2024 12th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW) | 979-8-3315-1645-1/24/$31.00 ©2024 IEEE | DOI: 10.1109/ACIIW63320.2024.00053

Enhancing Patient Intake Process in Mental Health Consultations Using RAG-Driven Chatbot

Minoo Shayaninasab
SIAT, Simon Fraser University
Surrey, B.C., Canada
minoo_shayaninasab@sfu.ca

Maryiam Zahoor
SIAT, Simon Fraser University
Surrey, B.C., Canada
maryiam_zahoor@sfu.ca

Özge Nilay Yalçın
SIAT, Simon Fraser University
Surrey, B.C., Canada
oyalcin@sfu.ca

***Abstract*—In this paper, we develop and evaluate an empathic chatbot, called EmoBot, for a mental healthcare patient intake scenario. EmoBot employs Retrieval-Augmented Generation (RAG) methods with Large Language Models (LLMs), to provide empathic support and advice, while ensuring the intake process is completed. We evaluate EmoBot's performance both by automated metrics using available datasets and through dialogues with a simulated patient model that mimics varying levels of depression severity. Our evaluations showed good agreement between EmoBot's categorization of depression severity levels and human raters. EmoBot's topic classification system achieved 78.48% accuracy on the Primate2022 dataset [1] without fine-tuning. EmoBot's responses showed a 0.93 similarity with patient inputs, proving contextual relevance. These findings highlight the effectiveness of RAG-based methods in guiding and providing safety guardrails for LLMs and their potential as supportive tools for health assessment and support.**

***Index Terms*—Conversational Agents, Mental Health, Retrieval Augmented Generation (RAG), Large Language Models (LLM)**

I. Introduction

Mental health services worldwide, especially on the aftermath of the COVID pandemic faced challenges in providing support, caregiver and counsellor shortages and extended wait times [2]. There are also barriers to accessing healthcare services due to the fear of stigma with mental health issues or substance use disorder [3], and cultural or language barriers that can affect certain demographics [4]. Conversational agents, dialogue systems or chatbots, with their ability to engage in natural conversations with users [5], can revolutionize the medical system by providing 24/7 personalized support, enabling access to information and helping with routine checks of patients, alleviating strain in the healthcare system where healthcare professionals are struggling with burnout. In recent years, conversational systems are increasingly being used in variety of healthcare tasks from answering medical exam questions to addressing patient inquiries [6]. Studies specifically show that users find comfort in expressing their concerns and frustrations to a conversational agent that is accessible, particularly if they believe the agent is "listening" and responding with care [7].

However, achieving personalized and empathic interaction can be challenging, especially in healthcare use cases where the accuracy and reliability of the responses are crucial [8]. Recent conversational agents using Large Language Models (LLMs) are shown promise in human-like and natural conversations, however, are found to be prone to generate hallucinations [9] and provide misinformation [10], which can have serious consequences in healthcare support.

In this paper, we introduce EmoBot, an empathic conversational agent designed to balance task completion and emotional support in healthcare patient intake scenarios. EmoBot uses a dialogue manager for guiding conversations, LLMs for response generation, and Retrieval Augmented Generation (RAG) methods to provide empathic responses, particularly addressing common symptoms of depression. We evaluate EmoBot's capabilities in dialogue management, task-oriented functionalities, and non-task-oriented interactions with a simulated patient model based on the PHQ-9 [11] documentation. Our automated and semi-automated evaluations, based on the PHQ-9, show a good agreement indicated by Intraclass Correlation Coefficients (ICC). Additionally, evaluations using BLEU and ROUGE metrics, with professional counselor responses as references, demonstrated the relevance and appropriateness of EmoBot's support. Notably, we achieved a high similarity score of 0.93 between EmoBot's responses and patient inputs, emphasizing the system's capability to provide contextually appropriate support.

II. Related Work

Conversational Agents in Healthcare. Conversational agents are designed to understand and generate human-like text, allowing them to engage in natural dialogue, compose poetry, solve puzzles, and even write codes [12]. Their increasingly improved capabilities have resulted in gaining popularity in healthcare use-cases as well, especially with a mental health focus, such as in post-treatment communications, facilitating support groups, offering counseling services, and assisting with administrative tasks [8], [13].

There has been multiple attempts at empathic chatbots, in use cases relevant to mental healthcare and emotional support. XiaoIce [14], an AI chatbot developed by Microsoft, is designed to engage in conversations with users, providing emotional support, companionship, and entertainment [15]. The system iteratively improves based on user feedback and utilizes machine learning and rule-based approaches for dialogue management. Wysa [16] includes a mood tracker that can detect negative emotions and recommends depression tests or professional help. Similarly, SERMO [17] utilizes Cognitive Behavioural Therapy (CBT) methods to interact with users and based on their emotions suggests pleasant activities or mindfulness exercises. SoulChat [18] an instruction-tuned LLM demonstrates empathy and understanding in counselling process. Woebot [19], developed by psychologists at Stanford University, also offers CBT to individuals grappling with depression and anxiety symptoms, and help reducing depression and anxiety symptoms over time [13]. M-Path [20], another conversational system designed for empathic patient intake scenarios in counselling services [20], is found to be engaging and preferred to a standard pen-and-paper intake process. However, these systems often can be limited to their specific tasks, and striking a balance between open ended, personalized and task oriented conversation has been a challenge in the empathic agents [21], which has been an active focus of research in conversational systems in general [22], [23].

LLMs for Healthcare Support. Recently conversational agents based on Large language models (LLMs), a revolutionary technology, have shown state-of-the-art performance in healthcare [6]. ChatGPT's achievement of passing grades on the United States Medical Licensing Examinations has garnered particular attention [24], [25]. Additionally, Med-PaLM2 [26] (PaLM2 fine-tuned on medical data) has achieved state-of-the-art accuracy, demonstrating near-expert human clinician levels [25]. Moreover, PubMedBERT and BioBERT [27] (derived from an LLM BERT) finetuned with domain-specifc data show exceptional performance in describing results in addition to rewriting for specified readers and audiences [25], [28]. To analyze vast amount of clinical text data, ClinicalBERT [29], GatorTron [30] and Galactica [31] demonstrate high efficiency [6], [25], [32].

