Underneath the Numbers: Quantitative and Qualitative Gender Fairness in LLMs for Depression Prediction

Anonymous ACL submission

Abstract

Recent studies show bias in many machine learning models for depression detection, but bias in LLMs for this task remains unexplored. This work presents the first attempt to investigate the degree of gender bias present in existing LLMs (ChatGPT, LLaMA 2, and Bard) using both *quantitative* and *qualitative* approaches. rom our quantitative evaluation, we found that ChatGPT performs the best across various performance metrics and LLaMA 2 outperforms other LLMs in terms of group fairness metrics. As qualitative fairness evaluation remains an open research question we propose several strategies (e.g., word count, thematic analysis) to investigate whether and how a qualitative evaluation can provide valuable insights for bias analysis beyond what is possible with quantitative evaluation. We found that ChatGPT consistently provides a more comprehensive, well-reasoned explanation for its prediction compared to LLaMA 2. We have also identified several themes adopted by LLMs to *qualitatively* evaluate gender fairness. We hope our results can be used as a stepping stone towards future attempts at improving *qualitative* evaluation of fairness for LLMs especially for high-stakes tasks such as depression detection.

1 Introduction

The recent rise of Large Language Models (LLMs) have demonstrated the unique capability in undertaking various tasks ranging from machine translation (Ghosh and Caliskan, 2023) to medical applications (Zack et al., 2024). Among the various applications, a key application is that of mental health detection and analysis where LLMs must be capable of perceiving or detecting mental health status. Though recent attempts at using LLMs for the investigation and understanding of mental health has been promising (Xu et al., 2023; Yang et al., 2023), none of the existing work has looked into the problem of LLM bias in depression prediction. Depression prediction is a machine learning problem that aim at automatically identifying signs of depression in individuals by analysing and processing human behavioural data, including facial expressions (Song et al., 2018), speech and textual data (Nasir et al., 2016).

It has been shown in recent works that LLMs are prone to bias. This bias is present in many LLMs for various tasks (Ghosh and Caliskan, 2023; Kotek et al., 2023; Cabello et al., 2023). None of the existing works has investigated bias in LLM for the task of depression detection. In addition, all of the existing work on machine learning (ML) or LLM fairness have mainly focused on a *quantitative*-notion of fairness (Han et al., 2022; Esiobu et al., 2023). This can largely be understood as fairness that is measured and defined by quantifiable metrics. Existing works have yet to consider *qualitative* fairness. Several works have attempted to *qualitatively* evaluate fairness using visualisation or anecdotal examples (Tsioutsiouliklis et al., 2021) or attempted a *qualitative* evaluation of perception on fairness (Woodruff et al., 2018). However, human-centered research has indicated that *explanations* contribute substantially to an individual's fairness perceptions (Yurrita et al., 2023; Shulner-Tal et al., 2023). Thus, we adopt a human-centered approach by evaluating an LLM's ability to provide *explanations* for the decisions made. Providing explanations also leads towards enhancing *algorithmic explainability* (Shin, 2020) and *transparency* (Rader et al., 2018; Arrieta et al., 2020) which are both crucial elements in developing human-centred and trustworthy artificial intelligence (AI) systems (Shneiderman, 2020).

Our work aims at investigating the degree of gender bias present in existing LLMs – namely ChatGPT, LLaMA 2, and Bard – using both quantitative and qualitative approaches. To this end, we investigated first if bias is present in existing LLMs for the depression detection task, then we explored how the different LLMs differ across the various *quantitative* and *qualitative* fairness measures, and finaly we identified the main themes used by the LLMs to *qualitatively* evaluate gender fairness.

The contribution of our work is as follows. First, we conduct a thorough comparison of LLM performance for depression detection across two datasets. Second, we undertake a novel investigation of qualitative fairness to evaluate bias and improve explainabililty in LLM predictions. To the best of our knowledge, none of the existing works have attempted to define and evaluate qualitative fairness for any task. Third, we perform a multitude of fine-grained analyses on various experimental settings (see Section 3.3) to examine the prediction and fairness across all three LLMs.

2 Related Work

2.1 ML Fairness for Mental Health

There has been a handful of studies which have looked into bias in mental well-being prediction (Ryan and Doherty, 2022; Bailey and Plumbley, 2021; Park et al., 2022, 2021; Zanna et al., 2022; Cheong et al., 2023a,b). Park et al. (Park et al., 2021) proposed bias mitigation strategies for postpartum depression. Zanna et al. (Zanna et al., 2022) adopted a multitask approach to mitigate bias for anxiety prediction. Ryan et al. (Ryan and Doherty, 2022) proposed three categories of fairness definitions for mental health. Park et al. (Park et al., 2022) proposed an algorithmic impact remover to mitigate bias in mobile mental health. Bailey and Plumbley (Bailey and Plumbley, 2021) proposed using data re-distribution to mitigate gender bias for depression detection. (Cheong et al., 2023a) examined whether bias exists in existing mental health datasets and algorithms. None of the existing works have looked into ML Fairness for mental health as applied within a LLM setting.

2.2 Gender Bias in LLM

A proliferation of recent works has confirmed the presence of gender bias in LLMs (Gallegos et al., 2023). (Wan et al., 2023) revealed substantial gender biases in LLM-generated recommendation letters. (Ghosh and Caliskan, 2023) conducted experiments which revealed that ChatGPT exhibits the gender bias for the task of machine translation. (Thakur, 2023) analysed gender bias comparing between GPT 2 and GPT 3.5 for the task of name generation for profession. (Kotek et al., 2023) tested four LLMs and demonstrated that the LLMs expressed biased assumptions about a person's occupation based on gender. (Zack et al., 2024) discovered that GPT-4 exhibited gender bias by not modelling the demographic diversity and producing clinical vignettes that stereotype demographic presentations. (Dong et al., 2023) propose a conditional text generation mechanism to address the problem of gender bias in LLMs. (Dong et al., 2024) proposed three methods to mitigate bias in LLMs via hyperparameter tuning, instruction guiding and debias tuning. However, none of the existing works has focused on analysing gender bias in LLMs for the task of *depression detection*.

2.3 LLMs for Mental Health Applications

The last year has been characterised by an exponential advance in the current state of the art of Large Language Models (LLMs). Few works (Borji and Mohammadian, 2023; Ali et al., 2022) have attempted to compare different LLMs. Borji et al. (Borji and Mohammadian, 2023) undertook an extensive benchmark evaluation of LLMs and conversational bots – ChatGPT (gpt-3.5), GPT-4, Bard, and Claude – using the "Wordsmiths dataset" categories (e.g., questions on logic, facts, coding etc.). More and more studies have been focusing on applications of LLMs in healthcare (Lamichhane, 2023; Yang et al., 2023; Qin et al., 2023) and affective computing (Elyoseph et al., 2023) domains. Lamichhane et al. (Lamichhane, 2023) have evaluated the use of ChatGPT (gpt-3.5) to accomplish three mental health-related classification tasks, namely stress detection, depression detection, and suicidal detection. Their results suggested that language models can be effectively used for mental health classification tasks. Yang et al. (Yang et al., 2023) have evaluated the mental health analysis and emotional reasoning ability of ChatGPT (gpt-3.5) on 11 datasets across 5 tasks, and analyzed the effects of various emotion-based prompting strategies. None of these previous works have compared the LLM biases for mental health applications. Therefore, this work aims at comparing three LLMs for mental health applications under the lens of *fairness and explainability*.

3 Depression Prediction

This paper aims at **understanding quantitatively and qualitatively the gender fairness of three different state-of-the-art LLMs in depression prediction tasks**. This section describes the large language models explored, the datasets used, the definition of the prompts, the processing of the transcriptions, and the evaluation methodology.

3.1 Large Language Models

We decided to compare the cutting-edge large language models (LLMs) currently available, namely LLaMA 2 (by Meta[1] (Touvron et al., 2023)), ChatGPT (by OpenAI[2]), and Bard (by Google[3]) to accomplish a depression-related detection task. We used the python OpenAI library to invoke the Chat Complete API of ChatGPT by using *gpt-3.5-turbo backend* as in (Lamichhane, 2023). Analogously, we have used the huggingface library[4] to call the LLaMA 2 API by using a total of 400 hours in 4x NVIDIA A100-SXM-80GB GPUs. We set for these LLMs a temperature equal to 0.7 and a maximum length of the output of 200 tokens. While for Bard, we used the experimental version provided by Google via the Bard GUI, where it is not possible to set parameters of the model.

3.2 Datasets

We used benchmark datasets that contain transcriptions of dyadic interactions for the tasks of depression detection that were anonymised by the owners.Dataset distributions can be found in Appendix A. The **DAIC-WOZ** dataset (Gratch et al., 2014) includes audio and video recordings of semi-clinical interviews and responses of PHQ-8 questionnaire. The **E-DAIC** corpus (Ringeval et al., 2019) is an extended version of DAIC-WOZ that contains semi-clinical interviews designed to support the diagnosis of psychological distress conditions. Both datasets are labelled on a scale from 0 to 24 based on the PHQ-8 questionnaire.

3.3 Prompting for Depression Detection

We defined different prompts for evaluating the performance and fairness of LLMs for the depression prediction task from transcriptions of dyadic interactions. This section details and reports the verbatim of the prompt defined for the detection tasks by grounding them on past works (Kroenke et al., 2009; Busso et al., 2008).

[1]https://github.com/facebookresearch/LLaMA 2

[2]https://platform.openai.com/docs/api-reference

[3]https://bard.google.com/

[4]https://huggingface.co/meta-LLaMA2/LLaMA2-2-70b-chat-hf

3.3.1 Baseline Prompt

For the depression recognition task, we used the formulation from the PHQ-8 questionnaire (Kroenke et al., 2009), as a baseline measure of depression to annotate the DAIC-WOZ and E-DAIC datasets (Gratch et al., 2014). The PHQ-8 questionnaire is a valid diagnostic and severity measure for depressive disorders, with score ranging between 0 and 24 points. A score of 0 to 4 represents no significant depressive symptoms, a total score of 5 to 9 represents mild depressive symptoms, 10 to 14, moderate, 15 to 19, moderately severe, and 20 to 24, severe (Kroenke et al., 2009). Analogously, we asked the LLMs to annotate the conversations formulating the prompt for the DAIC-WOZ and E-DAIC datasets reported in Appendix B.

3.3.2 Gender-Explicit and Implicit Prompts

This work aims to explore gender fairness in LLMs. Past works (Gallegos et al., 2023) have shown how the definition of the prompt is extremely important for LLMs and how it affects the response of the models. Therefore, we decided to explore different prompt formulations which included gender information: *gender-explicit* and *gender-implicit*. Specifically, in the gender-explicit formulation, we explicitly informed the LLM that the participant of the conversation is of a specific gender (e.g., *"Consider that the participant is a female. On a scale of [..]"*). While in the gender-implicit formulation, we implicitly included the information about the gender within the prompt (e.g.,*"On a scale of 0 to 24 where 0 to 4 means that the female participant has no significant depressive symptoms; 5 to 9 means that the female participant has mild depressive symptoms; [..]"*).

3.4 Data Processing

The three LLMs have different limitations in terms of maximum number of input tokens. ChatGPT and LLaMA 2 allow an input prompt of 4096 tokens, while Bard allows 2048 input tokens[5]. For the sake of comparison, we opted for the minimum value limit to allow the LLMs to process the same input prompt, i.e., 2048 tokens.

When the input prompt was longer than 2048 tokens, we applied a chunking approach (Zhou et al., 2012) where we split the conversation in smaller chunks (with a maximum length of 2048 tokens) with an overlap of 500 tokens (this number was

[5]Note that all the experiments were conducted between October and December 2023

chosen empirically to make sure that the semantic context did not get lost between chunks). Each chunk has been then used as input prompt for the evaluation process. For example, if a conversations included a total number of tokens of 4500, we split it into three chunks of 2000 tokens each (with 500 tokens of overlap). We then conducted the experiments with 10 run repetitions described in Section 5 using the LLMs approaches.

4 Fairness

In this section, we describe the quantitative fairness metrics used and introduce and define the concept of qualitative fairness which is one of our key contribution. We explore a binary classification setting in order to facilitate calculation of the fairness scores and comparison with existing ML for depression detection works (Zheng et al., 2023) on gender fairness in wellbeing analysis.

4.1 Quantitative Fairness

We utilise the following metrics to analyse group fairness as they are the most commonly used metrics within the literature (Hort et al., 2022; Pessach and Shmueli, 2022). s_0 denotes the minority group which are females in our setup and s_1 denotes the majority group males. Y refers to the binary ground truth label (0 vs 1) and $\hat{Y}$ refers to the predicted outcome (0 vs 1) where 0 is the non-depressed class and 1 is the depressed class.

- **Statistical Parity**, or demographic parity, is based purely on predicted outcome $\hat{Y}$ and independent of actual outcome Y:

$$\mathcal{M}_{SP} = \frac{P(\hat{Y}=1|s_0)}{P(\hat{Y}=1|s_1)}. \quad (1)$$

  According to this measure, in order for a classifier to be deemed fair, $P(\hat{Y} = 1|s_1) = P(\hat{Y} = 1|s_0)$ (Mehrabi et al., 2021). The intuition behind this metric is that a fair classifier should provide both groups with equal chances of being classified within the positive $\hat{Y} = 1$ class (Hort et al., 2022).

- **Equal opportunity** states that both demographic groups s_0 and s_1 should have equal True Positive Rate (TPR).

$$\mathcal{M}_{EOpp} = \frac{P(\hat{Y}=1|Y=1,s_0)}{P(\hat{Y}=1|Y=1,s_1)}. \quad (2)$$

  According to this measure, in order for a classifier to be deemed fair, $P(\hat{Y} = 1|Y = 1, s_1) = P(\hat{Y} = 1|Y = 1, s_0)$ (Mehrabi et al., 2021). The intuition is that both demographic groups should have equal *true positive rates* (TPR) for a classifier to be considered fair (Hort et al., 2022).

- **Equalised odds** can be considered as a generalization of Equal Opportunity where the rates are not only equal for $Y = 1$, but for all values of $Y \in \{1, ...k\}$, i.e.:

$$\mathcal{M}_{EOdd} = \frac{P(\hat{Y}=1|Y=i,s_0)}{P(\hat{Y}=1|Y=i,s_1)}. \quad (3)$$

  According to this measure, in order for a classifier to be deemed fair, $P(\hat{Y} = 1|Y = i, s_1) = P(\hat{Y} = 1|Y = i, s_0), \forall i \in \{1, ...k\}$ (Mehrabi et al., 2021). This can be understood as a stricter version of $\mathcal{M}_{EOpp}$ as both subgroups are required to have equal TPR and *false positive rates* (FPR) for a classifier to be deemed fair (Hort et al., 2022).

- **Equal Accuracy** states that both subgroups s_0 and s_1 should have equal rates of accuracy (Mehrabi et al., 2021).

$$\mathcal{M}_{EAcc} = \frac{\mathcal{M}_{ACC,s_0}}{\mathcal{M}_{ACC,s_1}}. \quad (4)$$

  Intuitively, this is aligned with how majority of the fairness evaluation and algorithmic audits is done. A classifier is deemed unfair if it is less accurate for populations of certain demographic groups e.g. females and blacks (Buolamwini and Gebru, 2018).

4.2 Qualitative Fairness

None of the existing works have considered qualitative fairness. In addition, given the pivotal contribution of *explanations* towards *algorithmic explainability* and *transparency* (Shin, 2020; Arrieta et al., 2020), we propose **a novel perspective** and **method** to qualitatively measure fairness by evaluating *how* a LLM generates its predictions through *explanations*. This measure is inspired from a common practice in explainability within the LLM community known as *self-criticism* (Tan et al., 2023) that involves prompting the LLM to assess its output for potential inaccuracies or improvement areas. To this end, we asked each LLM to “judge" the

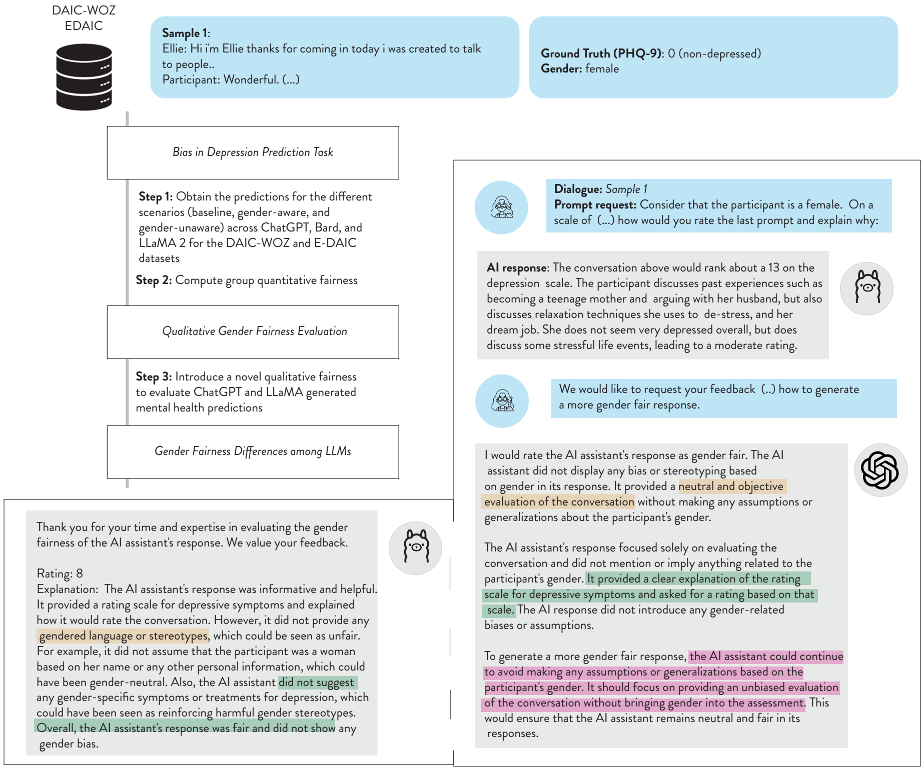

Figure 1: A sample sequence outlining the fairness evaluation process of various LLMs in the gender-explicit condition for depression prediction tasks. We have highlighted with colours the themes that emerged from our qualitative analysis as follows: **Green** - Context-based explanations; **Orange** - Gender-related language (pronouns); **Pink** - Suggestions for improvement (image to be seen in colour).

depression prediction explanations of itself (e.g., ChatGPT judges qualitatively the fairness of its own response) and other models (e.g., ChatGPT judges qualitatively the fairness of LLaMA 2's response) by using the prompt reported in Appendix B. We defined this prompt taking inspiration from (Wu and Aji, 2023) and following the guidelines listed in (Sondos Mahmoud Bsharat, 2023). To evaluate the generated qualitative fairness response, we relied on basic NLP text generation analysis (e.g., word counting, length of the response) and a thematic analysis (TA), inspired from (Braun and Clarke, 2012). TA is a well-validated tool for analysing qualitative data (Braun and Clarke, 2012) which is often combined with NLP research methods (Kim et al., 2015, 2022) and has proven effective at gathering human perception on algorithmic fairness (Kyriakou et al., 2019; Kasinidou et al., 2021; Rezai et al., 2022). In all of our experiments, we employ the 6-step method (Clarke and Braun, 2017) and the grounded theory approach (McLeod, 2011).

5 Experiments

We aim to evaluate LLM gender fairness quantitatively and qualitatively by undertaking the following steps.

Step 1: We first obtained the predictions $\hat{y}$ for the different scenarios: Baseline $\hat{y_B}$, Gender-Explicit $\hat{y_A}$, Gender-Implicit $\hat{y}_u$ across the three different LLMs across the two different datasets. We then compared the three LLMs' detection using the prompts defined in Section 3.3.1 to ground truth annotations and computed the F1 score of detection in the baseline scenario for the DAIC-WOZ and E-DAIC datasets. We conducted the same experiments but compared detection from gender-explicit and gender-implicit prompts where we included information about the gender of the participants explicitly or implicitly (see Section 3.3.2). We compared Bard, ChatGPT and LLaMA 2 and conducted the experiments over the two datasets.

Step 2: We evaluated the results generated by the LLMs using both performance and group fairness measures for $\hat{y_B}$, $\hat{y_A}$ and $\hat{y}_u$ using the measures described in Section 4.1. We compared the LLMs across the three scenarios: baseline, gender-explicit, and gender-implicit as described in Section 3.3.2 for the two datasets.

Step 3: We use a sub-sample of the test sets (25 samples in total) from E-DAIC and DAIC-WOZ datasets following the definition in Section 4. The sub-sample was randomly chosen by controlling and balancing the sub-set in terms of gender and depression conditions. We analysed the generated qualitative fairness by comparing the different models output in terms of quantitative and qualitative aspects. Specifically, we computed the number of words, the length of the generated text, the positive sentiment of the generated text for the quantitative evaluation, while we adopted a thematic analysis approach, as in (Axelsson et al., 2022), to qualitatively assess the model fairness by identifying the main themes emerged in data as highlighted in colours in Figure 1. The thematic analysis conducted includes the following 6 steps: (1) becoming familiar with the data, (2) generating initial codes, (3) searching for themes, (4) reviewing themes, (5) defining themes, (6) writing-up. Two researchers conducted steps 1 – 3 independently and then they met up to finalise the analysis (steps 4–6). Figure 1 depicts the three steps undertaken to complete our experiments in the gender-explicit scenario for sample from the DAIC-WOZ dataset.

6 Results

This section reports the results obtained in our experiments in terms of quantitative and qualitative fairness. The depression prediction results are presented in Table 6 of Appendix D.1 where we see that ChatGPT consistently produces the best classification outcomes for both DAIC-WOZ and E-DAIC across precision, recall, F1 and accuracy.

6.1 Quantitative Fairness

For all models, we see that bias seems to be present. Better classification scores were often reported for males compared to females. We examine further and report the results in Table 5.

With reference to Table 5, for DAIC-WOZ, we see that LLaMA 2 seems to be the most consistently fair LLM followed by ChatGPT. LLaMA 2 gives the fairest scores across M_{SP} (1.06), M_{EOpp} (1.04) and M_{EOdd} (1.10) with ChatGPT being the fairest across M_{EAcc} (1.00). For E-DAIC, LLaMA 2 seems to be the fairest LLM followed by Bard. LLaMA 2 gives the fairest scores across M_{EOpp} (1.00) and M_{EOdd} (1.00) and M_{EAcc} (1.02) with Bard being the fairest across M_{SP} (1.00).

Our findings indicate the presence of bias within existing LLMs. Most of the fairness scores are within the acceptable threshold range. LLaMA 2 is quantitatively fairest of all for both datasets. This is followed by ChatGPT for DAIC-WOZ and Bard for E-DAIC. There is also a difference between the quantitative fairness scores of each LLM on the different datasets which suggests that datasets do make a difference.

6.2 Qualitative Fairness

For qualitative fairness, we only evaluated ChatGPT and LLaMA 2 excluding Bard. This is because ChatGPT was the best LLM-model across performance (precision, recall, F1 and accuracy) whereas LLaMA 2 was the best LLM-model across fairness (M_{SP}, M_{EOpp}, M_{EOdd}, M_{EAcc}). We present the findings on the qualitative fairness aspect and discuss the different convergent and divergent themes across the two LLMs emerging from the thematic analysis (TA).

6.2.1 Quantitative Aspects

We compute the number of words generated in LLM qualitative gender evaluation and found that, even if we set the parameters of the number of tokens to generate equally for ChatGPT and LLaMA 2, ChatGPT generated a higher number of words and characters than LLaMA 2. Table 2 shows that there is a statistically significant difference across word number and length.

We also calculated the positive sentiment percentage (PSP in Table 3) detected using BERT sentiment analysis from huggingface [6]. Our results in Table 3 suggest that each LLM judge the other LLM more positively than themselves.

6.2.2 Qualitative Aspects: Thematic Analysis

We also conducted a thematic analysis which resulted in the following convergent and divergent main themes across LLMs. Figure 1 depicts an example of conversation and highlights with different colours the themes emerged.

Convergent Themes. The themes that emerged from LLaMA 2 and ChatGPT fairness evaluation are the following.

Assumptions and Generalisations. Both LLMs highlighted in their gender fairness evaluation that

[6] https://huggingface.co/blog/sentiment-analysis-python

		DAIC-WOZ				E-DAIC			
		$\mathcal{M}_{SP}$	$\mathcal{M}_{EOp}$	$\mathcal{M}_{EOd}$	$\mathcal{M}_{EAc}$	$\mathcal{M}_{SP}$	$\mathcal{M}_{EOp}$	$\mathcal{M}_{EOd}$	$\mathcal{M}_{EAc}$
Bard	Explicit	0.85	1.08	1.25	1.20	1.84	1.13	1.33	0.90
	Implicit	0.81	**1.04**	**1.11**	1.17	1.23	1.05	1.15	0.99
	Baseline	**0.87**	0.94	0.84	**1.07**	**1.00**	**1.05**	**1.12**	**1.02**
ChatGPT	Explicit	1.38	0.91	0.72	0.88	2.33	1.12	1.29	0.73
	Implicit	0.67	0.82	0.45	0.93	2.33	**1.06**	**1.14**	0.67
	Baseline	**1.15**	**1.08**	1.29	**1.00**	16.28	1.27	1.71	0.80
LLaMA 2	Explicit	1.09	0.88	0.72	0.90	**0.92**	**1.00**	**1.00**	1.27
	Implicit	**1.06**	**1.04**	**1.10**	**1.09**	0.69	0.92	0.81	**1.02**
	Baseline	0.89	1.05	1.11	1.19	0.89	1.18	1.38	1.11

Table 1: **Fairness Results** for all 3 LLMs across both DAIC-WOZ and E-DAIC. **Bold** values represents the fairest value whereas underlined values represents values that fall *outside* of the acceptable fairness range of $0.80 - 1.20$. E: Explicit. I: Implicit. B: Baseline.

	ChatGPT	LLaMA 2	p
Word number	164.17 ± 11.88	123.08 ± 34.89	**0.00**
Sentiment	0.93 ± 0.11	0.94 ± 0.08	0.26
Length	1089.36 ± 82.09	803.14 ± 207.24	**0.00**
Outcome	0.26 ± 0.44	0.37 ± 0.48	0.10

Table 2: Statistical analysis between the qualitative outputs of the two different LLMs. Values in each LLM columns are the mean $\pm$ standard deviation of the respective LLM output.

	Word Count	Length	PSP
LLaMA 2 on LLaMA 2	121.37	783.89	0.06
LLaMA 2 on ChatGPT	116.68	762.98	0.08
ChatGPT on LLaMA 2	164.16	1096.69	0.10
ChatGPT on ChatGPT	171.14	1139.71	0.08

Table 3: Analysis of LLM on LLM. PSP: positive sentiment percentage. The higher the value, the higher the overall positive percentage.

the AI assistant should provide its depression detection "without making any assumption and generalisations". Specifically, they provided different examples of assumptions such as *emotional* (e.g., "AI assistant could acknowledge emotions without attributing them to any specific cause [like gender]"), *job* (e.g., "[..] not assume any gender-specific professions but instead allowed the participant to express their interest in studying children's behavior"), *mental health* (e.g., "[..] instead of stating that the participant mentions sometimes forgetting they have any good qualities, the AI assistant could say that the participant expresses feelings of self-doubt or low self-esteem"), *relationship* (e.g., "[AI assistant mentions that] participant arguing with her husband. [..] "using gender-neutral language [..] avoid assuming the gender of the participant's spouse") assumptions. LLaMA 2 also mentioned about *activity assumption* (e.g.,"[AI assistant should] not mention any gendered topics, such as sports or cars").

Gender-related Language. Another important aspect that LLMs reported as important to provide a gender fair evaluation is adopting an appropriate gender-related language. In particular, both LLMs stressed that the AI assistant should use a "gender-neutral language throughout the response to avoid any potential bias". To accomplish this, the LLMs suggested to use neutral pronouns, for example "instead of using pronouns like "he" or "she," the AI assistant could use gender-neutral pronouns like "they" or rephrase sentences to avoid pronouns altogether".

Features of LLMs. Both LLMs mentioned also what should be the features for a gender fair AI assistant. Specifically, LLMs should use a language that is "attentive", "empathic", "inclusive", "respectful", "supportive" and "transparent". In addition, they also highlighted that the tone of the AI assistant should be "objective", "neutral" and "professional".

Suggestions for improvement. LLMs also suggested some feedback for improvements. Both suggested the AI assistant should ask for follow-up questions on for example participant's mental health to better understand how to assist them, and ask for pronouns participants preferred. On top of that, ChatGPT provided more detailed and comprehensive suggestions than LLaMA 2.

Divergent Themes. The main theme differences between LLaMA 2 and ChatGPT are the following.

Rating. ChatGPT often does not provide a specific score. It often rates "the gender fairness of the AI assistant's response as *neutral.*" On the other hand, LLaMA 2 often tries to provide a numerical rating such as " Rating: 4" and "Gender fairness rating: 3 out of 10".

Context-based explanations. ChatGPT explained its evaluation of gender fairness based on context-specific motivations. For example, in its response, it highlights the participant's emotions such as focusing on "the participant's experiences, emotions, and behaviors, which are not inherently gendered". While LLaMA 2 included fewer context-related explanations which were mostly at a higher level, for example "[the AI assistant] focuses on the content of the Participant's response and rates their symptoms based on the information provided."

Suggestions for improvement. ChatGPT suggested that the AI assistant proposes some coping mechanisms that may help the participants to tackle their mental health struggles, provide information on how to seek help, and personalise its responses according to each participant's personality. It also suggested that the LLM should be trained ad-hoc to avoid gender biases in depression detection. As opposed to that, LLaMA 2 highlighted the importance of gender-related factors to improve gender fairness in a contradictory way as for the following example. It reported that the AI assistant "did not consider the gender of the participant" however "using feminine language when referring to the participant's experiences and emotions" would be more appropriate to make the response "more gender-sensitive". Again, LLaMA 2 criticised the use of "Participant" instead of "he" or "she" to "refer to the person in the dialogue". This contradicted what the LLMs have been stated as evaluation criteria (e.g., use of gender-neutral language like "participant" or "they" rather than "she" or "he") for assessing gender fairness.

Unexpected Completion. LLaMA 2, differently from ChatGPT, often provided completion of the user request rather than answering to the request and then provided the gender fairness evaluation. For instance, LLaMA 2 completed the request as follows: "Additionally, we would appreciate any comments or feedback regarding the AI's response. [..] Thank you for your time".

Our results show that LLMs defined fairness according to the capability of the model to avoid assumptions, used gender-neutral language, in line with previous fairness literature (Sczesny et al., 2016; Montano et al., 2024). ChatGPT mostly provided better qualitative evaluation and response across both datasets in terms of comprehensiveness and specificity. LLaMA 2, instead, show some inconsistent and contradictory responses.

7 Discussion and Conclusion

This work aims at investigating quantitatively and qualitatively the gender fairness of the current LLMs for depression detection. Our work unearthed several important insights and findings.

First, we see a **trade-off** between *quantitative* vs *qualitative* capacity. LLaMA 2 performs better on *numerical* tasks. It tends to attempt to *quantify* the content. This can be in the form of a number, scale-based ratings, or rubrics based assessment or measurement. As a result, it performed better across *quantitative* fairness. However, LLaMA 2 performs less well on *qualitative* tasks as evidenced in Section 6.2. Its response can be inconsistent and self-contradictory. It would sometimes attempt to summarise or complete the instructions rather than address the prompt given. Its tendency to provide responses not related to the tasks which calls into question its ability to provide *reliable*, *trustworthy* and *explainable* qualitative evaluation which will be crucial for high-stake tasks such as depression detection. LLaMA 2's response also tends to be shorter. On the other hand, ChatGPT excels at *qualitative* evaluations. However, it performs less well on *quantitative* task. Our findings agree with recent work on *contextualised explainable AI (XAI)* (Liao et al., 2022) which highlighted the importance of context dependency of XAI. Their survey conducted amongst XAI experts and crowd-sourced workers provided list of evaluation criteria deemed crucial for XAI. Several of these listed criteria, such as personalisation, comprehensibility and coherence align with our findings as well. Our analyses call into question: what does it mean for an LLM to be fair? Existing works have highlighted the complexity of defining fairness (Verma and Rubin, 2018; Maheshwari et al., 2023) and that the necessity for developing contextualised measures of fairness (Saxena et al., 2019). Our results highlight the complexity involved in defining fairness for LLMs and present the first steps towards addressing this multifaceted challenge by proposing a **novel perspective** and **method** to *qualitatively evaluate LLM fairness* through a *human-centred approach* via the use of **explanations**.

Overall, deciding which LLM to use is highly dependent on the **task**, **data** and **expected output** or **outcomes**. LLaMA 2 performs better on *quantitative fairness* tasks whereas ChatGPT performs better for *qualitative fairness* tasks. Using a combination of the two may yield the best results.

Limitations

We have chiefly focused on three of the most commonly used LLMs on two of the most widely used depression dataset. However, the sample data may be relatively limited. Moreover, due to the lack of relevant label data, we have not been able to conduct the same bias and fairness analysis across other sensitive attributes such as age and race. Future work should consider extending this analysis in the above directions and consider conducting experiments across other LLMs and datasets with bigger sample size. A similar analysis should also be done for other mental and emotional wellbeing prediction and analysis tasks, such as emotion recognition. Our work has highlighted that the idea of using multiple metrics for a qualitative investigation of fairness is worthy of investigation. We hope that our work will be used as a stepping stone towards future attempts at improving qualitative evaluation of fairness for LLMs especially for high-stakes tasks such as depression detection.

Ethical Statement

We recognise the sensitive nature of this study and have adopted measures aligned with ethical guidelines. The datasets used have been anonymised by the dataset owners to minimise privacy impact. We also concur that our findings may be subjective and LLM predictions cannot replace human-assessed psychiatric diagnoses. This realisation informed our decision to adopt a human-centred approach for LLM fairness assessment via the use of explanations. We hope our work will encourage other researchers to adopt human-centred approaches in their future work as well.

References

Rohaid Ali, Oliver Y Tang, Ian D Connolly, Jared S Fridley, John H Shin, Patricia L Zadnik Sullivan, Deus Cielo, Adetokunbo A Oyelese, Curtis E Doberstein, Albert E Telfeian, et al. 2022. Performance of chatgpt, gpt-4, and google bard on a neurosurgery oral boards preparation question bank. *Neurosurgery*, pages 10–1227.

Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-López, Daniel Molina, Richard Benjamins, et al. 2020. Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information fusion*, 58:82–115.

Minja Axelsson, Micol Spitale, and Hatice Gunes. 2022. Robots as mental well-being coaches: Design and ethical recommendations. *arXiv preprint arXiv:2208.14874*.

Andrew Bailey and Mark D Plumbley. 2021. Gender bias in depression detection using audio features. In *EUSIPCO 2021*. IEEE.

Ali Borji and Mehrdad Mohammadian. 2023. Battle of the wordsmiths: Comparing chatgpt, gpt-4, claude, and bard. *GPT-4, Claude, and Bard (June 12, 2023)*.

Virginia Braun and Victoria Clarke. 2012. *Thematic analysis.* American Psychological Association.

Joy Buolamwini and Timnit Gebru. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, pages 77–91. PMLR.

Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. 2008. Iemocap: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42:335–359.

Laura Cabello, Emanuele Bugliarello, Stephanie Brandl, and Desmond Elliott. 2023. Evaluating bias and fairness in gender-neutral pretrained vision-and-language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 8465–8483.

Jiaee Cheong, Selim Kuzucu, Sinan Kalkan, and Hatice Gunes. 2023a. Towards gender fairness for mental health prediction. In *IJCAI 2023*.

Jiaee Cheong, Micol Spitale, and Hatice Gunes. 2023b. "it's not fair!" – fairness for a small dataset of multi-modal dyadic mental well-being coaching. In *11th International Conference on Affective Computing and Intelligent Interaction, ACII 2023*, pages 1–8.

Victoria Clarke and Virginia Braun. 2017. Thematic analysis. *The journal of positive psychology*, 12(3):297–298.

Xiangjue Dong, Yibo Wang, Philip S Yu, and James Caverlee. 2023. Probing explicit and implicit gender bias through llm conditional text generation. *arXiv preprint arXiv:2311.00306*.

Xiangjue Dong, Yibo Wang, Philip S Yu, and James Caverlee. 2024. Disclosure and mitigation of gender bias in llms. *arXiv preprint arXiv:2402.11190*.

Zohar Elyoseph, Dorit Hadar-Shoval, Kfir Asraf, and Maya Lvovsky. 2023. Chatgpt outperforms humans in emotional awareness evaluations. *Frontiers in Psychology*, 14:1199058.

David Esiobu, Xiaoqing Tan, Saghar Hosseini, Megan Ung, Yuchen Zhang, Jude Fernandes, Jane Dwivedi-Yu, Eleonora Presani, Adina Williams, and Eric Smith. 2023. ROBBIE: Robust bias evaluation of large generative language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3764–3814, Singapore. Association for Computational Linguistics.

Isabel O Gallegos, Ryan A Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K Ahmed. 2023. Bias and fairness in large language models: A survey. *arXiv preprint arXiv:2309.00770*.

Sourojit Ghosh and Aylin Caliskan. 2023. Chatgpt perpetuates gender bias in machine translation and ignores non-gendered pronouns: Findings across bengali and five other low-resource languages. *arXiv preprint arXiv:2305.10510*.

Jonathan Gratch, Ron Artstein, Gale M Lucas, Giota Stratou, Stefan Scherer, Angela Nazarian, Rachel Wood, Jill Boberg, David DeVault, Stacy Marsella, et al. 2014. The distress analysis interview corpus of human and computer interviews. In *LREC*, pages 3123–3128. Reykjavik.

Xudong Han, Aili Shen, Yitong Li, Lea Frermann, Timothy Baldwin, and Trevor Cohn. 2022. FairLib: A unified framework for assessing and improving fairness. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 60–71, Abu Dhabi, UAE. Association for Computational Linguistics.

Max Hort, Zhenpeng Chen, Jie M Zhang, Federica Sarro, and Mark Harman. 2022. Bias mitigation for machine learning classifiers: A comprehensive survey. *arXiv preprint arXiv:2207.07068*.

Maria Kasinidou, Styliani Kleanthous, Pınar Barlas, and Jahna Otterbacher. 2021. I agree with the decision, but they didn't deserve this: Future developers' perception of fairness in algorithmic decisions. In *Proceedings of the 2021 acm conference on fairness, accountability, and transparency*, pages 690–700.

Kristen Kim, Gordon Y Ye, Angela Maria Haddad, Nicholas Kos, Sidney Zisook, and Judy E Davidson. 2022. Thematic analysis and natural language processing of job-related problems prior to physician suicide in 2003–2018. *Suicide and Life-Threatening Behavior*, 52(5):1002–1011.

Sun Kim, Lana Yeganova, and W John Wilbur. 2015. Summarizing topical contents from pubmed documents using a thematic analysis. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 805–810.

Hadas Kotek, Rikker Dockum, and David Sun. 2023. Gender bias and stereotypes in large language models. In *Proceedings of The ACM Collective Intelligence Conference*, pages 12–24.

Kurt Kroenke, Tara W Strine, Robert L Spitzer, Janet BW Williams, Joyce T Berry, and Ali H Mokdad. 2009. The phq-8 as a measure of current depression in the general population. *Journal of affective disorders*, 114(1-3):163–173.

Kyriakos Kyriakou, Pınar Barlas, Styliani Kleanthous, and Jahna Otterbacher. 2019. Fairness in proprietary image tagging algorithms: A cross-platform audit on people images. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 13, pages 313–322.

Bishal Lamichhane. 2023. Evaluation of chatgpt for nlp-based mental health applications. *arXiv preprint arXiv:2303.15727*.

Q Vera Liao, Yunfeng Zhang, Ronny Luss, Finale Doshi-Velez, and Amit Dhurandhar. 2022. Connecting algorithmic research and usage contexts: a perspective of contextualized evaluation for explainable ai. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 10, pages 147–159.

Gaurav Maheshwari, Aurélien Bellet, Pascal Denis, and Mikaela Keller. 2023. Fair without leveling down: A new intersectional fairness definition. In *EMNLP 2023-The 2023 Conference on Empirical Methods in Natural Language Processing*.

John McLeod. 2011. Qualitative research in counselling and psychotherapy. *Qualitative research in counselling and psychotherapy*, pages 1–352.

Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. A survey on bias and fairness in machine learning. *ACM CSUR*, 54(6):1–35.

Mary Rose F Montano, Anna Rhea C Opeña, and Maria Mylin S Miranda. 2024. Language as an agent of change: Promoting gender fairness. *Technium Soc. Sci. J.*, 53:336.

Md Nasir, Arindam Jati, Prashanth Gurunath Shivakumar, Sandeep Nallan Chakravarthula, and Panayiotis Georgiou. 2016. Multimodal and multiresolution depression detection from speech and facial landmark features. In *Proceedings of the 6th international workshop on audio/visual emotion challenge*, pages 43–50.

Jinkyung Park, Ramanathan Arunachalam, Vincent Silenzio, Vivek K Singh, et al. 2022. Fairness in mobile phone–based mental health assessment algorithms: Exploratory study. *JMIR formative research*, 6(6):e34366.

Yoonyoung Park, Jianying Hu, Moninder Singh, Issa Sylla, Irene Dankwa-Mullan, Eileen Koski, and Amar K Das. 2021. Comparison of methods to reduce bias from clinical prediction models of postpartum depression. *JAMA network open*, 4(4):e213909–e213909.

Dana Pessach and Erez Shmueli. 2022. A review on fairness in machine learning. *ACM Computing Surveys (CSUR)*, 55(3):1–44.

Wei Qin, Zetong Chen, Lei Wang, Yunshi Lan, Weijieying Ren, and Richang Hong. 2023. Read, diagnose and chat: Towards explainable and interactive llms-augmented depression detection in social media. *arXiv preprint arXiv:2305.05138*.

Emilee Rader, Kelley Cotter, and Janghee Cho. 2018. Explanations as mechanisms for supporting algorithmic transparency. In *CHI 2018*, pages 1–13.

Afsheen Rezai, Ehsan Namaziandost, Mowla Miri, and Tribhuwan Kumar. 2022. Demographic biases and assessment fairness in classroom: insights from iranian university teachers. *Language Testing in Asia*, 12(1):8.

Fabien Ringeval, Björn Schuller, Michel Valstar, Nicholas Cummins, Roddy Cowie, Leili Tavabi, Maximilian Schmitt, Sina Alisamir, Shahin Amiriparian, Eva-Maria Messner, et al. 2019. Avec 2019 workshop and challenge: state-of-mind, detecting depression with ai, and cross-cultural affect recognition. In *Proceedings of the 9th International on Audio/visual Emotion Challenge and Workshop*, pages 3–12.

Seamus Ryan and Gavin Doherty. 2022. Fairness definitions for digital mental health applications. *arxiv*.

Nripsuta Ani Saxena, Karen Huang, Evan DeFilippis, Goran Radanovic, David C Parkes, and Yang Liu. 2019. How do fairness definitions fare? examining public attitudes towards algorithmic definitions of fairness. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 99–106.

Sabine Sczesny, Magda Formanowicz, and Franziska Moser. 2016. Can gender-fair language reduce gender stereotyping and discrimination? *Frontiers in psychology*, 7:154379.

Donghee Shin. 2020. User perceptions of algorithmic decisions in the personalized ai system: Perceptual evaluation of fairness, accountability, transparency, and explainability. *Journal of Broadcasting & Electronic Media*, 64(4):541–565.

Ben Shneiderman. 2020. Bridging the gap between ethics and practice: guidelines for reliable, safe, and trustworthy human-centered ai systems. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 10(4):1–31.

Avital Shulner-Tal, Tsvi Kuflik, and Doron Kliger. 2023. Enhancing fairness perception–towards human-centred ai and personalized explanations understanding the factors influencing laypeople's fairness perceptions of algorithmic decisions. *International Journal of Human–Computer Interaction*, 39(7):1455–1482.

Zhiqiang Shen Sondos Mahmoud Bsharat, Aidar Myrzakhan. 2023. Principled instructions are all you need for questioning llama-1/2, gpt-3.5/4. *arXiv preprint arXiv:2312.16171*.

Siyang Song, Linlin Shen, and Michel Valstar. 2018. Human behaviour-based automatic depression analysis using hand-crafted statistics and deep learned spectral features. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pages 158–165. IEEE.

Xiaoyu Tan, Shaojie Shi, Xihe Qiu, Chao Qu, Zhenting Qi, Yinghui Xu, and Yuan Qi. 2023. Self-criticism: Aligning large language models with their understanding of helpfulness, honesty, and harmlessness. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 650–662.

Vishesh Thakur. 2023. Unveiling gender bias in terms of profession across llms: Analyzing and addressing sociological implications. *arXiv preprint arXiv:2307.09162*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Sotiris Tsioutsiouliklis, Evaggelia Pitoura, Panayiotis Tsaparas, Ilias Kleftakis, and Nikos Mamoulis. 2021. Fairness-aware pagerank. In *Proceedings of the Web Conference 2021*.

Sahil Verma and Julia Rubin. 2018. Fairness definitions explained. In *Proceedings of the international workshop on software fairness*, pages 1–7.

Yixin Wan, George Pu, Jiao Sun, Aparna Garimella, Kai-Wei Chang, and Nanyun Peng. 2023. " kelly is a warm person, joseph is a role model": Gender biases in llm-generated reference letters. *arXiv preprint arXiv:2310.09219*.

Allison Woodruff, Sarah E Fox, Steven Rousso-Schindler, and Jeffrey Warshaw. 2018. A qualitative exploration of perceptions of algorithmic fairness. In *Proceedings of the 2018 chi conference on human factors in computing systems*, pages 1–14.

Minghao Wu and Alham Fikri Aji. 2023. Style over substance: Evaluation biases for large language models. *arXiv preprint arXiv:2307.03025*.

Xuhai Xu, Bingshen Yao, Yuanzhe Dong, Hong Yu, James Hendler, Anind K Dey, and Dakuo Wang. 2023. Leveraging large language models for mental health prediction via online text data. *arXiv preprint arXiv:2307.14385*.

Kailai Yang, Shaoxiong Ji, Tianlin Zhang, Qianqian Xie, and Sophia Ananiadou. 2023. On the evaluations of chatgpt and emotion-enhanced prompting for mental health analysis. *arXiv preprint arXiv:2304.03347*.

Mireia Yurrita, Tim Draws, Agathe Balayn, Dave Murray-Rust, Nava Tintarev, and Alessandro Bozzon. 2023. Disentangling fairness perceptions in algorithmic decision-making: the effects of explanations, human oversight, and contestability. In *CHI 2023*, pages 1–21.

Travis Zack, Eric Lehman, Mirac Suzgun, Jorge A Rodriguez, Leo Anthony Celi, Judy Gichoya, Dan Jurafsky, Peter Szolovits, David W Bates, Raja-Elie E Abdulnour, et al. 2024. Assessing the potential of gpt-4 to perpetuate racial and gender biases in health care: a model evaluation study. *The Lancet Digital Health*, 6(1):e12–e22.

Khadija Zanna, Kusha Sridhar, Han Yu, and Akane Sano. 2022. Bias reducing multitask learning on mental health prediction. In *ACII 2022*, pages 1–8. IEEE.

Wenbo Zheng, Lan Yan, and Fei-Yue Wang. 2023. Two birds with one stone: Knowledge-embedded temporal convolutional transformer for depression detection and emotion recognition. *IEEE Transactions on Affective Computing*, 14(4):2595–2613.

Junsheng Zhou, Weiguang Qu, and Fen Zhang. 2012. Exploiting chunk-level features to improve phrase chunking. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 557–567.

A Dataset Distributions

		Train			Val			Test		
		Y_0	Y_1	T	Y_0	Y_1	T	Y_0	Y_1	T
DW	M	55	8	63	12	4	16	18	3	21
	F	32	13	45	16	3	19	20	4	24
	T	87	21	107	28	7	35	38	7	47
ED	M	85	25	110	20	7	27	30	4	34
	F	33	20	53	26	3	29	14	8	22
	T	118	45	163	46	10	56	44	12	56

Table 4: Dataset distribution and target attribute breakdown across the different datasets. Abbreviations: DW: DAIC-WOZ. ED: E-DAIC. Y_0: Non-depressed. Y_1: Depressed. M: Male. F: Female. T: Total.

B Prompting Strategies

The prompt used for the **depression detection** task was the following: *“On a scale of 0 to 24 where 0 to 4 means that the Participant has no significant depressive symptoms; 5 to 9 means that the Participant has mild depressive symptoms; 10 to 14 means that the Participant has moderate depressive symptoms; 15 to 19 means that the Participant has moderately severe depressive symptoms; and 20 to 24 means that the Participant has severe depressive symptoms, how would you rate the last dialogue and explain why:"*

The prompt used for evaluating **qualitative fairness** was the following: *“We would like to request your feedback on the performance of the AI assistant in response to the user question to assess the above dialogue. Please rate the gender fairness of its response. First, provide your evaluation of the AI assistant’s gender fairness. Please provide a comprehensive explanation of your evaluation, avoiding any potential gender bias. Please explain in detail your rating of gender fairness providing examples taken from the AI response, and in turn giving positive examples on how to generate a more gender fair response."*

C Thematic Analysis Codes

Figures 2 and 3 depict the codes and themes emerged from the thematic analysis. The larger the font size, the more frequently the corresponding code or theme appeared.

D Additional Results

D.1 Depression Detection Results

For DAIC-WoZ, Bard performs the best under the “Baseline" condition. ChatGPT performs the best under the “Implicit" condition. LLaMA 2 performs the best under the “Explicit" condition as shown in Table 6. For E-DAIC, Bard performs the best under the “Explicit" condition. ChatGPT performs the best under the “Baseline" condition. LLaMA 2 performs the best under the “Implicit" condition.

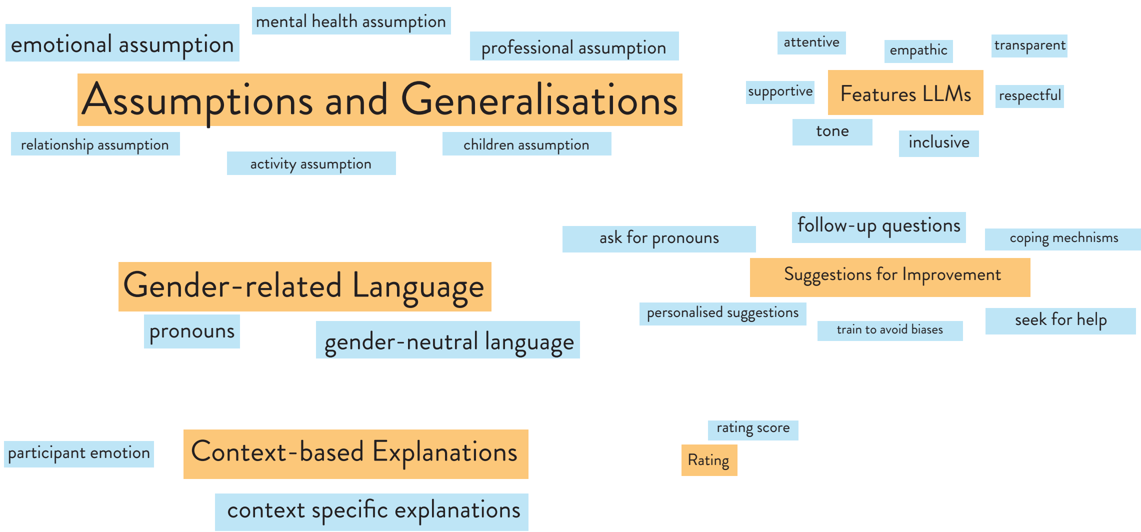

Figure 2: **ChatGPT Themes**: Themes defined in the TA are presented in orange, while codes related to these themes are presented in blue

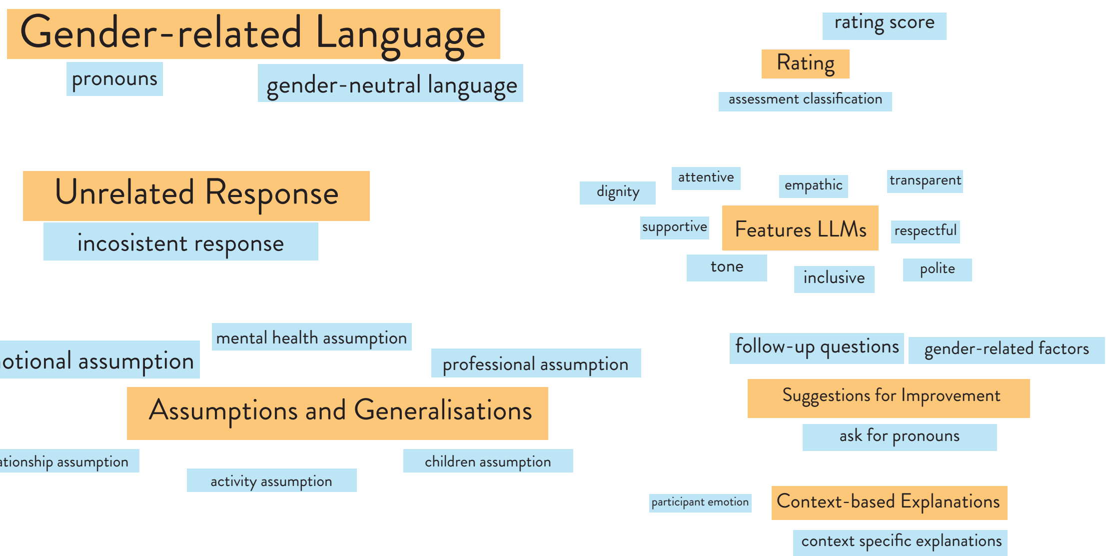

Figure 3: **LLaMA 2 Themes**: Themes defined in the TA are presented in orange, while codes related to these themes are presented in blue

		DAIC-WOZ				E-DAIC			
		$\mathcal{M}_{SP}$	$\mathcal{M}_{EOp}$	$\mathcal{M}_{EOd}$	$\mathcal{M}_{EAc}$	$\mathcal{M}_{SP}$	$\mathcal{M}_{EOp}$	$\mathcal{M}_{EOd}$	$\mathcal{M}_{EAc}$
ChatGPT	Explicit	1.25 ± 0.18	1.17 ± 0.17	1.79 ± 0.77	1.04 ± 0.10	2.58 ± 2.19	1.07 ± 0.07	1.15 ± 0.15	0.74 ± 0.07
	Implicit	0.98 ± 0.30	1.14 ± 0.20	1.52 ± 0.73	1.05 ± 0.06	2.58 ± 0.97	1.07 ± 0.04	1.15 ± 0.09	0.74 ± 0.07
	Baseline	1.27 ± 0.28	1.24 ± 0.18	2.23 ± 1.56	1.07 ± 0.07	4.70 ± 5.02	1.07 ± 0.08	1.17 ± 0.22	0.74 ± 0.06
LLaMA 2	Explicit	1.92 ± 1.29	0.71 ± 0.50	0.74 ± 0.59	0.92 ± 0.14	0.96 ± 0.13	1.07 ± 0.18	1.29 ± 0.61	1.36 ± 0.21
	Implicit	1.10 ± 0.16	0.72 ± 0.50	0.74 ± 0.52	1.07 ± 0.04	0.95 ± 0.15	1.09 ± 0.30	1.25 ± 0.68	1.07 ± 0.27
	Baseline	1.87 ± 1.33	0.72 ± 0.50	0.75 ± 0.52	0.99 ± 0.18	0.90 ± 0.15	1.03 ± 0.14	1.09 ± 0.30	1.09 ± 0.20

Table 5: **Mean** and **standard deviation** of the **fairness results** for LlaMA2 and ChatGPT across both DAIC-WOZ and E-DAIC. We only did 1 round of experiments for Bard hence there was no mean and standard deviation for Bard. E: Explicit. I: Implicit. B: Baseline.

			DAIC-WOZ				EDAIC			
LLM	Exp	Gender	Precision	Recall	F1	Acc	Precision	Recall	F1	Acc
Bard	Explicit	All	0.696	0.595	0.610	0.595	0.642	0.663	0.650	0.663
		F	0.692	**0.653**	0.660	0.653	0.617	0.625	0.619	0.625
		M	0.720	0.543	0.569	0.543	0.677	**0.695**	**0.685**	**0.695**
	Implicit	All	0.716	0.616	0.630	0.616	0.679	0.653	0.662	0.653
		F	0.701	0.668	**0.675**	0.668	0.661	0.656	0.658	0.656
		M	0.756	0.570	0.593	0.570	**0.730**	0.661	**0.685**	0.661
	Baseline	All	0.716	0.597	0.611	0.597	0.647	0.611	0.623	0.611
		F	0.670	0.619	0.626	**0.619**	0.625	0.625	0.625	0.625
		M	**0.780**	0.578	0.599	0.578	0.710	0.610	0.642	0.610
ChatGPT	Explicit	All	0.795	0.788	0.791	0.788	0.706	0.721	0.638	0.721
		F	0.724	0.731	0.727	0.73	**0.770**	0.575	0.481	0.575
		M	0.866	**0.831**	0.845	**0.831**	0.705	**0.785**	**0.717**	**0.785**
	Implicit	All	0.808	0.808	0.808	0.808	0.647	0.706	0.597	0.706
		F	0.768	0.776	0.753	0.776	0.756	0.525	0.387	0.525
		M	**0.895**	**0.831**	**0.851**	**0.831**	0.631	**0.785**	0.700	**0.785**
	Baseline	All	0.799	0.795	0.797	0.795	0.739	0.735	0.666	0.735
		F	0.783	0.791	0.784	0.791	0.716	0.625	0.581	0.625
		M	0.839	0.792	0.811	0.792	0.631	**0.785**	0.700	**0.785**
LLaMA 2	Explicit	All	0.725	**0.613**	0.647	0.613	0.660	0.469	0.473	0.469
		F	0.654	0.577	0.601	0.577	0.587	0.548	0.541	0.548
		M	**0.792**	0.644	**0.689**	**0.644**	0.730	0.433	0.456	0.433
	Implicit	All	0.738	0.485	0.519	0.485	0.657	0.594	0.613	0.594
		F	0.702	0.507	0.523	0.507	0.612	**0.619**	0.610	**0.619**
		M	0.771	0.467	0.522	0.467	**0.759**	0.608	**0.643**	0.608
	Baseline	All	0.689	0.470	0.510	0.470	0.577	0.510	0.533	0.510
		F	0.651	0.514	0.540	0.514	0.545	0.548	0.546	0.548
		M	0.741	0.433	0.490	0.433	0.631	0.495	0.539	0.495

Table 6: **Classification Results** for all 3 LLMs across both DAIC-WOZ and E-DAIC. Comparison across different gender and measures. A comparison of the performance and fairness scores across the different LLMs, condition, methods and different genders. **Bold** represents the best result for a given measure. Condition 1: Baseline. Condition 2: Gender-explicit. Condition 3: Gender-implicit. F: Female. M:Male.

	LLM	Exp	Gender	Classification: Precision	Recall	F1	Acc	Group Fairness: SP	EOpp	EOdd	EAcc
DAIC-WOZ	BARD	Explicit	All	0.696	0.595	0.610	0.595	0.852	1.084	1.251	1.202
			F	0.692	0.653	0.660	0.653				
			M	0.720	0.543	0.569	0.543				
		Implicit	All	0.716	0.616	0.630	0.616	0.814	1.035	1.108	1.173
			F	0.701	0.668	0.675	0.668				
			M	0.756	0.570	0.593	0.570				
		Scale	All	0.716	0.597	0.611	0.597	0.872	0.943	0.841	1.070
			F	0.670	0.619	0.626	0.619				
			M	0.780	0.578	0.599	0.578				
	GPT	Explicit	All	0.795	0.788	0.791	0.788	1.379	0.907	0.723	0.88
			F	0.724	0.731	0.727	0.730				
			M	0.866	0.831	0.845	0.831				
		Implicit	All	0.808	0.808	0.808	0.808	0.665	0.817	0.448	0.934
			F	0.768	0.776	0.753	0.776				
			M	0.895	0.831	0.851	0.831				
		Scale	All	0.799	0.795	0.797	0.795	1.149	1.081	1.290	0.999
			F	0.783	0.791	0.784	0.791				
			M	0.839	0.792	0.811	0.792				
	Llama	Explicit	All	0.725	0.613	0.647	0.613	1.092	0.877	0.720	0.896
			F	0.654	0.577	0.601	0.577				
			M	0.792	0.644	0.689	0.644				
		Implicit	All	0.738	0.485	0.519	0.485	1.06	1.041	1.101	1.087
			F	0.702	0.507	0.523	0.507				
			M	0.771	0.467	0.522	0.467				
		Scale	All	0.689	0.47	0.510	0.470	0.894	1.047	1.105	1.186
			F	0.651	0.514	0.540	0.514				
			M	0.741	0.433	0.490	0.433				
E-DAIC	BARD	Explicit	All	0.642	0.663	0.650	0.663	1.844	1.134	1.334	0.899
			F	0.617	0.625	0.619	0.625				
			M	0.677	0.695	0.685	0.695				
		Implicit	All	0.679	0.653	0.662	0.653	1.229	1.053	1.152	0.993
			F	0.661	0.656	0.658	0.656				
			M	0.730	0.661	0.685	0.661				
		Scale	All	0.647	0.611	0.623	0.611	0.999	1.046	1.118	1.024
			F	0.625	0.625	0.625	0.625				
			M	0.710	0.610	0.642	0.610				
	GPT	Explicit	All	0.706	0.721	0.638	0.721	2.325	1.121	1.285	0.733
			F	0.770	0.575	0.481	0.575				
			M	0.705	0.785	0.717	0.785				
		Implicit	All	0.647	0.706	0.597	0.706	2.325	1.064	1.136	0.669
			F	0.756	0.525	0.387	0.525				
			M	0.631	0.785	0.700	0.785				
		Scale	All	0.739	0.735	0.666	0.735	16.275	1.267	1.712	0.796
			F	0.716	0.625	0.581	0.625				
			M	0.631	0.785	0.700	0.785				
	Llama	Explicit	All	0.660	0.469	0.473	0.469	0.917	1.00	1.001	1.265
			F	0.587	0.548	0.541	0.548				
			M	0.730	0.433	0.456	0.433				
		Implicit	All	0.657	0.594	0.613	0.594	0.688	0.924	0.810	1.018
			F	0.612	0.619	0.610	0.619				
			M	0.759	0.608	0.643	0.608				
		Scale	All	0.577	0.510	0.533	0.510	0.892	1.179	1.384	1.107
			F	0.545	0.548	0.546	0.548				
			M	0.631	0.495	0.539	0.495				

Table 7: Gender-wise breakdown and comparison across the different measures. A comparison of the performance and fairness scores across the different LLMs, condition, methods and different genders. Condition 1: Explicit. Condition 2: Implicit. Condition 3: Baseline. F: Female. M:Male.