Generation, Distillation and Evaluation of Motivational Interviewing-Style Reflections with a Foundational Language Model

Anonymous EACL submission

Abstract

Large Foundational Language Models are capable of performing many tasks at a high level but are difficult to deploy in many applications because of their size and proprietary ownership. Many will be motivated to distill specific capabilities of foundational models into smaller models that can be owned and controlled. In the development of a therapeutic chatbot, we wish to distill a capability known as reflective listening, in which a therapist produces *reflections* of client speech. These reflections either restate what a client has said, or connect what was said to a relevant observation, idea or guess that encourages and guides the client to continue contemplation. In this paper, we present a method for distilling the generation of reflections from a Foundational Language Model (GPT-4) into smaller models. We first show that GPT-4, using zero-shot prompting, can generate reflections at near 100% success rate, superior to all previous methods. Using reflections generated by GPT-4, we fine-tune different sizes of the GPT-2 family. The GPT-2-small model achieves 83% success on a hold-out test set and the GPT-2 XL achieves 90% success. We also show that GPT-4 can help in the labor-intensive task of evaluating the quality of the distilled models, using it as a zero-shot classifier. Using triple-human review as a guide, the classifier achieves a Cohen-Kappa of 0.66, a substantial inter-rater reliability figure.

1 Introduction

Motivational Interviewing (MI) is a counselling technique that is used to guide people towards behaviour change (Miller and Rollnick, 2012). MI has seen success in smoking cessation (Lindson et al., 2019) and alcohol consumption reduction (Nyamathi et al., 2010), among other behaviours. Our long-term goal is to automate MI-based therapeutic conversations in smoking cessation.

A key technique in MI (and many other talk therapies) is *reflective listening*, a conversational approach in which a clinician which mirrors the client's thoughts and emotions, enabling them to recognize their own beliefs and contradictions (Miller and Rollnick, 2012). The core skill of reflective listening is to respond to client utterances with a *reflection*. Reflections are divided into two major types: *simple* reflections which rephrase what a client has said, and *complex* reflections which attempt to infer something based on a recent utterance, or to guess something based on general knowledge (Miller and Rollnick, 2012). Both types of reflections are illustrated in the conversation snippet in Table 1.

Conversation
MI Clinician: What are some things you don't like about your smoking addiction? **Client:** I don't like making other people uncomfortable with my smoking. **MI Clinician** (Simple Reflection): You don't enjoy making people feel uncomfortable with your smoking. **MI Clinician** (Complex Reflection): You might be feeling self-conscious about your smoking.

Table 1: Example of Simple vs Complex Reflection

There has been recent work to automate the generation and classification of MI reflections using GPT-2 (Radford et al., 2019) and GPT-3 (Brown et al., 2020). (Ahmed et al., 2022) showed that a few-shot prompted GPT-3 generates MI reflections scoring over 89% accuracy from human annotation and (Shen et al., 2020) demonstrated a fine-tuned GPT-2 generates reflections which are scored by human reviewers as nearly identical to clinician curated reflections. Furthermore, (Ahmed, 2022) showed that a fine-tuned BERT (Devlin et al., 2019) can classify reflections as acceptable at 80% accu-

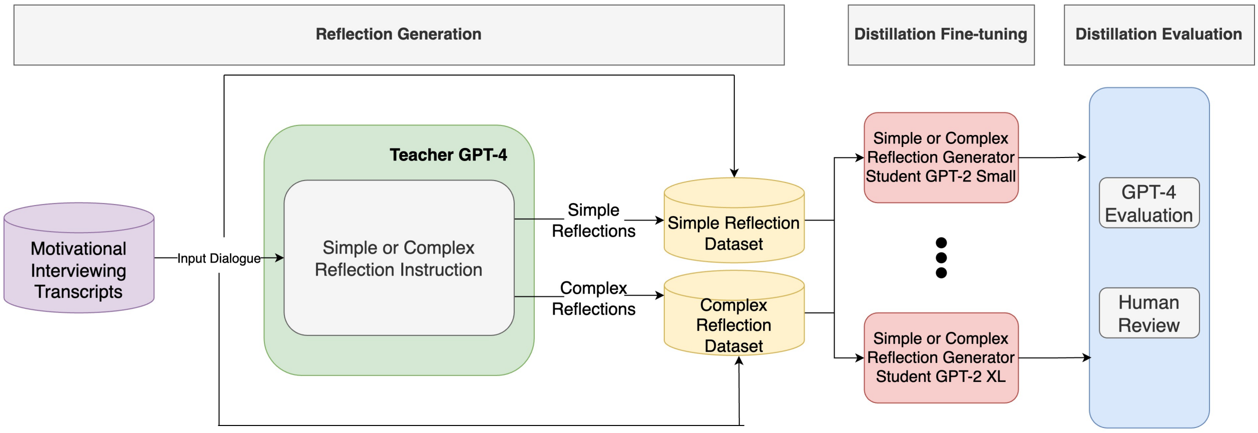

Figure 1: Knowledge Distillation Overview

racy. In this work we explore the use of zero-shot prompting of GPT-4 (OpenAI, 2023) to both generate and classify MI reflections. We use the high-quality reflections from the generator to fine-tune a smaller, proprietary models. The latter provides greater privacy for sensitive health communications since the information pathways can be fully controlled when a model is owned.

