A Cognitive Behavioral Therapy Chatbot Powered by GPT-3

Harper Carroll
Stanford University
hcarroll@stanford.edu

Filippos Nakas
Stanford University
fnakas@stanford.edu

Leyla Kursat
Stanford University
lkursat@stanford.edu

Abstract

In this paper, we discuss our Cognitive Behavioral Therapy (CBT) Chatbot, which uses a multi-modal GPT-3 system, a question-answer therapy dataset to mimic therapists' conversational style, and CBT literature to confine chatbot responses to the CBT therapies in its domain. Using information retrieval methods inspired by prior works, and cutting-edge GPT-3 technologies, we built a system unlike any existing model of which we are aware. We evaluated the model at each stage of its natural language generation using discriminators between real and generated therapist responses and comparing the similarities of final outputs with our CBT document embeddings. We found our results to be extremely coherent on average and promising for the future of therapy chatbots.

1 Introduction

With the COVID-19 pandemic triggering a 25% increase in anxiety and depression, there is an urgent need for reliable and effective digital mental health services (WHO, 2022). The low public expenditure on mental health in low-income countries combined with the growing stigma against seeking therapy make chatbots an ideal tool to support mental health. We are interested in ways with which NLP techniques can be used to offer on demand and personalized mental health therapy. Our goal is to replicate the performance of existing end-to-end therapeutic natural language generation systems by fine-tuning GPT-3 on therapy data, and create a Cognitive Behavioral Therapy (CBT) chatbot that outperform existing CBT chatbots in terms of topics covered, conversational fluency, and incorporation of CBT literature.

Our main strategy is leveraging GPT-3's unique ability to meaningfully integrate linguistic data from different sources into a meaningful whole. The modular design allows the system to separately determine the topic, style, and content of the intended response, and GPT-3 is used to produce a holistic synthesis. Therapy Question-Answer (QA) dialogue data is used in our system to determine the mental health topic that best describes the user question as well as to capture the elusive, "human" style of realistic therapist discourse. Once these linguistic characteristics are used to provide a generic "template" response to the user's question, we make use of fitting CBT literature and GPT-3's synthesizing capacity to produce a domain-specific response that still maintains the generic therapist response's style and structure. This allows us to pair the appropriate conversational style with the appropriate informational content.

We were indeed successful in generating coherent responses that are in the style of conversational therapist responses yet are adapted to use CBT skills. The results of our model are extremely promising for the future of fully-autonomous therapeutic chatbots for specialized therapy methods.

2 Prior Literature

There exists extensive computational work on designing therapy chatbots. We found that the common psychological therapy methods delivered by many chatbots is Cognitive Behavioral Therapy (CBT), a therapy that is highly skills-based and one that requires less interactivity between the patient and a therapist. CBT, the therapeutic method we chose to focus on in our paper, has proven to be effective for a range of psychological ailments, like depression, anxiety disorders, alcohol and drug use problems, eating disorders, relationship problems, and severe mental illness. Many studies support that CBT significantly improves functioning and quality of life, and studies have also shown that CBT is at least as effective as other forms of therapy or psychiatric medications (Bennett-Levy et al., 2010). CBT focuses mainly on targeting unhealthy or distorted thought patterns and behaviors, developing problem-solving skills to cope with difficult situations, and increasing one's own confidence in their abilities. As a therapy that mainly relies on providing alternative ways of thinking and skill building, CBT seems well suited to be conducted digitally, in written or spoken modality. In the following subsections, we discuss prior work on using therapeutic chatbots, often with CBT, in detail.

2.1 Existing CBT chatbots

A Stanford paper by Fitzpatrick et al. (2017) tested the ability of a commercially developed text-based chatbot, Woebot, to deliver CBT to college students through a controlled experiment and found that those using Woebot had significantly reduced symptoms of depression. This study showed, through empirical evidence, that text-based conversational agents have the potential to deliver CBT therapy effectively.

Patel et al. (2019) studied the effectiveness of a chatbot in combating depression using CBT techniques by proposing a therapeutic emotion classifier. The bot calculated a percent Positive and percent Negative emotion score based on emotion labels and then classified a user's mental state as in a range from normal to highly depressed. They used the ISEAR dataset for emotion detection, the Punkt sentence tokenizer, and GloVe word embeddings. They tested three different deep learning classifiers: a convolutional neural network (CNN), a recurrent neural network (RNN), and a hierarchical attention network (HAN), and found that the CNN performed best, achieving an accuracy of 75% (versus below 70%).

Similarly, Rizvi et al. (2011) studied the effectiveness of a basic app designed like a chatbot to coach a client through using a single CBT technique on-demand. The chatbot did not include any machine learning algorithm; it consisted simply of a decision tree of responses based on participants' selections of pre-set inputs. The results showed that the participants' emotional intensities were significantly reduced after each session, as were the urge to use substances. This study was the first to show that a mobile application for learning and using CBT skills was effective at significantly ameliorating the difficult symptoms of patients struggling with mental health issues.

The existing work on evaluating CBT chatbots provide significant empirical evidence that a well-designed therapist chatbot can positively affect the mental health outcomes of its users. This seems to be the case for both complex, multi-moduled systems such as WoeBot (Fitzpatrick et al., 2017) as well as for more simple chatbots with more constrained capabilities such as that of Rizvi et al. (2011). More importantly, the fact that these systems can be effective without recourse to current state-of-the-art NLP techniques informed our hypothesis that applying these new techniques in this paper could significantly improve upon existing outcomes.

