Deployment and Evaluation of a Motivational-Interviewing Chatbot for Moving Smokers Towards the Decision to Quit and Distillation of Large Foundational Language Models for Generating MI Reflections

by

Andrew Brown

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

Graduate Department of Electrical and Computer Engineering
University of Toronto

© Copyright 2023 by Andrew Brown

Deployment and Evaluation of a Motivational-Interviewing Chatbot for Moving Smokers Towards the Decision to Quit and Distillation of Large Foundational Language Models for Generating MI Reflections

Andrew Brown
Master of Applied Science
Graduate Department of Electrical and Computer Engineering
University of Toronto
2023

Abstract

Motivational Interviewing (MI) is a validated therapeutic approach that has been shown to help ambivalent people struggling with addiction move toward change. MI has been applied to many behaviours, including smoking cessation. If conversational agents could effectively apply MI, they may provide a scalable way to help more people access this therapy. Previous attempts to provide MI therapy through text-based conversational agents have typically employ scripted responses to client statements, but such non-specific responses have been shown to reduce effectiveness. A key technique in MI is to ask open-ended questions and then provide a reflection of the response to evoke contemplation in the client. Recent advances in Natural Language Processing provide a new way to create responses that are specific to client's statements, using a Transformer-based Language Model. We present the design, evolution and impact assessment of a chatbot that makes use of generated reflections, whose goal is to guide ambivalent smokers toward the decision to quit. Through four trials of 349 participants, we show that the chatbot significantly increases participants' confidence to quit smoking one week after the conversation compared to before the conversation (P=.001). As a key part of the chatbot is the language model that produces reflection, and it is difficult and sometimes impractical to run a clinical bot with a corporation's proprietary model, we explore methods of model distillation to train smaller, more practical language models to generate MI-adherent reflections. We present a method for distilling the specific tasks of generating MI reflections from a Large Foundational Language Model (GPT-4) into smaller models. We show that GPT-4 can generate MI-adherent reflections near 100% success and use output generated from that model to fine-tune the much smaller GPT-2 family as form of knowledge distillation. We also use GPT-4 as a zero-shot evaluator to classify the quality distilled student model outputs and validate that classifier with a triple human-review. We show that the GPT-2 small achieves an 83% success rate on a hold-out test set and the GPT-2 XL achieves 90% success. In addition, our GPT-4 zero-shot prompt evaluator achieves significantly high inter-rater reliability (.61 Cohen-Kappa) with triple-human review.

Acknowledgements

I would like to thank my supervisor, Professor Jonathan Rose, whose unwavering support, guidance, and insights helped me grow as a researcher and a person. Our weekly meetings were the highlight of my degree. You've taught me to embrace the truth in my work and question everything like a great engineer should. For that I'm grateful.

I also want to thank our collaborators at CAMH who supported this research with their domain experience and dedication to mental health. They have help this research immensely.

To our collaborative research group filled with passionate graduate and summer students, I thank all of you. The exchange of ideas and commitment to our shared goals hugely contributed to the success of this project.

Finally, to my family, thank you for the never-ending support and faith in me. I love all of you more than I could ever express in words.

Contents

1 Introduction **1**

1.1 Smoking Cessation and Motivational Interviewing . . . 1

1.2 Language Models and Distillation . . . 2

1.3 Focus and Goals . . . 3

1.4 Contribution . . . 3

1.5 Organization . . . 3

2 Background and Related Work **4**

2.1 Motivational Interviewing and Smoking Cessation . . . 4

2.1.1 Motivational Interviewing Reflections . . . 5

2.1.2 Readiness Ruler Survey . . . 5

2.1.3 Consultation and Relational Empathy Survey . . . 6

2.2 Natural Language Processing . . . 6

2.2.1 Word Embeddings . . . 6

2.2.2 Transformer-based Language Models . . . 7

2.2.3 Chatbots . . . 9

2.2.4 Language Model Knowledge Distillation . . . 10

2.3 Related Work on Motivational Interviewing Chatbots . . . 13

3 Chatbot Training and Study Method **15**

3.1 Overall Chatbot System Study Design . . . 15

3.2 Core Conversation . . . 17

3.3 Chatbot Versions . . . 18

3.3.1 Enhanced Conversation Structure . . . 18

3.4 Pre-Conversation Surveys and Screening . . . 19

3.5 Post-Conversation Survey . . . 20

3.6 One-Week Follow-up Survey . . . 21

3.7 Classifying Resolved Ambivalence . . . 22

3.8 Reflection Generation Training . . . 22

3.9 Software System . . . 23

3.10 Recruitment and Data Inclusion . . . 24

3.11 Statistical Analysis . . . 25

4 Chatbot Results and Discussion **26**
4.1 Recruitment Results . . . 26
4.2 Demographics . . . 26
4.3 Heaviness of Smoking . . . 27
4.4 Readiness Rulers . . . 27
4.5 Quit Attempts and Reduction of Smoking . . . 28
4.6 Consultation and Relational Empathy (CARE) Measure . . . 31
4.7 Did Ambivalence Change and in What Direction? . . . 32
4.8 Discussion . . . 34
4.8.1 Principle Findings . . . 34
4.8.2 Recruitment and Demographics . . . 34
4.8.3 Quit Attempts and Reduction of Smoking . . . 35
4.8.4 Resolution of Ambivalence . . . 35
4.9 Chatbot Contribution Attribution . . . 36

5 Methods for Generating and Evaluating Reflections and Distillation **37**
5.1 Generating Reflections from GPT-4 . . . 37
5.1.1 GPT-4 Prompting Schema and Formatting . . . 38
5.1.2 Gathering Motivational Interviewing Transcripts for Dataset Creation . . . 38
5.1.3 Prompt Engineering of GPT-4 for Reflection Generation . . . 39
5.2 Evaluation of Reflections using GPT-4, Human Review, and Automated Metrics . . . 39
5.2.1 GPT-4 MI-Adherence . . . 41
5.2.2 GPT-4 Task Classification . . . 41
5.2.3 GPT-4 Reflection Evaluation Overall Process . . . 42
5.2.4 Human Review . . . 43
5.2.5 Automated Metrics . . . 44
5.3 Knowledge Distillation Method . . . 44
5.3.1 Knowledge Distillation Overview . . . 45
5.3.2 Distillation Fine-tuning . . . 46
5.3.3 Student Model Selection . . . 47
5.4 Experiment . . . 47
5.5 Experimental Setup . . . 47

6 Reflection Generation, Evaluation and Distillation Results **48**
6.1 Performance of GPT-4 Generation of Reflections . . . 48
6.2 Performance of GPT-4-based Evaluation . . . 49
6.3 Performance of Distilled Student Reflection Generation Models . . . 50
6.3.1 Distilled Student MI-Adherence . . . 51
6.3.2 Distilled Student Task Classification . . . 51
6.4 Automated Metrics Results . . . 52
6.5 Knowledge Distillation Contribution Attribution . . . 52

7 Conclusions, Limitations, and Future Work **54**
7.1 Conclusions . . . 54
7.2 Limitations . . . 54
7.3 Future Work . . . 55

A **57**
A.0.1 MI Chatbot Transcript Version 4.7 - Questions but no reflections . . . 57
A.0.2 MI Chatbot Transcript Version 5.0 - Questions & Reflections with Generator Version 1 . . . 57
A.0.3 MI Chatbot Transcript Version 5.1 - Questions Reflections with Generator Version 2 . . . 59
A.0.4 MI Chatbot Transcript Version 5.2 - Enhanced/Extended Conversation based on v5.1 . . . 60

B **62**
B.0.1 Chatbot Feedback and Confidence Change with Ambivalence Resolution Label 62

C **83**
C.0.1 Consultation and Relational Empathy Survey . . . 83

Bibliography **88**

List of Tables

2.1 Simple vs Complex Reflections . . . 5
2.2 The Readiness Ruler . . . 5
2.3 GPT-2 Family Architectures . . . 9
2.4 Self-Instruct Dataset Examples . . . 12

3.1 Chatbot Versions . . . 18

4.1 Participant Count for each Chatbot Version Corresponding to Study Flowchart . . . 26
4.2 Demographics of Participants . . . 28
4.3 Heaviness of Smoking Measures . . . 29
4.4 Average Confidence, Before, After and 1 Week After Conversation . . . 29
4.5 Average Importance, Before, After and 1 Week After Conversation . . . 29
4.6 Average Readiness, Before, After and 1 Week After Conversation . . . 29
4.7 Number of Participants Whose Confidence who Increased, Decreased, or Stayed the Same 1 Week Later . . . 30
4.8 Percentage of Participants Who Made Quit Attempts Before and After Conversation 31
4.9 Count of Participants who Reduced/Did Not Reduce Smoking . . . 32
4.10 CARE Empathy Measure . . . 33
4.11 Counts of Quit, Smoke and Same Ambivalence classes . . . 33

5.1 MI Chatbot Transcript Excerpt . . . 38
5.2 GPT-4 Prompt for MI-Adherence . . . 42
5.3 GPT-4 Prompt for Task Classification . . . 42
5.4 Task 1 and Task 2 Dataset Entry Example . . . 46

6.1 MI-Adherence and Task Classification fractions of success using the GPT-4-based classifiers (The GPT-4 column) and using Human Review (HR). Columns 4 and 5 also include the count of candidate reflections which make it past MI-adherence in parenthesis. . . . 48
6.2 Inter-Rater Reliability Cohen Kappa scores between GPT-4 and Human Reviewers on three evaluation tasks. The last column, combines MI-Adherence and Task Classification into one evaluation task. . . . 49

6.3 MI-Adherence and Task Classification scores of distilled student models as well as reprise of results for teacher GPT-4 reflection generator. Columns 5 and 6 give the count of examples GPT-4 and human reviewers see for each model in parenthesis. HR stands for Human Review. 51
6.4 ROUGE and BERTScore Automated Metric scores for each distilled student model. ROUGE is broken into three types, ROUGE-1 (RG-1), ROUGE-2 (RG-2), and ROUGE-L (RG-L). For all automated metrics candidate sequences are the reflection generated by the distilled student model and reference sequences are the reflection which GPT-4 generated. 52

List of Figures

2.1 Generative Transformer-based Language Model Architecture 8

3.1 Overall Design of Each Study for Each Chatbot Version 16
3.2 Screen that measures HSI and Quit Attempts 19
3.3 The Readiness Ruler 20
3.4 One Week Later Quit Smoking Actions 21
3.5 Chatbot Architecture 24

4.1 Flowchart of Study Procedure 27
4.2 Distribution of all 4 versions' Confidence Values Pre, Post, and Week Later 30
4.3 Distribution of all 4 versions' Importance Values Pre, Post, and Week Later 31
4.4 Distribution of all 4 versions' Readiness Values Pre, Post, and Week Later 32
4.5 CARE Survey Distribution 33

5.1 Reflection Data Generation Format 40
5.2 Reflection Model Evaluation Pipeline 43
5.3 Knowledge Distillation Overview 45

Chapter 1

Introduction

1.1 Smoking Cessation and Motivational Interviewing

Tobacco use is the leading preventable cause of premature death in Canada, killing 45,000 Canadians every year [1] with 4.6 million Canadians ensnared by the addiction [2]. The harmful effects of tobacco use are well-documented and include significantly increased risks of cancer, heart disease, stroke, diabetes, and other harmful diseases [3]. Furthermore, individuals with constant exposure to secondhand smoking suffer a 30% increased risk of lung cancer, heart disease, and stroke [3]. Despite these well-known risks, tobacco use remains prevalent in Canada, with millions of individuals making the decision not to quit each year. Globally, 80% of all smokers are ambivalent toward their addiction [4] meaning that the positives and negatives of smoking balance out, and so they make little to no effort to stop [5].

Smokers can be guided towards the decision to quit by a widely used talk therapy approach known as Motivational Interviewing (MI) [6]. MI is a style of interaction between a clinician and participant that encourages a non-judgmental, open environment where participants can explore their thoughts, feelings, and motivations related to a behaviour. Specifically for smoking cessation, clinicians can use MI to guide individuals towards healthy behavior change by helping them to recognize the disadvantages of their smoking while recognizing the health and psychological benefits of changing it. This approach is effective in motivating individuals who may be hesitant, conflicted, ambivalent, or have made previous attempts to quit without success [7, 8]. The first goal of MI is distinct from smoking cessation efforts such as Nicotine Replacement Therapy [9], that assume the smoker is ready and willing to quit; however, the decision to quit is a necessary precursor of any quit attempt.

Since MI relies on highly trained clinicians working in hospitals and specialized clinics, it is both expensive and difficult to access. Clinicians are usually engaged only after a health issue occurs, whereas earlier engagement with a more accessible chatbot could improve health outcomes and even prevent illness or death. For every two smokers helped to quit, one life is saved from a tobacco-related death [10]. This motivates us to investigate the feasibility of automating an MI-style conversation which could be deployed directly to smokers online, helping more people, more easily and sooner than otherwise.

It is challenging, however, for a machine to achieve the level of understanding and facility needed to practice MI. Prior efforts at automated therapy, beginning with ELIZA [11] and proceeding through many generations of dialogue systems [12, 13, 14] suffer from two key issues: first, since most of the outgoing text is scripted, these systems have difficulty responding to the specific things that an individual says. These responses are often seen by users as either repetitive or too generic [15]. Second, many chatbots do not permit free-form text input, which prevents the user from expressing themselves fully. Recent dramatic advances in Natural Language Generation [12, 16, 17, 18, 19, 20] have produced language models that can generate very human-like responses that are more relevant to the free-form dialog of a human.

In this work we present the design of several versions of a chatbot, called MIBot, that makes use of these new kinds of language models to generate context-specific responses to users, in combination with scripted interactions. We also present a scientific infrastructure for measuring the impact of MIBot on recruited smokers.

1.2 Language Models and Distillation

As this work uses Transformer-based Language Models to generate MI reflections, and our experience has pushed us towards using the very largest and recent models as they have remarkably good performance. However, those models are both very large and quite proprietary within the organizations that create them. A chatbot, like many other clinical settings requires data privacy, a situation where we must guarantee that the data mental health clinicians and clients are sending is only seen by the correct parties. This led us to creating MIBot using our own software directly under our control so that data privacy can be maintained. Such model ownership requires the user to have a proper license to train and deploy the model, compute resources capable of hosting the model, and data for training the model. For MIBot, we selected OpenAI's Generative Pre-trained Transformer 2 (GPT-2) [17] as a model which satisfied our model ownership requirements. All MI reflections described in the MIBot intervention measurement in this work are generated with GPT-2.

After deploying and evaluating MIBot, we were motivated to explore using other, more advanced Transformer-based Language Models to generate MI reflections. Since the advent of GPT-2 in 2019, OpenAI has remained at the forefront of Transformer-based Language Models with GPT-3 [18] in 2020, GPT3.5 also known as ChatGPT [19] in 2022, and lastly GPT-4 [20] in 2023. GPT-4 is the most advanced language model to this date, and scores the highest on a breadth of tasks including 90th percentile scores on the United States Uniform Bar Examination, AP Psychology, SAT Math, and many other tasks. To solve these tasks, users prompt GPT-4 using few-shot [21] and zero-shot [22] learning, which are now widely-used techniques for prompt engineering a language model. Few-shot learning gives related examples to the task at hand, then finally an incomplete example, whereas zero-shot learning describes the task through natural language instruction without examples. These types of prompting are useful because they allow users to create useful generators and classifiers without the need for large amounts of data. For our usage of generating MI-adherent reflections, we were led to investigate if GPT-4 is capable of MI techniques.

In collaboration with MI-experts, we hand-engineered a zero-shot instruction to generate MI reflections with GPT-4 which satisfied human annotators. Unfortunately, we realized we could not deploy it due to privacy and ethical constraints, leaving us looking for a method to capture this performance with an own-able model.

A known way of transferring performance from an unreachable large model to a smaller one is through knowledge distillation, first achieved by Hinton et. al. [23]. Furthermore, recent research shows that Hinton's method can be applied to language models [24, 25]. In this work we present a second key goal of this research, to distill knowledge from GPT-4 to GPT-2 for the generation MI-adherent reflections.

1.3 Focus and Goals

The objective of this work is to explore methods for automating the MI talk therapy technique on a computer, and to measure its efficacy. Within this goal, we first focus on deploying and evaluating a smoking cessation chatbot which uses a Large Language Model for generating MI reflections. Second, we explore a method of knowledge distillation to transfer knowledge from GPT-4 to a range of student Language Models to generate MI adherent reflections. Within this distillation method, we evaluate how changing the type of reflection of knowledge and student architecture size affects performance.

1.4 Contribution

To achieve the goals stated above, this work makes the following contributions:

- Continuing the design and evolution of a chatbot intervention for hosting a Motivational Interviewing conversation about smoking cessation.
- Implementation of the Chatbot intervention using AWS cloud infrastructure.
- Evaluation and validation of the Chatbot intervention.
- Demonstration of a method for distilling teacher language models to smaller student language models for the generation of MI reflections.
- Evaluation of distilled student language Models using human review, automated metrics and the teacher language model.

1.5 Organization

This dissertation is structured as follows: Chapter 2 reviews relevant background on the Motivational Interviewing approach, transformer-based language models, chatbots, knowledge distillation, and finally related work on knowledge distillation and automated MI conversations. Chapter 3 presents the design of the MI chatbot and the method of how it is deployed and measured. Chapter 4 gives the results of its deployment with recruited smokers. Chapter 5 describes a method for generating MI reflections with a large foundational language model and knowledge distillation of MI reflections from a larger model into smaller ones. Chapter 6 presents the evaluation of our distilled student models. Chapter 7 concludes the work, alongside limitations and an overview of future work.

Chapter 2

Background and Related Work

This chapter reviews relevant background and prior work. First we review background on Smoking Cessation and the Motivational Interviewing (MI) behaviour change method. Next, we discuses relevant Natural Language Processing (NLP) techniques in Word Embeddings, Transformer-based Language Models, Chatbots, and Knowledge Distillation. Finally, we reviews prior work on MI-based chatbots.

2.1 Motivational Interviewing and Smoking Cessation

MI is a counselling approach that helps patients increase motivation towards changing unhealthy behaviours, including addiction [6]. Ambivalence is a conflicted state where opposing attitudes or feelings coexist in an individual toward changing a behaviour, resulting in no action taken. The goal of MI is to resolve this ambivalence, moving a person toward positive change. MI counsellors use a structured conversation including open-ended questions and reflective listening, which encourage a patient to contemplate the roots of a behaviour and guide them toward overcoming their ambivalence.

MI has been a widely used approach to motivate smoking cessation [6]. Globally, 80% of smokers are ambivalent about their smoking [4], and make no current effort to stop [26]. This makes smoking cessation a target for MI, as resolving ambivalence is the main goal of the talk therapy technique. In multiple studies, MI has been shown to be successful for smoking cessation [27, 28].

Early in the MI approach, clinicians often use four specific skills: **O**pen-ended Questions, **A**ffirmation, **R**eflective Listening, and **S**ummary Reflections in what is referred to as the OARS Model [6]. These four techniques help participants express themselves, allowing clinicians to understand and communicate a behaviour from the participant's perspective. Open-ended questions encourage the client to self-explore by sharing perspectives and experiences. Affirmations recognize the client's strengths to reinforce their confidence. Reflective listening mirrors the client's thoughts and emotions, enabling them to recognize their own beliefs and contradictions and encouraging them to continue contemplation. Lastly, Summaries remind clients of the most important parts of the intervention, providing a clear overview and giving another chance for contemplation. In the chatbot presented in this work, we employ concepts from the OARS model, specifically using open-ended questions and reflective listening. Below we discuss reflective listening in-depth, and then describe survey metrics which MI clinicians (and our chatbot) use to measure progress towards the decision to quit.

Table 2.1: Simple vs Complex Reflections

Participant Statement	Simple Reflection	Complex Reflection
I think reducing my smoking would improve my health.	You believe reducing your smoking would improve your health.	So, your health is really important to you.
Smoking allows me to relax and just think about everyday stuff.	Smoking allows you to relax.	Taking a break and having alone time to process your thoughts is valuable to you.
I don't like making other people uncomfortable with my smoking.	You don't enjoy making people feel uncomfortable with your smoking.	You might be feeling self-conscious about your smoking.

2.1.1 Motivational Interviewing Reflections

A core skill used by MI practitioners is reflective listening [6, 29, 30]. A therapist uses reflective listening to respond to a participant's statements with words that both repeat what is said and guide the participant toward continued exploration of their thoughts and feelings. Reflective listening responses are called *reflections* in MI, and can be either simple or complex. A *simple* reflection repeats or rephrases a participant statement, so as to convey understanding and invite continuation of the conversation [6]. A *complex* reflection attempts to infer something relevant between the recent utterance to either the prior utterances or to a guess of something generally relevant [6]. The distinction between simple and complex reflections is important to this work, as we create language models to generate both kinds of reflections. Table 2.1 shows three examples of simple and complex reflections in response to a participant's utterance. Notably, the same utterance can be responded to with either a simple or complex reflection.

Table 2.2: The Readiness Ruler

Measurement	Question
Importance	On a scale from 0 to 10, how important is it for you right now to stop smoking?
Confidence	On a scale from 0 to 10, how confident are you that you would succeed at stopping smoking if you start right now?
Readiness	On a scale from 0 to 10, how ready are you to start making a change at stopping smoking?

2.1.2 Readiness Ruler Survey

The Readiness Ruler [31] is a survey metric which asks participants their confidence that they could quit a behaviour now (which we will refer to as the confidence scale), their readiness to quit a behaviour now (the readiness scale), and how important they feel it is for them to quit a behaviour now (the importance scale). These three ratings are recorded on an 11-point scale at different times during the MI counselling process to gauge the specific aspects of a participant's feeling towards change. By tracking the progress along the ruler, both participants and clinicians can witness developments in overall readiness to quit and address any barriers. Table 2.2, shows the specific version of the Readiness Ruler we use for measuring smoking cessation readiness. This survey is used multiple times during the MIBot experiment to determine the effectiveness of the chatbot. For smoking cessation, the motivation of our chatbot, confidence to quit on the readiness ruler is known as the main predictor of success [32, 33, 34] so we use this metric as one of the main focuses.

2.1.3 Consultation and Relational Empathy Survey

The Consultation and Relational Empathy (CARE) Score [35] is a validated survey designed to assess the perceived empathy of a clinician from the client's perspective. The survey is has 10 questions, each a five option likert-scale from poor to excellent. The scores of the 10 questions are converted into numerical scores, and then summed. The CARE Score provides insight into the strength of the client-counselor connection, revealing areas for improvement or reinforcement. Originally, the CARE survey was developed and rigorously tested for use by general medical practitioners [35], but has since been successfully used by other medical staff, allied health professionals (AHPs) and nurses [36]. In this work, we use the CARE scale to measure the perceived empathy of the MI chatbot.

2.2 Natural Language Processing

Natural Language Processing (NLP) is a field touching linguistics, computer science, and machine learning that is concerned with the interactions between computers and human natural language [12]. Recently, there have been significant advances in the field of NLP [18, 19, 20, 37]. With the development of word embeddings [38, 39] and deep learning [40], various neural network based architectures have been used to solve NLP tasks such as Natural Language Generation (NLG) [17, 18, 19, 37, 20] and Natural Language Understanding (NLU) [41, 42, 43, 44]. Furthermore,Knowledge distillation [23], is an important topic in NLP, and has been successfully applied to language models [24, 25, 45, 46]. We review these concepts, beginning in Section 2.2.1 which discusses word embeddings and Section 2.2.2 which explains the architecture of Transformer-based Language Models and Section 2.2.3 which discusses chatbots and dialogue agents. Finally Section 2.2.4 explains techniques for applying knowledge distillation approaches to language models.

2.2.1 Word Embeddings

Word embeddings (or vectors) are one of the basic building blocks of modern deep-learning-based NLP [12]. A word embedding is an encoded representation of a word into a fixed-dimensional vector of real numbers. These embeddings represent the meaning of the words, and language models use them as inputs for prediction tasks [12].

Historically, word embeddings were created using statistical-based approaches beginning with individual word frequency counts and evolving into more complex techniques. An early example of this is the "Vector Space Model" [47] by Sparak in 1975, often known as "Term Frequency - Inverse Document Frequency" or TF-IDF, which uses a frequency count of terms in documents to represent meanings of words. Another more recent statistical approach is Global Vectors for Word Representation (GloVe) [39] which factorizes a word co-occurrence matrix to create word embeddings.

Alongside statistical approaches, Bengio et. al. showed the first neural network approach to word embeddings, proposing that a neural networks could be used to learn vectors that represent words, and used these representations as input to a neural network to predict the next word in a sentence [48]. This laid the groundwork for Word2Vec [38], a ground-breaking method which utilized shallow neural networks to create word embeddings.

All the above stated methods for word embeddings are non-context dependent (static) meaning they use the same vector regardless of the words surrounding them. Recently, new transformer-based language models like Bidirectional Encoder Representations from Transformers (BERT) [42] and Generative Pre-trained Transformer (GPT) [49] have shown that static word embeddings can be transformed through neural network layers to create a learned embedding which is contextual to an input token sequence. These learned embeddings represent the totality of the input tokens, and can be utilised for downstream tasks like GPT's text-generation [49] and BERT's language understanding [42]. Below we explain Transformer-based models in more detail.

2.2.2 Transformer-based Language Models

A Transformer-based Language Model is a neural-network based model which produces context dependent representations of tokenized words [16, 50]. Input embeddings of tokens are"transformed" through layers of encoder or decoder "blocks", creating one learned embedding of feature rich contextual information. Transformers use this embedding to make predictions about additional words/tokens or for downstream tasks, such as sentiment analysis [12].

To train a transformer, hidden (masked) tokens are predicted in input text sequences. This process of predicting masked tokens from a sequence gives the transformer understanding of how to represent language in meaningful latent space such that the missing words can be predicted. This style of training does not require data labelling, since there plentiful sources of written text on the internet. Choosing which tokens to mask is dependent on the architecture details of the transformer. In this work, we focus on text-generative language models, which typically use decoder only architectures [49, 17, 18, 19, 20]. Below we explain generative language models in more detail.

