LUNDS UNIVERSITET

Institutionen för psykologi
Psykologprogrammet

The Power of Words: ChatGPT's Assessments of Depression and Anxiety Using Responses to Open-ended Questions

Kraften av ord: ChatGPTs Bedömningar av Depression och Ångest med Svar på Öppna Frågor

Sophie Brunsberg, Linnea Holmlund Vidman & Ragna Jarbo

Kandidatexamensuppsats. 2024

Handledare: Sverker Sikström
Examinator: Bardur Hofgaard Joensen

Abstract

The aim of this study was to explore the accuracy of ChatGPT's numeric assessments of depression and anxiety. ChatGPT was prompted to estimate participants' levels of depression and anxiety using the participants' responses to an open-ended question about their symptoms during the past two weeks. In total, the study included responses from 876 participants. Estimated scores were compared to scores measured by the PHQ-9 and GAD-7 scale, respectively. Different prompts were tested to explore the impact of prompt design (H1), and the addition of three words describing participants' perceived reasons for their symptoms (H2). ChatGPT's performance was also compared to that of the BERT model (H3). The results demonstrated that ChatGPT successfully assessed levels of depression and anxiety, regardless of prompt design or the addition of reason. ChatGPT's performance was also shown to be comparable, and at times even surpass, the performance of BERT. Collectively the results demonstrate promising prospects for further applications of language-based assessments using ChatGPT, as an alternative to traditional rating scales. However, further research is necessary to ensure the reliability of ChatGPT's measures.

Keywords: *ChatGPT, BERT, Mental health assessments, Depression, Anxiety*

Sammanfattning

Syftet med den här studien var att undersöka ChatGPTs förmåga att uppskatta nivåer av depression och ångest. ChatGPT blev genom en prompt instruerad att uppskatta en deltagares nivå av depression och ångest, baserat på deltagarnas svar på öppna frågor om deras symptom de senaste två veckorna. Totalt inkluderade studien ordresponser från 876 deltagare. Uppskattade poäng från ChatGPT jämfördes med deltagarnas poäng på PHQ-9 respektive GAD-7 skalan. Olika prompts testades för att utforska inverkan av olika design på prompts (H1), samt av tillagt information kring orsaken bakom deltagarnas symptom (H2). ChatGPTs förmåga jämfördes med BERT-modellen (H3). Resultaten visade att ChatGPT framgångsrikt bedömde nivåer av depression och ångest, oberoende av designen på prompt eller information om orsaker kring symptomen. ChatGPTs förmåga visade sig också vara jämförbar, och ibland även överträffa BERTs prestation. Sammantaget visar resultaten lovande potential för att i framtiden använda ChatGPT som ett alternativ till traditionella skattningsskalor. Det krävs däremot vidare forskning för att säkerställa reliabiliteten av ChatGPTs bedömningar.

Nyckelord: *ChatGPT, BERT, Hälsobedömning, Depression, Ångest*

Thank you!

We want to express our deepest gratitude to our supervisor, Professor Sverker Sikström, for the opportunity to conduct our study within his research project. His guidance and mentorship helped us shape this essay to its current quality. Professor Sikströms's profound knowledge in the research area has forever inspired us. Sincerely, thank you.

The Power of Words: ChatGPT's Assessments of Depression and Anxiety Using Responses to Open-ended Questions

One out of eight people suffer from some kind of mental health disorder worldwide, of which the majority lack access to adequate treatment (WHO, 2022). In 2022, nearly one fifth (19.86%) of all Americans experienced some type of mental illness, and almost a fourth reported not receiving sufficient support (Mental Health America, 2022). For new patients, less than 20% of psychiatrists were available in 2023 and waiting times for an appointment in-person and over the phone were 67 and 43 days respectively (Sun et al., 2023). These findings emphasise the importance of accessible and accurate mental health assessments.

Commonly, rating scales are used to assess mental health. Rating scales offer a quantitative and efficient method of measuring mental health symptoms, and are used in both healthcare and psychological research. However, traditional rating scales have limitations. One criticism is that they cannot capture the full and complex nuances of the individual's experience, as they demand numeric or categorical responses (Kjell et al., 2019). Moreover, rating scales require the participants to perform the cognitive task of translating their experiences into a number on a scale. This is a burden that Kjell et al., (2019) argue should be placed on the method instead of the respondents. Comparatively, language is a more natural way of describing inner experiences and states (Kjell et al., 2022; Tausczik & Pennebaker, 2009). Language-based assessments are also experienced as more precise and are preferred by patients themselves (Sikström et al., 2023b). Recent advancements in AI technology, in particular Natural Language Processing (NLP), have presented an opportunity to make language-based assessments efficiently.

NLP is a field of computer science and pertains to a computer's ability to process, interpret and generate natural language (Chowdhary, 2020). NLP is applied by language models (LMs) such as ChatGPT, a type of machine learning model trained on extensive volumes of text data. Text input allows LM's to learn the probability distribution over words and to predict the coming word in a sentence (Ghojogh & Ghodsi, 2020; Wu et al., 2023). Using NLP, a model may be used to estimate levels of mental health numerically using a patient's description of their mental state in words. For instance, Kjell et al., (2019) have shown that responses to open-ended questions can be analysed using NLP to predict mental health scores. In Kjell et al.'s study participants were asked to respond to open-ended mental health questions using ten words. The same participants were also assessed using traditional rating scales. Using NLP, a model was trained on the participants' word-responses along with their scores on the rating scales. Predicted values generated by the model were shown to correlate strongly with participants' scores on the corresponding rating scales (Pearson's r = 0.58-0.72). In another study (Kjell et al., 2021), a model was trained on responses containing five words describing depression and anxiety. The model was able to generate numeric predictions which strongly correlated with individual criteria from the DSM (Diagnostic and Statistical Manual of Mental Disorders) (depression: r = 0.3-0.6, anxiety: r = 0.41-0.5).

Note that it is crucial to consider the amount, character, and context of the words or texts being used when leveraging assessments using natural language. While more information provided to the model is suggested to produce more accurate assessments, it depends on the model's ability to distinguish the provided words. For instance, Kjell et al. (2022) have suggested that a response containing ten words may be more informative than one containing 60 words.

BERT

Recent advancements in AI technology have further enhanced language-based assessments by leveraging so-called transformer architecture. Transformers are a type of deep-learning. Deep learning utilises so-called neural networks, which are structures consisting of layers of nodes, much like the human brain. "Deep" in deep learning refers to the amount of layers in the network. To process complex input many layers may be necessary, where each one processes a particular feature of the input data. Using training data, a network is able to learn patterns and adjust its structure to make better predictions (for more detail see Ghojogh & Ghodsi, 2020).

One feature which characterises transformers is a mechanism known as "self-attention", which enables LMs to weigh the importance of specific parts of a text, considering the context surrounding it (Ghojogh & Ghodsi, 2020). For instance, self-attention would allow a model to tell the difference between the word "blue" written in the sentence "I feel blue" and "The sky is blue", by attending to words in the sentence that provide important context (eg. "feel" and "sky"). Essentially, transformer architecture enables LMs to capture contextual and non-linear relationships in the data.

BERT (Bidirectional Encoder Representations from Transformers) (Devlin et al., 2018), created by researchers at Google in 2018, is currently the most cited transformer-based LM. In previous studies BERT and other transformer-based models have been shown to make highly accurate binary classifications of mental health states (eg. depressed vs. not depressed) and assessments of problem type (eg. depression or anxiety) (Greco et al., 2023). BERT-based AIs have been used to detect depression and/or anxiety using a variety of semantic material, such as social media posts (Jiang et al., 2020; Vajre et al., 2021) and clinical notes (Meng et al., 2021). Kjell et al. (2022) have further demonstrated that BERT is able to make estimates of psychological states (such as satisfaction and harmony in life) from word responses that correlate with rating scales and are reaching theoretical upper limits ($r = 0.85$). Considering this, BERT serves as a high standard to which future models can be compared.

ChatGPT

ChatGPT (Generative Pre-trained Transformer), developed by OpenAI, is one of the most used interactive LMs worldwide. While interpretative LM's such as BERT are characterised by their ability to understand and analyse text numerically, interactive LM's are designed to interpret and interact with the user through human-like dialogue.

ChatGPT operates through a chatbot interface where the user submits a prompt, which can be described as a question or instruction, which is then followed by a text response. The words in the input are translated into a set of vectors from where the model extracts their meaning. Text is then generated by predicting one word at a time using the probability distribution learned during training (Ghojogh & Ghodsi, 2020; Wu et al., 2023). OpenAI's models are trained on data from a variety of sources, including books, news articles, scientific journals, and data publicly available on the web (Ray, 2023). Although the exact size of ChatGPT's training corpus is undisclosed, it is assuredly one of the largest in the field (Ray, 2023). Using its accumulated knowledge in the domain of psychology, ChatGPT may pose as a tool for mental health assessment. While BERT would require the true scores of a sample of participants to predict scores, ChatGPT merely utilises its previous knowledge to perform a task specified by a prompt.