To improve the performance and generalizability of LLMs, the author of [33] efficiently tries to bridge the gap between AI technology and practical healthcare applications by investigating the incorporation of clinical questionnaire data, specifically Patient Health Questionnaire-9 (PHQ-9 [11]), thereby grounding the model's prediction to the symptoms outlined in PHQ-9 for detecting depression from social media text. Similarly, to generate clinically relevant follow-up questions for mental health triaging, Gupta et al. [1] developed PRIMATE dataset on extended PHQ-9 to better fine-tune their deep learning model and enhance model's ability. They evaluated their model's performance using BLEURT [34] and ROUGE-L [35] demonstrating effectiveness of their methodology.

While LLMs allowed for substantial advancements in conversational interaction, they are not without significant limitations. LLM-based agents often lack true comprehension of their content [10] and concerns exist regarding their transparency and explainability [36]. Moreover, LLMs are found to be prone to being deceived by misleading inputs, which can result in the generation of disinformation [10], generate hallucinations when responding to queries outside their training data [9] and most importantly, exhibit biases incorporated in their training data [10]. These shortcomings can be costly if not dangerous in healthcare settings, especially when used by vulnerable populations, pointing to the need of creating guardrails, testing and guiding the LLM outputs carefully.

RAG Method for Guiding LLMs. Retrieval Augmented Generation (RAG) have been recently suggested as a way to guide the response generation process of LLMs without computationally heavy training or fine-tuning. They utilize semantic similarity calculations to fetch relevant section of documents from external knowledge base thereby effectively improving factual accuracy of LLMs [9].

The development of RAGs can be broadly categorized into three paradigms: Naive RAG, Advanced RAG and Modular RAG [9]. Naive RAG, retrieves relevant document sections based on a query to generate responses [37]. Advanced RAG utilizes fine-grained segmentation, metadata and optimization strategies before and after retrieval to improve integration of retrieved data into the generation process [9]. To introduce new functionalities without redesigning the entire system, adjustable and replaceable modular components tailored to specific requirements were incorporated in Modular RAG [9].

The study by Siriwardhana et al. [38] demonstrated that fine-tuning RAG models on specific domain samples significantly enhances performance over GPT-3 based models, emphasizing the need for domain adaption in improving the effectivness of RAG for open-domain question answering. Meanwhile, Peng et al. [39] enhanced LLMs with external knowledge and automated feedback, showcasing potential advancements in chatbots.

The evaluation of RAG models, typically use task-specific metrics, such as EM and F1 scores for question answering, simple accuracy for fact checking [9], BLEU and ROUGE for assessing answer quality [9], [40] and lastly RALLE [41] performs automatic RAG evaluation. However, research on evaluating RAG's unique features is still sparse. Main evaluation areas include retrieval and generation quality [9], with new tools and benchmarks like RGB, RECALL, CRUD, RAGAS, and TruLens for a structured evaluation [42]–[44].

III. EmoBot: Architecture and Development

In this paper, we propose a dialogue management architecture, EmoBot, which is designed to provide empathic, targeted support during the patient intake process. We utilize RAGs as a way to integrate task-related information, and provide guardrails and guidance to the LLMs that are known to hallucinate. LLMs without restrictions or such guardrails can be potentially be dangerous in healthcare applications, where wrong information can pose serious risks to the lives and health of individuals. This section provides details on EmoBot's architecture and development, while focusing on a patient intake use case which is tailored to provide empathic responses to users' mental health needs or questions. In the patient intake use-case scenario, EmoBot represents a knowledge-based approach to mental health support via conversational agents, specifically designed to address symptoms of depression empathically while completing a patient intake questionnaire (PHQ-9 [11]) as its core task. However, it should be noted that EmoBot's architecture can be adapted to other

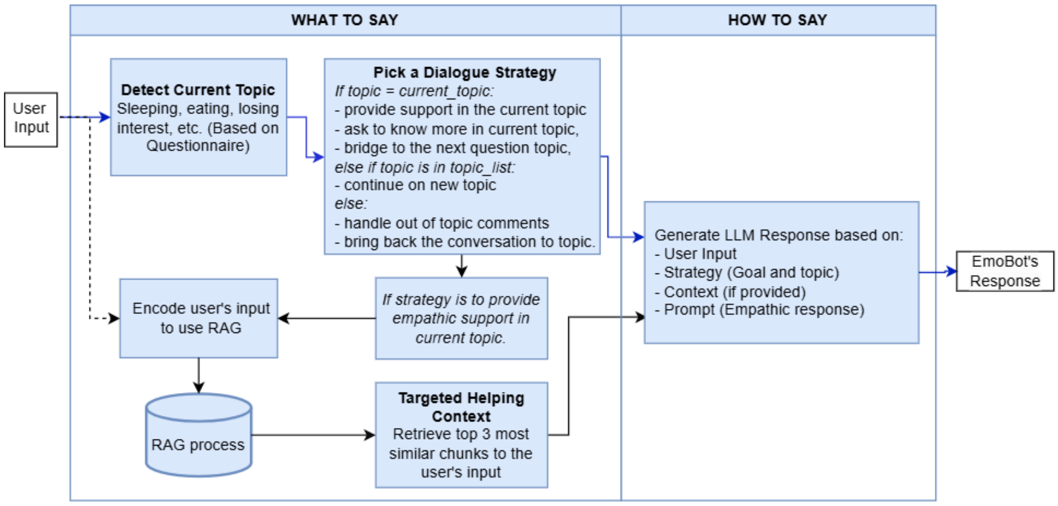

Fig. 1. Dialogue Management in EmoBot. LLM prompt is enriched depending on the topic, dialogue strategy and relevant content for generating response.

use cases that focuses on information gathering via surveys or questionnaires, while providing empathic support and advice.

EmoBot's capabilities can be categorized into two functions:

1) Administering PHQ-9 Intake Survey

- PHQ-9 Answers: EmoBot must ask and receive answer on all PHQ-9 questions.
- Evaluation of Answers: It should evaluate user responses based on depression severity scale.
- Provide Final Result of PHQ-9: EmoBot should summarize the results of the PHQ-9, offering users insights into their mental health status.

2) Providing Targeted Support and Advice

- Targeted Helping: It should offer empathic and targeted assistance, providing coping strategies, advice, or additional resources tailored to users' specific needs.

Figure 2 illustrates these different components of EmoBot's conversation process, where each component's contribution to EmoBot's response is highlighted in green and yellow and Figure 1 provides understanding on how the different components of EmoBot work together to manage a dialogue.

To ensure a natural conversation flow during the intake process, EmoBot initially employs a topic classification system which categorizes user input into ten categories, which includes nine topics relate to questions in PHQ-9 questionnaire, and an additional category used for "conversation flow" and out of topic conversations. Leveraging the zero-shot learning of LLMs (gpt-3.5-turbo in our implementation), this system enables a seamless transition between topics and minimizing repetitive exchanges. The classification system uses the user's query with the following prompt "classify the following statement into one of the relevant categories:", followed by the 10 categories and their corresponding integer value.

