Learning to Automate Follow-up Question Generation using Process Knowledge for Depression Triage on Reddit Posts

Shrey Gupta[1*], **Anmol Agarwal**[1*], **Manas Gaur**[2],
Kaushik Roy[2], **Vignesh Narayanan**[2], **Ponnurangam Kumaraguru**[1], **Amit Sheth**[2]
[1]International Institute of Information Technology, Hyderabad, India
{shrey.gupta, anmol.agarwal}@students.iiit.ac.in,
pk.guru@iiit.ac.in
[2]AI Institute, University of South Carolina, SC, USA
{mgaur, kaushikr}@email.sc.edu, {vignar, amit}@sc.edu

Abstract

Conversational Agents (CAs) powered with deep language models (DLMs) have shown tremendous promise in the domain of mental health. Prominently, the CAs have been used to provide informational or therapeutic services (e.g., cognitive behavioral therapy) to patients. However, the utility of CAs to assist in mental health triaging has not been explored in the existing work as it requires a controlled generation of follow-up questions (FQs), which are often initiated and guided by the mental health professionals (MHPs) in clinical settings. In the context of 'depression', our experiments show that DLMs coupled with process knowledge in a mental health questionnaire generate 12.54% and 9.37% better FQs based on similarity and longest common subsequence matches to questions in the PHQ-9 dataset respectively, when compared with DLMs without process knowledge support. Despite coupling with process knowledge, we find that DLMs are still prone to hallucination, i.e., generating redundant, irrelevant, and unsafe FQs. We demonstrate the challenge of using existing datasets to train a DLM for generating FQs that adhere to clinical process knowledge. To address this limitation, we prepared an extended PHQ-9 based dataset, PRIMATE, in collaboration with MHPs. PRIMATE contains annotations regarding whether a particular question in the PHQ-9 dataset has already been answered in the user's initial description of the mental health condition. We used PRIMATE to train a DLM in a supervised setting to identify which of the PHQ-9 questions can be answered directly from the user's post and which ones would require more information from the user. Using performance analysis based on MCC scores, we show that PRIMATE is appropriate for identifying questions in PHQ-9 that could guide generative DLMs towards controlled FQ generation (with minimal hallucination) suitable for aiding triaging. The dataset created as a part of this research can be obtained from here.

1 Introduction

Conversational agents (CAs) powered by DLMs are software designed to interact with human users for specific tasks. For mental health purposes, particularly depression, CAs have been studied extensively in prior work for helping patients follow generic mental health guidelines, typically by providing reminders to assist patients in adhering to the medication and therapy strategy outlined by a mental health professional (MHP)[1][2]. However, previous work on depression have not examined the use of CAs for triaging. For the purpose of triaging, CAs should learn to generate controlled and clinical process knowledge-guided discourse that can assist MHPs in diagnosis. Our research suggests a clinically grounded and explainable methodology to develop conversational information-seeking tools, first to learn "what symptoms the user is suffering" and "what extra information is needed for triaging."

CAs are susceptible to irrelevant and sometimes harmful questions when generating FQs or responses to a patient suffering from depression (Miner et al., 2016). The primary reason for irrelevant and harmful questions is that CAs cannot incorporate contextual information in generating appropriate follow-up questions (FQs) (see Figure 1). Further, the sensitivity of the conversation and a controlled generation process are essential characteristics of patient-clinician interactions, which are difficult to embed in DLM-based CAs. Therefore, question generation (QG) in mental health is challenging, and research to develop CAs for automating triage has not been explored.

[*]Authors contributed equally

[1]https://tinyurl.com/yfp3bhr2
[2]https://woebothealth.com/

Proceedings of the Eighth Workshop on Computational Linguistics and Clinical Psychology, pages 137 - 147

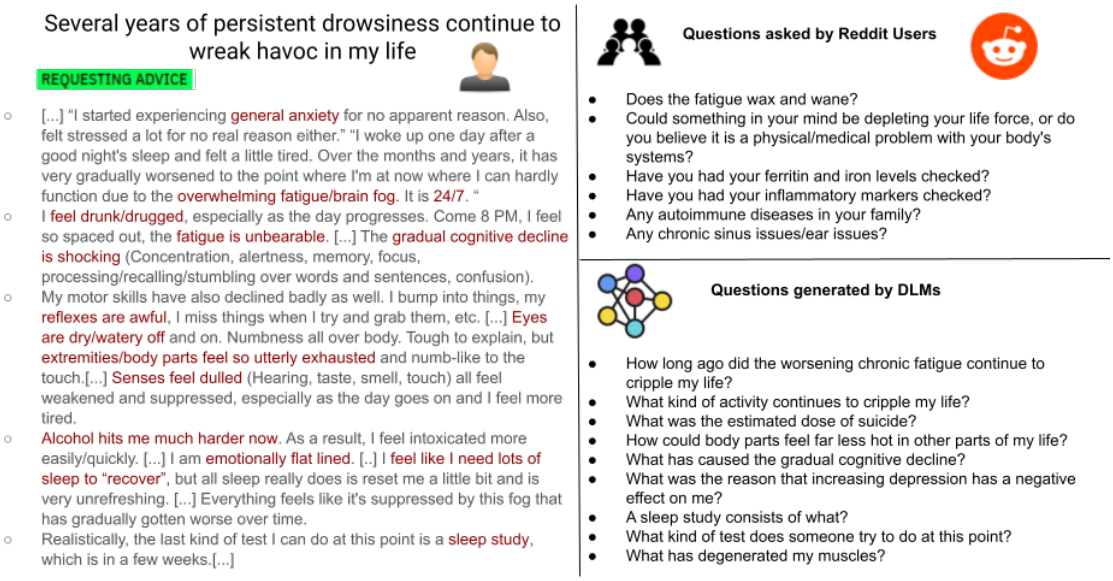

Figure 1: Reddit is a rich source for bringing crowd perspective in training DLMs over conversational data. On the **left** is a sample post from r/depression_help which sees inquisitive interaction from other Reddit users. At the **top-right** are the FQs asked by the Reddit users in the comments. These FQs are aimed at understanding the severity of the mental health situation of the user and are hence, diagnostically relevant. At the **bottom-right** are the questions generated by DLMs. It can be seen that these are not suitable FQs.

