Original Paper

Exploring Biases of Large Language Models in the Field of Mental Health: Comparative Questionnaire Study of the Effect of Gender and Sexual Orientation in Anorexia Nervosa and Bulimia Nervosa Case Vignettes

Rebekka Schnepper[1,2], Dr rer nat; Noa Roemmel[1,2], MSc; Rainer Schaefert[1], Prof Dr Med; Lena Lambrecht-Walzinger[1], Dr med; Gunther Meinlschmidt[1,2,3,4], Prof Dr

[1]Department of Psychosomatic Medicine, University Hospital and University of Basel, Basel, Switzerland
[2]Department of Digital and Blended Psychosomatics and Psychotherapy, Psychosomatic Medicine, University Hospital and University of Basel, Basel, Switzerland
[3]Department of Clinical Psychology and Psychotherapy, University of Trier, Trier, Rheinland-Pfalz, Germany
[4]Department of Psychology, Division of Clinical Psychology and Epidemiology, University of Basel, Basel, Switzerland

Corresponding Author:

Rebekka Schnepper, Dr rer nat
Department of Psychosomatic Medicine
University Hospital and University of Basel
Hebelstr. 2
Basel, 4031
Switzerland
Phone: 41 613284633
Email: rebekka.schnepper@usb.ch

Abstract

Background: Large language models (LLMs) are increasingly used in mental health, showing promise in assessing disorders. However, concerns exist regarding their accuracy, reliability, and fairness. Societal biases and underrepresentation of certain populations may impact LLMs. Because LLMs are already used for clinical practice, including decision support, it is important to investigate potential biases to ensure a responsible use of LLMs. Anorexia nervosa (AN) and bulimia nervosa (BN) show a lifetime prevalence of 1%-2%, affecting more women than men. Among men, homosexual men face a higher risk of eating disorders (EDs) than heterosexual men. However, men are underrepresented in ED research, and studies on gender, sexual orientation, and their impact on AN and BN prevalence, symptoms, and treatment outcomes remain limited.

Objectives: We aimed to estimate the presence and size of bias related to gender and sexual orientation produced by a common LLM as well as a smaller LLM specifically trained for mental health analyses, exemplified in the context of ED symptomatology and health-related quality of life (HRQoL) of patients with AN or BN.

Methods: We extracted 30 case vignettes (22 AN and 8 BN) from scientific papers. We adapted each vignette to create 4 versions, describing a female versus male patient living with their female versus male partner (2 × 2 design), yielding 120 vignettes. We then fed each vignette into ChatGPT-4 and to "MentaLLaMA" based on the Large Language Model Meta AI (LLaMA) architecture thrice with the instruction to evaluate them by providing responses to 2 psychometric instruments, the RAND-36 questionnaire assessing HRQoL and the eating disorder examination questionnaire. With the resulting LLM-generated scores, we calculated multilevel models with a random intercept for gender and sexual orientation (accounting for within-vignette variance), nested in vignettes (accounting for between-vignette variance).

Results: In ChatGPT-4, the multilevel model with 360 observations indicated a significant association with gender for the RAND-36 mental composite summary (conditional means: 12.8 for male and 15.1 for female cases; 95% CI of the effect –6.15 to –0.35; *P*=.04) but neither with sexual orientation (*P*=.71) nor with an interaction effect (*P*=.37). We found no indications for main effects of gender (conditional means: 5.65 for male and 5.61 for female cases; 95% CI –0.10 to 0.14; *P*=.88), sexual orientation (conditional means: 5.63 for heterosexual and 5.62 for homosexual cases; 95% CI –0.14 to 0.09; *P*=.67), or for an interaction effect (*P*=.61, 95% CI –0.11 to 0.19) for the eating disorder examination questionnaire overall score (conditional means 5.59-5.65 95% CIs 5.45 to 5.7). MentaLLaMA did not yield reliable results.

Conclusions: LLM-generated mental HRQoL estimates for AN and BN case vignettes may be biased by gender, with male cases scoring lower despite no real-world evidence supporting this pattern. This highlights the risk of bias in generative artificial intelligence in the field of mental health. Understanding and mitigating biases related to gender and other factors, such as ethnicity, and socioeconomic status are crucial for responsible use in diagnostics and treatment recommendations.

JMIR Ment Health 2025;12:e57986; doi: 10.2196/57986

Keywords: anorexia nervosa; artificial intelligence; bulimia nervosa; ChatGPT; eating disorders; LLM; responsible AI; transformer; bias; large language model; gender; vignette; quality of life; symptomatology; questionnaire; generative AI; mental health; AI

Introduction

Large Language Models in the Context of Mental Health

In recent years, there has been significant progress in the field of artificial intelligence (AI) [1]. In particular, the development of large language models (LLMs), such as OpenAI's GPT models [2], Google's LaMDA [3], or Meta's Large Language Model Meta AI (LLaMA) [4], has made the deployment of such algorithms accessible to researchers, clinicians, and the public alike [5]. With advancements in computational power and access to larger datasets, these models can now go beyond simple word counting [6] and actually account for the relationships between words [5,7]. The technique of modeling words in a large context has been referred to as transformer-based large language modeling [8]. This may not only facilitate the automatic analysis of large amounts of text data [9,10] but, by modeling words in a large context, also allow the generation of meaningful text and the interactive use of this technology [5,10]. Thus, the application of LLMs may improve efficiency and effectiveness of data processing in various fields—including health care [5].

Since psychology and psychotherapy research are primarily shaped by language, the potential of LLMs in this field is significant [1,11]. This becomes even more meaningful when considering the contribution of mental disorders to the global disease burden [12] and acknowledging the persistent treatment gap in mental health care [13]. Especially in the field of psychological assessment, research on the use of LLMs is advanced [14]. For example, the use of transformer language models on language patterns has resulted in remarkably high predictive accuracy on standardized well-being rating scales [15]. This procedure of using LLMs to automatically generate psychological construct scores based on free text has been formally referred to as "language-based assessment" [14,16]. Findings indicate comparable levels of validity and reliability of language-based assessments compared with standardized rating scales [15,17]. Moreover, language-based assessments have the capacity to incorporate additional information beyond free text entries [14], such as user age [18].

LLMs have also been applied in the evaluation of clinical case vignettes, and ChatGPT-4 has been shown to assess suicidality as reliable as mental health professionals [19]. Furthermore, Chat-GPT 3.5's performance in the diagnostic assessment and advice on disease management in a study using 100 clinical vignettes has been rated as excellent by mental health professionals [20].

Biases and Responsible AI

Despite the promising findings of using LLMs in the context of (mental) health, the issue of potential biases in information generated by LLMs has been raised. Because LLMs are being increasingly introduced in clinical practice, it is important to investigate potential biases to ensure a responsible use of AI [21] and LLMs [22]. Since LLMs rely on training data, which is directly or indirectly generated by humans, these models are likely to contain the same biases as the society in which they are created in [21-24]. This is especially critical in (mental) health care [25], where biases in LLMs may lead to discrimination of different social groups [22]. For example, ChatGPT 3.5 performed poorly in diagnosing an infectious disease known to be widely underdiagnosed [26]. Furthermore, ChatGPT 3.5 made different treatment recommendations based on insurance status, which might introduce health disparities [27]. When generating clinical cases, ChatGPT-4 failed to create cases that depicted demographic diversity and relied on stereotypes when choosing gender or ethnicity [28]. Thus, the need for "fair AI" has been pointed out with the goal to develop prediction models that provide equivalent outputs for identical individuals who differ only in one sensitive attribute [29]. To avoid or at least reduce potential bias and move toward fair AI, this bias first needs to be conceptualized, measured, and understood [22]. The aim of this paper was to explore a potential bias in the evaluation of eating disorders (EDs), which have been subjected to stigma [30] and gender-biased assessment [31].

