Leveraging Prompt Engineering and Large Language Models for Automating MADRS Score Computation for Depression Severity Assessment

Alessandro Raganato[1,*], Francesco Bartoli[2], Cristina Crocamo[2], Daniele Cavaleri[2], Giuseppe Carrà[2,3], Gabriella Pasi[1] and Marco Viviani[1,*]

[1]Department of Informatics, Systems, and Communication, University of Milano-Bicocca, Milan, Italy
[2]School of Medicine and Surgery, University of Milano-Bicocca, Monza, Italy
[3]Division of Psychiatry, University College London, London, UK

Abstract

This study ventures into the field of psychiatry by investigating the interactive dynamics between psychiatrists and their patients. The primary goal is to create an automated scoring mechanism using prompt engineering techniques applied to Large Language Models (LLMs) to assess the severity of depressive symptoms from these dialogues. In particular, the process of generating a depression severity score against MADRS, a rating scale widely used in psychiatry, is automated. This work aims to highlight the potential of using these techniques to improve traditional diagnostic approaches in psychiatry. The results that have emerged, while not optimal, are promising, including for the purpose of developing a full-fledged system in the future to enable the introduction of more targeted and timely interventions, thereby improving patient outcomes and improving the overall level of mental health.

Keywords

Mental Health, MADRS, Prompt Engineering, Large Language Models, Natural Language Processing

1. Introduction

The assessment of symptom severity plays a crucial role in the clinical management of mental disorders, being pivotal in diagnosing and monitoring the mental well-being of patients [1]. Traditionally, this evaluation has heavily relied on clinical experience, sometimes supported by questionnaires and rating scales during in-person visits. However, advancements in *Machine Learning* (ML) and *Natural Language Processing* (NLP) techniques offer the potential for automated systems that can support in assessing measures of symptom severity in dialogues between psychiatrists and the growing number of patients. In particular, the evolving landscape of *prompt engineering* techniques applied to *Large Language Models* (LLMs) presents a novel avenue for developing such kind of systems, to better support psychiatric assessment practices in the future.

This study, in particular, embarks on the task of automatically mapping psychiatrist-patient dialogue content to the *Montgomery-Åsberg Depression Rating Scale* (MADRS) [2], a widely accepted instrument for evaluating depression severity, through the potential of recently developed generative *Artificial Intelligence* (AI) models [3]. To establish a foundation, a manual mapping process performed by clinical experts is employed to establish connections between question-answers from some psychiatrist-patient dialogues and the corresponding *items* of the MADRS questionnaire, together with the corresponding scores (both at the individual item level and the global level). This manual mapping serves as a benchmark for subsequent comparison with results obtained from the considered AI-based approaches.

In a first approach, distinct prompt engineering techniques applied to LLMs are leveraged to compute depression severity scores for each MADRS item. Each item is devoted to assessing a different symptom domain, such as *sadness*, *inner tension*, *reduced sleep*, etc., rated on a scale from 0 to 6, with higher scores indicating more severe depressive symptoms. The computed scores are then further aggregated to provide an overall assessment, ranging from 0 to 60, with higher scores indicating more severe depression. In a second approach, we evaluate the effectiveness of using prompts to directly compute the overall depression severity score.

This study serves as a preliminary step to explore the feasibility, in the future, of creating an advanced conversational system that generates questions and analyses

Ital-IA 2024: 4th National Conference on Artificial Intelligence, organized by CINI, May 29-30, 2024, Naples, Italy
*Corresponding author.
alessandro.raganato@unimib.it (A. Raganato); francesco.bartoli@unimib.it (F. Bartoli); cristina.crocamo@unimib.it (C. Crocamo); d.cavaleri1@campus.unimib.it (D. Cavaleri); giuseppe.carra@unimib.it (G. Carrà); gabriella.pasi@unimib.it (G. Pasi); marco.viviani@unimib.it (M. Viviani)
0000-0002-7018-7515 (A. Raganato); 0000-0003-2612-4119 (F. Bartoli); 0000-0002-2979-2107 (C. Crocamo); 0000-0001-5342-9394 (D. Cavaleri); 0000-0002-6877-6169 (G. Carrà); 0000-0002-6080-8170 (G. Pasi); 0000-0002-2274-9050 (M. Viviani)