Generative Language Models

Generative Language Models are a variant of the transformer-based language model which use the learned embedding to generate a subsequent token, one step at a time. Specifically, a Generative Language Model calculates the probability of a word x_t given words x_1 through x_{t-1} as seen in Equation 2.1. This type of transformer utilises a decoder only architecture with a **C**asual **L**anguage **M**asking (CLM) training objective. CLM hides (masks) future tokens and uses them as labels for prediction. This type of training is auto-regressive in nature, meaning after every prediction we add the output of the last step (and correct it if incorrect) as input for the next prediction. Transformers which are trained using the CLM objective typically excel at text-generation tasks like story telling [17, 18] or Question Answering [19, 20]

$$P(x_t|x_1, x_2, ..., x_{t-1}) \tag{2.1}$$

A typical generative transformer-based language model is illustrated in Figure 2.1. First, The input text is tokenized, and then converted into an initial embedding with positional encoding. Next,

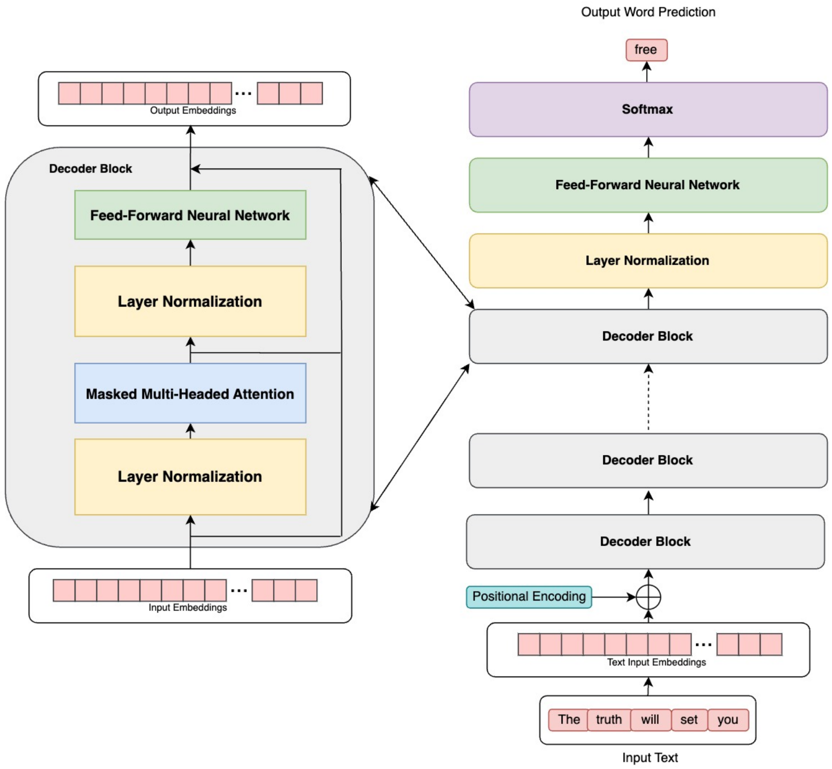

Figure 2.1: Generative Transformer-based Language Model Architecture

the embedding is transformed into a learned embedding through layers of transformer blocks. Within each block, input embeddings are passed through a Masked Multi-Headed Attention layer [16] and a Feed-Forward Neural-Network Layer [40] with Normalization layers in-between. Masked Multi-Headed Attention captures various contextual relationships within the input sequence by calculating relational scores between immediately preceding embeddings within the sequence concurrently [16]. This process is known as attention calculation. Next, the attended input is passed through a Feed-Forward layer to give the model for additional opportunity to expressing complex relationships between text. Layer normalization layers are placed in-between to stabilize the training process, reducing the impact of any irregularities. The end result of each decoder block is a more information rich embedding. After all decoder blocks, the learned embedding is passed through one last normalization layer, a feed-forward layer, then a softmax layer. The softmax layer outputs a probability for each word in the transformer's vocabulary, as part of an overall process to predict the next word. In Figure 2.1 we see the input text "the truth will set you" for which we predict "free".

The attention mechanism [16] is said to have allowed a transformer model to pay more attention to the parts of text that are more important to meaning of an input, regardless of text distance. This is in contrast to a Recursive Neural Network (RNN) which suffer from an informational bottleneck between an encoder-decoder architecture [16]. Additionally, transformer architectures benefit computationally from processing the entire input sequence in parallel, unlike RNNs, which require sequential processing. [16]. Below we review generative transformer-based language models that are relevant to this work.

Table 2.3: GPT-2 Family Architectures

Model Name	Number of Parameters	Number of Transformer Blocks	Embedding Length
GPT2-Small	124M	12	768
GPT2-Medium	355M	24	1024
GPT2-Large	774M	36	1280
GPT2-XL	1.5B	48	1600

Generative Pre-trained Transformer (GPT) Family of Language Models

One family of successful decoder-only generative language models are the Generative Pre-trained Transformer (GPT) architectures. In 2019, OpenAI released GPT-2 model [17] a generative language model ranging in size from 124 million parameters to 1.5 billion parameters, pre-trained on the WebText dataset. Table 2.3 gives the number of parameters, number of transformer blocks, and embedding length for the four sizes of GPT-2. Between GPT2 architectures, the only difference is the number of transformer blocks and embedding length, resulting in different numbers of parameters. GPT-2 can perform a variety of language generation tasks. At the time of its invention, OpenAI found that after pre-training, GPT-2 could be further trained in a process known as fine-tuning to achieve state of the art performance on Question Answering tasks [49]. In this work, we utilise GPT-2 to generate MI-reflections for our chatbot and in our knowledge distillation process.

OpenAI followed GPT-2 with GPT-3 [18] in 2020. GPT-3 has 175 billion parameters, which is more than 100 times the size of GPT-2 XL. This massively large architecture demonstrated that after pre-training and and a cleverly written input prompt, GPT-3 was capable of task specific prediction without fine-tuning. This style of learning is term *few-shot* learning [21], and *zero-shot* [22] learning. Few-shot learning providing an model with a few examples to guide its response generating process. The term ”shot” refers to the number of example prompts given to the model. On the other hand, zero-shot prompting refers to a technique where a model generates reasonable responses to prompts it has never seen before in its training data, thereby showing a certain level of understanding without any explicit pre-training.

Most recently, OpenAI has created GPT3.5 (ChatGPT) [19] (2022) and GPT-4 [20] (2023). GPT-4 exhibits human-level performance on various professional and academic benchmarks, including 90th percentile scores on the BAR Exam, AP Psychology, SAT Math, and many other tasks. Both GPT3.5 and GPT-4 are capable of few-shot and zero-shot learning at higher levels of performance than GPT-3 [20].

Notably, GPT-3, GPT-3.5 (ChatGPT), and GPT-4 are closed source models [18, 19, 20], with neither the code nor the trained parameters being publicly available. This means the learned embeddings and datasets used to train these models are not public. Furthermore, all three of these models are too large to host locally on hardware available to us. In this research, we use GPT-4 as a teacher model for knowledge distillation.

2.2.3 Chatbots

Chatbots are computer-based systems that converse with humans through natural language [12]. These systems can be task-oriented or extended-conversation oriented. Famous task-oriented chatbots are exemplified in Apple's Siri, Amazon's Alexa, or Microsoft's Cortana [51]. Generally, these systems don't have extended conversations and serve to accomplish a specific task. Examples of extended-conversation chatbots are the ELIZA Chatbot [11], Stanford's Chirpy Cardinal [52], and Microsoft's XiaoIce [53]. These chatbots aim to keep the user engaged through extended conversation. In the present work, our chatbot shares both the goals of a task-oriented chatbot and extended-conversation chatbot. MI interventions are usually extended but also have the goal of guiding a client toward behaviour change [12].

Building chatbots require three capabilities: extracting meaning out of utterances which is the domain of Natural Language Understanding (NLU), response generation, and keeping the conversation context. These strategies are addressed by chatbots which apply Rule-based and Corpus-based techniques [12], which are described in the next two sections.

Rule Based Chatbots

Rule-based chatbots use hard-coded rules to implement NLU, response generation, and conversation context. ELIZA [11] a rule-based chatbot from 1966, selected specific keywords in an utterance to extract meaning. After decomposing the utterance into rules, ELIZA combined user utterances with pre-written text to create a response. Both NLU and response generation work together to keep the conversation context. ELIZA is a strictly rule-based chatbot, however rule-based chatbots in the present era typically include more complex corpus-based techniques. For example, in the NLU stage and response generation, they combine rule-based and corpus-based methods.

Corpus-Based Chatbots

Corpus-based chatbots requires the three capabilities mentioned above to be learned from an external corpus [12]. This corpus should contain text relating to the domain of which the chatbot is conversing in. These systems are enormously data-intensive, requiring hundreds of millions or even trillions of words for training [54]. Incoming responses are classified with statistical or neural-based models. After this, responses are generated by using either retrieval methods or generation methods. Information retrieval finds a response from a corpus that is appropriate given the dialogue context while generation methods use a language model to generate the response given the dialogue context [12]. In this work, our chatbot uses a combination of rule-based and corpus-based chatbot techniques for NLU, conversation context, and response generation.

2.2.4 Language Model Knowledge Distillation

Knowledge distillation [23] is a technique in machine learning where a student model is trained to reproduce the behaviour of a teacher model. Generally, the goal is to achieve similar performance by a student architecture, but with a smaller model size. Knowledge distillation systems are composed of three components: the type of knowledge to be distilled, a distillation algorithm, and the choice of a student and teacher architectures [55]. Below we explain these three components in further detail, followed with a review of related work in knowledge distillation.

Types of Knowledge

Generally, there are three types of knowledge that are distilled: Response-based knowledge, Feature-based knowledge, and Relation-based knowledge [55]. Response-based knowledge refers to the relationship between the last output layer of the teacher model and the inputs [55]. Here, the goal is to teach a student to behave like the teacher by mimicking the final predictions of the teacher model with the same inputs. As a training label, student models are trained on either the teacher's final prediction labels, known as a hard-target or on all the teachers logits (a distribution), known as a soft-target [23]. Choosing which type of target to distill is a commonly discussed design choice. Originally, Hinton et. al. [23] suggested that soft-targets contained more information, thus was best for distillation, but more recent evidence shows that when the student-teacher architecture is different, hard-targets are more effective [56]. In this work, we distill Response-based knowledge using hard-targets from a teacher language model to a smaller student language model.

Feature-based knowledge refers to the information stored in the intermediate activations of a teacher model [55]. Usually, this type of knowledge is combined with Response-based knowledge, and has been shown to give the student architecture more insight into to how the teacher reasons [57, 58].

Relation-based knowledge further explores the relationships between different layers or data samples [55]. In this type of knowledge, mutual relations are distilled between similar data examples, giving the student architecture an idea of how the teacher groups together data points. An example could be teaching a student computer vision model to represent dogs and cats in the same latent space which the teacher represents them.

Alongside the type of knowledge we wish to distill, it is also important to consider the scope of knowledge. The scope of knowledge defines the prediction task we wish to distill between architectures. Hinton et. al. [23] defined the knowledge scope as all knowledge of the teacher. However, more recent research defines a narrower *task-specific* distillation of knowledge [59, 46]. In this research we use task-specific knowledge distillation to transfer the prediction task of generating MI reflections from a teacher model to student models.

Distillation Algorithms

Distillation algorithms are the training techniques we use to distil our specified knowledge from a teacher to a student model [55]. Generally, we define three techniques: Offline distillation, Online distillation, and Self-distillation [55]. Offline distillation is the most standard method [23, 60], defined as a two stage process. First, the teacher is pre-trained before distillation, then used to distill knowledge to a student [55]. In this work, we make use of offline distillation, as our teacher model has already been pre-trained. Online distillation trains the teacher and student together, with hope that the teacher can guide the student during training and end up in a similar state [55]. Self-distillation uses a pre-trained teacher model as the student model. A pre-trained teacher model generates labels from curated inputs to generate a knowledge dataset. A student with the same architecture of the student is then trained on the curated dataset as an attempt to create a student with refined knowledge of the teacher [55].

Teacher Student Architecture

Knowledge distillation requires selection of a teacher-student architecture. A teacher model contains knowledge we wish to distill to a student model. Typically this decision is very objective-dependent, motivated by a teacher model which has desirable performance but is unfeasible for deployment. Since we wish to emulate the teacher, the student model is often a similar, more accessible version of the teacher. In this work, we distil GPT-4 [20] into GPT-2 [17] models, which share architectural similarities to GPT-4 while being much smaller.

Related Work on Language Model Distillation

There have been successful attempts to apply knowledge distillation to language models. Sanh et. al. [24] and Jiao et. al. [25] created DistilBERT and TinyBERT which showed that the same distilling techniques Hinton et. al. [23] used can be applied to distill transformer-based language models. DistilBERT [24] compresses the original BERT architecture by 40% while retaining 97% of performance on all the benchmarks BERT was tested on.

Alongside the above methods which aim to distill all knowledge of a teacher model, Task-specific knowledge distillation has also been explored for transformer-based language models [59, 61]. These works use the same distilling methods as [23] but differ in the amount and type of knowledge distilled. For example, as a task, Tang et. al. [59] use sentiment analysis and Liu et. al. [61] use the GLUE dataset benchmark task. Below are some relevant examples of task-specific knowledge distillation with related knowledge distillation techniques to this work.

He et al [62] show a method for task-specific knowledge distillation for the generation of synthetic text. Teacher models generate a dataset of prompt-completion pairs using both fine-tuned and few-shot learning. This dataset is then annotated for data quality and used for the training of smaller student models. The end result is a student model which learns to replicate the prompt-response behaviour of the teacher model, at a smaller, more manageable architecture size. He calls this method: “Generate, Annotate, and Learn”, an appropriately named framework distilling large language models through text.

Table 2.4: Self-Instruct Dataset Examples

Instruction	Input	Output
Given an address and city, come up with the zip code.	Address: 6 King’s College Road, City: Toronto	M5S 3H5
Which genre does this song belong to?	Song: Well, I was born one mornin’ when the sun didn’t shine. I picked up my shovel and I walked to the mine.	Country
Generate a response to a chat message using previous messages.	Hi, how are you?	I’m fine. How about you?

Self-instruct [63] is an application of self-distillation for GPT-3 [18]. Synthetic data is generated from GPT-3 and used to fine-tune itself, creating a version of GPT-3 which is task-oriented. First, GPT-3 is prompted with 175 seed instructions on many diverse tasks and instructed to create more instructions. Next, GPT-3 generates inputs for the instructions and the corresponding output. An example of instruction, input and output dataset entries is exemplified in Table 2.4. After fine-tuning GPT-3 on the synthetic dataset, performance rivals InstructGPT [64], a known language model which excels at instructional tasks.

Motivated by Self-instruct, Taori et. al. created Alpaca [65], a distilled instruction following LLaMA [66] language model relased by Meta. Alpaca replicates the self-instruct method but changes the student-teacher architecture. As a teacher, Alpaca also uses GPT-3 to generate a knowledge distillation dataset, but shrinks the student architecture to a LLaMA 7B, a compression of 25 times. Alpaca produces generations that rival its teacher model, showing that this method of knowledge distillation through synthetic text can be used to create models a fraction of the size with competitive performance.

2.3 Related Work on Motivational Interviewing Chatbots

There have been several attempts to automate an MI conversation using a chatbot across different domains, including stress management, sex health education, and smoking cessation [27, 28, 67, 68, 69]. Below we explain relevant examples with details about method and results.

Park et. al. [67] designed a conversational sequence utilising MI to aid in stress management. This conversation was deployed to 30 graduate students to compare its efficacy with human-to-human MI. It posed thought-provoking questions combined with scripted reflections. Participants reported they were satisfied with the evocative open-ended questions but were dissatisfied with the pre-written reflective statements.

Almusharraf et. al. [68, 69], the predecessor of the present work, designed and tested an MI chatbot used for motivating smoking cessation. The chatbot NLP classifiers to select scripted responses that guide a client through the conversation. Both the questions and reflections used in this chatbot were scripted. Almusharraf found that in a study of 97 participants [68], the average confidence to quit on the 11-point readiness ruler scale increased one week after the conversation by 0.8, ($P < .001$). The contents of Almusharraf's work heavily motivate the design of MIBot.

He et. al. [27] created both an MI and non-MI chatbot to investigate if chatbots can motivate smoking cessation. In an experiment with 153 participants, differences in motivation to quit smoking and perceived empathy were compared between a chatbot which utilised MI and one that did not. Both chatbots used evoking open-ended questions and pre-written statements for reflections. There were no significant differences between chatbots on engagement, therapeutic alliance, or perceived empathy. Notably, for both chatbots, participants reported significantly increased motivation to quit smoking.

The chatbots mentioned above responded with scripted statements based on keyword detection or neural net-based classification of the users' utterances. Scripted responses are often interpreted by humans as generic and repetitive [15], making it possible that the scripted reflections employed in prior work contributed to poor user satisfaction and perceived empathy. We hypothesize that a chatbot capable of delivering context-specific MI reflections will better motivate smokers to move towards the decision to quit.

Some work exploring the use of generated MI reflections within chatbots demonstrates this capacity. For example, Shen et. al.. [70] showed how transformer-based models can produce context-specific generative reflections. Although these reflections were used to train practitioners and were not patient facing, they highlight the capacity of Transformer-based models to produce good quality context-specific reflections. Similarly, Saiyed et. al. [28] created a Technology Assisted Motivational Interviewing (TAMI) chatbot for smoking cessation. The chatbot was designed to onboard participants, utilise MI, and refer participants to human-to-human treatment. It used intent classifiers and transformers to understand and generate utterances, including MI reflections. In a pilot trial of 34 smokers, participants reported the chatbot had a strong competency in MI, but only scored a 3/5 on user satisfaction, leaving room for improvement. Lastly, Ahmed et. al. [71], another predecessor of this work, investigated the use of Transformer-based Language Models to generate and classify MI reflections. With data curated by MI professionals, Ahmed found that GPT-2 and GPT-3 could generate acceptable simple and complex reflections. The present work builds off much of Ahmed's work.

Chapter 3

Chatbot Training and Study Method

The first goal of this work is to automate a MI conversation through a chatbot. In this chapter we describe the overall structure and implementation details of the chatbot experiment. We describe the four versions of the chatbot that are studied and provide the specifics of how the generative model is trained to provide MI-style reflections. Finally, we describe all the survey metrics used to measure effectiveness of the conversation.

The work in this chapter and the next describes the work of a number of people, including the present author, who contributed to the MIbot project. Section 4.9 provides attribution for the contributions.

3.1 Overall Chatbot System Study Design

Figure 3.1 illustrates the procedure used to evaluate each version of the chatbot, and some details of the steps are as follows:

1. Participants are recruited online through the Prolific [72] paid-recruitment system, the details of which are provided in Section 3.10. Each participant that is offered the opportunity to participate (by Prolific) is asked to review an informed consent document. The interaction proceeds if the participant provides consent.

2. Participants are taken to a custom website, are asked several questions, and then fill out three surveys on readiness to quit, heaviness of smoking, and number of quit attempts, as described in the Section 3.4.

3. The participant is then presented with a text chat window in which the chatbot, called 'MIBot' begins to interact with them. The conversation begins with an introductory section, and then the participant is asked if they wish to chat about smoking, as part of an MI-style permission-asking approach [6]. If the participant agrees, the conversation continues, or otherwise it terminates.

4. The core conversation, described in Section 3.2, ensues.

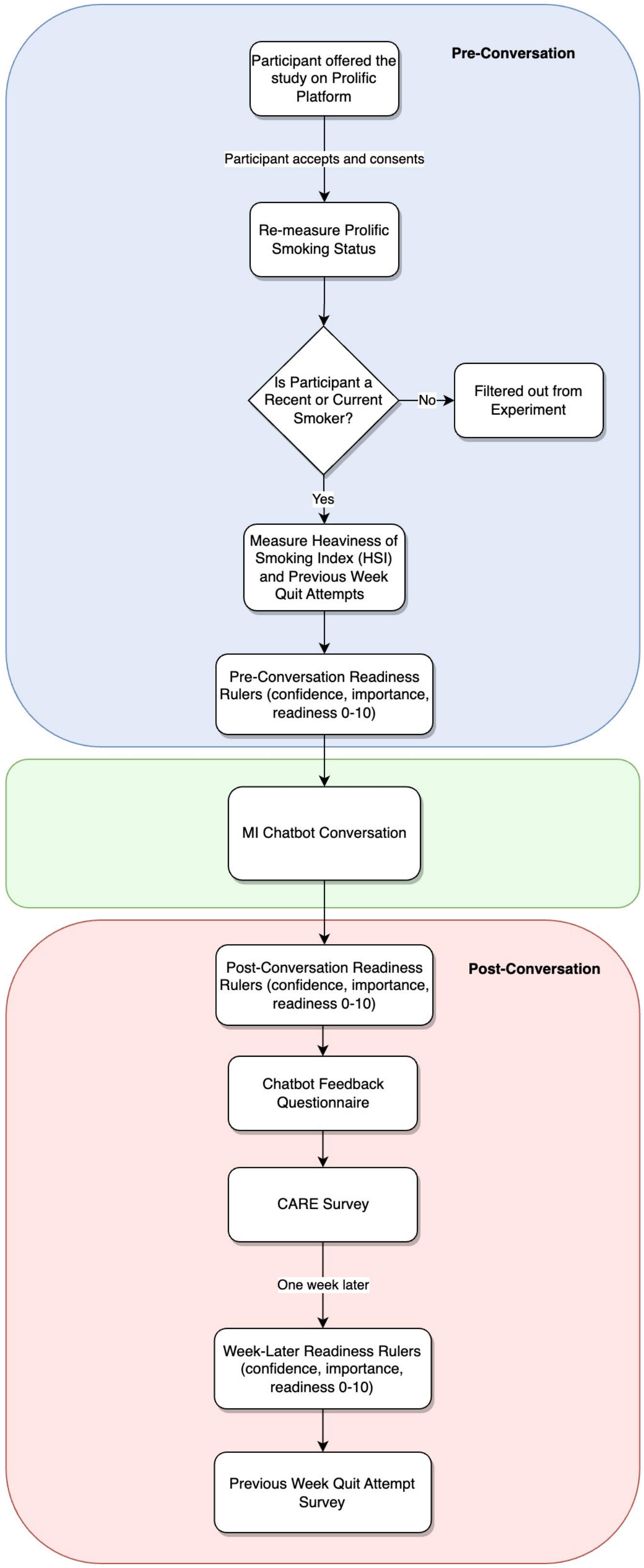

Figure 3.1: Overall Design of Each Study for Each Chatbot Version

5. After the conversation, the participant is asked to respond to another readiness to quit survey, the CARE measure and other qualitative questions that is described in Section 3.5.

6. Finally, the participant is directed back to the Prolific system so that the Prolific system can record the successful conclusion of the tasks.

7. One week after the end of the conversation, the participant is invited to a second task (also within the Prolific system), which is to answer the survey and questions described in Section 3.6. Participants are not paid unless they fill out the week-later survey and their submission is reviewed for data quality.

The sections below provide the details of the interaction and evaluation process.

3.2 Core Conversation

The core conversation consists of five open-ended questions that make use of the "running head start" [6] method of MI. The underlying theory of the running head start method is that ambivalent smokers spend very little time thinking about their smoking addiction, and do it by habit [73]. So, a key first step is to bring the habit to their attention and ask them to contemplate it. Here a clinician inquires what participants like about smoking, what they don't like, and then uses these reasons as a basis for further discussion and contemplation. This approach is realized through these five questions:

1. To start, what is the thing you like most about smoking?

2. Now, what is the thing you like least about smoking?

3. Now, what is one thing about your smoking habit that you would like to change?

4. What will it look like when you have made this change in your smoking habit?

5. Finally, what are the steps you need to take to make this change?

The first two questions are based on the running head start approach, and the subsequent three questions attempt to stimulate contemplation around the addiction.

One important aspect of this chatbot is that the participants respond with free-form text. That means they can provide any English textual response as compared to making a selection from scripted responses [14]. In allowing free-form responses, the conversation may more closely align with a human-clinical conversation.

In this work, we will describe and evaluate several versions of the conversation. For most of the versions, the chatbot generates an MI-style reflection (as described in the Section 2.1.1) of the free-form response to each question that is specific to the words of the response. Here we employ a fine-tuned Transformer-based neural network [17] to generate that reflection. The data and training of that neural network is described below in Section 3.8.

After each reflection is provided, the chatbot (in most versions) asks "Did that make sense?" If the participant responds in a way that indicates 'yes', the chatbot offers a thank you. If the participants' answer is equivalent to a 'no', the chatbot thanks the participant for helping it improve. Since the participant can write a free-form text answer, they may respond to the chatbot's question in many ways, for example, offering a longer, possibly corrective answer.

3.3 Chatbot Versions

The goal of the overall project is to continuously improve the system through iteration, and in this work, we report on the impact that four different versions have on readiness to quit in recruited participants. Each new version increases the complexity of the interaction. Table 3.1 provides a short description of each version. The differences between the versions are more readily seen in Appendix A, which provides one example of the full conversation for each version, taken from actual conversations with participants.

Table 3.1: Chatbot Versions

Version	Description/Differences	Date of Experiment
MI v4.7	Asks *just* the 5 questions shown in Section Core Conversation but does *not* provide reflections. (Instead, responds 'thank you')	July 26, 2022 – August 2, 2022
MI v5.0	Asks the 5 questions and provides MI-style reflective answers, as described in Section Core Conversation. It uses early version of the generator described in Section 3.8.	August 12, 2022 – August 19, 2022
MI v5.1	Same as v5.0, but uses the significantly improved Generator, described in Section 3.8	August 16, 2022 – August 23, 2022
MI v5.2	Based on v5.1, with extensions to the sequence of Questions 1 and 2. Also, if the answer to Question 3 relates to 'reduction' of smoking, changes Question 4 to be specifically about reduction. Also extends the interaction around question 3, as described in Section 3.3.1	November 22, 2022 – November 29, 2022

3.3.1 Enhanced Conversation Structure

Above we described the four versions of the chatbot, but version 5.2 requires more detail because of its increased complexity. The structure of the enhanced conversation follows:

- After the reflection is generated for Question 1 and Question 2, the chatbot asks what else the subject likes (or dislikes in the case of Question 2) about smoking. If the answer provides more reasons, these are also reflected. If there are no further reasons, then the chatbot moves onto the next question without generating a reflection or asking for validation of the previous reflections.

- If the participant response to Question 3 is related to the reduction of smoking – a common answer – then the chatbot switches to a different dialogue stream. The subsequent question becomes: "It's great to hear you want to reduce your smoking. What would it look like when you have reduced your smoking addiction?" After the reflection of the response to this question, the chatbot waits silently for 30 sections, to encourage the participant to respond on their own. For the minority of participants (12%) who did not respond by 30 seconds, the chatbot prompts a response by stating “Could you elaborate on what I said?”

We were motivated to design and add this enhanced conversation structure from two factors. First, our MI expert collaborators suggested after deployment of MIBot v5.0, we could invoke more contemplation through extension of the interaction. Furthermore, feedback from Prolific experiments suggested that a longer conversation would have led to a more fulfilling experience.

Thus, we looked for seamless methods of adding meaningful extension to the conversation. The addition of “What else do you like/dislike about smoking?” was added when an MI expert noted that this part of the conversation could be extended, and switching dialogue stream when users mentioned “reduction of smoking” was also mentioned by a MI expert.

3.4 Pre-Conversation Surveys and Screening

To confirm the participant’s smoking status and ensure this status had not changed since the administration of Prolific’s own screening survey, each participant is first asked to respond (again) to the same screening question administered by Prolific prior to interacting with the chatbot. Participants who made an inconsistent response with their prior Prolific-administered survey, indicating that they do not identify as smokers, were not allowed to proceed with the study.

Next, the participant is asked to respond to a Heaviness of Smoking Index (HSI) [74] and also to indicate how many quit attempts they had made the previous week, as shown in Figure 3.2. HSI and a count of quit attempts gives a baseline indication of the stage of addiction the smokers are currently in. Higher levels of HSI and lower quit attempts suggest the smoker has a higher level of addiction, thus is harder to motivate to quit. We also use the count of quit attempts in comparison with the count of quit attempts in a week later survey as explained in Section 3.5.

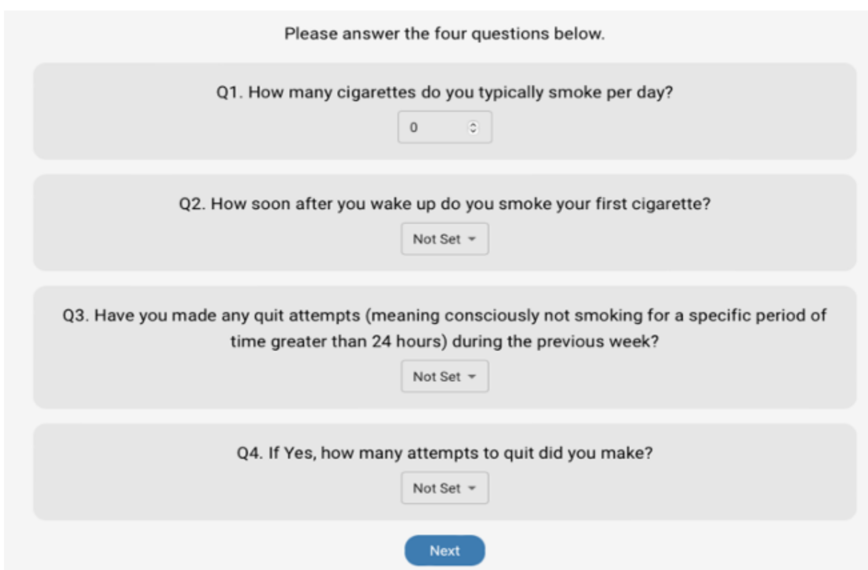

Figure 3.2: Screen that measures HSI and Quit Attempts

Finally, participants are asked to fill out the Readiness Ruler [31] survey, shown in Figure 3.3, which rates on an 11 point scale: their confidence that they could quit now (which will be referred to as confidence), their readiness to quit now (readiness), and how important they feel it is for them to quit smoking (importance).

