Journal of Health Informatics

J Health Inform. 2026;18:e1715
DOI: https://doi.org/10.59681/2175-4411.v18.2026.1715

ORIGINAL ARTICLE

Do large language models possess clinical reasoning? Perspectives from a real-world sepsis challenge

Grandes modelos de linguagem possuem raciocínio clínico? Perspectivas de um desafio sobre sepse no mundo real

¿Poseen razonamiento clínico los grandes modelos de lenguaje? Perspectivas a partir de un desafío sobre sepsis en el mundo real

Rildo Pinto da Silva[1] ✉, Antonio Pazin-Filho[1]

ORCID: 0000-0001-5718-2747 · 0000-0001-5242-329X

[1] Departamento de Clínica Médica, Faculdade de Medicina de Ribeirão Preto, Universidade de São Paulo (FMRP-USP), Ribeirão Preto (SP), Brasil.

ARTICLE INFORMATION

TYPE
Original Article

IDENTIFIER
e1715

DOI
10.59681/2175-4411.v18.2026.1715

RECEIVED
16 July 2026

ACCEPTED
19 November 2026

REVIEW PROCESS
Double Blind Review

HANDLING EDITOR
Juliano de Souza Gaspar

ACCESS
Open, CC BY 4.0

CORRESPONDING AUTHOR
Rildo Pinto da Silva
rildo.silva@alumni.usp.br

LICENSE

This is an open access article published under the Creative Commons CC BY 4.0 license, which permits use, distribution, and reproduction in any medium, provided the original work is properly cited.

CC BY 4.0

ABSTRACT

Background: Large Language Model (LLM) evaluations often rely on simplified datasets and metrics, masking deficits in complex clinical scenarios. **Objective:** To evaluate LLM limitations in sepsis classification using large-scale real-world data and determine, through robustness analysis, whether performance reflects genuine clinical reasoning. **Methods:** In a retrospective study of 65,523 hospitalizations from a tertiary emergency hospital, MedGemma was evaluated across administrative (ICD-Only), clinical (Sepsis-3-Only), and Non-Sepsis cohorts. Performance was assessed with the Matthews Correlation Coefficient (MCC) and sensitivity analysis; adversarial testing with clinically incoherent text probed reasoning. **Results:** A severity gradient confirmed Sepsis-3-Only cases as significantly more severe. Performance was inversely related to clinical complexity and dependent on the sepsis definition: sensitivity was higher for administrative codes than clinical cases (0.86 vs. 0.59; MCC 0.42 vs. 0.21). Inserting adversarial text triggered reclassification of 97.1% of cases, revealing keyword-matching behavior. Still, as a screening tool, the model identified 512 underreported sepsis cases explicitly from the Sepsis-3-Only cohort. Conservatively, this yields 1 confirmed true positive per 10.5 flagged cases, outperforming existing alerts. **Conclusion:** LLMs lack the clinical reasoning needed for autonomous sepsis diagnosis, but add value in human-in-the-loop roles such as retrospective record quality auditing rather than high-stakes autonomous decisions.

Keywords: Sepsis; Natural Language Processing; Clinical Reasoning.

RESUMO

Contexto: Avaliações de Grandes Modelos de Linguagem (LLMs) se apoiam em dados e métricas simplificados, mascarando deficiências em cenários clínicos complexos. **Objetivo:** Avaliar limitações de LLMs na classificação de sepse utilizando dados do mundo real e determinar se o desempenho reflete raciocínio clínico. **Métodos:** Estudo retrospectivo de 65.523 internações avaliando o modelo MedGemma em coortes administrativa (ICD-Only), clínica (Sepsis-3-Only) e Não-Sepse (Non-Sepsis). O desempenho foi aferido pelo Coeficiente de Correlação de Matthews (MCC) e por análise de sensibilidade. Testes adversariais investigaram o raciocínio do modelo. **Resultados:** O desempenho mostrou-se inversamente relacionado à complexidade clínica e dependente da definição de sepse: a sensibilidade foi maior na coorte administrativa que na clínica (0,86 vs. 0,59; MCC 0,42 vs. 0,21). A inserção de texto adversarial provocou a reclassificação de 97,1% dos casos, revelando comportamento de correspondência por palavras-chave. Contudo, como ferramenta de triagem, o modelo identificou 512 casos subnotificados (1 verdadeiro positivo a cada 10,5 alertas), superando os sistemas vigentes. **Conclusão:** LLMs carecem do raciocínio clínico necessário para o diagnóstico autônomo de sepse, mas agregam valor em aplicações com supervisão humana como a auditoria retrospectiva da qualidade dos registros, em vez de decisões autônomas de alto risco.

Descritores: Sepse; Processamento de Linguagem Natural; Raciocínio Clínico.

RESUMEN

Antecedentes: Las evaluaciones de Grandes Modelos de Lenguaje (LLM) se basan en datos simplificados, ocultando deficiencias en escenarios clínicos complejos. **Objetivo:** Evaluar las limitaciones de LLMs al clasificar sepsis con datos reales y determinar si su rendimiento refleja un razonamiento clínico. **Métodos:** Estudio retrospectivo de 65.523 hospitalizaciones, evaluando MedGemma en las cohortes administrativa (ICD-Only), clínica (Sepsis-3-Only) y No-Sepsis (Non-Sepsis). El rendimiento se midió mediante Coeficiente de Correlación de Matthews (MCC) y análisis de sensibilidad. Pruebas adversarias investigaron el razonamiento del modelo. **Resultados:** El rendimiento fue inversamente proporcional a la complejidad clínica y dependió de la definición de sepsis: la sensibilidad fue mayor en la cohorte administrativa que en la clínica (0,86 vs. 0,59; MCC 0,42 vs. 0,21). Insertar texto adversarial reclasificó el 97,1% de los casos, evidenciando correspondencia por palabras clave. Sin embargo, como triaje, identificó 512 casos subnotificados (1 verdadero positivo cada 10,5 alertas), superando los sistemas actuales. **Conclusión:** Los LLM carecen del razonamiento clínico necesario para el diagnóstico autónomo de sepsis, pero aportan valor con supervisión humana, en la auditoría retrospectiva de registros, en lugar de decisiones autónomas de alto riesgo.

Descriptores: Sepsis; Procesamiento de Lenguaje Natural; Razonamiento Clínico.

HOW TO CITE THIS ARTICLE: Pinto da Silva R, Pazin-Filho A. Do large language models possess clinical reasoning? Perspectives from a real-world sepsis challenge. J Health Inform 2026;18:e1715. doi: https://doi.org/10.59681/2175-4411.v18.2026.1715

1 INTRODUCTION

Electronic health records (EHR) are a crucial source of health data. Researchers have postulated that Natural Language Processing (NLP) is the ideal tool for analyzing this unstructured data, offering the potential to optimize healthcare delivery, medical reasoning, patient education, and clinical outcomes[1–3].

