Evaluating Diagnostic Accuracy and Clinical Reasoning of Multiple Large Language Models in Psychiatry

Kevin W. Jin, BS[1,2,+], Yasna Rostam-Abadi, MD[3], Pooja Chaudhary, MD[4], Margaret A. Garrett, MD[4], Ashley S. Huang, MD[4], Mario Montelongo, MD[4], Caesa Nagpal, MD[4], Jasperina Shei, MD[4], Judah Weathers, DPhil, MD[5], Juliana S. Zhang, MD[4], Qingyu Chen[2], PhD, Jiyeong Kim, PhD[6], Matteo Malgaroli, PhD[7,8], Walter S. Mathis, MD[3], Carolyn I. Rodriguez, MD, PhD[9,10], Salih Selek, MD[4], Manu S. Sharma, MD[3,11], Christopher Pittenger, MD, PhD[3,5,12-15], Sarah W. Yip, PhD, MSc[3,5,14,15], Brian A. Zaboski, PhD[3], and Hua Xu, PhD[2,14]

[1] Program in Computational Biology and Biomedical Informatics, Yale University, New Haven, CT
[2] Department of Biomedical Informatics and Data Science, Yale School of Medicine, New Haven, CT
[3] Department of Psychiatry, Yale School of Medicine, New Haven, CT
[4] Louis A. Faillace, MD, Department of Psychiatry and Behavioral Sciences, McGovern Medical School at The University of Texas Health Science Center at Houston, Houston, TX
[5] Yale Child Study Center, Yale School of Medicine, New Haven, CT
[6] Stanford Center for Digital Health, Department of Medicine, Stanford University, Stanford, CA
[7] Department of Psychiatry, New York University Grossman School of Medicine, New York, NY
[8] New York University Center for Data Science, New York, NY
[9] Department of Psychiatry and Behavioral Sciences, Stanford School of Medicine, Stanford, CA
[10] Veterans Affairs Palo Alto Health Care System, Palo Alto, CA
[11] The Institute of Living, Hartford Healthcare, Hartford, CT
[12] Department of Psychology, Yale University, New Haven, CT
[13] Department of Neuroscience, Yale School of Medicine, New Haven, CT
[14] Wu Tsai Institute, Yale University, New Haven, CT
[15] Center for Brain & Mind Health, Yale School of Medicine, New Haven, CT

+ Address correspondence to Kevin W. Jin (kevin.jin@yale.edu)

Research in Context

Evidence before this study: We searched PubMed and Scopus for studies evaluating large language models for psychiatric diagnosis and/or differential diagnosis from text vignettes. Searches were run from database inception to February 6, 2026, using terms including ("large language model" OR LLM "artificial intelligence" OR "generative AI" OR AI OR ChatGPT OR GPT OR Claude OR Gemini OR DeepSeek OR Llama) AND (psychiatr* OR mental OR DSM OR "differential diagnosis" OR diagnos*) AND (vignette OR case OR "case report"). We included empirical studies that evaluated model diagnostic outputs using psychiatric cases/vignettes and excluded editorials/commentaries and studies that did not report case-level diagnostic performance. We did not formally assess risk of bias/quality; studies were heterogeneous in vignette sources, model access, and outcome definitions, precluding quantitative pooling. Prior studies have typically examined small vignette sets, focused on narrow diagnostic domains, evaluated single models, or relied primarily on outcome-based accuracy metrics. When diagnostic reasoning has been assessed, it has usually been inferred from post hoc explanations accompanying model outputs rather than evaluated as a distinct, clinician-visible artifact. Clinician-grounded evaluations of diagnostic reasoning across multiple contemporary models remain limited.

Added value of this study: This study provides a large-scale, clinician-grounded evaluation of diagnostic accuracy and diagnostic reasoning quality across four contemporary large language models using a diverse dataset of psychiatric case vignettes. Rather than relying solely on outcome-based explanations, we directly evaluated model-generated diagnostic reasoning traces as clinician-visible artifacts using structured clinician ratings and qualitative analysis. By

integrating multiple accuracy metrics with clinician assessment of reasoning coherence, flexibility, and plausibility, we demonstrate that clinician-rated reasoning quality is strongly associated with diagnostic correctness, whereas data extraction quality alone is not. Our analysis also identifies recurrent reasoning failure modes not captured by accuracy metrics, highlighting psychiatry as a stringent testbed for evaluating reasoning in narrative-driven clinical domains.

Implications of all the available evidence: Evaluations of large language models for clinical decision support should extend beyond accuracy to include systematic assessment of clinician-visible diagnostic reasoning. Mixed-methods, clinician-grounded evaluation frameworks that examine both diagnostic outcomes and reasoning artifacts may be critical for responsible assessment of LLMs in psychiatry and other areas of medicine where diagnosis depends on interpretation, judgment, and tolerance of uncertainty.

Summary

Background: Existing large language model (LLM) evaluations rely on accuracy benchmarks that fail to capture whether models reason well while making diagnoses. Studies that do analyse reasoning focus on post hoc explanations accompanying model outputs rather than distinct, clinician-visible artifacts such as detailed reasoning traces. This creates a translational gap in domains such as psychiatry, where diagnosis relies on narrative interpretation, diagnostic reasoning, and clinical judgment under uncertainty.

Methods: We conducted a mixed-methods evaluation of four state-of-the-art LLMs using a clinician-curated dataset of 196 psychiatric case vignettes, including 135 published cases and 61 novel clinician-authored vignettes. Diagnostic accuracy was assessed using multiple metrics (top-1 accuracy, top-5 accuracy, recall@5, and mean reciprocal rank) based on ranked lists of five differential diagnoses per vignette. Clinical reasoning quality was evaluated on a randomly selected subset of 30 vignettes through clinician assessment of model-generated diagnostic reasoning traces alongside qualitative commentary from board-certified psychiatrists. We examined the association between clinician-rated reasoning quality and diagnostic correctness and included an illustrative comparison with psychiatry residents.

Findings: Clinician-rated diagnostic reasoning quality was strongly associated with diagnostic correctness in mixed-effects logistic regression analyses ($\beta = 1{\cdot}80$; $p < 0{\cdot}001$), whereas data extraction quality alone was not. Across the full vignette set, models demonstrated moderate to high diagnostic accuracy. The highest-performing model achieved a top-5 accuracy of 0·801 and

also received the highest clinician-rated reasoning scores. In an illustrative comparison, model diagnostic accuracy fell within the range observed for psychiatry residents.

