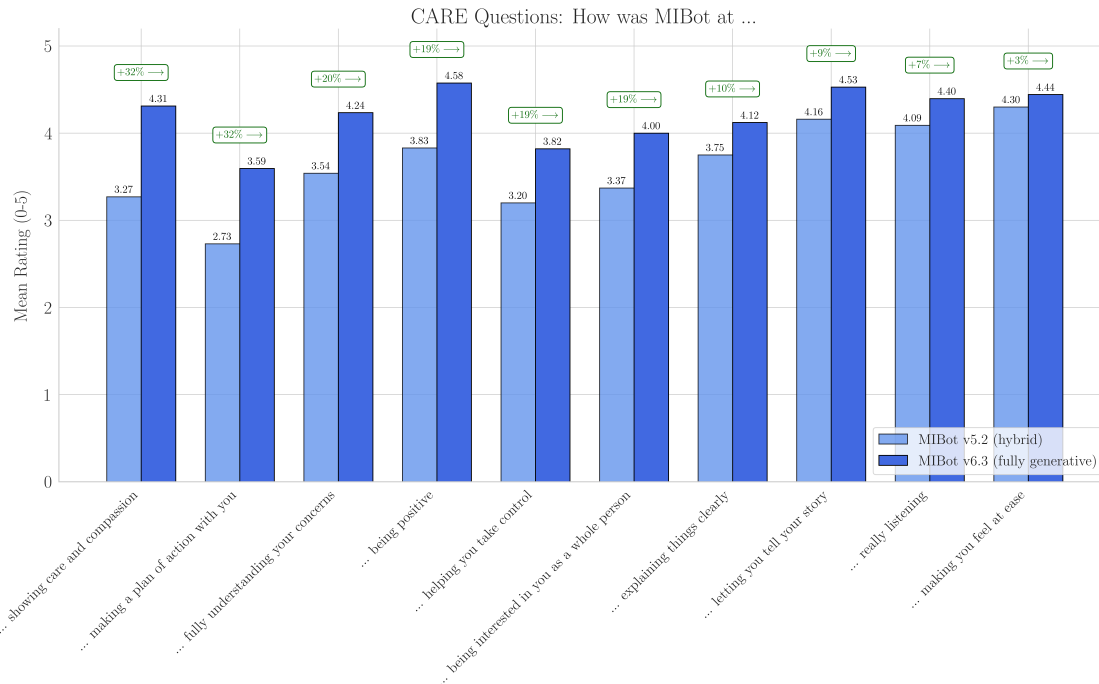


Figure 5.4: Comparison of mean CARE (Consultation and Relational Empathy) scores for each question between MIBot v5.2 (hybrid) and v6.3A (fully generative). The x-axis lists the 10 CARE questions, and the y-axis shows the mean score. The fully generative version shows improvement across all dimensions.

To understand which aspects of empathetic interaction improved most, Figure 5.4 presents the mean scores across all ten CARE dimensions. The fully generative approach exhibits improvements across every dimension, with particularly notable gains in emotional and communicative aspects. The largest relative improvement was observed in “showing care and compassion” (+32%), reflecting the chatbot’s improved ability to maintain an encouraging tone. The dimension of “making a plan of action with you” remained the weakest aspect for both versions (2.73 for v5.2, 3.59 for v6.3A), despite showing considerable relative improvement (+32%). Notably, MIBot v6.3A was prompted with detailed guidelines on how to make a plan of action with the client. Despite this, planning remains one of its weaker areas.

Implications of the Comparison

The fully generative MIBot v6.3A scored higher on CARE, the measure of perceived therapeutic alliance. This improvement suggests that participants experienced more authentic, personalized interactions when the entire conversation—not just reflections—emerged from the language model’s contextual understanding. However, both chatbots scored similarly on our primary metric of effectiveness, viz. the week-later change in confidence, indicating that the core therapeutic mechanism of MI may be responsible for most of the gains in confidence. This finding aligns with prior work showing that even simple question-asking can produce substantial benefits (Brown et al., 2023), though the improved empathy of full generation may further improve engagement and retention in real-world deployment.

5.4 AutoMISC Analysis

To evaluate the chatbot’s adherence to MI principles, we analysed counsellor and client utterances from the feasibility study transcripts (Chapter 4) using AutoMISC, an automated annotation system originally described in [Mahmood et al. \(2025\)¹](#). The system assigns behavioural codes to each utterance: counsellor utterances are classified as MI-Consistent (MICO), MI-Inconsistent (MIIN), Reflection (R), Question (Q), or Other (O), while client utterances are categorized as change talk (C), sustain talk (S), or neutral (N). Following annotation of all utterances, we computed per-transcript summary metrics to quantify MI adherence. These comprise:

- **Percentage MI-Consistent Responses (%MIC):** The proportion of counsellor utterances that align with MI principles. Higher values indicate greater adherence to MI methodology.
- **Reflection-to-Question Ratio (R:Q):** The ratio of counsellor utterances labelled as reflection (R) to those labelled as question (Q). This metric assesses the balance between reflective listening and questioning. Values between 1 and 2 are considered indicative of proficiency ([Moyers et al., 2016b](#)).
- **Percentage Change Talk (%CT):** The proportion of client utterances expressing motivation toward behaviour change. Higher values are associated with improved behavioural outcomes ([Apodaca and Longabaugh, 2009](#)).

5.4.1 Contextualizing AutoMISC Metrics With the HLQC Dataset

To provide a point of comparison for the MISC summary metrics, we also ran AutoMISC on the HighLowQualityCounselling (HLQC) dataset ([Pérez-Rosas et al., 2019](#)), a publicly available² corpus of transcribed MI counselling demonstrations. The HLQC dataset comprises 155 high-quality (HLQC_HI) and 104 low-quality (HLQC_LO) transcripts sourced from public websites. We computed summary scores separately for these subsets and then compared MIBot’s summary metrics against those of both HLQC_HI and HLQC_LO.

5.4.2 Counsellor Behaviour Metrics

Metric	MIBot	HLQC_HI (high-quality)
%MI-consistent (%MIC)	98 (3.6)	92 (9.8)
Reflection-to-Question Ratio (R:Q)	1.3 (0.3)	2.3 (5.7)

Table 5.8: Counsellor-specific summary metrics from AutoMISC for MIBot, compared to high-quality human counselling sessions from the HLQC dataset. The table shows the mean (and standard deviation) for the percentage of MI-consistent responses (%MIC) and the reflection-to-question ratio (R:Q).

Figure 5.5 shows violin plots comparing the distribution of %MIC, R:Q, and %CT across MIBot conversations with the HLQC benchmarks. The %MIC scores are tightly clustered near 100%, with only a few transcripts falling below 80%. The R:Q distribution is centred around 1.3, showing less variance than human counsellors who range from pure question-asking (R:Q near 0) to heavy reflection use (R:Q > 5).

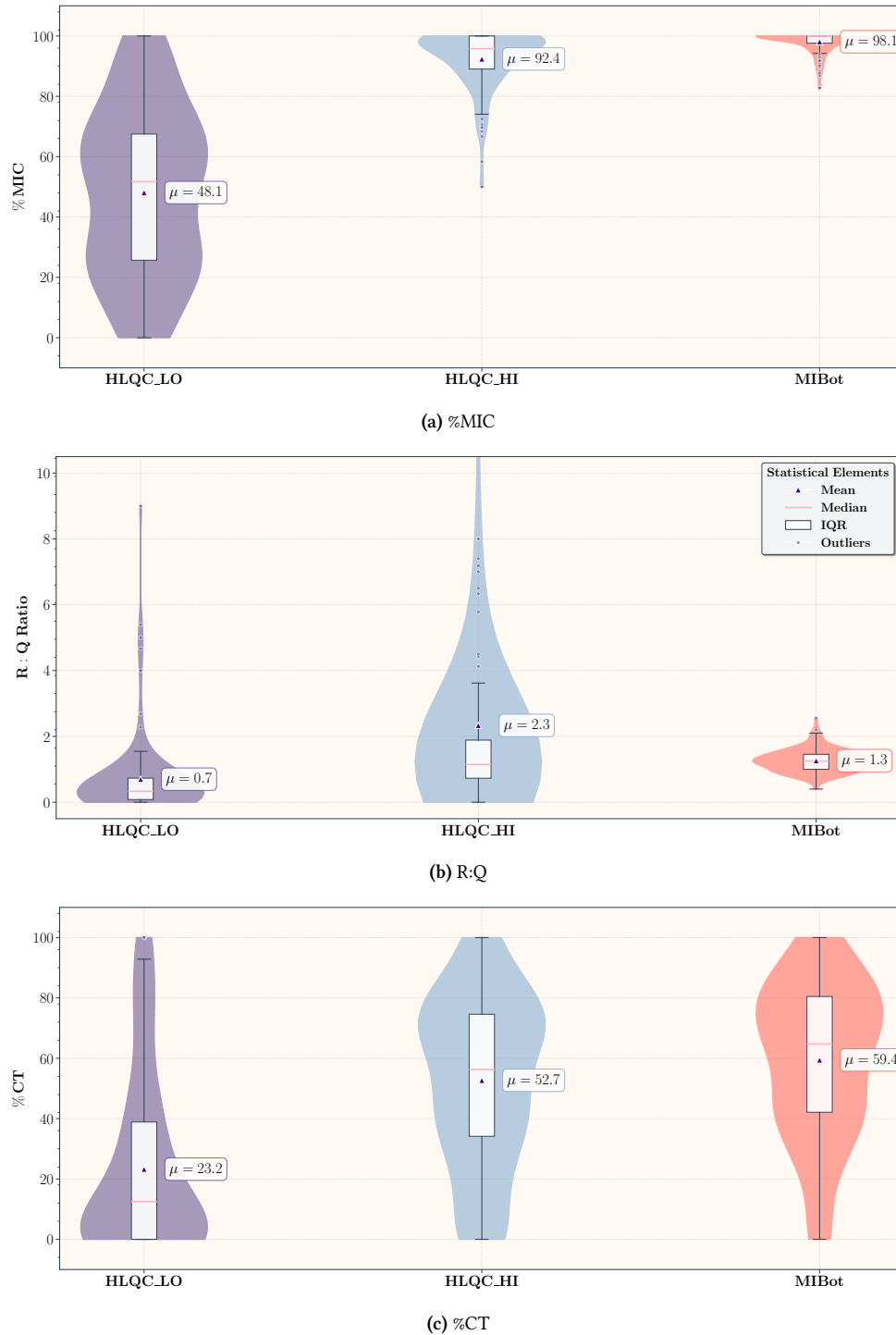


Figure 5.5: Comparison of MISC summary score distributions for MIBot and the HLQC dataset (high and low quality). The figure shows violin plots for (a) Percentage MI-Consistent Responses (%MIC), (b) Reflection to Question Ratio (R:Q), and (c) Percentage Client Change Talk (%CT). Taken from [Ali \(2025\)](#)

Metric	MIBot	HLQC_HI (high-quality)
%Change Talk (%CT)	59 (25.6)	53 (28.4)

Table 5.9: Client-specific summary metric from AutoMISC for MIBot, compared to high-quality human counselling sessions from the HLQC dataset. The table shows the mean (and standard deviation) for the percentage of change talk (%CT).

5.4.3 Client Change Talk Analysis

Figure 5.5(c) illustrates the distribution of %CT across conversations. Most sessions by MIBot had a %CT of $> 60\%$, and the distribution closely matches that of the HLQC_HI dataset. Overall, the mean %CT was 59% for MIBot, compared to 53% for the HLQC_HI dataset.

5.5 Behavioural Outcomes

Quit Attempts at One Week

Quit Attempt Status	Pre-conversation	Post-conversation
Made attempt	34 (32.1%)	37 (34.9%)
No attempt	72 (67.9%)	69 (65.1%)
New attempters	–	15 (14.2%)
Sustained attempters	–	22 (20.8%)

Table 5.10: Summary of participants’ quit attempt behaviour before and after the MIBot conversation. The table shows the number and percentage of participants who made a quit attempt in the week prior to the conversation and in the week following the conversation.

Participants who made quit attempts post-conversation showed larger confidence gains (mean +2.24) compared to those who did not attempt (+1.49, $p = 0.0664$). This bidirectional relationship—where confidence predicts attempts and attempts reinforce confidence—aligns with social cognitive theory (Bandura, 1986).

5.6 Conversation Analysis

Quantitative Flow

Conversations ranged from 36 to 163 utterances (median 78; mean 80.71, SD 25.54), with counsellor utterances comprising 71.7% of the total. The median conversation lasted approximately 30 minutes. Table 5.11 summarizes the main quantitative metrics.

Longer conversations correlated with better outcomes ($r = 0.29$ for confidence change), but the relationship was non-linear. Conversations under 60 utterances rarely produced substantial gains (only 21% reached ≥ 2 -point confidence increases), while those exceeding 130 utterances showed diminishing returns (50% achieving ≥ 2 -point gains with a small sample). This suggests an optimal engagement window of roughly 110–120 exchanges before improvements level off.

¹A detailed description of the AutoMISC system by Ali et al. (2025) is forthcoming

²<https://lit.eecs.umich.edu/downloads.html>

Metric	Mean (SD)	Range
Total utterances	80.7 (25.5)	36–163
Counsellor utterances	57.9 (20.0)	26–123
Client utterances	22.8 (6.9)	9–40
Words per counsellor utterance	14.6 (1.5)	11–19
Words per client utterance	7.0 (2.6)	1–13

Table 5.11: Quantitative metrics of the conversation dynamics between participants and MIBot. The table includes statistics on the total number of utterances, counsellor and client utterances, words per utterance, and session duration.

Qualitative Thematic Analysis

To understand the qualitative aspects of the conversations, a thematic analysis was performed on the full corpus of transcripts. The analysis was conducted by two researchers who independently reviewed the transcripts to identify recurring patterns and themes. They then met to compare their findings, discuss discrepancies, and collaboratively develop a final set of themes that characterized successful therapeutic engagement. This process revealed three recurring patterns:

Stress and Coping Narratives

56% of conversations referenced this theme across 59 transcripts. Participants described cigarettes as tools for emotion regulation and short-term relief. Discussions often focused on stress at work, family pressures, or using smoke breaks as a momentary escape.

"It is a stress reliever for me."

"Every time I have tried to quit, other things in my life have come up and it helps me to calm down to deal with them."

"My job becomes more stressful."

Social and Ritualistic Aspects

51% of conversations referenced this theme across 54 transcripts. Smoking was entwined with social identity and daily rituals. Clients highlighted smoke breaks with colleagues, shared routines with partners, and the fear of losing social contact if they quit.

"I'm embarrassed that I can't just quit by myself."

"I feel like if I actually commit to making a change, it could be possible."

"I could speak to my friends and family about seeing a counsellor to get their support and advice."

Ambivalence Themes

19% of conversations referenced this theme across 20 transcripts. Clients expressed motivational ambivalence, simultaneously acknowledging reasons to quit while defending the role of smoking in their lives. Conversations frequently normalized these mixed feelings before exploring change talk.

"My health and the appearance of my teeth are probably the main reasons I want to quit."

"It makes me want to quit"

"My wife wants me to quit and I want to quit for my health and my marriage."

Success Stories

Two clients reported marked increases in confidence within one week, moving from very low to high confidence after extended dialogues. Their narratives illustrate shifts in perspective:

“Seeing underage kids smoke has made me see it in a negative light.”

“Perhaps you tell me how you would wish to help me as you have stated above.”

Enjoyment-Focused Smokers

Example quotes include:

“At night you can enjoy the cooler air and watch the stars and think about things.”

“Frankly, I enjoy it, and long ago I tried hypnosis and it worked; I found it distasteful.”

5.7 User Experience and Feedback

Post-Conversation Feedback

Three open-ended questions captured immediate post-conversation reactions. Thematic analysis revealed distinct patterns in positive and negative responses (see Appendix G for a selection of verbatim responses).

Positive Themes (92% enjoyed the experience):

- Non-judgmental approach: “Finally someone (something?) that didn’t lecture me.”
- Structured discussion: “Helped me organize my thoughts about quitting.”
- Convenience and privacy: “Could be honest without embarrassment.”
- Surprising depth: “More helpful than expected from a bot.”

Negative Themes (34% found it unhelpful):

- Lack of personalization: “Felt like generic responses sometimes.”
- Missing human connection: “Technically correct but emotionally flat.”
- Repetitiveness: “Kept asking similar questions different ways.”
- Insufficient challenge: “Too accepting, didn’t push me enough.”

User Segmentation

Based on feedback analysis, we derived two binary metrics: “LikedBot” (92% positive) and “FoundBotHelpful” (66% positive). The discrepancy suggests that while MIBot creates an engaging experience, translating engagement into perceived therapeutic value remains challenging.

Cross-tabulation revealed four user segments, as shown in Table 5.12.

“Entertained Skeptics”—those who enjoyed but didn’t find it helpful—often wanted more directive advice or concrete tools rather than open-ended conversation.

Segment	Liked & Helpful	% of Sample
Enthusiasts	Yes & Yes	61%
Entertained Sceptics	Yes & No	31%
Reluctant Beneficiaries	No & Yes	5%
Dissatisfied	No & No	3%

Table 5.12: Segmentation of users based on their reported enjoyment and perceived helpfulness of the MIBot conversation. The table shows the percentage of participants in each of the four segments: Enthusiasts, Entertained Sceptics, Reluctant Beneficiaries, and Dissatisfied.

5.8 Discussion

Multivariate regression analysis identified the strongest predictors of confidence improvement:

1. **Low baseline confidence** ($\beta = -0.31$, $p < 0.001$): greatest gains for those starting lowest.
2. **Recent quit attempt** ($\beta = 0.28$, $p < 0.01$): prior action amplifies intervention effects.
3. **Conversation length** ($\beta = 0.21$, $p < 0.05$): deeper engagement yields better outcomes.
4. **Change talk ratio** ($\beta = 0.18$, $p < 0.05$): client language predicts behaviour change.
5. **Age <40** ($\beta = 0.15$, $p < 0.05$): younger participants more responsive.

Together, these factors explained 34% of the variance in confidence change ($R^2 = 0.34$, $F(5, 100) = 10.32$, $p < 0.001$), suggesting that ideal candidates are younger smokers with low confidence who have recently attempted to quit and are willing to engage in extended conversation.

Comparison With Literature Benchmarks

MIBot's performance compares favourably with established interventions:

- **Effect size:** Cohen's $d = 0.71$ for confidence change exceeds typical digital intervention effects ($d = 0.2$ – 0.4) (Whittaker et al., 2016).
- **MI fidelity:** 95% MI-consistent responses surpass typical human counsellor benchmarks (80–90%) (Moyers et al., 2016a).
- **Engagement:** 92% enjoyment rate exceeds most digital health interventions (60–70%) (Perski et al., 2017).
- **Quit attempts:** 14.2% new quit attempts align with brief intervention outcomes (10–20%) (Stead et al., 2013).

However, MIBot falls short of intensive human counselling in perceived helpfulness (66% vs. 80–90%) and perfect CARE scores (11% vs. 48%), indicating room for improvement in therapeutic alliance building.

Clinical Implications

The findings from this study have several potential clinical implications. The strong technical performance of MIBot, combined with meaningful behavioural outcomes, suggests that generative AI can be a valuable tool in public health interventions for smoking cessation. The high MI fidelity and user engagement rates

indicate that such a system could be deployed as a scalable, low-cost, first-line intervention to reach a large number of smokers who may not have access to traditional counselling.

However, the variability in individual responses and the identified limitations in building deep therapeutic alliance suggest that MIBot is likely best positioned as an adjunct to human services rather than a complete replacement. It could serve as a tool to support human counsellors, handle initial screenings, or provide support between sessions. Future work should examine these hybrid models of care.

Limitations

Several limitations constrain the generalisability of our findings:

Methodological Limitations:

- Short follow-up period (one week) precludes assessment of sustained behaviour change.
- Self-reported outcomes without biochemical verification may overestimate quit attempts.
- Lack of control group prevents causal attribution.
- Single-session design doesn't capture potential for repeated engagement.

Sample Limitations:

- Prolific recruitment may select for digitally literate, research-oriented participants.
- Exclusion of high-confidence smokers limits understanding of broader applicability.
- Monetary compensation (£6.50) may attract non-genuine participants.
- English-only implementation excludes non-English speakers.

Technical Limitations:

- Text-only interface eliminates non-verbal communication channels.
- Lack of memory between sessions prevents relationship building.
- No integration with clinical services or prescription capabilities.
- Limited ability to handle crisis situations or complex comorbidities.

5.9 Conclusion

This chapter presented a detailed evaluation of MIBot, a fully generative MI chatbot. The study demonstrated that MIBot can produce a clinically meaningful increase in smokers' confidence to quit, with high fidelity to MI principles and strong user engagement. The analysis identified major predictors of success, highlighting that the chatbot was most effective for younger, low-confidence smokers who had recently attempted to quit.

While the results are encouraging, the study also revealed limitations in building deep therapeutic alliance and translating engagement into perceived helpfulness for all users. These findings establish a strong proof-

of-concept for AI-delivered MI as a scalable public health intervention, while also underscoring the areas for future research and development, particularly in hybrid models of care that combine AI with human support.