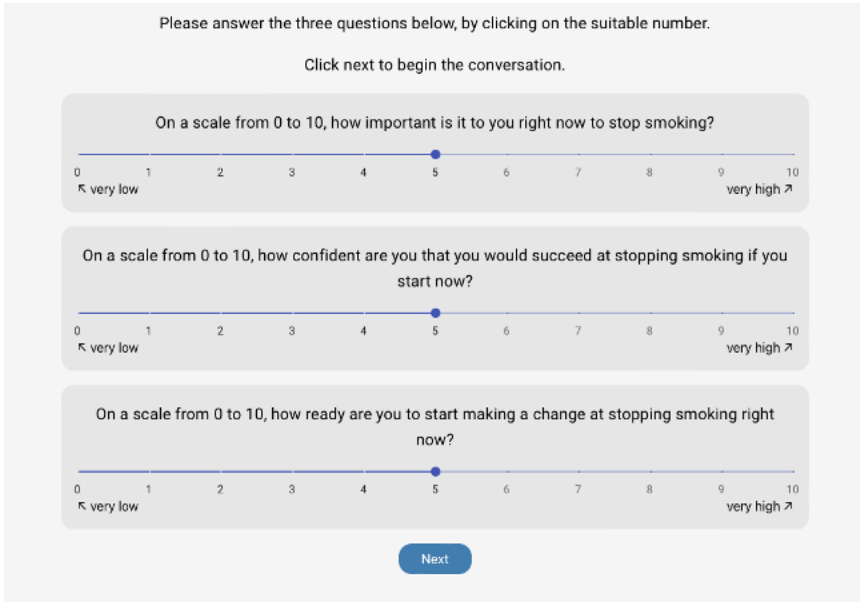

Figure 3.3: The Readiness Ruler

Our protocol imposes a second screen that is based on Readiness Ruler responses and the fact that an MI conversation is targeted towards ambivalent smokers: If the participant is already very confident that they will be able to quit, then they have already achieved the goal of this conversation, and so should not be part of the study. To achieve this, participants are only included if they have a confidence level less than or equal to 5, with one exception: if they have confidence greater than 5, but they also rate the *importance* more than 5 points below that confidence level, there is a contradiction that implies the presence of ambivalence. While these participants have confidence they could quit, but the fact that they don't think it is important means that they could benefit from a conversation that may raise the importance.

3.5 Post-Conversation Survey

After the conversation, the participant is asked to fill out the same readiness ruler survey [31] as prior to the conversation, and then to respond to the CARE survey [35]. The latter is a validated tool developed to assess empathy in a primary care patient-provider relationship. Empathy in the therapeutic encounter is linked with patient satisfaction and positive health outcomes [75]. The CARE survey examines empathy in the encounter by asking patients to rate provider (in this case the chatbot's) ability to a) appreciate their perspective, b) communicate back this understanding and c) given this understanding, be helpful to them. The CARE measure has 10 statements that are rated using a 6-point Likert-scale with total scores ranging from 0-50. The exact CARE survey we use in the chatbot conversation can be seen in Appendix C.

Finally, the participants are asked to respond to the following qualitative questions:

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

The purpose of these questions is to evoke feedback about the chatbot in different modalities. The first two questions ask for explicit feedback by asking how the user would describe the bot, then asking what they would change. Answers to these questions are considered for future improvement to the chatbot experience or simply to spot any technical difficulties. The last question asks the user for a qualitative sense of the impact the conversation had on their smoking addiction. We use the answer to this question in combination with other survey metrics to attempt to measure if users resolved any ambivalence in their smoking addiction. This is further explained in Section 3.7.

3.6 One-Week Follow-up Survey

Participants are contacted one week after engaging in the conversation, through the Prolific platform to do two more surveys: The first is a reprisal of the Readiness Ruler [31] to determine if the effect of the conversation on their responses to the ruler have changed. The second is three questions relating to quit attempts made during the preceding week. The first two questions are the same as the last two questions (Q3 and Q4) shown in Figure 3.2, and the third is given in Figure 3.4.

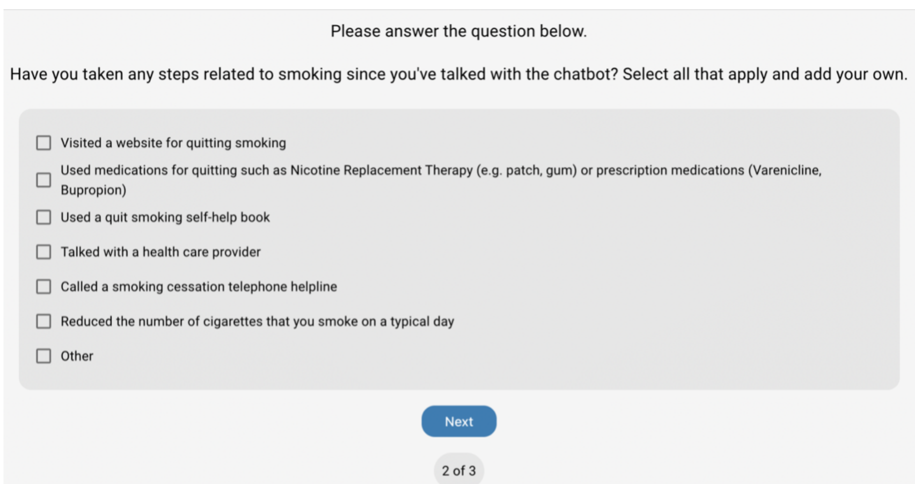

Figure 3.4: One Week Later Quit Smoking Actions

As mentioned in Section 3.4, we use this re-measurement of quit attempts as an indicator of impact the conversation had on the smoker. If the smoker had more quit attempts the following week after the conversation in comparison to the week before, there may be an indication that the chatbot inspired this change.

The question depicted in Figure 3.4 allows us to get more information into the steps that smokers are taking toward quitting smoking. This question came from a suggestion of a MI expert who noted that reduction of smoking often comes before quitting, and with many facets (hence the different options).

3.7 Classifying Resolved Ambivalence

An underlying goal of this work is to help smokers resolve their ambivalence towards smoking. It is possible to create a metric of ambivalence resolution which classifies the data from each participant as belonging to one of three outcome categories: the participant moved toward changing their smoking addition (which we will refer to as the *quit class*), they moved towards maintaining their smoking addiction (*smoke class*) or had no change (*same class*).

To place participants into one of these three classes, we make use of two outcome data: First we compute the pre-conversation to one-week later change in confidence from the readiness ruler, which can range in value from +10 to -6. A more positive value suggests a stronger move towards the quit class, whereas a negative value suggests a move towards the smoke class. To gain a more accurate signal, we combined this number with a subjective evaluation of the participant's answer to the third 1-week later question: "Did the conversation help you realize anything about your smoking behavior? Why or why not?" If the participant states that they realized something that helps them change their smoking addiction toward quitting, and had a positive change in confidence, then they are placed in the quit class. If the participant states that they realized they wish to sustain their smoking addiction and the confidence change was negative, then they are placed in smoke class. If these two specific signals are not both in the same direct, we place the participant into the same class. The coding of the qualitative response was done through human review.

3.8 Reflection Generation Training

One of the key contributions of this work is the novel way that MI-consistent reflections are generated in response to participant responses to the five questions shown in Section 3.2. Here we make use of recent advances in Natural Language Processing, and specifically in text generation, as described in the background chapter. The reflection generation neural network evolved from the one described in [76, 71]. It makes use of the pre-trained GPT-2 XL Transformer-based neural network model [17], which is fine-tuned, as described in Section 2.2.2. In this section, we provide more detail on how the generators used were trained.

There are two versions of the reflection generator that are evaluated in this work. The GPT-2-based generators are fine-tuned using example sequences of text, which we call a triplet, consisting of a question (from Section 3.2) , a response (from the participant), and a reflection (a known-good quality reflection).

In Version 1 of the generator, the fine-tuning question and response dataset came from two sources: the first was our prior work [68, 69], and the second was from earlier deployments of MIBot, prior to version 4.7. The reflections used came from a variety of sources – previous versions of this chatbot that were deemed to be acceptable MI reflections by MI-literate researchers, or actual reflections produced by MI-literate researchers or MI-expert clinicians.

We used the "hit rate" metric to evaluate the quality of a generator, which is the number of MI-consistent reflections generated, divided by the total number of reflections generated. Hit rate was measured on a validation set of question prompts and human responses that did not overlap with the training (fine-tuning) set. The MI-consistency of the reflection was judged by a single human rater, trained in MI literacy. The hit rate of the Version 1 generator was roughly 76% on a validation set of 33 prompt/responses. It is important to note that a hit rate less than 100% means that some fraction of the generated responses will not be consistent with MI and may indeed make counter-intuitive or simply wrong statements. In our experience, the most common type of error was a misstatement of the clear intent of the human. For example, when a user suggested that they would like to quit smoking, the chatbot would sometimes generate a reflection that implies the user would like to continue smoking.

To address the rate of poor reflections, we developed Version 2 of the generator with two significant enhancements: first, a larger set of 301 fine-tuning triplets were collected over roughly 10 months of deploying the chatbot, making use of the various responses from smokers that were recruited in a manner similar to that described in Section Recruitment. This second dataset did not include any of the data from the earlier chatbot [68, 69]. Only MI-consistent reflections were used, which were sourced from MI clinicians, MI-literate researchers, or the Version 1 generator. The labelling and selection of the MI-consistent reflections was improved by using multiple human raters, and a carefully controlled decision tree to determine validity of the reflections. The new rating scheme itself was stricter than the one used in Version 1, and so this caused the hit rate to go down – not because the generation was worse, but because of the stricter rating. The hit rate of the new generator was measured to be 55% on a set of 300 reflections.

The second enhancement was the implementation of a separate classifier neural network, trained to determine if a reflection is MI-consistent or not, given the triplet prompt, response, and reflection. The classifier was used to filter out poor reflections, and therefore increase the overall hit rate of generation. This makes use of the fact that all modern neural-network-style generators can easily generate many reflections, as the generation process is done through sampling from a probability distribution [77]. The classifier is referred to as the Reflection Quality Classifier (RQC), and an earlier version of it is described in [71, 76]. This RQC, based on the BERT [42] pre-trained neural network, was trained using a dataset of 740 examples, both positive and negative. Using our validation data, we achieved an accuracy of 70% on question, prompt, response triplets.

3.9 Software System

Figure 3.5 illustrates the structure of the software system used in the studies. Once the Prolific system transfers a participant to our system, they are brought to our chatbot frontend, which exists on a web page. That web page connects to back-end database (based on Google Firebase [78]) that records the entire conversation, and all data associated with the surveys and information the participant provides. It also connects to the chatbot backend. The chatbot backend exists as multiple servers provisioned by Amazon Web Services Elastic Compute 2 [79]). For all versions of the chatbot, we provision three g4dn.4xlarge type servers each equipped with a NVIDIA T4 GPU (16GB of GPU memory which is enough to inference GPT-2 XL). The frontend routes calls to the backend through a load balancer which distributes calls evenly to the three g4dn.4xlarge servers. This distributes the load of reflection generation among three servers instead of one, speeding up the service time of the chatbot conversation and preventing timeouts.

Each backend is split into a Dialog Management Engine and Dialogue Generation Engine. The Dialogue Management Engine, uses a Yes or No classifier, Content or No classifier, and an Intent Classifier to classify incoming user utterances using Natural Language Understanding (NLU) techniques and controls the current state and direction of the of conversation. The Yes or No classifier and Content or No Classifier are both rule based, meaning they use hard-coded rules to make a decision on the input. The intent classifier used is off-the-shelf from a third party service called Wit.ai (an online NLU service) [80]. Responses are constructed using a combination of the Dialogue Generation Engine and Response and Question Database. The Dialogue Generation Engine uses the GPT-2-XL neural network to generate MI reflections as described in Section 3.8.

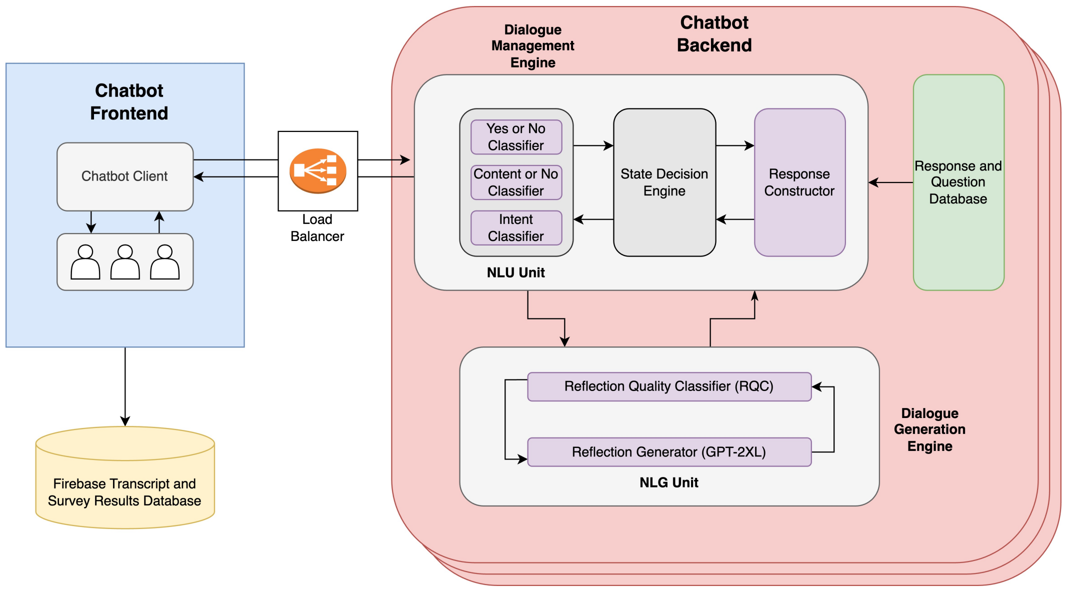

Figure 3.5: Chatbot Architecture

3.10 Recruitment and Data Inclusion

Participants were recruited through the Prolific [72] online recruitment system, and were paid a total of 5 British pounds for completing two tasks one week apart. The inclusion criteria based on Prolific's screening filters was:

- Can be located in any country
- Minimum age 18
- Fluent in English
- Smoking Status is one of these two choices:
  - "I am a current smoker (smoke at least 5 cigarettes a day and have smoked this amount for at least one year)"
  - "I am a recent smoker (smoke at least 5 cigarettes a day and have smoked this amount for less than one year)"
- Minimum approval rate on participant's prior Prolific studies of 90%

In addition, the Prolific system was set to recruit an equal number of men and women. However, since there is a subsequent screening of participants as described in Section 3.4, the final number of participants is not perfectly balanced between men and women, but is close.

All participants provided consent to the study participation. This research was approved by the University of Toronto Research Ethics board under Protocol # 35567, as amended, on June 29, 2022.

Data from the participants who completed part one and two of the study were manually reviewed for data inclusion on the following criteria:

1. The participant properly filled out each survey metric with realistic values.
2. The participant responded to the chatbot with apparent honesty (e.g., no toxic language or apparent ulterior motives).
3. The participant met the additional screening criteria as described in Section 3.4.

3.11 Statistical Analysis

Significance testing was completed within and across chatbot versions. Within each chatbot version, we compared readiness ruler responses, quit attempts, and ambivalence resolution counts before and after the conversation. For readiness rulers, a two-tailed t-test was applied to examine changes in readiness ruler attributes (i.e., readiness, confidence, importance) pre-conversation to one week later. A Fisher's exact test was used to evaluate significant changes in pre-conversation quit attempts compared to one week later. We also compared changes in readiness ruler attributes, CARE survey, and reduction of smoking across chatbot versions. To compare readiness rulers and CARE survey data, a Welch's t-test was used. For reduction of smoking, we used a two-sample proportion test (z-test)

For all tests, a significance level of $P < .05$ was considered statistically significant. Statistical analysis was completed using the SciPy library for the Python programming language [81].

Chapter 4

Chatbot Results and Discussion

This chapter reports the results of the interaction of the participants with the four versions of the chatbot described in Chapter 3, together with a discussion. We begin with the recruitment yield and data inclusion and then provide the demographics of participants and heaviness of smoking index (HSI). Then, we present the readiness rulers, quit attempts, CARE survey, and ambivalence resolution counts. Finally, a discussion section shows our principle findings and other relevant observations.

This chapter concludes with a section describing the contributions of the author and other contributors to this work.

4.1 Recruitment Results

Figure 4.1 depicts study procedures (also described in Section 3.1), showing the points at which participants enter and (may) exit the study. Table 4.1 gives the specific numbers of exit and entry for each version of the Chatbot that was deployed. Of the total 517 participants who completed part one and two of the study, 168 were filtered out using the Pre-conversation Survey criteria described in Section 3.4, across all four chatbot versions, and so a total of 349 participants were included in the analysis.

Table 4.1: Participant Count for each Chatbot Version Corresponding to Study Flowchart

Exp/Version	AC	WD	NC	CF	NWL	CWL	FSS	N
MI v4.7	119	17	9	93	0	93	41	52
MI v5.0	171	23	3	145	4	141	43	98
MI v5.1	169	24	4	141	1	140	41	99
MI v5.2	195	36	6	153	10	143	43	100

4.2 Demographics

Table 4.2 provides participants' demographic data, collected by Prolific when participants enroll on the platform. Prolific allows for these data to be revoked or changed (e.g., see attribute "Student Status" in Table 4.2).

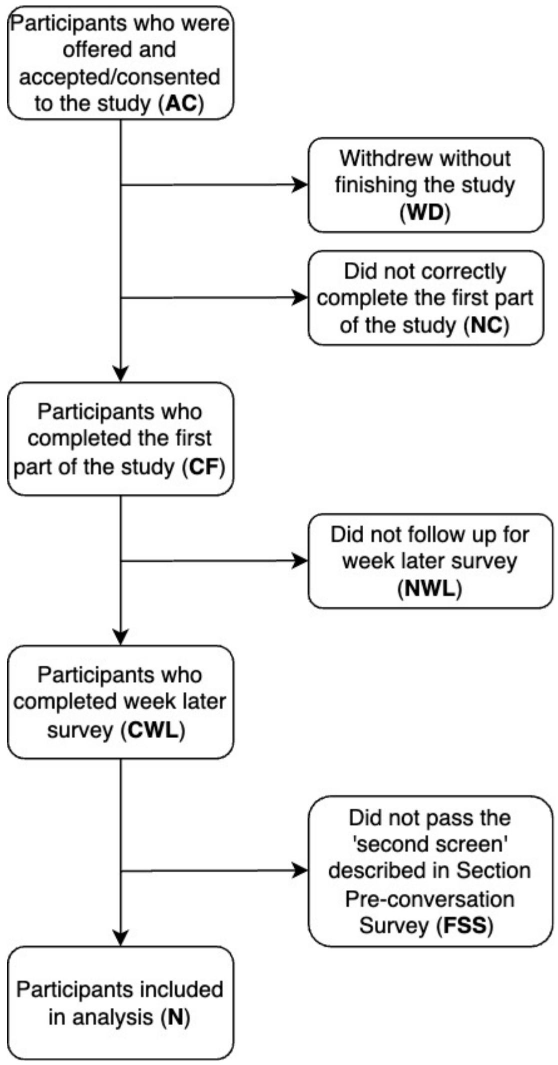

Figure 4.1: Flowchart of Study Procedure

4.3 Heaviness of Smoking

Table 4.3 provides the data from the Heaviness of Smoking Survey that was applied to the most recent versions of the chatbot, not including chatbot version 4.7, for which this data was not collected.

4.4 Readiness Rulers

Participant responses to readiness rulers were collected before, immediately after, and one week after the chatbot conversation, as described in the Sections Methods, Pre-conversation Surveys, Post-Conversation Survey, and One-Week Follow up Survey.

Table 4.4 provides mean and standard deviation (SD) of confidence to quit smoking for each chatbot version pre- and post-conversation and one week later, as well as the average change after one week and its statistical significance.

Table 4.5 shows the participants' mean value of the importance to quit smoking across each

Table 4.2: Demographics of Participants

Demographic Factors Count, n (%)	MIV4.7	MIV5.0	MIV5.1	MIV5.2
Sex				
Female	30 (57.7)	46 (47)	54 (54.5)	51 (51)
Male	22 (42.3)	52 (53)	45 (45.5)	49 (49)
Age				
18 to 19	2 (3.8)	5 (5.1)	1 (1)	2 (2)
20 to 29	35 (67.3)	58 (59.2)	50 (50.5)	58 (58)
30 to 39	8 (15.5)	16 (16.3)	22 (22.2)	26 (26)
40 to 49	5 (9.6)	14 (14.3)	12 (12.1)	10 (10)
50 to 59	2 (3.8)	5 (5.1)	11 (11.1)	3 (3)
≥ 60	0 (0)	0	3 (3.1)	1 (1)
Student Status				
Yes	27 (51.9)	45 (45.9)	32 (32.3)	48 (48)
No	25 (48.1)	52 (53.1)	59 (59.6)	45 (45)
Data Revoked	0 (0)	1 (1)	8 (8.1)	7 (7)
Employment Status				
Full-Time	23 (44.2)	39 (39.8)	42 (45.2)	41 (41)
Part-Time	10 (19.2)	22 (22.5)	16 (17.2)	19 (19)
Unemployed (and job seeking)	9 (17.4)	19 (19.4)	16 (17.2)	21 (21)
Not Paid in Work	5 (9.6)	6 (6.1)	12 (12.9)	0 (0)
Other	5 (9.6)	12 (12.2)	7 (7.5)	19 (19)
Average Total Approvals (for all participant studies)	233.8	262.5	360.8	254.8
Average Approval Rate (for all participant studies)	99%	99%	99%	$\sim$

experiment at each collection time, with the same format as Table 4.4.

Table 4.6 gives the participants' mean value of the readiness to quit smoking across each experiment at each collection time, with the same format as Table 4.4.

Figure 4.2 provides the distribution of confidence scores in each of the four versions of the chatbot for each of the three collection points.

Figure 4.3 provides the distribution of importance scores in each of the four versions of the chatbot for each of the three collection points.

Figure 4.4 provides the distribution of readiness scores in each of the four versions of the chatbot for each of the three collection points.

Table 4.7 presents the number of participants whose confidence increased from prior to the conversation to the week later, as well decreased, and stayed the same.

4.5 Quit Attempts and Reduction of Smoking

Table 4.8 provides the percentage of participants, for each version, who made at least one quit attempt (defined as going 24 hours without smoking a cigarette) in the week prior to engaging in

Table 4.3: Heaviness of Smoking Measures

Exp/Version	Mean Daily Num Cigarettes	Time to First Cigarette	Mean HSI
MI v5.0	11.9 (9.3)	Within 5 minutes: 26 6 to 30 minutes: 15 31 to 60 minutes: 31 More than 60 minutes: 26	1.8 (1.5)
MI v5.1	11.1 (7.9)	Within 5 minutes: 26 6 to 30 minutes: 27 31 to 60 minutes: 29 More than 60 minutes: 17	1.8 (1.5)
MI v5.2	9.9 (6.1)	Within 5 minutes: 15 6 to 30 minutes: 32 31 to 60 minutes: 17 More than 60 minutes: 35	1.6 (1.4)

Table 4.4: Average Confidence, Before, After and 1 Week After Conversation

Exp/Version	Pre-Conv Avg (SD)	Post-Conv Avg (SD)	1 Week Later Avg (SD)	Change from Pre to Week Later Avg (SD)	Pre to Week Later *P* value from paired t-test
MI v4.7	3.6 (2.2)	4.5 (2.4)	4.7 (2.6)	1.0 (2.0)	0.0001
MI v5.0	3.5 (2.7)	4.1 (2.9)	4.7 (2.9)	1.2 (2.0)	<.001
MI v5.1	3.2 (2.2)	3.9 (2.1)	4.4 (2.4)	1.3 (2.3)	<.001
MI v5.2	3.3 (2.3)	4.1 (2.5)	4.7 (2.7)	1.3 (2.0)	<.001

Table 4.5: Average Importance, Before, After and 1 Week After Conversation

Exp/Version	Pre-Import Avg (SD)	Post-Import Avg (SD)	1 Week Later Avg (SD)	Change from Pre to Week Later Avg (SD)	Pre to Week Later *P* value from paired t-test
MI v4.7	5.1 (3.1)	5.5 (3.1)	5.3 (3.1)	0.3 (1.6)	0.41
MI v5.0	5.2 (3.0)	5.7 (3.0)	5.6 (2.8)	0.4 (1.5)	0.033
MI v5.1	5.2 (2.8)	5.7 (2.8)	5.5 (2.9)	0.3 (1.3)	0.17

Table 4.6: Average Readiness, Before, After and 1 Week After Conversation

Exp/Version	Pre-Readi Avg (SD)	Post-Readi Avg (SD)	1 Week Later Avg (SD)	Change from Pre to Week Later Avg (SD)	Pre to Week Later *P* value from paired t-test
MI v4.7	4.3 (2.7)	4.6 (2.6)	4.8 (2.8)	0.4 (1.5)	0.085
MI v5.0	4.3 (2.7)	4.4 (2.8)	4.4 (2.7)	0.1 (1.8)	0.75
MI v5.1	4.4 (2.4)	4.6 (2.4)	4.6 (2.6)	0.2 (1.5)	0.14
MI v5.2	4.9 (2.8)	5.3 (2.7)	5.4 (2.9)	0.4 (1.7)	<.05

the conversation, and in the week after the conversation. Note that our evaluation of chatbot version 4.7 did not include a survey for the number of quit attempts prior to the conversation.

Table 4.9 shows the count of participants who did and did not reduce the number of cigarettes they smoke after talking to the chatbot. The binary result of 'reduce'/ 'did not reduce' was determined by setting the result to 'reduce' if any one of the conditions in Figure 3.4 was selected. Rows two through five show the count for each experiment respectively.

