iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 14 **Issue:** VII **Month of publication:** July 2026

DOI: https://doi.org/10.22214/ijraset.2026.84386

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue VII July 2026- Available at www.ijraset.com

Evaluating Large Language Models Against Clinical Assessment Frameworks for Early Sepsis Detection in the ICU

Anvit More[1], Dr. Vishala Bodetti[2], Kishan Gor[3], Nrip Nihalani[4], Aditya Patkar[5]
[1]AI Engineer, [2]AI Medical Lead, [3]VP, AI Solution Architect, [4]Director, [5]CEO, Plus91 Technology, Pune, Maharashtra, India

Abstract: ***Timely recognition of sepsis remains difficult when early physiological abnormalities are subtle or incomplete. This study examined whether general-purpose large language models could discriminate sepsis risk from an initial ICU vital-sign snapshot as effectively as established clinical scoring approaches. We performed a retrospective benchmark using the de-identified PhysioNet Sepsis Prediction Dataset. The eligible cohort included 39,234 adults, of whom 2,733 were sepsis-positive. Two zero-shot language models and two modified clinical scores were evaluated on the same patients. Discrimination was measured by the area under the receiver operating characteristic curve (AUROC), with bootstrap confidence intervals and DeLong tests for paired comparisons. AUROC was 0.597 for GPT-5.5, 0.592 for modified NEWS2, 0.591 for Claude Sonnet 5, and 0.570 for modified SOFA. Neither language model differed significantly from modified NEWS2, whereas both produced higher AUROC values than modified SOFA. A pre-onset-only sensitivity analysis, restricted to patients whose snapshot preceded their first positive sepsis label (achieved median lead time 39 hours), showed all four estimators converging (AUROC 0.579–0.591) with no significant pairwise differences, indicating the primary-analysis gap over SOFA was partly attributable to post-onset records. A stricter analysis limited to a minimum 6-hour lead time reversed the ranking: the SOFA-derived score obtained the highest AUROC (0.597), with the LLMs and NEWS2-derived score converging lower (0.573–0.578), again with no significant pairwise differences. Their alerting behaviour was not interchangeable: GPT-5.5 produced a more balanced sensitivity-specificity profile, while Claude Sonnet 5 identified more positive cases at the cost of additional false alerts. These findings show that zero-shot language models can provide discrimination comparable to the strongest modified clinical score in this dataset. The modest absolute AUROC values, use of modified scores, and retrospective single-dataset design mean that the models should be considered complementary research tools, not stand-alone clinical detectors.***

Keywords: sepsis prediction, large language models, clinical assessment scores, intensive care, early warning systems, clinical decision support, AUROC, sensitivity analysis.

I. INTRODUCTION

Sepsis arises when a dysregulated response to infection is accompanied by life-threatening organ dysfunction. Its global burden is substantial: an analysis of 2017 data estimated 48.9 million cases and about 11 million associated deaths [1]. For clinicians, the challenge is to recognise a concerning trajectory before deterioration becomes unmistakable. In septic shock, longer delays between the onset of recurrent or persistent hypotension and effective antimicrobial therapy have been associated with lower survival [2].

Bedside scores provide a consistent way to organise physiological observations. NEWS2 combines routinely measured variables and links the total to escalation guidance [3]. SOFA quantifies organ dysfunction and contributes to the Sepsis-3 definition [4]. Both are transparent, but they serve different purposes and require specific inputs. When those inputs are missing or replaced with proxies, the resulting calculation is no longer the standard instrument.

Large language models have been evaluated on medical knowledge, diagnostic reasoning, and triage tasks [5]–[9]. They can turn a compact patient description into an ordinal judgement without a bespoke feature-engineering pipeline. That convenience is not evidence of clinical usefulness. A credible benchmark must use the same outcome and cohort for every estimator, disclose exactly what each estimator receives, and report threshold-dependent errors as well as a summary discrimination statistic.

We therefore evaluated two frontier language models against two modified clinical scoring systems using the same 39,234-patient ICU cohort. Each model received only one early cross-sectional assessment. The primary analysis compared AUROC values; secondary analyses examined sensitivity, specificity, predictive values, and performance across early and later ICU assessment windows. The study asks a deliberately narrow question: how well can an unadapted language model rank sepsis risk when it sees the same limited physiological information available to a simple bedside score?

II. RELATED WORK

A. Clinical Scoring Systems for Sepsis Detection

NEWS2 assigns points for respiratory rate, oxygen saturation, supplemental oxygen, systolic blood pressure, pulse, temperature, and consciousness [3]. It is a general deterioration score, not a sepsis-specific diagnostic test. Its reported performance varies with setting, outcome, case mix, and assessment time [10], [11].

SOFA scores respiratory, coagulation, hepatic, cardiovascular, neurological, and renal dysfunction from 0 to 4 in each domain [4], [12]. It is central to Sepsis-3 but was not designed as a rapid triage score. Several components require laboratory values, vasopressor information, or a neurological assessment. Any implementation that omits domains or substitutes measurements should therefore be described as a study-specific approximation, not as standard SOFA.

B. Machine-Learning Approaches

Conventional machine-learning and sequential models have also been applied to sepsis prediction. Studies using related data have reported high discrimination after feature engineering, ensembling, or longitudinal modelling [13], [14]. Those systems address a different problem from this benchmark. We did not train a specialised predictor on the study cohort; instead, we tested whether general-purpose LLMs could recover a useful risk ordering from a single sparse snapshot.