In collaboration with MI-experts, we designed prompts to generate both simple and complex reflections and classify them with GPT-4. We present a method to distill the reflection generation capability from GPT-4. A dataset was created consisting of questions (that were presented to clients), their answers and generated GPT-4 reflections. These are used to fine-tune smaller language models, and we sought to determine the the trade-off between size of the smaller model and its performance.

In the larger context of a smoking-cessation chatbot that would use the generated reflections, there are situations when a simple reflection is called for, and other times when complex reflection is appropriate (Miller and Rollnick, 2012). For this reason we will distill two models, one for each type of reflection.

Figure 1 illustrates the overall approach used in this work. The fine-tuning datasets are created based on portions of transcripts from a previous chatbot created by the authors (which will be cited upon acceptance of the work) and simple or complex reflections generated by GPT-4. Next, as a form of knowledge distillation, we fine-tune the GPT-2 (Radford et al., 2019) family of models on the simple reflection or complex reflection dataset. To evaluate the student models we employ both human reviewers and use the GPT-4 model itself as a zero-shot classifier. That classification is done in two stages, the first to check for adherency to the principles of MI (Miller and Rollnick, 2012), and then to classify acceptable MI-adherent reflections as simple or complex. The idea of using a large foundational model as an zero-shot evaluator has just begun to appear in the literature (Kamalloo et al., 2023; Chiang and Lee, 2023) and is not yet well studied. If it can be shown to be successful, it will reduce the costly human effort in determining the effectiveness of distilled and other models. Previous works in MI reflection generation such as (Shen et al., 2020) and (Ahmed, 2022) have used human curated datasets to train classifiers.

The contributions of this paper are: (1) State-of-the-art success rate in generation of reflections (2) An example of end-to-end task-specific distillation from a foundational language model and (3) demonstration of the effectiveness of using a foundational language models to evaluate reflections, which has the potential to reduce the amount of human labour in generative model work.

2 Related Work

Generative Reflections

There have been past attempts to generate MI reflections using transformer-based language models. The work in (Shen et al., 2020) showed that GPT-2 could generate counseling-style MI reflections by fine-tuning on the dialogue context and responses retrieved from similar counseling sessions. Human reviewers scored a test set of generated reflections at 4.13 on a 5-point likert scale while scoring known-good reflections at 3.84, suggesting that the human reviewers preferred the quality of generated reflections over known-good ones.

These reflections were proposed to be used in clinician training, allowing for easier access to context specific reflections. This work was subsequently improved (Shen et al., 2022) by including commonsense and domain specific knowledge while generating responses, similar to what counselors do. These generated reflections scored lower on human review scores. On reflection coherence, accuracy and preference, human reviewers scored ground-truth reflections higher than generated domain specific reflections.

(Ahmed et al., 2022) investigated the use of prompting and fine-tuning transformer-based language models to generate and classify MI reflections for smoking cessation. Human reviewers scored reflection acceptability on a prompted GPT-2 XL as 54%, a prompted GPT-3 as 89%, and a fine-tuned GPT-2 XL at 80%. For reflection classification, (Ahmed, 2022) fine-tuned a BERT model to achieve 81% accuracy in classifying reflections.

We view the previous work in MI reflection generation and classification as preliminary and seek to build upon it. With GPT-4, our goal is to create an improved reflection generation which scores higher with human reviewers than that of (Shen et al., 2022) and (Ahmed, 2022), and create a more accurate reflection classifier than (Ahmed, 2022) which agrees with human decisions.

Knowledge Distillation

Knowledge distillation is a technique in machine learning where a student model is trained to reproduce the behaviour of a teacher model, typically to achieve model compression (Gu et al., 2023). (Hinton et al., 2015) showed the first method of knowledge distillation in which a student neural network was trained to mimic a teacher model's performance on MNIST and speech recognition. The student was trained using a loss function which optimized a combined objective of minimizing the loss of the ground-truth labels and the teacher model's output logits as labels.

Knowledge distillation has since been successfully applied to language models, with DistilBERT (Sanh et al., 2020), a transformer-based language model trained using a loss function for the student model similar to (Hinton et al., 2015) for the purpose of compressing BERT (Devlin et al., 2019). Subsequently, researchers have also considered Task-specific knowledge distillation, which seeks to distill a subset of the teacher model's capability into the student. Two examples of this are (Tang et al., 2019) which sought to distill only sentiment analysis, and (Liu et al., 2022) which focused on the tasks specific to the GLUE dataset benchmark (Wang et al., 2019).