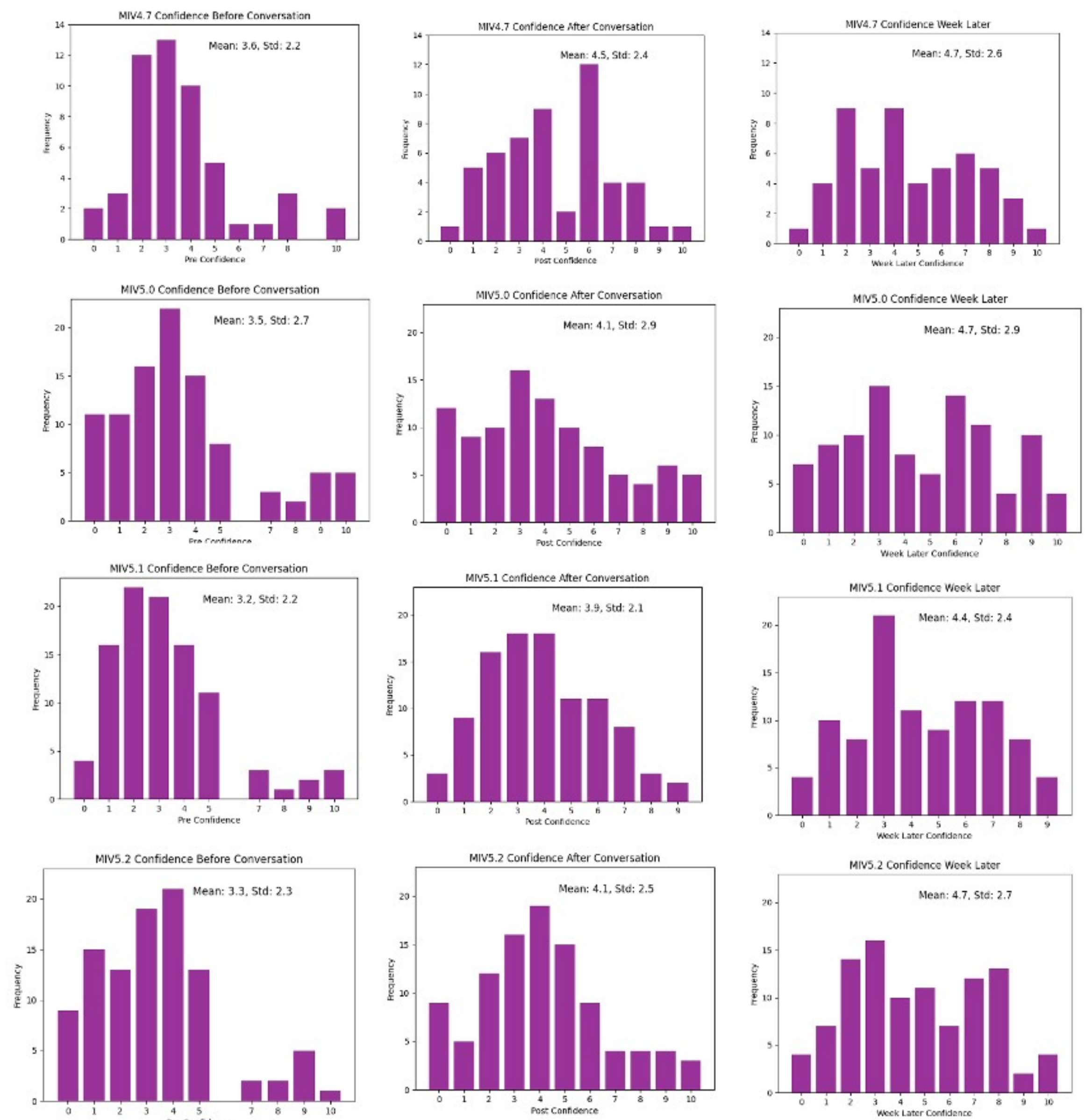

Figure 4.2: Distribution of all 4 versions' Confidence Values Pre, Post, and Week Later

Table 4.7: Number of Participants Whose Confidence who Increased, Decreased, or Stayed the Same 1 Week Later

Exp/Version Count, n (%)	Confidence Increased	Confidence Decreased	Confidence Stayed Same	Total
MI v4.7	31 (59.6)	8 (15.4)	13 (25)	52
MI v5.0	52 (53)	12 (12.3)	34 (34.7)	98
MI v5.1	50 (50.5)	21 (21.2)	28 (28.3)	99
MI v5.2	61 (61)	16 (16)	23 (23)	100

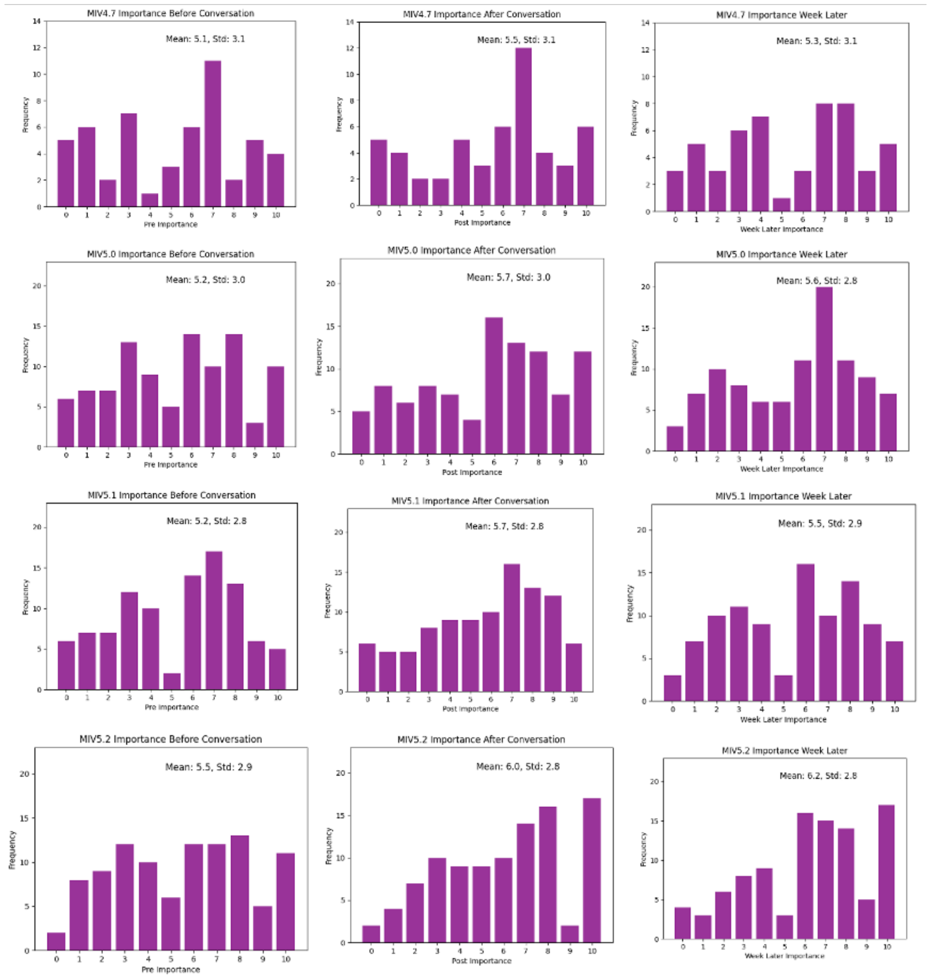

Figure 4.3: Distribution of all 4 versions' Importance Values Pre, Post, and Week Later

Table 4.8: Percentage of Participants Who Made Quit Attempts Before and After Conversation

Exp/Version	% of users with pre-conv Quit Attempt	% of users with after-conv Quit Attempt	P value from Fisher's Exact Test
MI v4.7	N/A	35%	N/A
MI v5.0	39%	34%	0.55
MI v5.1	26%	25%	1
MI v5.2	40%	38%	0.88

4.6 Consultation and Relational Empathy (CARE) Measure

Table 4.10 shows the count of participants who reduced the number of cigarettes they smoke after talking to the chatbot. Rows two through five show the count for each experiment respectively.

Figure 4.5 shows a histogram for each average CARE score distribution. Within each plot, the mean, median, and standard deviation are given.

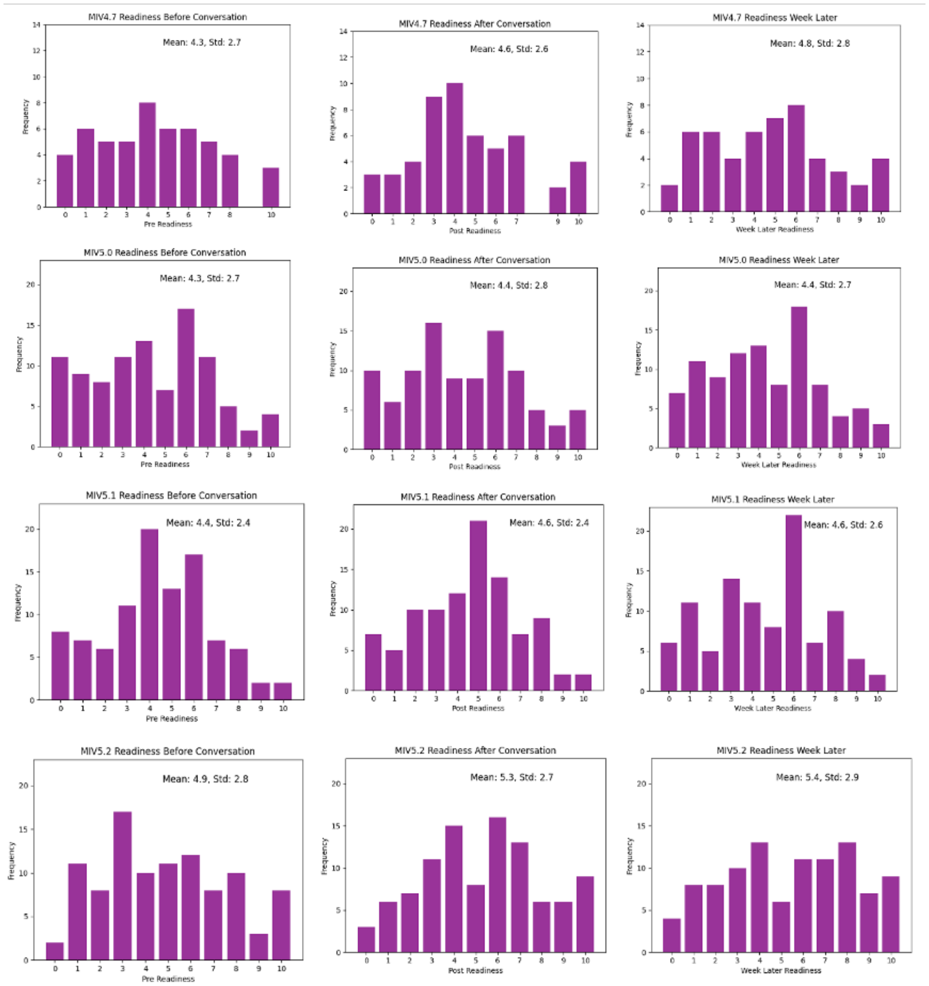

Figure 4.4: Distribution of all 4 versions' Readiness Values Pre, Post, and Week Later

Table 4.9: Count of Participants who Reduced/Did Not Reduce Smoking

Exp/Version Count, n (%)	Reduced smoking after talking to chatbot	Did not reduce Total smoking after talking to chatbot	Total	*P* Value from two-sample proportion test, z-test against MI v4.7
MI v4.7	37 (71.2%)	15 (28.8%)	52	N/A
MI v5.0	68 (69.4%)	30 (30.6%)	98	0.82
MI v5.1	67 (67.7%)	32 (32.3%)	99	0.66
MI v5.2	74 (74%)	26 (26%)	100	0.7

4.7 Did Ambivalence Change and in What Direction?

Table 4.11 presents the counts of participants who were classified as moving in the direction toward quitting, toward smoking and staying the same as described in Section 3.7. Across all 4 experiments, none of the values in each class were statistically significant. Appendix B provides the classification

Table 4.10: CARE Empathy Measure

Exp Version	Mean (SD)	P Value test against MI v4.7
MI v4.7	31.5 (9.6)	-
MI v5.0	33.1 (9.1)	0.24
MI v5.1	35.3 (9.4)	0.02
MI v5.2	36.2 (9.1)	0.004

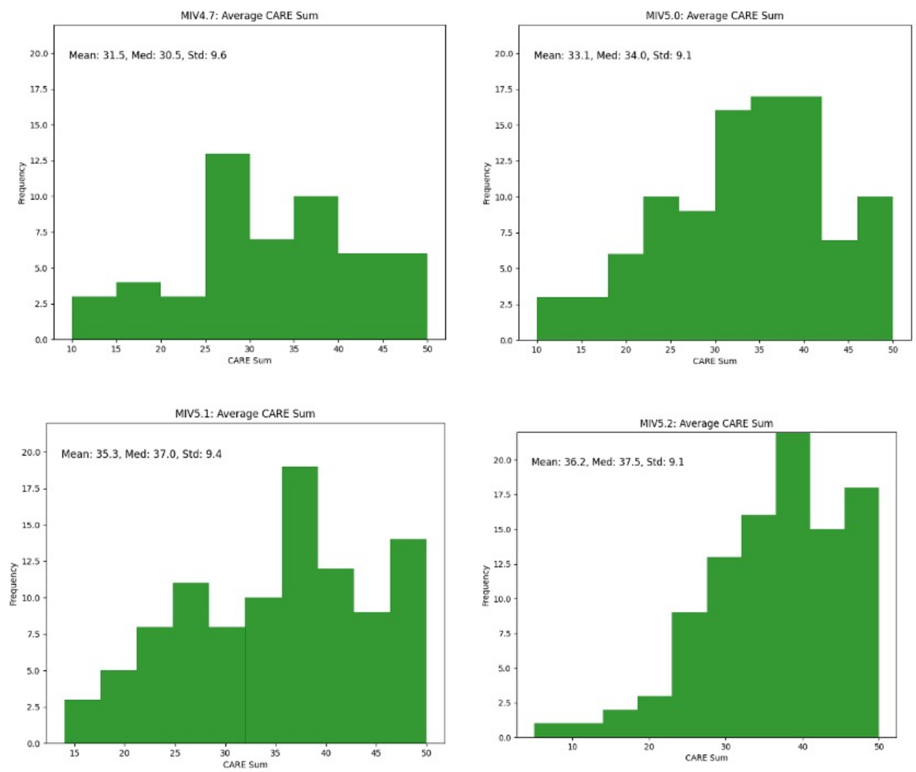

Figure 4.5: CARE Survey Distribution

for each participant and the raw data the classification was based upon.

Table 4.11: Counts of Quit, Smoke and Same Ambivalence classes

Exp/Version	Quit Class N, (%)	Smoke Class N, (%)	Same Class N, (%)	Total
MI v4.7	17 (32.7)	2 (3.8)	33 (63.5)	52
MI v5.0	26 (27)	1 (1)	71 (72)	98
MI v5.1	20 (20.2)	4 (4)	75 (75.8)	99
MI v5.2	30 (30)	6 (6)	64 (64)	100

4.8 Discussion

4.8.1 Principle Findings

The long-term goal of this work is to evolve a chatbot to have impact on a smoker's readiness to quit, with a focus in this study on whether generative reflections can improve chatbot efficacy. It is important to note that, although the readiness ruler has three attributes - confidence, importance, and readiness – it is the confidence measure which most successfully predicts quitting success. It has been shown that more confident someone is, the more likely they are to make a quit attempt and succeed [32, 33, 34]. Table 4.4 shows that all four versions of the conversation achieved a statistically significant improvement in confidence one week after the conversation takes place. The average increase in confidence ranges from 1 to 1.3 on the 11-point scale. This finding is consistent with He et al. [27] who found that a short chatbot intervention about smoking cessation can significantly impact quitting intentions and behaviours. Also, while the average increase is greater for the later versions (v5.0 through v5.2), these are not statistically significant changes between versions of the chatbot (P = 0.43 for v5.2 vs. 4.7, for example).

Although we hypothesized that generative responses that are specific to what a smoker says would lead to better outcomes, this result suggests that simply asking questions is sufficient to evoke most of the impact on confidence that we observed. However, there is other evidence to suggest that the improvements to the conversation beyond v4.7 (i.e., generative reflections and extended dialogue) have a positive impact on participants' readiness to quit, with respect to increases in importance and readiness: Tables 4.5 and 4.5, show that v5.2 is the only version associated with significant increases in these two attributes. In addition, the perceived empathy of the MIBot v5.2 is significantly higher compared to version 4.7 (see Table 10). This makes intuitive sense because a response that addresses what a person says would likely be perceived as more empathetic compared to a response that only says, 'Thank you for answering.' Our result contrasts He et al.'s finding of no significant difference in perceived empathy between a chatbot performing MI and one that does not [27]. It is possible that our use of generative reflections (vs. the scripted reflections and responses of [27]) are the cause of the difference.

A study by Bikker et al contrasts these CARE scores with those achieved by human practitioners. It shows that nurses receive high scores on the CARE survey (i.e., 46, with 48% of nurses achieving a perfect score) [82], which is much higher than the score achieved by version v5.2 (i.e., 36, with only 3% of interactions receiving a perfect score). So, although much of the benefits for confidence in quit readiness can be attributed to simply asking MI questions, there may be other benefits from producing generative reflections on the importance and readiness metrics, as well as perceived empathy of the chatbot. These findings are encouraging, and support further evolution of this capability.

4.8.2 Recruitment and Demographics

The demographic characteristics of participants in our study notably differ from participants in prior MI intervention studies in two ways: first, the mean age of our cohort was 30, and is somewhat lower than that of prior studies of human to human MI interventions which were roughly 35 [7]. Second, we recruited a balanced sample of men and women, whereas in many MI studies, approximately 68%

of participants tend to be women [7]. Third, based on the HSI (Table 4.3), participants in our study tend to consume fewer cigarettes (i.e., a mean of 10.8 daily) than participants in studies of human to human MI interventions (i.e., 16 on average) [7]. These findings suggest that the participants in our study are younger and overall lighter smokers than in the typical MI studies.

Figure 4.1 and Table 4.1 show the number entering each study and the quantity of exits from the study. The FSS column of Table 4.1 shows that roughly a third (34%) did not meet the secondary screening criteria, which is that they are not already confident that they can quit smoking and thought that doing so was important. Carpenter et al show that, globally, 20% of smokers are in a similar state, already motivated to quit [83]. It may be that the younger demographic of the present study account for this difference.

4.8.3 Quit Attempts and Reduction of Smoking

The number of quit attempts related to interacting with each version of the chatbot did not significantly change from the week before to the week after (see Table 4.8). However, the percentages of participants who attempted to change across all versions is in the 30-35% range, which is much higher than the 11% that is reported occurring 4-8 weeks after human MI interventions [7]. This difference in quit attempts may be related to the demographic differences we observed in our sample, as compared to other MI studies. We speculate that the groups in the present studies were more likely to make quit attempts because they are a younger and less addicted population, as discussed in Section 4.8.2.

Across all conversations, Table 4.9 shows that a large fraction of the participants did make some kind of smoking reduction attempt – meaning that they clicked one of the boxes in Figure 3.4. However, the differences in percentages are not significant between the different groups/chatbot versions.

4.8.4 Resolution of Ambivalence

We employed an alternative measure of the chatbot's impact by classifying participants based on their ambivalence status, as moving towards quitting, towards smoking, or staying the same. There was no significant difference between chatbot versions in the percentage of participants belonging to each category, as shown in Table 4.11.

It is possible that the participants who resolved towards quitting were just ready to do that, and were going to do it anyway, or the conversation was just the push that they needed to go there.

It is important to consider the possibility that the roughly 2 to 5% of the participants who resolved to continue smoking were hurt by the interaction with the conversation. We manually reviewed each of these conversations, and for roughly 85% of the conversations, we did not see evidence of harmful statements made by the chatbot that could have contributed to this resolution. For the other 15% of conversations, the chatbot produced poor reflections, which may have caused participants to be less likely to quit or believe they had less of a chance to do so. For example, in response to a participant expressing the idea that 'cold turkey' quitting is their best approach to quitting, the bot responded 'A smoker can't really do that,' which is quite inappropriate.

4.9 Chatbot Contribution Attribution

The work presented in Chapters 3 and 4 represent is part of a long-term project with quite a few collaborators making contributions. In this section we describe the various contributions and attribute these to specific people, as well as the present author. Below are the contributions ordered alphabetically by last name, followed by the contributions of the author.

- Imtihan Ahmed - built a MI reflection generation micro-service [71], a project which served as part of the foundation for this chatbot. This was the first attempt at hosting an online application which gave generated MI reflections to users. Ahmed created a Javascript front-end and Python back-end which were both hosted on AWS [79]. He also created the initial versions of the Reflection Quality Classifer, described in Section 3.8.
- Arnaud Deza - created the first version of the back-end as described in Section 3.9, including the basic finite state machine. This back-end utilized the same five questions as explained in Section 3.2, generated MI reflections like Section 5.1.
- Tanuj (Ash) Kumar - created software to extract data from the Firebase data storage system. This extractor software was used many times for data collection during analysis. Kumar also trained a reflection generator which was used for chatbot version MI v5.0
- Marc Morcos - created the first version of the front-end as described in Section 3.9. This front-end used the readiness rulers for pre- and post-conversation measurement as explained in Section 3.4 and 3.5. He worked with Deza to bring up the very first version of the bot.
- Angus Wang - created data and trained reflection generation models as explained in Section 3.8. Chatbot versions MI v5.1 and MI v5.2 used reflection generators trained by Wang. He also trained the version of the RQC that was used in these versions.
- Leon Zhu - built additional survey components on the chatbot front-end after Marcos. The one-week follow-up survey was built by Zhu, as explained in Section 3.6

The author of this work is the main contributor and organizer of each of the chatbot experiments, from version 4.7 through 5.2. This includes the multi-week process of recruitment, week-later recruitment and vetting of the responses for inclusion. The author was took the raw material of development work of Ahmed, Deza, Morcos, and Kumar and built it into a functioning scientific ecosystem. He re-implemented a new version of the load-sharing system that used multiple servers to serve the large language models that generated the reflections. Work included making developments individually and supervising Wang and Zhu's contributions. The author added additional functionality such as survey metrics as explained in Section 3.4 and Section 3.5, conversation improvements as explained in Section 3.3.1, which included a complete re-design and implementation of the Finite State Machine infrastructure. He improved MI reflection generation as shown Section 3.8. Furthermore, all deployment infrastructure was designed and created by the author as explained in Section 3.1, and all experiment deployments on Prolific [72] were overseen and managed by the author. The proceeding data analysis and discussion were completed by the author. This included the manual analysis required to do the ambivalence analysis. This work was published here [84].

Chapter 5

Methods for Generating and Evaluating Reflections and Distillation

The previous chapters showed some benefit of using a large language model as part of a therapeutic chatbot, but currently the best transformer-based language models are proprietary and so large that it is expensive and difficult to serve them, and they don't come with sufficient privacy guarantees. In the next two chapters we explore the use of the current very-best transformer-based language model (GPT-4) [20] for generating MI reflections, evaluating reflections, and then distilling that capability into smaller student models. We believe that this work will be of general interest to others who would like the capabilities of the larger model for specific tasks, but also suffer from downsides of large proprietary language models.

In this chapter we describe the method we use to generate reflections with GPT-4 and then how we also use GPT-4 to evaluate MI reflections, for use in evaluating the quality of smaller distilled models. We also describe other evaluation techniques we use to validate GPT-4's evaluation and the distilled models performance. Next, we describe our knowledge distillation process from GPT-4 into a range of smaller pre-trained language models. Finally, we describe the details of our experimental setup.

Similarly to Chapter 3 and 4, we conclude Chapter 5 and 6 with a with a description of contributions attribution in Section 6.5.

5.1 Generating Reflections from GPT-4

This section describes our method in utilising GPT-4 to generate simple and complex MI reflections. The MIBot chatbot described in Chapters 3 and 4 generated reflections using a fine-tuned version of GPT-2 [17], a model which is no longer considered state-of-the-art, having been surpassed by GPT-3 [18] and GPT-4 [20]. Thus, we were motivated to investigate reflection generation using GPT-4, the most advanced language model to this date which only requires few-shot [21] or zero-shot learning [22]. If the newer model can produce improved simple and complex reflections it could have a more beneficial clinical impact than the ones generated from GPT-2 in MIBot. Furthermore, if the reflections are better, those generated reflections can be used as part of a model distillation process.

This section is structured as follows: First we explain how OpenAI instructs users to do zero-shot prompting with GPT-4. Next, we describe how we use the transcripts from the chatbot deployed in Chapter 3 and 4 as part of the inputs to GPT-4 for reflection generation. Finally, we describe the prompts we create to generate simple and complex reflections and show how we format a complete example.

5.1.1 GPT-4 Prompting Schema and Formatting

To generate any output from GPT-4, OpenAI uses a specific API message formatting protocol known as chat-complete [20]. This protocol divides the prompting text into two blocks of information [20]. We note that OpenAI has not explained the structure or reasoning behind this separation, but gives guidance on how to use each block. This information is relevant to our use of GPT-4 reflection generation and also evaluation of reflections.

The input format divides the prompt into a *System Role*, and a *User Message*. The *System Role* should contain the instruction of the task containing information about how to produce an output given an input. The *User Message* are inputs coming from the User, to be processed by the model, which in our case will be utterances coming from a client. The API also produces a *System Message* which is an output coming from GPT-4. In our case, these are statements made by the chatbot acting as a therapist.

Below we describe the MI transcripts from which we extract data for developing the zero-shot prompt for reflection generation in Section 5.1.2, then Section 5.1.3 describes the development of the prompt itself.

Context
Bot: (QUESTION) To start, what is the thing you like most about smoking?
Client: (ANSWER) Stress relief.
Bot: (REFLECTION) You enjoy smoking because it helps you cope with stressful situations.
Bot: Did that make sense?
Client: Yes.
Bot: That's great to hear, thanks for letting me know!
Bot: (QUESTION) Now, what is the thing you like least about smoking?
Client: (ANSWER) I spend a lot of money on cigarettes.
Bot: (REFLECTION) You dislike spending money on cigarettes.
· · · (more turns)

Table 5.1: MI Chatbot Transcript Excerpt

5.1.2 Gathering Motivational Interviewing Transcripts for Dataset Creation

An essential task in prompting or tuning a model to generate MI reflections is to source realistic questions and answers as input to the reflection model. For this purpose we used the 349 transcripts from the deployed chatbot in Chapter 3 and 4. Table 5.1 shows an excerpt of a conversation transcript. The MIBot chatbot adopts a pattern of: asking open-ended questions (QUESTION), retrieving answers (ANSWER), and generating reflections (REFLECTION). These three utterances are marked in Table 5.1. We collect question and answers without the reflection from those transcripts as a dataset. Using the chat-complete format, questions and answers represent *System Messages* and *User Messages*, respectively. The generation process is set up to have GPT-4 generate the reflection where GPT-2 originally did in the MIBot experiment. In total, 4194 question-answer pairs are collected for the dataset.

Next, in Section 5.1.3 we explain the complete prompt we format for reflection generation.

5.1.3 Prompt Engineering of GPT-4 for Reflection Generation

Recall, from Chapter 2, that there are two types of reflections we may wish to generate: simple and complex reflections. We will refer to simple reflections as Task 1 and complex reflections as Task 2. Each task requires a different prompt (*System Role* in GPT-4 terminology). The full input to GPT-4 for reflection generation would consist of the a *System Role* combined with a Question and an Answer like the those seen in Table 5.1. A complete formatted example is shown in Figure 5.1. In the Figure, GPT-4 generates a *System Message* which represents either a simple (left-side of the Figure) or complex (right-side) reflection by using a *System Role*, *System Message*, and *User Message*. The reflection is collected into a dataset. This creates two datasets which share the same questions and answers, but differ in the task instruction and output reflection. More specifically, each dataset entry is a triplet containing:

- **Instruction**: A *System Role* instruction for task 1 (simple reflection) or task 2 (complex reflection)
- **Input**: An open-ended question from a therapist and a Client's answer
- **Output**: A simple or complex reflection

In Section 5.3.1, we describe how we use the Task 1 or Task 2 dataset for knowledge distillation.

The prompt engineering of the *System Role* for reflection generation was designed using an iterative process on a private test-set in collaboration with MI-experts. We hand-engineered an initial prompt until we were able to reach an acceptable accuracy on a test-set, then we increased the size of the test-set. This process repeated until we were satisfied with the overall performance of each *System Role*. Figure 5.1 shows the prompts that were created for simple and complex reflection generation using GPT-4. For our knowledge distillation method and results, we refer to GPT-4's reflection generation as the GPT-4 Reflection Generator.

In this work, we perform a separate validation of GPT-4's reflection generation through a human review of reflections it generates. This validation is described in Section 5.2.4.

