Generation and Classification of Motivational-Interviewing-Style Reflections for Smoking Behaviour Change Using Few-Shot Learning with Transformers

Jonathan Rose [1], Imtihan Ahmed [2], Eric Keilty [2], Carolynne Cooper [2], and Peter Selby [2]

[1]University of Toronto
[2]Affiliation not available

October 30, 2023

Abstract

If conversational agents can take on a therapeutic role, they may provide a scalable way to help many people suffering from addictions. Motivational Interviewing (MI) is a validated therapy for behaviour change that can be applied to addiction, including smoking cessation. A core technique in MI (and many other kinds of talk therapy) is to pose an open-ended question concerning a negative behaviour, and then to provide a reflection of the response. Reflections can be a simple restatement of the response, or a more complex inference from prior statements or general knowledge, and they help someone contemplate the behaviour more deeply. We describe a method to generate reflections that uses few-shot priming of the GPT-2 and GPT-3 language models. These produce very promising simple and complex reflections, but also some that are off-topic or irrelevant. To filter these, we train a classifier to detect poor reflections, employing samples labeled by an MI expert. Its accuracy is 81%, sensitivity 90% and specificity 71%. We show that GPT-2 can generate acceptable reflections at a 54% success rate, and when combined with the classifier/filter produces acceptable reflections 73% of the time. The GPT-3 model has a native success rate of 89%.

Generation and Classification of Motivational-Interviewing-Style Reflections for Smoking Behaviour Change Using Few-Shot Learning with Transformers

Imtihan Ahmed, Eric Keilty, Carolynne Cooper, Peter Selby and Jonathan Rose, *Fellow, IEEE*

***Abstract*— If conversational agents can take on a therapeutic role, they may provide a scalable way to help many people suffering from addictions. Motivational Interviewing (MI) is a validated therapy for behaviour change that can be applied to addiction, including smoking cessation. A core technique in MI (and many other kinds of talk therapy) is to pose an open-ended question concerning a negative behaviour, and then to provide a *reflection* of the response. Reflections can be a simple restatement of the response, or a more complex inference from prior statements or general knowledge, and they help a person contemplate the behaviour more deeply. We describe a method to generate reflections that uses few-shot priming of the GPT-2 and GPT-3 language models. These produce very promising simple and complex reflections, but also some that are off-topic or irrelevant. To filter these, we train a classifier to detect poor reflections, employing samples labeled by an MI expert. Its accuracy is 81%, sensitivity 90% and specificity 71%. We show that GPT-2 can generate acceptable reflections at a 54% success rate, and when combined with the classifier/filter produces acceptable reflections 73% of the time. The GPT-3 model has a native success rate of 89%.**

***Index Terms*— addiction, mental health, motivational interviewing, dialogue systems, natural language processing, natural language generation, transformers.**

I. Introduction

Mental health disorders are the least served of all the medical fields as there are insufficient practitioners to serve the world-wide need [1]–[3]. Therapeutic conversations [4]–[6] form an important part of the treatments available for such disorders. If these conversations could be automated and shown to be efficacious, they could provide far greater access to care than is currently possible, through any internet portal.

An automated conversational agent that can faithfully mimic a psychotherapist is likely far beyond current capabilities, as this task seems equivalent to building a General AI [7], [8]. However, some widely used (and validated) therapies have a structure that may lend themselves to automation through software that makes use of natural language processing (NLP) approaches. This structure, together with recent advances in language models that have exhibited an ability to carry a conversation [9]–[12] may provide a pathway to creating an effective therapeutic conversation.

In this paper, we focus on a specific therapy called Motivational Interviewing (MI) [6] and address a specific condition: addiction to tobacco smoking. Tobacco is the number one preventable cause of death in the US and Canada and results in more than 8 million deaths per year worldwide. There are effective treatments but most current smokers are *ambivalent* about quitting smoking and therefore make no attempt to quit. Motivational Interviewing is effective in helping ambivalent smokers make quit attempts.

We narrow our focus to the generation of a specific kind of conversational response within MI, called a *reflection*, that is broadly used as a part of MI therapy but also has application in many other kinds of talk therapy [13], [14]. An MI-style conversational reflection has two main types: a *simple* reflection, which is a summary and re-expression of the content expressed and helps encourage the patient to keep talking. A *complex* reflection makes an inference of the patient's values, feelings, and intentions. This can be done by connecting what they have said to something relevant and useful conveyed previously, or to a helpful insight based on general knowledge. The purpose of reflections in a smoking addiction MI conversation is to help the patient *continue* to contemplate their smoking behaviour, and to engage in self-reflection. This can help them resolve their ambivalence towards making an attempt to quit [6], [15]. The automatic generation of appropriate reflections in automated conversational agents is one key to this process. In addition, this capability could be useful in the many other therapies that also require the patient to engage in self-reflection [13], [16].

There have been several previous attempts to create automated MI conversations [17]–[24]. However, in most cases the reflections that are presented to the subjects are explicitly scripted by humans, and the choice of reflection is based on either an automated classification of the subject's input [17], [24] or through topic modelling based on keywords [21]. One contribution of this paper is to describe a method to create reflections that are generated *directly* from what a subject says, rather than being scripted. These more specific-to-what-is-said responses have been shown to improve the therapeutic alliance between patient and conversational system [NEED REF]. The challenge of this approach, however, is that the automatically generated responses are not explicitly controlled, and so may be counter-therapeutic, by being factually wrong, racist, sexist or judgemental. Our work most closely relates to Shen et. al's [25] work on automatically generating reflections, which we describe in more detail in Section 2.

MI conversations begin with open-ended questions from a counselor which produce free-form responses from patients. The goal of this work is to generate an appropriate reflection, automatically, given an input that is the question together with the response. We are partly motivated by recent progress in language models has exhibited remarkable conversational fluency in the generation of text [8]–[12], [26]. The few-shot approach to language tasks that was suggested in [9] may allow such models to produce good reflections when conditioned with only a few examples of good reflections.

In this paper, we show that both OpenAI's GPT-2 [12] and GPT-3 [9] language models produce good simple and complex reflections that could be used in an MI-style smoking cessation conversation. However, some of the reflections are inappropriate, motivating the development of a classifier that can select relevant and good quality reflections. The classifier is trained and validated with data that is labelled by an expert MI counselor. We measure the difference in

This work was supported in part by NSERC Discovery Grant RGPIN-2019-04395

I.Ahmed was with the Edward S. Rogers Sr. Department of Electrical and Computer Engineering, University of Toronto, and is now with Wish Corporation

E. Keilty and J. Rose are with the Edward S. Rogers Sr. Department of Electrical and Computer Engineering, University of Toronto, Toronto, Ontario, Canada.

C. Cooper, P. Selby and J. Rose are with Centre for Addiction and Mental Health, Toronto, Ontario, Canada.

TABLE I: Examples of Simple and Complex Reflections

Patient Statement	Simple Reflection	Complex Reflection
I don't think I have a problem with smoking, I don't smoke that much, and I can control my urges when I want.	You don't believe smoking is a problem for you.	If you smoked more and the urges forced you to smoke you would have a problem with smoking.
I really want to quit, but I get headaches if I don't have a few each day.	You want to stop smoking.	You feel caught between wanting to quit and the side effects of quitting.
I can't just suddenly stop smoking; it would make my friends feel awkward when we hang out and they all smoke.	You're worried about ruining the mood when you hang out with friends.	Your friends mean a lot to you. You don't want to disrupt the vibe when you hang out with them by quitting smoking suddenly

quality of reflections generated by GPT-2 and GPT-3, using both hand labels and the classifier. We show that GPT-3 is more likely to produce a good reflection, but that with the use of the classifier as a filter, the GPT-2 model can be significantly improved, in a much easier-to-use, smaller model.

This paper is organized as follows: Section II reviews relevant background on the Motivational Interviewing approach, the language models we use to generate and classify reflections, and related work on automatic MI conversations. Section III describes the dataset that is used to condition/train the generators and the classifier. Section IV describes the reflection generation methodology, and Section V describes the training and performance of the classifier. Section VI presents the native success rate of reflection generation by the GPT-2 and GPT-3 models, as well as the success rate of GPT-2 combined with the classifier as a filter. It also presents a qualitative discussion of successful and unsuccessful reflections. Section VII discusses the limitations of the presented work, and Section VIII concludes.

II. Background

In this section we review the Motivational Interviewing (MI) therapy approach, and the language models used in the reflection generator system. We also review prior work relating to the automation of MI.

A. Reflections and MI

MI [6] is a widely used counseling approach used to help people move towards positive behavior change. It has been used to target many health behaviours, including smoking [15]. A key goal in the approach is to help move a person away from their *ambivalence* about changing a behavior, a state which leaves them unmotivated to change. MI counselors use a structured conversation that encourages a patient to contemplate the roots of the behaviour and guides them to overcome this ambivalence.