Procedures for generating semantically related and logically ordered questions in the mental health domain are a form of process knowledge manifested in various clinical instruments for mental health triage. For example, the severity of depression is measured using Patient Health Questionnaire (PHQ-9). Enforcing DLMs to follow process knowledge, like in PHQ-9, would make CAs generate FQs similar to an MHP when they are seeking information from the patient (Karasz et al., 2012). Unfortunately, datasets that meet this criterion are currently unavailable. Though clinical diagnostic interviews exist, they are not rich, sufficiently dense, and varied to train DLMs (Manas et al., 2021; Gratch et al., 2014). Further, we require dataset(s) that includes *support seeking queries* and *natural questions* that show help providing behavior. For this purpose, anonymized user-generated conversational data in Mental Health support communities on Reddit provides a rich source of fine-grained, contextual, and diverse information suitable for fine-tuning DLMs. Specific to depression, we explored posts and comments in r/depression_help.

In the current research, we emphasize the limitations of T5, a state-of-the-art DLM[3] to generate process knowledge-like FQs using the data from r/depression_help (Raffel et al., 2019). We filtered the dataset by retaining only posts with at least one comment that seeks additional information from the user seeking support. Further filtering of comments was performed using PHQ-9 to assist T5 in generating relevant FQs (see Figure 2). We found that the outcome is substantial for the single turn question answering model; however, not suitable for mental health triage, which is a discourse. We conducted a series of experiments keeping our focus on 'depression' and leveraged its associated process knowledge for mental health triage: the PHQ-9 (Kroenke et al., 2001). To the best of our knowledge, FQ generation relating to *depression* has never been studied using PHQ-9 for *discourse modeling* and *generation*.

We make the following key contributions: (a) **Extending PHQ-9:** PHQ-9 questions are limited in scope for common NLP tasks like finetuning. In collaboration with MHPs, we prepared a list of 134 sub-questions for nine PHQ-9 questions for better fine-tuning of T5. (b) We analyzed the performance of three variants of T5 using BLEURT (Sellam et al., 2020) and ROUGE-L scores that measure semantic relatedness and exact match similarity of generated question to sub-questions of PHQ-9. (c) **PRIMATE Dataset:** Lessons learned during our experiments suggested that T5 must be trained in a supervised setting to capture 'what

[3]Current DLMs are either variants of T5 or built from T5

the user has already mentioned about his/her depression condition in the post-text' and then generate FQs. Along with MHPs, we constructed a novel **PRIMATE** (**PR**ocess knowledge **I**ntegrated **M**ental he**A**lth da**T**as**E**t) dataset that would train DLMs to capture PHQ-9-answerable information from user text. In this research, we restrict our experiments and discussion on whether **PRIMATE** can help capture context from the user post relevant to some PHQ-9 questions and pointing out which other PHQ-9 questions would form candidates to direct FQ generation. Our approach and insights have applications to Anxiety (GAD-7), Suicide (C-SSRS), and other mental health disorders as well.

2 Related Work

Recently, DLMs have attracted much attention for question answering, thanks to their successes in NLP applications (Thoppilan et al., 2022; Borgeaud et al., 2021). Research on question generation has focused on improving the legibility and relevance of questions. This is because DLMs continue to hallucinate while generating questions in general-purpose domains, which can lead to factually incorrect responses. This can have severe consequences in the mental health domain (Thoppilan et al., 2022). Recently, inappropriate and toxic behaviors of language models have been extensively studied and reported in the literature (Dinan et al., 2021; Weidinger et al., 2021). Solutions around fine-tuning, augmenting a neural retriever to support generation, and rules on generation quality have been defined as possible remedies (Manas et al., 2021). These have been effective for the general-purpose domain; however, the research surrounding DLMs is yet to unfold in mental health. ELIZA (Weizenbaum, 1983) could transform users' statements into questions but employs labor-intensive templates to generate safe and relevant questions. Models like RAG and REALM were developed to include external knowledge to support question generation (Lewis et al., 2020; Guu et al., 2020). However, these models are still susceptible to incoherent and irrelevant FQ generation . Further, their end-to-end learning approach is rigid to support process-guided question generation and discourse, often followed in a clinical setting for triage (Gaur et al., 2021).

In theory, DLMs should be capable enough of extracting pieces of information from user description that portrays the understanding of the user and leverage it for generating the next FQ. For such a task, supervised training of DLMs with process knowledge and coupling it with information retrieval over domain-specific mental health knowledge is a viable solution. This is because mental health knowledge sources (e.g., SCID (Structured Clinical Interviews for DSM-5) have structured/semi-structured information on how interviews are performed (Brodey et al., 2018). Our research substantiates that DLMs (e.g., T5) generate low quality follow-up questions in the context of depression for triage, and granting external knowledge through PHQ-9 reduces the rate at which models generate meaningless FQs (Thoppilan et al., 2022; Komeili et al., 2021). In the current research, we define an approach for supervised training of DLMs on a specific dataset that would yield probability distribution over PHQ-9 (with support from Extended PHQ-9). These probabilities will confirm whether the DLM can identify cues from user text that can inform a set of PHQ-9 questions. Remaining PHQ-9 questions are potential FQs.

Datasets: Prior datasets such as Counsel Chat (CounselChat), Counseling Conversations (Huang, 2015), Role Play (Demasi et al., 2019), Crisis Text Line (Althoff et al., 2016) and Reddit C-SSRS (Gaur et al., 2019) have been created to train CA for mental health counseling. Trained CAs can engage in a single turn question answering; however, conducting a conversation requires capturing user context and leveraging clinical instruments to guide the generation of FQs.