In previous research, ChatGPT has mainly been explored as a conversational agent, and few articles have examined its ability to identify mental health conditions. While ChatGPT seems less appropriate as a virtual clinician (Dergaa et al., 2024), it has been found to successfully identify clinical mental health conditions when given case descriptions (D'Souza et al., 2023). ChatGPT has also been shown to successfully prescribe appropriate treatments for mild and severe depression at a higher rate than primary care physicians (Levkovich & Elyoseph, 2023). However, Grabb (2023) has argued that previous studies on

ChatGPT have overlooked the importance of prompt design. ChatGPTs responses rely on the formulation of the prompt, making prompt design a variable of performance (Grabb, 2023; Ekin, 2023; Jiang et al., 2020). Aspects of a prompt that may influence ChatGPT are for example length, verbiage, detail and specificity (Grabb, 2023; Ekin, 2023).

While previous research has studied ChatGPT's conversational abilities, there have not been, to our knowledge, any studies conducted to explore its ability to assess mental states numerically. Even less so ones that take prompt design into account. This brings us to the present study.

Present study

We aim to research how accurately ChatGPT can estimate scores for depression and anxiety using answers to open-ended questions. In accordance with the previously presented findings, three hypotheses have been formulated. Since studies have underlined the importance of prompt design (Grabb, 2023; Ekin, 2023), an initial hypothesis will address the impact of different conceptualizations of depression and anxiety instructed in the prompt. The second hypothesis will explore the impact of information provided in the word response. We hypothesise that ChatGPT's responses will improve with the addition of a three-word description of the reason behind their depression or anxiety. The third hypothesis will compare ChatGPT's performance with that of the BERT model. This is done to evaluate ChatGPT's overall performance.

How accurately can ChatGPT estimate scores for depression and anxiety using answers to open-ended questions?

We hypothesise that:

H1: ChatGPT's performance will depend on the design of the prompts.

H2: ChatGPT's performance will improve with the addition of reasons for perceived depression or anxiety.

H3: ChatGPT's performance will be comparable to that of BERT.

Method

Participants

This study used data collected by Sikström, Kelmendi & Persson (2023a) in their study titled "Assessment of depression and anxiety in young and old with a question-based computational language approach". The final dataset consisted of 876 participants with English as their native language. 7 participants had formerly been removed because they either failed to follow the instructions or failed to answer the control question correctly. 457 participants were recruited from the Mechanical Turk (www.mTurk.com) platform, and the rest (N=419) were recruited from the Prolific Academic (https://prolific.co/) platform. Half of the participants were recruited by screening for major depressive disorder (MDD) or Generalised Anxiety Disorder (GAD) using the screening tool SDAS. The other half were not screened, but were also assessed using SDAS after recruitment to control for an eventual MDD or GAD diagnosis. In total, 61 participants had MDD alone, 137 had GAD alone, and 259 had both diagnoses according to the SDAS. The participants' ages ranged from 18 to 70 (M=35.5, SD=11.9). 538 of the participants identified as female (61.4%), 327 as male (37.3%), and 11 as "other gender" (1.3%). The participants' partaking lasted for approximately 20 minutes and they received 4 USD for their participation. Before data collection, the participants were obliged to sign a declaration of informed written consent. They were informed that their responses would be anonymized before analysis and that they had the right to withdraw from the study at any time. (For details see Sikström et al, 2023a.)

Material

Open-ended questions and word responses

The participants were asked a series of open-ended questions about mental health (for details about the data and its collection see; Sikström et al. 2023a) of which four were used in the current study. Two questions were asked regarding symptoms of mental health: "During the last two weeks describe in words whether you felt depressed or not", and "During the last two weeks, describe in words whether you have felt worry or not". Note that worry is used interchangeably with anxiety in this study, since worry was the term used in the initial data collection. Frequent worrying is also considered a symptom of GAD, and worry/anxiety has therefore commonly been used to refer to the same construct (Kjell et al., 2021; Kumar et al., 2019). Additionally, participants were asked two questions concerning the reason for these symptoms: "Describe the reason for your depression/worry in descriptive words". The respondents were asked to answer with one word per text box. For the descriptive questions there were five text boxes, and for the reason questions there were three. To summarise, eight words total were gathered for each participant; one five-word-response describing symptoms of depression/worry and one three-word response describing the perceived reasons.

Rating scales

To measure each participant's degree of depression, the 9-item Patient Health Questionnaire (PHQ-9) was used. The PHQ-9 scale is a widely used measure for diagnosing and assessing the severity of major depression, based on criteria from the DSM-5. The questionnaire includes nine self-administered questions assessing the frequency of depression symptoms during the past two weeks. Each question is scored on a scale from 0 (not at all) to 3 (nearly every day), resulting in a total score between 0 and 27. Scores between 0-4 indicate no depressive symptoms, 5-9 indicate mild depression, 10-14 indicate moderate depressive symptoms, 15-19 indicate moderate-severe depressive symptoms and 20-27 indicate severe depressive symptoms. (Kroenke et al., 2001)

To measure each participant's degree of anxiety, the 7-item Generalised Anxiety Disorder Questionnaire (GAD-7) was used. The GAD-7 scale is commonly used to evaluate the severity of anxiety and to diagnose Generalised Anxiety Disorder. The scale assesses the frequency of anxiety symptoms experienced over the past two weeks through seven self-administered questions. Each question is scored on a scale from 0 (not at all) to 3 (nearly every day), resulting in a total score ranging from 0 to 21. Scores between 0-4 indicate minimal anxiety, 5-9 indicate mild anxiety, 10-14 indicate moderate anxiety, and lastly, 15-21 indicate severe anxiety symptoms. (Spitzer et al., 2006)

ChatGPT-4 turbo

The ChatGPT-4 turbo version is the most recent model released by OpenAI and is trained on data from up to December 2023 (OpenAI, n.d.). The model is available for paying customers on OpenAI's website.

Prompts

To explore the influence of prompt design, three different prompts using three different conceptualisations of depression/worry were formulated. An initial prompt simply referring to depression and worry as constructs was created (Prompt 1):

> "In a study, a participant used five words to answer the following question: During the last two weeks, describe in words whether you felt depressed/have felt worry or not. Estimate the degree of depression/worry based on this five-word response on a scale from 0 (not at all depressed/worried) to 10 (maximally depressed/worried). Only provide a single number in your response."

This initial prompt aims to explore if ChatGPT can accurately conceptualise these constructs and make estimations that correlate with the rating scales. As this prompt does not explicitly provide ChatGPT with a definition of depression or worry, this prompt is expected to produce a less accurate response. A scale of 0-10 was used for simplicity. The instruction to "only provide a single number in your response" was used to avoid descriptive word responses and scores presented as a range.

Secondly, the initial prompt was altered to instruct ChatGPT to make the estimation based on the definition of depression and worry as described in the DSM-5. The following prompt was formulated (Prompt 2):

> "In a study, a participant used five words to answer the following question: During the last two weeks, describe in words whether you felt depressed/have felt worry or not. Based on this five-word-response, estimate the degree of depression/worry as defined in the DSM-5. The estimation should be on a scale from 0 (not at all depressed/worried) to 10 (maximally depressed/worried). Only provide a single number in your response."

The instruction provides a source to how depression and worry should be interpreted, which may improve ChatGPT's performance. The DSM also includes the definition of depression/worry used in the rating scales, against which ChatGPT's estimates will be compared.

Thirdly, a prompt was constructed instructing ChatGPT to make their depression/worry estimates based on the rating scales used in this study, PHQ-9 and GAD-7. The following prompt was formulated (Prompt 3):

> "In a study, a participant used five words to answer the following question: During the last two weeks, describe in words whether you felt depressed/have felt worry or not. Based on this five-word-response, estimate the degree of depression/worry according to the PHQ-9/GAD-7 scale. The estimation should be on a scale from 0 (not at all depressed/worried) to 27/21 (maximally depressed/worried). Only provide a single number in your response."

This third prompt refers directly to the rating scales with which ChatGPT's performance will be evaluated. This provides important context specific to the task, and is expected to further improve ChatGPT's estimations. To avoid confusion, ChatGPT was instructed to follow the same scales as the rating scales (0-27 respectively 0-21) in this prompt. We expect the prompt referring to the PHQ-9/GAD-7 to perform the best, and the prompt referring simply to the undefined constructs to perform the worst.

In addition to prompt design, ChatGPT's performance may be influenced by the word responses provided as input. All three prompts were thus also altered to include the addition of words describing reasons for one's depression/worry:

> "In a study, a participant used five words to answer the following question: During the last two weeks, describe in words whether you felt depressed/have felt worry or not. The participant also used three words to answer the following question: Describe the reason for your depression/worry in descriptive words…"

Using this prompt, ChatGPT's estimations are expected to be more accurate.

With three different prompt designs (construct, DSM-5 and PHQ-9/GAD-7) and four versions of each (depression with/without reason, worry with/without reason), this resulted in twelve prompts in total.

Application Programming Interface

An API (Application Programming Interface) was set up to automate the submission of each prompt and each participants' word-response. The API was programmed to submit one word-response at a time along with the prompt, until a full set (n=876) was generated.