However, this premise has faced significant criticism. Major concerns include: evaluating models on data that do not reflect real-world patient care[4], inadequate assessment of hallucinations[5], poor quality of training data[6], reliance on retrospective data without external validation[7] and lack of standardized outcome measures[8]. Additional critiques highlight the limited scope of practical healthcare applications[9], underrepresentation of medical specialties[10], and low complexity of the evaluated tasks[11]. Consequently, authors advocate for the establishment of best practices in research regarding the application of large language models (LLMs) in healthcare[12].

Amidst this complex landscape, an emerging research frontier investigates the clinical reasoning capabilities of LLMs, primarily benchmarking performance against medical licensing exams and question-answering datasets[13,14]. However, this methodological reliance has drawn criticism[15].

Within this context, sepsis serves as a critical case study. Sepsis is associated with substantial mortality, although estimates vary according to case definition, severity, and care setting. A systematic review of hospital-treated adults estimated pooled in-hospital mortality at 26.7% (95% CI, 22.9–30.7%); among ICU-treated cases, the estimate was 41.9% (95% CI, 36.2–47.7%)[16]. The Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3 Consensus) highlight the diagnostic challenges posed by the diverse manifestations of sepsis[17]. Underreporting further complicates accurate estimation of incidence, morbidity, and mortality[16], thereby hindering effective surveillance, policy development, and quality improvement[18].

While classifying hospitalized patients with suspected sepsis using real-world data presents a complex challenge, it also offers an opportunity to elevate validation standards. By integrating established healthcare metrics with rigorous robustness testing, we can move beyond superficial evaluation towards a deeper assessment of model reliability.

This study aims to identify the limitations and failure modes of applying LLMs to clinical sepsis classification, utilizing real-world data from complex cases. To evaluate performance, we employed established healthcare metrics alongside a robustness analysis to determine whether the model relies on genuine clinical reasoning or keyword association. This study seeks to demonstrate the limitations of deploying LLMs without specialized human supervision and to contribute to future research in applied LLM.

2 METHODS

The study reporting adheres to the TRIPOD+LLM statement, and the corresponding checklist was completed to ensure comprehensive and transparent reporting of the model evaluation[19].

2.1 Design and settings

This retrospective, observational study utilized secondary data from the Emergency Unit of the Hospital das Clínicas, Ribeirão Preto Medical School, University of São Paulo (HCFMRP-USP), Ribeirão Preto, Brazil, a single hospital specializing in emergency medicine. This 190-bed facility, part of a tertiary hospital complex, serves as the main referral center.

From an initial dataset of 1,009,055 encounters (64,525 patients) collected between July 2013 and July 2024, we included only inpatient hospitalizations with physician-authored discharge summaries in Brazilian Portuguese. All clinical data were documented in a custom-developed Electronic Health Record (EHR) system. This system utilizes the International Classification of Diseases, Tenth Revision (ICD-10), for diagnosis coding, which is maintained and validated by an internal auditing team.

2.2 Datasets

We created four datasets for this study.

a – Sepsis ICD-10 dataset: Comprised 1,940 hospitalizations (1,757 patients) with explicit sepsis diagnoses documented in discharge summaries via ICD-10 codes.

b – Non-Sepsis ICD-10 dataset: Initially comprised all remaining hospitalizations. To ensure patient-level exclusivity and prevent confounding, all 1,757 patients from the Sepsis ICD-10 dataset were excluded. The final dataset contained 63,583 hospitalizations (35,765 patients).

c – Organ dysfunction criterion (partial SOFA): To operationalize the Sepsis-3 consensus, we first identified hospitalizations of patients with organ dysfunction, calculating a partial Sequential Organ Failure Assessment (SOFA) score as follows:

Data Extraction and Cleaning: We extracted structured data from serum creatinine (mg/dL), platelet count ($10^3/\mu L$), total bilirubin (mg/dL), mean arterial pressure (MAP; mmHg), and Glasgow Coma Scale (GCS). The PaO_2/FiO_2 ratio and vasopressor use (e.g., dobutamine, norepinephrine) were excluded due to the lack of structured data; to preserve validity, we avoided extraction from unstructured text, acknowledging potential score underestimation. A physician reviewed

outliers, a MAP ceiling of 146 mmHg was applied, and GCS scores were restricted to the valid range of 3–15.

As the missing components could only add points to each daily score, hospitalizations meeting the criterion had documented dysfunction in the measured organ systems, whereas sepsis with predominantly respiratory or vasopressor-dependent dysfunction may not have been detected and may have been assigned to the ICD-Only or Non-Sepsis cohorts.

Score calculation: Data recorded at the minute/hour level were aggregated daily, using the mode (defaulting to the lowest value for ties). Components were scored using standard SOFA cut-offs (MAP < 70 mmHg scored as 1 point), and the daily partial SOFA was the sum of available components, with unmeasured components scored as zero. Inclusion criteria: baseline was defined as the lowest daily partial SOFA during the hospitalization, and the dataset included any hospitalization with an increase of two or more points from baseline to a subsequent peak, at any time during the stay.

d – Positive Infection Criterion (confirmed cultures): Identified using positive culture results derived exclusively from structured data. We required microbiological confirmation, rather than suspected infection (cultures sampled plus antibiotics), to increase the certainty of an active infection coexisting with organ dysfunction, thereby approximating the Sepsis-3 construct with higher specificity at the cost of excluding culture-negative sepsis. To maintain criterion integrity, we did not extract suspected infection terms from unstructured text. The dataset included positive results for urine (183 tests; 129 hospitalizations), blood (3,448 tests; 2,394 hospitalizations), and other specimens (e.g., catheter tips; 431 tests; 356 hospitalizations).

