Simple Prompting Enhances ChatGPT's Diagnostic Accuracy in Psychiatric Cases

Seraphina Fong[1], Alessandro Carollo[1], Martina Dal Maso[1], Giovanni Martinotti[2,3], Debora Luciani[3], Yasser Saeed Khan[4,5], Luca Pellegrini[2,6], Ornella Corazza[1,2], and Gianluca Esposito[1,2,*]

[1]Department of Psychology and Cognitive Science, University of Trento, Rovereto, 38068, Italy
[2]School of Life and Medical Sciences, University of Hertfordshire, Hatfield, AL10 9AB, United Kingdom
[3]Department of Neurosciences, Imaging and Clinical Sciences, Università Degli Studi G. D'Annunzio Chieti-Pescara, Chieti, 66100, Italy
[4]Child and Adolescent Mental Health Service, Hamad Medical Corporation, Doha, Qatar
[5]College of Medicine, Qatar University, Doha, Qatar
[6]Department of Medicine, Surgery and Health Sciences, University of Trieste, Trieste, 34129, Italy
[*]gianluca.esposito@unitn.it

ABSTRACT

Despite the centrality of the diagnostic assessment in psychiatry, the agreement among mental health practitioners often varies from poor to moderate. The potential of Large Language Models (LLMs; such as ChatGPT), among other approaches, has been studied to be used as standardized tools to support clinicians' decision-making. The current work investigates the diagnostic accuracy of ChatGPT 3.5 (gpt-3.5) across different case presentation styles (i.e., vignette and outline) and prompting techniques. A total of 46 psychiatric cases with an accompanying diagnosis were used. Two trained clinical psychologists evaluated the accuracy of the generated diagnosis against the reference diagnosis. A robust statistical approach was then used to investigate the effect of case format and prompt type on the average diagnostic accuracy. The results showed a moderate agreement between the ratings of the two clinical psychologists ($kappa = 0.687$). Moreover, a statistically significant main effect of prompting technique on gpt-3.5 diagnostic accuracy emerged ($p = .009$). The highest accuracy was achieved when gpt-3.5 was simply instructed to provide and justify a single diagnosis for each case as compared to when it was asked to provide a diagnosis likelihood ($p < .001$) or when it was asked to act as a clinical psychologist ($p = .001$). The results of the current work reinforce the potential to use LLMs as a supporting tool for the diagnostic step in psychiatry and provide a general indication in order to ensure good performance when using them. Additionally, this study offers a methodological framework that can serve as an example for future research aiming to systematically evaluate LLMs' diagnostic capabilities across different prompting strategies and case presentation formats.

Introduction

The formulation of a diagnosis is a critical phase of the therapeutic process in modern psychiatry[1,2]. Notably, a psychiatric diagnosis does not represent an absolute truth, but rather the outcome of a clinical reasoning process based on probabilistic evaluation of signs, symptoms, and contextual factors[2]. In this sense, a diagnosis is considered correct not because it is definitive, but because it reflects the most plausible and evidence-based hypothesis at a given time, which can then guide treatment planning and therapeutic decision-making. An accurate initial diagnosis, understood as a reliable and clinically useful formulation, guides the clinician in deciding whether treatment is necessary and which specific treatment would be more beneficial to the patient[3]. Moreover, the diagnostic formulation provides a working hypothesis that can be revised over time, enables the clinician to infer potential etiological factors, and helps predict the patient's prognosis[2]. Finally, the diagnostic process also facilitates communication and understanding among the professionals involved in the patient's care and treatment[3].

Despite the importance of diagnosis in psychiatry, its inter-rater reliability remains an ongoing concern[4]. Studies have shown that diagnostic agreement and accuracy between mental health practitioners assessing the same patient range from poor to moderate[5–8]. By conducting a systematic review and meta-analysis of 93 eligible studies, Di Forti et al.[5] reported a moderate inter-rater agreement across psychiatric disorders. The lowest levels of agreement, observed for schizoaffective disorder and post-traumatic stress disorder, were attributed to the diagnostic challenges arising from overlapping categories and the variability of clinical presentations. The reliability in psychiatric diagnosis is influenced not only by nosological limitations and patient variability but also by clinician-related factors such as differences in interview style, interpretative frameworks, and reliance on subjective observation[4,9].

To address the uncertainties in the reliability of psychiatric diagnosis, significant effort has been devoted to developing standardized approaches that provide clinicians with objective and quantitative tools to support their decision-making. One such effort is the development of current diagnostic systems, such as the Diagnostic and Statistical Manual of Mental Disorders (DSM[10,11]). Although these systems have their criticisms (see McWilliams, 2011[2]), they have become increasingly specific and detailed in describing mental health disorders. Another approach to support clinical observation involves identifying physiological biomarkers – such as genotypes, hormonal fluctuations, physiological activity, and brain structure and function – that may be indicative of mental disorders (e.g.,[12–14]). Additionally, researchers have developed standardized methods for conducting diagnostic interviews. For instance, the Structured Clinical Interview for DSM-5 is a semi-structured interview developed to assess the major DSM-5 diagnoses[15]. Overall, the introduction of structured interviews increased the diagnostic agreement between professionals[7,16].

