

# Conversational Bots for Psychotherapy: A Study of Generative Transformer Models Using Domain-specific Dialogues

Avisha Das<sup>1</sup>, Salih Selek<sup>2</sup>, Alia R. Warner<sup>2</sup>,  
Xu Zuo<sup>1</sup>, Yan Hu<sup>1</sup>, Vipina Kuttichi Keloth<sup>1</sup>, Jianfu Li<sup>1</sup>, W. Jim Zheng<sup>1</sup>, and Hua Xu<sup>1</sup>

<sup>1</sup>School of Biomedical Informatics, University of Texas Health Science Center at Houston, Houston, TX

<sup>2</sup>Department of Psychiatry and Behavioral Sciences, UTHealth McGovern Medical School, Houston, TX  
{Avisha.Das, Salih.Selek, Alia.R.Warner, Xu.Zuo, Yan.Hu, Vipina.KuttichiKeloth, Jianfu.Li, Wenjin.J.Zheng, Hua.Xu}@uth.tmc.edu

## Abstract

Conversational bots have become non-traditional methods for therapy among individuals suffering from psychological illnesses. Leveraging deep neural generative language models, we propose a deep trainable neural conversational model for therapy-oriented response generation. We leverage transfer learning methods during training on therapy and counseling based data from Reddit and AlexanderStreet. This was done to adapt existing generative models – GPT2 and DIALOGPT – to the task of automated dialog generation. Through quantitative evaluation of the linguistic quality, we observe that the dialog generation model - DIALOGPT (345M) with transfer learning on video data attains scores similar to a human response baseline. However, human evaluation of responses by conversational bots show mostly signs of generic advice or information sharing instead of therapeutic interaction.

## 1 Introduction

Psychological and mental disorders, such as depression and anxiety, are a growing concern worldwide. About an estimated 5% of the global adult population suffer from depression.<sup>1</sup> The National Alliance on Mental Illness (NAMI) reports 1 in 5 adults in the U.S. were diagnosed with mental health issues.<sup>2</sup>

Psychological ailments are complicated and challenging to diagnose and can manifest in an individual, in any form, regardless of their age, race, and gender. In extreme situations, lack of diagnosis and proper treatment can also be fatal.<sup>2</sup> However, only a fraction of the suffering individuals seek proper treatment from a mental health profes-

sional due to the existing stigma surrounding mental health. Additionally, the growing dearth in the current clinical workforce also adds to the problem. This impending crisis has led to a growing interest in automated conversational bots as non-traditional methods of receiving treatment for mental health (Ali et al., 2020; Vaidyam et al., 2019). A majority of the available conversational agents generate responses based on predefined rules or tree-based dialog flows, and may not be useful for therapeutic counseling (Mousavi et al., 2021) due to shallow and ineffective conversations (Abd-Alrazaq et al., 2021). Recent developments in massive language modeling through deep learning has resulted in successful outcomes in natural language understanding and generation tasks. Transformer architectures like OpenAI’s GPT-2 (Radford et al., 2019) have been used in conversational modeling and dialog generation with great empirical success (Zhang et al., 2020; Wolf et al., 2019).

While deep neural learning has helped improve the cognitive capability of these conversational agents, training such chatbots for a particular task require massive amounts of in-domain conversational data. Currently available massive pre-trained models have been trained on a motley of web-scraped articles and conversations on social media platforms (Devlin et al., 2018; Radford et al., 2019) that may contain toxic and aggressive content (Anderson, 2015). Therefore, dialog models pre-trained on such text can often generate responses that are harmful and callous, making them unsuitable for conversational psychotherapy (Pérez-Rosas et al., 2018; Harrigian et al., 2021).

Existing research on neural response generation has generated multiple data sets for evaluating dialog responses, however data related to mental health counseling is very limited (Harrigian et al., 2021; Pérez-Rosas et al., 2018). Additionally, a majority of these data sets have been collected through crowd-sourced human-human conversa-

<sup>1</sup>Institute of Health Metrics and Evaluation - Global Health Data Exchange (GHDx), <http://ghdx.healthdata.org/gbd-results-tool?params=gbd-api-2019-permalink/d780dffbe8a381b25e1416884959e88b>

<sup>2</sup><https://www.nami.org/mhstats>

tions (Rashkin et al., 2019), video transcripts of motivational interviewing (Pérez-Rosas et al., 2018), through text messaging (Gupta et al., 2020), etc. Mental health and psychological counseling data is sensitive with limited access and availability – restricting the improvement of dialog agents in the domain of psychotherapy counseling.

This study leverages existing generative architectures like the DialoGPT (Zhang et al., 2020), an open-domain dialog model based on OpenAI’s GPT-2 (Radford et al., 2019), for therapeutic conversational modeling for mental and emotional support. We additionally explore model fine-tuning through transfer learning on therapy counseling videos for therapy-based response generation (Wolf et al., 2019). For pre-training and fine-tuning, we use Subreddit threads that contain submissions on therapy and counseling, mental disorders and ailments (De Choudhury and De, 2014; Sharma et al., 2020), and transcripts of English therapy and counseling videos from AlexanderStreet website. Selecting the top models through a metric-based quantitative evaluation (Sedoc et al., 2019), we perform a task-based effectiveness study through human evaluation setup.

## 2 Related Work

While conversational agents have been used for multiple reasons, one major area of application is the diagnosis and treatment of psychological illnesses. From the first simple conversational agent ELIZA developed in 1966 by Joseph Weizenbaum to act as a Rogerian psychotherapist, current chatbots have undergone major improvements with the advancements in the field of artificial intelligence (AI) like natural language processing (NLP) and machine learning (ML) (Sharma et al., 2017).

### 2.1 Therapy-based Conversational Systems

Chatbots can generate human-like social and emotional responses, however the effectiveness of such automated agents have not been thoroughly investigated. Previous researchers have examined the key considerations and usefulness for incorporating conversation AI in psychotherapy (Miner et al., 2019; De Gennaro et al., 2020; Pham et al., 2022). Pacheco-Lorenzo et al. (2021) study review how smart conversational agents have been used to detect neuropsychiatric disorders by researchers – of which, (Mallol-Ragolta et al., 2019; Tsai and Lin, 2018) applied deep neural learning models for psy-

chiatric oriented response generation. Vaidyam et al. (2019) also report studies showing the potential of conversational agents in psycho-education and self-adherence. Most of the systems like (Bickmore et al., 2010b,a; Tielman et al., 2017a,b) used three-dimensional setups or interfaces to interact with the users.