Interpretation: Diagnostic reasoning quality captures clinically meaningful variation in LLM performance beyond accuracy metrics. Psychiatry may represent a stringent testbed for evaluating reasoning in narrative-driven clinical domains. Evaluation frameworks for LLM-based clinical decision support should incorporate structured assessment of reasoning processes, not accuracy alone.

Funding: API credits from the OpenAI Researcher Access Program and Google Gemini Academic Program.

Main Text

Introduction

Psychiatric disorders represent a substantial global health burden, with over one billion people living with at least one mental disorder worldwide,[1] yet diagnosis in psychiatry remains uniquely challenging. Unlike many other areas of medicine, psychiatric diagnosis relies heavily on the interpretation of patient-reported symptoms, contextual information, and clinical judgment rather than definitive biological markers. As a result, diagnostic decisions often involve uncertainty and may vary across clinicians, particularly in time-constrained settings or where clinical experience differs.

Psychiatric diagnosis is grounded in phenomenological assessment using classificatory frameworks such as the Diagnostic and Statistical Manual of Mental Disorders and the International Classification of Diseases (ICD). While these frameworks have improved diagnostic reliability, the absence of objective biomarkers means that diagnosis depends on clinicians' ability to synthesise narrative information, weigh competing hypotheses, and tolerate ambiguity. This interpretative process is central to psychiatric practice but also presents challenges for standardisation and evaluation. Consequently, there is growing interest in whether artificial intelligence (AI) systems might support psychiatric diagnostic reasoning, provided their performance and reasoning are rigorously assessed.

Large language models (LLMs) have demonstrated strong performance across a range of medical tasks, including answering clinical questions and generating differential diagnoses. In several non-psychiatric domains, LLMs have been shown to perform comparably to clinicians in

controlled diagnostic evaluations.[2–5] However, relatively few studies have examined LLM performance in psychiatry, despite widespread public use of these models for mental health-related queries and concerns about potential harms.[6–8] Existing psychiatric evaluations of LLMs have been limited by small vignette sets, narrow diagnostic scope, reliance on older models, or a primary focus on outcome-based accuracy metrics.[9–12]

Importantly, diagnostic accuracy alone provides limited insight into how models arrive at their conclusions. Recent commentaries have highlighted the limitations of accuracy-only benchmarks for evaluating LLMs in clinical settings and called for clinician-grounded assessment of reasoning and safety.[13] Contemporary advances enabling LLMs to produce explicit, clinician-visible reasoning traces create an opportunity to evaluate not only diagnostic outputs but also the quality of model-generated diagnostic reasoning. In psychiatry, where reasoning coherence, flexibility, and plausibility are critical, such evaluation is particularly informative. To date, no study has examined whether clinician-assessed reasoning quality predicts diagnostic correctness or characterised the failure modes that accuracy metrics miss.

In this study, we conducted a mixed-methods evaluation of four state-of-the-art LLMs using a diverse set of psychiatric case vignettes. We integrated automated, multi-metric assessment of diagnostic accuracy with a clinician-grounded evaluation of model-generated diagnostic reasoning across two clinically meaningful dimensions, alongside an illustrative comparison with psychiatry residents. Together, these analyses clarify how LLMs perform in psychiatric diagnosis and identify strengths and limitations relevant to their potential role as clinical decision support within narrative-driven clinical domains.

Methods

We constructed a novel clinician-curated dataset of 196 psychiatric case vignettes to evaluate four LLMs. Models were assessed across three aspects of diagnostic performance: diagnostic accuracy, clinician-rated diagnostic reasoning quality, and an illustrative comparison with psychiatry residents (**Figure 1**). This study involved analysis of de-identified published case reports and fictitious clinician-authored vignettes and did not involve prospective recruitment or interaction with human participants, and no new patient data were collected; therefore, institutional review board approval was not sought.

Dataset

We assembled a dataset of psychiatric case vignettes from two sources: published diagnostic case reports and clinician-authored fictitious vignettes.

An initial pool of 183 published diagnostic case reports was identified from the medical literature, including *Case Reports in Psychiatry*, *The New England Journal of Medicine*, *NEJM Evidence*, *The Journal of the American Medical Association*, *JAMA Psychiatry*, *The Lancet*, and the *Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition, Text Revision* (*DSM-5-TR) Clinical Cases*. These case reports were manually collected and independently reviewed by seven psychiatrists (two board-certified psychiatrists and five psychiatry residents) using clinician-defined inclusion criteria (as detailed in the Supplementary Material), yielding 135 curated case reports.

To complement published material and assess performance on non-public cases, six doctoral-level clinicians (two board-certified psychiatrists, three psychiatry residents, and one board-

certified clinical psychologist) authored 61 fictitious vignettes based on their real-world clinical experience. Vignettes were paragraph-length, conformed to the same inclusion criteria as the published case reports, and used standardised *DSM-5-TR* diagnoses. Authors were encouraged to include diagnostically challenging cases. Each clinician authored approximately 10 vignettes.

The final dataset comprised 196 vignettes (135 published; 61 fictitious), spanning a wide range of psychiatric diagnoses and diagnostic complexity.

Models and Prompts

We evaluated four contemporary large language models: three proprietary models (Claude Opus 4·5, Gemini 3 Pro, and GPT-5·2) and one open-source model, DeepSeek-V3·2. Models were selected for state-of-the-art performance and their ability to generate explicit, clinician-visible diagnostic reasoning alongside diagnostic outputs, enabling clinical evaluation of reasoning quality.[14–17] Full model specifications are provided in the Supplementary Material.

Each model was prompted to generate the five most probable *DSM-5-TR* diagnoses for each vignette, with the primary diagnosis in the topmost position and secondary diagnoses in the remaining positions. For evaluative purposes, the model was instructed to provide only diagnoses codified in the *DSM-5-TR* and include the associated ICD-10 code for disambiguation.