5.2 Evaluation of Reflections using GPT-4, Human Review, and Automated Metrics

This section explains the different techniques we use to evaluate MI reflections. For the purposes of this work, we use this evaluation to judge the quality of distilled models. To evaluate MI reflec-

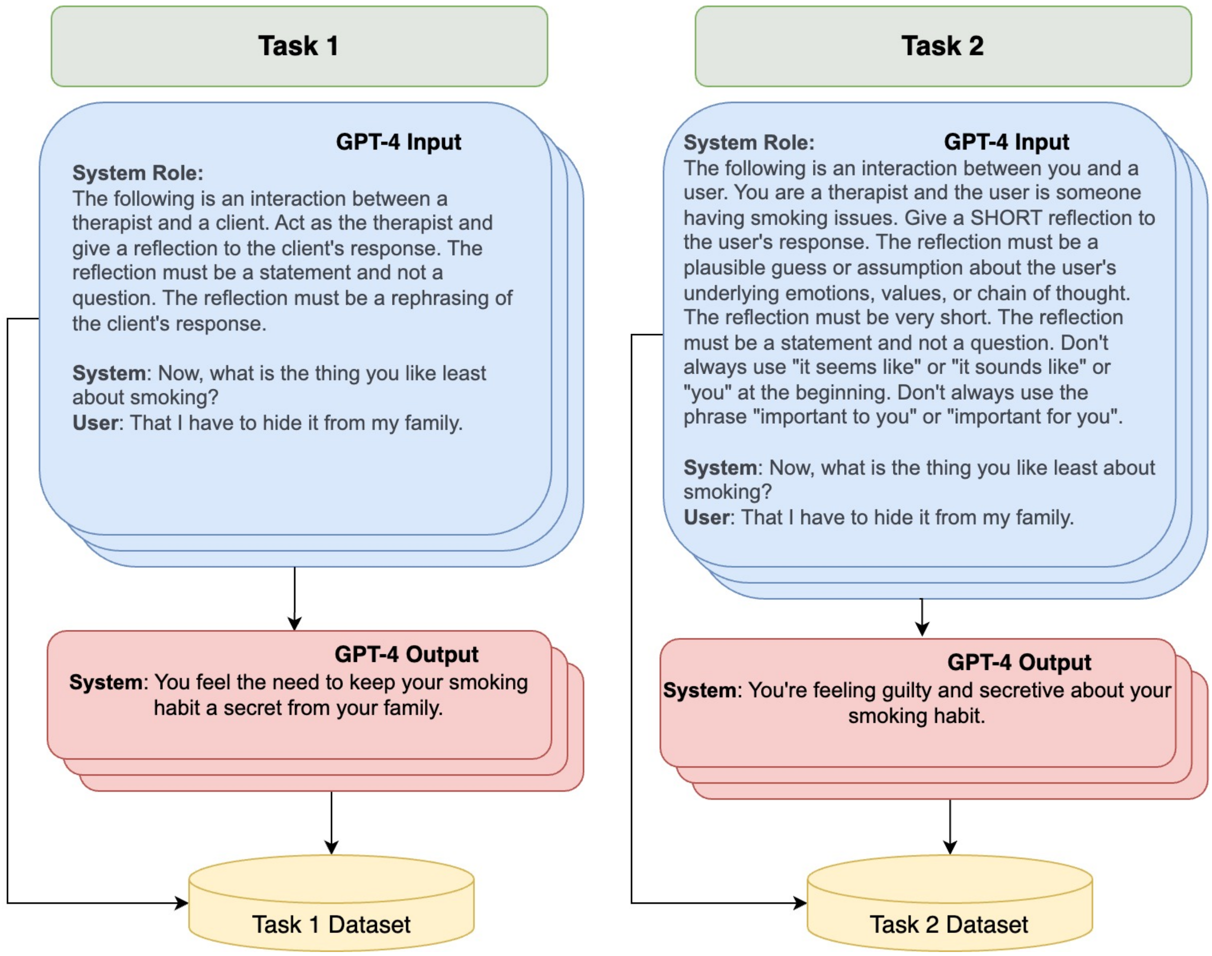

Figure 5.1: Reflection Data Generation Format

tions, we use three techniques: using GPT-4 as a reflection classifier, human reviewers, and classic automated metrics like ROUGE [85] and BERTScore [86].

Reflection evaluation is a difficult task, and has typically been done by humans [87, 88]. While human review is the most trusted method of evaluation for MI reflections they are time and cost-intensive. Furthermore, typical text generation evaluation metrics like ROUGE [85] and BERTScore [86] may not capture the context-dependent nature of evaluating MI reflections. If GPT-4 can be shown to do high-quality evaluation, it allows us to have a task-specific evaluation at much lower cost and effort. Since GPT-4 has often been used in a few-shot or zero-shot mode to perform well as a text classifier, this may be a far easier way to create a classifier.

Our method uses human review for two purposes. First, it is the best-known ground truth to evaluate if GPT-4 generates reflections which satisfy MI-adherence and task classification. Second, it is needed to validate the GPT-4-based evaluation described above. Due to the subjective nature of this kind of classification, it is typical to measure the quality of a new rating system by calculating an inter-rater reliability score between human review and GPT-4's evaluation.

We will also use the traditional automated metrics ROUGE and BERTScore to explore how the results contrast with GPT-4-based evaluation and human review.

This section first explains how we use GPT-4 for MI-adherence, then task classification, followed by an explanation of how they work together. Next, we explain the human review process, Finally, we conclude by explaining our use of automated metrics.

5.2.1 GPT-4 MI-Adherence

MI-Adherence refers to the classification task of determining if a given statement abides by the principles of motivational interviewing for a reflection. This is the most basic qualification of a MI reflection and gives an indication of how well a reflection model is performing. Miller et. al. [6] states that the four principles of MI are to: express empathy, develop discrepancy between the client's goals and their behaviors, roll with resistance from the client, and support self-efficacy. Reflections should not infringe on any of these qualities.

To concretize our definition of MI-adherence, we define a six point list of our own MI-adherence rules for reflections, which are inspired from Miller et. al. [6] and from working in MI in collaboration with MI experts, specifically for smoking cessation.

A reflection should:

1. Be a statement, not a question.
2. Not be MI-inconsistent in the following ways: giving advice or information without permission, or confronting the person by disagreeing, arguing, correcting, shaming, blaming, criticizing, labeling, ridiculing, or questioning the person's honesty, or directing the person by giving orders, commands, or imperatives, or otherwise challenging the person's autonomy.
3. Not incentivize people to smoke more, or discourage people from quitting smoking
4. Not exaggerate or understate the sentiment of the sentence to be reflected
5. Not be factually wrong about smoking
6. Be grammatically correct

This six point list is used to design the zero-shot prompt given to GPT-4 for MI-adherence classification. We show the complete prompt that we use for MI-adherence in Table 5.2.

Similar to Section 5.1.3, the design of the MI-adherence classifier prompt using GPT-4 was created using a step-by-step process on a private test-set in collaboration with MI-experts. Each prompt was hand-engineered and changed until we were able to reach an acceptable accuracy on a test-set, then the size of the test-set was increased. This process repeated until we were satisfied with the overall performance of each *System Role*.

5.2.2 GPT-4 Task Classification

Task classification refers to classifying a reflection as either simple or complex, with the assumption that the reflection has already been determined to be MI-adherent. Since we generate both simple and complex reflections, we will use this classifier to measure how well a model is generating either simple or complex reflections. We use the following working definition to define a simple vs complex reflection: A simple reflection must be a rephrasing of the client's response. In contrast, a complex

Table 5.2: GPT-4 Prompt for MI-Adherence

Prompt
Decide whether the "reflection" sentence in the following smoking-related conversation meets the standards for Motivational Interviewing. If it does, output "True"; otherwise, output "False". Additionally, a good reflection must: 1. Be a statement, not a question. 2. Not be MI-inconsistent in the following ways: giving advice or information without permission, or confronting the person by disagreeing, arguing, correcting, shaming, blaming, criticizing, labeling, ridiculing, or questioning the person's honesty, or directing the person by giving orders, commands, or imperatives, or otherwise challenging the person's autonomy. 3.Not incentivize people to smoke more, or discourage people from quitting smoking. 4.Not exaggerate or understate the sentiment of the sentence to be reflected. 5. Not be factually wrong about smoking. 6. Be grammatically correct.

reflection must not be just a rephrasing of the client's response, but instead a plausible guess or assumption about the user's underlying emotions, values, or chain of thought.

Using this working definition, we create a zero-shot prompt for GPT-4 to classify reflections as simple or complex. We show the complete prompt that we use for task classification in Table 5.3.

Similar to Section 5.1.3 and Section 5.2.1, the design of the Task Classification prompt using GPT-4 was created using a step-by-step process on a private test-set in collaboration with MI-experts. Each prompt was hand-engineered and changed until we were able to reach an acceptable accuracy on a test-set, then the size of the test-set was increased. This process repeated until we were satisfied with the overall performance of each *System Role*.

Table 5.3: GPT-4 Prompt for Task Classification

Prompt
Decide whether the "reflection" sentence in the following smoking-related conversation is a SIMPLE or COMPLEX reflection. If it is simple, output "simple"; otherwise, output "complex". A simple reflection must be a rephrasing of the client's response. In contrast, a complex reflection must not be just a rephrasing of the client's response, but instead a plausible guess or assumption about the user's underlying emotions, values, or chain of thought.

5.2.3 GPT-4 Reflection Evaluation Overall Process

The complete method to evaluate a given reflection with GPT-4 combines the classifier made for MI-adherence and simple/complex reflections.

Figure 5.2 shows a complete example of the evaluation pipeline. On the left side of Figure 5.2, either the GPT-4 Reflection Generator or distilled models generate a candidate reflection with either

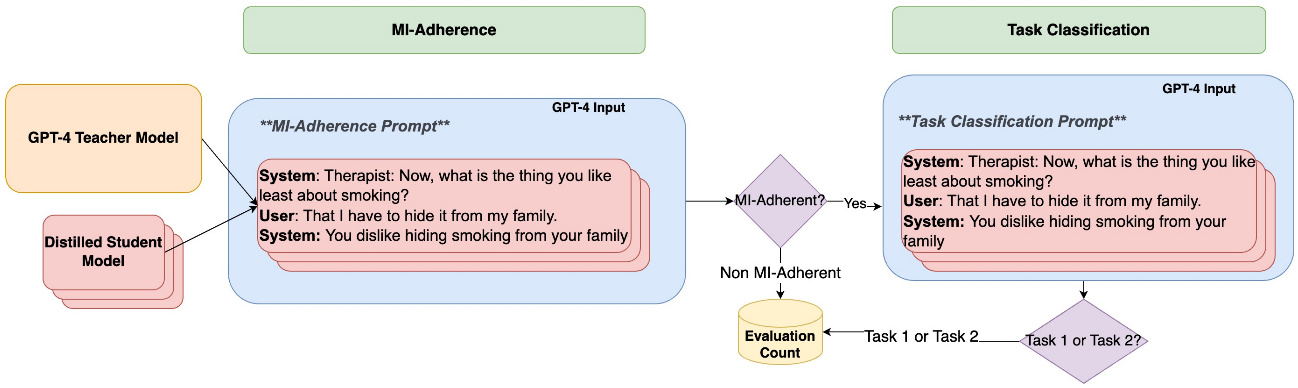

Figure 5.2: Reflection Model Evaluation Pipeline

the Task 1 or Task 2 instruction. A holdout-set of question-answer pairs is used as the input in the evaluation process. The candidate output reflection is passed through the two classifiers, the first one using the MI-adherence prompt shown in Table 5.2. If the candidate reflection passes the MI-adherence test, it is sent to the simple/complex reflection classifier to determine which kind of reflection it is using the prompt shown in Table 5.3. The end result is a reflection which is classified as not MI-adherent, simple, or complex (with the implication that both simple and complex reflections are MI-adherent).

We provide a validation of the performance of each GPT-4 evaluation classifier in Section 6.2. The Cohen Kappa [89], a validated metric to measure inter-rater reliability is calculated between GPT-4's evaluation and the human review. Since we aim to automate costly human review, the reliability score gives a sense of how well these classifiers agrees with human review.

5.2.4 Human Review

We recruited five annotators to evaluate holdout test set reflections from the GPT-4 Reflection Generator and each distilled student model. Each annotator has a basic understanding of MI having read [6] and taken coursework [1]. We take inspiration from Wu et. al. [87] who showed that lay-people are able to label MI reflections with consistent inter-group correlation.

The GPT-4 Reflection Generator creates 1201 reflections (a process later explained in Section 5.3.1 and 61 of them (∼5%) are randomly sampled with stratification[2] from each distilled model for human review. We review 10 models total: the teacher GPT-4 Reflection Generator for Task 1/simple and Task 2/complex reflections and four student GPT-2 models of different sizes and each type of reflection. This gives a total of 610 review examples. Below we describe the human review process which closely follows the same MI-adherence and task classification process as the one used in the GPT-4 Classifiers described in Section 5.2.3.

For MI-adherence, the annotators classified reflections using their own understanding of MI. We aim to capture the subjective opinions of MI-adherence and contrast them with the results of the GPT-4-based classifiers.

For task classification, the annotators classified reflections as either Task 1 (simple) or Task

[1] http://test.teachdev.ca/ola/index.html
[2] Reflections are stratified by the question asked, to ensure there is diverse context

2 (complex). Reflections are assumed as simple unless there is a plausible assumption about the client's underlying emotions, values, or chain of thought, similar to the instructions as specified in Figure 5.1, Task Classification.

Among the five annotators, each reflection reviewed is subjected to triple-blind decision, meaning three annotators make independent decisions. For the binary/two-way classifications being determined, the majority result, from the three, is chosen. We use this majority voted aggregate decision to calculate the agreement score explained in Section 5.2.3.

5.2.5 Automated Metrics

In addition to using a GPT-4 based classifier and human review, we use also made use of traditional (but often problematic) automated metrics to evaluate the student models. Distilled student model candidate reflections were automatically compared to the GPT-4 Reflection Generator reference reflections using the following metrics:

ROUGE metrics: ROUGE calculates word overlap metric between candidate and reference generations [85]. It is used frequently in the evaluation of text generation tasks. We use ROUGE-1 (1-gram matching), ROUGE-2 (2-gram matching), and ROUGE-L (longest common sub-sequence matching) similarly to how [70] evaluates candidate MI reflections.

BERTScore: BERTScore computes a similarity score for each token in the candidate sentence with each token in the reference sentence using contextual embedding similarity [86]. We include BERTScore to ensure we evaluate reflections based on N-gram (ROUGE) and an embedding approach.

5.3 Knowledge Distillation Method

This section presents a knowledge distillation process used to distill the zero-shot prompted reflection generation from the GPT-4 Reflection Generator to smaller student models. Above, we described a method for generating simple and complex reflections with GPT-4. We expected, and the results chapter will show that the quality and success rate of the reflection generation is far superior to our previous methods (used in Chapters 3 and 4), but since we can not guarantee data privacy with GPT-4 in a deployed experiment like MIBot requires, we must use a method like knowledge distillation to transfer GPT-4's reflection performance to a smaller model which we can own.

In Section 2.2.4, we reviewed prior attempts to distill language models, with investigation into the algorithm of distillation and evaluation. In our knowledge distillation method, we investigate using fine-tuning as the algorithm for distillation and attempt to use GPT-4 as an evaluator of distilled student model performance. Furthermore, we investigate how changing student model size and the type of reflection distilled (simple or complex) changes the distillation outcome.

First, we give a general overview of the knowledge distillation process, making use of reflection generation and evaluation described above. Then, we describe the fine-tuning process we use for knowledge distillation in more detail. Finally, we discuss the details of our student model selection.

5.3.1 Knowledge Distillation Overview

We present a knowledge distillation method using generated MI reflections from the GPT-4 Reflection Generator as a knowledge dataset for fine-tuning smaller students. Figure 5.3 illustrates the knowledge distillation process we use. We divide our distillation into Knowledge Extraction, Distillation Fine-tuning, and Distillation Evaluation. Below, we explain the these three sub-processes in further detail.

First, for Knowledge Extraction, we use the GPT-4 Reflection Generator as explained in Section 5.1.3. In total the 4194 question-answer pairs (of the type described in Section 5.1.3) is used to generate a reflection for each task. This creates a dataset with 4194 question-answer-reflection triplets where the reflection is meant to be a simple reflection, created from the Task 1 prompt and 4194 triplets with complex reflections, created from the Task 2 prompt. As these are going to be used in a gradient-descent-based fine tuning, they need to be split into training, validation and testing sets. The 4194 examples are divided into 2394 training set examples, 599 validation set examples, and 1201 holdout testing set examples. These datasets contain some GPT-4's "knowledge" of how to generate reflections, and are used in the next step, fine-tuning of the student models.

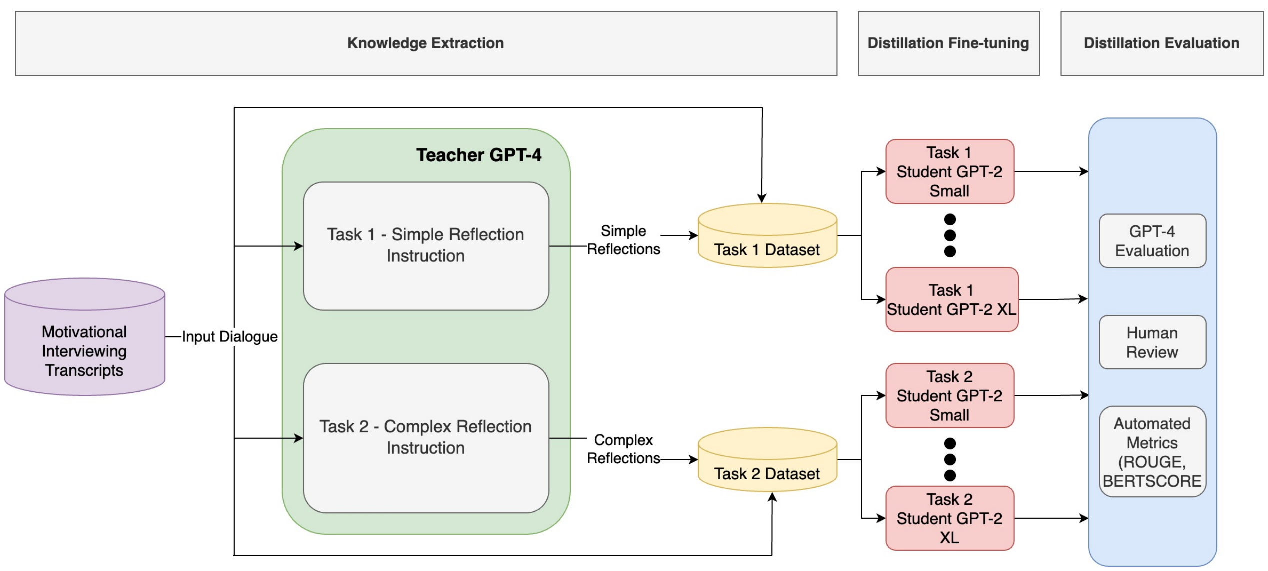

Figure 5.3: Knowledge Distillation Overview

Next, for Distillation Fine-tuning, we use the Task 1 or Task 2 dataset to fine-tune the GPT-2 [17] family of models. The end result is a distilled student reflector which generates either simple or complex reflections. We explore the effect of model size (GPT-2 Small to GPT-2 XL) on the quality of the results, as well as the effect changing the type of reflections (simple/Task 1 or complex/Task 2). Distillation Fine-tuning is further explained in Section 5.3.2 and further details of student model selection are explained in Section 5.3.3.

The right-most part of Figure 5.3 illustrates the full process of the evaluation of the results. After the student models have been fine-tuned they are evaluated using three methods: Using GPT-4 for evaluation, using human review, and using the traditional automated metrics ROUGE and BERTScore. GPT-4's evaluation follows the same process as described in Section 5.2.3.

Alongside, GPT-4's evaluation, human reviewers are employed to do the same evaluation process as GPT-4. This review process, along with the inter-rater reliability score which validates GPT-4's evaluation is the same process as explained in Section 5.2.4.

Finally, we compute automated metrics (ROUGE, BERTScore) of distilled student models to contrast with the results of GPT-4's evaluation and human review.Our use of automated metrics is the same process as explained in Section 5.2.

5.3.2 Distillation Fine-tuning

Our knowledge distillation method uses fine-tuning as the algorithm for transferring knowledge from a teacher to a student model. This differs from the traditional method of knowledge distillation, where the student model begins completely untrained, and is trained on the output logits (soft-targets) of a teacher model [23, 24, 25]. In the present method, we investigate the efficacy of instead using a pre-trained model and fine-tuning using hard-targets in the form of generated text. This style of distillation is similar to He et. al. [62], who showed that knowledge distillation can be achieved through generating synthetic text and using it to fine-tune language models.

We motivate this method by noting that state-of-the-art Foundational Language Models such as ChatGPT [19] and our teacher, GPT-4 [20] do not give users access to the output logits or probabilities used in next word prediction. Furthermore, it has been shown in recent research [56] that hard-target distillation can more effective when the student-teacher architectures are very different, which is likely true between GPT-4 and GPT-2. Below we describe the text formatting used and details of fine-tuning.

Example Task 1 Entry
Instruction: The following is an interaction between a therapist and a client. Act as the therapist and give a reflection to the client's response. The reflection must be a statement and not a question. The reflection must be a rephrasing of the client's response. ### Conversation: Therapist: Now, what is the thing you like least about smoking? Client: That I have to hide it from my family. Therapist: You feel the need to keep your smoking habit a secret from your family.
Example Task 2 Entry
Instruction: The following is an interaction between you and a user. You are a therapist and the user is someone having smoking issues. Give a SHORT reflection to the user's response. The reflection must be a plausible guess or assumption about the user's underlying emotions, values, or chain of thought. The reflection must be very short. The reflection must be a statement and not a question. Don't always use "it seems like" or "it sounds like" or "you" at the beginning. Don't always use the phrase "important to you" or "important for you". ### Conversation: Therapist: Now, what is the thing you like least about smoking? Client: That I have to hide it from my family. Therapist: You're feeling guilty and secretive about your smoking habit.

Table 5.4: Task 1 and Task 2 Dataset Entry Example

After text has been generated using the process illustrated in Figure 5.1 and described in Section 5.1 it must be formatted for the fine-tuning of the student model. Table 5.4 shows an example entry from the simple reflection/Task 1 dataset as well as the complex reflection/Task 2 dataset. The fine-tuning training takes each dataset entry and predicts the next word given a sequence (or sub-sequence) of the dataset entry. We use the word predicted vs the actual correct word to calculate a loss for gradient descent.

Our dataset entries includes the prompt (GPT-4's *System Role*) used to generate the reflection and the conversation of a question, answer, and reflection. We use a triple # sign to separate the Instruction and Conversation, inspired by how Taori et. al. [65] formatted fine-tuning data for the Alpaca language model as mentioned in Section 2.2.4.

5.3.3 Student Model Selection

We selected the family of GPT-2 [17] transformer-based language models as the student models. All models have been pre-trained on the WebText dataset, a 40GB corpus of diverse text. Table 2.3 shows the parameter count, transformer block count, and embedding length for each GPT-2 model.

We selected the family of GPT-2 transformer-based language models for the variety in size among models, and familiarity. For variety in size, we set out to investigate how different model sizes would correspond to the knowledge distillation outcome. The GPT-2 family has a large variety of sizes, with the smallest to the largest being an increase of 12 times (but all models significantly smaller than GPT-4). For familiarity, since we used the GPT-2 XL model to generate MI reflections in the chatbot experiment, we already knew how to train and inference the model.

5.4 Experiment

5.5 Experimental Setup

All models the student GPT-2 models are implemented using PyTorch [90] and downloaded from the HuggingFace Transformers library [91]. For training and inference, we used 4 NVIDIA A10G Tensor Core GPUs with DeepSpeed ZeRO [92] parallelism and CPU offloading. Several key model hyper-parameters were determined using a hyperparameter search. We searched for Batch Size in [8, 16, 32, 64] and Learning Rate in [0.00005, 0.0005, 0.001]. The fine-tuning described in Section 5.3.2 is executed with 4 epochs and early stopping [93]. We use the Adam Optimizer [94] with zero weight decay. For inference, we use a decoding strategy of temperature=0.6 with top-k=100 and top-p=1.0. All code is released online here [3].

[3] https://github.com/andrewmbrown/transformer-fine-tune

Chapter 6

Reflection Generation, Evaluation and Distillation Results

This chapter presents the results of the various methods presented in Chapter 5. First, report the quality of the reflections generated by the GPT-4 Reflection Generator against human reviewer results. Next, we present our Inter-Reliability agreement between GPT-4's evaluation and human review in an attempt to validate GPT-4's evaluation. These measurement methods, together with others are used to report on the quality of the distilled student models.

We conclude Chapter 6 by providing a section to attribute contributions.

6.1 Performance of GPT-4 Generation of Reflections

	MI-Adherence		Classified as Simple		Classified as Complex	
Model - Task	GPT-4	HR	GPT-4	HR	GPT-4	HR
GPT-4 - Task 1	**0.99**	**1.00**	**0.91** (1196)	**0.97** (61)	0.08	0.03
GPT-4 - Task 2	**0.98**	**1.00**	0.26 (1183)	0.13 (61)	**0.74**	**0.87**

Table 6.1: MI-Adherence and Task Classification fractions of success using the GPT-4-based classifiers (The GPT-4 column) and using Human Review (HR). Columns 4 and 5 also include the count of candidate reflections which make it past MI-adherence in parenthesis.

Section 5.1 described a method for prompting GPT-4 to produce simple or complex reflections, and Section 5.2 described several methods of evaluating those reflections, including using GPT-4 itself as a classifier, and human review. Table 6.1 presents the metrics from these two methods. The test set used for the GPT-4-based classifiers is the 1201 holdout-set described in Section 5.3.1. The human review test set is smaller (due to the labour required) and is a randomly sampled 61 examples as explained in Section 5.2.4. Each of the values in Table 6.1 gives the fraction of the test set that was deemed acceptable by the evaluation method. For example, the 0.99 for the GPT-4 MI-Adherence column of the Table indicates that 99% of the 1201 set of generated reflections for Task 1 was judged as MI-Adherent by the GPT-4 classifier. The right-most four columns of Table 6.1 give the fraction of the reflections that were deemed, by GPT-4 Evaluation or the human review, to be simple/Task 1 adherent or complex/Task 2. In columns 4 and 5, we also include the number of examples from the holdout set which were used to calculate the evaluation score (in parenthesis). Since reflections have to first be classified as MI-adherent before task/reflection type classification, some reflections do not make it to the second step, hence the smaller count. This count of reflections which pass MI-adherence also apply to columns 6 and 7, for complex reflection classification.

It is notable that, for MI-adherence, the human reviewers scored GPT-4 Reflection Generation at 100% success for both simple and complex reflections. Furthermore, also for human review, GPT-4 Reflection Generation on Task 1 scores over 97% success in generating simple reflections and GPT-4 Reflection Generation Task 2 scores 87% in complex reflections. These results suggest that the GPT-4 Reflection Generator is very consistent in generating acceptable simple and complex reflections.

For task classification, the GPT-4 Reflection Generator demonstrates reliable Task 1/simple reflection generation. We note that the GPT-4 Reflection Generator does not always succeed in creating a Task 2/complex reflection. We hypothesize that because an effective complex reflection is often context dependent [6], GPT-4 is not always able to generate a coherent complex reflection, and opts for a simple reflection instead. We note that our knowledge distillation goal is to capture the performance of GPT-4, regardless of how well it performs.

6.2 Performance of GPT-4-based Evaluation

Task	MI-Adherence	Task Classification	Combined
Task 1	**0.671**	0.604	0.61
Task 2	0.429	**0.711**	0.312
Average	0.54	0.66	**0.61**

Table 6.2: Inter-Rater Reliability Cohen Kappa scores between GPT-4 and Human Reviewers on three evaluation tasks. The last column, combines MI-Adherence and Task Classification into one evaluation task.

In Section 5.2 we presented a method for using GPT-4 itself to evaluate the quality of reflections produced by models, as a possible alternative to the laborious human review. In this section we measure its performance against human review, using the Cohen kappa Inter-Rater Reliability [89] coefficient. Recall that the human review result is determined by majority vote across the three reviewers. The Cohen kappa metric ranges from -1 to 1 representing perfect disagreement and agreement, respectively. Notably, any score of above 0.6 is considered substantial agreement [89]. The results in Table 6.2 presents the Cohen kappa coefficient between the GPT-4-based evaluation and the majority human evaluation for the adherence and simple/complex classification classifiers. Each value in the table gives the Cohen kappa agreement score between GPT-4's evaluation and human review for the type of reflection specified in the left-most column, and type of evaluation in the top row. The right-most column is the Cohen kappa between combining the decisions of GPT-4 MI-adherence and task classification and human reviewers into one task.