Zhang et al. (2020) proposes DIALOGPT, a large-scale, tunable conversational model trained and fine-tuned on Reddit conversational threads. The model was built using OpenAI’s GPT-2 architecture as the base model (Radford et al., 2018, 2019). To better tune a massive generative model for task-specific performance, (Wolf et al., 2019) used an architecture called TransferTransfo to fine-tune the transformer-based BERT model on conversational data for the ConvAI2 challenge. (Huang et al., 2020) present a graph-based automated coherence metric for evaluating open-domain didactic conversations. Sedoc et al. (2019) combines multiple popular metrics like lexical diversity, BLEU scores, mean cosine similarity between generated and ground-truth responses, system perplexity, etc. into an evaluation tool called ChatEval.

### 2.2 Psychotherapy Dialog Datasets

Pérez-Rosas et al. (2018) proposes a novel dataset that consists of high and low quality counseling conversations collected from publicly available sources. Along with collection procedure the authors also describe the annotation procedure involving counseling skills like reflective listening and questioning. Harrigan et al. (2021) analyze the state and impact of social media resources as data for mental health research. Such sources like Reddits have been used in systems proposed by (Sharma and De Choudhury, 2018) and (Sharma et al., 2020) for studying empathy in human-human conversation threads. Researchers in (Rashkin et al., 2019) and (Mousavi et al., 2021) have proposed corpora on empathetic and therapeutic dialogs collected through real-life human-human conversations. Rashkin et al. (2019) uses crowdsourcing for building the corpora, while conversations between therapists and human participants are used by Mousavi et al. (2021) in their study. The authors in (Campillos-Llanos et al., 2020) address the task of varied terminologies and ontologies in medical domain by designing a knowledge-based patient record model using frame- and rule-based approach and terminology-rich resources like struc-

tured thesauri with linguistic, terminological and ontological knowledge. Similar task of term-based adaptivity in the clinical domain was also studied by Nirenburg et al. (2008), who used a multi-agent network model as a solution.

### 3 Data Collection

To train and incorporate the attributes of therapy in conversation - we collect the data by scraping Subreddit threads on mental health and transcripts of videos on psychotherapy and counseling. We extract the conversations about mental health and therapy from online sources and platforms such as Reddit (Baumgartner et al., 2020) and Alexander Street Press.<sup>3</sup>

#### 3.1 Mental Health Subreddits

Reddit ([www.reddit.com](http://www.reddit.com)) is an online platform that hosts multiple sub-communities or *sub-reddits* where people share their comments and views through posts and comments. Prior research (Sharma et al., 2020; De Choudhury and De, 2014; Sharma and De Choudhury, 2018) has proposed the use of Reddit to facilitate conversations and support for mental health and wellness through different subreddits like depression, anxiety, therapy and counseling, etc. In this study, we collected a total of 68,835 posts and 809,646 comments from mental-health related subreddits. These publicly available subreddit threads have been curated and used in previous research for mental health and empathetic textual modeling. We use the Pushshift Reddit API to periodically scrape Reddit for the data.<sup>4</sup>

We followed the pre-processing steps outlined by Zhang et al. (2020) to prepare our scraped subreddit submission data. These include removal of submissions – (a) with a URL in source or target, (b) not containing at least one of the most frequent English words (like "the", "a", etc.), (c) empty or upvoting comments, (d) with less than five words or more than 200 words. To convert the thread-based structure (posts and comments) to a conversational dialog-like input, we model them as *tree-structured reply chains* (Zhang et al., 2020). Table 1 shows the statistics of the data collected. The subreddits scraped can be divided to the following broad categories based on (Sharma and De Choudhury, 2018): (a) *Coping and Therapy (C-Th)*: 7Cup-

sofTea, Existential\_crisis, getting\_over\_it, Grief-Support, helpmeco, hardshipmates, HereToHelp, itgetsbetter, LostALovedOne, offmychest, MMFB, Miscarriage, reasonstolive, SuicideBereavement, therapy; (b) *Mood Disorders (MD)*: depression, depressed, lonely, mentalhealth; (c) *Psychosis and Anxiety (P-An)*: anxiety, BipolarReddit, socialanxiety; and (d) *Trauma and Abuse (Tr-A)*: abuse, survivors, Anger, emotionalabuse, PTSDcombat.

**Training setup.** We used the subreddit data to fine-tune the GPT-2 (345M) model and it served as a baseline. This was done because we compared its performance with the DialoGPT model (Zhang et al., 2020), which has already been trained on a huge Reddit dump. We discuss the models in detail in Section 4.

#### 3.2 Psychotherapy Videos

Alexander Street Press<sup>5</sup> is a website with a large collection of video transcripts and video recordings of therapy and counseling sessions on topics like depression, abuse, trauma, mental disorders, etc. The video transcript dataset was collected from the *Counseling and Therapy* channel on the website.<sup>6</sup> Of the 2,253 videos on the channel, we extracted videos on counseling and therapy training sessions. Collecting only sessions recorded in English ranging from the years 1980 to 2018, the collected set consists of 1,284 videos and transcripts. We removed some short-length non-informative videos, the final set has 1,130 video transcripts with a total of 180,765 dialog turns. After cleaning the data to remove unicode characters, pauses, etc., the data consists in total 2,914,307 words with a vocabulary size of 30,438.

**Training setup.** We divide the set of video transcripts into different subsets, with 80% for training (904 videos), 10% for development (112 videos) and hyperparameter fine-tuning, and 10% for testing (114 videos).

#### 3.3 Independent Test Data

To further evaluate, the performance of the dialog generation models, we built an independent conversational dataset by collecting responses from a different source. This is an *out-of-domain* dataset of synthetic human-human conversations. The *Empathic Conversation* dataset was created by collecting 25 conversations written by a group of research

<sup>3</sup><https://video.alexanderstreet.com/>

<sup>4</sup><https://github.com/pushshift/api>

<sup>5</sup><https://alexanderstreet.com/>

<sup>6</sup><https://video.alexanderstreet.com/channel/counseling-and-therapy-in-video>