Each model interaction consisted of a system prompt specifying the model's role and response constraints and a user prompt providing the vignette and diagnostic task. Prompts were iteratively engineered to instruct the model to assume the role of a psychiatrist adhering to *DSM-5-TR* diagnostic criteria and generate a ranked list of diagnoses. All models were evaluated in a zero-shot setting without task-specific examples. Clinical reasoning guidance adapted from the

Accreditation Council for Graduate Medical Education (ACGME) Psychiatry Residency Milestones was included in the user prompt.[18,19] Prompts are reported in the Supplementary Material. Model hyperparameters were configured according to provider recommendations to elicit optimal reasoning behavior (as detailed in the Supplementary Material).[20–23] All models were accessed between December 2025 and January 2026.

Evaluation

We designed three tasks to evaluate the LLMs along three aspects of psychiatric diagnostic performance:

Task 1: Diagnostic Accuracy

Diagnostic accuracy was assessed on all 196 vignettes using top-1 accuracy, top-5 accuracy, recall@5, and mean reciprocal rank (MRR), based on the five predicted diagnoses per vignette. Top-1 accuracy provides insight into the model's ability to prioritise the correct primary diagnosis within its differential, and top-5 accuracy gives the model's hit rate for any correct diagnoses given five chances. Recall@5 measures diagnostic coverage, assessing the percentage of correct diagnoses successfully predicted in the differential. MRR measures the model's tendency to rank the primary diagnosis towards the top of the list.

The foremost diagnosis in each vignette was treated as the primary reference diagnosis. For all metrics, each of the diagnoses in the model's predicted differential was matched against the reference diagnosis using a separate LLM prompted to act as an independent psychiatric adjudicator. LLM-based adjudication has been shown to achieve agreement comparable to human raters in prior medical evaluation studies.[24] We used GPT-5-mini as our adjudicator in this study based on its high concordance with human raters in medical adjudicative tasks.[25,26]

Task 2: Human Evaluation of Diagnostic Reasoning

To complement our automated diagnostic accuracy metrics, we conducted a clinician-grounded evaluation of model-generated diagnostic reasoning using a mixed-methods approach. This analysis assessed whether clinician judgements of reasoning quality aligned with diagnostic correctness (whether a model's predicted primary diagnosis was included anywhere in its differential), and characterised common reasoning strengths and failure modes.

A randomly selected subset of 30 vignettes was evaluated by five board-certified psychiatrists ("human judges"). For each vignette-model pair, human judges performed three assessments: (1) Determine whether the model's predicted primary diagnosis was included anywhere in the generated differential diagnosis; (2) Provide two numerical Likert scores of clinical reasoning quality from 0 (worst) to 4 (best), based on a clinical reasoning rubric adapted from the ACGME Psychiatry Residency Milestones, with the first score rating the data extraction and organizational ability and the second rating diagnostic reasoning and differential diagnosis ability;[18,19] and (3) Provide brief free-text commentary guided by three predefined prompts addressing (i) logical coherence, (ii) presence of unsafe, stigmatizing, or hallucinated content, and (iii) flexibility of diagnostic reasoning under diagnostic ambiguity. Each comment corresponded to a single clinician's evaluation of a specific vignette-model pair.

Using assessments (1) and (2), we calculated average reasoning quality scores for all four LLMs and modeled the association between diagnostic correctness and reasoning quality scores using a mixed-effects logistic regression model in R (version 4·5·2) using the "lme4" package.

Free-text commentary from assessment (3) was analysed qualitatively to identify recurring reasoning patterns and failure modes. Semantic grouping was used solely to facilitate manual thematic review; all themes were defined and assigned through direct human interpretation (as detailed in the Supplementary Material).

Human judges were blinded to model identity throughout the evaluation. Judges also provided background information regarding clinical experience and familiarity with large language models to assess potential contributions to annotation variability.

Interrater agreement was assessed using Fleiss's κ for diagnostic correctness and intraclass correlation coefficients for reasoning scores (as detailed in the Supplementary Material). Substantial agreement was observed for diagnostic correctness, with good to excellent agreement for reasoning dimensions, supporting the reliability of downstream quantitative and qualitative analyses.

Task 3: Illustrative Benchmark with Trainee Clinicians

We took the best-performing model defined based on top-5 accuracy from Task 1 and illustratively compared its diagnostic accuracy with the mean performance of two psychiatry residents, using the same subset of 30 vignettes from Task 2. Residents generated five differential diagnoses per vignette, and diagnostic accuracy was measured using the same adjudicator from Task 1. This comparison was intended solely to provide contextual reference rather than to support statistical inference.

Role of the Funding Source

The funders had no role in study design, data collection, data analysis, data interpretation, or writing of the report. All authors had full access to all the data in the study and accept responsibility for the decision to submit for publication.

Results

Task 1: Diagnostic Accuracy

Table 1 summarises the psychiatric diagnostic accuracy of the four evaluated LLMs across 196 vignettes, reported as mean top-5 accuracy, top-1 accuracy, recall@5, and mean reciprocal rank (MRR). All models demonstrated comparable performance across metrics, with top-5 accuracy ranging from 0·730 to 0·801, top-1 accuracy from 0·566 to 0·638, recall@5 from 0·662 to 0·731, and MRR from 0·630 to 0·710.

To assess whether diagnostic accuracy might be influenced by memorization of published case reports, we conducted a robustness analysis by comparing top-1 diagnostic accuracy on clinician-authored fictitious vignettes and vignettes derived from the medical literature. Across all four models, top-1 accuracy estimates were similar across vignette sources, with overlapping 95% confidence intervals (**Supplementary Figure 1** and **Supplementary Table 1A**). No model demonstrated consistently higher accuracy on medical literature vignettes. In mixed-effects logistic regression models accounting for case-level variability, vignette source was not significantly associated with diagnostic correctness for any model (all $p > 0{\cdot}05$; **Supplementary Table 1B**).

Task 2: Human Evaluation of Diagnostic Reasoning

Across all evaluated cases, mean clinician-rated diagnostic reasoning scores ranged from 2·37 to 3·60 across models, and mean extraction scores ranged from 2·21 to 3·67 (**Table 2**). Interrater agreement for diagnostic correctness was substantial (Fleiss's κ = 0·626; average pairwise agreement = 0·767). Agreement for human-judged reasoning dimensions was higher, with intraclass correlation coefficients indicating good agreement for data extraction (ICC = 0·773; 95% CI: 0·702-0·831) and excellent agreement for diagnostic reasoning quality (ICC = 0·84; 95% CI: 0·790-0·881).

