Human Behavior and Emerging Technologies

WILEY

RESEARCH ARTICLE OPEN ACCESS

Diagnostic Accuracy of GPT-Based Large Language Models Across Versions, Prompting Techniques, and Case Presentation Formats

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] | 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, UK | [3]Department of Neurosciences, Imaging and Clinical Sciences, Universita 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

Correspondence: Gianluca Esposito (gianluca.esposito@unitn.it)

Received: 18 December 2024 | **Revised:** 9 February 2026 | **Accepted:** 27 March 2026

Academic Editor: Nikhat Kaura

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 gpt-based models), 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 gpt-based LLMs (gpt-3.5 and gpt-5.1) 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 proximity 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. Importantly, accuracy in this context reflects alignment with a reference label under constrained vignette-based inputs, rather than equivalence with comprehensive clinical diagnostic practice. The results showed a strong agreement between the ratings of the two clinical psychologists ($kappa = 0.798$), with moderate agreement for gpt-3.5's diagnoses and almost perfect for gpt-5.1's diagnoses. Overall, gpt-5.1 showed higher diagnostic accuracy and proximity to human diagnostic evaluations than gpt-3.5 ($p < 0.001$). For gpt-3.5, a small but statistically significant main effect of prompting technique on diagnostic accuracy emerged ($p = 0.009$). The highest proximity to the reference diagnosis 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 ($p < 0.001$) or when it was asked to act as a clinical psychologist ($p = 0.001$). Conversely, gpt-5.1 showed high performance independent of the prompting technique and case format. Under these experimental conditions, the results of the current work provide preliminary evidence supporting the potential use of LLMs as tools to assist the diagnostic process in psychiatry and provide general indication for slightly optimizing their performance. 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.

Seraphina Fong and Alessandro Carollo contributed equally to this work.

1 | 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 (i.e., the extent to which different clinicians assign the same diagnosis to the same patient) and accuracy (i.e., the correspondence between a clinician's diagnosis and a reference or gold-standard diagnosis) 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 have been explored for tasks such as generating diagnostic suggestions or supporting clinical reasoning in structured settings (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 toward various applications (e.g., [21–31]). These studies differ substantially in their primary objectives, ranging from the evaluation of diagnostic accuracy to treatment recommendation, clinical management support, and summarization of medical records. Importantly, only a subset of this literature explicitly evaluates diagnostic accuracy, whereas many studies focus on downstream tasks such as therapeutic recommendations or general clinical assistance. 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 [19] suggested 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. [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. [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. At the same time, the application of LLMs in psychiatry raises important concerns. Prior work has highlighted risks related to hallucinations and bias propagation in generated outputs [35, 36]. LLMs may reflect biases present in their training data, including cultural, linguistic, and diagnostic biases, which can be especially problematic in psychiatric contexts characterized by subjective interpretation and contextual sensitivity. Moreover, LLMs lack grounded clinical understanding and may generate plausible-sounding but clinically inappropriate or inconsistent responses, underscoring the need for careful evaluation and human oversight when considering their use in diagnostic settings. Beyond downstream clinical risks, these phenomena are also methodologically relevant to the present study, as systematic biases or hallucinated clinical features can directly affect diagnostic category selection and therefore influence agreement-based diagnostic accuracy metrics.

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. Unlike traditional standardized instruments, LLMs do not operate independently of how information is presented to them. Their outputs are inherently sensitive to prompt design, including both the structure of the clinical case and the instructions used to elicit a diagnosis. For instance, Pagano et al. [27] provided gpt-4 with comprehensive

anonymized patient information, which included descriptions of symptoms, results of physical examinations, and radiographic interpretations. Pagano et al. [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. [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.

Alongside the growing interest in applying LLMs to psychiatric diagnosis, there is currently limited systematic investigation into how variations in prompting strategies affect diagnostic accuracy and clinical applicability. As a result, it remains unclear which prompt designs are most suitable for supporting psychiatric diagnostic reasoning. From the reviewed literature, 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 diagnostic accuracy. 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 evaluates how LLMs (i.e., gpt-3.5 gpt-5.1) perform on structured psychiatric diagnostic tasks relevant to clinical contexts. In this work, we distinguish between epistemic uncertainty inherent to psychiatric diagnosis, which reflects the probabilistic and evolving nature of clinical knowledge, and measurement uncertainty, which reflects methodological constraints rather than properties of clinical diagnosis itself. Importantly, the present study evaluates the alignment of LLM-generated diagnostic outputs with clinician-referenced diagnoses under controlled vignette-based conditions, rather than full clinical diagnostic reasoning in real-world practice. These design choices introduce measurement-related constraints that should not be conflated with the inherent epistemic uncertainty of psychiatric diagnosis. Specifically, we will use clinical cases in English and Italian and ask gpt-3.5 and gpt-5.1 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.

2 | Results

2.1 | Inter-Rater Reliability Among Clinical Psychologists

Two independent raters evaluated the psychiatric diagnostic accuracy of gpt-3.5 and gpt-5.1 against a reference diagnosis provided alongside the case. In this study, diagnostic accuracy refers to agreement with the reference diagnosis provided for each vignette under controlled input conditions. The inter-rater reliability between the two clinical psychologist raters was assessed using Cohen's kappa. Across the two gpt models, the analysis resulted in a kappa value of 0.798, indicating strong agreement between the two raters. Specifically, Cohen's kappa values of 0.687 and 0.974 were obtained on accuracy judgments for diagnoses generated by gpt-3.5 and gpt-5.1, respectively. Cohen's kappa scores across models, prompts, and case types are reported in Table 1.

TABLE 1 | Cohen's kappa values for inter-rater agreement between the two clinical psychologists across models, prompt types, and case formats.

Model	Level	Cohen's kappa
Gpt-3.5	Prompt Simple	0.658
Gpt-3.5	Prompt Diagnosis Likelihood	0.640
Gpt-3.5	Prompt Acting	0.752
Gpt-3.5	Format Vignette	0.675
Gpt-3.5	Format Outline	0.694
Gpt-5.1	Prompt Simple	0.963
Gpt-5.1	Prompt Diagnosis Likelihood	1.000
Gpt-5.1	Prompt Acting	0.959
Gpt-5.1	Format Vignette	0.975
Gpt-5.1	Format Outline	0.972

2.2 | Statistical Comparison Between Gpt Versions

To compare accuracy scores between gpt versions, we used Yuen's trimmed mean *t*-test for paired samples. The results indicated that diagnoses generated by gpt-5.1 had significantly higher accuracy as compared to the ones generated by gpt-3.5 ($t(497) = 6.68$, $p < 0.001$), with a trimmed mean difference of 0.08 (95% CI[0.06, 0.11]) and an exploratory measure of effect size of 0.19.

2.3 | Statistical Analysis Across Prompts and Case Types for Gpt-3.5

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., 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 ($F(2, 133.68) = 4.92$, $p = 0.009$). Conversely, there was no statistically significant main effect for the case format ($F(1, 153.06) = 2.59$, $p = 0.110$). The interaction between prompt type and case format was also not significant ($F(2, 133.68) = 0.50$, $p = 0.609$). Table 2 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

TABLE 2 | Descriptive statistics of gpt-3.5's diagnostic accuracy across case types and prompt formats.

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

likelihood" prompts ($t(165) = 3.66$, $p < 0.001$), 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 ($t(165) = 3.26$, $p = 0.001$), 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 ($t(165) = 0.30$, $p > 0.050$; 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 slightly more accurate and more aligned to the diagnoses provided by the cases as compared to the "diagnosis likelihood" and the "acting" prompts.

2.4 | Statistical Analysis Across Prompts and Case Types for Gpt-5.1

The same statistical approach was applied to the averaged diagnostic accuracy scores for diagnoses generated by gpt-5.1. The results indicated no statistically significant main effects or interaction between case format and prompt type on average diagnostic accuracy (case format: $F(1, 83.00) = 0.19$, $p > 0.05$; prompt type: $F(1, 73.78) = 0.89$, $p > 0.05$; interaction: $F(1, 73.78) = 0.89$, $p > 0.05$). Table 3 reports the descriptive statistics for gpt-5.1's diagnostic accuracy across case formats and prompt types.

3 | Discussion

Despite the critical importance of diagnosis in psychiatry, studies highlight significant concerns regarding inter-rater reliability, showing that diagnostic agreement among mental health practitioners often ranges from poor to moderate [4, 6–8]. Toward improving the reliability of psychiatric diagnosis, this study explores the potential of using gpt-based LLMs 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 and gpt-5.1'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

