

Diagnostic Accuracy and Alignment with Human Reader Responses across Five Large Language Models on Complex Clinical Case Challenges

Rino Bellocco^(1,2), Luca Soraci⁽³⁾, Lorenzo Lo Cicero⁽⁴⁾, Salvatore Silipigni⁽⁵⁾, Giuseppe Lanfranchi⁽⁶⁾, Savino Sciascia⁽⁷⁾, Wisit Cheungpasitporn⁽⁸⁾, Andrea Corsonello^(3,9), Domenico Santoro⁽⁴⁾, Guido Gembillo⁽⁴⁾

 Luca Soraci 0000-0002-0171-3358

(1) Department of Statistics and Quantitative Methods, University of Milano-Bicocca, Milan, Italy

(2) Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden

(3) Unit of Geriatric Medicine and Clinical Epidemiology, Italian National Research Center on Aging (IRCCS INRCA), Cosenza, Italy

(4) Unit of Nephrology, Department of Clinical and Experimental Medicine, University of Messina, Messina, Italy

(5) Radiology Unit, Biomorf Department, University of Messina, Messina, Italy

(6) MIFT Department, University of Messina, Messina, Italy

(7) University Center of Excellence on Nephrologic, Rheumatologic and Rare Diseases, ASL Città di Torino, San Giovanni Bosco Hub Hospital, Turin, Italy

(8) Division of Nephrology and Hypertension, Mayo Clinic, Rochester, Minnesota, USA

(9) Department of Pharmacy and Health and Nutritional Sciences, University of Calabria, 87036 Rende, Italy

CORRESPONDING AUTHOR: Luca Soraci, Unit of Geriatric Medicine and Clinical Epidemiology, Italian National Research Center on Aging (IRCCS INRCA), Cosenza, Italy. E-Mail: l.soraci@inrca.it

INTRODUCTION

Large language models (LLMs) match or exceed physician accuracy on benchmark diagnostic tasks. Whether higher accuracy reflects human-like diagnostic behavior, which bears on how safely clinicians can supervise these systems, is rarely assessed.

OBJECTIVES

To determine whether diagnostic accuracy and alignment with human reader responses move together across widely accessible LLMs under default, point-of-care conditions.

METHODS

This cross-sectional evaluation submitted the 29 most recent New England Journal of Medicine (NEJM) case challenges to 5 publicly accessible LLMs (GPT-5, Claude, Google Gemini, DeepSeek, and Grok) through each model's web interface (November 5-8, 2025) at default settings, one query per case, no system prompt. Each case offered a clinical vignette and 6 candidate diagnoses; published hu-

man reader response distributions (4386-17 714 per case; median, 7732) were the comparator. Diagnostic accuracy against the published gold-standard diagnosis, compared with the human reader majority by the exact McNemar test; agreement by Cohen κ ; alignment by a human-error similarity index (the share of model errors on cases the reader majority also missed) and by the Jensen-Shannon divergence (JSD) between each model's response and the reader distribution; and ordinary least squares regression of per-case JSD on normalized reader-disagreement entropy.

RESULTS

Reader majority accuracy was 48.3% (95% CI, 31.4-65.6). All 5 reached or exceeded this value, from 51.7% for DeepSeek (95% CI, 34.4-68.6) to 82.8% for Grok (95% CI, 65.5-92.4); Grok ($P = .02$), GPT-5 ($P = .004$), and Gemini ($P = .04$) significantly exceeded the reader majority, whereas Claude and DeepSeek did not. Grok, the most accurate model, showed the weakest alignment (similarity index, 0.40; mean JSD, 0.50 bits), with 3 of 5 errors on cases readers answered correctly. GPT-5 showed the strongest alignment (index, 1.000), failing only on cases that also challenged readers. For 4 of 5 models, divergence rose with reader-disa-

greement entropy; DeepSeek was the exception.

CONCLUSIONS

In this evaluation, diagnostic accuracy and alignment with human reader responses diverged across LLMs. Because a model can outperform clinicians in aggregate yet fail where oversight is least likely to detect it, evaluation of clinical AI should weigh error structure alongside accuracy.