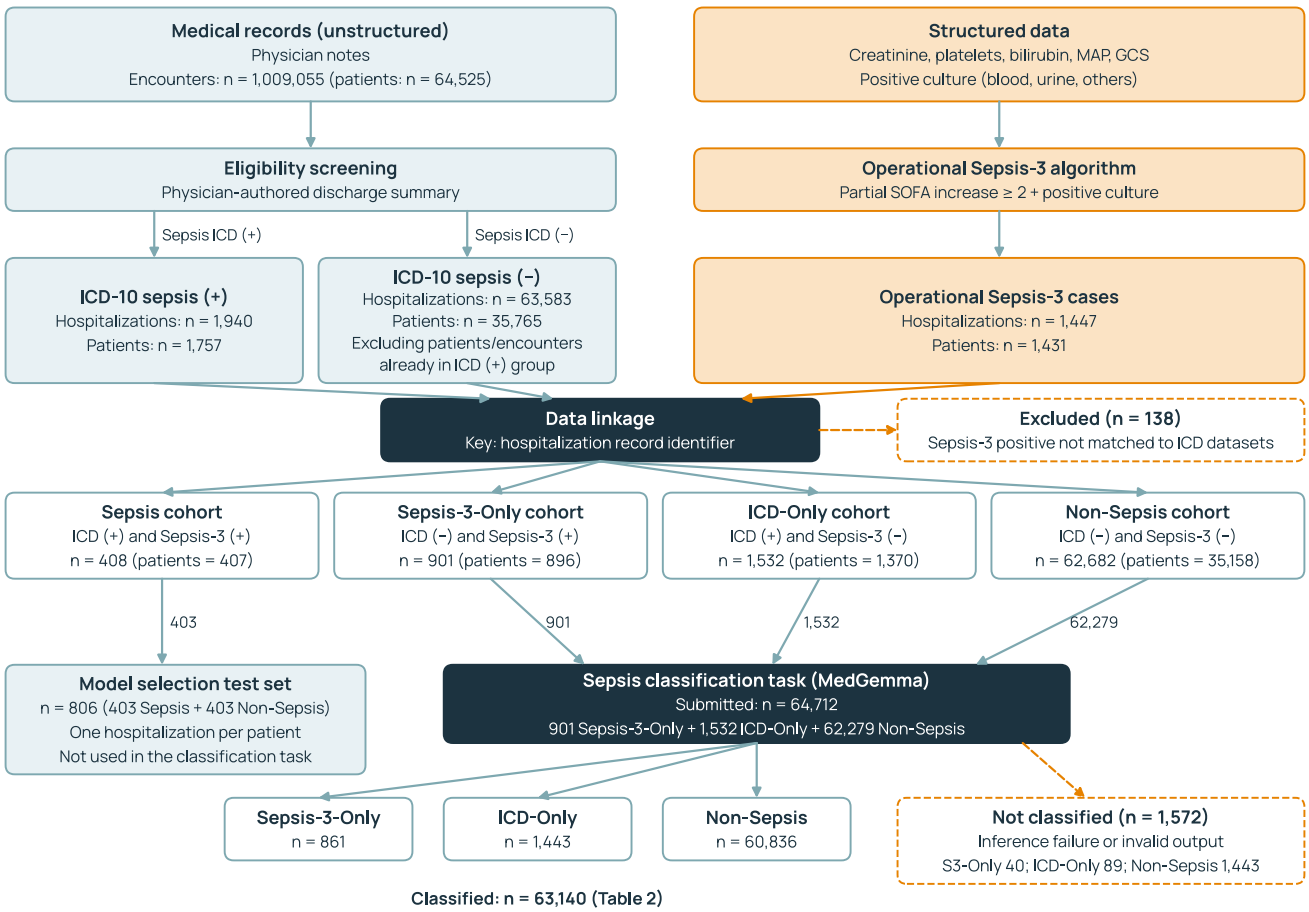
Operational Definition of Sepsis-3: Hospitalizations meeting both organ dysfunction (partial SOFA) and positive infection criteria were classified as "operational Sepsis-3 cases group" (1,447 hospitalizations; 1,431 patients). Of these, 138 were not matched to the eligible ICD-10 datasets, leaving 1,309 operational Sepsis-3 cases (1,303 patients) for cohort assignment.

2.3 Study population cohorts

Hospitalization record identifiers were used to link datasets, creating four mutually exclusive cohorts:

a – Sepsis cohort: The intersection group, composed of hospitalizations present in both the Sepsis ICD-10 dataset and our operational Sepsis-3 cases group (408

Figure 1. Selection flowchart



Source: Prepared by the authors (2026).

hospitalizations; 407 patients).

b – ICD-Only cohort: Hospitalizations from the Sepsis ICD-10 dataset that were not present in the Sepsis-3 operational cases group (1,532 hospitalizations; 1,370 patients).

c – Sepsis-3-Only cohort: Hospitalizations from the Sepsis-3 operational cases group that were not present in the Sepsis ICD-10 dataset (901 hospitalizations; 896 patients).

d – Non-Sepsis cohort: Hospitalizations meeting neither criterion (62,682 hospitalizations; 35,158 patients). The selection flowchart is detailed in Figure 1.

2.4 Data extraction and processing

We extracted the following variables: length of stay in days, number of specialties involved in the treatment (specialties), intensive care unit (ICU) admission, palliative care utilization and in-hospital mortality. The Charlson Comorbidity Index was calculated based on the ICD codes using the *comorbidipy* Python library. To reflect the conditions of real-world LLM application, missing data imputation was not performed. Text was normalized (lowercasing, removal of special characters and redundant whitespace), and non-ambiguous medical abbreviations were expanded using a custom dictionary[20] ; extracted variables were linked to the cohorts by hospitalization identifier.

2.5 Modeling

The initial pipeline stage involved selecting an optimal prompt using a development dataset and the Llama model (3.1-8B-Instruct)[21], selected for its open-access nature and documented performance[22]. The prompt was optimized for sensitivity using a development dataset of 200 randomly selected hospitalizations, comprising 100 from each of the Sepsis and Non-Sepsis cohorts.

The optimized one-shot prompt was subsequently evaluated across six distinct models (Box 1). This comparative analysis used a test dataset of 806 hospitalizations. Sepsis cases (n=403) were obtained by selecting, for each patient with a sepsis ICD-10 code, the hospitalization with the longest discharge summary and retaining those meeting the operational Sepsis-3 criteria. Non-Sepsis controls (n=403) were a random sample stratified by in-hospital death, drawn from Non-Sepsis hospitalizations with complete length-of-stay and outcome data, one hospitalization per patient (the one with the longest discharge summary).

To manage computational constraints, the models were loaded in a 4-bit quantized format using the *bitsandbytes* library with *bfloat16* computation data type. Context parameters were set to a maximum input length of 6,000 tokens and a maximum output length of 40 tokens. MedGemma 27b-text-it (MedGemma)[23] demonstrated the highest Matthews Correlation Coefficient (MCC), indicating a superior balance between sepsis and non-sepsis classification[24]. Consequently, it was selected for the sepsis clinical classification task. MedGemma is an instruction-tuned LLM based on Gemma architecture and specialized for the medical domain.

2.6 Sepsis clinical classification task

For this task, the ICD-Only, Sepsis-3-Only and Non-Sepsis cohorts were used, excluding the 403 Non-Sepsis hospitalizations of the test dataset (the Sepsis cohort was not used), which left 64,712 hospitalizations. Their processed discharge summaries were classified by MedGemma, using the same one-shot prompt, as sepsis or non-sepsis.