C. Language Models in Clinical Evaluation

Previous studies have tested LLMs on medical questions, published clinical cases, and emergency triage [5]–[9]. Triage evaluations often compare a model-assigned acuity level with a clinician's decision. Such agreement is informative, but it is not the same as discrimination against a patient-level outcome. Model-level AUROC can also conceal operational differences: two models with similar ranking performance may miss different numbers of cases or create very different alert burdens.

III. MATERIALS AND METHODS

A. Study Design

This retrospective benchmark compared four estimators: a NEWS2-derived score, a SOFA-derived score, GPT-5.5, and Claude Sonnet 5. Reporting was organised with reference to TRIPOD+AI [15]. Only the public, de-identified PhysioNet Sepsis Prediction Dataset was used; no directly identifiable patient information was accessed, and no new human-subjects data were collected. Under the authors' institutional research governance policy, secondary analysis of a fully de-identified, publicly released dataset of this kind does not require separate institutional ethics board review.

B. Dataset

We used the public training release of the 2019 PhysioNet/Computing in Cardiology Challenge [16]. It contains 40,336 ICU records from two hospital systems, with one file per patient and observations arranged by ICU hour. Variables include vital signs, laboratory measurements, demographics, and SepsisLabel. The Challenge defines sepsis using a Sepsis-3-based rule and shifts positive labels six hours earlier than the estimated clinical onset to support early-prediction research [16].

C. Cohort and Assessment Time

We excluded records for patients younger than 18 years and records with fewer than two ICU hours. For each remaining record, we selected the earliest hour containing heart rate, oxygen saturation, temperature, systolic blood pressure, mean arterial pressure, and respiratory rate. The outcome was record-level sepsis positivity, defined as at least one SepsisLabel value of 1 anywhere in the record. This procedure yielded 39,234 records. Because the selected snapshot was not required to precede the first positive label, the analysis does not establish a fixed-horizon forecast of incident sepsis.

D. Derivation of the Clinical Comparators

The available variables did not permit complete NEWS2 or SOFA calculations at the chosen assessment. We therefore evaluated study-specific score approximations. These scores should not be interpreted as validated implementations of the source instruments.

1) NEWS2-derived score: The calculation used respiratory rate, oxygen saturation, systolic blood pressure, heart rate, and temperature. Consciousness and supplemental-oxygen status were unavailable. Mean arterial pressure was added as a study-specific hypotension term even though it is not a NEWS2 component [3]. Published NEWS2 cut points were retained only for variables that directly corresponded to the original score.

2) SOFA-derived score: Five available fields were used: oxygen saturation as a respiratory proxy, platelet count for coagulation, total bilirubin for hepatic function, mean arterial pressure for cardiovascular status, and creatinine for renal function. Glasgow Coma Scale and vasopressor data were unavailable. Oxygen saturation is not the standard SOFA respiratory measure, which relies on the PaO2/FiO2 ratio; the resulting score was therefore a heuristic approximation [4], [12].

E. Language-Model Evaluation

GPT-5.5 was evaluated with the high-reasoning option and Claude Sonnet 5 with the medium-reasoning option available in their respective chat products. Neither model was fine-tuned on this cohort. The use of consumer-facing chat interfaces was intended to represent an accessible workflow, but it also limits reproducibility because product routing and inference settings may change over time.

Both LLMs were run on all 39,234 eligible records. Records were submitted in sequential batches of about 4,900 rows, with instructions that each row represented an independent patient assessment and that information must not carry across rows. The same row template and 0–10 output scale were used for both models.

The zero-shot prompt was as follows:

> You are a clinical decision support system in an ICU setting.
>
> Patient vitals at assessment:
> Heart Rate: [HR] bpm; Respiratory Rate: [RR] breaths/min; Systolic BP: [SBP] mmHg; Mean Arterial Pressure: [MAP] mmHg; SpO2: [O2Sat]%; Temperature: [Temp] °C; Age: [Age] years; Sex: [Gender].
>
> What is this patient's risk of developing sepsis? Reply with a single integer from 0 (no risk) to 10 (highest risk). No explanation. Just the number.

The returned integer was analysed as an ordinal score rather than a calibrated probability. Explanations, diagnostic labels, and chain-of-thought text were neither requested nor used.

F. Statistical Analysis

AUROC was calculated for each estimator in the full eligible cohort. Ninety-five per cent confidence intervals were obtained from 1,000 bootstrap resamples with seed 42. Paired AUROCs were compared using DeLong's non-parametric test [17] with a two-sided alpha of 0.05. Sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV) were calculated at the Youden-optimal threshold selected in the same cohort; these threshold-based estimates are therefore descriptive. A secondary timing analysis compared snapshots from ICU hours 1–6 with those from hours 7–24. Snapshots after hour 24 were retained in the full-cohort analysis but excluded from this timing analysis. P values were not adjusted for multiple comparisons.

IV. RESULTS

A. Cohort Characteristics

The final cohort contained 39,234 adult ICU records, of which 2,733 (7.0%) were sepsis-positive. Median age was 63 years (interquartile range [IQR], 51–74), and 21,934 patients (55.9%) were male. The first snapshot meeting the completeness rule occurred at a median of four ICU hours (IQR, 2–5). Table I summarises the cohort.