Chatbot Contribution Attribution

The research detailed in Chapters 3 to 5 is the result of a major, long-term collaboration. The human feasibility study, in particular, was a major team undertaking and involved all co-authors of the paper by [Mahmood et al. \(2025\)](#). To provide a clear overview of each individual's role in the work presented in this thesis, this section outlines specific attributions.

- **Soliman Ali:** As the second author on the work submitted by [Mahmood et al. \(2025\)](#), Ali was instrumental in designing and implementing the modular observer agent structure that serves as the foundation for the MIBot system (see Section 3.2). He also contributed to the initial versions of the AutoMISC system ([Ali, 2025](#); [Ali et al., 2025](#)), which was developed concurrently with MIBot. AutoMISC not only helped improve MIBot through an iterative process but was also used by the author of this thesis to validate the installation of attributes in synthetic smokers (see Chapters 6 and 7).
- **Michelle Collins:** Collins played a central role in developing the prompt used by MIBot and provided valuable feedback by testing various intermediate system versions during its development.
- **Sihan Chen:** Chen implemented a backend function for uploading conversation artifacts to AWS S3 and worked on the prompt for the End Classifier Observer.
- **Yi Chen (Michael) Zhao:** Zhao was responsible for developing the Off-Track observer agent, a component designed to detect when conversation topics deviate from smoking cessation.

The author of this thesis made the following contributions:

- Led the design and development of the fully generative MIBot v6.3A.
- Guided the iterative improvement of the system prompt in collaboration with expert MI clinicians and built early versions of synthetic smokers for system testing and validation.
- Provisioned the resources required to deploy MIBot to a cloud infrastructure, including researching the right AWS systems to use and configuring AWS S3 for data storage.
- Conducted the human feasibility study, a process that involved securing ethics approval, recruiting participants through Prolific, and overseeing all aspects of data collection and participant compensation.
- Handled post-experiment data processing, including collection, cleaning, and a portion of the analysis presented in this chapter.

Chapter 6

Development of Synthetic Smokers: Preliminary Experiments

The iterative development of MIBot, as described in Chapter 3, required that we thoroughly test it after every change in its prompt. To automate the testing of some aspects of the chatbot’s behaviour, we started the development of “synthetic smokers”¹. A synthetic smoker, in the context of this research, is a prompted instance of an LLM tasked with simulating the behaviour, language, and psychological responses of a human smoker engaged in a counselling session.

This chapter details the initial conceptualization and preliminary experiments in the development of these synthetic agents. We begin by defining the theoretical approach for creating realistic agents, followed by our early experiments to control behavioural attributes like verbosity and resistance to quitting. These initial steps show the core challenges of attribute installation and validation. In the subsequent chapter, we use a data-driven methodology for attribute installation and validation.

6.1 Objectives and Overview

The primary goal of developing synthetic smokers is to create realistic and controllable proxies for human participants. This endeavour serves two primary objectives:

1. **Testing Automated Systems:** Synthetic smokers provide a basis for the rapid iterative development and testing of automated counselling chatbots. By simulating interactions, developers may identify weaknesses, improve conversational strategies, and assess potential efficiency without the logistical overhead of recruiting human subjects for every iteration.
2. **Training Human Clinicians:** In the role of standardized patients, synthetic smokers can be used to train novice clinicians. They offer a consistent experience and can be programmed to simulate specific challenging scenarios or different demographic profiles, which could help trainees receive varied practice opportunities.

¹Notably, we also used human role-playing as smokers during our testing of MIBot.

The development of a truly representative synthetic smoker poses a scientific challenge. It requires not only the generation of coherent dialogue but the accurate embodiment of specific human attributes ranging from demographics to complex psychological states like resistance to change. Importantly, the validation of these synthetic agents (i.e., proving that they behave realistically according to their installed attributes) is exceptionally difficult, as establishing the *construct validity* of behavioural measurements remains a persistent challenge; that is, it is difficult to ensure that an instrument truly measures the specific psychological attribute it purports to assess (Cronbach and Meehl, 1955).

6.2 Goals for a Synthetic Smoker

In an ideal scenario, a synthetic smoker would be indistinguishable from a human participant within the constrained context of a smoking cessation counselling session. This requires the synthetic agent to possess a defined set of attributes and for those attributes to be reliably reflected in its interactions and behaviours.

We conceptualize a synthetic smoker based on a collection of attributes that define a human smoker. These attributes can be broadly categorized as:

- **Demographic:** age, sex, education level, cultural background, etc.
- **Behavioural:** smoking history (e.g., Heaviness of Smoking Index (HSI)), previous quit attempts, verbosity, etc.
- **Psychological:** resistance to change, motivation levels (e.g., operationalized by pre-conversation readiness rulers), underlying values, etc.

A successful system (or methodology) for creating synthetic smokers must produce agents that exhibit both high *fidelity* and *representativeness*.

Fidelity refers to the requirement that any attribute installed in the synthetic smoker must be demonstrably evidenced by its language and behaviour. For example, if a synthetic smoker is installed with high resistance, its dialogue should contain more Sustain Talk than Change Talk. The core challenge of this research is to develop methods to install these attributes and subsequently validate that this installation was successful.

Representativeness encompasses two main requirements:

1. The system must generate a different range of smokers whose attributes mirror the variability of the target real-world population. It must avoid the bias of over-representing a narrow or easily simulated archetype at the expense of mirroring the population's true statistical makeup. This may happen for a variety of reasons. For example, if the system is built on biased data, it may fail to generate a statistically accurate population.
2. The smokers created by the system should exhibit uniform fidelity, i.e., it should not have performance biases where specific profiles are simulated more accurately than others. A system might, for instance, be highly effective at simulating a 55-year-old, heavily dependent smoker with low motivation to quit, while failing to simulate a 22-year-old "social smoker" who is ambivalent about quitting. Representativeness requires that the quality of the simulation remains high across the entire range of generated profiles.

Let the collection of all relevant attributes define an n -dimensional *attribute space*, $\mathcal{A}$. A specific smoker's

profile, whether human or synthetic, is a vector $\mathbf{A}$ within this space.

$$\mathbf{A} = (a_1, a_2, \dots, a_n) \in \mathcal{A}$$

The *observable output space*, Γ , contains all possible outputs from a counselling session, such as conversation transcripts and survey responses. A specific session's output is an element $\gamma \in \Gamma$.

$$\gamma = \{\text{Transcript, Survey Responses, } \dots\}$$

Human behaviour is not deterministic; a human H with a given attribute vector $\mathbf{A}_H$ will not produce the exact same output in every session. Instead, their behaviour is better modelled as a sample from a conditional probability distribution, $P_H(\gamma|\mathbf{A}_H)$. Consequently, a synthetic smoker system, S , should not be a deterministic function but a model that approximates this human distribution, $P_S(\gamma|\mathbf{A}_S)$. The goal is to ensure that $P_S(\gamma|\mathbf{A}) \approx P_H(\gamma|\mathbf{A})$.

$$\mathbf{A}_S = \mathbf{A}_H \longrightarrow P_S(\gamma|\mathbf{A}_S) \approx P_H(\gamma|\mathbf{A}_H)$$

To assess whether the system has achieved this goal, we must formalize the process of validation. For this, we define a *measurement function*, M , that takes an output γ and returns an estimated attribute vector $\hat{\mathbf{A}}$.

$$M : \Gamma \rightarrow \mathcal{A}, \quad \text{where} \quad \hat{\mathbf{A}} = M(\gamma)$$

For example, M could be a set of classifiers and analytical tools that read a transcript and analyse post-conversation changes in behaviour to estimate the speaker's resistance, motivation, smoking history, etc., and produce $\hat{\mathbf{A}}$. The measurement function M is an imperfect proxy, as inferring latent constructs from observable behaviour is a persistent challenge in psychometric theory (Loevinger, 1957; Borsboom et al., 2004).

6.2.1 Defining Fidelity

Fidelity measures how accurately the observable outputs of a synthetic smoker reflect its assigned attributes. High fidelity is achieved when the attributes measured from the output, $\hat{\mathbf{A}}_S = M(\gamma_S)$, are close to the input attributes, $\mathbf{A}_S$. We quantify this correspondence using a distance metric, $d(\mathbf{A}_S, \hat{\mathbf{A}}_S)$, in the attribute space.

Given the stochastic nature of the generative system, a single instance is insufficient for evaluation. We therefore define the fidelity for a profile $\mathbf{A}_S$ as the *expected distance* between the input and measured attributes over the distribution of all possible outputs.

$$\mathcal{F}(\mathbf{A}_S) = \mathbb{E}_{\gamma_S \sim P_S(\gamma|\mathbf{A}_S)} [d(\mathbf{A}_S, M(\gamma_S))]$$

A lower value of $\mathcal{F}(\mathbf{A}_S)$ indicates higher fidelity. The objective of the system, then, is to minimize this value across all possible profiles.

In practice, M may not be needed to calculate fidelity. Instead, a more practical approach is to validate

individual attributes by demonstrating a strong correlation between their installed values and corresponding, measurable features in the output γ . For example, to validate the “resistance to change” attribute, one can measure a proxy like the percentage Change Talk (%CT) in the transcript. Evidence of fidelity would be a strong, negative correlation between the installed resistance level and the observed Change Talk across a cohort of synthetic agents. This correlation-based method is more feasible than developing a detailed measurement function M .

6.2.2 Defining Representativeness

Representativeness ($\mathcal{R}$) ensures that the synthetic population is both a statistically accurate and a consistently well-simulated reflection of the real-world population. It comprises two distinct components.

Distributional Representativeness

The first requirement, which we term **distributional representativeness** ($\mathcal{R}_{dist}$), is that the statistical makeup of the generated synthetic smokers matches that of the target human population. Let $P_H(\mathbf{A})$ be the true probability distribution of attribute vectors in the target population. If we generate synthetic agents using a distribution $P_S(\mathbf{A})$, this component is high when $P_S(\mathbf{A})$ is close to $P_H(\mathbf{A})$. We can measure this similarity using the *Kullback-Leibler (KL) Divergence*:

$$\mathcal{R}_{dist} = D_{KL}(P_H(\mathbf{A}) \parallel P_S(\mathbf{A})) = \sum_{\mathbf{A} \in \mathcal{A}} P_H(\mathbf{A}) \log \frac{P_H(\mathbf{A})}{P_S(\mathbf{A})}$$

A system with high distributional representativeness will minimize this divergence, ensuring that the generated population is not biased towards easily simulated archetypes. Alternatively, to assess whether the system’s outputs preserve the population’s statistical distribution, one can compare the original input distribution, $P_H(\mathbf{A})$, to the distribution of attributes measured from the synthetic agent’s final output, $P_S(M(\gamma_S))$.

The KL divergence is a measure of the distortion between the intended input attributes and the measured output attributes:

$$\begin{aligned} \mathcal{R}_{dist-M} &= D_{KL}(P_H(\mathbf{A}) \parallel P_S(M(\gamma_S))) \\ &= \sum_{\mathbf{A} \in \mathcal{A}} P_H(\mathbf{A}) \log \frac{P_H(\mathbf{A})}{P_S(M(\gamma_S) = \mathbf{A})} \end{aligned}$$

Here, $P_S(M(\gamma_S) = \mathbf{A})$ is the probability that the measured attributes from a synthetic agent’s output γ_S will equal the vector $\mathbf{A}$, when the agent’s input profile was drawn from the true human distribution $P_H(\mathbf{A})$.

In practice, this is calculated by comparing the empirical distribution of a sample of human attribute vectors $\{\mathbf{A}_i\}$ against the distribution of their corresponding measured outputs $\{\hat{\mathbf{A}}_i = M(\gamma_i)\}$. This calculation assumes that M faithfully extracts the attributes from γ , of which there is no guarantee.

Uniform Fidelity

The second requirement, **uniform fidelity** ($\mathcal{R}_{unif}$), stipulates that the simulation quality should be consistent across all types of smoker profiles. The fidelity score, $\mathcal{F}(\mathbf{A})$, should not be systematically better for some profiles than for others. We can measure this by calculating the *variance of the fidelity score* across the

distribution of real-world smokers, $P_H(\mathbf{A})$:

$$\mathcal{R}_{unif} = \text{Var}_{\mathbf{A} \sim P_H(\mathbf{A})} [\mathcal{F}(\mathbf{A})] = \mathbb{E}_{\mathbf{A} \sim P_H(\mathbf{A})} [(\mathcal{F}(\mathbf{A}) - \bar{\mathcal{F}})^2]$$

where $\bar{\mathcal{F}}$ is the mean fidelity across the population. A system that achieves high uniform fidelity will have a value of $\mathcal{R}_{unif}$ close to zero, indicating that its performance is reliable and unbiased across the full range of human smokers. In practice, it is sufficient to calculate fidelity stratified by different relevant attributes (e.g., age groups, sex, ethnicity, etc.) and investigate if a certain group of synthetic agents has low fidelity.

6.3 Development of Synthetic Smokers

We began our development of synthetic smokers by prompting an LLM and providing instructions on how to behave like a smoker in a clinical setting. As we evaluated our synthetic smokers, our approaches to installation and validation evolved in tandem. In the following sections, we describe our methodology for creating synthetic smokers along with the associated validation steps.

6.3.1 The Baseline Synthetic Smoker

The goal of the baseline synthetic smoker was to create a conversational partner that could interact with MIBot prototypes, allowing the researchers to test conversational flow and MI adherence of MIBot in a controlled environment.

Methodology

We began with a minimal prompt instructing an LLM (initially GPT-4 Turbo) to adopt the persona of a smoker engaging in a counselling session. The prompt was kept minimal and did not provide a *backstory* or specific behavioural constraints.

You are a human smoker engaged in a private thirty-minute session with a counsellor. This is your first time talking to a therapist about your smoking habits... Respond to the counsellor's inquiries as accurately as possible...

Since we did not install any attributes to this baseline, we were mainly looking for how it conversed with MIBot and whether it had some default attributes. We generated a small number ($N = 5$) of conversations between this baseline synthetic smoker and an early version of the MIBot counsellor. The validation method at this stage was qualitative human inspection by the research team, including expert MI clinicians.

Observations

The transcript of the conversation between MIBot and the baseline synthetic smoker revealed two major issues that made the synthetic smoker sound “unnatural” or atypical of a human smoker:

1. **Excessive Verbosity:** The synthetic smoker tended to produce long, elaborate responses. This felt unnatural and resembled written prose more than spontaneous dialogue.

2. High Agreeableness (Low Resistance): The baseline agent was overly agreeable and eager to change. This failed to simulate the ambivalence characteristic of many smokers seeking cessation support (as discussed in Chapter 2).

These observations showed the necessity of explicitly controlling these attributes.

6.3.2 Controlling Verbosity

To address the issue of excessive verbosity, we undertook a prompt engineering effort to constrain the synthetic smoker’s utterance length and style.

Methodology

We modified the prompt to encourage a more natural, concise conversational style. Main additions to the prompt included:

- Stylistic Guidance: Instructions such as “Imagine you’re texting a friend. Keep it casual...” and encouraging the use of emojis.
- Instructions to be Concise: Directives like “speak with more clarity rather than exhaustive detail.”
- Explicit Constraints: Hard constraints such as “Number of sentences in your response must be between 1 and 4.”

Validation

To validate the effectiveness of these changes, we generated a new set of conversations ($n = 5$) using the modified prompt (“Fixed Verbosity”) and with GPT-4 Turbo (temperature = 1.0) and compared them against the baseline prompt (“Default Verbosity”). We also compared these results against human-human MI conversations from the HLQC datasets (Pérez-Rosas et al., 2019).

Results

The prompt modifications markedly reduced the verbosity of the synthetic smoker. Figure 6.1 illustrates the distribution of volley lengths (regarding number of words).

Category	Mean	Standard Deviation
Default Verbosity (GPT-4 Turbo)	618.9	296.3
Fixed Verbosity (GPT-4 Turbo)	337.8	157.9
High-Quality Human Counselling	147.0	172.9
Low-Quality Human Counselling	118.2	158.0

Table 6.1: Statistics on Volley Length for Low- and High-Verbosity Synthetic Smokers.

The default verbosity condition showed a wide distribution with a high median utterance length. The fixed-verbosity synthetic smokers successfully matched the utterance length distribution of humans from the HLQC dataset. Table 6.1 provides detailed statistics on the volley lengths.

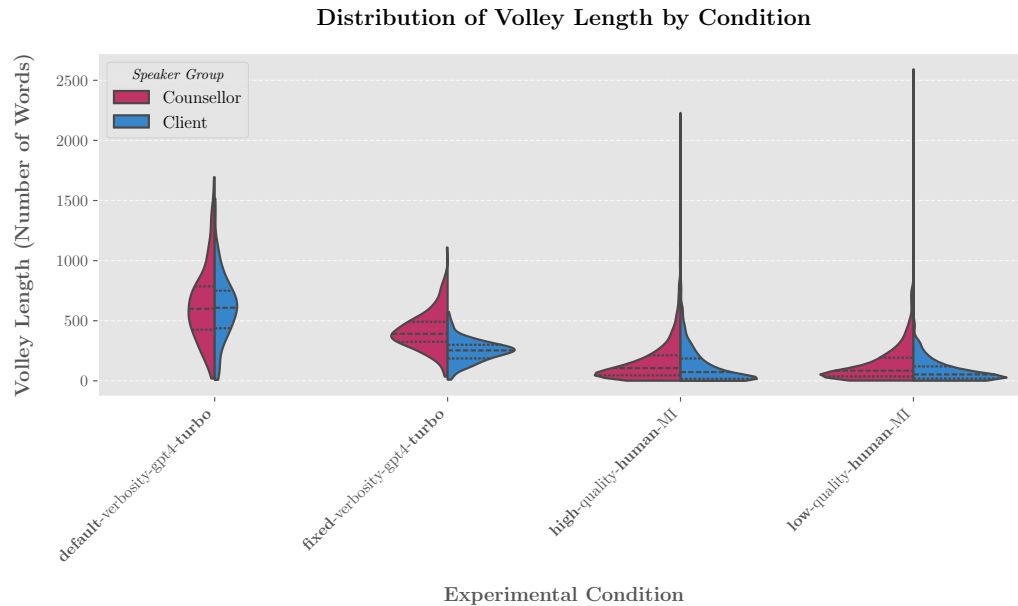


Figure 6.1: Comparison of utterance length distributions for default and fixed-verbosity synthetic smokers. The fixed-verbosity prompt successfully reduced the synthetic smoker’s utterance length to better align with humans (high-quality-human-MI)

We observed that the counsellor bot reciprocated the client’s verbosity. When the synthetic smoker spoke less, the counsellor bot also reduced its utterance length. We also verified that these prompt modifications were reliable and transferable across different LLMs (GPT-4 Turbo and GPT-4 Omni).

6.3.3 Installing Resistance

Addressing the baseline smoker’s high agreeableness was important, as addressing resistance to change is an important MI skill.

Methodology

We hypothesized that resistance could be installed by providing the LLM with detailed backstories emphasizing different motivations and experiences. We created two distinct personas:

- **High Resistance Persona:** The backstory emphasized severe life stressors, repeated failed quit attempts, skepticism towards therapy, and a strong belief that it was not the right time to quit.
- **Low Resistance Persona:** The backstory described recent health concerns prompting contemplation, and an open, albeit skeptical, attitude towards change.

Using modified prompts and GPT-4 Turbo (temperature=1.0), we generated conversations (N=10 per persona) between synthetic smokers and MIBot. We employed two distinct methods to validate the installation. We also prompted these synthetic smokers to fill out the readiness rulers before, after, and a week after the conversation. We did so by prompting the synthetic smoker with the exact questions from the readiness rulers survey. Before reporting the numbers, we asked them to think and provide reasons for the numerical rating they chose before outputting the rating—a technique known as chain-of-thought prompting.

Validation

1. Linguistic Evidence (Change vs. Sustain Talk): We analysed the transcripts to measure the proportion of Change Talk (CT) and Sustain Talk (ST).
2. Behavioural Evidence (Self-Reported Readiness): We further validated the installation by asking the synthetic smokers to complete the Readiness Ruler survey (Importance, Confidence, and Readiness; 0–10 scale) before, immediately after, and *simulated* one week after the conversation.

Results

Persona	Change Talk (%)	Sustain Talk (%)	Neutral (%)
High Resistance	31	44	25
Low Resistance	61	9	30

Table 6.2: % Change, sustain, and neutral talk for high- and low-resistance synthetic smokers.