Other knowledge distillation works use different loss functions during training, while others employed pre-trained models as the student. (He et al., 2022) showed a method for task-specific knowledge distillation using pre-trained transformer language models as the student and fine-tuning for training. First, a teacher language model is instructed to generate a dataset of additional prompts and output text using an initial set of prompts. Next, this dataset is annotated for data quality and used to fine-tune the smaller student models.

The Self-Instruct approach (Wang et al., 2023) is another application of knowledge distillation which fine-tuned a pre-trained language model. First, a set of 175 seed prompts (describing text instructions for many tasks) were created and used to generate more instructions using GPT-3 (Brown et al., 2020). Next, GPT-3 also generates inputs for the instructions and then the corresponding output. This creates a text dataset of instructions, inputs and outputs. Finally, the dataset is used to fine-tune GPT-3, the same model which generated the dataset. Motivated by Self-instruct, (Taori et al., 2023) created Alpaca, an instruction following LLaMA (Touvron et al., 2023) language model created through fine-tuning on text generated by InstructGPT. The Alpaca method also uses GPT-3 to generate a knowledge distillation dataset, but shrinks the student architecture to the LLaMA-7B model (Touvron et al., 2023), a compression of 25 times. Alpaca's quality of generation were shown to be close to the GPT-3 teacher model, showing that this method of knowledge distillation through generated text can be used to create models a fraction of the size with competitive performance.

The present work combines ideas from previous research in generative MI reflections and knowledge distillation. We use a style of zero-shot prompting similar to (Wang et al., 2023) with GPT-4 to generate MI reflections with the same goal as (Shen et al., 2020) and (Ahmed, 2022). Next, we distill knowledge by fine-tuning smaller transformer-based language models similar to (He et al., 2022).

3 Method

The goals of this paper are to generate high-success rate reflections using GPT-4, to distill that capability into smaller models and measure their success rate, and to determine how well a zero-shot prompt-based GPT-4 model can evaluate the quality of reflections. This section describes the methods for each of these steps.

3.1 Dataset Collection

To generate MI reflections from GPT-4, we need input questions and answers from a MI conversation. Mentioned previously, we use transcripts from the smoking cessation MI chatbot created by the authors. Table 2 shows an excerpt of a conversation transcript. The chatbot adopts a pattern of asking open-ended questions (QUESTION), retrieving answers (ANSWER), and generating reflections (REFLECTION) as shown in Table 2. We gather question and answers without the reflection as inputs to generate a reflection with GPT-4. In total, 4194 question-answer pairs are divided into 2394 training set examples, 599 validation set examples, and 1201 holdout testing set examples.

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 2: MI Chatbot Transcript Excerpt

3.2 Reflection Generation with GPT-4

Reflection generation is done using zero-shot prompting with GPT-4. We use the question-answer pairs described in Section 3.1 with a prepended instruction to generate either a simple or complex reflection. The input prompt and reflection are gathered into a dataset, and used to fine-tune student models, as discussed in Section 3.3.

The instruction for simple and complex reflection generation prompts were developed iteratively on a private test set. First, we hand-wrote an initial prompt and tested it on just a few (1-5) examples. We then increased the size of the test set, noting the examples in which the prompt generated non-MI-adherent reflections, and made modifications accordingly. While evolving the prompt we prioritized maintaining its generality, ensuring that the language use would accommodate many examples, rather than just a few specific ones. For example, in one of the iterations, we noticed that a few generated reflections included questions rather than statements, making these reflections non MI-adherent. The prompt was modified by adding the sentence "The reflection must be a statement and not a question", which is a general instruction. Throughout this iterative design process, we also consulted with MI-experts to get feedback and suggestions on the wording of the prompt.

The full prompt for generating reflections with GPT-4 uses OpenAI's chat-complete (OpenAI, 2023) format, which divides the input prompt into three segments: *System Role*, *System Message*, and *User Message*. The *System Role* is the instruction of the desired task, which in this work is the prompt for generating a simple or complex reflection. The *System Message* and *User Message* are questions and answers, respectively, from our MI dataset like the one seen in Table 2. Figure 2 shows the full prompt for simple and complex reflection generation, with an example for each. Additionally, the prompt for simple and complex reflection generation can be viewed by itself in Table 5, Appendix A. Hereinafter, we refer to GPT-4 for reflections as the GPT-4 Reflection Generator.

We perform a separate validation of the GPT-4 Reflection Generator through a human review. This is described in Section 3.6.