In a mixed-effects logistic regression model accounting for case- and judge-level variability, higher clinician-rated diagnostic reasoning quality was strongly associated with diagnostic correctness ($\beta = 1{\cdot}80$, $p < 0{\cdot}001$), corresponding to an approximately six-fold increase in the odds of diagnostic correctness per one-point increase on the reasoning scale (**Figure 2**). This relationship persisted after controlling for model identity. In contrast, data extraction quality was not independently associated with diagnostic correctness after adjustment ($p = 0{\cdot}36$).

Clinician free-text commentary provided additional insight into model reasoning beyond numerical scores. Across all comments, concerns related to reasoning coherence and flexibility under diagnostic ambiguity were more frequently noted than safety-related concerns, which were infrequent across models (**Supplementary Table 2A**). The prevalence of these concerns varied by model (**Supplementary Table 2B**). Analysis of non-boilerplate clinician comments identified recurring qualitative themes, including conclusion-driven (post hoc) reasoning, coherent reasoning with incorrect diagnosis, and coherent but sparse reasoning. **Supplementary Table**

2C summarises the prevalence of these themes by model, reported as *n* (%) using non-boilerplate comments as the denominator for each model.

Qualitative observations aligned with quantitative findings: clinicians frequently identified reasoning deficiencies, such as post hoc justification, even when the final diagnosis was correct, underscoring that diagnostic accuracy alone may not capture clinically relevant limitations in reasoning processes.

Task 3: Illustrative Comparison with Trainee Clinicians

Table 3 presents an illustrative comparison of diagnostic accuracy between Claude Opus 4·5 and two psychiatry residents on the 30-vignette subset. Claude Opus 4·5 exhibited higher observed diagnostic accuracy than the averaged trainee performance (top-5 accuracy, 0·867 versus 0·767; top-1 accuracy, 0·733 versus 0·617).

Discussion

In this mixed-methods evaluation of LLMs on psychiatric case vignettes, we assessed diagnostic accuracy, clinician-rated diagnostic reasoning quality, and qualitative reasoning characteristics across four contemporary models. Across a diverse set of 196 vignettes, models demonstrated moderate to high diagnostic accuracy. On a clinician-evaluated subset, higher clinician-rated diagnostic reasoning quality was strongly associated with diagnostic correctness, whereas data extraction quality alone was not. Together, these findings suggest that diagnostic reasoning quality captures clinically meaningful variation in model performance beyond accuracy alone.

This association is particularly salient in psychiatry, where diagnosis relies heavily on the synthesis and interpretation of narrative information under uncertainty, and is complicated by a lack of confirmatory biomarkers, high comorbidity, diagnostic instability, and cultural variation in symptom presentation, requiring a high tolerance for ambiguity. Our findings suggest that models capable of articulating coherent, flexible reasoning are more likely to arrive at correct diagnoses, underscoring the importance of evaluating reasoning processes in addition to accuracy. Psychiatry may therefore serve as a particularly stringent testbed for evaluating diagnostic reasoning in LLMs. The properties that make psychiatry diagnostically demanding are shared by other narrative-driven clinical domains, suggesting that reasoning evaluation methods validated in psychiatry may transfer broadly.

More broadly, our findings suggest that regulatory or institutional evaluation of LLM-based clinical decision support should require structured assessment of reasoning traces, not just accuracy on benchmark datasets. The strong reasoning-correctness association observed suggests that clinician assessment of reasoning traces is both feasible (given the high interrater agreement) and informative (given its predictive relationship with correctness). Together, these pieces of evidence support the inclusion of reasoning assessments in pre-deployment evaluation standards.

Prior evaluations of LLMs in psychiatry have often focused on narrow diagnostic domains, small vignette sets, or single performance metrics. For example, existing studies have examined model performance on specific conditions or have relied primarily on outcome-based accuracy without a structured assessment of clinical reasoning quality.[9–12] In contrast, our study integrates multiple accuracy metrics with clinician-grounded evaluation of diagnostic reasoning across a broader

range of psychiatric presentations, enabling the identification of clinically relevant failure modes that may not be apparent from accuracy alone.

Qualitative analysis of clinician commentary revealed several recurring reasoning patterns across models. Clinicians frequently identified post hoc or conclusion-driven reasoning, in which models applied correct diagnostic labels with limited or circular justification. Such patterns pose clinical risk by creating a misleading appearance of diagnostic competence. Other common patterns included coherent reasoning supporting incorrect diagnoses and insufficient flexibility in the face of diagnostic ambiguity. Safety-related concerns were infrequent, suggesting that overtly unsafe or stigmatizing content was uncommon in this evaluation. These qualitative findings align with the observed association between reasoning quality and diagnostic correctness, highlighting that correct diagnostic outputs do not necessarily reflect robust clinical reasoning.

Our evaluation relied on model-generated clinician-visible reasoning traces produced through publicly accessible interfaces.[21,27–29] Although these summaries may not necessarily reflect full internal reasoning processes, they represent the primary artifacts available for clinical evaluation of model reasoning. Clinicians were nonetheless able to consistently assess coherence, flexibility, and diagnostic plausibility using these summaries, as reflected by substantial interrater agreement. This suggests that, despite their limitations, model-generated clinician-visible reasoning traces provide an auditable artifact for comparative evaluation of reasoning quality across models, even if they are imperfect proxies for internal model computation.

The illustrative comparison with psychiatry trainees should be interpreted cautiously and was not designed to establish equivalence or superiority but rather to contextualise model performance relative to human reference points on a small subset of cases. While the highest-performing model demonstrated diagnostic accuracy within the range observed for the trainee clinicians, this finding does not imply readiness for clinical use and should not be extrapolated beyond the constrained evaluation setting.

Several limitations warrant consideration. First, diagnostic performance was assessed using case vignettes rather than real-world clinical encounters, which may not fully capture the complexity and longitudinal nature of psychiatric diagnosis. Second, the clinician-evaluated subset was limited to 30 vignettes, constraining the precision of human-rated analyses. Third, although efforts were made to mitigate memorization through vignette curation and inclusion of fictitious cases, residual exposure effects cannot be fully excluded. Fourth, the use of an LLM judge for the accuracy tasks may introduce a stylistic bias favoring models with similar training data or architectural origins to the judge.[30] Fifth, all models were evaluated using a single prompting strategy; performance may vary under alternative prompt designs, and the sensitivity of reasoning quality to prompt engineering warrants further investigation. Finally, qualitative theme identification involved manual interpretation by the study authors, reflecting interpretive judgment applied to clinician-generated commentary.

Despite these limitations, this study provides an empirical baseline for evaluating diagnostic reasoning in LLMs within psychiatry and highlights the need for evaluation frameworks that

extend beyond accuracy metrics to include clinician-grounded assessment of reasoning processes as LLMs continue to be explored for potential roles in clinical decision support.

Conclusions

In this mixed-methods evaluation of psychiatric case vignettes, large language models demonstrated moderate to high diagnostic accuracy and generated diagnostic reasoning that clinicians judged to be largely coherent, with infrequent safety concerns in this sample. Diagnostic accuracy was strongly associated with clinician-rated reasoning quality, underscoring that performance evaluation based on outputs alone may overlook clinically meaningful limitations in model reasoning. These findings support the need for clinician-grounded evaluation frameworks that assess both diagnostic outcomes and clinician-visible reasoning artifacts when considering large language models for potential roles in psychiatric decision support and other narrative-driven clinical domains.

Contributors

Conception or study design: KWJ, MSS, CP, SWY, BAZ, and HX. Funding acquisition: KWJ and HX. Figure and table creation: KWJ. Data curation: CN, JW, JSZ, MAG, ASH, MM, and JS. Fictitious vignette writing: JW, CN, ASH, JSZ, PC, and BAZ. Clinical reasoning evaluation: CIR, SS, WSM, CN, and PC. Illustrative human comparison: PC and JSZ. Acquisition, analysis, and interpretation of the data: KWJ. KWJ and QC have directly accessed and verified the underlying data reported in the manuscript. All authors drafted or reviewed the manuscript, approved the final version, and had responsibility for the decision to submit for publication.

Data sharing

Analysis code, evaluation prompts, and derived evaluation outputs will be publicly available in a repository upon publication at https://github.com/BIDS-Xu-Lab/psychiatry-frontier-llm-evaluation. Clinician-authored fictitious vignettes will also be publicly available. We will not publicly redistribute text derived from published case reports or verbatim model reasoning traces; citations to original sources will be provided, and access to restricted materials may be provided under controlled conditions (eg, to qualified researchers under a data-use agreement and/or institutional approval).

Declaration of interests

In the last 3 years, CIR has served as a consultant for Biohaven Pharmaceuticals and Osmind; and receives research grant support from Biohaven Pharmaceuticals, a stipend from American Psychiatric Association Publishing for her role as Deputy Editor at The American Journal of Psychiatry, a stipend from Springer Nature for her role as Deputy Editor for Neuropsychopharmacology, and book royalties from American Psychiatric Association Publishing. The remaining authors report no financial or other relationship relevant to the subject of this manuscript.

Acknowledgments

The authors thank Daniela Weng for creating Figure 1 and Qianqian Xie, Lingfei Qian, and Jeffrey Zhang for their advice on the study design. This study received material support in the form of API credits from the OpenAI Researcher Access Program and the Google Gemini Academic Program. KWJ is supported by the National Science Foundation's Graduate Research Fellowship Program and was formerly supported by the National Library of Medicine's T15

University-based Biomedical Informatics and Data Science Training Program. MM was supported by the National Institute of Mental Health under grant K23MH134068. JK is supported by the National Institutes of Health under grant K01MH137386. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Figures

Figure 1: Study workflow

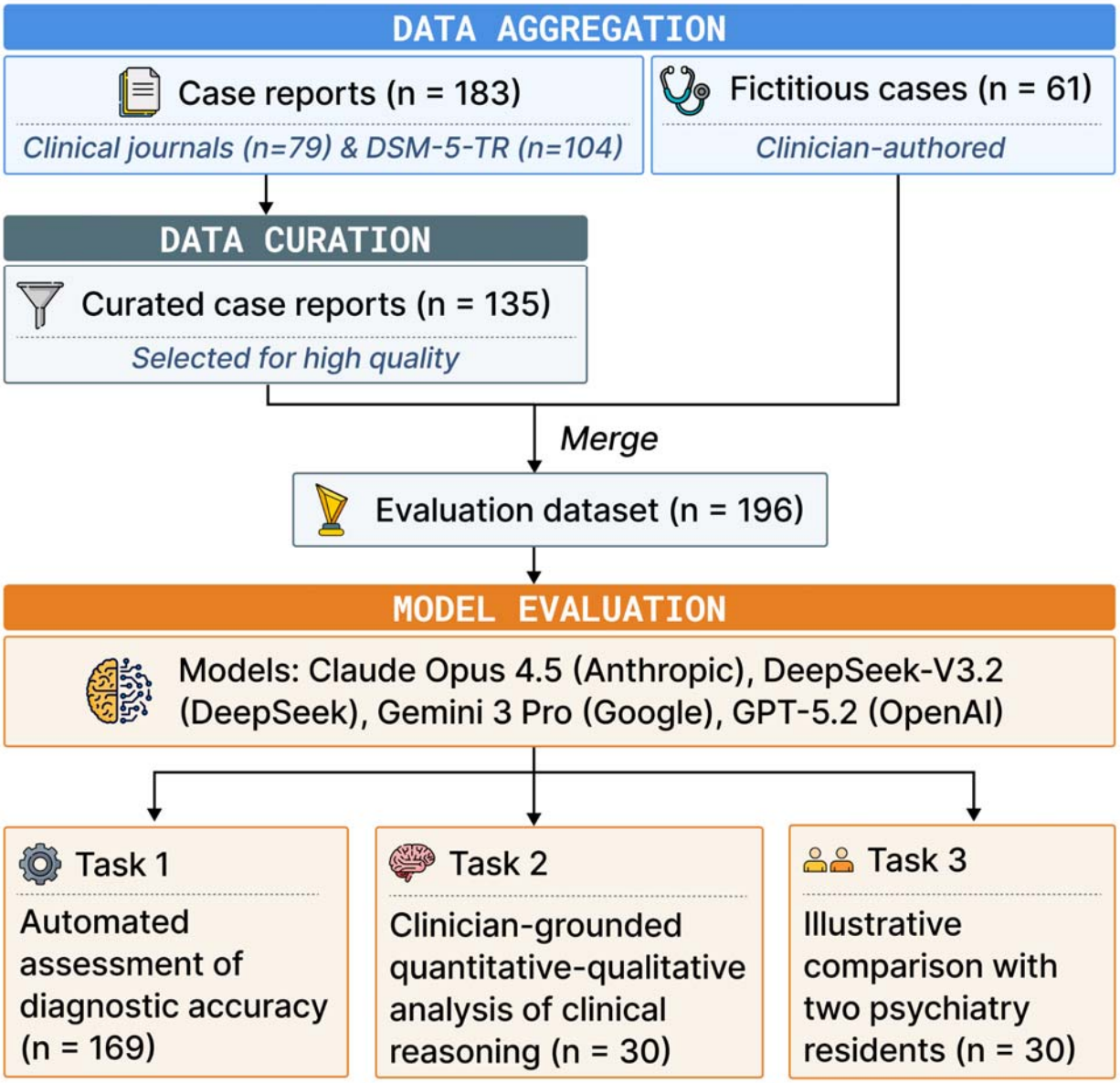
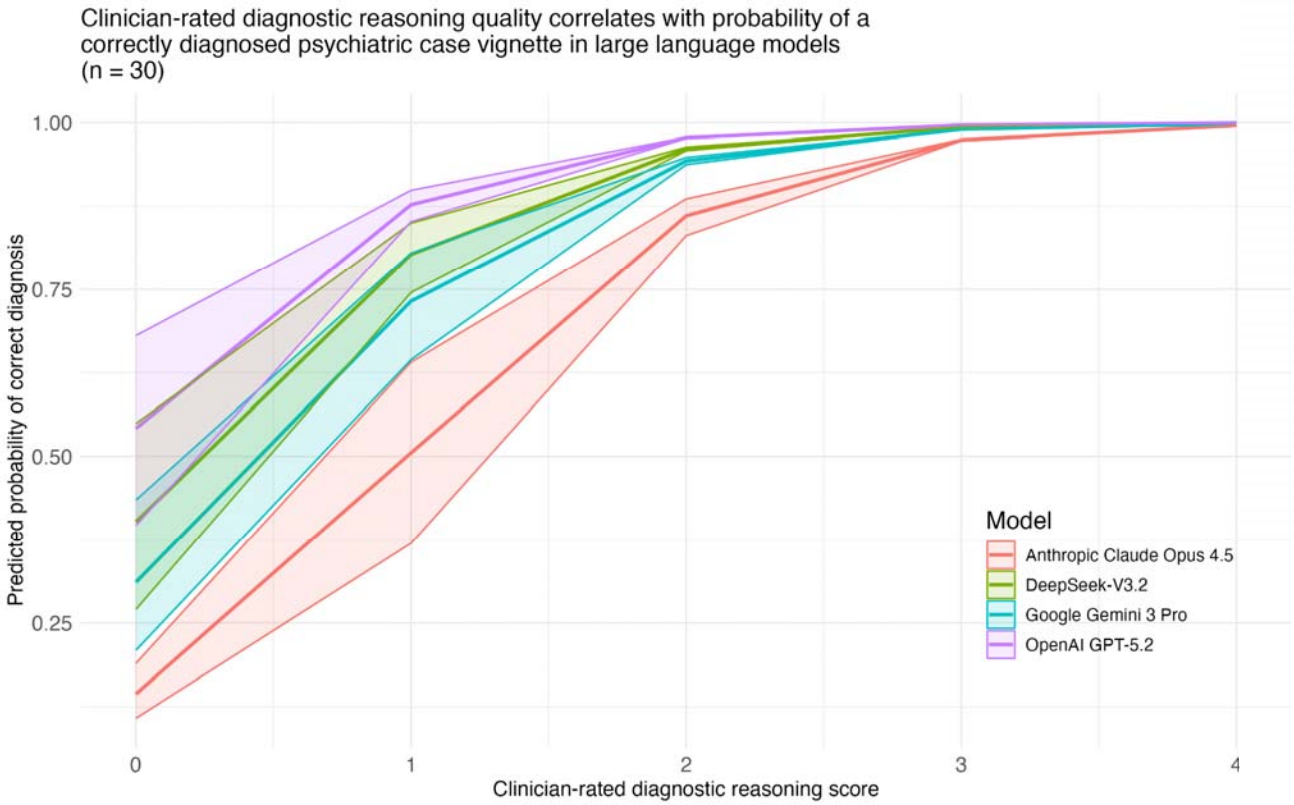


Figure 2: Predicted probability of diagnostic correctness given clinician-rated diagnostic reasoning quality, using mixed-effects logistic regression (*n* = 30)



Tables

Table 1: Psychiatric diagnostic accuracy of large language models (n = 196)

	Top-1 Accuracy	**Top-5 Accuracy**	**Recall@5**	**Mean Reciprocal Rank**
Claude Opus 4.5	0·638	0·801	0·731	0·710
DeepSeek-V3.2	0·566	0·770	0·691	0·657
GPT-5.2	0·607	0·735	0·684	0·665
Gemini 3 Pro	0·566	0·730	0·662	0·630

Table 2: Clinician-rated data extraction and diagnostic reasoning quality of large language models (n = 30)

	Clinician-Rated Scores (out of 4; higher is better)	
Model Name	**Data Extraction and Organization (mean ± SD)**	**Diagnostic Reasoning and Differential Diagnosis (mean ± SD)**
Claude Opus 4.5	3·67 ± 0·384	3·6 ± 0·446
Gemini 3 Pro	3·15 ± 0·687	3·12 ± 0·603
DeepSeek-V3.2	2·97 ± 0·613	2·9 ± 0·630
GPT-5.2	2·21 ± 0·501	2·37 ± 0·515

Table 3: Illustrative comparison of diagnostic accuracy between Claude Opus 4·5 and two psychiatry residents

	Top-1 Accuracy	**Top-5 Accuracy**	**Recall@5**	**Mean Reciprocal Rank**
Claude Opus 4.5	0·733	0·867	0·776	0·800
Psychiatry residents (n = 2)	0·617	0·767	0·8656	0·683

References

1 Over a billion people living with mental health conditions – services require urgent scale-up. https://www.who.int/news/item/02-09-2025-over-a-billion-people-living-with-mental-health-conditions-services-require-urgent-scale-up (accessed Nov 12, 2025).

2 Singhal K, Azizi S, Tu T, *et al.* Large language models encode clinical knowledge. *Nature* 2023; **620**: 172–80.

3 Shyr C, Cassini TA, Tinker RJ, *et al.* Large Language Models for Rare Disease Diagnosis at the Undiagnosed Diseases Network. *JAMA Netw Open* 2025; **8**: e2528538.

4 McDuff D, Schaekermann M, Tu T, *et al.* Towards accurate differential diagnosis with large language models. *Nature* 2025; : 1–7.

5 Cabral S, Restrepo D, Kanjee Z, *et al.* Clinical Reasoning of a Generative Artificial Intelligence Model Compared With Physicians. *JAMA Internal Medicine* 2024; **184**: 581–3.

6 Rubin R. Millions Turn to AI Chatbots for Mental Health Support. *JAMA* 2026; published online Jan 9. DOI:10.1001/jama.2025.23965.

7 Siddals S, Torous J, Coxon A. "It happened to be the perfect thing": experiences of generative AI chatbots for mental health. *npj Mental Health Res* 2024; **3**: 1–9.

8 Angus DC, Khera R, Lieu T, *et al.* AI, Health, and Health Care Today and Tomorrow: The JAMA Summit Report on Artificial Intelligence. *JAMA* 2025; **334**: 1650–64.

9 Sarma KV, Hanss KE, Halls AJM, *et al.* Integrating expert knowledge into large language models improves performance for psychiatric reasoning and diagnosis. *Psychiatry Research* 2026; **355**: 116844.

10 Gargari OK, Fatehi F, Mohammadi I, Firouzabadi SR, Shafiee A, Habibi G. Diagnostic accuracy of large language models in psychiatry. *Asian Journal of Psychiatry* 2024; **100**: 104168.

11 Fouda AE, Hassan AA, Hanafy RJ, Fouda ME. PsychiatryBench: A Multi-Task Benchmark for LLMs in Psychiatry. 2025; published online Nov 23. DOI:10.48550/arXiv.2509.09711.

12 Kim J, Leonte KG, Chen ML, *et al.* Large language models outperform mental and medical health care professionals in identifying obsessive-compulsive disorder. *npj Digit Med* 2024; **7**: 1–5.

13 Wang X, Xiong Z, Zou K, *et al.* Reasoning-driven large language models in medicine: opportunities, challenges, and the road ahead. *The Lancet Digital Health* 2026; **0**. DOI:10.1016/j.landig.2025.100931.

14 DeepSeek. DeepSeek-V3.2 Release | DeepSeek API Docs. 2025; published online Dec 1. https://api-docs.deepseek.com/news/news251201 (accessed Jan 8, 2026).

15 Google. Gemini 3 Pro. Google DeepMind. 2025. https://deepmind.google/models/gemini/pro/ (accessed Jan 8, 2026).

16 OpenAI. Introducing GPT-5.2. 2025; published online Dec 11. https://openai.com/index/introducing-gpt-5-2/ (accessed Jan 8, 2026).

17 Anthropic. Introducing Claude Opus 4.5. 2025; published online Nov 24. https://www.anthropic.com/news/claude-opus-4-5 (accessed Jan 8, 2026).

18 The Accreditation Council for Graduate Medical Education. Psychiatry Milestones. 2020; published online March. https://www.acgme.org/specialties/psychiatry/milestones/ (accessed Jan 8, 2026).

19 The Accreditation Council for Graduate Medical Education. Supplemental Guide: Psychiatry. 2020; published online March. https://www.acgme.org/globalassets/pdfs/milestones/psychiatrysupplementalguide.pdf (accessed Jan 8, 2026).

20 DeepSeek. The Temperature Parameter | DeepSeek API Docs. DeepSeek API Docs. 2025. https://api-docs.deepseek.com/quick_start/parameter_settings (accessed Jan 8, 2026).

21 Anthropic. Building with extended thinking. Claude Docs. 2025. https://platform.claude.com/docs/en/build-with-claude/extended-thinking (accessed Jan 8, 2026).

22 OpenAI. Using GPT-5.2 | OpenAI API. OpenAI Platform. 2025. https://platform.openai.com (accessed Jan 8, 2026).

23 Google. Gemini 3 Developer Guide | Gemini API. Google AI for Developers. 2025. https://ai.google.dev/gemini-api/docs/gemini-3 (accessed Jan 8, 2026).

24 Gu J, Jiang X, Shi Z, *et al.* A Survey on LLM-as-a-Judge. 2025; published online Oct 19. DOI:10.48550/arXiv.2411.15594.

25 Mondillo G, Abbate FG, Masino M, Colosimo S, Perrotta A, Frattolillo V. Closing the Pediatric Divide: A Performance Analysis of the GPT-5 Family in Medical Diagnostics. 2025; : 2025.08.28.25334657.

26 Wang S, Hu M, Li Q, Safari M, Yang X. Capabilities of GPT-5 on Multimodal Medical Reasoning. 2025; published online Aug 13. DOI:10.48550/arXiv.2508.08224.

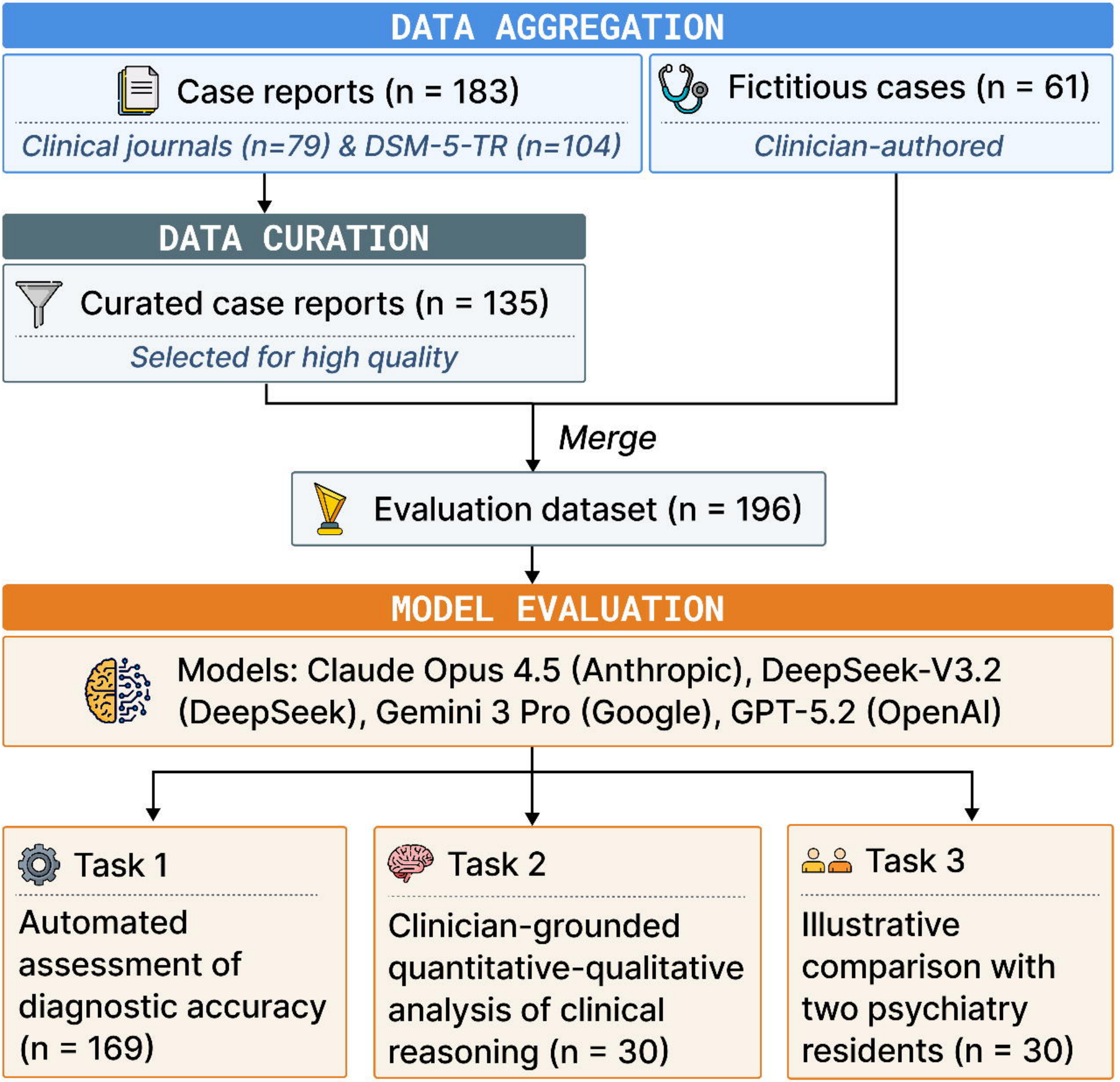
27 Google. Gemini thinking | Gemini API. Google AI for Developers. 2025. https://ai.google.dev/gemini-api/docs/thinking (accessed Jan 8, 2026).

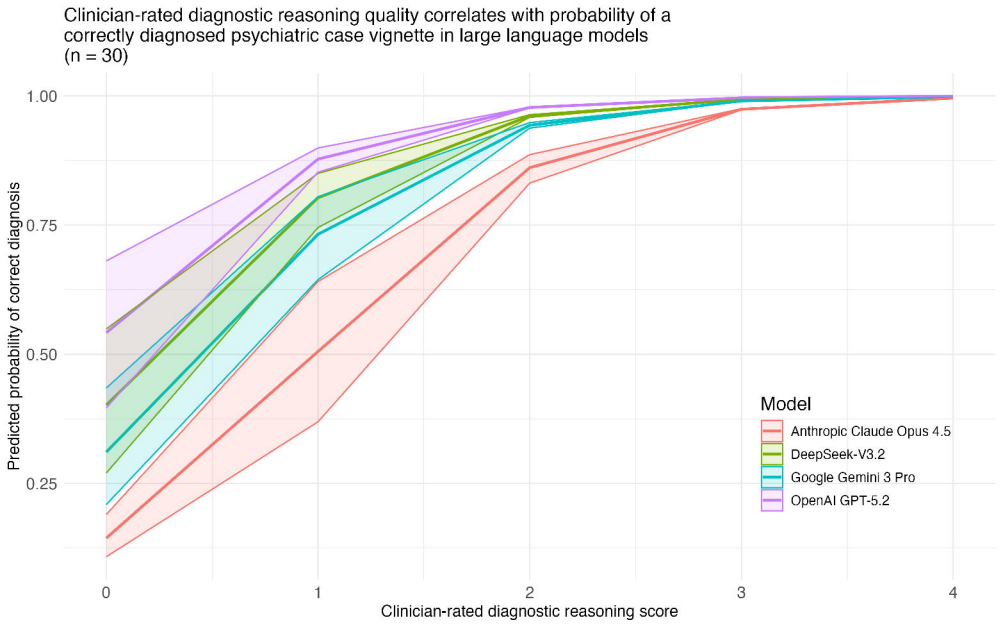
28 DeepSeek. Reasoning Model (deepseek-reasoner) | DeepSeek API Docs. 2025. https://api-docs.deepseek.com/guides/reasoning_model (accessed Jan 8, 2026).

29 OpenAI. Reasoning models | OpenAI API. 2025. https://platform.openai.com (accessed Jan 8,

2026).

30 Zheng L, Chiang W-L, Sheng Y, *et al.* Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. 2023. https://openreview.net/forum?id=uccHPGDlao (accessed Jan 29, 2026).


DATA AGGREGATION
Case reports (n = 183)
Clinical journals (n=79) & DSM-5-TR (n=104)
Fictitious cases (n = 61)
Clinician-authored
DATA CURATION
Curated case reports (n = 135)
Selected for high quality
Merge
Evaluation dataset (n = 196)
MODEL EVALUATION
Models: Claude Opus 4.5 (Anthropic), DeepSeek-V3.2 (DeepSeek), Gemini 3 Pro (Google), GPT-5.2 (OpenAI)
Task 1
Automated assessment of diagnostic accuracy (n = 169)
Task 2
Clinician-grounded quantitative-qualitative analysis of clinical reasoning (n = 30)
Task 3
Illustrative comparison with two psychiatry residents (n = 30)


Clinician-rated diagnostic reasoning quality correlates with probability of a correctly diagnosed psychiatric case vignette in large language models
(n = 30)
Predicted probability of correct diagnosis
1.00
0.75
0.50
0.25
0
1
2
3
4
Clinician-rated diagnostic reasoning score
Model
Anthropic Claude Opus 4.5
DeepSeek-V3.2
Google Gemini 3 Pro
OpenAI GPT-5.2