Box 1. Optimized one-shot prompt used across the models

Clinical Text Example: ```1 respiratory infection by coronavirus symptom onset 06/07 dyspnea, myalgia, hyporexia worsening of symptoms 06/09 drop in oxygen saturation positive rtpcr on 06/09 pao2/fio2 ratio on admission 174 rt pcr negative on 06/26 2 tracheostomy on 06/30 schedule procedure for first cannula change with head and neck surgery in 10 days extubation failure on 06/16 extubation on 06/19 with need for reintubation on 06/23 tracheostomy on 06/30 3 candidemia blood culture and catheter tip culture positive for candida albicans multi s on 06/23 echocardiogram 06/29 no signs of vegetation micafungin from 06/27 to 06/30 fluconazole d0 06/30 expected 14 days after negative blood culture 4 hf profile b? discarded ntprobnp 06/09 368 5 acute kidney injury kdigo 1 resolved increase of 0.6 in creatinine secondary to dehydration? furosemide administered considering a diagnosis of hf profile b, later discarded cr 1.07 1.82 1.12 0.99 06/11 0.73 06/14 0.67 06/17 0.5 06/9 6 hydroelectrolytic disorder hypokalemia resolved magnesium replacement administered from 06/25 to 07/03 7 previous urinary tract infection treated treatment with ceftriaxone between 05/18 and 05/24 in a primary care unit. admission {omitted}, discharge {omitted}, stay 25 d 10 h, bed {omitted}, insurance sus. patient admitted to ward on {omitted}, post-discharge from the coronavirus intensive care unit, already outside the coronavirus transmission period. remained hemodynamically stable, ecg 11t. remained two days without bipap use however, on 07/06 she developed drowsiness, with hypercapnia seen on blood gas test. patient connected to bipap and quetiapine suspended. patient's drowsiness improved, returning to a baseline glasgow of 11t vital signs blood pressure 132 x 80 heart rate 80 beats per minute oxygen saturation 100% at 3.5 l/min via tracheostomy respiratory rate 16 breaths per minute regular general state, pale 1/4, acyanotic, anicteric, miotic isocoric pupils photoreactive. glasgow 11t cardiac auscultation 2 regular normophonetic sounds without murmurs respiratory auscultation bilateral vesicular murmur with sparse rhonchi. no signs of respiratory distress. Abdomen semiglobose, normoactive hydroaerial sounds, flaccid, apparently painless, without masses or visceromegaly, tympanic extremities without lower limb edema, full and symmetric pulses, capillary refill time less than 3 seconds, without signs of deep vein thrombosis. attachments in the system. not applicable. patient with symptom onset on 6/7, already has a negative pcr on 06/26, remaining afebrile thus, according to institutional protocol, patient is in condition to be removed from coronavirus isolation plan to keep patient on bipap during the night with progressive weaning of oxygen fluconazole maintained for treatment of candidemia planned treatment of 14 days post-negative blood culture dose increase to 200 mg q8h schedule procedure for head and neck surgery for tracheostomy cannula change on 07/10. patient in regular general condition.

Source: Prepared by the authors (2026).

2.7 Statistical analysis

Statistical analyses were performed using IBM SPSS (v29.0). Categorical and continuous variables were compared using Pearson Chi-square and Kruskal-Wallis tests, respectively, with significance set at $p < 0.05$. Model performance was assessed via sensitivity, specificity, PPV, NPV, F1-score, accuracy, and MCC; 95% confidence intervals were generated using bootstrapping (1,000 iterations). A sensitivity analysis was conducted to assess the impact of sepsis classification on MedGemma performance. Classification metrics were evaluated under three distinct scenarios by altering the cohorts: 1) defining both the ICD-Only and Sepsis-3-Only cohorts as sepsis; 2) defining only the ICD-Only cohort as sepsis; and 3) defining only the Sepsis-3-Only cohort as sepsis.

2.8 Manual validation of model classification

We manually reviewed random samples of two discordant subsets: 348 of the 3,638 Non-Sepsis hospitalizations classified as sepsis and 185 of the 349 Sepsis-3-Only hospitalizations classified as non-sepsis. Sample sizes were calculated with finite population correction (expected prevalence 20%, 95% confidence level, 4% margin of error). A case was considered sepsis when the discharge summary documented both suspected infection (culture request, new antibiotic or explicit mention) and concomitant acute organ dysfunction. The review was conducted by the principal investigator (a physician with a specialization in medical informatics) and, as the records had already been classified by the model, it was not blinded.

2.9 Robustness analysis

Cases classified as non-sepsis by MedGemma but originating from the Sepsis-3-Only cohort were subjected to a robustness analysis by adversarial testing. A series of purposefully incorrect or clinically inconsistent statements were incrementally added to each original record, and the modified records were subsequently re-classified by the model. This methodological approach was adapted from the Script Concordance Test (SCT), a tool established in medical education to assess the organization of clinical knowledge[25].

The incrementally added statements were categorized into two distinct types to probe different failure modes. Type 1 (Insufficient Infection Triggers): 1) "Positive blood culture in one sample for *Staphylococcus aureus*." and 2) "Antibiotic therapy escalated to bacitracin." Type 2 (Normal or Stable Data Masked as Abnormal): 3) "Patient with stable chronic renal failure, no need for urgent dialysis during this admission.", 4) "Patient developed hypotension without severe hemodynamic repercussion and vasoactive drugs - saline solution.", and 5) "Thrombocytopenia of 200,000, hyperbilirubinemia of 1.0, and respiratory failure measured by PaO2/FiO2 of 410 mmHg."

Statements were designed around the Sepsis-3 construct (infection plus organ dysfunction) and phrased as information commonly found in discharge summaries. Type 1 statements suggest infection without indicating organ dysfunction and therefore should not, alone, support a sepsis classification. Type 2 statements use terms associated with organ dysfunction while reporting stable conditions or normal values (e.g., platelets 200,000/µL, bilirubin 1.0 mg/dL, PaO_2/FiO_2 410 mmHg) and therefore should not support sepsis either.

Table 1. Cohort characteristics

	Non-Sepsis (n=62,682)	ICD-Only (n=1,532)	Sepsis-3-Only (n=901)	Sepsis (n=408)	p-value
Length of stay days median [IQR]	3 [1 - 7]	15 [7 - 32]	22 [13 - 37]	32 [20 - 46]	<0.001 [a]
Specialties median [IQR]	2 [1 - 3]	5 [3 - 8]	6 [4 - 9]	7 [4 - 10]	<0.001 [a]
Charlson Score, n (%)					
Score 0	38,877 (62.0%)	691 (45.1%)	394 (43.7%)	198 (48.5%)	<0.001 [b]
Score 1-2	19,121 (30.5%)	529 (34.5%)	367 (40.7%)	147 (36.0%)	<0.001 [b]
Score 3-4	2,559 (4.1%)	166 (10.8%)	97 (10.8%)	41 (10.0%)	<0.001 [b]
Score >4	2,125 (3.4%)	146 (9.5%)	43 (4.8%)	22 (5.4%)	<0.001 [b]
ICU, n (%)	3,234 (5.2%)	552 (36.0%)	294 (32.6%)	171 (41.9%)	<0.001 [b]
Palliative care, n (%)	1,655 (2.6%)	250 (16.3%)	219 (24.3%)	130 (31.9%)	<0.001 [b]
Mortality, n (%)	2,312 (3.7%)	561 (36.6%)	383 (42.5%)	243 (59.6%)	<0.001 [b]

Source: Prepared by the authors (2026). IQR: interquartile range. [a] Kruskal-Wallis test. [b] Chi-square test. Death missing values: Non-Sepsis (1,150; 1.8%), ICD-Only (19; 1.2%), Sepsis-3-Only (0), Sepsis (0). Mortality percentages are calculated over the total cohort.

2.10 Ethical considerations

This study was approved by the Research Ethics Committee of the Hospital das Clínicas, Ribeirão Preto Medical School, University of São Paulo (CAAE 75360523.7.0000.5440; opinions 6.505.957 and 7.244.248, the latter for amendment 1), with a waiver of informed consent.

3 RESULTS

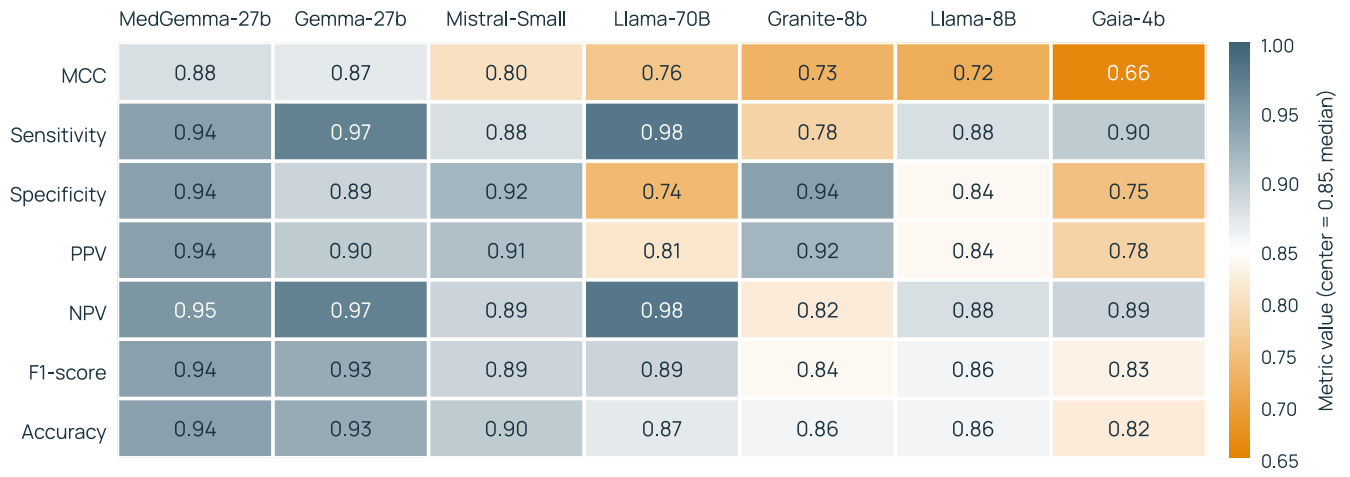
The final study cohort consisted of 65,523 hospitalizations. A total of 1,940 hospitalizations had an ICD-10 code for sepsis; of these, 408 (0.6%) were classified as Sepsis (also meeting operational Sepsis-3 criteria), while 1,532 (2.3%) were classified as ICD-Only. Additionally, 901 (1.4%) hospitalizations were classified as Sepsis-3-Only. The Non-Sepsis cohort comprised the remaining 62,682 (95.7%) hospitalizations (Table 1).

Table 1 shows a severity gradient across the Non-Sepsis, ICD-Only, Sepsis-3-Only and Sepsis cohorts in length of stay, number of specialties, palliative care and mortality (3.7%, 36.6%, 42.5% and 59.6%, respectively). Sepsis-3-Only hospitalizations were more severe than ICD-Only ones, except for ICU admission and a Charlson score above four, which were more frequent in the ICD-Only cohort (36.0% vs. 32.6% and 9.5% vs. 4.8%).

The initial pipeline stage involved selecting an optimal prompt using the development dataset and the Llama (8B) model. The one-shot prompt, in which the model is given a single example to guide its response, demonstrated the highest MCC (0.82, 95% CI 0.78–0.91) and sensitivity (0.81, 95% CI 0.76–0.91). All prompt types tested exhibited specificity and precision of 1.0.

Figure 2 compares the models on the test dataset. MedGemma achieved the highest MCC (0.88, 95% CI 0.85–0.91) and specificity (0.94, 95% CI 0.91–0.96), followed by Gemma (MCC 0.87, 95% CI 0.84–0.90). Llama-70B had the highest sensitivity (0.98, 95% CI 0.97–0.99) but also the highest prompt error rate (27.67% vs. 4.84% for Gemma). Llama-8B performed worse on the test set than on the prompt-tuning dataset (MCC 0.72 vs. 0.82).

Figure 2. Performance comparison of models in the test dataset



Source: Prepared by the authors (2026).

Table 2. MedGemma model classification results on the underreporting dataset

Cohort	Model classification results				
	Non-sepsis	Sepsis	%	Total Classified	Not classified (%)
Non-Sepsis	57,198	3,638	68%	60,836	1,443 (2.3%)
ICD-Only	209	1,234	23%	1,443	89 (5.8%)
Sepsis-3-Only	349	512	10%	861	40 (4.4%)
Total	57,756	5,384		63,140	1,572 (2.4%)

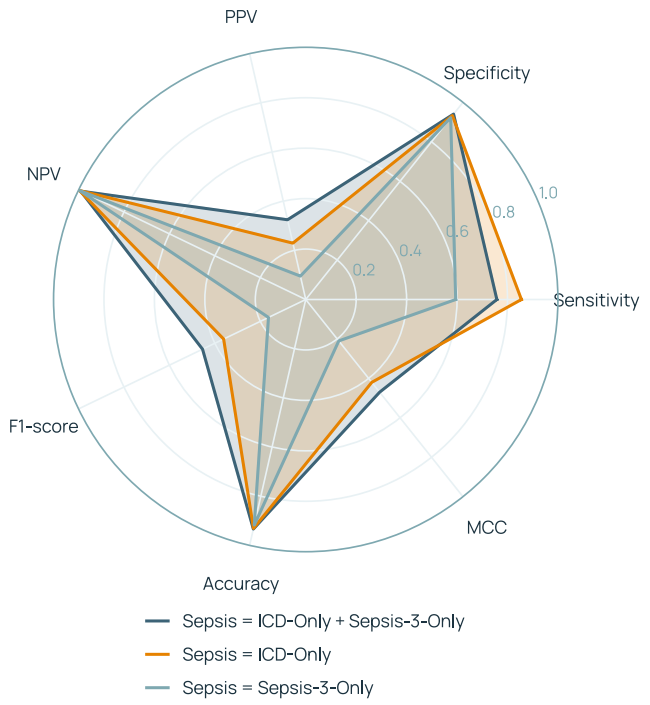
Source: Prepared by the authors (2026). The Sepsis cohort (n=408) and the Non-Sepsis test set (n=403) were not included in this task. %: share of the 5,384 sepsis classifications. Not classified (inference failure or invalid output): % of hospitalizations submitted per cohort.

Table 3. Robustness analysis

Phase / Step	Added Statement	Reclassified as sepsis (n=349)	%
Baseline	Step 0: Original text (False Negatives)	0	0.0%
Type 1 Statements	Insufficient Infection Triggers		
Step 1	+ Blood Culture (Inconsistent)	193	55.3%
Step 2	+ Antibiotic (Incorrect Treatment)	258	73.9%
Type 2 Statements	Normal or Stable Data Masked as Abnormal		
Step 3	+ CKD (Irrelevant)	250	71.6%
Step 4	+ Hypotension (Mild)	319	91.4%
Step 5	+ Labs (Contradictory)	339	97.1%
Type 2 Result	Final cumulative reclassification after Type 2	339	97.1%

Source: Prepared by the authors (2026). The statements were added incrementally.