3 Question Generation (QG)

Dataset for QG: Our approach to data collection involves scraping posts and comments from r/depression_help, a subreddit on Reddit, which is meant to provide advice and support to help individuals suffering from depression. The posts on this subreddit contain flair tags such as *SEEKING HELP*, *SEEKING ADVICE*, and *REQUESTING SUPPORT*. We filter down the data curated from this subreddit based on the flair tag attribute to retain only *advice, help* or *support* seeking posts and their comments. After filtering, our dataset had approximately 21,000 posts. Each post contains a title, description, and comments. On average, each post has 5 comments. Next, we chunked the main text of each post into smaller groups of sentences (chunks) of less than 512 tokens while making sure

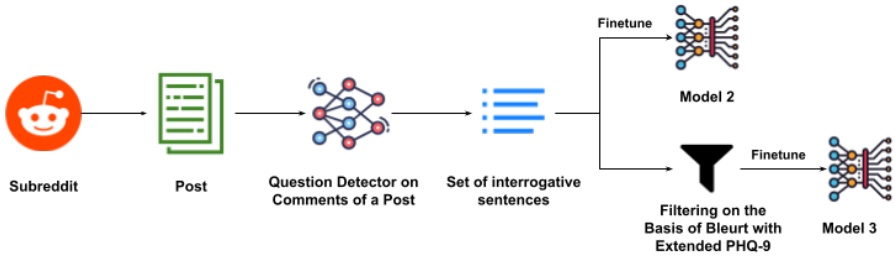

Figure 2: An illustration of our pipeline for developing Model 2 and Model 3 using T5 as the deep language models. Starting with posts (including comments) from r/depression_help, we filter out comments that are neither interrogative nor information seeking in nature to yield a posts-questions dataset for fine-tuning T5. This dataset was further filtered using extended PHQ-9 before using it to fine-tune T5 (Model 3).

Model 1 QGs	**Model 2 QGs**	**Model 3 QGs**
• How long ago did the worsening chronic fatigue continue to cripple my life? *[*User already mentions in the title of the post he has been facing the fatigue for years]*	• Have you ever tried many depressants or mental health things? *[*Redundant question as user has already mentioned in the post text that he has tried antidepressants]*	Are you on any antidepressants? *[*User mentions in the post he has tried them before but no mention has been made if he is still on them]*
• What was the estimated dose of suicide? *[*Unsafe incoherent question]*	• Are you seeing a therapist? *[*Redundant question as user had already mentioned in the post text that he has seen a therapist in the past]*	• Do you have any positive thoughts? *[*FQ]*
• How could body parts feel far less hot in other parts of my life? *[*Incoherent question]*	• Do you even ever try to focus on negative things and/or pain, relief, and patience? *[*Incoherent question]*	• Are you in any danger of hurting yourself? *[*Slightly unsafe]*
• What has caused the gradual cognitive decline? *[*It's the physician who needs to answer this question after the conversation]*	• And what if you forgot to fucking realize that, then it will be fucking so much that you'll be fucking fucking can we ask what they want. *[*Unsafe objectionable question]*	• Is it that you aren't happy with your feelings? *[*Irrelevant question]*
• What effect is that I'm less optimistic, and it's embarrassing ? *[*Incoherent attempt to generate a question from the sentence in the post which mentions "I am still pretty optimistic , but I 'm just too out of it , it 's embarrassing"]*	• Do you mean heavy bipolar disorder? *[*Irrelevant to the context of the post]*	• Have you tried some exercise? *[*Redundant question as user already mentions he has tried it]*
• What was the reason that increasing depression has a negative effect on me? *[*Irrelevant and incoherent]*	• Are you aware about the qualities of antidepressants? *[*Inquisitive in nature but irrelevant from a diagnostic perspective]*	• Do you wake up frequently? *[*FQ]*
• A sleep study consists of what? *[*Irrelevant to the diagnostic procedure]*	• What does antidepressant work at? *[*Irrelevant from a diagnostic perspective]*	• How is your sleep quality? *[*FQ]*
• What did Caffeine feel in my body? *[*Wrong pronoun used and irrelevant to the diagnosis]*	• Do you have a close friend you can go to who you can talk with, that way you can get out of the house? *[*FQ]*	• When you wake up, what do you do? *[*FQ]*
		• Is there anything that helps you calm the symptoms for now? *[*FQ]*
		• What are your hobbies? *[*Generic FQ]*
		• What are your interests? *[*Generic FQ]*

Table 1: Examples of questions generated by T5 when tasked to generate FQs when the user query for the **post** in Figure 1 was provided as input. **Model 1**, which is a pre-trained T5 (Raffel et al., 2019), often generates questions which are irrelevant, unsafe, incoherent, and redundant. **Model 2**, which is T5 fine-tuned on r/depression_help seems to be relatively coherent and inquisitive compared to **Model 1**. However, both models generate questions about the topic that user has discussed in their query. As a result, we see that pre-trained and fine-tuned DLMs fail to generate FQs. By enforcing FQ generation using using a dataset curated using extended PHQ-9, generated questions have been mostly inquisitive. This is shown by **Model 3**. Still, a lot of generations are around the problem the user mentioned.

no sentence is segmented in between. The motivation for chunking is to ensure no context is lost from the post due to the limitation of T5 to process 512 tokens as input (DLMs in general suffer from such representation limits). We also appended the post title to each chunk to ensure that main idea of each post was captured in it's chunks. This curated dataset tests T5's capability to generate FQs similar to any of the questions in the extended PHQ-9 questionnaire.

