ECAI 2024
U. Endriss et al. (Eds.)

doi:10.3233/FAIA240785

# Counseling Responses for Mental Health Forum Questions with Early Maladaptive Schema Prediction

**Sujatha Das Gollapalli**[1], **Beng Heng Ang**[2], **Mingzhe Du**[1,3] **and See-Kiong Ng**[1]

[1]Institute of Data Science, National University of Singapore
[2]Integrative Sciences and Engineering Programme, National University of Singapore
[3]College of Computing and Data Science, Nanyang Technological University
ORCID (Sujatha Das Gollapalli): https://orcid.org/0000-0002-4567-8937, ORCID (Beng Heng Ang): https://orcid.org/0009-0008-3096-195X, ORCID (Mingzhe Du): https://orcid.org/0000-0001-7832-0459, ORCID (See-Kiong Ng): https://orcid.org/0000-0001-6565-7511

**Abstract.** State-of-the-art Large Language Models (LLMs) have shown remarkable capabilities for general Question Answering (QA) tasks. However, their practical use for answering mental health questions has been limited due to the missing link between LLM-generated answer responses and well-established theories and guiding principles from Psychology and Counseling. We present a first step in this direction with *STeer*, an AI-based method that supports **S**chema **T**herapy-**e**nabl**e**d **r**esponses for mental health questions on community QA forums. *STeer* uses Early Maladaptive Schemas (EMSs), a fundamental concept from Schema Therapy that characterizes "self-defeating, unhealthy patterns of thought and behavior" in individuals, to effectively prompt state-of-the-art LLMs to generate specific, theoretically-grounded, empathetic counseling responses to mental health questions. We present *EMSRank*, a novel method based on the Personalized PageRank algorithm, to automatically predict the EMSs from mental health forum question texts. We show that *EMSRank* is computationally scalable and can be further combined with textual entailment to obtain high precision, explainable EMS labels for mental health forum questions. To address the current lack of annotated datasets, we also leveraged on *EMSRank* to create a first-of-its-kind, large dataset of about 23K EMS-annotated mental health questions from three diverse, currently operating, peer-support community forums for mental health. With the global rise in mental health issues, our work is a timely step towards enabling the use of AI-based assistive tools for counseling support on mental health community forums.

## 1 Motivation

**M**ental **h**ealth (MH) and well-being **c**ommunity **q**uestion **a**nswering (cQA) forums are attracting increased volumes of traffic in the current age of digital healthcare [23, 13, 51, 6]. Amidst the global rise in mental health issues [65], cQA forums present convenient platforms for cost-effective and speedy access to both community support as well as professional advice. Indeed, this has resulted in a push towards blended healthcare and a rise in research into AI-based assistive tools for both peer and professional support [52, 2, 28].

At the same time, recent breakthrough research has shown that Large Language Models (LLMs) can be trained "to act in accordance with the user's intentions" and as a consequence be "prompted" to perform a range of AI and NLP tasks including question answering (QA) [44, 12, 61, 43]. In particular, due to their exceptional QA performance, state-of-the-art LLMs are now incorporated into practical QA bots on knowledge-sharing and problem-solving cQA forums [53, 59].[1] However, LLMs are yet to be fully espoused with mental health related QA tasks due to several reasons including the known shortcomings of LLMs such as privacy concerns, ethical implications, as well as the crucial requirement in the MH domain for machine learning model outputs to be theoretically-grounded and explainable in view of potential risk to the target care-seekers [10, 18, 25]. We focus on addressing the latter concern in this work: *How can we incorporate relevant theories from Psychology and guiding principles from Counseling for grounding response generation with LLMs?*

We present *STeer*, an AI-based method to address this precise question. *STeer* is designed to complement assistive systems such as CARE [28] and SAHAR [7] that support peer counseling, as an effective scaling mechanism for handling the large numbers of incoming counseling questions on cQA forums.[2] Our objective in *STeer* is to generate empathetic as well as therapeutic suggestions for efficient editing by a peer counselor to respond effectively to counseling questions. Response generation in *STeer* is enabled via Schema Therapy (ST), an integrative therapy approach and theoretical framework introduced by Young [68, 17], for treating clients with personality disorders, characterological issues, and various other mental health problems [5]. Schema Therapy-based counseling has recently seen increasing adoption in practice due to its effectiveness in treating a range of individual and relationship problems [60, 49, 39].

**Early Maladaptive Schemas (EMS)** comprise a fundamental concept in Schema Therapy [67, 68]. EMS labels characterize negative and enduring cognitive patterns that develop from childhood experiences and affect an individual's perception of reality, influencing their emotions, thoughts, and personal, societal, and professional behavior. A crucial step in the practice of Schema Therapy includes

[1] Examples: https://www.quora.com/ and https://community.cisco.com/t5/cisco-cafe-blogs/ai-in-cisco-community-an-experiment-in-evolution/ba-p/4818048

[2] Trained individuals provide help and support to care seekers as opposed to rigorously-qualified and licensed professional counselors. For example: https://www.7cups.com/listener/become-a-volunteer-listener.php

**Table 1.** The 18 Early Maladaptive Schema names and sample statements from Young's Schema Questionaire

| |
|---|
| ***List of EMS Labels*** 1. Abandonment/Instability (AB), 2. Mistrust/Abuse (MA), 3. Emotional Deprivation(ED), 4. Defectiveness/Shame (DS), 5. Social Isolation/Alienation (SI), 6. Dependence/Incompetence (DI), 7. Vulnerability to harm or illness (VH), 8. Enmeshment/Undeveloped Self (EM), 9. Failure to achieve (FA), 10. Entitlement/Grandiosity (ET), 11. Insufficient Self-control/Self-discipline (IS), 12. Subjugation (SB), 13. Self-sacrifice (SS), 14. Approval-seeking/Recognition-seeking (AS), 15. Negativity/Pessimism (NP), 16. Emotional Inhibition (EI), 17. Unrelenting Standards/Hypercriticalness (US), 18. Punitiveness (PU) |
| Sample "MISTRUST/ABUSE" related questionaire items<br>• It is only a matter of time before someone betrays me. • I have a great deal of difficulty trusting people.<br>• I feel that I cannot let my guard down in the presence of other people, or else they will intentionally hurt me . . . |

uncovering individuals' EMSs through the administration of Young's Schema Questionaire (**YSQ**), where 232 statements about oneself are rated by the individual on a scale of 1 ("Completely untrue of me") to 6 ("Describes me perfectly").[3] The responses from this questionaire are used to identify one or more labels from a list of 18 EMS labels listed in Table 1.

Consider a question from the CounselChat dataset [8] shown in Table 2 that we use for illustration in this paper. Here, the person is seeking advice on handling the "insecurity" he was feeling regarding his girlfriend. One of the EMS labels identified for this question text (by a professional counselor) is "MISTRUST/ABUSE (MA)".[4] Schema Therapy [68] provides counseling guidelines (shown in Table 7) to address the feelings of disconnection and rejection in individuals with the "MISTRUST/ABUSE" schema [30, 16].

**Table 2.** (Anecdote) Example question with an EMS label[4]

| |
|---|
| *My relationship feels off and I feel insecure.* My girlfriend's grandma passed away 5 months ago. They were very close. She took care of her till she died. Things kinda returned to normal few weeks later. Last month it feels like we hit a brick wall. Intimacy . . . *Now I'm very insecure about us and have thoughts of her cheating.* She says otherwise, but . . . It just feels like something is really off. |
| **EMS Label**: 2. MISTRUST / ABUSE (MA)<br>Label Definition: The expectation that others will hurt, abuse, humiliate, cheat, lie, manipulate, or take advantage. Usually involves the perception that the harm is intentional or the result of unjustified and extreme negligence. May include the sense that one always ends up being cheated relative to others or "getting the short end of the stick." |

**Contributions**: In this paper, we propose methods for generating Schema Therapy-grounded counseling responses to mental health questions using their EMS labels. Our contributions are:

1. We propose *EMSRank*, a novel technique for computing EMS labels for mental health cQA posts using the Personalized PageRank algorithm. The underlying graph for applying *EMSRank* is obtained by innovatively applying Sentence Embeddings on Young's Schema Questionaire and converting questions texts into "personalization" vectors for PageRank.
2. We demonstrate the performance of *EMSRank* on a dataset of about 434 cQA questions that we compiled from three operational, prominent mental health forums BeyondBlue,[5] 7-Cups,[6] and PatientInfo.[7] We illustrate that *EMSRank* is able to predict EMS labels efficiently at web-scale and can be further combined with textual entailment to obtain high-precision, explainable EMS labels. This crowd-annotated dataset complements existing small-scale expert-annotated datasets for the EMS prediction task.
3. Applying *EMSRank* on the entire collection of ∼70K question texts crawled from the aforementioned MH forums, we compile a large, first-of-its-kind dataset (**CQEMS**) of ∼23K **EMS**-annotated mental health **C**ommunity **Q**uestions, and release it for academic research as part of our contributions.[8]
4. We describe *STeer*, our technique for generating **S**chema **T**herapy-**e**nabl**e**d **r**esponses for mental health questions. In *STeer*, treatment guidelines from Schema Therapy for specific EMSs are used to construct suitable LLM prompts for generating responses. We compare *STeer* responses to responses from professional counselors on a subset of questions from the CounselChat dataset [8] and illustrate that *STeer* responses are empathetic as well as relevant in addition to being grounded on Schema Therapy.

**Organization**: In Section 2, we describe our novel algorithm *EMSRank* for predicting EMS labels using Personalized PageRank. The use of EMS labels in *STeer* for generating grounded responses is also covered in this section. In Section 3, we describe our datasets, experiment settings, and our findings. Recent work on topics closely-related to this study are described in Section 4, while Section 5 concludes the paper with a summary and directions for future research.

## 2 Methods

### 2.1 EMS Prediction with PageRank

Due to a lack of labeled datasets for learning supervised models, previous works applied sentence similarity as well as zero-shot approaches with LLMs using YSQ statements and schema definitions, respectively, for predicting EMS labels [22]. Against these approaches, we consider unsupervised PageRank-based algorithms for similar problems such as emotion detection and textspan ranking when specialized dictionaries for class labels are available [21, 62]. Given the availability of Schema Therapy resources, we therefore explore the design of a PageRank-based variant for EMS prediction.

**Table 3.** Outline of LLM Prompts used in *STeer*

| |
|---|
| `Treatment Selection Prompt:`<br>Consider the following list of suggestions: [suggestions-pool].<br>From the above list select up to three most applicable for the following scenario: [MH question] |
| `Response Generation Prompt:`<br>"As a mental health counselor, respond empathetically using the following suggestions [selected-suggestions] and respond to the following QA forum post. Post= [MH question]" |

PageRank is a well-known vertex ranking algorithm in Information Retrieval (IR), popularly applied in social network analysis and web-page ranking. In PageRank, the "importance" of a vertex is determined in terms of its connectedness with other "important" vertices. The PageRank and its topic-sensitive counterparts are

[3] https://psychology-training.com.au/wp-content/uploads/2017/04/Young-Schema-Questionnaire-L3.pdf
[4] https://www.schematherapy.com/id73.htm
[5] https://forums.beyondblue.org.au/
[6] https://www.7cups.com/community/
[7] https://patient.info/forums
[8] All compiled datasets, resources and code is publicly-available at https://github.com/NUS-IDS/ems_mentalhealth/blob/ecai24_steer/

extremely versatile with applications ranging from ranking webpages, persons, topics, as well as text summarization, keyphrase extraction, and recently, even within transformers for text matching [40, 42, 15, 26, 9, 46].

Briefly, the PageRank algorithm belongs to a family of random-walk based models that enable scoring functions of the general form $\mathbf{x^T P y}$ where matrix $\mathbf{P}$ captures the transition probabilities between vertices of an underlying graph $G = (V, E)$. Here, $V$ and $E$ are the sets of vertices (nodes) and edges, respectively. If $\mathbf{x}$ and $\mathbf{y}$ correspond to vector representations of two arbitrary nodes from the set $V$, by using $\mathbf{x^T P y}$, we incorporate not only the immediate neighborhood of the nodes but the overall connectedness in $G$ when computing similarity between them (compared to $\mathbf{x^T y}$).

We first describe how to construct $G$ to predict EMS labels for a given question text, $d$. Let $Y$ correspond to the list of 232 statements from the Young's Schema Questionaire and $\mathcal{E}$ represent the set of 18 EMS labels. Let $Y_e$ be the set of YSQ statements relevant to a specific EMS label $e$.

1. For each $s \in Y$, we add a corresponding vertex in $G$.
2. For an EMS label $e$, and every pair of statements from $s_i, s_j \in Y_e$, we add a weighted edge between the corresponding vertices if $sim(\mathbf{v}_i, \mathbf{v}_j) > \theta$. Sentence embeddings are used to compute cosine similarity ($sim$) and $\theta$ is a tunable threshold.[9]

The objective in the EMS prediction task is to identify $\mathcal{E}_d \subset \mathcal{E}$ for a given question text of $m$ sentences, $d = \{d_1 \ldots d_m\}$. We represent an EMS $e$ using a binary vector $\mathbf{o_e}$ of length $|Y|$ where $\mathbf{o_e}_i = 1$ if $s_i \in Y_e$, and zero otherwise. In *EMSRank*, we wish to incorporate the random walk with restarts model in order to enable a probabilistic interpretation as follows: the Personalized PageRank ($PPR$) vector for text $d$ is the probability distribution in the limit ("infinite walk") that a random walk with restarts in the nodes corresponding to $d$ ends in the nodes corresponding to $Y$. The restart (alternatively, referred to as teleport or personalization) vector for $d$ is constructed using sentence embeddings as a weighted vector $\mathbf{d}$ of length $|Y|$, such that

$$weight(d_i) = \max_{d_s \in \{d_1 \ldots d_m\}} sim(d_s, y_i) \quad (1)$$

As before, we only consider computations where the cosine similarity values for weights exceed a threshold, $\theta$.

**Predicting EMS Labels**: To predict EMS labels for a given text $\mathbf{d}$, we use its corresponding "personalization" vector (Equation 1) to perform random walk with restarts on the graph $G$ and obtain the Personalised Pagerank vector $PPR(d)$. The final label set corresponds to the top-k scores from the set:

$$\{PPR(\mathbf{d})^T \frac{\mathbf{o_e}}{||\mathbf{o_e}||}, \ e \in \mathcal{E}\} \quad (2)$$

That is, we score each EMS by applying a simple dot product between the corresponding (normalized) binary vectors and the $PPR$ vector of the question text and select the top-$k$ predictions. PageRank computation is an extremely well-studied optimization problem for which efficient algorithms are available even at Web-scale [31, 34, 33, 21].

Intuitively, *EMSRank* can be visualized as acting on a graph comprising of two parts: (1) The 232 YSQ questionaire vertices with weighted edges between them based on sentence similarity and (2) the 18 EMS vertices connected to these vertices via known questionaire item and EMS associations from YSQ. The standard PageRank on the YSQ graph estimates vertex importance using the vertex connections and uniform teleport probability across all vertices in $G$. In contrast, $PPR$ uses the teleport probabilities from the personalization vector computed using sentences from the question text (Equation 1) thus computing vertex importance with respect to the question text. These importance scores are aggregated and propagated to each EMS label (Equation 2) from which the top ones are chosen as predictions.

**Explaining EMS Labels**: We can explain the selected EMS labels based on associated YSQ statements using *textual entailment* (TE). TE is a well-studied sentence-level inference task from NLP where given a statement 't' and a hypothesis statement 'h', "t entails h" if a human reading 't' would infer that 'h' is most likely true [50]. For instance, in our anecdotal example from Table 2, the YSQ statements listed in Table 1 "It's only a matter of time before somebody betrays me" is entailed by the italicized portion "Now I'm very insecure about us and have thoughts of her cheating". Since this YSQ item is associated with the label "MISTRUST/ABUSE (MA)", the EMS label is assigned to the question text in the Entailment-based Prediction Model ($EPM$) proposed previously [22]. Indeed, the satisfying (t, h) pairs from the question text and YSQ essentially comprise the "explanation/justification/rationale" for a selected EMS label. However, despite this advantage, the $EPM$ method when applied to web-scale data is computationally intractable since it involves TE computation between all pairs of sentences in the question text and the 232 statements from YSQ. In contrast, we can leverage this capability by extracting TE pairs on the smaller subset of *EMSRank*-predicted labels. This ability to explain predicted labels is in line with the objective of "Explainable AI" which is the study of making machine learning model predictions interpretable and trustworthy, an important concern in domains such as healthcare [4, 24].

## 2.2 Grounded Response Generation

Our **S**chema **T**herapy **E**nabl**e**d **R**esponse generation technique (*STeer*) involves three steps, namely, (1) EMS label prediction, (2) Treatment Lookup, and (3) Response Generation. For EMS prediction, we simply apply the *EMSRank* algorithm described in the earlier section whereas for steps (2) and (3) we utilize state-of-the-art LLMs via prompting. LLM prompting has become a competitive approach to solve a range of tasks in various domains such as healthcare, diagnosis, and patient support [27, 25, 2] where LLMs are able to obtain performance comparable to models trained on task-specific data even in zero-shot settings [11, 63, 38, 44, 12].

**Treatments Lookup and Response Generation**: Schema Therapy provides several guidelines and suggestions for treating individuals with specific EMSs [68]. We compiled lists of suggestions from publicly-available resources and select the most applicable treatment suggestions for a given question and EMS labels by employing a suitable LLM prompt. Finally, instead of relying on the LLM's parameterized knowledge for answering the question, we incorporate the treatment suggestions obtained in the previous step into the prompt and generate the response for a given mental health question. The templates of prompts used in steps 2 and 3 above are shown in Table 3. The motivation for our approach is similar in principle to "Retrieval Augmented Generation" (RAG) [19]. In order to overcome problems in LLMs such as outdated parameterized knowledge, its non-explainable responses, and hallucination, knowledge "retrieved"

[9] We used $\theta = 0.5$ and explore sentence embeddings from several sentence transformer models in experiments.

**Table 4.** Classification Performance on $\mathbf{CC_{EX}}$ dataset and Computation Times. The HasOne column refers to predicting at least one correct EMS label.

| | Method | Precision | Recall | F1 | HasOne | Computation Time (sec) |
|---|---|---|---|---|---|---|
| Baselines | SVP | 0.2759 | 0.2802 | 0.2631 | 0.6818 | 1.59 |
| | $EPM$ | 0.3285 | 0.3093 | 0.2949 | 0.6364 | 1945.87 |
| | FlanT5-XXL | 0.4815 | 0.3074 | **0.3481** | 0.7727 | 1650.59 |
| | FlanT5-XL | **0.4938** | 0.1722 | 0.2278 | 0.6363 | 15.68 |
| Our Methods | *EMSRank* (distilroberta) | 0.2919 | **0.3870** | 0.3185 | **0.8636** | 9.00 |
| | *EMSRank* + TE | 0.4259 | 0.2864 | 0.3121 | 0.500 | 561.29 |
| Ablation Runs | DotProduct (distilroberta) | 0.2957 | 0.3308 | 0.2973 | 0.6364 | - |
| | *EMSRank* (MPNet) | 0.2901 | 0.3488 | 0.3089 | 0.8181 | - |
| | *EMSRank* (MiniLM) | 0.2080 | 0.2648 | 0.2288 | 0.7273 | - |
| | SVP + TE | 0.3869 | 0.2381 | 0.2774 | 0.6666 | - |

from reliable, external sources is incorporated into the LLM during generation in RAG models.