Recall that the human review set is 61 holdout reflections long, as discussed in Section 5.2.4, the agreement calculation of MI-adherence is calculated on 610 reflections, based off the 61 reflections from the sub-set human annotators review, for 10 models total. Furthermore, the bottom row average MI-adherence review score in Table 6.2 is calculated for 1220 examples (610 from task 1 and 610 from task 2). The agreement of task classification between GPT-4 and human reviewers is done on a sub-set of reflections where both GPT-4 and humans agree that the reflection is already MI-adherent. For Task 1/simple reflections, agreement is calculated on 272 examples and for Task 2/complex reflections, agreement is calculated on 261 examples. Altogether, the task classification average on the bottom row is 533 reflections.

For the classification of MI-adherence, the second column in the table, we observe relatively high levels of agreement, between human and GPT-4 based evaluation. The Task 1/simple reflection generation scores 0.671, and the Task 2/complex reflection generation scores lower at 0.429. This score suggests that GPT-4's MI-adherence evaluation follows more similar evaluation trends to human review in task 1, but not as well in task 2. Later we show in Section 6.3 that GPT-4's MI-adherence evaluation follows a similar pattern in Task 1/simple reflections, but not as well in Task 2/complex reflections. Overall, average MI-adherence score of 0.54 suggests that there is near substantial agreement between the GPT-4 classifier and human review, validating our use of GPT-4 for MI-Adherence.

The third column in Table 6.2 shows agreement between the GPT-4 Simple/Complex classifier and human review. Notably, GPT-4 and human reviewers score higher on reflections generated for Task 2/complex than those generated for Task 1/simple. We speculate this is because the GPT-4 Reflection Generator and distilled student models when instructed or trained to generate Task 2/complex reflections generate more simple reflections than vice versa. This more uniform distribution from complex reflectors leads to a higher agreement score as there are more chances for GPT-4's evaluation and human reviewers to agree in either direction. Overall, the average score of 0.66 shows that GPT-4 and human review have substantial agreement on task classification, validating our use of GPT-4 for task Classification.

The last column of Table 6.2 combines decisions made by GPT-4's MI-adherence and task classification and computes agreement with human review. These scores show that human reviewers agree more with the predictions made by task 1/simple reflection models than predictions made by task 2 models.

6.3 Performance of Distilled Student Reflection Generation Models

Section 5.3.1 describes a method for distilling the GPT-4-based reflection generation model into a smaller set of student models based on GPT-2. An earlier section described both the GPT-4 based method for evaluating these student models as well as a human based review method. Table 6.3 presents the fraction of generated samples that were deemed successful by each evaluation method.

Student models are listed in each row of the table, in order of increasing model size, and are grouped by which reflection generation task they performed - Task 1/Simple or Task 2/Complex. The table includes the results from GPT-4 reflection generation itself (already presented in Table 6.1) for ease of comparison, in blue. Similar to Section 6.1 higher fractions of MI-adherence are better. Columns 5 and 6 similarly to Section 6.1 and 6.2 give the task classification scores for task 1/simple reflections and beside the score is the number of examples which GPT-4 or human reviewers evaluated for that model in parenthesis. Because we only classify reflections as simple or complex which are MI-adherent, the count is less than 1201 for GPT-4's evaluation and 61 for human review (assuming some reflections are classified as non MI-adherent). This count of examples

		MI-Adherence		Classified as Simple		Classified as Complex	
Model - Task	Size	GPT-4	HR	GPT-4	HR	GPT-4	HR
GPT-2 Small - Task 1	124M	0.76	0.90	0.78 (895)	0.69 (55)	0.22	0.31
GPT-2 Medium - Task 1	355M	0.91	0.87	0.77 (1083)	0.81 (53)	0.23	0.19
GPT-2 Large - Task 1	774M	0.93	0.90	0.79 (1112)	0.71 (55)	0.21	0.29
GPT-2 XL - Task 1	1.5B	**0.93**	**0.92**	**0.80** (1117)	**0.82** (56)	0.20	0.18
GPT-4 - Task 1	>>>	**0.99**	**1.00**	**0.91** (1196)	**0.97** (61)	0.08	0.03
GPT-2 Small - Task 2	124M	0.83	0.85	0.25 (1004)	0.17 (52)	0.76	0.83
GPT-2 Medium - Task 2	355M	0.86	0.92	0.25 (1029)	0.05 (56)	0.75	**0.95**
GPT-2 Large - Task 2	774M	0.86	**0.97**	0.23 (1026)	0.17 (59)	**0.77**	0.83
GPT-2 XL - Task 2	1.5B	**0.90**	0.92	0.26 (1086)	0.11 (56)	0.74	0.89
GPT-4 - Task 2	>>>	**0.98**	**1.00**	0.26 (1183)	0.13 (61)	**0.74**	**0.87**

Table 6.3: MI-Adherence and Task Classification scores of distilled student models as well as reprise of results for teacher GPT-4 reflection generator. Columns 5 and 6 give the count of examples GPT-4 and human reviewers see for each model in parenthesis. HR stands for Human Review.

which GPT-4 evaluation and human reviewers see is the same in columns 7 and 8, for reflections classified as complex (so the counts are only given once in the table).

6.3.1 Distilled Student MI-Adherence

The third and fourth column of Table 6.3 shows MI-adherence fractions the two tasks. In student models, we observe that in both tasks, the GPT-4-based evaluation of MI-adherence score increases in accuracy as model size increases, with both GPT-2 XLs scoring the highest. For human review MI-adherence, we see the same trend of larger models scoring higher on MI-adherence but with less predictability. Human review scores GPT-2 XL Task 1 as the highest MI-Adherence but for Task 2, GPT-2 Large, a smaller model scores the highest MI-adherence. Notably, the non-predictability in human review MI-adherence could be due to the smaller sample size. For MI-adherence, GPT-4 reviews 1201 examples per model, while reviewers evaluate 61 examples per model.

6.3.2 Distilled Student Task Classification

The 5th, 6th, 7th, and 8th columns of Table 6.3 give the fraction of the reflections processed that were classified as either Task 1/simple 1 or Task2/complex. For student models, we observe that models generating Task 1/simple reflections on average generate more simple reflections as model sizes increase, with GPT-2 XL scoring higher than 80% simple reflections generated from both GPT-4's evaluation task classification and human review. Notably, Task 2/complex reflection classification does not improve with model size, and we observe no apparent trend. GPT-4's evaluation scores and human review do not score higher accuracy of Task 2/complex reflections as model size increases for task 2 models. We conclude from task classification that generating and classifying simple reflections is an easier task, with both GPT-4 and human review having similar scores. Complex reflections are more subjective, and GPT-4's evaluation and human review do not always agree.

6.4 Automated Metrics Results

We were interested to see how well traditional automated language metrics would compare to the two metrics - GPT-4 based and human, presented above. As described in Section 5.2.5 we use the ROUGE-1, ROUGE-2, ROUGE-L, and BERTScore as alternative automated metrics. Table 6.4 reports scores on automated metrics for all the distilled models.

		ROUGE			
Model - Task	Model Size	RG-1	RG-2	RG-L	BERTScore
GPT-2 Small - Task 1	124M	0.414	0.196	0.377	0.902
GPT-2 Medium - Task 1	355M	0.446	0.216	0.406	0.909
GPT-2 Large - Task 1	774M	**0.466**	**0.237**	**0.426**	**0.914**
GPT-2 XL - Task 1	1.5B	0.459	0.232	0.421	0.913
GPT-2 Small - Task 2	124M	0.34	0.127	0.297	0.885
GPT-2 Medium - Task 2	355M	0.339	0.129	0.293	0.885
GPT-2 Large - Task 2	774M	0.336	0.132	0.294	0.887
GPT-2 XL - Task 2	1.5B	**0.358**	**0.144**	**0.313**	**0.89**

Table 6.4: ROUGE and BERTScore Automated Metric scores for each distilled student model. ROUGE is broken into three types, ROUGE-1 (RG-1), ROUGE-2 (RG-2), and ROUGE-L (RG-L). For all automated metrics candidate sequences are the reflection generated by the distilled student model and reference sequences are the reflection which GPT-4 generated.

ROUGE scores range between 0 and 1 with higher scores representing closer word overlap. BERTScore also scores between 0-1 but higher scores represent word semantic similarity. Recall that the distilled model outputs are compared to the GPT-4 Reflection Generation outputs, which although good, are not the only way to achieve a good reflection. Below we discuss Task 1 generation models, Then task 2.

For Task 1 models, the scores have the same order in magnitude across each automated metric. Task 1 models increase in score as model size increases, but peak at GPT-2 Large Task 1 with GPT-2 XL task 1 scoring lower. Notably, Table 6.3 suggests that GPT-2 XL task 1 is the best performing model in MI-adherence and task classification, which disagrees with Table 6.4.

Task 2 models score the same order in magnitude across each automated metric. For Task 2 models, GPT-2 XL task 2 scores the highest across all metrics. This result also disagrees with Table 6.3 which shows that GPT-2 XL task 2 only scores the highest for GPT-4's evaluation of MI-adherence, and nothing else. We conclude from these observations that automated metrics like ROUGE and BERTScore can be misleading for subjective generative tasks that require more sensitivity than semantic similarity.

6.5 Knowledge Distillation Contribution Attribution

Several parts of work presented in Chapters 5 and 6 was completed by other contributors. These are listed below alphabetically by last name, followed by the contributions of the author.

- Mohamed Abdelwahab - created the system role for simple and complex GPT-4 reflections as described in Section 5.1.3. Abdelwahab also contributed to data labelling as explained in Section 5.2.4.

- Alec Dong - contributed to data labelling as explained in Section 5.2.4.
- Cindy Wang - contributed to data labelling as explained in Section 5.2.4.
- Jiading Zhu - created the system role for GPT-4's evaluation of MI-Adherence and Scope Classification as explained in Section 5.2.3. also contributed to data labelling as explained in Section 5.2.4.

The author of the present work is the main contributor and organizer of the knowledge distillation work. All collection of MI transcript data was done by the experiments deployed by the author as explained in Section 5.1.2. All data collection, model fine-tuning, and evaluation was done by the author, as well as the organization of results and discussion was done by the author.

Chapter 7

Conclusions, Limitations, and Future Work

7.1 Conclusions

In this work, we presented a chatbot conversation and a knowledge distillation process which have a common research goal, using computers with Natural Language Processing to automate the Motivational Interviewing (MI) therapeutic approach.

Chapter 3 presented a scientific and engineering framework for measuring the effect of an automated conversation on a smoker's readiness to quit smoking. Using this framework, Chapter 4, presented how four versions of the conversation affect that readiness. We found that simply asking relevant questions about smoking is sufficient to confer benefits on readiness confidence, whereas generated reflections may additionally increase other readiness attributes, while making the chatbot appear more empathetic.

Chapter 5 presented a method for generating simple and complex MI reflections using GPT-4, then distilling that task to smaller student models via fine-tuning on synthetic text. We investigated how changing model size and the type of reflection changed evaluation results of distilled models, including using GPT-4 as an evaluator of the distillation success. Chapter 6 showed that distilled student models were observed to be successful at reflection generation through a variety of evaluations. Furthermore, we found that a re-purposed GPT-4 evaluating student models shows substantial agreement with human reviewers and has different results from automated metrics like ROUGE and BERTScore. This motivates the goal of changing the way we do automated evaluation of text generation, toward more task-dependent, low-data evaluators like GPT-4.

7.2 Limitations

We divide limitations into chatbot (Chapter 3 and 4) and distillation works (Chapter 5 and 6).

Our findings should be considered in the context of several limitations. First, self-reported measures used in our study to evaluate the various chatbot versions (i.e., readiness ruler, HSI, CARE, self-reported cigarette consumption and change) are potentially less accurate than a clinician-administered survey. Research suggests that participants in health studies tend to under-report unhealthy behaviours and over-report intentions to improve [95]. This tendency may account for some differences in smoking behaviours observed in our sample and other studies of MI (e.g., quit attempts). In addition, the data suggested that the comprehension of the metric number of quit attempts was interpreted differently by different participants, and so that less reliable. Furthermore, participants were informed that the aim of the conversation was to help improve the chatbot, potentially leading them respond in what they believed to be a desirable way after the conversation, rather than their true feelings. Participants recruited in this study were also financially compensated contingent on a review of their responses, which may have led them to agree with statements on surveys even if they did not (i.e., Acquiescence Bias [96]). Although such tendencies would apply to all chatbot versions (and not apply to comparisons between them), they limit conclusions drawn about pre- to post-conversation comparisons. Nevertheless, one purpose of the survey administered one week later was to give participants time to forget their answers to the initial surveys and to see if the impact holds over time. There is also variance in characteristics among populations in our chatbot versions. This effect is known as the cohort effect [97] and can be seen in Table 4.4 and Table 4.5 where we see variation in participant starting values on the readiness scales. Each population sample has different characteristics and thus have different starting values. This makes comparison among studies difficult because we lose relative significance. In this study, the iterative nature of the chatbot motivated evaluations at different temporal periods. However, to draw appropriate conclusions about the impacts of different versions, future research should randomize smokers to interact with one of the various versions, or with a control in an Randomized Control Trial (RCT), to eliminate such cohort effects.

The knowledge distillation findings should also be considered with several limitations.

First is our use of evaluation techniques for distilled models. As explained in Section 5.5, during holdout-set reflection generation, we only allow for one try using the same decoding strategy of temperature=0.6 with top-k=100 and top-p=1.0. The randomness in this generation could cause effects that do not accurately represent the performance of the distilled models. Nevertheless, GPT-3 [18], a notable work which evaluates generated samples uses the same method of evaluation, only allowing one generation with similar decoding techniques.

Another limitation that should be considered is our use of hard-target text labels for knowledge distillation. Typically, in knowledge distillation students are trained on the output logits (entire distribution) of a teacher model [23]. This is argued to extract more information from the teacher model as we can see more of the teachers thought process. Our distillation method which only uses the hard-target text labels may have limited the amount of information transferred from teacher to student, impacting the results.

Finally, is the discrepancy between the size of the holdout test set between GPT-4's evaluation and human review. GPT-4 is given 1201 reflections for review while Human Reviewers are given 61 ($\sim$5% of 1201). This was done in order to reduce the labour of labelling, but results in a smaller sub-set which may not accurately represent the original holdout test set.

7.3 Future Work

Moving forward, we propose three avenues for future work. First is to design a more complex, longer conversation which employs more aspects of a clinician-delivered MI conversation. We hypothesize that a longer, more nuanced interaction will help to invoke more contemplation in participants, and we plan to build a much more complex conversation. Second, we wish to include trained models from knowledge distillation in the chatbot conversation. This work has shown that it is feasible to transfer the task of generative MI reflections from a large, private foundational model to a smaller, user-owned model. If we can successfully deploy these smaller models, we could reap the benefits of knowledge distillation by achieving similar performance but with accessible resources. Lastly, we intend to improve the quality of the reflections generated. GPT-4 [20] has been shown to be very powerful at language generation, thus we feel investigation of more clever prompt-engineering techniques could create more effective responses and reflections, with more clinical impact. These generations could serve as the next task we wish to capture with knowledge distillation, improving our automation of MI further.

Appendix A

A.0.1 MI Chatbot Transcript Version 4.7 - Questions but no reflections

BOT: Hello! My name is MIBot, I am a computer chatbot and part of a research project at the University of Toronto. I was made by a group of researchers and doctors, with the goal of exploring your smoking behaviours.
BOT: I will ask a few questions, listen to your answers, and then respond.
BOT: However, I am learning how to have this conversation. So, I may ask you if the things I say make sense.
BOT: Would you like to chat with me about your smoking?
USER: yes
BOT: Great!
BOT: To start, what is the thing you like most about smoking?
USER: I like how it makes me feel relaxed
BOT: Thank you for your answer.
BOT: Now, what is the thing you like least about smoking?
USER: it being bad for my health
BOT: Thanks for your response.
BOT: Now, what is one thing about your smoking habit that you would like to change?
USER: being able to control better how much I actually smoke
BOT: Thank you for answering.
BOT: What will it look like when you have made this change in your smoking habit?
USER: I'll be healthier and waste less money on cigarettes
BOT: Thank you for your response.
BOT: Finally, what are the steps you need to take to make this change?
USER: be emotionally stable and less stressed overall
BOT: Thanks for your answer.
BOT: Thank you very much for this conversation!

A.0.2 MI Chatbot Transcript Version 5.0 - Questions & Reflections with Generator Version 1

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

your smoking behaviours.
BOT: I will ask a few questions, listen to your answers, and then respond.
BOT: However, I am learning how to have this conversation. So, I may ask you if the things I say make sense.
BOT: Would you like to chat with me about your smoking?
USER: sure thing
BOT: Great!
BOT: To start, what is the thing you like most about smoking?
USER: the quiet moments
BOT: You feel that smoking helps you relax.
BOT: Did that make sense?
USER: yes
BOT: That's great to hear, thanks for letting me know!
BOT: Now, what is the thing you like least about smoking?
USER: it gives me too much motivation to keep doing it
BOT: You feel as if your smoking habit is holding you back.
BOT: Did what I said make sense to you?
USER: no
BOT: I am sorry to hear this, thank you for clarifying!
BOT: Now, what is one thing about your smoking habit that you would like to change?
USER: i'd like to be less judged by doing it
BOT: You don't want people to judge you for your habit.
BOT: Did that make sense?
USER: yes
BOT: Great, let's continue
BOT: What will it look like when you have made this change in your smoking habit?
USER: i feel that when i do end up quitting smoking, it's out of necessity or out of caring too much about someone
BOT: You're worried that if you smoke, someone will be hurt or killed by the effects of your smoking.
BOT: Does this make sense to you?
USER: more or less
BOT: Great, this lets me know I'm on the right track
BOT: Finally, what are the steps you need to take to make this change?
USER: i guess meet the right person ?
BOT: It's not always easy to find the right person.
BOT: Did that make sense?
USER: yes
BOT: Thanks, that is good to hear!
BOT: Thank you very much for this conversation, it was really helpful!

A.0.3 MI Chatbot Transcript Version 5.1 - Questions Reflections with Generator Version 2

BOT: Hello! My name is MIBot, I am a computer chatbot and part of a research project at the University of Toronto. I was made by a group of researchers and doctors, with the goal of exploring your smoking behaviours.
BOT: I will ask a few questions, listen to your answers, and then respond.
BOT: However, I am learning how to have this conversation. So, I may ask you if the things I say make sense.
BOT: Would you like to chat with me about your smoking?
USER: Sure.
BOT: Great!
BOT: To start, what is the thing you like most about smoking?
USER: stress relief
BOT: You enjoy the relief that smoking provides you.
BOT: Did that make sense?
USER: Yes.
BOT: That's great to hear, thanks for letting me know!
BOT: Now, what is the thing you like least about smoking?
USER: Health issues.
BOT: You are aware of the health risks associated with smoking and are concerned about your own health.
BOT: Did what I said make sense to you?
USER: Yes.
BOT: Thanks, I'm glad, this helps my listening skills
BOT: Now, what is one thing about your smoking habit that you would like to change?
USER: Smoking among friends.
BOT: You would like to stop smoking among your friends.
BOT: Did that make sense?
USER: Yeah.
BOT: Great, let's continue
BOT: What will it look like when you have made this change in your smoking habit?
USER: Stop smoking or smoking less
BOT: You will stop smoking completely.
BOT: Does this make sense to you?
USER: Yes.
BOT: Great, this lets me know I'm on the right track
BOT: Finally, what are the steps you need to take to make this change?
USER: Taking nicotine tablets.
BOT: You realize that you need to replace smoking with nicotine replacement therapy (NRT).
BOT: Did that make sense?
USER: Yea
BOT: Thanks, that is good to hear!
BOT: Thank you very much for this conversation, it was really helpful!

A.0.4 MI Chatbot Transcript Version 5.2 - Enhanced/Extended Conversation based on v5.1

BOT: Hello! My name is MIBot, I am a computer chatbot and part of a research project at the University of Toronto. I was made by a group of researchers and doctors, with the goal of exploring your smoking behaviours.
BOT: I will ask a few questions, listen to your answers, and then respond.
BOT: However, I am learning how to have this conversation. So, I may ask you if the things I say make sense.
BOT: Would you like to chat with me about your smoking?
USER: Yes please
BOT: Great!
BOT: To start, what is the thing you like most about smoking?
USER: stress relief
BOT: You enjoy smoking because it helps you cope with stressful situations.
BOT: Did that make sense?
USER: yes
BOT: That's great to hear, thanks for letting me know!
BOT: What else do you like about smoking?
USER: alone time
BOT: You enjoy spending time by yourself with cigarettes.
BOT: Did that make sense?
USER: yes
BOT: That's great to hear, thanks for letting me know!
BOT: Now, what is the thing you like least about smoking?
USER: the money
BOT: You dislike spending money on cigarettes.
BOT: Did what I said make sense to you?
USER: yes
BOT: That's great to hear, thanks for letting me know!
BOT: What else do you dislike about smoking?
USER: the addiction
BOT: You are aware that cigarettes can be addictive and you worry about how much you smoke.
BOT: Did that make sense?
USER: yes
BOT: Thanks, I'm glad, this helps my listening skills
BOT: Now, what is one thing about your smoking addiction that you would like to change?
USER: the amount I smoke
BOT: You want to reduce the amount of cigarettes you smoke.
BOT: Did that make sense?
USER: yes
BOT: Thanks, that is good to hear!

BOT: It's great to hear you want to reduce your smoking. What would it look like when you have reduced your smoking addiction?
USER: id be smoking 10 cigs per day
BOT: You'd smoke less than 10 cigarettes per day.
USER: yes
BOT: Understood, thank you for sharing.
BOT: What might be helpful towards reducing your smoking?
USER: not sure
BOT: You are unsure about what your life will look like after making this change.
BOT: Did that make sense?
USER: yes
BOT: Great, this lets me know I'm on the right track
BOT: Finally, what are the steps you need to take to make this change?
USER: try and reduce by one cig per day
BOT: You want to reduce your smoking by one cigarette per day.
BOT: Did that make sense?
USER: yes
BOT: Thanks, that is good to hear!
BOT: Thank you very much for this conversation, it was really helpful!

Appendix B

B.0.1 Chatbot Feedback and Confidence Change with Ambivalence Resolution Label

Experiment Name	Confidence Quit	Feedback	Final Label
MIV4.7	-6	Just made me realise i dont what to stop smoking right now	Smoke
MIV4.7	-3	It helped me understand that in order to Quit something one should plan and know the steps that one needs to take	Same
MIV4.7	-2	No. It was pretty fair	Same
MIV4.7	-1	It didn't because they were just normal questions and I already knew the answers	Same
MIV4.7	-1	No, because the questions asked are already things that i know and think about frequently, nothing really made me re-think my behaviour regarding smoking	Same
MIV4.7	-1	Oh yeah surerly that im addicted to it because i may hate myself internally as well as im "used" to it already in a way	Same
MIV4.7	-1	Not really. I already know (before the chat) that I have to smoke less.	Same
MIV4.7	-1	Not really. I am conscious about what it implies to smoke and I do it regard this. I think the conversation can help people who actually want to quit smoking, but It's not my case.	Smoke
MIV4.7	0	Yes it did. It made me think about my behaviour and how it impacts my health	Same
MIV4.7	0	It brought me closer to the idea that quitting smoking would do me good	Same
MIV4.7	0	No. The questions didn't really make me think. Kinda just made me feel irritated since it's the same questions everyone asks smokers.	Same

MIV4.7	0	Honestly no, the bot didn't do or said anything new nor remotely helpful	Same
MIV4.7	0	Somethings I already knew, it's a topic I think about monthly at least, but yes	Same
MIV4.7	0	no, too short and impersonal	Same
MIV4.7	0	Not really. Felt like an automatic response. Like I was speaking to myself and Ive done that a thousand times	Same
MIV4.7	0	Not actually, because they were basically just collecting data, but not giving me any feedback.	Same
MIV4.7	0	It helped me to realise I probably smoke more cigarettes than I truly need to, although it didn't help in any practical way for methods on how to cut down.	Same
MIV4.7	0	I don't really think it did. The conversation felt more like monologue, chatbot was asking questions and I was answering them. He didn't say anything interesting or new for me at all.	Same
MIV4.7	0	Yes..because the chatbox makes me feel uncomfortable	Same
MIV4.7	0	It did not. I was already very aware of the pros and cons that smoking brings into my life, as I have had a lot of introspection regarding that concern, in the past.	Same
MIV4.7	0	No, every thing that I have said I already have thought about	Same
MIV4.7	1	nope, it was just few questions	Same
MIV4.7	1	It makes me think that maybe it is better to stop, it points out to me but it is difficult to do so	Quit
MIV4.7	1	Not really. I know the Quits i need to make. its just hard to do them	Same
MIV4.7	1	Acutally it did. It brought out 2 main problems I have with smoking (money & drink-smoking)	Quit
MIV4.7	1	Yes. It made me realize that even though I am not ready to give up smoking yet, that there is in fact something I don't like about my smoking habit and that I already have the knowledge of how to Quit that.	Same

MIV4.7	1	Yes, i need to think more about stopping smoking for my family's sake as well as my own.	Quit
MIV4.7	1	This is how I began to analyze how much I smoke, how much I spend on cigarettes and under what conditions I smoke less or not at all.	Quit
MIV4.7	1	Not really but only because this is something I have been aware for a while now and I think about it quite regularly but it feels like for some reason I am not willing to do it.	Same
MIV4.7	1	nothing i didn't know before, it is a good opportunity for reflection. It helps you to think about things and say things you might avoid in fear of others expectations or failure.	Same
MIV4.7	1	Bot only asked questions and didnt provide any solutions so no, the chatbot didnt make me realize anything about my smoking behavior	Same
MIV4.7	1	not really	Same
MIV4.7	1	I realized a little that I should try harder to smoke less.	Quit
MIV4.7	1	It highlighted concerns that have been sitting in the back of my mind about my smoking behavior, it brought them to the forefront of my mind.	Quit
MIV4.7	1	Only what I already knew: that there's no advantages that overpower the disadvantages	Same
MIV4.7	2	Yes, it made me reflect about it and decide where to start changing my habit	Quit
MIV4.7	2	Yes, because from the questions he asked me, I realized that I can try to quit smoking and that it is willpower.	Quit
MIV4.7	2	The conversation makes you think about what you need to do to find the willpower to do good things for your health...	Quit
MIV4.7	2	Not much but I would think about that	Same
MIV4.7	2	Yes, because i had to face that im an addict	Quit
MIV4.7	2	not really, I already feel aware of my smoking behaviors	Same
MIV4.7	2	Not really.	Same
MIV4.7	3	That I'm aware that I have a problem.	Quit

MIV4.7	3	I believe a bit yes because nowadays people (including me also) don't think about the addiction so it's an excellent method (like this chat) to remind ourselves about the existing problem	Quit
MIV4.7	3	yes it made me understand even more that wanting is power	Quit
MIV4.7	3	yes, how bad it is and the purpose there of	Quit
MIV4.7	4	Yes, it helps to talk to someone or the Bot about addiction	Quit
MIV4.7	4	yes, it did because to be honest I was smoking when I received this survey	Quit
MIV4.7	4	No , my answers were things I've already realized.	Same
MIV4.7	5	Yes, sometimes a simple question can really put things into perspective	Quit
MIV4.7	6	It didn't make any difference, It was more of a questionnaire I think it should have given me more reasons to help me quit smoking but I do get were it's coming from. Hopefully, I'll be able to Quit my behavior in a week.	Same
MIV4.7	7	Nothing more than I already knew	Same
MIV5.0	-4	I hadn't questioned why i've been chain smoking lately and it might me confront that.	Same
MIV5.0	-3	It did. It made me think about my smoking behavior and I think I'm going to spend more time on wondering how to stop smoking.	Same
MIV5.0	-3	Yes i need to stop and i need to start now	Same
MIV5.0	-2	It made me realise that like smoking for some reasons and hate it for others. It all comes down on my own priorites	Same
MIV5.0	-2	No, because I'm fully aware of my smoking behavior/habit	Same
MIV5.0	-1	Yes and no. I already pretty much knew everything that I spoke about. It just was funny me answering aloud (via text) what I keep in the back recesses of my mind mostly.	Same
MIV5.0	-1	Not a lot ,Just a reminder that is a bad habit and do a lot of harm , but there isn t a new info	Same