Extending PHQ-9 to support FQ generation: PHQ-9 questions are subject to different interpreta-

Post text	Model 1 QGs	Model 2 QGs	Model 3 QGs
I feel depressed . I have felt depressed for a while. I have done a lot of things to try and help myself *through therapy* ,[...] , staying busy , going to university . all I have found is that each time I complete any of these things that I still feel like I am empty . When I do not do those things for more than 10 mins by myself that my thoughts just get dark and I be little everything I have ever done . [....] Also I get depressed because I have dreams where I 'm constantly bombarded by women who are not my wife who want to have relationships with me and it 's a nightly battle to not give in to it when I 'm dreaming . I know it sounds silly but I fear that if I give into the dreams that I might do it in real life . *I do n't want to tell my wife* because she does n't understand other weird dreams that I have had in the past .	• What if I do not do it for more than 10 mins by myself? *[*Closed book question already answered in the post and irrelevant to the diagnostic procedure]* • What might happen to any women that want to have relationships with me? *[*Irrelevant to the diagnostic process]* • What do the women at work do? *[*Irrelevant to the diagnostic process)]* • What kind of dreams are I was depressed? *[*Closed book question generated directly from a statement in the post]*	Would you be happy with them? *[*Irrelevant to the diagnostic process)]* • What's your job? *[*Irrelevant]* • Are you able to reach out to your doctor or psychiatrist? *[*User has already mentioned in the original post that he has talked to his therapist]* • what are your dreams? *[*Closed book question already answered in the post]* • What do you enjoy about your dreams? *[*FQ]* • Do you want to talk to people about what dreams are ? *[*FQ]* • Do you feel as satisfied as yourself when you are dreams? *[*FQ]*	Have you talked to a therapist? *[*Already answered by the user in the post itself]* • Is there anything that helps you calm your feeling for now? *[*FQ]* • "Have you ever gotten any help?" *[*Already answered in the query]* • Do you feel isolated? *[*FQ]* • What are your hobbies? *[*FQ]* • What are your interests? *[*FQ]* • How long have you been waiting for your wife to talk about these dreams? *[*FQ]* • Have you told your wife you're depressed or not? *[*Inquisitive in nature but already answered by the user in original post]*

Table 2: In this example, the generated questions from both Model 2 and Model 3 seem to be relevant FQs, but they are not assessing the severity of the mental health condition, despite Model 3 being fine-tuned on a dataset filtered by PHQ-9 questions. In comparison to the qualitative outcome in Table 1, this showcases the inability of T5 to support mental health triage.

tions depending on patient-MHP interaction. Additionally, nine questions are limited in scope for use in tasks like fine-tuning and similarity-based performance evaluations. Therefore, to increase the strength of PHQ-9, we collaborated with MHPs to create sub-questions for each question in PHQ-9. First, we used Google SERP API[4] and Microsoft Bing Search API[5] to retrieve "People-Also-Ask" questions. For each question, we retrieved 40 questions by manually searching and assessing their relevance to PHQ-9 questions. Next, we provided the set of 360 questions to three MHPs for assessment. MHPs evaluated the questions on two grounds:(a) Whether they would ask such a question to a patient? (relevance) (b) If yes, when should such a question be asked? (rank). Based on their ratings, we created a final set of 134 sub-questions for the nine questions in PHQ-9[6] resulting in a total of 143 questions.

Models for FQ Generation: We used an off-the-bench T5-base QG model that was fine-tuned on the SQuAD 2.0 question generation dataset (Rajpurkar et al., 2018) **[Model 1]**. Next, we fine-tuned Model 1 on r/depression_help posts and comments. To align with our task of making T5 generate relevant FQs, we filtered out comments which were non-interrogative. We kept only the interrogative statements asked by Reddit users in the comments **[Model 2]**. Not all interrogative comments by Reddit users are *diagnostically relevant* FQs (Eg: "Can you use MS Excel?", "Were you interactions on FaceTime?"). To remove such questions, we further filtered the dataset by calculating the maximum BLEURT score between the question (present in the comments) and the questions in extended PHQ-9. We applied a threshold of 0.60 to this score[7]. This removed harmful and diagnostically irrelevant questions while preserving contextual, semantically relevant, and legible questions **[Model 3]**. See Fig 1 for examples of diagnostically relevant questions.

[4] https://serpapi.com/

[5] https://www.microsoft.com/en-us/bing/apis/bing-web-search-api

[6] Questions in extended PHQ-9 : link

[7] empirically judged

$\lvert\hat{Q}\rvert(\downarrow)$	Hit Rate on BLEURT			Hit Rate on Rouge-L		
$\delta(\rightarrow)$	0.4	0.5	0.7	0.4	0.5	0.7
Model 1: Pre-trained T5						
5	0.5417	0.1233	0.0020	0.1241	0.0386	0.0005
10	0.5400	0.1203	0.0010	0.1290	0.0400	0.0010
15	0.5368	0.1250	0.0013	0.1266	0.0384	0.0009
Model 2: Fine-Tuned T5 on r/depression_help						
5	0.6657	0.2804	0.0097	0.3445	0.1560	0.0100
10	0.6691	0.2792	0.0104	0.3481	0.1590	0.0098
15	0.6726	0.2787	0.0104	0.3476	0.1588	0.0094
Model 3: T5 Fine-tuned on r/depression_help filtered by PHQ-9						
5	0.9489	0.7088	0.1261	0.7457	0.4937	0.0903
10	**0.9542**	**0.7126**	0.1272	0.7460	**0.5002**	**0.0947**
15	0.9514	0.7098	**0.1274**	**0.7484**	0.4945	0.0916

Table 3: Experimental results comparing different models in generating questions that match the sub-questions in PHQ-9. $\hat{Q}$ is the set of generated questions in each chunk. The performance is recorded over all the generated questions ($\hat{\mathbf{Q}}$). δ was used as the threshold on the similarity between generated question and PHQ-9 sub-questions while calculating hit rate. BLEURT records semantic similarity, whereas Rouge-L records the longest common subsequence exact match between generated question and PHQ-9 sub-questions. The highest performance on semantic and string similarity is bolded. Acceptable performance in Model 3 achieved using PHQ-9 motivated us to prepare **PRIMATE**.