3.3 Fine-tuning Knowledge Distillation Process

After gathering the dataset of MI conversation questions, answers, and GPT-4 generated reflections, we use fine-tuning to distill that reflection capability in a student model. We motivate this method by noting that state-of-the-art foundational language models such as GPT-4 (OpenAI, 2023) do not provide

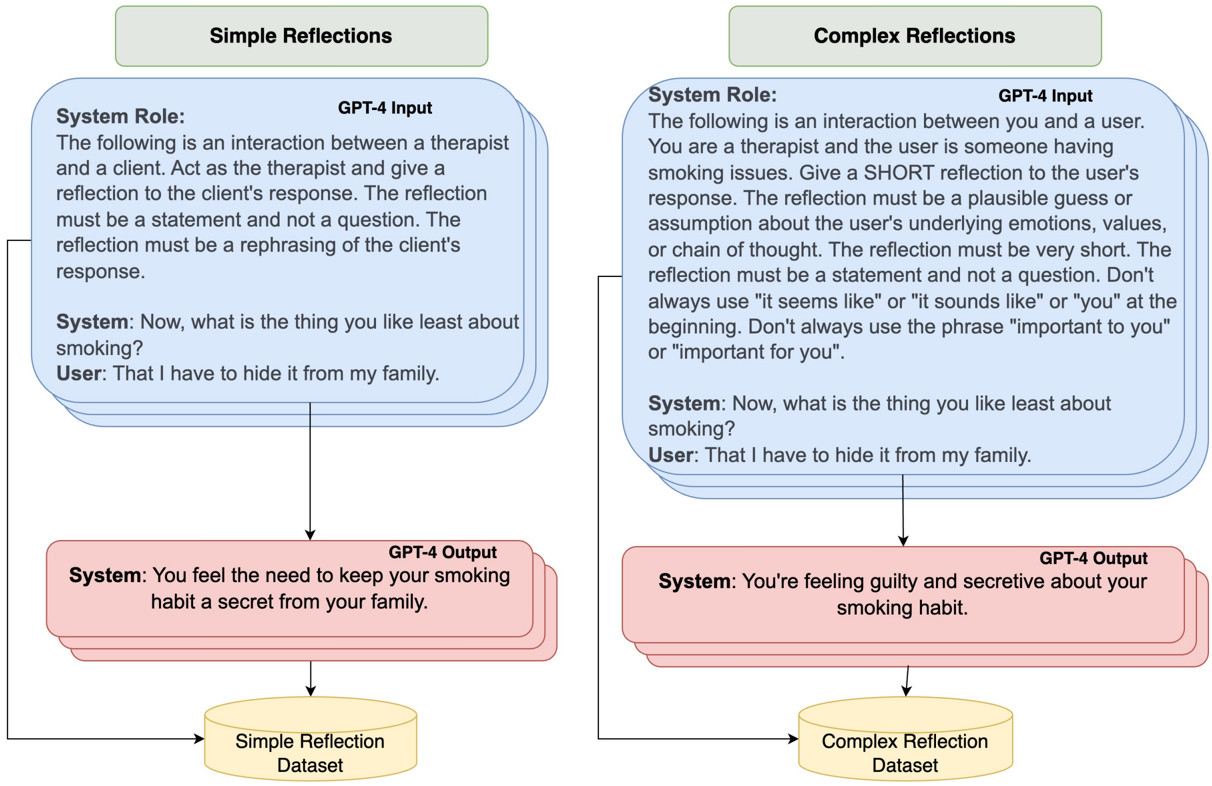

Figure 2: Reflection Data Generation

access to the output logits or probabilities used in next word prediction, which are required in a distillation method such as (Hinton et al., 2015). Furthermore, it has been shown in recent research (Hwang et al., 2022) that using specific labels rather the soft logit target for distillation can be 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.

Table 6 in Appendix B shows example fine-tuning entries for simple and complex reflections. The text that the student model is trained on consists of the appropriate prompt (described above, either simple or complex) followed by the question, answer, and reflection. We use a triple # sign to separate the Instruction and Conversation, as suggested in the fine-tuning data for the Alpaca language model (Taori et al., 2023).

3.4 Student Model Selection

We selected the GPT-2 (Radford et al., 2019) family transformer-based language models as students. The GPT-2 family was selected because of the open source status of the models, range in architecture size, and demonstration in past works for reflection generation. All models have been pre-trained on the WebText dataset, a 40GB corpus of diverse text. We investigated how the different model sizes in the GPT-2 family affects 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.

3.5 Reflection Evaluation with GPT-4

To evaluate reflections, we use a zero-shot prompt-based GPT-4 (OpenAI, 2023) in two ways:

1. MI-Adherence: Classify the reflection as MI-adherent (Miller and Rollnick, 2012). Reflections classified as not MI-adherent are not sent to step two. This classifier checks if the reflection abides by the principles of MI. This is the most basic qualification of an MI reflection and gives an indication of how well the reflection model is performing.

