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

Kevin W. Jin, BS[1,2], Yasna Rostam-Abadi, MD, MPH[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 Hua Xu (hua.xu@yale.edu; 203-432-4771)

Word count: 3000

# Key Points

**Question:** Can multiple large language models accurately diagnose psychiatric conditions and generate diagnostic reasoning that clinicians judge as coherent, safe, and clinically meaningful?

**Findings:** Across 196 psychiatric case vignettes, four large language models demonstrated high diagnostic accuracy. In a clinician-evaluated subset of 30 vignettes, model diagnostic accuracy fell within the range observed for psychiatry residents. Clinicians judged model-generated diagnostic reasoning to be largely coherent and safe. Higher clinician-rated reasoning quality was strongly associated with diagnostic correctness, independent of data extraction quality.

**Meaning:** Evaluating diagnostic reasoning, in addition to accuracy, may be important when assessing large language models for potential clinical decision support in psychiatry.

# Abstract

**Importance:** Large language models (LLMs) have demonstrated diagnostic potential in several medical specialties, but their application to psychiatry – where diagnosis relies heavily on clinical judgment, narrative interpretation, and reasoning under uncertainty – remains insufficiently evaluated.

**Objective:** To evaluate diagnostic accuracy and clinician-judged reasoning quality of multiple large language models using psychiatric case vignettes.

**Design:** Mixed-methods evaluation study of diagnostic accuracy across four LLMs using 196 psychiatric case vignettes (135 published and 61 novel). Clinical reasoning quality was evaluated on a randomly selected subset of 30 vignettes using structured clinician ratings along two reasoning dimensions. The highest-performing model was illustratively compared with psychiatry trainees on the same subset. Diagnostic correctness for the full vignette set was assessed by a separate adjudicator LLM.

**Setting:** Publicly available model interfaces, December 2025.

**Participants:** Five board-certified psychiatrists evaluated model-generated clinical reasoning. Two psychiatry residents served as the illustrative human comparison.

**Main Outcomes and Measures:** Diagnostic accuracy and clinician-rated clinical reasoning quality. Diagnostic accuracy was assessed using 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 assessed using two 5-point Likert scales adapted from the American Council of Graduate Medical Education Psychiatry Residency Milestones, evaluating data extraction and diagnostic reasoning.

**Results:** Across 196 psychiatric case vignettes, Claude Opus 4.5 (Anthropic) achieved the highest diagnostic accuracy (top-1 accuracy, 0.638; top-5 accuracy, 0.801; recall@5, 0.731; mean reciprocal rank, 0.710) and clinician-rated reasoning scores. Higher clinician-rated diagnostic reasoning quality was strongly associated with diagnostic correctness in mixed-effects logistic regression analyses ($\beta = 1.80$; $p < 0.001$), corresponding to an approximately six-fold increase in odds of a correct diagnosis per 1-point increase in reasoning score. In an illustrative comparison, diagnostic accuracy of Claude Opus 4.5 fell within the range observed for psychiatry trainees.

**Conclusions and Relevance:** LLMs demonstrated high diagnostic accuracy and generated clinical reasoning that clinicians judged to be largely coherent and safe. Diagnostic reasoning quality was more strongly associated with diagnostic correctness than data extraction quality, underscoring the importance of evaluating reasoning alongside accuracy when assessing LLMs for clinical decision support in psychiatry.

# Main Text (3000 words)

## Introduction

The United States is experiencing a mental health crisis. In 2022, more than one in five individuals lived with at least one mental disorder, and nearly half received no treatment.[1,2] Among those who did, treatment costs exceeded $100 billion annually.[3] Demand for psychiatric services has increased while workforce capacity has not kept pace, with a projected deficit of more than 30,000 psychiatrists by 2036 and high rates of clinician burnout.[2,4,5] These pressures exacerbate a longstanding challenge in psychiatry: substantial variability in clinical judgment arising from diagnostic complexity and reliance on patient report.

Psychiatric diagnosis is grounded in clinicians' assessment of patient-reported symptoms against classificatory frameworks such as the Diagnostic and Statistical Manual of Mental Disorders and the International Classification of Diseases (ICD). Although this phenomenological approach has improved diagnostic reliability, the absence of definitive biomarkers means that diagnosis depends heavily on clinical interpretation. While this interpretative component is essential for capturing the complexity of psychiatric presentations, it may also contribute to diagnostic variability, particularly in time-constrained settings or when clinicians have differing levels of experience. Accordingly, there is growing interest in whether artificial intelligence (AI) systems might serve as diagnostic aids to support clinical decision-making. While clinical deployment of such systems remains limited, systematic evaluation of their diagnostic performance and reasoning is increasingly important.