For prompts using both descriptions of and reasons for perceived depression/worry, two word-responses, instead of one, were submitted along with the prompt. Each run of the API would be equivalent to submitting one prompt and one/two word-responses into the ChatGPT-4 turbo interface 876 times, each time in a new chat. Ensuring that each submission was made in a new chat was done to prevent ChatGPT from being influenced by previous submissions and responses. Each of the twelve prompts was run three times, and an average was calculated. This was done to account for potential variations between runs, as ChatGPT does not produce identical results for runs of the same prompt. With the same purpose, the temperature parameter, a setting adjusting the randomness of ChatGPT's responses, was set to 0.

Procedure

The API was run three times for all twelve prompts, generating a total of 36 sets of 876 estimated scores. Each of the twelve prompts was also run once only for the five first participants, this time omitting the instruction to provide only a single number. When this is done, ChatGPT provides a text-response, describing its approach. This was done to allow ChatGPT to demonstrate its reasoning for each prompt and to ensure that it follows the given instructions.

To compare ChatGPT's estimations with the BERT model, the online software Semantic Excel (https://semanticexcel.com/) was used. Under the train function, the model BERT-base-uncased was chosen along with the setting BERT preprocessed with SVD. BERT-base-uncased is pre-trained on a large corpus of English text from BookCorpus and English Wikipedia (Geetha & Renuka, 2021). The SVD preprocessing is a matrix factorization technique and was applied to reduce the risk of overfitting (Hsu et al., 2022). In Semantic Excel, BERT was trained on the PHQ-9/GAD-7 scores and participants' word-responses using 10%-leave-out cross-validation. Four sets of cross-validated predicted values (n=876) were gathered using five-word responses of symptoms of depression (1) or worry (2), and using both the five-word responses and the three-word responses of reasons for depression (3) or worry (4).

Data analysis

For the data analysis the open-source computer program Jamovi was used. The accuracy of each prompt and model's performance was measured as their correlation to participants' true PHQ-9/GAD-7 scores. Correlations were calculated between participants' true PHQ-9/GAD-7 scores and (1) ChatGPT's estimated depression/worry scores for each prompt design, (2) between ChatGPT's estimated depression/worry scores with/without the addition of reasons, and (3) between BERT's predictions with/without the addition of reason.

A Steiger's z-test (Steiger, 1980) was performed to measure whether the correlations from different models and prompts were significantly different from each other. Steiger's Z was chosen to correct for dependency, since every correlation being compared has one variable (participants' true PHQ-9/GAD-7 scores) in common. Comparisons using Steiger's Z were made (1) between the three different prompt designs, (2) between ChatGPT's estimations with/without reason, and (3) between ChatGPT's estimations and BERT's.

Ethics

In the original study (Sikström, Kelmendi & Persson, 2023a), participants were informed beforehand about the purpose of the study and the potentially sensitive nature of the questions they would be required to answer. Participants also signed a declaration of informed consent, informing participants that they could withdraw from the study at any time, that their responses would be anonymized and may be used in future projects. The study was reviewed by the Swedish Ethical Review Authority (EPN), who determined no ethical approval was needed since the participants were anonymously recruited and tested (Dnr 2020-007). No further ethical approval was needed to reuse the data for the current analysis.

Results

Descriptive statistics

Table 1 and 2 show descriptives for ChatGPT's estimates of depression/worry using the three different prompt designs with and without the addition of reasons, as well as BERT's predicted PHQ-9/GAD-7 scores with and without reason. ChatGPT's estimated scores are within the range of values allowed by the prompt (0-10, 0-27 or 0-21). BERT's predicted scores fall into negative numbers in most minimums, and are thus not within the correct ranges demanded by the rating-scales (0-27 or 0-21).

Table 1

Descriptives of participants' true PHQ-9-scores, Chat-GPT's estimations of depression using the three different prompts with/without reason, and BERT's predicted PHQ-9 scores with/without reason.

	PHQ-scores	Dep. con*	Dep.con .reas.*	Dep. DSM*	Dep. DSM. reas.*	Dep. PHQ	Dep. PHQ. reas	BERT. PHQ	BERT. PHQ. reas.
N	876	876	876	876	876	876	876	876	876
Mean	12.0	5.30	5.07	5.19	4.91	11.6	9.85	12.0	12.0
Median	12.0	7.00	7.00	7.00	7.00	12.8	13.3	13.4	12.9
Mode	0.00	8.00	7.00	8.00	7.00	0.00	14.0	17.6	15.9
SD	7.78	3.28	2.97	3.29	3.01	9.55	7.03	5.09	5.10
Min.	0	0.00	0.00	0.00	0.00	0.00	0.00	-0.516	-0.266
Max.	27	10.0	9.00	10.0	9.00	27.0	27.0	22.3	24.2

Note: PHQ-score: participants' actual PHQ-9 scores. Dep.con/Dep.con.reas.: ChatGPT-4's estimations generated by prompt 1 referring to constructs without/with reason. Dep.DSM/Dep.DSM.reas.: ChatGPT-4's estimations generated by prompt 2 referring to DSM-5 without/with reason. Dep.PHQ/Dep.PHQ.reas.: ChatGPT-4's estimations generated by prompt 3 referring to PHQ-9 without/with reason. BERT.PHQ/BERT.PHQ.reas.: BERT's predicted PHQ-9 scores without/with reason. There are no missing values.

*Note that these variables are on a scale of 0-10, whereas "PHQ-scores", "Dep-PHQ/Dep.PHQ.reas." and "BERT.PHQ/BERT.PHQ.reas." are on a scale of 0-27.

Table 2

Descriptives of participants' true GAD-7 scores, Chat-GPT's estimations of worry using the three different prompts with/without reason, and BERT's predicted GAD-7 scores with/without reason.

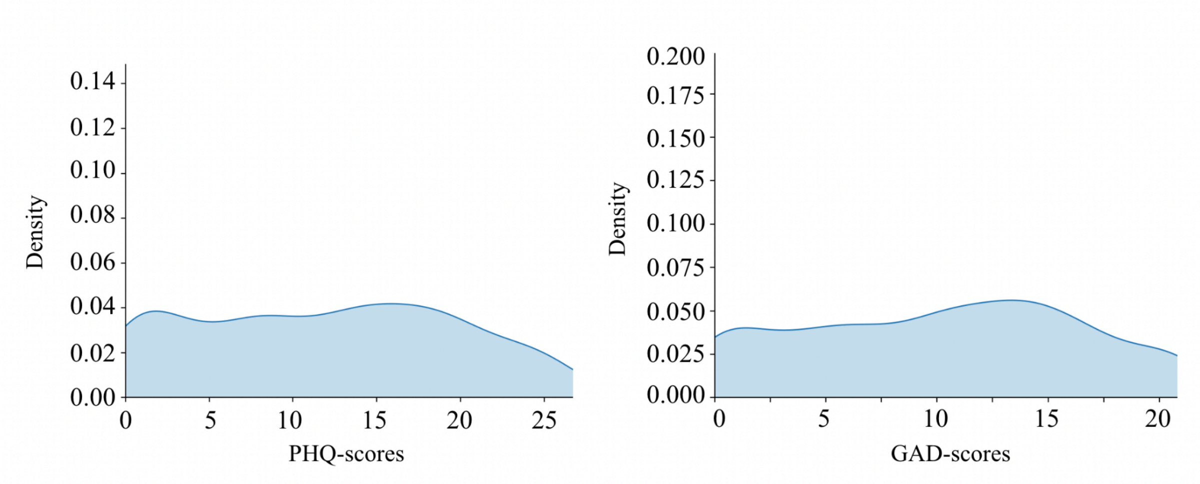
	GAD-scores	Wor. con*	Wor.con .reas.*	Wor. DSM*	Wor. DSM. reas.*	Wor. GAD	Wor. GAD. reas	BERT. GAD	BERT. GAD. reas.
N	876	876	876	876	876	875	875	876	876
Mean	10.2	6.61	6.31	6.43	6.13	13.0	11.4	10.2	10.1
Median	11.0	8.00	7.00	8.00	7.00	15.0	13.3	11.0	11.1
Mode	0.00	8.00	7.00	8.00	7.00	21.0	15.0	10.4[b]	10.6[b]
SD	6.25	2.86	2.31	3.03	2.55	7.10	5.34	3.24	3.52
Min.	0	0.00	0.00	0.00	0.00	0.00	0.00	0.178	-0.976
Max.	21	10.0	9.00	10.0	9.00	21.0	21.0	16.3	16.2

Note: GAD-score: participants' actual GAD-7 scores.Wor.con/Wor.con.reas.: ChatGPT-4's estimations generated by prompt 1 referring to constructs without/with reason.Wor.DSM/Wor.DSM.reas.: ChatGPT-4's estimations generated by prompt 2 referring to DSM-5 without/with reason. Wor.GAD/Wor.GAD.reas.: ChatGPT-4's estimations generated by prompt 3 referring to GAD-7 without/with reason. BERT.GAD/BERT.GAD.reas.: BERT's predicted GAD-7 scores without/with reason. There are no missing values.
*Note that these variables are on a scale of 0-10, whereas "GAD-scores", "Wor-GAD/Wor.GAD.reas." and "BERT.GAD/BERT.GAD.reas." are on a scale of 0-21.
[b]More than one mode exists, only the first is reported.