MIV5.0	-1	Not really. I know is not the healthiest thing, but is not something that I'm looking to Quit right now. I feel like there are other things more important to work on in my life.	Smoke
MIV5.0	-1	I realized what's my least favorite part in my habit, thanks to some specific questions about it.	Same
MIV5.0	-1	no, i already know	Same
MIV5.0	-1	No, in the sense that I was already aware of the answers to the questions asked regarding motivation etc	Same
MIV5.0	-1	It made me realize I haven't really thought a lot about the reasons why I smoke besides that is something that I have been doing for years. Also, it kinda make me feel bad when it refered to my smoking as "your problem" or something like that. It was like I DON'T HAVE A PROBLEM. But maybe I do?	Same
MIV5.0	0	Yes. By asking what I liked the most about smoking and what I less liked, it made be think consider if what I like the most is more important than what I like less.	Same
MIV5.0	0	I didn't help me really, but it opened my eyes on how easily i could get help quitting if i really wanted to	Same
MIV5.0	0	It made me realize that I already know some of the steps needed to quit. Now it is up to me to follow through.	Same
MIV5.0	0	It didn't help me understand anything better because it was only trying to get me to stop smoking, not actually being introspective at all. The questions were extremely open-ended and it struggled to understand complex sentences anyway, and since people are complex it's pretty essential that it can understand things. If its purpose is to simply reframe what somebody says into a biased "oh so you're trying to stop for this reason right!!!?!?!???1!!!" then it will never reveal anything the person doesn't already know about themselves.	Same

MIV5.0	0	Not really, it would've felt better to know i was speaking to a real person. Maybe changing the name of the bot could help a little bit.	Same
MIV5.0	0	Kind of, it helped my question myself on why do i actually smoke, what makes me want a cigarette	Same
MIV5.0	0	yes, that I can Quit my smoking habit if I really want to.	Same
MIV5.0	0	no I can't really say I realized anything	Same
MIV5.0	0	It helped me reafirm that I'm not ready to quit somking because when the chatbot asked me if I knew where to begin changing my habits, I didn't know how to answer.	Same
MIV5.0	0	I don't think it did because sometimes the chatbot didnt understood what i meant but maybe that was my problem maybe cuz i wasnt clear sometimes	Same
MIV5.0	0	No, I still feel the same.	Same
MIV5.0	0	In the moment yes, but i think quitting smoking is a longer process in your mind than the conversation i just had with the lovely chatbot. Again the question raised were interesting and made me reflect on my cigarette count, but that is something i already think about on the daily and try to manage	Same
MIV5.0	0	no, because i was aware of my smoking behavior before hand.	Same
MIV5.0	0	It didn't	Same
MIV5.0	0	not really but that's because i already interiorized the faults in my smoking behaviour	Same
MIV5.0	0	Yes, because I know its not a good habit	Same
MIV5.0	0	No, the questions were mostly things that i have thought about before	Same
MIV5.0	0	It made me realise that if I really wanted to, I could perhaps substantially reduce the frequency of my smoking simply by cutting out every other cigarette.	Same
MIV5.0	0	No, these are things I've already considered.	Same
MIV5.0	0	it as not, i have wanting to stop smoking a long time ago but i can never do it	Same
MIV5.0	0	yes	Same

MIV5.0	0	Not really, as I mentioned in the previous answer it would just point out what I described myself using more words. But on the other hand, it felt good in a weird way admitting my concerns to someone-something about smoking.	Same
MIV5.0	0	Not really, it didn't give me any new information.	Same
MIV5.0	0	No, because, while it made me reflect on it, I didn't find out anything surprising	Same
MIV5.0	0	It made me think about why I smoked and why I would want to give up.	Same
MIV5.0	0	Yes. It reminded me why I want to quit.	Same
MIV5.0	0	No. It only gave him information, it didn't provide much for me	Same
MIV5.0	0	not really, just highlights what i already know	Same
MIV5.0	0	The questions were rather very simple and typical, nothing exceptional. All I could write was what I already know and become more aware of the problem. This is also important.	Same
MIV5.0	0	No, I already had those conclusions in mind	Same
MIV5.0	0	Not really, it just listened to everything i said, just that	Same
MIV5.0	0	no, because of the reason I've just written. also: there's nothing to realize about smoking. we know it's costly, we know it's damaging our health. you won't find a time traveler from the 20es that will go "oh wow, what are you saying??? smoking is actually BAD for you??!?"	Same
MIV5.0	0	it did not Quit anything. I am aware that smoking is bad for my health and I would quit it if I wanted to. I quit it for a year about 2 years ago and I could do it again only if I wanted to.	Same
MIV5.0	0	No. Because I'm the type of user who doesn't Quit her opinion on smoking after a random conversation, be it online or in real life	Same
MIV5.0	1	no. i already know the problem i have	Same
MIV5.0	1	Nothing precise. Only that I don't want to quit smoking right now.	Same

MIV5.0	1	it did, that i sohuld seek someone i care for so much im willing to stop smoking jsut so i can life better with that person, which at the end of the day is a good food for thought, thanks	Quit
MIV5.0	1	no, it was to basic	Same
MIV5.0	1	No. It was a real short conversations, and I feel the bot did not understand me very well.	Same
MIV5.0	1	yes, that my smoking habit is linked with social events and conforming wiyh my peers	Quit
MIV5.0	1	yes, because analyze what could be the reasons why it would be worth quitting than health and expenses	Quit
MIV5.0	1	no, didnt offer any solutions/methods	Same
MIV5.0	1	I just reminded myself to cut down the number of cigarettes I consume daily.	Quit
MIV5.0	1	not particularly, stressful situations do bring up my anxiety which does make me smoke more often	Quit
MIV5.0	1	Yes it did. Namely what I like and don't like about smoking.	Quit
MIV5.0	1	Yes, it did not mention anything specific but the fact that it repeated what I was saying in a different manner really brought my goals to my attenttion	Quit
MIV5.0	1	No because i already know what's wrong with my smoking problem. Bot didn't tell me anything innovative	Same
MIV5.0	1	yes it did ,because it helped me realize and remember most importantly the bad thing smoking does to my health it putted from unconscious to conscious	Quit
MIV5.0	1	yes, similar suggestions	Quit
MIV5.0	1	I think it kinda made me understand that if I met with people who don't smoke as much I would not be smoking as much either	Quit
MIV5.0	1	No - it made it clear I had no clue where to start	Same
MIV5.0	1	Not so much, as I have been going over these things in my mind lately	Same
MIV5.0	1	It did. While thinking about the answers to the questions, i realised that my smoking is problematic for me.	Quit

MIV5.0	2	Yes, i understand that i know what i have to do. I'm just lazy	Quit
MIV5.0	2	Yes it made me realize that I need to quit smoking and find better ways to cope with life	Quit
MIV5.0	2	Yes, it reminded me of how unhealthy smoking is to me	Quit
MIV5.0	2	yes	Same
MIV5.0	2	It did because I'm actively talking about the issue in itself but the convo in hand didn't help much in that regards. (as i mentioned in the previous answer)	Same
MIV5.0	2	No, the responses weren't in depth enough	Same
MIV5.0	2	No, I already understood why I smoke, for the robot I just had to formulate it	Same
MIV5.0	2	Not really. I think I had the information in my head already and the bot just gives you a means to reflect on it.	Same
MIV5.0	2	Yes, I need to start changing my habits. I need to stop smoking or work towards reducing my intake.	Quit
MIV5.0	2	Not really, but had very optimistic and enthusiastic responses which made me feel a little more optimistic to quiting	Same
MIV5.0	2	YES, MADE BE MORE AWARE ABOUT MY SMOKING HABIT	Quit
MIV5.0	3	It actually made me realise I can quit smoking if I am determined enough because I really want to be able to participate in activities I enjoy.	Quit
MIV5.0	3	It wasn't an enlightening chat, honestly	Same
MIV5.0	3	nothing especially new but yes, I really need to use my "fake cigarette" which is really the best thing I ever tried till now to give up smoking; be always motivated and not only time to time - I don't like this character in other people but, finally, regarding giving up smoking, I am exactly the same! I am almost disgusted by that discovery	Quit
MIV5.0	3	That I need to stop smoking	Quit
MIV5.0	3	not really. it didnt give me any tips	Same

MIV5.0	3	Yes, because I have never asked myself those questions with the intention of changing my habit.	Quit
MIV5.0	4	Yes. It made me realise that I need to find a substitute for smoking that doesn't affect my health negatively.	Quit
MIV5.0	4	Not really, I was already aware that I'm choosing to behave in a way that won't benefit me in the future and I already knew what are the reasons behind my habit.	Same
MIV5.0	4	Not that much becouse I know I can't do it	Same
MIV5.0	4	Yes, because I really want to quit smoking	Quit
MIV5.0	4	i don't know, maybe a little bit	Same
MIV5.0	4	Totally. The answers that I gave to the bot were not concise and didn't have any meaningful background to me, but after reading the suggestions I started thinking that those suggestions were a kind of thought that I had but at this moment I hadn't had.	Quit
MIV5.0	4	Not really. The bot seemed to just summarize what I was saying to it. This is impressive on a technical level, but it didn't offer me anything constructive with regards to my habit.	Same
MIV5.0	4	Not much	Same
MIV5.0	4	Yes, i realised i smoke a lot	Quit
MIV5.0	5	yes, it makes it real , you have to see the real problem	Same
MIV5.0	5	It reminded me that I should keep going because I still smoke too much	Same
MIV5.0	6	No	Same
MIV5.0	6	Not much, I told things that I knew before.	Same
MIV5.0	6	In part, yes, although the chatbot did not try to motivate me, but asked questions. On the other hand, this conversation made me realize that I could do it.	Quit
MIV5.0	7	I realize that i can saves a lot of money if i start to smoke less	Quit
MIV5.0	9	No not really, I already know smoking is horrible for me	Same
MIV5.1	-7	not at all. i love smoking, but was funny	Same
MIV5.1	-3	not really, i knew all that already	Same

MIV5.1	-2	No, because everyting about this conversation is on mind since ever	Same
MIV5.1	-2	Nothing new that I didn't know before. I'm aware of the fact that most of the cigarettes in the week aren't enjoyable. But I'm smoking because 1 of 10 feels good, especially at the parties etc.	Smoke
MIV5.1	-2	Yes, for a long time I didn't have conversation about my smoking habits and how many cigarettes I smoke daily.	Same
MIV5.1	-2	No. It's just a personal decision, no chatbot can help	Same
MIV5.1	-2	No not really - it did make me want to smoke though	Smoke
MIV5.1	-1	Not particularly. Because it just repeated my answers back to me.	Same
MIV5.1	-1	No, the chatbot repeated information I gave back to me. All the things I wrote were already conclusions I had reached on my own.	Same
MIV5.1	-1	no, i already know i should make Quits but not ready now	Smoke
MIV5.1	-1	I think i should find another device to relax instead of smoking	Same
MIV5.1	-1	The conversation was pleasant, however, the chatbot repeated my theses. It was a positive experience, but it did not add much to the perception of my smoking.	Same
MIV5.1	-1	helped me realize how quickly I reach for a cigarette after waking up	Same
MIV5.1	-1	Yes :) it helped me think about the steps I would have to take if I wanted to quit smoking	Same
MIV5.1	-1	yes, it made me realize that if i stop smoking ont he pc it will stop rather quick	Same
MIV5.1	-1	I already knew why I smoke and also why should I stop. Didnt realize anything new.	Same
MIV5.1	-1	no, i just repeated the things i already knew	Same
MIV5.1	-1	Yes. It made me realize that I need more control over my smoking habits.	Same
MIV5.1	-1	It only made me remember stronger my vice that I have.	Smoke

MIV5.1	-1	maybe would make me think of looking into the costs. but i wouldnt say it made me commit to stop smoking or cut down	Same
MIV5.1	-1	No because I have discussed and thought about my smoking for decades. The chatbox didn't enlighten me further.	Same
MIV5.1	0	Yes ... it confirmed that it's a habit I need to break.	Same
MIV5.1	0	not really. it was quite short	Same
MIV5.1	0	I was already aware of all of the things I stated however it was helpful to write it down	Same
MIV5.1	0	Nothing I didn't already know I must say. I know smoking is a very bad habit and the faster you quit, the better. Still, it didn't make me realize anything in particular, but it was a reminder that I should quit smoking as soon as possible.	Same
MIV5.1	0	not really	Same
MIV5.1	0	Yes. I have put into words what I think almost every day.	Same
MIV5.1	0	No, because I already knew all those things	Same
MIV5.1	0	No,, i already knew everything i said	Same
MIV5.1	0	It just reminded me of things I had already considered	Same
MIV5.1	0	No, but I understand it's purpose in trying to help people stop smoking.	Same
MIV5.1	0	yes it did i have to be the one to do it	Same
MIV5.1	0	yes, gave me an insight of the dangers and made me realize that i am capable of quitting this behavior	Same
MIV5.1	0	No, the bot wasn't very profound or intuitive.	Same
MIV5.1	0	yes - that is my copying mechanism	Same
MIV5.1	0	I have thought about quiting lots of times so i probably did not lealize anything new.	Same
MIV5.1	0	no	Same
MIV5.1	0	Noy really. Noy anything I already know.	Same
MIV5.1	0	A little, seeing words written made me realize that i am a slave of myself	Same
MIV5.1	0	No, as the bot only confirmed what I've already written down, so no breakthrough.	Same
MIV5.1	0	Yes, it put it in another words so I could see it differently	Same

MIV5.1	0	No. It wasn't able to understand my point of view and it didn't add anything, nor did it provide me with any new information.	Same
MIV5.1	0	Made me realise the government will never make it legal	Same
MIV5.1	0	YES I REALIZED THAT I DON'T WANT TO QUIT SMOKING. AS BAD AS IT IS	Same
MIV5.1	0	Yes, it made me realize that I smoke for stress, to relax.	Same
MIV5.1	0	Yes smoking is linked to stress	Same
MIV5.1	0	Maybe, I just tried quit smoking a little while and I couldnt	Same
MIV5.1	0	No, I have a feeling that nothing Quitd.	Same
MIV5.1	0	Yes, it helped me to realize that I have a couple of priorities in changing my smoking behaviour.	Same
MIV5.1	1	Honestly? It was just nice to talk to someone who understood what I was saying for once. But I didn't make any realizations. The bot is amazing though.	Same
MIV5.1	1	Momentarily, yes. I ask myself similar questions from time to time, but lack the strong will and motivation to quit smoking cigarettes	Quit
MIV5.1	1	It didn't because I don't believe a chatbot could motivate me	Same
MIV5.1	1	no because it just reiterated what I had said previously	Same
MIV5.1	1	It made me realize that I shouldn't and that i can't because it is an addiction with certain habits	Quit
MIV5.1	1	It didn't, It just asked me questions I already knew the answer to and proceeded to say the same things with other words	Same
MIV5.1	1	conversation did not help, because the bot only "understood" my problems but did not help in any way to solve them	Same
MIV5.1	1	so helped me to see more clearly how this phenomenon negatively affects my life	Quit
MIV5.1	1	Yes, it helped me see the bad consequences and effects of it	Quit
MIV5.1	1	Wasn't very detailed, so didn't really make me think about quitting	Same

MIV5.1	1	It didn't, I just answered questions I was asked	Same
MIV5.1	1	no. it was, for me, just a questionaire	Same
MIV5.1	1	quite honestly not, as I was only asked about my smoking habits. Since I work in the health area, I know perfectly well the harm that tobacco does to me and to make any Quit in this habit it has to come from myself because I have always known that it is not good for me.	Same
MIV5.1	1	how much of my "peer pressure" was from me and not from my friends themselves	Quit
MIV5.1	1	Nope, nothing was highlighted really that I didn't already know. I was mostly informing the chatbot	Same
MIV5.1	2	Not really, I thought about those questions before	Same
MIV5.1	2	not really, it did not include anything that I did not already know or give me ideas about quitting smoking	Same
MIV5.1	2	yes because the chatbot said logical things for my health	Quit
MIV5.1	2	Not really, im aware of How my whole situation looks and for now its the best option for me	Same
MIV5.1	2	Yes it thought me that I need to reduce smoking.	Quit
MIV5.1	2	No, nothing at all. The chatbot only really responded in order to clarify, it suggested nothing to help with quitting.	Same
MIV5.1	2	Yes, it helped me realize that I often reach for a smoke when I get filled with anxiety. I really only smoke to help me cope with daily anxiety and restlessness. It gives me something to do with my hands which makes me feel like I am accomplishing something.	Quit
MIV5.1	3	It kinda did but it was mostly just confirming the things I said	Same
MIV5.1	3	Yes. Self analysis has made me understand.	Quit
MIV5.1	3	A little bit. I know more what to do.	Quit
MIV5.1	3	no because i alrwady knew all these	Same
MIV5.1	3	yes, i am more concious about my health	Quit
MIV5.1	3	no, the bot just rephrase my thinking	Same

MIV5.1	3	not so much, they were things I knew and realized before	Same
MIV5.1	3	Nothing I haven't already put together myself, I've already done a lot of self reflection over this so it has nothing to do with the bot.	Same
MIV5.1	3	its sunk in more how addicted i am	Quit
MIV5.1	3	it didn't , because the chatbox tried to understand my behavior, but didn't comment on it	Same
MIV5.1	3	Not really, there were things that i have already known, so i hadn't learnt anything new about my smoking behavior	Same
MIV5.1	4	It made me realise what I like and don't like about smoking and that I can do small things to reduce smoking	Quit
MIV5.1	4	Not really, there was not any advice	Same
MIV5.1	4	yes, that I deep down do not want to stop smoking yet.	Same
MIV5.1	4	Not really, because this is a conversation that I have with myself daily	Same
MIV5.1	4	Yes, he has a remorse that I smoke	Same
MIV5.1	4	Yes of course it helps, he gave me advice to get what I want.	Quit
MIV5.1	4	Yes, that I would feel better if I stopped smoking	Quit
MIV5.1	5	One on hand , it did help me to listen to my thoughts and realise them. On the other hand, other than that, it did not provide any actual and useful solutions to my smoking problem.	Same
MIV5.1	5	It was helpful just to talk and realize how I feel about my own smoking.	Same
MIV5.1	5	I need time to endure I follow through	Same
MIV5.1	5	Kind of. It put what I want and need into perspective.	Quit
MIV5.1	6	It made me think more seriously about my desire to quit	Quit
MIV5.1	6	It made me think I have to stop smoking. Look for ways to do so.	Quit
MIV5.1	6	The conversation didnt Quit how I felt however thinking about smoking more and thinking about quitting makes me feel more guilty about the fact i havent made an effort to reduce my smoking habits	Quit

MIV5.1	7	Not alot like I said I was just starting to think about it	Same
MIV5.1	7	Yes. That I can set a goal on how many I can smoke per day.	Quit
MIV5.1	7	Partly, because i already know what to do, it is just hard to stop smoking.	Same
MIV5.2	-6	That i might not have the willpower and the mentality to cut it off so easily	Smoke
MIV5.2	-4	Yes that I want to make a Quit, that I am not ready to quit and that it is sad that one of the few things in my life at the moment that make me feel comfort is smoking.	Smoke
MIV5.2	-3	It just helped me to think about quitting smoking again, but although I want to, at this point in my life it is very difficult.	Smoke
MIV5.2	-2	Yes. That I smoke even when I don't need it.	Same
MIV5.2	-2	nope, it was just simple smalltalk on what could be done for me to stop smoking	Same
MIV5.2	-1	Not really, the conversation merely confirmed what I knew already, it did not bring up anything new	Same
MIV5.2	-1	It seemed to repeat things i already knew but in a medical context and i'm hyper aware already so its hard to know what to do next, it tried to help but it didnt tell me things i wasn't already thinking about, my anxiety and pain are severe and without guidance im lost	Smoke
MIV5.2	-1	it didn't make me realise anything new, the chat bot simply repeated my points and gave some basic input.	Same
MIV5.2	-1	no, im fully aware of my addiction and behaviors connected with it	Smoke
MIV5.2	-1	Nothing at all.	Same
MIV5.2	-1	yes, because it made me rethink some of my behaviors torwards smoking	Same
MIV5.2	-1	ye a bit, bot countinously asked me questions that made me think about that for a couple of seconds and make good arguments based on my responses	Same
MIV5.2	-1	i already thought about that but telling it made me think more	Same

MIV5.2	-1	I think yes. I found out that probably i smoke more than regular smoker	Same
MIV5.2	-1	No, because even though I would like to smoke less, I don't want to quit because it helps me	Smoke
MIV5.2	-1	Not really, because those were all conclusions I had already reached by myself	Same
MIV5.2	0	It helped me realize I need to start to try and understand myself better so I can quit bad habits such as smoking	Same
MIV5.2	0	that I smoke a little bit more than I thought	Same
MIV5.2	0	It did not, I know my smoking is a terrible habit and I know that I need to stop, I have tried multiple times with a variety of methods without luck. There was no further information given to me through the chat bot that I don't already know.	Same
MIV5.2	0	No, It only helped to confirm what I already knew about my smoking, but did not give me any new insight	Same
MIV5.2	0	A little, it made me think that I smoke too much	Same
MIV5.2	0	Yes, that the central problem is my anxiety and that I should prioritize that.	Same
MIV5.2	0	I'm pretty self aware of my smoking habits, so I don't think I gained any new insight into it. I already know quite well what I should do and even how to achieve my goal of reducing smoking, but I feel like I lack the motivation to do so. I'm not sure if the chatbot actually suggested anything, it only asked questions, which to be fair, can be quite useful for others.	Same
MIV5.2	0	It did so in a relevant way.	Same
MIV5.2	0	Nothing I am not already thinking.	Same
MIV5.2	0	It helped me realize that smoking offers me nothing at all really	Same
MIV5.2	0	not really, there are things that already know	Same
MIV5.2	0	I don't think so. I know I wanna quit smoking, the conversation didn't help me to realise that	Same
MIV5.2	0	Yes, it made me realize the emphasis stress plays in my smoking habits.	Same
MIV5.2	0	I think so. It made me think about my smoking	Same

MIV5.2	0	Yes. It helped me realize I need a strategic plan to stop smoking for good.	Same
MIV5.2	0	No, im very self aware of my habbit, im aware of the health problems it may give me and the waste of money it is	Same
MIV5.2	0	not really	Same
MIV5.2	0	Yes, it made me realize that i don't know which steps to follow in order to help with quitting my smoking	Same
MIV5.2	0	yes, that I actually go outside to get some fresh air and a Quit of scenery	Same
MIV5.2	0	yes, i should smoke less	Same
MIV5.2	0	It helped me to realize that smoking is having an effect on my health and on my finances because I had to take the time to think of ways that smoking is affecting my life and things I can do to quit smoking	Same
MIV5.2	0	yes made me realize some things about my smoking habit	Same
MIV5.2	0	I'm generally very aware about my smoking habit, but it was interesting conversing about it	Same
MIV5.2	1	Not so much	Same
MIV5.2	1	More or less. I already know that I should stop smoking, but the conversation didn't gave me any reasons or ideias to stop	Same
MIV5.2	1	No, as I'm already aware of the cons of smoking.	Same
MIV5.2	1	Yes it was helpful to rationalise why I smoke and what should I do to start smoking less	Quit
MIV5.2	1	Not really, I am mindfull of my addiction	Same
MIV5.2	1	Not much; there were questions inquiring about habits I have already acknowledged	Same
MIV5.2	1	Yes it did, I had to think about the reasons of my addiction	Quit
MIV5.2	1	The conversation makes me realize that i am very addicted.	Quit
MIV5.2	1	That just reminds me why I want to quit.	Same
MIV5.2	1	it did not help	Same
MIV5.2	1	Yes, it made me realize I need to Quit	Quit
MIV5.2	1	no, I already knew	Same
MIV5.2	1	no i did not realize anything new	Same

MIV5.2	1	yes, it made me give deeper thought into my habits and the effects it has in other parts of my life	Quit
MIV5.2	1	i realized im surrounded by the wrong crowd	Quit
MIV5.2	1	Not really.	Same
MIV5.2	1	No, I have been a regular smoker for 64 years and know all about it	Same
MIV5.2	1	I think it could give me some courage to do something and finally manage to stop smoking. But it didn't really give me any information that I didn't knew already.	Quit
MIV5.2	1	Not really , I knew I need to set goals	Same
MIV5.2	1	No. I am aware I should quit. I am aware of what I like and what I don't like. Moreover, my addiction is not about nicotine, purely about the act - I didn't light a single cig for a week now because it's cold and smoking would be unpleasant.	Same
MIV5.2	1	no, because I already knew this behavior	Same
MIV5.2	2	Yes, it helped me understand why do I smoke	Quit
MIV5.2	2	Not really, as I stated before i felt like the chat bot just repeated whatever i told it adding no extra input, in a way it just felt like I was talking in the mirror.	Same
MIV5.2	2	Yes, if I do things gradually it is probably achievable.	Quit
MIV5.2	2	Yes, I have realized that I smoke too much and if I don't quit I might die of lung cancer or disease.	Quit
MIV5.2	2	No, because I am already aware of my smoking habit	Same
MIV5.2	2	It didn't help me realize anything because i was aware of consequences of my habit, but it definetly make me think about quitting smoking	Quit
MIV5.2	2	No. I had already realised that I have a smoking addiction	Same
MIV5.2	2	He just copied most of the things i wrote, so I felt like I was talking to myself	Same
MIV5.2	2	Yes, it made me realise things I was not really taking into consideration when I am smoking.	Quit

MIV5.2	2	Perhaps, the conversation begins to make me think about this vice.	Same
MIV5.2	2	Yes, that quitting smoking is a life-long process and it will need to be a gradual Quit.	Quit
MIV5.2	2	Not really, it's something I have been monitoring for a while now	Same
MIV5.2	3	Yes it made me realize that smoking is a bad habit through the questions that I was asked	Quit
MIV5.2	3	Yes. It helped incorporate the ideal that the Quit I'll have in my life from quitting smoke is for the best.	Quit
MIV5.2	3	Not really, I am already aware of my bad habits and the health factors associated with it, nothing really new was stated	Same
MIV5.2	3	Yes it did, I have realized that now I smoke more than I did a month ago.	Quit
MIV5.2	3	Yes, it made me realise I do it for the constant need of pleasure, which hadn't occurred to me	Quit
MIV5.2	3	Yes a bit, i noticed i don't have many things that i like about smoking, it's only addiction.	Quit
MIV5.2	3	maybe to reflect about that and aspire to improve and to a better situation	Quit
MIV5.2	3	Kinda, now i still think about my girlfriend that i want her to quit smoking too. It's make it kinda easier mentally.	Quit
MIV5.2	3	It helped me realise that as soon as i quit the habit of social smoking , i will be one step closer to quit smoking once and for all.	Quit
MIV5.2	3	No. It didnt gave me any advice how to stop smoking	Same
MIV5.2	3	I was completely aware of my behavior already	Same
MIV5.2	3	Yes	Same
MIV5.2	3	no, because it suggested nothing it was just my own answers	Same
MIV5.2	3	It was nice to type out my thoughts and be asked about my habits.	Quit
MIV5.2	3	Yeah, I have the courage to quit it, but after some time I get back by the feeling of calm that the nicotine prodeces on me.	Same
MIV5.2	4	yes it made me realise that i can quit and that im not really keen on smoking anyway im just addicted	Same

MIV5.2	4	it elaborated what I meant in a kind way	Same
MIV5.2	4	Yes. It did help me realizing how bad I should leave this addiction behind.	Quit
MIV5.2	4	No, not really the conversation was not that deep, but afterwards i was more determined to stop and to give it a try.	Quit
MIV5.2	5	Well since I am trying to quit, I already make sense of my smoking behaviors.	Same
MIV5.2	5	Yeah, mine is very addictive and it needs to stop	Quit
MIV5.2	5	Yes because now i have realized that it is easy stop smoking if you willing to	Quit
MIV5.2	5	Yes. I only realized now that smoking is a expensive habit.	Quit
MIV5.2	6	Yes. The questions made me think about smoking as an addiction and that's something I don't usually consider	Quit
MIV5.2	6	It was interesting to see it in writing but the chat bot (as lovely as they were) simply rephrased/clarified what I had just told them so I didn't learn anything new.	Same
MIV5.2	6	No, it just aske me questions	Same
MIV5.2	8	yes, It made me realise my smoking habits are not good and need to be Quitd.	Quit
MIV5.2	10	Yes it did because it helped me do some self introspection .	Quit

Appendix C

C.0.1 Consultation and Relational Empathy Survey

Feedback Questionnaire

Please rate the following statements about today's conversation.