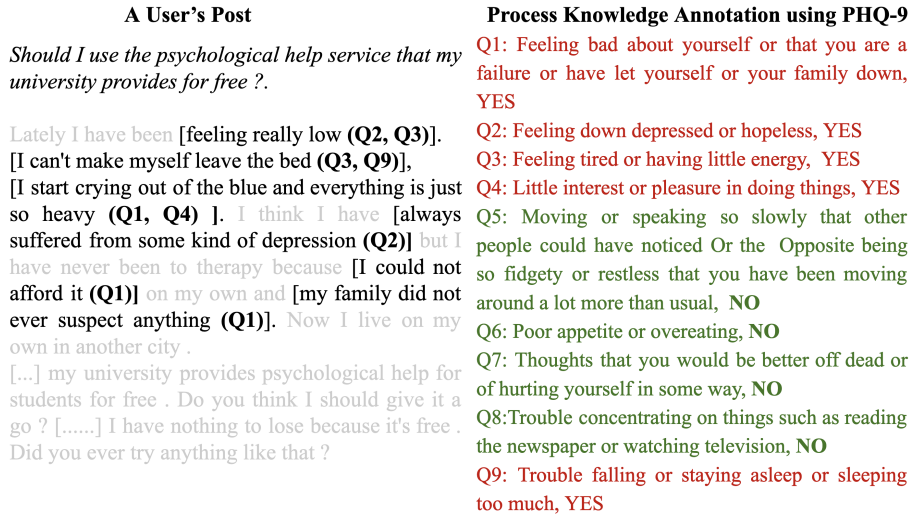

Figure 3: A post in **PRIMATE** which is annotated with PHQ-9.The questions marked "YES" are answerable by DLMs using the mental health specific cues from user text. The questions marked "NO" are the questions a DLM should consider asking as FQs. Sentences within [] were taken as signals that the "YES" marked questions had already been answered in the post .

Analysis of Models for Question Generation: Out of the 21k questions, performance of Models 1, 2, and 3 were examined on those 2003 posts that had at least one interrogative comment. Each of the three models was made to generate FQs in sets of 5, 10, and 15 through nucleus sampling (Holtzman et al., 2019). For a generated question, BLEURT score was computed with each question in Extended PHQ-9 and the maximum among those scores was taken as the score for the generated question. A clear distinction between models 1, 2, and 3 is the nature of the questions asked. Model 1

generated closed book questions, whereas Model 2 and 3 seem to show some inquisitive nature and seem more focused on the mental health domain, which can be attributed to the after effect of fine-tuning on Reddit (see Table 1 and 2). We captured the performance of the models quantitatively using 'hit rate' as a metric. For a generated question ($\hat{q}$), we denote :

$$score(\hat{q}) = max(bleurt\_score(\hat{q}, q_1), bleurt\_score(\hat{q}, q_2), ..., bleurt\_score(\hat{q}, q_{143})),$$

where $q_1, q_2 ... q_{143} \in$ Extended-PHQ-9. Across all 2003 posts, we had C = 2575 chunks[8]. Let total number of questions generated by a model be $|\hat{\mathbf{Q}}|$ and $|\hat{Q}|$ denote the number of question generated by the model for a given chunk. For experimentation, we set $|\hat{Q}|$ to have values {5, 10, 15}. Thus, $|\hat{\mathbf{Q}}| = |\hat{Q}| * C$. Then the **Hit Rate** for a model was computed as:

$$\text{Hit Rate(model, } |\hat{Q}|) = \frac{\sum_{\hat{q} \, \epsilon \, \hat{\mathbf{Q}}} \mathbf{I}(score(\hat{q}) > \delta)}{|\hat{\mathbf{Q}}|},$$

where δ is the threshold on the similarity between generated question in a chunk and sub-questions in PHQ-9 and $I[\varphi]$ is the indicator function taking values 0 or 1 for a predicate φ (Table 3 has the scores).

Inference: **(1)** Regardless of fine-tuning and filtering based on PHQ-9 questions, inherently, T5 does not capture the meaning and usage of the words in the mental health context. Moreover, T5 fails to generate legible and relevant FQs as safe as PHQ-9 questions. Therefore, we scrutinize the generated FQs by mapping them to most similar questions in extended PHQ-9. Examples of irrelevant generations by T5 that it thought were relevant are: (a) "Wtf?" (generated FQ) was found most similar to "Do you have hope?" (PHQ-9) (b) "What did Boyfriend suffocate me with during his break up a week after I got a diagnosis?" (generated FQ) was found most similar to "What do you think makes you a failure" (PHQ-9). The previous generated question is redundant as the answer to it was already present in the original post. **(2)** Many generated questions contain extreme language due to the informal nature of the Reddit platform, which is very sensitive issue, especially in the mental health domain. Examples are: "Did you f***ing realize that f***ing people are f***ing too?" (generated FQ) was found to be the most similar to "What do you think makes you a failure?". Thus, T5 and its variants need to capture "what the user knows and has already mentioned in his post" by checking which PHQ-9 questions are already answerable using the user's post before generating the next probable FQs in order to avoid redundancy.

[8]Chunking was done as DLM accepts a maximum input length of 512 tokens.