responses to automatically assess symptom severity levels. The obtained results illustrate that the proposed approaches and the best models tested have an accuracy of about 70% in making the mapping between conversation and MADRS scores, with a pretty high correlation. While not optimal, this result appears encouraging in the belief that refinements on the models (via fine-tuning) and prompts could lead to higher results and pursuit of the goal of developing a fully automated system.

2. Related Work

The urgent need for innovation around access and quality of mental health care has become clear in the last few years [4]. More and more mental health-related digital strategies for therapeutic approaches have been offered via ML and, in general, AI models, thus contributing to the development of detection systems for mental disorders, e.g., [5, 6, 7].

However, although significant progress has been made in the field, there are several barriers in the implementation of detection systems in real-world applications, including a need for increased transparency and replication [8]. Moreover, the literature is sparse with a high degree of heterogeneity between studies and the use of non-standardized metrics reporting [9]. In addition, several areas remain understudied, including the use of these approaches among people suffering from mental disorders such as depression. Nonetheless, a few studies analyzed automated approaches for evaluating depression.

A recent study trained ML models to diagnose depression from spontaneous responses of 113 outpatients using interviews by experienced physicians that were first audio-recorded and transcribed verbatim. The study showed automated depression diagnosis based on interviews as a feasible approach [10]. The use of transcribed autobiographical memory interviews was also considered for patients with treatment-resistant depression treated with psilocybin [11]. Quantitative speech measures were computed using the interview data from 17 patients and 18 untreated age-matched healthy control subjects, and an ML algorithm was developed to classify between controls and patients and predict treatment response. Results showed that speech analytics and ML successfully differentiated individuals with depression from healthy controls and identified treatment responders from non-responders with a significant level of accuracy and precision. More generally, question-based computational language assessment, based on self-reported and freely generated word responses, analyzed with AI, has been shown as a potential tool that may complement rating scales and evaluate mental health issues in clinical settings [12]. A recent systematic review highlighted preliminary favorable evidence about the use of conversational agents (i.e., tools providing feedback to user input related to well-being and mental health queries) and their promising role in screening, assessment, diagnosis, and treatment of mental disorders, including the effective identification of people with depressive symptoms [8, 13, 14]. For instance, discreet text interfaces possibly allowed participants to feel more comfortable using conversational agents in public [15].

Although these approaches appear to ensure optimal control over conversation flow and topics benefiting users and providers, a pre-defined response range may decrease usability in a diverse range of clinical settings with different risks such as possibly disrupting the therapeutic alliance [15]. Indeed, a feasible option for developing a mass screening integrated approach for early detection of depression is intended as a means of assisting with automation and concealed communication with verified scoring systems rather than replacing clinical interviews [16]. Moreover, the diversity of outcomes and the choice of outcome measurement instruments employed in studies on conversational agents for mental health point to the need for an established minimum core outcome set and greater use of validated instruments [17]. Therefore, an enhanced personalization of conversational agents leveraging the interdisciplinary use of NLP techniques to better understand the context of the conversation about vulnerable experiences related to depressive symptoms – with a more human-like approach – appears desirable [18].

3. Guiding LLMs to Automate MADRS Score Computation

LLMs are advanced AI systems [19], which possess the capability to generate human-like text across a wide range of topics, and thus seem to be the most suitable tool for solving the literature problem enunciated above. However, to accomplish a particular task, there is the need for a process for crafting specific instructions or *prompts* to guide these models; such a process is known as *prompt engineering* [20], and is gauging importance in recent years in medicine [21].