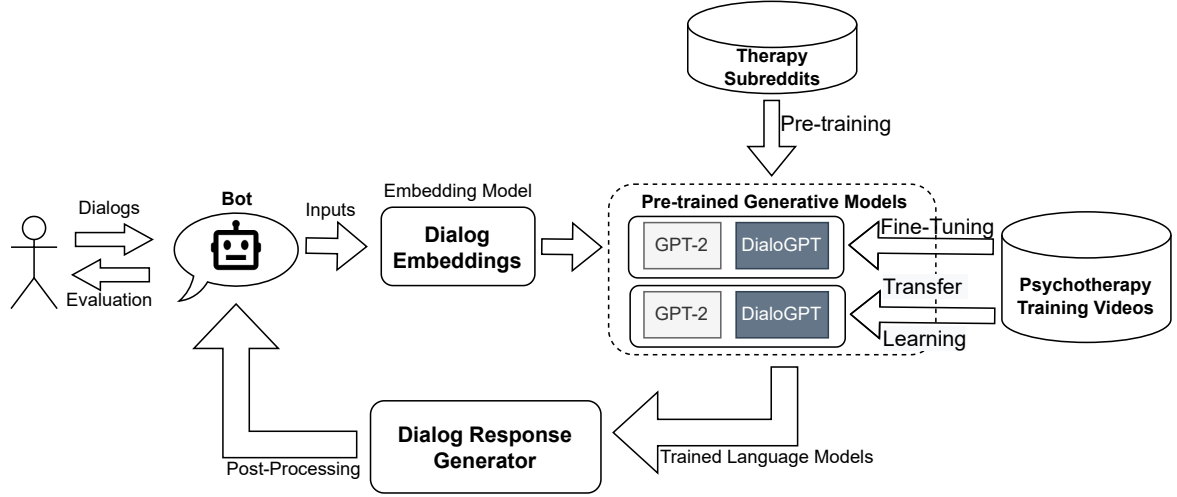


Figure 1: The schematic of the model training and dialog response generation

Category	Collected Posts	Filtered Posts	Comments	Avg. CL	Vocab Size
C-Th	49,426	38,750	625,149	16.13	102,458
MD	6,000	3,519	37,350	10.61	28,833
P-An	3,048	1,925	31,112	16.17	28,116
Tr-A	10,361	6,758	116,035	17.17	56,542

Table 1: Statistics for Scraped Subreddit Data. Avg. CL: Average Conversation Length or Number of Turns in Conversation, Vocab Size: Vocabulary Size

scientists, post-doctoral scholars and doctoral students for a given set of empathic prompts corresponding to stress inducing situations like health, work, trauma or abuse (Zhang et al., 2019; Du et al., 2018). The group consisted of 5 participants, each of whom wrote 5 conversations. Each conversation consisted of a total of 5 or more utterances.<sup>7</sup> The average conversation length was 7.2 utterances with each utterance has an average of 12.3 words.

## 4 Model

### 4.1 Model Architecture

We use the Generative Pretrained Transformer (GPT-2) (Radford et al., 2019) architecture as our baseline model. The massive pre-trained generative language models like GPT-2 can generate realistic looking text from a given prompt (Radford et al., 2019) – but the text may be noisy or unrelated to the nature of the task (Wolf et al., 2019). This becomes more challenging in the case of automatic coherent response generation during didactic conversations. The other baseline, Zhang et al. (2020)’s DialoGPT was trained (fine-tuning and training from scratch)

on subreddit threads to capture task-oriented dialogue generation.

Additionally, to adapt the model for specialized textual content, like *therapeutic counseling*, we need to build a generative system that is more goal-oriented, topical and coherent. While fine-tuning has been widely used for domain adaptation - the technique *may not be adequate* for our task for two main reasons – (a) the nature of the dataset we use for training/fine-tuning our base models is different, video transcripts vary in nature from community posts like Reddit (DialoGPT) and general web text pages (GPT-2), and (b) dialog response generation tasks combines multiple linguistic aspects such as co-reference resolution, common-sense knowledge, and long-range dependency. Therefore, following the technique proposed by (Wolf et al., 2019), we add **transfer learning during fine-tuning** the pre-trained baseline models for our dialog generation task. The experimental setup and details are further explained in the following sections.

#### 4.1.1 GPT-2

OpenAI’s GPT-2 (Radford et al., 2019, 2018) is a large transformer-based network trained on web-scraped textual content (8 million pages of web-

<sup>7</sup>An uninterrupted sequence of words spoken in a speech is an *utterance*.



text). The generative pre-trained architecture is based on transformer decoder-only blocks with attention modeling (Vaswani et al., 2017) and has outperformed previous state-of-the-art approaches on natural language understanding based tasks. Of the three model configurations, we use the GPT-2 *medium* (345M, 24, 1024, 64)<sup>8</sup> The models use byte-pair encoding (BPE) scheme (Sennrich et al., 2015) to encode the input text allowing the architecture to handle a wider range of vocabulary. Since the GPT-2 model was trained on web text, we pre-train a baseline model on the collected subreddit data.

#### 4.1.2 DialoGPT

Zhang et al. (2020) proposes the DialoGPT model that adapts the GPT-2 (Radford et al., 2018) for dialog generation. The implementation of the DialoGPT architecture along with the pre-trained models have been provided by (Zhang et al., 2020). In our implementation, we use the DialoGPT model based on a PyTorch adaptation made available by the HuggingFace team.<sup>9</sup> The model was trained on 147M multi-turn dialogue from Reddit discussion thread, collected over a span of 2005 to 2017.

### 4.2 Model Fine-tuning

Retraining the pre-trained models is necessary to condition the model for dialog generation – fine-tuning the generative architecture produce stylistically and linguistically better content from a given prompt (Das and Verma, 2020). We use the Python implementation of the GPT-2 models made available by OpenAI.<sup>10</sup>

The traditional fine-tuning experiment on the video transcripts and Reddit threads resulted in two sets of models. We fine-tune the GPT2 model, pre-trained on subreddit threads, on the psychotherapy training videos. Since DialoGPT was already trained on a huge dump of Reddit data, we only use the video transcripts for fine-tuning it. Therefore, we have two sets of models as a result of traditional fine-tuning. Here, the transformer model size varies from small, medium, and large, here we focus on the *medium* (345M) size of the generative models. The set of fine-tuned models to evaluate:

(a) GPT2 (345M)-FT-V, and (b) DialoGPT (345M)-FT-V.<sup>11</sup> The batch size and the learning rate were chosen based on the computation capability of the A100-SXM4 GPUs used to fine-tune the models.

### 4.3 Fine-tuning with Transfer Learning

We use the *TransferTransfo* architecture proposed by Wolf et al. (2019) for the Conversational Intelligence Challenge 2<sup>12</sup> (ConvAI2). *TransferTransfo* uses the multi-layer transformer encoder model GPT-2 (Radford et al., 2018, 2019) along with positional and segment embeddings extracted from a set of dialog conversations to incorporate speaker personality into conversations. The architecture uses transfer learning to adapt a content generation model like GPT-2 to a dialog generation task.