The result of this topic classification is used to pick a dialogue strategy. If user's input is deemed relevant to one of the nine PHQ-9 topics, then EmoBot retrieves relevant information from professional documents using the Retrieval Augmented Generation (RAG) system. If the topic was also brought up before, EmoBot strategizes to bridge to the next PHQ-9 question to ensure task completion. If the user input is unrelated to the questions, EmoBot initiates non-task oriented conversation by providing empathic response and steering the conversation back to the task.

EmoBot carefully monitors the dialogue to address all topics outlined in the PHQ-9 questionnaire and guides the conversation to empathically respond to the user's mental health concerns. Once the list of topics is exhausted, EmoBot prompts to conclude the conversation, signaling the completion of the intake process. EmoBot permits users to revisit a topic multiple times during conversation, however, once a topic has been addressed and removed from discussion points, EmoBot refrains from reintroducing it.

The selected strategy, along with the detected topic, user input and context (retrieved through RAG) are then used to engineer a prompt for a guided LLM response. Notably, these prompts consistently request empathic responses. Figure 2 give insights of these instructions. By integrating these elements, EmoBot enhances the user experience through the intake process while addressing mental health needs with empathy and understanding. The upcoming sub-sections will detail the two main functions of EmoBot's dialogue processes.

A. *PHQ-9 Intake Survey Implementation*

EmoBot's primary goal is set to administer the PHQ-9 during a patient intake process in mental health counselling services. The PHQ-9 includes 9 questions to assess the frequency of certain symptoms related to depression, including sleeping and eating habits, social interactions, concentration, and suicidal thoughts. In its original survey format, users respond in a 4-item Likert scale scored from "Not at all" as

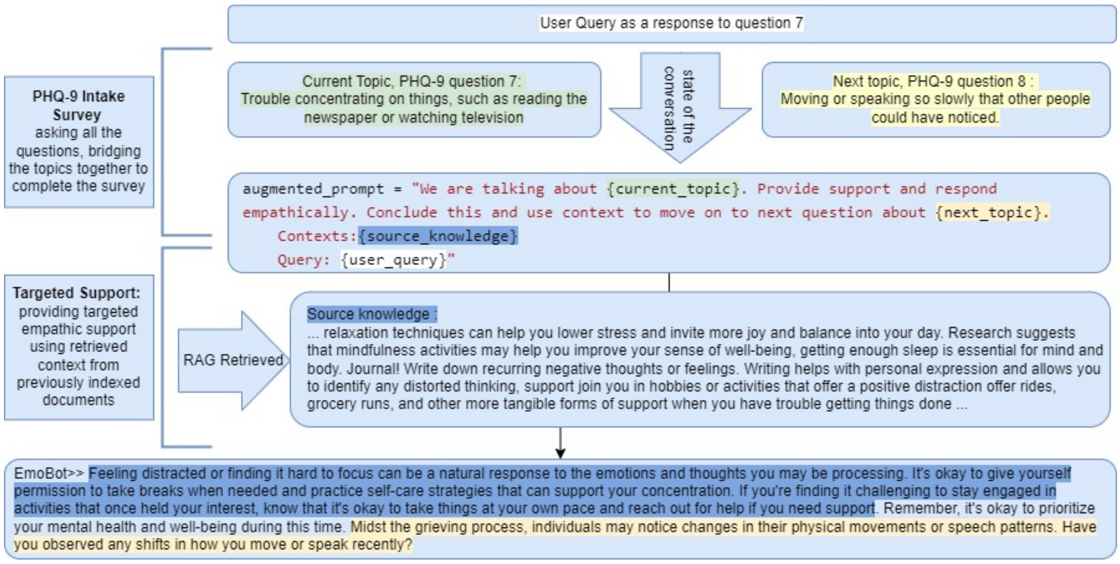

Fig. 2. An illustrative example of a conversation flow, showing how the intake questions are used while providing empathic support and advice to the user.

0, to "Nearly every day" as 3. These scores then summed up to determine the severity level of depression, categorized as minimal (0-4), mild (5-9), moderate (10-14), moderately severe (15-19) and severe (20-27). This assessment is then used to decide on the urgency of care required for the patient.

There are several tasks a conversational system needs to complete for the patient intake use-case: 1) asking and receiving answers for all 9 questions of the PHQ-9 2) evaluating each reply based on the Likert scale 3) providing a final depression severity assessment. This structure mirrors a typical survey protocol and can be adapted to other use-cases. However, given the mental healthcare context, additional precautions are necessary in case a user is experiencing suicidal thoughts: 4) the system needs to immediately direct the user to a healthcare professional, or suggest to dial 9-1-1.

EmoBot stores all 9 questions of the PHQ-9, a list to track previously asked questions, and each reply the user gave corresponding to the questions asked in its memory. If a question was asked and received at least one reply in topic, this question number is appended in this list and is no longer asked again. However, if the user wants to continue discussing a topic, EmoBot responds empathically and stores that information. Once the dialogue concludes and all questions are asked, the stored replies of the user per each topic are utilized to complete the PHQ-9 assessment. EmoBot rates the severity of user statement on each topic on the Likert scale of 0 to 3, by prompting LLMs (gpt-3.5-turbo), see Figure 3 for the final prompt used for this step after multiple rounds of prompt engineering. If the context is not provided, the assessment for that question will not be made. The model used for this task, is set up to respond with maximum 1 token, in order to make sure we get only the rating that we want rather than a full textual description of the patient's state.

```
prompt = f"Here are given some questions and answers related to {topic}
        as a factor of mental health well-being.
        how would you rate the following question about this person?\n
        questions and answers:" + {statements} + f"\n  On scale of 0 to 3,
        over the last 2 weeks,  rate how often has this person been bothered by {topic}.\n
        0: Not at all, 1: Several days, 2:More than half the days, 3: Nearly every day"
```

Fig. 3. The prompt used for zero-shot classification is; "statements" are all the user's lines about this topic and "topic" is one of the symptoms in PHQ-9.