TABLE I COHORT CHARACTERISTICS (N = 39,234)

Characteristic	Value
Total eligible patients	39,234
Sepsis positive, n (%)	2,733 (7.0%)
Age, years, median (IQR)	63 (51–74)

Characteristic	Value
Male sex, n (%)	21,934 (55.9%)
Time to first complete snapshot, hours, median (IQR)	4 (2–5)
NEWS2-derived score calculable, n (%)	39,234 (100%)
SOFA-derived score calculable, n (%)	39,234 (100%)

B. Discriminative Performance

GPT-5.5 had the highest AUROC point estimate at 0.597 (95% confidence interval [CI], 0.586–0.607). The NEWS2-derived score followed at 0.592 (95% CI, 0.581–0.603), Claude Sonnet 5 at 0.591 (95% CI, 0.580–0.603), and the SOFA-derived score at 0.570 (95% CI, 0.559–0.581). Although the confidence intervals were narrow in this large matched cohort, discrimination was modest for every estimator.

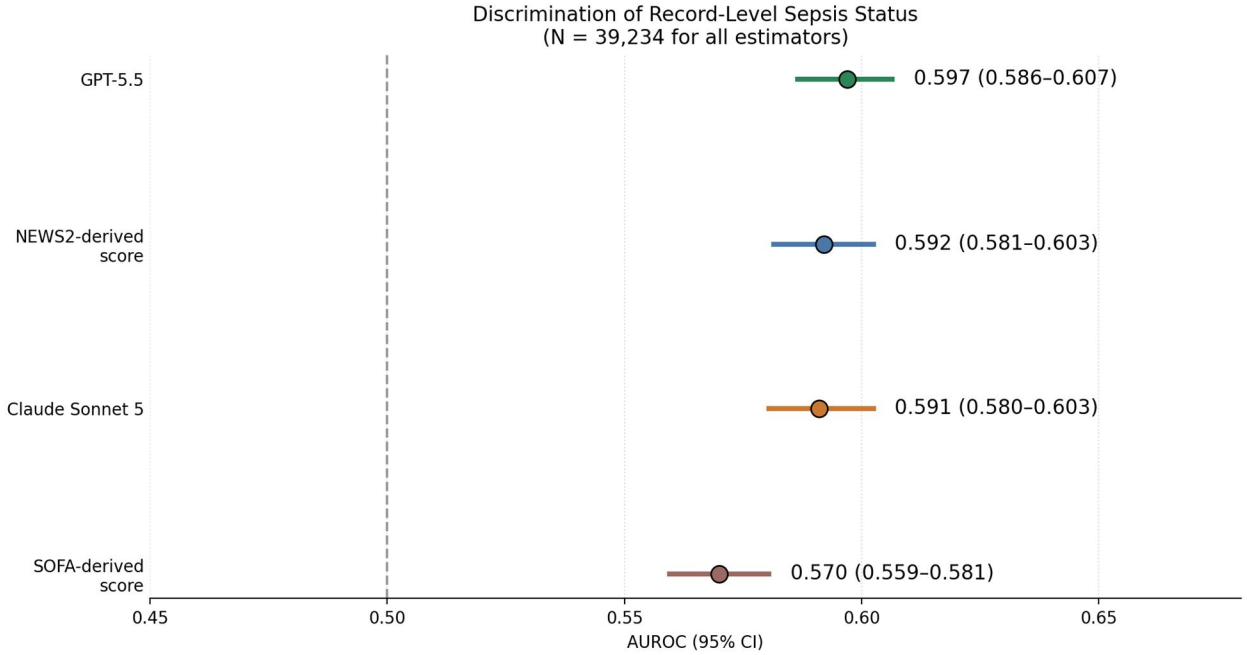


Fig. 1. AUROC estimates and 95% bootstrap confidence intervals for record-level sepsis discrimination in the full cohort. The dashed line denotes chance discrimination (AUROC = 0.50).

TABLE II DISCRIMINATION IN THE FULL COHORT

Model	AUROC (95% CI)	N	p vs. SOFA-derived score	p vs. NEWS2-derived score
GPT-5.5	0.597 (0.586–0.607)	39,234	0.002**	0.656
NEWS2-derived score	0.592 (0.581–0.603)	39,234	0.010*	Reference
Claude Sonnet 5	0.591 (0.580–0.603)	39,234	0.005**	0.827
SOFA-derived score	0.570 (0.559–0.581)	39,234	Reference	0.010*

* $p < 0.05$; ** $p < 0.01$. GPT-5.5 versus Claude Sonnet 5: $z = 0.23$, $p = 0.816$. P values are unadjusted.

Each LLM had a higher AUROC than the SOFA-derived score (p = 0.002 for GPT-5.5; p = 0.005 for Claude Sonnet 5). Neither LLM differed significantly from the NEWS2-derived score, and the LLMs did not differ from each other (p = 0.816). These are difference tests; a nonsignificant result does not demonstrate equivalence.

C. Operating Characteristics

Similar AUROCs did not produce similar alert patterns. At its Youden-optimal threshold, GPT-5.5 yielded 51.0% sensitivity and 63.0% specificity. Claude Sonnet 5 yielded 67.9% sensitivity and 44.9% specificity. Claude therefore identified more sepsis-positive records while also flagging more records that remained sepsis-negative.