Please click the box for each statement and answer every statement.

How was MIBot at ...

1. Making you feel at ease...

(being friendly and warm towards you, treating you with respect; not cold or abrupt)

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

2. Letting you tell your "story"...

(giving you time to fully describe your illness in your own words; not interrupting or diverting you)

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

3. Really listening...

(paying close attention to what you were saying)

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

4. Being interested in you as a whole person...

(asking/knowing relevant details about your life, your situation, not treating you as "just a number")

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

5. Fully understanding your concerns...

(communicating that your concerns were accurately understood; not overlooking or dismissing anything)

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

6. Showing care and compassion...

(seeming genuinely concerned, connecting with you on a human level; not being indifferent or "detached")

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

7. Being Positive...

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

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

8. Explaining things clearly...

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

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

9. Helping you take control...

(exploring with you what you can to to improve your health yourself; encouraging rather than "lecturing" you)

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

10. Making a plan of action with you...

(discussing the options, involving you in decisions as much as you want to be involved; not ignoring your views)

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

Finish

4 of 4

Bibliography

[1] Health Canada. *Overview of Canada's Tobacco Strategy*. eng. education and awareness. Last Modified: 2022-12-19. May 2018. URL: https://www.canada.ca/en/health-canada/services/publications/healthy-living/canada-tobacco-strategy/overview-canada-tobacco-strategy.html (visited on 01/26/2023).

[2] Health Canada. *Canadian Tobacco Alcohol and Drugs (CTADS) Survey: 2017 summary*. eng. surveys. Last Modified: 2021-08-12. Oct. 2018. URL: https://www.canada.ca/en/health-canada/services/canadian-alcohol-drugs-survey/2017-summary.html (visited on 01/26/2023).

[3] Health Canada. *Risks of smoking*. eng. education and awareness. Last Modified: 2016-05-17. Dec. 2011. URL: https://www.canada.ca/en/health-canada/services/smoking-tobacco/health-effects-smoking-second-hand-smoke/risks-smoking.html (visited on 06/15/2023).

[4] Stephen J. Wilson et al. "Ambivalence about smoking and cue-elicited neural activity in quitting-motivated smokers faced with an opportunity to smoke". eng. In: *Addictive Behaviors* 38.2 (Feb. 2013), pp. 1541–1549. ISSN: 1873-6327. DOI: 10.1016/j.addbeh.2012.03.020.

[5] W. F. Velicer et al. "Distribution of smokers by stage in three representative samples". eng. In: *Preventive Medicine* 24.4 (July 1995), pp. 401–411. ISSN: 0091-7435. DOI: 10.1006/pmed.1995.1065.

[6] William R. Miller and Stephen Rollnick. *Motivational Interviewing: Helping People Change*. en. Guilford Press, Sept. 2012. ISBN: 978-1-60918-227-4.

[7] Carolyn J. Heckman, Brian L. Egleston, and Makary T. Hofmann. "Efficacy of motivational interviewing for smoking cessation: a systematic review and meta-analysis". eng. In: *Tobacco Control* 19.5 (Oct. 2010), pp. 410–416. ISSN: 1468-3318. DOI: 10.1136/tc.2009.033175.

[8] Nicola Lindson et al. "Motivational interviewing for smoking cessation". In: *The Cochrane Database of Systematic Reviews* 2019.7 (July 2019), p. CD006936. ISSN: 1469-493X. DOI: 10.1002/14651858.CD006936.pub4. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6699669/ (visited on 06/15/2023).

[9] *Nicotine Replacement Therapy to Help You Quit Tobacco*. en. URL: https://www.cancer.org/cancer/risk-prevention/tobacco/guide-quitting-smoking/nicotine-replacement-therapy.html (visited on 07/01/2023).

[10] Robert West. “The clinical significance of ‘small’ effects of smoking cessation treatments”. en. In: *Addiction* 102.4 (2007). _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1360-0443.2007.01750.x, pp. 506–509. ISSN: 1360-0443. DOI: 10.1111/j.1360-0443.2007.01750.x. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1360-0443.2007.01750.x (visited on 01/26/2023).

[11] Joseph Weizenbaum. “ELIZA—a computer program for the study of natural language communication between man and machine”. In: *Communications of the ACM* 9.1 (Jan. 1966), pp. 36–45. ISSN: 0001-0782. DOI: 10.1145/365153.365168. URL: https://doi.org/10.1145/365153.365168 (visited on 01/26/2023).

[12] Daniel Jurafsky and James Martin. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Vol. 3. Feb. 2023.

[13] Laura Martinengo et al. “Conversational Agents in Health Care: Scoping Review of Their Behavior Change Techniques and Underpinning Theory”. EN. In: *Journal of Medical Internet Research* 24.10 (Oct. 2022). Company: Journal of Medical Internet Research Distributor: Journal of Medical Internet Research Institution: Journal of Medical Internet Research Label: Journal of Medical Internet Research Publisher: JMIR Publications Inc., Toronto, Canada, e39243. DOI: 10.2196/39243. URL: https://www.jmir.org/2022/10/e39243 (visited on 01/26/2023).

[14] Kathleen Kara Fitzpatrick, Alison Darcy, and Molly Vierhile. “Delivering Cognitive Behavior Therapy to Young Adults With Symptoms of Depression and Anxiety Using a Fully Automated Conversational Agent (Woebot): A Randomized Controlled Trial”. eng. In: *JMIR mental health* 4.2 (June 2017), e19. ISSN: 2368-7959. DOI: 10.2196/mental.7785.

[15] Timothy Bickmore, Amanda Gruber, and Rosalind Picard. “Establishing the computer-patient working alliance in automated health behavior change interventions”. eng. In: *Patient Education and Counseling* 59.1 (Oct. 2005), pp. 21–30. ISSN: 0738-3991. DOI: 10.1016/j.pec.2004.09.008.

[16] Ashish Vaswani et al. *Attention Is All You Need*. arXiv:1706.03762 [cs]. Dec. 2017. DOI: 10.48550/arXiv.1706.03762. URL: http://arxiv.org/abs/1706.03762 (visited on 01/26/2023).

[17] Alec Radford et al. “Language Models are Unsupervised Multitask Learners”. In: (2019). URL: https://www.semanticscholar.org/paper/Language-Models-are-Unsupervised-Multitask-Learners-Radford-Wu/9405cc0d6169988371b2755e573cc28650d14dfe (visited on 01/26/2023).

[18] Tom B. Brown et al. *Language Models are Few-Shot Learners*. arXiv:2005.14165 [cs]. July 2020. DOI: 10.48550/arXiv.2005.14165. URL: http://arxiv.org/abs/2005.14165 (visited on 01/26/2023).

[19] OpenAI. *Introducing ChatGPT*. en-US. URL: https://openai.com/blog/chatgpt (visited on 06/12/2023).

[20] OpenAI. *GPT-4 Technical Report*. arXiv:2303.08774 [cs]. Mar. 2023. URL: http://arxiv.org/abs/2303.08774 (visited on 06/12/2023).

[21] Archit Parnami and Minwoo Lee. *Learning from Few Examples: A Summary of Approaches to Few-Shot Learning*. arXiv:2203.04291 [cs]. Mar. 2022. URL: http://arxiv.org/abs/2203.04291 (visited on 06/26/2023).

[22] Yongqin Xian et al. *Zero-Shot Learning – A Comprehensive Evaluation of the Good, the Bad and the Ugly*. arXiv:1707.00600 [cs]. Sept. 2020. URL: http://arxiv.org/abs/1707.00600 (visited on 07/01/2023).

[23] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. *Distilling the Knowledge in a Neural Network*. arXiv:1503.02531 [cs, stat]. Mar. 2015. URL: http://arxiv.org/abs/1503.02531 (visited on 06/12/2023).

[24] Victor Sanh et al. *DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter*. arXiv:1910.01108 [cs]. Feb. 2020. URL: http://arxiv.org/abs/1910.01108 (visited on 06/28/2023).

[25] Xiaoqi Jiao et al. *TinyBERT: Distilling BERT for Natural Language Understanding*. Oct. 2020. URL: http://arxiv.org/abs/1909.10351 (visited on 06/28/2023).

[26] Wayne F. Velicer et al. "Smoking cessation and stress management: Applications of the transtheoretical model of behavior change". In: *Homeostasis in Health and Disease* 38 (1998). Place: Czech Republic Publisher: CIANS-Homeostasis, pp. 216–233. ISSN: 0960-7560.

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

[28] Ahson Saiyed et al. "Technology-Assisted Motivational Interviewing: Developing a Scalable Framework for Promoting Engagement with Tobacco Cessation Using NLP and Machine Learning". en. In: *Procedia Computer Science*. International Society for Research on Internet Interventions 11th Scientific Meeting 206 (Jan. 2022), pp. 121–131. ISSN: 1877-0509. DOI: 10.1016/j.procs.2022.09.091. URL: https://www.sciencedirect.com/science/article/pii/S1877050922009644 (visited on 01/29/2023).

[29] Alain Braillon and Françoise Taiebi. "Practicing "Reflective listening" is a mandatory prerequisite for empathy". en. In: *Patient Education and Counseling* 103.9 (Sept. 2020), pp. 1866–1867. ISSN: 0738-3991. DOI: 10.1016/j.pec.2020.03.024. URL: https://www.sciencedirect.com/science/article/pii/S0738399120301828 (visited on 01/29/2023).

[30] Ina Diener, Mark Kargela, and Adriaan Louw. "Listening is therapy: Patient interviewing from a pain science perspective". In: *Physiotherapy Theory and Practice* 32.5 (July 2016). Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/09593985.2016.1194648, pp. 356–367. ISSN: 0959-3985. DOI: 10.1080/09593985.2016.1194648. URL: https://doi.org/10.1080/09593985.2016.1194648 (visited on 01/29/2023).

[31] *Readiness Ruler*. en. Aug. 2021. URL: https://case.edu/socialwork/centerforebp/resources/readiness-ruler (visited on 01/26/2023).

[32] Eline Suzanne Smit et al. "Predictors of successful and unsuccessful quit attempts among smokers motivated to quit". en. In: *Addictive Behaviors* 39.9 (Sept. 2014), pp. 1318–1324. ISSN: 0306-4603. DOI: 10.1016/j.addbeh.2014.04.017. URL: https://www.sciencedirect.com/science/article/pii/S0306460314001191 (visited on 08/07/2023).

[33] Chad J. Gwaltney et al. "Self-Efficacy and Smoking Cessation: A Meta-Analysis". In: *Psychology of addictive behaviors : journal of the Society of Psychologists in Addictive Behaviors* 23.1 (Mar. 2009), 10.1037/a0013529. ISSN: 0893-164X. DOI: 10.1037/a0013529. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3829471/ (visited on 08/07/2023).

[34] Diane Von Ah et al. "Factors Related to Cigarette Smoking Initiation and Use among College Students". In: *Tobacco Induced Diseases* 3.1 (Dec. 2005), p. 27. ISSN: 2070-7266. DOI: 10.1186/1617-9625-3-1-27. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2643419/ (visited on 08/07/2023).

[35] Stewart W. Mercer et al. "The consultation and relational empathy (CARE) measure: development and preliminary validation and reliability of an empathy-based consultation process measure". eng. In: *Family Practice* 21.6 (Dec. 2004), pp. 699–705. ISSN: 0263-2136. DOI: 10.1093/fampra/cmh621.

[36] *The CARE Measure Website*. URL: https://caremeasure.stir.ac.uk/ (visited on 08/03/2023).

[37] Wayne Xin Zhao et al. *A Survey of Large Language Models*. arXiv:2303.18223 [cs]. May 2023. URL: http://arxiv.org/abs/2303.18223 (visited on 06/12/2023).

[38] Tomas Mikolov et al. *Efficient Estimation of Word Representations in Vector Space*. Sept. 2013. URL: http://arxiv.org/abs/1301.3781 (visited on 01/29/2023).

[39] Jeffrey Pennington, Richard Socher, and Christopher Manning. "Glove: Global Vectors for Word Representation". en. In: (2014), pp. 1532–1543. DOI: 10.3115/v1/D14-1162. URL: http://aclweb.org/anthology/D14-1162 (visited on 01/29/2023).

[40] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. http://www.deeplearningbook.org. MIT Press, 2016.

[41] Matthew E. Peters et al. *Deep contextualized word representations*. arXiv:1802.05365 [cs]. Mar. 2018. URL: http://arxiv.org/abs/1802.05365 (visited on 06/22/2023).

[42] Jacob Devlin et al. *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. arXiv:1810.04805 [cs]. May 2019. DOI: 10.48550/arXiv.1810.04805. URL: http://arxiv.org/abs/1810.04805 (visited on 01/26/2023).

[43] Yinhan Liu et al. *RoBERTa: A Robustly Optimized BERT Pretraining Approach*. arXiv:1907.11692 [cs]. July 2019. URL: http://arxiv.org/abs/1907.11692 (visited on 06/22/2023).

[44] Tianyu Gao, Xingcheng Yao, and Danqi Chen. *SimCSE: Simple Contrastive Learning of Sentence Embeddings*. arXiv:2104.08821 [cs]. May 2022. DOI: 10.48550/arXiv.2104.08821. URL: http://arxiv.org/abs/2104.08821 (visited on 06/22/2023).

[45] Yuxian Gu et al. *Knowledge Distillation of Large Language Models*. arXiv:2306.08543 [cs]. June 2023. URL: http://arxiv.org/abs/2306.08543 (visited on 06/28/2023).

[46] Cheng-Yu Hsieh et al. *Distilling Step-by-Step! Outperforming Larger Language Models with Less Training Data and Smaller Model Sizes*. arXiv:2305.02301 [cs]. May 2023. URL: http://arxiv.org/abs/2305.02301 (visited on 06/26/2023).

[47] Jones Karen Sparak. "A STATISTICAL INTERPRETATION OF TERM SPECIFICITY AND ITS APPLICATION IN RETRIEVAL". In: *Journal of Documentation* 28.1 (Jan. 1972). Publisher: MCB UP Ltd, pp. 11–21. ISSN: 0022-0418. DOI: 10.1108/eb026526. URL: https://doi.org/10.1108/eb026526 (visited on 08/21/2023).

[48] Yoshua Bengio et al. "A Neural Probabilistic Language Model". en. In: ().

[49] Alec Radford et al. "Improving Language Understanding by Generative Pre-Training". en. In: ().

[50] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. "Sequence to Sequence Learning with Neural Networks". In: 27 (2014). URL: https://proceedings.neurips.cc/paper/2014/hash/a14ac55a4f27472c5d894ec1c3c743d2-Abstract.html (visited on 06/23/2023).

[51] Veton Këpuska and Gamal Bohouta. "Next-generation of virtual personal assistants (Microsoft Cortana, Apple Siri, Amazon Alexa and Google Home)". In: (Jan. 2018), pp. 99–103. DOI: 10.1109/CCWC.2018.8301638.

[52] Ashwin Paranjape et al. "HINDSIGHT: POSTERIOR-GUIDED TRAINING OF RE- TRIEVERS FOR IMPROVED OPEN-ENDED GENERATION". en. In: (2022).

[53] Li Zhou et al. *The Design and Implementation of XiaoIce, an Empathetic Social Chatbot*. arXiv:1812.08989 [cs]. Sept. 2019. URL: http://arxiv.org/abs/1812.08989 (visited on 06/27/2023).

[54] Iulian Vlad Serban et al. *A Survey of Available Corpora for Building Data-Driven Dialogue Systems*. arXiv:1512.05742 [cs, stat]. Mar. 2017. URL: http://arxiv.org/abs/1512.05742 (visited on 06/27/2023).

[55] Jianping Gou et al. "Knowledge Distillation: A Survey". In: *International Journal of Computer Vision* 129.6 (June 2021). arXiv:2006.05525 [cs, stat], pp. 1789–1819. ISSN: 0920-5691, 1573-1405. DOI: 10.1007/s11263-021-01453-z. URL: http://arxiv.org/abs/2006.05525 (visited on 06/28/2023).

[56] Dongseong Hwang et al. *Comparison of Soft and Hard Target RNN-T Distillation for Large-scale ASR*. arXiv:2210.05793 [cs, eess]. Oct. 2022. URL: http://arxiv.org/abs/2210.05793 (visited on 06/28/2023).

[57] Adriana Romero et al. *FitNets: Hints for Thin Deep Nets*. arXiv:1412.6550 [cs]. Mar. 2015. URL: http://arxiv.org/abs/1412.6550 (visited on 06/28/2023).

[58] Sergey Zagoruyko and Nikos Komodakis. *Paying More Attention to Attention: Improving the Performance of Convolutional Neural Networks via Attention Transfer*. arXiv:1612.03928 [cs]. Feb. 2017. URL: http://arxiv.org/abs/1612.03928 (visited on 06/28/2023).

[59] Raphael Tang et al. *Distilling Task-Specific Knowledge from BERT into Simple Neural Networks*. arXiv:1903.12136 [cs]. Mar. 2019. URL: http://arxiv.org/abs/1903.12136 (visited on 06/28/2023).

[60] Lei Jimmy Ba and Rich Caruana. *Do Deep Nets Really Need to be Deep?* arXiv:1312.6184 [cs]. Oct. 2014. URL: http://arxiv.org/abs/1312.6184 (visited on 06/28/2023).

[61] Chang Liu et al. "Rethinking Task-Specific Knowledge Distillation: Contextualized Corpus as Better Textbook". In: (Dec. 2022), pp. 10652–10658. URL: https://aclanthology.org/2022.emnlp-main.729 (visited on 06/29/2023).

[62] Xuanli He et al. *Generate, Annotate, and Learn: NLP with Synthetic Text*. arXiv:2106.06168 [cs]. May 2022. DOI: 10.48550/arXiv.2106.06168. URL: http://arxiv.org/abs/2106.06168 (visited on 06/26/2023).

[63] Yizhong Wang et al. *Self-Instruct: Aligning Language Models with Self-Generated Instructions*. arXiv:2212.10560 [cs]. May 2023. URL: http://arxiv.org/abs/2212.10560 (visited on 06/12/2023).

[64] Long Ouyang et al. *Training language models to follow instructions with human feedback*. arXiv:2203.02155 [cs]. Mar. 2022. URL: http://arxiv.org/abs/2203.02155 (visited on 07/18/2023).

[65] Rohan Taori et al. *Stanford Alpaca: An Instruction-following LLaMA model*. https://github.com/tatsu-lab/stanford_alpaca. 2023.

[66] Hugo Touvron et al. *LLaMA: Open and Efficient Foundation Language Models*. arXiv:2302.13971 [cs]. Feb. 2023. URL: http://arxiv.org/abs/2302.13971 (visited on 06/30/2023).

[67] SoHyun Park et al. "Designing a Chatbot for a Brief Motivational Interview on Stress Management: Qualitative Case Study". EN. In: *Journal of Medical Internet Research* 21.4 (Apr. 2019), e12231. DOI: 10.2196/12231. URL: https://www.jmir.org/2019/4/e12231 (visited on 01/29/2023).

[68] Fahad Almusharraf, Jonathan Rose, and Peter Selby. "Engaging Unmotivated Smokers to Move Toward Quitting: Design of Motivational Interviewing-Based Chatbot Through Iterative Interactions". eng. In: *Journal of Medical Internet Research* 22.11 (Nov. 2020), e20251. ISSN: 1438-8871. DOI: 10.2196/20251.

[69] Fahad Almusharraf. "Motivating Smokers to Quit Through a Computer-Based Conversational System". en. In: ().

[70] Siqi Shen et al. "Counseling-Style Reflection Generation Using Generative Pretrained Transformers with Augmented Context". In: (July 2020), pp. 10–20. URL: https://aclanthology.org/2020.sigdial-1.2 (visited on 01/29/2023).

[71] Imtihan Ahmed. "Automatic Generation and Detection of Motivational Interviewing-style Reflections for Smoking Cessation Therapeutic Conversations using Transformer-based Language Models". en. Accepted: 2022-06-29T15:11:56Z. Thesis. June 2022. URL: https://tspace.library.utoronto.ca/handle/1807/123170 (visited on 01/26/2023).

[72] Stefan Palan and Christian Schitter. "Prolific.ac—A subject pool for online experiments". en. In: *Journal of Behavioral and Experimental Finance* 17 (Mar. 2018), pp. 22–27. ISSN: 2214-6350. DOI: 10.1016/j.jbef.2017.12.004. URL: https://www.sciencedirect.com/science/article/pii/S2214635017300989 (visited on 01/26/2023).

[73] Y. K. Bartlett et al. ""It's my business, it's my body, it's my money": experiences of smokers who are not planning to quit in the next 30 days and their views about treatment options". eng. In: *BMC public health* 15 (Aug. 2016), p. 716. ISSN: 1471-2458. DOI: 10.1186/s12889-016-3395-0.

[74] R. Borland et al. "The reliability and predictive validity of the Heaviness of Smoking Index and its two components: Findings from the International Tobacco Control Four Country study". In: *Nicotine & Tobacco Research* 12.Suppl 1 (Oct. 2010), S45–S50. ISSN: 1462-2203. DOI: 10.1093/ntr/ntq038. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3307335/ (visited on 07/04/2023).

[75] Stewart W. Mercer et al. "General practitioner empathy, patient enablement, and patient-reported outcomes in primary care in an area of high socio-economic deprivation in Scotland–a pilot prospective study using structural equation modeling". eng. In: *Patient Education and Counseling* 73.2 (Nov. 2008), pp. 240–245. ISSN: 0738-3991. DOI: 10.1016/j.pec.2008.07.022.

[76] Imtihan Ahmed et al. *Generation and Classification of Motivational-Interviewing-Style Reflections for Smoking Behaviour Change Using Few-Shot Learning with Transformers*. en. June 2022. DOI: 10.36227/techrxiv.20029880.v1. URL: https://www.techrxiv.org/articles/preprint/Generation_and_Classification_of_Motivational-Interviewing-Style_Reflections_for_Smoking_Behaviour_Change_Using_Few-Shot_Learning_with_Transformers/20029880/1 (visited on 01/26/2023).

[77] Ari Holtzman et al. *The Curious Case of Neural Text Degeneration*. arXiv:1904.09751 [cs]. Feb. 2020. URL: http://arxiv.org/abs/1904.09751 (visited on 08/24/2023).

[78] *Firebase*. en. URL: https://firebase.google.com/ (visited on 01/29/2023).

[79] *AWS Documentation*. URL: https://docs.aws.amazon.com/ (visited on 01/29/2023).

[80] *Wit.ai*. URL: https://wit.ai/ (visited on 08/24/2023).

[81] *SciPy*. URL: https://scipy.org/ (visited on 01/29/2023).

[82] Annemieke P. Bikker et al. "Measuring empathic, person-centred communication in primary care nurses: validity and reliability of the Consultation and Relational Empathy (CARE) Measure". In: *BMC Family Practice* 16.1 (Oct. 2015), p. 149. ISSN: 1471-2296. DOI: 10.1186/s12875-015-0374-y. URL: https://doi.org/10.1186/s12875-015-0374-y (visited on 02/07/2023).

[83] Matthew J. Carpenter et al. "Both smoking reduction with nicotine replacement therapy and motivational advice increase future cessation among smokers unmotivated to quit". eng. In: *Journal of Consulting and Clinical Psychology* 72.3 (June 2004), pp. 371–381. ISSN: 0022-006X. DOI: 10.1037/0022-006X.72.3.371.

[84] Andrew Brown et al. *A Motivational-Interviewing Chatbot with Generative Reflections for Increasing Readiness to Quit Among Smokers: Iterative Development Study (Preprint)*. en. preprint. JMIR Mental Health, May 2023. DOI: 10.2196/preprints.49132. URL: http://preprints.jmir.org/preprint/49132 (visited on 07/12/2023).

[85] Chin-Yew Lin. "ROUGE: A Package for Automatic Evaluation of Summaries". In: (July 2004), pp. 74–81. URL: https://aclanthology.org/W04-1013 (visited on 07/13/2023).

[86] Tianyi Zhang et al. *BERTScore: Evaluating Text Generation with BERT*. arXiv:1904.09675 [cs]. Feb. 2020. URL: http://arxiv.org/abs/1904.09675 (visited on 07/10/2023).

[87] Zixiu Wu et al. “Are Experts Needed? On Human Evaluation of Counselling Reflection Generation”. In: (July 2023), pp. 6906–6930. URL: https://aclanthology.org/2023.acl-long.382 (visited on 07/13/2023).

[88] Zixiu Wu et al. “Anno-MI: A Dataset of Expert-Annotated Counselling Dialogues”. In: (May 2022). ISSN: 2379-190X, pp. 6177–6181. DOI: 10.1109/ICASSP43922.2022.9746035.

[89] Mary L. McHugh. “Interrater reliability: the kappa statistic”. In: *Biochemia Medica* 22.3 (Oct. 2012), pp. 276–282. ISSN: 1330-0962. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3900052/ (visited on 08/06/2023).

[90] Adam Paszke et al. “PyTorch: An Imperative Style, High-Performance Deep Learning Library”. In: 32 (2019). URL: https://papers.nips.cc/paper_files/paper/2019/hash/bdbca288fee7f92f2bfa9f7012727740-Abstract.html (visited on 07/17/2023).

[91] Thomas Wolf et al. *HuggingFace’s Transformers: State-of-the-art Natural Language Processing*. arXiv:1910.03771 [cs]. July 2020. URL: http://arxiv.org/abs/1910.03771 (visited on 07/17/2023).

[92] Samyam Rajbhandari et al. *ZeRO: Memory Optimizations Toward Training Trillion Parameter Models*. arXiv:1910.02054 [cs, stat]. May 2020. URL: http://arxiv.org/abs/1910.02054 (visited on 07/28/2023).

[93] Lutz Prechelt. “Early Stopping - But When?” In: (Mar. 2000). ISSN: 978-3-540-65311-0. DOI: 10.1007/3-540-49430-8_3.

[94] Diederik P. Kingma and Jimmy Ba. *Adam: A Method for Stochastic Optimization*. arXiv:1412.6980 [cs]. Jan. 2017. URL: http://arxiv.org/abs/1412.6980 (visited on 07/28/2023).

[95] Timothy P. Johnson et al. “An investigation of the effects of social desirability on the validity of self-reports of cancer screening behaviors”. eng. In: *Medical Care* 43.6 (June 2005), pp. 565–573. ISSN: 0025-7079. DOI: 10.1097/01.mlr.0000163648.26493.70.

[96] Ozan Kuru and Josh Pasek. “Improving social media measurement in surveys: Avoiding acquiescence bias in Facebook research”. en. In: *Computers in Human Behavior* 57 (Apr. 2016), pp. 82–92. ISSN: 0747-5632. DOI: 10.1016/j.chb.2015.12.008. URL: https://www.sciencedirect.com/science/article/pii/S0747563215302788 (visited on 02/10/2023).

[97] R. C. Willets. “The Cohort Effect: Insights and Explanations”. en. In: *British Actuarial Journal* 10.4 (Oct. 2004). Publisher: Cambridge University Press, pp. 833–877. ISSN: 2044-0456, 1357-3217. DOI: 10.1017/S1357321700002762. URL: https://www.cambridge.org/core/journals/british-actuarial-journal/article/cohort-effect-insights-and-explanations/798CF62CB590829E247D4A72484518AE (visited on 02/10/2023).