Neither the true scores nor the estimated depression and worry scores for any model or prompt meet the criteria for normality (Shapiro Wilk's $p < 0.05$, non-linear Q-Q plots). Since none of the data sets follow a normal distribution, Spearman's rho was used for all correlation matrices. Both participants' true PHQ-9 and GAD-7 scores follow a more or less uniform distribution, with a thinner right tail (Figure 1). The number of participants with the mode score on the PHQ-9 (Mo=0) and GAD-7 (Mo=0) scale is 66 and 75 respectively.

Figure 1

Density plots for PHQ-9 and GAD-7 scores as measured by questionnaire.



Distributions for depression scores generated by prompts one and two without reason are bimodal with peaks towards zero and between 7-8, whereas prompt three is multimodal, with peaks towards zero, 14, and the max (Figure 2). The number of participants with the mode depression score using prompt one (Mo=8), two (Mo=8) and three (Mo=0) without reason is 248, 207 and 208 respectively. With reasons added, distributions are slimmer. The number of participants with the mode score using prompt one (Mo=7), two (Mo=7) and three (Mo=14) with reason is 426, 387 and 267 respectively.

Similarly, distributions for worry scores generated by prompts one and two are bimodal with peaks towards zero and 8-7, whereas prompt three is multimodal, with peaks towards zero, 15, and the max (Figure 2). When reasons are added, distributions are again slimmer. The number of participants with the mode worry score using prompt one (Mo=8), two (Mo=8) and three (Mo=21) is 363, 306 and 203 respectively. With reasons added, the number of participants with the mode score using prompt one (Mo=7), two (Mo=7) and three (Mo=15) with reasons is 448, 380 and 299 respectively. Distributions for BERT are bimodal, with one peak between 0-5 and another between 17-18 (PHQ) or 10-11(GAD) (Figure 3).

Figure 2

Density plots for ChatGPT's estimated depression and worry scores using the three different prompts.

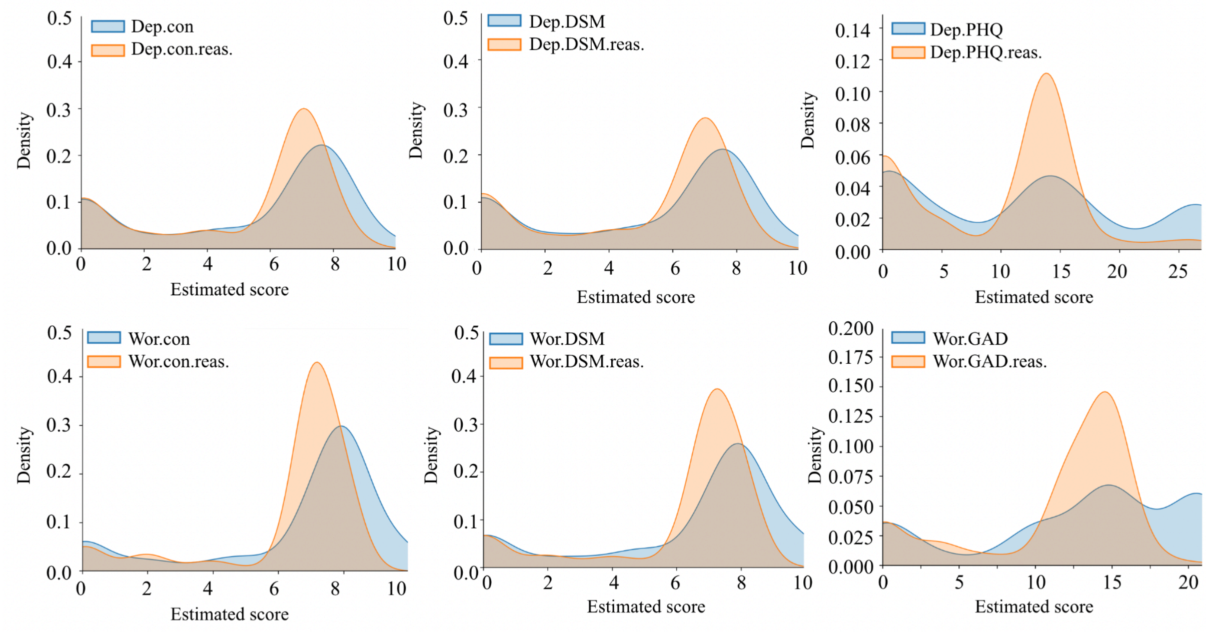
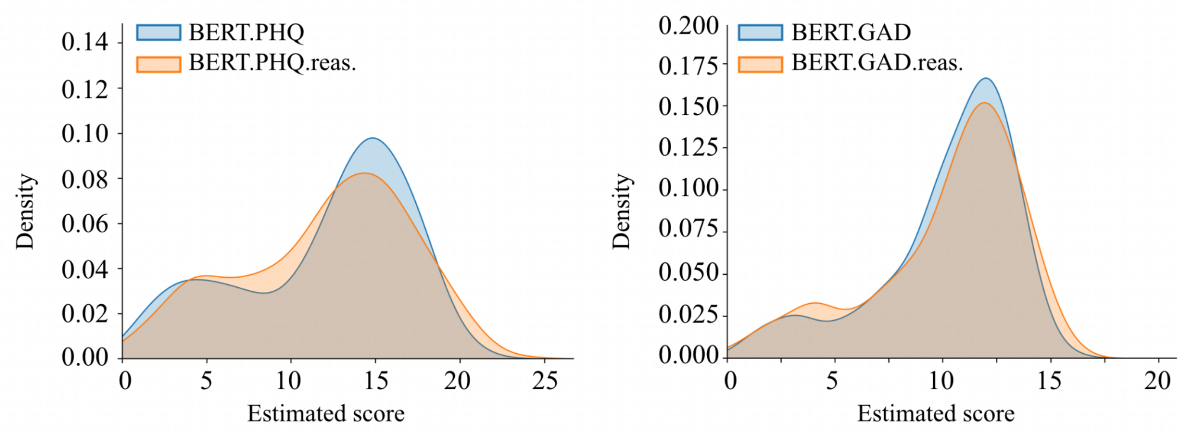


Figure 3

Density plots for BERT's predicted PHQ-9 and GAD-7 scores.



H1: Prompt Design

Chat-GPT's estimations for all three prompts are shown to strongly correlate to the participants' true PHQ-9 ($\rho = 0.612$-0.619, $p < 0.001$) and GAD-7 scores ($\rho = 0.514$-0.527, $p < 0.001$) (Tables 3 & 4). All three prompts correlate strongly with each other ($\rho = 0.912$-0.979, $p < 0.001$) and differences between them are not significant according to Steiger's z-tests ($p > 0.05$). This applies to both depression and worry.

Table 3

Spearman's rho correlation matrix between Chat-GPT-4's estimations of depression using the three different prompts and participants' true PHQ-9-scores as measured by questionnaire.

	PHQ-score	Dep.con	Dep.DSM	Dep.PHQ
PHQ-score	–			
Dep.con	0.619**	–		
Dep.DSM	0.616**	0.979**	–	
Dep.PHQ	0.612**	0.953**	0.963**	–

Note: **: p= <.001

Table 4

Spearman's rho correlation matrix between Chat-GPT-4's estimations of worry using the three different prompts and participants' true GAD-7-scores as measured by questionnaire.

	GAD-score	Wor.con	Wor.DSM	Wor.GAD
GAD-score	–			
Wor.con	0.527**	–		
Wor.DSM	0.521**	0.957**	–	
Wor.GAD	0.514**	0.912**	0.932**	–

Note: **: p= <.001

H2: Added Reasons

Correlations between participant's true PHQ-9/GAD-7 scores, ChatGPT with/without reason and BERT with/without reason can be found in table 5 and 6. The hypothesis investigates whether correlations to the true scores are higher with the addition of words describing the reasons behind participants' mental health. The results show no significant difference in performance with and without reason, for neither depression nor worry (Table 7). The same applies for the BERT model in the case of depression (Steiger's Z = -1.6, p = 0.11), but not for worry (Steiger's Z = -1.96, p = 0.049) (Table 7).

Table 5

Spearman's rho correlation matrix between Chat-GPT-4's scores with the addition of reason, BERT's predicted PHQ-9 scores with/without reason and participants' true PHQ-9-scores.

	PHQ-score	Dep.con. reas.	Dep.DSM .reas.	Dep.PHQ. reas.	BERT. PHQ. reas.	BERT. PHQ
PHQ-score	–					
Dep.con.reas.	0.624**	–				
Dep.DSM.reas.	0.621**	0.961**	–			
Dep.PHQ.reas	0.596**	0.909**	0.919**	–		
BERT.PHQ. reas.	0.606**	0.822**	0.826**	0.794**	–	
BERT.PHQ	0.579**	0.800**	0.809**	0.781**	0.797**	–

Note: **: p= <.001

Table 6

Spearman's rho correlation matrix between Chat-GPT-4's estimated GAD-7 scores, BERT's predicted GAD-7 scores and participants' true GAD-7 scores.