Large language models (LLMs) have demonstrated strong performance across a range of medical tasks, including answering clinical questions and generating differential diagnoses in multiple specialties. Prior studies have shown that LLMs can perform comparably to clinicians in controlled diagnostic evaluations.[6–12] However, most existing evaluations have focused on non-psychiatric conditions, and relatively few studies have examined LLM performance in psychiatry.[13,14] This gap is notable given the widespread public use of LLMs for mental health-related queries, sometimes with harmful consequences, underscoring the need for clinician-grounded assessment of these systems.[15–21]

Existing psychiatric evaluations of LLMs have been limited by small sample sizes, narrow diagnostic scope, reliance on older models, or a lack of structured assessment of diagnostic reasoning.[19,22–25] To date, no study has comprehensively evaluated LLM diagnostic performance in psychiatry by integrating multiple accuracy metrics, clinician-rated reasoning quality, and human comparison.

Recent advances enabling LLMs to produce explicit reasoning rationales offer an opportunity to evaluate not only diagnostic accuracy but also the quality of model-generated diagnostic reasoning. We conducted a mixed-methods evaluation of four state-of-the-art LLMs using a diverse set of psychiatric case vignettes. Our study integrated automated, multi-metric assessment of diagnostic accuracy with a clinician-grounded evaluation of diagnostic reasoning across two clinically meaningful dimensions, alongside an illustrative comparison with psychiatry trainees. Together, these analyses aim to clarify how LLMs perform in psychiatric

diagnosis and to identify strengths and limitations relevant to their potential role as clinical decision support tools.

# 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 of diagnostic accuracy with two psychiatry residents (**Figure 1**). Case vignettes are commonly used to evaluate LLM diagnostic ability in other specialties.[9,26–29]

## 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 (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 standardized *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 three proprietary models - Claude Opus 4.5 (claude-opus-4-5-20251101, Anthropic), Gemini 3 Pro (gemini-3-pro-preview, Google), and GPT-5.2 (gpt-5.2-2025-12-11, OpenAI) - and one open-source model, DeepSeek-V3.2 (DeepSeek). Models were selected based on their state-of-the-art performance across a wide variety of benchmarks, their ability to generate explicit reasoning rationales alongside diagnostic outputs, and the accessibility of those rationales for clinical evaluation.[30–33]

Each model was prompted (i.e., provided a standardized diagnostic instruction) 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, which specified the model's role and response constraints, and a user prompt, which provided the vignette and diagnostic task. We

iteratively engineered prompts, crafting a system prompt that instructs the model to assume the role of a psychiatrist adhering to *DSM-5-TR* diagnostic criteria, and a user prompt that provides the vignette and asks the model to generate the corresponding list of diagnoses. All models were evaluated in a zero-shot setting, meaning they were asked to perform diagnostic tasks without being provided examples or additional training data.[34,35] We also provided clinical reasoning guidance adapted from the Accreditation Council for Graduate Medical Education (ACGME) Psychiatry Residency Milestones within the user prompt.[36,37] Prompts are reported in the Supplementary Material.

Model hyperparameters (fixed model configurations) were configured according to provider recommendations to elicit optimal reasoning behavior (Supplementary Material).[34,38–40] All models were accessed in December 2025.

## 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 prioritize the correct primary diagnosis within its differential, and top-5 accuracy gives the model's hit rate for any correct diagnoses given five chances, telling us if the model is headed in the right direction.[9] Recall@5 measures diagnostic coverage, assessing the percentage of correct diagnoses successfully predicted in the differential.[41] MRR measures the model's tendency to rank the primary

diagnosis towards the top of the list, which helps further assess the model's ability to prioritize the topmost diagnosis as the primary diagnosis.[42]

For all metrics, each of the diagnoses in the model's predicted differential is matched against the reference diagnoses by an LLM judge prompted to act as an independent medical adjudicator. LLM-based adjudication is a scalable alternative to human evaluators that has achieved over 80% agreement with humans, matching human-to-human agreement rates.[43] We used GPT-5-mini (gpt-5-mini-2025-08-07, OpenAI) as our adjudicator in this study based on its high concordance with human raters in medical adjudicative tasks.[44,45]

### 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 quantitative-qualitative 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 characterized 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;[36,37] 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 analyzed qualitatively to identify recurring reasoning patterns and failure modes (Supplementary Material). Non-boilerplate comments containing substantive critique were grouped using semantic similarity to facilitate manual thematic review. All themes were defined and assigned through direct human review of the underlying comments, and each comment was assigned a single primary theme reflecting its dominant qualitative feature. Theme prevalence was calculated by model, using non-boilerplate comments as the denominator.