A key skill used by MI counselors is called **reflective listening** [6], [13], [16]. This includes careful listening to a patient and responding with words that both reflect what is said and guides the patient, in a safe and non-judgemental way, towards continued exploration of their thoughts and feelings about change. These responses, called *reflections* can be *simple* or *complex*. A simple reflection repeats or rephrases the patient's words to convey understanding of the content and signals that the person was heard. A complex reflection makes a relevant and helpful inference based either on previous statements by the patient or on the meaning within a patient's words. A good reflection promotes continued exploration of the ambivalence and ways to change the behaviour. The thesis of MI is that doing so will help the patient come to their own conclusion to make the change in behaviour. In addition, it helps build trust between the therapist and patient. Table I shows three examples of simple and complex reflections that might be given in response to a patient statement.

An MI counselor requires significant training in MI that would typically be in addition to training in a social science or medical discipline. The small number of counselors can help only a limited number of people, and tend to be concentrated in urban areas [27]. If an automated and effective counselor could be created, it would allow far greater scale of access to help at much lower cost. The automatic generation of good quality reflections is a first step towards this goal.

B. Language Models

An autoregressive Language Model (LM) generates text by taking a sequence of tokens (that represent words or portions of words) as an input and producing subsequent tokens/words [9], [12]. Recent models have had significant success in text generation and classification tasks [28]–[30]. The training objective of an autoregressive LM optimizes the likelihood of the next token given a series of input tokens. Autoregressive LMs can also be fine-tuned on downstream tasks, which is helpful in domains that have limited data, such as the therapeutic conversation arena of this work.

One family of autoregressive languages models that have been very successful in various language modelling tasks is OpenAI's GPT-2 (Generative Pre-trained Transformer 2) [12], and its more recent larger variant, GPT-3 [9]. GPT-2 employs the multi-headed self-attention decoder Transformer block described by Vaswani et al [31]. A key feature of self-attention is the efficient encoding of long-range dependencies in text, compared to earlier text-generation models based on recurrent neural networks [32], [33]. The attention layers do this because they have visibility over a large context window of the input.

The GPT-2 model [12] has 1.5 billion parameters and was pre-trained using the WebText [34] language corpus internal to OpenAI, comprised of 40GB of internet text. The dataset was curated with an emphasis on document quality, only including websites that are outbound links from Reddit with significant user engagement. It comprises text from over eight million documents but does not include text from websites that are present in other popular textual datasets such as Wikipedia.

The GPT-3 model [9] has 175 billion parameters and is trained on 499GB of text data from multiple sources: Common Crawl from 2016-2019 [35], WebText, Books1 and Books2 (collection of books and movie script text data) [36], and English-language Wikipedia [37].

In this paper, we make use of the pre-trained GPT-2 and GPT-3 models to generate reflections using few-shot learning [9]. In [9], [10] the term 'few-shot learning' is defined differently than the original notion [38], [39] of having few training examples that fine-tune the model parameters. Here, rather, the input to the model (as opposed to the training samples) are several complete examples of the text generation task itself. Each example consists of the context and desired completion (which is what the generated text should look like). After the examples, the model is given a final context, and from which it generates a completion. The presentation of the examples to the model input is referred to as 'priming' or 'conditioning' or 'prompting' and techniques for doing this effectively are explored in [40].

GPT-2 and GPT-3 are fixed-length models, meaning the amount of conditioning presented to these models is limited by the maximum size of the input available to the model. The maximum size of the inputs to GPT-2 is 1024 tokens, and for GPT-3 it is 2048 tokens. In general, larger/more conditioning generates better completions [9]. Adjustments to the GPT-2 architecture have been proposed to address this limitation by turning it into a model with unbounded context by implementing a recurrence mechanism [41]. There are other models such as the Transformer-XL [26] and XLNet [42] that can handle much larger context windows, however these models require fine-tuning to perform well on domain-specific tasks.

Other factors that influence the quality of the generated text in any autoregressive language model is the choice of decoding method, which is the algorithm used to select the next word to be generated given the probabilities across the vocabulary [9]. There are significant differences in the quality of the text generated between different decoding methods. A greedy approach may pick the most probable word given all the previous words but typically does not produce a high probability global solution. More complex search-based methods such as beamforming, are better, but require significantly more computation due to many more invocations of the model inference. It has also been shown that the volatility inherent in human dialogue is best mimicked with a sampling technique such as top-k, or top-p (nucleus) sampling [43]. GPT-2 has a choice between greedy decoding (with no parameters), search-based decoding with beam search (which uses the number of beams as the parameter), and sampling decoding methods (which uses a combination of top_k and top_p sampling). A repetition penalty [44] is also used to reduce the likelihood of a previously input or generated token to be repeated.

C. Text Sequence Classification Models

Of additional relevance to this work is the use of a second model to classify generated text for the purpose of filtering inappropriate or poor quality generation [45]. Here the DialoGPT system uses a maximum mutual scoring function to rank generated text to filter bland and generic outputs. In Section V we describe a simpler approach using a fine-tuned text classifier to filter out low quality reflections.

There are two categories of text sequence classifiers: rule-based and machine learning-based. Rule-based models make use of hand-crafted features that are then used in a classification model to classify text sequences. One popular method of deriving hand-crafted features is using the bag of words (BoW), combined with a simple classifier based on one of Naïve Bayes, Support Vector Machine, or hidden Markov Model approach [46]. Although rule-based methods had early success with text sequence classification, they required extensive domain knowledge and feature engineering. The features were also static; thus, they could not be improved upon by using the growing number of text corpora that are available [46].

The machine learning-based methods address many of the weaknesses of rule-based methods, with the caveat that large text corpora training data are a required for good performance. These model types are used to derive continuous low-dimensional representation of features using vectors, commonly referred to as embeddings, that can be used in downstream tasks for text sequence classification. Early machine learning-based methods, such as latent semantic analysis (LSA) [47], suffered from low parameter sizes and small training corpora, thus they did not perform as well as rule-based methods [46]. More recent methods used significantly larger models and corpora with billions of words and now perform significantly better than rule-based methods [46]. Embedding models such as Word2Vec [48] and ELMo [49] have been very successfully used for text-sequence classification tasks using Long-Short Term Memory (LSTM) models [50], [51]. Although these models performed well generally, they needed to be fully trained from scratch for each task, reducing these models' utility in low-data domains.

A newer approach based on the transformer architecture [31] has been shown to be significantly more powerful for text sequence classification tasks than its predecessors. The Bidirectional Encoder Representation from Transformers (BERT) [11] model uses Masked Language Modeling (MLM) as a pretraining step to learn its representation of words. MLM works by masking a subset of the input, and then having the model predict those masked words. In addition, BERT combines this task with the task of predicting if a second sentence logically or naturally follows from the first. This combined training objective allows BERT to encode relationships between tokens and their positions in a text sequence and to learn entailment [46] relationships from text sequences.

Once the pre-training step is complete, the BERT model can be augmented by appending various neural network layers to its outputs. The augmented model can then be further fine-tuned on a small dataset for specific natural language understanding (NLU) tasks, including classifying sequences of texts based on specific characteristics. For example, BERT can be fine-tuned to detect if question-answer pairs make sense. It can also do general textual entailment between two sequences of texts [46]. In Section V we use a fine-tuned BERT model to filter the generated reflections.

D. Related Work on Chatbots and Motivational Interviewing

There have been a few prior chatbots that used Motivational Interviewing approaches for behaviour change. That prior work was in several domains, including stress management [21], sexual health education [52], smoking cessation [17] and substance abuse [53]. There are also a two other studies of the use of MI chatbots, that are clinically-oriented and so do not provide details on the design and engineering of the system itself [23], [54].

Park et al. designed a qualitative case study for a chatbot that conducts a brief MI session with university students [21]. The researchers constructed pre-written responses in four MI categories of counselor statements: giving information, asking questions, providing reflections, and MI-adherent statements. The responses were designed to be generic enough that they could be used in many conversational contexts. The chatbot responded to statements by selecting from the set of responses based on specific keywords present in the statements. The researchers found that the students responded well to the chatbot's open-ended questions but did not relate strongly to the reflective and affirming statements, since these statements did not directly address what was said. The participants felt the chatbot could be improved if it were able to give more contextual replies and informational support. The present work addresses this issue directly by generating a unique reflection from the preceding question and response.

A second prior effort where MI chatbots have had some success is providing sexual health information through Facebook [52]. This system is designed to address the 'giving information' component of the MI conversation, specifically providing information about

HIV/AIDS. It uses a question-answer corpus to train a response classifier to rank potential responses to a user's questions and presents the most likely response while accounting for repetitiveness. Here the responses are also pre-written, but because there is a larger set to select from, it is more likely to be responsive. The context of providing information is also more straightforward than that of making reflections, the subject of the present work.

A complete automated MI chatbot targeting a specific behaviour was created by Almusharraf el al. [17], [55]. The researchers designed a fully automatic MI-style chatbot capable of conducting a conversation relating to an individual's reasons for smoking. Their system followed a special form of MI known as the running head start technique [6]. The chatbot identified a person's reasons for and against smoking by asking open-ended questions and classifying their responses into one of twenty-one categories. The system was evaluated with 121 participants that produced over 6568 responses to the chatbot. Though a significant portion of participants (35%) found their interactions with the chatbot helpful in getting them to think about their smoking, there were some weaknesses also identified. The conversation was specific to the running head start method and does not generalize beyond that; it also cannot respond to participant statements outside of the scope given by the smoking reasons. In addition, the structure of the responses was often felt by the participants to be repetitive, as they were also scripted and had a repetitive structure.