B. *Targeted Support and Advice*

In addition to administering the PHQ-9 questionnaire as a main goal, EmoBot is designed to provide targeted support to user inputs in an empathic way. EmoBot uses prompt engineering and RAGs for providing empathic support. Initially, the empathic behavior was defined only in initial system prompt, but this resulted in the system's empathy waned as the conversation extended, as it is a known limitation of LLMs. Using prompt engineering, which requires continually refining prompts through iterative experimentation, the current prompts for receiving empathic acknowledgement are explicit commands to the LLM to " ... provide support and respond empathically." and "... if the user needs comfort, keep talking and respond empathically." with the initial system setting of "You are a helpful mental health chat bot ...". The low-level empathic response of EmoBot therefore is achieved through appending prompts in each turn to ensure that the conversation maintains its tone and EmoBot sustains its empathic feature throughout interactions.

Targeted Helping as an empathic behavior requires providing help and support related to the emotional state and response of the user [45]. However, any incorrect responses or advice can cause serious harm, up to a point that it might be better not to provide any help than to provide a wrong type of help. Thus, it is essential to restrict chatbot responses to prevent potential harms caused by hallucination of LLMs, we leverage Retrieval Augmented Generation (RAG) methods, and information extracted from relevant documents.

EmoBot integrates RAGs to provide targeted helping response tailored to the PHQ-9 questionnaire and user's particular challenges. We do so by incorporating insights that are extracted from blog posts of professional healthcare providers [46]–[51]. The extracted documents mainly discuss how to deal with symptoms of depression which are related to or seen frequent in depression cases, in order to make it more manageable. Note that, EmoBot's capability of "Targeted Helping" is therefore tied to the intake questions and user response to ensure user and topic specific empathic behavior, while ensuring guardrails for LLM's response. However, "Targeted Helping", which is achieved via RAGs, can include any other resources that might be relevant and can also be generated on the fly depending on user's input [9].

The three-level RAG method used in EmoBot includes:

- **Indexing:** The initial step involves processing and indexing the documents containing relevant information, by splitting them into smaller and meaningful chunks of size 30 words each, and encoding them into embedding vectors using ext-embedding-ada-002 model. These vector representations are then stored in a vector database for indexing. This process is completed before conversation starts to ensure fast retrieval during interaction.
- **Retrieval:** Upon receiving user input, EmoBot encodes the user prompt to compare with indexed chunks using cosine similarity to determine relevance. EmoBot retrieves the top-3 chunks that are deemed most relevant to the user's query.
- **Generation:** Retrieved information chunks are then combined with the user query to be sent to the LLM model, which then generates the final response.

These three steps of RAG process allows EmoBot to enhance its conversational capabilities, by delivering personalized and contextually relevant responses. The augmented responses via RAG process serve a dual purpose. Firstly, they provide empathic support by acknowledging the user's feelings and validating their experiences. Additionally, they include targeted helping, such as practical advice, coping strategies, or resources aimed at addressing specific concerns raised by the user. This multifaceted approach ensures that EmoBot not only empathizes with the user's emotions but also offers actionable assistance to help them navigate their challenges effectively. Figure 1 illustrates EmoBot's overall strategy in response generation through the integration of RAGs. Figure 2 further details the final step of utilizing context in response generation.

IV. EmoBot in Action: Evaluation and Results

Our initial evaluation of EmoBot focuses on its abilities to perform the intake process, and providing relevant targeted support. We provide both automated and semi-automated evaluations of EmoBot, along with dialogue examples that showcases EmoBot's functionality.

To evaluate EmoBot's performance, a simulated patient model was created to engage in conversations mimicking individuals with various symptoms of depression, as suggested in other LLMs based dialogue systems [52], [53]. We define 20 patient models using prompts, each corresponding to different levels of depression severity described in PHQ-9 documentation [11]. Below are 3 example patient model prompts:

- **Mild depression:** You are a university student who is experiencing mild depression. You're having trouble sleeping and feeling tired less than half of the days. You're having trouble concentrating only a few days. But your eating habits and other activities remain normal.
- **Moderate depression:** You are a working parent struggling with moderate depression. You have been having trouble ... [list PHQ-9 symptoms] for about half of the days for the past two weeks. You don't have thoughts of hurting yourself.
- **Severe depression:** You are a retired individual facing severe depression. You have been having trouble sleeping, ... [list PHQ-9 symptoms] almost every day.

We then tie EmoBot with each of the patient model for automated conversation, and log all interactions in JSON format to evaluate EmoBot's functionality. We ran a total of 20 conversational examples (4 examples x 5 depression levels) that completed with an average of 28 conversation turns, and a total of 562 conversation turns.

A. Evaluation of the Intake Process

1) Topic Accuracy: As mentioned in Section III, EmoBot initially classifies the user input into one of 10 determined topics (PHQ-9 + conversation flow). To evaluate the accuracy of this classification, we used the Primate2022 dataset [1], which has a total of 2003 anonymized posts and their relevance to all nine PHQ-9 topics. We evaluated the accuracy of our topic categorization system using this dataset without any training, and achieved 78.48 accuracy over all records, and 81.09 over the test-set.

2) Task Completion: In a total of 20 conversation examples, EmoBot finished asking all questions while responding to user prompt. However, some inputs from the Patient model were misclassified, resulting in an incomplete assessment of the PHQ-9. This conversation was related to an instance of patient model with no depression symptoms, where the lack of context in inputs were misclassifed as a general "conversation flow" class. Future work needs to address this by improving topic classification system.

3) PHQ-9 Depression Level Accuracy: To evaluate EmoBot's ability to accomplish its task of detecting depression levels, we compared the PHQ-9 ratings of EmoBot with those of a human rater using 20 example dialogue cases. Human rater followed the instructions of PHQ-9 to decide the depression level categories. We calculated the Intraclass Correlation Coefficients (ICC) [54] to assess the agreement between EmoBot and the human rater using Pingouin statistical package [55]. The ICC score (average raters absolute) was 0.86 ($F = 6.94$, $p < 0.001$). The 95% confidence interval

(CI 95%) ranged from 0.61 to 1.0, which corresponds to good agreement between the human rater and EmoBot. Additionally, the Cohen's kappa coefficient for absolute agreement at the question level of PHQ-9 was 0.62, showing moderate agreement. As shown in Figure 4 in the cases where EmoBot disagreed with the human rater, the total score indicated one level more severe than the human rating. These cases were scored as "moderately severe" and "moderate" by the human rater, while EmoBot scored them as "severe" and "moderately severe", respectively. Another source of disagreement was among "none/minimal depression" class where Emobot classified into "mild depression", when the patient model talked about occasional symptoms but the statements were not strong enough to convince the human rater as depressive symptoms. However, it must be noted that there was only one human rater who was not a healthcare professional, and we need to establish the ground truth with multiple raters in future analyses.