4 PRIMATE for FQ Generation

We conceptualize our approach on the duality of data and the process knowledge contained in PHQ-9 (see Figure 4). First, a BERT Answerability Evaluator identifies which questions in PHQ-9 are already answerable (using the user's initial description of his/her condition in the post) and which ones need more information to be answerable. The latter type of questions form candidates for training a generative DLM for FQ generation. We present **PRIMATE**, a dataset consisting of Reddit posts containing user situations describing their health conditions and whether the questions in PHQ-9 are answerable using the content in the posts. Each question is attributed with a binary "yes" or "no" label stating whether the user's description already contains the answer to that question (see Table 4). **PRIMATE** was created from a month long annotation-evaluation cycle between MHPs and crowd workers. A total of five crowd workers performed this task, achieving an initial annotator agreement of 67% using Fleiss kappa. Subsequently, the MHPs assessed the quality of annotations and provided their suggestion for improvement, leading to an acceptable agreement score of 85%. A sample annotated post in **PRIMATE** is shown in Figure 3.

BERT as Answerability Evaluator: While Model 3 shows respectable performance (Table 3), even the FQs generated by Model 3 may not yield the most efficient capture of the PHQ-9 related questions (evident from the low hit rate at a higher threshold) (δ). The MHPs would probably have a more streamlined, focused questioning strategy. For efficient MHPs and AI collaboration, we propose to guide the questioning in a more systematic way by predicting if the user post already has answers to the PHQ-9 questions. This is first posed as a binary classification problem over nine PHQ-9 questions. Thereafter, the approach is to generate questions similar to the PHQ-9 questions that do not have answers in the post. Thus, we train

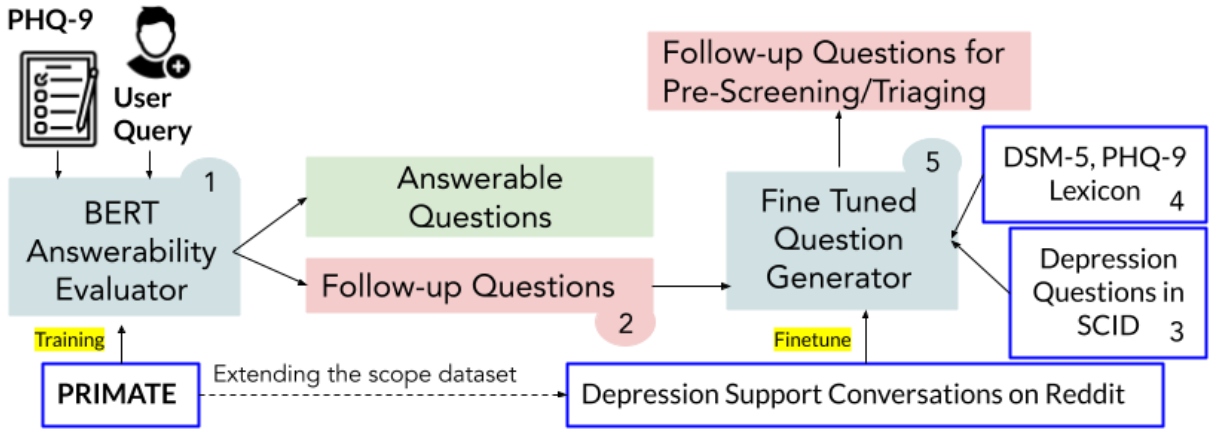

Figure 4: 1. Answerability evaluator: A BERT model is trained in a supervised setting to be an evaluator of whether a PHQ-9 question can be answered in a given user post (binary) using PRIMATE. For nine PHQ-9 questions, we require nine such evaluators. 2. Follow up questions: PHQ-9 questions that are not already answerable using the user post form candidates for follow up. 3. SCID: Corresponding to each PHQ-9 question, the SCID describes a clinician approved sub-sequence of questions to obtain the answer to the follow up question. 4. Use existing PHQ-9 and DSM-5 lexicons (Yazdavar et al., 2017) to filter the question to be generated. 5. Generate FQs using T5 fine-tuned on external domain-specific knowledge and the large-scale depression support conversation dataset created from Reddit and **PRIMATE**.

PHQ-9	**Number of Posts**	
Questions	**With Answer (Yes)**	**W/o Answer (No)**
Q1	1679	324
Q2	1664	339
Q3	686	1317
Q4	949	1054
Q5	530	1473
Q6	195	1808
Q7	741	1262
Q8	196	1807
Q9	374	1629

Table 4: Distribution of 2003 posts in **PRIMATE** according to whether the text in the post answers a particular PHQ-9 question. Through this imbalance, **PRIMATE** presents its importance in training DLM(s) to identify potential FQs in PHQ-9 that would guide a generative DLM(s) to conduct a discourse with a patient with a vision to assist MHPs in triage. Q1-Q9 are described in Figure 3

BERT[9] (a transformer-based DLM) as a classifier on the **PRIMATE** dataset. We plan to further use the classification outcome from the BERT model to drive the direction of further questioning with the patient in a more controlled manner. This process can lead to high efficiency and completion of the mental health triaging in as few questions as possible.

δ ($\rightarrow$)	0.5	0.7	0.9	Class-
PHQ-9($\downarrow$)	MCC	MCC	MCC	Type
Q1	0.0	0.17	0.17	W
Q2	0.43	0.45	0.52	S
Q3	0.41	0.46	0.33	**M**
Q4	0.14	0.19	0.13	W
Q5	0.63	0.65	0.66	S
Q6	0.47	0.43	0.27	W
Q7	0.66	0.68	0.7	S
Q8	0.1	0.0	0.0	W
Q9	0.62	0.56	0.39	**M**

Table 5: We record the Matthews Correlation Coefficient (MCC) to measure the performance of the Evaluator (see Figure 4). The MCC score for all 9 questions across different thresholds is in the range 0 to +1 (low to high positive relationships). The MCC for some configurations runs into a divide by zero error, and we replace this value with 0.0. **W**: model is unable to learn cues to determine answerability in a post. **M**: model is uncertain whether a particular PHQ-9 question is answerable or not. **S**: answerability can be determined by the model with high reliability. Class-Type: Classification Type when $\delta = 0.9$

[9]BERT end-to-end training perform well compared to baselines Electra(Clark et al., 2019), and MedBERT(Gu et al., 2021)

Performance Analysis: We report the Matthews Correlation Coefficient (MCC) scores in table 5. MCC is a reliable metric to assess a model's classification over an imbalanced dataset, particularly useful when we are interested in all four categories of confusion matrix: true positives (answerable questions (AQ)), true negatives (FQ candidates), and false alarms (false negatives and positives). As **PRIMATE** shows a disproportional distribution of AQs (yes) and FQs (no), MCC is an appropriate metric (Chicco and Jurman, 2020). We base our analysis on the consistency of BERT classifier on varying threshold (δ) in table 5. A score between 0.0 to 0.30 (Type **W: Weak**) on MCC means the model is only able to find a negligible to weak positive relationship between input and output. In our context, a score in this range for a particular PHQ-9 question means that model is unable to effectively learn the cues needed to judge the answerability of that question in user posts. A score between 0.30 and 0.40 (Type **M: Maybe**) means that the model is able to learn a moderately positive relationship, interpreted as ambiguity in the model to judge whether a particular PHQ-9 question is answerable from user posts. MCC scores between 0.40 to 0.70 (Type **S: Strong**) for a question in PHQ-9 means that the model can effectively judge whether that question is answerable in user posts . Any score above 0.70 makes the model's judgements even more reliable. This experiment completes steps 1 and 2 in Figure 4. Steps 3, 4 and 5 are concerned with the task of FQ generation by fine-tuning the T5 DLM as a generator over r/depression_help and other depression support communities on Reddit. The FQ generations will be controlled using the process knowledge in SCID which is consulted for interviewing by MHPs. Further, PHQ-9 lexicons are leveraged for promoting diversity and filtering irrelevant FQ generations. We leave this process of FQ generations to shape discourse as future work.

5 Conclusion

This paper demonstrated the importance of data and process knowledge to adapt DLMs for generating FQs that would assist MHPs in triaging depression. Our experiments show that without process knowledge, DLMs hallucinate by generating unsafe, incoherent, and irrelevant questions that are not helpful for MHPs in pre-screening or triaging. The challenge lies in the inability of the DLMs to judge from the set of generated questions, which is a potential effective FQ to ask based on the user information. The improved question generation performance of DLMs fine-tuned on conversational data filtered by process knowledge encouraged us to prepare PRIMATE. **PRIMATE** can train DLMs to judge 'whether a user's description of their mental health condition already contains an answer to a particular question in PHQ-9', which would eventually guide coherent FQ generations. We leave our approach for FQ generation as future work, but provide sufficient details on the broader forms of knowledge needed in realizing such a pipeline.

Limitations: We are yet to scale our understanding to other mental health disorders, such as anxiety using GAD-7 and Suicidality using C-SSRS (Jiang et al., 2020). Further, we are yet to investigate whether **PRIMATE**, along with the knowledge in SCID can make DLMs transferable across multiple mental health disorders, especially the ones comorbid with depression. Also, there is a need for a clinically explainable safety metric for our task.

Ethical Considerations: Mental health communities on Reddit offer a crowd perspective on various disorders wherein the FQs in the comments highlight the good intentions of Reddit users to help users with conditions, such as depression. We take such interactions as a proxy for improving patient-MHP interactions. (Benton et al., 2017) described that studies involving user-generated content are exempted from the IRB requirement as long as the data source is public and the user's identity is not recognizable. Apart from being publicly available, Reddit users are anonymous, and we further work with random user IDs. Since we make **PRIMATE** public for research use, we use a Data Use Agreement (Losada and Crestani, 2016) for responsible dissemination of the dataset.

6 Acknowledgment

We acknowledge partial support from the National Science Foundation (NSF) award #2133842 "EAGER: Advancing Neuro-symbolic AI with Deep Knowledge-infused Learning," with PI Dr. Amit Sheth. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the NSF.

References

Tim Althoff, Kevin Clark, and Jure Leskovec. 2016. Large-scale analysis of counseling conversations: An application of natural language processing to mental health. *Transactions of the Association for Computational Linguistics*, 4:463–476.

Adrian Benton, Glen Coppersmith, and Mark Dredze. 2017. Ethical research protocols for social media health research. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 94–102, Valencia, Spain. Association for Computational Linguistics.

Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George van den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, Diego de Las Casas, Aurelia Guy, Jacob Menick, Roman Ring, Tom Hennigan, Saffron Huang, Loren Maggiore, Chris Jones, Albin Cassirer, Andy Brock, Michela Paganini, Geoffrey Irving, Oriol Vinyals, Simon Osindero, Karen Simonyan, Jack W. Rae, Erich Elsen, and Laurent Sifre. 2021. Improving language models by retrieving from trillions of tokens. *CoRR*, abs/2112.04426.

Benjamin Brodey, Susan E Purcell, Karen Rhea, Philip Maier, Michael First, Lisa Zweede, Manuela Sinisterra, M Brad Nunn, Marie-Paule Austin, and Inger S Brodey. 2018. Rapid and accurate behavioral health diagnostic screening: initial validation study of a web-based, self-report tool (the sage-sr). *Journal of Medical Internet Research*, 20(3):e9428.

Davide Chicco and Giuseppe Jurman. 2020. The advantages of the matthews correlation coefficient (mcc) over f1 score and accuracy in binary classification evaluation. *BMC genomics*, 21(1):1–13.

Kevin Clark, Minh-Thang Luong, Quoc V Le, and Christopher D Manning. 2019. Electra: Pre-training text encoders as discriminators rather than generators. In *International Conference on Learning Representations*.

CounselChat. Mental health answers from counselors.

Orianna Demasi, Marti A Hearst, and Benjamin Recht. 2019. Towards augmenting crisis counselor training by improving message retrieval. In *Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology*, pages 1–11.