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 (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.

## Results

### Task 1: Diagnostic Accuracy

**Table 1** summarizes 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. Claude Opus 4.5 (Anthropic) achieved the highest scores, while Gemini 3 Pro (Google) achieved the lowest.

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 between 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 $p > 0.05$; **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**). Claude Opus 4.5 (Anthropic) achieved the highest mean scores on both dimensions. GPT-5.2 (OpenAI) achieved the lowest mean scores on both dimensions. Substantial interrater agreement was observed for diagnostic match (Fleiss's $\kappa = 0.626$; average pairwise agreement = 0.767). Agreement for human-judged reasoning dimensions was higher, with good agreement for data extraction (ICC = 0.773; 95% CI: 0.702-0.831) and excellent agreement for diagnostic ability (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 ($\beta = 1.80$, $p < 0.001$), 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 ($p = 0.36$).

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

Qualitative observations were consistent with quantitative findings. Clinicians frequently identified reasoning deficiencies, such as post hoc justification, even in cases where 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. Unlike specialties with definitive biomarkers, psychiatric diagnosis depends on clinical judgment informed by symptom patterns, contextual factors, and differential diagnosis. 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 stringent testbed for evaluating diagnostic reasoning in LLMs.

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 diagnostic reasoning quality.[22–25,46] 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 reasoning summaries produced through publicly accessible interfaces.[34,35,47,48] 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 reasoning summaries are viable proxies for comparative evaluation of reasoning quality across models.

The illustrative comparison with psychiatry trainees should be interpreted cautiously and was not designed to establish equivalence or superiority but rather to contextualize 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.[49] 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, LLMs demonstrated moderate to high diagnostic accuracy and generated diagnostic reasoning that clinicians judged to be largely coherent and safe. We found that diagnostic accuracy was inextricably linked to the quality of clinical reasoning, suggesting that future models be evaluated not just on their outputs, but on the interpretability of their logic. These findings provide an empirical foundation for future clinician-grounded evaluations of LLMs applied to psychiatric diagnosis.

# 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 NIH (K01MH137386). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. (Disclosure): 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.

# Figures and 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 | 0.656 | 0.683 |

Figure 1: Study workflow

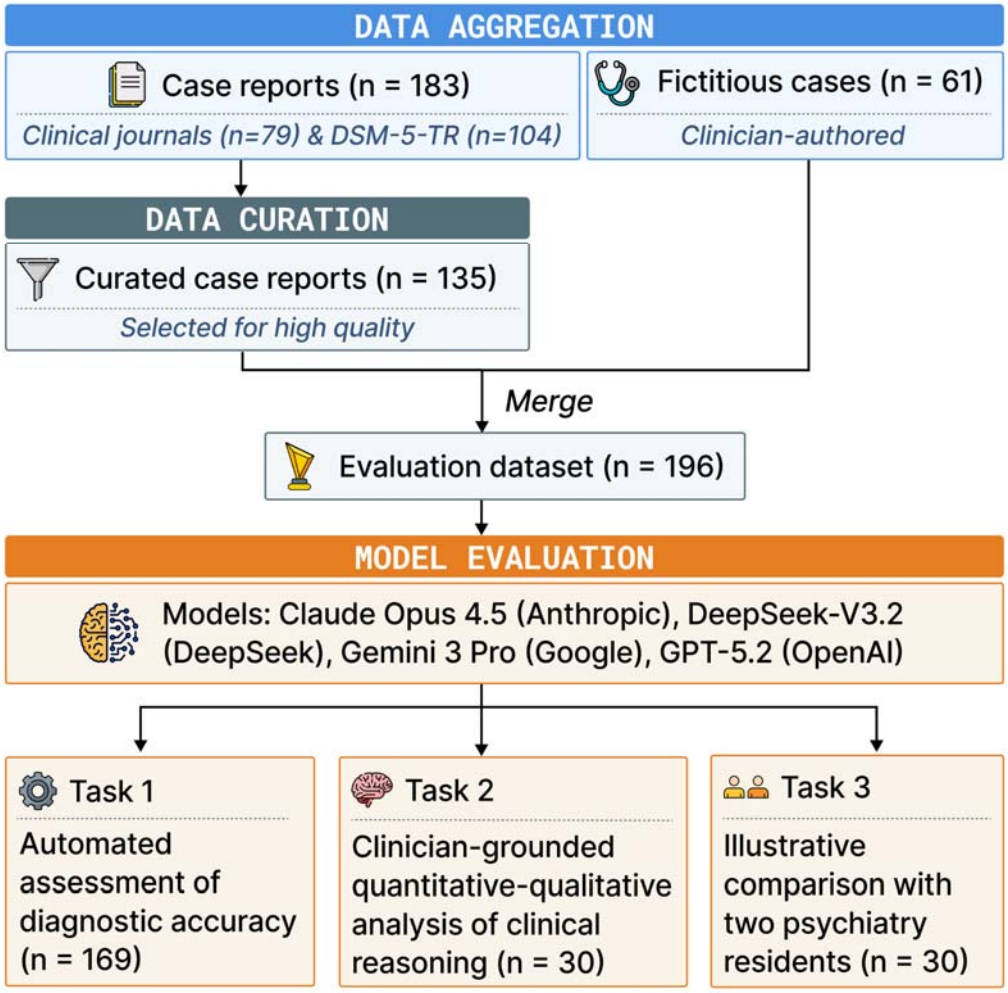
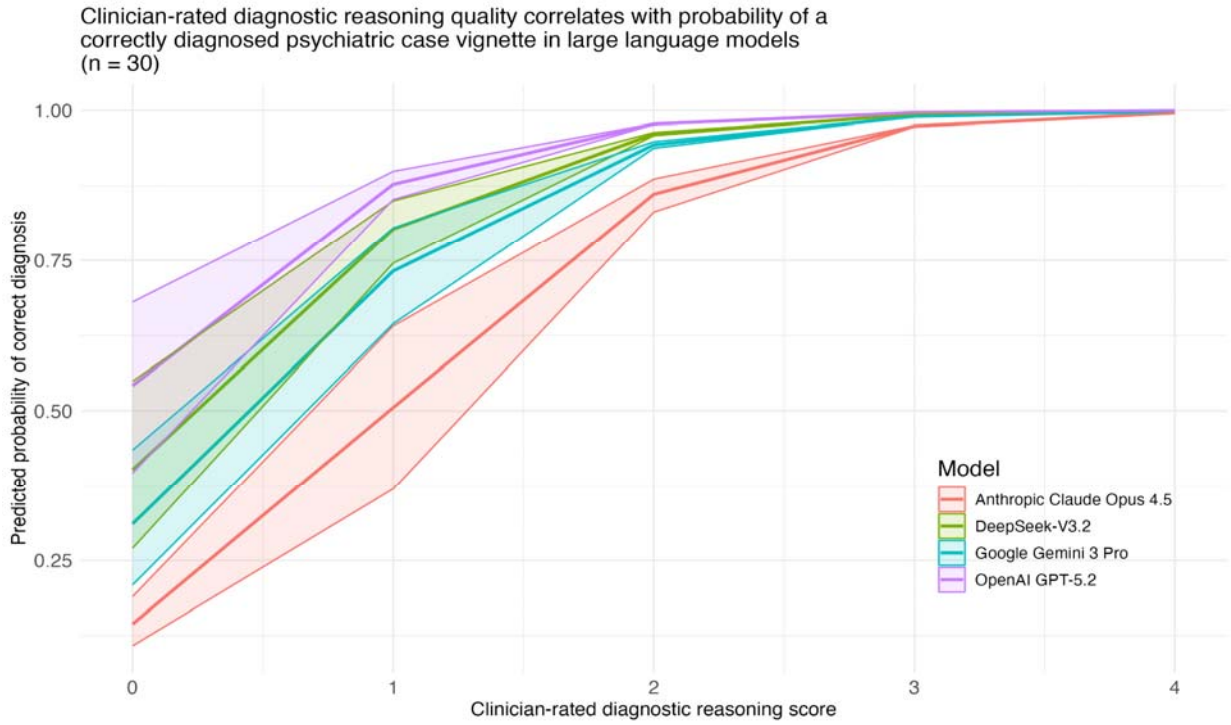


Figure 2: Predicted probability of diagnostic correctness given clinician-rated diagnostic reasoning quality, using mixed-effects logistic regression (*n* = 30)



# References


1. Mental Illness - National Institute of Mental Health (NIMH). Accessed November 12, 2025. https://www.nimh.nih.gov/health/statistics/mental-illness

2. National Center for Health Workforce Analysis. Behavioral Health Workforce, 2023. Bureau of Health Workforce. December 2023. Accessed December 18, 2025. https://bhw.hrsa.gov/data-research/review-health-workforce-research

3. Soni A. Healthcare Expenditures for Treatment of Mental Disorders: Estimates for Adults Ages 18 and Older, U.S. Civilian Noninstitutionalized Population, 2019. In: *Statistical Brief (Medical Expenditure Panel Survey (US))*. Agency for Healthcare Research and Quality (US); 2001. Accessed January 21, 2026. http://www.ncbi.nlm.nih.gov/books/NBK581181/

4. Summers RF, Gorrindo T, Hwang S, Aggarwal R, Guille C. Well-Being, Burnout, and Depression Among North American Psychiatrists: The State of Our Profession. *American Journal of Psychiatry*. Published online October 1, 2020. doi:10.1176/appi.ajp.2020.19090901

5. Summers R, Aggarwal R, Bailey T, et al. Psychiatrist Well-being and Burnout. *American Psychiatric Association*. Published online 2018.

6. Singhal K, Azizi S, Tu T, et al. Large language models encode clinical knowledge. *Nature*. 2023;620(7972):172-180. doi:10.1038/s41586-023-06291-2

7. Ayers JW, Poliak A, Dredze M, et al. Comparing Physician and Artificial Intelligence Chatbot Responses to Patient Questions Posted to a Public Social Media Forum. *JAMA Intern Med*. 2023;183(6):589-596. doi:10.1001/jamainternmed.2023.1838

8. 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(8):e2528538. doi:10.1001/jamanetworkopen.2025.28538

9. McDuff D, Schaekermann M, Tu T, et al. Towards accurate differential diagnosis with large language models. *Nature*. Published online April 9, 2025:1-7. doi:10.1038/s41586-025-08869-4

10. Nori H, Daswani M, Kelly C, et al. Sequential Diagnosis with Language Models. *arXiv*. Preprint posted online July 2, 2025:arXiv:2506.22405. doi:10.48550/arXiv.2506.22405

11. Cabral S, Restrepo D, Kanjee Z, et al. Clinical Reasoning of a Generative Artificial Intelligence Model Compared With Physicians. *JAMA Internal Medicine*. 2024;184(5):581-583. doi:10.1001/jamainternmed.2024.0295

12. Schaye V, Miller L, Kudlowitz D, et al. Development of a Clinical Reasoning Documentation Assessment Tool for Resident and Fellow Admission Notes: a Shared Mental Model for Feedback. *J GEN INTERN MED*. 2022;37(3):507-512. doi:10.1007/s11606-021-06805-6

13. Xie Q, Chen Q, Chen A, et al. Medical foundation large language models for comprehensive text analysis and beyond. *npj Digit Med*. 2025;8(1):1-10. doi:10.1038/s41746-025-01533-1

14. Liu X, Liu H, Yang G, et al. A generalist medical language model for disease diagnosis assistance. *Nat Med*. 2025;31(3):932-942. doi:10.1038/s41591-024-03416-6

15. Rubin R. Millions Turn to AI Chatbots for Mental Health Support. *JAMA*. Published online January 9, 2026. doi:10.1001/jama.2025.23965

16. Cunningham M. ChatGPT served as "suicide coach" in man's death, lawsuit alleges - CBS News. CBS News. January 15, 2026. Accessed January 15, 2026. https://www.cbsnews.com/news/chatgpt-lawsuit-colordo-man-suicide-openai-sam-altman/

17. Lavandera RK Allison Gordon, Ed. ChatGPT encouraged college graduate to commit suicide, family claims in lawsuit against OpenAI. CNN. November 20, 2025. Accessed January 15, 2026. https://www.cnn.com/2025/11/06/us/openai-chatgpt-suicide-lawsuit-invs-vis

18. Blease C, Worthen A, Torous J. Psychiatrists' experiences and opinions of generative artificial intelligence in mental healthcare: An online mixed methods survey. *Psychiatry Research*. 2024;333:115724. doi:10.1016/j.psychres.2024.115724

19. Hua Y, Na H, Li Z, et al. Applying and Evaluating Large Language Models in Mental Health Care: A Scoping Review of Human-Assessed Generative Tasks. *arXiv*. Preprint posted online August 21, 2024:arXiv:2408.11288. doi:10.48550/arXiv.2408.11288

20. 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):1-9. doi:10.1038/s44184-024-00097-4

21. 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(18):1650-1664. doi:10.1001/jama.2025.18490

22. 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. doi:10.1016/j.psychres.2025.116844

23. 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. doi:10.1016/j.ajp.2024.104168

24. Fouda AE, Hassan AA, Hanafy RJ, Fouda ME. PsychiatryBench: A Multi-Task Benchmark for LLMs in Psychiatry. *arXiv*. Preprint posted online November 23, 2025:arXiv:2509.09711. doi:10.48550/arXiv.2509.09711

25. 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):1-5. doi:10.1038/s41746-024-01181-x

26. Han T, Adams LC, Bressem KK, Busch F, Nebelung S, Truhn D. Comparative Analysis of Multimodal Large Language Model Performance on Clinical Vignette Questions. *JAMA*. 2024;331(15):1320-1321. doi:10.1001/jama.2023.27861

27. Xu L, Zhao W, Huang X. Diagnosis and Triage Performance of Contemporary Large Language Models on Short Clinical Vignettes. *J Med Syst*. 2025;49(1):141. doi:10.1007/s10916-025-02284-y

28. Zhang Z, Ma Y, Xiao Y, et al. Benchmarking large language models against clinicians across hospital levels in cardiovascular decision-making: a cross-sectional vignette-based study. *Sci Rep*. Published online December 15, 2025. doi:10.1038/s41598-025-32602-w

29. Johri S, Jeong J, Tran BA, et al. An evaluation framework for clinical use of large language models in patient interaction tasks. *Nat Med*. 2025;31(1):77-86. doi:10.1038/s41591-024-03328-5

30. DeepSeek. DeepSeek-V3.2 Release | DeepSeek API Docs. December 1, 2025. Accessed January 8, 2026. https://api-docs.deepseek.com/news/news251201

31. Google. Gemini 3 Pro. Google DeepMind. 2025. Accessed January 8, 2026. https://deepmind.google/models/gemini/pro/

32. OpenAI. Introducing GPT-5.2. December 11, 2025. Accessed January 8, 2026. https://openai.com/index/introducing-gpt-5-2/

33. Anthropic. Introducing Claude Opus 4.5. November 24, 2025. Accessed January 8, 2026. https://www.anthropic.com/news/claude-opus-4-5

34. Anthropic. Building with extended thinking. Claude Docs. 2025. Accessed January 8, 2026. https://platform.claude.com/docs/en/build-with-claude/extended-thinking

35. OpenAI. Reasoning models | OpenAI API. 2025. Accessed January 8, 2026. https://platform.openai.com

36. The Accreditation Council for Graduate Medical Education. Psychiatry Milestones. Published online March 2020. Accessed January 8, 2026. https://www.acgme.org/specialties/psychiatry/milestones/

37. The Accreditation Council for Graduate Medical Education. Supplemental Guide: Psychiatry. Published online March 2020. Accessed January 8, 2026. https://www.acgme.org/globalassets/pdfs/milestones/psychiatrysupplementalguide.pdf

38. DeepSeek. The Temperature Parameter | DeepSeek API Docs. DeepSeek API Docs. 2025. Accessed January 8, 2026. https://api-docs.deepseek.com/quick_start/parameter_settings

39. OpenAI. Using GPT-5.2 | OpenAI API. OpenAI Platform. 2025. Accessed January 8, 2026. https://platform.openai.com

40. Google. Gemini 3 Developer Guide | Gemini API. Google AI for Developers. 2025. Accessed January 8, 2026. https://ai.google.dev/gemini-api/docs/gemini-3

41. Ma MD, Wang X, Xiao Y, et al. Memorize and Rank: Elevating Large Language Models for Clinical Diagnosis Prediction. *Proceedings of the AAAI Conference on Artificial Intelligence*. 2025;39(23):24786-24794. doi:10.1609/aaai.v39i23.34660

42. Yu H, Chen T, Zhang X, et al. Performance of Large Language Models in Diagnosing Rare Hematologic Diseases and the Impact of Their Diagnostic Outputs on Physicians: Combined Retrospective and Prospective Study. *Journal of Medical Internet Research*. 2025;27(1):e77334. doi:10.2196/77334

43. Gu J, Jiang X, Shi Z, et al. A Survey on LLM-as-a-Judge. *arXiv*. Preprint posted online October 19, 2025:arXiv:2411.15594. doi:10.48550/arXiv.2411.15594

44. 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. *medRxiv*. Preprint posted online August 29, 2025:2025.08.28.25334657. doi:10.1101/2025.08.28.25334657

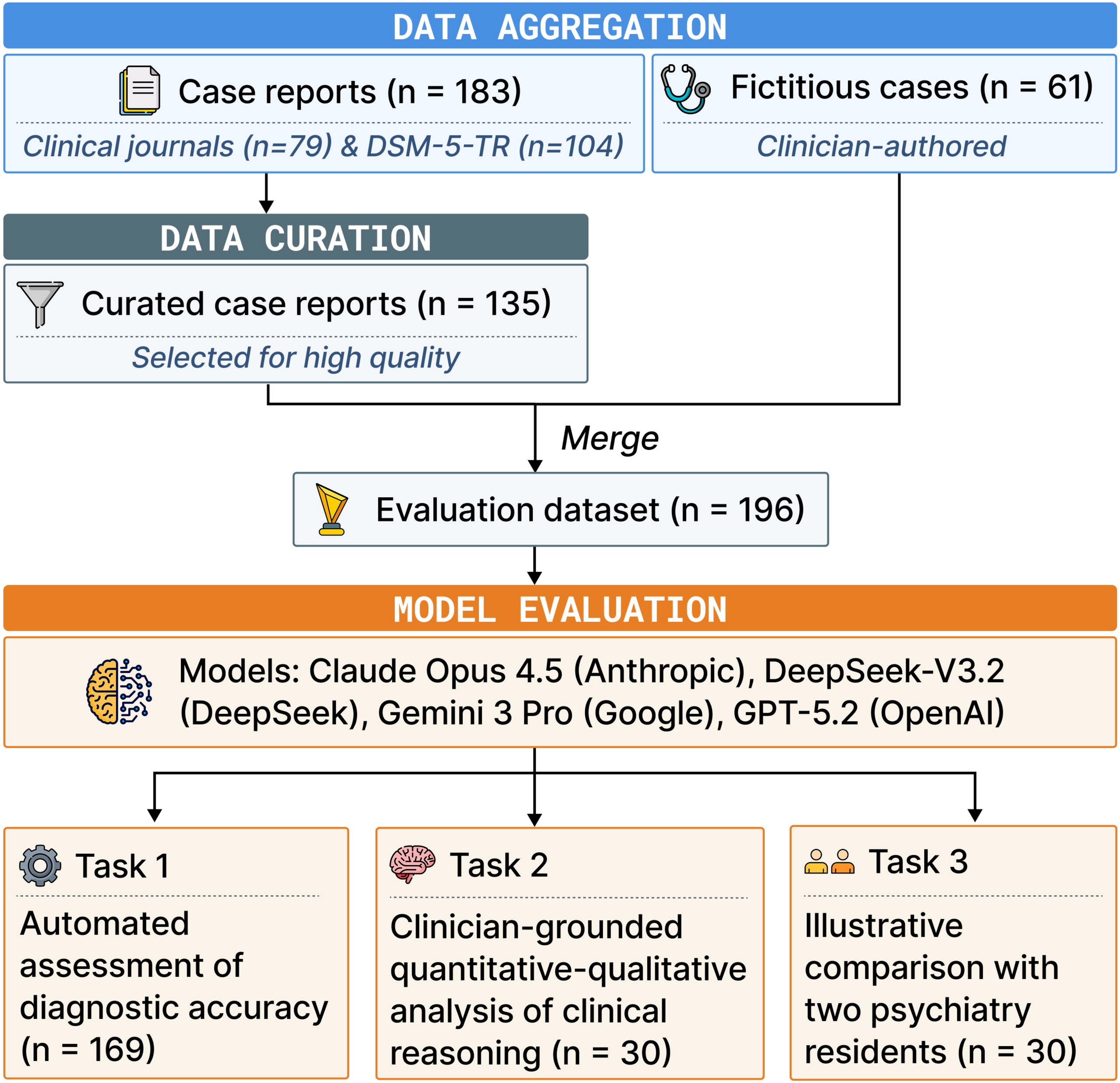
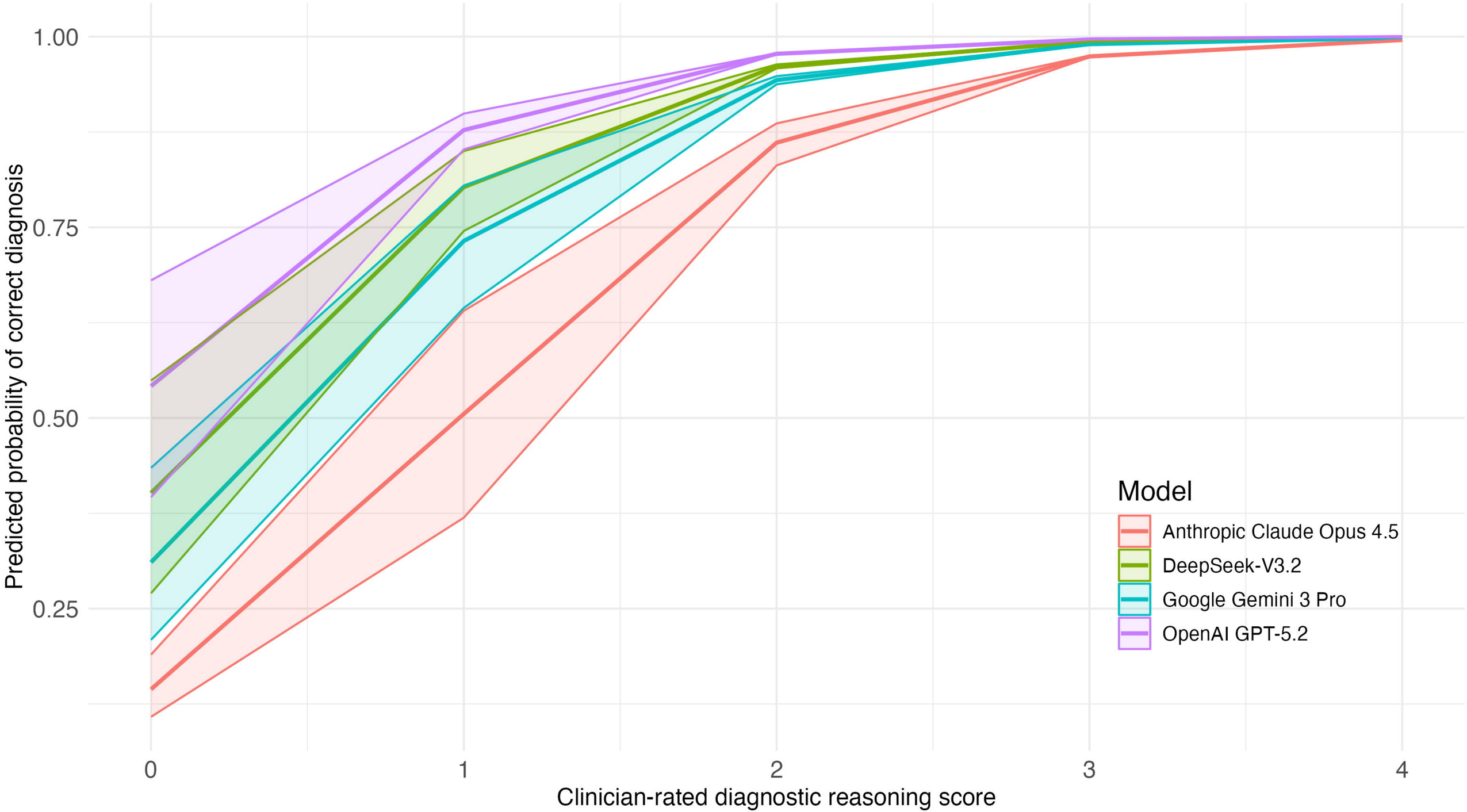
45. Wang S, Hu M, Li Q, Safari M, Yang X. Capabilities of GPT-5 on Multimodal Medical Reasoning. *arXiv*. Preprint posted online August 13, 2025:arXiv:2508.08224. doi:10.48550/arXiv.2508.08224

46. Hua Y, Liu F, Yang K, et al. Large Language Models in Mental Health Care: a Scoping Review. *arXiv*. Preprint posted online January 1, 2024:arXiv:2401.02984. doi:10.48550/arXiv.2401.02984

47. Google. Gemini thinking | Gemini API. Google AI for Developers. 2025. Accessed January 8, 2026. https://ai.google.dev/gemini-api/docs/thinking

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

49. Zheng L, Chiang WL, Sheng Y, et al. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. In: 2023. Accessed January 29, 2026. https://openreview.net/forum?id=uccHPGDlao


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)