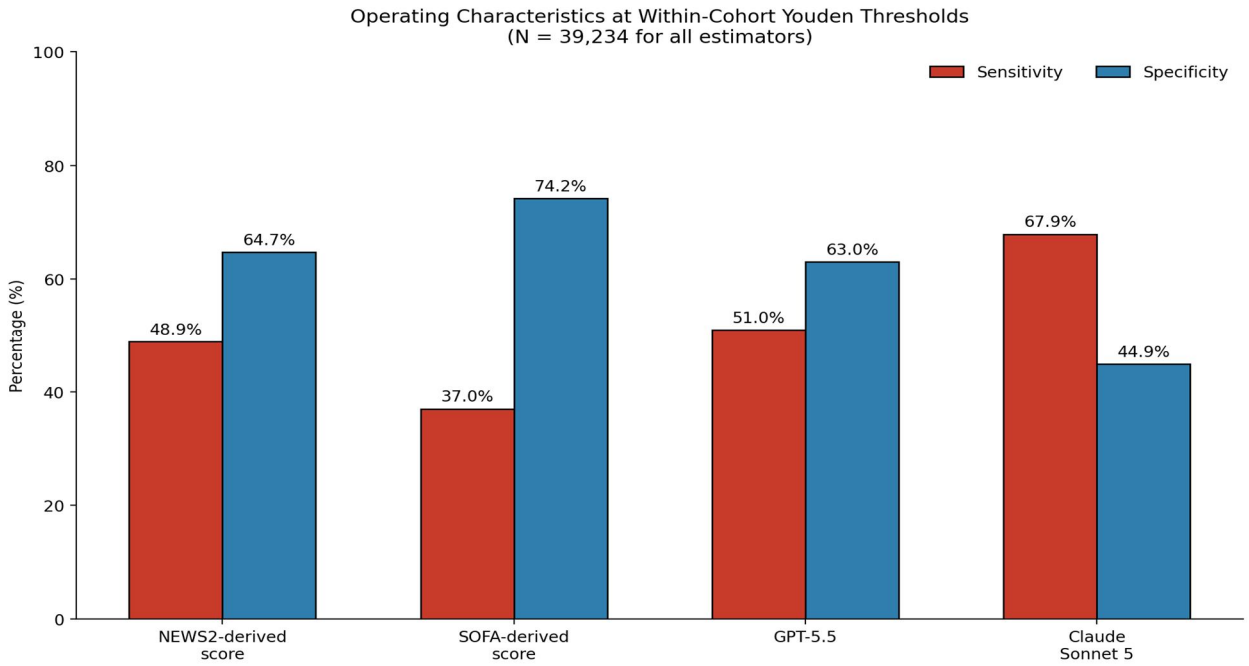


Fig. 2. Sensitivity and specificity at the within-cohort Youden-optimal threshold for each estimator.

TABLE III OPERATING CHARACTERISTICS AT WITHIN-COHORT YOUDEN THRESHOLDS

Model	Threshold	Sensitivity	Specificity	PPV	NPV
NEWS2-derived score	≥3	48.9%	64.7%	9.4%	94.4%
SOFA-derived score	≥2	37.0%	74.2%	9.7%	94.0%
GPT-5.5	≥4	51.0%	63.0%	9.4%	94.5%
Claude Sonnet 5	≥2	67.9%	44.9%	8.4%	94.9%

PPV was below 10% for every estimator, whereas NPV was approximately 94%–95%. These values are strongly influenced by the cohort's 7.0% sepsis prevalence and do not support use of any estimator as a stand-alone rule-in test.

D. Early and Later ICU Windows

The 1–6-hour subgroup contained 32,594 assessments, and the 7–24-hour subgroup contained 6,088. A further 552 assessments occurred after hour 24 and were omitted only from this timing analysis. AUROC point estimates were slightly lower in the 7–24-hour group for all four estimators, although the confidence intervals overlapped. The rank order was unchanged.

TABLE IV AUROC BY ICU ASSESSMENT WINDOW

Model	Early (1–6 h) AUROC (95% CI)	Late (7–24 h) AUROC (95% CI)	N early	N late
GPT-5.5	0.599 (0.587–0.611)	0.587 (0.560–0.614)	32,594	6,088
NEWS2-derived score	0.593 (0.579–0.605)	0.585 (0.558–0.613)	32,594	6,088
Claude Sonnet 5	0.593 (0.579–0.605)	0.586 (0.558–0.612)	32,594	6,088
SOFA-derived score	0.572 (0.559–0.585)	0.562 (0.534–0.592)	32,594	6,088