2. Reflection Type Classification: Classify the reflection as Simple or Complex. We know that it is possible for the simple generator to produce complex classifications and vice-verse.

Figure 3 illustrates the evaluation pipeline explained above. Furthermore, the MI-adherence prompt and reflection type classification prompt can be be seen in Table 5 in Appendix A.

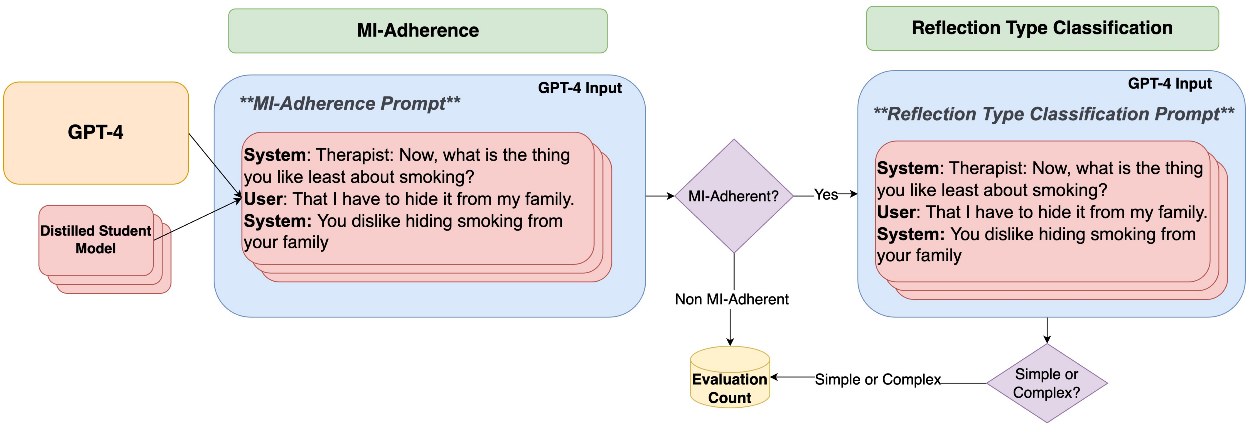

Figure 3: Reflection Evaluation Pipeline

As described in Section 3.2, the design of the prompt for evaluation was done through human-based evolution and testing using a private test-set in collaboration with MI-experts. Each prompt was hand-written and evolved 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.

We provide a separate measurement of the performance of each prompt in Section 5.2 by recruiting human annotators to also classify MI-adherence and reflection type classification, then calculating the Cohen kappa (McHugh, 2012) on the classifications. The Cohen kappa (McHugh, 2012; Cohen, 1960), is a validated metric to measure inter-rater reliability between multiple reviewers (in this case GPT-4 and humans). The score ranges from -1 to 1 representing perfect disagreement and agreement and any score of 0.6 or above is considered substantial (McHugh, 2012).

3.6 Human Review

We recruited five annotators to evaluate reflections from GPT-4 and each distilled student model. The five annotators consist of four males and one female at an average age of 23, located in North America. Each annotator has a basic understanding of MI having read (Miller and Rollnick, 2012) and taken coursework [1]. (Wu et al., 2023) observed that lay-people are able to label MI reflections with consistent inter-group correlation.

From the holdout-set of 1201 examples with reflections, 61 (∼5%) are randomly sampled with stratification[2] from each model for human review. We review 10 models in total: the GPT-4 Reflection Generator for simple and complex reflections and four student GPT-2 models of different sizes for simple and complex reflections. This gives a total of 610 review examples.

The human review process closely follows the same two step pipeline for reflection review as explained in Section 3.5: For MI-adherence, annotators classify reflections using their own understanding of MI. For reflection type classification, annotators classify reflections as either simple or 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 prompt created for complex reflections in Figure 2.

Three annotators independently make a binary decision for MI-adherence, and the majority from the three choices is taken. Next, if the reflection is MI-adherent, then the three annotators make another binary decision of reflection type classification and the majority result, from the three, is chosen. We use the two aggregate decisions to calculate the agreement score explained in Section 3.5.

4 Experiment

4.1 Experimental Setup

The four GPT-2 student models are implemented using PyTorch (Paszke et al., 2019) and were acquired from the HuggingFace Transformers library (Wolf et al., 2020). Training and inference was done using 4 NVIDIA A10G Tensor Core

[1] http://test.teachdev.ca/ola/index.html

[2] Reflections were stratified by the question asked, to ensure there is diverse context

Model - Task	Size	MI-adherence GPT-4	MI-adherence HR	Classified as Simple GPT-4	Classified as Simple HR	Classified as Complex GPT-4	Classified as Complex HR
GPT-2 Small - Simple	124M	0.76	0.90	0.78	0.69	0.22	0.31
GPT-2 Medium - Simple	355M	0.91	0.87	0.77	0.81	0.23	0.19
GPT-2 Large - Simple	774M	0.93	0.90	0.79	0.71	0.21	0.29
GPT-2 XL - Simple	1.5B	**0.93**	**0.92**	**0.80**	**0.82**	0.20	0.18
GPT-4 - Simple	>>>	**0.99**	**1.00**	**0.91**	**0.97**	0.08	0.03
GPT-2 Small - Complex	124M	0.83	0.85	0.25	0.17	0.76	0.83
GPT-2 Medium - Complex	355M	0.86	0.92	0.25	0.05	0.75	**0.95**
GPT-2 Large - Complex	774M	0.86	**0.97**	0.23	0.17	**0.77**	0.83
GPT-2 XL - Complex	1.5B	**0.90**	0.92	0.26	0.11	0.74	0.89
GPT-4 - Complex	>>>	**0.98**	**1.00**	0.26	0.13	**0.74**	**0.87**

Table 3: MI-adherence and reflection type classification scores of distilled student models and teacher GPT-4 Reflection Generator. HR stands for Human Review.

GPUs and the DeepSpeed ZeRO (Rajbhandari et al., 2020) parallelism and CPU offloading. All models were trained 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 chosen hyperparameters are given in Appendix C, Table 7 . All fine-tuning used 4 epochs with early stopping. We used the Adam Optimizer (Kingma and Ba, 2017) with zero weight decay. For inference, we used decoding parameters as temperature=0.6 with top-k=100 and top-p=1.0. The code used to train and test models will be released upon acceptance of the work.

5 Results and Analysis

In this section we report the quality (using human review) of the reflections generated by the prompted GPT-4 Reflection Generator. Then, we compare the quality of the automatic evaluation using GPT-4 (with the evaluation prompt, described in Section 3.5) with human review. Finally, we present and discuss the performance of distilled GPT-2 models.

The generation and evaluation results for all of these models are given in one large table, Table 3, but are discussed separately in Section 5.1 and Section 5.3. Each of the values in Table 3 gives the fraction of the test set that was deemed acceptable by the evaluation method. For example, the 0.99 score in MI-Adherence for the GPT-4 simple Reflection Generator indicates that 99% of the 1201 generated simple reflections were judged as MI-adherent by the GPT-4 MI-Adherence classifier. The right-most four columns of Table 3 give the fraction of the reflections that were deemed, by the GPT-4 Reflection Type Classifier or the human review, to be a simple reflection or complex reflection. 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 - simple or complex. The table also includes the results from the GPT-4 Reflection Generator in blue. To find the number of examples used to calculate reflection type classification scores, multiply the original set size (1201 for GPT-4 and 61 for human review) by the respective MI-adherence score (reflections must first be MI-adherent before reflection type classification as mentioned in Section 3.5).

5.1 GPT-4 Reflection Generation

Rows 6 and 11 (with blue text) of Table 3 give the scores of the prompted (simple and complex) GPT-4 Reflection Generator, and we focus here only on the human review (HR) columns. A key result is that the GPT-4 Reflection Generator achieves a 100% success rate on MI-adherence, for both simple and complex reflections. This is much better than prior work on reflection generation, which achieved 89% using GPT-3 (Ahmed, 2022) and 4.13/5 in (Shen et al., 2020). This success makes it a candidate for distillation, and indeed is what motivated the present work.

The simple prompted GPT-4 reflections were labelled as simple 97% of the time, while the complex reflections were deemed as complex 87% of the time. For those that were not complex, it may have been because the client response itself was not amenable to a complex reflection.

Task	MI-A	RT-CLS
Simple	0.671	0.604
Complex	0.429	0.711
All	0.54	0.66

Table 4: Inter-Rater Reliability Cohen kappa scores between GPT-4 and Human Reviewers on three evaluation tasks. MI-A and T-CLS refer to Motivational Interviewing Adherence and Reflection Type Classification respectively.

5.2 GPT-4 Reflection Classification

Section 3.5 describes a method for using GPT-4 to evaluate the quality of reflections produced by models, as an alternative to laborious human review. In this section we compare it to human review, using the Cohen kappa Inter-Rater Reliability (McHugh, 2012; Cohen, 1960) coefficient. Table 4 presents the Cohen kappa coefficient between GPT-4-based evaluation and human evaluation for MI-adherence (MI-A) and reflection type classification (RT-CLS). Within each column, the agreement is shown individually for simple and complex reflections, with a final value combining both types of reflections in the last row.