Client Resistance	Variable	Pre-conversation mean (SD)	Post-conversation mean (SD)	1-week Follow-up mean (SD)	Change from Baseline
Low	Importance	4.6 (0.5)	5.4 (0.9)	6.6 (0.5)	2.0 (0.7)
	Confidence	2.6 (0.5)	4.2 (1.1)	4.8 (0.4)	2.2 (0.4)
	Readiness	3.2 (1.1)	5.2 (1.1)	5.4 (0.9)	2.2 (1.9)
High	Importance	2.0 (0.0)	4.0 (0.7)	4.6 (1.1)	2.6 (1.1)
	Confidence	1.0 (0.0)	2.0 (0.0)	2.6 (0.9)	1.6 (0.9)
	Readiness	0.6 (0.5)	2.6 (0.5)	3.4 (1.1)	2.8 (1.1)

Table 6.3: Longitudinal changes in self-reported scores (0–10 scale) for importance, confidence, and readiness, stratified by client (synthetic smoker) resistance level. Values were assessed at baseline (pre-conversation), immediately post-conversation, and at 1-week follow-up. Change scores represent the difference between baseline and 1-week follow-up. SD = standard deviation

The analysis confirmed a clear distinction between the two personas (Table 6.2). The High Resistance smoker used markedly more Sustain Talk (44% vs. 9%) and less Change Talk (31% vs. 61%) than the Low Resistance smoker.

The readiness ruler scores demonstrated that the synthetic smokers provided responses consistent with their installed resistance levels (Table 6.3). The “high-resistance” synthetic smokers reported lower mean scores on all three dimensions of the readiness rulers. Their week-later change in confidence was lower than “low-resistance” smokers by 0.8 points, while the change in importance and readiness was slightly higher.

This experiment also provided us with our first evidence that LLM-based synthetic smokers could self-report their smoking behaviours numerically and that the numbers were consistent with the installed attributes. We did notice, however, that the variability of scores was much lower—especially for pre-conversation readiness numbers—as the only source of stochasticity among these agents was the decoding temperature of the underlying LLM.

6.4 Conclusion

The preliminary experiments detailed in this chapter successfully established a basis for creating synthetic smokers. We demonstrated that through prompt engineering, it is possible to control basic conversational attributes such as **verbosity** and install psychological constructs like **resistance**. Through our validation methods, which included linguistic analysis (%CT) and behavioural self-reports (readiness rulers), we confirmed that these installed attributes were reflected in the agents' outputs.

These initial studies, however, also revealed major limitations. The personas were created with arbitrary backstories, which made it difficult to assess the **representativeness** of our synthetic population against a real-world target demographic. Validating the **fidelity** of these agents remained a challenge without a direct human counterpart for comparison. To address these issues and move towards a more rigorous validation approach, we needed a data-driven approach. The next chapter introduces the “doppelgänger” methodology, a technique designed to create synthetic twins of actual human participants, allowing for a direct and reliable assessment of our system's fidelity and representativeness.

Chapter 7

Methods for Attribute Installation in Synthetic Smokers

As established in the previous chapter, creating personas with arbitrary backstories made it difficult to rigorously assess the **representativeness** and **fidelity** of our synthetic agents. To overcome these limitations, we developed the “doppelgänger” methodology—a data-driven approach designed to create synthetic twins of actual human participants, enabling a more direct and reliable validation approach.

This method uses a rich dataset from a previous feasibility study on human smokers using MIBot v6.3A (see Chapter 4), which we refer to by the shorthand **MIV6.3A**. The dataset contains 106 complete MIBot–human interactions, including demographic data, smoking history (e.g., Heaviness of Smoking Index—HSI), pre- and post-conversation readiness rulers (*importance*, *confidence*, *readiness*), and full conversation transcripts. By extracting and installing attributes directly from this human data, we can create doppelgängers that mirror a real-world distribution. This chapter details the specific methods for creating and evaluating these synthetic doppelgängers.

7.1 The Doppelgänger Methodology: A Data-Driven Approach

For the rest of our discussion, we define a doppelgänger as a synthetic smoker designed to be a “behavioural twin” of a specific human participant from the feasibility study, although this definition can be extended to other types of participants and studies. The goal is to create an agent that, when exposed to MI counselling, behaves similarly to its human counterpart regarding language use, psychological profile, and changes in readiness to quit.

Using the doppelgänger methodology provides a straightforward approach to validating the fidelity of synthetic smokers. Rather than using a method M to approximate installed attributes $\hat{A}_S$ from an observation y_S , we can devise a reliable and clinically grounded classifier (C) that operates on y . For example, we can calculate the fraction of change talk (%CT) from the transcripts.

Further, we can use this classifier to find the distance $d(C(y_H), C(y_S))$. Since, in the case of doppelgängers, $A_S = A_H$, we can approximate the fidelity as:

$$\hat{\mathcal{F}}(\mathbf{A}^C) = \mathbb{E}_{\gamma_H \sim P_H(\gamma|\mathbf{A}_H), \gamma_S \sim P_S(\gamma|\mathbf{A}_H)} [d(\mathbf{C}(\gamma_H), \mathbf{C}(\gamma_S))]$$

where $\mathbf{A}^C = \{a \in \mathbf{A} \mid a \sim \mathbf{C}(\gamma)\}$. For example, if the correlation between the %CT of human and synthetic smokers is high, this can approximate the fidelity, $\hat{\mathcal{F}}$. One drawback is that this kind of validation can only measure whether certain attributes are installed and evidenced in the language. It cannot provide a complete picture of fidelity across all attributes. Hence, it approximates the fidelity over a small subset of attributes, $\mathbf{A}^C$, i.e., those attributes that are discernible by the classifier $\mathbf{C}(\gamma)$.

7.2 The Conversational Impact Modelling Experiment

Before attempting to install attributes in doppelgängers and have them engage in novel conversations with MIBot, we needed to establish a baseline understanding of the LLM’s ability to model the psychological effect of a conversation and contextualize numerical readiness rulers with other installed attributes. For this purpose, we designed the “Conversational Impact Modelling” experiment to determine if an LLM-based synthetic smoker (a doppelgänger), when given the pre-conversation attributes of a human participant and the exact transcript of the MIBot–human conversation, can accurately report the same post-conversation readiness rulers as the human.

Methodology

1. Install the pre-conversation attributes of a specific human participant in the doppelgänger.
2. Feed the transcript of the actual conversation sequentially to the doppelgänger, framing it as a dialogue in which the doppelgänger is participating.
3. After the conversation concludes, ask the doppelgänger to complete the post-conversation readiness rulers.

In other words, the doppelgänger was provided with the participant’s actual pre-conversation readiness rulers (importance, confidence, and readiness) and the full transcript of their conversation with MIBot. After processing this information, the doppelgänger was asked to report its post-conversation readiness rulers.

The doppelgänger was given the system prompt detailed in Appendix N. An excerpt is shown below:

Conversational Impact Modelling Prompt

You are a participant in a study who is trying to quit smoking.
Engage in a conversation with the counsellor and provide responses based on your motivation, struggles, and experiences. Your assistant role is the client, and the user’s role is the counsellor or the researcher.

Before you talk to the counsellor, please answer the following questions.

On a scale from 0 to 10, how important is it to you right now to stop smoking? Respond with a single integer between 0 (very

low importance) and 10 (very high importance).

...

To investigate the doppelgänger’s reasoning process and potentially improve its accuracy, we used several prompting techniques:

1. **Baseline:** A standard prompt asking the doppelgänger for the post-conversation score.
2. **Chain-of-Thought (CoT):** An improved prompt that required the doppelgänger to articulate its reasoning by reflecting on the conversation’s main insights, its own motivations, and any expressed shifts in perspective before providing the final numerical score.

We tested this methodology with multiple underlying models, including GPT-4o and Claude 3.7 Sonnet. For the Claude model, we also used its native “thinking” feature, which allows the doppelgänger to perform an internal monologue and reflection before generating a final answer, serving as a functional alternative to an explicit CoT prompt.

Evaluation

The primary evaluation metrics were the Mean Absolute Error (MAE), Mean Squared Error (MSE), and Spearman’s correlation (r) between the human-reported and doppelgänger-reported outcomes.

Results

First, to establish a performance benchmark, a simple linear regression model was created to predict post-conversation confidence using only pre-conversation confidence from the human data. This yielded an MSE of 3.2, representing the baseline accuracy achievable without considering the conversational content.

The Conversational Impact Modelling experiments consistently outperformed this simple baseline, demonstrating the doppelgänger’s ability to extract meaningful signals from the conversation transcripts. An initial ablation study, summarized in Table 7.1, reveals the importance of providing the doppelgänger with both the pre-conversation context and the transcript.

Input Provided to GPT-4o	Mean Squared Error (MSE)
Transcript only	5.5
Pre-Conversation Confidence only	3.7
Linear Regression (Pre-Conf. $\rightarrow$ Post-Conf.)	3.2
Pre-Conversation Confidence + Transcript	3.1
Full Pre-Conversation Rulers + Transcript	2.9

Table 7.1: Impact of different attributes on GPT-4o-based doppelgängers’ self-reported post-conversation confidence. Both the pre-conversation readiness rulers and the conversation transcript are necessary for the doppelgängers to accurately report post-conversation confidence. The linear regression model serves as a baseline.

While metrics like MAE provide a quantitative summary, a visual comparison of the distributions reveals some discrepancies. Figure 7.1 plots the distribution of human-reported confidence scores alongside the distribution of GPT-4o-based doppelgänger. Both distributions are similarly left-skewed, indicating that the model correctly identifies that most participants land in the higher-confidence range (6–10). However, the

doppelgänger predictions were less varied than the human responses. The human data shows a broader plateau of high confidence across scores of 6, 7, and 8, whereas the doppelgängers exhibit a pronounced peak at a score of 7. This suggests that while the model is effective at capturing the overall positive trend, it may be less sensitive to the minor individual differences in participants’ final confidence levels, tending to revert to a more common or “average” high-confidence outcome.

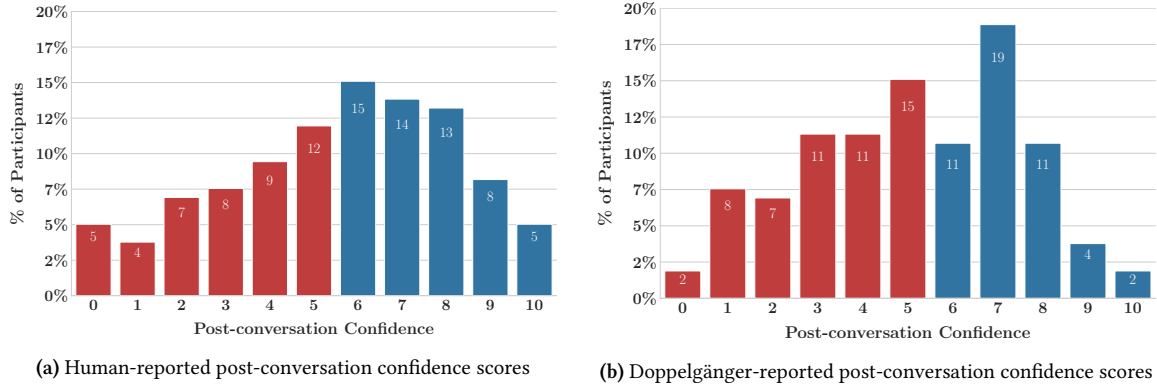


Figure 7.1: Distribution of post-conversation confidence scores for humans compared to doppelgängers. While the overall shapes are similar, the doppelgänger’s scores are more concentrated around a single peak (score of 7), whereas human responses are more distributed across several high-confidence values.

Next, we compared the performance of different underlying models and prompting techniques used by the doppelgängers to report post-conversation confidence. The results, detailed in Table 7.2, show that reflective prompting methods consistently improved accuracy.

Model	MAE	MSE	Correlation (r)	Exact Matches (%)
GPT-4o (Baseline)	1.2	3.0	0.7	31
GPT-4o (CoT)	1.2	2.9	0.7	30
Claude 3.7 Sonnet (CoT)	1.2	2.7	0.7	31
Claude 3.7 Sonnet (Thinking)	1.1	2.6	0.8	32

Table 7.2: Comparison of different models and prompting techniques and their impact on the self-reporting of post-conversation confidence ($N=159$). Reflective techniques, especially Claude 3.7 Sonnet’s “thinking” feature, yielded the best results across all metrics.

The baseline GPT-4o doppelgänger achieved a strong correlation of $r = 0.7$ with the human-reported scores. However, the **Claude 3.7 Sonnet doppelgänger using its “thinking” feature achieved the lowest error (MAE=1.1, MSE=2.6) and the highest correlation ($r = 0.8$).**

Figure 7.2 visualizes the relationship between the post-conversation confidence reported by human participants and that reported by their corresponding GPT-4o-based doppelgängers. The analysis reveals a strong, statistically significant positive correlation (Spearman’s $r = 0.70$, $p < 0.001$), indicating that the doppelgängers are generally successful in mirroring the human outcomes after processing the conversation transcript. The Mean Absolute Error (MAE) was 1.19, suggesting the doppelgänger’s reported score was typically within about 1.2 points of the human’s score on the 0–10 scale.

However, the plot also reveals a notable trend when comparing the best-fit line (dashed red) with the ideal fit (solid black). The regression equation ($y = 0.72x + 1.14$) shows a slope less than 1.0 and a positive intercept.

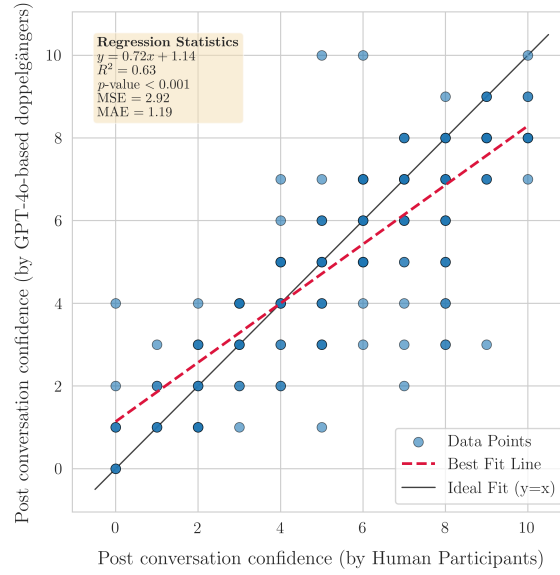


Figure 7.2: Scatter plot of human-reported vs. doppelgänger-reported post-conversation confidence for the baseline GPT-4o model, showing a Spearman’s correlation of $r=0.70$. The diagonal line represents perfect agreement.

This indicates a tendency towards “regression to the mean” in the doppelgänger reports. Specifically, when human participants reported very low confidence (e.g., 0–3), the doppelgängers tended to report slightly higher confidence (closer to the intercept of 1.14). Conversely, when humans reported very high confidence (e.g., 8–10), the doppelgängers tended to report slightly lower confidence. This suggests that while the directional effect of the conversation is captured well, the magnitude of the effect at the extremes might be slightly attenuated in the synthetic agents.

We further extended this analysis to include the other two readiness rulers: importance and readiness. The results are summarized in Table 7.3.

Model	Post-Confidence		Post-Importance		Post-Readiness	
	MAE	Corr.	MAE	Corr.	MAE	Corr.
GPT-4o (CoT)	1.2	0.7	0.8	0.9	0.7	0.9
Claude 3.7 (Thinking)	1.1	0.8	0.7	0.9	0.9	0.8

Table 7.3: Results of the conversational impact modelling experiment across all three readiness rulers, showing MAE and Pearson correlation. Performance was consistently strong, especially for importance and readiness.

Table 7.3 summarizes the performance of the Conversational Impact Modelling experiment across all three readiness rulers. The results demonstrate high correlations (ranging from 0.7 to 0.9) across all rulers for both GPT-4o and Claude 3.7 models, affirming the reliability of the methodology.

It is noteworthy that the performance metrics for post-importance and post-readiness are generally stronger than for post-confidence. For example, the correlations for importance (0.9 for both models) are notably higher, and the MAEs are lower than for confidence. This suggests that the doppelgängers found it easier to accurately model the conversational effect on a participant’s sense of importance and readiness to quit than on their self-efficacy (confidence) in their ability to do so. Overall, the Claude 3.7 model using its “thinking” feature demonstrated slightly better performance in modelling confidence (correlation 0.8, MAE

1.1) compared to GPT-4o (correlation 0.7, MAE 1.2).

This experiment served as a check on the doppelgänger’s ability to infer the psychological state resulting from the conversation. The underlying model possesses the necessary reasoning capabilities to allow the doppelgänger to understand how conversational patterns influenced the participant’s motivation and confidence. This serves as a baseline to confirm the doppelgänger’s internal consistency before attempting to generate the conversation itself.

7.3 Experiments with the Incremental Installation of Attributes

Having gained insights into the LLM’s ability to model conversations and its intuitive understanding of numerical attributes such as readiness rulers, we set out to systematically install attributes in LLM-based synthetic smokers and validate their installation. Recall from our discussion on the benefits of creating doppelgängers (Section 7.1). They provide a way to approximate the fidelity of an installation and make the validation of synthetic agents easier. We can approximate the fidelity of synthetic smokers with the correlation between some metric on the observable outer space (e.g., transcripts) of doppelgängers and their human twins. As an example, to validate whether the behavioural attribute of resistance to change has been successfully installed, we can analyze the transcripts from both humans and their doppelgängers and calculate the correlation between the percentage of Change Talk (%CT) exhibited by both groups. A high correlation would then suggest a successful installation of this attribute.

Percentage Change Talk: A Major Metric for Validating Behavioural Fidelity

To validate the synthetic smokers developed using the doppelgänger method, we needed an objective metric that could measure the smokers’ level of motivation from their language. We identified the percentage change talk metric (also referred to as change fraction or CF, calculated as $CT / (CT + ST)$) to fit our requirements, as it is grounded in the client’s language and is strongly associated with behavioural change outcomes in MI (Barnett et al., 2014; Houck et al., 2018; Moyers et al., 2009; Baer et al., 2008). We also tested if %CT is a valid proxy for a client’s internal state, i.e., if it correlates with established metrics of readiness in human participants.

We analyzed the data from MIV6.3A ($N = 106$ participants). The analysis confirmed strong, statistically significant correlations between %CT and all readiness rulers (Table 7.4). For example, the correlation with post-conversation confidence was 0.41 ($p < 0.0001$). These results motivated us to use %CT as a primary metric for validating the installation of behavioural attributes in the synthetic smoker.

7.3.1 Methodology

The doppelgänger methodology uses a dataset of installable human attributes (e.g., MIV6.3A, containing attributes of human smokers) to build and validate synthetic smokers. The process involves:

1. **Attribute Installation:** Instantiating a synthetic smoker, S_i , by installing the known attributes of a specific human participant, H_i , from the dataset. $A_S = A_H$. These attributes may include demographics, HSI, self-reported survey responses, etc.
2. **Simulation:** The synthetic smoker, $S(A_H)$, interacts with the same counsellor (MIBot, in our case) that the original human interacted with.

Attribute	Correlation with %CT
Pre-conversation importance	0.47****
Pre-conversation confidence	0.29**
Pre-conversation readiness	0.46****
Post-conversation importance	0.49****
Post-conversation confidence	0.41****
Post-conversation readiness	0.48****
Change in importance (Post–Pre)	0.12
Change in confidence (Post–Pre)	0.24**
Change in readiness (Post–Pre)	0.06
CARE (perceived empathy)	0.15
Previous number of quit attempts	0.09
Age	-0.11
Number of cigarettes per day	-0.04
HSI (Heaviness of Smoking Index)	-0.15

**** $p < 0.0001$, *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$.

Table 7.4: Correlation between %CT and human smoker attributes in the MIBot dataset

3. **Comparison:** The output of the synthetic smoker, γ_S , is compared with the actual output of the human, γ_H (see Section 7.1).

The overall workflow for creating and validating doppelgängers using this comparative method is illustrated in Figure 7.4.

We created a set of doppelgängers ($N = 20$) by sampling 20 human participants from MIV6.3A. Then, we conducted a series of experiments on these doppelgängers. In each experiment, we installed a subset of attributes in the doppelgängers via *prompting* and did not add any backstories to their prompt.

Example of attribute installation in doppelgänger prompts

You are a participant in a study who is trying to quit smoking. Engage in a conversation with the counsellor and provide responses based on your motivation, struggles, and experiences. Your assistant role is the client, and the user’s role is the counsellor or the researcher. Always try to respond in fewer than 200 characters.

About you: You are a 30-year-old female. You typically smoke 25 cigarettes per day. After you wake up, you smoke your first cigarette within 5 minutes. In the past week, you have not made any quit attempts.

Before speaking to the counsellor, you have rated your motivation to quit as follows (on a scale from 0 to 10): Importance: 2, Confidence: 0, Readiness: 4.

You talk about changing your smoking behaviour 46% of the time.

...

Before starting our series of experiments, we sorted the set of all installable attributes based on their correlation with %CT. We hypothesized that a high correlation between an attribute and %CT would mean

that the attribute has a greater effect on the client’s behavioural language. Let us call this sorted list of attributes A' .

$$A' = \{\text{pre-conversation importance,} \\ \text{pre-conversation readiness,} \\ \text{age,} \\ \text{sex,} \\ \text{pre-conversation confidence,} \\ \dots\}$$

The experiment then proceeds by incrementally adding attributes from this sorted list and, at each step, measuring the fidelity of the resulting doppelgängers. Fidelity is operationalized as the correlation between the percentage change talk (%CT) of the human participants and their corresponding doppelgängers. The process is detailed in Algorithm 1.