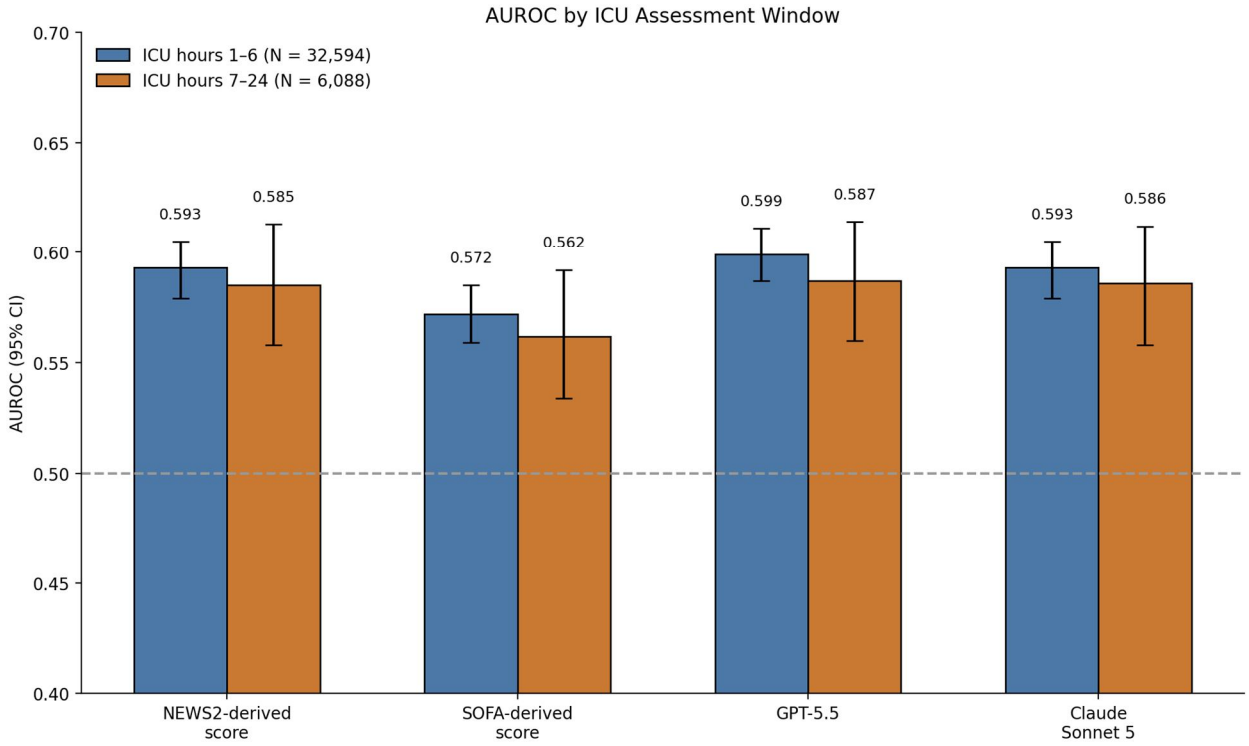


Fig. 3. AUROC estimates by timing of the first complete ICU assessment. Error bars show 95% bootstrap confidence intervals. The analysis excludes 552 assessments obtained after ICU hour 24.

E. Sensitivity Analysis: Pre-Onset-Only Records

Because the primary analysis did not require the assessment snapshot to precede the first positive SepsisLabel, a planned sensitivity analysis was conducted to isolate a genuinely pre-onset subset. Using the hourly SepsisLabel trajectory, the first hour of sepsis positivity was identified for every sepsis-positive patient. Records for which the selected snapshot occurred at or after this hour (576 of 2,733 sepsis-positive patients, 21.1%) were excluded; all sepsis-negative records were retained unchanged. This yielded a pre-onset-only cohort of 38,658 patients (2,157 sepsis-positive, 5.58% prevalence).

In this pre-onset-only cohort, AUROC point estimates for all four estimators converged: 0.591 (95% CI, 0.579–0.604) for the SOFA-derived score, 0.586 (95% CI, 0.576–0.600) for GPT-5.5, 0.584 (95% CI, 0.574–0.597) for Claude Sonnet 5, and 0.579 (95% CI, 0.568–0.592) for the NEWS2-derived score. No pairwise DeLong comparison reached significance (all $p > 0.5$), including both LLMs versus the SOFA-derived score. Precision–recall AUC, reflecting the lower prevalence in this subset, was 0.082 for the SOFA-derived score, 0.074 for GPT-5.5, 0.073 for the NEWS2-derived score, and 0.072 for Claude Sonnet 5.

TABLE V PRE-ONSET-ONLY SENSITIVITY ANALYSIS (N = 38,658)

Model	AUROC (95% CI)	PR-AUC	p vs. SOFA-derived score
SOFA-derived score	0.591 (0.579–0.604)	0.082	Reference
GPT-5.5	0.586 (0.576–0.600)	0.074	0.898
Claude Sonnet 5	0.584 (0.574–0.597)	0.072	0.807
NEWS2-derived score	0.579 (0.568–0.592)	0.073	0.688

PR-AUC = area under the precision–recall curve. No pairwise DeLong comparison reached significance at $\alpha = 0.05$. This convergence indicates that the SOFA-derived score's underperformance in the primary, full-cohort analysis was driven in meaningful part by the inclusion of post-onset sepsis-positive records, on which physiological derangement—and therefore all four estimators' scores—would be expected to be more pronounced and easier to rank correctly.