3.1. Basics of Prompt Engineering

The main prompting techniques employed today in the literature are known as *Zero-Shot* (ZS), *Few-Shot* (FS), and *Chain-of-Thought* (CoT) *learning*. In ZS learning, the LLM is provided with a prompt (describing the task to be accomplished) without any examples or specific training data for that task. Despite this, the model attempts to generate a suitable response based solely on its understanding of the task description. FS learning extends ZS by providing the model with a small number of examples

or demonstrations for the task at hand. These examples serve as additional context for the model to understand the task better. Finally, CoT prompts guide the model to generate coherent and logically connected responses by sequentially structuring the prompt. Each step of the prompt builds upon the previous one, creating a chain of thoughts that guide the model's generation process.

3.2. Automated Score Computation

Having made this necessary premise about prompt engineering, we can illustrate the two different approaches proposed in this article to perform the considered task, denoted as *local* and *global*. For both approaches, we consider ZS and CoT prompting techniques, being insufficient in the number of available examples in the considered dataset (detailed in Section 4.1) to perform FS. This means designing appropriate *prompt templates* for each prompting technique with respect to each approach.

3.2.1. Local Computation Approach

We ask LLMs appropriately guided by prompts to generate a score for each item of the MADRS. Such items and their descriptions are illustrated in Figure 1, while ZS and CoT prompt templates are detailed in the following.

1. Apparent sadness	Representing despondency, gloom, and despair (more than just ordinary transient low spirits), reflected in speech, facial expression, and posture. Rate by the depth and inability to brighten up.
2. Reported sadness	Representing reports of depressed mood, regardless of whether it is reflected in appearance or not. Includes low spirits, despondency or the feeling of being beyond help and without hope.
3. Inner tension	Representing feelings of ill defined discomfort, edginess, inner turmoil, mental tension mounting to either panic, dread or anguish. Rate according to intensity, frequency, duration and the extent of reassurance called for.
4. Reduced sleep	Representing the experience of reduced duration or depth of sleep compared to the subject's own normal pattern when well.
5. Reduced appetite	Representing the feeling of a loss of appetite compared with when well. Rate by loss of desire for food or the need to force oneself to eat.
6. Concentration difficulties	Representing difficulties in collecting one's thoughts mounting to an incapacitating lack of concentration. Rate according to intensity, frequency, and degree of incapacity produced.
7. Lassitude	Representing difficulty in getting started or slowness in initiating and performing everyday activities.
8. Inability to feel	Representing the subjective experience of reduced interest in the surroundings, or activities that normally give pleasure. The ability to react with adequate emotion to circumstances or people is reduced.
9. Pessimistic thoughts	Representing thoughts of guilt, inferiority, self-reproach, sinfulness, remorse and ruin.
10. Suicidal thoughts	Representing the feeling that life is not worth living, that a natural death would be welcome, suicidal thoughts, and preparations for suicide. Suicide attempts should not in themselves influence the rating.

Figure 1: A detail on the 10 items, with related descriptions, that constitute the MADRS.

Zero-Shot Learning. The model is simply asked to generate a score for each item of the MADRS. These items are specified in the template, as follows:

> *Given the following document containing a conversation between a physician and a patient, denoted by M and P respectively, following the Montgomery-Åsberg Depression Rating Scale (MADRS), answer me with the severity score, from a minimum of 0 (symptom absent) to a maximum of 6 (extremely severe), for the following item only:* [item title, description]. *Answer me only with a value between a minimum of 0 and a maximum of 6 related only to the described label. Below is the document to be analyzed:* [document].

This template is repeated for each of the 10 items of MADRS, and [item title, description] contains the title and description shown in Figure 1 for each item, for example: *Reduced sleep, representing the experience of reduced duration or depth of sleep compared to the subject's own normal pattern when well.* Once the scores for each item are obtained, they are simply added together to obtain the overall score.

CoT Learning. In this preliminary work, the CoT approach is based on simply asking the model to provide a motivation before performing the task. This helps the model make a more informed decision than the ZS scenario. Therefore, the CoT template used is as follows:

> [ZS "local" template] + *Provide the rationale before answering.*

Also in this case, the scores for each item are summed up to obtain the overall score.