	GAD-score	Wor.con. reas.	Wor.DSM .reas.	Wor.GAD .reas.	BERT. GAD. reas.	BERT. GAD
GAD-score	–					
Wor.con.reas.	0.511**	–				
Wor.DSM.reas.	0.530**	0.932**	–			
Wor.GAD.reas	0.530**	0.847**	0.882**	–		
BERT.GAD. reas.	0.466**	0.675**	0.693**	0.692**	–	
BERT.GAD	0.427**	0.651**	0.669**	0.654**	0.779**	–

Note: **: p= <.001

Table 7

Comparisons of correlations to participant's true PHQ-9/GAD-7 scores between prompts with/without reason, and between BERT and ChatGPT with/without reason, using Steiger's Z.

Depression		Test Statistic z	p-value
No reason/Reason			
Dep.con.	– Dep.con.reas.	-0.6	0.552
Dep.DSM	– Dep.DSM.reas.	-0.6	0.549
Dep.PHQ	– Dep.PHQ.reas.	1.12	0.262
BERT.PHQ	– BERT.PHQ.reas.	-1.6	0.11
BERT/ChatGPT			
BERT.PHQ	– Dep.con	-3.1	0.002*
	– Dep.DSM	-2.85	0.004*
	– Dep.PHQ	-2	0.045*
BERT.PHQ.reas.	– Dep.con.reas.	-1.42	0.155
	– Dep.DSM.reas.	-1.02	0.309
	– Dep.PHQ.reas.	0.63	0.53
Worry		**Test Statistic z**	**p-value**
No reason/Reason			
Wor.con.	– Wor.con.reas.	0.89	0.373
Wor.DSM	– Wor.DSM.reas.	-0.53	0.595
Wor.GAD	– Wor.GAD.reas.	-0.94	0.345
BERT.GAD	– BERT.GAD.reas.	-1.96	0.049*
BERT/ChatGPT			
BERT.GAD	– Wor.con	-4.19	<0.001**
	– Wor.DSM	-4.28	<0.001**
	– Wor.GAD	-3.78	<0.001**
BERT.GAD.reas.	– Wor.con.reas.	-2.78	0.005*
	– Wor.DSM.reas.	-3.72	<0.001**
	– Wor.GAD.reas.	-3.72	<0.001**

(Steiger, 1980; Hoerger, 2013)

H3: ChatGPT vs. BERT

The results indicate that the accuracy of ChatGPT's estimated scores are comparable, and at times even surpass those of BERT. Without the addition of reason, ChatGPT's correlations ($\rho = 0.514$-0.619) are significantly higher than BERT's ($\rho = 0.427$, 0.579) for both depression ($p < 0.05$) and worry ($p < 0.001$) (Table 7). With reasons, ChatGPT's correlations are only significantly higher than BERT's for worry ($p \leq 0.005$), but not for depression ($p \geq 0.155$) (Table 7).

Text-responses

The text responses gathered for the five first participants for each prompt were examined. Across all prompts for both depression and worry, ChatGPT's reasoning remains similar in its responses. Example responses for prompts one, two, and three with/without reasons are provided in appendix A and B. ChatGPT appears to make its estimates mainly by assessing the type and severity of the sentiment expressed by the five-word response at hand. For prompt two, the DSM-manual is indeed often mentioned, but its specific criterias or contents are not as often reviewed or applied. In prompts referring to PHQ-9 and GAD-7 however, ChatGPT not only mentions the questionnaires, but occasionally also attempts to assess the *frequency* of the sentiments expressed by the five-word-responses. This is attempted by assigning a value from "not at all" (0) to "nearly every day" (3) to the five-word response, according to the questions asked in PHQ-9 and GAD-7 (Appendix A). PHQ-9 and GAD-7 questions that are not answered by the five-word response are judged as a 0 ("not at all") by ChatGPT, and when a score of 1 or 2 is set, it is unclear how ChatGPT was able to interpret the frequency of the symptom. These numbers of frequency are subsequently added together by ChatGPT, resulting in a final score. For prompts including reasons, ChatGPT successfully addresses the additional three words and considers their impact (Appendix B). Scores provided in all text responses generally match the scores provided in the full runs where no text-response was required. Since ChatGPT does not provide identical scores between runs, some variance is to be expected.

Discussion

H1: Prompt Design

Our first hypothesis explores if and how ChatGPT's performance varies across prompts: (1) referring to depression or worry as undefined constructs, (2) referring to the DSM, and (3) referring to the PHQ-9 or the GAD-7 scale. According to our hypothesis, prompt one is expected to produce the least accurate responses, and prompt three is expected to produce the most accurate responses. This was based on previous research which indicates that ChatGPT's text responses depend on, among other things, the amount of relevant context given in the prompt (Grabb, 2023; Ekin, 2023). Our results, however, demonstrate no significant differences between prompts' performance, and the alternative hypothesis is thereby rejected.

However, while all prompts correlate similarly to the true scores, possible differences can be observed by examining the distributions of scores generated by the prompts. As seen in figure 2 and 3, distributions for prompt one and two are similar in shape, whereas prompt three differs significantly for both depression and worry. Indeed, prompt three also correlates less strongly with the two others (Table 3 & 4). When reasons are added, this is more evident (Table 5 & 6).

To explore possible explanations for this finding, ChatGPT's text responses were examined. As seen in Appendix A-B, ChatGPT's reasoning remains similar across prompts. Although the DSM-5 is sometimes mentioned in responses to prompt two, little meaningful information is extracted from it. This may explain why ChatGPT's performance does not improve when the instruction to refer to the DSM-5 is included, and why prompt one and two's distributions are similar. This may also imply that ChatGPT is not using DSM-5 as instructed, which could be a source of error in this study.

In responses to prompt three, ChatGPT sometimes interprets the instruction to mean that each word in a word-response should be matched to specific items listed in the PHQ or GAD questionnaire. ChatGPT approaches this by matching individual words to a specific item or by attempting to assign a frequency (from 0 "never" to 3 "nearly every day") to the word. This is irrational since the word response is not guaranteed to include this type of information. For instance, some symptoms assessed by PHQ-9, such as sleep-patterns and appetite, are less likely to appear in the five-word response. Ultimately, it remains unclear why the distribution of scores using prompt three differs from prompt one and two. Although

ChatGPT's text responses indicate that prompt three produces inconsistent approaches to the task, we are not able to conclude that the observed differences are explained by this.

In opposition to our hypothesis, it appears that prompt two and three do not improve ChatGPT's understanding of the task, and in the case of prompt three, even produces erroneous types of reasoning at times. ChatGPT's performance may have benefitted from further instructions as to how PHQ-9/GAD-7 and the DSM-5 should be considered when estimating a score. Despite these observed differences, no significant difference is found in the prompts' correlation to participant's true scores.

H2: Added Reasons

The second hypothesis explores whether the performance of ChatGPT improves with additional information (Table 5 & 6). As evaluated by the correlations between estimated and actual scores, there is no significant difference between ChatGPT's performance with or without adding the reason behind the depression/worry (Table 7). Our alternative hypothesis is thereby rejected. In line with Kjell's suggestion (2023) that the impact of additional words depends on their unique contribution, a possible explanation could be that the words describing reasons behind depression/worry do not provide any meaningful additional information. However, ChatGPT does indeed seem to recognize and consider the impact of the provided reasons in its text-responses (Appendix A & B).

Despite non-significant differences between correlations, differences in the average scores and standard deviations between prompts with/without reasons were observed (see Table 1 & 2). These differences are also visible in the distribution plots (Figure 2). A non-parametric ANOVA (Friedman) was performed to test differences in averages for significance. The Freidman's ANOVA confirms that ChatGPT's estimated scores are comparatively lower with the addition of reason. This applies to all prompts ($p<0.001$, see Appendix C & D), indicating that ChatGPT underestimates participants' degree of depression/worry when reasons are provided. This is in contrast to the BERT model, where no significant differences were found (PHQ: $p= 0.589$, GAD: $p=0.787$, Appendix C & D).

It remains unclear why ChatGPT underestimates the average and variance of participants' depression/worry scores when reasons are added. No indication of a possible explanation is found in ChatGPT's text responses. To explore this further, a more thorough exploration of ChatGPT's reasoning should be conducted. In addition, future research should explore how prompts can be better formulated to incorporate the addition of reasons for the purpose of these types of assessments.

H3: ChatGPT vs. BERT

Lastly, in agreement with the third hypothesis, ChatGPT's performance is comparable, and even succeeds that of BERT. According to comparisons using Steiger's Z, ChatGPT performs better than BERT for all prompts without reasons added (Table 7). However, when descriptions of reason are added, this significant difference is no longer found. This may be explained by BERT benefiting more, although insignificantly, from reasons than ChatGPT (Table 7), resulting in a smaller difference in performance. In other words, ChatGPT only performs significantly better than BERT when reasons are not given. Future research should therefore explore the optimal type and amount of input for each model.