Figure 3. Metrics comparison between analysis groups



Source: Prepared by the authors (2026).

The MedGemma model's classification results on the sepsis clinical classification task are detailed in Table 2. In total, 63,140 records were classified (97.6% of the 64,712 submitted). From the Sepsis-3-Only cohort, 349 (40.5%) were classified as non-sepsis. As for the ICD-Only cohort, 209 (14.5%) were classified as non-sepsis. On the other hand, 3,638 (6.0%) of the Non-Sepsis cohort records were classified as sepsis (Table 2).

The sensitivity analysis revealed that model metrics varied depending on the sepsis definition used. A sequential decrease in MCC was observed depending on the sepsis definition: 0.47 (when considering ICD-Only + Sepsis-3-Only), 0.42 (ICD-Only), and 0.21 (Sepsis-3-Only). The same pattern occurred with specificity (0.94, 0.93, and 0.92, respectively). In contrast, sensitivity showed a different behavior: it increased to 0.86 for the ICD-Only group (compared to 0.76 for the combined ICD-Only + Sepsis-3-Only group) but dropped to its lowest value (0.59) for the Sepsis-3-Only group.

Table 3 shows the robustness analysis of the 349 Sepsis-3-Only records classified as non-sepsis by MedGemma. Type 1 statements reclassified 258 records (73.9%); among the 91 records not reclassified after them, 82 (90.1%) were reclassified after Type 2 statements, reaching 97.1% (339/349). Classification was not monotonic: between consecutive steps some records changed from sepsis to non-sepsis (e.g., 38 after the CKD statement). Inference errors (30 at Step 1 and 1 at Steps 2 and 4) were counted as not reclassified.

Among the 3,638 hospitalizations from Non-Sepsis cohort but classified as sepsis by the model, a random sample of 348 cases was manually reviewed. The review identified that 29.9% (104/348) of these cases met the textual sepsis criteria. Regarding the Sepsis-3-Only cohort classified as non-sepsis, a review of 185 cases showed that only 22 (11.9%) contained textual information sufficient to support a sepsis diagnosis. Consequently, the model's non-sepsis classification was consistent with the available discharge summary text in 88.1% of these cases.

4 DISCUSSION

In this study, the best-performing LLM in the test phase was applied to a real-world dataset comprising ICD-Only, Sepsis-3-Only and Non-Sepsis cohorts, and evaluated with established clinical metrics and robustness testing.

Studies on the application of artificial intelligence in ICUs show promise but are constrained by limitations such as biases stemming from small sample sizes, inconsistent variable definitions, a lack of clear rationale for model selection, and reliance on ICD codes as a primary information source[26]. In contrast,

the dataset used in this study comprises large-scale, real-world data, intentionally curated to challenge the model's capabilities.

A distinct severity gradient was observed across sepsis cohorts. Our approach, distinguishing between the ICD-Only (human coding) and Sepsis-3-Only cohorts, revealed marked differences in clinical profiles, with the latter demonstrating significantly higher severity. As expected, performance, particularly MCC, sensitivity and PPV, degraded from the test environment to the real-world dataset.

Robust metrics are as critical as dataset selection. Shool et al. demonstrated that accuracy is the most common metric for LLM evaluation, despite its evident flaws in imbalanced data, while sensitivity and specificity appear in only 2% of reviewed studies[10]. We observed that accuracy, NPV, and specificity remained static in our analysis, masking underlying issues. In contrast, MCC, sensitivity, and PPV proved to be sensitive indicators, exposing performance degradation tied to data composition. This confirms the necessity of using robust, multi-dimensional metrics to accurately assess model behavior in healthcare[27].

Moreover, the sensitivity analysis revealed that performance is contingent upon the sepsis definition: results were notably superior when human-assigned ICD text was present compared to Sepsis-3-Only cases. This suggests the model may have been exposed to similar ICD patterns during its original pre-training[28]. Additionally, the sensitivity analysis highlights a critical issue: data selection bias. The model's reported results fluctuate depending on the initial ground truth definition of sepsis (i.e., whether it encompasses ICD-Only or Sepsis-3-Only groups). This challenge is well-recognized in medicine and rigorously addressed in health technology assessment studies, standards from which AI model evaluation could significantly benefit.

We observed that robustness testing with adversarial data triggered the model to reclassify the 349 Sepsis-3-Only cases previously missed by the model. Overall, the addition of misleading text increased sepsis classifications, although not monotonically (the stable CKD statement slightly reduced them). First, the model demonstrated a bias toward infection-related keywords (Type 1 statements), reclassifying 73.9% of cases when presented with suggestive but clinically insufficient data (e.g., a single positive blood culture). The evidence against clinical reasoning emerged from Type 2 statements: the model completely failed to process normal numerical values or stable clinical contexts when they were preceded by alarming keywords (e.g., misinterpreting a normal platelet count of 200,000 as severe thrombocytopenia). This inability to override semantic keywords with logical numerical assessment drove the final reclassification rate to 97.1%.

The two types of statements exposed complementary failures. In Type 1, the model accepted infection cues (a positive culture, an antibiotic change) as sufficient for sepsis without evidence of organ dysfunction; in Type 2, it accepted organ dysfunction terms as abnormal even when the accompanying values or context were normal or stable, except for the stable CKD statement. In both cases, the model responded to the presence of the clinical term rather than integrating it with the remaining data.

These findings challenge discussions regarding the clinical reasoning capabilities of large language models[13,14]. Clinical reasoning is a complex cognitive process inexorably linked to critical thinking; grounded in foundational cognitive theories, such as System 1 and 2 processing and the management of cognitive biases[29].

The models failed when presented with errors identifiable by medical students. The robustness analysis suggests that the LLM is operating as a probabilistic pattern identifier rather than following a hypothetico-deductive approach. It is possible that the model exhibits a bias derived from the statistical patterns ingested during its original pre-training, effectively mimicking human cognitive shortcuts without the associated metacognition.

While our methodology differs from McCoy et al.[15], our conclusions align: the model demonstrates lack of reasoning capabilities. We argue that this emergent research field would benefit from more rigorous definitions and metrics for evaluating clinical reasoning in LLMs.

One might argue that one-shot learning could be improved by more sophisticated techniques, such as chain-of-thought (CoT) prompting. This technique aims to simulate System 2 thinking, encouraging the model to pause and "think" through the problem stepwise. However, studies have shown inconsistent results with this technique[30]. An interesting finding is that the CoT performance improves when examples from the dataset are included in the prompt[31]. This dependency raises concerns regarding the independence of the tests[12] suggesting that performance gains may stem from pattern replication rather than genuine deductive reasoning.

Notwithstanding the identified model limitations, a clinically relevant finding emerged regarding underreported cases. The model successfully detected 512 underreported sepsis cases originating explicitly from the Sepsis-3-Only cohort (i.e., cases objectively meeting structured partial SOFA and culture criteria but lacking ICD codes). Against a total of 5,384 cases flagged as sepsis across all cohorts, this yields a strict, objective ratio of 1 confirmed Sepsis-3 true positive for every 10.5 cases flagged. In the manual review, 29.9% of the Non-Sepsis hospitalizations flagged by the model

met the textual sepsis criteria, corresponding to approximately 1,088 hospitalizations; as this estimate derives from a single-reviewer, unblinded review of discharge summaries, it should be interpreted as indicative. This performance is superior in the context of medical record quality audit to that reported by Wong et al. (1 in 109)[32] for a widely utilized EHR alert system.

Regarding the Sepsis-3-Only cases classified as non-sepsis, 88% of the model's predictions were consistent with the available discharge summary text, highlighting potential clinical documentation deficiencies rather than model failure. An alternative explanation involves distinct documentation practices for palliative care patients, where the clinical narrative often shifts focus away from acute diagnostic labeling and specific treatment interventions.

These findings underscore the need for a strategic approach to clinical implementation, aligning with frameworks that advocate for responsible integration: the data-centric approach [33,34] and the human-in-the-loop approach[35]. Rather than autonomous high-stakes tasks, LLMs should currently be directed to specific, safe applications, such as retrospective medical record quality auditing.

Study limitations and future work

This study has limitations. First, the operational Sepsis-3 definition was incomplete: PaO_2/FiO_2 and vasopressor use were unavailable as structured data (maximum partial SOFA of 17 of 24 points); this absence is a limitation of the study, consistent with retrospective analyses of structured real-world data. No time limit was applied between baseline and peak, and culture-negative sepsis was not captured. Organ dysfunction was therefore likely underestimated, and some sepsis cases may have been assigned to the ICD-Only or Non-Sepsis cohorts and counted as false positives; performance metrics should be interpreted relative to this reference, whereas the robustness analysis, which compares the same records before and after adding adversarial text, is less dependent on it. Second, sepsis diagnosis is inherently challenging, even for experts, and less ambiguous tasks should also be studied. Third, the manual review was performed by a single, unblinded reviewer (the principal investigator is a physician with a specialization in medical informatics), precluding assessment of inter-rater agreement; its results should be considered indicative. Fourth, 4-bit quantization may have degraded performance and is a potential confounder. Finally, only generative models were evaluated and adversarial statements were added only incrementally; future studies should evaluate encoder-based (e.g., BERT), fine-tuned and agent-based approaches, and test adversarial statements in isolation across a broader range of models.

5 CONCLUSIONS

Our evaluation using a curated, complex real-world dataset demonstrates that LLMs currently lack the clinical reasoning required for autonomous sepsis diagnosis. By applying robust performance metrics, we revealed that these models operate primarily as probabilistic keyword matchers susceptible to training data patterns. At the current stage of development, clinical reasoning capabilities cannot yet be attributed to this LLM on the basis of the present study; this would need to be demonstrated across independent datasets and future model iterations. However, despite these limitations, the model proved valuable as a retrospective screening tool, identifying underreported cases with precision superior to existing alert systems. Therefore, rather than pursuing high-stakes automation, we advocate for a strategic deployment focused on human-in-the-loop applications, specifically medical record quality auditing, where the model's capacity to flag missed diagnoses can be leveraged safely to improve institutional outcomes.

Author contributions. All authors contributed significantly to the study.

Funding. CNPq Research Productivity Scholarship, Level 2 (318645/2025-4), to Antonio Pazin-Filho; FAPESP (process 25/15172-9).

Data and materials availability. The supplementary material is available at https://doi.org/10.5281/zenodo.24366361.

Conflicts of interest. The authors declare no conflict of interest.

Use of artificial intelligence. The authors used the Nature Research Assistant tool to assist with the linguistic review and textual organization of the manuscript. All content was reviewed and validated by the authors, who assume full responsibility for the work presented.

REFERENCES

1. Alafari F, Driss M, Cherif A. Advances in natural language processing for healthcare: A comprehensive review of techniques, applications, and future directions. Comput Sci Rev. 2025;56:100725. doi: https://doi.org/10.1016/j.cosrev.2025.100725

2. Biesheuvel LA, Workum JD, Reuland M, van Genderen ME, Thoral P, Dongelmans D, et al. Large language models in critical care. J Intensive Med. 2025;5(2):113-8. doi: https://doi.org/10.1016/j.jointm.2024.12.001