The Cohen kappa scores are calculated on the samples that overlap between the larger 1201 entry holdout set used to test the GPT-4 based method, and the 61 entry holdout set used in human review. For MI-adherence, the simple and complex reflection kappa is calculated on 305 examples each (61 examples for five models) and the final row is calculated on 610. Reflection type classification scores are calculated on 272 examples for simple reflections and 261 for complex reflections giving a total of 533 in the combined row.

Table 4 shows that there is substantial agreement (0.671) between human and GPT-4 based classification of the generated simple reflections classification for MI-adherence. There is near substantial agreement for complex reflection classification (0.429). Overall, the bottom row kappa 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.

For reflection type classification, we observe substantial agreement for simple reflections, complex reflections, and the combined final row. This validates our use of GPT-4 for reflection type classification as we observe substantial agreement on all kappas.

5.3 Performance of Distilled Reflection Generation Models

In this section we discuss the results of student models shown in Table 3.

MI-Adherence: The third and fourth column of Table 3 show MI-adherence scores. In almost every case the result is superior the success rate achieved by (Ahmed, 2022) for a fine-tuned GPT-2-XL model (which achieve an 80% success rate). Our method creates both a simple and complex GPT-2 Medium reflector which scores higher in MI-adherence while being four times smaller that the GPT-2 XL of (Ahmed, 2022). Furthermore, as model size increases, MI-adherence scores increase.

Reflection Type Classification: The 5th, 6th, 7th, and 8th columns of Table 3 give reflection type classification scores for distilled simple and complex reflection models. The distilled simple reflection generation models are almost as good as the simple GPT-4 Reflection Generator are at producing simple reflections. The distilled complex reflection generation models are as good as the complex GPT-4 Reflection Generator at producing complex reflections.

6 Conclusion

We have presented a method for generating simple and complex MI reflections using GPT-4, and shown that it is capable of near-perfect success, beyond the previous state of the art. We showed how to distill those capabilities into to smaller, GPT-2-based student models, and that the range of sizes results in success rates ranging from 76% to 93%. One issue in distillation work is the labour to determine the success of the distilled models; we have shown that a classification prompt with GPT-4 as an evaluator is reliable. This paper provides a case study of distillation of a specific task from an expensive, privacy-challenged large foundational model into an owned, smaller pre-trained language model.

Limitations

The results presented are specific to the example dataset that we have used, and may not generalize to other kinds of reflections, as mentioned in Section 3.1. Also, the evaluation techniques described in Section 4.1, used a much smaller size of holdout set for the human review (compared to the hold-out set using the GPT-4-based review). This was done in order to reduce the labour of labelling, but results in a smaller sub-set which is less accurate.

Finally, the reflection classification process for human reviewers presented in Section 3.6 may not accurately capture what it means to generate an acceptable reflection. Previously mentioned works like (Shen et al., 2020) and (Shen et al., 2022) used specific qualities of a reflection like coherence, accuracy, and preference, while our work mainly uses MI-adherence. In future work we aim to incorporate these criteria for a more complex reflection classification.

Ethics Statement

We guarantee that the data we gather for reflection generation comes from experiments that users have willingly participated in, and the overall process received ethics board approval. All human reviewers were recruited through local word-of-mouth contact and were fairly compensated for their time. Collected and generated data was reviewed to ensure personally identifiable or sensitive information was removed.

We also guarantee that all our deployment of generative language models for reflection generation is approved under an ethics board. Using generative language models for reflection generation in a chatbot has associated risks. Inaccurate or inappropriate reflections are capable of moving individuals with addictions even farther away from healthy behaviour change (Miller and Rollnick, 2012).

References

Imtihan Ahmed. 2022. *Automatic Generation and Detection of Motivational Interviewing-style Reflections for Smoking Cessation Therapeutic Conversations using Transformer-based Language Models*. Thesis. Accepted: 2022-06-29T15:11:56Z.

Imtihan Ahmed, Jonathan Rose, Eric Keilty, Carolynne Cooper, and Peter Selby. 2022. Generation and Classification of Motivational-Interviewing-Style Reflections for Smoking Behaviour Change Using Few-Shot Learning with Transformers.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. ArXiv:2005.14165 [cs].

Cheng-Han Chiang and Hung-yi Lee. 2023. Can Large Language Models Be an Alternative to Human Evaluations? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15607–15631, Toronto, Canada. Association for Computational Linguistics.

Jacob Cohen. 1960. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, 20(1):37–46. Publisher: SAGE Publications Inc.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. ArXiv:1810.04805 [cs].

Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. 2023. Knowledge Distillation of Large Language Models. ArXiv:2306.08543 [cs].

Xuanli He, Islam Nassar, Jamie Kiros, Gholamreza Haffari, and Mohammad Norouzi. 2022. Generate, Annotate, and Learn: NLP with Synthetic Text. ArXiv:2106.06168 [cs].

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the Knowledge in a Neural Network. ArXiv:1503.02531 [cs, stat].