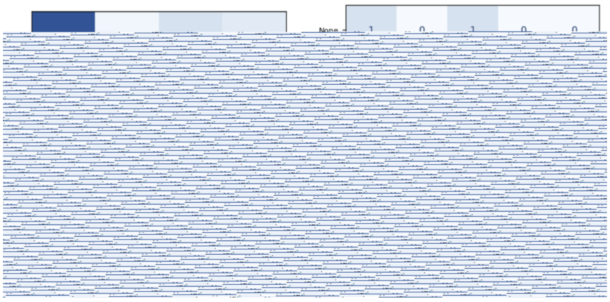

Fig. 4. Confusion matrices show the agreement between the human rater and Emobot in (A) question-level (B) depression level assessments.

B. Evaluation of Targeted Support

TABLE I
EXAMPLE INTERACTION WITH A PATIENT MODEL.

User	In addition to the fatigue and low energy levels, I've also been feeling more irritable and on edge lately. Small things that didn't bother me before now seem to trigger intense feelings of frustration and irritability. I find myself snapping at people or withdrawing even more when I'm feeling overwhelmed. It's like I'm on an emotional rollercoaster, and it's hard to predict how I'll react to situations. I know it's not healthy, and I want to find ways to manage these intense emotions better.
EmoBot	It sounds like you've been dealing with not just physical fatigue, but also emotional challenges like irritability and feeling on edge. It's tough when small things start triggering intense emotional reactions and impact your interactions with others. Recognizing these changes in your mood and behavior is an important step towards seeking support and finding healthier ways to manage these intense emotions. Have you noticed any specific triggers or patterns that lead to these feelings of frustration and irritability?

Interactions with EmoBot showed that EmoBot was able to acknowledging the user's distress and offers validation and empathy (Table I), and expressing its availability and willingness to listen whenever the user feels ready to engage (Table II). Through these examples, we see how EmoBot employs active listening and empathy to navigate conversations with users, in a pace that is comfortable for them.

TABLE II
EXAMPLE INTERACTION WITH HUMAN USER.

User	I don't go out much lately. Also don't feel like talking to my friends anymore.
EmoBot	I'm sorry to hear that you've been feeling this way. It's completely understandable to feel like withdrawing when you're going through a tough time. It's okay to take things at your own pace, but it's also important to reach out for support when you need it. Is there a particular reason why you don't feel like talking to your friends anymore?
User	I don't wanna talk about it right now.
EmoBot	I understand if you're not ready to talk about it. It's important to take things at your own pace and only share when you feel comfortable. Just know that I'm here whenever you're ready to open up. In the meantime, remember to take care of yourself and practice self-care. Can I ask if you've been experiencing little interest or pleasure in doing things you normally enjoy recently?

We use two similarity evaluations to evaluate the targeted helping capability of EmoBot, as suggested for RAG evaluation [9]: the similarity between EmoBot responses and topics of conversation, and similarity between EmoBot responses and user input. We could not utilize other metrics such as accuracy, as there is no ground truth data for our task. Future work could address this issue by manual rating of the retrieved information and EmoBot responses in the human interaction scenario by professional mental health counsellors.

TABLE III
COSINE SIMILARITY BETWEEN THE TOPIC AND EMOBOT'S TARGETED SUPPORT RESPONSES PER EACH PHQ-9 QUESTION SHOWN AS ROWS. COLUMNS SHOW THE DEPRESSION LEVELS OF THE PATIENT MODEL (0= NO DEPRESSION, 4= SEVERE DEPRESSION).

PHQ-9	Depression Level					
Questions	0	1	2	3	4	AVG
1	0.75	0.76	0.78	0.78	0.78	0.77
2	0.80	0.81	0.83	0.82	0.81	0.814
3	0.79	0.81	0.82	0.80	0.82	0.808
4	0.82	0.83	0.85	0.84	0.84	0.836
5	0.78	0.78	0.79	0.79	0.80	0.788
6	0.79	0.80	0.79	0.83	0.81	0.804
7	0.79	0.80	0.81	0.79	0.81	0.80
8	0.73	0.76	0.79	0.81	0.79	0.776
9	0.73	0.75	0.73	0.79	0.81	0.762
AVG	0.776	0.789	0.799	0.806	0.808	**0.795**

Table III shows cosine similarity between the topic and the turns of the conversation and each of the PHQ-9 questions that they are categorized into. Each cell represents the average similarity of all EmoBot's responses. Last row and column represents the averages of one topic or depression level conversations. This was measured using the embedding vector space using text-embedding-ada-002 [56] from OpenAI. Using the same embedding model, we further computed the cosine similarity between the patient model's inputs and EmoBot's following responses to calculate how relevant and appropriate

EmoBot's response was given user input. Our results showed on average 0.93 ($\pm$0.03) similarity in the embedding space.

In order to evaluate EmoBot's response quality for providing healthcare advice, we further used BLEU [57] and ROUGE [58] metrics over a dataset that includes answers to user questions from healthcare professionals, called Counsel-chat [59]. Counsel-chat consists of questions from users, Q_i, and a set of counsellor responses to it, $R_{i1}, R_{i2}, ...R_{im}$. Our chatbot dialogue consists of a set of $(UserInput_j, EmoBotResponse_j)$ pairs following each other in a dialogue. In order to pick a set of references for each $EmoBotResponse_j$ to calculate BLEU and ROUGE for, we compared similarity between the Q_is and $UserInput_j$s, and picked all the R_is paired with the most similar Q_i as gold labels for that specific turn of EmoBot's response. To achieve this set of crafted gold labels, we used cosine similarity between embeddings of $UserInput_j$ and Q_i extracted using text-embedding-ada-002 model. EmoBot's 562 responses across 20 conversations, resulted in an average cosine similarity of 0.85 between $UserInput_j$s and selected Counsel-chat questions, Q_is, and an average of 5.5 references per $EmoBotResponse_j$. The overall BLEU and ROUGE are average of these metrics across all $EmoBotResponse_j$s.

Using the NLTK library [60] in Python, we calculated BLEU scores for each response. The average BLEU score for unigrams and bigrams was 0.30, while the average BLEU score for up to 4-grams was 0.01, highlighting the challenges in generating longer coherent responses. We also computed ROUGE scores using the rouge-score library. The average ROUGE-1 (unigram) scores were 0.36 for precision, 0.21 for recall, and 0.23 for F-measure. ROUGE-2 (bigram) scores averaged 0.04 for precision, 0.04 for recall, and 0.021 for F-measure. For ROUGE-L, which considers the longest common sub-sequence, the scores were 0.20 for precision, 0.11 for recall, and 0.12 for F-measure. Using the same reference sets from Counsel-chat, we calculated the cosine similarity score for each EmoBot response within the embedding space. The average similarity score of 0.81 indicates a resemblance between EmoBot's and counselors' responses.

