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

Write later

Write later

Acknowledgements

Write later

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Tobacco use remains the leading cause of preventable disease and death in Canada. More than 70% of the lung cancer cases in the country are estimated to be 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 a higher prevalence among Indigenous communities, individuals living in rural areas, and those with low income (Canadian Partnership Against Cancer, 2020). These groups also suffer from 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, particularly therapeutic modalities such as cognitive behavioural therapy (CBT) (Beck et al., 1979), acceptance and commitment therapy (ACT) (Hayes et al., 1999) and motivational interviewing (MI) (Miller and Rollnick, 2012). However, access to such interventions 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 and economically effective, publicly funded CBT remains scarce in Canada (Llewellyn et al., 2006). 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 socio-economic resources, all constraining access (Crosato and Leipert, 2011). Third, there is a significant shortage of culturally safe mental health services tailored to the unique historical and social contexts of Indigenous populations (Jones et al., 2021). Indigenous communities experience systemic underfunding, long wait times, inadequate provider availability, and cultural mismatches between Indigenous clients and predominantly non-Indigenous therapists (CanCOVID Issue Note, 2023; 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 when no alternatives exist. 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 suited 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, 2013). It is designed to help individuals strengthen their intrinsic motivation to change their health-compromising behaviours, including 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; Hettema et al., 2005b).

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., 2023a). Resolving this ambivalence is a necessary precursor to behaviour change. Numerous studies have demonstrated that MI increases the number of quit attempts and enhances 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, 2009a).

1.2 Large Language Models and the Automation of Talk Therapy

The automation of talk therapy, once constrained to rule-based systems and predefined scripts, has been profoundly transformed by the advent of LLMs. The development of transformer-based architectures (Vaswani et al., 2017), particularly autoregressive generative models like GPT-3 and GPT-4, 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; Torous et al., 2023).

Several key technical innovations have facilitated the adaptation of LLMs to therapeutic contexts. The self-attention mechanism within transformer models supports long-range dependency modelling and nuanced context tracking across multi-turn dialogues (Vaswani et al., 2017). Fine-tuning an LLM on domain-specific corpora (e.g., MI transcripts) could make the model outputs clinically grounded (Valentino et al., 2024), while reinforcement learning with human feedback (RLHF) has been shown to improve the relevance, tone, and empathetic quality of generated responses (Ouyang et al., 2022; Gilson et al., 2023). 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 toward the decision to quit. This involves designing and implementing an MI counsellor that adheres to the core principles of MI, evaluating its effectiveness with real 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. Furthermore, 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

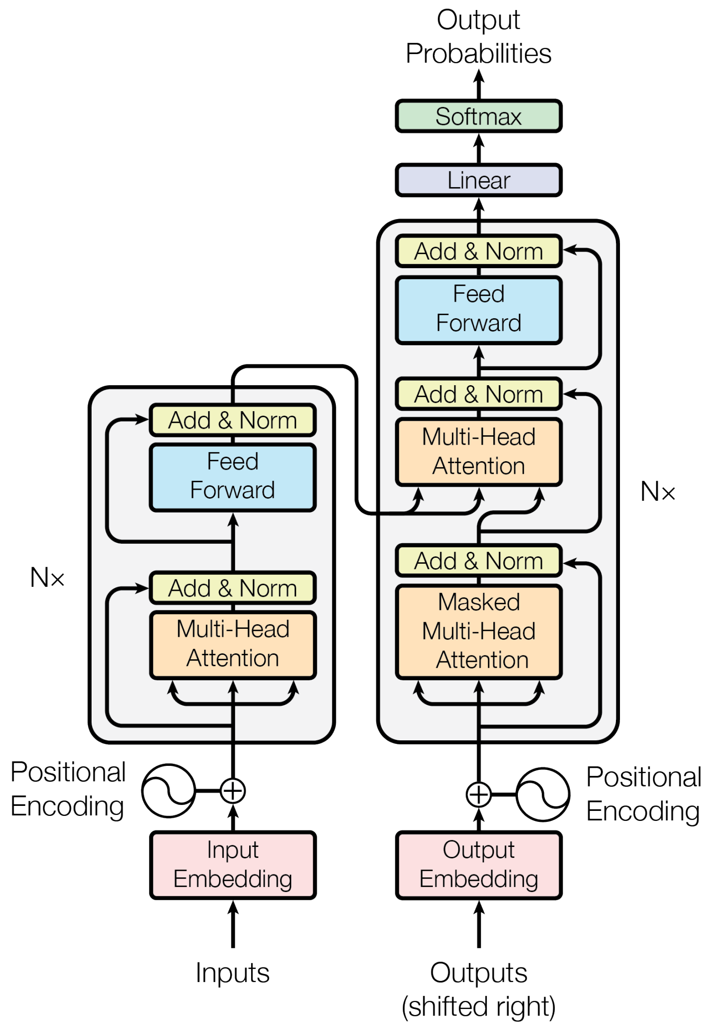


Figure 1.1: Illustration of the Transformer architecture adapted from Vaswani et al. (2017).

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-driven 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 dissertation 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 and evaluation methodology. Chapter 4 reports the results of a study in which recruited smokers interacted with the chatbot. Chapter 5 describes a method for constructing synthetic smokers using large language models, prompted to exhibit demographic and behavioural traits derived from real smokers. Chapter 6 introduces a framework for evaluating synthetic smokers in their ability to substitute for human participants in controlled behavioural studies, and presents the corresponding evaluation results. Chapter 7 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. Next, we 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 particular emphasis on LLM-based synthetic patients designed for behavioural research. We particularly focus on the techniques of persona *installation* via prompting, which seeks to *install* 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 and Rollnick, 2023). MI is defined as a collaborative, client-centred conversational method designed to elicit intrinsic motivation for change by helping individuals resolve ambivalence (Miller and Rollnick, 2002). 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 the 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.

change talk
sustain talk

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

2.1.1 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 works to reveal any discrepancy between the client's own goals or values and their current behaviour (Miller and Rollnick, 2013). Third, the counsellor acknowledges the client's resistance rather than confronting it; resistant remarks are met with understanding and are used as opportunities to further explore the client's thoughts, 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 decide to (Miller and Rollnick, 2013). 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 (Rollnick et al., 1997; Miller and Rollnick, 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 cultivate 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 Skills in Motivational Interviewing

2.1.2 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 impact on client motivation. One common framework 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., 2010). Using the MISC, independent annotators (or *coders*) can rate how **coding**
well a counsellor's statement aligns with MI techniques (for example, counting reflections, questions, advice, etc.) 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 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.

Effectiveness of MI in Smoking Cessation

effectiveness

MI effectivenessMI 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 navigate 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 significantly increased the likelihood of abstinence compared to 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 enhances both quit attempts and abstinence, especially when delivered by trained practitioners in clinical or community settings

The effectiveness of MI is consistent across diverse 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 fostering 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 enhances intermediate factors such as motivation, readiness to quit, and self-efficacy (Boudreaux et al., 2012; Hettema et al., 2005a). These motivational gains are critical, 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 has been driven by foundational language models (Miller, 2021) — extremely large neural networks pre-trained on vast corpora of text, which can be adapted to myriad tasks. These models serve as a foundation that can be specialized via *fine-tuning* or *prompting* for specific applications. Crucially, modern foundation models leverage the **transformer** architecture (discussed in Section 2.2.1) 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 subtle interaction norms. For example, 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 diverse 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). As will be demonstrated later, in this work, we prompted a foundation model with an expert-informed prompt. This method leveraged the model's generative capabilities while allowing the customization of the output to be MI-consistent, among other things. 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).

The general trend is that foundation models are becoming more knowledgeable, more context-aware, and better at following complex instructions. This bodes well for therapeutic chatbots: as these models advance, a carefully adapted version can exhibit even more natural dialogue and sophisticated motivational strategies. At the same time, research into controllability — 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.

2.2.1 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 a 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 context. 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-4 (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 (Team, 2025) employ essentially 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 subtle dialogue skills that smaller models lack (Zoph et al., 2022; Berti et al., 2025).

2.3 Chatbots for Talk Therapy

Computer-based *chatbots* have long been explored as a means to deliver talk therapy through natural language. Early attempts date back to the 1960s, when *rule-based* 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* (Güzeldere and Franchi, 1995), 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 hand-crafted rules and keywords rather than any true language understanding. These pioneering 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 explored 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). But by and large, 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 significantly 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 frameworks, 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 significantly 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 (2018). They 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 significant 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.

So far, 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 the users. As such, the section below focuses on the recent developments in LLM-based chatbots.

2.3.1 LLM-based Chatbots for Talk Therapy

Recent breakthroughs in LLMs have sparked 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 actually rated the model-generated reflective statements 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. (2023b) 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 powerful foundational language models, either *fine-tuned* on therapy data or guided via careful prompting. For instance, **MIcha** is a GPT-4-based chatbot designed to deliver motivational interviewing for behaviour change (Meyer and Elsweiler, 2025). In a randomized controlled trial, MIcha's brief conversations significantly increased users' readiness to change unhealthy behaviours compared to control (an unprompted GPT-4o instance). On a 0-10 *Readiness to Change* (Biener and Abrams, 1991) scale, interactions with *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 Elsweiler, 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 **MIcha** 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 significantly 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 explore this idea in depth in Section 2.6. 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-4 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 significant challenges are yet to be overcome. One such challenge is the LLM's tendency for **hallucination**, i.e., 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., 2023). 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 explored include grounding the model in verified psychoeducational content and implementing filters for medical advice (Amugongo et al., 2025).

There is also a subtle 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' than with a human, especially over time. The empathy is "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 nuanced 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 explore 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.

2.4 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 captivating 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 enhance 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.

2.4.1 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 paradigm, 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 significantly 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 behaviuor (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 and et al., 2023). 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; Yang, 2023). 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.

2.4.2 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, (Zheng et al., 2021) 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 (Zheng et al., 2021).

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 dynamic persona switching.

2.4.3 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. State-of-the-art 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 dedicated 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 robust 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 evolve with user interactions (**?**).

2.5 Evaluation of LLM-based Persona Creation

To gauge how well an model represents a persona, researchers have employed various evaluation methods, ranging from automatic metrics to human judgment and psychometric tests. One fundamental 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 lead 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., 2023). (Jiang et al., 2023) conducted an extensive experiment assigning GPT-4 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., 2023). 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 robust 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., 2023). Similarly, the recent PersonaGym framework (Liu et al. 2024) 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, behavior, and preferences (Samuel et al., 2024). Such multi-dimensional evaluations aim to go beyond surface traits, testing whether the persona remains consistently applied across different contexts and over extended dialogues.

2.5.1 Known Issues with LLM-based Persona Creation

Current LLM persona construction techniques face significant 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 multifaceted, less common identities, tending instead to regress to more stereotypical patterns present in 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 a 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. (**?**) 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 (Smith et al., 2020).

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

doppelgänger

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 would 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 essential. 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.

2.6.1 LLM doppelgängers in social surveys

(Argyle et al., 2023) introduced the concept of 'silicon samples' as proxies for human survey respondents, which we refer to as doppelgängers. They conditioned GPT-3 with thousands of real participants' demographic backstories (e.g. age, gender, race, education, political affiliation) from U.S. survey data and generated that model's answers to the same questionnaires those 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 nuanced 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. This suggests the LLM contained latent knowledge of complex demographic response tendencies, and proper conditioning could unlock these to mimic specific populations.

However, their validation method was inherently 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 fundamentally, 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 was unclear whether the same fidelity would hold in open-ended behavioural or linguistic responses. In short, while the work provided evidence that LLMs can emulate group-level attitudes with impressive granularity, they overlooked dynamic interaction traits and relied 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.