Despite performing equally as well, and at times better than BERT, ChatGPT appears to underestimate the variance of scores to a higher degree than BERT. This is especially clear when reasons are added (Table 1 & 2). At most, ChatGPT assigns more than 50% of the participants (n=448) the same score (Mo=7), whereas BERT distribution appears less slim. While this does not impair ChatGPT's performance when assessing mental health in a larger participant sample, as measured by our correlations, it could be a limitation if ChatGPT were applied for individual assessments.

Implications, Limitations & Future Research

The results collectively demonstrate that ChatGPT can estimate levels of depression and anxiety with high accuracy. In all instances, it remains robust against variations in prompt design and form of input, achieving accuracies comparable to those of BERT. Using word-responses as input, assessments using ChatGPT may lift the burden off of the patient to convert their experiences into numerical values. Word-responses may also more accurately convey the complexity of mental health symptoms, and are preferred by patients (Sikström et al., 2023b). Not only do the results align with previous studies demonstrating the potential of transformer-based LMs for mental health assessments (Kjell et al., 2022; Greco et al., 2023), but they are also unique in demonstrating the potential of interactive language models for assessments using word-responses. With no need for task-specific training, this suggests promising potential for utilising interactive language models for more accessible assessments in the future.

However, these results should be interpreted in the context of their potential limitations. The first limitation is due to non-random recruitment of the participants, which limits the generalizability of this study. A second limitation is that ChatGPT appears to not consistently follow the instructions provided in the prompts. For instance, it sometimes overlooks the reference to the DSM manual, which could lead to unreliable assessments and is a problem of validity (Appendix A & B). The third limitation, as mentioned in the method, is the risk of inconsistencies between runs. Currently, ChatGPT is not able to produce identical responses to the same prompt submitted in different chats at different times. This limitation was minimised by doing three runs for each prompt and calculating a mean score. However, this remains a limitation, should ChatGPT be applied for assessments clinically. A last limitation to this study, as previous research has pointed out (Kjell et al., 2023), is that the underlying architecture and training data behind ChatGPT are not publicly accessible and often change. This limits a full understanding of how the model functions, which brings challenges in employing ChatGPT clinically. It is therefore desirable to gain further insight into the architecture behind the model in order to ensure its reliability and ethical use.

Future research should continue the exploration of how ChatGPT can be used to assess mental health numerically, and what impact prompt design has in that process. To assess this, qualitative research may be beneficial. Furthermore, the optimal amount and type of input for ChatGPT should be explored. Although our results show that added reasons, and subsequently more words, do not improve ChatGPT's performance, it may benefit from other types of input. As shown by ANOVAs, the averages of ChatGPT's estimates are still impacted by added reasons. In addition to this, it appears ChatGPT only out-performs BERT for depression if no reasons are added. Reasons for this should be explored further. In summary, ChatGPT could pose as an alternative or complement to rating scales and previous language models, provided the optimal prompt and type of input.

Conclusion

This study has demonstrated that ChatGPT can, with high accuracy, assess levels of depression and anxiety using answers to open-ended questions. The results have highlighted both the potential and the limitations of using ChatGPT for mental health assessments. Further research is encouraged to further explore the potential of ChatGPT, and is needed to ensure reliable and ethical assessments.

References

Dergaa, I., Fekih-Romdhane, F., Hallit, S., Loch, A. A., Glenn, J. M., Fessi, M. S., ... & Ben Saad, H. (2024). ChatGPT is not ready yet for use in providing mental health assessment and interventions. Frontiers in Psychiatry, 14, 1277756. https://doi.org/10.3389/fpsyt.2023.1277756

D'Souza, R. F., Amanullah, S., Mathew, M., & Surapaneni, K. M. (2023). Appraising the performance of ChatGPT in psychiatry using 100 clinical case vignettes. *Asian Journal of Psychiatry, 89*, 103770. https://doi.org/10.1016/j.ajp.2023.103770

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv:1810.04805v2. https://doi.org/10.48550/arXiv.1810.04805

Ekin, S. (2023). Prompt Engineering For ChatGPT: A Quick Guide To Techniques, Tips, And Best Practices. Institute of Electrical and Electronics Engineers (IEEE). https://doi.org/10.36227/techrxiv.22683919

Geetha, M. P., & Karthika Renuka, D. (2021). Improving the performance of aspect based sentiment analysis using fine-tuned Bert Base Uncased model. In International Journal of Intelligent Networks (Vol. 2, pp. 64–69). Elsevier BV. https://doi.org/10.1016/j.ijin.2021.06.005

Ghojogh, B., & Ghodsi, A. (2020). Attention Mechanism, Transformers, BERT, and GPT: Tutorial and Survey. Center for Open Science. https://doi.org/10.31219/osf.io/m6gcn

Grabb, D. (2023). The impact of prompt engineering in large language model performance: a psychiatric example. *Journal of Medical Artificial Intelligence, 6*. https://doi.org/10.21037/jmai-23-71

Greco, C. M., Simeri, A., Tagarelli, A., & Zumpano, E. (2023). Transformer-based language models for mental health issues: a survey. Pattern Recognition Letters, 167, 204-211. https://doi.org/10.1016/j.patrec.2023.02.016

Hoerger, M. (2013). Z_H: An updated version of Steiger's Z and web-based calculator for testing the statistical significance of the difference between dependent correlations. http://www.psychmike.com/dependent_correlations.php

Hsu, Y.-C., Hua, T., Chang, S., Lou, Q., Shen, Y., & Jin, H. (2022). Language model compression with weighted low-rank factorization. arXiv. https://doi.org/10.48550/ARXIV.2207.00112

Jiang, Z., Levitan, S. I., Zomick, J., & Hirschberg, J. (2020). Detection of Mental Health from Reddit via Deep Contextualized Representations. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.louhi-1.16

Kjell, K., Johnsson, P., & Sikström, S. (2021). Freely Generated Word Responses Analyzed With Artificial Intelligence Predict Self-Reported Symptoms of Depression, Anxiety, and Worry. In Frontiers in Psychology (Vol. 12). Frontiers Media SA. https://doi.org/10.3389/fpsyg.2021.602581

Kjell, O. N. E., Kjell, K., Garcia, D. & Sikström, S. Semantic measures: using natural language processing to measure, differentiate, and describe psychological constructs. Psychol. Methods 24, 92–115 (2019). https://doi.org/10.1037/met0000191

Kjell, O. N. E., Kjell, K, & Schwartz, H.A. (2023). Beyond Rating Scales: With Targeted Evaluation, Language Models are Poised for Psychological Assessment. *Psychiatry Research*, 115667–115667. https://doi.org/10.1016/j.psychres.2023.115667

Kjell, O. N. E., Sikström, S., Kjell, K., & Schwartz, H. A. (2022). Natural language analyzed with AI-based transformers predict traditional subjective well-being measures approaching the theoretical upper limits in accuracy. *Scientific Reports*, *12*(1), 3918. https://doi.org/10.1038/s41598-022-07520-w

Kumar, V., Avasthi, A., Grover, S. (2019) Correlates of worry and functional somatic symptoms in generalized anxiety disorder. Ind Psychiatry J 2019;28:29-36.

Kroenke, K., Spitzer, R. L., & Williams, J. B. (2001). The PHQ-9: validity of a brief depression severity measure. Journal of general internal medicine, 16(9), 606–613.

Levkovich, I., & Elyoseph, Z. (2023). Identifying depression and its determinants upon initiating treatment: ChatGPT versus primary care physicians. Family Medicine and Community Health, 11(4). https://doi.org/10.1136%2Ffmch-2023-002391

Meng, Y., Speier, W., Ong, M. K., & Arnold, C. W. (2021). Bidirectional Representation Learning From Transformers Using Multimodal Electronic Health Record Data to Predict Depression. In IEEE Journal of Biomedical and Health Informatics (Vol. 25, Issue 8, pp. 3121–3129). Institute of Electrical and Electronics Engineers (IEEE). https://doi.org/10.1109/jbhi.2021.3063721

OpenAI. (n.d.). *GPT-4 Turbo and GPT-4*: OpenAI. Retrieved May 20, 2024, from https://platform.openai.com/docs/models/gpt-4-turbo-and-gpt-4

Ray, P. P. (2023). ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. In Internet of Things and Cyber-Physical Systems (Vol. 3, pp. 121–154). Elsevier BV. https://doi.org/10.1016/j.iotcps.2023.04.003

Sikström, S., Kelmendi, B., & Persson, N. (2023a). Assessment of depression and anxiety in young and old with a question-based computational language approach. Npj Mental Health Research, 2(1), 2-3 https://doi.org/10.1038/s44184-023-00032-z

Sikström, S., Pålsson Höök, A., & Kjell, O. (2023b). Precise language responses versus easy rating scales-Comparing respondents' views with clinicians' belief of the respondent's views. PloS one, 18(2), e0267995. https://doi.org/10.1371/journal.pone.0267995

Spitzer, R. L., Kroenke, K., Williams, J. B. W., & Löwe, B. (2006). A brief measure for assessing generalized anxiety disorder: The GAD-7. Archives of Internal Medicine, 166, 1092–1097.

Sun, C. F., Correll, C. U., Trestman, R. L., Lin, Y., Xie, H., Hankey, M. S., ... & Kablinger, A. S. (2023). Low availability, long wait times, and high geographic disparity of psychiatric outpatient care in the US. General Hospital Psychiatry, 84, 12-17. https://doi.org/10.1016/j.genhosppsych.2023.05.012

Tausczik, Y. R., & Pennebaker, J. W. (2009). The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods. Journal of Language and Social Psychology, 29(1), 24–54. https://doi.org/10.1177/0261927x09351676

Vajre, V., Naylor, M., Kamath, U., & Shehu, A. (2021). PsychBERT: A Mental Health Language Model for Social Media Mental Health Behavioral Analysis. In 2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). https://doi.org/10.1109/bibm52615.2021.9669469

WHO. (2022, June 8). *Mental disorders*. World Health Organization; World Health Organization. https://www.who.int/news-room/fact-sheets/detail/mental-disorders

Wu, T., He, S., Liu, J., Sun, S., Liu, K., Han, Q.-L., & Tang, Y. (2023). A Brief Overview of ChatGPT: The History, Status Quo and Potential Future Development. In IEEE/CAA Journal of Automatica Sinica (Vol. 10, Issue 5, pp. 1122–1136). Institute of Electrical and Electronics Engineers (IEEE). https://doi.org/10.1109/jas.2023.123618

Appendix A

Examples of ChatGPT's text responses without reasons added.

Word-response	ChatGPT's response
	Dep.con
Suicidal Scared Panicked Lonely Broken	Based on the five words provided: "Suicidal," "Scared," "Panicked," "Lonely," and "Broken," the degree of depression can be estimated to be very high. These words indicate severe emotional distress and a critical mental state. I would estimate the degree of depression at 9 or 10 on a scale from 0 to 10, where 10 represents maximal depression. It is important to consider immediate professional help and support for someone expressing these feelings.
	Dep.DSM
anxious apprehensive sad happy elated	Based on the five words provided: "anxious," "apprehensive," "sad," "happy," and "elated," the participant's emotional state appears mixed, with signs of both negative and positive emotions. The words "anxious" and "apprehensive" suggest feelings of unease or worry, while "sad" indicates a level of unhappiness. However, the words "happy" and "elated" indicate positive emotions and a sense of joy or high spirits. Given this mix, it is challenging to estimate a precise degree of depression without more context or detail about the intensity and duration of each emotion. However, the presence of both positive and negative emotions suggests a fluctuating mood rather than a consistently depressed state. Estimating on a scale from 0 (not at all depressed) to 10 (maximally depressed), and considering the DSM-5 criteria which emphasize the presence of depressed mood or loss of interest or pleasure most of the day, nearly every day, the response does not strongly indicate a high level of depression. The positive emotions might counterbalance the negative ones to some extent. Therefore, I would estimate the degree of depression to be around a 3 on the scale, reflecting the presence of some depressive emotions but moderated by significant positive feelings. This is a rough estimate and should be treated with caution, as a more detailed assessment would be necessary for a more accurate understanding.
	Dep.PHQ
Neutral Tired Indifferent Fluctuating Lazy	Based on the five-word response provided: "Neutral Tired Indifferent Fluctuating Lazy," we can attempt to estimate the degree of depression on the PHQ-9 scale. Here's a breakdown of the possible implications of each word in relation to depressive symptoms: 1. **Neutral** - This suggests a lack of strong feelings, which could be interpreted as neither depressed nor happy. It might indicate a baseline or normal mood, but in the context of depression, it could also suggest a flat affect, which is sometimes associated with depression. 2. **Tired** - Fatigue is a common symptom of depression. Feeling consistently tired or having low energy is a relevant indicator. 3. **Indifferent** - This could suggest a lack of interest or pleasure in activities, which is a key symptom of depression. 4. **Fluctuating** - This implies variability in mood, which could be related to episodes of feeling more or less depressed. Mood swings can be associated with different types of depressive disorders, including major depressive disorder or bipolar disorder. 5. **Lazy** - Often, people describe themselves as lazy when they are experiencing low energy and motivation, which are symptoms of depression. Given these interpretations, the response indicates some symptoms that align with depression, such as tiredness, indifference, and low energy/motivation. However, the use of "neutral" and the absence of more explicit depressive feelings like sadness or hopelessness might suggest a lower intensity of depressive symptoms. Estimating on the PHQ-9 scale, which typically quantifies frequency of symptoms as "not at all" (0 points) to "nearly every day" (3 points) for each of the 9 symptoms, we can infer: - **Little interest or pleasure in doing things (Indifferent)**: This could be scored as 1 (several days). - **Feeling down, depressed, or hopeless**: Not explicitly mentioned, so possibly 0. - **Trouble falling or staying asleep, or sleeping too much**: Not mentioned, so 0. - **Feeling tired or having little energy (Tired, Lazy)**: Likely scored as 1 (several days).

	- **Poor appetite or overeating**: Not mentioned, so 0. - **Feeling bad about yourself - or that you are a failure or have let yourself or your family down**: Not mentioned, so 0. - **Trouble concentrating on things**: Not mentioned, so 0. - **Moving or speaking so slowly that other people could have noticed. Or the opposite - being so fidgety or restless that you have been moving around a lot more than usual**: Not mentioned, so 0. - **Thoughts that you would be better off dead, or of hurting yourself**: Not mentioned, so 0. Based on this rough analysis, the score might be around 2 out of a possible 27, indicating a mild level of depressive symptoms. However, it's important to note that a thorough clinical evaluation would be necessary for an accurate diagnosis.
	Wor.con
Worried Anxious Scared Panicky On-edge	Based on the five words provided: "Worried," "Anxious," "Scared," "Panicky," "On-edge," the degree of worry appears to be very high. Each word indicates a strong feeling of unease or fear. Therefore, I would estimate the degree of worry to be around 9 on a scale from 0 (not at all worried) to 10 (maximally worried).
	Wor.DSM
worried calm scared anxious happy	Based on the five words provided: "worried," "calm," "scared," "anxious," and "happy," it appears there is a mix of both positive and negative emotions related to worry. The words "worried," "scared," and "anxious" indicate a presence of worry or anxiety. However, the inclusion of "calm" and "happy" suggests that there are also periods of relief or positive feelings. To estimate the degree of worry on a scale from 0 to 10, where 0 is not at all worried and 10 is maximally worried, we need to consider the balance of these emotions. Since there are three words indicating worry or anxiety and two that suggest less worry or more positive states, it seems the participant experiences a significant amount of worry, but it is not overwhelming or constant. Estimation: 6/10. This score reflects a moderate level of worry, with some fluctuations in mood and moments of calmness and happiness.
	Wor.GAD
occasionally often frequently sometimes routinely	The words provided in the response — "occasionally," "often," "frequently," "sometimes," and "routinely" — suggest varying levels of worry. To estimate the degree of worry on the GAD-7 scale, we can analyze the implications of each word: 1. **Occasionally** - This implies worry happens from time to time but is not constant. 2. **Often** - Indicates a more regular occurrence of worry. 3. **Frequently** - Suggests worry happens very regularly, almost a part of daily routine. 4. **Sometimes** - Implies worry occurs at certain times but not consistently. 5. **Routinely** - Indicates worry is a regular part of life, almost habitual. Given these descriptions, the participant experiences worry at varying frequencies, but the use of words like "often," "frequently," and "routinely" suggests a higher frequency and possibly intensity of worry. On the GAD-7 scale: - Scores of 5, 10, and 15 are the cut-off points for mild, moderate, and severe anxiety, respectively. - The response suggests a presence of worry that is more than just "several days" (which would score 1 per symptom on the GAD-7). The use of "often," "frequently," and "routinely" aligns more closely with "more than half the days" (which scores 2 per symptom) or possibly "nearly every day" (which scores 3 per symptom). If we were to hypothetically assign a GAD-7 score based on the qualitative description: - Assigning a score of 2 or 3 for each descriptor (since the terms suggest a frequency of at least "more than half the days" and possibly "nearly every day"), and considering there are multiple descriptors suggesting frequent worry, the score could potentially be in the moderate to severe range (10-21). Therefore, based on the response, it is reasonable to estimate that the participant might fall into the moderate to severe category of worry on the GAD-7 scale. However, a precise score and a definitive clinical assessment would require a more detailed evaluation including all seven items of the GAD-7 questionnaire, each specifically addressing different aspects of anxiety.

Appendix B

Examples of ChatGPT's text responses with reasons added.