While these scores provide insights into EmoBot's performance, it is crucial to consider the contextual differences between tasks and datasets. BLEU and ROGUE are typically used for translation and summarization tasks [61], respectively, but are recently suggested to be used for LLM and chatbot evaluation [9], [61]. Moreover, the counsel-chat dataset consists of open-domain single-turn QA from expert counselors, not specifically related to the depression symptoms or PHQ-9 questions asked by EmoBot. Therefore, the relatively low scores are expected, as they do not directly represent ground truth for our task. Nonetheless, this evaluation underscores the need for suitable datasets or evaluations from healthcare professionals to refine EmoBot's responses.

V. Conclusion and Future Directions

In this paper, we explored the development of a healthcare chatbot, EmoBot, to provide targeted support during the mental health patient intake process. We focused on both task oriented and non task-oriented conversation, using LLMs, dialogue management strategies and RAG systems. EmoBot's use of RAGs is aimed to address common symptoms of depression with empathic responses, while guiding and providing guardrails for LLMs which are known to hallucinate. Our evaluations showed promising results, with a good agreement level between EmoBot and the human rater on PHQ scores and a high similarity score between user questions and RAG-based content retrieved by EmoBot. However, current EmoBot model cannot classify a single input into multiple topics, posing limitations for a real-life use-case. Future work needs to address this limitation by improving the RAG system.

The current RAG implementation relies on carefully selected resources to provide targeted helping and empathy, ensuring that EmoBot's responses are supervised and aligned with best practices in mental health support. However, these resources could be improved by collaborating with healthcare professionals to refine the content and guidance, and further enhance EmoBot's ability to deliver effective and appropriate support to users in need. Moreover, EmoBot's flexible and adaptable features can be utilized to administer surveys beyond PHQ-9, such as GAD-7 [62], or BDI [63]. By customizing topic definitions, and slight changes in prompts, this method can be applied to other healthcare use-cases to handle an empathic dialogue while administering surveys.

Main limitation for the current method was the lack of ground truth datasets for a comprehensive evaluation of EmoBot in patient intake scenario. Future research efforts should include developing datasets for mental health counselling intake using the PHQ-9 questionnaire or targeted support examples for more comprehensive evaluations. Additionally, utilizing user-generated data, such as questionnaire responses and dialogue interactions, can lead to training more robust and accurate models for EmoBot, particularly emphasizing the final questionnaire filling task. This strategy signifies a departure from the current dependence on few-shot learning methods, thereby laying the groundwork for the development of a more reliable system for mental health assessment based on its accuracy. At present, there is no established measure of accuracy for this task. Finally, EmoBot's performance needs to be examined during human interaction, while involving individuals with diverse backgrounds, stages, and severity of depression symptoms. Real-world usage by a diverse user population will provide valuable insights into EmoBot's performance and areas for improvement. Gathering feedback from users will be essential for enhancing EmoBot's responses and tailoring its support to meet the diverse needs of individuals grappling with mental health challenges.

Overall, EmoBot holds promise for advancing the field of mental health support through RAG-driven conversational agents. By providing guardrails to LLMs through RAGs and guiding empathic responses through dynamic prompt generation, EmoBot can provide well-informed, relevant, and guided support while ensuring task completion in the intake process. However, further evaluation and refinements are necessary to ensure its usability in real-world scenarios.

VI. Ethical Impact Statement

In developing EmoBot, a chatbot for both task-oriented and non-task-oriented conversations using LLMs, dialogue management strategies, and RAG systems, several ethical considerations arise, particularly due to its applications in mental health support. Currently, we have used only automatic evaluations and neutral judgments to build and ensure the functionality of the system. However, EmoBot must undergo rigorous testing to qualify for public use. Continuous evaluations with real users and incorporating their feedback are essential to address ethical concerns and improve the chatbot's effectiveness. We propose initial user testing with professional counselors and psychotherapists mimicking depressed individuals, rather than actual patients, to evaluate EmoBot's responses for appropriateness and safety. Additionally, EmoBot's effectiveness may vary across demographic groups and contexts. Comprehensive evaluations with diverse user populations are planned to understand these limitations better and reduce bias that might lead to unfair treatment or misdiagnosis.

The RAG system, designed to enhance response accuracy, still has limitations, particularly with complex or ambiguous inputs. Continuous improvement of the algorithms and feedback from mental health professionals are essential to enhance EmoBot's reliability. Collaboration with mental health experts will ensure that EmoBot's content aligns with best practices, involving regular updates with validated and current information. It is crucial to anticipate scenarios where EmoBot might fail to provide appropriate support, especially in critical situations like dealing with users at risk of suicide. Establishing clear guidelines for handling such failures, including redirecting users to human professionals, is a priority.

There is a risk that the technology could be misused for surveillance or unauthorized data analysis. Robust data security measures, including encryption, secure storage, and limited access to authorized personnel, are crucial. Users must be informed about how EmoBot works, its limitations, and how their data is used to make informed decisions about their engagement with the chatbot.

If EmoBot's patient intake capability is used to assess mental health in institutions, strict privacy policies must govern access to user data while keeping chat content confidential. EmoBot's assessment accuracy must also be high to be trusted. Our current system, implemented with zero-shot learning, achieves acceptable accuracy but is not yet suitable for public use. Trust can be built only with systems trained on large datasets, which were unavailable during EmoBot's development.

References

[1] S. Gupta, A. Agarwal, M. Gaur, K. Roy, V. Narayanan, P. Kumaraguru, and A. Sheth, "Learning to automate follow-up question generation using process knowledge for depression triage on reddit posts," *arXiv preprint arXiv:2205.13884*, 2022.

[2] W. H. Organization *et al.*, "World mental health report: Transforming mental health for all," 2022.

[3] F. A. Gary, "Stigma: Barrier to mental health care among ethnic minorities," *Issues in mental health nursing*, vol. 26, no. 10, pp. 979–999, 2005.

[4] Y. Zhao, N. Segalowitz, A. Voloshyn, E. Chamoux, and A. G. Ryder, "Language barriers to healthcare for linguistic minorities: The case of second language-specific health communication anxiety," *Health communication*, vol. 36, no. 3, pp. 334–346, 2021.