3.2.2. Global Computation Approach

Here, LLMs are appropriately guided to directly generate the overall depression score with respect to MADRS.

Zero-Shot Learning. The ZS template employed in this global approach to computation is as follows:

> *Given the following document containing a conversation between a physician and a patient, denoted by M and P respectively, following the Montgomery-Åsberg Depression Rating Scale (MADRS), answer me with what would be the severity score with respect to depression that you would assign. The threshold values are: 0 to 6 no depression, 7 to 19 mild depression, 20 to 34 moderate depression, and 35 to 60 severe depression. Answer only with a value between*

the minimum of 0 and a maximum of 60. Below is the document to be analyzed: [document].

CoT Learning. CoT learning in the global approach uses the ZS "global" template in which reasoning is required before providing the answer:

> [ZS "global" template] + *Provide the rationale before answering.*

4. Comparative Evaluation

In this section, we present the results of the comparative evaluation of the local and global approaches, in relation to the various proposed prompt engineering techniques (and thus, regarding the different templates used). Firstly, we introduce the dataset employed in the evaluations and the technical characteristics of the implemented models.

4.1. The Conversation Dataset

It is well understood, especially in such a delicate field as psychiatry, that dealing with patient data is rather complex and ethically sensitive. For this reason, for this preliminary study, a team of medical experts generated a small dataset in which clinicians took on the roles of both the doctor and the patient. This was done to create typical conversations regarding various levels of depression severity, namely: *severe depression*, *moderate depression*, *mild depression*, and *absence of depression.* In total, 10 doctor-patient conversations were generated in Italian, with at least 3 conversations for the first three previously outlined severity levels. Clinicians also labeled the questions and answers against the corresponding items of the MADRS and provided both item-level and global scores for the entire conversation.[1]

4.2. Technical Details

To assess the effectiveness of generative models in addressing the considered problem, various LLMs were tested. These models were trained on diverse datasets, tailored for a multilingual context, given that our psychiatrist-patient conversations are in Italian. In particular, the following models were used: **GPT-3.5**: `GPT-3.5-turbo-0613`, it is an iteration of the *Generative Pre-trained Transformer* (GPT) model developed by OpenAI. It is an advanced version of its predecessor, GPT-3, with improvements in various aspects such as model architecture, training data, and fine-tuning techniques. **GPT-4**: `GPT-4-0613`, it is a large multimodal model (accepting image and text inputs, emitting text outputs).[2] **Mistral**: `Mistral-7B-Instruct-v0.2`, it is an instruct fine-tuned 7B LLM, trained mainly on English data, but also acquainted with Italian during its pretraining phase [22]. **Mixtral**: `Mixtral-8x7B-Instruct-v0.1`, it is a pretrained generative *Sparse Mixture of Experts* model, trained mainly on 5 languages including Italian. It has 46.7B total parameters but only uses 12.9B parameters per token.[3] **Dante**: `DanteLLM_instruct_7b-v0.2-boosted`, it is a recent state-of-the-art Italian LLM based on the 7B Mistral model.[4] **Hermes**: `Hermes7b_ITA`, it is a 7B LLM trained on a 120K instruction/answer dataset in Italian. It is based on `Nous-Hermes-llama-2-7b` LLM, a version of `meta/Llama-2-7b` fine-tuned to follow instructions.[5]

4.3. Results

The results obtained measure the effectiveness of the above-mentioned models, in conjunction with the appropriate prompting templates, in correctly predicting the item-level scores and overall score of each conversation compared with those assigned by the medical experts. They are illustrated in terms of *accuracy* (Acc.), *Pearson* (P.), and *Spearman* (S.) correlation coefficients.

4.3.1. Local Computation Results

Tables 1 and 2 show some results of the prompts and LLMs models applied to the local computation approach.

Table 1
Overall results for the local computation approach.