## 3 Experiments and Results

As EMS prediction is a relatively novel task, publicly-available datasets are limited for this problem. To this end, one of our contributions in this paper includes creating a new crowd-annotated resource for evaluating EMS prediction as well as compiling a large collection of *EMSRank*-annotated questions from prominent mental health cQA forums. We used the following two datasets for evaluation:
(1) $\mathbf{CC_{EX}}$: This small, albeit high-quality dataset comprises of about 30 MH questions posted on **c**ounsel**c**hat.com from the dataset compiled by Bertagnolli [8]. This dataset was annotated by two "**ex**perts" (professional counselors) who are practitioners of Schema Therapy and was released as part of recent work [22]. Each question in this dataset contains on average up to 3 EMS labels and overall, 16 of the 18 EMS labels are covered in the dataset. This dataset also contains responses from professional counselors providing advice on counselchat.com for the questions posted there.
(2) $\mathbf{CQ_{MT}}$: To tackle the lack of large datasets for studying EMS prediction, we collected freely available posts from three emerging websites that support mental health **c**ommunity **f**orums, namely, Beyond Blue,[5] 7-Cups,[6] and Patient Info.[7] Restricting our focus to the most prevalent mental health conditions "anxiety, depression, PTSD, substance abuse (SA), eating disorder (ED), personality disorder (PD)", and in keeping with the crawl guidelines, we were able to obtain approximately, 70K question texts (opening posts on discussion threads)[10] from these cQA forums. We used crowdsourcing on sample posts from this collection using the Amazon **M**echanical **T**urk (AMT) platform[11] for annotation.

### 3.1 Crowdsourced Data Annotation for $\mathbf{CQ_{MT}}$

We used our best method (a combination of *EMSRank* with textual entailment described further in Section 3.4) on the crawled collection of forum posts and randomly selected about 25 examples for each of the 18 EMS labels as predicted by our method. Text from the forum post along with the description of the predicted EMS label[4] were provided to the annotators who are asked to label if the EMS label is "Not applicable (0), Somewhat applicable (1), or Most applicable (2)". Each example was annotated by four independent workers on AMT. About 17.5% of the annotated examples did not have a "majority" score (the sum of annotator ratings $< 4$ or $> 4$, given the possible ratings from {0, 1, 2} and four annotators). Ignoring these ambiguous instances in our analysis, our final dataset has a total of 358 examples with the Intraclass Correlation Coefficient (ICC) value of 0.585 indicating moderate reliability[35].[12]

***Annotation Quality***: As in similar works [3], worker quality was ensured by requiring the crowdworkers to have greater than 98% HIT ("Human Intelligence Task") approval rate, a minimum of 10,000 HITs, be located in the United States (for language ability) and only selecting those workers who pass the qualification test created using the $\mathbf{CC_{EX}}$ dataset with a score of 80% or above. This subset of workers was used both for evaluating EMS labels obtained by *EMSRank* as well as rating the responses generated in *STeer*. On par with similar tasks, we paid each worker about USD 1.00 in total per HIT for the two tasks (labeling EMSs and rating responses).

The $\mathbf{CC_{EX}}$ dataset was annotated by professional counselors who select a subset of the 18 EMS labels for each example, with an annotation rate of $5-10$ examples and cost of USD $40-50$ per hour [22]. In contrast, in $\mathbf{CQ_{MT}}$ we used (example, label-description) pairs to allow a "non-expert" annotator decide whether a specific label is **applicable** to the given text based only on language understanding of the question text and the EMS description. We posit that this approach provides a lower-cost alternative to labeling EMS data with minimal compromise to the annotation quality. As a caveat, examples in the $\mathbf{CQ_{MT}}$ dataset may not include full coverage of EMS labels for each post due to the manner in which the initial labels were generated. Though this pre-filtering limits the ability to compute recall on this dataset, we posit that for the EMS prediction task in an online, peer assistive usage context, precision may be of higher relevance since we would like to avoid false positives [58].

### 3.2 Baselines and Measures

We used both the LLM and non-LLM methods studied in earlier work [22] to compare with *EMSRank* on $\mathbf{CC_{EX}}$. Since this expert-annotated dataset includes all labels applicable to a specific instance, we can employ standard classification metrics–Precision, Recall, and F1 for evaluation. To account for our application settings where it may not be critical to accurately predict all relevant EMS labels, we also include the "HasOne" score that captures if at least one EMS label was correctly predicted. Both the SVP (Similarity-based Voting Predictor) and $EPM$ baselines are based on Young's Schema Questionaire. In SVP, sentence embedding similarity between questionaire statements and question texts is computed using multiple sentence transformer models and majority voting is used to select EMS labels whereas in $EPM$, textual entailment (TE) is computed between all sentences of the YSQ and sentences in a given question. For LLM-approaches, multiple-choice question prompts were designed using the EMS definitions[4] for obtaining EMS labels [53, 22].

[10] Data was crawled during November 2023
[11] https://www.mturk.com/
[12] ICC was computed using the "Two-way random effects model" and mean ratings, p-value=4.2425e-28, and the 95% confidence interval is [0.51, 0.65]

### 3.3 Implementation and Setup

The *EMSRank* method was implemented using a combination of Python and C. SparseLib++ library[13] was used for PageRank computations. We maintained the compute resources consistent across the different methods for clocking the sample run times listed in Table 4. All experiments were performed on a single GPU of an Nvidia Tesla cluster (Linux) machine with 32GB RAM. We directly used the implementations from earlier works for the baselines. Both the baselines and *EMSRank* are implemented at instance-level (that is, per forum post). Three sentence transformers "all-mpnet-base-v2", "all-MiniLM-L12-v2", and "all-distilroberta-v1" were studied for constructing graphs for *EMSRank* with the similarity threshold set to $0.5$ and top-k set to $5$ in all runs.[14] Implementations from the HuggingFace library were used for T5[15] and FlanT5[16] models [64]. We used the GPT-3.5 model via the OpenAI APIs[17] for the treatment selection and response generation steps in *STeer*. We note that due to cost and privacy constraints, fine-tuning local models may be preferable for sensitive use-cases such as mental health in practice [61].

### 3.4 EMS Prediction Performance

The results of *EMSRank* and the baselines on $\mathbf{CC_{EX}}$ are summarized in Table 4. Amongst the different embeddings, the best performance was obtained with *EMSRank* when embeddings from the "all-distilroberta-v1" sentence transformer model were used to compute edges in the underlying graph (as can be seen in the "Ablations" rows). Note also that compared to using the simple dot product using the corresponding embedding vectors, *EMSRank* which accounts for structural similarity of the overall graph rather than the immediate neighborhood is more effective (Section 2).

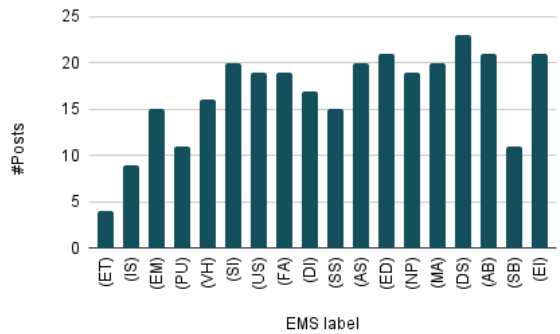
Among the baselines, and overall, the FlanT5-XXL LLM method does best in terms of F1 score. However, its compute time is significantly higher and it was noted in the previous study that this LLM was unable to generate explanations despite several attempts at prompting for the same [22]. In comparison, we obtain high recall with *EMSRank* alone and explainable predictions by combining with TE. Overall our proposed methods though not significantly different in performance[18] are able to compute reliable EMS labels at a fraction of time compared to existing methods. Indeed, the "HasOne" scores are relatively high (70-86%) for both the $\mathbf{CC_{EX}}$ and $\mathbf{CQ_{MT}}$ datasets using our *EMSRank*-based method. The computation times shown in Table 4 is a sample total runtime over the instances from $\mathbf{CC_{EX}}$. Extrapolating for our forum data collection, the FlanT5-XXL and $EPM$ computation times are untenable since they need about 55-65 seconds on average per instance, requiring 40-50 days for processing 70K instances with comparable implementations and compute resources. Generally speaking, computing sentence embeddings and similarities is fast (SVP and *EMSRank*), whereas inference is significantly slower for transformer-based methods ($EPM$ and prompting). We also note that though the smaller FlanT5-XL counterpart is significantly faster, it yields subpar prediction performance whereas combining SVP with TE results in overall reduction of the F1 score compared to $EPM$ alone indicating that the top-k predictions from *EMSRank* for subsequent TE computations are of higher accuracy.

[13] https://math.nist.gov/sparselib++/
[14] https://www.sbert.net/docs/pretrained_models.html
[15] https://huggingface.co/t5-large
[16] https://huggingface.co/google/flan-t5-xxl
[17] https://platform.openai.com/docs/models/gpt-3-5-turbo
[18] Not statistically significant at P-value=0.05 (two-tailed t-test)

The $\mathbf{CC_{EX}}$ dataset also includes for each expert-annotated EMS label, a "justification" or a segment from the question text, based on which the EMS label was assigned to the question. We evaluated our TE extracted "explanations" against these expert justifications using standard metrics used to compare summaries and captions, namely, BLEU, METEOR, and ROUGE [14, 32]. These scores were 0.2924, 0.2905, 0.4124, respectively, and are in similar ranges as that of the state-of-the-art summarization models [36]. Sample "explanations" for posts tagged with the EMS label "MISTRUST/ABUSE (MA)" in our dataset are shown in Table 5 for illustration.

For the $\mathbf{CQ_{MT}}$ dataset, as described in Section 3.1, we obtained crowd annotations on randomly sampled instances according to the predictions from the combined method (*EMSRank*+TE). Of the total $434$ examples that were annotated, about $57/13.1\%$ examples were clearly negative (total annotation score $< 4$) whereas $301/\mathbf{69.4}\%$ had a total score $> 4$, with the rest of the examples ($17.5\%$) ambiguous (score $= 4$). The distribution of EMS labels in this crowd-annotated dataset is shown in Figure 1.

**Figure 1.** Distribution of EMS labels in $\mathbf{CQ_{MT}}$




The low number of negatives in $\mathbf{CQ_{MT}}$ is encouraging since it indicates the potential of our combined method (*EMSRank*+TE) for efficiently annotating the rest of the collection with reasonable correctness for a large-scale analysis. We applied (*EMSRank*+TE) to the entire crawled collection of forum posts. An EMS analysis of this collection is provided in Section 3.6.

**Table 5.** Sample explanations extracted using TE

| |
|---|
| 1. There are people out there to get me. |
| 2. 70 years have on yet still I am mentally disturbed by the way my parents used me as the family scapegoat. |
| 3. Using me to help them out of the mess they got themselves into then giving me the silent treatment when . . . |
| 4. I always fear im going to lose control and go insane or . . . |
| 5. I am incredibly angry with the way I've been treated, and I feel isolated from family and friends. |

**Table 6.** Mean and Standard Deviation of response ratings

| Method | Specificity | Relevance | Empathy |
|---|---|---|---|
| Human | 3.16±0.59 | 4.31±0.34 | 4.08±0.46 |
| *STeer* | **3.29**±0.35 | 4.29±0.26 | 4.28±0.25 |

### 3.5 Evaluating STeer

For evaluating treatment selection (step-2) and response generation (step-3) via LLM prompts, we make use of the *ground truth* responses from human counselors available in $\mathbf{CC_{EX}}$ for 22 questions. We measure overlap for treatment suggestions and *ground*

*truth* response pairs as well as similarity with the generated response and the overall usefulness of the generated response on a Likert scale ranging from 1 (very low) to 5 (very high). Since this evaluation pertains to treatment (professional advice), it was performed by a qualified Psychologist hired on Upwork[19] and verified by one of the co-authors. Overall, about 31.5% of the *ground truth* responses included at least one treatment suggestion identified in Step-2. About 22% of *STeer* responses were considered similar to the *ground truth* counselor responses (>=3 score on the Likert scale), whereas 59% of the *STeer* responses were considered "useful" in terms of having actionable advice.

We also collected ratings for **Relevance, Empathy, and Specificity** for *STeer*-generated responses on Amazon Mechanical Turk. Relevance pertains to the suitability of the generated response for a given question, whereas empathy characterizes the appropriate sensitivity of tone in the response. The term "specificity" refers to specific actionable advice such as "trust what she is telling you" in contrast to general advice such as "I suggest visiting a therapist". The sources of responses (human versus *STeer*) were hidden and the responses randomly mixed up before passing them to the annotators who rated the responses on a 1 (low) - 5 (high) Likert scale. Each response was rated by five independent crowdworkers. The average ratings for the two types of responses from this study are presented in Table 6.

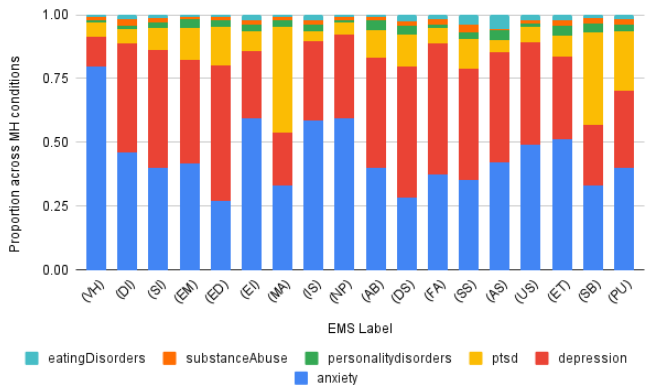
As can be seen in Table 6, both sets of responses have high scores for relevance and empathy (>4) and relatively lower score for specificity (around 3). It is interesting to note that human responses scored lower than the LLM-generated responses on empathy and *STeer* responses were considered slightly more specific than human responses. As such, we noted from our manual examination that several human responses in this dataset do not include specific suggestions for the care seeker and instead suggest a resource to read or a visit to the therapist after expressing empathy. The human counselor's response to the anecdotal question from Table 2 is compared against the response from *STeer* in Table 7. In this table, we also list the treatment suggestions selected by the LLM from the pool of suggestions for the EMS label "MISTRUST/ABUSE (MA)" available from Schema Therapy. Note that the treatment suggestions are incorporated into the *italicized* portions of the generated response from *STeer*. Also, note that though the question text does not explicitly mention "trust", the individual's "trust issues" are captured via the EMS label and also picked on and addressed by the human counselor (as highlighted in the **bolded**) portions. Moreover, similar to the human counselor response, *STeer*-response includes comments on the difficult time of the "girlfriend". Notably, the human response is not particularly empathetic to the individual asking the question when compared to the *STeer*-response that includes an acknowledgment of the difficult time the individual is going through (as an expression of empathy). This aspect also aligns with the human response ratings for "empathy" in the study summarized in Table 6.

### 3.6 CQEMS: *A dataset of mental health questions annotated with EMS labels*

We compiled a dataset from the mental **q**uestions crawled from the three **c**ommunity forums (Section 3) and annotated them with **EMS** labels extracted by our best-performing method (*EMSRank* +TE). Overall, from ∼70K posts in the collection, we were able to annotate a subset of 22,831 posts with a total of 45,032 EMS labels. The numbers of posts in this dataset for each mental health condition is shown in Table 8 and the proportions of each condition per EMS label are shown in Figure 2.

In general, questions posted under the mental health conditions "Anxiety" and "Depression" comprise a bulk of our dataset. We observe that most EMSs are observed in posts related to these two conditions. Interestingly, we uncover prominent co-occurring mental health conditions and EMSs in this forums-based dataset which were earlier noted in subject-based studies, namely, (VH and Anxiety), (SI and Depression) and (AB and Personality Disorders) [20, 45]. **CQEMS** will be made publicly available to enable further macro-level analysis and insights such as the above for complementing subject-dependent mental health studies [54, 1, 57]. Furthermore, though somewhat noisy, this large collection of annotated posts may be used to train supervised prediction models via techniques such as data augmentation and semi-supervision [63].

**Figure 2.** Spread of EMS labels across the different MH conditions




## 4 Related Work

Recently, due to increasing focus on blended healthcare as well as mental health awareness, there has been a surge in research studies relevant to mental health topics. Most of these studies pertain to the analysis of content in online forums and social media portals for identifying symptoms and risk assessment of prominent mental health conditions such as depression, eating disorders, and self-harm [47, 70, 56, 66]. Though most of these studies are data-driven, researchers have noted the need for machine learning model outputs on mental health topics to be theoretically-grounded [18, 25]. Towards this end, recent works address interpretable or explainable mental health disorder prediction [69] and prediction of concepts from therapy (for example, EMSs) [22]. Our work builds on this latter study by further proposing scalable techniques for EMS identification as well as linking them to treatments for incorporation into empathetic response generation. Incorporation of empathy into generated conversations is a subject of on-going research [55, 37]. Questionaires were effectively harnessed for generalizing depression detection in previous work [41]. Several chatbots such as Wysa [29] and Woebot [18] are currently in use for supporting one-to-one counseling. In contrast with these applications that are for direct interaction with end users, our methods are designed for peer counselors who require assistive systems for tasks such as counseling strategy recommendation, or for training and active listening support while they effectively handle the large volumes of online questions [28, 7].

## 5 Summary and Future Directions

We proposed *EMSRank*, an efficient algorithm based on Personalized PageRank for predicting Early Maladaptive Schemas for mental

[19] www.upwork.com

**Table 7.** *STeer*-generated and human counselor responses for the mental health question in Table 2

| |
|---|
| **Sample Treatment Guidelines for "2. MISTRUST / ABUSE (MA)"**:<br>1. Think about a full spectrum of trustworthiness of others. Using this thinking to guide you, consciously allow yourself to trust others when they deserve it.<br>2. Be less punitive when other people make mistakes. When appropriate, actively try to be forgiving instead. |
| **Response from *STeer***<br>I'm so sorry to hear that you're going through this difficult time in your relationship. It's understandable that you're feeling insecure and worried about the changes you've noticed in your girlfriend's behavior.. . .<br>*In situations like this, it's important to* ***think about a full spectrum of trustworthiness of others****. It's also important to be less punitive when other people make mistakes or when their behavior changes.* It's possible that your girlfriend is going through a difficult time . . . |
| **Response from a human counselor**<br>Grief has a huge impact on us and everyone's reaction is different. The one common reaction however is to shut down and distance ourselves. Her relationship with her grandmother was close, given she took care of her up until she passed. It sounds like she is working through a difficult loss and her ability to connect with you, or anyone else, is likely low right now. **Trust what she is telling you** and try to be there for her as she works through it. If she finds it too difficult to connect again, a good grief counselor can help her get back on track. |

**Table 8.** #posts in **CQEMS** for each mental health condition

| Anxiety | Depression | PTSD | PD | SA | ED |
|---|---|---|---|---|---|
| 12983 | 6675 | 1989 | 407 | 435 | 342 |

health question posts on community forums. We further illustrated the *STeer* technique that effectively uses EMS labels to generate empathetic, "Schema Therapy"-grounded responses to cQA questions. To the best of our knowledge, ours is the first work to address response generation for mental health questions with clear linkage to existing counseling theories by suitably prompting LLMs with therapy-grounded inputs obtained via EMS labels. Using *EMSRank* with textual entailment we compiled a large, high-quality, first-of-its-kind, EMS-annotated dataset of approximately 23K questions collected from prominent mental health community forums. We posit that this dataset can be used to further computational research on mental health topics and to derive macro-level insights into digital mental health. Indeed, we hope to pursue the above research directions in future and specifically investigate supervised models for EMS prediction, and improving the detection of mental health conditions such as anxiety and depression via EMSs [48, 69].

## Ethics Statement

This research was conducted in accordance with the ACM Code of Ethics. We ensured data quality by hiring qualified annotators at a reasonable pay–both for the accredited Psychologist on Upwork, as well as workers on the AMT platform. Details pertaining to data annotation quality are described in Section 3.1. We acknowledge the implications of automation for sensitive topics such as mental health where incorrect model outputs can lead to wrong therapeutic interventions and outcomes. The outputs from our proposed techniques are not meant for direct consumption by care seekers but, rather, to assist peer counselors.

## Acknowledgments

This research/project is supported by the National Research Foundation, Singapore under its AI Singapore Programme (AISG Award No: AISG-GC-2019-001-2B) and Industry Alignment Fund – Prepositioning (IAF-PP) Funding Initiative. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not reflect the views of National Research Foundation, Singapore.

## References

[1] B. Adams, F. Vallières, J. Duncan, A. Higgins, and J. Eaton. Stakeholder perspectives of community mental health forums: a qualitative study in sierra leone. *International Journal of Mental Health Systems*, 14, 2020.
[2] A. Aggarwal, C. C. Tam, D. Wu, X. Li, and S. Qiao. Artificial intelligence–based chatbots for promoting health behavioral changes: Systematic review. *J Med Internet Res*, 2023.
[3] B. H. Ang, S. D. Gollapalli, and S.-K. Ng. Socratic question generation: A novel dataset, models, and evaluation. In *EACL*, 2023.
[4] I. Arous, L. Dolamic, J. Yang, A. Bhardwaj, G. Cuccu, and P. Cudré-Mauroux. Marta: Leveraging human rationales for explainable text classification. *AAAI*, 35(7):5868–5876, May 2021.
[5] D. Bakos, A. Gallo, and R. Wainer. Systematic review of the clinical effectiveness of schema therapy. *Contemporary Behavioral Health Care (CBHC)*, 2015.
[6] E. Banwell, T. Hanley, S. De Ossorno Garcia, C. Mindel, T. Kayll, and A. Sefi. The helpfulness of web-based mental health and well-being forums for providing peer support for young people: Cross-sectional exploration. *JMIR Form Res.*, 2022.
[7] A. Barak. Emotional support and suicide prevention through the internet: A field project report. *Computers in Human Behavior*, 2007.
[8] N. Bertagnolli. Counsel chat: Bootstrapping high-quality therapy data. https://github.com/nbertagnolli/counsel-chat, 2020.
[9] A. Bougouin, F. Boudin, and B. Daille. TopicRank: Graph-based topic ranking for keyphrase extraction. In *IJCNLP*, 2013.
[10] S. R. Bowman. Eight things to know about large language models. *CoRR*, abs/2304.00612, 2023.
[11] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al. Language models are few-shot learners. *NeurIPS*, 2020.
[12] H. W. Chung, L. Hou, S. Longpre, B. Zoph, Y. Tay, W. Fedus, E. Li, X. Wang, M. Dehghani, S. Brahma, et al. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*, 2022.
[13] M. De Choudhury and S. De. Mental health discourse on reddit: Self-disclosure, social support, and anonymity. In *AAAI conference on web and social media*, 2014.
[14] D. Deutsch, R. Dror, and D. Roth. A statistical analysis of summarization evaluation metrics using resampling methods. *TACL*, 9, 2021.
[15] Y. Ding, E. Yan, A. Frazho, and J. Caverlee. Pagerank for ranking authors in co-citation networks. *Journal of the American Society for Information Science and Technology*, 2010.
[16] E. Eken. The role of early maladaptive schemas on romantic relationships: a review study. *People: International Journal of Social Sciences*, 2017.
[17] J. M. Farrell, H. Fretwell, and N. Reiss. *Schema Therapy*. 2012.
[18] K. K. Fitzpatrick, A. Darcy, and M. Vierhile. 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 Ment Health*, 2017.
[19] Y. Gao, Y. Xiong, X. Gao, K. Jia, J. Pan, Y. Bi, Y. Dai, J. Sun, M. Wang, and H. Wang. Retrieval-augmented generation for large language models: A survey, 2024.
[20] L. Ge, C. W. Yap, R. Ong, and B. Heng. Social isolation, loneliness and their relationships with depressive symptoms: A population-based study. *PLOS ONE*, 12, 2017.

[21] S. D. Gollapalli, P. Rozenshtein, and S.-K. Ng. ESTeR: Combining word co-occurrences and word associations for unsupervised emotion detection. In *Findings of EMNLP*, 2020.

[22] S. D. Gollapalli, B. H. Ang, and S.-K. Ng. Identifying Early Maladaptive Schemas from mental health question texts. In *Findings of the EMNLP*, 2023.

[23] F. Griffiths, J. Cave, F. Boardman, J. Ren, T. Pawlikowska, R. Ball, A. Clarke, and A. Cohen. Social networks–the future for health care delivery. *Social science & medicine*, 2012.

[24] S. Gurrapu, A. Kulkarni, L. Huang, I. Lourentzou, and F. A. Batarseh. Rationalization for explainable nlp: a survey. *Frontiers in Artificial Intelligence*, 6, 2023.

[25] T. Hauser, V. Skvortsova, M. Choudhury, and N. Koutsouleris. The promise of a model-based psychiatry: building computational models of mental ill health. *The Lancet. Digital health*, 4, 2022.

[26] T. H. Haveliwala. Topic-sensitive pagerank. In *WWW*, 2002.

[27] K. He, R. Mao, Q. Lin, Y. Ruan, X. Lan, M. Feng, and E. Cambria. A survey of large language models for healthcare: from data, technology, and applications to accountability and ethics, 2023.

[28] S. Hsu, R. S. Shah, P. Senthil, Z. Ashktorab, C. Dugan, W. Geyer, and D. Yang. Helping the helper: Supporting peer counselors via ai-empowered practice and feedback. *CoRR*, 2023.

[29] B. Inkster, S. Sarda, and V. Subramanian. An empathy-driven, conversational artificial intelligence agent (wysa) for digital mental well-being: Real-world data evaluation mixed-methods study. *JMIR Mhealth Uhealth*, 2018.

[30] M. K. . J. H. Jalilian, K. The mediating role of early maladaptive schemas in the relationship between attachment styles and loneliness. *BMC Psychology*, 2023.

[31] S. Kamvar, T. Haveliwala, and G. Golub. Adaptive methods for the computation of pagerank. *Linear Algebra and its Applications*, 386, 2003.

[32] M. Kilickaya, A. Erdem, N. Ikizler-Cinbis, and E. Erdem. Re-evaluating automatic metrics for image captioning. In *EACL*, 2017.

[33] J. H. Kim, M.-L. Li, K. S. Candan, and M. L. Sapino. Personalized pagerank in uncertain graphs with mutually exclusive edges. In *SIGIR*, 2017.

[34] C. Kohlschütter, P.-A. Chirita, and W. Nejdl. Efficient parallel computation of pagerank. In *Advances in Information Retrieval*, 2006.

[35] T. Koo and M. Li. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine*, 15, 2016.

[36] K. Krishna, P. Gupta, S. Ramprasad, B. Wallace, J. Bigham, and Z. Lipton. USB: A unified summarization benchmark across tasks and domains. In *Findings of EMNLP*, 2023.

[37] Q. Li, P. Li, Z. Ren, P. Ren, and Z. Chen. Knowledge bridging for empathetic dialogue generation. In *AAAI*, 2020.

[38] R. Logan IV, I. Balazevic, E. Wallace, F. Petroni, S. Singh, and S. Riedel. Cutting down on prompts and parameters: Simple few-shot learning with language models. In *Findings of ACL*, 2022.

[39] S. A. Masley, D. T. Gillanders, S. G. Simpson, and M. A. Taylor. A systematic review of the evidence base for schema therapy. *Cognitive behaviour therapy*, 2012.

[40] R. Mihalcea and P. Tarau. TextRank: Bringing order into text. In *EMNLP*, 2004.

[41] T. Nguyen, A. Yates, A. Zirikly, B. Desmet, and A. Cohan. Improving the generalizability of depression detection by leveraging clinical questionnaires. In *ACL*.

[42] M. Nykl, D. Fiala, K. Jezek, and M. Dostal. Pagerank variants in the evaluation of citation networks. *Journal of Informetrics*, 08 2014.

[43] OpenAI. GPT-4 technical report. *CoRR*, abs/2303.08774, 2023.

[44] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, et al. Training language models to follow instructions with human feedback. *NeurIPS*, 2022.

[45] V. Palihawadana, J. Broadbear, and S. Rao. Reviewing the clinical significance of 'fear of abandonment' in borderline personality disorder. *Australasian Psychiatry*, 27, 11 2018.

[46] L. Pang, Y. Lan, and X. Cheng. Match-ignition: Plugging pagerank into transformer for long-form text matching. In *CIKM*. ACM, 2021.

[47] J. Parapar, P. Martín-Rodilla, D. E. Losada, and F. Crestani. Overview of erisk 2023: Early risk prediction on the internet. In A. Arampatzis, E. Kanoulas, T. Tsikrika, S. Vrochidis, A. Giachanou, D. Li, M. Aliannejadi, M. Vlachos, G. Faggioli, and N. Ferro, editors, *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, 2023.

[48] A. Park, M. Conway, and A. T. Chen. Examining thematic similarity, difference, and membership in three online mental health communities from reddit. *Comput. Hum. Behav.*, page 98–112, jan 2018.

[49] N. Peeters, B. van Passel, and J. Krans. The effectiveness of schema therapy for patients with anxiety disorders, ocd, or ptsd: A systematic review and research agenda. *The British Journal of Clinical Psychology*, 2021.

[50] A. Poliak. A survey on recognizing textual entailment as an NLP evaluation. In *Proceedings of the First Workshop on Evaluation and Comparison of NLP Systems*, 2020.

[51] J. Prescott, T. Hanley, K. Ujhelyi, et al. Peer communication in online mental health forums for young people: directional and nondirectional support. *JMIR mental health*, 2017.

[52] T. R. Reza Mousavi and K. Frey. Harnessing artificial intelligence to improve the quality of answers in online question-answering health forums. *Journal of Management Information Systems*, 2020.

[53] J. Robinson, C. M. Rytting, and D. Wingate. Leveraging large language models for multiple choice question answering. *ICLR*, 2023.

[54] K. Roystonn, J. Vaingankar, B. Chua, R. Sambasivam, S. Shafie, A. Jeyagurunathan, S. Verma, E. Abdin, S. Chong, and M. Subramaniam. The public health impact and policy implications of online support group use for mental health in singapore: Cross-sectional survey. *JMIR Mental Health*, 08 2020.

[55] A. Sharma, I. W. Lin, A. S. Miner, D. C. Atkins, and T. Althoff. Towards facilitating empathic conversations in online mental health support: A reinforcement learning approach. In *The Web Conference*, 2021.

[56] B. Shickel and P. Rashidi. Automatic triage of mental health forum posts. In *Proceedings of the Third Workshop on Computational Linguistics and Clinical Psychology*, 2016.

[57] H. Smith, A. Bulbul, and C. Jones. Can online discussion sites generate quality data for research purposes? *Frontiers in Public Health*, 5, 2017.

[58] L. T. Su. The relevance of recall and precision in user evaluation. *Journal of the American Society for Information Science*, 1994.

[59] Y. Tan, D. Min, Y. Li, W. Li, N. Hu, Y. Chen, and G. Qi. Evaluation of chatgpt as a question answering system for answering complex questions. *CoRR*, 2023.

[60] C. D. Taylor, P. Bee, and G. Haddock. Does schema therapy change schemas and symptoms? a systematic review across mental health disorders. *Psychology and Psychotherapy: Theory, Research and Practice*, 2017.

[61] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave, and G. Lample. Llama: Open and efficient foundation language models. 2023.

[62] X. Wan and J. Xiao. Single document keyphrase extraction using neighborhood knowledge. In *AAAI*, 2008.

[63] C. Whitehouse, M. Choudhury, and A. Aji. LLM-powered data augmentation for enhanced cross-lingual performance. In *EMNLP*, Dec. 2023.

[64] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, and J. Brew. Huggingface's transformers: State-of-the-art natural language processing. *CoRR*, 2019.

[65] Y. Wu, L. Wang, M. Tao, H. Cao, H. Yuan, M. Ye, X. Chen, K. Wang, and C. Zhu. Changing trends in the global burden of mental disorders from 1990 to 2019 and predicted levels in 25 years. In *Epidemiol Psychiatr Sci.*, 2023.

[66] A. Yates, A. Cohan, and N. Goharian. Depression and self-harm risk assessment in online forums. In *EMNLP*, 2017.

[67] J. E. Young and J. S. Klosko. *Reinventing Your Life: The Breakthough Program to End Negative Behavior... and Feel Great Again*. Penguin, 1994.

[68] J. E. Young, J. S. Klosko, and M. E. Weishaar. *Schema therapy: A practitioner's guide*. guilford press, 2006.

[69] Z. Zhang, S. Chen, M. Wu, and K. Zhu. Symptom identification for interpretable detection of multiple mental disorders on social media. In *EMNLP*, 2022.

[70] Z. Zhang, S. Chen, M. Wu, and K. Q. Zhu. Psychiatric scale guided risky post screening for early detection of depression. In *IJCAI*, 2022.