[5] S. Singh and H. Beniwal, "A survey on near-human conversational agents," *Journal of King Saud University-Computer and Information Sciences*, vol. 34, no. 10, pp. 8852–8866, 2022.

[6] 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," *arXiv preprint arXiv:2310.05694*, 2023.

[7] H. Prendinger and M. Ishizuka, "The empathic companion: A character-based interface that addresses users'affective states," *Applied artificial intelligence*, vol. 19, no. 3-4, pp. 267–285, 2005.

[8] L. Laranjo, A. G. Dunn, H. L. Tong, A. B. Kocaballi, J. Chen, R. Bashir, D. Surian, B. Gallego, F. Magrabi, A. Y. Lau *et al.*, "Conversational agents in healthcare: a systematic review," *Journal of the American Medical Informatics Association*, vol. 25, no. 9, pp. 1248–1258, 2018.

[9] Y. Gao, Y. Xiong, X. Gao, K. Jia, J. Pan, Y. Bi, Y. Dai, J. Sun, and H. Wang, "Retrieval-augmented generation for large language models: A survey," *arXiv preprint arXiv:2312.10997*, 2023.

[10] A. Tamkin, M. Brundage, J. Clark, and D. Ganguli, "Understanding the capabilities, limitations, and societal impact of large language models," 2021.

[11] K. Kroenke and R. L. Spitzer, "The phq-9: a new depression diagnostic and severity measure," pp. 509–515, 2002.

[12] 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," *Advances in neural information processing systems*, vol. 35, pp. 27 730–27 744, 2022.

[13] P. Rathnayaka, N. Mills, D. Burnett, D. De Silva, D. Alahakoon, and R. Gray, "A mental health chatbot with cognitive skills for personalised behavioural activation and remote health monitoring," *Sensors*, vol. 22, no. 10, p. 3653, 2022.

[14] L. Zhou, J. Gao, D. Li, and H.-Y. Shum, "The design and implementation of xiaoice, an empathetic social chatbot," *Computational Linguistics*, vol. 46, no. 1, pp. 53–93, 2020.

[15] H.-Y. Shum, X.-d. He, and D. Li, "From eliza to xiaoice: challenges and opportunities with social chatbots," *Frontiers of Information Technology & Electronic Engineering*, vol. 19, pp. 10–26, 2018.

[16] B. Inkster, S. Sarda, V. Subramanian *et al.*, "An empathy-driven, conversational artificial intelligence agent (wysa) for digital mental well-being: real-world data evaluation mixed-methods study," *JMIR mHealth and uHealth*, vol. 6, no. 11, p. e12106, 2018.

[17] K. Denecke, S. Vaaheesan, and A. Arulnathan, "A mental health chatbot for regulating emotions (sermo)-concept and usability test," *IEEE Transactions on Emerging Topics in Computing*, vol. 9, no. 3, pp. 1170–1182, 2020.

[18] Y. Chen, X. Xing, J. Lin, H. Zheng, Z. Wang, Q. Liu, and X. Xu, "Soulchat: Improving llms' empathy, listening, and comfort abilities through fine-tuning with multi-turn empathy conversations," in *Findings of the Association for Computational Linguistics: EMNLP 2023*, 2023, pp. 1170–1183.

[19] 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 mental health*, vol. 4, no. 2, p. e7785, 2017.

[20] Ö. N. Yalçın and S. DiPaola, "M-path: A conversational system for the empathic virtual agent," in *Biologically Inspired Cognitive Architectures 2019: Proceedings of the Tenth Annual Meeting of the BICA Society 10*. Springer, 2020, pp. 597–607.

[21] A. K. Wardhana, R. Ferdiana, and I. Hidayah, "Empathetic chatbot enhancement and development: A literature review," in *2021 International Conference on Artificial Intelligence and Mechatronics Systems (AIMS)*. IEEE, 2021, pp. 1–6.

[22] S. Roller, Y.-L. Boureau, J. Weston, A. Bordes, E. Dinan, A. Fan, D. Gunning, D. Ju, M. Li, S. Poff *et al.*, "Open-domain conversational agents: Current progress, open problems, and future directions," *arXiv preprint arXiv:2006.12442*, 2020.

[23] H. Chen, X. Liu, D. Yin, and J. Tang, "A survey on dialogue systems: Recent advances and new frontiers," *Acm Sigkdd Explorations Newsletter*, vol. 19, no. 2, pp. 25–35, 2017.

[24] T. H. Kung, M. Cheatham, A. Medenilla, C. Sillos, L. De Leon, C. Elepaño, M. Madriaga, R. Aggabao, G. Diaz-Candido, J. Maningo

et al., "Performance of chatgpt on usmle: potential for ai-assisted medical education using large language models," *PLoS digital health*, vol. 2, no. 2, p. e0000198, 2023.

[25] A. J. Thirunavukarasu, D. S. J. Ting, K. Elangovan, L. Gutierrez, T. F. Tan, and D. S. W. Ting, "Large language models in medicine," *Nature medicine*, vol. 29, no. 8, pp. 1930–1940, 2023.

[26] K. Singhal, T. Tu, J. Gottweis, R. Sayres, E. Wulczyn, L. Hou, K. Clark, S. Pfohl, H. Cole-Lewis, D. Neal *et al.*, "Towards expert-level medical question answering with large language models," *arXiv preprint arXiv:2305.09617*, 2023.

[27] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, and J. Kang, "Biobert: a pre-trained biomedical language representation model for biomedical text mining," *Bioinformatics*, vol. 36, no. 4, pp. 1234–1240, 2020.

[28] Y. Gu, R. Tinn, H. Cheng, M. Lucas, N. Usuyama, X. Liu, T. Naumann, J. Gao, and H. Poon, "Domain-specific language model pretraining for biomedical natural language processing," *ACM Transactions on Computing for Healthcare (HEALTH)*, vol. 3, no. 1, pp. 1–23, 2021.

[29] K. Huang, J. Altosaar, and R. Ranganath, "Clinicalbert: Modeling clinical notes and predicting hospital readmission," *arXiv preprint arXiv:1904.05342*, 2019.

[30] X. Yang, N. PourNejatian, H. C. Shin, K. E. Smith, C. Parisien, C. Compas, C. Martin, M. G. Flores, Y. Zhang, T. Magoc *et al.*, "Gatortron: A large language model for clinical natural language processing," *medRxiv*, pp. 2022–02, 2022.