We use a similar setup for our transfer learning approach during the fine-tuning step. Using the pre-trained GPT-2 and DialoGPT (Radford et al., 2018; Zhang et al., 2020) as the base models, we fine-tune the language models on the collected therapy-specific conversation data from Reddit and AlexanderStreet. Unlike the original model, we remove the persona inputs, but keep the conversation history of a pre-specified fixed sequence length. Additionally, similar to Wolf et al. (2019)’s implementation, we combine the ‘gold’ human response (or correct response) as well as sampled ‘distractor’ response from the dataset. The distractor responses are actually randomly chosen responses from different conversation sequences in the dataset. The combined set of inputs (conversation history, gold response, and distractor response) are then used to create the set of input embeddings using the GPT-2 tokenizer model. The model schematic has been shown in Figure 2. To learn a global representation of the given context (nature of dialog conversations), we use a double headed model implementation called *OpenAIGPTDoubleHeadsModel*,<sup>13</sup> trained using a multi-task loss function that models both the generative and predictive loss functions.

To adapt the process of transfer learning for our purpose, we take a sequence of  $N = 5$  sentences during training as *conversation history*. The  $N + 1$ -th response is added as *gold response* and randomly sampled a *distractor response* from the utterances from other conversations. The OpenAI’s DoubleHead GPT2 Language Model was

<sup>8</sup>We report the model configurations in the following order: *Model-name (number of parameters, number of layers, embedding dimension, batch size/GPU)*

<sup>9</sup>[https://huggingface.co/docs/transformers/model\\_doc/dialogpt](https://huggingface.co/docs/transformers/model_doc/dialogpt)

<sup>10</sup><https://github.com/nshepperd/gpt-2>

<sup>11</sup>FT - Fine-tuning based, V - Video

<sup>12</sup><http://convai.io/>

<sup>13</sup>[https://huggingface.co/docs/transformers/model\\_doc/gpt2#transformers.GPT2DoubleHeadsModel](https://huggingface.co/docs/transformers/model_doc/gpt2#transformers.GPT2DoubleHeadsModel)

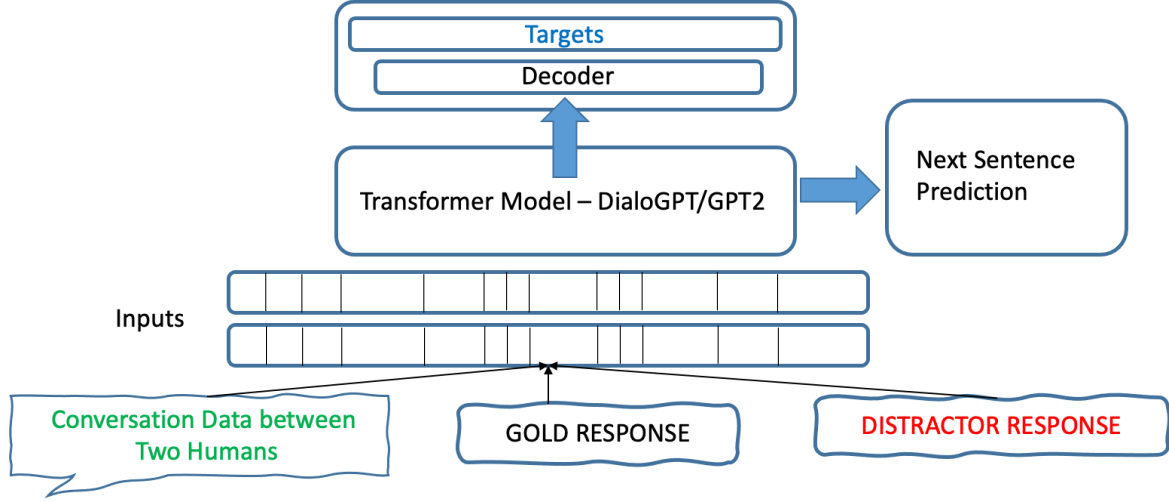


Figure 2: Transfer Learning Architecture

used for tokenization and embedding the inputs. Using the *medium* (345M) configuration for our evaluation models, we evaluate the following models for performance – (a) GPT2 (345M)-TL-V and (b) DialoGPT (345M)-TL-V.<sup>14</sup> Similar to the fine-tuning, DialoGPT is fine-tuned using only video data, while GPT-2 baseline model was used after pre-training on the collected subreddit data.

#### 4.4 Generation Methods

The most popular and widely used decoding techniques for text generation include top- $k$  sampling, greedy decoding and beam search (Das and Verma, 2020). Of these, *top- $k$  sampling* (Holtzman et al., 2019) has shown the best results for coherent long-form text generation. We select the value of  $k$  as 10 in our experiments. The softmax temperature  $t$  was chosen as 1.0, which is the default value.

Greedy decoding is also not a good decoding algorithm often generating incoherent textual content. *Beam search (BS)* method of sampling improves upon greedy sampling by selecting  $k$  most probable responses at each step of decoding the output and finally repeats the step iteratively until the most probable sequence is selected. Here,  $k$  is the beam length. DialoGPT models with 345M parameters and beam search of length 10, showed the best results (Zhang et al., 2020) in comparison to top- $k$  sampling method.

#### 4.5 Experiments and Evaluation

##### 4.5.1 Experimental Setup

We evaluate the performance of the **GPT-2** and **DialoGPT** models (Zhang et al., 2020; Radford et al., 2018, 2019) on the dialog response generation task with the goal of therapeutic counseling. Based on computational ease and prior performance (Das and Verma, 2020; Zhang et al., 2020), we select the generative model with the **medium size with 345M parameters** to demonstrate the findings for the task. Finally, we also test two decoding/sampling techniques – top- $k$  and *beam search*, we set  $k$  and the beam length to 10 in both cases. This was chosen empirically through a hyperparameter tuning setup.