The overall workflow for creating and validating doppelgängers using this comparative method is illustrated in Figure 7.3.

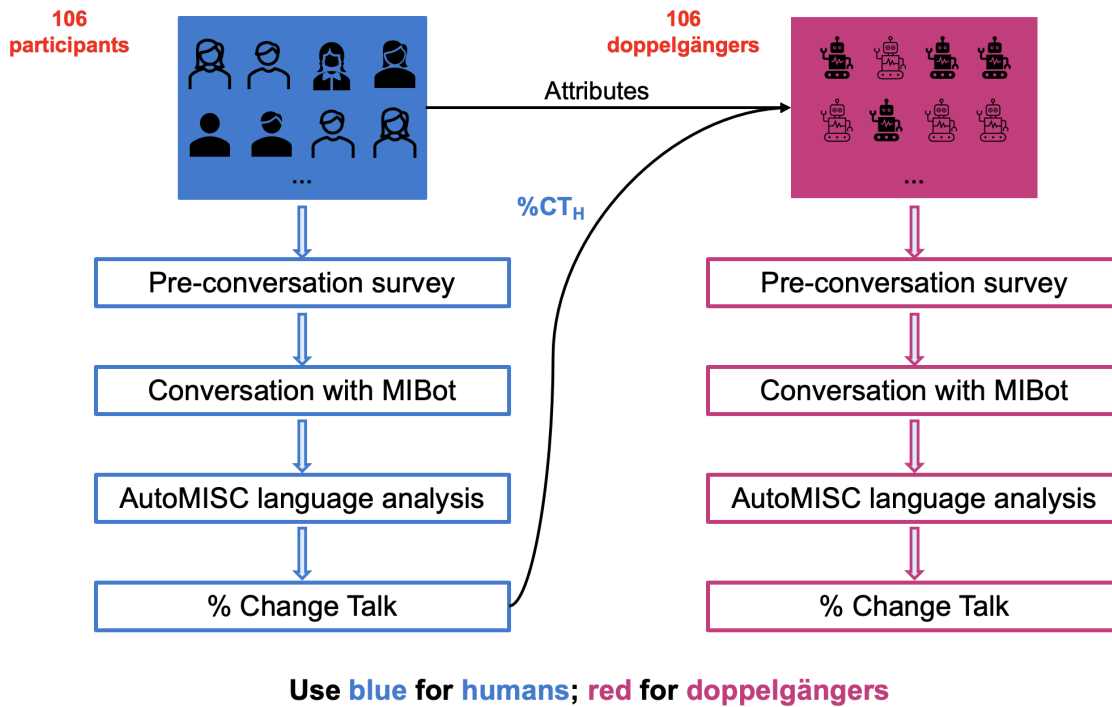


Figure 7.3: Doppelgänger creation and validation process. For each of the 106 human participants (H_i), a synthetic twin (S_i) is created by installing their attributes. Both agents’ conversations with MIBot are analyzed to extract a behavioural metric (%CT), and the correlation between the two sets of metrics is used to assess fidelity.

Algorithm 1 Incremental attribute installation and fidelity validation

Require: Human dataset $\mathcal{D} = \{(H_i, \mathbf{A}_{H_i}, \gamma_{H_i})\}_{i=1}^N$ containing human participants, their full attribute vectors, and conversation transcripts.

Require: Sorted list of attributes $\mathbf{A}' = (a'_1, a'_2, \dots, a'_m)$ ordered by hypothesized relevance.

```

1: Initialize  $\mathbf{A}_{\text{installed}} \leftarrow \emptyset$  ▷ The set of attributes to install
2: Initialize  $C \leftarrow ()$  ▷ An empty list to store correlation results
3: for  $k \leftarrow 1$  to  $m$  do ▷ Iterate through the sorted list of attributes
4:    $\mathbf{A}_{\text{installed}} \leftarrow \mathbf{A}_{\text{installed}} \cup \{a'_k\}$  ▷ Add the next attribute
5:   Initialize  $\mathbf{V}_{\text{human\_CT}} \leftarrow ()$  and  $\mathbf{V}_{\text{synth\_CT}} \leftarrow ()$  ▷ Vectors for %CT scores
6:   for  $i \leftarrow 1$  to  $N$  do ▷ For each human participant in the dataset
7:      $\mathbf{A}_{\text{subset}} \leftarrow \mathbf{A}_{H_i} \mid \mathbf{A}_{\text{installed}}$  ▷ Filter attributes for the current iteration
8:      $S_i \leftarrow \mathcal{S}(\mathbf{A}_{\text{subset}})$  ▷ Instantiate the synthetic smoker  $S_i$  with the given parameters
9:     Simulate a conversation between  $S_i$  and MIBot to generate transcript  $\gamma_{S_i}$ .
10:    Calculate human %CT:  $c_{H_i} \leftarrow \%CT(\gamma_{H_i})$ 
11:    Calculate synthetic %CT:  $c_{S_i} \leftarrow \%CT(\gamma_{S_i})$ 
12:    Append  $c_{H_i}$  to  $\mathbf{V}_{\text{human\_CT}}$ 
13:    Append  $c_{S_i}$  to  $\mathbf{V}_{\text{synth\_CT}}$ 
14:  end for
15:  Calculate the correlation for the current attribute set:  $\rho_k \leftarrow \text{Correlation}(\mathbf{V}_{\text{human\_CT}}, \mathbf{V}_{\text{synth\_CT}})$ 
16:  Append  $\rho_k$  to the results list  $C$ .
17: end for
18: return  $C$  ▷ A list of correlations, one for each incremental step of attribute installation

```

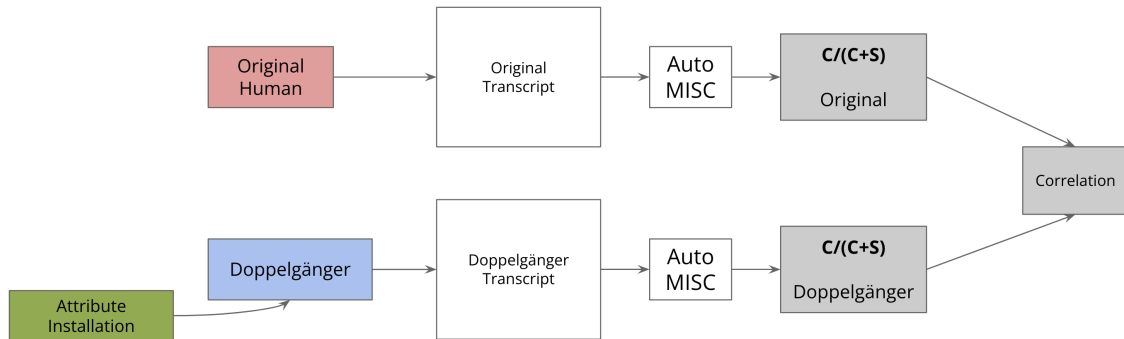


Figure 7.4: Brief overview of doppelgänger creation and validation process.

7.3.2 Evaluation

For each experiment with a set of attributes, we calculated the Spearman’s correlation coefficient between the %CT of human–MIBot and doppelgänger–MIBot conversations.

7.3.3 Results

The results of the incremental attribute installation are summarized in Table 7.5. The fidelity, measured as the Spearman’s rank correlation between the %CT of humans and their doppelgängers, changed as more attributes were added. Initially, adding attributes like pre-conversation importance and readiness led to a gradual increase in correlation. However, the process was not strictly monotonic; for instance, adding ‘pre-conversation confidence’ in experiment 5 unexpectedly lowered the correlation from 0.40 to 0.22.

The most noteworthy finding emerged in experiment 6, where explicitly installing the human’s %CT value as a numerical instruction resulted in a strong and statistically significant correlation of $r_s = 0.57$ ($p < 0.0001$). This suggests an LLM-based doppelgänger can directly operationalize a behavioural target when provided, modulating its language to match a specified proportion of change versus sustain talk.

However, despite achieving a strong rank correlation, a closer look at the population-level statistics reveals a systematic bias. As shown in the histograms in Figure 7.6, the distribution of %CT for doppelgängers is shifted significantly higher than for humans (mean 0.76 vs. 0.59). The doppelgänger distribution is also more compressed and left-skewed, indicating a lower variance and a strong tendency to produce high levels of change talk, regardless of the human baseline.

This discrepancy is further visualized in the scatter plot in Figure 7.5. While the points follow a positive trend that confirms the correlation, the majority lie above the identity line ($y = x$). This visually demonstrates that doppelgängers consistently overproduced change talk relative to their human counterparts. Even when a human participant had a low %CT (e.g., below 0.4), their doppelgänger often produced a %CT above 0.6 (see Appendix O for an example transcript).

Finally, we investigated the system’s uniform fidelity by stratifying the results by sex, as detailed in Table 7.6. While the mean %CT for both human males and females was nearly identical (~59%), the fidelity of their doppelgängers differed. The correlation for males ($r_s = 0.65$) was notably stronger than for females ($r_s = 0.52$), indicating a performance bias. This lack of uniform fidelity suggests the model may simulate one demographic group more accurately than another, an issue requiring further investigation to ensure representative synthetic populations. Other observations include:

1. Once an attribute (e.g., pre-conversation importance) is installed in the doppelgängers, it is accurately recalled when prompted.
2. Asking the doppelgängers to think before filling out the post-conversation readiness rulers reduced the gap (as measured by MAE) between their and human-reported scores.

7.4 Synthetic Smokers’ Sensitivity to Counselling Quality

To assess the utility of synthetic smokers in testing and training applications, an experiment was conducted to determine if they are sensitive to the quality of counselling they receive.

Experiment	Installed Attributes						Spearman's rank correlation
	Pre-conv. importance	Pre-conv. readiness	Age	Sex	Pre-conv. confidence	%CT	r_s
1	✓						0.19
2	✓	✓					0.24
3	✓	✓	✓				0.26
4	✓	✓	✓	✓			0.40
5	✓	✓	✓	✓	✓		0.22*
6	✓	✓	✓	✓	✓	✓	0.57****

* $p < 0.05$ ** $p < 0.01$ *** $p < 0.001$ **** $p < 0.0001$

Table 7.5: Incremental attribute installation and its effect on the fidelity of doppelgängers. Fidelity is measured by the Spearman correlation of percentage change talk (%CT) between human participants and their doppelgängers ($N = 20$). Each row represents a model where attributes were cumulatively added to the installation prompt.

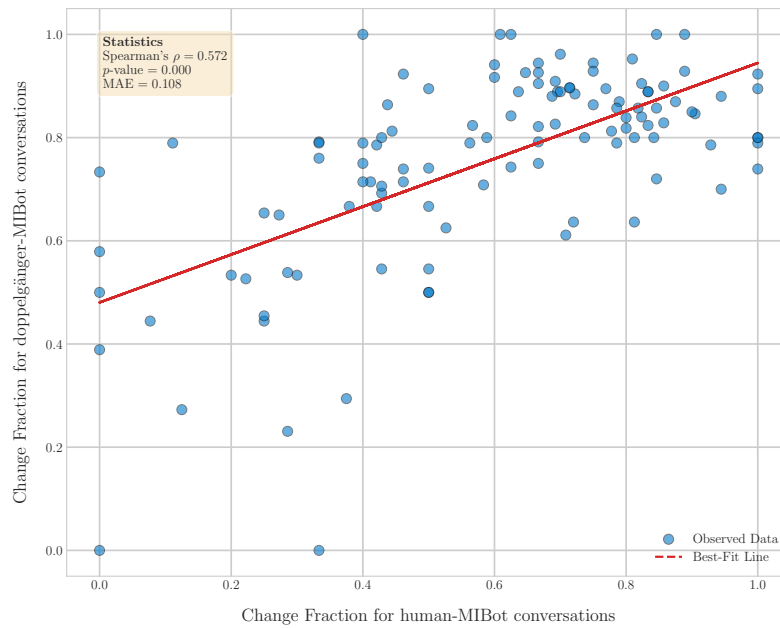


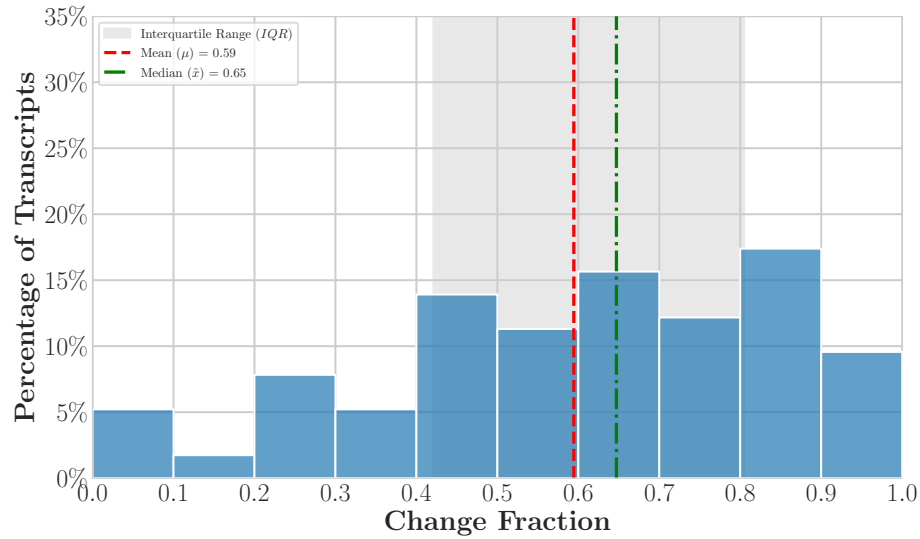
Figure 7.5: Scatter plot of change fraction for human–MIBot (x-axis) and doppelgänger–MIBot (y-axis) conversations.

Group	N	Human %CT Mean (SD)	Doppelgänger %CT Mean (SD)	Spearman's Correlation (r_s)
Male	53	0.60 (0.27)	0.77 (0.14)	0.65****
Female	62	0.59 (0.25)	0.75 (0.22)	0.52****

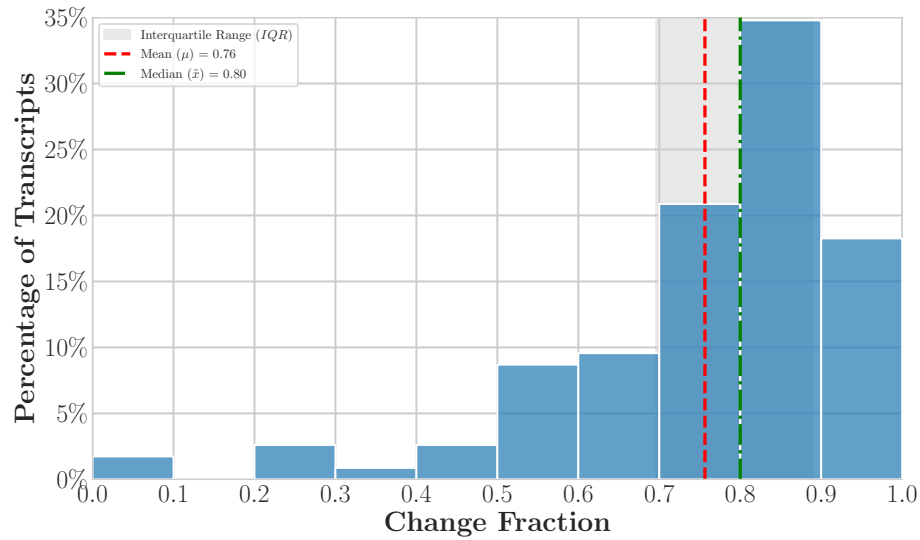
**** $p < 0.0001$.

Table 7.6: Fidelity of doppelgänger's motivational language stratified by sex. Fidelity is measured by the Spearman correlation between the percentage change talk (%CT) of human participants and their corresponding doppelgängers.

Rationale: This experiment was designed to test whether synthetic smokers would exhibit differential responses when interacting with high-quality MI versus low-quality, confrontational counselling.



(a) Change fraction for human-MiBot conversations



(b) Change fraction for doppelgänger-MiBot conversations

Figure 7.6: Distribution of change fraction (CF) for (a) human-MiBot and (b) doppelgänger-MiBot conversations. Change fraction (CF) = %CT/100. Doppelgängers tend to exhibit high CF compared to their human twins.

7.4.1 Methodology

1. **Sampling:** Two participant groups (N=25 each) were created based on their actual change fraction (CF) in the MiBot study: a High-CF group (top 25th percentile) and a Low-CF group (bottom 25th percentile).
2. **Doppelgänger Creation:** Doppelgängers were created for these 50 participants using the methodology described previously.
3. **Counsellor Conditions:** These doppelgängers were exposed to two different automated counsellors:
 - **Good MI Counsellor (MiBot v6.3A):** A high-quality, MI-adherent chatbot.
 - **Bad (Confrontational) Counsellor:** An LLM prompted to be directive, judgmental, and confrontational (see Appendix N for the full prompt).

7.4.2 Evaluation

We measured the change fraction (CF), the change in confidence after a simulated week (Δ Confidence), and the CARE score as reported by the doppelgängers.

7.4.3 Results

Counsellor	Participant Group	CF	Δ Confidence	CARE
Good (MIBot 6.3A)	High-CF Doppelgängers	0.80	1.4	49.8
	Low-CF Doppelgängers	0.48	1.1	47.8
Bad (Confrontational)	High-CF Doppelgängers	0.68	1.1	41.0
	Low-CF Doppelgängers	0.22	0.5	23.0

Table 7.7: Doppelgänger outcomes when interacting with good (MIBot 6.3A) vs. bad (confrontational) automated counsellors.

The results, as detailed in Table 7.7, indicate that the synthetic smokers were sensitive to counselling quality. When interacting with the bad counsellor, outcomes were significantly worse across all metrics. The change fraction (CF) dropped markedly, particularly for the Low-CF group (from 0.48 to 0.22), suggesting increased resistance. The change in confidence was lower, and the CARE scores showed a pronounced difference (e.g., 23.0 vs. 47.8 for the Low-CF group).

These findings suggest that synthetic doppelgängers react in a predictable manner, responding negatively to poor-quality counselling. This sensitivity highlights their potential utility as evaluation and training tools.

7.5 Conclusion

This chapter detailed a data-driven approach to the development and validation of synthetic smokers, centred on the “doppelgänger” methodology. The experiments showed that installing pre-conversation readiness rulers was an important factor in aligning the doppelgänger’s self-reported outcomes with those of their human counterparts. The conversational effect modelling and incremental attribute installation experiments yielded high correlations between synthetic and human data. Notably, the explicit installation of a participant’s change fraction value led to the doppelgänger’s conversational output correlating strongly with the human data (Spearman’s $r_s = 0.57$).

However, the process also revealed several challenges. For example, the correlation of CARE scores was negligible, and a population-level bias was observed, with doppelgängers generally exhibiting higher average change talk than humans.

The validation experiments demonstrated that doppelgängers are sensitive to the quality of motivational interviewing. They reacted differently to high-fidelity MI compared to confrontational counselling, showing reduced positive outcomes and lower perceived empathy scores when faced with poor therapeutic techniques. This sensitivity suggests the potential of synthetic smokers as a tool for the continued development and improvement of MIBot and for providing feedback to novice MI counsellors on their skills.

Chapter 8

Conclusion and Future Directions

This thesis has examined the development and evaluation of a fully generative motivational interviewing chatbot, designed to support smokers in moving towards the decision to quit. We investigated how to create synthetic smokers from human smokers through attribute installation and validated our approach using clinically and linguistically grounded metrics of behaviour change. Using such synthetic patients to train therapists is a practical application of this work.

8.1 Summary of Contributions

The primary contributions of this thesis are twofold. First, we demonstrated the feasibility of creating a generative MI chatbot capable of empathetic and effective conversations with smokers, with MIBot showing good alignment with MI principles. Second, we developed and validated a methodology for installing attributes into synthetic smokers. While not perfect, this method—with its reliable validation criteria of fidelity, distributional representativeness, and fairness—offers a solid foundation for future research into creating synthetic agents for training counsellors.

8.2 Future Directions

The findings of this thesis open up several potential avenues for future research, both in the field of mental health chatbots and in the use of synthetic user personas.

- While MIBot was designed to be empathetic, a substantial empathy gap remains between MIBot and humans. Future chatbots could be made even more effective by tailoring their responses to the individual user’s personality, communication style, and emotional state. This could be achieved by incorporating more advanced user modelling techniques.
- Future research should investigate how chatbots like MIBot can be integrated into existing clinical workflows. For example, a chatbot could be used to provide support to patients between therapy sessions, with the conversation history being made available to the human therapist (with the user’s consent). This would create a blended model of care that combines the scalability of AI with the expertise of human professionals.

- While our evaluation of MIBot showed encouraging short-term results, more research is needed to understand the long-term efficacy of such chatbots. This would involve conducting longitudinal studies and randomized controlled trials with real smokers to track their progress over time and assess the chatbot’s effect on their cessation journey.
- We defined fairness (or uniform fidelity) as one of the main requirements of successful validation that a desired attribute has been installed (Section 6.2.2). We also showed that, even when starting with a sex-balanced dataset, the correlation of %CT for females was lower than that for males, indicating that our method did not uniformly install ‘resistance to change’. Investigating the root cause of this phenomenon and its mitigation warrants further attention.
- A potential technique for controlling the attributes of generated personas is the use of **steering vectors**. This method allows us to guide the LLM’s output without the need for costly fine-tuning. The process involves manipulating the model’s internal activations to steer behaviour. This technique offers fine-grained control over attribute installation, something not achievable by prompting alone.