	ZS			CoT		
Model	Acc.	P.	S.	Acc.	P.	S.
GPT-3.5	0.30	0.81	0.81	0.30	0.86	0.83
GPT-4	0.30	**0.92**	0.88	0.40	**0.93**	0.90
Mistral	0.30	0.70	0.69	0.40	0.85	**0.91**
Mixtral	0.40	**0.92**	**0.91**	0.40	0.86	0.87
Dante	0.30	0.47	0.42	0.40	0.27	0.16
Hermes	0.40	0.51	0.54	**0.60**	0.31	0.15

It can be seen that from the results in Table 1, especially in terms of accuracy, the local approach does not provide satisfactory overall results. However, a substantial improvement can be appreciated when models are asked to explain the reasons for their choices (CoT), and in particular for the Hermes model. Regarding the correlation coefficients of Person and Spearman, we can observe how these are globally quite high, improving in the CoT scenario for models trained on larger amounts of data and decreasing on smaller ones.

[1] The dataset used and the respective labels and scores can be downloaded at the following address: https://drive.google.com/file/d/18HL5v8Hh2GBm1l0dt9Z8cHW0Opy8JgA7/view?usp=sharing.

[2] https://platform.openai.com/docs/models/overview

[3] https://mistral.ai/news/mixtral-of-experts/

[4] https://github.com/RSTLess-research/DanteLLM

[5] https://huggingface.co/raicrits/Hermes7b_ITA

Table 2
Correlation results for each MADRS item in the local CoT scenario.

	#1.		#2.		#3.		#4.		#5.		#6.		#7.		#8.		#9.		#10.	
Model	P.	S.	P.	S.	P.	S.	P.	S.	P.	S.	P.	S.	P.	S.	P.	S.	P.	S.	P.	S.
GPT-3.5	0.61	0.80	0.35	0.24	0.48	0.56	0.73	0.81	0.74	0.79	0.60	0.66	0.54	0.58	0.17	0.24	0.31	0.41	0.83	0.87
GPT-4	0.65	0.51	0.61	0.50	0.70	0.67	0.89	0.79	0.90	0.89	0.18	0.36	0.83	0.76	0.47	0.37	0.84	0.83	0.95	0.96
Mistral	0.15	0.20	0.64	0.78	0.53	0.21	0.71	0.79	0.21	0.20	0.40	0.54	-0.34	-0.37	0.31	0.31	0.82	0.82	0.94	0.93
Mixtral	0.46	0.48	0.91	0.88	0.73	0.43	0.76	0.69	0.84	0.90	0.21	0.35	0.72	0.64	-0.52	-0.36	0.36	0.39	0.83	0.87
Dante	-0.32	-0.49	0.49	0.66	0.68	0.75	0.47	0.50	-0.78	-0.76	-0.08	-0.08	-0.25	-0.05	-0.04	0.09	0.11	0.11	0.24	0.25
Hermes	0.57	0.56	-0.25	-0.61	0.06	0.24	0.07	0.01	-0.16	-0.22	-0.25	-0.32	0.30	0.17	0.24	0.16	0.18	0.29	-0.02	0.22

4.3.2. Global Computation Results

Table 3 shows the results of the prompts and LLMs models applied to the global computation approach.

Table 3
Overall results for the global computation approach.

	ZS			CoT		
Model	Acc.	P.	S.	Acc.	P.	S.
GPT-3.5	**0.70**	0.66	0.62	0.60	0.79	0.71
GPT-4	0.60	**0.96**	**0.94**	0.40	**0.87**	0.82
Mistral	0.20	0.47	0.23	0.60	0.22	0.51
Mixtral	0.50	0.43	0.57	0.50	0.33	0.20
Dante	0.30	-0.03	0.13	**0.70**	0.68	**0.86**
Hermes	0.30	0.31	0.47	0.50	0.76	0.64

The results in this case show that an accuracy of around 70% can be achieved. It is particularly interesting to note how the best models are the GPT-based in the ZS case, while it is Dante in the CoT case, which instead turns out to be one of the worst using a ZS technique. Person and Spearman correlation coefficient results illustrate a significant increase in correlation in the smaller models in the CoT scenario, with variable fluctuations in the case of the larger models.