3. Bilal M, Hamza A, Malik N. NLP for Analyzing Electronic Health Records and Clinical Notes in Cancer Research: A Review. J Pain Symptom Manage. 2025;69(5):e374-e394. doi: https://doi.org/10.1016/j.jpainsymman.2025.01.019
4. Bedi S, Liu Y, Orr-Ewing L, Dash D, Koyejo S, Callahan A, et al. Testing and Evaluation of Health Care Applications of Large Language Models. JAMA. 2025;333(4):319. doi: https://doi.org/10.1001/jama.2024.21700
5. Chakraborty N, Ornik M, Driggs-Campbell K. Hallucination Detection in Foundation Models for Decision-Making: A Flexible Definition and Review of the State of the Art. ACM Comput Surv. 2025;57(7):1-35. doi: https://doi.org/10.1145/3716846
6. Holmes G, Tang B, Gupta S, Venkatesh S, Christensen H, Whitton A. Applications of Large Language Models in the Field of Suicide Prevention: Scoping Review. J Med Internet Res. 2025;27:e63126. doi: https://doi.org/10.2196/63126
7. Liang EN, Pei S, Staibano P, van der Woerd B. Clinical applications of large language models in medicine and surgery: A scoping review. J Int Med Res. 2025;53(7):3000605251347556. doi: https://doi.org/10.1177/03000605251347556
8. Lee J, Park S, Shin J, Cho B. Analyzing evaluation methods for large language models in the medical field: a scoping review. BMC Med Inform Decis Mak. 2024;24(1):366. doi: https://doi.org/10.1186/s12911-024-02709-7
9. Poon EG, Lemak CH, Rojas JC, Guptill J, Classen D. Adoption of artificial intelligence in healthcare: survey of health system priorities, successes, and challenges. J Am Med Inform Assoc. 2025;32(7):1093-100. doi: https://doi.org/10.1093/jamia/ocaf065
10. Shool S, Adimi S, Saboori Amleshi R, Bitaraf E, Golpira R, Tara M. A systematic review of large language model (LLM) evaluations in clinical medicine. BMC Med Inform Decis Mak. 2025;25(1):117. doi: https://doi.org/10.1186/s12911-025-02954-4
11. Klug K, Beckh K, Antweiler D, Chakraborty N, Baldini G, Laue K, et al. From admission to discharge: a systematic review of clinical natural language processing along the patient journey. BMC Med Inform Decis Mak. 2024;24(1):238. doi: https://doi.org/10.1186/s12911-024-02641-w
12. Gallo RJ, Baiocchi M, Savage TR, Chen JH. Establishing best practices in large language model research: an application to repeat prompting. J Am Med Inform Assoc. 2025;32(2):386-90. doi: https://doi.org/10.1093/jamia/ocae294
13. Savage T, Nayak A, Gallo R, Rangan E, Chen JH. Diagnostic reasoning prompts reveal the potential for large language model interpretability in medicine. NPJ Digit Med. 2024;7(1):20. doi: https://doi.org/10.1038/s41746-024-01010-1
14. Tordjman M, Liu Z, Yuce M, Fauveau V, Mei Y, Hadjadj J, et al. Comparative benchmarking of the DeepSeek large language model on medical tasks and clinical reasoning. Nat Med. 2025;31(8):2550-5. doi: https://doi.org/10.1038/s41591-025-03726-3
15. McCoy LG, Swamy R, Sagar N, Wang M, Bacchi S, Fong JMN, et al. Assessment of Large Language Models in Clinical Reasoning: A Novel Benchmarking Study. NEJM AI. 2025;2(10). doi: https://doi.org/10.1056/aidbp2500120
16. Fleischmann-Struzek C, Mellhammar L, Rose N, Cassini A, Rudd KE, Schlattmann P, et al. Incidence and mortality of hospital- and ICU-treated sepsis: results from an updated and expanded systematic review and meta-analysis. Intensive Care Med. 2020;46(8):1552-62. doi: https://doi.org/10.1007/s00134-020-06151-x
17. Singer M, Deutschman CS, Seymour CW, Shankar-Hari M, Annane D, Bauer M, et al. The Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3). JAMA. 2016;315(8):801. doi: https://doi.org/10.1001/jama.2016.0287
18. Silva AR, Hoffmann NG, Fernandez-Llimos F, Lima EC. Data quality review of the Brazilian nosocomial infections surveillance system. J Infect Public Health. 2024;17(4):687-95. doi: https://doi.org/10.1016/j.jiph.2024.02.013
19. Gallifant J, Afshar M, Ameen S, Aphinyanaphongs Y, Chen S, Cacciamani G, et al. The TRIPOD-LLM reporting guideline for studies using large language models. Nat Med. 2025;31(1):60-9. doi: https://doi.org/10.1038/s41591-024-03425-5
20. Da Silva RP, Pazin-Filho A. Dataset of anonymized discharge summaries of sepsis patients from a Brazilian tertiary hospital for NLP applications. Data Brief. 2025;61:111804. doi: https://doi.org/10.1016/j.dib.2025.111804
21. Meta Llama. Llama-3.1-8B-Instruct; 2024 [cited 2026 Nov 15]. Available from: https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct.
22. Grattafiori A, Dubey A, Jauhri A, Pandey A, Kadian A, Al-Dahle A, et al. The Llama 3 Herd of Models [Preprint]. arXiv: 2407.21783. 2024. doi: https://doi.org/10.48550/arXiv.2407.21783
23. Google. MedGemma 27B text-it: (medgemma-27b-text-it); 2025 [cited 2026 Nov 15]. Available from: https://huggingface.co/google/medgemma-27b-text-it.
24. Chicco D, Jurman G. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. BMC Genomics. 2020;21(1):6. doi: https://doi.org/10.1186/s12864-019-6413-7
25. Charlin B, Roy L, Brailovsky C, Goulet F, van der Vleuten C. The Script Concordance Test: A Tool to Assess the Reflective Clinician. Teach Learn Med. 2000;12(4):189-95. doi: https://doi.org/10.1207/s15328015tlm1204_5
26. Tungushpayev M, Suleimenova D, Sarria-Santamerra A, Aimyshev T, Gaipov A, Viderman D. The value of machine and deep learning in management of critically ill patients: An umbrella review. Int J Med Inform. 2025;204:106081. doi: https://doi.org/10.1016/j.ijmedinf.2025.106081
27. van der Vegt AH, Campbell VK, Webb R, Venkatesh B, Lane PJ, Wilks K, et al. A novel, standardised approach to balancing effectiveness, efficiency and utility of surveillance AI prediction models for hospitalised patients using sepsis prediction as an exemplar. J Am Med Inform Assoc. 2026;33(2):304-15. doi: https://doi.org/10.1093/jamia/ocaf192
28. Chen H, Fang Z, Singla Y, Dredze M. Benchmarking Large Language Models on Answering and Explaining Challenging Medical Questions. In: Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers). 2025. p. 3563-99. doi: https://doi.org/10.18653/v1/2025.naacl-long.182

29. Richards JB, Hayes MM, Schwartzstein RM. Teaching Clinical Reasoning and Critical Thinking. Chest. 2020;158(4): 1617-28. doi: https://doi.org/10.1016/j.chest.2020.05.525

30. Jeon S, Kim HG. A comparative evaluation of chain-of-thought-based prompt engineering techniques for medical question answering. Comput Biol Med. 2025;196:110614. doi: https://doi.org/10.1016/j.compbiomed.2025.110614

31. Park C, Kim C. A Novel Chain-of-Thought Reasoning Approach for Alzheimer's Disease Detection Using Large Language and Vision-Language Models. IEEE Trans Neural Syst Rehabil Eng. 2025;33:4386-95. doi: https://doi.org/10.1109/tnsre.2025.3624762

32. Wong A, Otles E, Donnelly JP, Krumm A, McCullough J, DeTroyer-Cooley O, et al. External Validation of a Widely Implemented Proprietary Sepsis Prediction Model in Hospitalized Patients. JAMA Intern Med. 2021;181(8):1065. doi: https://doi.org/10.1001/jamainternmed.2021.2626

33. Kumar S, Datta S, Singh V, Singh SK, Sharma R. Opportunities and Challenges in Data-Centric AI. IEEE Access. 2024;12:33173-89. doi: https://doi.org/10.1109/access.2024.3369417

34. Bhatt N, Bhatt N, Prajapati P, Sorathiya V, Alshathri S, El-Shafai W. A Data-Centric Approach to improve performance of deep learning models. Sci Rep. 2024;14(1):22329. doi: https://doi.org/10.1038/s41598-024-73643-x

35. Mandreoli F, Ferrari D, Guidetti V, Motta F, Missier P. Real-world data mining meets clinical practice: Research challenges and perspective. Front Big Data. 2022;5:1021621. doi: https://doi.org/10.3389/fdata.2022.1021621

EDITORIAL NOTES

The Journal of Health Informatics is a scientific journal maintained by the Brazilian Society of Health Informatics (Sociedade Brasileira de Informática em Saúde), a non-profit scientific association dedicated to the development of Health Informatics and Digital Health in Brazil. The editorial process is conducted with scientific independence, free from institutional, commercial, or sponsor interference, and follows international best practices in publication ethics.

This article was double-blind peer reviewed by specialists in the field. The editorial team used artificial intelligence tools, under human supervision, in the initial screening, in checking the text, references and numbers and, when needed, in redrawing figures, charts and tables, which the authors approved in the proof. The tools and safeguards are described in the journal's Policy on the use of artificial intelligence. Editorial decisions were made by people.

SBIS Sociedade Brasileira de Informática em Saúde