Supplementary Table S1. Codebook and operational definitions	
Variable	**Definition/allowed values**
PMID	PubMed identifier (numeric).
Year	Publication year (YYYY).
Journal	Journal title as indexed in Web of Science.
Journal tier	SCI-E or ESCI (Web of Science-Emergency Medicine category).
Study type	Original research/review /commentary/other.
Model identification	Report of model name (e.g., GPT-4), release/version (e.g., 2024-04-09), and access path (API or web UI). Encoded as: full/partial/none.
Prompt disclosure	Availability of full prompt text or templates used in evaluation. encoded as: full/partial/none/unclear.
Input fidelity	Primary data realism. categories: real clinical (EHR, clinician notes, imaging reports); simulated (synthetic/vignette); exam proxy (standardized test/Q-bank); no input (free-form or general outputs).
Human expert benchmarking	Direct comparison against domain experts or adjudicated labels. Encoded as: Yes/No.
External validation	Evaluation on an independent dataset distinct from development/tuning material. Encoded as: Yes/No.
Calibration reporting	Any calibration metric reported: slope, intercept, Brier score, calibration plot. Encoded as: Yes/No; optional detail field for metric type.
Decision-curve analysis	Net benefit or decision-curve analysis reported. Encoded as: Yes/No.
Use case	Primary functional role: clinical decision support (CDS)/documentation–communication/education-training/other.

Supplementary Table 2. Comparison of Our Framework with TRIPOD-LLM and MI-CLEAR-LLM Guidelines			
Criteria	**MI-CLEAR-LLM**	**TRIPOD-LLM**	**Current Three-Axis Framework (This Study)**
Model version disclosure	Mandatory	Mandatory	Assessed; compliance rate analyzed
Prompt disclosure	Mandatory (full prompt)	Recommended	Assessed; full/partial/none coded and reported
Access pathway reporting	Mandatory	Recommended	Coded (e.g., web interface, API)
Use of real clinical data	Mandatory	Recommended	Core criterion; proportion reported
Human expert benchmarking	Recommended	Mandatory	Core criterion; proportion reported
Avoidance of synthetic tasks	Recommended	Recommended	Mandatory (only studies using real data and human benchmarking considered “rigorous”)
Functional use case classification	Not included	Not included	Comprehensive (e.g., decision support, education, documentation, etc.)
Empirical rigor scoring	Not included	Not included	Systematically coded across three axes
Open data/coding transparency	Recommended	Recommended	Included in assessment

MI-CLEAR-LLM and TRIPOD-LLM are recent reporting guidelines for large language model (LLM) studies in medicine. The current study’s three-axis framework extends these standards by systematically coding studies according to input fidelity, evaluation strategy, and functional use case, enabling more granular analysis of methodological rigor.
MI-CLEAR-LLM: MInimum reporting items for CLear Evaluation of Accuracy Reports of Large Language Models, TRIPOD-LLM: Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis Large Language Models