4.3.3. Further Investigating Best Results

Compared to the approaches, prompt engineering techniques, and LLMs considered, it is clear that the use of the global approach is superior to the local one. This would seem to suggest that LLMs have a greater chance of success with respect to the task considered when the conversation is considered to produce the global MADRS score, without the model being asked to generate MADRS item-based scores to be later aggregated. However, we operated in a context in which we did not provide specific examples of the model according to a Few-Shot strategy, which need to be investigated in the future.

As it emerges from Table 2, referring to the local computation approach in the CoT scenario, the correlation with respect to the scores predicted in the individual items is generally not very high, although it is objectively better in some specific items such as #4 (i.e., *reduced sleep*, for the models trained on more data), #10 (i.e., *suicidal thoughts*, again for larger models). The smaller, Italian-specific models do not correlate well on this task.

Concerning Figure 2, illustrating the confusion matrix referring to the global computation approach for the Dante model performed in the CoT scenario, we can observe how the model does not confuse depression severity classes that are too distant from each other.

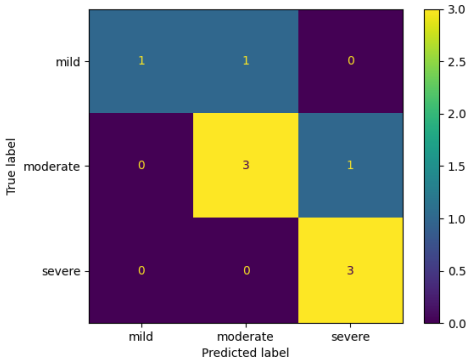

Figure 2: Dante's CoT global confusion matrix.

5. Conclusion and Future Research

This study explored the utilization of generative Artificial Intelligence (AI) models for automatically mapping psychiatrist-patient dialogue content to the Montgomery-Åsberg Depression Rating Scale (MADRS). Two distinct approaches were investigated: the application of prompt engineering techniques to compute symptom severity scores for each MADRS item, and the direct calculation of the overall depression severity score. The results demonstrated that the proposed approaches, coupled with the best-performing models, achieved an accuracy of approximately 70% in mapping conversations to MADRS scores.

Though the current accuracy shows promise, there is room for improvement. Future studies could refine models, improve prompt techniques, explore new methods, and use more data sources. This could lead to an automated system that generates questions and evaluates symptom severity from dialogue analysis.

References

[1] J. J. Silverman, M. Galanter, M. Jackson-Triche, D. G. Jacobs, J. W. Lomax, M. B. Riba, L. D. Tong, K. E. Watkins, L. J. Fochtmann, R. S. Rhoads, et al., The american psychiatric association practice guidelines for the psychiatric evaluation of adults, American Journal of Psychiatry 172 (2015) 798–802.

[2] B. Fantino, N. Moore, The self-reported montgomery-åsberg depression rating scale is a useful evaluative tool in major depressive disorder, BMC psychiatry 9 (2009) 1–6.

[3] K.-B. Ooi, G. W.-H. Tan, M. Al-Emran, M. A. Al-Sharafi, A. Capatina, A. Chakraborty, Y. K. Dwivedi, T.-L. Huang, A. K. Kar, V.-H. Lee, et al., The potential of generative artificial intelligence across disciplines: Perspectives and future directions, Journal of Computer Information Systems (2023) 1–32.

[4] J. Torous, K. J. Myrick, N. Rauseo-Ricupero, J. Firth, et al., Digital mental health and covid-19: using technology today to accelerate the curve on access and quality tomorrow, JMIR mental health (2020).

[5] M. Fokkema, D. Iliescu, S. Greiff, M. Ziegler, Machine learning and prediction in psychological assessment, European Journal of Psychological Assessment (2022).