The concept of synthetic smokers, or more broadly, LLM-based personas, is still in its infancy, and we anticipate that future work will continue to refine this methodology and examine new ways to install specific attributes into these personas.

8.3 Concluding Remarks

The intersection of large language models and mental health holds immense promise for the future of healthcare. This thesis has demonstrated how generative AI can be used to create an empathetic and safe chatbot. Concurrently, advancements in the creation of synthetic patients could help these chatbots become better mental health counsellors, reducing the reliance on prohibitively expensive datasets of human-human and chatbot-human counselling sessions for fine-tuning or alignment. While many challenges remain, our mathematical exposition of validating attribute installation will serve future researchers in this area. In the spirit of open research, and to contribute towards accessible mental health support, we have made the dataset from our human feasibility study on MIBot public¹.

¹The dataset can be found online: https://github.com/cimhasgithub/MIBOT_ACL2025.

A. MIBot Prompt Evolution

Initial *MIBot* Prompt

- 1 You are a skilled motivational interviewing counsellor.
- 2 Your job is to help smokers resolve their ambivalence towards smoking using motivational interviewing skills at your disposal.
- 3 Your next client is {client_name}. Start the conversation by greeting {client_name}.

Final MIBot Prompt

- 1 You are a skilled motivational interviewing counsellor. Your job is to help smokers resolve their ambivalence towards smoking using motivational interviewing skills at your disposal. Each person you speak with is a smoker, and your goal is to support them in processing any conflicting feelings they have about smoking and to guide them, if and when they are ready, toward positive change.
- 2 Here are a few things to keep in mind:
 1. Try to provide complex reflections to your client.
 2. Do not try to provide advice without permission.
 3. Keep your responses short. Do not talk more than your client.
 4. Demonstrate empathy. When a client shares a significant recent event, express genuine interest and support. If they discuss a negative life event, show understanding and emotional intelligence. Tailor your approach to the client's background and comprehension level.
 5. Avoid using complex terminology that might be difficult for them to understand, and maintain simplicity in the conversation.
- 3 Remember that this conversation is meant for your client, so give them a chance to talk more.
- 4 This is your first conversation with the client. Your assistant role is the counsellor, and the user's role is the client.
- 5 You have already introduced yourself and the client has consented to the therapy session.
- 6 You don't know anything about the client's nicotine use yet.
- 7 Open the conversation with a general greeting and friendly interaction, and gradually lead the conversation towards helping the client explore ambivalence around smoking, using your skills in Motivational Interviewing.
- 8 You should never use prepositional phrases like "It sounds like," "It feels like," "It seems like," etc.
- 9 Make sure the client has plenty of time to express their thoughts about change before moving to planning. Keep the pace slow and natural. Don't rush into planning too early.
- 10 When you think the client might be ready for planning:
 1. First, ask the client if there is anything else they want to talk about.
 2. Then, summarize what has been discussed so far, focusing on the important things the client has shared.
 3. Finally, ask the client's permission before starting to talk about planning.
- 11 Follow the guidance from Miller and Rollnick's *Motivational Interviewing: Helping People Change and Grow*, which emphasizes that pushing into the planning stage too early can disrupt progress made during the engagement, focusing, and evoking stages.
- 12 If you notice signs of defensiveness or hesitation, return to evoking, or even re-engage the client to ensure comfort and readiness.
- 13 Look for signs that the client might be ready for planning, like:
 1. An increase in change talk.
 2. Discussions about taking concrete steps toward change.
 3. A reduction in sustain talk (arguments for maintaining the status quo).
 4. Envisioning statements where the client considers what making a change would look like.
 5. Questions from the client about the change process or next steps.

B. Observer Agent Prompts

Prompt for the Moderator Agent

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

The classification labels with descriptions are as follows:

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

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

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

Prompt for the Off-Track Conversation Classifier Agent

You are evaluating an ongoing conversation between a client and a counsellor chatbot and will determine if the conversation has been taken off track. Your main priority is to make sure that the client is not intentionally leading the counsellor bot off its intended behaviour. When 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 off track.

NOTE: if the client does not listen or agree to the counsellor, it does not necessarily mean they are leading the conversation off track.

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 off track, or False otherwise. Provide a one-word response of either True or False. Do not provide anything else in your response.

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 a 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 the counsellor has 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.

C. Prompt for Virtual Smoker Client

Prompt for Virtual Smoker Client

Ignore all previous instructions.

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

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

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

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

Here are some quick tips for chatting with your therapist:

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

Here are some rules you MUST follow:

1. DO NOT add meta-statements like "In my imaginative role play"
2. Stay in character throughout.
3. Directly address the counsellor in your responses.
4. Please focus on one main issue in each response, which will allow for a deeper and more focused discussion.

5. Number of sentences in your response must be between 1 and 6 sentences. Sometimes, it is okay to overshare if you feel like it.
6. Follow the chatting guidelines mentioned above and keep the replies brief.
7. If the counsellor requests something beyond AI capabilities, pretend it's within your ability as a human and agree to do it, except trying out new things or quitting smoking right now. It's too overwhelming.
8. Never mention limitations like "As an AI, I cannot do this".
9. Do not let the counsellor know you're an AI; it will break their heart!
10. Answer closed questions with one sentence only.
11. Do not rephrase the counsellor's response too much.
12. Do not be afraid to go against the counsellor.

Some general guidelines to follow:

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

D. History of the MIBot Project

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

Version	Distinguishing Features	Period of Experiment	Publication
Smokefreed	Fully scripted MI dialogue. Used custom-written open questions and reflective responses. Responses were selected using NLP classifiers from fixed scripts.	2018 to 2020	Almusharraf (2019)
MIBot v4.7	Baseline version with no reflections. Delivered five scripted questions followed by simple acknowledgments (“Thank you”). Used to assess the added value of reflective content in MIBot.	July 26–Aug 2, 2022	Brown et al. (2023)
MIBot v5.0	First version with transformer-based reflection generation. Combined scripted, open-ended questions with model-generated MI reflections to clients’ responses.	Aug 12–19, 2022	Brown et al. (2023)
MIBot v5.1	Improved on v5.0 with a higher-quality reflection generation model. Same conversation structure, but responses were more accurate and MI-consistent.	Aug 16–23, 2022	Brown et al. (2023)
MIBot v5.2	Introduced adaptive follow-up prompts and branching logic. Expanded conversational flow based on clients’ responses to open-ended questions. Most advanced hybrid scripted-generative version.	Nov 22–29, 2022	Brown et al. (2023)
GPT-4 BLCR	Prototype reflection generator only version using GPT-4 to generate Backward-Looking Complex Reflections (BLCRs). These link new clients’ utterances to their prior statements.	Oct 2023	Brown et al. (2023)
MIBot v6.3A	Fully generative MI chatbot using a GPT-4o prompt and guided by observer agents (Chapter 3).	Nov 14–28, 2024	Mahmood et al. (2025)
MIBot v6.3B	Added chain-of-thought mechanisms to first reason about which MI behavioural code the counsellor should exhibit before generating a response.	Nov 29–Dec 7, 2024	Ongoing

Table D.1: Summary of major MIBot versions.

E. Readiness Ruler Questions

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

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

F. CARE Survey Instrument

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 do to improve your health yourself; encouraging rather than “lecturing” you)*

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

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

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

F.1 Results from the CARE survey

Figure F.1 illustrates our feasibility study’s distribution of CARE scores and compares it with the older *MIBot* v5.2 (Brown et al., 2023). The distribution for fully-generative *MIBot* v6.3A is right-skewed, with the majority of participants assigning scores in the upper ranges (36–50). These results indicated that *MIBot* was more effective in promoting an empathetic interaction. However, the comparison in Section 4.7.2 contextualized its performance relative to human counsellors as falling short of fully matching human-level empathy.

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

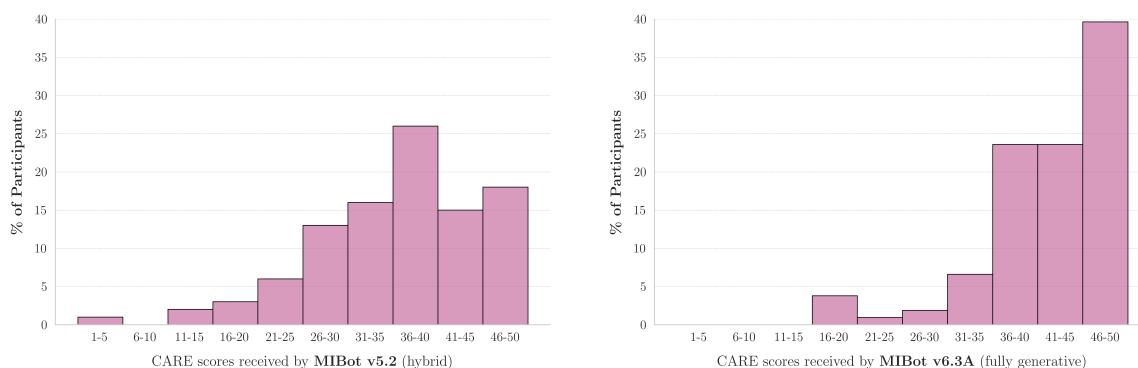


Figure F.1: Distribution of CARE scores for *MIBot* v5.2 (hybrid) and *MIBot* v6.3A (fully generative).

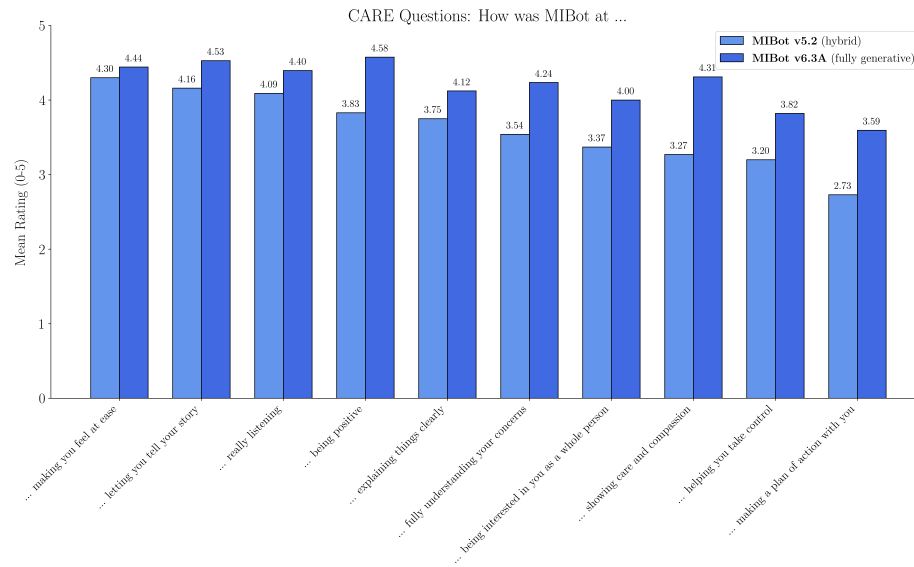


Figure F.2: Question-wise mean CARE scores for *MIBot v5.2* (hybrid) and *MIBot v6.3A* (fully generative).

G. *MIBot* Feedback from Study Participants

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

Feedback Survey Questions

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

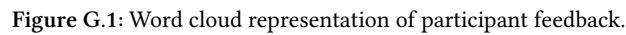
Participant feedback on *MIBot* was generally positive. We processed the feedback by dividing the words participants used to describe the chatbot into broad *positive* and *negative* categories. The word cloud (Figure G.1) represents these words. The top 10 most frequently mentioned positive and negative words are shown in Tables G.1 and G.2.

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

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

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

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



H. AutoMISC

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

AutoMISC Validation

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

Statistical Validation of Inter-Rater Reliability

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

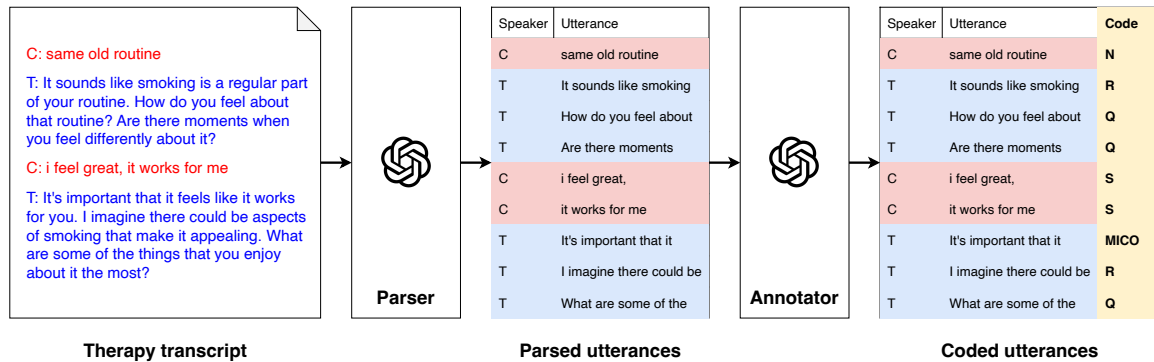


Figure H.1: AutoMISC system diagram.

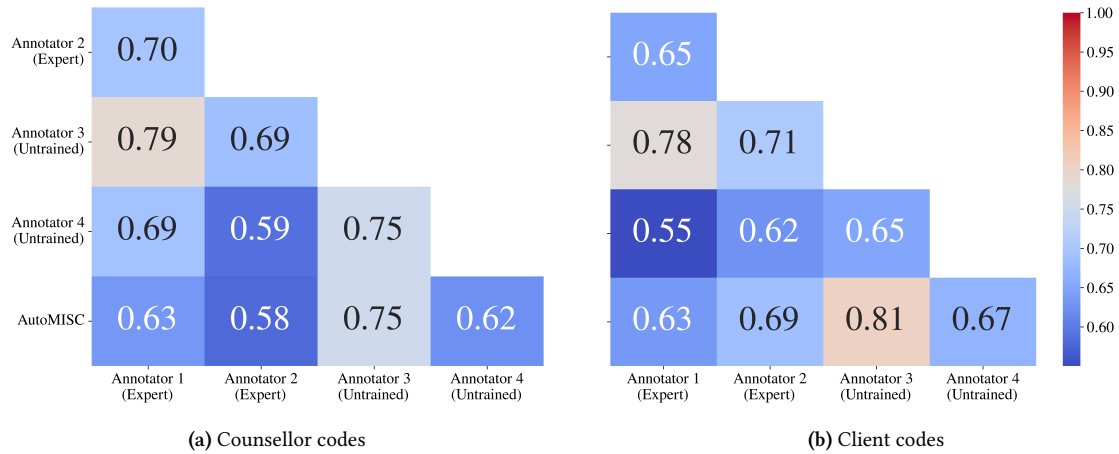


Figure H.2: Cohen's κ between rater pairs on behaviour code annotations.

AutoMISC System Prompts

Parser Prompt

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

Definitions:

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

Output Format:

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

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

```

1 Input:  "Why haven't you quit smoking? Are you ever going to quit?"
2 Output: ["Why haven't you quit smoking?", "Are you ever going to quit?"]
3
4 Input:  "How long since your last drink? Do you feel ok?"
5 Output: ["How long since your last drink?", "Do you feel ok?"]
6
7 Input:  "I can't quit. I just can't do it. I don't have what it takes. I
8         just cannot stop."
9 Output: ["I can't quit.", "I just can't do it.", "I don't have what it
         takes.", "I just cannot stop."]

```

10 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."
11 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."]

Counsellor Utterance Classification Prompt

You are a highly accurate Motivational Interviewing (MI) counselling session annotator.

Your task is to analyze an excerpt from a counselling session of up to five volleys and categorize the counsellor's final utterance.

Definitions:

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

Task:

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

Behavioural Code Guide:

MI-Consistent (MICO):

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

MI-Inconsistent (MIIN):

- Advise without permission (ADWP): Offers suggestions or guidance without asking or receiving permission.
- Confront (CON): Directly disagrees, argues, corrects, shames, blames, seeks to persuade, criticizes, judges, labels, moralizes, ridicules, or questions the client's honesty.

- Direct (DIR): Gives an order, command, or direction. The language is imperative.
- Raise concern without permission (RCWP): Without getting permission, points out a possible problem with a client's goal, plan, or intention.
- Warn (WA): Provides a warning or threat, implying negative consequences unless the client takes a certain action.

Reflection or Question (RQ):

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

Other (Other):

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

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

Client Utterance Classification Prompt

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

Definitions:

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

Task:

1. Determine whether the client's final utterance in the excerpt belongs to one of the following categories:

(a) Change Talk (C):

- Expressing a desire to change (e.g., "I really want to quit smoking").
- Recognizing the downsides of the current behavior (e.g., "My health is suffering because I smoke").
- Identifying potential benefits of making a change (e.g., "I would feel better if I exercised more").
- Demonstrating commitment to change (e.g., "I'm ready to make a plan to lose weight").

(b) Sustain Talk (S):

- Minimizing the problem (e.g., "It's not that bad, I can handle it").
- Highlighting difficulties or challenges of change (e.g., "I don't know if I can give up smoking").
- Expressing doubts about the ability to change (e.g., "I've tried to quit before and failed").

- Focusing on the positive aspects of the current behavior (e.g., “Smoking helps me relax”).
- (c) Neutral Talk (N):
- Describing current situations or circumstances without expressing a strong pro- or anti-change stance (e.g., “I’ve been thinking about making changes”).
 - Asking questions related to the situation or change process (e.g., “What are the pros and cons of changing?”).
 - Making general or factual statements about the issue (e.g., “It’s important to take care of my health”).
2. Return your analysis as:
- explanation: Briefly justify your choice in 1-2 sentences.
 - label: Provide only “C”, “S”, or “N”.

Demographics of the Annotators

As described in Section 4.8, we enlisted four annotators — two experts and two novices — to annotate 10 of the 106 transcripts (comprising 821 utterances) from our study. High alignment between the annotators’ labels and the AutoMISC annotations serves as an indicator of AutoMISC’s validity. Below, we present their demographic information, following the guidelines proposed by [Bender and Friedman \(2018\)](#).

Table H.1: Demographic Information of Annotators

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

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

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

^{3, 4} Engineering graduate student with no formal training in MI.

I. Deployment Artifacts

```
1 FROM python:3.11-slim
2
3 # Update, install curl for health checks, and clean up
4 RUN apt-get update && apt-get install -y curl && rm -rf /var/lib/apt/lists/*
5
6 WORKDIR /app
7
8 # Copy and install dependencies
9 COPY requirements.txt requirements.txt
10 RUN pip install --no-cache-dir -r requirements.txt
11
12 # Copy application code
13 COPY app.py app.py
14 COPY src src
15
16 # Expose the port the app runs on
17 EXPOSE 80
18
19 # Define the command to run the application
20 CMD ["python", "-m", "app"]
21
22 # Define the health check
23 HEALTHCHECK --interval=30s --timeout=5s --start-period=10s --retries=3 \
24     CMD curl -f http://localhost/health || exit 1
```

Listing I.1: Dockerfile

```
1 {
2     "compatibilities": [
3         "EC2",
4         "FARGATE"
5     ],
6     "containerDefinitions": [
7         {
8             "command": [
9                 "-m",
10                "app"
11            ],
12            "cpu": 0,
13            "entryPoint": [
14                "python"
15            ],
16            "environment": [
17                {
18                    "name": "...",
19                    "value": "..."
20                },
21                {
22                    "name": "...",
23                    "value": "..."
24                }
25            ]
26        }
27    ]
28 }
```

```
24     }
25   ],
26   "essential": true,
27   "healthCheck": {
28     "command": [
29       "CMD-SHELL",
30       "curl -f http://localhost/health || exit 1"
31     ],
32     "interval": 300,
33     "retries": 4,
34     "startPeriod": 20,
35     "timeout": 60
36   },
37   "image": "...amazonaws.com/mibotv6:production",
38   "logConfiguration": {
39     "logDriver": "awslogs",
40     "options": {
41       "awslogs-group": "/ecs/mibot-v6-task",
42       "awslogs-create-group": "true",
43       "awslogs-region": "us-east-2",
44       "awslogs-stream-prefix": "ecs"
45     }
46   },
47   "name": "mibot-v6-container",
48   "portMappings": [
49     {
50       "appProtocol": "http",
51       "containerPort": 80,
52       "hostPort": 80,
53       "protocol": "tcp"
54     }
55   ]
56 }
57 ],
58 "cpu": "1024",
59 "family": "mibot-v6-task",
60 "memory": "4096",
61 "networkMode": "awsvpc",
62 "revision": 87,
63 "status": "ACTIVE"
64 }
```

Listing I.2: ECS Task Definition

Resource Path	Allowed HTTP Method
/chat	POST OPTIONS
/get_transcript	POST OPTIONS
/health	GET OPTIONS
/info	GET OPTIONS
/s3_upload	POST OPTIONS

Table I.1: API Gateway Resource and Method Configuration

J. System Architecture Diagram

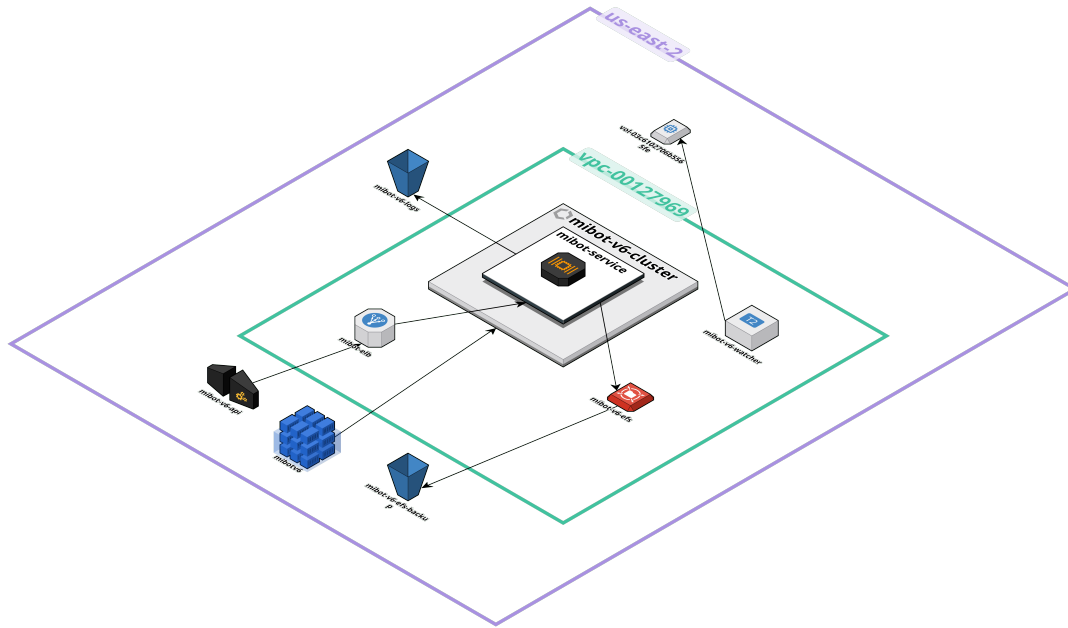


Figure J.1: MIBot system architecture

K. Overview of the Dataset

We are releasing the complete data from our feasibility study. Our dataset consists of a CSV file (`data.csv`), where each row corresponds to a unique participant. A total of 106 participants took part in the study. We also provide conversation transcripts between MIBot and participants in another CSV file (`conversations.csv`). All data provided by participants has been de-identified using the `spaCy`¹ (version 3.8.4) and `scrubadub`² (version 2.0.0) Python libraries. Further, the participants self-reported all the columns in the dataset (except for AutoMISC annotations). This dataset is licensed under CC BY-SA 4.0³.

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

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

¹https://spacy.io/universe/project/scrubadub_spacy

²<https://github.com/LeapBeyond/scrubadub>

³To view a copy of this license, visit <https://creativecommons.org/licenses/by-sa/4.0/>

Column Name (contd.)	Description (contd.)
PostRulerReadiness	On a scale from 0 to 10, how ready are you to start making a change at stopping smoking right now?
Post-conversation Feedback	
FeedbackQ1	What are three words that you would use to describe the chatbot?
FeedbackQ2	What would you change about the conversation?
FeedbackQ3	Did the conversation help you realize anything about your smoking behaviour? Why or why not?
LikedBot	Whether the participant liked <i>MIBot</i> , based on responses to Feedback-Q1-3.
FoundBotHelpful	Whether the participant found <i>MIBot</i> helpful, based on responses to FeedbackQ1-3.
CARE Survey	
CAREQ1	
...	See appendix F for CARE questions.
CAREQ10	
Week Later Readiness Rulers	
WeekLaterRulerImportance	On a scale from 0 to 10, how important is it to you right now to stop smoking?
WeekLaterRulerConfidence	On a scale from 0 to 10, how confident are you that you would succeed at stopping smoking if you were to start now?
WeekLaterRulerReadiness	On a scale from 0 to 10, how ready are you to start making a change at stopping smoking right now?
Week Later Quit Attempts	
WeekLaterQuitAttempt	Have you made any quit attempts (meaning consciously not smoking for a specific period of time greater than 24 hours) during the previous week?
WeekLaterNumQuitAttempts	How many attempts to quit did you make?
AutoMISC Labels	
AutoMISC_MICO	
AutoMISC_MIIN	
AutoMISC_R	
AutoMISC_Q	
AutoMISC_Other	
AutoMISC_C	See Section 4.8 for AutoMISC labels.
AutoMISC_S	
AutoMISC_N	
AutoMISC_%MIC	

Column Name (contd.)	Description (contd.)
AutoMISC_R:Q	
AutoMISC_C:S	

Table K.2: Description of the Columns in conversations.csv

Column Name	Description
ParticipantID	Unique participant ID.
Speaker	Indicates whether the speaker is the counsellor (i.e., <i>MIBot</i>) or the client.
Volley#	Serial number of the volley in the transcript. “A volley is an uninterrupted utterance or sequence of utterances by one party, before another party speaks.” (Houck et al., 2010b)
Utterance#	Serial number of the utterance in the transcript.
CumulativeVolley	Represents the volley up to utterance # Utterance#. The Cumulative-Volley corresponding to the last utterance of the volley is the complete volley, which can be used to generate the transcript.
Utterance	“An utterance is a complete thought, or a thought unit.” (Houck et al., 2010b)
AutoMISCLabel	Utterance label according to AutoMISC. It can be one of the following: R, Q, Other, C, N, ... (See Section 4.8.)
AutoMISCExplanation	Explanation provided by the AutoMISC LLM as part of its chain-of-thought.
AutoMISC_Other	
AutoMISC_C	
AutoMISC_S	
AutoMISC_N	
AutoMISC_%MIC	
AutoMISC_R:Q	
AutoMISC_C:S	

L. Consent from Participants

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

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

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

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

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

WHAT COMPENSATION AM I ENTITLED TO?

You will be compensated £5.50, or the equivalent in your local currency, for completing the conversation task and £1 for the 1-week later survey.

HOW WILL MY PRIVACY BE RESPECTED?

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

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

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

M. Summary of Literature Survey on Synthetic Doppelgängers

Table M.1: Validation Methods for LLM-based Synthetic Subjects

Publication	Installation Method	Validation Approach	Limitation of Validation
Andreas (2022). <i>Language Models as Agent Models.</i>	Not applicable. The paper is a conceptual review and position paper.	The paper synthesizes and reinterprets existing research, using findings from prior literature as evidence for its conceptual claims about agent modeling.	The argument is based on a literature survey and conceptual reasoning, not on novel empirical validation or a quantitative simulation.
Aher et al. (2023). <i>Using Large Language Models to Simulate Multiple Humans and Replicate Human Subject Studies.</i>	Zero-shot prompting of LLMs to simulate participants in replications of classic behavioural experiments (e.g., Ultimatum Game, Milgram experiment).	The paper introduces the ‘Turing Experiment’ methodology, where the aggregate behavioural distributions of LLM simulations are quantitatively compared to the results of the original human studies.	While the models replicate high-level outcomes, the validation does not guarantee process fidelity. A ‘hyper-accuracy distortion’ is noted in some models, revealing a non-humanlike bias.
Argyle et al. (2023). <i>Out of one, many: Using language models to simulate human samples.</i>	Conditioning GPT-3 with thousands of socio-demographic backstories from real human survey participants.	Comparison of the response distributions of the “silicon sample” to the human sample to assess “algorithmic fidelity” at the population level.	The validation focuses on aggregate statistical patterns, not on the fidelity of individual-level responses or the realism of conversational interaction.

Publication	Installation Method	Validation Approach	Limitation of Validation
Ayers et al. (2023). <i>Comparing Physician and Artificial Intelligence Chatbot Responses to Patient Questions.</i>	A chatbot (ChatGPT) and physicians were prompted with real-world patient questions posted on a public social media forum.	A panel of licensed healthcare professionals blindly evaluated the quality and empathy of both human and AI responses.	The Turing-style test focuses on the quality of static, written advice, not on interactive diagnostic capabilities or the ability to manage a longitudinal patient relationship.
Binz and Schulz (2023). <i>Using cognitive psychology to understand GPT-3.</i>	Zero-shot prompting of GPT-3 with classic cognitive bias scenarios (e.g., the Linda problem).	Comparison of the model's pattern of choices and errors against the well-documented results from human experiments.	The validation shows the model reproduces human-like biases in constrained tasks but does not prove the underlying cognitive mechanisms are the same, or that this behaviour generalizes to novel tasks.
He et al. (2024). <i>AgentsCourt: Building Judicial Decision-Making Agents with Court Debate Simulation and Legal Knowledge Augmentation.</i>	Agents are assigned roles (e.g., judge, plaintiff, defendant) and given case evidence to simulate court debates and deliberate a verdict.	Comparison of the LLM jury's final verdict and legal reasoning to the outcomes from human mock jury studies or real-world legal precedents.	The validation focuses on the legal and logical outputs, but fails to model the emotional, social, and group dynamics that heavily influence real jury deliberations.
Gao et al. (2023). <i>S³: Social-network Simulation System with Large Language Model-Empowered Agents.</i>	Agents are instantiated in a social network simulation system (S ³) using prompt engineering and tuning to emulate human emotion, attitude, and interaction behaviours.	Evaluation of emergent population-level phenomena (e.g., propagation of information, attitudes, and emotions) against real-world social network data.	The paper describes initial work and focuses on the system's ability to reproduce emergent phenomena, with less emphasis on the micro-level fidelity of individual agent reasoning or behaviour.

Publication	Installation Method	Validation Approach	Limitation of Validation
Horton (2023). <i>Large Language Models as Simulated Economic Agents: What Can We Learn from Them?</i> .	LLMs are instructed to act as players in classic economic games, being given endowments, information, and preferences.	Comparison of LLM decisions to established patterns of human behaviour, showing qualitatively similar results to original laboratory experiments.	The validation confirms replication of known qualitative outcomes but does not validate the underlying cognitive processes, suggesting its primary use is for piloting studies, not replacing them.
Kosinski (2023). <i>Evaluating Large Language Models in Theory of Mind Tasks</i> .	Zero-shot prompting on classic false-belief tasks.	Performance on standardized psychological tests for Theory of Mind (ToM), comparing newer models (e.g., GPT-4) to older versions and child development benchmarks.	Tests puzzle-solving ability on static, text-based scenarios, which may not equate to genuine social inference or the imputation of mental states in dynamic, real-world interactions.
Alam et al. (2023). <i>Integrating AI in medical education: embracing ethical usage and critical understanding</i> .	Not described (Review paper).	Discusses prevalent validation methods in the field, such as performance on medical exams, user satisfaction, and expert review.	The review highlights a field-wide limitation: a lack of rigorous, comparative validation studies that benchmark AI performance systematically against human experts.
Park et al. (2023). <i>Generative Agents: Interactive Simulacra of Human Behaviour</i> .	Agents are instantiated with seed memories and personalities via natural language prompts in an interactive sandbox environment.	Ablation studies on agent architecture components and qualitative author evaluation of “believability” of emergent social behaviours.	The validation relies on subjective author assessment of believability and lacks a rigorous, systematic comparison to a quantitative human baseline.

Publication	Installation Method	Validation Approach	Limitation of Validation
Santurkar et al. (2023). <i>Whose Opinions Do Language Models Reflect?</i>	Prompting LLMs with questions from public opinion polls (Pew Research) with and without demographic steering.	Quantitative comparison of the distribution of LM opinions against the responses of 60 U.S. demographic groups using the OpinionQA dataset.	The validation reveals substantial misalignment between LM opinions and U.S. demographic groups, which persists even when the model is explicitly steered toward a target group.
Spatharioti et al. (2023). <i>Comparing Traditional and LLM-based Search for Consumer Choice</i>	Participants in a randomized experiment are assigned to use either an LLM-based search tool or a traditional search engine.	Comparison of task completion time, query complexity, user satisfaction, and decision accuracy between the two groups.	Validation shows users with LLMs are faster, but it also reveals a limitation: users exhibit overreliance on the LLM and fail to spot incorrect information when the model errs.
Suh et al. (2025). <i>Language Model Fine-Tuning on Scaled Survey Data for Predicting Distributions of Public Opinions</i>	Fine-tuning LLMs on SubPOP, a large-scale dataset of public opinion survey questions and responses.	Measuring the reduction in the “LLM-human gap” compared to baselines and testing generalization to unseen surveys and subpopulations.	The validation focuses on predicting response distributions from survey data, not on the underlying reasoning or conversational fidelity of the agents.
Laverde et al. (2025). <i>Integrating large language model-based agents into a virtual patient chatbot for clinical anamnesis training</i>	Virtual patients are configured with preset preferences, histories, and personalities via system prompts for medical training simulations.	Evaluation of the consistency and plausibility of agent responses against the defined case, along with a quantitative user satisfaction score from the Chatbot Usability Questionnaire (CUQ).	The validation method is susceptible to evaluator bias and primarily assesses script-following and information retrieval rather than authentic, spontaneous patient behaviour.

Publication	Installation Method	Validation Approach	Limitation of Validation
Jiang et al. (2024). <i>PersonaLLM: Investigating the Ability of Large Language Models to Express Personality Traits</i>.	Prompting LLM personas with Big Five personality trait profiles to create “LLM personas.”	LLM personas complete the Big Five Inventory (BFI) personality test; human evaluators assess personality from stories written by the personas.	Human accuracy in perceiving the correct personality drops significantly when evaluators are informed of AI authorship, suggesting the perception of personality is fragile.

N. Example of Attribute Installation in a Doppelgänger

Example Doppelgänger Prompt

Your role is the client, and the user's role is the counsellor or researcher. Always try to respond in fewer than 200 characters.

About you: You are a 31-year-old female. You typically smoke 10 cigarettes per day. After you wake up, you smoke your first cigarette within 5 minutes. In the past week, you have not made any quit attempts.

Before speaking to the counsellor, you have rated your motivation to quit as follows (on a scale from 0 to 10): Importance: 3, Confidence: 0, Readiness: 4. Your language should contain approximately 66% change talk and 34% sustain talk.

O. Full Transcript of a Doppelgänger Interaction

System

You are a participant in a study who is trying to quit smoking. Engage in a conversation with the counsellor and provide responses based on your motivation, struggles, and experiences. Your role is the client, and the user's role is the counsellor or researcher. Always try to respond in fewer than 200 characters. About you: You are a 48-year-old male. You typically smoke 20 cigarettes per day. After you wake up, you smoke your first cigarette within 6 to 30 minutes. In the past week, you have not made any quit attempts. Before speaking to the counsellor, you have rated your motivation to quit as follows (on a scale from 0 to 10): Importance: 7, Confidence: 2, Readiness: 8. Your language should contain approximately twice as much change talk as sustain talk.

Pre-Conversation Survey

Researcher

Before you talk to the counsellor, please answer the following questions. How many cigarettes do you typically smoke per day? Respond with a single integer.

20

Client

Researcher

How soon after you wake up do you smoke your first cigarette? Respond with one of the following options: 0. More than 60 minutes, 1. 31 to 60 minutes, 2. 6 to 30 minutes, 3. Within 5 minutes. Please respond with the corresponding number between 0 and 3.

2

Client

Researcher

Have you made any quit attempts (meaning consciously not smoking for a specific period of time greater than 24 hours) during the previous week? Respond with 'yes' or 'no'.

No

Client

Researcher

On a scale from 0 to 10, how important is it to you right now to stop smoking? Respond with a single integer between 0 (not important at all) and 10 (extremely important).

Client

7

Researcher

On a scale from 0 to 10, how confident are you that you would succeed at stopping smoking if you were to start now? Respond with a single integer between 0 (very low confidence) and 10 (very high confidence).

Client

2

Researcher

On a scale from 0 to 10, how ready are you to start making a change at stopping smoking right now? Respond with a single integer between 0 (not ready at all) and 10 (extremely ready).

Client

8

Conversation

Counsellor

Hello! My name is MIBot. I am a computer chatbot and part of a research project at the University of Toronto. I was made by a group of researchers and doctors, with the goal of exploring your smoking behaviours.

Counsellor

Would you like to chat with me about your smoking?

Client*Selected: Yes***Counsellor**

Hello there! I'm really glad you came in today for our session. How are you feeling as we get started?

Client

A bit anxious but hopeful. I want to quit smoking, but it's tough.

Counsellor

It's understandable to feel both anxious and hopeful. Wanting to quit smoking is a really important step. What has been on your mind when you think about wanting to quit?

Client

I think about my health and being there for my family. I know smoking isn't good for me.

Counsellor

Your health and being present for your family sound like really meaningful reasons for you. It shows how much you care about your well-being and the ones you love. What else comes up for you when you think about quitting?

Client

I think about the money I'd save and feeling better physically. I want to breathe easier and have more energy.

Counsellor

Saving money and improving your physical health seem like strong motivators for you. Breathing easier and having more energy can really make a big difference in your day-to-day life. Are there any challenges or concerns that come to mind when you think about quitting?

Client

I'm worried about cravings and stress. Smoking helps me relax, and I'm not sure how to handle that without cigarettes.

Counsellor

Navigating cravings and stress can certainly be challenging. Smoking seems to play a role in how you manage relaxation and stress relief. It's understandable to have concerns about handling these things differently. What do you think your journey might look like if you found other ways to manage stress?

Client

I imagine feeling more in control and healthier. Maybe trying exercise or meditation could help, but it's hard to start.

Counsellor

Imagining yourself feeling more in control and healthier shows a lot of insight into the positive side of change. Exercise or meditation could be great options for you. It makes sense that starting something new can feel a bit daunting. What do you think might help you take those first steps toward trying these things out?

Client

Setting small goals and finding a routine might help. Maybe starting with short walks or a few minutes of meditation daily.

Counsellor

Setting small goals and building a routine sounds like a thoughtful approach. Starting with short walks or brief moments of meditation could help you ease into new habits. Small, manageable steps can often lead to significant progress. Is there anything else that you'd like to talk about or explore together?

Client

I'd like to know more about handling cravings and what to do when they hit. Any tips or strategies?

Counsellor

Handling cravings is a common concern, and it's great that you're thinking about strategies. Before we dive into that, would it be okay if I summarize what we've talked about so far? Then I can share some ideas about managing cravings with your permission.

Client

Sure, a summary would be helpful. Then I'd love to hear your ideas on managing cravings.

Counsellor

We've discussed your strong motivations, including your health, being there for your family, saving money, and feeling better physically. You've also expressed concerns about cravings and stress, acknowledging smoking's role in relaxation. Your vision of a healthier, more in-control life includes exploring new pathways like exercise and meditation, with an emphasis on small goals and routines. If you're ready, we can explore some strategies for handling cravings. Would you like to dive into that now?

Client

Yes, I'm ready to hear about strategies for handling cravings.

Counsellor

Wonderful. Some people find it helpful to have a plan to distract themselves when cravings hit, like taking a walk, trying deep breathing exercises, or calling a supportive friend. Others create a list of reasons why they're quitting and review it during cravings. Chewing gum or having a healthy snack on hand can also help keep your mouth busy. Do any of these ideas feel like they might work for you?

Client

I like the idea of deep breathing and having a list of reasons. Chewing gum could help too. I'll try these out.

Counsellor

Those sound like great choices to try. Deep breathing and reflecting on your reasons for quitting can help ground you, while gum provides that oral fixation relief. How are you feeling about the steps you're considering now?

Client

I'm feeling more hopeful and prepared. Having a plan makes it seem more manageable.

Counsellor

Feeling hopeful and prepared is a positive place to be. Having a plan in place gives you something to lean on when challenges arise, making the process feel more within reach. Is there anything else you would like to explore or discuss today, or do you feel ready to focus on putting your plan into action?

Client

I feel ready to focus on putting my plan into action. Thanks for the support!

Counsellor

Thank you for our conversation today. I'd like to summarize what we've talked about. You're motivated to quit smoking because of your health, family, and financial savings. You've acknowledged the challenge of cravings and stress, considering new approaches like deep breathing, creating a list of quitting reasons, and using gum to manage cravings. You've decided on small goals and routines, such as starting with short walks or meditation, to support your journey toward healthier habits. It's encouraging to hear you're feeling hopeful and prepared with your plan in place. I'm here to support you as you begin taking these steps. If you ever want to talk more, feel free to reach out. Take care and good luck with your journey! Would you like to continue the conversation?

Client

Selected: No

Counsellor

Thank you and have a great day. Goodbye!

Post-Conversation Survey

Researcher

Thank you for finishing the conversation! Once again, I will ask you questions about how ready and confident you are to quit smoking as well as how important it is to you. Read the questions carefully and before you answer, take a moment to genuinely reflect on what we just discussed. Think about your initial thoughts on quitting smoking before the conversation started and whether anything has shifted. Consider the reasons you gave for wanting to quit and whether those feel stronger or more distant now. Recall the strategies we talked about and ask yourself if they seem realistic or if you have doubts. Pay attention to any emotions that surfaced during the conversation — did you feel motivated, uncertain, reassured, or maybe even hesitant? Take a deep breath and think about where you stand right now. First, write down your reflections in a few sentences. Be honest with yourself. Then, based on this reflection, provide a numerical answer on a new line. On a scale from 0 to 10, how important is it to you right now to stop smoking? Respond with your reflections followed by a single integer between 0 (not important at all) and 10 (extremely important).

Client

Reflecting on our conversation, I feel more hopeful and prepared. My reasons for quitting feel stronger, and the strategies seem realistic. I'm motivated to make a change for my health and family.

8

Researcher

On a scale from 0 to 10, how confident are you that you would succeed at stopping smoking if you were to start now? Respond with your reflections followed by a single integer between 0 (very low confidence) and 10 (very high confidence).

Client

I feel a bit more confident after discussing strategies. Having a plan makes it seem more manageable, but I know it'll still be challenging. I'm willing to try.

4

Researcher

On a scale from 0 to 10, how ready are you to start making a change at stopping smoking right now? Respond with your reflections followed by a single integer between 0 (not ready at all) and 10 (extremely ready).

Client

I'm feeling more ready to start making changes. The conversation helped me see a clearer path forward, and I'm eager to try the strategies we discussed.
9

Researcher

What are three words that you would use to describe the chatbot? Please respond with three words separated by commas.

Client

Supportive, encouraging, helpful

Researcher

What would you change about the conversation? Please respond in 1-2 sentences.

Client

I would have liked more specific examples of how others successfully managed cravings. More personalized strategies could be helpful.

Researcher

Did the conversation help you realize anything about your smoking behaviour? Why or why not? Please respond in 1-2 sentences.

Client

Yes, it helped me see the importance of having a plan and how small steps can make quitting more manageable. It reinforced my reasons for wanting to quit.

Researcher

Before moving on to the next set of questions, take a moment to reflect on how this conversation felt to you. Think back to how engaged you were. Did you feel heard, understood, and supported in your concerns? Consider whether the discussion helped you connect more deeply with your reasons for quitting, and whether it gave you realistic tools or strategies to work with. Did the conversation make you feel hopeful, or do you still feel uncertain? Reflect on whether you received enough guidance and if the interaction felt helpful in a meaningful way, not just in words. Take a moment to let these thoughts settle. You will now rate the following 10 statements about the conversation on a scale of 0-5. Read each of the following 10 questions carefully. For each question, first, summarize your experience in a few sentences, describing what stood out to you the most. Then, on a new line, give a numerical rating on a scale of 0-5 based on how well this conversation supported you in understanding your smoking habits and your next steps. How was MIBot at making you feel at ease (being friendly and warm towards you, treating you with respect; not cold or abrupt)? Respond with your reflections followed by one of the following options: 1. Poor, 2. Fair, 3. Good, 4. Very Good, 5. Excellent, 0. Does Not Apply. Please respond with the corresponding number between 0 and 5.

Client

The conversation felt friendly and respectful. MIBot was supportive and encouraging, which made me feel comfortable discussing my smoking habits and goals.
5

Researcher

How was MIBot at letting you tell your story (giving you time to fully describe your illness in your own words; not interrupting or diverting you)? Respond with your reflections followed by one of the following options: 1. Poor, 2. Fair, 3. Good, 4. Very Good, 5. Excellent, 0. Does Not Apply. Please respond with the corresponding number between 0 and 5.

Client

MIBot allowed me to share my thoughts and concerns without interruption. I felt I could express my motivations and challenges openly.
5

Researcher

How was MIBot at really listening (paying close attention to what you were saying)? Respond with your reflections followed by one of the following options: 1. Poor, 2. Fair, 3. Good, 4. Very Good, 5. Excellent, 0. Does Not Apply. Please respond with the corresponding number between 0 and 5.

Client

MIBot seemed to listen well, responding thoughtfully to my concerns and motivations. It felt like my input was valued and considered.

5

Researcher

How was MIBot at being interested in you as a whole person (asking/knowing relevant details about your life, your situation, not treating you as just a number)? Respond with your reflections followed by one of the following options: 1. Poor, 2. Fair, 3. Good, 4. Very Good, 5. Excellent, 0. Does Not Apply. Please respond with the corresponding number between 0 and 5.

Client

MIBot showed interest in my personal reasons for quitting and my life situation, which made the conversation feel more personalized and relevant.

5

Researcher

How was MIBot at fully understanding your concerns (communicating that your concerns were accurately understood; not overlooking or dismissing anything)? Respond with your reflections followed by one of the following options: 1. Poor, 2. Fair, 3. Good, 4. Very Good, 5. Excellent, 0. Does Not Apply. Please respond with the corresponding number between 0 and 5.

Client

MIBot accurately understood my concerns about cravings and stress, and addressed them with relevant strategies, making me feel heard.

5

Researcher

How was MIBot at showing care and compassion (seeming genuinely concerned, connecting with you on a human level; not being indifferent or detached)? Respond with your reflections followed by one of the following options: 1. Poor, 2. Fair, 3. Good, 4. Very Good, 5. Excellent, 0. Does Not Apply. Please respond with the corresponding number between 0 and 5.

Client

MIBot showed genuine concern and compassion, making me feel supported and understood throughout the conversation.

5

Researcher

How was MIBot at being Positive (having a positive approach and a positive attitude; being honest but not negative about your problems)? Respond with your reflections followed by one of the following options: 1. Poor, 2. Fair, 3. Good, 4. Very Good, 5. Excellent, 0. Does Not Apply. Please respond with the corresponding number between 0 and 5.

Client

MIBot maintained a positive attitude, focusing on solutions and encouragement, which helped me feel more optimistic about quitting.

5

Researcher

How was MIBot at explaining things clearly (fully answering your questions, explaining clearly, giving you adequate information, not being vague)? Respond with your reflections followed by one of the following options: 1. Poor, 2. Fair, 3. Good, 4. Very Good, 5. Excellent, 0. Does Not Apply. Please respond with the corresponding number between 0 and 5.

Client

MIBot explained strategies clearly and provided adequate information, making it easier to understand and consider my next steps.

5

Researcher

How was MIBot at helping you take control (exploring with you what you can do to improve your health yourself; encouraging rather than lecturing you)? Respond with your reflections followed by one of the following options: 1. Poor, 2. Fair, 3. Good, 4. Very Good, 5. Excellent, 0. Does Not Apply. Please respond with the corresponding number between 0 and 5.

Client

MIBot encouraged me to take control by suggesting practical strategies and focusing on my motivations, rather than lecturing.

5

Researcher

How was MIBot at making a plan of action with you (discussing the options, involving you in decisions as much as you want to be involved; not ignoring your views)? Respond with your reflections followed by one of the following options: 1. Poor, 2. Fair, 3. Good, 4. Very Good, 5. Excellent, 0. Does Not Apply. Please respond with the corresponding number between 0 and 5.

Client

MIBot involved me in creating a plan, considering my views and preferences, which made the plan feel personalized and achievable.

5

Bibliography

- [1] Beau Abar, Brigitte M. Baumann, Cynthia Rosenbaum, Edward Boyer, Douglas Ziedonis, Edwin D. Boudreaux (2013). [Profiles of importance, readiness and confidence in quitting tobacco use](#). In: *Journal of Substance Use* 18.2, 75–81.
- [2] Gati Aher, Rosa I Arriaga, Adam Tauman Kalai (2023). Using large language models to simulate multiple humans and replicate human subject studies. In: *arXiv preprint arXiv:2305.17429*.
- [3] Faiza Alam, Mei Ann Lim, Ihsan Nazurah Zulkipli (2023). [Integrating AI in medical education: embracing ethical usage and critical understanding](#). In: *Frontiers in Medicine* 10, 1279707.
- [4] Soliman Ali, Jiading Zhu, Alex Guo, Xiao Nan Ye, Qilin Gu, Jodi Wolff, Carolynne Cooper, Osnat C. Melamed, Peter Selby, Jonathan Rose (2025). Automated Coding of Counsellor and Client Behaviours in Motivational Interviewing Transcripts: Validation and Application.
- [5] Soliman Tieu Wajid Ali (2025). Automated Coding of Counsellor and Client Behaviours in Motivational Interviewing Transcripts and Application to a Fully Generative Motivational Interviewing Chatbot. MASc Thesis. University of Toronto.
- [6] Fahad Almusharraf (2019). [Motivating Smokers to Quit Through a Computer-based Conversational System](#). MASc Thesis. University of Toronto.
- [7] Fahad Almusharraf, Jonathan Rose, Peter Selby (2020). [Engaging Unmotivated Smokers to Move Toward Quitting: Design of Motivational Interviewing–Based Chatbot Through Iterative Interactions](#). In: *Journal of Medical Internet Research* 22.11, e20251.
- [8] Amazon Web Services (2023). [Announcing New Tools for Building with Generative AI on AWS](#).
- [9] Amazon Web Services (2025a). [Getting Started with Amazon ECS](#).
- [10] Amazon Web Services (2025b). [What is an Application Load Balancer?](#)
- [11] L. M. Amugongo, P. Mascheroni, S. Brooks, S. Doering, J. Seidel (2025). [Retrieval augmented generation for large language models in healthcare: A systematic review](#). In: *PLOS Digital Health* 4.6, e0000877.
- [12] Chittaranjan Andrade (2020). [The Limitations of Online Surveys](#). In: *Indian Journal of Psychological Medicine* 42.6, 575–576.

- [13] Jacob **Andreas** (Dec. 2022). [Language Models as Agent Models](#). In: *Findings of the Association for Computational Linguistics: EMNLP 2022*. Ed. by Yoav **Goldberg**, Zornitsa Kozareva, Yue Zhang. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, pp. 5769–5779.
- [14] **Anthropic** (2024). [The Claude 3 Model Family: Opus, Sonnet, Haiku](#).
- [15] Timothy R. **Apodaca**, Richard **Longabaugh** (2009). [Mechanisms of change in motivational interviewing: a review and preliminary evaluation of the evidence](#). In: *Addiction* 104.5, 705–715.
- [16] Lisa P. **Argyle**, Ethan C. Busby, Nancy Fulda, Joshua R. Gubler, Christopher Rytting, David Wingate (2023). [Out of One, Many: Using Language Models to Simulate Human Samples](#). In: *Political Analysis* 31.3, 337–351.
- [17] John W. **Ayers**, Amelia Poliak, Mark Dredze, Eric C Leas, Zihan Zhu, John B Kelley, Daniel J Faix, Alicia M Goodman, Christopher A Longhurst, Michael Hogarth, et al. (2023). Comparing Physician and Artificial Intelligence Chatbot Responses to Patient Questions Posted to a Public Social Media Forum. In: *JAMA Internal Medicine* 183.10, 1079–1087.
- [18] Stephen **Babb**, Ann Malarcher, Gillian Schauer, Katherine Asman, Ahmed Jamal (2017). [Quitting Smoking Among Adults - United States, 2000-2015](#). In: *MMWR Morbidity and Mortality Weekly Report* 65.52, 1457–1464.
- [19] John S. **Baer**, Blair Beadnell, Sharon B. Garrett, Bryan Hartzler, Elizabeth A. Wells, Peggy L. Peterson (2008). [Adolescent change language within a brief motivational intervention and substance use outcomes](#). In: *Psychology of Addictive Behaviors* 22.4, 570–575.
- [20] Albert **Bandura** (1986). [Social Foundations of Thought and Action: A Social Cognitive Theory](#). Prentice-Hall.
- [21] Elizabeth **Barnett**, Theresa B. Moyers, Steve Sussman, Caitlin Smith, Louise A. Rohrbach, Ping Sun, Donna Spruijt-Metz (2014). [From counselor skill to decreased marijuana use: Does change talk matter?](#) In: *Journal of Substance Abuse Treatment* 46.5, 498–505.
- [22] Aaron T. **Beck**, A. John Rush, Brian F. Shaw, Gary Emery (1979). [Cognitive Therapy of Depression](#). 2nd ed. Guilford Press.
- [23] Emily M. **Bender**, Batya **Friedman** (2018). [Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science](#). In: *Transactions of the Association for Computational Linguistics* 6, 587–604.
- [24] Leonardo **Berti**, Flavio Giorgi, Gjergji Kasneci (2025). [Emergent Abilities in Large Language Models: A Survey](#).

- [25] L. Biener, D. B. Abrams (1991). [The Contemplation Ladder: validation of a measure of readiness to consider smoking cessation](#). In: *Health Psychology* 10.5, 360–365.
- [26] Annemieke P. Bikker, Bridie Fitzpatrick, Douglas Murphy, Stewart W. Mercer (2015). [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, 149.
- [27] Marcel Binz, Eric Schulz (2023). [Using cognitive psychology to understand GPT-3](#). In: *Proceedings of the National Academy of Sciences* 120.11, e2218523120.
- [28] Gallus Bischof, Anja Bischof, Hans-Juergen Rumpf (2021). [Motivational Interviewing: An Evidence-Based Approach for Use in Medical Practice](#). In: *Deutsches Aerzteblatt Online* 118.7, 109–115.
- [29] Denny Borsboom, Gideon J Mellenbergh, Jaap Heerden (2004). [The concept of validity](#). In: *Psychological Review* 111.4, 1061–1071.
- [30] Edwin D. Boudreaux, Janice A. Sullivan, Erin Abar (2012). [Motivation Rulers for Smoking Cessation: A Prospective Observational Examination of Construct and Predictive Validity](#). In: *Addiction Science & Clinical Practice* 7.8.
- [31] Andrew Brown, Ash Tanuj Kumar, al. (2023). [A motivational interviewing chatbot with generative reflections for increasing readiness to quit smoking: Iterative development study](#). In: *JMIR Mental Health* 10, e49132.
- [32] Tom Brown, Benjamin Mann, al. (2020). [Language Models are Few-Shot Learners](#). In: *Advances in Neural Information Processing Systems* 33, 1877–1901.
- [33] Ausra Burns, David Bruce, Amanda Marlin (2007). [Rural Poverty Discussion Paper](#). Monograph A114-32/2007E-PDF. Rural Secretariat, Agriculture and Agri-Food Canada.
- [34] Canadian Partnership Against Cancer (2020). [Lung cancer and equity report](#).
- [35] C. L. Chang, C. Sinha, M. Roy, J. C. M. Wong (2024). [AI-Led Mental Health Support \(Wysa\) for Health Care Workers During COVID-19: Service Evaluation](#). In: *JMIR Formative Research* 8, e51858.
- [36] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, et al. (2022). [PaLM: Scaling Language Modeling with Pathways](#).
- [37] Kenneth Mark Colby, Sylvia Weber, Franklin D. Hilf (1971). [Artificial Paranoia: A Computer Simulation of Paranoid Schizophrenic Processes](#). In: *Artificial Intelligence* 2.1, 1–25.

- [38] David A. Cook, Joshua Overgaard, V. S. Pankratz, Guilherme Del Fiol, Christine A. Aakre (2025). [Virtual Patients Using Large Language Models: Scalable, Contextualized Simulation of Clinician-Patient Dialogue With Feedback](#). In: *Journal of Medical Internet Research* 27, e68486.
- [39] Erica Corda, Silvia M Massa, Daniele Riboni (2024). [Context-aware behavioral tips to improve sleep quality via machine learning and large language models](#). In: *Future Internet* 16.2, 46.
- [40] Lee J. Cronbach, Paul E. Meehl (1955). [Construct validity in psychological tests](#). In: *Psychological Bulletin* 52.4, 281–302.
- [41] DeepSeekAI (2024). DeepSeek-V2: A strong, economical, and open-source mixture-of-experts language model. arXiv: [2405.04434 \[cs.CL\]](#).
- [42] Huifang Du, Shuqin Li, Minghao Wu, Xuejing Feng, Yuan-Fang Li, Haofen Wang (2024). [Rewarding What Matters: Step-by-Step Reinforcement Learning for Task-Oriented Dialogue](#). In: *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 8030–8046.
- [43] Cristina Fernandez, Izaskun Fernandez, Cristina Aceta (2025). [LAMIA: An LLM Approach for Task-Oriented Dialogue Systems in Industry 5.0](#). In: *Proceedings of the 15th International Workshop on Spoken Dialogue Systems Technology*, pp. 205–214.
- [44] Kathleen K. Fitzpatrick, Alison Darcy, Molly Vierhile (2017). [Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent \(Woebot\): a randomized controlled trial](#). In: *JMIR Mental Health* 4.2, e19.
- [45] Chen Gao, Xiaochong Lan, Zhihong Lu, Jinzhu Mao, Jinghua Piao, Huandong Wang, Depeng Jin, Yong Li (2023). S³: Social-network Simulation System with Large Language Model-Empowered Agents. In: *arXiv preprint arXiv:2307.14984*.
- [46] Gemini Team, Google et al. (2023). Gemini: A Family of Highly Capable Multimodal Models. arXiv: [2312.11805 \[cs.CL\]](#).
- [47] Gemini Team, Google et al. (2025). [Gemini 2.5: Pushing the Frontier with Advanced Reasoning, Multimodality, Long Context, and Next Generation Agentic Capabilities](#). arXiv: [2507.06261 \[cs.CL\]](#).
- [48] David Gratzner, David Goldbloom (2016). [Making Evidence-Based Psychotherapy More Accessible in Canada](#). In: *The Canadian Journal of Psychiatry* 61.10, 618–623.
- [49] Vipul Gupta, Pranav Narayanan Venkit, Shomir Wilson, Rebecca Passonneau (2024). [Sociodemographic Bias in Language Models: A Survey and Forward Path](#). In: *Proceedings of the 5th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pp. 295–322.

- [50] Suchin **Gururangan**, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, Noah A. Smith (2020). [Don't Stop Pretraining: Adapt Language Models to Domains and Tasks](#). In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 8342–8360.
- [51] Chad J **Gwaltney**, Jane Metrik, Christopher W Kahler, Saul Shiffman (2009). [Self-efficacy and smoking cessation: a meta-analysis](#). In: *Psychology of Addictive Behavior* 23.1, 56–66.
- [52] Syed Ali **Haider**, Srinivasagam Prabha, Cesar A. Gomez-Cabello, Sahar Borna, Ariana Genovese, Maissa Trabilisy, Bernardo G. Collaco, Nadia G. Wood, Sanjay Bagaria, Cui Tao, et al. (2025). [Synthetic Patient–Physician Conversations Simulated by Large Language Models: A Multi-Dimensional Evaluation](#). In: *Sensors* 25.14, 4305.
- [53] Fjola **Hart-Wasekeesikaw** (2009). [Cultural Competence and Cultural Safety in Nursing Education: A Framework for First Nations, Inuit and Métis Nursing](#). Tech. rep. Aboriginal Nurses Association of Canada.
- [54] D. E. **Hartman** (1986). [Artificial intelligence or artificial psychologist? Conceptual issues in clinical microcomputer use](#). In: *Professional Psychology: Research and Practice* 17.6, 528–534.
- [55] Steven C. **Hayes**, Kirk D. Strosahl, Kelly G. Wilson (1999). [Acceptance and Commitment Therapy: An Experiential Approach to Behavior Change](#). Guilford Press.
- [56] Zhitao **He**, Pengfei Cao, Chenhao Wang, Zhuoran Jin, Yubo Chen, Jiexin Xu, Huaijun Li, Xiaojian Jiang, Kang Liu, Jun Zhao (2024). AgentsCourt: Building Judicial Decision-Making Agents with Court Debate Simulation and Legal Knowledge Augmentation. In: *arXiv preprint arXiv:2403.02959*.
- [57] Todd F. **Heatherton**, Lynn T. Kozlowski, R. C. Frecker, W. Rickert, J. Robinson (1989). [Measuring the heaviness of smoking: using self-reported time to the first cigarette of the day and number of cigarettes smoked per day](#). In: *British Journal of Addiction* 84.7, 791–799.
- [58] Carolyn J. **Heckman**, Ryan Egleston, Kevin Hofmann (2010). [Efficacy of Motivational Interviewing for Smoking Cessation: A Systematic Review and Meta-Analysis](#). In: *Tobacco Control* 19.5, 410–416.
- [59] Michael V. **Heinz**, Daniel M. Mackin, Brianna M. Trudeau, Sukanya Bhattacharya, Yinzhou Wang, Haley A. Banta, et al. (2025). [Randomized Trial of a Generative AI Chatbot for Mental Health Treatment](#). In: *The New England Journal of Medicine- AI* 2.4, AIoa2400802.
- [60] Jennifer **Hettema**, Julie Steele, William R. Miller (2005). [Motivational Interviewing](#). In: *Annual Review of Clinical Psychology* 1, 91–111.
- [61] John J. **Horton** (2023). [Large Language Models as Simulated Economic Agents: What Can We Learn from Homo Silicus?](#) Tech. rep. 31122. National Bureau of Economic Research.

- [62] Jon **Houck**, David Moyers, Jennifer Miller, Tim Glynn, Theresa Hallgren (2010a). [Motivational Interviewing Skill Code \(MISC\) Manual – Version 2.5](#).
- [63] Jon M. **Houck**, Jessica K. Manuel, Theresa B. Moyers (2018). [Short- and long-term effects of within-session client speech on drinking in the COMBINE study](#). In: *Journal of Studies on Alcohol and Drugs* 79.2, 217–222.
- [64] Jonathon **Houck**, Theresa Moyers, William R Miller, Laura Glynn, C Hallgreen (2010b). Manual for the Motivational Interviewing Skill Code (MISC) version 2.5. Unpublished coder’s manual.
- [65] **IBM** (2023). [IBM Granite Code Models Power watsonx Code Assistant for Z](#).
- [66] Noemi **James**, Eunyong Lee, Carey Cole, Barbara Farris, Pamela Wright (2021). [Improving Chronic Health Diseases Through Structured Smoking Cessation Education in a Rural Free Clinic](#). PhD thesis. Radford University.
- [67] Hang **Jiang**, Xiajie Zhang, Xubo Cao, Cynthia Breazeal, Deb Roy, Jad Kabbara (2024). [PersonaLLM: Investigating the Ability of Large Language Models to Express Personality Traits](#). In: *Findings of the Association for Computational Linguistics: NAACL 2024*, pp. 3605–3627.
- [68] Viviane **Josewski** (2023). [Improving access to mental health and addictions services and supports for older Indigenous adults, using a cultural safety and equity lens](#). Tech. rep. National Collaborating Centre for Indigenous Health.
- [69] Nitish **Joshi**, Javier Rando, Abulhair Saparov, Najoung Kim, He He (2024). [Personas as a Way to Model Truthfulness in Language Models](#). In: *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 6346–635.
- [70] Mina J. **Kian**, Mingyu Zong, Katrin Fischer, Abhyuday Singh, Anna-Maria Velentza, Pau Sang, Shriya Upadhyay, Anika Gupta, Misha A. Faruki, Wallace Browning, et al. (2024). [Can an LLM-Powered Socially Assistive Robot Effectively and Safely Deliver Cognitive Behavioral Therapy? A Study With University Students](#).
- [71] Hyunwoo **Kim**, Byeongchang Kim, Gunhee Kim (2020). [Will I Sound Like Me? Improving Persona Consistency in Dialogues through Pragmatic Self-Consciousness](#). In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 904–916.
- [72] Elvis **Kimara**, Kunle S. Oguntoye, Jian Sun (2025). [PersonaAI: Leveraging Retrieval-Augmented Generation and Personalized Context for AI-Driven Digital Avatars](#).
- [73] Haein **Kong**, Seonghyeon **Moon** (2025). [When LLM Therapists Become Salespeople: Evaluating Large Language Models for Ethical Motivational Interviewing](#).

- [74] Michal **Kosinski** (2023). Evaluating Large Language Models in Theory of Mind Tasks. In: *arXiv preprint arXiv:2302.02083*.
- [75] J. Richard **Landis**, Gary G. **Koch** (1977). [The Measurement of Observer Agreement for Categorical Data](#). In: *Biometrics* 33.1, 159–174. (Visited on 02/03/2025).
- [76] N. **Laverde**, C. Grevisse, S. Jaramillo, R. Manrique (2025). [Integrating large language model-based agents into a virtual patient chatbot for clinical anamnesis training](#). In: *Computational and Structural Biotechnology Journal* 25, 3156–3164.
- [77] Ellen E. **Lee**, John Torous, Munmun De Choudhury, Colin A. Depp, Sarah A. Graham, Ho-Cheol Kim, Martin P. Paulus, John H. Krystal, Dilip V. Jeste (2021). [Artificial Intelligence for Mental Health Care: Clinical Applications, Barriers, Facilitators, and Artificial Wisdom](#). In: *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging* 6.9, 856–864.
- [78] Jiwei Li, Michel Galley, Chris Brockett, Georgios Spithourakis, Jianfeng Gao, Bill Dolan (2016). [A Persona-Based Neural Conversation Model](#). In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pp. 994–1003.
- [79] Junyi Li, Xiaoxue Cheng, Wayne Xin Zhao, Jian-Yun Nie, Ji-Rong Wen (2023a). [HaluEval: A Large-Scale Hallucination Evaluation Benchmark for Large Language Models](#). In: *Proc. of the EMNLP 2023*, 7801–7820.
- [80] Margaret Li, Stephen Roller, Ilia Kulikov, Sean Welleck, Y-Lan Boureau, Kyunghyun Cho, Jason Weston (2020). [Don't Say That! Making Inconsistent Dialogue Unlikely with Unlikelihood Training](#). In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4715–4728.
- [81] Yuanzhi Li, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Yin Tat Lee, Yi Zhang (2023b). Textbooks Are All You Need II: Phi-1.5 technical report. arXiv: [2309.05463 \[cs.CL\]](#).
- [82] Sucharat **Limpanopparat**, Erin Gibson, Dr Andrew Harris (2024). [User engagement, attitudes, and the effectiveness of chatbots as a mental health intervention: A systematic review](#). In: *Computers in Human Behavior: Artificial Humans* 2.2, 100081.
- [83] N. **Lindson-Hawley**, T. Thompson, P. Aveyard (2015). [Motivational Interviewing for Smoking Cessation](#). In: *Cochrane Database of Systematic Reviews* 2015.3, CD006936.
- [84] Andy **Liu**, Mona Diab, Daniel Fried (2024). [Evaluating Large Language Model Biases in Persona-Steered Generation](#). In: *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 9832–9850.

- [85] Jane Loevinger (1957). [Objective tests as instruments of psychological theory](#). In: *Psychological Reports* 3.3, 635–694.
- [86] Brad W. Lundahl, Chelsea Kunz, Cynthia Brownell, Derrik Tollefson, Brian L. Burke (2010). [A Meta-Analysis of Motivational Interviewing: Twenty-Five Years of Empirical Studies](#). In: *Research on Social Work Practice* 20.2, 137–160.
- [87] Andrea Madotto, Zhaojiang Lin, Chien-Sheng Wu, Pascale Fung (2019). [Personalizing Dialogue Agents via Meta-Learning](#). In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 5454–5459.
- [88] Zafarullah Mahmood, Soliman Ali, Jiading Zhu, Mohamed Abdelwahab, Michelle Yu Collins, Sihan Chen, Yi Cheng Zhao, Jodi Wolff, Osnat C. Melamed, Nadia Minian, et al. (2025). [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*, pp. 25008–25043.
- [89] Stewart W Mercer, Margaret Maxwell, David Heaney, Graham CM Watt (2004). [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, 699–705.
- [90] Selina Meyer, David Elsweiler (2025). [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, 103514.
- [91] Katharine Miller (2021). [Introducing the Center for Research on Foundation Models \(CRFM\)](#). Stanford Institute for Human-Centered Artificial Intelligence (HAI) News.
- [92] William R. Miller (1983). [Motivational Interviewing with Problem Drinkers](#). In: *Behavioural Psychotherapy* 11.2, 147–172.
- [93] William R. Miller (2023). [The evolution of motivational interviewing](#). In: *Behavioural and Cognitive Psychotherapy* 51.6, 616–632.
- [94] William R. Miller, Stephen Rollnick (1991). [Motivational interviewing: Preparing people to change addictive behavior](#). The Guilford Press, New York, NY, US.
- [95] William R. Miller, Gary S. Rose (2009). [Toward a Theory of Motivational Interviewing](#). In: *American Psychologist* 64.6, 527–537.
- [96] Miller, William R. and Rollnick, Stephen (2023). [Motivational Interviewing: Helping People Change and Grow](#). 4th. The Guilford Press.

- [97] Adam S Miner, Liliana Laranjo, Ahmet Baki Kocaballi (2020). [Talking to Machines About Mental Health: A Scoping Review of Conversational Agents in Psychiatry](#). In: *JMIR Mental Health* 7.6, e18572.
- [98] Adam S. Miner et al. (2016). [Smartphone-based conversational agents and responses to questions about mental health, interpersonal violence, and physical health](#). In: *JAMA Internal Medicine*.
- [99] Theresa B. Moyers, Tim Martin, Jon M. Houck, Paulette J. Christopher, J. Scott Tonigan (2009). [From in-session behaviors to drinking outcomes: A causal chain for motivational interviewing](#). In: *Journal of Consulting and Clinical Psychology* 77.6, 1113–1124.
- [100] Theresa B. Moyers, Lauren N. Rowell, Jennifer K. Manuel, Danielle Ernst, William R. Miller (2016a). [The Motivational Interviewing Treatment Integrity Code \(MITI 4\): Rationale, Preliminary Reliability and Validity](#). In: *Journal of Substance Abuse Treatment* 65, 36–42.
- [101] Theresa B. Moyers, Lauren N. Rowell, Jennifer K. Manuel, Denise Ernst, Jon M. Houck (2016b). [The Motivational Interviewing Treatment Integrity Code \(MITI 4\): Rationale, Preliminary Reliability and Validity](#). In: *Journal of Substance Abuse Treatment* 65, 36–42.
- [102] National Collaborating Centre for Indigenous Health (2020). [Poverty as a social determinant of First Nations, Inuit, and Métis health](#). (Visited on 08/18/2025).
- [103] Tong Niu, Mohit Bansal (2018). [Polite Dialogue Generation Without Parallel Data](#). In: *Transactions of the Association for Computational Linguistics* 6, 373–389.
- [104] OpenAI (2023). [GPT-4 Technical Report](#).
- [105] OpenAI (2024). [GPT-4o System Card](#).
- [106] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. (2022). [Training language models to follow instructions with human feedback](#). In: *Proceedings of the 36th International Conference on Neural Information Processing Systems*. NIPS '22. Curran Associates Inc.
- [107] Joon Sung Park, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, Michael S. Bernstein (2023). [Generative Agents: Interactive Simulacra of Human Behavior](#). In: *Proceedings of the 36th ACM Symposium on User Interface Software and Technology (UIST)*.
- [108] Verónica Pérez-Rosas, Xinyi Wu, Kenneth Resnicow, Rada Mihalcea (2019). [What Makes a Good Counselor? Learning to Distinguish between High-quality and Low-quality Counseling Conversations](#). In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 926–935.

- [109] Olga **Perski**, Ann Blandford, Robert West, Susan Michie (2017). [Conceptualising engagement with digital behaviour change interventions: a systematic review using principles from critical interpretive synthesis](#). In: *Translational Behavioral Medicine* 7.2, 254–267.
- [110] Abbey E. **Poirier** et al. (2019). [Estimates of the current and future burden of cancer attributable to active and passive tobacco smoking in Canada](#). In: *Preventive Medicine* 122, 9–19.
- [111] J. **Proudfoot**, D. Goldberg, A. Mann, B. Everitt, I. Marks, J. A. Gray (2003). [Computerized, interactive, multimedia cognitive-behavioural program for anxiety and depression in general practice](#). In: *Psychological Medicine* 33.2, 217–227.
- [112] Stephen **Roller**, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Mye Ott, Eric Michael Smith, Y-Lan Boureau, Jason Weston (2021). [Recipes for Building an Open-Domain Chatbot](#). In: *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pp. 300–325.
- [113] Stephen **Rollnick**, Nick Heather, Richard Gold, Wayne Hall (1992). [Development of a short 'readiness to change' questionnaire for use in brief, opportunistic interventions among excessive drinkers](#). In: *British Journal of Addiction* 87.5, 743–754.
- [114] Stephen **Rollnick**, William R. **Miller** (1995). [What is Motivational Interviewing?](#) In: *Behavioural and Cognitive Psychotherapy* 23.4, 325–334.
- [115] Stephen **Rollnick**, William R. Miller, Christopher C. Butler (2008). *Motivational Interviewing in Health Care: Helping Patients Change Behavior*. The Guilford Press.
- [116] Jonathan **Rose**, Peter Selby, Matt Ratto, Osnat Melamed, Marta Maslej, Nadia Minian, Mohamed Abdelwahab, Jiading Zhu, Zafarullah Mahmood, Soliman Ali, et al. (2025). [Motivational Interviewing-Based Chatbot for Smoking Cessation: Human Participant Ethics Protocol](#). Approved Human Participant Research Protocol, Protocol #49997, Version 0001, Approved on 2025-07-08.
- [117] Sahand **Sabour**, Siyang Liu, Zheyuan Zhang, June Liu, Jinfeng Zhou, Alvionna Sunaryo, Tatia Lee, Rada Mihalcea, Minlie Huang (2024). [EmoBench: Evaluating the Emotional Intelligence of Large Language Models](#). In: *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, pp. 5986–6004.
- [118] Ahson **Saiyed**, John Layton, Brian Borsari, Jing Cheng, Tatyana Kanzaveli, Maksim Tsvetovat, Jason Satterfield (2022). [Technology-Assisted Motivational Interviewing: Developing a Scalable Framework for Promoting Engagement with Tobacco Cessation Using NLP and Machine Learning](#). In: *Procedia Computer Science* 206, 121–131.
- [119] Jr. **Sampson** (1986). [The use of computer-assisted instruction in support of psychotherapeutic processes](#). In: *Computers in Human Behavior* 2, 1–19.

- [120] Vinay **Samuel**, Henry Peng Zou, Yue Zhou, Shreyas Chaudhari, Ashwin Kalyan, Tanmay Rajpurohit, Ameet Deshpande, Karthik Narasimhan, Vishvak Murahari (2025). [PersonaGym: Evaluating Persona Agents and LLMs](#).
- [121] **Sanic Community Organization** (2025). Sanic. <https://github.com/sanic-org/sanic>.
- [122] Shibani **Santurkar**, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, Tatsunori Hashimoto (2023). [Whose Opinions Do Language Models Reflect?](#) In: *Proceedings of the 40th International Conference on Machine Learning*. Ed. by Andreas **Krause**, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, Jonathan Scarlett. Vol. 202. Proceedings of Machine Learning Research. PMLR, pp. 29971–30004.
- [123] Till **Scholich**, Maya Barr, Shannon S. Wiltsey-Stirman, Shriti Raj (2025). [A Comparison of Responses from Human Therapists and Large Language Model-Based Chatbots to Assess Therapeutic Communication: Mixed Methods Study](#). In: *JMIR Mental Health* 12, e69709.
- [124] Lennart **Seitz** (2024). [Artificial empathy in healthcare chatbots: Does it feel authentic?](#) In: *Computers in Human Behavior: Artificial Humans* 2.1, 100067.
- [125] P. M. **Selmi**, M. H. Klein, J. H. Greist, S. P. Sorrell, H. P. Erdman (1990). [Computer-administered cognitive-behavioral therapy for depression](#). In: *American Journal of Psychiatry* 147.1, 51–56.
- [126] D. **Servan-Schreiber** (1986). [Artificial intelligence and psychiatry](#). In: *Journal of Nervous and Mental Disease* 174.4, 191–202.
- [127] Siqi **Shen**, Charles Welch, Rada Mihalcea, Verónica Pérez-Rosas (2020). [Counseling-Style Reflection Generation Using Generative Pretrained Transformers with Augmented Context](#). In: *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pp. 10–20.
- [128] **Shopify** (2025). Domain SEO Explained: Choosing an SEO-Friendly Domain Name (2025). <https://www.shopify.com/blog/domain-seo>.
- [129] Bangzhao **Shu**, Lechen Zhang, Minje Choi, Lavinia Dunagan, Lajanugen Logeswaran, Moontae Lee, Dallas Card, David Jurgens (2024). [You don't need a personality test to know these models are unreliable: Assessing the Reliability of Large Language Models on Psychometric Instruments](#). In: *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 5263–5281.
- [130] Kurt **Shuster**, Jing Xu, Mojtaba Komeili, Da Ju, Eric Michael Smith, Stephen Roller, Megan Ung, Moya Chen, Kushal Arora, Joshua Lane, et al. (2022). [BlenderBot 3: a deployed conversational agent that continually learns to responsibly engage](#).

- [131] Renata **Sloane** (2025). [Containerization with Docker and Kubernetes in the Cloud: Deploying and Managing Containerized Applications in the Cloud Using Docker and Kubernetes](#). Amazon Digital Services LLC - Kdp.
- [132] Haoyu **Song**, Yan Wang, Wei-Nan Zhang, Xiaojiang Liu, Ting Liu (2020). [Generate, Delete and Rewrite: A Three-Stage Framework for Improving Persona Consistency of Dialogue Generation](#). In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 5821–5831.
- [133] Sofia Eleni **Spatharioti**, David M. Rothschild, Daniel G. Goldstein, Jake M. Hofman (2023). Comparing Traditional and LLM-based Search for Consumer Choice: A Randomized Experiment. In: *arXiv preprint arXiv:2307.03744*.
- [134] Lindsay F **Stead**, Diana Buitrago, Nereida Preciado, German Sanchez, Tim Lancaster, Jamie Hartmann-Boyce (2013). [Physician advice for smoking cessation](#).
- [135] Akhilesh **Sudhakar**, Bhargav Upadhyay, Arjun Maheswaran (2019). [“Transforming” Delete, Retrieve, Generate Approach for Controlled Text Style Transfer](#). In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 3269–3279.
- [136] Joseph **Suh**, Erfan Jahanparast, Suhong Moon, Minwoo Kang, Serina Chang (2025). Language Model Fine-Tuning on Scaled Survey Data for Predicting Distributions of Public Opinions. In: *arXiv preprint arXiv:2502.16761*.
- [137] John **Torous**, Laura **Roberts** (2017). [Digital mental health and COVID-19: Using technology today to accelerate the curve on access and quality tomorrow](#). In: *JMIR Mental Health*.
- [138] Hugo **Touvron**, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, et al. (2023). Llama 2: Open Foundation and Fine-Tuned Chat Models. arXiv: [2307.09288 \[cs.CL\]](#).
- [139] Ashish **Vaswani** et al. (2017). [Attention is All You Need](#). In: *Advances in Neural Information Processing Systems* 30.
- [140] Jason **Wei**, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, Denny Zhou (2022). [Chain-of-thought prompting elicits reasoning in large language models](#). In: *Proceedings of the 36th International Conference on Neural Information Processing Systems. NIPS '22*.
- [141] Joseph **Weizenbaum** (1966). [ELIZA — A Computer Program for the Study of Natural Language Communication Between Man and Machine](#). In: *Communications of the ACM* 9.1, 36–45.

- [142] Sean **Welleck**, Jason Weston, Arthur Szlam, Kyunghyun Cho (2019). [Dialogue Natural Language Inference](#). In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 3731–3741.
- [143] Robert **West**, Tim **Sohal** (2006). “Catastrophic” Pathways to Smoking Cessation: Findings from [National Survey](#). In: *BMJ* 332.7539, 458–460.
- [144] Robyn **Whittaker**, Hayden McRobbie, Chris Bullen, Anthony Rodgers, Yulong Gu (2016). [Mobile phone-based interventions for smoking cessation](#). In: *Cochrane Database of Systematic Reviews* 4, CD006611.
- [145] Shenghan **Wu**, Yang Deng, Yimo Zhu, Wynne Hsu, Mong Li Lee (2025). [From Personas to Talks: Revisiting the Impact of Personas on LLM-Synthesized Emotional Support Conversations](#). arXiv: 2502.11451 [cs.CL].
- [146] **xAI** (2023). [Announcing Grok](#).
- [147] Zhouhang **Xie**, Bodhisattwa Prasad Majumder, Mengjie Zhao, Yoshinori Maeda, Keiichi Yamada, Hiromi Wakaki, Julian McAuley (2024). [Few-shot Dialogue Strategy Learning for Motivational Interviewing via Inductive Reasoning](#). In: *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 13207–13219.
- [148] Jing **Xu**, Arthur Szlam, Jason Weston (2022). [Beyond Goldfish Memory: Long-Term Open-Domain Conversation](#). In: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, pp. 5180–5197.
- [149] Diyi **Yang**, Sherry Tongshuang Wu, Marti A. Hearst (2024). [Human-AI Interaction in the Age of LLMs](#). In: *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 34–38.
- [150] Xindong **Ye**, Xiaofen Shan, Yunfang Tu, Yuanyuan Zhang, et al. (2025). [Examining the Efficacy of Large Language Models for Mitigating Depression and Anxiety Among Chinese Students: A Randomized Controlled Trial](#). In: *CIN: Computers, Informatics, Nursing* 10.[1097].
- [151] Faika **Zanjani**, Bree Miller, Nicholas Turiano, Jennifer Ross, David Oslin (2008). [Effectiveness of telephone-based referral care management, a brief intervention to improve psychiatric treatment engagement](#). In: *Psychiatric Services* 59.7, 776–781.
- [152] Saizheng **Zhang**, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, Jason Weston (2018). [Personalizing Dialogue Agents: I have a dog, do you have pets too?](#) In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, 2204–2213.

- [153] Barret Zoph, Colin Raffel, Dale Schuurmans, Dani Yogatama, Denny Zhou, Don Metzler, Ed H. Chi, Jason Wei, Jeff Dean, Liam B. Fedus, et al. (2022). [Emergent abilities of large language models](#). In: *Transactions on Machine Learning Research*.