Symptoms	Reason	ChatGPT's response
		Dep.con.reas.
Suicidal Scared Panicked Lonely Broken	Pressure Stress Anxiety	Based on the words provided by the participant to describe their feelings and the reasons for their depression, it is possible to estimate a high degree of depression. The words "Suicidal," "Scared," "Panicked," "Lonely," and "Broken" indicate severe emotional distress and a profound negative impact on mental health. The reasons provided, "Pressure," "Stress," and "Anxiety," further support the severity of the emotional state as these factors are known to significantly contribute to depressive feelings. Given the intensity and nature of the descriptive words, the estimated degree of depression on a scale from 0 to 10 would be around 8 to 10. This suggests a critical level of depression that likely requires immediate professional intervention and support. It is important for the individual to seek help from mental health professionals to address these feelings and work on coping strategies and possibly medical treatment.
		Dep.DSM.reas.
anxious apprehensive sad happy elated	work study future	Based on the words provided to describe feelings over the last two weeks — "anxious," "apprehensive," "sad," "happy," "elated" — it appears there is a mix of both negative and positive emotions. The negative emotions (anxious, apprehensive, sad) suggest some level of distress, while the positive emotions (happy, elated) indicate periods of positive mood. The reasons provided for the depression — "work," "study," "future" — suggest stressors that are contributing to feelings of anxiety and apprehension, likely related to performance and uncertainty about outcomes. Given the mix of both positive and negative emotions, and the nature of the stressors, the degree of depression might not be severe but indicates some emotional fluctuations possibly linked to situational stressors. On a scale from 0 (not at all depressed) to 10 (maximally depressed), this scenario might be estimated around a 3 or 4, reflecting mild depression or possibly adjustment issues with specific stress factors but not pervasive low mood.
		Dep.PHQ.reas.
Neutral Tired Indifferent Fluctuating Lazy	Little Insignificant Negligible	Based on the words used by the participant to describe their feelings over the last two weeks ("Neutral," "Tired," "Indifferent," "Fluctuating," "Lazy") and the reasons for their depression ("Little," "Insignificant," "Negligible"), it appears that the participant is experiencing some symptoms of depression, but these symptoms are mild and not severely impacting their daily functioning. The words suggest a state of low energy and a lack of strong emotional responses ("Indifferent"), but not a profound or pervasive sense of sadness or hopelessness. The fluctuation in mood ("Fluctuating") indicates variability in their emotional state, which can be a characteristic of mild depression. The reasons provided for feeling depressed are described as minor ("Little," "Insignificant," "Negligible"), suggesting that the causes of these feelings are not significantly distressing or overwhelming to the participant. Given this information, on the PHQ-9 scale, which ranges from 0 (not at all depressed) to 27 (maximally depressed), the participant's responses align more with a lower score. They indicate some symptoms of depression, particularly tiredness and mood fluctuations, but these symptoms seem to have a limited impact. An estimated score might be around 5 to 6 on the PHQ-9 scale. This score reflects mild depression, where the individual experiences some symptoms that are noticeable but not dominant or disabling. It's important to note that a precise score should ideally be determined through a complete and structured clinical assessment.

		Wor.con.reas.
Worried Anxious Scared Panicky On-edge	Stress Pressure Relationships	Based on the words used to describe the feelings during the last two weeks ("Worried," "Anxious," "Scared," "Panicky," "On-edge") and the reasons provided ("Stress," "Pressure," "Relationships"), it appears that the participant is experiencing a high level of worry. The words chosen to describe their feelings are all associated with significant anxiety and fear, indicating a heightened emotional state. The reasons for worry mentioned—stress, pressure, and relationships—are common triggers for anxiety and can contribute to an overall sense of being overwhelmed or unable to cope effectively. Given the intensity of the descriptive words for both the feelings and reasons, I would estimate the degree of worry on a scale from 0 to 10 to be around 8. This suggests a high level of worry that might be significantly impacting the participant's daily functioning and emotional well-being. It might be beneficial for the participant to seek professional help or support to manage these feelings and address the underlying causes.
		Wor.DSM.reas.
worried calm scared anxious happy	university family christmas	Based on the words provided to describe feelings over the last two weeks — "worried," "calm," "scared," "anxious," and "happy" — it appears there is a mix of both positive and negative emotions. The presence of words like "worried," "scared," and "anxious" indicates feelings of worry, but these are somewhat balanced by "calm" and "happy." The reasons for worry mentioned — "university," "family," and "Christmas" — suggest specific stressors related to personal responsibilities, relationships, and possibly seasonal pressures. Given the mixture of both worry-inducing and positive emotions, and considering the reasons for worry which are significant but common life stressors, I would estimate the degree of worry on a scale from 0 to 10 to be around 5. This suggests a moderate level of worry, where the participant experiences significant worry but also periods of calm and happiness.
		Wor.GAD.reas.
occasionally often frequently sometimes routinely	work study homelife	To estimate the degree of worry or anxiety on the GAD-7 scale based on the words provided, we can analyze the frequency and context of the words used: 1. **Words Describing Worry**: "occasionally," "often," "frequently," "sometimes," "routinely" - "Occasionally" and "sometimes" suggest infrequent or irregular periods of worry. - "Often," "frequently," and "routinely" indicate a more regular or common occurrence of worry. Given the mixture of terms indicating both sporadic and regular worry, it suggests a moderate to high level of worry. The presence of both higher frequency terms ("often," "frequently," "routinely") and lower frequency terms ("occasionally," "sometimes") points towards a more consistent presence of worry, though not at the highest possible level. 2. **Reasons for Worry**: "work," "study," "homelife" - These are significant areas of life, and issues in these areas can lead to substantial worry or anxiety. Considering the GAD-7 scale ranges from 0 to 21, where: - 0-4 indicates minimal anxiety, - 5-9 indicates mild anxiety, - 10-14 indicates moderate anxiety, - 15-21 indicates severe anxiety, **Estimation**: The participant's use of words and the areas of concern suggest a moderate to potentially high level of anxiety. A reasonable estimate might be in the range of 10-16 on the GAD-7 scale. This suggests moderate anxiety, potentially bordering on severe, depending on the intensity and impact of the worries on daily functioning, which isn't fully clear from the words alone. A precise score would require more detailed information about the frequency and impact of these feelings.

Appendix C

Table 1

Non-parametric ANOVA (Friedman) comparing means between ChatGPT's estimated depression scores using prompt 1 referring to the construct with/without the addition of reasons.

Friedman			
X^2	df	p-value	
112	1	<0.001**	
Pairwise Comparisons (Durbin-Conover)			
		Statistic	p-value
Dep.con	– Dep.con.reas.	11.3	<0.001**

Table 2

Non-parametric ANOVA (Friedman) comparing means between ChatGPT's estimated depression scores using prompt 2 referring to the DSM with/without the addition of reasons.

Friedman			
X^2	df	p-value	
146	1	<0.001**	
Pairwise Comparisons (Durbin-Conover)			
		Statistic	p-value
Dep.DSM	– Dep.DSM.reas.	13.2	<0.001**

Table 3

Non-parametric ANOVA (Friedman) comparing means between ChatGPT's estimated depression scores using prompt 3 referring to PHQ-9 with/without the addition of reasons.

Friedman			
X^2	df	p-value	
55.9	1	<0.001**	
Pairwise Comparisons (Durbin-Conover)			
		Statistic	p-value
Dep.PHQ	– Dep.PHQ.reas.	7.73	<0.001**

Table 4

Non-parametric ANOVA (Friedman) comparing means between BERT's predicted PHQ-9 scores with/without the addition of reasons.

Friedman			
X^2	df	p-value	
0.292	1	0.589	
Pairwise Comparisons (Durbin-Conover)			
		Statistic	p-value
BERT.PHQ	– BERT.PHQ.reas.	0.540	0.589

Appendix D

Table 1

Non-parametric ANOVA (Friedman) comparing means between ChatGPT's estimated anxiety scores using prompt 1 referring to the construct with/without the addition of reasons.

Friedman			
X^2	df	p-value	
509	1	<0.001	
Pairwise Comparisons (Durbin-Conover)			
		Statistic	p-value
Wor.con	– Wor.con.reas.	34.8	<0.001

Table 2

Non-parametric ANOVA (Friedman) comparing means between ChatGPT's estimated anxiety scores using prompt 2 referring to the DSM with/without the addition of reasons.

Friedman			
X^2	df	p-value	
128	1	<0.001	
Pairwise Comparisons (Durbin-Conover)			
		Statistic	p-value
Wor.DSM	– Wor.DSM.reas.	12.2	<0.001

Table 3

Non-parametric ANOVA (Friedman) comparing means between ChatGPT's estimated anxiety scores using prompt 3 referring to GAD-7 with/without the addition of reasons.

Friedman			
X^2	df	p-value	
25.2	1	<0.001	
Pairwise Comparisons (Durbin-Conover)			
		Statistic	p-value
Wor.GAD	– Wor.GAD.reas.	5.10	<0.001

Table 4

Non-parametric ANOVA (Friedman) comparing means between BERT's predicted GAD-7 scores with/without the addition of reasons.

Friedman			
X^2	df	p-value	
0.0731	1	0.787	
Pairwise Comparisons (Durbin-Conover)			
		Statistic	p-value
BERT.GAD	– BERT.GAD.reas.	0.270	0.787