[6] S. S. Panicker, P. Gayathri, A survey of machine learning techniques in physiology based mental stress detection systems, Biocybernetics and Biomedical Engineering 39 (2019) 444–469.

[7] M. Viviani, C. Crocamo, M. Mazzola, F. Bartoli, G. Carrà, G. Pasi, Assessing vulnerability to psychological distress during the covid-19 pandemic through the analysis of microblogging content, Future Generation Computer Systems (2021).

[8] A. N. Vaidyam, H. Wisniewski, J. D. Halamka, M. S. Kashavan, J. B. Torous, Chatbots and conversational agents in mental health: a review of the psychiatric landscape, The Canadian Journal of Psychiatry 64 (2019) 456–464.

[9] A. Viduani, V. Cosenza, R. M. Araújo, C. Kieling, Chatbots in the field of mental health: challenges and opportunities, Digital Mental Health: A Practitioner's Guide (2023) 133–148.

[10] K. Mao, Y. Wu, J. Chen, A systematic review on automated clinical depression diagnosis, npj Mental Health Research 2 (2023) 20.

[11] F. Carrillo, M. Sigman, D. F. Slezak, P. Ashton, L. Fitzgerald, J. Stroud, D. J. Nutt, R. L. Carhart-Harris, Natural speech algorithm applied to baseline interview data can predict which patients will respond to psilocybin for treatment-resistant depression, Journal of affective disorders (2018).

[12] K. Kjell, P. Johnsson, S. Sikström, Freely generated word responses analyzed with artificial intelligence predict self-reported symptoms of depression, anxiety, and worry, Frontiers in Psychology (2021).

[13] P. Philip, J.-A. Micoulaud-Franchi, P. Sagaspe, E. D. Sevin, J. Olive, S. Bioulac, A. Sauteraud, Virtual human as a new diagnostic tool, a proof of concept study in the field of major depressive disorders, Scientific reports 7 (2017) 42656.

[14] G. Dosovitsky, B. S. Pineda, N. C. Jacobson, C. Chang, E. L. Bunge, et al., Artificial intelligence chatbot for depression: descriptive study of usage, JMIR Formative Research 4 (2020) e17065.

[15] A. N. Vaidyam, D. Linggonegoro, J. Torous, Changes to the psychiatric chatbot landscape: A systematic review of conversational agents in serious mental illness: Changements du paysage psychiatrique des chatbots: une revue systématique des agents conversationnels dans la maladie mentale sérieuse, The Canadian Journal of Psychiatry 66 (2021) 339–348.

[16] P. Kaywan, K. Ahmed, A. Ibaida, Y. Miao, B. Gu, Early detection of depression using a conversational ai bot: A non-clinical trial, Plos one (2023).

[17] A. I. Jabir, L. Martinengo, X. Lin, J. Torous, M. Subramaniam, L. Tudor Car, Evaluating conversational agents for mental health: Scoping review of outcomes and outcome measurement instruments, J Med Internet Res 25 (2023).

[18] A. Ahmed, A. Hassan, S. Aziz, A. A. Abd-Alrazaq, N. Ali, M. Alzubaidi, D. Al-Thani, B. Elhusein, M. A. Siddig, M. Ahmed, et al., Chatbot features for anxiety and depression: a scoping review, Health informatics journal 29 (2023) 14604582221146719.

[19] Y. Chang, X. Wang, J. Wang, Y. Wu, L. Yang, K. Zhu, H. Chen, X. Yi, C. Wang, Y. Wang, et al., A survey on evaluation of large language models, ACM Transactions on Intelligent Systems and Technology (2023).

[20] P. Sahoo, A. K. Singh, S. Saha, V. Jain, S. Mondal, A. Chadha, A systematic survey of prompt engineering in large language models: Techniques and applications, 2024. arXiv:2402.07927.

[21] B. Meskó, Prompt engineering as an important emerging skill for medical professionals: tutorial, Journal of Medical Internet Research (2023).

[22] A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. d. l. Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier, et al., Mistral 7b, arXiv preprint arXiv:2310.06825 (2023).