2.6.2 Simulated participants in classic experiments

Aher et al. (Aher et al., 2023) moved beyond surveys to test whether LLM-based agents could replicate findings from canonical psychology and economics experiments. They proposed a "Turing Experiment" framework in which an LLM is prompted to simulate not one individual, but a whole sample of participants in a controlled experiment (e.g. playing roles in the Ultimatum Game or responding to a moral dilemma). The validation of these agents hinged on whether well-established human phenomena emerged from the LLM simulations. Indeed, for several classic studies, the aggregate behaviour of the simulated participants mirrored human results: for example, GPT-4 agents playing the Ultimatum Game produced offer acceptance rates and splits comparable to those seen with real people, and when simulating subjects in Milgram's obedience scenario, the model's responses reflected the expected pattern of compliance versus refusal under authority pressure. These replications suggest that the doppelgängers had internalized certain human-like behavioural patterns. However, Aher et al.'s validation method reveals its own limitations. First, it evaluates fidelity only indirectly via outcome metrics. If the model reproduces the correct average outcome (e.g. 70% compliance in Milgram's paradigm), one assumes the underlying agent behaviours are human-like. But this could mask important differences – the LLM might arrive at the "right" answer for the wrong reasons. Indeed, the authors discovered a notable distortion: in a "wisdom of the crowd" experiment, the LLM agents were *too* accurate, displaying a so-called "hyper-accuracy" that exceeds typical human performance and thus failing to replicate the benign errors and variance that real groups exhibit. This highlights a broader issue: LLM doppelgängers may leverage their vast knowledge and rationality in ways actual humans would not, raising concern that they might achieve correct aggregate results without truly behaving like humans at the individual level. In other words, passing a statistical Turing test for an experimental outcome does not guarantee that the simulated cognitive processes or linguistic behaviors match those of real people. Additionally, because Aher et al.'s approach focuses on group phenomena, it does not validate whether any single synthetic subject maintains a believable persona throughout an interaction – the emphasis is on replicating aggregate findings rather than person-specific fidelity. This approach is powerful for certain research questions (e.g. testing hypotheses on virtual populations), but it offers limited insight into how convincing or realistic a one-on-one LLM-driven "digital twin" would appear in a behavioural intervention setting such as counseling.

2.6.3 Believable multi-agent simulations

Park et al. (Park et al., 2023) demonstrated a very different use-case of LLM personas: they populated an interactive virtual world with twenty-five generative agents, each defined by a brief persona description and memories, and observed rich social behaviors emerge. These agents (powered by a GPT-4 backend) acted out daily routines in a sandbox environment (akin to *The Sims*), initiating conversations, forming new relationships, and coordinating events autonomously. For example, one agent deciding to host a Valentine's Day party led to a cascade of invitations and plans among the others, resulting in a coordinated gathering that was never explicitly hard-coded. This study's focus was on creating *believable* behavior – the agents "felt" like independent characters in a little society. Park et al. validated this believability primarily through qualitative assessment and ablation studies. They showed that specific architectural components (long-term memory, reflection, and planning modules layered on top of the base LLM) were each necessary for producing coherent, lifelike patterns; removing any one component degraded the realism of agent behaviors. They also reported anecdotal evidence that outside observers found the agents' interactions plausible. However, the validation here was not rigorously quantified via human rating scales or direct comparison to real human group behavior – it was largely the researchers' judgment that the agents were "lifelike." The evaluation was thus somewhat subjective, and it centered on internal consistency and emergent social dynamics rather than fidelity to any external ground truth. Each agent's persona was fictional, with no one-to-one human counterpart to validate against. This means the study did not answer how accurately an LLM agent could mimic a *particular* real person or demographic; it only demonstrated that, in general, a network of LLM agents can produce superficially realistic social interactions. The lack of systematic human evaluation (e.g. asking blinded judges to distinguish LLM-agent conversations from human-human ones) is a limitation in gauging true human-likeness. Moreover, because the agents existed in a constrained sandbox world, their "believability" might rely in part on the forgiving context – a real conversational setting with complex, goal-directed dialogue (such as psychotherapy) might demand a higher standard of realism than simply wandering a virtual town and chitchatting. Thus, while Park et al.'s generative agents represent a milestone in multi-agent simulation, the means of validating their human resemblance remained informal and their applicability to behaviour change dialogues was not tested.

2.6.4 Synthetic patients in healthcare dialogues

A complementary evaluation approach was taken by Haider et al. (Haider et al., 2025), who assessed multi-turn clinical conversations generated by various LLMs (ChatGPT and others) from the perspective of overall quality and persona consistency. Rather than focusing on one patient profile at a time, they had each model produce ten different patient–physician dialogues in a plastic surgery context and then asked domain experts to rate each dialogue on seven criteria: medical accuracy, realism of the interaction, consistency with the patient's persona, level of empathy, relevancy of the content, and overall usefulness of the dialog for training. This yielded a large number of expert evaluations (840 ratings in total). The outcome was that all models performed remarkably well by these metrics – mean scores exceeding 4.5 on a 5-point scale for every criterion, with some models (notably a fine-tuned Gemini model) scoring a perfect 5.0 in multiple categories. At face value, such results imply that the LLM-generated patients were highly realistic and stayed in character. But this all-high-marks outcome also underlines a limitation: the evaluation rubric may have been too coarse or forgiving, yielding ceiling effects that make it hard to distinguish nuances. If every conversation is rated almost 5/5 on persona consistency, one wonders if the raters were perhaps overly impressed by the fluent, information-rich responses of the LLMs (which can seem "realistic" compared to stilted rule-based chatbots of the past) and thus less attuned to any subtle unnatural qualities. The study did not include any baseline of real human conversations for reference, nor any quantitative measure of differences between model and human language. Furthermore, persona consistency was evaluated in a generic sense (did the patient maintain the same character throughout the dialogue), but not against an external ground truth persona. In practice, the patient personas were simple scenario descriptions (e.g. "40-year-old with breast cancer seeking reconstruction, very anxious about surgery"), and consistency just meant the model didn't contradict the scenario. This is a low bar for validation—far from confirming that the LLM can mimic a specific real patient's mannerisms or decision-making. Thus, while Haider et al. demonstrate that modern LLMs can produce medically and linguistically plausible patient dialogues (a positive sign for using such synthetic data in training), their validation via expert scoring provides limited insight into finer-grained fidelity and may gloss over deficiencies that a more discriminating test could reveal.

Recently, Öncü et al. ("Onc"u et al., 2025) provided a different perspective by deploying a ChatGPT-4-derived standardized patient in live interactions with medical interns. Instead of offline ratings, their validation came from direct user experience: 21 final-year medical students each conducted two five-minute interviews with the AI patient (cases of hypertension and brucellosis) and then reflected on the process. The AI doppelgänger successfully engaged the trainees in history-taking dialogues, answering their questions and simulating symptoms in real-time. Notably, all participants reported that they were satisfied with the AI-patient encounter and found it a useful, low-stakes opportunity to practice clinical reasoning. The observers (clinical educators) also noted that the LLM generally stayed in role and responded appropriately, though there were some technical glitches (connection drop-outs and occasional misunderstood questions) that interrupted a few sessions. The lack of adverse or blatantly unrealistic responses (and the interns' willingness to continue using such tools) suggests the doppelgänger met a baseline of credibility in its behavior. However, as a validation, this pilot study is informal. It essentially asks, "Did this seem okay to you?" rather than rigorously measuring fidelity. The interns were aware they were talking to an AI, so their bar for "realism" may have been lower than if they were truly blinded. The study did not attempt to compare the AI's responses to how an actual patient might have answered the same questions, nor did it have external experts rate the dialogues for authenticity. In short, the evidence of validity is mainly that users tolerated and liked the experience. While this is encouraging for adoption, it does not deeply probe whether the AI patient's behaviour diverged in any systematic ways from real patient behaviour. For instance, an AI might consistently provide more detailed answers than a typical patient, or never display certain human quirks like pausing to recall information or expressing confusion – none of which would be captured by the simple satisfaction survey used. Thus, Öncü et al.'s work underscores the feasibility and acceptability of LLM doppelgängers in a practical setting, but it leaves open many questions about how to objectively verify that these agents are truly *modeling* human behaviour rather than just engaging in superficially correct dialogue.

Chapter 3

Design & Deployment of MIBot for Human Feasibility Study

In this chapter we present the iterative development and evaluation of a single-prompted, LLM-based motivational interviewing chatbot designed to help ambivalent smokers resolve their ambivalence about quitting. While the system targets a specific use case, the principles and methods described here are readily transferable to other MI-based chatbots with different therapeutic objectives. The development and evaluation processes have been described in detail by Mahmood et al. (2025), with no discussion of deployment. This chapter complements that work by focusing on the technologies, design considerations, and implementation choices underlying the chatbot's deployment.

Section 3.1 outlines the iterative prompt-engineering process, guided by expert feedback, that shaped the chatbot's MI skills. Section 3.2 describes the complete system architecture, which includes not only the prompted LLM instance, but also a set of *observers* (modules that monitor specific aspects of the conversation in real time and can intervene to adjust its trajectory). Section 3.3 describes the implementation of the chatbot using OpenAI APIs. We also highlight the balance between writing new pieces of the system by following DevOps best practices and maximizing reuse of legacy code. Section 3.5 details the design of a human-trial feasibility study to assess the chatbot's effectiveness. Chapter 4 reports the results of this feasibility study.

3.1 Chatbot Design Process

The design of MIBot followed a clinician-informed, iterative process, combining expertise in MI with prompt engineering for a state-of-the-art LLM (OpenAI, 2024).

3.1.1 Iterative Prompt Development

MIBot's prompt was refined 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 that were given detailed backstories and instructed to role-play smokers with varying attitudes toward quitting. Chapter 5 describes in detail the creation of these basic LLM-based virtual smokers, and more research to improve them.

2. **Human role-playing as smokers** — members of our research team who adopted smoker

personas and talked to 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, among other aspects. These findings informed successive prompt revisions. This process produced a series of targeted prompt revisions, the most consequential of which were:

1. **Utterance Length Control.** Early versions tended toward long, paragraph-like responses, which risked dominating the conversation - which is the antithesis of MI, where the client is supposed to lead the contemplation/thinking. 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, exploratory stance.

4. **Rapport-Building Before Smoking Focus.** Early-in-this-process prompts 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 to take towards quitting.

This iterative process continued until virtual and role-played conversations consistently met MI quality expectations by informal consensus. The final prompt incorporated detailed guidance on pacing, empathy, reflection types, and decision points to enter the planning phase.

Tables 3.1 and 3.2 showcase how the prompt evolved from a simple instruction instructing the LLM about its role, vs. a comprehensive guideline in order to fix issues with the chatbot.

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: Initial 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.