We fine-tune our models on a single A100-SXM4 GPU and select the training epoch size as 10<sup>15</sup> for both sets of models. The batch size and initial learning rate for the fine tuning experiments are chosen as 16 and  $2e^{-5}$ . As used by Zhang et al. (2020) for their implementation, we use the Noam learning rate (selected based on validation loss) scheduler with 2000 warm-up steps.<sup>16</sup> For an *accelerated training* of the DialoGPT models, (Zhang et al., 2020) first compress the training data to a lazy-loading file type for faster loading during fine-tuning. We convert our datasets (video and reddit conversations) to an HDF5 file format to reduce computation load and accommodate GPU memory limitations. For the transfer learning setup, we use

<sup>15</sup>chosen through hyperparameter tuning on held-out development dataset

<sup>16</sup>[https://docs.allennlp.org/main/api/training/learning\\_rate\\_schedulers/noam/](https://docs.allennlp.org/main/api/training/learning_rate_schedulers/noam/)

<sup>14</sup>TL - Transfer learning based, V - Video

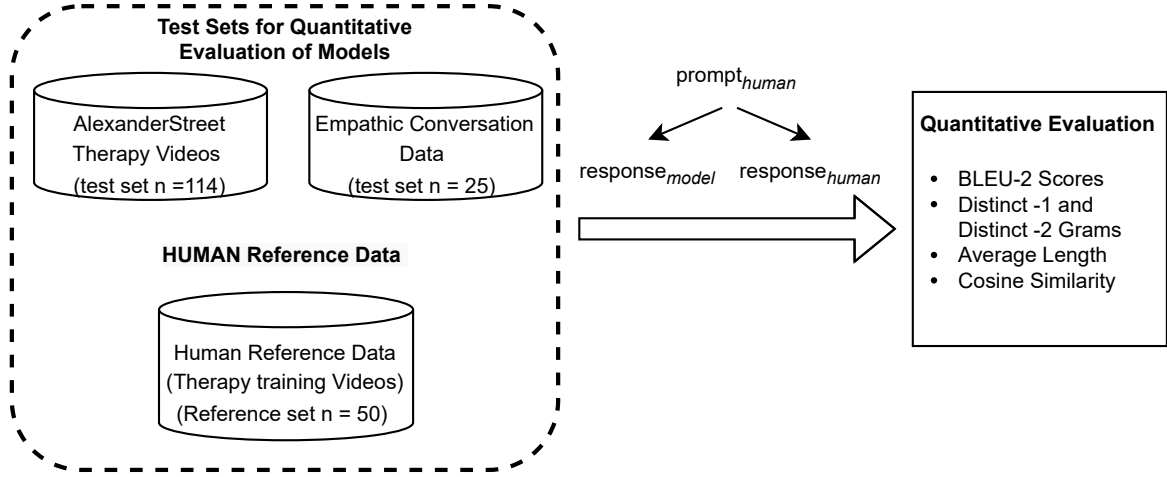


Figure 3: Quantitative Evaluation Setup

the  $6.25e^{-5}$  as the learning rate with Adam optimizer and PiecewiseLinear as the rate scheduler. We use the same number of epochs for fine-tuning the models with transfer learning. For the GPT2 model, the fine-tuning setup without the transfer learning took about 27 hours on the GPU, and with transfer learning the fine-tuning took approximately 38 hours. The traditional fine-tuning of the DialoGPT model took an average of roughly 34 hours and with transfer learning it took about 49 hours.

#### 4.5.2 Evaluation Metrics

The ChatEval toolkit (Sedoc et al., 2019) is a collection of automatic quantitative metrics used for evaluating standard machine translation task performance. Using the metrics toolkit, we report the following: (a) *Lexical diversity (Distinct-n)*: The number of unique  $n$ -grams in the models' response normalized by the token length. We consider the value of  $n = 1, 2$ , (b) *Average cosine-similarity* between the word embedding vectors of a generated response with the ground-truth human written response. We use a Word2Vec model trained on the conversation data to create the embedding vectors, (c) *Sentence average BLEU-2 score*. Additionally, we also report the *average length* of the generated responses.

## 5 Results

We discuss in detail the qualitative and quantitative evaluation metrics and performance of the different generative language models for therapy-based dialog response generation. We show our evaluation setup in Figure 3. We first perform a **quantitative**

**tative metric-based evaluation** on an in-domain test dataset of therapy videos. The test dataset for model evaluation consists of 114 videos collected from AlexanderStreet Therapy and Counseling training videos. The dataset has a total of 18,237 dialog utterances and 312,276 words. To compare the generated response with the test reference set, we select conversation length of at least 5 utterances as the maximum sequence length for the next utterance generation. Given a sequence of  $N$  responses, the generated  $N + 1^{th}$  response is compared to the corresponding human response in the original conversation. Here, we evaluate the model performance for a single utterance given a prompt. For calculating the **HUMAN** system metrics, we created a held-out subset of 50 conversations to calculate the metrics. These 50 conversations were selected from the AlexanderStreet video test dataset. We calculate the metrics by using the  $N$ -th response from these conversations and comparing them with the human ground truth response.

For the **qualitative and task-based effectiveness analysis**, we selected responses generated by the top three models from the quantitative evaluation, and showed them to our psychotherapy domain experts to judge. We further calculate the inter-annotator agreement between two judges to measure the effectiveness of the systems and their feasibility in automated therapy-based counseling.

We present the scores of the automated quantitative metrics provided by the ChatEval tool (Sedoc et al., 2019) on the video test dataset in Table 2. The different metric scores are denoted by "Vid." in the table. We see that the DialoGPT model usu-

ally scores closer to the human response baseline scores. With transfer learning on the video data, the scores like BLEU-2 and Distinct-uni and bigrams in the generated responses are relatively higher – even than the human baselines. Beam search also shows to generate responses with improved BLEU and Distinct- $n$  scores. As explained in (Zhang et al., 2020), we also conclude that typically ‘higher’ scores from the generative models do not mean that these systems generate responses largely better in quality than human speakers. The quantitative evaluation is a measure of *semantic distance* between a set of preceding responses and the generated content. So, a higher automatic BLEU-2 score shows a lower semantic distance and thus provides an understanding of the model’s performance with respect to human baselines. Based on the models’ performance across the Video test dataset, we see that DialoGPT is the best model with techniques like transfer learning and beam search decoding helping in achieving better quantitative scores overall.

## 5.1 Additional Evaluation

We also compare the quantitative evaluation metrics on the independent dataset described in Section 3. The results on the independent *synthetic Empathic conversation data* have been summarized in Table 2 under "Emp". Although the quantitative scores are typically lower in this case, due to the shorter nature of the conversations, we see a similar trend in the model performance – DialoGPT fine-tuned with transfer learning and beam search decoding generated responses that score quantitatively higher compared to other models.

Based on the results shown in Table 2, we select the top three dialog-based generative systems for our human evaluation setup. The three models selected were – DialoGPT-BS (beam search), DialoGPT-top $k$ , and GPT-2-BS – all the models trained using the transfer learning (TL setup).

### 5.1.1 Human Evaluation

We evaluate the performance of dialog generation models based on qualitative scores to measure the linguistic and task-based effectiveness of the generated samples. We sample a set of 10 source statements from Therapy and counseling videos on AlexanderStreet. Each of these prompts belong to a psychotic stressor-inducing situation like addiction, alcohol consumption, cyberbullying, etc (Zhang et al., 2019; Du et al., 2018). To anonymize

the system names and to remove any existing bias, we refer to them as Systems A, B, and C respectively. Given a starting conversation prompt, each system generates the statement in response to the prompt – generating a total of 30 samples to be evaluated.

We present the samples to two judges – a psychiatrist and a psychologist, experts in psychotherapy and acquainted with the required resources for the evaluation. For every stressor situation, we present three generated conversations to each judge to help them compare the context across all three systems. The average length (i.e., the number of utterances) of each generated conversation was 5 – we choose to present longer conversations to ensure the judges have more context and content and look at the conversation more globally before making a decision – instead of evaluating the system at a single utterance generation level (Zhang et al., 2020).

The qualitative metrics used to measure the therapeutic effectiveness were taken from a *standard set of assessment questions used to score psychotherapy resident performance* – (a) *Communication*: Did the bot ask any relevant questions to understand the situation – reason for session, suggestions, response of patient, plans for future, etc.? (b) *Basic Psychotherapy Skills*: Which conversation showed the signs of basic psychotherapy skills like active listening, open-ended inquiries, restatement/reflection/summarization, empathy? (c) *Overall Psychotherapy Competence*: Which conversation would you say is better on overall psychotherapeutic competence? Each sample is rated using Likert scale-based system from 1 to 5 for each metric, with 1 denoting the conversation as ‘*Not effective at all*’ and 5 being ‘*Extremely effective*’.

We calculated the Inter-Annotator Agreement scores using both Cohen’s  $\kappa$  and Krippendorff’s  $\alpha$ , with 0.286 and 0.34 respectively – this demotes fair agreement between the judges. Although we observe a strong preference for the DialoGPT generated responses – the judges comment on the unhealthy and non-therapeutic advice from the chat agents, typically discouraged in psychotherapy practice, indicating the need for further improvements in automated psychotherapy. Such an example has been shown in Table 5 in the Appendix Section. The skewness in the scores could also mean the confusing nature of the conversations causing the judges to not fully comprehend the actual purpose of the bot. Along with some sample



Method	Type	BLEU-2 (%)		Dist-1 (%)		Dist-2 (%)		Avg. Len.		Cos. Sim.	
		Vid.	Emp.	Vid.	Emp.	Vid.	Emp.	Vid.	Emp.	Vid.	Emp.
GPT-2 (345M)-top-k	-	7.14	8.34	8.3	9.1	11.4	14.3	7.4	8.3	0.412	0.586
	FT	8.02	9.68	7.8	8.2	17.2	12.7	6.2	7.6	0.507	0.632
	TL	11.17	10.03	9.1	8.7	16.8	15.2	8.1	10.2	0.618	0.643
GPT-2 (345M)-BS	-	6.85	7.05	9.5	8.3	14.2	15.6	7.4	7.8	0.603	0.587
	FT	10.73	8.96	10.4	10.9	14.7	18.4	9.5	9.1	0.702	0.732
	TL	10.18	11.76	13.7	9.6	17.6	15.7	8.8	9.3	0.674	0.613
DialoGPT (345M)-top-k	-	9.04	9.34	7.4	6.3	8.9	11.3	8.1	8.3	0.505	0.519
	FT	12.03	12.61	11.3	9.7	13.3	14.9	9.0	10.1	0.640	0.678
	TL	16.83	17.71	15.1	7.4	16.4	17.7	10.6	11.4	0.719	0.701
DialoGPT (345M)-BS	-	9.06	9.13	9.2	7.9	19.2	20.3	8.7	9.4	0.752	0.658
	FT	13.16	12.02	15.4	9.1	23.1	26.5	9.2	10.3	0.730	0.771
	TL	16.31	17.15	18.3	10.3	25.7	28.8	11.4	10.2	0.683	0.662
<b>HUMAN</b>	-	12.19	10.43	14.9	15.8	41.2	33.2	7.3	11.6	0.801	0.726

Table 2: Evaluation Results on the AlexanderStreet Video test dataset (*Vid.*) and Empathic Conversation dataset (*Emp.*). Here, *BS* - Beam Search, *TL* - Transfer learning based, *FT* - Fine tuning based, *Dist-1* - Distinct - 1, *Dist-2* - Distinct-2, *Avg. Len.* - Average Length and *Cos. Sim.* - Cosine Similarity.

conversations, we also include an example of the human evaluation template used for judging in the Appendix.

## 6 Conclusion

We propose a novel technique to adapt existing open-domain pre-trained generative models, DialoGPT (dialog-based) and GPT-2, for therapeutic conversation modeling. To fine-tune the model to the specific task of didactic conversations, we use a transfer learning technique to model aspects of therapist-patient counseling extracted from therapy video transcripts. On existing automated quantitative evaluation metrics, DialoGPT performed better than the GPT-2. Furthermore, a task-based effectiveness evaluation by human judges of the generated samples show that the models, pre-trained on a subsequently larger content of Subreddit data, generate responses with unhealthy advice or generic information usually discouraged in traditional therapeutic counseling. This indicates the need for further improvements in automated therapy counseling like domain-specific clinical and therapy data.

## 7 Future Work

We used a subset of the massive Reddit community to train our models – but there exists a large number of subreddit threads like *r/therapy*, *r/askatherapist*, *r/schizophrenia*, etc., that can capture contents and discussions on psychotherapy and psychosis. Inclusion of a variety of mental health related content

will enrich our conversational model incrementally. Our primary goal is to build an automated system that can provide therapeutic guidance and advice to individuals through conversation. The usefulness of such a bot-based therapist is the accessibility, availability, and convenience of a conversational assistant during emergencies. Therapist-patient conversational transcripts have not been used in previous AI-based bots and is a novel contribution of this study and using further sophisticated post-processing and abstraction techniques will help alleviate the problems the bot is suffering from.

## 8 Acknowledgements

This work is partly supported by the Cancer Prevention and Research Institute of Texas through grant RP170668 (Dr. Wenjin J. Zheng), the National Institutes of Health (NIH) through grants 1UL1TR003167 and R01AG066749 (Dr. Wenjin J. Zheng), and NIA R56AG069880 (Dr. Hua Xu).

## 9 Conflict of Interest Statement

Dr. Hua Xu and the University of Texas Health Science Center at Houston have research financial interest at Melax Technologies Inc.

## References

Alaa A Abd-Alrazaq, Mohannad Alajlani, Nashva Ali, Kerstin Denecke, Bridgette M Bewick, and Mowafa Househ. 2021. Perceptions and opinions of patients

- about mental health chatbots: Scoping review. *Journal of medical Internet research*, 23(1):e17828.
- Mohammad Rafayet Ali, Seyedeh Zahra Razavi, Raina Langevin, Abdullah Al Mamun, Benjamin Kane, Reza Rawassizadeh, Lenhart K Schubert, and Ehsan Hoque. 2020. A virtual conversational agent for teens with autism spectrum disorder: Experimental results and design lessons. In *Proceedings of the 20th ACM International Conference on Intelligent Virtual Agents*, pages 1–8.
- Katie Elson Anderson. 2015. Ask me anything: what is reddit? *Library Hi Tech News*.
- Jason Baumgartner, Savvas Zannettou, Brian Keegan, Megan Squire, and Jeremy Blackburn. 2020. The pushshift reddit dataset. In *Proceedings of the international AAAI conference on web and social media*, volume 14, pages 830–839.
- Timothy W Bickmore, Suzanne E Mitchell, Brian W Jack, Michael K Paasche-Orlow, Laura M Pfeifer, and Julie O’Donnell. 2010a. Response to a relational agent by hospital patients with depressive symptoms. *Interacting with computers*, 22(4):289–298.
- Timothy W Bickmore, Kathryn Puskar, Elizabeth A Schlenk, Laura M Pfeifer, and Susan M Sereika. 2010b. Maintaining reality: Relational agents for antipsychotic medication adherence. *Interacting with Computers*, 22(4):276–288.
- Leonardo Campillos-Llanos, Catherine Thomas, Eric Bilinski, Pierre Zweigenbaum, and Sophie Rosset. 2020. Designing a virtual patient dialogue system based on terminology-rich resources: Challenges and evaluation. *Natural Language Engineering*, 26(2):183–220.
- Avisha Das and Rakesh M. Verma. 2020. [Can machines tell stories? a comparative study of deep neural language models and metrics](#). *IEEE Access*, 8:181258–181292.
- Munmun De Choudhury and Sushovan De. 2014. Mental health discourse on reddit: Self-disclosure, social support, and anonymity. In *Eighth international AAAI conference on weblogs and social media*.
- Mauro De Gennaro, Eva G Krumhuber, and Gale Lucas. 2020. Effectiveness of an empathic chatbot in combating adverse effects of social exclusion on mood. *Frontiers in psychology*, 10:3061.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.
- Jingcheng Du, Yaoyun Zhang, Jianhong Luo, Yuxi Jia, Qiang Wei, Cui Tao, and Hua Xu. 2018. Extracting psychiatric stressors for suicide from social media using deep learning. *BMC medical informatics and decision making*, 18(2):77–87.
- Itika Gupta, Barbara Di Eugenio, Brian Ziebart, Aiswarya Baiju, Bing Liu, Ben Gerber, Lisa Sharp, Nadia Nabulsi, and Mary Smart. 2020. [Human-human health coaching via text messages: Corpus, annotation, and analysis](#). In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 246–256, 1st virtual meeting. Association for Computational Linguistics.
- Keith Harrigian, Carlos Aguirre, and Mark Dredze. 2021. [On the state of social media data for mental health research](#). In *Proceedings of the Seventh Workshop on Computational Linguistics and Clinical Psychology: Improving Access*, pages 15–24, Online. Association for Computational Linguistics.
- Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2019. The curious case of neural text degeneration. *arXiv preprint arXiv:1904.09751*.
- Lishan Huang, Zheng Ye, Jinghui Qin, Liang Lin, and Xiaodan Liang. 2020. [GRADE: Automatic graph-enhanced coherence metric for evaluating open-domain dialogue systems](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9230–9240, Online. Association for Computational Linguistics.
- Adria Mallol-Ragolta, Ziping Zhao, Lukas Stappen, Nicholas Cummins, and Björn Schuller. 2019. A hierarchical attention network-based approach for depression detection from transcribed clinical interviews.
- Adam S Miner, Nigam Shah, Kim D Bullock, Bruce A Arnow, Jeremy Bailenson, and Jeff Hancock. 2019. Key considerations for incorporating conversational ai in psychotherapy. *Frontiers in psychiatry*, page 746.
- Seyed Mahed Mousavi, Alessandra Cervone, Morena Danieli, and Giuseppe Riccardi. 2021. [Would you like to tell me more? generating a corpus of psychotherapy dialogues](#). In *Proceedings of the Second Workshop on Natural Language Processing for Medical Conversations*, pages 1–9, Online. Association for Computational Linguistics.
- Sergei Nirenburg, Marjorie McShane, Stephen Beale, and Bruce Jarrell. 2008. Adaptivity in a multi-agent clinical simulation system. In *Proceedings of AKRR’08, International and Interdisciplinary Conference on Adaptive Knowledge Representation and Reasoning*, pages 17–19.
- Moisés R Pacheco-Lorenzo, Sonia M Valladares-Rodríguez, Luis E Anido-Rifón, and Manuel J Fernández-Iglesias. 2021. Smart conversational agents for the detection of neuropsychiatric disorders: A systematic review. *Journal of Biomedical Informatics*, 113:103632.
- Verónica Pérez-Rosas, Xuetong Sun, Christy Li, Yuchen Wang, Kenneth Resnicow, and Rada Mihalcea. 2018. [Analyzing the quality of counseling conversations:](#)

- the tell-tale signs of high-quality counseling. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).
- Kay T Pham, Amir Nabizadeh, and Salih Selek. 2022. Artificial intelligence and chatbots in psychiatry. *Psychiatric Quarterly*, pages 1–5.
- Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training.
- Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.
- Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2019. [Towards empathetic open-domain conversation models: A new benchmark and dataset](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5370–5381, Florence, Italy. Association for Computational Linguistics.
- João Sedoc, Daphne Ippolito, Arun Kirubakaran, Jai Thirani, Lyle Ungar, and Chris Callison-Burch. 2019. [ChatEval: A tool for chatbot evaluation](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 60–65, Minneapolis, Minnesota. Association for Computational Linguistics.
- Rico Sennrich, Barry Haddow, and Alexandra Birch. 2015. Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909*.
- Ashish Sharma, Adam Miner, David Atkins, and Tim Althoff. 2020. [A computational approach to understanding empathy expressed in text-based mental health support](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5263–5276, Online. Association for Computational Linguistics.
- Eva Sharma and Munmun De Choudhury. 2018. Mental health support and its relationship to linguistic accommodation in online communities. In *Proceedings of the 2018 CHI conference on human factors in computing systems*, pages 1–13.
- Vibhor Sharma, Monika Goyal, and Drishti Malik. 2017. An intelligent behaviour shown by chatbot system. *International Journal of New Technology and Research*, 3(4):263312.
- Myrthe L Tielman, Mark A Neerincx, Rafael Bidarra, Ben Kybartas, and Willem-Paul Brinkman. 2017a. A therapy system for post-traumatic stress disorder using a virtual agent and virtual storytelling to reconstruct traumatic memories. *Journal of medical systems*, 41(8):1–10.
- Myrthe L Tielman, Mark A Neerincx, Marieke Van Meggelen, Ingmar Franken, and Willem-Paul Brinkman. 2017b. How should a virtual agent present psychoeducation? influence of verbal and textual presentation on adherence. *Technology and Health Care*, 25(6):1081–1096.
- Yin-Te Tsai and Wei-An Lin. 2018. Design of an intelligent cognition assistant for people with cognitive impairment. In *2018 IEEE 20th International Conference on High Performance Computing and Communications; IEEE 16th International Conference on Smart City; IEEE 4th International Conference on Data Science and Systems (HPCC/SmartCity/DSS)*, pages 1207–1212. IEEE.
- Aditya Nrusimha Vaidyam, Hannah Wisniewski, John David Halamka, Matcheri S Kashavan, and John Blake Torous. 2019. Chatbots and conversational agents in mental health: a review of the psychiatric landscape. *The Canadian Journal of Psychiatry*, 64(7):456–464.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.
- Thomas Wolf, Victor Sanh, Julien Chaumond, and Clement Delangue. 2019. [Transfertransfo: A transfer learning approach for neural network based conversational agents](#). *CoRR*, abs/1901.08149.
- Yaoyun Zhang, Olivia R Zhang, Rui Li, Aaron Flores, Salih Selek, Xiang Y Zhang, and Hua Xu. 2019. Psychiatric stressor recognition from clinical notes to reveal association with suicide. *Health informatics journal*, 25(4):1846–1862.
- Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2020. [DIALOGPT : Large-scale generative pre-training for conversational response generation](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 270–278, Online. Association for Computational Linguistics.

## A Appendix

Role	Response
Human	Hello! I have been feeling very down lately.
System	Was there anything you do for fun or anything you enjoy doing or interests hobbies.
Human	Yeah, like diving. We actually me and Lisa we went to Hawaii a couple weeks ago. I guess she wanted to pull me out of this funk. She saw it coming and wanted to do something that we both love.
System	Did you feel better.
Human	No. No. I just I don't know. This time it just didn't help. And that really scares me. Couldn't even I didn't even end up actually diving while I was there. Stayed on the boat.
System	Maybe you need to find something new to do to bring you joy. Maybe try some new activities or maybe take on a new hobby. Do you have any ideas?

Table 3: Conversation with a DialoGPT trained through Transfer Learning on Video transcripts to assess **depression**

Role	Response
Human	Hi. I wanted to talk about something .
System	Sure! What would you like to talk about?
Human	I experienced some recent losses... the loss of my mother about a year ago, just over a year ago.. then the loss of my partner in September. My father also moved to a home. So, really, uh, an enormous amount of loss to cope with, to have to deal with in, in a relatively short time.
System	That sounds very difficult. Can you tell me more?
Human	My mother passed away so quickly and unexpectedly... I cannot deal with it
System	I'm so sorry to hear that. Losing a loved one is never easy, no matter how long they've been gone. You may find it helpful to talk to someone about how you're feeling, or to write down your thoughts and feelings. There are also support groups available for people who have lost a loved one. Don't hesitate to reach out for help if you need it.

Table 4: Conversation with a DialoGPT trained through Transfer Learning on Video transcripts to counsel **loss in family**

#### Prompt #1: Addiction Counseling

Human: Hello. How are you?  
System A: Hi there! How can I help you today?  
Human: How are you today?  
System A: I'm feeling great today. How about you?  
Human: Hmm, I don't even know.  
System A: That's okay! Every day is a new opportunity for new experiences.

	Not effective at all	Slightly effective	Moderately effective	Very effective	Extremely effective
Communication	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Basic Psychotherapy Skills	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Overall Psychotherapy Competence	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Any additional comments

Human: Hello. How are you?  
System B: I'm doing well today. How are you?  
Human: Well I know that we met for the first time last week and, you know, we talked about a lot of stuff and -- I think most of -- you know, we were talking about, you know, my drug use.  
System B: Yes, I remember that we talked about your drug use. What would you like to talk about today?  
Human: I don't even know.  
System B: I'm sorry to hear that! I am here if you need me.

	Not effective at all	Slightly effective	Moderately effective	Very effective	Extremely effective
Communication	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Basic Psychotherapy Skills	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Overall Psychotherapy Competence	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Figure 4: Sample Questionnaire



Role	Response
Human	Hi! I am here seeking help with my past history with various substances and alcohol. I drink sometimes. Alcohol, beer, vodka. Personal favorite is rum.
System	Okay. About... How often when you say you drink?
Human	I mean, I used to drink uh... more often. Uh... Like, you know, three times a week, when I've tried to cut back.
System	How about marijuana? Have you ever used marijuana?
Human	Before. I mean I've... I've tried. It wasn't really my thing though. the last time I used marijuana was couple of years ago...
System	Okay. How about other drugs?
Human	No not much.
System	Okay. So marijuana and alcohol are the only two substances you've tried?
Human	Yes
System	All right. Well, from what you've said it sounds like you're trying to cut back on your drinking. That's really good. You're taking steps to improve your health and that's important. <b>As far as marijuana goes, it sounds like you've tried it before and it wasn't really your thing. That's okay. Not everyone likes marijuana.</b> <b>There are plenty of other things to try out there. If you ever want to try marijuana again, there are plenty of ways to do so safely.</b> But if you're not interested, that's perfectly fine too. Thanks for talking with me.

Table 5: Conversation with a DialoGPT trained through Transfer Learning on Video transcripts to counsel **Opioid Use Disorder**. This is a *negative* example.