Another complete MI chatbot targeting smoking was created by He et al [56], which engaged 78 smokers and compared the impact with a more neutral (non-MI) chatbot on 75 smokers. Their principle result was that the addition of MI-style conversation did *not* improve the effectiveness of the bot, compared to the neutral bot. It is possible that this lack of improvement stemmed from the used of scripted reflections and summaries, rather than the specific reflections we seek in this work.

Olafsson et al [53] describe a system that interacts through an audio interface that engages in an MI-style dialogue through the use of dialogue trees to engage and help subjects with activities such as breathing exercises. Tests of the system indicated good user engagement, but also the need to have more relevant response to free-form input speech, such as the generated reflections that are the goal of the present paper. The second part of the work describes a system for predicting the nature of the next "move" by the agent, which is the kind of MI interaction (such as a reflection, a summarization, or an affirmation) that would best used at that point in the conversation. The authors suggest that, by successfully predicting the next style of MI statement, a generative approach could be conditioned to produce the next utterance.

Most of these full-chatbot studies cited a need for the chatbot to provide more free-form, contextual responses to the user's statements or queries, described below in Section VI.

Shen et al [25] explored, as we do in this work, how generative models can be used to create reflections that could be used in actual MI conversations. Their work focused on reproducing reflections from pre-existing clinical conversations. The goal was to see how closely a GPT-2 based language model can reproduce the actual reflections given by the therapists and counselors from those conversation sessions. The authors demonstrate that it is possible to produce compelling free-form reflections in a therapeutic context using a transformer-based generative language model. They showed that it performs better than standard seq2seq-type models.

Their approach was to first domain-adapt a GPT-2 based model on transcripts of videos of therapy and then to fine-tune the model using an MI counseling dataset [57]. The authors then evaluate the model's ability to generate reflections given a subset of dialogue history. They also explored how adding more context, by adding reflections from similar conversations as part of the input, changes the quality of the generated reflections. The generated reflections were evaluated on how similar they were to the actual reflections from the context conversations, as well as the quality of the grammar, reflection-likeness, and relevance. The GPT-2 based models performed better than the baseline seq2seq models on the generation task. The addition of similar reflections to the context was also shown to slightly improved the similarity of the generated reflections to the ground truth, though not in all cases. Their qualitative analysis also show that the generated reflections were relevant, high quality (in terms of grammar and structure) and had high reflection-likeness.

III. Methods: Dataset and Relevant Therapeutic Context

The goal of this work is to automatically generate a *reflection*, in the context of behaviour change therapy for smoking cessation. The input to this process is an open-ended question/prompt, together with the response from a patient. The output is a reflection as described in Section II-A. We have also noted the need for a classifier that can detect good quality reflections, ultimately for use in filtering the generations, but also useful for measuring the quality of the unfiltered generation process.

For both tasks, a dataset is required that contains prompts, responses, and reflections. Ideally these would come from real counselors and their patients engaged in MI counseling sessions, but such counseling transcripts are protected by medical privacy laws and regulations. Fortunately, our previous study, in which we prototyped a different MI-oriented chatbot, provides us with 204 conversations to mine for appropriate prompt-response pairs [17], [55]. The data collected for that study was approved under University of Toronto Health Science Research Ethics Board (REB) protocol number 35962 on May 28, 2018 and REB protocol number 36639 approved on September 10, 2018.

Section III-A presents a review of that study and gives a description of the textual data. Section III-B details how the data is used for the reflection generation process and how the data is augmented to provide training data for the reflection classification task.

A. Dataset

We collected transcripts from the deployment of a previous MI-oriented chatbot that explored an individual's reasons for smoking [17], [55]. The transcripts contain open-ended questions from the chatbot asking for the participant's reasons for and against smoking. The bot then classified each reason into one of 21 categories, and then selected a reflection based on that classification from a human-authored set. Subsequently, the chatbot asks the participant to recall situations where the reason was in play – for example, 'recall a time when you were stressed.' It asked the participants to describe a time that this reason caused them to smoke. In addition, they were asked to describe a time that they *did not* smoke with that reason in play. Finally, the participant was asked to say what enabled them to resist smoking in the case that they smoked, compared to the time that they did smoke. This last question was a key part of the conversation, as the intention was to evoke contemplation within the participant, on what gave them the ability to resist the addiction [17], [55].

From each transcript in this dataset created by Almusharraf et al. [17], [55], we extracted pairs of utterances. The first utterance was the question asked by the chatbot itself, with the second utterance being the response from the participant. We then manually reviewed this collection of utterance pairs in two steps: The first step was to discard any utterance pairs where the first utterance was not an open-ended question (where the possible answers were limited to specific responses – for example 'yes' or 'no'). Secondly, we reviewed the remaining utterance pairs based on whether the participant responses could (and should) be reflected upon, discarding any utterance pairs that do not warrant a reflection.

The resulting dataset consists of 1665 utterance pairs where the first utterance is the question prompt, the second being the participant response. However, this dataset does not have the (required) reflections that are needed for few-shot generation (as described in Sections II-B and IV, which requires a *completion*) and classifier training (described in Section V). The prior study did not have the capability of generating a free-form reflection. In the next Section we describe how this dataset was modified and enhanced for both the few-shot reflection generation task and for the classifier training task.

B. Enhancing Dataset for Few-Shot Reflection Generation & Classification

The dataset described above contains utterance pairs of prompts and responses. To create high quality conditioning examples that can be used in the few-shot reflection generation task, we first manually constructed a set of high-quality reflections for each prompt-response pair in the dataset, to form the requisite completions.

The creation of good-quality reflections requires expertise in the field of motivational interviewing, which we gained in two ways: first, the authors undertook a training course in MI, and also read and reviewed a fundamental text in the field [6]. We then created 20 reflections, manually, to go along with 20 of the prompt-responses. Using the few-shot priming approach we generated 369 reflections (based on 123 different prompt-response pairs, reflected three times each) using the few-shot approach described in Section IV.

Each of these 369 reflections were then labelled by an MI expert practitioner at the Centre for Addiction and Mental Health in Toronto. The MI expert also provided adjustments to poor quality generated reflections to improve the dataset.

The reflections were given a binary label with '0' meaning that the reflection is not appropriate to the prompt and response, and '1' meaning that the reflection is consistent with MI. The reflections and the corresponding prompts and responses that received a label of 1 were exclusively used as the set from which examples were drawn for the few-shot reflection generation task. This dataset will be referred to as the *priming* set.

Using the priming set, we were able to generate and label many more prompt-response-reflection triplets that were of good quality over multiple experiments. Due to the generative nature of the task and the fact that the overall dataset was collected over many reflection generation experiments with varying decoding parameter choices (described in Section IV), we were able to produce multiple distinct reflections for each prompt-response pair. These reflections were then labelled by non-expert (student) labellers and left us with a total of 3301 prompt-response-reflection triplets, 39% of which were positive and 61% were negative. This set of prompt-response-reflection-label collection was then used as the training data for the reflection classification task, described in Section V.

IV. Few-Shot Generation of Reflections Using GPT-2 and GPT-3

As described in Section II-A, a *reflection* is a response to a prior open-ended question (the prompt) and the response from a patient. The reflection generation task is to produce a reflection given the prompt and response.

We employ the few-shot learning method (described in Section II-B) using the GPT-2 and GPT-3 models for the reflection generation task. The GPT-2 pretrained models are easily accessible from the Huggingface library [58]. We were able to access the GPT-3 pretrained models from OpenAI due to a generous program from OpenAI that enabled several projects at the University of Toronto to access and run the large models, but within a fixed budget. As such, we decided to explore the various parameters and conditions needed to create good reflections mostly with the GPT-2 model given the constrained access to GPT-3.

The *context* of the input is the prompt and the response whereas the desired *completion* is the reflection. The few-shot learning method [9] requires several examples of the context and completion to be prepended to the specific input context for which a completion is desired. Since the GPT-2 and GPT-3 models have finite context windows, the maximum number of example context-completions is limited. For GPT-2 the approximate example limit is eight, and for GPT-3 is it roughly 12. These numbers are determined by dividing the context window size by the average size of context-completions in the dataset. However, the exact amount of examples may vary because the size of the examples themselves vary. So, in the case the input token was greater the context size limit (which did happen occasionally) the input was truncated from the start until the input was equal to the maximum number of tokens.

The prepended examples are drawn from the priming set described in Section III-B. According to Brown et al. [9], the examples and the final context must be constructed in a specific way, with delineators between the elements of the context and completion. We chose to prepend the prompt with the identifier 'Prompt: ', and on the next line we prepend the response with 'Response: ', followed by another line where the reflection is prepended with 'Reflection: '. To distinguish between each example, they are separated with a blank line. The final prompt has one line for the prompt, and a second line for the response. The response line is appended with a newline character. In the case of GPT-2, the entire input is then passed through the pre-trained GPT-2 tokenizer from the HuggingFace library, based on byte-level byte-pair-encoding [58]. The generator then follows the examples to produce a completion, which is usually (but not always) of the form 'Reflection: some text'. If the generator fails to produce output in this form, we reject the output and record it as a blank string. Fig. 1 presents an example of the full few-shot generation input with three example context-completion priming pairs and a final context without the completion.

A. Choosing the Number of Primers

A key parameter of the few-shot generation method is the number of context-completion examples to use to condition/prime the models. To determine this number, 100 randomly selected prompt-response pairs were used to generate reflections once each with different numbers of primers, from 2 to 8. The model used was GPT2-XL from the HuggingFace model library [58]. The primers were resampled from the primer set for each generated completion. The specific decoding parameters used were top-k=100, top-p=1.0 and temperature=0.4. The high value of top-k and top-p of 1.0 ensure that a sufficient number of high probability words are considered for the completions, while a temperature of 0.4 ensures some creativity in the generated responses without letting the text become too random [43]. Other combinations of decoding values are explored in Section IV-B.

In some cases, the input size was larger than the context window of GPT2-XL, which is 1024 tokens. To be able to use these larger inputs to produce reflections, the input was truncated during tokenization, removing the input text from the beginning until the input size was

Prompt: Please describe a time where you were worried about the smell of cigarettes and you didn't smoke
Response: When I was going to my doctor. I refrained from smoking on the way so she wouldn't smell it and be reminded to lecture me about smoking.
Reflection: You are afraid of being reminded by your doctor to talk about smoking.

Prompt: Think back to the time when you were able to prevent yourself from smoking. What made it different from when you did smoke?
Response: Because of the effect it was having on my daughter
Reflection: You're worried how smoking affects your daughter.

Prompt: Please describe a time where you had a compelling feeling to smoke but didn't end up smoking
Response: I was compelled to have a cigarrette before an interview but refrained
Reflection: You felt like you could do without cigarettes.

Prompt: Please describe a time where you smoked because you needed to feel good
Response: When I am feeling stressed out at work.

Fig. 1: Example Few-Shot Input Sequence

reduced to 1024 tokens. In the case that the input was smaller than the context window size, the end of the input was padded with 'end-of-sentence' tokens.

Every reflection generated by this process was hand-labeled according to the binary labeling criteria described in Section III-B. We will refer to the fraction of reflections that were acceptable as the *hit rate*.

Fig. 2 is a bar chart of hit rate plotted against the number of priming examples (which range from 2 to 8.). These results suggest that, as the number of primers is increased, the likliehood that GPT2-XL can produce acceptable reflections reaches a plateau with seven and eight primers. We will use eight primers for the remainder of the experiments with GPT-2 in this work. We can observe that using less than three priming examples significantly reduces the likelihood of producing an acceptable reflection.

B. Generative Decoding Parameters

A second factor that has significant effect on the quality of the generated output from GPT-2 are the decoding strategy and parameters, as described in Section II-B. We focus on finding good decoding parameters for sampling-based decoding as they have been shown to be more effective than search or greedy strategies for dialogue generation [43]. We explored different combination of values for temperature, top_k, and top_p. To select the best decoding parameters for these experiments and datasets, we first produced 50 reflections using each combination of parameters, and then manually labeled the generated reflections to determine the hit-rate. The results are presented in Fig. 3, which gives several curves for different combinations of parameters, with temperature on the x-axis. The hit rate is somewhat noisy as a function of temperature (because there are only 50 examples per point), but the trend is quite clear: a low temperature achieves the highest hit rate. For the remainder of this paper, we chose the following decoder parameter values: temperature = 0.1, top_k = 100, and top_p = 0.8, which had the highest hit rate of 0.66.

GPT-3 has many more options for decoding parameters (than GPT-2) that impact the quality of the generated output. The only GPT-3 parameters/options that were the same as the decoding parameters of GPT-2 are temperature, top_p, and max_tokens. The other decoding parameters relevant to generating a single completion were kept at their default values (as recommended by OpenAI) which are: presence_penalty = 0, frequency_penalty = 0, best_of = 1, logit_bias = null. A full description of these parameters and others can be found in the OpenAI API documentation page [59]. When modifying temperature and top_p, the OpenAI API documentation suggests only modifying either one or the other but not both when trying to find good values for decoding parameters for GPT-3, so we chose to vary the temperature but keep top_p at its default 1.0 value. Ten reflections were then generated with eight primers for temperature values of 0.1, 0.4, 0.5, 0.75, and 1.0. The reflections were then manually reviewed for coherence and repetition. We observed that the temperature value of 0.4 provided the least repetitive and most coherent reflections and so this value was used to generate reflections for the remainder of the experiments using GPT-3 in this paper.

V. Reflection Quality Classifier (RQC)

In this section we describe the training of a classifier to determine if a candidate reflection is acceptable or not. We call this the *Reflection Quality Classifier* (RQC), and it will be used to filter poor-quality reflections, and as a fast, approximate way to label the quality of generated reflections. Section V-A describes the model. Section V-B describes the training process of the RQC and reports its performance.

A. Model Description

The RQC classifier which takes two strings as inputs: a context followed by a completion and produces a binary label indicating if the completion is an acceptable reflection to the context. The output is 0 if the reflection is not acceptable and 1 if it is acceptable. The inputs to the RQC are built by concatenating a prompt, a newline character, and a response as the context and using the reflection as the completion.

The RQC is trained by fine-tuning a pretrained uncased BERT Base model from the HuggingFace library [58]. We use BERT because it has been shown to be successful at text sequence classification tasks when fine-tuned with a small dataset. The input text is tokenized using the corresponding pretrained BertTokenizer from HuggingFace tokenizers [58], based on the WordPiece tokenization algorithm [60].

B. Training the RQC

The model was fine-tuned in two stages: the first step used the Microsoft Research Paraphrase Corpus (MSRP) [61], which captures

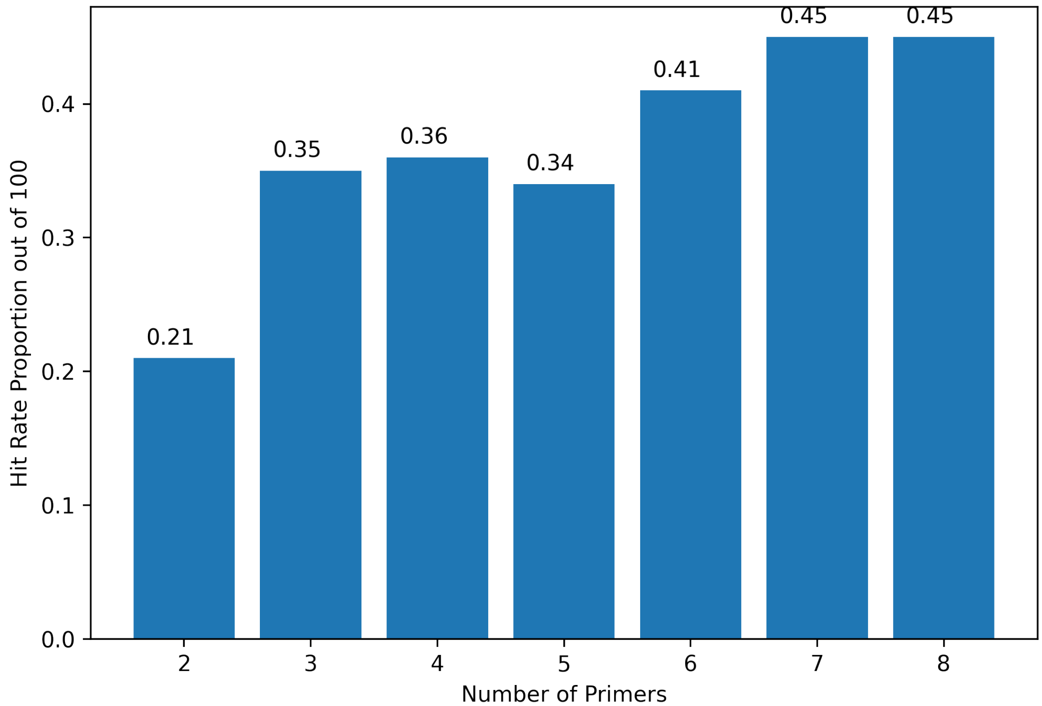

Fig. 2: Reflection Generation Hit Rate for Different Number of Primers for GPT-2

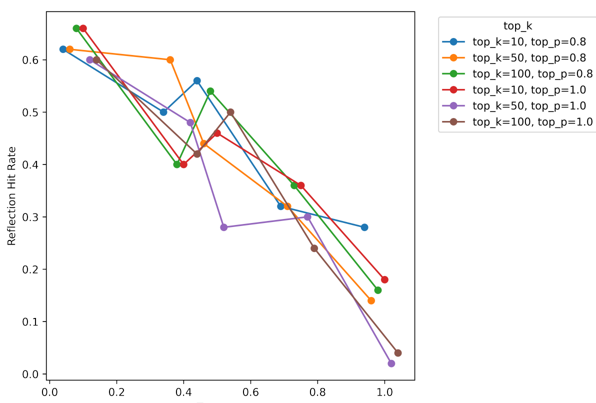

Fig. 3: Reflection Hit Rate at Different Decoding Parameter Combinations for GPT-2

a text entailment [62] relationship between a pair of texts. Each pair has a binary label which indicates whether the second text is a paraphrase of the first, or not. This is similar to the relationship between the MI prompt/response and reflection for simple reflections. Using the MSRP dataset increased the total amount of data that the model can be trained on, without increasing the burden of collecting and labelling the difficult-to-collect MI-style conversations.

The MSRP dataset comes pre-split into a 4077 training samples and 1726 test samples. We further subdivided the training samples into 80% training and a 20% validation set.

The model was further fine-tuned with the dataset described in Section 3.2, containing prompts, responses, reflections, and a binary label for the acceptability of the reflection. Since the dataset is unbalanced, a sample of 1075 positive and 1075 negative examples were randomly selected for the training task, giving a total of 2150 samples. We then shuffle and split this dataset into 80% training, 10% validation, and 10% test sets.

For both fine-tuning steps training was done using the tf.Keras functionality within the Tensorflow library [63]. An initial learning rate of 3x10-5 was used over 20 epochs, with early stopping enabled if the validation accuracy does not change over 3 epochs. The Adam optimizer was used due to its ability to adapt the learning rate over the training period, and it has also been shown to work well with large parameter models [64]. The loss function was the sparse categorical cross entropy loss which is recommended by the Tensorflow library when there are two label classes and they are provided as integers, and not one-hot encoded [63].

Table II reports the accuracy, sensitivity, and specificity of the model at each stage of the fine-tuning process, across the 220-sample test set described in Section 3.2. Each row of the table gives the accuracy impact with different levels of finetuning: first without any finetuning (the row labeled 'Bert Not Finetuned') and when the two different datasets are used to fine-tune BERT separately ('BERT-base MSRP Fine-tuned' and 'BERT Reflection Finetuned'), and when they are used together ('BERT MSRP+Reflection Finetuned').

Table II shows that the base pure BERT MODEL without finetuning is no better than an uninformative model (at 50% accuracy), but that both training datasets improve this when used to train the model separately and do even better when used together. This approach shows that even in the absence of a large dataset in a target domain, we can still achieve improved classification performance by drawing on datasets designed for similar tasks in different target domains.

TABLE II: RQC Model Accuracy, Sensitivity and Specificity Pre- and Post- Fine-Tuning

Model	**Accuracy**	**Sensitivity**	**Specificity**
BERT Not Finetuned	50%	0.07	0.95
BERT MSRP Finetuned	55%	0.13	0.97
BERT Reflection Finetuned	73%	0.82	0.59
BERT MSRP+ Reflection Finetuned	81%	0.90	0.71

Since one role of the RQC is to filter generative outputs, it is important to know how likely it is that the RQC will produce a false positive classification, and thus fail in its filtering function. For the 220-sample test set used above, Table III provides the confusion matrix for the BERT MSRP+Reflection fine-tuned model. From these we can calculate the specificity, or true negative rate, which is 71% for the final model, meaning that the chance that an unacceptable reflection will be classified as acceptable is 29%. The false negatives are less important in the filtering application as the cost of a false negative is the time and effort required to generate another reflection. However, we must be careful to select a model for the RQC that also has a good sensitivity, as practically speaking, we need an RQC filter that eventually lets a positive class through. Table II shows that the sensitivity improves with more of the fine-tuning steps included, while specificity decreases. The best balance of specificity and sensitivity is achieved with BERT-base MSRP+Reflection Fine-tuned, and this model is used as the RQC in the following sections.

When used as a reflection dataset labeling tool, the RQC's performance in labeling both the positive and negative classes are important. For this case the accuracy and F1 scores will give us an idea of the RQC's performance as a labeling aid. The overall accuracy of the RQC is 81%, and the F1 score is 83%.

TABLE III: BERT MSRP+Reflection Fine-tuned Confusion Matrix

Class Labels	**Predicted Positive**	**Predicted Negative**
Actual Positive	103 (True Positive)	12 (False Negative)
Actual Negative	30 (False Positive)	75 (True Negative)

C. Automatic Reflection Labeling with the RQC

A secondary use of the RQC is as an automatic reflection labeling aid to help parse and label large collections of MI transcripts to identify good quality reflections. This has application as both an aid for new MI practitioners evaluating MI conversations, or to help build large, labeled corpora of MI utterances with labeled reflections.

We evaluated the performance of the RQC as a labeling aid by using it to label 300 reflections produced natively by GPT-2 and GPT-3 models (these are all the examples described in the next section). We then compare the RQC labels to the hand-labeling, using the Cohen-Kappa Interrater correlation [65]. It measures inter-rater reliability between labelers who are labeling qualitative items. It considers the likelihood that two labelers agreed on the label of a sample randomly, which makes this correlation values more conservative than simple percentage agreement between the two labelers. McHugh [66] reports that a Cohen-Kappa score of 0 implies no agreement between the two labelers, 0.1-0.20 implies slight agreement, 0.21-0.40 implies fair agreement, 0.41-0.60 as moderate agreement, 0.61-0.80 as substantial agreement, and 0.81-1.0 as almost perfect agreement. The agreement between the hand-labeling and the RQC labeling is shown in Table IV with both the Cohen-Kappa correlation as well as the percentage of matching labels.

TABLE IV: Inter-rater Correlation Between Hand-Labeling and RQC

Cohen-Kappa Correlation	**Percentage Agreement (%)**
0.33	75

The RQC agrees with the human hand-labeling 75% of the time, with a Cohen-Kappa score of 0.33. The Cohen-Kappa score indicates 'fair agreement' between the hand labeling and the RQC in this dataset.

VI. Generating Reflections with GPT-2, GPT-3, and Filtering GPT-2 with the RQC

In Section VI-A we present the performance of the reflection generation system using the GPT-2 model with and without the RQC acting as a filter, and the GPT-3 model without the filter. In Section VI-B we discuss, qualitatively, a sample of the acceptable and unacceptable generated reflections from each model.

A. Reflection Generation

Generation Set-up

We first present the conditions and parameters used for the models during the few-shot generation process: First, the output sequence length was limited to a maximum of 200 tokens for both models. This causes generation to stop when this limit is reached and prevents the model from generating extremely long sequences of text, which was observed on occasion. In almost all cases, a very long output has little value as a reflection, both because the output tended to be nonsensical when that long, and in MI, it is good practice to let a patient speak more than the therapist [6]. We chose the 200 token limit because the average response was roughly 14 words with the longest response at 73 words.

The primers used in the few-shot generation were randomly selected from the priming set described in Section III-B. The generation is launched by appending the primers with the input prompt and response, for which a completion was generated.

When generating reflections with the GPT-2-XL model, eight primers were used (as described in Section IV-A). The decoding parameters were set to the best-performing combination that was explored in Section IV-B, which are as follows: temperature = 0.1, top_k = 100, top_p = 0.8. The repetition penalty was set to 1.3.

Reflections generated with the GPT-3 model were done using both eight and 12 primers. The reflections generated with eight primers provides a direct way to compare the few-shot reflection generation performance of GPT-3 and GPT-2, as eight was the maximum allowed for GPT-2. The 12-primer method was used to explore how well GPT-3 can generate reflections when taking advantage of its much larger context window by adding more priming examples. The GPT-3 engine used for these experiments were run using the DaVinci engine from the OpenAI API [59]. The decoding parameters were set as described in Section IV-B

Reflection Generation

A set of 100 prompt-response pairs from the dataset described in Section III-A was used to evaluate the generative models on their ability to produce acceptable reflections. When selecting the prompt-response pairs, we ensured that the same prompt-response pair did not exist as part of the priming set described in Section III-B. This ensures that the model is not given an unfairly helpful priming set.

TABLE V: Reflection Hit Rate (for 100 trials) by Model

Model Type	**Human-Label Hit Rate**
GPT-2 - 8 primers	0.54
GPT-3 - 8 primers	0.89
GPT-3 - 12 primers	0.74

Table V gives the human-label hit rate when generating reflections using the GPT-2 model with eight primers, and with the GPT-3 model using both eight and 12 primers. The GPT-2 model achieves a 54% hit rate, which is very encouraging – it shows that with just a little priming, the language models can be influenced to perform an important counseling function quite often. It can form the basis of a system that includes the output filtering function, as discussed in Section V.

Table V also shows that both versions of the generation from the GPT-3 model outperform the GPT-2 model, achieving 74% and 89% hit rates. That said, it is more difficult to use the GPT-3 model at the present time, due to its very large size and computational requirements, and thus the hardware systems needed to run it. By contrast, the GPT-2 based models can be run locally on consumer-grade hardware, and smaller versions of it can even be run on lightweight machines.

Table V also shows a surprising performance difference between the GPT-3 model with 8 primers and the GPT-3 model with 12 primers, with the 12-primer version performing worse. This result highlights the importance of tuning input and decoding parameters for each model to the target task, as performance can vary significantly (and not in an intuitive way) based on the parameter values.