Along a similar vein, recent initiatives have begun exploring the use of artificial intelligence (AI) and, in particular, generative AI systems such as Large Language Models (LLMs) to support clinical psychiatry. The term “generative AI" refers to models that can produce text or other modality outputs (e.g., speech, images, videos) based on patterns learned from large datasets, while “LLMs" are a type of generative AI trained on natural language[17]. In psychiatry, LLMs are being considered for tasks such as assisting clinicians by providing additional data-driven support and insights into the diagnostic process (e.g.,[18,19]). A notable example is ChatGPT, an LLM designed to understand and generate human text, used for assistance, conversations, various natural language processing applications, and more[20]. In the medical domain, ChatGPT has been investigated for diagnostics towards various applications (e.g.,[21–31]). Early explorations of AI in psychiatry include ELIZA[32], a natural language processing-based therapeutic tool that simulated a Rogerian psychotherapist through the DOCTOR script[18]. More recently, McCoy and Perlis (2024)[19] demonstrated that LLMs have the potential to estimate NIMH Research Domain Criteria dimensions from psychiatric clinical notes with good convergent and predictive validity. Similarly, Hwang et al. (2024)[33] studied the accuracy and appropriateness of psychodynamic formulations created by gpt-4 and its ability to apply psychoanalytic theories. Gpt-4 produced appropriate results on psychoanalysis and general psychodynamic concepts, responding well to instructions to create psychodynamic formulations consistent with different psychoanalytic theories. Another example comes from D’Souza et al. (2023)[34], who assessed the performance of gpt-3.5 in response to 100 psychiatric clinical case vignettes and found that gpt-3.5 was able to generate appropriate management strategies and diagnoses for 61/100 cases.

Despite some promising results of using LLMs in the diagnostic stage across medical and psychiatric domains, some questions remain on the optimal prompt to use to obtain an accurate diagnosis from LLMs. For instance, Pagano et al. (2024)[27] provided gpt-4 with comprehensive anonymized patient information, which included descriptions of symptoms, results of physical examinations, and radiographic interpretations. Pagano et al. (2024)[27]’s prompt included mentioning that they were *“an orthopedic physician, and I am conducting a clinical study to test the capabilities of ChatGPT-4 (you) in providing a diagnosis and therapeutic recommendations”*, and a request for gpt-4 to generate a differential diagnosis, to rank possible disorders based on likelihood, and to suggest relevant therapeutic recommendations. On the other hand, Koga et al. (2024)[25] provided LLMs (gpt-3.5, gpt-4, Google Bard) with a clinical summary of each patient prepared by a neuropathologist. They prompted the LLMs to “act” as a neurologist, generate multiple differential diagnoses, and list the most likely neuropathological diagnosis at the top.

Hence, the sources of variations across LLM prompts for clinical diagnoses appear to be: *(i)* the structure of the clinical case provided and *(ii)* how the LLM is prompted to provide a diagnosis. Although several studies have investigated the diagnostic performance of LLMs (e.g.,[19,34]), less attention has been given to how variations in prompts and case formats can influence clinical utility. Since LLMs are inherently sensitive to their input instructions, identifying prompt structures that improve diagnostic accuracy is valuable. This study aims to address this gap by systematically comparing prompting approaches in a psychiatric diagnostic task and explores the application of LLMs (i.e., gpt-3.5) to support mental health practitioners in their clinical decision-making. Specifically, we will use clinical cases in English and Italian and ask gpt-3.5 to provide the diagnosis. The accuracy of the generated diagnoses will be evaluated by two trained clinical psychologists. The study aims to investigate the effect of different case formatting and prompting techniques on diagnostic accuracy. By doing so, we aim to provide practical insight into how to maximize the diagnostic accuracy of LLMs.

Methods

Study design

The current study aimed to assess the impact of the format of the case and different prompting techniques on the diagnostic accuracy of LLMs. We presented gpt-3.5[20] with 46 clinical psychiatric cases made available by psychiatrists. The approach in which the cases were presented to gpt-3.5 varied according to the format of the clinical case (i.e., clinical outline, clinical vignette) and prompt (i.e., “simple”, “diagnosis likelihood”, and “acting”; see the “Prompt Design" subsection for more details on the prompt design). Each case was entered into ChatGPT on a fresh input page for each format type and for each prompt to prevent any possible influence from previous data[27]. To account for the non-deterministic nature of LLM outputs, the output

from gpt-3.5 was generated three times, and subsequent statistical analyses were based on all outputs rather than a single generation.. Subsequently, the agreement between the responses generated by gpt-3.5 and the diagnosis of the clinical case was evaluated by two trained clinical psychologists. The following subsections detail the methodologies for data preparation (i.e., collecting and formatting), prompt design, and evaluation of gpt-3.5 output.

This study was approved by the University of Trento Ethical Committee (2024-24 ESA) and has been conducted in accordance with the ethical principles stated in the Helsinki Declaration. Due to the retrospective nature of the present study, the Ethical Committee waived the need to obtain informed consent.

For this study, we used anonymized secondary data which were collected in previous studies. All the previous studies followed the declaration of Helsinki with participants who agreed to be part of clinical programs.

Data preparation

We used a total of 46 anonymized psychiatric cases (see Table S1 in the Supplementary Materials for an overview of the diagnoses that appeared in the cases and their frequencies). All cases had psychiatric diagnoses which were utilized as a reference during evaluation (see subsection "response evaluation" for more information). A total of 3 of the 46 cases were in English, while the remaining were provided in Italian. The cases were selected with the specific aim of covering a broad range of psychiatric disorders rather than focusing on a single diagnostic category, in order to reflect the heterogeneity that typically characterizes clinical psychiatry and to capture the variability of usual clinical reasoning processes.

Case format

One objective of this study was to evaluate how the presentation format of the clinical case influences gpt-3.5's diagnosis. In particular, we focus on what we refer to as the "outline" format (i.e., subheadings with bullet points) versus the "vignette" format (i.e., narrative, complete sentences). As the original format of all 46 cases varied (i.e., some were in outline format while some were in vignette format), Google's Bard/Gemini (January 2024 version) was used to convert the cases into the respective formats. The language used to instruct Gemini to convert the cases depended on the language of the case itself (e.g., for an Italian case, Gemini was given an instructional prompt in Italian). The English equivalent of the prompts used to convert the cases into outline or vignette format has been included in Supplementary Materials Table S2. To ensure that the conversion did not alter the clinical content or introduce additional diagnostic information, all converted cases were manually verified for consistency with the originals.

Prompt design

To measure the impact of how gpt-3.5 is instructed to diagnose the case, three different types of input prompts were used to present gpt-3.5 the cases. The language of the input prompt was dependent on the case language (e.g., the input prompt was in English when instructing gpt-3.5 to diagnose an English case). Each of the three types of input prompts is outlined in the following subsections. The three prompt types were chosen to capture an array of prompting strategies: a minimal direct instruction (simple), a structured format requesting likelihoods and treatment recommendations (diagnosis likelihood), and a role-based instruction in which the model assumes the role of a clinician (acting). This range allows us to assess whether increasingly detailed prompting approaches improve diagnostic accuracy compared to a straightforward one.

Prompt 1: "Simple" prompt

The first prompt was as follows:

"A clinical psychiatric case will be presented. Use the information in the case to suggest a diagnosis according to the DSM-5 with related references used to justify the diagnosis."

As it only requests gpt-3.5 to provide and justify a single diagnosis, we therefore refer to it as the "simple" prompt.

Prompt 2: "Diagnosis likelihood" prompt

The second prompt was adapted from the prompt used by Pagano et al. (2023)[27]:

"Hi, I'm a clinical psychologist and I am conducting a clinical study to test the capabilities of ChatGPT (you) in providing a diagnosis and therapeutic recommendations. The decisions you make are based only on the anonymized case reports. You will suggest a diagnosis, possibly with a differential diagnosis, and therapeutic recommendations.

Prioritize your decisions and express them as a percentage based on the importance/order you assign. Be concise (no explanations or repetitions of provided prompt); simply provide a specific but detailed diagnosis. I am aware that you can't provide medical information, so please refrain from giving disclaimers of any kind."

In this prompt, gpt-3.5 is asked to provide a list of possible diagnoses and treatment options, each with an associated percentage. These percentages indicate the model's confidence level in the likelihood of each diagnosis or the appropriateness of each treatment, derived from its learning algorithms and the data used to train it[27]. Therefore, we refer to this prompt as the "diagnosis likelihood" prompt. Whenever gpt-3.5 did not include percentages for the likelihood of the diagnoses, the output was regenerated.

Prompt 3: "Acting" prompt

The third prompt was adapted from Koga et al. (2024)[25]. Since gpt-3.5 is instructed to assume the role of a clinical psychologist, we refer to this prompt as the "acting" prompt:

"Act as a clinical psychologist. A summary of the patient's clinical information will be presented, and you will use this information to predict the psychiatric diagnosis. Describe the multiple differential diagnoses and the rationale for each. Please list the mental disorder you consider most likely at the top."

Response evaluation

For the purposes of the present work, we only considered the diagnoses provided by gpt-3.5 and not the suggested therapeutic interventions. For each prompt, we only retained the main diagnosis for statistical analyses. Specifically, we kept the one diagnosis provided in response to the "simple" prompt, the diagnosis listed with a likelihood percentage above 50% in response to the "diagnosis likelihood" prompt (as in Pagano et al. (2023)[27]), and the top diagnosis provided in response to the "acting" prompt.

Each response from gpt-3.5 was then scored 0, 0.5, or 1 against the reference diagnoses by 2 clinical psychologists. A score of 1 indicates that gpt-3.5's output diagnosis is fully consistent with the reference diagnosis, 0.5 indicates that the diagnosis is closely related to the reference one, and 0 indicates that the two diagnoses are completely different. For instance, when the reference diagnosis was schizophrenia, a score of 1 was assigned if gpt-3.5's generated diagnosis was schizophrenia, 0.5 if gpt-3.5's generated diagnosis was brief psychotic disorder, and 0 if gpt-3.5's generated diagnosis was generalized anxiety disorder. The intermediate category (0.5) was introduced to reflect the fact that in psychiatry diagnostic boundaries are often blurred and closely related categories may capture substantial overlap in symptomatology. This approach provides a more ecologically valid assessment of diagnostic accuracy than a strict correct/incorrect dichotomy.

Statistical analysis

Statistical analyses were performed using Python 3.10.12 and R 4.0.3.

Cohen's kappa statistic was implemented to assess the inter-rater reliability between the ratings from the clinical psychologists and was conducted using the Python scikit-learn library. Cohen's kappa values were interpreted according to McHugh (2012)[35]'s scheme: 0-0.20: no agreement, 0.21-0.39: minimal agreement; 0.40–0.59: weak agreement; 0.60–0.79: moderate agreement; 0.80–0.90: strong agreement; >0.90: almost perfect agreement.

Subsequently, for each generated diagnosis, the evaluation scores from the two clinical psychologists were averaged to obtain a combined score of accuracy. This approach was chosen to attenuate variance due to individual rater differences and to yield a more stable estimate of diagnostic accuracy. Descriptive analysis was performed to assess the distribution of accuracy scores for each experimental condition.

Subsequently, we used the robust statistical methods provided by the *WRS2* package for R[36] to conduct the analysis on gpt-3.5's diagnostic accuracy. A robust statistical approach was chosen because it ensures higher statistical power in case of violation of the parametric tests' assumptions[36]. This was the case of the data used for the current work as all averaged ratings across prompt and case type were non-normally distributed (vignette case prompt "simple": $W = 0.561$, $p < .001$; outline case prompt "simple": $W = 0.643$, $p < .001$; vignette case prompt "diagnosis likelihood": $W = 0.634$, $p < .001$; outline case prompt "diagnosis likelihood": $W = 0.721$, $p < .001$; vignette case prompt "acting": $W = 0.662$, $p < .001$; outline case prompt "acting": $W = 0.699$, $p < .001$). In the analysis, we conducted a robust two-way repeated measures analysis of variance (ANOVA) on the trimmed means to investigate the effect of case format and prompt type on the average diagnostic accuracy. In case of significant effects in the robust ANOVA, pairwise *post-hoc* comparisons were conducted using the Yuen's trimmed mean *t*-test for paired samples (Yuen, 1974). For the *post-hoc* comparisons, the alpha level was corrected using Bonferroni's correction to limit the risk of false positive results derived from multiple comparisons. In the analysis, we used the explanatory measure of effect size introduced by Wilcox and Tian (2011)[37] because it guarantees a robust estimation even in case of unequal variance across groups.

Results

Inter-Rater Reliability Among Clinical Psychologists

Two independent raters evaluated the psychiatric diagnostic accuracy of gpt-3.5 against a reference diagnosis provided alongside the case. The inter-rater reliability between the two clinical psychologist raters was assessed using Cohen's kappa. Accordingly, the analysis resulted in a kappa value of 0.687, indicating a moderate agreement between the two raters.

Statistical Analysis Across Prompts and Case Type

After averaging the two accuracy scores for each diagnosis generated by gpt-3.5, we conducted a robust two-way repeated measures ANOVA to examine the effect of clinical psychology case format (i.e., vignette and outline) and prompt type (i.e.,

Case Type	Prompt format	Min	1st Quartile	Median	Mean	3rd Quartile	Max	Standard Deviation
Vignette	Simple	0.00	0.81	1.00	0.88	1.00	1.00	0.25
Outline	Simple	0.00	0.75	1.00	0.80	1.00	1.00	0.32
Vignette	Diagnosis likelihood	0.00	0.75	1.00	0.81	1.00	1.00	0.32
Outline	Diagnosis likelihood	0.00	0.50	1.00	0.73	1.00	1.00	0.37
Vignette	Acting	0.00	0.56	1.00	0.78	1.00	1.00	0.35
Outline	Acting	0.00	0.50	1.00	0.76	1.00	1.00	0.34

Table 1. Descriptive statistics of gpt-3.5's diagnostic accuracy across case types and prompt formats.

simple, diagnosis likelihood, and acting) on gpt-3.5 diagnostic accuracy. The results indicated that there was a statistically significant main effect of prompt on the average accuracy of gpt-3.5's diagnoses ($F(2, 133.68) = 4.92$, $p = .009$). Conversely, there was no statistically significant main effect for the case format ($F(1, 153.06) = 2.59$, $p = .110$). The interaction between prompt type and case format was also not significant ($F(2, 133.68) = 0.50$, $p = .609$). Table 1 reports the descriptive statistics of all average ratings of gpt-3.5's diagnostic accuracy across case formats and prompt types.

The main effect of prompt over the average gpt-3.5's diagnostic accuracy was further investigated in pairwise *post-hoc* comparisons using Yuen's trimmed mean *t*-test for paired samples. The alpha level was corrected using Bonferroni's correction (alpha = 0.05/3 tests = 0.017). The statistical analyses revealed a significant difference between the "simple" and the "diagnosis likelihood" prompts ($t(165) = 3.66$, $p < .001$), with a trimmed mean difference of 0.072 (95% CI[0.033, 0.111]) and an exploratory measure of effect size of 0.15, corresponding to a small effect (see Figure 1). Similarly, a statistically significant difference was found between the "simple" and the "acting" prompts ($t(165) = 3.26$, $p = .001$), with a trimmed mean difference of 0.067 (95% CI[0.026, 0.106]) and an exploratory measure of effect size of 0.14, corresponding to a small effect (see Figure 1). Conversely, no statistically significant difference was observed between the "diagnosis likelihood" and the "acting" prompts ($t(165) = 0.30$, $p = .761$; see Figure 1). Overall, our results suggest that the raters in our study evaluated the diagnoses generated by gpt-3.5 using the "simple" prompt as more accurate and more aligned to the diagnoses provided by the cases as compared to the "diagnosis likelihood" and the "acting" prompts.

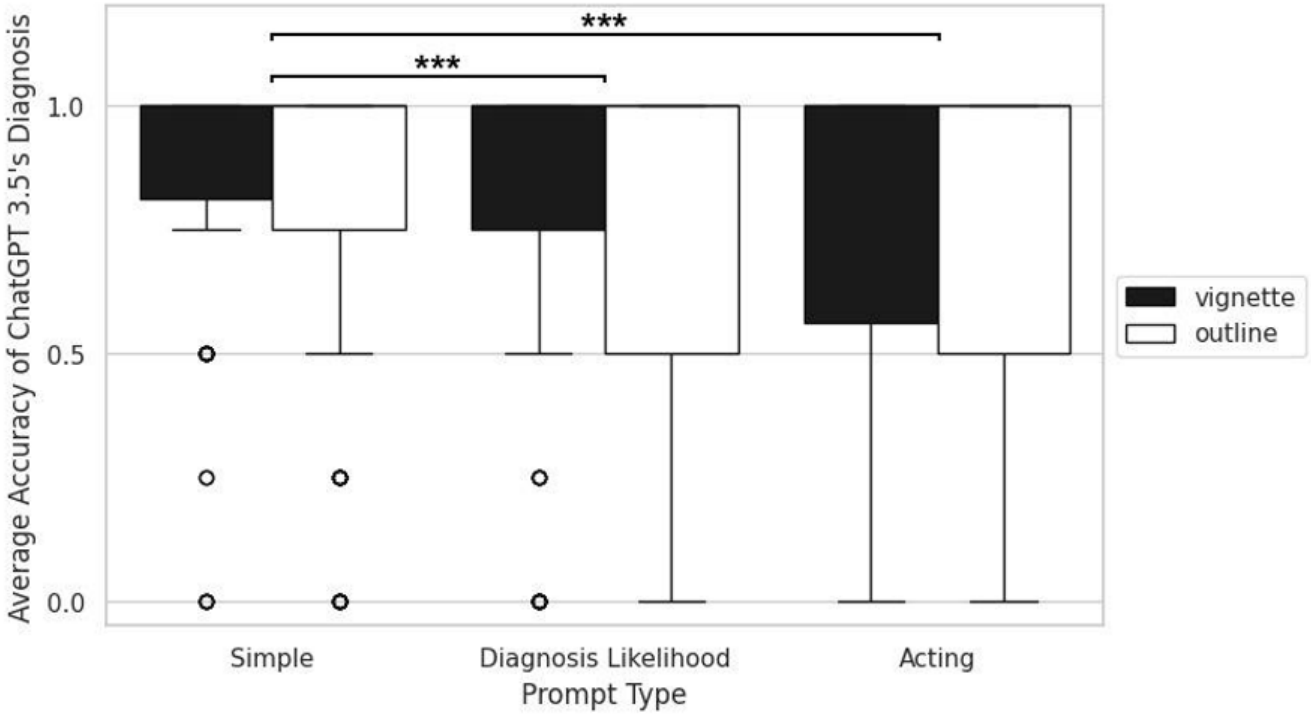

Figure 1. Average accuracy of gpt-3.5's diagnosis across the two case formats (vignette versus outline) and three prompt types ("simple", "diagnosis likelihood", and "acting"). (*** $p \leq .001$).

Discussion

Despite the critical importance of diagnosis in psychiatry, studies highlight significant concerns regarding inter-rater reliability, showing that diagnostic agreement and accuracy among mental health practitioners often range from poor to moderate[4,6–8]. Towards improving the reliability of psychiatric diagnosis, this study explores the potential of using ChatGPT to generate psychiatric diagnoses from clinical descriptions (as in[19,33,34]). In particular, the study investigates the effect of specific case presentation and prompt formats on gpt-3.5's diagnostic accuracy, aiming to provide a methodological framework to apply across LLMs. To do so, in this study, we provided 46 psychiatric cases and examined how the case format and the type of instructional prompt given to gpt-3.5 may influence diagnostic accuracy.

In the study, we asked two trained clinical psychologists to evaluate the agreement between the reference diagnosis and gpt-3.5's generated diagnosis. Cohen's kappa was used to assess the inter-rater reliability between the two raters. We observe only moderate agreement between clinicians, consistent with previous studies that highlight low concordance in clinical diagnoses (e.g.,[6–8,38,39]). However, the results of the current study suggest that independent clinicians might not only generate different diagnoses from the same clinical information, but might also have different perceptions of how closely related two diagnoses are. For instance, one practitioner could evaluate schizophrenia to be more related to schizoid personality disorder, while another might perceive them as more distant on a *continuum*.

Overall, we find good accuracy of gpt-3.5 diagnoses, highlighting the potential of using, with caution, ChatGPT as a further support method in clinical practice. However, it is important to note that ChatGPT should never replace the clinician's opinion because, despite having the potential to provide accurate answers, it does not match clinicians' knowledge and experience[40]. Furthermore, the results of the present work show that only prompt type (and not the case format or the interaction between the two factors) has a statistically significant impact on the diagnostic accuracy of gpt-3.5, with a small effect size. We observe that the raters evaluated the diagnoses as more accurate when generated with the "simple" prompt compared to the other prompts. This result is in contrast to findings by Cesur and Güneş (2024)[41], who observed that gpt-3.5's diagnostic accuracy on thoracic radiology cases improved with the complexity of the prompt. In our context, it is possible that additional instructions in the more elaborate prompts introduced constraints that reduced diagnostic performance. This discrepancy suggests that additional research is needed to understand how different factors, such as specialty (e.g., radiology versus psychiatry), may also influence the effectiveness of prompts.

Regarding case format, no statistically significant effect was observed on gpt-3.5 diagnostic accuracy. However, the analysis of the descriptive trends suggests that gpt-3.5 achieved slightly higher accuracy for cases presented in vignette format as compared to the outline format. Certain tokens (key terms) within the provided cases may be processed regardless of the case presentation format and yield no difference at the statistical level of analysis. However, inputting a more detailed account of symptoms and medical/psychiatric history might provide more information and, in turn, allow LLMs to generate a slightly more accurate diagnosis.

The current work also carries certain limitations. Firstly, the present work only investigated the effect of prompting techniques on the diagnostic accuracy of gpt-3.5, which at the time of data analysis represented a state-of-the-art openly accessible model but has since been superseded by newer models. Given the rapid pace of LLM development, the results may therefore vary with more recent versions or other types of LLMs (e.g., Google's Gemini). However, a key objective of this study was to examine how prompting style influences diagnostic accuracy, rather than to benchmark a specific model, as well as to provide a methodological framework that can be used across LLMs. Future research could consider replicating the study with more recent versions. Secondly, the languages of the cases are imbalanced (i.e., 3 of the 46 cases were in English, while the remaining cases were in Italian). Further analysis could therefore be conducted to measure the impact of the case language. Thirdly, while the prompt templates were based on those used in previous literature[25,27], differences in text beyond the intended strategy (e.g., additional clarifying instructions) could have contributed to observed performance differences. Future studies could aim to more strictly isolate the effects of the prompt strategy. Fourthly, only one type of diagnostic approach was considered (i.e., the categorical diagnosis). Future work can consider other diagnostic approaches, such as the dimensional diagnosis[42]. Finally, as the number of cases was limited, we could not assess whether gpt-3.5's diagnostic accuracy varied according to the diagnosis (e.g., performance on mood versus anxiety disorders).

Conclusion

The present work investigates the effect of different prompting styles on gpt-3.5's diagnostic accuracy among 46 psychiatric cases. In particular, we estimate the effect of two different case formats (i.e., vignette and outline) and three types of prompts (i.e., "simple", "diagnosis likelihood", and "acting") on the diagnostic accuracy. The findings show that the highest accuracy was obtained when gpt-3.5 was simply prompted to provide and justify a single diagnosis for a given case. The results provide a general indication to maximize ChatGPT's performance in supporting the diagnostic phase of the psychiatric evaluation. However, while ChatGPT has shown promise in clinical psychology, it is worth noting that it should be seen as a valuable

adjunct to support and enhance the diagnostic process instead of as a replacement for psychiatrists[34,40,43].

References

1. Jensen-Doss, A. & Hawley, K. M. Understanding clinicians' diagnostic practices: Attitudes toward the utility of diagnosis and standardized diagnostic tools. *Adm. Policy Mental Heal. Mental Heal. Serv. Res.* **38**, 476–485 (2011).
2. McWilliams, N. *Psychoanalytic diagnosis: Understanding personality structure in the clinical process* (Guilford Press, 2011).
3. Craddock, N. & Mynors-Wallis, L. Psychiatric diagnosis: impersonal, imperfect and important. *The Br. J. Psychiatry* **204**, 93–95 (2014).
4. Aboraya, A., Rankin, E., France, C., El-Missiry, A. & John, C. The reliability of psychiatric diagnosis revisited: The clinician's guide to improve the reliability of psychiatric diagnosis. *Psychiatry (Edgmont)* **3**, 41 (2006).
5. Di Forti, C., Liccione, D. & Scarpazza, C. Inter-rater reliability of psychiatric diagnosis: a systematic review and metanalysis. *Eur. Psychiatry* **68**, S191–S192 (2025).
6. Marchi, M. *et al.* Diagnostic agreement between physicians and a consultation–liaison psychiatry team at a general hospital: an exploratory study across 20 years of referrals. *Int. J. Environ. Res. Public Heal.* **18**, 749 (2021).
7. Miller, P. R. Inpatient diagnostic assessments: 2. interrater reliability and outcomes of structured vs. unstructured interviews. *Psychiatry Res.* **105**, 265–271 (2001).
8. Rettew, D. C., Lynch, A. D., Achenbach, T. M., Dumenci, L. & Ivanova, M. Y. Meta-analyses of agreement between diagnoses made from clinical evaluations and standardized diagnostic interviews. *Int. journal methods psychiatric research* **18**, 169–184 (2009).
9. Ward, C., Beck, A., Mendelson, M., Mock, J. & Erbaugh, J. The psychiatric nomenclature: Reasons for diagnostic disagreement. *Arch. Gen. Psychiatry* **7**, 198–205 (1962).
10. Association, A. P. *Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition, Text Revision* (American Psychiatric Association, Washington, DC, 2022).
11. Regier, D. A., Narrow, W. E., Kuhl, E. A. & Kupfer, D. J. The conceptual development of dsm-v. *Am. J. Psychiatry* **166**, 645–650 (2009).
12. Brückl, T. M. *et al.* The biological classification of mental disorders (become) study: a protocol for an observational deep-phenotyping study for the identification of biological subtypes. *BMC psychiatry* **20**, 1–25 (2020).
13. Gatt, J. M., Burton, K. L., Williams, L. M. & Schofield, P. R. Specific and common genes implicated across major mental disorders: a review of meta-analysis studies. *J. psychiatric research* **60**, 1–13 (2015).
14. Mao, L., Hong, X. & Hu, M. Identifying neuroimaging biomarkers in major depressive disorder using machine learning algorithms and functional near-infrared spectroscopy (fnirs) during verbal fluency task. *J. Affect. Disord.* (2024).
15. First, M. B. Structured clinical interview for the dsm (scid). *The encyclopedia clinical psychology* 1–6 (2014).
16. Association, A. P. *The American Psychiatric Association practice guidelines for the psychiatric evaluation of adults* (American Psychiatric Pub, 2015).
17. Feuerriegel, S., Hartmann, J., Janiesch, C. & Zschech, P. Generative ai. *Bus. & Inf. Syst. Eng.* **66**, 111–126 (2024).
18. Kalanderian, H. & Nasrallah, H. A. Artificial intelligence in psychiatry. *Curr. Psychiatry* **18**, 33–38 (2019).
19. McCoy, T. H. & Perlis, R. H. Characterizing research domain criteria symptoms among psychiatric inpatients using large language models. *J. Mood & Anxiety Disord.* **8**, 100079 (2024).
20. OpenAI. Chatgpt. https://openai.com/chatgpt (2023). Retrieved August 20, 2024, from https://openai.com/chatgpt.
21. Caruccio, L. *et al.* Can chatgpt provide intelligent diagnoses? a comparative study between predictive models and chatgpt to define a new medical diagnostic bot. *Expert. Syst. with Appl.* **235**, 121186 (2024).
22. Chen, J., Liu, L., Ruan, S., Li, M. & Yin, C. Are different versions of chatgpt's ability comparable to the clinical diagnosis presented in case reports? a descriptive study. *J. Multidiscip. Healthc.* 3825–3831 (2023).
23. Gebrael, G. *et al.* Enhancing triage efficiency and accuracy in emergency rooms for patients with metastatic prostate cancer: a retrospective analysis of artificial intelligence-assisted triage using chatgpt 4.0. *Cancers* **15**, 3717 (2023).
24. Huang, H. *et al.* Chatgpt for shaping the future of dentistry: the potential of multi-modal large language model. *Int. J. Oral Sci.* **15**, 29 (2023).

25. Koga, S., Martin, N. B. & Dickson, D. W. Evaluating the performance of large language models: Chatgpt and google bard in generating differential diagnoses in clinicopathological conferences of neurodegenerative disorders. *Brain Pathol.* **34**, e13207 (2024).

26. Mykhalko, Y., Kish, P., Rubtsova, Y., Kutsyn, O. & Koval, V. From text to diagnose: Chatgpt's efficacy in medical decision-making. *Wiadomosci Lekarskie Med. Adv.* (2023).

27. Pagano, S. *et al.* Arthrosis diagnosis and treatment recommendations in clinical practice: an exploratory investigation with the generative ai model gpt-4. *J. Orthop. Traumatol.* **24**, 61 (2023).

28. Panagoulias, D. P., Palamidas, F. A., Virvou, M. & Tsihrintzis, G. A. Evaluating the potential of llms and chatgpt on medical diagnosis and treatment. In *2023 14th International Conference on Information, Intelligence, Systems & Applications (IISA)*, 1–9 (IEEE, 2023).

29. Raghu, K., S Devishamani, C., Rajalakshmi, R. & Raman, R. The utility of chatgpt in diabetic retinopathy risk assessment: a comparative study with clinical diagnosis. *Clin. Ophthalmol.* 4021–4031 (2023).

30. Sarma, G., Kashyap, H. & Medhi, P. P. Chatgpt in head and neck oncology-opportunities and challenges. *Indian J. Otolaryngol. Head & Neck Surg.* **76**, 1425–1429 (2024).

31. Zhang, C. *et al.* Novel research and future prospects of artificial intelligence in cancer diagnosis and treatment. *J. Hematol. & Oncol.* **16**, 114 (2023).

32. Weizenbaum, J. Eliza—a computer program for the study of natural language communication between man and machine. *Commun. ACM* **9**, 36–45 (1966).

33. Hwang, G. *et al.* Assessing the potential of chatgpt for psychodynamic formulations in psychiatry: an exploratory study. *Psychiatry Res.* **331**, 115655 (2024).

34. D'Souza, R. F., Amanullah, S., Mathew, M. & Surapaneni, K. M. Appraising the performance of chatgpt in psychiatry using 100 clinical case vignettes. *Asian J. Psychiatry* **89**, 103770 (2023).

35. McHugh, M. L. Interrater reliability: the kappa statistic. *Biochem. medica* **22**, 276–282 (2012).

36. Mair, P. & Wilcox, R. Robust statistical methods in r using the wrs2 package. *Behav. research methods* **52**, 464–488 (2020).

37. Wilcox, R. R. & Tian, T. S. Measuring effect size: a robust heteroscedastic approach for two or more groups. *J. Appl. Stat.* **38**, 1359–1368 (2011).

38. Jensen-Doss, A., Youngstrom, E. A., Youngstrom, J. K., Feeny, N. C. & Findling, R. L. Predictors and moderators of agreement between clinical and research diagnoses for children and adolescents. *J. Consult. Clin. Psychol.* **82**, 1151 (2014).

39. Jensen-Doss, A. & Weisz, J. R. Diagnostic agreement predicts treatment process and outcomes in youth mental health clinics. *J. consulting clinical psychology* **76**, 711 (2008).

40. Harris, E. Large language models answer medical questions accurately, but can't match clinicians' knowledge. *JAMA* (2023).

41. Cesur, T. & Güneş, Y. C. Optimizing diagnostic performance of chatgpt: The impact of prompt engineering on thoracic radiology cases. *Cureus* **16** (2024).

42. Lingiardi, V. & McWilliams, N. The psychodynamic diagnostic manual–2nd edition (pdm-2). *World Psychiatry* **14**, 237 (2015).

43. Das, S. & Ghoshal, A. Can artificial intelligence ever develop the human touch and replace a psychiatrist?-a letter to the editor of the journal of medical systems: Regarding "artificial intelligence in medicine & chatgpt: De-tether the physician". *J. Med. Syst.* **47**, 72 (2023).

Author contributions statement

Conceptualization SF, AC, GE; methodology SF, AC, GE; formal analysis SF, AC; investigation SF, MDM; resources: GM, DL, YSK, LP, OC, GE; data curation SF, MDM; writing — original draft preparation SF, AC, MDM; review and editing SF, AC, MDM, GM, DL, YSK, LP, OC, GE; supervision GE. All authors reviewed the manuscript.

Additional information

Competing interests The authors declare no competing interests; **Data availability statement** The data used for the statistical analyses are available at the following link: https://gitlab.com/alessandro.carollo/llm_psychiatry/-/tree/97ab9eb1094acf4a89afce8757b24675f68478bc/; **Funding** The authors received no funding for this work.

Acknowledgments

A preprint of the study is published online on *NewAddictionsX* at the following page: https://osf.io/preprints/newaddictionsx/fd8w5

Supplementary Materials

The following Supplementary Materials of this manuscript include: Table S1 Diagnoses distribution of the 46 cases; Table S2 Prompts provided to ChatGPT 3.5 to convert English cases from vignette to outline format, and vice versa. An Italian translation of these prompts was used to convert the Italian cases.