Among the 2,157 pre-onset-only sepsis-positive patients, the achieved lead time between the assessment snapshot and first sepsis positivity had a median of 39 hours (IQR, 14–85 hours); 89.1% had a lead time of at least 6 hours and 94.9% had a lead time of at least 3 hours. A stricter sensitivity cohort was therefore constructed, retaining only sepsis-positive patients with a lead time of 6 hours or more (1,922 of 2,733 patients, 70.3% of all sepsis-positive patients; 38,423 total records, 5.00% prevalence). In this stricter cohort, the SOFA-derived score obtained the highest point estimate (AUROC 0.597, 95% CI 0.585–0.608), followed by GPT-5.5 (0.578, 95% CI 0.566–0.591), Claude Sonnet 5 (0.577, 95% CI 0.564–0.590), and the NEWS2-derived score (0.573, 95% CI 0.559–0.585). No pairwise DeLong comparison against the SOFA-derived score reached significance (NEWS2: $p = 0.071$; GPT-5.5: $p = 0.219$; Claude Sonnet 5: $p = 0.223$). Taken together, the three cohort definitions used in this study—full record-level, pre-onset-only, and strict ≥6-hour lead time—were deliberately reported without selecting the definition most favourable to any single estimator, and the ranking of estimators was not stable across them: the SOFA-derived score performed worst under the record-level definition and best under the strict lead-time definition, while the LLMs and the NEWS2-derived score showed the reverse pattern.

V. DISCUSSION

A. *Main Findings*

In this matched-cohort benchmark, both zero-shot LLMs ranked sepsis-positive records about as well as the NEWS2-derived score and better than the SOFA-derived score in the primary, full-cohort analysis. The numerical differences were small. Critically, a pre-onset-only sensitivity analysis (Section IV.E) showed all four estimators converging, with no significant pairwise differences, indicating that the SOFA-derived score's apparent underperformance was driven in part by the inclusion of post-onset records in which physiological derangement was already more pronounced. The result is therefore not evidence that an LLM solved sepsis detection; it shows that, from the sparse snapshot supplied, the LLMs recovered an amount of rank-order signal broadly comparable to the clinical-score approximations across both the full cohort and the stricter pre-onset subset.

The two clinical comparators are not equally matched to the LLM input in feature content. The NEWS2-derived score and the LLM prompts draw on an overlapping set of vital-sign and demographic variables, making the LLM-versus-NEWS2 comparison approximately feature-matched. The SOFA-derived score additionally incorporates laboratory values (platelets, bilirubin, creatinine) that were not supplied to either LLM. The LLMs' numerical parity with the SOFA-derived score in the primary analysis, and their convergence with it in both sensitivity analyses, was therefore achieved without access to laboratory data that the SOFA-derived score used directly. This asymmetry should be read as working against, not for, the LLMs: had laboratory values been included in the LLM prompt, discrimination might have improved further, or might not, and this remains untested. The feature-matched NEWS2 comparison is accordingly the more interpretable primary benchmark in this study, with the SOFA comparison retained for clinical context rather than as a like-for-like test.

Using one cohort and one outcome for all four estimators makes the AUROC comparison easier to interpret. It does not, however, create a feature-matched contest. The LLM prompt contained age, sex, and vital signs, whereas the SOFA-derived score also used laboratory measurements. The analysis should be read as a cohort-level benchmark of four estimators, not as a controlled comparison of identical inputs.

B. What Comparable AUROC Does and Does Not Mean

A bedside score is deterministic: the same inputs produce the same total, and its thresholds can be inspected. An LLM is more flexible, but its output depends on model version, product routing, prompt construction, and inference settings. A similar AUROC does not make the approaches operationally interchangeable. Clinical evaluation would require a locked implementation, calibration, repeatability testing, audit logs, and prospective assessment in the intended workflow.

The absolute AUROCs, which ranged from 0.570 to 0.597, also call for restraint. They are above chance but well below what would be expected from a stand-alone clinical prediction tool. The sparse snapshot made the task deliberately difficult, yet that does not lessen the performance needed for safe clinical use. At most, these findings motivate further study of an additional decision-support signal.

C. Model Choice Changes the Alert Burden

The most visible practical difference concerned sensitivity and specificity. Claude Sonnet 5 identified roughly two thirds of sepsis-positive records, but its lower specificity would create substantially more false alerts. GPT-5.5 produced a more even trade-off. Neither pattern is automatically preferable: a screening pathway may prioritise sensitivity, whereas an ICU already burdened by alarms may be unable to tolerate poor specificity.

This is why AUROC should not stand alone in clinical LLM reporting. Models can rank patients similarly overall yet behave differently at the threshold that prompts action. Threshold selection should reflect the intended use, staffing and alert capacity, outcome prevalence, and the relative consequences of false-negative and false-positive decisions. PPV and NPV are also necessary because prevalence changes the meaning of an alert.

D. Timing of Assessment

All four estimators had slightly higher AUROC point estimates in the 1–6-hour subgroup than in the 7–24-hour subgroup. The overlapping confidence intervals do not support a firm timing effect, but the unchanged ordering suggests that the overall comparison was not driven by one of these windows. Patients whose first complete observations appeared after 24 hours were not included in this subgroup analysis, so the timing result does not cover the entire cohort.

E. Limitations

Several limitations affect interpretation. First, this was a retrospective analysis of one public ICU dataset, and the results may not generalise to other hospitals or to emergency, ward, and community settings. Second, each estimator was reduced to one cross-sectional assessment. Trends, treatments, symptoms, clinical notes, and laboratory trajectories, which all influence real sepsis assessment were not available to the LLMs.