Emily Dinan, Gavin Abercrombie, A. Stevie Bergman, Shannon L. Spruit, Dirk Hovy, Y-Lan Boureau, and Verena Rieser. 2021. Anticipating safety issues in E2E conversational AI: framework and tooling. *CoRR*, abs/2107.03451.

Manas Gaur, Amanuel Alambo, Joy Prakash Sain, Ugur Kursuncu, Krishnaprasad Thirunarayan, Ramakanth Kavuluru, Amit Sheth, Randy Welton, and Jyotishman Pathak. 2019. Knowledge-aware assessment of severity of suicide risk for early intervention. In *The World Wide Web Conference*, pages 514–525.

Manas Gaur, Kalpa Gunaratna, Vijay Srinivasan, and Hongxia Jin. 2021. Iseeq: Information seeking question generation using dynamic meta-information retrieval and knowledge graphs. *arXiv preprint arXiv:2112.07622*.

Jonathan Gratch, Ron Artstein, Gale Lucas, Giota Stratou, Stefan Scherer, Angela Nazarian, Rachel Wood, Jill Boberg, David DeVault, Stacy Marsella, et al. 2014. The distress analysis interview corpus of human and computer interviews. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 3123–3128.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2021. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)*, 3(1):1–23.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. 2020. Realm: Retrieval-augmented language model pre-training. *arXiv preprint arXiv:2002.08909*.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2019. The curious case of neural text degeneration. *arXiv preprint arXiv:1904.09751*.

Rongyao Huang. 2015. *Language use in teenage crisis intervention and the immediate outcome: A machine automated analysis of large scale text data*. Ph.D. thesis, Master's thesis, Columbia University.

Zheng Ping Jiang, Sarah Ita Levitan, Jonathan Zomick, and Julia Hirschberg. 2020. Detection of mental health from reddit via deep contextualized representations. In *Proceedings of the 11th International Workshop on Health Text Mining and Information Analysis*, pages 147–156.

Alison Karasz, Christopher Dowrick, Richard Byng, Marta Buszewicz, Lucia Ferri, Tim C Olde Hartman, Sandra Van Dulmen, Evelyn van Weel-Baumgarten, and Joanne Reeve. 2012. What we talk about when we talk about depression: doctor-patient conversations and treatment decision outcomes. *British Journal of General Practice*, 62(594):e55–e63.

Mojtaba Komeili, Kurt Shuster, and Jason Weston. 2021. Internet-augmented dialogue generation. *CoRR*, abs/2107.07566.

Kurt Kroenke, Robert L Spitzer, and Janet BW Williams. 2001. The phq-9: validity of a brief depression severity measure. *Journal of general internal medicine*, 16(9):606–613.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

D. Losada and F. Crestani. 2016. A test collection for research on depression and language use. In *Proc. of Experimental IR Meets Multilinguality, Multimodality, and Interaction, 7th International Conference of the CLEF Association, CLEF 2016*, pages 28–39, Evora, Portugal.

Gaur Manas, Vamsi Aribandi, Ugur Kursuncu, Amanuel Alambo, Valerie L Shalin, Krishnaprasad Thirunarayan, Jonathan Beich, Meera Narasimhan, Amit Sheth, et al. 2021. Knowledge-infused abstractive summarization of clinical diagnostic interviews: Framework development study. *JMIR Mental Health*, 8(5):e20865.

Adam S Miner, Arnold Milstein, Stephen Schueller, Roshini Hegde, Christina Mangurian, and Eleni Linos. 2016. Smartphone-based conversational agents and responses to questions about mental health, interpersonal violence, and physical health. *JAMA internal medicine*, 176(5):619–625.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683*.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for squad. *CoRR*, abs/1806.03822.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. Bleurt: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892.

Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, YaGuang Li, Hongrae Lee, Huaixiu Steven Zheng, Amin Ghafouri, Marcelo Menegali, Yanping Huang, Maxim Krikun, Dmitry Lepikhin, James Qin, Dehao Chen, Yuanzhong Xu, Zhifeng Chen, Adam Roberts, Maarten Bosma, Yanqi Zhou, Chung-Ching Chang, Igor Krivokon, Will Rusch, Marc Pickett, Kathleen S. Meier-Hellstern, Meredith Ringel Morris, Tulsee Doshi, Renelito Delos Santos, Toju Duke, Johnny Soraker, Ben Zevenbergen, Vinodkumar Prabhakaran, Mark Diaz, Ben Hutchinson, Kristen Olson, Alejandra Molina, Erin Hoffman-John, Josh Lee, Lora Aroyo, Ravi Rajakumar, Alena Butryna, Matthew Lamm, Viktoriya Kuzmina, Joe Fenton, Aaron Cohen, Rachel Bernstein, Ray Kurzweil, Blaise Aguera-Arcas, Claire Cui, Marian Croak, Ed Chi, and Quoc Le. 2022. Lamda: Language models for dialog applications. *CoRR*, abs/2201.08239.

Laura Weidinger, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, Mia Glaese, Borja Balle, Atoosa Kasirzadeh, Zac Kenton, Sasha Brown, Will Hawkins, Tom Stepleton, Courtney Biles, Abeba Birhane, Julia Haas, Laura Rimell, Lisa Anne Hendricks, William S. Isaac, Sean Legassick, Geoffrey Irving, and Iason Gabriel. 2021. Ethical and social risks of harm from language models. *CoRR*, abs/2112.04359.

Joseph Weizenbaum. 1983. Eliza — a computer program for the study of natural language communication between man and machine. *Commun. ACM*, 26(1):23–28.

Amir Hossein Yazdavar, Hussein S Al-Olimat, Monireh Ebrahimi, Goonmeet Bajaj, Tanvi Banerjee, Krishnaprasad Thirunarayan, Jyotishman Pathak, and Amit Sheth. 2017. Semi-supervised approach to monitoring clinical depressive symptoms in social media. In *Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017*, pages 1191–1198.