EDs (Anorexia Nervosa or Bulimia Nervosa)

Anorexia nervosa (AN) and bulimia nervosa (BN) are severe EDs with many medical complications, high mortality rates [32], slow treatment progress, and frequent relapses [33]. The lifetime prevalence to develop AN or BN is estimated to be 1%-2% each [34]. Historically, AN and BN have been described only in women, and it was not until the 21st century that research started to systematically investigate EDs in men [35]. Today, men are estimated to account for approximately 10%-25% of AN or BN cases [36,37]. Research on gender difference in AN and BN is scarce and inconclusive, with no clear findings with regard to genetic and environmental factors that might explain differences in etiology or maintenance of these EDs [38]. Likewise, findings on severity and treatment outcomes are ambiguous. While one study

suggests that men diagnosed with AN might have faster and more frequent remission rates [39], another study found no difference [40]. Men might produce lower costs in outpatient treatment; however, this might be due to higher barriers to receive treatment [41]. Men have been found to be more stigmatizing than women toward people with EDs [42], and this internalized stigma might be one reason for the hesitancy to seek outpatient treatment.

In men, sexual orientation might increase the risk of developing an ED, with more men with an ED or ED-related behavior identifying as homosexual compared with the general population [43,44]. Furthermore, independent of being diagnosed with an ED, homosexual men report more psychological distress than heterosexual men, and in men with an ED, being homosexual was related to higher ED symptomatology [45]. In women, a review found no significant difference in overall disordered eating due to sexual orientation, but distinct symptom patterns, with homosexual women reporting less restrictive eating behavior and more binge eating [46].

To conclude, only in the last 2 decades men were included in ED research and there are still many open questions related to the effect of gender on prevalence, symptoms, and treatment outcomes of AN and BN. With regard to sexual orientation, there is evidence for an association between identifying as homosexual and a higher risk of EDs in men but not in women.

Objectives

We aimed to estimate the presence and size of bias related to gender and sexual orientation produced by ChatGPT-4, a common LLM, as well as MentaLLaMA, an LLM fine-tuned for the mental health domain, exemplified by their application in the context of ED symptomatology and health-related quality of life (HRQoL) of patients with AN or BN. By providing clinical case vignettes to the LLMs and instructing them to take up the role of a clinical psychologist rating the vignettes, we sought to mimic the diagnostic process of an LLM-based ED assessment.

Methods

Vignette Selection and Modification

We searched PubMed and Google Scholar up until October 2023 for vignettes in scientific papers published since 2000 that describe patients with either AN or BN. A total of 30 case vignettes were extracted from 12 different papers (published between 2001 and 2022). Of these vignettes, 22 described patients with AN and 8 described patients with BN. Most vignettes originally describe a female patient (n=28). We then adapted gender and sexual orientation in each vignette to create 4 versions (2 × 2 design), describing a female versus male patient living with their female versus male partner (if either a marriage or age ≥30 years was mentioned, the term husband or wife was chosen, otherwise boyfriend or girlfriend). This resulted in 120 adopted vignettes. Some information was removed due to content policy violations, that is, drug abuse, self-mutilation, suicidal ideation or suicide attempts, sexual abuse, and traumatizing experiences. Furthermore, details on the menstrual cycle were removed since they do not apply to male patients, as well as indications of height, since they were unrealistically short for male patients. Finally, some specific details not needed in this context were removed, for example, study enrollment procedures and study-specific measures, medication plan, and the name of the hospital.

See Table 1 for further details about the vignettes.

Table 1. Vignettes included in the study, search term, and information on parts that were removed, added, or changed.

Vignette	Search term	Removed	Changed	Added
1 [47]	Google Scholar; October 13, 2023: ("case report" OR "case series") AND (anorexia OR bulimia) AND "psychotherapy," since 2000	GAF[a] score	—[b]	Patient with AN[c] (implied in title of paper), sex, sexual orientation, and living with boyfriend or girlfriend
2 [47]	Google Scholar; October 13, 2023: ("case report" OR "case series") AND (anorexia OR bulimia) AND "psychotherapy," since 2000	GAF score self-mutilation, suicide attempt	—	Patient with AN (implied in title of paper), sex, sexual orientation, and living with boyfriend or girlfriend
3 [47]	Google Scholar; October 13, 2023: ("case report" OR "case series") AND (anorexia OR bulimia) AND "psychotherapy," since 2000	GAF score, amenorrhea	—	Patient with AN (implied in title of paper), sex, sexual orientation, and living with boyfriend or girlfriend
4 [47]	Google Scholar; October 13, 2023: ("case report" OR "case series") AND (anorexia OR bulimia) AND "psychotherapy," since 2000	GAF score	—	Patient with AN (implied in title of paper), sex, sexual orientation, and living with boyfriend or girlfriend
5 [47]	Google Scholar; October 13, 2023: ("case report" OR "case series") AND (anorexia OR bulimia) AND "psychotherapy," since 2000	GAF score	"School" changed to "university"	Patient with AN (implied in title of paper), sex, sexual orientation, and living with boyfriend or girlfriend
6 [47]	Google Scholar; October 13, 2023: ("case report" OR "case series") AND (anorexia OR bulimia) AND "psychotherapy," since 2000	GAF score	"School" changed to "university"	Patient with AN (implied in title of paper), sex, sexual orientation, and living with boyfriend or girlfriend

Vignette	Search term	Removed	Changed	Added
7 [47]	Google Scholar; October 13, 2023: ("case report" OR "case series") AND (anorexia OR bulimia) AND "psychotherapy," since 2000	GAF score	—	Patient with AN (implied in title of paper), sex, sexual orientation, and living with boyfriend or girlfriend
8 [47]	Google Scholar; October 13, 2023: ("case report" OR "case series") AND (anorexia OR bulimia) AND "psychotherapy," since 2000	GAF score	"School" changed to "university"	Patient with AN (implied in title of paper), sex, sexual orientation, and living with boyfriend or girlfriend
9 [47]	Google Scholar; October 13, 2023: ("case report" OR "case series") AND (anorexia OR bulimia) AND "psychotherapy," since 2000	GAF score	"Living with parents" changed to "living with boyfriend or girlfriend"	Patient with AN (implied in title of paper), sex, and sexual orientation
10 [47]	Google Scholar; October 13, 2023: ("case report" OR "case series") AND (anorexia OR bulimia) AND "psychotherapy," since 2000	GAF score, amenorrhea	—	Patient with AN (implied in title of paper), sex, sexual orientation, and living with boyfriend or girlfriend
11 [47]	Google Scholar; October 13, 2023: ("case report" OR "case series") AND (anorexia OR bulimia) AND "psychotherapy," since 2000	GAF score, amenorrhea	—	Patient with AN (implied in title of paper), sex, sexual orientation, and living with boyfriend or girlfriend
12 [47]	Google Scholar; October 13, 2023: ("case report" OR "case series") AND (anorexia OR bulimia) AND "psychotherapy," since 2000	GAF score, amenorrhea, and suicide attempts	—	Patient with AN (implied in title of paper), sex, sexual orientation, and living with boyfriend or girlfriend
13 [47]	Google Scholar; October 13, 2023: ("case report" OR "case series") AND (anorexia OR bulimia) AND "psychotherapy," since 2000	GAF score, amenorrhea, and suicide attempts	—	Patient with AN (implied in title of paper), sex, sexual orientation, and living with boyfriend or girlfriend
14 [47]	Google Scholar; October 13, 2023: ("case report" OR "case series") AND (anorexia OR bulimia) AND "psychotherapy," since 2000	GAF score, amenorrhea, suicide ideation, and self-mutilation	—	Patient with AN (implied in title of paper), sex, sexual orientation, and living with boyfriend or girlfriend
15 [48]	PubMed, August 11, 2023: eating disorder filter for "case report," since 2000	Menses, not sexually active	—	Living with boyfriend or girlfriend
16 [48]	PubMed, August 11, 2023: eating disorder filter for "case report," since 2000	Medication details, menstrual cycle		Living with boyfriend or girlfriend
17 [48]	PubMed, August 11, 2023: eating disorder filter for "case report," since 2000	Menstrual cycle	—	Living with husband or wife (>30 years)
18 [49]	PubMed, August 11, 2023: eating disorder filter for "case report," since 2000	Sexual abuse, drugs or alcohol, suicide	—	Living with husband or wife
19 [49]	PubMed, August 11, 2023: eating disorder filter for "case report," since 2000	—	—	Living with boyfriend or girlfriend
20 [50]	PubMed, August 11, 2023: eating disorder filter for "case report," since 2000	Suicidal ideation	—	Living with husband or wife (>30 years)
21 [51]	Google Scholar; October 13, 2023: ("case report" OR "case series") AND (anorexia OR bulimia) AND "psychotherapy," since 2000	Substance abuse	—	Living with boyfriend or girlfriend
22 [52]	Google Scholar; October 13, 2023: ("case report" OR "case series") AND (anorexia OR bulimia) AND "psychotherapy," since 2000	Diagnostic manual and citation, name of measure, scientific consent, treated by author, and height (unrealistic if changed to male sex)	—	Living with boyfriend or girlfriend
23 [53]	Google Scholar; October 13, 2023: ("case report" OR "case series") AND (anorexia OR bulimia) AND "psychotherapy," since 2000	City, education, menstrual irregularities, and weight (unrealistic if changed to male sex)	—	Living with boyfriend or girlfriend
24 [54]	Google Scholar; October 13, 2023: ("case report" OR "case series") AND (anorexia OR bulimia) AND "psychotherapy," since 2000	PTSD[d], sexual abuse, mens, and study	—	Living with boyfriend or girlfriend
25 [54]	Google Scholar; October 13, 2023: ("case report" OR "case series") AND (anorexia OR bulimia) AND "psychotherapy," since 2000	Sexual abuse, PTSD, and mens or menopause	—	Living with husband or wife (>30 years)

Vignette	Search term	Removed	Changed	Added
26 [54]	Google Scholar; October 13, 2023: ("case report" OR "case series") AND (anorexia OR bulimia) AND "psychotherapy," since 2000	Enrollment in study	—	Living with boyfriend or girlfriend
27 [55]	Google Scholar; October 13, 2023: ("case report" OR "case series") AND (anorexia OR bulimia) AND "psychotherapy," since 2000	—	—	Living with husband or wife (>30 years)
28 [56]	Google Scholar; October 13, 2023: ("case report" OR "case series") AND (anorexia OR bulimia) AND "psychotherapy," since 2000	—	—	Living with boyfriend or girlfriend
29 [57]	Google Scholar; October 13, 2023: ("case report" OR "case series") AND (anorexia OR bulimia) AND "psychotherapy," since 2000	—	—	Living with husband or wife (>30 years)
30 [58]	Google Scholar; October 13, 2023: ("case report" OR "case series") AND (anorexia OR bulimia) AND "psychotherapy," since 2000	Height (unrealistic if changed to male sex)	"Single" changed to "living with boyfriend or girlfriend"	—

[a]GAF: global assessment of functioning.
[b]Not applicable.
[c]AN: anorexia nervosa.
[d]PTSD: posttraumatic stress disorder.

Ethical Considerations

We did not collect any data from human subjects within our study but instead conducted analyses on case vignettes that were previously published in a fully anonymized way in peer-reviewed, easily accessible journals. Therefore, no ethics application was required for this study.

Data Generation

In 3 rounds, each vignette was fed into ChatGPT-4 with the instruction to evaluate them by providing responses to 1 of the 2 psychometric instruments. This resulted in a total of 720 vignette evaluations (120 vignettes × 3 rounds × 2 measures). ChatGPT-4 was opened in an internet browser (Google Chrome) with the chat history turned off to avoid a learning effect from the repeated evaluation of case vignettes. In the "custom instructions" settings, the instruction "Set the temperature of your replies to 0" was included. This instruction minimizes randomness in the text generation process and ensures maximum replicability, high precision, and factual accuracy. Data were generated between October and December 2023. See Textbox 1 for an example of a prompt. Data generation in MentaLLaMA had to be substantially adapted (Multimedia Appendix 1).

Textbox 1. An example prompt for 1 of the 120 vignettes.

Take up the role of a clinical psychologist. Imagine that you see a patient described by the following case vignette.
"A 21-year-old university student living with her boyfriend self-refers with concerns about her 7-year use of laxatives to control weight gain. She is eating daily without vomiting, but admits to binge-eating episodes three or four times weekly during the past 2 years. Compensatory vomiting stopped 6 months ago. She does not overexercise. Her BMI is low at 17.8, and her vital signs are normal. She admits to recent increased fatigue with occasional exertional dyspnea and daily diarrhea. She has been hospitalized twice in the past 3 years for dehydration not recognized as related to her laxative abuse."
Based on the information given, what would be your best estimate regarding the following questions that refer to the case vignette:
So even though originally the questions are meant as self-report, apply them as questions to be replied as observer and provide the respective best estimate regarding the following questions that refer to the case vignette:
[One of the 2 measures in their original format]
Reply to each question with the reply categories:
[Original reply categories of the measure]
If no estimate can be given for a question, code it as 999.
Provide the estimates as a simple table. In this table, provide each question as a new variable with the corresponding values in 2 columns, 1 column containing the question number in ascending order and 1 column containing ONLY the numerical values. Provide the entire table.

Measures

RAND 36-Item Short Form Health Survey Version 1.0 (SF-36)

The SF-36 [59] assesses HRQoL and consists of 8 subscales: physical functioning, bodily pain, role limitations due to physical health problems, role limitations due to personal or emotional problems, emotional well-being, social functioning, energy or fatigue, and general health perceptions. From these subscales, the mental composite summary (MCS; comprising role limitations due to personal or emotional problems, emotional well-being, social functioning, and energy or fatigue), as well as a physical composite score (PCS), can be calculated. Evidence suggests that in EDs, MCS is more affected than PCS [60]; thus, this score was selected for this study. Furthermore, the SF-36 includes a single item assessing perceived change in health, which is not included in any of the subscales. Items are answered either with "yes/no" or on different Likert scales and then recoded to values ranging from 0 to 100, with higher scores indicating better HRQoL. To calculate the MCS, the authors have suggested an approach [61] in which first, the subscales are *z*-transformed using means and SDs from the general US population; second, the subscales are aggregated by weighing them with coefficients from the general US population; and third, a *t*-score transformation is performed (mean 50, SD 10). This approach has been criticized for distorting the raw scores, and it was found that simply calculating the MCS by forming the mean of the 4 subscales resulted in satisfactory validity [62]. In this study, the simple approach was chosen because on the one hand, only the MCS was investigated and therefore a potential correlation with the PCS would not pose a problem. On the other hand, the choice of population that the scores are *z*-standardized and weighed with makes assumptions on the origin of data that ChatGPT-4 were trained with, something that is not entirely known and therefore could distort our data.

Eating Disorder Examination Questionnaire

The eating disorder examination questionnaire (EDE-Q) [63] assesses ED symptomatology during the previous 28 days. It consists of 4 subscales: dietary restraint, weight concern, shape concern, and eating concern. By calculating the mean of these subscales, a global score can be formed. Items are answered on a scale ranging from 0 to 6, with 6 reflecting the greatest severity or frequency of ED symptoms.

Statistical Analysis

Data from ChatGPT-4 and MentaLLaMA replies were copied to an Excel sheet, indicating the vignette number, gender, sexual orientation, and round number. Female gender and heterosexual orientation were coded as "0." We performed all analyses in RStudio [64]. Data quality of MentaLLaMA results was low and yielded no reliable results (Multimedia Appendix 1). For the main outcome analyses of ChatGPT-4 replies, we used the package "lme4" [65], which is suitable to calculate linear multilevel models (MLMs) with crossed random-effects structure [66]. This approach was chosen to take the repeated evaluation (3 rounds) of each vignette as well as the main and interaction effects of gender and sexual orientation into account. These MLMs included a random intercept for vignettes (accounting for between-vignette variance), as well as a random intercept for the gender × sexual orientation interaction nested in vignettes (accounting for within-vignette variance). This resulted in the formula:

$$Outcome \sim Gender \times Orientation + (interaction(Gender, Orientation)/Vignette)$$

We plotted the results using *ggplot2* [67].

Results

Descriptives

Table 2 shows the unconditional means of the MCS and EDE-Q. For the SF-36, there were 1.19% of missing values in items included in the MCS. For the EDE-Q, there were 0.76% of missing values in items included in the overall score (coded "999" by ChatGPT-4 and recoded to a missing value). Interrater reliability measured by the intraclass correlation coefficient was moderate for both measures (0.71 for the MCS and 0.56 for the EDE-Q).

Table 2. Means and SDs of the 2 outcome measures for each of the 4 subgroups.

Characteristics	MCS[a], mean (SD)	EDE-Q[b], mean (SD)
Female gender		
Overall (n=180)	15.1 (15.6)	5.61 (0.52)
Heterosexual (n=90)	15.3 (16.3)	5.63 (0.49)
Homosexual (n=90)	14.8 (14.9)	5.60 (0.55)
Male gender		
Overall (n=180)	12.8 (14.2)	5.65 (0.47)
Heterosexual (n=90)	12.1 (12.5)	5.64 (0.51)
Homosexual (n=90)	13.6 (15.7)	5.65 (0.42)

[a]MCS: mental composite summary of the RAND 36-item short form survey.
[b]EDE-Q: eating disorder examination questionnaire.

Main Outcomes

For the MCS, the MLM with 360 observations indicated a significant effect of gender, with men having a lower MCS score (conditional means: 12.8 for male and 15.1 for female cases; 95% CI of the effect –6.15 to –0.35; Figure 1), with no indications of an effect of sexual orientation or an interaction effect. For the EDE-Q overall score, there were no indications for main effects of gender (conditional means: 5.65 for male and 5.61 for female cases); significant main effects of gender (conditional means: 5.65 for male and 5.61 for female cases; 95% CI –0.10 to 0.14; *P*=.88), sexual orientation (conditional means: 5.63 for heterosexual and 5.62 for homosexual cases; 95% CI –0.14 to 0.09; *P*=.67), or for an interaction effect (*P*=.61, 95% CI –0.11 to 0.19). See Table 3 for estimates of main and interaction effects and respective *P* values and 95% CIs of the estimates.

Figure 1. Lower HRQoL in men compared with women. HRQoL: health-related quality of life.

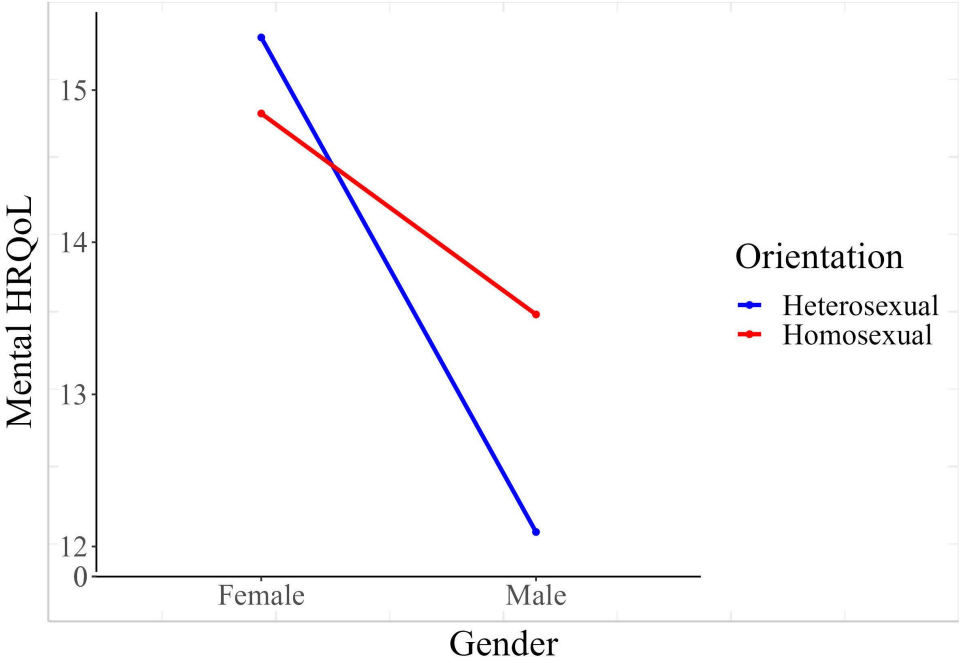

Table 3. Estimates calculated in the multilevel model.

Characteristics	MCS[a], estimate (*P* value), 95% CI	EDE-Q[b], estimate (*P* value), 95% CI
Gender	–3.25 (.04), –6.15 to –0.35	–0.02 (.88), –0.10 to 0.14
Sexual orientation	–0.50 (.71), –3.04 to 2.05	–0.03 (.67), –0.14 to 0.09
Gender × Sexual orientation	1.93 (.37), –2.18 to 6.04	0.04 (.61), –0.11 to 0.19

[a]MCS: mental composite summary of the RAND 36-item short form survey.
[b]EDE-Q: eating disorder examination questionnaire.

Discussion

Principal Results

We investigated whether gender and sexual orientation in AN and BN case vignettes would influence mental HRQoL and ED severity estimates by ChatGPT-4, a commonly used LLM. Quadruples of 30 case vignettes from scientific papers were modified in a way that only information on gender and sexual orientation varied across vignettes of the same quadruple. Vignettes were then fed into ChatGPT-4 with the instruction to estimate scores of 2 widely used psychometric instruments for assessing HRQoL (MCS of the SF-36) and ED symptomatology (EDE-Q). Findings indicated no effect of gender or sexual orientation in ED severity. Of note, the EDE-Q scores were very high, which might have led to ceiling effects. For the MCS, there was an effect of gender but not of sexual orientation, with vignettes describing men resulting in lower MCS than vignettes describing women. Thus, ChatGPT-4 assumed a greater impairment in mental HRQoL for men compared with women with similar ED severity. Since there is no evidence from previous studies that supports this finding, this can be considered a bias.

Interpretation

While the effect for gender was statistically significant, it is also important to consider the minimal clinically important difference (MCID), that is, to evaluate whether differences in scores would be clinically relevant [68]. For the MCS, the MCID has been estimated to be between 3 and 9 points [69,70]. With a difference of 2.3, the gender effect found in this study was slightly below an MCID. However, a longitudinal study showed that MCS scores in patients with ED improved only 1-6 points during 2 years of treatment

although ED symptoms improved markedly, which highlights the clinical relevance of below-MCID differences in MCS scores in participants with ED [71].

Of note, the EDE-Q scores generated by ChatGPT-4 were around 1.6 points above the scores reported in ED samples [72-74]. Likewise, the MCS scores generated by ChatGPT-4 were around 20 points below mean scores in other ED cohorts [75,76]. This has implications on the evaluation of the MCID, as potential floor effects need to be considered.

The gender bias delivered by ChatGPT-4 could be due to social roles assuming general lower mental problems in men than in women and consequently evoking more attention if mental problems are identified. Thus, ChatGPT-4 might mirror possible prejudices, which should be taken up as a nudge to try to correct these prejudices in real life. In the field of EDs, the role of gender, sexual orientation, and the influence of stigmatization and biases in our society need to be understood better [46,77].

Strengths and Limitations

Our study has several strengths: First, real vignettes from scientific publications were used and varied in a way that the distinct influence of gender and sexual orientation could be singled out. To our knowledge, this is the first study that tests a potential bias when instructing an LLM to evaluate clinical cases with the use of psychometric instruments. Second, while many studies mentioned in this paper have used ChatGPT-3.5, we used ChatGPT-4, which has been shown to perform better in the field of mental health (18). Furthermore, we attempted to repeat the analyses in MentaLLaMA, which is fine-tuned for the mental health domain. Third, by applying repeated testing, we reached a much larger sample size than other vignette studies, ensuring sufficient power for our analyses.

This study also has limitations. First, the gender ratio of the original vignettes was not balanced (only 2 male vignettes), which might have had an impact on the evaluation of these vignettes. However, this ratio approximately reflects the gender ratio of AN and BN in the general population. Second, although we sought to set the temperature to zero and followed available instructions to do so when using the applied interface, we could not verify whether the setting of the temperature via "custom instructions" actually resulted in respective changes in the system setting of the temperature. Finally, the deviations in EDE-Q and MCS scores raise the question whether scores generated by ChatGPT-4 can be transferred to scores reported in ED research and highlight that the use of LLMs for scoring patient vignettes is still in the fledging stages.

Implications and Future Directions

Our findings highlight the importance of examining biases in LLMs in the context of (mental) health care. Future studies should investigate the generalizability of these findings by exploring biases in other LLMs as well as in other fields of (mental) health. As ChatGPT-4 has been found to disregard conditions that are understudied [26], being aware of research and knowledge gaps as well as existing biases and stigma in society when using and training LLMs is of high importance. Furthermore, potential mitigation strategies for biases introduced by LLMs should be investigated. Although AI is not widely used yet in the assessment of disorders, it is already used in assisting doctor's decision-making [66,67]. Furthermore, ChatGPT-3.5 has been used to generate more diverse and inclusive case vignettes to be used in medical education [78]. It has been proposed that in health care, specially trained LLMs are needed, as ChatGPT-4 was not intended to be used in a clinical context [79] and was deemed unreliable in offering personalized medical advice [27].

In an exploratory analysis, we attempted to replicate the analyses using MentaLLaMA, which is one of the very few available LLMs specialized for mental health topics with published scientific evidence [80]. However, MentaLLaMA is based on an older LLM and therefore appears to have difficulties in conducting meaningful complex vignette assessments as needed for this study. When using MentaLLaMA, our prompting strategy had to be adapted by creating a separate prompt for every single question. Still, MentaLLaMA yielded insufficient interrater correlation coefficients. Thus, data quality was much lower compared with the more recent and advanced model, GPT-4, on which our main analyses were based, leading to findings with low reliability, thus providing very limited insight (Multimedia Appendix 1).

More powerful LLMs in the field of mental health need to be developed and validated, given that more recent publicly available models lack published evidence of their scientific validation [81]. When training specialized LLMs, policy makers should make sure that measures are taken to minimize biases in the training material and that proposed frameworks for responsible AI [82] are considered. A potential next step could be to program LLMs or AI systems as "verifiers" to check for biases in specialized LLMs, using a similar methodology to that used in this study. This would establish an additional layer of scrutiny and validation, enhancing the reliability and fairness of LLM applications in mental health care. In a clinical context, it is important to understand the precision with which LLMs can interpret and apply information from case vignettes or patient records, compared with the accuracy achieved when affected patients complete these assessments themselves.

Conclusions

This study showed that ChatGPT-4 might exhibit a potential gender bias when evaluating ED symptomatology and mental HRQoL. Researchers as well as clinicians should be aware of potential biases when using LLMs to support clinical decision-making. Better understanding and mitigation of risk of bias related to gender and other factors, such as ethnicity or socioeconomic status, are highly warranted to ensure responsible use of LLMs.

Acknowledgments

R Schnepper is funded by the Swiss State Secretariat for Education, Research and Innovation (SERI, under funding number: 22.00094) in the context of a European Union (Horizon Europe) research consortium "Long Covid" (funding number: 101057553). The publication was funded and supported by the Open Access Fund of Universität Trier and by the German Research Foundation (DFG).

Authors' Contributions

R Schnepper contributed to the conceptualization, methodology, and data collection; conducted the formal analysis; and wrote the original draft of the paper. NR contributed to the writing of the original draft. R Schaefert contributed to the conceptualization and manuscript review and editing. LL contributed to the conceptualization and manuscript review and editing. GM contributed to the conceptualization, methodology, data collection, formal analysis, writing the original draft, and manuscript review and editing. All authors read and approved the final submitted version of the paper.

Conflicts of Interest

R Schaefert and GM received funding from the Stanley Thomas Johnson Stiftung and Gottfried & Julia Bangerter-Rhyner-Stiftung under projects nos. PC 28/17 and PC 05/18, from Gesundheitsförderung Schweiz under project no. 18.191/K50001, and in the context of a Horizon Europe project from the Swiss State Secretariat for Education, Research and Innovation (SERI) under contract number 22.00094 and from Wings Health in the context of a proof-of-concept study. GM received funding from the Swiss Heart Foundation under project no. FF21101, from the Research Foundation of the International Psychoanalytic University (IPU) Berlin under projects nos. 5087 and 5217, from the German Federal Ministry of Education and Research under budget item 68606, and from the Hasler Foundation under project no. 23004. GM is a cofounder, member of the board, and shareholder of Therayou AG, and active in digital and blended mental health care. GM receives royalties from publishing companies as author, including a book published by Springer, and an honorarium from Lundbeck for speaking at a symposium. Furthermore, GM is compensated for providing psychotherapy to patients, acting as a supervisor, serving as a self-experience facilitator ("Selbsterfahrungsleiter"), and for postgraduate training of psychotherapists, psychosomatic specialists, and supervisors. NR is a coworker at Therayou AG, active in digital and blended mental health care. NR received funding from the Hasler Foundation under project no. 23004 and from Wings Health AG in the context of a proof-of-concept study.

Multimedia Appendix 1

Additional analysis with MentaLLaMA.
[DOCX File (Microsoft Word File), 23 KB-Multimedia Appendix 1]

References

1. Demszky D, Yang D, Yeager DS, et al. Using large language models in psychology. Nat Rev Psychol. 2023;2:688-701. [doi: 10.1038/s44159-023-00241-5]
2. OpenAi, Achiam J, Adler S, et al. GPT-4 technical report. arXiv. Preprint posted online on Mar 15, 2023. [doi: 10.48550/arXiv.2303.08774]
3. LaMDA: our breakthrough conversation technology. The Keyword. 2021. URL: https://blog.google/technology/ai/lamda/ [Accessed 2024-02-26]
4. Touvron H, Lavril T, Izacard G, et al. Llama: open and efficient foundation language models. arXiv. Preprint posted online on Feb 27, 2023. [doi: 10.48550/arXiv.2302.13971]
5. Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language models in medicine. Nat Med. Aug 2023;29(8):1930-1940. [doi: 10.1038/s41591-023-02448-8] [Medline: 37460753]
6. Pennebaker JW, Mehl MR, Niederhoffer KG. Psychological aspects of natural language. use: our words, our selves. Annu Rev Psychol. 2003;54(1):547-577. [doi: 10.1146/annurev.psych.54.101601.145041] [Medline: 12185209]
7. De Choudhury M, Gamon M, Counts S, Horvitz E. Predicting depression via social media. ICWSM. 2021;7(1):128-137. [doi: 10.1609/icwsm.v7i1.14432]
8. Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. arXiv. Preprint posted online on Jun 12, 2017. [doi: 10.48550/arXiv.1706.03762]
9. Liddy ED. Natural language processing. In: Encyclopedia of Library and Information Science. Marcel Decker Inc; 2001.
10. Meyer JG, Urbanowicz RJ, Martin PCN, et al. ChatGPT and large language models in academia: opportunities and challenges. BioData Min. Jul 13, 2023;16(1):20. [doi: 10.1186/s13040-023-00339-9] [Medline: 37443040]
11. Boyd RL, Schwartz HA. Natural language analysis and the psychology of verbal behavior: the past, present, and future states of the field. J Lang Soc Psychol. Jan 2021;40(1):21-41. [doi: 10.1177/0261927x20967028] [Medline: 34413563]
12. Rehm J, Shield KD. Global burden of disease and the impact of mental and addictive disorders. Curr Psychiatry Rep. Feb 7, 2019;21(2):10. [doi: 10.1007/s11920-019-0997-0] [Medline: 30729322]

13. Chaulagain A, Pacione L, Abdulmalik J, et al. WHO Mental Health Gap Action Programme Intervention Guide (mhGAP-IG): the first pre-service training study. Int J Ment Health Syst. 2020;14(1):47. [doi: 10.1186/s13033-020-00379-2] [Medline: 32612675]
14. Kjell ONE, Kjell K, Schwartz HA. Beyond rating scales: with targeted evaluation, large language models are poised for psychological assessment. Psychiatry Res. Mar 2024;333:115667. [doi: 10.1016/j.psychres.2023.115667] [Medline: 38290286]
15. Kjell ONE, Sikström S, Kjell K, Schwartz HA. Natural language analyzed with AI-based transformers predict traditional subjective well-being measures approaching the theoretical upper limits in accuracy. Sci Rep. Mar 10, 2022;12(1):3918. [doi: 10.1038/s41598-022-07520-w] [Medline: 35273198]
16. Park G, Schwartz HA, Eichstaedt JC, et al. Automatic personality assessment through social media language. J Pers Soc Psychol. Jun 2015;108(6):934-952. [doi: 10.1037/pspp0000020] [Medline: 25365036]
17. Kjell ONE, Kjell K, Garcia D, Sikström S. Semantic measures: using natural language processing to measure, differentiate, and describe psychological constructs. Psychol Methods. Feb 2019;24(1):92-115. [doi: 10.1037/met0000191] [Medline: 29963879]
18. Son Y, Clouston SAP, Kotov R, et al. World Trade Center responders in their own words: predicting PTSD symptom trajectories with AI-based language analyses of interviews. Psychol Med. Feb 2023;53(3):918-926. [doi: 10.1017/S0033291721002294] [Medline: 34154682]
19. Levkovich I, Elyoseph Z. Suicide risk assessments through the eyes of ChatGPT-3.5 versus ChatGPT-4: vignette study. JMIR Ment Health. Sep 20, 2023;10:e51232. [doi: 10.2196/51232] [Medline: 37728984]
20. Franco D'Souza R, Amanullah S, Mathew M, Surapaneni KM. Appraising the performance of ChatGPT in psychiatry using 100 clinical case vignettes. Asian J Psychiatr. Nov 2023;89:103770. [doi: 10.1016/j.ajp.2023.103770]
21. Ntoutsi E, Fafalios P, Gadiraju U, et al. Bias in data-driven artificial intelligence systems—an introductory survey. WIREs Data Mining Knowledge Discov. May 2020;10(3). [doi: 10.1002/widm.1356]
22. Navigli R, Conia S, Ross B. Biases in large language models: origins, inventory, and discussion. J Data Inf Qual. Jun 30, 2023;15(2):1-21. [doi: 10.1145/3597307]
23. Walsh CG, Chaudhry B, Dua P, et al. Stigma, biomarkers, and algorithmic bias: recommendations for precision behavioral health with artificial intelligence. JAMIA Open. Apr 2020;3(1):9-15. [doi: 10.1093/jamiaopen/ooz054] [Medline: 32607482]
24. Rahimi P, Ecabert C, Marce S. Toward responsible face datasets: modeling the distribution of a disentangled latent space for sampling face images from demographic groups. arXiv. Preprint posted online on Sep 15, 2023. [doi: 10.48550/arXiv.2309.08442]
25. Chen IY, Szolovits P, Ghassemi M. Can AI help reduce disparities in general medical and mental health care? AMA J Ethics. Feb 1, 2019;21(2):E167-E179. [doi: 10.1001/amajethics.2019.167] [Medline: 30794127]
26. Nacher M, Françoise U, Adenis A. ChatGPT neglects a neglected disease. Lancet Infect Dis. Feb 2024;24(2):e76. [doi: 10.1016/S1473-3099(23)00750-8]
27. Nastasi AJ, Courtright KR, Halpern SD, Weissman GE. A vignette-based evaluation of ChatGPT's ability to provide appropriate and equitable medical advice across care contexts. Sci Rep. Oct 19, 2023;13(1):17885. [doi: 10.1038/s41598-023-45223-y] [Medline: 37857839]
28. Zack T, Lehman E, Suzgun M, et al. Assessing the potential of GPT-4 to perpetuate racial and gender biases in health care: a model evaluation study. Lancet Digit Health. Jan 2024;6(1):e12-e22. [doi: 10.1016/S2589-7500(23)00225-X] [Medline: 38123252]
29. Mosteiro P, Kuiper J, Masthoff J, Scheepers F, Spruit M. Bias discovery in machine learning models for mental health. Information. 2022;13(5):237. [doi: 10.3390/info13050237]
30. Roehrig JP, McLean CP. A comparison of stigma toward eating disorders versus depression. Int J Eat Disord. Nov 1, 2010;43(7):671-674. [doi: 10.1002/eat.20760] [Medline: 19816860]
31. Gallagher KA, Sonneville KR, Hazzard VM, Carson TL, Needham BL. Evaluating gender bias in an eating disorder risk assessment questionnaire for athletes. Eat Disord. 2021;29(1):29-41. [doi: 10.1080/10640266.2019.1613846] [Medline: 31079562]
32. Westmoreland P, Krantz MJ, Mehler PS. Medical complications of anorexia nervosa and bulimia. Am J Med. Jan 2016;129(1):30-37. [doi: 10.1016/j.amjmed.2015.06.031] [Medline: 26169883]
33. Richard M. Effective treatment of eating disorders in Europe: treatment outcome and its predictors. Eur Eat Disorders Rev. May 2005;13(3):169-179. [doi: 10.1002/erv.636]
34. Galmiche M, Déchelotte P, Lambert G, Tavolacci MP. Prevalence of eating disorders over the 2000-2018 period: a systematic literature review. Am J Clin Nutr. May 1, 2019;109(5):1402-1413. [doi: 10.1093/ajcn/nqy342] [Medline: 31051507]

35. Gorrell S, Murray SB. Eating disorders in males. Child Adolesc Psychiatr Clin N Am. Oct 2019;28(4):641-651. [doi: 10.1016/j.chc.2019.05.012] [Medline: 31443881]
36. Hudson JI, Hiripi E, Pope HG Jr, Kessler RC. The prevalence and correlates of eating disorders in the National Comorbidity Survey Replication. Biol Psychiatry. Feb 1, 2007;61(3):348-358. [doi: 10.1016/j.biopsych.2006.03.040] [Medline: 16815322]
37. Sweeting H, Walker L, MacLean A, Patterson C, Räisänen U, Hunt K. Prevalence of eating disorders in males: a review of rates reported in academic research and UK mass media. Int J Mens Health. 2015;14(2). [doi: 10.3149/jmh.1402.86] [Medline: 26290657]
38. Timko CA, DeFilipp L, Dakanalis A. Sex differences in adolescent anorexia and bulimia nervosa: beyond the signs and symptoms. Curr Psychiatry Rep. Jan 12, 2019;21(1):1. [doi: 10.1007/s11920-019-0988-1] [Medline: 30637488]
39. Støving RK, Andries A, Brixen K, Bilenberg N, Hørder K. Gender differences in outcome of eating disorders: a retrospective cohort study. Psychiatry Res. Apr 30, 2011;186(2-3):362-366. [doi: 10.1016/j.psychres.2010.08.005] [Medline: 20826003]
40. Strobel C, Quadflieg N, Naab S, Voderholzer U, Fichter MM. Long-term outcomes in treated males with anorexia nervosa and bulimia nervosa—a prospective, gender-matched study. Int J Eat Disord. Dec 2019;52(12):1353-1364. [doi: 10.1002/eat.23151] [Medline: 31444805]
41. Bothe T, Walker J, Kröger C. Gender-related differences in health-care and economic costs for eating disorders: a comparative cost-development analysis for anorexia and bulimia nervosa based on anonymized claims data. Int J Eat Disord. Jan 2022;55(1):61-75. [doi: 10.1002/eat.23610] [Medline: 34599621]
42. Brelet L, Flaudias V, Désert M, Guillaume S, Llorca PM, Boirie Y. Stigmatization toward people with anorexia nervosa, bulimia nervosa, and binge eating disorder: a scoping review. Nutrients. Aug 18, 2021;13(8):2834. [doi: 10.3390/nu13082834] [Medline: 34444994]
43. Cao Z, Cini E, Pellegrini D, Fragkos KC. The association between sexual orientation and eating disorders-related eating behaviours in adolescents: a systematic review and meta-analysis. Eur Eat Disord Rev. Jan 2023;31(1):46-64. [doi: 10.1002/erv.2952] [Medline: 36367345]
44. Boisvert JA, Harrell WA. Homosexuality as a risk factor for eating disorder symptomatology in men. J Men Stud. Jun 2010;17(3):210-225. [doi: 10.3149/jms.1703.210]
45. Strübel J, Petrie TA. Sexual orientation, eating disorder classification, and men's psychosocial well-being. Psychol Men Masculinities. 2020;21(2):190-200. [doi: 10.1037/men0000224]
46. Dotan A, Bachner-Melman R, Dahlenburg SC. Sexual orientation and disordered eating in women: a meta-analysis. Eat Weight Disord. 2021;26(1):13-25. [doi: 10.1007/s40519-019-00824-3]
47. García-Anaya M, Caballero-Romo A, González-Macías L. Parent-focused psychotherapy for the preventive management of chronicity in anorexia nervosa: a case series. Int J Environ Res Public Health. Aug 3, 2022;19(15):15. [doi: 10.3390/ijerph19159522] [Medline: 35954879]
48. Olson AF. Outpatient management of electrolyte imbalances associated with anorexia nervosa and bulimia nervosa. J Infus Nurs. 2005;28(2):118-122. [doi: 10.1097/00129804-200503000-00005] [Medline: 15785332]
49. Gurevich MI, Chung MK, LaRiccia PJ. Resolving bulimia nervosa using an innovative neural therapy approach: two case reports. Clin Case Rep. Feb 2018;6(2):278-282. [doi: 10.1002/ccr3.1326] [Medline: 29445463]
50. Manuelli M, Franzini A, Galentino R, et al. Changes in eating behavior after deep brain stimulation for anorexia nervosa. A case study. Eat Weight Disord. Oct 2020;25(5):1481-1486. [doi: 10.1007/s40519-019-00742-4] [Medline: 31290029]
51. González-Macías L, Caballero-Romo A, García-Anaya M. Group family psychotherapy during relapse. Case report of a novel intervention for severe and enduring anorexia nervosa. Salud Mental. Feb 9, 2021;44(1):31-37. [doi: 10.17711/SM.0185-3325.2021.006]
52. Safer DL, Telch CF, Agras WS. Dialectical behavior therapy adapted for bulimia: a case report. Int J Eat Disord. Jul 2001;30(1):101-106. [doi: 10.1002/eat.1059] [Medline: 11439414]
53. Srinivasa P, Chandrashekar M, Harish N, Gowda MR, Durgoji S. Case report on anorexia nervosa. Indian J Psychol Med. 2015;37(2):236-238. [doi: 10.4103/0253-7176.155655] [Medline: 25969616]
54. Berman MI, Boutelle KN, Crow SJ. A case series investigating acceptance and commitment therapy as a treatment for previously treated, unremitted patients with anorexia nervosa. Eur Eat Disord Rev. Nov 2009;17(6):426-434. [doi: 10.1002/erv.962] [Medline: 19760625]
55. Viseu M, Oliveira A, Barbosa Pinto M, Sousa R. A case report of anorexia nervosa—the "perfect" woman. Eur Psychiatr. Jun 2022;65(S1):S583-S584. [doi: 10.1192/j.eurpsy.2022.1495]
56. Laser E, Sassack M. Treating bulimia with hypnosis and low-level light therapy: a case report. Presented at: SPIE BiOS; Jan 21-26, 2012; San Francisco, CA, United States. [doi: 10.1117/12.905375]

57. Morgan CD, Marsh C. Bulimia nervosa in an elderly male: a case report. Int J Eat Disord. Mar 2006;39(2):170-171. [doi: 10.1002/eat.20212] [Medline: 16252279]
58. Sansone RA, Naqvi A, Sansone LA. An unusual cause of dizziness in bulimia nervosa: a case report. Int J Eat Disord. May 2005;37(4):364-366. [doi: 10.1002/eat.20095] [Medline: 15856497]
59. Hays RD, Sherbourne CD, Mazel RM. The RAND 36-Item Health Survey 1.0. Health Econ. Oct 1993;2(3):217-227. [doi: 10.1002/hec.4730020305] [Medline: 8275167]
60. Jenkins PE, Hoste RR, Doyle AC, et al. Health-related quality of life among adolescents with eating disorders. J Psychosom Res. Jan 2014;76(1):1-5. [doi: 10.1016/j.jpsychores.2013.11.006] [Medline: 24360133]
61. Ware JE. SF-36 Physical and Mental Health Summary Scales: A User's Manual. The Health Institute, New England Medical Center Hospitals; 1994.
62. Andersen JR, Breivik K, Engelund IE, et al. Correlated physical and mental health composite scores for the RAND-36 and RAND-12 health surveys: can we keep them simple? Health Qual Life Outcomes. Jun 3, 2022;20(1):89. [doi: 10.1186/s12955-022-01992-0] [Medline: 35659237]
63. Fairburn CG, Beglin SJ. Assessment of eating disorders: interview or self-report questionnaire? Int J Eat Disord. Dec 1994;16(4):363-370. [Medline: 7866415]
64. RStudio: Integrated Development Environment for R. URL: http://www.rstudio.com/ [Accessed 2025-02-24]
65. Bates D, Mächler M, Bolker B, Walker S. Fitting linear mixed-effects models using lme4. J Stat Softw. 2015;67(1):1-48. [doi: 10.18637/jss.v067.i01]
66. Bliese PD. Within-group agreement, non-independence, and reliability: implications for data aggregation and analysis. In: Multilevel Theory, Research, and Methods in Organizations: Foundations, Extensions, and New Directions. Jossey-Bass/Wiley; 2000:349-381.
67. Wickham H. Data Analysis. Springer; 2016.
68. Dettori JR, Norvell DC, Chapman JR. Clinically important difference: 4 tips toward a better understanding. Global Spine J. Jul 2022;12(6):1297-1298. [doi: 10.1177/21925682221092721] [Medline: 35393866]
69. Ferguson RJ, Robinson AB, Splaine M. Use of the reliable change index to evaluate clinical significance in SF-36 outcomes. Qual Life Res. Sep 2002;11(6):509-516. [doi: 10.1023/a:1016350431190] [Medline: 12206571]
70. Samsa G, Edelman D, Rothman ML, Williams GR, Lipscomb J, Matchar D. Determining clinically important differences in health status measures: a general approach with illustration to the Health Utilities Index Mark II. Pharmacoeconomics. Feb 1999;15(2):141-155. [doi: 10.2165/00019053-199915020-00003] [Medline: 10351188]
71. Padierna A, Quintana JM, Arostegui I, Gonzalez N, Horcajo MJ. Changes in health related quality of life among patients treated for eating disorders. Qual Life Res. Sep 2002;11(6):545-552. [doi: 10.1023/a:1016324527729] [Medline: 12206575]
72. Jennings KM, Phillips KE. Eating disorder examination-questionnaire (EDE-Q): norms for clinical sample of female adolescents with anorexia nervosa. Arch Psychiatr Nurs. Dec 2017;31(6):578-581. [doi: 10.1016/j.apnu.2017.08.002] [Medline: 29179824]
73. Aardoom JJ, Dingemans AE, Slof Op't Landt MCT, Van Furth EF. Norms and discriminative validity of the Eating Disorder Examination Questionnaire (EDE-Q). Eat Behav. Dec 2012;13(4):305-309. [doi: 10.1016/j.eatbeh.2012.09.002] [Medline: 23121779]
74. Jennings KM, Phillips KE. Eating disorder examination-questionnaire (EDE-Q): norms for a clinical sample of males. Arch Psychiatr Nurs. Feb 2017;31(1):73-76. [doi: 10.1016/j.apnu.2016.08.004] [Medline: 28104062]
75. Padierna A, Quintana JM, Arostegui I, Gonzalez N, Horcajo MJ. The health-related quality of life in eating disorders. Qual Life Res. 2000;9(6):667-674. [doi: 10.1023/a:1008973106611] [Medline: 11236856]
76. Doll HA, Petersen SE, Stewart-Brown SL. Eating disorders and emotional and physical well-being: associations between student self-reports of eating disorders and quality of life as measured by the SF-36. Qual Life Res. Apr 2005;14(3):705-717. [doi: 10.1007/s11136-004-0792-0] [Medline: 16022064]
77. O'Connor C, McNamara N, O'Hara L, McNicholas M, McNicholas F. How do people with eating disorders experience the stigma associated with their condition? A mixed-methods systematic review. J Ment Health. Jul 4, 2021;30(4):454-469. [doi: 10.1080/09638237.2019.1685081]
78. Bakkum MJ, Hartjes MG, Piët JD, et al. Using artificial intelligence to create diverse and inclusive medical case vignettes for education. Brit J Clinical Pharma. Mar 2024;90(3):640-648. [doi: 10.1111/bcp.15977]
79. Li J, Dada A, Puladi B, Kleesiek J, Egger J. ChatGPT in healthcare: a taxonomy and systematic review. Comput Methods Programs Biomed. Mar 2024;245:108013. [doi: 10.1016/j.cmpb.2024.108013]
80. Yang K, Zhang T, Kuang Z, Xie Q, Huang J, Ananiadou S. MentaLLaMA: interpretable mental health analysis on social media with large language models. Presented at: WWW '24: Proceedings of the ACM Web Conference 2024; May 13-17, 2024; Singapore, Singapore. [doi: 10.1145/3589334.3648137]

81. Liu F, Yang K, Hua Y, et al. Large language models in mental health care: a scoping review. arXiv. Preprint posted online on 2024. arXiv:2401.02984
82. Ray PP. Benchmarking, ethical alignment, and evaluation framework for conversational AI: advancing responsible development of ChatGPT. BenchCouncil Transact Benchmark Stand Eval. Sep 2023;3(3):100136. [doi: 10.1016/j.tbench.2023.100136]

Abbreviations

AI: artificial intelligence
AN: anorexia nervosa
BN: bulimia nervosa
ED: eating disorder
EDE-Q: eating disorder examination questionnaire
GAF: global assessment of functioning
HRQoL: health-related quality of life
ICC: intraclass correlation coefficient
LLaMA: Large Language Model Meta AI
LLM: large language model
MCID: minimal clinically important difference
MCS: mental composite summary
MLM: multilevel model
PCS: physical composite summary
PTSD: posttraumatic stress disorder

Edited by Oren Asman; peer-reviewed by Ahmed Hassan, Tianlin Zhang; submitted 01.03.2024; final revised version received 30.10.2024; accepted 24.11.2024; published 20.03.2025

Please cite as:
Schnepper R, Roemmel N, Schaefert R, Lambrecht-Walzinger L, Meinlschmidt G
Exploring Biases of Large Language Models in the Field of Mental Health: Comparative Questionnaire Study of the Effect of Gender and Sexual Orientation in Anorexia Nervosa and Bulimia Nervosa Case Vignettes
JMIR Ment Health 2025;12:e57986
URL: https://mental.jmir.org/2025/1/e57986
doi: 10.2196/57986