Third, the clinical comparators were nonstandard. The NEWS2-derived score omitted consciousness and supplemental oxygen and added mean arterial pressure; the SOFA-derived score omitted domains and replaced the standard respiratory component with oxygen saturation. The feature sets also differed across estimators. These results must not be presented as validation of complete NEWS2 or SOFA.

Fourth, the primary outcome was positivity anywhere in the ICU record, and the selected snapshot was not required to fall before the first positive SepsisLabel, so the primary analysis measures record-level discrimination rather than a prespecified prediction horizon. The pre-onset-only sensitivity analysis (Section IV.E) partially addresses this by excluding the 21.1% of sepsis-positive records for which the snapshot occurred at or after onset, but it still uses a variable, data-driven lead time rather than a fixed horizon (e.g., 6 hours pre-onset for every patient); a prespecified fixed-horizon design remains a priority for future work. Fifth, the LLMs were accessed through evolving chat products and were not tested repeatedly for within-model variability. Finally, Youden thresholds were selected and evaluated in the same cohort, p values were unadjusted for multiplicity, and the analysis did not address calibration, fairness, clinical utility, alert fatigue, or patient outcomes.

VI. CONCLUSION

In 39,234 ICU records, two zero-shot LLMs assigned ordinal scores that discriminated record-level sepsis status about as well as a NEWS2-derived score and better than a SOFA-derived score. A pre-onset-only sensitivity analysis on 38,658 records (median achieved lead time 39 hours) showed all four estimators converging, with no significant pairwise differences, indicating that the primary-analysis gap over the SOFA-derived score was driven in part by post-onset records rather than a stable advantage under stricter early-prediction conditions. A further, stricter sensitivity analysis restricted to a minimum 6-hour lead time reversed the

ranking, with the SOFA-derived score obtaining the highest point estimate; this instability across cohort definitions is itself an informative finding, indicating that the relative ranking of these estimators is not robust to the prediction horizon considered and should not be treated as a fixed property of any single model. Across both analyses, the two LLMs' overall AUROCs were similar to each other, but their sensitivity-specificity trade-offs differed enough to imply very different alert burdens.

These results support further evaluation of LLM-derived signals within critical-care research, not stand-alone bedside use. Future studies should define a fixed, prespecified pre-onset prediction horizon applied uniformly across all patients, use complete and feature-matched comparators, lock model versions, publish scoring and missing-data rules, assess repeatability and calibration, and validate prespecified thresholds prospectively before considering clinical integration.

REFERENCES

[1] K. E. Rudd et al., "Global, regional, and national sepsis incidence and mortality, 1990–2017: Analysis for the Global Burden of Disease Study," Lancet, vol. 395, no. 10219, pp. 200–211, 2020, doi: 10.1016/S0140-6736(19)32989-7.

[2] A. Kumar et al., "Duration of hypotension before initiation of effective antimicrobial therapy is the critical determinant of survival in human septic shock," Crit. Care Med., vol. 34, no. 6, pp. 1589–1596, 2006, doi: 10.1097/01.CCM.0000217961.75225.E9.

[3] Royal College of Physicians, National Early Warning Score (NEWS) 2: Standardising the Assessment of Acute-Illness Severity in the NHS: Updated Report of a Working Party. London, U.K.: RCP, 2017. [Online]. Available: https://www.rcp.ac.uk/resources/national-early-warning-score-news-2/

[4] M. Singer et al., "The Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3)," JAMA, vol. 315, no. 8, pp. 801–810, 2016, doi: 10.1001/jama.2016.0287.

[5] K. Singhal et al., "Large language models encode clinical knowledge," Nature, vol. 620, no. 7972, pp. 172–180, 2023, doi: 10.1038/s41586-023-06291-2.

[6] Z. Kanjee, B. Crowe, and A. Rodman, "Accuracy of a generative artificial intelligence model in a complex diagnostic challenge," JAMA, vol. 330, no. 1, pp. 78–80, 2023, doi: 10.1001/jama.2023.8288.

[7] J. Achiam et al., "GPT-4 Technical Report," arXiv:2303.08774, 2023, doi: 10.48550/arXiv.2303.08774.

[8] G. B. Haim et al., "Evaluating large language model-assisted emergency triage: A comparison of acuity assessments by GPT-4 and medical experts," J. Clin. Nurs., published online Nov. 28, 2024, doi: 10.1111/jocn.17490.

[9] A. Zaboli et al., "Chat-GPT in triage: Still far from surpassing human expertise: An observational study," Am. J. Emerg. Med., vol. 92, pp. 165–171, 2025, doi: 10.1016/j.ajem.2025.03.028.

[10] O. A. Usman, A. A. Usman, and M. A. Ward, "Comparison of SIRS, qSOFA, and NEWS for the early identification of sepsis in the emergency department," Am. J. Emerg. Med., vol. 37, no. 8, pp. 1490–1497, 2019, doi: 10.1016/j.ajem.2018.10.058.

[11] R. Goulden et al., "qSOFA, SIRS and NEWS for predicting inhospital mortality and ICU admission in emergency admissions treated as sepsis," Emerg. Med. J., vol. 35, no. 6, pp. 345–349, 2018, doi: 10.1136/emermed-2017-207120.