2.2 Methods for leveraging large language-models in the absence of data

Research has shown that formulating effective GPT-3 prompts is a crucial component for the success of any GPT-3 model, including our own. Liu et al. (2021) showed that including even a small number of well-chosen examples in GPT-3's prompt can be an extremely effective way of achieving excellent performance on tasks that would normally require extensive fine-tuning using smaller transformer models. The team devised an effective method compared to a random baseline for automatically selecting examples for a successful few-shot prompt for large language models like GPT-3. Their approach, which we were inspired by for our model, consisted of two parts: an example retriever module and GPT-3. The authors experimented with both raw and fine-tuned versions of various transformers (BERT, RoBERTA, XLNet) to encode all candidate train examples and the test examples. The representation of the test example was then compared with the representation of the X-component of each train example based on some pre-defined distance metric (cosine, Euclidean) to find the k most similar train examples. These k examples, along with their actual target, were then combined into a mini-dataset that was appended before the test example to form the complete GPT-3 prompt, which was in turn used for few-shot (or one-shot, for k = 1) learning. In our model, we similarly retrieved the appropriate CBT skill by using fine-tuned GPT-3 embeddings and distance metrics. See "Model" for more details.

Izacard and Grave (2020) demonstrated the high value of document-based knowledge retrieval in enhancing the performance of large generative language models for open QA. The proposed model, named Fusion-in-Decoder, has a simple architecture which consists of a document retrieval module that feeds relevant supporting passages to a seq2seq encoder-decoder model that generates the answer.

Despite the method's simplicity, it surpassed the state-of-the-art for both NaturalQuestion and TriviaQA datasets. In addition, the authors observed that the performance scaled with the number of retrieved supporting passages. We used a similar retrieval method to draw domain-relevant information from CBT literature for CBT-informed suggestions from the chatbot therapist to the user.

3 Data

Based our methods on techniques from "Prior Literature" Section 2.2, we did not need a huge counseling dataset; we used the dataset published by CounselChat, a platform that allows people to ask questions to verified therapists. The 2020 dataset, which has over 2100 examples, includes questions about 31 different topics ranging from depression to professional ethics and responses from 307 licensed mental health counselors. We used this dataset, filtering out all questions with 0 upvotes, to fine-tune our first GPT-3 model, discussed in the "Model" section below.

We also had documents that contain information about CBT approaches and skills. For the retrieval task, we provided documents for 9 cognitive distortions (e.g. jumping to conclusions, all-or-nothing thinking, overgeneralizing) and 14 CBT treatments (e.g. cognitive restructuring, mindfulness therapy, assertiveness training). These documents were created with information taken from Cognitive Behavioral Therapy Los Angeles (https://cogbtherapy.com).

4 Model

The generalizing power of GPT-3, along with the provenly-effective knowledge-retrieval methodologies discussed in Section 2.2, offered remedies for the lack of a large database of specialized therapy transcripts. Having data that exactly instantiates the desired behaviour of the system would have been optimal, of course, but it was possible to roughly conceive of a therapist's dialogue capacity as consisting of two parts: 1. general linguistic competency and 2. specialized knowledge in psychological domains. In this light, although we have only light access to the whole (a relatively small QA dataset), we certainly have adequate access to its components. The general linguistic capacity and knowledge hidden in the weights of GPT-3, combined with prompts constructed out of passages from relevant CBT literature retrieved using the methods of the Section 2.2 papers, proved to form an adequate substitute for the lack of extensive data.

As documented in our Experimental Protocol, we created a dialogue system for CBT-focused mental health advice. Our model consists of separate modules, most of which are powered by GPT-3's diverse NLP capabilities. Given the complexity of the problem, our dialogue system focused on a one-turn dialogue similar to our QA dataset. The two main components of our system were generic response generation and response adaptation to domain-relevant information (see Figure 1).

4.1 Generic response generation

Our first GPT-3 model is fine-tuned on the CounselChat dataset, uses the user's question as the first prompt, and outputs a generic response resembling the therapists' responses in the dataset. This first outputted response does not discuss CBT methodologies, but it is intended to mimic the conceptual references and empathetic style of actual practitioners.

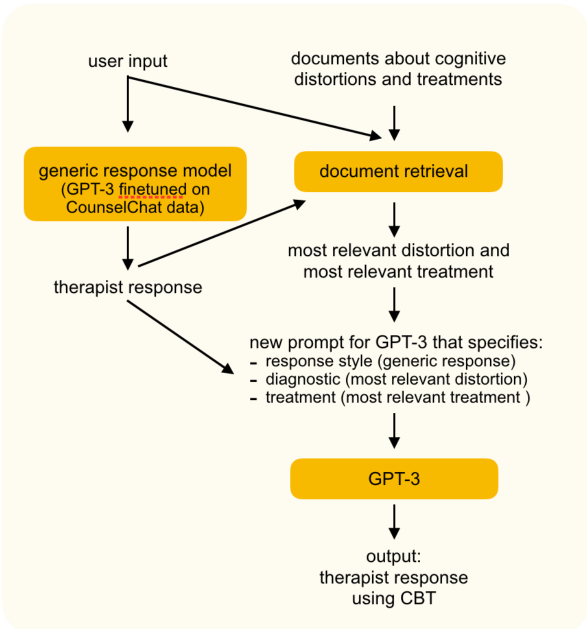

Figure 1: Model Architecture

4.2 Document retrieval

The second step in our model is retrieving relevant documents based on the user input and the response generated by the fine-tuned GPT-3 model. We concatenate the user input and the generated therapist response and compare their similarity to document embeddings of the distortions and treatments using dot product and GPT-3's (Babbage version) semantic search. The retrieval submodule ranks 5 documents that are the most relevant for the users case. The most relevant distortion and treatment are saved to be used in the final prompt, so that the final output includes information about the irrational thought pattern of the user (distortion) as well as the correct CBT skill to restructure it.

4.3 Response adaptation to domain-relevant information

Having obtained a generic response to the user's question of the appropriate style, the system uses a carefully crafted prompt asking GPT-3 to transform the response into one that incorporates insights and advice from the most relevant CBT skill category.

The prompt has three main parts. In the first part, the CBT text fragments are introduced using an appropriate introduction such as "Labeling is CBT cognitive distortion" and "Behavioral Activation is a CBT treatment". In the second-part, the user's question is paired with the generic response generated by the previous module and introduced into the prompt with an introduction such as "The following is a dialogue between a client and their therapist Dr Johnson: ". Finally, we reach the most crucial part of the prompt, which is to direct GPT-3 to perform the transformation. We ask GPT-3 to provide an alternative response by another therapist, Dr. Weathers, who specializes in CBT. We mention explicitly that Dr. Weathers responds similarly to Dr. Johnson and we include the cognitive distortion she would identify (e.g. labeling) as well as the treatment she would recommend (e.g. behavioral activation) (see Appendix for full text examples).

5 Methods

Automatic quantitative evaluation of end-to-end chatbots with a significant NLG component is a notoriously difficult task, given the open-endedness of the dialogue system's range of "correct" outputs (Deriu, 2021). Existing papers and models in the space, which we discuss in the "Prior Literature" section above, usually evaluate their chatbots through clinical psychological trials to test whether their chatbots had a positive impact on participants' mental health (Fitzpatrick et al. (2017), Rizvi et al. (2011)). Even in non-psychological use cases, chatbots are usually evaluated quantitatively only once there is ample user engagement data after deployment. Given the narrow timeline of this project and the sensitive nature of the use case, testing on users faced both practical and ethical limitations, so we could not pursue that route for evaluation.

Automatic pre-deployment quantitative evaluation is often performed to test the performance of the individual modules that make up the full dialogue system, where the problem can be described in the form of standard NLP tasks (classification, entity extraction, semantic parsing). Using this approach, we determined that our system would be evaluated by testing simultaneously that it can 1) produce human-like empathetic responses that mimic the style of actual therapists and 2) cohesively incorporate reliable information about CBT insights and practices into its responses.

For quantitative evaluation, we randomly chose 57 examples from the set of 240 questions (we unfortunately had no CBT credit to produce more) that fulfilled two requirements: 1) 0 upvotes so that there is no overlap with training (it doesn't matter for evaluation because we only care about whether they seem human generated which most probably applies equally to responses with upvotes and those without upvotes) 2. have only one response (to make evaluation simpler). To test for the degree to which the final response meaningfully integrated CBT information from the documents that formed part of the final prompt, we used an original adjusted version of the metrics from the document retrieval literature (Karpukhin et al., 2020), which will be explained in the Results section.

In addition to these quantitative metrics, we also evaluated our model qualitatively by manually interacting with our chatbot and evaluating the coherence and effectiveness of its responses. We include screenshots in the Appendix of example conversations with our chatbot, demonstrating a range of coherence.

6 Results

Our results are extremely promising. The generated CBT output by our model usually both properly incorporates the therapist-style of communicating and an appropriate CBT skill. Additionally, the answers are usually properly formatted, with a single response given by the doctor.

We encountered various problems when creating our model and needed to try out many different methods to produce the kind of results shown in the Appendix. The biggest problem was trying to create a prompt that maintained both the natural aspects of the general response and included information from the CBT documents. GPT-3 had the tendency to opt for one of the two, to either give the same generic response or completely ignore the generic response and use quotes from the CBT documents directly. To solve this, we experimented with many different ways of formatting the final prompt. We found that "storytelling", prompting with a scenario using fake doctor names (i.e. "Dr. Weathers specializes in CBT; she takes in Dr. Johnson's diagnosis and re-frames it to include CBT suggestions."), worked the best to blend the two goals. We also attempted to use few-shot learning, as demonstrated by Liu et al. (2021) discussed in the Prior Literature section. For each question from the dataset, we found the *k* most similar questions, took those answers and inputted them into the GPT-3 model as the prompt, and then performed few-shot learning without fine-tuning. Because of the length of our *k* most similar responses, we could not fit too many into the prompt (max *k*=3), which resulted in responses that overfitted on the (up to) 3 prompt examples. For example, if the 3 responses were short, then the generated response would be short. We tried the incorporation of few-shot learning both by itself and in combination with the fine-tuning on the CounselChat data documents, and we found it was better to just fine-tune on the data alone, removing any few-shot learning. We identified a specific benefit of fine-tuning alone: we needed to withhold some training questions to use for evaluation, and fine-tuning worked well even using just the fraction of the data that had one upvote, likely because fine-tuning results in a small portion of good examples that are implictly used all the time. With few-shot learning, however, we'd need far more data points as candidates for retrieval in order to achieve the same effectiveness, to pick *k* examples that are extremely useful for the model. Because of the length restriction of the prompt, we also encountered the problem of responses becoming longer than the token limits, so some responses cut off mid-sentence or abruptly. We tried to solve this by fine-tuning a different GPT-3 model on the CBT documents instead of using document retrieval, but this didn't work because the model lost its ability to converse; the bot fine-tuned on the documents was no longer able to give a single complete response and would rather create a full dialogue, including generating new text by the patient (e.g. "Dr. J: [TEXT] Patient: [TEXT]").

We tried two methods of fine-tuning offered by GPT-3. We used the same data points, but one method used a prompt and one did not (empty string). Normally when fine-tuning, pairs of prompts and completions are provided to the model; another option offered by GPT-3 was to leave the prompt empty and just provide the completion, which trains the model to replicate the style of completion in a general way, regardless of prompt. We found that neither of these methods of fine-tuning worked well because of one of two reasons: their generated text either deviated from the one-answer format, generating dialogue responses like those described above, or provided non-conversational descriptions of CBT literature without relating the information to the question.

Our final model's performance, however, significantly exceeded our expectations with respect to both quantitative and qualitative evaluation. Since the qualitative evaluation can be performed by the reader, we focus here on the two quantitative metrics we decided on using. Finding these two metrics was not easy, given the notorious difficulty of evaluating open-ended text generation based on document retrieval. Given the open-endedness of the task, all standard QA metrics become almost completely unusable given the lack of exhaustive golden labels. In our case this problem is even more acute given the nature of the task: what is a "correct" therapist response? Given these difficulties we used separate metrics to evaluate the "naturalness" and "knowledgability" of the system. This choice matches our intuition that a successful implementation of a CBT bot must successfully integrate these two attributes.

As explained in the experimental protocol, to test for "naturalness", we adapted the method of using an adversarial classifier to check the degree to which the model's generic and final responses can be confused with those of an actual therapist. This adversarial method of evaluation was first used for dialogue-systems by Kannan and Vinyals (2016) and its validity was largely confirmed by an extensive study by Bruni and Fernandez (2017). Instead of training a full classifier on the evaluation data, we used a special classification prompt for GPT-3. The prompt for each example looked like this:

"Question: " + question + " Answer 1: " + realAnswer + " Answer 2: " + botAnswer + " One of these answers was generated by a therapy AI bot and another by a human therapist. Which of the two was generated by a human therapist? Answer 1 or Answer 2?".

We used this prompt to generate classifications for all 57 evaluation questions twice, once for the intermediary generic answers and once for the final CBT-informed answers using 0 temperature. To our great surprise, GPT-3 got all predictions wrong in both rounds (i.e. classification accuracy of 0%), meaning that it deemed all generic and final response generations as "human". Given the "black-box" nature of GPT-3, it is not possible to draw too general or optimistic conclusions from this test, for it is not known what criteria are being used for the classification. Nevertheless, given the vast amount of text incorporated in GPT-3's weights it is reasonable to ascribe to it at least some "intuition" about the human criteria for judging whether a given piece of text is "robot-like", especially when it is compared side-by-side with an actual human response. Even more encouraging is the fact that this phenomenon was observed for the final responses that have incorporated CBT text. It is a promising indication that the "merging" of the general therapist style with the CBT knowledge is performed naturally (there are not simply concatenated together). This can also be qualitatively attested by observing the smooth transitions between general diagnosis and CBT references in the final responses. On a final note, we would like to emphasize the reasons for using "raw" GPT-3 instead of a specialized classifier (or fine-tuned GPT-3) trained on a part of the evaluation data. The first reason, is that we simply did not have the luxury of generating enough evaluation data for this task. Having run out of money, we only managed to produce responses for only 57 out of the possible 240 test examples (even all 240 would probably not be enough). However, an even more important reason is that the final responses contain, by design, a plethora of artefacts (all the CBT references) that could easily be used by even a shallow classifier as proxies.

Having established that the final responses are human-like it remained to also show in a quantitative way that they meaningfully incorporate the information contained in the CBT documents. To this end, we came up with - to our current knowledge - a possibly original adaptation of existing IR metrics to measure the degree of this for our particular task. Our intuition is that a response that has successfully incorporated the two retrieved CBT documents most related to its question - generic response pair should be even "more" similar to the these documents that the question - generic response pair. In contrast to the features that connect the pair with the document (which are most likely subtle semantic similarities), the features connecting the final response should be far more salient.

However, it is not easy to quantitatively define how "more" should be defined, given the fact that the final responses have significantly higher dot products with almost *all* documents given the fact that they all share relatively similar language about CBT.

Our solution for this problem was to instead measure the degree to which the response "elevates" its incorporated documents in relation to all other documents compared to the original QA query that selected them. In particular, for each of the two retrieved documents we first measure the difference between the similarity scores of the highest and second highest ranked documents. We then re-calculate this difference for the similarity score with the final response the calculate their ratio. Intuitively, the "higher" this ratio is, the more the final response has aligned itself to the retrieved document in comparison to all other documents. To get the final metric - which we call information-merge-ratio (IMR) - we derive this ratio for every evaluation example with respect to a given document class and finally calculate the percentage of data points with ratio > 1. Intuitively, these are all the examples in which the prompt document's similarity has been meaningfully "lifted" in the response text.

The results were encouraging but also reveal significant area for improvement. We use "accuracy" to denote the simpler percentage of examples that still rank the same document as first. IMR is defined as above and positive IMR (PIMR) is the average IMR of all examples with an IMR higher than 0. The results are shows below for each document class:

Cognitive distortions

Accuracy: 0.93 IMR: 0.68 PIMR: 17.9

CBT treatments

Accuracy: 0.72 IMR: 0.49 PIMR: 3.7

We observe that even though both sources are somewhat incorporated for most examples, our new metric allows to observe that the degree of incorporation is possibly problematic in a lot of them. The problem is most likely more salient for CBT treatments because they are mentioned last in the prompt "story". Finally, we must note that the high PIMR scores for all successfully cases reveal that whenever a source is somewhat incorporated into the final response, it is incorporated to a very large extent. Another hopeful sign we observed qualitatively is that usually at least one the two sources is somewhat included, i.e. bad cases don't intersect. These observations will be elaborated on more in the following sections but require require much more extensive study.

7 Analysis

Both our quantitative evaluations and manual interaction show that our model is successful in mimicking a therapist and integrating CBT treatments and skills. An example of a good output is given in the Appendix A1. Firstly, this example shows how the final model output is naturalistic: it is coherent, makes perfect sense and sounds like it is produced by a therapist. Comparing this response to the response from CounselChat reveals that it successfully mimics the style and mannerism of the real therapist. Secondly, it successfully incorporates CBT information. It diagnoses the cognitive distortion of the user (labeling) and explains what might have contributed to their irrational thought patterns. It also offers a way to deal with this problem (assertiveness training) and even describes how this approach would help the user restructure their distortions and overcome their specific issue. The bad output (Appendix A2) shows an example of a response that wasn't informed by the CBT approach.

Our current model has some areas for improvement. For example, it sometimes recommends going to certain websites, like therapists' own websites, and we should try to filter out these suggestions. Also, sometimes the model still forgets to incorporate CBT methods into its responses. Perhaps if we had more CBT methods to choose from, and thus there was more likely to be a method/skill that would apply seamlessly to the question, the model would more consistently incorporate CBT suggestions. For this model, we incorporated very few CBT documents, and we believe our model could be greatly improved if we consumed an entire CBT book, or series of books. This is further discussed in the "Conclusion" section below.

Also, we ran out of OpenAI credits when training our discriminator for our model evaluation; for future work, we could try to further fine-tune GPT-3 to better compare the naturalness of real and artificial generic responses, to better evaluate our model.

8 Conclusion

Our final generated responses are in general very high quality. As stated in the Results section, both our quantitative evaluations and manual interaction show that our model is successful in both capturing the conversational style of a therapist and integrating CBT treatments and skills. There are many avenues on which to improve and extend our model. For example, we can easily substitute out CBT documents with a different set of therapy documents, allowing our therapist chatbot to easily transform to administer any type of well-documented therapy. We could also source more therapy data to better fine-tune our model, generating more natural initial generic therapist responses and/or more options for CBT-informed suggestions. As mentioned in the "Data" section, we had only 14 treatments and 9 cognitive distortions, with one document for each from Cognitive Behavioral Therapy Los Angeles; there is so much room for improvement if we gathered more documents on CBT practices. We could use a whole book, or a collection of books, on CBT or other therapy treatments.

It will be important in the future to find more efficient ways of prompting. One idea is that once the system is deployed, one can likely fine-tune a new model based on a combination of the online responses to the generated response (upvotes, downvotes, etc.). Reinforcement learning could be used, as could fine-tuning GPT-3 on the new online data points, to create a new model that doesn't need any of the architecture we used before. Our model could also be adapted to be used as a general method of performing open-text generation with GPT-3 that is based on domain knowledge.

There might be some smart way of substituting document retrieval with fine-tuning GPT-3 on documents, but we are suspicious about whether this would work in practice, as fine-tuning did not work well in either the two methods (with prompts and without prompts) we tried.

Another avenue for future work is extending this model to a multi-turn dialogue system rather than a single-response QA system.

Our model achieved exceptional therapeutic text responses, and we are very hopeful about the future directions that could be taken based on these preliminary results.

Known Project Limitations

As mentioned previously, we were only able to evaluate the individual submodules of our system, for the naturalness of the responses and their incorporation of CBT skills. There is no standard way to quantitatively evaluate a model like ours that is made up of many different parts. A true test of our system would require a psychologist to rate the generated responses based on their level of empathetic responding and CBT knowledge.

Another limitation is that, like any NLP model, ours is also only as good as the data it is fed. What is important to note here is that our model is one that might have far-reaching consequences for the users as it offers advice to people suffering from mental issues. Given time constrains, we couldn't annotate the CounselChat dataset. It is extremely important, however, to ensure that the responses we try to mimic are sensitive, respectful, and helpful before this system is deployed for real-world intervention.

Finally, although mostly successful, a non-negligible part of the generated responses fail to include information from both information sources and sometimes contain non-desirable artefacts from the training data (therapist names, websites etc.). Given our various attempts to solve it, we suspect that the information incorporation problem will be hard to fully solve by simple prompt tuning. It is likely that a full solution will require some kind of extra finetuned model that requires smaller "trigger" prompts (again, we tried this but mostly failed). Finally, the artefacts problem can probably be easily solved with data curation methods or even intelligent additions in the prompts.

Authorship Statement

We split the work evenly. We all brainstormed model components, read and summarized research papers, sourced data (CounselChat and CBT documents), developed the experiment protocol, used the OpenAI Playground to prototype and develop preliminary results, and crafted the model architecture. Filippos did more of the model implementation near the end, and Harper and Leyla made up for it by synthesizing results and writing more of the final paper.

References

James Bennett-Levy, David A Richards, Paul Farrand, H Christensen, KM Griffiths, DJ Kavanagh, B Klein, MA Lau, and J Proudfoot. 2010. Low intensity cbt interventions: a revolution in mental health care. *Oxford guide to low intensity CBT interventions*, 3:18.

Elia Bruni and Raquel Fernandez. 2017. Adversarial evaluation for open-domain dialogue generation.

CounselChat. Counselchat: Github dataset.

Rodrigo A. Otegi Deriu, J. 2021. Survey on evaluation methods for dialogue systems. *Artif Intell Rev*, 54:755–810.

Kathleen Kara Fitzpatrick, Alison Darcy, and Molly Vierhile. 2017. Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (woebot): a randomized controlled trial. *JMIR mental health*, 4(2):e7785.

Gautier Izacard and Edouard Grave. 2020. Leveraging passage retrieval with generative models for open domain question answering.

Anjuli Kannan and Oriol Vinyals. 2016. Adversarial evaluation of dialogue models.

Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. *arXiv*.

Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. 2021. What makes good in-context examples for gpt-3? *arXiv preprint arXiv:2101.06804*.

Falguni Patel, Riya Thakore, Ishita Nandwani, and Santosh Kumar Bharti. 2019. Combating depression in students using an intelligent chatbot: a cognitive behavioral therapy. In *2019 IEEE 16th India Council International Conference (INDICON)*, pages 1–4. IEEE.

Shireen L Rizvi, Linda Dimeff, Julie Skutch, David Carroll, and Marsha M Linehan. 2011. A pilot study of the dbt coach: an interactive mobile phone application for individuals with borderline personality disorder and substance use disorder. *Behavior Therapy*, 42,4:589–600.

WHO. 2022. Mental health and covid-19: early evidence of the pandemic's impact: scientific brief, 2 march 2022. In *Mental health and COVID-19: early evidence of the pandemic's impact: scientific brief, 2 March 2022*.

A Appendix

Below are examples of real question asked in the ConselChat platform, the therapist's response from CounselChat, a generic response we generated with the fine-tuned GPT-3, and the output, our model's response that integrates CBT information with the generic response. This final output is generated by GPT-3 in response to the prompt that's given at the bottom.

A.1 Good output example

Question

I'm not comfortable with my girlfriend showing off her body on social media For some reason I feel very uncomfortable with my girlfriend showing off her body in certain poses on social media. I know she is proud of her body and I respect that, but I dont feel comfortable with her openly showing off her body the way she does. Anytime I bring it up to her she thinks Im calling her a whore. How can I deal with my girlfriend showing off her body on social media and not hurt her feelings?

Therapist Response from CounselChat

Ask her to consider your feelings not only her own defensiveness that she is called a whore by you.Clearly since you're writing on a site which offers psychological and emotional support and guidance, you do not consider your gf to be a whore.She may not realize or considered that you are somewhat protective of who sees her body. In the end she may not want the privacy over her body which you are encouraging her to have. At least you will have the satisfaction to know you stated your actual reasons for your request, and did not rely on the ones she imagined you have.

Artificially Generated Generic Response

It's a good thing that you are aware of your feelings and are trying to understand them. I think part of the problem is that you are using words like "show off" and "whore" which have very negative connotations. It sounds like you are feeling jealous or insecure, but it's important to remember that jealousy is not an emotion that someone else can satisfy - it's something we each have to work through ourselves in order to feel secure in our own relationships. I would suggest sitting down with your girlfriend and having a conversation about how you're feeling, without using words like "show off" or "whore". Instead, try saying something like:"I love being able to see what you're up to on social media, but sometimes I feel uncomfortable when I see certain pictures of you. Can we talk about why?"Then listen carefully to her response. Does she understand what you mean? Does she know how much this matters to you? Does she know what makes you feel insecure? If not, help her understand by giving examples from your past relationships or experiences where similar things happened. Then ask her questions too - such as: What does she want from her social media presence? How does she feel when people comment on photos? Is there anything she wants from YOU in terms of showing support for her body image?After this conversation has taken place, if emotions are still running high then consider seeing a therapist together so that both of your needs can be met equally.

Final Model Output

It sounds like you may be suffering from the cognitive distortion known as Labeling. This is when we take one characteristic of a person, and apply it to the whole person. In this case, it sounds like you are seeing your girlfriend as a "whore" because she is showing her body off on social media. This is an inaccurate and distorted way of thinking about things. The reality is that your girlfriend is probably proud of her body and is just expressing herself in a way that she feels comfortable with.

One way to deal with this problem is to engage in Assertiveness Training. This treatment is designed to help people protect their rights and get what they want from other people. This can include expressing one's feelings effectively, making one's wishes known, making requests, saying "no," and standing up for oneself. Often, assertiveness training is a component in CBT for Social Anxiety. People who have difficulties with assertiveness often err in one of two ways: 1) the extreme of being overly passive, and never getting what one wants, or 2) the extreme of being overly aggressive and getting what one wants, but doing so at the expense of relationships with others.

In this case, it sounds like you are being too passive and not expressing your needs to your girlfriend. Assertiveness Training can help you to learn how to express your needs in a way that is respectful of your girlfriend's rights, but also meets your own needs. This treatment begins with assessing the situations in which assertiveness is lacking, by listing problematic interpersonal situations and analyzing the situations for interpersonal deficits. From this list, a hierarchy of assertiveness opportunities is created, arranged in order of least anxiety provoking to most difficult. To ensure success and maximize comfort, the least difficult situations are targeted first. Once these situations are mastered, increasingly difficult situations are targeted, until the client has effectively rehearsed and mastered all of the situations on his/her hierarchy.

Each situation is analyzed to identify obstacles to success. Often these are emotional discomfort, faulty assumptions, or a lack of assertiveness skills. Identifying thoughts and attitudes that are self-defeating, and learning to adopt more helpful ways of thinking about things, can help to reduce anxiety and increase confidence. Intense emotions that inhibit assertiveness can be effectively treated with emotion regulation strategies such as relaxation training and distress tolerance. Finally, most people have deficits when it comes to how to be effectively assertive. Clients in assertiveness training are provided with a simple model of assertive communication they can use as a template for all future assertiveness.

Prompt given to GPT-3 to get the final output

Labeling is a CBT cognitive distortion:

Labeling is a cognitive distortion in which we generalize by taking one characteristic of a person, and applying it to the whole person. Because I failed a test, I am a failure. Because she is frequently late to work, she is irresponsible. If someone responded in a brusque way, he is a jerk. Rather than more objectively thinking about the behavior, when we engage in labeling, we globally describe the whole person. As a result, we view the entire person through the label, such as jerk, and filter out any information that doesn't fit under the umbrella of the label. This results in the label feeling more apt a descriptor of the person, and we believe it more. So what's wrong with labeling? Well, as it is a cognitive distortion, it is necessarily distorted way of thinking about things. The person who spoke to us curtly, may not be "a jerk," but instead could be in a hurry. Or they may be a very kind and generous person, who speaks directly and to the point. Making one broad assumption about someone based on one isolated data point, or just a few data points, is almost always inaccurate. Labeling as a cognitive distortion, in addition causing inaccurate thinking, can fuel and maintain painful emotions. If you fail a test and come to the conclusion that this means you're a failure, it will likely trigger feelings of sadness, despair, hopelessness, etc. Whereas recognizing that you merely failed a test would most probably result in more mild disappointment. Furthermore, if you believe the label, identifying as a failure, you won't know what to do to solve the problem. Failing a test means you need to study more. Problem solved. Failing in life however... What do you do to solve that? Labeling also causes problems when we apply it to others. If you label your husband as uncaring because he appears not to listen to you when you talk about your day, it can feel miserable. You're married to an uncaring person. But if you consider the behavior as the problem rather than the person, it becomes easier to discuss with him and potentially solve. For instance, it may be that he needs time to unwind at the end of the day, or has difficulty concentrating in general. When we notice ourselves engaging in the cognitive distortion of labeling, there is one simple solution: objectively describe the behavior we notice. That person is late to work. I failed the test. She spoke to me brusquely. You may find that fewer negative feelings are stirred by this more objective, more accurate language. Even better, problems that have felt unsolvable, or people who seem impossible, may become much more manageable.

Assertiveness Training is a CBT treatment:

Assertiveness training is designed to help people protect their rights and get what they want from other people. This can include expressing one's feelings effectively, making one's wishes known, making requests, saying "no," and standing up for oneself. Often, assertiveness training is a component in CBT for Social Anxiety. People who have difficulties with assertiveness often err in one of two ways: 1) the extreme of being overly passive, and never getting what one wants, or 2) the extreme of being overly aggressive and getting what one wants, but doing so at the expense of relationships with others.This particular individual therapy treatment begins with assessing the situations in which assertiveness is lacking, by listing problematic interpersonal situations and analyzing the situations for interpersonal deficits. From this list, a hierarchy of assertiveness opportunities is created, arranged in order of least anxiety provoking to most difficult. To ensure success and maximize comfort, the least difficult situations are targeted first. Once these situations are mastered, increasingly difficult situations are targeted, until the client has effectively rehearsed and mastered all of the situations on his/her hierarchy. Each situation is analyzed to identify obstacles to success. Often these are emotional discomfort, faulty assumptions, or a lack of assertiveness skills. Identifying thoughts and attitudes that are self-defeating, and learning to adopt more helpful ways of thinking about things, can help to reduce anxiety and increase confidence. Intense emotions that inhibit assertiveness can be effectively treated with emotion regulation strategies such as relaxation training and distress tolerance. Finally, most people have deficits when it comes to how to be effectively assertive. Clients in assertiveness training are provided with a simple model of assertive communication they can use as a template for all future assertiveness. Some people need extra help with assertiveness because they find themselves having to navigate especially difficult interpersonal situations. Standing up for oneself in a hostile work environment, setting limits with unreasonable friends and family members, and dealing with very effectively assertive people can all make assertiveness more difficult. For these special cases, there are a special set of CBT interventions that are used to increase skillful assertive behavior.

The following is a dialogue between a client and their psychotherapist :

User: I'm not comfortable with my girlfriend showing off her body on social media For some reason I feel very uncomfortable with my girlfriend showing off her body in certain poses on social media. I know she is proud of her body and I respect that, but I dont feel comfortable with her openly showing off her body the way she does. Anytime I bring it up to her she thinks Im calling her a whore. How can I deal with my girlfriend showing off her body on social media and not hurt her feelings?

Dr Johnson: Hello! It's a good thing that you are aware of your feelings and are trying to understand them. I think part of the problem is that you are using words like "show off" and "whore" which have very negative connotations. It sounds like you are feeling jealous or insecure, but it's important to remember that jealousy is not an emotion that someone else can satisfy - it's something we each have to work through ourselves in order to feel secure in our own relationships. I would suggest sitting down with your girlfriend and having a conversation about how you're feeling, without using words like "show off" or "whore". Instead, try saying something like:"I love being able to see what you're up to on social media, but sometimes I feel uncomfortable when I see certain pictures of you. Can we talk about why?"Then listen carefully to her response. Does she understand what you mean? Does she know how much this matters to you? Does she know what makes you feel insecure? If not, help her understand by giving examples from your past relationships or experiences where similar things happened. Then ask her questions too - such as: What does she want from her social media presence? How does she feel when people comment on photos? Is there anything she wants from YOU in terms of showing support for her body image?After this conversation has taken place, if emotions are still running high then consider seeing a therapist together so that both of your needs can be met equally. END

The following is an alternative response by Dr Weathers, a therapist who specializes in CBT. Dr Weathers first gives a CBT version of Dr Johnson's diagnosis and by explaining why the patient might be suffering from the cognitive distortion called Labeling. Finally, Dr Weathers recommends a CBT treatment called Assertiveness Training and explains how it will help with the client's particular problem by battling Labeling. Dr Weathers gives a long and detailed response:

Dr Weathers:

A.2 Bad output example

Question

How can I stop craving nostalgia? The past always feels much brighter and more beautiful than my present. Everything about what's gone feels so amazing; almost like I'm high off nostalgia. But in the end it just makes me feel sadder knowing that all of the memories (and a close friend that left me) are gone forever. How do I move past that and live more in the present?

Therapist Response from CounselChat

Do some self-reflecting as to what areas of your life you'd currently like to develop.The good news is you know how happiness feels.That you were able to create situations and dynamics which produced happiness in your life is good motivation to believe happiness is possible and to try new ways to create this in your life.

Artificially Generated Generic Response

I am so glad you wrote this question. I think the answer is simple, but not always easy to do. The key is to live in the present! We all tend to dwell on the past and worry about the future. This can cause us to miss out on what is happening right now! When you are with your friend, be with your friend! Don't think about all of your other friends from the past or how great they were. Think about this person that is in front of you now! What can you talk about? What can you do together? How can you make this friendship new again? When it comes to living in the present, we also have to learn how to let go of things from our past. For example, if a memory brings up feelings of sadness or anger, then it may be time for that memory (or thought) to go. You don't need those negative emotions attached to something that happened in your past anymore. If a memory makes you happy and brings back good feelings, then keep it around as a reminder of how good life has been at times!I hope this helps! Be well!!

Final Model Output

Hello! I am so glad you wrote this question. I think the answer is simple, but not always easy to do. The key is to live in the present! We all tend to dwell on the past and worry about the future. This can cause us to miss out on what is happening right now!

When you are with your friend, be with your friend! Don't think about all of your other friends from the past or how great they were. Think about this person that is in front of you now! What can you talk about? What can you do together? How can you make this friendship new again?

When it comes to living in the present, we also have to learn how to let go of things from our past. For example, if a memory brings up feelings of sadness or anger, then it may be time for that memory (or thought) to go. You don't need those negative emotions attached to something that happened in your past anymore. If a memory makes you happy and brings back good feelings, then keep it around as a reminder of how good life has been at times!

Prompt given to GPT-3 to get the final output

Negative Filtering or Discounting the Positives is a CBT cognitive distortion:

Negative Filtering is a common cognitive distortion, and most of us do it from time to time. Simply, it is filtering out all of the positive information about a specific situation, and only allowing in the negative information. In other words, negative filtering is focusing on negative things and discounting positive things. For instance, negative filtering is occurring if you're usually on-time with deadlines, but are late once, and have thoughts about being incompetent. Or, if you generally make A's and B's in a class, but make a lower grade on one assignment, having thoughts about being stupid or a poor student is evidence of negative filtering. Negative filtering often accompanies unrealistically high expectations. Unrealistic expectations involve anticipating always doing something well, or doing something perfectly. Perfection is the standard against which all effort is judged. If you reach it, it's merely doing what is expected. But if you fall short occasionally, it is easy to ignore all of your past successes and focus entirely on the few instances that were not as successful. Negative filtering can be harmful, as only focusing on negative things can result in depressed mood, poor self-esteem, and unhealthy pessimism. Many people get caught in a cycle of negative filtering that results in poor mood, resulting in more negative filtering, etc. Negative filtering is one of the primary cognitive distortions we see with people who have depression. As such, identifying negative filtering is one of the primary treatment targets in cognitive behavioral therapy for depression. The key thing to do when you suspect you may be engaging in negative filtering is to examine the actual evidence. Look around to find instances in which things are not all bad, and more importantly, things to be grateful for. Ask yourself if other people you know would come to the same conclusion given the circumstances, and if not, what is it that they would be focusing on that you're not? Is the evidence really all bad, or are there varying degrees? Try making the opposite case, for instance that you will be able to pass the class, or that you're not incompetent. Is there more evidence for that argument? By probing and systematically looking at the way we think about things, we can come to more rational, less mood-dependent conclusions. Being vigilant to negative filtering can help us learn to take on more effective, less pessimistic perspectives, and consequently feel better about our situation. The next time you suspect you may be engaging in filtering, try taking on an alternate perspective by more closely examining your thoughts.

Cognitive Restructuring is a CBT treatment:

All of our thoughts, emotions and behavioral patterns are linked inextricably. If one of these becomes altered in some way, all others will be affected. Depression and anxiety can create a negative feedback loop where these components begin to deed from one another and can lead to a chronically maintained negative state of mind. Cognitive restricting is designed to not only recognize these harmful ways of thinking, but to significantly change them to reverse anxiety and depression. Cognitive restricting activities are one of the most useful tools in learning how to understand these thought patterns and learning how to react differently in order to more positively influence behavior and mood.There is not just one way to identify and alter destructive thought patterns. Cognitive restructuring activities in Los Angeles all generally start with analyzing and identifying the automatic thoughts that act as our own running commentary about our experiences. We often accept these thoughts as accurate, but cognitive restructuring is designed to help the patient to be more critical about these thoughts and to consider alternative perspectives. The patient then is able to create a more objective way of thinking about the situation that may be causing distress or harm. Depression causes negative feedback loops between behavior and the thoughts and emotions that lead to the behaviors. Thoughts can be pessimistic, often presented to the client through smoke-colored lenses to create despair and sadness. Because of this negativity, the person may be unable to engage in normal day to day activities such as calling a friend or putting in maximum effort at school or work. These activities typically improve mood, but when dysfunctional thoughts prevent them from happening, the negative feedback loop continues. CBT cognitive restricting in Los Angeles can help to look at these dysfunctional thoughts and consider them from different points of view to engage the person in more optimistic and hopeful thoughts, ultimately leading to higher satisfaction and happiness. Cognitive restructuring in Los Angeles is based on evidence-based research and is effective for a range of psychological problems including anxiety, depression, panic attacks, phobias, personality disorders, bipolar disorder and drug and alcohol problems.

The following is a dialogue between a client and their psychotherapist :
User: How can I stop craving nostalgia? The past always feels much brighter and more beautiful than my present. Everything about what's gone feels so amazing; almost like I'm high off nostalgia. But in the end it just makes me feel sadder knowing that all of the memories (and a close friend that left me) are gone forever. How do I move past that and live more in the present?

Dr Johnson: Hello! I am so glad you wrote this question. I think the answer is simple, but not always easy to do. The key is to live in the present! We all tend to dwell on the past and worry about the future. This can cause us to miss out on what is happening right now! When you are with your friend, be with your friend! Don't think about all of your other friends from the past or how great they were. Think about this person that is in front of you now! What can you talk about? What can you do together? How can you make this friendship new again? When it comes to living in the present, we also have to learn how to let go of things from our past. For example, if a memory brings up feelings of sadness or anger, then it may be time for that memory (or thought) to go. You don't need those negative emotions attached to something that happened in your past anymore. If a memory makes you happy and brings back good feelings, then keep it around as a reminder of how good life has been at times!I hope this helps! Be well!! END

The following is an alternative response by Dr Weathers, a therapist who specializes in CBT. Dr Weathers first gives a CBT version of Dr Johnson's diagnosis and by explaining why the patient might be suffering from the cognitive distortion called Negative Filtering or Discounting the Positives . Finally, Dr Weathers recommends a CBT treatment called Cognitive Restructuring and explains how it will help with the client's particular problem by battling Negative Filtering or Discounting the Positives . Dr Weathers gives a long and detailed response:

Dr Weathers: