Simulated Reasoning and Self-Verification in Generalist Large Language Models for Psychiatric Diagnostic Performance: Cross-Sectional Study

Karthik V Sarma MD PhD[1], Kaitlin E Hanss MD MPH[1], Andrew J M Halls MD[1], Daniel F Becker MD[1], Anne L Glowinski MD MPE[1], Andrew Krystal MD[1]

[1]Department of Psychiatry and Behavioral Sciences, University of California San Francisco, San Francisco, CA, USA

Abstract

Background. Large language models (LLMs) have rapidly garnered significant interest for application in psychiatry and behavioral health. However, recent studies have identified significant shortcomings and potential risks in the performance of LLM-based systems, complicating their application to psychiatric diagnosis. Two promising approaches to addressing these challenges and improving the efficacy of these models are simulated reasoning and self-verification, in which additional “reasoning tokens” are used to guide model output, either during or after inference.

Objectives. We aimed to explore how the use of simulated reasoning (via the use of large reasoning models, or LRMs) and self-verification (via supplemental prompting) affect the psychiatric diagnostic performance of language models.

Methods. 106 case vignettes and associated diagnoses were extracted from the DSM-5-TR Clinical Cases book, with permission. An LLM and an LRM model were selected from each of the two vendors studied (OpenAI and Google). Two inference approaches were developed, a Basic approach that directly prompted models to provide diagnoses, and a Self-Verification approach that augmented the Basic approach with additional prompts. All case vignettes were processed by the four models and two approaches, and diagnostic performance was evaluated using the sensitivity and positive predictive value (PPV). Linear mixed effect models were used to test for significant differences between the vendors, model types, and inference approaches.

Results. All vignettes were successfully processed by each model and inference approach. Sensitivity ranged from 0.732 to 0.817, and PPV ranged from 0.534 to 0.779. The best overall performance was found in *o3-pro* using SV, with a sensitivity of 0.782 and a PPV of 0.779. No statistically significant fixed effects were found for sensitivity. For PPV, a statistically significant effect was found for prompt type (SV – coefficient 0.09, p=0.002), model type (LRM – coefficient 0.09, p=0.003), and the interaction between model type and vendor (LRM:OpenAI – coefficient 0.10, p=0.021).

Conclusions. We found that both simulated reasoning and self-verification yielded statistically significant improvements in the PPV, without significant differences in the sensitivity. The addition of the manually specified self-verification prompt improved the PPV even when simulated reasoning was used. This suggests that future efforts to apply language models in behavioral health may benefit from a combination of manually crafted reasoning prompts and automated simulated reasoning.

Keywords
Large language models, artificial intelligence, diagnosis, DSM, simulated reasoning, self-verification

1. Introduction

Large language models (LLMs) have rapidly garnered significant interest for application in psychiatry and behavioral health. LLMs, a type of deep artificial intelligence (AI) system, are trained as “next token” (i.e., next word) predictors on very large-scale unsupervised text corpora, with some estimates suggesting that the most recent models are trained on the majority of all text written by humans that is publicly available. These models have demonstrated unexpected and remarkable emergent capabilities in the areas of information retrieval and processing and problem-solving, suggestive of a capability to engage with the semantic content of unstructured text not possible with previous technologies.

These emergent capabilities have enabled the demonstration of both clinical and consumer-facing LLM-based systems with a broad set of applications across fields of medicine. In behavioral health, projects have demonstrated their use in automated diagnostic and therapeutic interactive agents (i.e., chatbots)[1], question-answering[2], and diagnostic reasoning[3]. However, recent studies have also identified significant shortcomings and potential risks in the performance of LLM-based systems, including hallucination, bias, sycophancy, and the generation of associated misaligned and dangerous content for users experiencing mental health challenges.[4–7] These challenges complicate the application of large language models to psychiatric diagnosis, especially in the context of the limitations in interrater reliability of the DSM diagnoses.[8]

Two promising approaches to addressing these challenges and improving the efficacy of these models are simulated reasoning (SR) and self-verification (SV). In these methods, a structured approach is used to generate additional tokens (known as *reasoning tokens*), either generated by the model or provided by the system designer, that increase the probability of correct response. In SR, tokens are generated to create a reasoning “chain of thought” that leads to the model’s final response. In self-verification, once a preliminary response is generated by the model, reasoning tokens are added to lead the model to re-evaluate the correctness of this answer before generating a final result. LLMs that automatically incorporate the use of reasoning tokens are often known as large reasoning models (LRMs), and the major vendors of large language models have now produced LRM versions of their latest-generation tools (i.e., OpenAI o3, Google Gemini 2.5 Pro). Some recent works, however, have suggested that the use of reasoning tokens in LRM may not lead to reasoning-like processes[9–11] in some settings.

Though the use of LLM-based clinical tools continues to grow rapidly, few studies are available to guide the choice of model or reasoning approach for behavioral health, or to provide estimation of how these models will perform in this setting. In prior work, we

demonstrated that fixed prompt decision trees improve LLM psychiatric diagnostic performance, but the performance of SR remains unexplored. To address this evidence gap, we aim here to evaluate whether the use of SR (i.e., LRMs) and/or SV (i.e., standardized post-processing) enhances model performance on this task.

2. Methods

2.1 Standardized Task and Dataset.

We extracted 106 case vignettes written by 133 discussants across 19 primary DSM-5-TR diagnostic categories from the DSM-5-TR Clinical Cases book. Each vignette is a multi-paragraph description of the particular case; the vignettes are written for educational purposes; the author's intended final diagnoses are also provided separate from the vignette. The standardized task was to generate diagnoses from the case vignette, with the author-designated diagnoses used as the gold-standard comparison.

2.2 Model Selection and Configuration.

For this study, we selected one LLM and one LRM from two major vendors from whom an enterprise-grade LLM application programming interface was available, and which had agreed terms of use that were 1) compatible with the IP rights of the study researchers and rightsholders associated with the study data, and 2) prohibited the use of study data for model training. The vendors and models used for this study were: OpenAI (LLM: *gpt-4.1*, LRM: *o3-pro*), and Google (LLM: *Gemini 2.5 Flash*, LRM: *Gemini 2.5 Pro*). The LLMs were used at their default settings, with an output token limit of 2000. LRMs were either set to medium reasoning effort (o3-pro) or to a maximum of 16,000 reasoning tokens (Gemini 2.5 Pro). Reasoning was disabled for both LLMs. Safety settings were disabled or set to the minimum level of filtering. Models were constrained to respond with JSON-parsable outputs either through the use of API flags or by re-prompting until the output was in valid JSON.

2.3 Inference Approaches.

We compared two inference approaches previously developed by the authors[3] using each of the available models. In the first approach (the "basic" approach), the model was directly prompted to assign diagnoses to the vignette using a standardized prompt. In the second approach (the "SV" or self-verification approach), diagnoses generated by the basic approach were then re-presented pairwise to the models using a standardized prompt to select which of the diagnoses were necessary. See Table 1 for prompt templates.

2.4 Scoring and Analysis.

<u>Diagnosis Matching and Simplification</u>

To facilitate ease of comparison of model-generated and author-designated diagnoses, a previously developed semi-automated standardized diagnosis matching and

simplification system was applied to all diagnoses. In this system, all specifiers and modifiers were removed, neurocognitive disorders were collapsed into a single diagnosis, and then diagnoses were systematically matched to DSM-5-TR diagnoses, Z codes, or other categories. See Multimedia Appendix 1 for full detail on this matching system.

Statistical Analysis.
For every case, we calculated the sensitivity and positive predictive value (PPV) based on each model's predicted diagnoses and the author-designated diagnoses. We then fit linear mixed effect models to both sensitivity and PPV to examine the fixed effects of model vendor (OpenAI, Google), type (LLM, LRM) and use of the SV prompt. A threshold of $p < 0.05$ was used to evaluate for significant fixed effects.

3. Results

All 106 vignettes were successfully processed by the four models, without unresolved errors or context window length overruns. Simplification and matching were completed and all model-predicted diagnoses were matched to a DSM-5-TR diagnosis, Z code, or a non-psychiatric diagnosis. Mean values for sensitivity and PPV for each experiment are available in Table 2. Sensitivity ranged from 0.732 to 0.817, PPV ranged from 0.534 to 0.779. Per-case performance is depicted in Figure 1.

Linear mixed effects models were successfully fit, including a random effect, fixed effects for model type, vendor, and inference approach, and interactions. The intraclass correlation (ICC) was 0.69 for sensitivity and 0.63 for PPV. No statistically significant fixed effects were found for sensitivity. For PPV, a statistically significant effect was found for prompt type (SV – coefficient 0.09, p=0.002), model type (LRM – coefficient 0.09, p=0.003), and the interaction between model type and vendor (LRM:OpenAI – coefficient 0.10, p=0.021).

4. Discussion

In this work, we sought to evaluate the impact of simulated reasoning and self-verification on the psychiatric diagnostic performance of language models from two different vendors. We successfully proceed 106 psychiatric case vignettes using the selected approaches, and found that both simulated reasoning and self-verification yielded statistically significant improvements in the positive predictive value, without significant differences in the sensitivity. The best overall performance of models was found in *o3-pro* using SV, with a sensitivity of 0.782 and a PPV of 0.779. The high ICC found for both metrics suggests that most of the variance found in our experiments was explained by the intrinsic difficulty of the individual vignettes, a finding also clear in the per-vignette results depicted in Figure 1.

Interestingly, we found no statistically significant interaction effect between the two approaches, suggesting that the use of both yielded additional benefit over any one alone. This suggests that the addition of explicit reasoning suggestions in prompts (such

as that in the SV prompt) may continue to be helpful even when using automated simulating reasoning systems, and that future efforts aimed at using language models for diagnostic application should continue to integrate expert insight. This finding may, however, be due to a lack of sufficient power from our study to detect these differences and future evaluation is required. Similarly, we found no significant impact of the use of either approach on sensitivity; however, a non-significant improvement from the use of LRMs and impairment from the use of SV was found. We also found that the OpenAI LRM performed statistically significantly better than the Google LRM; however, this may be confounded by underlying differences in the language model used for *gpt-4.1* and *o3-pro*; the two Google models use a similar underlying model. This difference represents a limitation of our work.

The semi-automated simplification and matching process we used was intended to given models the best opportunity to be scored as correct, and this may have the impact of overestimating the performance of the models. Additionally, the vignettes and diagnoses used for our experiment may be in the training set of these models (though they have not been licensed for such use); our results are not suggestive that the models have memorized our data but it is possible that this biases our evaluation towards better performance. Since the vignettes were written for teaching, the prevalence of diagnoses is higher in the dataset than in the general population and more comprehensive information is provided in the vignettes than is generally available in the clinical setting. This may serve to inflate the performance metrics that we found.

In conclusion, we found that both simulated reasoning and self-verification improved psychiatric diagnostic performance in language models. Future efforts could make use of larger datasets and more broadly examine the impact of model selection and prompt optimization to better elucidate how best to configure language models for psychiatric use.

Acknowledgements. This work was supported by the National Institute of Mental Health of the National Institutes of Health [grant number R25 MH060482]. The authors acknowledge the mentorship and scientific vision provided by Dr. Atul J Butte, MD PhD, director of the UCSF Bakar Computational Health Sciences Institute, who passed away in June 2025. We thank the UCSF AI Tiger Team, UCSF Academic Research Services, UCSF Research Information Technology, and the UCSF Chancellor’s Task Force for Generative AI for their support in developing LLM resources used for this project. Generative AI was used as described in the methods section for the execution of this research, but was not otherwise used in the writing of this manuscript. The authors appreciated the opportunity to submit an earlier component of this work to the 2026 ACNP Annual Meeting.

Conflicts of Interest. Karthik V Sarma reports a relationship with SimX, Inc. that includes: board membership, employment, equity or stocks, funding grants, and travel reimbursement, and a relationship with OpenEvidence, Inc. that includes: equity or stocks. Andrew Krystal reports a relationship with Big Health that includes: consulting or advisory and equity or stocks. The disclosed organizations had no role in any

component of the conduct of the research. The authors disclose no other interests related to the content of this manuscript.

Data Availability. The dataset used for this study was extracted from the DSM-5-TR Clinical Cases textbook under license from the American Psychiatric Association (as follows), and cannot be reproduced by the authors of this study:

> This research was made possible through the use of content belonging to the American Psychiatric Association; express permission was obtained from the American Psychiatric Association for the use of such content (DSM-5-TR Clinical Cases. Copyright © 2023. American Psychiatric Association. All Rights Reserved, including rights for text and data mining (TDM), Artificial Intelligence (AI) training, and similar technologies).

Authors' Contributions.
Karthik V Sarma: conceptualization, methodology, software, validation, formal analysis, investigation, resources, data curation, writing – original draft, writing – review & editing, visualization, supervision, project administration
Kaitlin E Hanss: methodology, software, data curation, writing – review & editing
Andrew J M Halls: methodology, writing – review & editing, supervision
Daniel F Becker: methodology, writing – review & editing, supervision
Anne L Glowinski: methodology, writing – review & editing, supervision
Andrew Krystal: methodology, resources, writing – review & editing, supervision, funding acquisition

Abbreviations.
AI – Artificial Intelligence
DSM – Diagnostic and Statistical Manual
DSM-5-TR – Diagnostic and Statistical Manual, Version 5, Text Revision
JSON – JavaScript Object Notation
LLM – Large Language Model
LRM – Large Reasoning Model
PPV – Positive Predictive Value
SR – Simulated Reasoning
SV – Self-Verification

References

1. Heinz MV, Mackin DM, Trudeau BM, et al. Randomized Trial of a Generative AI Chatbot for Mental Health Treatment. *NEJM AI*. 2025;2(4):AIoa2400802. doi:10.1056/AIoa2400802

2. Hanss K, Sarma KV, Glowinski AL, et al. Assessing the Accuracy and Reliability of Large Language Models in Psychiatry Using Standardized Multiple-Choice Questions: Cross-Sectional Study. *Journal of Medical Internet Research*. 2025;27(1):e69910. doi:10.2196/69910

3. Sarma KV, Hanss KE, Halls AJM, et al. Integrating Expert Knowledge into Large Language Models Improves Performance for Psychiatric Reasoning and Diagnosis.

medRxiv. Preprint posted online July 21, 2025:2025.07.19.25331840. doi:10.1101/2025.07.19.25331840

4. King DR, Nanda G, Stoddard J, et al. An Introduction to Generative Artificial Intelligence in Mental Health Care: Considerations and Guidance. *Curr Psychiatry Rep*. 2023;25(12):839-846. doi:10.1007/s11920-023-01477-x

5. Obradovich N, Khalsa SS, Khan WU, et al. Opportunities and risks of large language models in psychiatry. *NPP—Digit Psychiatry Neurosci*. 2024;2(1):1-8. doi:10.1038/s44277-024-00010-z

6. Volkmer S, Meyer-Lindenberg A, Schwarz E. Large language models in psychiatry: Opportunities and challenges. *Psychiatry Research*. 2024;339:116026. doi:10.1016/j.psychres.2024.116026

7. Orrù G, Melis G, Sartori G. Large language models and psychiatry. *International Journal of Law and Psychiatry*. 2025;101:102086. doi:10.1016/j.ijlp.2025.102086

8. Clarke DE, Narrow WE, Regier DA, et al. DSM-5 Field Trials in the United States and Canada, Part I: Study Design, Sampling Strategy, Implementation, and Analytic Approaches. *AJP*. 2013;170(1):43-58. doi:10.1176/appi.ajp.2012.12070998

9. Zhao C, Tan Z, Ma P, et al. Is Chain-of-Thought Reasoning of LLMs a Mirage? A Data Distribution Lens. *arXiv*. Preprint posted online August 13, 2025. doi:10.48550/arXiv.2508.01191

10. Chen Y, Benton J, Radhakrishnan A, et al. Reasoning Models Don’t Always Say What They Think. *arXiv*. Preprint posted online May 8, 2025. doi:10.48550/arXiv.2505.05410

11. Shojaee*† P, Mirzadeh* I, Alizadeh K, Horton M, Bengio S, Farajtabar M. The Illusion of Thinking: Understanding the Strengths and Limitations of Reasoning Models via the Lens of Problem Complexity. Published online 2025. https://ml-site.cdn-apple.com/papers/the-illusion-of-thinking.pdf

Table 1. Prompt Templates for Inference Approaches

Basic Prompt I am going to give you an academic psychiatry clinical case that describes a patient with one or more psychiatric DSM-5-TR diagnoses, and you are going to answer questions about that case that I provide. The clinical case is as follows: <X> Please provide me with a list of DSM-5-TR diagnoses, without specifiers or modifiers, that you believe apply to this patient based solely on the clinical case. Please format them as a JSON list titled "diagnoses" with one diagnosis per entry. Do not include any other text in your response. Do not include any incorrect, inappropriate, or candidate diagnoses. For example, if the diagnoses are "Insomnia Disorder" and "Bipolar I Disorder", you would reply with: {"diagnoses": ["Insomnia Disorder", "Bipolar I Disorder"]}
Self-Verification Prompt The following two DSM-5-TR diagnoses are candidate diagnoses for this patient: <X> and <Y>. We are interested in deciding if both diagnoses are necessary for the patient, or if one of these two diagnoses is better explained by the other. Please respond with a list of which of these two diagnoses are necessary for the patient in JSON format. For example, if the candidate diagnoses are 'major depressive disorder' and 'adjustment disorder', respond with ["major depressive disorder", "adjustment disorder"] if both diagnoses are necessary, or with either ["major depressive disorder"] or ["adjustment disorder"] if one diagnosis is better than the other.

Table 2. LLM Diagnostic Performance Metrics by Experiment.

Vendor	Model Type	Inference Approach	Sensitivity	PPV
Google	LLM	Basic	0.767	0.534
		SV	0.750	0.628
	LRM	Basic	0.793	0.625
		SV	0.770	0.673
OpenAI	LLM	Basic	0.750	0.569
		SV	0.732	0.640
	LRM	Basic	0.817	0.759
		SV	0.782	0.779

LLM = Large Language Model; LRM = Large Reasoning Model; SV = Self-Verification; PPV = Positive Predictive Value

Figure 1. **Per-case results of evaluation of diagnostic performance, by experiment and diagnostic category.** This color maps displays the per-case sensitivity and positive predictive value (PPV) for diagnosis for every vignette, grouped by DSM-5-TR category. Each row consists of experimental results in one specific category for one specific experiment (i.e., combination of model and inference approach). A legend is displayed for the color map. Each column within each metric represents a single specific case vignette.

SENSITIVITY
PPV
SENSITIVITY
PPV
Depressive Disorders
Personality Disorders
Trauma- and Stressor-Related Disorders
Bipolar and Related Disorders
Neurocognitive Disorders
Substance-Related and Addictive Disorders
Schizophrenia Spectrum and Other Psychotic Disorders
Somatic Symptom and Related Disorders
Anxiety Disorders
Developmental Disorders
Obsessive-Compulsive and Related Disorders
Feeding and Eating Disorders
Sleep-Wake Disorders
Paraphilic Disorders
Dissociative Disorders
Sexual Dysfunctions
Disruptive, Impulse-Control, and Conduct Disorders
Elimination Disorders
Gender Dysphoria
Gemini 2.5 Flash LLM
Gemini 2.5 Pro LRM
gpt-4.1 LLM
o3-pro LRM
Basic
SV
LEGEND
0
0.25
0.5
0.75
1