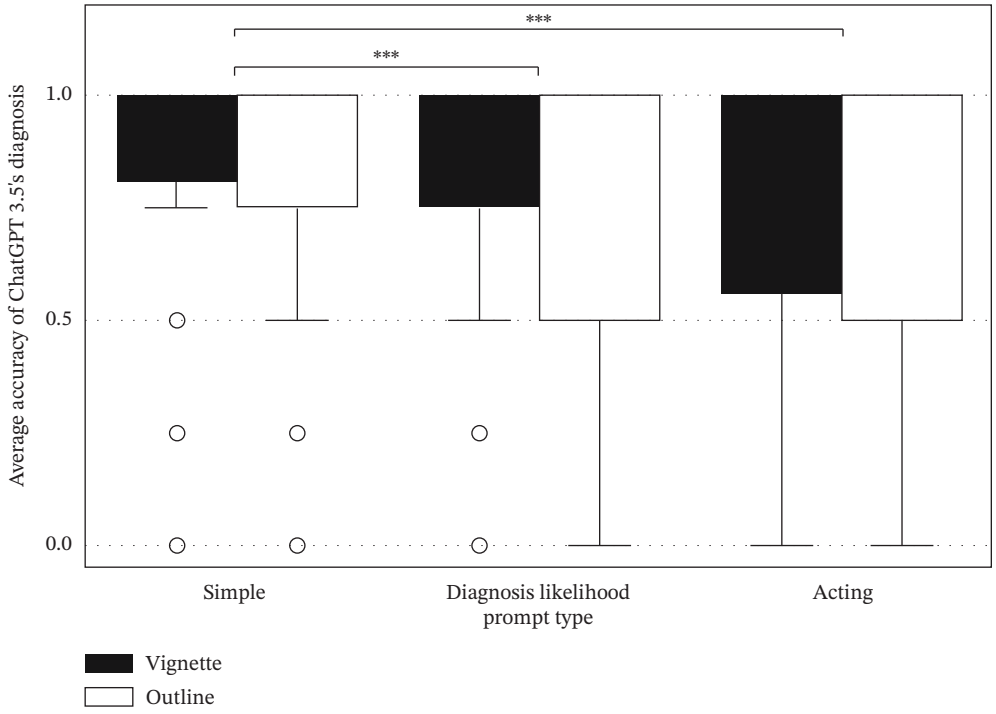

FIGURE 1 | Average accuracy of gpt-3.5's diagnosis across the two case formats (vignette versus outline) and three prompt types ("simple," "diagnosis likelihood," and "acting") (***$p \leq 0.001$).

TABLE 3 | Descriptive statistics of gpt-5.1's diagnostic accuracy across case types and prompt formats.

Case type	Prompt format	Min	1st quartile	Median	Mean	3rd quartile	Max	Standard deviation
Vignette	Simple	0.00	1.00	1.00	0.92	1.00	1.00	0.22
Outline	Simple	0.00	1.00	1.00	0.86	1.00	1.00	0.29
Vignette	Diagnosis likelihood	0.00	1.00	1.00	0.93	1.00	1.00	0.21
Outline	Diagnosis likelihood	0.00	1.00	1.00	0.88	1.00	1.00	0.26
Vignette	Acting	0.50	1.00	1.00	0.95	1.00	1.00	0.15
Outline	Acting	0.00	1.00	1.00	0.88	1.00	1.00	0.24

and the type of instructional prompt given to gpt-3.5 and gpt-5.1 may influence diagnostic accuracy.

In the study, we asked two trained clinical psychologists to evaluate the proximity between the reference diagnosis and gpt-3.5 and gpt-5.1's generated diagnosis. Cohen's kappa was used to assess the inter-rater reliability between the two raters. We observed overall strong inter-rater agreement, with moderate agreement for evaluations of diagnoses generated by gpt-3.5 and almost perfect agreement for those generated by gpt-5.1. Previous studies highlight low concordance in clinical diagnoses (e.g., [6–8, 37, 38]). 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*. This variability likely reflects the overlapping nature of symptomatology across different psychiatric disorders within the DSM framework, where shared features and dimensional characteristics can blur categorical boundaries and contribute to differences in clinical judgment.

Overall, we find that newer versions of gpt demonstrate higher diagnostic accuracy and proximity to the reference diagnosis as compared to older models, highlighting the potential of using, with caution, gpt-based LLMs as a further support method in clinical practice. However, it is important to note that LLMs should never replace clinical judgment. Despite their potential to provide accurate answers, they do not match the depth of clinicians' training, contextual understanding, and expertise [39]. Rather than functioning as autonomous diagnostic agents, LLMs may be more appropriately conceptualized within a human–AI collaboration framework, in which they serve as decision-support tools. In this role, LLMs could assist clinicians by summarizing clinical information, generating differential diagnoses, or highlighting potential diagnostic considerations, while the final diagnostic decision remains firmly grounded in human expertise, clinical reasoning, and ethical responsibility.

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 small but statistically significant impact on the diagnostic accuracy of gpt-3.5. We observe that the raters evaluated the diagnoses as slightly more accurate and proximal to the reference diagnosis when generated with the "simple" prompt compared to the other prompts. This result is in contrast to findings by Cesur and Günes [40], 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. One possible explanation for this discrepancy lies in domain-specific differences in the nature of diagnostic reasoning. Radiological diagnosis often involves mapping relatively well-defined visual patterns to specific disease entities, where structured and explicit prompts may help constrain the model's reasoning toward salient features. In contrast, psychiatric diagnosis relies heavily on overlapping and context-dependent symptom descriptions, where additional prompt constraints or role-based instructions may inadvertently limit the model's ability to flexibly integrate ambiguous or multidimensional clinical information. In this context, one possible interpretation is that simpler prompts may allow greater latitude for pattern integration, resulting in higher perceived diagnostic accuracy. However, this remains a hypothesis that requires further empirical validation across different psychiatric datasets and experimental designs. Importantly, this slight advantage of simpler prompts does not necessarily imply enhanced reasoning capabilities. Rather, reduced prompt complexity may help minimize instruction overload or distraction from irrelevant tokens, allowing the model to focus more directly on diagnostically salient information without being constrained by additional procedural or role-based instructions.

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 and proximity to the reference diagnosis 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.

When analyzing gpt-5.1 performance, neither prompt type nor case format showed a statistically significant effect on diagnostic accuracy and proximity to the reference diagnosis. This finding suggests that newer generations of gpt-based models may be less sensitive to variations in prompt formulation and case presentation, maintaining consistently high performance across

TABLE 4 | Structural and linguistic differences across prompt formulations and their theoretical rationale.

Dimension	Simple prompt	Diagnosis likelihood prompt	Acting prompt
Role assignment	None	Explicit human role (clinical psychologist conducting a study)	Explicit role-play (act as a clinical psychologist)
Perspective/framing	Neutral, task-oriented	Conversational, first-person framing	Instructional, imperative role-based framing
Output structure	Single diagnosis with justification	Multiple diagnoses with likelihood percentages	Ranked differential diagnoses with rationale
Degree of instruction specificity	Minimal	High (percentages, prioritization, conciseness, no disclaimers)	Moderate (differential diagnosis and ranking)

different prompting and formatting conditions. Such robustness may reflect improvements in model architecture and training, enabling more effective extraction and integration of clinically relevant information regardless of surface-level input differences. However, given the high overall accuracy levels observed for GPT-5.1, ceiling effects may have reduced sensitivity to detect subtle differences across prompting conditions. Although robust statistical methods based on trimmed means were employed, clustering of scores near the upper bound of the 0-1 scale may limit discrimination between conditions. Therefore, the apparent invariance across prompts may reflect true robustness, metric saturation, or a combination of both.

The current work also carries certain limitations. Firstly, the languages of the cases are imbalanced (i.e., 3 of the 46 cases were in English, while the remaining cases were in Italian). This asymmetry limits the generalizability of the findings across languages. Language-specific linguistic structures (e.g., differences in syntax, semantic ambiguity, idiomatic expressions, and clinical terminology) may influence how LLMs parse and represent clinical information, potentially affecting diagnostic performance. Future studies should therefore systematically examine the impact of case language by employing balanced multilingual datasets. Secondly, 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. Thirdly, 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 [41]. Fourthly, due to the limited number of cases and the highly uneven distribution across diagnostic categories, we were unable to examine whether diagnostic accuracy varied systematically by disorder type (e.g., psychotic versus mood disorders). Future work employing larger and more diagnostically balanced datasets will be required to assess diagnostic-category–specific performance differences. An additional limitation concerns the graded similarity-based scoring procedure used to evaluate diagnostic accuracy. While this approach was intended to better reflect the dimensional and overlapping nature of psychiatric diagnoses, it introduces an element of measurement subjectivity that may affect precision and replicability. Another limitation concerns the use of Gemini to convert cases between vignette and outline formats. Although all converted cases were manually verified to ensure content fidelity, the involvement of a second generative system may have introduced subtle shifts in phrasing or emphasis. Such variations could potentially influence how clinically relevant information is processed by LLMs and should be considered when interpreting the findings. Finally, a limitation concerns the use of clinician-assigned reference diagnoses as fixed ground truth. Although all cases were originally diagnosed by board-certified psychiatrists, psychiatric diagnosis inherently involves epistemic uncertainty and probabilistic reasoning. In this study, reference diagnoses served as operational benchmarks for evaluating model outputs rather than definitive or infallible clinical truths.

4 | Conclusion

The present work investigated the effects of different prompting styles on the diagnostic accuracy and proximity to a reference diagnosis of gpt-3.5 and gpt-5.1 across 46 psychiatric cases. In particular, we examined the impact of two case presentation formats (i.e., vignette and outline) and three prompting strategies (i.e., "simple," "diagnosis likelihood," and "acting") on the diagnostic performance of gpt-based LLMs. The findings show that newer gpt-based models exhibit higher diagnostic accuracy and generated diagnoses that were closer to the reference ones as compared to earlier versions. For gpt-3.5, a slightly higher diagnostic accuracy was observed when the model was prompted to provide and justify a single diagnosis using a simple instruction. In contrast, gpt-5.1 demonstrated consistently high performance across different prompting strategies and case formats, suggesting that more recent LLMs may be less sensitive to variations in prompt formulation and input structure and more effective at extracting clinically relevant information irrespective of surface-level differences. The results provide preliminary guidance for optimizing the use of LLMs as supportive tools during the psychiatric diagnostic process. However, while LLMs have shown promise in clinical psychology, they should be conceptualized within a human–AI collaboration framework: they may assist clinicians by supporting clinical reasoning and decision-making, but should not be regarded as replacements for psychiatrists or clinical judgment [34, 39, 42]. These findings primarily contribute to methodological optimization in controlled settings and should not be construed as validation for autonomous clinical implementation.

5 | Methods

5.1 | 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 gpt-based LLMs. We presented gpt-3.5 and gpt-5.1

[20] with 46 clinical psychiatric cases made available by psychiatrists. The approach in which the cases were presented to gpt-based models 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 the LLM model 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 nondeterministic nature of LLM outputs, the output from gpt-based models was generated three times, and subsequent statistical analyses were based on all outputs rather than a single generation. This procedure operationalizes performance as agreement with reference diagnoses in a structured evaluation task, rather than assessment of interactive clinical reasoning processes. Subsequently, the agreement between the responses generated by gpt-based models 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 LLMs 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. The cases consisted of real clinical cases that were fully deidentified prior to analysis. All cases were originally assessed and diagnosed by board-certified psychiatrists working in university psychiatric hospitals and registered with their respective national professional boards. All the previous studies followed the Declaration of Helsinki with participants who agreed to be part of clinical programs.

5.2 | Data Preparation

We used a total of 46 anonymized psychiatric cases (see Table S1 in the Supporting Information 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.

5.3 | Case Format

One objective of this study was to evaluate how the presentation format of the clinical case influences gpt-based models' 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 Supporting Information 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.

5.4 | Prompt Design

To measure the impact of how gpt-based models are instructed to diagnose the case, three different types of input prompts were used to present gpt-3.5 and gpt-5.1 cases. The language of the input prompt was dependent on the case language (e.g., the input prompt was in English when instructing gpt-based models to diagnose an English case). Each of the three types of input prompts is outlined in the following subsections (see Table 4 for a summary of their structural and linguistic characteristics). 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 approach allowed us to compare the effect of the prompting instructions that are used in the literature on the LLM diagnostic accuracy.

5.4.1 | 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 LLMs to provide and justify a single diagnosis, we therefore refer to it as the "simple" prompt.

5.4.2 | Prompt 2: "Diagnosis Likelihood" Prompt

The second prompt was adapted from the prompt used by Pagano et al. [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, LLMs are 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 LLMs did not include percentages for the likelihood of the diagnoses, the output was regenerated.

5.4.3 | Prompt 3: "Acting" Prompt

The third prompt was adapted from Koga et al. [25]. Since LLMs are 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."

5.5 | Response Evaluation

For the purposes of the present work, we only considered the diagnoses provided by LLMs 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. [27]), and the top diagnosis provided in response to the "acting" prompt.

Each response from the LLMs was then scored 0, 0.5, or 1 against the reference diagnoses by two trained clinical psychologists. Based on a shared rubric, the raters independently evaluated the correctness of the model-generated diagnosis and its proximity to the reference diagnosis assigned by the psychiatrist. Raters were blinded to each other's evaluations and to the experimental conditions under which the diagnoses were generated (i.e., prompting strategy and case format).

A score of 1 indicates that the model'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 the model's generated diagnosis was schizophrenia, 0.5 if the model's generated diagnosis was a brief psychotic disorder, and 0 if the model's generated diagnosis was a 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 a substantial overlap in symptomatology. This approach provides a more ecologically valid assessment of diagnostic accuracy than a strict correct/incorrect dichotomy. This graded scoring scheme was preferred over a rigid category-based approach (e.g., assigning full credit only to identical DSM diagnoses and partial credit to diagnoses within the same diagnostic category) because diagnostic proximity in psychiatry does not reliably align with the hierarchical structure of DSM classifications. Psychiatric diagnoses often exhibit a substantial phenomenological overlap across formal categories, while conditions grouped within the same category may differ markedly in their core symptomatology.

5.6 | 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 [43]'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. Importantly, in our case, Cohen's kappa reflects agreement between raters on the application of the graded similarity-based scoring rubric (0–0.5–1), rather than agreement exclusively on binary diagnostic correctness. Thus, the reliability estimate captures consistency in proximity-based evaluative judgments as operationalized in this study.

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 [44] to conduct the analysis on gpt-based models' diagnostic accuracy. A robust statistical approach was chosen because it ensures higher statistical power in case of violation of the parametric tests' assumptions [44]. This was the case of the data used for the current work as all averaged ratings across prompt and case types were non-normally distributed (gpt 3.5: vignette case prompt "simple": $W = 0.561$, $p < 0.001$; outline case prompt "simple": $W = 0.643$, $p < 0.001$; vignette case prompt "diagnosis likelihood": $W = 0.634$, $p < 0.001$; outline case prompt "diagnosis likelihood": $W = 0.721$, $p < 0.001$; vignette case prompt "acting": $W = 0.662$, $p < 0.001$; outline case prompt "acting": $W = 0.699$, $p < 0.001$; gpt 5.1: vignette case prompt "simple": $W = 0.381$, $p < 0.001$; outline case prompt "simple": $W = 0.545$, $p < 0.001$; vignette case prompt "diagnosis likelihood": $W = 0.347$, $p < 0.001$; outline case prompt "diagnosis likelihood": $W = 0.515$, $p < 0.001$; vignette case prompt "acting": $W = 0.345$, $p < 0.001$; outline case prompt "acting": $W = 0.524$, $p < 0.001$). Yuen's trimmed mean *t*-test for paired samples was initially used to test the difference in diagnostic accuracy between gpt-based models. In the analysis for each gpt model, 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. Specifically, a 20% trimming was applied, as implemented in the *bwtrim* function, to reduce the influence of outliers and departures from normality and homoscedasticity. Repeated model generations were not averaged prior to analysis; instead, each generation-level accuracy score was treated as an independent observation. While this approach allows stochastic variability across generations to be modeled explicitly, it may increase the effective sample size relative to the number of unique cases. This choice, therefore, represents a trade-off between modeling generative variability and maintaining a strictly case-level analytic structure. In case of significant effects in the robust ANOVA, pairwise *post hoc* comparisons were conducted

using 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 (alpha = 0.05/3 tests = 0.017). In the analysis, we used the explanatory measure of effect size introduced by Wilcox and Tian [45] because it guarantees a robust estimation even in case of unequal variance across groups.

Author Contributions

Conceptualization: Seraphina Fong, Alessandro Carollo, and Gianluca Esposito; methodology: Seraphina Fong, Alessandro Carollo, and Gianluca Esposito; formal analysis: Seraphina Fong and Alessandro Carollo; investigation: Seraphina Fong and Martina Dal Maso; resources: Giovanni Martinotti, Debora Luciani, Yasser Saeed Khan, Luca Pellegrini, Ornella Corazza, and Gianluca Esposito; data curation: Seraphina Fong and Martina Dal Maso; writing–original draft preparation: Seraphina Fong, Alessandro Carollo, and Martina Dal Maso; review and editing: Seraphina Fong, Alessandro Carollo, Martina Dal Maso, Giovanni Martinotti, Debora Luciani, Yasser Saeed Khan, Luca Pellegrini, Ornella Corazza, and Gianluca Esposito; supervision: Gianluca Esposito. All authors reviewed the manuscript.

Funding

The authors received no funding for this work. Open access publishing was facilitated by the Universita degli Studi di Trento, as part of the Wiley-CRUI-CARE agreement.

Disclosure

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

Conflicts of Interest

The authors declare no conflicts of interest.

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/. The provided CSV file includes the scores assigned by human raters, organized by the gpt model version, prompting technique, and case presentation formats.

References

1. A. Jensen-Doss and K. M. Hawley, "Understanding Clinicians' Diagnostic Practices: Attitudes Toward the Utility of Diagnosis and Standardized Diagnostic Tools," *Administration and Policy in Mental Health and Mental Health Services Research* 38, no. 6 (2011): 476–485, https://doi.org/10.1007/s10488-011-0334-3.

2. N. McWilliams, *Psychoanalytic Diagnosis: Understanding Personality Structure in the Clinical Process* (Guilford Press, 2011).

3. N. Craddock and L. Mynors-Wallis, "Psychiatric Diagnosis: Impersonal, Imperfect and Important," *British Journal of Psychiatry* 204, no. 2 (2014): 93–95, https://doi.org/10.1192/bjp.bp.113.133090.

4. A. Aboraya, E. Rankin, C. France, A. El-Missiry, and C. John, "The Reliability of Psychiatric Diagnosis Revisited: The Clinician's Guide to Improve the Reliability of Psychiatric Diagnosis," *Psychiatry* 3, no. 1 (2006): 41–50.

5. C. Di Forti, D. Liccione, and C. Scarpazza, "Inter-Rater Reliability of Psychiatric Diagnosis: A Systematic Review and Metanalysis," *European Psychiatry* 68, no. S1 (2025): S191–S192, https://doi.org/10.1192/j.eurpsy.2025.457.

6. M. Marchi, F. M. Magarini, G. Mattei, et al., "Diagnostic Agreement Between Physicians and a Consultation–Liaison Psychiatry Team at a General Hospital: An Exploratory Study Across 20 Years of Referrals," *International Journal of Environmental Research and Public Health* 18, no. 2 (2021): 749, https://doi.org/10.3390/ijerph18020749.

7. P. R. Miller, "Inpatient Diagnostic Assessments: 2. Interrater Reliability and Outcomes of Structured vs. Unstructured Interviews," *Psychiatry Research* 105, no. 3 (2001): 265–271, https://doi.org/10.1016/s0165-1781(01)00318-3.

8. D. C. Rettew, A. D. Lynch, T. M. Achenbach, L. Dumenci, and M. Y. Ivanova, "Meta-Analyses of Agreement Between Diagnoses Made from Clinical Evaluations and Standardized Diagnostic Interviews," *International Journal of Methods in Psychiatric Research* 18, no. 3 (2009): 169–184, https://doi.org/10.1002/mpr.289.

9. C. Ward, A. Beck, M. Mendelson, J. Mock, and J. Erbaugh, "The Psychiatric Nomenclature: Reasons for Diagnostic Disagreement," *Archives of General Psychiatry* 7 (1962): 198–205, https://doi.org/10.1001/archpsyc.1962.01720030044006.

10. American Psychiatric Association, *Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition, Text Revision* (American Psychiatric Association, 2022).

11. D. A. Regier, W. E. Narrow, E. A. Kuhl, and D. J. Kupfer, "The Conceptual Development of dsm-v," *American Journal of Psychiatry* 166, no. 6 (2009): 645–650, https://doi.org/10.1176/appi.ajp.2009.09020279.

12. T. M. Brückl, V. I. Spoormaker, P. G. Sämann, 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 (2020): 1–25, https://doi.org/10.1186/s12888-020-02541-z.

13. J. M. Gatt, K. L. Burton, L. M. Williams, and P. R. Schofield, "Specific and Common Genes Implicated Across Major Mental Disorders: A Review of Meta-Analysis Studies," *Journal of Psychiatric Research* 60 (2015): 1–13, https://doi.org/10.1016/j.jpsychires.2014.09.014.

14. L. Mao, X. Hong, and M. Hu, "Identifying Neuroimaging Biomarkers in Major Depressive Disorder Using Machine Learning Algorithms and Functional Near-Infrared Spectroscopy (Fnirs) During Verbal Fluency Task," *Journal of Affective Disorders* (2024).

15. M. B. First, "Structured Clinical Interview for the dsm (Scid)," *The Encyclopedia Clinical Psychology* (2014): 1–6.

16. American Psychiatric Association, *The American Psychiatric Association Practice Guidelines for the Psychiatric Evaluation of Adults* (American Psychiatric Pub, 2015).

17. S. Feuerriegel, J. Hartmann, C. Janiesch, and P. Zschech, "Generative Ai," *Business & Information Systems Engineering* 66 (2024): 111–126.

18. H. Kalanderian and H. A. Nasrallah, "Artificial Intelligence in Psychiatry," *Current Psychiatry Reports* 18 (2019): 33–38.

19. T. H. McCoy and R. H. Perlis, "Characterizing Research Domain Criteria Symptoms Among Psychiatric Inpatients Using Large Language Models," *Journal of Mood & Anxiety Disorders* 8 (2024): 100079, https://doi.org/10.1016/j.xjmad.2024.100079.

20. OpenAI, "Chatgpt," (2023), https://openai.com/chatgpthttps://openai.com/chatgpt.

21. L. Caruccio, S. Cirillo, G. Polese, G. Solimando, S. Sundaramurthy, and G. Tortora, "Can Chatgpt Provide Intelligent Diagnoses? A Comparative Study Between Predictive Models and Chatgpt to Define a New Medical Diagnostic Bot," *Expert Systems With Applications* 235 (2024): 121186, https://doi.org/10.1016/j.eswa.2023.121186.

22. J. Chen, L. Liu, S. Ruan, M. Li, and C. Yin, "Are Different Versions of Chatgpt's Ability Comparable to the Clinical Diagnosis Presented in Case Reports? A Descriptive Study," *Journal of Multidisciplinary Healthcare* 16 (2023): 3825–3831, https://doi.org/10.2147/jmdh.s441790.

23. G. Gebrael, K. K. Sahu, B. Chigarira, 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, no. 14 (2023): 3717, https://doi.org/10.3390/cancers15143717.

24. H. Huang, O. Zheng, D. Wang, et al., "Chatgpt for Shaping the Future of Dentistry: The Potential of Multi-Modal Large Language Model," *International Journal of Oral Science* 15, no. 1 (2023): 29, https://doi.org/10.1038/s41368-023-00239-y.

25. S. Koga, N. B. Martin, and D. W. Dickson, "Evaluating the Performance of Large Language Models: Chatgpt and Google Bard in Generating Differential Diagnoses in Clinicopathological Conferences of Neurodegenerative Disorders," *Brain Pathology* 34, no. 3 (2024): e13207, https://doi.org/10.1111/bpa.13207.

26. Y. Mykhalko, P. Kish, Y. Rubtsova, O. Kutsyn, and V. Koval, "From Text to Diagnose: Chatgpt's Efficacy in Medical Decision-Making," *Wiadomości Lekarskie Medical Advances* (2023).

27. S. Pagano, S. Holzapfel, T. Kappenschneider, et al., "Arthrosis Diagnosis and Treatment Recommendations in Clinical Practice: An Exploratory Investigation With the Generative Ai Model gpt-4," *Journal of Orthopaedics and Traumatology* 24, no. 1 (2023): 61, https://doi.org/10.1186/s10195-023-00740-4.

28. D. P. Panagoulias, F. A. Palamidas, M. Virvou, and G. A. Tsihrintzis, "Evaluating the Potential of llms and Chatgpt on Medical Diagnosis and Treatment," in *2023 14th International Conference on Information, Intelligence, Systems & Applications (IISA)* (IEEE, 2023), 1–9.

29. K. Raghu, C. S Devishamani, R. Rajalakshmi, and R. Raman, "The Utility of Chatgpt in Diabetic Retinopathy Risk Assessment: A Comparative Study With Clinical Diagnosis," *Clinical Ophthalmology* 17 (2023): 4021–4031, https://doi.org/10.2147/opth.s435052.

30. G. Sarma, H. Kashyap, and P. P. Medhi, "Chatgpt in Head and Neck Oncology-Opportunities and Challenges," *Indian Journal of Otolaryngology and Head & Neck Surgery* 76, no. 1 (2024): 1425–1429, https://doi.org/10.1007/s12070-023-04201-6.

31. C. Zhang, J. Xu, R. Tang, et al., "Novel Research and Future Prospects of Artificial Intelligence in Cancer Diagnosis and Treatment," *Journal of Hematology & Oncology* 16, no. 1 (2023): 114, https://doi.org/10.1186/s13045-023-01514-5.

32. J. Weizenbaum, "Eliza: A Computer Program for the Study of Natural Language Communication Between Man and Machine," *Communications of the ACM* 9, no. 1 (1966): 36–45, https://doi.org/10.1145/365153.365168.

33. G. Hwang, D. Y. Lee, S. Seol, et al., "Assessing the Potential of Chatgpt for Psychodynamic Formulations in Psychiatry: An Exploratory Study," *Psychiatry Research* 331 (2024): 115655, https://doi.org/10.1016/j.psychres.2023.115655.

34. R. F. D'Souza, S. Amanullah, M. Mathew, and K. M. Surapaneni, "Appraising the Performance of Chatgpt in Psychiatry Using 100 Clinical Case Vignettes," *Asian Journal of Psychiatry* 89 (2023): 103770.

35. A. Bouguettaya, E. M. Stuart, and E. Aboujaoude, "Racial Bias in ai-Mediated Psychiatric Diagnosis and Treatment: A Qualitative Comparison of Four Large Language Models," *npj Digital Medicine* 8, no. 1 (2025): 332, https://doi.org/10.1038/s41746-025-01746-4.

36. R. Schnepper, N. Roemmel, R. Schaefert, L. Lambrecht-Walzinger, and G. Meinlschmidt, "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 Mental Health* 12 (2025): e57986, https://doi.org/10.2196/57986.

37. A. Jensen-Doss, E. A. Youngstrom, J. K. Youngstrom, N. C. Feeny, and R. L. Findling, "Predictors and Moderators of Agreement Between Clinical and Research Diagnoses for Children and Adolescents," *Journal of Consulting and Clinical Psychology* 82, no. 6 (2014): 1151–1162, https://doi.org/10.1037/a0036657.

38. A. Jensen-Doss and J. R. Weisz, "Diagnostic Agreement Predicts Treatment Process and Outcomes in Youth Mental Health Clinics," *Journal of Consulting and Clinical Psychology* 76, no. 5 (2008): 711–722, https://doi.org/10.1037/0022-006x.76.5.711.

39. E. Harris, "Large Language Models Answer Medical Questions Accurately, But Can'T Match Clinicians' Knowledge," *JAMA* (2023).

40. T. Cesur and Y. C. Günes, "Optimizing Diagnostic Performance of Chatgpt: The Impact of Prompt Engineering on Thoracic Radiology Cases," *Cureus* 16 (2024): https://doi.org/10.7759/cureus.60009.

41. V. Lingiardi and N. McWilliams, "The Psychodynamic Diagnostic Manual–2nd Edition (pdm-2)," *World Psychiatry* 14, no. 2 (2015): 237–239, https://doi.org/10.1002/wps.20233.

42. S. Das and A. Ghoshal, "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," *Journal of Medical Systems* 47, no. 1 (2023): 72, https://doi.org/10.1007/s10916-023-01974-9.

43. M. L. McHugh, "Interrater Reliability: The Kappa Statistic," *Biochemia Medica* 22 (2012): 276–282, https://doi.org/10.11613/bm.2012.031.

44. P. Mair and R. Wilcox, "Robust Statistical Methods in R Using the wrs2 Package," *Behavior Research Methods* 52, no. 2 (2020): 464–488, https://doi.org/10.3758/s13428-019-01246-w.

45. R. R. Wilcox and T. S. Tian, "Measuring Effect Size: A Robust Heteroscedastic Approach for Two or More Groups," *Journal of Applied Statistics* 38, no. 7 (2011): 1359–1368, https://doi.org/10.1080/02664763.2010.498507.

Supporting Information

Additional supporting information can be found online in the Supporting Information section. *(Supporting Information)*

The following Supporting Information of this manuscript includes the following: Table S1: Diagnoses distribution of 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.