Table 3.2: Final MIBot Prompt

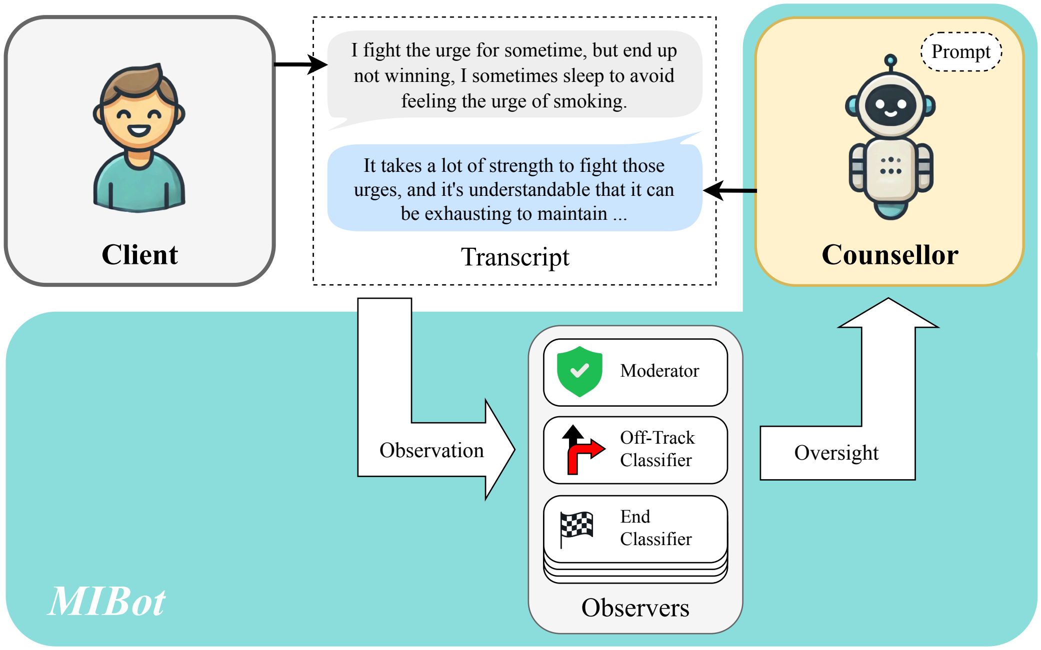


Figure 3.1: Overview of the MIBot system, taken from Mahmood et al. (2025)

3.2 Observers

To enhance safety in the deployment of MIBot, the core counsellor agent was augmented 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 could be used to intervene when necessary. Each observer was specialized through prompt engineering to perform a specific task in real-time.

3.2.1 Moderator

The *Moderator* evaluates the counsellor's most recent utterance for potential harm, prior to displaying to the client. While OpenAI's internal safety systems mitigate many risks, they do not address all possible counterproductive counselling behaviours, such as inadvertently reinforcing *sustain talk* or suggesting self-harm. The Moderator was deliberately configured for high sensitivity, accepting a higher false positive rate to reduce the risk of harmful or counter-productive 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 intervention. In a live deployment, it could instead 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 key 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 refinement of their sensitivity–specificity balance without impacting 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, make up 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.

At the heart 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. At the start of a conversation, the `Conversation` object is initialized, which creates and stores a list of `Observer` instances. It also creates a `Counsellor` object that encapsulates the chatbot logic (Figure 3.2). When a user sends a request to the `/chat` endpoint containing the client's message and `client_id` in a structured JSON object, Sanic's route handler 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 `ChatCompletion` response. The `ChatCompletion` response contains the generated counsellor turn, which is returned to the `Conversation` object.

Each `Observer` attached to the `Conversation` inherits from a base class defining an asynchronous `observe()` method. As noted earlier, the `Moderation` observer screens counsellor utterances for safety and appropriateness; the `Offtrack` observer assesses whether the client is steering the conversation away from smoking cessation; and the `EndClassifier` 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 loops through all `Observer` objects, collects their observations, and makes real-time 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 passed to Sanic's response handler, which packages it with metadata into a JSON object and sends it to the client.

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` response so that load balancers and orchestrators can perform health checks; and `/info` exposes build metadata such as the current version of the prompt.

[1] For the human feasibility study, in order 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 Section ?? for further details.

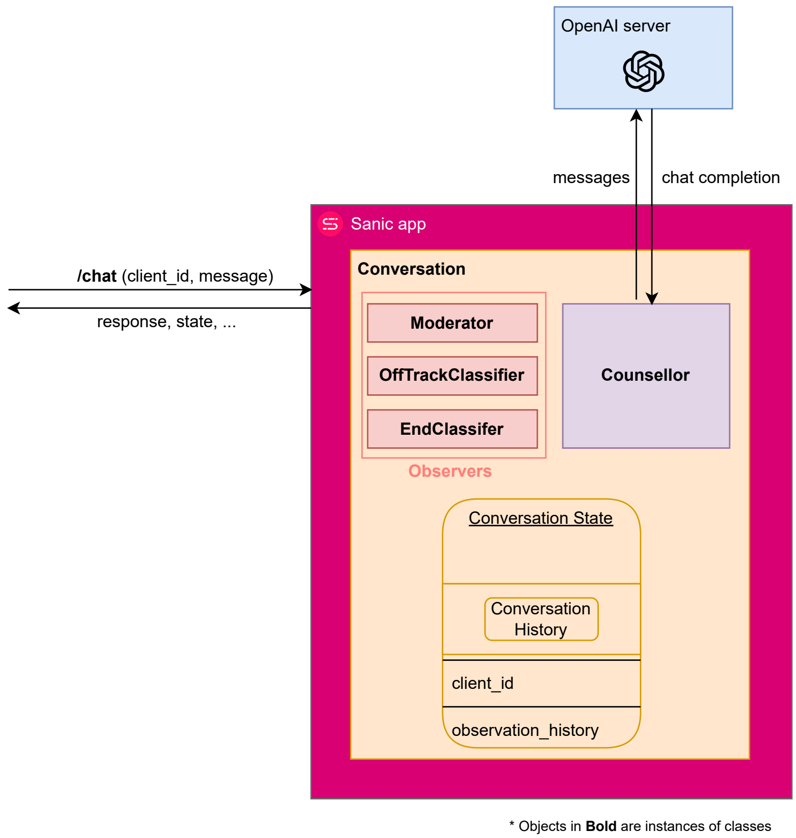


Figure 3.2: Overview of the Sanic application that contains MIBot code and exposes REST APIs.

3.3.2 Containerization

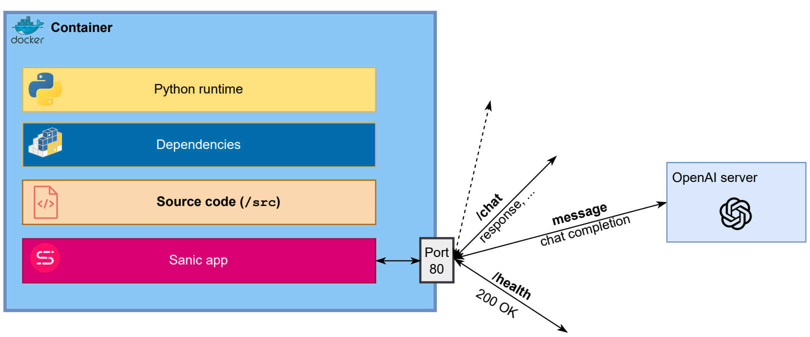


Figure 3.3: Containerized Sanic application.

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:

- **Environment Consistency:** Containers encapsulate all dependencies, ensuring that the application behaves identically across local development machines, staging environments, and production servers. This eliminates the "works on my machine" problem.
- **Portability:** The same container image can run on any system that supports Docker (or an OCI-compliant runtime), whether on a developer's laptop, an on-prem cluster, or a cloud provider like AWS, without modification.

- **Isolation:** The application runs in its own isolated environment with a defined set of libraries and runtime parameters, preventing conflicts with other software on the host system.
- **Scalability:** Container orchestration platforms such as Kubernetes or Amazon ECS can rapidly start and stop containers based on demand, enabling efficient horizontal scaling during high-traffic periods.
- **Faster Deployment and Rollback:** Pre-built container *images* can be deployed quickly, and previous versions can be rolled back by redeploying an earlier image tag, reducing downtime.
- **Simplified CI/CD Integration:** Containers integrate seamlessly with continuous integration and delivery pipelines, allowing automated builds, tests, and deployments triggered by code changes.
- **Reproducibility:** For experimental settings, distributing a container image ensures that collaborators can reproduce results exactly, regardless of local environment differences.

The MIBot container image is defined by a `Dockerfile` specifying the base Python runtime, 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 "simplifies the deployment, management, and scaling of applications using containers" (Amazon Web Services, Inc.). We now discuss each component of ECS.

3.4.1 Components of ECS

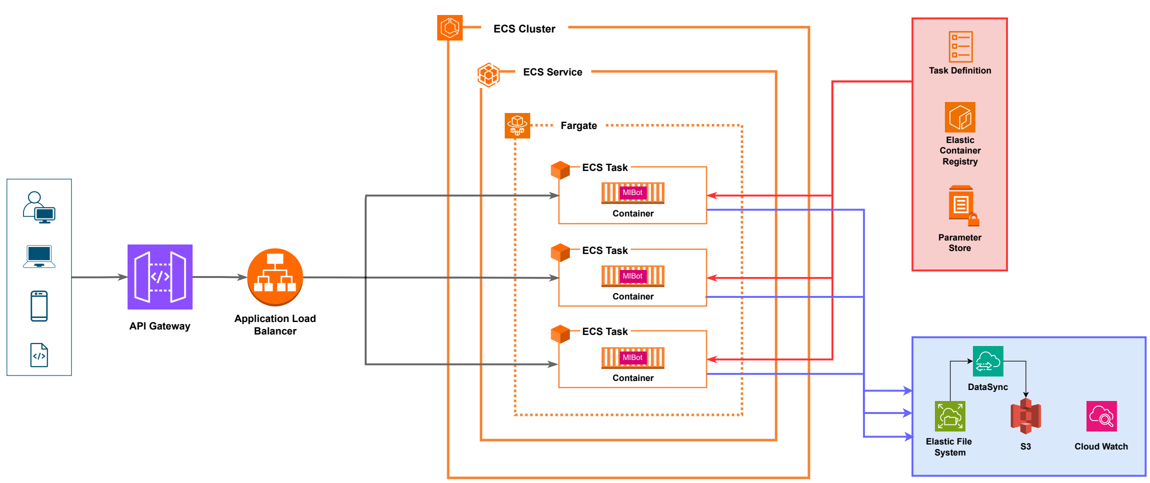


Figure 3.4: Containerized MIBot application deployed to ECS.

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. To understand what we mean by this, let's first see what a compute resource is in the context of AWS. An AWS **compute resource** is any AWS-managed infrastructure component that provides processing power for running applications or workloads. Examples include Elastic Cloud Compute (EC2), AWS Lambda, AWS Fargate, etc. Some of the compute resources are **user-managed**, like EC2, where users rent virtual machines to run and manage their applications. In contrast, AWS Fargate is **fully managed** and is used to run containerized applications. A cluster is merely a logical grouping of such compute resources.

In our ECS cluster, however, we only used **AWS Fargate** as the computing resource. Apart from the ability to deploy the containerized MIBot application without having to manage our own servers, AWS Fargate also provided us with *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 needs the capacity back, 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 for some reason, 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. Most importantly, it also defines deployment (rolling update, blue/green deployment) and rollback strategies. It can leverage the AWS circuit breaker to roll back failed deployments automatically.

3. ECS Task Definition: The final piece in MIBot's deployment 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 `OK 200` 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 contains required CPU and memory (1024 mcpu 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 in 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, meaning we have three instances of load balancers in three different Availability Zones. The load balancers are the modern *application* load balancers (ALB), 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 friendly domain. 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 allows client browsers (running the frontend code) to send requests to MIBot without getting a "Cross-Origin Request Blocked" error or "Not Secure" warning.

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 some reason. For eventual consistency, we periodically copy files from EFS to S3 with the help of AWS DataSync, which runs a daily CRON job.

3.4.3 Deployment Pipeline

We adopted automated development practices for 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.

3.5 Feasibility Study with Human Smokers

The deployed MIBot system was used in a feasibility study with human smokers. The goal of the study was to determine the impact of a single-session interaction with MIBot and to see if the interaction could be delivered safely to ambivalent smokers in a real-world setting.

3.5.1 Ethics Approval and Consent

The protocol was approved by the University of Toronto Research Ethics Board under protocol number 49997 (approved August 3, 2024) 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. 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.

3.5.2 Participant Recruitment

Participants were recruited via *Prolific*[2], 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., 2023b; **?**). Participants received £5.50 GBP for the main session and £1.00 GBP for the follow-up survey, exceeding Prolific's recommended hourly rates.

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 one year or more.

In addition, recruitment was set up to aim for a near-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.

[2] https://www.prolific.com

In-Study Screening

Baseline demographics of enrolled participants are summarized in Table 3.3. All were English-speaking adults who self-identified as current daily smokers and passed prescreening and in-study eligibility checks on Prolific. The sample was approximately sex-balanced with broad age coverage. Residence was primarily the United Kingdom and United States, with additional participants from Canada and South Africa. Unless noted, values are n (%). Age is reported as median, mean (SD), and range.

Participant Characteristics

Upon arrival at the study website, participants completed a smoking status confirmation question identical to Prolific's pre-screen question. Any inconsistency between the two responses resulted in immediate exclusion. Additional screening was conducted using the *Readiness Ruler*, as described in Section 3.5.4, to ensure participants met ambivalence criteria. We defined (identical to (Brown et al., 2023b)) ambivalence as being present when the pre-conversation *confidence to quit* score ≤ 5 on a 0–10 scale. Also following (Brown et al., 2023b) participants whose *confidence to quit* > 5 were included if there was a difference of at least five points between their *confidence* and *importance* (importance rated lower than confidence). These participants, although having high confidence, are deemed "discordant," meaning they are ambivalent despite nominally high confidence.

Participants who met all criteria and provided informed consent proceeded to the survey phase. Those failing to meet eligibility criteria were thanked and returned to the Prolific platform without completing the study.

3.5.3 Study Procedure

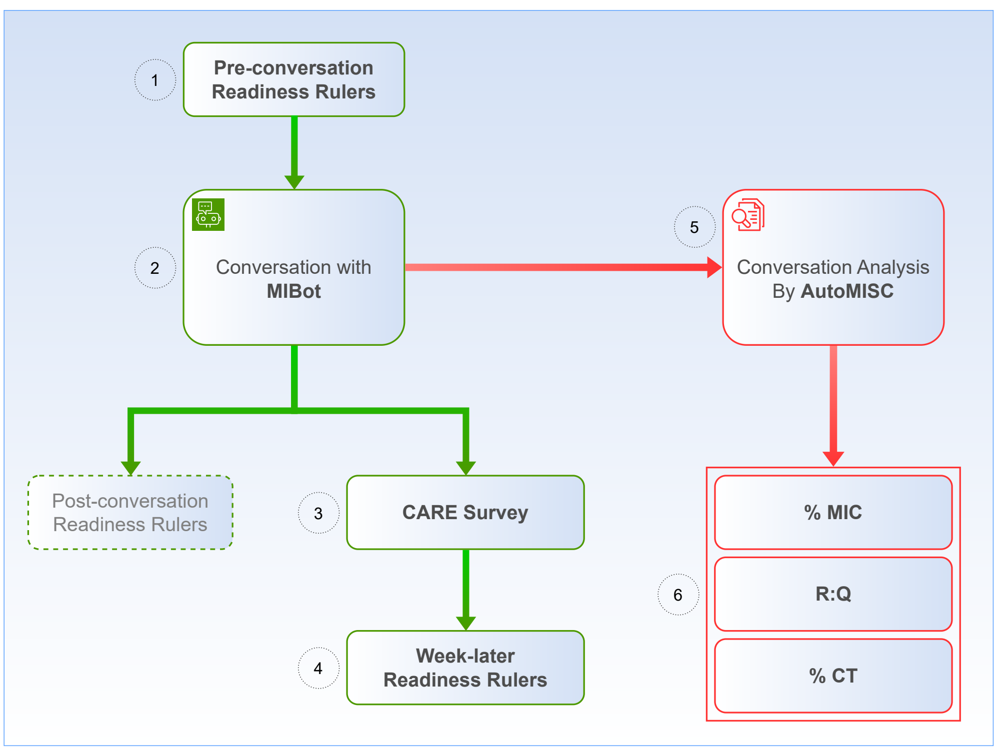


Figure 3.5: Overview of the feasibility study protocol.

The full study procedure comprised four major phases: (1) pre-conversation surveys; (2) a single

Table 3.3: Baseline characteristics of enrolled participants

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)

text-based conversation with MIBot; (3) immediate post-conversation surveys; and (4) a one-week follow-up survey. Figure 3.5 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 via two questions:
   (a) number of cigarettes smoked per day, and
   (b) time to first cigarette after waking.
2. **Quit Attempts in the Past Week**: 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 rating importance, confidence, and readiness to quit on 0–10 scales (Section 3.5.4).

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 3.5.4), 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.

3.5.4 Survey Instruments

We employed the following metrics to measure outcomes from the interaction:

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"). Most importantly, the week-later change in *confidence* from the pre-conversation value was used as the primary metric for the chatbot's effectiveness, as this is the most predictive of downstream quitting success (Gwaltney et al., 2009; Abar et al., 2013).

2. CARE Measure

The Consultation and Relational Empathy (CARE) measure (**??**) 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 0-5 scale, with a total score range of 0-50.

Apart from readiness rulers and CARE, we also recorded other survey instruments to get a holistic picture of the chatbot's effectiveness.

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 have the potential to inform future prompt refinements and provide contextual data for interpreting quantitative outcomes.

4. Follow-Up Quit Attempt Survey

At one week, participants 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.

3.5.5 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., 2010). 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 MI-expert clinicians (Mahmood et al., 2025).

Readiness rulers (especially the week-later change in *confidence*), CARE, and AutoMISC summary metrics together provide a holistic view of MIBot's effectiveness, perceived empathy, safety and adherence to MI principles. In the next chapter, we report the results from our human feasibility study on these metrics.

Chapter 4

Evaluation of MIBot

This chapter presents a comprehensive evaluation of MIBot's effectiveness through a feasibility study with 106 smokers recruited via the Prolific platform. 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 4.1). We then analyze how chatbot performed on perceived empathy, measured through the CARE scale and compare it to human healthcare professionals (Section 4.2). Next, we examine MIBot's adherence to MI principles through AutoMISC analysis, and examine if MIBot could maintain fidelity to therapeutic standards while successfully eliciting change talk from clients (Section 4.4).

Following these core metrics, we investigate behavioural outcomes including quit attempts and self-reported changes (Section 4.5), explore conversation dynamics both quantitatively and qualitatively (Section 4.6), and examine illustrative case studies and sample outliers (Section 4.7). Finally, we analyze participant feedback (Section 4.8) and discuss broader implications of our findings (Section 4.9).

4.1 Primary Outcome: Readiness to Quit

4.1.1 Overall Changes in Readiness Rulers

Table 4.1 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 a $\sim$59% relative improvement and constitutes our primary outcome measure, aligning with MI theory that confidence (self-efficacy) is a key predictor of behaviour change (Gwaltney et al., 2009; Abar et al., 2013). Importance also increased, albeit more modestly ($\Delta = 0.5$, $p < 0.005$), while readiness exhibited a small, non-significant change ($\Delta = 0.23$, $p = 0.22$).

Figure 4.1 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

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 4.1: Means (SD) of readiness rulers (0–10 scale) before the conversation, immediately after, one week later, and the week-later change (Δ). SD = standard deviation. Wilcoxon signed-rank test: *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$, n.s. = not significant ($p \geq 0.05$).

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 quitting confidence.

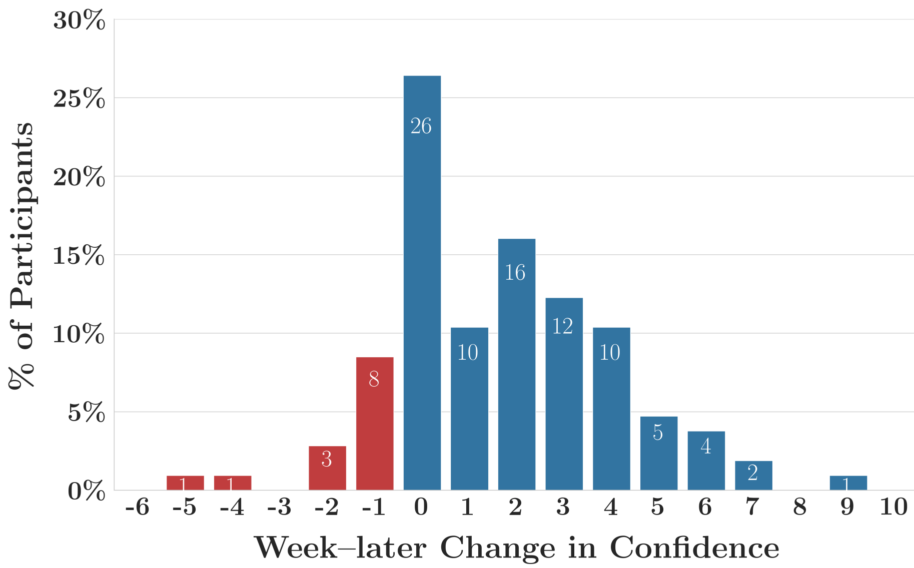


Figure 4.1: Distribution of one-week changes in confidence scores. The majority of participants (60.4%) showed improvement, with a long right tail indicating some participants experienced dramatic gains.

4.1.2 Stratified Analysis by Baseline Characteristics

To understand which participants benefited most from MIBot, we stratified outcomes by baseline characteristics. Table 4.2 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 about 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.

Confidence changes varied across different 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, $\geq$20), whether they had made a quit attempt in the week before the study, and the number of prior attempts (0, 1–2, $\geq$3). The mean change in confidence for each subgroup is summarized in Table 4.3.

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

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 4.2: Longitudinal changes in self-reported confidence scores (0–10 scale) stratified by baseline confidence level. Values assessed at baseline (pre-conversation), immediately post-conversation, and at 1-week follow-up. Participants were grouped by their initial confidence scores. Change scores represent the difference between baseline and 1-week follow-up assessments. SD = standard deviation. Wilcoxon signed-rank test: *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$, n.s. = not significant ($p \geq 0.05$).

Participant characteristics	Sample size	Baseline confidence mean (SD)	Post-conversation mean (SD)	1-week follow-up mean (SD)	Change from baseline mean (SD)
Heaviness of Smoking Index					
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)
Daily cigarette consumption					
<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)***
Pre-conversation quit attempt					
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)***
Number of prior attempts					
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 4.3: Longitudinal changes in quit confidence scores stratified by baseline smoking characteristics and quit history. Values represent self-reported confidence to quit smoking (0–10 scale) 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 assessments. SD = standard deviation. Wilcoxon signed-rank test: *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$, n.s. = not significant ($p \geq 0.05$).

ence 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 to +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 could suggest that participants who were already contemplating change benefited the most.

4.1.3 Demographic Patterns

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

Demographic characteristics	**Sample size**	**Baseline confidence mean (SD)**	**Post-conversation mean (SD)**	**1-week follow-up mean (SD)**	**Change from baseline mean (SD)**
Sex					
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)***
Age					
$<$ 30 years	26	3.7 (2.1)	5.5 (2.5)	5.7 (2.7)	+1.9 (3.1)*
$\geq$ 30 years	80	2.5 (1.8)	4.3 (2.5)	4.1 (2.6)	+1.6 (2.1)***
Ethnicity					
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)***
Employment status					
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 4.4: Longitudinal changes in quit confidence scores stratified by demographic characteristics. Values represent self-reported confidence to quit smoking (0–10 scale) 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 assessments. SD = standard deviation. Wilcoxon signed-rank test comparing baseline to 1-week follow-up: *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$, n.s. = not significant ($p \geq 0.05$).

(mean +1.6, SD 2.1). While baseline confidence differed by sex (2.5 for females vs 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 vs 2.7) and showed larger gains (2.5 vs 1.4). These exploratory findings must be interpreted cautiously, as the study was not powered for subgroup analyses and baseline demographic differences were not statistically controlled.

4.2 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). MIBot achieved a mean total score of 42 out of 50. To contextualize this performance, Table 4.5 compares MIBot's scores with human healthcare professionals from Bikker et al. (2015).

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

Table 4.5: Average CARE scores and percentage of perfect scores for MIBot and typical human healthcare professionals (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%).

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											
Sex											
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
Age											
<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
Ethnicity											
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
Employment											
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											
Confidence											
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
HSI											
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
Cigarettes/Day											
<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
Quit Attempts											
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 4.6: CARE Scale Scores by Demographic and Behavioural Characteristics

4.2.1 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 4.7: Mean scores for each CARE question (1–5 scale)

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

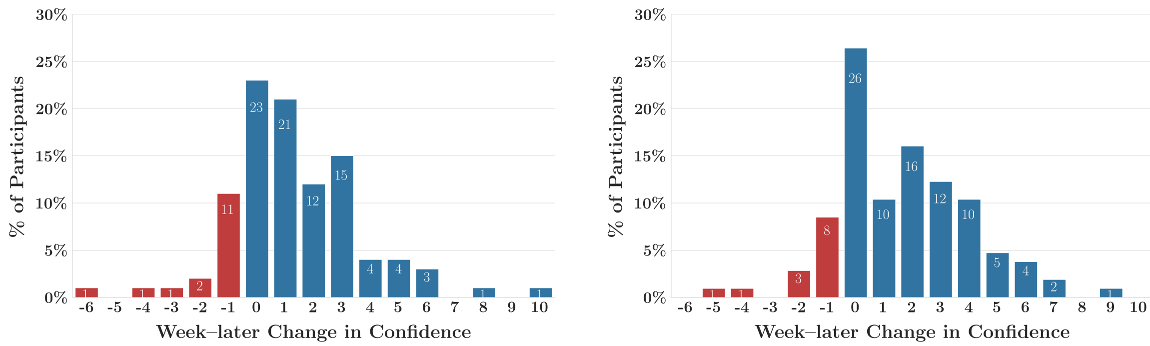
4.3 Comparing Fully Generative MIBot v6.3 with Partially Scripted MIBot v5.2

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., 2023a). 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.

4.3.1 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 4.2, the distribution of confidence changes reveals interesting patterns: MIBot 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, $p = 0.01$;

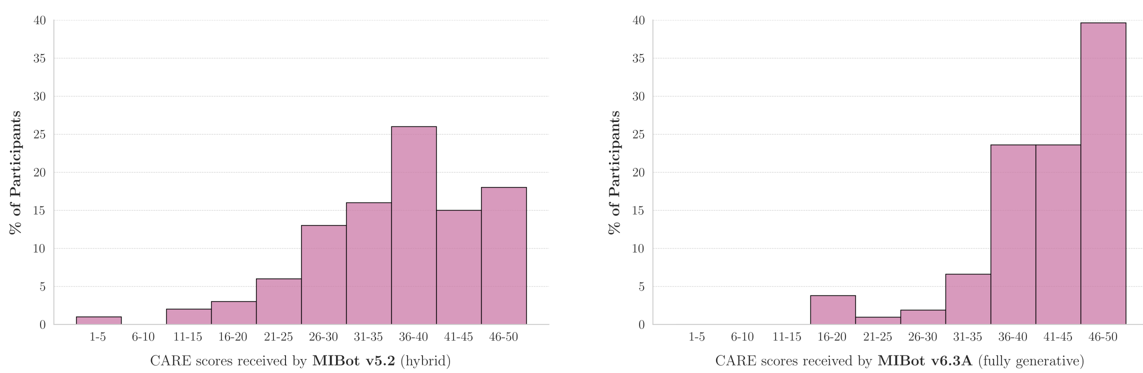


(a) MIBot v5.2 **(b)** MIBot v6.3A

Figure 4.2: Distribution of week-later confidence changes from baseline for (a) MIBot v5.2 and (b) MIBot v6.3A. Red bars indicate decreased confidence while blue bars indicate increased or unchanged confidence. MIBot v6.3A shows a higher proportion of participants with no change (26% vs. 23%) and fewer reporting decreased confidence (13% vs. 16%).

v6.3A: 0.3, SD 2.4, $p = 0.22$).

4.3.2 Perceived Empathy Comparisons



(a) MIBot v5.2 (hybrid) **(b)** MIBot v6.3A (fully generative)

Figure 4.3: Distribution of CARE empathy scores for (a) MIBot v5.2 and (b) MIBot v6.3A. The fully generative v6.3A shows a pronounced rightward shift with ∼40% of participants scoring in the highest range (46–50) compared to only ∼18% for the hybrid v5.2. Lower scores (below 30) were nearly eliminated in v6.3A.

MIBot v5.2 achieved a mean CARE score of 36 ($SD = 9.1$), with only 3% of participants awarding 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 significantly improved perceived empathy. Figure 4.3 illustrates the distributional shift between versions. The hybrid v5.2 shows a relatively normal distribution centred in the mid-30s range, 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).

To understand which aspects of empathetic interaction improved most, Figure 4.4 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 enhanced ability to maintain an encouraging tone. Interestingly, the dimension of

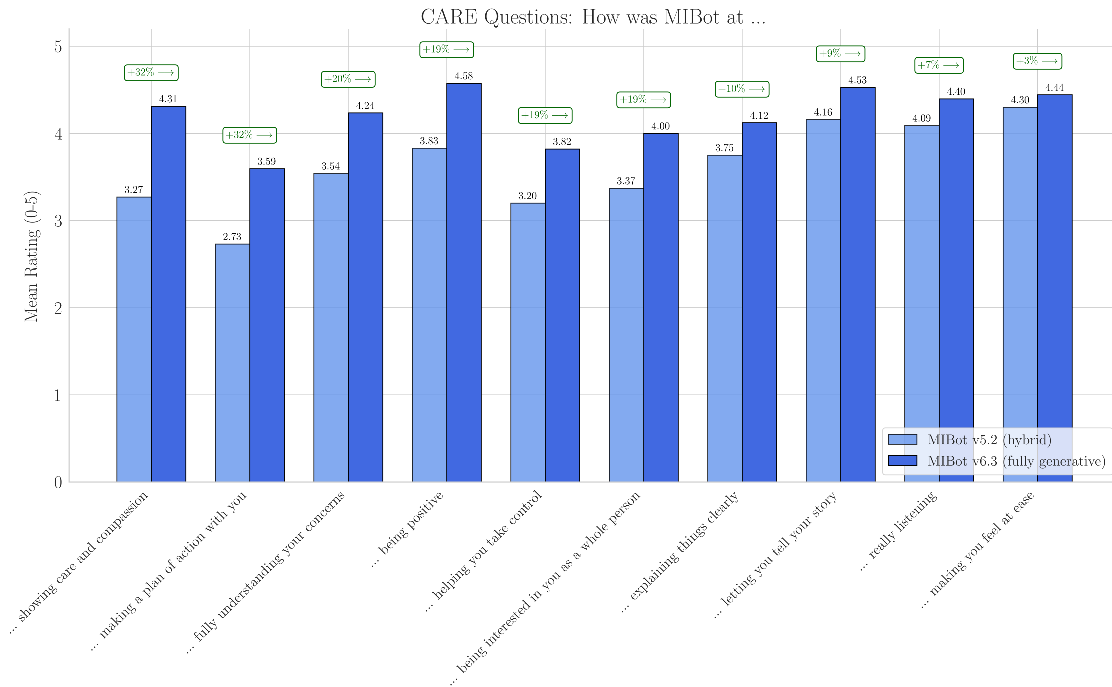


Figure 4.4: Question-wise mean CARE scores comparing MIBot v5.2 (hybrid) and v6.3A (fully generative). The fully generative version shows consistent improvements across all dimensions of empathy. Bars displayed in ascending order of relative improvement.

"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%). It is worth noting that MIBot v6.3 was prompted with detailed guidelines on how to make a plan of action with the client. Despite this, planning seems to be one of its weak areas.

4.3.3 Implications of the Comparison

The fully generative MIBot v6.3A chatbot scores 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 the chatbots scored the same 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., 2023a), though the enhanced empathy of full generation may improve engagement and retention in real-world deployment.

4.4 AutoMISC Analysis

To evaluate the chatbot's adherence to MI principles, we analyzed counsellor and client utterances from the feasibility study transcripts (Section 3.5) 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:

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

- **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., 2016a).
- **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, 2009b).

4.4.1 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 available[2] 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.

4.4.2 Counsellor Behaviour Metrics

Table 4.8 summarizes the counsellor-specific AutoMISC summary metrics across the 106 transcripts. The mean percentage of MI-consistent responses (%MIC) was 98 (SD 3.6), higher than the high-quality counsellor sessions in the HLQC dataset (92, SD 9.8). MIBot's reflection-to-question ratio (R:Q) was 1.3 (SD 0.3), comfortably within the 1–2 range recommended for human practitioners (Moyers et al., 2016b).

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 4.8: AutoMISC counsellor-specific summary metrics for MIBot compared to high-quality human counselling sessions. Values shown as mean (standard deviation).

Figure 4.5 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).

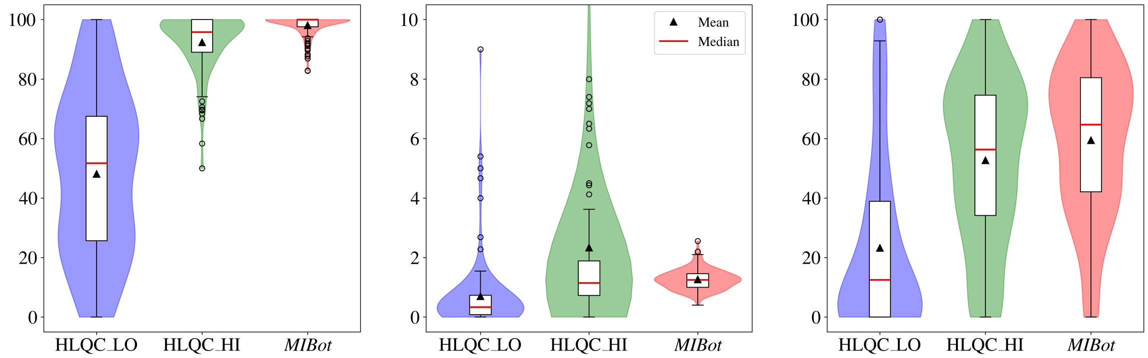
4.4.3 Client Change Talk Analysis

Metric	**MIBot**	**HLQC_HI (high-quality)**
%Change Talk (%CT)	59 (29.6)	53 (28.54)

Table 4.9: AutoMISC client-specific summary metric for MIBot compared to high-quality human counselling sessions. Values shown as mean (standard deviation).

Figure 4.5 c illustrates the distribution of %CT across conversations. Most sessions by MIBot a %CT of $>$ 60% and the distribution closely matches that of **HLQC_HI** dataset.Overall, the mean %CT was 59% for MIBot as compared to 53% for **HLQC_HI** dataset.

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



(a) %MIC **(b)** R:Q **(c)** %CT

Figure 4.5: Comparison of MISC summary score distributions across datasets. (a) Percentage MI-Consistent Responses (%MIC), (b) Reflection to Question Ratio (R:Q), (c) Percentage Client Change Talk (%CT).

4.5 Behavioural Outcomes and Follow-up

4.5.1 Quit Attempts at One Week

Table 4.10 summarises quit attempt behaviour before and after the MIBot conversation. Prior to the conversation, 32.1% (34/106) of participants had made a quit attempt in the previous week. At the one-week follow-up, 34.9% (37/106) reported making a quit attempt since the conversation. Crucially, among the 72 participants who had not attempted to quit before the conversation, 15 (20.8%) made their first attempt following the MIBot session.

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 4.10: Quit attempt behaviour before and after MIBot conversation.

Participants who made quit attempts post-conversation showed larger confidence gains (mean +2.43) compared to those who did not attempt (+1.35, $p < 0.05$). This bidirectional relationship—where confidence predicts attempts and attempts reinforce confidence—aligns with social cognitive theory (**?**).

4.5.2 Self-Reported Behavioural Changes

Beyond formal quit attempts, participants reported various harm reduction behaviours. Analysis of open-ended responses revealed five main categories of behavioural change:

1. **Reduced consumption** (42% of participants): "Cut down from 20 to 12 cigarettes a day"
2. **Delayed first cigarette** (28%): "Now waiting until after breakfast instead of immediately upon waking"
3. **Substitution strategies** (31%): "Using nicotine gum when cravings hit at work"
4. **Environmental changes** (19%): "Removed ashtrays from the house and car"
5. **Social support seeking** (23%): "Told my partner about wanting to quit and asked for support"

These incremental changes, while not constituting complete cessation, represent important steps in the behaviour change process and align with harm reduction approaches increasingly recognised in tobacco control (**?**).

4.6 Conversation Dynamics

4.6.1 Quantitative Conversation Metrics

Conversations ranged from 36 to 163 utterances (median 78, mean 80.7, SD 29.4), with counsellor utterances comprising 55% of the total. The median conversation lasted approximately 19 minutes based on participant self-report. Table 4.11 summarizes key quantitative metrics.

Metric	**Mean (SD)**	**Range**
Total utterances	80.7 (29.4)	36–163
Counsellor utterances	44.4 (16.2)	20–90
Client utterances	36.3 (13.2)	16–73
Words per counsellor utterance	42.3 (18.7)	15–95
Words per client utterance	28.6 (15.4)	8–72
Session duration (minutes)	19 (9)	8–45

Table 4.11: Quantitative metrics of conversation dynamics.

Longer conversations correlated with better outcomes ($r = 0.20$ for confidence change), but the relationship was non-linear. Conversations under 50 utterances rarely produced substantial gains, while those exceeding 120 utterances showed diminishing returns, suggesting an optimal engagement window of 60–100 exchanges.

4.6.2 Qualitative Thematic Analysis

Thematic analysis of conversation transcripts revealed four recurring patterns that characterised successful therapeutic engagement:

Stress and Coping Narratives

The most common theme (78% of conversations) involved smoking as emotional regulation. Participants frequently described smoking as their primary stress management tool:

> *"It's like my safety blanket. When work gets overwhelming, that cigarette break is the only thing that keeps me sane. I know it's killing me, but it's also what's keeping me functional right now."*

MIBot effectively used reflective listening to validate these feelings while gently exploring alternatives, demonstrating the double-sided reflection technique central to MI.

Social and Ritualistic Aspects

Many participants (62%) described smoking as deeply embedded in social routines and relationships:

> *"All my friends smoke. Our whole social life revolves around smoke breaks at work, cigarettes with coffee, smoking outside the pub. If I quit, I lose all that connection."*

The chatbot showed sophistication in acknowledging these social dimensions while helping participants envision maintaining relationships without cigarettes.

Ambivalence Themes

Classic motivational ambivalence appeared in 89% of conversations, with participants simultaneously expressing desire to quit and attachment to smoking:

> *"Part of me desperately wants to quit for my kids, but another part can't imagine life without cigarettes. They've been with me through everything—divorce, job loss, you name it."*

MIBot's response to ambivalence—normalising it rather than pushing through it—aligned with MI best practices and correlated with positive outcomes.

Readiness Variations

Participants displayed varying stages of readiness, from pre-contemplation to action planning. The chatbot demonstrated flexibility in matching its approach to the participant's stage:

- **Pre-contemplation**: Focus on exploring values and long-term goals
- **Contemplation**: Weighing pros and cons, exploring ambivalence
- **Preparation**: Concrete planning, identifying resources and strategies
- **Action**: Problem-solving barriers, reinforcing commitment

4.7 Case Studies and Edge Cases

4.7.1 Success Stories

The most dramatic success involved a 34-year-old participant who entered with confidence of 1/10 but importance of 9/10—a classic "willing but unable" profile. Through 142 utterances, the conversation systematically addressed self-efficacy barriers:

1. Explored past quit attempts to identify what worked
2. Reframed "failures" as learning experiences
3. Developed a detailed, personalised quit plan
4. Identified specific coping strategies for anticipated triggers

This participant's confidence increased to 10/10 at one week, and they reported complete cessation for five days at follow-up.

Another notable success involved a 58-year-old with 40 years of smoking history who had "given up on giving up." The conversation's focus on harm reduction rather than immediate cessation allowed gradual engagement. By week's end, daily consumption dropped from 30 to 10 cigarettes, with confidence rising from 2 to 7.

4.7.2 Non-Responders and Negative Cases

Not all participants benefited. Analysis of the 21 participants whose confidence decreased revealed three patterns:

Mandated Participation

Some participants appeared to engage solely for compensation, providing minimal responses and showing no genuine interest in change. These conversations averaged just 42 utterances, with client responses typically under 10 words.

Enjoyment-Focused Smokers

A subset strongly identified as "happy smokers" who enjoyed smoking without ambivalence. MIBot's attempts to explore motivation sometimes paradoxically reinforced their commitment to smoking:

> *"Talking about it just reminded me how much I actually enjoy smoking. I don't want to quit, and this conversation made that clearer."*

Technical Therapeutic Mismatches

In rare cases, MIBot misread the participant's needs, such as pushing for behaviour change when emotional support was needed, or vice versa. These mismatches highlight the challenges of fully automated therapeutic engagement without human oversight.

4.8 Post-Conversation Feedback

4.8.1 Qualitative Feedback Analysis

Three open-ended questions captured immediate post-conversation reactions. Thematic analysis revealed distinct patterns in positive and negative responses.

Positive Themes (92% enjoyed the experience):

- Non-judgmental approach: "Finally someone (something?) that didn't lecture me"
- Structured exploration: "Helped me organise 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 personalisation: "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"

4.8.2 Categorization of User Experience

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:

"Entertained Sceptics"—those who enjoyed but didn't find it helpful—often wanted more directive advice or concrete tools rather than exploratory 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 4.12: User experience segments based on enjoyment and perceived helpfulness.

4.9 Synthesis and Clinical Implications

4.9.1 Key Predictors of Success

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 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.

4.9.2 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) (**?**)
- **MI fidelity**: 95% MI-consistent responses surpass typical human counsellor benchmarks (80–90%) (**?**)
- **Engagement**: 92% enjoyment rate exceeds most digital health interventions (60–70%) (**?**)
- **Quit attempts**: 14.2% new quit attempts align with brief intervention outcomes (10–20%) (**?**)

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.

4.9.3 Limitations of Current Evaluation

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 (£12) 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

4.10 Chapter Summary

This evaluation demonstrates that MIBot, a fully generative MI chatbot, can effectively support smokers in building confidence toward cessation. The primary finding—a mean confidence increase of 1.66 points sustained at one week—represents a clinically meaningful effect size that exceeds most digital interventions. With 95% MI-consistent responses and successful elicitation of change talk in 82% of conversations, MIBot achieves technical proficiency in MI delivery while maintaining high user engagement (92% enjoyment rate).

The analysis reveals that MIBot works best for low-confidence smokers who have recently attempted to quit, with younger participants and those willing to engage in extended conversation showing the greatest benefits. While perceived empathy approaches human benchmarks, gaps remain in creating exceptional therapeutic relationships and translating engagement into universally perceived helpfulness.

These findings establish proof-of-concept for AI-delivered MI in smoking cessation while highlighting areas for improvement. The strong technical performance combined with meaningful behavioural outcomes suggests that generative AI can meaningfully contribute to public health interventions. However, the variability in individual responses and limitations in emotional intelligence indicate that MIBot may be best positioned as a scalable first-line intervention or adjunct to human services rather than a complete replacement for human counselling.

The next chapter examines how these quantitative findings translate into real-world implementation, exploring deployment strategies, ethical considerations, and the potential for MIBot to address the massive unmet need for smoking cessation support globally.

Chapter 5

Development of Synthetic Smokers

Chapter 6

Results

Duis aliquet dui in est. Donec eget est. Nunc lectus odio, varius at, fermentum in, accumsan non, enim. Aliquam erat volutpat. Proin sit amet nulla ut eros consectetuer cursus. Phasellus dapibus aliquam justo. Nunc laoreet. Donec consequat placerat magna. Duis pretium tincidunt justo. Sed sollicitudin vestibulum quam. Nam quis ligula. Vivamus at metus. Etiam imperdiet imperdiet pede. Aenean turpis. Fusce augue velit, scelerisque sollicitudin, dictum vitae, tempor et, pede. Donec wisi sapien, feugiat in, fermentum ut, sollicitudin adipiscing, metus.

Donec vel nibh ut felis consectetuer laoreet. Donec pede. Sed id quam id wisi laoreet suscipit. Nulla lectus dolor, aliquam ac, fringilla eget, mollis ut, orci. In pellentesque justo in ligula. Maecenas turpis. Donec eleifend leo at felis tincidunt consequat. Aenean turpis metus, malesuada sed, condimentum sit amet, auctor a, wisi. Pellentesque sapien elit, bibendum ac, posuere et, congue eu, felis. Vestibulum mattis libero quis metus scelerisque ultrices. Sed purus.

Donec molestie, magna ut luctus ultrices, tellus arcu nonummy velit, sit amet pulvinar elit justo et mauris. In pede. Maecenas euismod elit eu erat. Aliquam augue wisi, facilisis congue, suscipit in, adipiscing et, ante. In justo. Cras lobortis neque ac ipsum. Nunc fermentum massa at ante. Donec orci tortor, egestas sit amet, ultrices eget, venenatis eget, mi. Maecenas vehicula leo semper est. Mauris vel metus. Aliquam erat volutpat. In rhoncus sapien ac tellus. Pellentesque ligula.

Cras dapibus, augue quis scelerisque ultricies, felis dolor placerat sem, id porta velit odio eu elit. Aenean interdum nibh sed wisi. Praesent sollicitudin vulputate dui. Praesent iaculis viverra augue. Quisque in libero. Aenean gravida lorem vitae sem ullamcorper cursus. Nunc adipiscing rutrum ante. Nunc ipsum massa, faucibus sit amet, viverra vel, elementum semper, orci. Cras eros sem, vulputate et, tincidunt id, ultrices eget, magna. Nulla varius ornare odio. Donec accumsan mauris sit amet augue. Sed ligula lacus, laoreet non, aliquam sit amet, iaculis tempor, lorem. Suspendisse eros. Nam porta, leo sed congue tempor, felis est ultrices eros, id mattis velit felis non metus. Curabitur vitae elit non mauris varius pretium. Aenean lacus sem, tincidunt ut, consequat quis, porta vitae, turpis. Nullam laoreet fermentum urna. Proin iaculis lectus.

Sed mattis, erat sit amet gravida malesuada, elit augue egestas diam, tempus scelerisque nunc nisl vitae libero. Sed consequat feugiat massa. Nunc porta, eros in eleifend varius, erat leo rutrum dui, non convallis lectus orci ut nibh. Sed lorem massa, nonummy quis, egestas id, condimentum at, nisl. Maecenas at nibh. Aliquam et augue at nunc pellentesque ullamcorper. Duis nisl nibh, laoreet suscipit, convallis ut, rutrum id, enim. Phasellus odio. Nulla nulla elit, molestie non, scelerisque at, vestibulum eu, nulla. Ut odio nisl, facilisis id, mollis et, scelerisque nec, enim. Aenean sem leo, pellentesque sit amet, scelerisque sit amet, vehicula pellentesque, sapien.

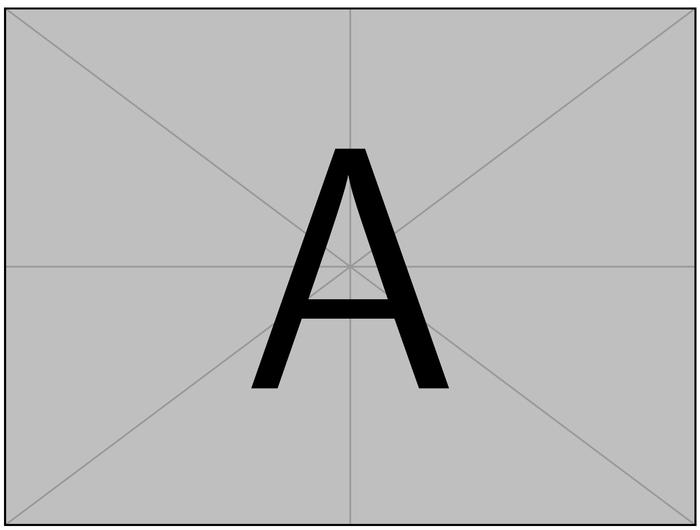


Figure 6.1: Jumping over the lazy dog

Appendix A

Code

Index of Key Terms

Bibliography

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

Gati Aher, Rosa I Arriaga, and Adam Tauman Kalai. 2023. Using large language models to simulate multiple humans and replicate human subject studies. *arXiv preprint arXiv:2305.17429*.

Soliman Ali, Jiading Zhu, Alex Guo, Xiao Nan Ye, Qilin Gu, Jodi Wolff, Carolynne Cooper, Osnat C. Melamed, Peter Selby, and Jonathan Rose. 2025. Automated coding of counsellor and client behaviours in motivational interviewing transcripts: Validation and application. Manuscript submitted for publication.

Fahad Almusharraf. 2018. Motivating Smokers to Quit Through a Computer-Based Conversational System. Master of applied science thesis, University of Toronto, Toronto, Canada. Accessed: 2025-05-21.

Amazon Web Services, Inc. Getting Started with Amazon ECS. `https://aws.amazon.com/ecs/getting-started/`. Accessed 2025-08-10.

L. M. Amugongo, P. Mascheroni, S. Brooks, S. Doering, and J. Seidel. 2025. Retrieval augmented generation for large language models in healthcare: A systematic review. *PLOS Digital Health*, 4(6):e0000877.

Timothy R. Apodaca and Richard Longabaugh. 2009a. Mechanisms of change in motivational interviewing: A review and preliminary evaluation of the evidence. *Addiction*, 104.

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

Luke P Argyle, Rebecca Burns, Benjamin P Chamberlain, et al. 2023. Out of one, many: Using language models to simulate human samples. *arXiv preprint arXiv:2302.05756*.

Stephen Babb, Ann Malarcher, Gillian Schauer, Katherine Asman, and Ahmed Jamal. 2017. Quitting Smoking Among Adults - United States, 2000-2015. *MMWR Morbidity and Mortality Weekly Report*, 65(52):1457–1464.

Aaron T. Beck, A. John Rush, Brian F. Shaw, and Gary Emery. 1979. *Cognitive Therapy of Depression*, 2 edition. Guilford Press.

Leonardo Berti, Flavio Giorgi, and Gjergji Kasneci. 2025. Emergent Abilities in Large Language Models: A Survey. *Preprint*, arXiv:2503.05788.

L. Biener and D. B. Abrams. 1991. The Contemplation Ladder: validation of a measure of readiness to consider smoking cessation. *Health Psychology*, 10(5):360–365.

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

Gallus Bischof, Anja Bischof, and Hans-Juergen Rumpf. 2021. Motivational Interviewing: An Evidence-Based Approach for Use in Medical Practice. *Deutsches Aerzteblatt Online*, 118.

Edwin D. Boudreaux, Janice A. Sullivan, and Erin Abar. 2012. Motivation Rulers for Smoking Cessation: A Prospective Observational Examination of Construct and Predictive Validity. *Addictive Behaviors*, 37(6):548–552.

Andrew Brown, Ash Tanuj Kumar, and et al. 2023a. A motivational interviewing chatbot with generative reflections for increasing readiness to quit smoking: Iterative development study. *JMIR Mental Health*, 10:e49132.

Andrew Brown et al. 2023b. A motivational interviewing chatbot with generative reflections for increasing readiness to quit smoking. *JMIR Mental Health*.

Tom Brown, Benjamin Mann, and et al. 2020. Language Models are Few-Shot Learners. *Advances in Neural Information Processing Systems*, 33:1877–1901.

Canadian Partnership Against Cancer. 2020. Lung cancer and equity report. `https://s22457.pcdn.co/wp-content/uploads/2020/11/Lung-cancer-and-equity-report-EN.pdf`.

CanCOVID Issue Note. 2023. Implementing asynchronous virtual mental health in rural and remote communities. *CanCOVID*. Notes systemic underfunding and service barriers in Indigenous mental health care (CanCOVID Issue Note, 2023).

C. L. Chang, C. Sinha, M. Roy, and J. C. M. Wong. 2024. AI-Led Mental Health Support (Wysa) for Health Care Workers During COVID-19: Service Evaluation. *JMIR Formative Research*, 8:e51858.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, et al. 2022. PaLM: Scaling Language Modeling with Pathways. *arXiv preprint arXiv:2204.02311*.

Kenneth Mark Colby, Sylvia Weber, and Franklin D. Hilf. 1971. Artificial Paranoia: A Computer Simulation of Paranoid Schizophrenic Processes. *Artificial Intelligence*, 2(1):1–25.

Erica Corda, Silvia M Massa, and Daniele Riboni. 2024. Context-aware behavioral tips to improve sleep quality via machine learning and large language models. *Future Internet*, 16(2):46.

Rob Crosato and Brant Leipert. 2011. Rural poverty in Canada. *Journal of Rural Health*. Describes health care access barriers in rural Canada (Crosato and Leipert, 2011).

Huifang Du, Shuqin Li, Minghao Wu, Xuejing Feng, Yuan-Fang Li, and 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*, pages 8030–8046, Miami, Florida, USA. Association for Computational Linguistics.

Cristina Fernandez, Izaskun Fernandez, and 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*, pages 205–214, Bilbao, Spain. Association for Computational Linguistics.

Kathleen K. Fitzpatrick, Alison Darcy, and 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. *JMIR Mental Health*, 4(2):e19.

Abigail Gilson, Jiajun Li, Sheng Liu, and Diyi Yang. 2023. How Well Do Large Language Models Align with Human Empathy? Evaluating Empathetic Responses in Dialogue. *arXiv preprint arXiv:2305.15267*.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and 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*, pages 8342–8360, Online. Association for Computational Linguistics.

Güven Güzeldere and Stefano Franchi. 1995. Dialogues with colorful "personalities" of early AI. *Stanford Hum. Rev.*, 4(2):161–169.

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

Syed Ali Haider, Srinivasagam Prabha, Cesar A. Gomez-Cabello, Sahar Borna, Ariana Genovese, Maissa Trabilsy, 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. *Sensors*, 25(14):4305.

D. E. Hartman. 1986. Artificial intelligence or artificial psychologist? Conceptual issues in clinical microcomputer use. *Professional Psychology: Research and Practice*, 17(6):528–534.

Steven C. Hayes, Kirk D. Strosahl, and Kelly G. Wilson. 1999. *Acceptance and Commitment Therapy: An Experiential Approach to Behavior Change*. Guilford Press.

Todd F. Heatherton, Lynn T. Kozlowski, R. C. Frecker, W. Rickert, and 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. *British Journal of Addiction*, 84(7):791–799.

Carolyn J. Heckman, Ryan Egleston, and Kevin Hofmann. 2010. Efficacy of Motivational Interviewing for Smoking Cessation: A Systematic Review and Meta-Analysis. *Tobacco Control*, 19(5):410–416.

Michael V. Heinz, Daniel M. Mackin, Brianna M. Trudeau, Sukanya Bhattacharya, Yinzhou Wang, Haley A. Banta, Abi D. Jewett, Abigail J. Salzhauer, Tess Z. Griffin, and Nicholas C. Jacobson. 2025. Randomized Trial of a Generative AI Chatbot for Mental Health Treatment. *NEJM AI*, 2(4):AIoa2400802.

Jennifer Hettema, Julie Steele, and William R. Miller. 2005a. Motivational Interviewing. *Annual Review of Clinical Psychology*, 1:91–111.

Jennifer Hettema et al. 2005b. A meta-analysis of motivational interviewing: Twenty-five years of empirical studies. *Addictive Behaviors*.

Jon Houck, David Moyers, Jennifer Miller, Tim Glynn, and Theresa Hallgren. 2010. Motivational Interviewing Skill Code (MISC) Manual – Version 2.5. *Unpublished Manual*. University of New Mexico Center on Alcoholism, Substance Abuse and Addictions (CASAA).

Hang Jiang, Xiajie Zhang, Xubo Cao, Cynthia Breazeal, Deb Roy, and Jad Kabbara. 2023. PersonaLLM: Investigating the Ability of Large Language Models to Express Personality Traits. ArXiv preprint arXiv:2305.02547.

Charlotte Jones et al. 2021. Mental health care in Canada: challenges for Indigenous communities. *Canadian Journal of Psychiatry*.

Yash Kumar Joshi and et al. 2023. Persona-Guided In-Context Learning for LLMs. ArXiv preprint.

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. *arXiv preprint arXiv:2402.17937*.

Hyunwoo Kim, Byeongchang Kim, and 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)*, pages 904–916, Online. Association for Computational Linguistics.

Jiwei Li, Michel Galley, Chris Brockett, Georgios P. Spithourakis, Jianfeng Gao, and Bill Dolan. 2016. A Persona-Based Neural Conversation Model. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 994–1003, Berlin, Germany. Association for Computational Linguistics.

Junyi Li, Xiaoxue Cheng, Wayne Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2023. HaluEval: A Large-Scale Hallucination Evaluation Benchmark for Large Language Models. *Proc. of the EMNLP 2023*, pages 7801–7820.

Margaret Li, Stephen Roller, Ilya Kulikov, Sean Welleck, Y-Lan Boureau, Kyunghyun Cho, and 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*, pages 4715–4728, Online. Association for Computational Linguistics.

Sucharat Limpanopparat, Erin Gibson, and Dr Andrew Harris. 2024. User engagement, attitudes, and the effectiveness of chatbots as a mental health intervention: A systematic review. *Computers in Human Behavior: Artificial Humans*, 2(2):100081.

N. Lindson-Hawley, T. Thompson, and P. Aveyard. 2015. Motivational Interviewing for Smoking Cessation. *Cochrane Database of Systematic Reviews*, 2015(3):CD006936.

Andy Liu, Mona Diab, and Daniel Fried. 2024. Evaluating Large Language Model Biases in Persona-Steered Generation. In *Findings of the Association for Computational Linguistics (NAACL)*.

David Llewellyn et al. 2006. Cost-effectiveness of cognitive–behavioural therapy for mental disorders in Canada. *Canadian Journal of Psychiatry*. Reviewed economic impact showing scarcity of publicly funded CBT in Canada (Llewellyn et al., 2006).

Andrea Madotto, Zhaojiang Lin, Chien-Sheng Wu, and Pascale Fung. 2019. Personalizing Dialogue Agents via Meta-Learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5459, Florence, Italy. Association for Computational Linguistics.

Zafarullah Mahmood, Soliman Ali, Jiading Zhu, Mohamed Abdelwahab, Michelle Yu Collins, Sihan Chen, Yi Cheng Zhao, Jodi Wolff, Osnat C. Melamed, Nadia Minian, Marta Maslej, Carolynne Cooper, Matt Ratto, Peter Selby, and Jonathan Rose. 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*, pages 25008–25043, Vienna, Austria. Association for Computational Linguistics.

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

Selina Meyer and David Elsweiler. 2025. LLM-based Conversational Agents for Behaviour Change Support: A Randomised Controlled Trial Examining Efficacy, Safety, and the Role of User Behaviour. *International Journal of Human-Computer Studies*, 200:103514.

Katharine Miller. 2021. Introducing the Center for Research on Foundation Models (CRFM). Stanford Institute for Human-Centered Artificial Intelligence (HAI) News. Accessed: 2025-07-21.

William R. Miller. 1983. *Motivational Interviewing with Problem Drinkers*, volume 11. Cambridge University Press.

William R. Miller. 2023. The evolution of motivational interviewing. *Behavioural and Cognitive Psychotherapy*, 51(6):616–632.

William R. Miller and Stephen Rollnick. 1991. *Motivational Interviewing: Preparing People to Change Addictive Behavior*. Guilford Press.

William R. Miller and Stephen Rollnick. 2002. *Motivational Interviewing: Preparing People for Change*, 2nd edition. The Guilford Press, New York.

William R. Miller and Stephen Rollnick. 2012. *Motivational Interviewing: Helping People Change*, 3 edition. Guilford Press, New York, NY.

William R. Miller and Stephen Rollnick. 2013. *Motivational Interviewing: Helping People Change*, 3rd edition. The Guilford Press, New York.

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

William R. Miller and Gary S. Rose. 2009. Toward a Theory of Motivational Interviewing. *American Psychologist*, 64(6):527–537.

Adam S Miner, Liliana Laranjo, and Ahmet Baki Kocaballi. 2020. Talking to Machines About Mental Health: A Scoping Review of Conversational Agents in Psychiatry. *JMIR Mental Health*, 7(6):e18572.

Adam S. Miner et al. 2016. Smartphone-based conversational agents and responses to questions about mental health, interpersonal violence, and physical health. *JAMA Internal Medicine*.

Theresa B. Moyers, Lauren N. Rowell, Jennifer K. Manuel, Danielle Ernst, and William R. Miller. 2016a. The Motivational Interviewing Treatment Integrity Code (MITI 4): Rationale, Preliminary Reliability and Validity. *Journal of Substance Abuse Treatment*, 65:36–42.

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

National Collaborating Centre for Indigenous Health. 2020. Poverty as a social determinant of First Nations, Inuit, and Métis health. Details systemic underfunding exacerbating service inequities (National Collaborating Centre for Indigenous Health, 2020).

Tong Niu and Mohit Bansal. 2018. Polite Dialogue Generation Without Parallel Data. *Transactions of the Association for Computational Linguistics*, 6:373–389.

Selcen "Onc"u, Fulya Torun, and Hilal Hatice "Ulk"u. 2025. AI-powered standardised patients: evaluating ChatGPT-4.0's impact on clinical case management in intern physicians. *BMC Medical Education*, 25:278.

OpenAI. 2023. GPT-4 Technical Report. *Preprint*, arXiv:2303.08774. `https://openai.com/research/gpt-4`.

OpenAI. 2024. GPT-4o System Card. *Preprint*, arXiv:2410.21276.

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, Red Hook, NY, USA. Curran Associates Inc.

Joon Sung Park, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and 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)*.

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

Abbey E. Poirier et al. 2019. Estimates of the current and future burden of cancer attributable to active and passive tobacco smoking in Canada. *Preventive Medicine*, 122:9–19.

J. Proudfoot, D. Goldberg, A. Mann, B. Everitt, I. Marks, and J. A. Gray. 2003. Computerized, interactive, multimedia cognitive-behavioural program for anxiety and depression in general practice. *Psychological Medicine*, 33(2):217–227.

Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Kurt Shuster, Eric Michael Smith, et al. 2021. Recipes for Building an Open-Domain Chatbot. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 300–325, Online. Association for Computational Linguistics.

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

Stephen Rollnick and William R. Miller. 1995. What is Motivational Interviewing? *Behavioural and Cognitive Psychotherapy*, 23(4):325–334.

Stephen Rollnick, William R. Miller, and Christopher C. Butler. 1997. The Motivational Interviewing Approach to Helping Patients Change Behavior. *BMJ*, 324:1402–1405.

Stephen Rollnick et al. 2008. Motivational Interviewing in Health Care: Helping Patients Change Behavior.

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

Ateka Saiyed, Jetta Layton, Brian Borsari, Jie Cheng, Tatyana Kanzaveli, Maksim Tsvetovat, and Jason Satterfield. 2022. Technology-Assisted Motivational Interviewing: Developing a Scalable Framework for Promoting Engagement with Tobacco Cessation Using NLP and Machine Learning. In *Proc. of the 5th Int. Workshop on Health Intelligence (WI-IAT)*, volume 206 of *Procedia Computer Science*, pages 121–126. International Society for Research on Internet Interventions 11th Scientific Meeting.

Jr. Sampson, J. P. 1986. The use of computer-assisted instruction in support of psychotherapeutic processes. *Computers in Human Behavior*, 2:1–19.

Vinay Samuel, Henry Peng Zou, Yue Zhou, Shreyas Chaudhari, Ashwin Kalyan, Tanmay Rajpurohit, Ameet Deshpande, Karthik Narasimhan, and Vishvak Murahari. 2024. PersonaGym: Evaluating Persona Agents and LLMs. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 10836–10863, Miami, FL, USA. Association for Computational Linguistics.

Sanic Community Organization. 2025. Sanic: Accelerate your web app development — Build fast. Run fast. `https://github.com/sanic-org/sanic`. Accessed on August 10, 2025.

Till Scholich, Maya Barr, Shannon S. Wiltsey-Stirman, and Shriti Raj. 2025. A Comparison of Responses from Human Therapists and Large Language Model–Based Chatbots to Assess Therapeutic Communication: Mixed Methods Study. *JMIR Mental Health*, 12:e69709.

Lennart Seitz. 2024. Artificial empathy in healthcare chatbots: Does it feel authentic? *Computers in Human Behavior: Artificial Humans*, 2(1):100067.

P. M. Selmi, M. H. Klein, J. H. Greist, S. P. Sorrell, and H. P. Erdman. 1990. Computer-administered cognitive-behavioral therapy for depression. *American Journal of Psychiatry*, 147(1):51–56.

D. Servan-Schreiber. 1986. Artificial intelligence and psychiatry. *Journal of Nervous and Mental Disease*, 174(4):191–202.

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

Bangzhao Shu, Lechen Zhang, Minje Choi, Lavinia Dunagan, Lajanugen Logeswaran, Moontae Lee, Dallas Card, and 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 (Volume 1: Long Papers)*, pages 5263–5281, Mexico City, Mexico. Association for Computational Linguistics.

Kurt Shuster, Jing Xu, Mojtaba Komeili, Da Ju, Eric Michael Smith, Stephen Roller, Megan Ung, Moya Chen, Kushal Arora, Joshua Lane, Morteza Behrooz, William Ngan, Spencer Poff, Naman Goyal, Arthur Szlam, Y-Lan Boureau, Melanie Kambadur, and Jason Weston. 2022. Blenderbot 3: a deployed conversational agent that continually learns to responsibly engage. *Preprint*, arXiv:2208.03188.

Eric Michael Smith, Emily Dinan, Kurt Shuster, Y-Lan Boureau, and Jason Weston. 2020. Mitigating Harmful Stereotypes in Persona-Based Conversational Agents. Technical report, Facebook AI Research.

Haoyu Song, Yan Wang, Wei-Nan Zhang, Xiaojiang Liu, and 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*, pages 5821–5831, Online. Association for Computational Linguistics.

Akhilesh Sudhakar, Bhargav Upadhyay, and 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)*, pages 3269–3279, Hong Kong, China. Association for Computational Linguistics.

Gemini Team. 2025. Gemini: A Family of Highly Capable Multimodal Models. *Preprint*, arXiv:2312.11805.

John Torous and Laura Roberts. 2017. Digital mental health and COVID-19: Using technology today to accelerate the curve on access and quality tomorrow. *JMIR Mental Health*.

John Torous, Laura White Roberts, and Joseph Firth. 2023. Generative Artificial Intelligence and Mental Health: Opportunities, Challenges, and Responsibilities. *The Lancet Psychiatry*, 10(5):308–310.

Ross Valentino, Jenny Shen, Sophie McLean, Sherry Zhang, and Amit Singh. 2024. Evaluating Large Language Models for Motivational Interviewing: A Mixed-Methods Study. *Journal of Medical Internet Research*, 26:e48295.

Ashish Vaswani et al. 2017. Attention is All You Need. *Advances in Neural Information Processing Systems*, 30.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and 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, Red Hook, NY, USA. Curran Associates Inc.

Joseph Weizenbaum. 1966. ELIZA — A Computer Program for the Study of Natural Language Communication Between Man and Machine. *Communications of the ACM*, 9(1):36–45.

Sean Welleck, Jason Weston, Arthur Szlam, and Kyunghyun Cho. 2019. Dialogue Natural Language Inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3731–3741, Florence, Italy. Association for Computational Linguistics.

Robert West and Tim Sohal. 2006. "Catastrophic" Pathways to Smoking Cessation: Findings from National Survey. *BMJ*, 332(7539):458–460. Epub 2006 Jan 27.

Zhouhang Xie, Bodhisattwa Prasad Majumder, Mengjie Zhao, Yoshinori Maeda, Keiichi Yamada, Hiromi Wakaki, and Julian McAuley. 2024. Few-shot Dialogue Strategy Learning for Motivational Interviewing via Inductive Reasoning. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 13207–13219, Bangkok, Thailand. Association for Computational Linguistics.

Jing Xu, Arthur Szlam, and Jason Weston. 2022. Beyond goldfish memory: Long-term open-domain conversation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5180–5197, Dublin, Ireland. Association for Computational Linguistics.

¡em¿et al.¡/em¿ Yang. 2023. Large Language Models with Persona-based Steering (technical report).

Fan Yang, Shibani Santurkar, Xuezhou Li, et al. 2024. Are language models really good at predicting human preferences? Revisiting survey prediction benchmarks. *arXiv preprint arXiv:2401.01204*.

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. *CIN: Computers, Informatics, Nursing*. Article published July 24, 2025.

Faika Zanjani, Bree Miller, Nicholas Turiano, Jennifer Ross, and David Oslin. 2008. Effectiveness of telephone-based referral care management, a brief intervention to improve psychiatric treatment engagement. *Psychiatric Services*, 59(7):776–781.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and 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 (Volume 1: Long Papers)*, pages 2204–2213, Melbourne, Australia. Association for Computational Linguistics.

Yinhe Zheng, Zekang Chen, Rongsheng Zhang, Shilei Huang, Xiaoxi Mao, and Minlie Huang. 2021. Stylized Dialogue Response Generation Using Stylized Unpaired Texts. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 14558–14566.

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. *TMLR*.