[31] R. Taylor, M. Kardas, G. Cucurull, T. Scialom, A. Hartshorn, E. Saravia, A. Poulton, V. Kerkez, and R. Stojnic, "Galactica: A large language model for science," *arXiv preprint arXiv:2211.09085*, 2022.

[32] M. Agrawal, S. Hegselmann, H. Lang, Y. Kim, and D. Sontag, "Large language models are few-shot clinical information extractors," *arXiv preprint arXiv:2205.12689*, 2022.

[33] T. Nguyen, A. Yates, A. Zirikly, B. Desmet, and A. Cohan, "Improving the generalizability of depression detection by leveraging clinical questionnaires," *arXiv preprint arXiv:2204.10432*, 2022.

[34] T. Sellam, D. Das, and A. P. Parikh, "Bleurt: Learning robust metrics for text generation," *arXiv preprint arXiv:2004.04696*, 2020.

[35] C.-Y. Lin and F. Och, "Looking for a few good metrics: Rouge and its evaluation," in *Ntcir workshop*, 2004.

[36] E. A. Van Dis, J. Bollen, W. Zuidema, R. Van Rooij, and C. L. Bockting, "Chatgpt: five priorities for research," *Nature*, vol. 614, no. 7947, pp. 224–226, 2023.

[37] X. Ma, Y. Gong, P. He, H. Zhao, and N. Duan, "Query rewriting for retrieval-augmented large language models," *arXiv preprint arXiv:2305.14283*, 2023.

[38] S. Siriwardhana, R. Weerasekera, E. Wen, T. Kaluarachchi, R. Rana, and S. Nanayakkara, "Improving the domain adaptation of retrieval augmented generation (rag) models for open domain question answering," *Transactions of the Association for Computational Linguistics*, vol. 11, pp. 1–17, 2023.

[39] B. Peng, M. Galley, P. He, H. Cheng, Y. Xie, Y. Hu, Q. Huang, L. Liden, Z. Yu, W. Chen *et al.*, "Check your facts and try again: Improving large language models with external knowledge and automated feedback," *arXiv preprint arXiv:2302.12813*, 2023.

[40] Z. Ke, W. Kong, C. Li, M. Zhang, Q. Mei, and M. Bendersky, "Bridging the preference gap between retrievers and llms," *arXiv preprint arXiv:2401.06954*, 2024.

[41] Y. Hoshi, D. Miyashita, Y. Ng, K. Tatsuno, Y. Morioka, O. Torii, and J. Deguchi, "Ralle: A framework for developing and evaluating retrieval-augmented large language models," *arXiv preprint arXiv:2308.10633*, 2023.

[42] J. Chen, H. Lin, X. Han, and L. Sun, "Benchmarking large language models in retrieval-augmented generation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 16, 2024, pp. 17 754–17 762.

[43] Y. Liu, L. Huang, S. Li, S. Chen, H. Zhou, F. Meng, J. Zhou, and X. Sun, "Recall: A benchmark for llms robustness against external counterfactual knowledge," *arXiv preprint arXiv:2311.08147*, 2023.

[44] Y. Lyu, Z. Li, S. Niu, F. Xiong, B. Tang, W. Wang, H. Wu, H. Liu, T. Xu, and E. Chen, "Crud-rag: A comprehensive chinese benchmark for retrieval-augmented generation of large language models," *arXiv preprint arXiv:2401.17043*, 2024.

[45] Ö. N. Yalçın and S. DiPaola, "Modeling empathy: building a link between affective and cognitive processes," *Artificial Intelligence Review*, vol. 53, no. 4, pp. 2983–3006, 2020.

[46] U. D. O. HEALTH and H. SERVICES. Depression. [Online]. Available: https://www.nimh.nih.gov/health/publications/depression

[47] NHS. How to cope with depression. [Online]. Available: https://www.nhs.uk/mental-health/self-help/tips-and-support/cope-with-depression/

[48] E. C. Ann Pietrangelo. Depression and anxiety: How to identify and treat coexisting symptoms. [Online]. Available: https://www.healthline.com/health/mental-health/depression-and-anxietyWatch-more-from-the-Youth-in-Focus-video-series

[49] Anxiety and D. A. of America. Tips to manage depression. [Online]. Available: https://adaa.org/understanding-anxiety/depression/tips

[50] K. Holland. 22 ways to cope with your depression. [Online]. Available: https://www.healthline.com/health/depression/how-to-fight-depressionfaq

[51] L. R. Melinda Smith, M.A. and P. Jeanne Segal. Coping with depression. [Online]. Available: https://www.helpguide.org/articles/depression/coping-with-depression.htm

[52] L. Zheng, W.-L. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, Z. Lin, Z. Li, D. Li, E. Xing *et al.*, "Judging llm-as-a-judge with mt-bench and chatbot arena," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[53] J. Bozic, O. A. Tazl, and F. Wotawa, "Chatbot testing using ai planning," in *2019 IEEE International Conference On Artificial Intelligence Testing (AITest)*. IEEE, 2019, pp. 37–44.

[54] T. K. Koo and M. Y. Li, "A guideline of selecting and reporting intraclass correlation coefficients for reliability research," *Journal of chiropractic medicine*, vol. 15, no. 2, pp. 155–163, 2016.

[55] R. Vallat, "Pingouin: statistics in python." *J. Open Source Softw.*, vol. 3, no. 31, p. 1026, 2018.

[56] A. Name, "New and improved embedding model," 2022, accessed on June 9, 2024. [Online]. Available: https://openai.com/index/new-and-improved-embedding-model/

[57] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, 2002, pp. 311–318.

[58] C.-Y. Lin, "Rouge: A package for automatic evaluation of summaries," in *Text summarization branches out*, 2004, pp. 74–81.

[59] N. Bertagnolli, "Counsel chat: Bootstrapping high-quality therapy data," 2020.

[60] S. Bird, E. Klein, and E. Loper, *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.", 2009.

[61] W. Maroengsit, T. Piyakulpinyo, K. Phonyiam, S. Pongnumkul, P. Chaovalit, and T. Theeramunkong, "A survey on evaluation methods for chatbots," in *Proceedings of the 2019 7th International conference on information and education technology*, 2019, pp. 111–119.

[62] N. Williams, "The gad-7 questionnaire," *Occupational medicine*, vol. 64, no. 3, pp. 224–224, 2014.

[63] A. T. Beck, N. Epstein, G. Brown, and R. Steer, "Beck anxiety inventory," *Journal of consulting and clinical psychology*, 1993.