Generating Reflections with GPT-2 followed by an RQC Filter

Here we show the results of a system using the GPT-2 reflection generator, in combination with the RQC as a filter, to reduce poor-quality reflections from being presented. In this system the RQC is responsible for providing a label as to whether a reflection is acceptable given the context. It takes as input the same prompt and response given to the model, as well as the generated completion (which is the reflection produced by GPT-2). If the RQC deemed the reflection as not acceptable, the generation system was then re-invoked to generate a new candidate reflection for the model. To increase the likelihood that a different reflection is generated, the primers are resampled from the priming set. This method exploits the generator's ability to produce distinct reflections due to the stochastic nature of the sequence-to-sequence process, including the probabilistic sampling done at the output stage. We place a limit of five attempts to generate an acceptable reflection, to limit computational cost and avoid a possible infinite loop. If no generated reflection was deemed acceptable by the RQC, the fifth reflection generated is passed through as the output.

It is possible to compute an approximation of the anticipated hit rate of the combined generation and RQC filtering process. Under the (inexact) assumption that there is no limit of 5 filtering operations (and the generation/reflection loop would continue until successful), the approximate hit rate can be determined based on the generator's native hit rate (as shown in the GPT_2 row of Table 3) and the true negative rate (the specificity) of the RQC as follows:

$$FHitRate = 1 - (1 - NativeHitRate)(1 - RQC\_Specificity) \quad (1)$$

The *NativeHitRate* when using the GPT-2 model is 0.54 (from Table 3), while the *RQC_Specificity* for the RQC model with *BERT-base MSRP+Reflection Fine-tuned* is 0.71 (as given in Table 1). Using these two values we obtain an estimated *FHitRate* of 0.87.

Table VI gives the results of one run of an experiment using the filtered generation process, in which the RQC filtering loop was limited to a maximum of five cycles for each input. The same 100 examples as were used to produce Table 3 were used as the input for this experiment. The hand-labelled reflection generation hit rate was improved from 0.54 to 0.73. This is a significant improvement on the hit rate and makes the use of a reflections more practical.

This is less than the estimated value of 0.87, perhaps in part due to the 5 filter-cycle limit, which contributed to three poor quality reflections in the results. If those three poor quality results had been acceptable, the hit rate would increase to 0.77.

TABLE VI: GPT-2 Reflection Generation Hit Rate with and without Filtering

Model/Process	**Human Labelled Reflection Hit Rate**
GPT-2	0.54
GPT-2 w. RQC Filtering loop	0.73

B. Qualitative Discussion of Reflection Generated Reflections

The quantitative results above indicate that we can achieve some success generating MI-consistent reflections. Here we review specific examples of the generated outputs qualitatively, to provide a sense of what is promising about the results, and what still needs work.

Table VII shows a set of six acceptable (by human labeling) reflections that were generated from one of the three models listed in Table V, when producing the data in that table. Each row of the table provides the input prompt and response, as well as the generated reflection, and which specific model produced it.

One can informally observe that the reflections in Table VII do read as if a human created them and could be appropriate statements that a therapist could use. It can also be observed that the model appears robust when dealing with grammatically incorrect inputs - Examples 1 and 6 have spelling errors in the responses, but the reflections are still produced with correct grammar and address the responses appropriately.

Although all examples in Table VII are acceptable MI-consistent reflections, there is still a variation in the level of each reflection, with a mix of simple and complex reflections. Recall that complex reflections not only follow from the response but also have a link to an insight or a suggestion. Consider Example 1, where the patient expresses the desire to not embarrass their children by smoking in front of the children's friends. The reflection suggests a deeper struggle the patient may be facing about their smoking. This kind of reflection often encourages a patient to contemplate such issues further which is one of the key goals of MI-based therapy. We also know that, even if the link (in this case to a struggle) is incorrect, it is still therapeutic for the patient to correct the suggestion and to perhaps contemplate the real link, for them.

The reflections in Examples 2,4, and 6 also draw from an idea that was not present in the patient's response, and so could help the patient contemplate the prompt further. Consider Example 2, which connects the idea of making a "conscious decision not to smoke" to the dynamic of an interaction during driving that is about buying cigarettes. Similarly, Example 4 makes a guess that the person was working "long hours" recently. A good therapist will make a guess like this to keep the patient contemplating cause and effect about their habit.

Sometimes a reflection does not need to be complex to have a therapeutic effect [6]. A simple restatement of the response, in different words, is sufficient to build rapport with a patient and move the conversation forward. Examples 3 and 5 in Table VII show reflections that such simple reflections.

These examples demonstrate that within the context of these prompts and responses, few-shot generation with GPT-2 and GPT-3 can produce reflections consistent with the ones a therapist may use in an MI session. The language sounds natural, follows many of the principles of MI, and could fulfill the goal of moving the conversation forward.

However, not every reflection produced by these models are as successful. Several examples of poor reflections are shown in Table VIII. There are were two general categories of failure: The first are statements that are coherent and follow from the context prompt and response but fail to adhere to MI principles, as illustrated in Examples

TABLE VII: Acceptable and Better Reflection Examples

	Prompt	Response	Reflection	Model
1	Think back to the time when you were able to prevent yourself from smoking. What made it different from when you did smoke?	i dont want to embarse my kids in front of their friends	You didn't want them to see how much you struggled with this problem.	GPT-2
2	Iwant you to describe a time or situation where you experienced the reason "Relationship Problem" and it did not lead you to smoke	I refused her request to stop at the gas station to buy cigarettes on the way to work. I didn't want one and wanted to demonstrate it to her.	You are making a conscious decision not to smoke.	GPT-3 - 8 Primers
3	Let me see if I understand, smoking is helpful because it provides a useful distraction for you. Are there other things that are good about smoking? If so, please tell me	I'm able to meet new people	You enjoy meeting new people when you smoke.	GPT-3 - 8 Primers
4	Think back to the time when you were able to prevent yourself from smoking. What made it different from when you did smoke?	I was very stressed at work	You'vebeen working long hours lately.	GPT-2 with RQC Filter
5	Think back to the time when you were able to prevent yourself from smoking. What made it different from when you did smoke?	I felt like I was affecting my chances at having childnre	You're worried about the impact on your fertility.	GPT-3- 8 Primers
6	Let me see if I understand, smoking is helpful because it provides a useful distraction for you. Are there other things that are good about smoking? If so, please tell me	I'm able to meet new people	You met someone interesting.	GPT-2 with RQC Filter

TABLE VIII: Not Acceptable Reflection Examples

	Prompt	Response	Reflection	Model	RQC Label
1	Think back to the time when you were able to prevent yourself from smoking. What made it different from when you did smoke?	The difference was that I was paying important bills	When you have a choice between spending money on cigarettes vs other things, you will choose cigarettes over those other things.	GPT-2 with RQC Filter	0
2	Please describe a time where you experienced stress and didn't end up smoking	i cant really its been so long	Stress causes anxiety which leads to depression.	GPT-2 with RQC Filter (Case where 5-cycle limit was hit)	0
3	Think back to the time when you were able to prevent yourself from smoking. What made it different from when you did smoke?	it feels good to know i can resist if i try	The only way to stop smoking is to quit completely	GPT-2	0
4	Please describe a time where you experienced stress and didn't end up smoking	i cant really its been so long	You feel stressed about smoking.	GPT-3 8 Primers	1
5	Please describe a time where you dreaded tasting the flavour of smoking a cigarette and you didn't end up smoking	plenty of times just sat at home, especially when i already dont feel well	you know that there are many flavours out there	GPT-2 with RQC Filter	1
6	Please describe a situation where you thought about being a good role model and thatcause you not to smoke	onmy first date with my current boyfriend	You found someone who smokes and he/she doesn't judge you	GPT-2	1

1 and 3 in Table VIII. Example 1 responds to the patient's statement about having important bills to pay by stating that they would choose to buy cigarettes rather than paying bills. This directly misrepresents the patient's intent which stated the opposite. It is also phrased in an accusatory tone, something that therapists are trained to avoid as much as possible as it produces discord in the patient-therapist relationship.

The generated reflection in Example 3 dismisses the patient's response about their attempt to resist smoking, stating rather unhelpfully that the patient needs to stop smoking completely to quit. This kind of reflection falls into the 'expert trap' [6], where the therapist offers direct guidance instead of letting the patient draw their own conclusions.

The second common category of failure occurs when the generated completion does not follow from the context prompt and response, and thus cannot be considered a reflection. This kind of failure occurs in Examples 2,4,5, and 6 in Table VIII. Examples 2 and 4 have the same context prompt and response with the reflection being generated by two different models, but both models failed to generate an acceptable reflection. In both instances, the generated reflections focused on addressing stress whereas the patient response was about not being able to remember an example where they experienced stress without it leading them to smoke.

Example 5 is a reflection that appears to focus on a specific word ('flavour') in the prompt and ignores the essence of the patient's response. Finally, the reflection in Example 6 fails to capture the context of the response and makes a leap about the patient's boyfriend for which there is no evidence.

Beyond these two categories of failure, there are some rare instances where the models would produce incoherent reflections or just produce a blank. These types of failure case were not present in the reflections produced from this test set.

The failure cases show that the reflection generation is still imperfect and there is significant room to improve. One positive aspect of these failure cases is that they are less about actual grammar or language issues, and more about the nuance of human conversation. Including a filter like the RQC can help identify these failure cases, although as we see from the 'RQC Label' column in Table VIII, it is only able to capture some of the failures and thus also has significant room to be improve.

VII. Limitations

In this work we have attempted to show how few-shot generation might be used to generate reflections within MI-style smoking cessation conversations. The most significant limitation in this work is that the reflections were evaluated within the scope of the specific questions and responses from the smoking cessation dataset described in Section III-A, and we cannot conclude that the approach is generalizable beyond that dataset. However, since we have employed few-shot learning there is reason to believe that primers specific to other questions would also succeed.

Another limitation of these results is the low test-set sample size, which was necessary due to the significant labour required to label the reflections produced by the generative models. Since trained MI-experts are difficult to find and their availability is limited, only the primer set was labeled and edited by a trained MI expert. All other labeling was done by the authors, whose experience with labeling MI came from readings [6] and consultations with MI practitioners.

VIII. Conclusions and Future Work

This study demonstrates a first step towards moving therapeutic chatbots from responding repetitively and predictably, to a form where they can respond with human-like specificity and understanding. We demonstrated that transformer-based language models can be effective at producing MI-consistent reflections when responding to participant statements within a specific domain, even when working in a low-data environment. We also showed that the larger GPT-3 model produces acceptable reflections more frequently than the GPT-2 XL based model. As one method for mitigating unacceptable reflections, we trained a reflection quality classifier (RQC) to determine whether a reflection is acceptable and showed that it has fair correlation to human labels. We also demonstrated a GPT-2 model can be used with the RQC as a filter to build a system that can produce reflections at a better rate than just using GPT-2 with few-shot learning. This comes at the cost of slightly higher memory usage and longer execution time needed to generate extra reflections.

In the future, we plan to fine-tune the generative models to see if this can improve the hit rate. We would also like to be able to control whether a simple or more complex reflection will be produced, as different situations in a full MI conversation may call for one or the other.

Finally, the reflection generation system was evaluated solely on its ability to produce acceptable MI-consistent reflections and not on therapeutic impact. To evaluate the latter, the reflection generation system needs to be integrated into a full chatbot and measured as an intervention.

References

[1] S. Davis, *Community mental health in Canada, revised and expanded edition: Theory, policy, and practice*. UBC Press, 2013.

[2] G. L. Fricchione, C. P. C. Borba, A. Alem, T. Shibre, J. R. Carney, and D. C. Henderson, "Capacity building in global mental health: Professional training," *Harvard Review of Psychiatry*, vol. 20, no. 1, pp. 47–57, 2012, pMID: 22335182. [Online]. Available: https://www.tandfonline.com/doi/abs/10.3109/10673229.2012.655211

[3] N. Kates, G. Mazowita, F. Lemire, A. Jayabarathan, R. Bland, P. Selby, T. Isomura, M. Craven, M. Gervais, and D. Audet, "The evolution of collaborative mental health care in canada: A shared vision for the future," *Canadian Journal of Psychiatry*, vol. 56, no. 5, p. I1, 2011.

[4] G. Whitfield and A. Davidson, *Cognitive Behavioural Therapy Explained (1st ed.)*. CRC Press., 2007.

[5] L. Nisbet, "Dialectical behavioural therapy," in *Child and Adolescent Mental Health*. CRC Press, 2021, pp. 448–456.

[6] W. R. Miller and S. Rollnick, *Motivational interviewing: Helping people change*. Guilford press, 2012.

[7] D. D. Luxton, "Chapter 1 - an introduction to artificial intelligence in behavioral and mental health care," in *Artificial Intelligence in Behavioral and Mental Health Care*, D. D. Luxton, Ed. San Diego: Academic Press, 2016, pp. 1–26. [Online]. Available: https://www.sciencedirect.com/science/article/pii/B9780124202481000015

[8] D. Adiwardana, M.-T. Luong, D. R. So, J. Hall, N. Fiedel, R. Thoppilan, Z. Yang, A. Kulshreshtha, G. Nemade, Y. Lu, and Q. V. Le, "Towards a human-like open-domain chatbot," 2020.

[9] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei, "Language models are few-shot learners," 2020.

[10] A. Madotto, Z. Liu, Z. Lin, and P. Fung, "Language models as few-shot learner for task-oriented dialogue systems," 2020.

[11] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," 2019.

[12] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language models are unsupervised multitask learners," *OpenAI blog*, vol. 1, no. 8, p. 9, 2019.

[13] A. Braillon and F. Taiebi, "Practicing "reflective listening" is a mandatory prerequisite for empathy," *Patient Education and Counseling*, vol. 103, no. 9, pp. 1866–1867, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0738399120301828

[14] E. Rautalinko, H.-O. Lisper, and B. Ekehammar, "Reflective listening in counseling: effects of training time and evaluator social skills," *American journal of psychotherapy*, vol. 61, no. 2, pp. 191–209, 2007.

[15] C. J. Heckman, B. L. Egleston, and M. T. Hofmann, "Efficacy of motivational interviewing for smoking cessation: a systematic review and meta-analysis," *Tobacco Control*, vol. 19, no. 5, pp. 410–416, 2010. [Online]. Available: https://tobaccocontrol.bmj.com/content/19/5/410

[16] I. Diener, M. Kargela, and A. Louw, "Listening is therapy: Patient interviewing from a pain science perspective," *Physiotherapy Theory and Practice*, vol. 32, no. 5, pp. 356–367, 2016, pMID: 27351690. [Online]. Available: https://doi.org/10.1080/09593985.2016.1194648

[17] F. Almusharraf, J. Rose, and P. Selby, "Engaging unmotivated smokers to move toward quitting: Design of motivational interviewing–based chatbot through iterative interactions," *J Med Internet Res*, vol. 22, no. 11, p. e20251, Nov 2020. [Online]. Available: https://www.jmir.org/2020/11/e20251

[18] F. Jódar-Sánchez, L. Carrasco Hernández, F. J. Núñez-Benjumea, M. A. Mesa González, J. Moreno Conde, C. L. Parra Calderón, L. Fernandez-Luque, S. Hors-Fraile, A. Civit, P. Bamidis, and F. Ortega-Ruiz, "Using the social-local-mobile app for smoking cessation in the smokefreebrain project: Protocol for a randomized controlled trial." *JMIR research protocols*, vol. 7, p. e12464, Dec 2018.

[19] D. M. Low, K. H. Bentley, and S. S. Ghosh, "Automated assessment of psychiatric disorders using speech: A systematic review," *Laryngoscope Investigative Otolaryngology*, vol. 5, no. 1, pp. 96–116, 2020. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/lio2.354

[20] J. Nurmi, K. Knittle, T. Ginchev, F. Khattak, C. Helf, P. Zwickl, C. Castellano-Tejedor, P. Lusilla-Palacios, J. Costa-Requena, N. Ravaja, and A. Haukkala, "Engaging users in the behavior change process with digitalized motivational interviewing and gamification: Development and feasibility testing of the precious app," *JMIR Mhealth Uhealth*, vol. 8, no. 1, p. e12884, Jan 2020. [Online]. Available: https://mhealth.jmir.org/2020/1/e12884

[21] S. Park, J. Choi, S. Lee, C. Oh, C. Kim, S. La, J. Lee, and B. Suh, "Designing a chatbot for a brief motivational interview on stress management: Qualitative case study," *J Med Internet Res*, vol. 21, no. 4, p. e12231, Apr 2019. [Online]. Available: https://www.jmir.org/2019/4/e12231/

[22] S. I. Pratt, S. J. Bartels, K. T. Mueser, J. A. Naslund, R. Wolfe, H. S. Pixley, and L. Josephson, "Feasibility and effectiveness of an automated telehealth intervention to improve illness self-management in people with serious psychiatric and medical disorders." *Psychiatric rehabilitation journal*, vol. 36, pp. 297–305, Dec 2013.

[23] T. N. Stephens, A. Joerin, M. Rauws, and L. N. Werk, "Feasibility of pediatric obesity and prediabetes treatment support through tess, the ai behavioral coaching chatbot," *Translational behavioral medicine*, vol. 9, no. 3, pp. 440–447, 2019.

[24] I. Gulenko, "Chatbot for it security training: Using motivational interviewing to improve security behaviour." in *AIST (Supplement)*, 2014, pp. 7–16.

[25] S. Shen, C. Welch, R. Mihalcea, and V. Pérez-Rosas, "Counseling-style reflection generation using generative pretrained transformers with augmented context," in *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*. 1st virtual meeting: Association for Computational Linguistics, Jul. 2020, pp. 10–20. [Online]. Available: https://aclanthology.org/2020.sigdial-1.2

[26] Z. Dai, Z. Yang, Y. Yang, J. Carbonell, Q. V. Le, and R. Salakhutdinov, "Transformer-xl: Attentive language models beyond a fixed-length context," 2019.

[27] T. Olmstead, K. M. Carroll, M. Canning-Ball, and S. Martino, "Cost and cost-effectiveness of three strategies for training clinicians in motivational interviewing," *Drug and alcohol dependence*, vol. 116, no. 1-3, pp. 195–202, 2011.

[28] P. Budzianowski and I. Vulić, "Hello, it's gpt-2–how can i help you? towards the use of pretrained language models for task-oriented dialogue systems," *arXiv preprint arXiv:1907.05774*, 2019.

[29] K. Elkins and J. Chun, "Can gpt-3 pass a writer's turing test," *Journal of Cultural Analytics*, vol. 2371, p. 4549, 2020.

[30] T. Klein and M. Nabi, "Learning to answer by learning to ask: Getting the best of gpt-2 and bert worlds," *arXiv preprint arXiv:1911.02365*, 2019.

[31] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," *arXiv preprint arXiv:1706.03762*, 2017.

[32] T. Mikolov, M. Karafiát, L. Burget, J. Černockỳ, and S. Khudanpur, "Recurrent neural network based language model," in *Eleventh annual conference of the international speech communication association*, 2010.

[33] T. Mikolov, S. Kombrink, L. Burget, J. Černockỳ, and S. Khudanpur, "Extensions of recurrent neural network language model," in *2011 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2011, pp. 5528–5531.

[34] V. Liu and J. R. Curran, "Web text corpus for natural language processing," in *11th Conference of the European Chapter of the Association for Computational Linguistics*, 2006.

[35] J. M. Patel, "Introduction to common crawl datasets," in *Getting Structured Data from the Internet*. Springer, 2020, pp. 277–324.

[36] Y. Zhu, R. Kiros, R. Zemel, R. Salakhutdinov, R. Urtasun, A. Torralba, and S. Fidler, "Aligning books and movies: Towards story-like visual explanations by watching movies and reading books," in *The IEEE International Conference on Computer Vision (ICCV)*, December 2015.

[37] L. Denoyer and P. Gallinari, "The wikipedia xml corpus," in *International Workshop of the Initiative for the Evaluation of XML Retrieval*. Springer, 2006, pp. 12–19.

[38] L. Yan, Y. Zheng, and J. Cao, "Few-shot learning for short text classification," *Multimedia Tools and Applications*, vol. 77, no. 22, pp. 29 799–29 810, 2018.

[39] B. Peng, C. Zhu, C. Li, X. Li, J. Li, M. Zeng, and J. Gao, "Few-shot natural language generation for task-oriented dialog," *arXiv preprint arXiv:2002.12328*, 2020.

[40] P. Liu, W. Yuan, J. Fu, Z. Jiang, H. Hayashi, and G. Neubig, "Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing," 2021. [Online]. Available: https://arxiv.org/abs/2107.13586

[41] D. Yoshida, A. Ettinger, and K. Gimpel, "Adding recurrence to pretrained transformers for improved efficiency and context size," 2020.

[42] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. Salakhutdinov, and Q. V. Le, "Xlnet: Generalized autoregressive pretraining for language understanding," 2020.

[43] A. Holtzman, J. Buys, L. Du, M. Forbes, and Y. Choi, "The curious case of neural text degeneration," *arXiv preprint arXiv:1904.09751*, 2019.

[44] N. S. Keskar, B. McCann, L. R. Varshney, C. Xiong, and R. Socher, "Ctrl: A conditional transformer language model for controllable generation," 2019.

[45] Y. Zhang, S. Sun, M. Galley, Y. Chen, C. Brockett, X. Gao, J. Gao, J. Liu, and B. Dolan, "Dialogpt: Large-scale generative pre-training for conversational response generation," *CoRR*, vol. abs/1911.00536, 2019. [Online]. Available: http://arxiv.org/abs/1911.00536

[46] S. Minaee, N. Kalchbrenner, E. Cambria, N. Nikzad, M. Chenaghlu, and J. Gao, "Deep learning–based text classification: A comprehensive review," *ACM Comput. Surv.*, vol. 54, no. 3, Apr. 2021. [Online]. Available: https://doi.org/10.1145/3439726

[47] S. T. Dumais, G. W. Furnas, T. K. Landauer, S. Deerwester, and R. Harshman, "Using latent semantic analysis to improve access to textual information," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ser. CHI '88. New York, NY, USA: Association for Computing Machinery, 1988, p. 281–285. [Online]. Available: https://doi.org/10.1145/57167.57214

[48] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Advances in neural information processing systems*, 2013, pp. 3111–3119.

[49] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer, "Deep contextualized word representations," *arXiv preprint arXiv:1802.05365*, 2018.

[50] L. Xiao, G. Wang, and Y. Zuo, "Research on patent text classification based on word2vec and lstm," in *2018 11th International Symposium on Computational Intelligence and Design (ISCID)*, vol. 1. IEEE, 2018, pp. 71–74.

[51] Y. Sung, S. Jang, Y.-S. Jeong, J. Hyuk *et al.*, "Malware classification algorithm using advanced word2vec-based bi-lstm for ground control stations," *Computer Communications*, vol. 153, pp. 342–348, 2020.

[52] J. Brixey, R. Hoegen, W. Lan, J. Rusow, K. Singla, X. Yin, R. Artstein, and A. Leuski, "Shihbot: A facebook chatbot for sexual health information on hiv/aids," in *Proceedings of the 18th annual SIGdial meeting on discourse and dialogue*, 2017, pp. 370–373.

[53] S. Olafsson, B. Wallace, and T. Bickmore, "Towards a computational framework for automating substance use counseling with virtual agents," in *Proceedings of the 19th International Conference on Autonomous Agents and MultiAgent Systems*, ser. AAMAS '20. Richland, SC: International Foundation for Autonomous Agents and Multiagent Systems, 2020, p. 966–974.

[54] R. Fulmer, A. Joerin, B. Gentile, L. Lakerink, and M. Rauws, "Using psychological artificial intelligence (tess) to relieve symptoms of depression and anxiety: randomized controlled trial," *JMIR mental health*, vol. 5, no. 4, p. e64, 2018.

[55] F. Almusharraf, "Motivating smokers to quit through a computer-based conversational system," Ph.D. dissertation, University of Toronto (Canada), 2019.

[56] L. He, E. Basar, R. W. Wiers, M. L. Antheunis, and E. Krahmer, "Can chatbots help to motivate smoking cessation? a study on the effectiveness of motivational interviewing on engagement and therapeutic alliance," *BMC Public Health*, vol. 22, no. 1, p. 726, Apr 2022. [Online]. Available: https://doi.org/10.1186/s12889-022-13115-x

[57] V. Pérez-Rosas, R. Mihalcea, K. Resnicow, S. Singh, and L. An, "Building a motivational interviewing dataset," in *Proceedings of the Third Workshop on Computational Linguistics and Clinical Psychology*. San Diego, CA, USA: Association for Computational Linguistics, Jun. 2016, pp. 42–51. [Online]. Available: https://aclanthology.org/W16-0305

[58] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. L. Scao, S. Gugger, M. Drame, Q. Lhoest, and A. M. Rush, "Huggingface's transformers: State-of-the-art natural language processing," 2020.

[59] OpenAI, "Openai api," Nov 2021. [Online]. Available: https://openai.com/api/

[60] Y. Wu, M. Schuster, Z. Chen, Q. V. Le, M. Norouzi, W. Macherey, M. Krikun, Y. Cao, Q. Gao, K. Macherey, J. Klingner, A. Shah, M. Johnson, X. Liu, L. Kaiser, S. Gouws, Y. Kato, T. Kudo, H. Kazawa, K. Stevens, G. Kurian, N. Patil, W. Wang, C. Young, J. Smith, J. Riesa, A. Rudnick, O. Vinyals, G. Corrado, M. Hughes, and J. Dean, "Google's neural machine translation system: Bridging the gap between human and machine translation," *CoRR*, vol. abs/1609.08144, 2016. [Online]. Available: http://arxiv.org/abs/1609.08144

[61] W. B. Dolan and C. Brockett, "Automatically constructing a corpus of sentential paraphrases," in *Proceedings of the Third International Workshop on Paraphrasing (IWP2005)*, 2005. [Online]. Available: https://www.aclweb.org/anthology/I05-5002

[62] I. Dagan, B. Dolan, B. Magnini, and D. Roth, "Recognizing textual entailment: Rational, evaluation and approaches–erratum," *Natural Language Engineering*, vol. 16, no. 1, pp. 105–105, 2010.

[63] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, and X. Zheng, "TensorFlow: Large-scale machine learning on heterogeneous systems," 2015, software available from tensorflow.org. [Online]. Available: https://www.tensorflow.org/

[64] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014.

[65] G. Becker, "Creating comparability among reliability coefficients: the case of cronbach alpha and cohen kappa," *Psychological reports*, vol. 87, no. 3_suppl, pp. 1171–1182E, 2000.

[66] M. L. McHugh, "Interrater reliability: the kappa statistic," *Biochemia medica*, vol. 22, no. 3, pp. 276–282, 2012.