[12] J. L. Vincent et al., "The SOFA (Sepsis-related Organ Failure Assessment) score to describe organ dysfunction/failure," Intensive Care Med., vol. 22, no. 7, pp. 707–710, 1996, doi: 10.1007/BF01709751.

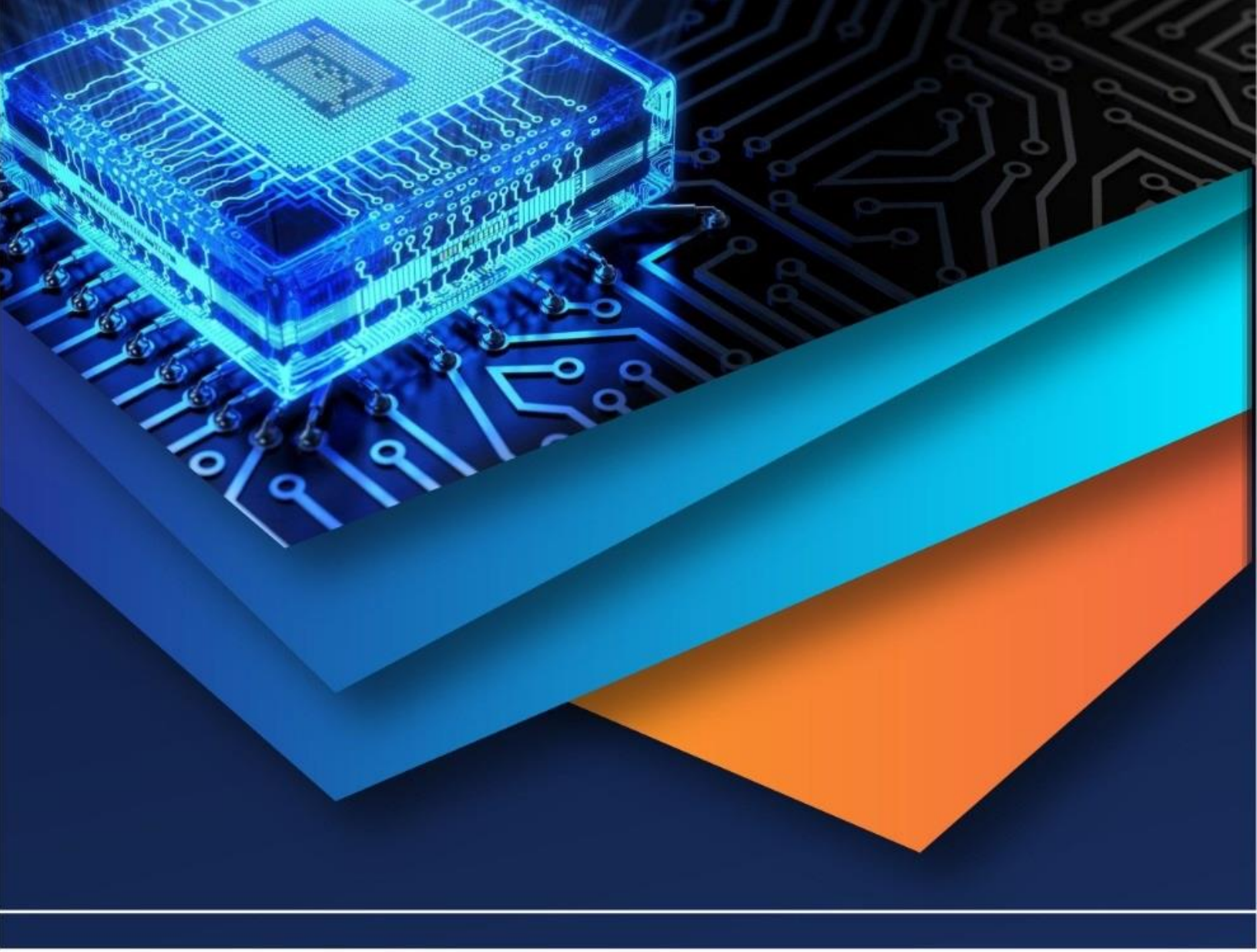
[13] M. A. Ansari Khoushabar and P. Ghafariasl, "Advanced meta-ensemble machine learning models for early and accurate sepsis prediction to improve patient outcomes," arXiv:2407.08107, 2024, doi: 10.48550/arXiv.2407.08107.

[14] Q. Mao et al., "Multicentre validation of a sepsis prediction algorithm using only vital sign data in the emergency department, general ward and ICU," BMJ Open, vol. 8, no. 1, e017833, 2018, doi: 10.1136/bmjopen-2017-017833.

[15] G. S. Collins et al., "TRIPOD+AI statement: Updated guidance for reporting clinical prediction models that use regression or machine learning methods," BMJ, vol. 385, e078378, 2024, doi: 10.1136/bmj-2023-078378.

[16] M. A. Reyna et al., "Early prediction of sepsis from clinical data: The PhysioNet/Computing in Cardiology Challenge 2019," Crit. Care Med., vol. 48, no. 2, pp. 210–217, 2020, doi: 10.1097/CCM.0000000000004145.

[17] E. R. DeLong, D. M. DeLong, and D. L. Clarke-Pearson, "Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach," Biometrics, vol. 44, no. 3, pp. 837–845, 1988, doi: 10.2307/2531595.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY