

RESEARCH PROTOCOL

Research protocol: Evaluating Chain-of-Thought Reasoning in LLMs for Complex Clinical Psychiatric Cases

Vittorio De Vita^{1,2}, Bianca Destro Castaniti^{2,3}, Antonio Cristiano^{1,2}, Nicolò Lentini¹,
Mariapia Vassalli^{2,4}, Lorenzo De Mori^{2,5}, Doriana Lacalaprice^{1,2}, Chiara Battipaglia^{2,6},
Giulia Vojvodic^{1,2}, Pietro Eric Risuleo^{1,2}, Alessandra Piscitelli^{2,7}, Tommaso Dionisi^{2,8},
Francesco Andrea Causio^{1,2}

¹ Department of Life Sciences and Public Health, Section of Hygiene, Università Cattolica del Sacro Cuore, Rome, Italy.

² SIAM (Italian Society of Artificial Intelligence in Medicine)

³ Department of Diagnostic Imaging, Oncological Radiotherapy and Hematology, Università Cattolica del Sacro Cuore, Rome, Italy.

⁴ Department of Nutritional Sciences and Human Nutrition, Università La Sapienza, Rome, Italy.

⁵ Department of mental health, Addiction treatment unit, ASL RM 4, Bracciano, Italy.

⁶ Istituto Paritario di Istruzione Secondaria Superiore "G. Leopardi", Battipaglia, Italy.

⁷ Department of Mental and Physical Health and Preventive Medicine, section of Psychiatry, Azienda Ospedaliera Universitaria Luigi Vanvitelli, Naples, Italy.

⁸ Internal Medicine and Alcohol Related Disease Unit, Columbus-Gemelli Hospital, Fondazione Policlinico Universitario A. Gemelli IRCCS, Rome, Italy.

Received August 17, 2026 · Accepted August 31, 2026 · Published August 31, 2026



Keywords: Artificial Intelligence, General Practitioners, Mental Health, Clinical Reasoning, Psychiatry.

ABSTRACT

Background

Psychiatric medicine presents unique diagnostic and therapeutic challenges, often involving multimorbidity, polypharmacy, and atypical presentations requiring complex reasoning. Artificial Intelligence, particularly Large Language Models (LLMs), is emerging as a support tool in these settings. However, the clinical validity, interpretability, and reliability of LLMs in psychiatry remain largely unexplored, particularly their ability to generate transparent, guideline-consistent reasoning.

Aims

This study evaluates the clinical reasoning capabilities of LLMs in complex psychiatric scenarios. The primary aim is to assess the validity of their long chain-of-thought (CoT) reasoning. A secondary aim is to determine whether LLM assistance improves clinicians' diagnostic accuracy.

Method

Three LLMs, Gemini 2.5, Grok, and DeepSeek R1, will be assessed using ten complex psychiatric cases sourced from a non-public clinical manual and rated with the Amsterdam

Clinical Challenge Scale. Each model's CoT response will be evaluated by blinded panels of psychiatrists, residents, and general practitioners using standardized metrics for factual accuracy, coherence, and medical plausibility. In a second phase, clinicians will answer thirty diagnostic questions with and without support from the best-performing LLM. The study uses step-by-step reasoning prompts and few-shot examples to elicit detailed responses and includes bias mitigation strategies such as randomization, blinding, and statistical controls.

Results

Analyses will assess inter-rater reliability, metric redundancy, and reasoning quality. Closed-source models are expected to outperform open-source ones. LLM assistance is anticipated to improve diagnostic accuracy, especially among non-specialists.

Conclusions

This study provides a framework for evaluating LLMs in psychiatry and supporting their safe, evidence-based integration into mental health practice.

INTRODUCTION

Background

Psychiatric medicine presents unique and multifaceted challenges due to multimorbidity, polypharmacy, and atypical disease presentations, all of which require highly nuanced and context-sensitive clinical reasoning (1,2). Artificial Intelligence (AI) demonstrates considerable potential in addressing these challenges, particularly within complex, uncertain, and high-stakes domains such as psychiatry (3,4). However, it is crucial that the selected AI solutions are not only accurate and technically feasible but also explainable, transparent, and interpretable by physicians and other healthcare professionals who depend on these tools in real-world clinical practice (5-9). Large Language Models (LLMs), such as Google's Gemini and xAI's Grok, are beginning to show promising capabilities in explicit reasoning, which may significantly enhance their applicability in demanding medical contexts. In such settings, model interpretability and the capacity of users to critically evaluate and contextualize the outputs are particularly essential for safe and effective integration into clinical workflows (10).

Rationale and Study Aim

Despite these encouraging developments, the reliability of LLMs consistently adhering to clinical guidelines, ensuring reproducible outputs, and adapting to novel or atypical psychiatric scenarios remains largely unexplored (11,12). While LLMs have demonstrated potential across broader clinical applications (13-15), their performance in psychiatry-specific contexts has yet to be thoroughly validated. Current limitations include hallucination phenomena in medical outputs, inconsistent guideline compliance, and restricted validation in non-English medical contexts (16). The present study aims to systematically address these limitations by providing a structured and evidence-informed framework for physicians, healthcare professionals, decision-makers, and AI developers, with the goal of guiding the safe, reliable, and context-appropriate implementation and fine-tuning of AI agents in psychiatric care.

The **primary objective** of this study is to rigorously assess the clinical validity of LLMs' extended chain-of-thought reasoning when applied to complex psychiatric cases, evaluated through blinded specialist ratings using a set of standardized metrics. The **secondary objective** is to compare the medical accuracy of responses to a 30-item dataset, derived from the Italian medical licensing examination, across specialist clinicians, psychiatry residents, and general practitioners, both with and without AI assistance, employing the LLM that demonstrates the highest performance in the primary

objective. This two-tiered approach is designed to provide a comprehensive evaluation of LLMs' diagnostic reasoning capabilities and their potential impact on clinical decision-making in real-world psychiatric practice.

Trial design

The study is a non-interventional cross-sectional experiment that evaluates the performance of LLMs on pre-existing clinical cases (from DSM-5-TR), without randomization, assignment of interventions, or enrollment of human participants in a clinical setting. It does not alter medical practices or test health outcomes on patients; it is an in vitro evaluation of AI tools with blinded human evaluators (15-30 professionals) on predefined metrics.

Eligibility criteria per evaluator

Inclusion: Practicing psychiatrists (≥ 5 years post-residency experience in psychiatry); psychiatry residents (enrolled in accredited Italian psychiatry residency program); general practitioners (≥ 3 years independent practice). All must be fluent in Italian, have access to secure online platform for blinded evaluation, and provide informed consent.

Exclusion: Prior involvement in case selection/rating; conflicts of interest (e.g., employment by LLM developers); inability to complete evaluations within 2 weeks.

Informed consent will be obtained electronically by study coordinator (VDV).

Sample size justification

For the primary objective, the study will enrol a total of 15 clinical evaluators, comprising 5 psychiatrists, 5 psychiatry residents, and 5 general practitioners. Each evaluator will rate 10 clinical cases for all three LLMs across six predefined clinical reasoning metrics, yielding 180 individual ratings per metric (10 cases $\times$ 3 LLMs $\times$ 6 metrics). This sample size is sufficient to estimate inter-rater reliability with an expected intraclass correlation coefficient (ICC) greater than 0.7, and provides 80% power to detect a difference of 0.2 points on the rating scale, assuming a standard deviation of 0.8, a two-sided alpha of 0.05, and parameters derived using G*Power.

For the secondary objective, the total sample will include 45 clinicians, namely the 15 evaluators from the first phase plus 30 additional participants, resulting in 15 individuals per professional group (psychiatrists, residents, and general practitioners). Each participant will answer 30 multiple-choice questions, producing 1,350 responses overall (45 $\times$ 30 items). This sample size is powered to detect an absolute improvement in diagnostic accuracy of approximately 10% between AI-assisted and non-assisted conditions, corresponding to an odds ratio of about 1.5, with 90% power and a two-sided alpha of 0.05 in mixed-effects logistic regression models. No formal interim analyses are planned. The calculations assume a dropout rate of around 10%; additional participants will be recruited if needed to compensate for any losses to follow-up and preserve the planned statistical power.

Recruitment strategies

Evaluators recruited via SIIAM network, Università Cattolica alumni, Italian psychiatry associations. Target: 20% response rate; reminders after 1 week. Phase 2 adds new recruits post-primary results via same methods.

MATERIALS AND METHODS

Primary Objective

To assess the clinical validity of LLM reasoning, we will perform a cross-sectional, experimental evaluation of model performance.

Endpoints

I: Mean score assigned by clinical evaluators for each Chain-of-Thought (CoT) response across six predefined metrics. II: Inter-rater reliability measured by intraclass correlation coefficients (ICC 2,k and ICC 2,1) for consistency across raters and intra-rater scoring patterns. III: Metric redundancy and sensitivity assessed through Spearman correlations, Kaiser-Meyer-Olkin (KMO) statistics, principal component analysis (PCA), and exploratory factor analysis (EFA). IV: Performance rankings of LLMs by mean scores across all six metrics, with comparisons among models using linear mixed-effects models. V: Group differences in scoring patterns between evaluator types (psychiatrists, residents, general practitioners) using Tukey's post-hoc tests. VI: Effect sizes (Cohen's d) for relevant comparisons to convey practical significance.

Dataset Preparation

To prepare the dataset, 20 complex psychiatric clinical cases will be retrieved by LDM from the book "DSM-5-TR® Clinical Cases" by Barnhill (2024) (17), a reputable medical textbook that is not publicly available online. This reduces the likelihood of overlapping with LLM training data. To ensure that the clinical cases are appropriately complex, we will apply **the Amsterdam Clinical Challenge Scale (ACCS)** (18) to the retrieved cases. Five independent psychiatrists (not involved in the subsequent evaluation phases of the primary objective) will assess the complexity of each case using the ACCS, scoring them on a scale from 1 (least complex) to 5 (most complex). We will select the 10 cases with the highest average complexity score for inclusion in the dataset (Figure 1).

Long CoT generation

The following LLMs will be assessed for the clinical validity of their reasoning skills: two Closed-source models (Gemini 2.5 by Google and Grok by xAI) and one Open-source model (DeepSeek R1). These models were selected because, as of June 2025, they are among the few that provide users with direct access to their internal chain-of-thought reasoning. This feature is particularly relevant, as LLMs are known to exhibit sycophantic behavior and may suppress or distort their actual reasoning when interacting with users. This concern has been well-documented in recent studies by Turpin M et al. and by Chen Y et al. (19,20). Such issues are of high relevance in the healthcare domain, where transparency and fidelity of reasoning are essential for clinical decision support.

To ensure independence between evaluations and prevent contextual contamination, each clinical case will be submitted to each LLM in a fresh conversation session (i.e., with no prior conversation history).

The selected clinical cases will be translated into English - using DeepL (22) and subsequently validated by CB, a researcher with a master's degree in modern Languages and Literatures (English and French) - and then administered to the LLMs. We will use explicit step-by-step reasoning prompts that instruct the model to generate intermediate reasoning steps before providing a final answer, for example, prompts that say: "Explain your reasoning step-by-step before answering" (see Supplementary Material). This approach encourages the model to produce a transparent and logically structured chain of thoughts, which has been shown to improve accuracy on complex tasks (5,13,19,21).

The prompt will also include few-shot examples (three in total) that demonstrate detailed reasoning to further enhance performance by guiding the model toward the expected reasoning style (5,13). Examples will be drawn from the same book used to select the clinical cases (see supplementary material).

The CoT responses will be translated into Italian using DeepL (22) and subsequently validated by CB, a researcher with a master's degree in modern Languages and Literatures (English and French).

Evaluation

We will create a dataset consisting of the 10 selected standardized clinical questions paired with the corresponding long CoT responses from each LLM and present it to three groups of evaluators, five practicing psychiatrists, five psychiatry residents, and five general practitioners. The dataset will be presented to the physicians in Italian - translated using DeepL (22) and subsequently validated by CB. Physicians will be asked to evaluate each CoT response using a 6-metric scale ranging from 0 to 5 (where 0 indicates the poorest performance and 5 the best) for each metric. This evaluation will be conducted for every CoT response (one per question), across all 10 questions, for each of the three LLMs, resulting in a total of 180 evaluation items. To minimize bias, the order of LLMs presentation will be randomized and all evaluations conducted by physicians will be performed in a blinded manner. This means that the evaluators will not be aware of which LLM generated the responses they are assessing. By removing this knowledge, we aim to prevent any preconceived notions or biases related to specific models from influencing the results (6,11,23).

The following six metrics will be used to assess the quality of each Chain-of-Thought (CoT) response: **(1) Factual Correctness:** Are the medical facts and concepts mentioned in the CoT accurate and up to date? **(2) Consideration of Relevant Factors:** Does the CoT consider all the pertinent information from the clinical question (e.g., patient history, symptoms, lab results)? **(3) Logical Flow of Reasoning:** Does the CoT follow a logical and coherent sequence of steps that a medical professional would typically take? **(4) Sufficiency of Steps:** Are there enough intermediate steps in the CoT to justify the final answer, or are there significant logical gaps? Does the CoT provide an appropriate level of detail for the clinical context and the expertise of the intended user (e.g., a specialist might expect more nuanced reasoning than a general practitioner for certain questions)? **(5) Relevance to the Clinical Question:** Are all the steps in the CoT directly relevant to answering the specific clinical question asked? **(6) Medical Plausibility:** Are the individual steps and the overall reasoning medically plausible and sound? Would a medical professional consider the reasoning steps to be appropriate?

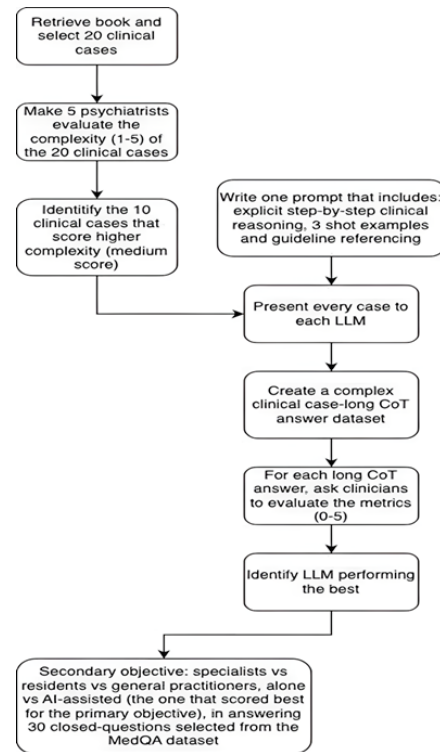


Figure 1- Study workflow for evaluating LLM performance in complex psychiatric cases. The process includes selecting high-complexity clinical cases, creating prompts with step-by-step reasoning and guideline references, generating long chain-of-thought (CoT) responses from each LLM, clinician-based evaluation of CoT metrics, identifying the best-performing LLM, and comparing the accuracy of AI-assisted versus non-assisted responses across specialists, residents, and general practitioners using a 30-item closed-question dataset.

Scoring and data analysis

After completion of the evaluations, we will calculate descriptive statistics (mean, standard deviation, quartiles, range) for each evaluation metric to characterize score distributions. Group comparisons will be performed using parametric tests (e.g., t-tests or ANOVA) when normality assumptions are met, or non-parametric tests (Mann-Whitney U or Kruskal-Wallis), given the ordinal nature of the scales. We will assess metric reliability using intraclass correlation coefficients [ICC (2,k)] to evaluate consistency across raters. Each case will be scored by the same set of raters considered a random sample from a broader population of possible raters. ICC(2,k) is appropriate because we are interested in the reliability of the mean ratings across multiple raters for each metric, rather than the reliability of a single rater's score (24,25). In a complementary analysis, we will also compute ICC (2,1) to explore intra-rater consistency, the stability of individual rater scoring patterns across different cases. LLM performance will be ranked by mean scores across all six metrics. Comparisons among LLMs, as well as between closed-source (e.g., Gemini, Grok) and open-source (e.g., DeepSeek-R1), will be performed linear mixed-effects models, both overall and stratified by metrics, accounting for clustering by metric (in overall model), rater and case as random effects. Holm's corrections will control for multiple comparisons, and Tukey's post-hoc tests will identify significant pairwise differences between evaluator groups. Degrees of freedom for the linear mixed-effects models will be estimated using the Kenward-Roger method. Metric sensitivity and redundancy will first be assessed using Spearman correlations and KMO statistics to evaluate metric relationships and suitability for factor analysis. Principal component analysis (PCA) and exploratory factor analysis (EFA) will then be performed to identify dominant or overlapping dimensions, with scree plots and explained variance used to guide factor selection. Cohen's d will be calculated from the linear mixed-effects models for

relevant comparisons, when appropriate, to convey practical significance alongside p-values, facilitating interpretation and cross-study comparison (26). All statistical analysis will be performed using R software (version 4.4.0, version 2024-04-24).

Secondary Objective:

Comparing the medical accuracy of responses among medical practitioners, both with and without AI assistance.

Endpoints

I: Accuracy rates with 95% confidence intervals compared to AI-assisted and non-assisted groups. II: Odds ratios with 95% confidence intervals for correct responses derived from mixed-effects logistic regression models. III: Order effects as a sensitivity analysis including position of answer options as fixed effect.

Dataset Preparation

A dataset of 30 closed, single-best-answer psychiatry questions will be carefully selected from the Italian national exam used to access medical residency (SSM). Each item will consist of a detailed clinical scenario followed by five possible answer options, of which only one is correct. This question format was deliberately chosen to align with standardized medical assessment practices, ensuring both consistency and comparability across participants. Additionally, it allows for a clear and objective evaluation of diagnostic reasoning and decision-making performance in a controlled experimental setting.

Participants and Procedure

For this objective, we will enroll the 15 participants from the previous phase, along with an additional 15 new participants: five psychiatrists, five psychiatry residents, and five general practitioners, resulting in a total of 15 participants per professional group. Each group will be randomly divided into two equal subgroups: one responding without AI assistance and the other responding with AI assistance. The AI-assisted subgroups will utilize the LLM that demonstrated the overall highest performance during the first phase of the study. Participants in the non-assisted subgroup will provide their responses independently, relying solely on their clinical knowledge and reasoning. Those in the AI-assisted subgroup will interact directly with the selected LLM, basing their responses with the help of its recommendations. To minimize potential bias related to the order of answer options, the presentation order will be randomized for each participant. This design ensures a balanced comparison of diagnostic performance with and without AI support while controlling potential confounding factors.

Statistical Analysis

We will calculate and compare accuracy rates, along with 95% confidence intervals (CIs), between the AI-assisted and non-assisted groups using chi-square tests. To account for the nested structure of the data, mixed-effects logistic regression models will be employed, incorporating fixed effects for assistance condition and participant profession, and random effects for participant ID and case ID. As a sensitivity analysis, and to verify the absence of potential order effects, the position of each answer option will be included as a fixed effect within the linear mixed-effects models. Odds ratios and their corresponding 95% CIs for correct responses will be derived from the model coefficients, providing interpretable estimates of the effect of AI assistance across professional groups. All statistical analyses will be conducted using R software (version 4.4.0, released 2024-04-24), ensuring reproducibility and transparency of the results.

EXPECTED RESULTS

Preliminary results of this study were presented at the SIIAM 3rd Annual Meeting (October 2025) in Naples (27).

Primary objective: LLM Chain of Thought Reasoning Validity

We expect that each large language model, Gemini 2.5, Grok, and DeepSeek R1, will demonstrate varying levels of performance across the six clinical reasoning metrics. We also anticipate that closed-source models (Gemini 2.5 and Grok) may achieve higher average scores compared to the open-source model (DeepSeek R1), although this difference will be formally tested for statistical significance (11,13). Among the evaluated dimensions, metrics such as logical flow and factual correctness are expected to show higher inter-rater reliability, while more subjective criteria like sufficiency of steps and medical plausibility may produce lower agreement across evaluators (11,13). Models will be ranked according to the mean of their metric scores, and 95% confidence intervals will be reported to convey the precision of these estimates. The top-performing model is expected to display consistent, high-quality reasoning across most cases, although some variability is likely, particularly in response to more complex clinical scenarios (2). Measures of inter-rater reliability, including intraclass correlation coefficients (ICC), are expected to fall in the moderate to high range, suggesting an acceptable level of consensus among evaluators. Through thematic analysis of evaluator comments and structured error classification using the Amsterdam Clinical Challenge Scale (ACCS), we anticipate identifying common reasoning errors such as hallucinated medications, premature diagnostic closure, and failure to account for relevant comorbidities (12). In terms of metric structure, principal component analysis may reveal high correlations between certain dimensions (e.g., logical flow and medical plausibility), suggesting potential redundancy in how evaluators assess reasoning quality (11). Finally, diagnostic trajectory mapping using Sankey diagrams is expected to show that high-scoring chain-of-thought responses follow more standard and complete diagnostic pathways. In contrast, low-scoring responses may reflect more erratic, disorganized, or incomplete reasoning patterns (3).

Secondary Objective: Accuracy With and Without LLM Assistance

We expect that medical professionals completing the 30-question dataset with the support of the best-performing LLM identified in the primary objective will achieve higher mean accuracy scores compared to those responding without AI assistance (16). The magnitude of this improvement is expected to vary across professional groups. General practitioners and psychiatry residents, who may have lower baseline diagnostic accuracy, are likely to derive greater benefit from LLM support, whereas psychiatrists, with higher baseline expertise, may experience comparatively smaller gains. Statistical comparisons are expected to reveal significant differences between the AI-assisted and non-assisted groups, particularly for questions that are complex, ambiguous, or diagnostic, challenging. These findings will provide valuable insights into the impact of LLM assistance across varying levels of clinical difficulty and cognitive demand. Additionally, we expect to observe higher inter-rater agreement within AI-assisted groups, reflecting the standardizing effect of consistent reasoning support provided by the model (11), and highlighting its potential role in promoting more uniform and reliable clinical decision-making.

Bias Mitigation

We expect that the implemented bias mitigation strategies will enhance the internal consistency, reliability, and overall robustness of the study findings. The randomization procedures, combined with appropriate statistical adjustments, are anticipated to demonstrate minimal order effects related to question sequence or LLM presentation, thereby confirming the absence of systematic bias in response ordering (11,16). These measures will help ensure that observed differences in performance can be attributed primarily to the experimental conditions rather than confounding procedural factors.

DISCUSSION

Clinical implication of LLMs in Psichiatria

The present research rigorously evaluates the clinical reasoning capabilities of LLMs within complex psychiatric contexts, addressing a critical and longstanding gap in AI applications in mental health care. Psychiatric medicine is inherently characterized by diagnostic complexity, stemming from multimorbidity, polypharmacy, and atypical presentations, all of which necessitate nuanced and context-sensitive clinical judgment (1,2). Advanced LLMs, such as Google's Gemini 2.5 and xAI's Grok, offer promising developments in explicit chain-of-thought (CoT) reasoning, potentially enhancing clinical decision support by generating interpretable, systematic, and traceable reasoning pathways (5,6). Our structured evaluation protocol incorporates blinded, randomized assessments of LLM-generated CoTs across multiple clinically meaningful metrics. Prior research highlights the importance of such multidimensional evaluation frameworks to adequately capture the intricacies of medical reasoning beyond conventional accuracy metrics (3,11). Based on current evidence, we anticipate that closed-source LLMs will generally outperform open-source counterparts, consistent with recent findings demonstrating differential performance profiles among leading biomedical LLMs (10,11).

Error analysis and interpretability

The use of thematic coding to systematically identify recurrent reasoning errors - such as hallucinated medications, premature diagnostic closure, or neglect of comorbidities - aligns with previously reported challenges in AI-driven psychiatric assessments and will provide valuable insights to guide targeted model improvements and prompt refinement (19,20). Visualization of diagnostic trajectories through methods like Sankey diagrams further complements recent advances in AI interpretability research, enhancing understanding of how LLM reasoning aligns with or diverges from established clinical heuristics (10). Importantly, a key secondary objective of the study is to evaluate the real-world impact of LLM assistance on diagnostic accuracy among clinicians with varying levels of expertise. Consistent with findings from randomized clinical trials in diagnostic reasoning, we anticipate that LLM support will substantially enhance performance, particularly for less-experienced practitioners, while also promoting greater inter-rater agreement- outcomes emphasized by Goh et al. (16) and corroborated by Omar et al. (4). These results contribute to the growing consensus that AI tools, when integrated thoughtfully with attention to reliability, interpretability, and user experience, can meaningfully augment clinical decision-making (7,10). Methodologically, the protocol incorporates rigorous strategies to mitigate bias, including blinding, randomization, and multilevel statistical controls that account for case complexity (6,23). These design choices align with best practices in medical AI evaluation guidelines, ensuring the validity and reproducibility of findings (11).

LIMITATIONS

Limitations include potential sampling bias due to reliance on psychiatric cases from a single textbook, which we address through validated complexity scoring (17,18). Additional challenges arise from cross-linguistic nuances in the Italian psychiatric context, as LLMs are predominantly trained on English-language data (6,16). Moreover, the rapidly evolving nature of LLMs requires periodic reassessment to maintain clinical relevance (20).

CONCLUSIONS

In summary, this research represents a significant step toward translating the theoretical potential of LLMs into practical applications in psychiatry. By combining rigorous clinical evaluation, detailed error analysis, and real-world performance testing, it supports the responsible integration of AI tools that can enhance clinical reasoning, promote adherence to guidelines, and ultimately contribute to improved psychiatric care outcomes.

DECLARATIONS

Acknowledgements

We would like to thank SIIAM for believing in this research.

Authors' contributions

Conceptualization: VDV and FAC; methodology: VDV, FAC, DL AC and NL; validation: MV, LDM, PER and DL; investigation: BDC, AC, MV, GV, PER and MDM; data curation: TD, LDM, CB and FAC; writing-original draft preparation: GV, BDC, PER AC, NL, MDM; writing--review and editing: BDC, DL, AP and GV; supervision: VDV, TD, AP and CB; project administration: VDV, AP, TD and FAC. All authors have read and agreed to the published version of the manuscript.

Funding

No funding was received.

Data availability

All data will be made available through the final article publication.

Ethical approval and book use authorization

Ethics committee approval was not required for this study, as confirmed by the Activation Trial Team (ATT) at Fondazione Policlinico Universitario "A. Gemelli" IRCCS (FPG).

Use of "DSM-5-TR® Clinical Cases" by Barnhill (2024) was authorized for research purposes with permission granted by the editor through e-mail.

Consent

Signed informed consent of the participants will be obtained electronically by study coordinator (VDV).

Competing interests

None of the authors has any competing interests to disclose.

REFERENCES

1. Thammasitboon S, Sur M, Rencic JJ, Dhaliwal G, Kumar S, Sundaram S, et al. Psychometric validation of the reconstructed version of the assessment of reasoning tool. *Med Teach* [Internet]. 2021 [cited 2025 Jun 24];43(2):168-73. Available from: <https://pubmed.ncbi.nlm.nih.gov/33073665/>
2. Daniel M, Rencic J, Durning SJ, Holmboe E, Santen SA, Lang V, et al. Clinical Reasoning Assessment Methods: A Scoping Review and Practical Guidance. *Academic Medicine* [Internet]. 2019 Jun 1 [cited 2025 Jun 24];94(6):902-12. Available from: <https://pubmed.ncbi.nlm.nih.gov/30720527/>
3. Jin Y, Liu J, Li P, Wang B, Yan Y, Zhang H, et al. The Applications of Large Language Models in Mental Health: Scoping Review. *J Med Internet Res* 2025;27:e69284 <https://www.jmir.org/2025/1/e69284> [Internet]. 2025 May 5 [cited 2025 Jun 24];27(1):e69284. Available from: <https://www.jmir.org/2025/1/e69284>
4. Omar M, Soffer S, Charney AW, Landi I, Nadkarni GN, Klang E. Applications of large language models in psychiatry: a systematic review. *Front Psychiatry* [Internet]. 2024 [cited 2025 Jun 24];15. Available from: <https://pubmed.ncbi.nlm.nih.gov/38979501/>

5. Liu J, Wang Y, Du J, Zhou JT, Liu Z. MedCoT: Medical Chain of Thought via Hierarchical Expert. 2024 Dec 18 [cited 2025 Jun 24]; Available from: <https://arxiv.org/pdf/2412.13736>
6. Kim BH, Wang C. Large Language Models for Interpretable Mental Health Diagnosis. 2025 Jan 13 [cited 2025 Jun 24]; Available from: <https://arxiv.org/pdf/2501.07653>
7. Graham S, Depp C, Lee EE, Nebeker C, Tu X, Kim HC, et al. Artificial Intelligence for Mental Health and Mental Illnesses: An Overview. *Curr Psychiatry Rep* [Internet]. 2019 Nov 1 [cited 2025 Jun 25];21(11):116. Available from: <https://pmc.ncbi.nlm.nih.gov/articles/PMC7274446/>
8. Sacchi FA, Cascini F, Conditì N, Ravizza A, Daverio M, Causio FA, De Vita V, Pivetta A, Maio P, De Angelis L, Baglivo F, Diedenhofen G, Di Pumpo M, Belpiede A, Ferro D, Bolognini L. The path to trustworthy medical AI: the evolving role of explainability. *Recenti Prog Med*. 2025 Oct;116(10):546-550. doi: 10.1701/4573.45774. PMID: 41037358.
9. Causio FA, De Angelis L, Diedenhofen G, Talio A, Baglivo F; Workshop Participants. Perspectives on AI use in medicine: views of the Italian Society of Artificial Intelligence in Medicine. *J Prev Med Hyg*. 2024 Aug 31;65(2):E285-E289. doi: 10.15167/2421-4248/jpmh2024.65.2.3261. PMID: 39430984; PMCID: PMC11487733.
10. Sankar BS, Gilliland D, Rincon J, Hermjakob H, Yan Y, Adam I, et al. Building an Ethical and Trustworthy Biomedical AI Ecosystem for the Translational and Clinical Integration of Foundation Models. *Bioengineering* [Internet]. 2024 Oct 1 [cited 2025 Jun 24];11(10):984. Available from: <https://pmc.ncbi.nlm.nih.gov/articles/PMC11504392/>
11. Hanss K, Sarma K V, Glowinski AL, Krystal A, Saunders R, Halls A, et al. Assessing the Accuracy and Reliability of Large Language Models in Psychiatry Using Standardized Multiple-Choice Questions: Cross-Sectional Study. *J Med Internet Res* [Internet]. 2025 May 20 [cited 2025 Jun 24];27(1):e69910. Available from: <https://www.jmir.org/2025/1/e69910>
12. Hanafi A, Saad M, Zahran N, Hanafy RJ, Fouda ME. A Comprehensive Evaluation of Large Language Models on Mental Illnesses. 2024 Sep 24 [cited 2025 Jun 24]; Available from: <https://arxiv.org/pdf/2409.15687>
13. Yang HS, Li J, Yi X, Wang F. Performance evaluation of large language models with chain-of-thought reasoning ability in clinical laboratory case interpretation. *Clin Chem Lab Med* [Internet]. 2025 [cited 2025 Jun 24]; Available from: <https://pubmed.ncbi.nlm.nih.gov/40023838/>
14. Di Pumpo M, Riccardi MT, De Vita V, Damiani G. Evaluation of a large language model (ChatGPT) versus human researchers in assessing risk-of-bias and community engagement levels: a systematic review use-case analysis. *Eur J Public Health*. 2025 Jun 10:ckaf072. doi: 10.1093/eurpub/ckaf072. Epub ahead of print. PMID: 40493519.
15. Cocolicchio D, De Vita V, Cuoco F, Tona D, Porcelli M, Postorino D, Scaglioso F, Pascucci D, Regazzi L, Pastorino R, Laurenti P. Care-Vax: integrazione digitale ospedale-territorio per la tutela dei pazienti fragili e il potenziamento della copertura vaccinale secondo i principi della value-based healthcare. *Recenti Prog Med*. 2025 Oct;116(10):617-618. Italian. doi: 10.1701/4573.45803. PMID: 41037387.
16. Goh E, Gallo R, Hom J, Strong E, Weng Y, Kerman H, et al. Large Language Model Influence on Diagnostic Reasoning: A Randomized Clinical Trial. *JAMA Netw Open* [Internet]. 2024 Oct 1 [cited 2025 Jun 24];7(10):e2440969-e2440969. Available from: <https://jamanetwork.com/journals/jamanetworkopen/fullarticle/2825395>
17. Barnhill JW, Nicolò G, Pompili E, curatori. *DSM-5-TR: casi clinici*. Milano: R. Cortina; 2024.
18. Gercama AJ, De Haan M, Van der Vleuten CPM. Reliability of the Amsterdam Clinical Challenge Scale (ACCS): A new instrument to assess the level of difficulty of patient cases in medical education. *Med Educ* [Internet]. 2000 Jul 1 [cited 2025 Jun 24];34(7):519-24. Available from: [/doi/pdf/10.1046/j.1365-2923.2000.00663.x](https://doi.org/10.1046/j.1365-2923.2000.00663.x)
19. Turpin M, Michael J, Perez E, Bowman SR. Language Models Don't Always Say What They Think: Unfaithful Explanations in Chain-of-Thought Prompting. *Adv Neural Inf Process Syst* [Internet]. 2023 May 7 [cited 2025 Jun 27];36. Available from: <https://arxiv.org/pdf/2305.04388>
20. Chen Y, Benton J, Radhakrishnan A, Uesato J, Denison C, Schulman J, et al. Reasoning Models Don't Always Say What They Think. 2025 May 8 [cited 2025 Jun 27]; Available from: <http://arxiv.org/abs/2505.05410>

21. Wei J, Wang X, Schuurmans D, Bosma M, Ichter B, Xia F, et al. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. *Adv Neural Inf Process Syst* [Internet]. 2022 Jan 28 [cited 2025 Jun 24];35. Available from: <https://arxiv.org/pdf/2201.11903>
22. DeepL Translator. DeepL SE; 2025. Disponibile su: <https://www.deepl.com/translator> [Ultimo accesso: 15 ottobre 2025]
23. Feraco A, Armani A, Amoah I, Guseva E, Camajani E, Gorini S, et al. Assessing gender differences in food preferences and physical activity: a population-based survey. *Front Nutr* [Internet]. 2024 [cited 2025 May 15];11. Available from: <https://pubmed.ncbi.nlm.nih.gov/38445208/>
24. Koo TK, Li MY. A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. *J Chiropr Med* [Internet]. 2016 Jun 1 [cited 2025 Jun 26];15(2):155-63. Available from: <https://pubmed.ncbi.nlm.nih.gov/27330520/>
25. Schober P, Vetter TR. Adjustments for Multiple Testing in Medical Research. *Anesth Analg* [Internet]. 2020 Jan 1 [cited 2025 Jun 26];130(1):99. Available from: <https://pubmed.ncbi.nlm.nih.gov/31842175/>
26. Maher JM, Markey JC, Ebert-May D. The other half of the story: Effect size analysis in quantitative research. *CBE Life Sci Educ* [Internet]. 2013 Sep 4 [cited 2025 Jun 26];12(3):345-51. Available from: <https://pubmed.ncbi.nlm.nih.gov/24006382/>
27. De Vita V, Destro Castaniti B, Vassalli M, De Mori L, Lacalaprice D, Arcà E, Cristiano A, Battipaglia C, Risuleo PE, Dionisi T, Causio FA. Valutazione del ragionamento clinico dei reasoning large language models su casi clinici complessi. *Recenti Prog Med*. 2025 Oct;116(10):599-600. Italian. doi: 10.1701/4573.45794. PMID: 41037378.