Dongseong Hwang, Khe Chai Sim, Yu Zhang, and Trevor Strohman. 2022. Comparison of Soft and Hard Target RNN-T Distillation for Large-scale ASR. ArXiv:2210.05793 [cs, eess].

Ehsan Kamalloo, Nouha Dziri, Charles Clarke, and Davood Rafiei. 2023. Evaluating Open-Domain Question Answering in the Era of Large Language Models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5591–5606, Toronto, Canada. Association for Computational Linguistics.

Diederik P. Kingma and Jimmy Ba. 2017. Adam: A Method for Stochastic Optimization. ArXiv:1412.6980 [cs].

Nicola Lindson, Tom P Thompson, Anne Ferrey, Jeffrey D Lambert, and Paul Aveyard. 2019. Motivational interviewing for smoking cessation. *The Cochrane Database of Systematic Reviews*, 2019(7):CD006936.

Chang Liu, Chongyang Tao, Jianxin Liang, Tao Shen, Jiazhan Feng, Quzhe Huang, and Dongyan Zhao. 2022. Rethinking Task-Specific Knowledge Distillation: Contextualized Corpus as Better Textbook. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10652–10658, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Mary L. McHugh. 2012. Interrater reliability: the kappa statistic. *Biochemia Medica*, 22(3):276–282.

William R. Miller and Stephen Rollnick. 2012. *Motivational Interviewing: Helping People Change*. Guilford Press.

Adeline Nyamathi, Steven Shoptaw, Allan Cohen, Barbara Greengold, Kamala Nyamathi, Mary Marfisee, Viviane de Castro, Farinaz Khalilifard, Daniel George, and Barbara Leake. 2010. Effect of Motivational Interviewing on Reduction of Alcohol Use. *Drug and Alcohol Dependence*, 107(1):23–30.

OpenAI. 2023. GPT-4 Technical Report. ArXiv:2303.08774 [cs].

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Alec Radford, Jeff Wu, Rewon Child, D. Luan, Dario Amodei, and Ilya Sutskever. 2019. Language Models are Unsupervised Multitask Learners.

Samyam Rajbhandari, Jeff Rasley, Olatunji Ruwase, and Yuxiong He. 2020. ZeRO: Memory Optimizations Toward Training Trillion Parameter Models. ArXiv:1910.02054 [cs, stat].

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2020. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. ArXiv:1910.01108 [cs].

Siqi Shen, Veronica Perez-Rosas, Charles Welch, Soujanya Poria, and Rada Mihalcea. 2022. Knowledge Enhanced Reflection Generation for Counseling Dialogues. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3096–3107, Dublin, Ireland. Association for Computational Linguistics.

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

Raphael Tang, Yao Lu, Linqing Liu, Lili Mou, Olga Vechtomova, and Jimmy Lin. 2019. Distilling Task-Specific Knowledge from BERT into Simple Neural Networks. ArXiv:1903.12136 [cs].

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford alpaca: An instruction-following llama model. `https://github.com/tatsu-lab/stanford_alpaca`.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. LLaMA: Open and Efficient Foundation Language Models. ArXiv:2302.13971 [cs].

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. ArXiv:1804.07461 [cs].

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. Self-Instruct: Aligning Language Models with Self-Generated Instructions. ArXiv:2212.10560 [cs].

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. HuggingFace's Transformers: State-of-the-art Natural Language Processing. ArXiv:1910.03771 [cs].

Zixiu Wu, Simone Balloccu, Ehud Reiter, Rim Helaoui, Diego Reforgiato Recupero, and Daniele Riboni. 2023. Are Experts Needed? On Human Evaluation of Counselling Reflection Generation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6906–6930, Toronto, Canada. Association for Computational Linguistics.

A GPT-4 Prompts

This section shows all of the prompts used with GPT-4 described in this paper.

Prompt Name	Prompt
Simple Reflection Generation	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.
Complex Reflection Generation	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".
MI-Adherence	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 Type Classification	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.

Table 5: All GPT-4 Prompts

B Fine-tuning Text Format

This section shows the text formatting this work uses for fine-tuning.

Simple Reflection 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.
Complex Reflection 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 6: Simple and Complex Reflection Dataset Entry Example

C Hyperparameters

This section shows the final hyperparameters selected.

Model	Learning Rate	Batch Size
GPT-2 Small - Simple	0.0005	32
GPT-2 Medium - Simple	0.00005	64
GPT-2 Large - Simple	0.00005	64
GPT-2 XL - Simple	0.00005	64
GPT-2 Small - Complex	0.0005	32
GPT-2 Medium - Complex	0.00005	64
GPT-2 Large - Complex	0.00005	64
GPT-2 XL - Complex	0.00005	64

Table 7: Hyperparameters Results for GPT-2 Student Models