A Prospective Comparison of Large Language Models for Early Prediction of Sepsis[1]

Supreeth P. Shashikumar* and Shamim Nemati
Division of Biomedical Informatics, University of California San Diego
La Jolla, California, USA
Email: spshashikumar, snemati@health.ucsd.edu

We present a comparative study on the performance of two popular open-source large language models for early prediction of sepsis: Llama-3 8B and Mixtral 8x7B. The primary goal was to determine whether a smaller model could achieve comparable predictive accuracy to a significantly larger model in the context of sepsis prediction using clinical data.

Our proposed LLM-based sepsis prediction system, COMPOSER-LLM, enhances the previously published COMPOSER model, which utilizes structured EHR data to generate hourly sepsis risk scores. The new system incorporates an LLM-based approach to extract sepsis-related clinical signs and symptoms from unstructured clinical notes. For scores falling within high-uncertainty prediction regions, particularly those near the decision threshold, the system uses the LLM to draw additional clinical context from patient notes; thereby enhancing the model's predictive accuracy in challenging diagnostic scenarios.

A total of 2,074 patient encounters admitted to the Emergency Department at two hospitals within the University of California San Diego Health system were used for model evaluation in this study. Our findings reveal that the Llama-3 8B model based system (COMPOSER-LLM$_{Llama}$) achieved a sensitivity of 70.3%, positive predictive value (PPV) of 32.5%, F-1 score of 44.4% and false alarms per patient hour (FAPH) of 0.0194, closely matching the performance of the larger Mixtral 8x7B model based system (COMPOSER-LLM$_{mixtral}$) which achieved a sensitivity of 72.1%, PPV of 31.9%, F-1 score of 44.2% and FAPH of 0.020. When prospectively evaluated, COMPOSER-LLM$_{Llama}$ demonstrated similar performance to the COMPOSER-LLM$_{mixtral}$ pipeline, with a sensitivity of 68.7%, PPV of 36.6%, F-1 score of 47.7% and FAPH of 0.019 vs. sensitivity of 70.5%, PPV of 36.3%, F-1 score of 47.9% and FAPH of 0.020. This result indicates that, for extraction of clinical signs and symptoms from unstructured clinical notes to enable early prediction of sepsis, the Llama-3 generation of smaller language models can perform as effectively and more efficiently than larger models. This finding has significant implications for healthcare settings with limited resources.

Keywords: Large language model, Unstructured clinical notes, Clinical decision support systems

[1*] This work is supported by the National Institutes of Health (#R01LM013998, #R35GM143121, #R42AI177108). The opinions or assertions contained herein are the private ones of the author and are not to be construed as official or reflecting the views of the NIH or any other agency of the US Government.

© 2024 The Authors. Open Access chapter published by World Scientific Publishing Company and distributed under the terms of the Creative Commons Attribution Non-Commercial (CC BY-NC) 4.0 License.

1. Introduction

Sepsis is a life-threatening condition that arises when the body's response to infection causes systemic inflammation, leading to tissue damage and organ failure. It is a major cause of mortality and morbidity worldwide, accounting for a significant portion of hospital deaths[1–3]. Early detection and timely intervention are critical to improving patient outcomes, as delayed treatment can lead to severe complications and increased mortality[4–6]. Recent advancements in artificial intelligence (AI) have enabled the development of predictive models that utilize electronic health record (EHR) data to identify early signs of sepsis[7]. These AI-driven models can analyze vast amounts of structured data, such as laboratory results and vital signs, to predict sepsis risk and prompt early clinical intervention. Despite the success of these models, they often overlook the rich contextual information embedded in unstructured clinical notes, which can provide additional insights into a patient's condition.

Large language models (LLMs) have emerged as powerful tools for processing and interpreting unstructured text data, making them valuable assets in predictive analytics for healthcare[8]. LLMs, such as GPT-3, Claude and their variants, are pre-trained on extensive text corpora and fine-tuned for specific tasks. In healthcare, LLMs have shown promise in tasks ranging from generating clinical notes and summarizing patient histories to identifying clinical entities and predicting patient outcomes[9–12]. The integration of LLMs with traditional AI models has the potential to improve predictions by incorporating nuanced information from unstructured data, thereby providing a more comprehensive view of a patient's health status.

However, the deployment of large LLMs in clinical settings presents significant challenges. Large models, such as the Mixtral 8x7B model with 47 billion parameters, require substantial computational resources for training and inference, which can be prohibitive in resource-constrained environments. The motivation for this study was to explore the feasibility of using a new generation smaller LLM, specifically the Llama-3 model with 8 billion parameters, to achieve comparable performance for extraction of clinical signs and symptoms from unstructured clinical notes. By reducing the model size, we aim to address the issues of computational efficiency, scalability, and cost, while maintaining or even improving predictive accuracy.

2. Methods

2.1. *Data*

This study utilized de-identified data from the electronic health records (EHR) of patient encounters in the Emergency Department (ED) at two University of California San Diego (UCSD) Health hospitals, using FHIR and HL7v2 standards. Patients were identified as having sepsis according to the Sepsis-3 international consensus definition for sepsis[2]. The onset time of sepsis was established by following previously published methodology, using evidence of organ dysfunction and suspicion of clinical infection[13–15]. Patients aged 18 and older were monitored throughout their stay until either their first episode of sepsis, transition to comfort care, or transfer out of the ED. To ensure a sufficient quantity of predictor data, we focused on sequential hourly predictions of sepsis starting two hours after ED triage. While the previously established COMPOSER model[14,16] used a decision threshold of 0.6, the COMPOSER-LLM model adopted a lower threshold of 0.5 to enhance sensitivity. To mitigate the potential increase in false alarms, the model incorporated additional contextual information from clinical notes for predictions in the high-uncertainty range of 0.5-0.75. To assess the impact of lowering the decision threshold and to explore the advantages of using an LLM for uncertain predictions, all patients with at least one COMPOSER risk score above 0.5 were included for further analysis. Exclusions were made for patients identified as having sepsis before the prediction start-time or those lacking heart rate or blood pressure measurements prior to this time. Predictions were considered if the following criteria were met: 1) At least one vital sign and lab measurement within the past 24 hours; 2) No antibiotics received; and 3) Availability of an "ED provider note" or "H&P note."

The *retrospective cohort* included ED patient encounters from October 1, 2023 to December 31, 2023. A total of 1320 ED encounters (16.3% septic) met the inclusion criteria for the *retrospective cohort*. Additionally, the COMPOSER-LLM pipeline was prospectively deployed in silent mode for real-time sepsis prediction in the two EDs within the UCSD Health system starting from May 1, 2024. The prospective data collected during the time period of May 1 - June 15 2024 will be referred to as *prospective cohort*. A total of 754 ED encounters (18.4% septic) met the inclusion criteria for the *prospective cohort*. Table T1 in the appendix shows baseline characteristics and summary characteristics for the *retrospective* and *prospective* cohorts.

This investigation was conducted according to University of California San Diego IRB approved protocol #805726 with a waiver of informed consent.

2.2. *COMPOSER-LLM*

The schematic diagram of the entire COMPOSER-LLM pipeline is shown in Figure 1. Starting from the time of ED admission, COMPOSER[14,16] generated a sepsis risk score at an hourly resolution. We direct the reader to Shashikumar et al.[16] for more details regarding the input features (structured data) of COMPOSER. If the risk score exceeded a primary decision threshold (θ_1=0.75), an alert was fired. Risk scores closer to a secondary decision threshold (θ_2=0.50) were often associated with false alarms. Consequently, for risk scores within the high-uncertainty region ($\theta_1 \geq$ risk score $>\theta_2$), an LLM-based sepsis likelihood tool was utilized to enhance diagnostic accuracy. Specifically, if the likelihood score surpassed a predetermined likelihood-based decision threshold (α) and the LLM indicated a 'suspicion of bacterial infection,' an alert was fired.

2.2.1. *Sepsis likelihood tool*

The sepsis likelihood tool was designed to improve diagnostic accuracy by confirming the presence of sepsis-related clinical signs and symptoms documented in clinical notes. It first utilized a large language model (LLM) to extract these signs or symptoms from the notes. The extracted symptoms were then processed through a likelihood calculator to assess the probability of sepsis. This calculated likelihood was subsequently used to confirm diagnosis of sepsis.

A Bayesian likelihood calculator was used to compute the likelihood of sepsis based on the clinical signs or symptoms identified by the LLM. The posterior probability of sepsis given a set of clinical signs or symptoms ($\left\{CS_i\right\}$, $i \in 1....M$), $P(D|CS)$, was calculated as follows: $P(D|CS) \ = \ \frac{P(D)\cdot P(CS|D)}{P(CS)}$. Where, $CS_i \ = \ 1$ corresponded to the scenario under which the clinical signs or symptom CS_i was identified to be present by the LLM pipeline. The set of sepsis-related clinical signs or symptoms used in this study ($\left\{CS_i\right\}$, $i \in 1....9$) were as follows: *fever, hypotension, tachypnea, tachycardia, altered mental status, elevated inflammatory markers, positive blood culture, suspicion of bacterial infection, and organ dysfunction syndrome.*

The likelihood values for each of the clinical signs or symptoms conditioned on sepsis have been tabulated in Table T2 of Appendix.

2.2.2. *LLM-based clinical sign or symptom extractor:*

The LLM-based clinical signs or symptoms extraction pipeline was designed to accept a prompt and clinical notes (all notes generated from admission to the time of prediction) as input and generate a JSON-formatted text output. Given that clinical notes can sometimes exceed the predefined input length (context size) of the LLM, we employed the retrieval augmented generation (RAG) technique[17] to extract smaller, relevant text chunks (context) for the queried clinical sign or symptom. These extracted text chunks were then appended to the input prompt for the LLM. The prompt used in our analysis was as follows:

"You are an ED doctor. Your task is to identify the following abnormal clinical signs and symptoms: {clinical sign or symptom}. Think step-by-step and provide your response in the following JSON format: {<clinical sign or symptom> : ["Yes or No", "Concise justification?"]}
Medical note: {RAG context}."

To minimize hallucinations and to maintain consistency in text generation, the temperature parameter of the LLM was set to 0.3. Additionally, for each clinical sign or symptom, the LLM pipeline was run three times and the majority outcome (clinical sign or symptom present or not) across the multiple runs was used for downstream tasks.

Recent advancements, including enhanced training data, advanced architecture (such as group query attention), improved tokenization, and refined training techniques, have enabled newer generations of smaller parameter LLMs (such as Llama-3 8B[18]) to match or even surpass the performance of older, larger models (such as Llama 2 70B[19] and Mixtral 8x7B[20]). In this study, we used the open-source Llama-3 8B and Mixtral 8x7B models to investigate whether the newer, smaller LLM (Llama-3 8B) can achieve comparable effectiveness in extracting clinical signs and symptoms from unstructured clinical notes to the much larger Mixtral 8x7B model. Specifically, the Mixtral 8x7B, developed by Mistral AI, is a sparse mixture of experts LLM with 46.7 billion total parameters, referred to as COMPOSER-LLM$_{Mixtral}$ in this study. The Llama-3 8B, the latest state-of-the-art LLM developed by Meta with 8 billion parameters, is referred to as COMPOSER-LLM$_{Llama}$ in our analysis.

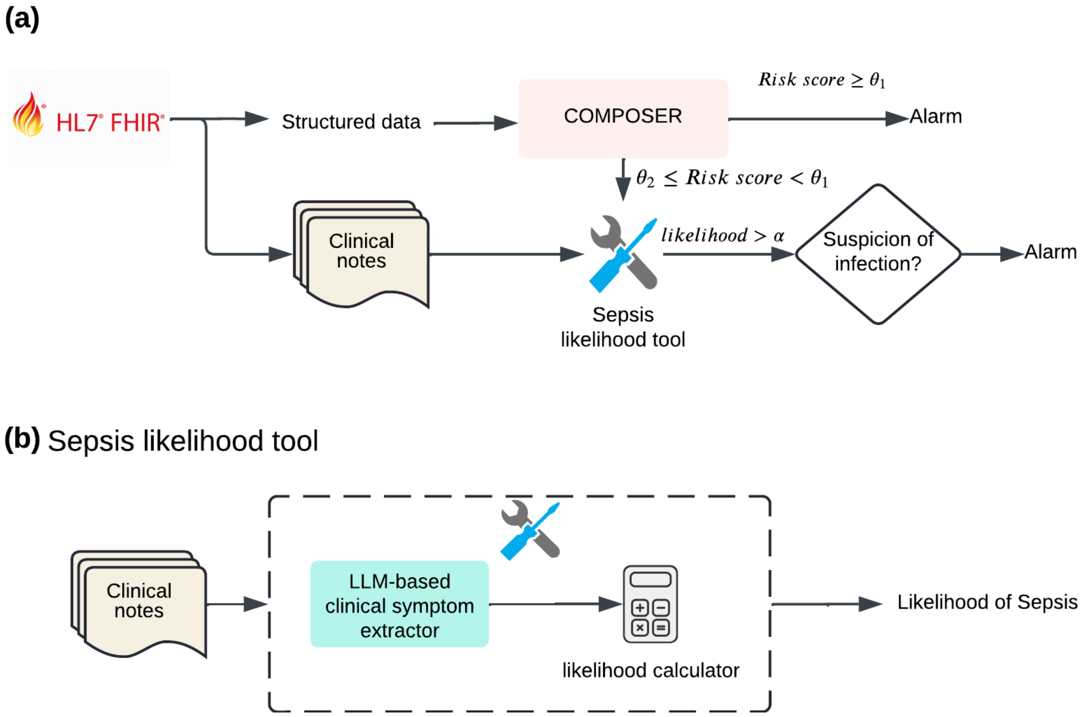

Figure 1. Schematic Diagram of the COMPOSER-LLM pipeline.

2.3. *Experimental setup and evaluation*

For all continuous variables, we have reported medians ([25th–75th percentile]). For binary variables, we have reported percentages. Differences between the septic and non-septic cohort were assessed with Wilcoxon rank sum tests on continuous variables and Pearson's chi-squared tests on categorical variables and significance was assessed at a p-value of 0.05. Sensitivity (SEN), positive predictive value (PPV), and F-1 score at a fixed decision threshold have been reported at the encounter level. SEN, PPV and F-1 score were reported under an end-user clinical response policy in which alarms fired up to 48 hours prior to onset of sepsis were considered as true alarms, and the model was silenced for six hours after an alarm was fired. Additionally, we have reported false alarms per patient hour (FAPH) which can be used to calculate the expected number of false alarms per unit of time in a typical care unit (e.g., a FAPH of 0.025 translates to roughly 1 alarm every 2 h in a 20-bed care unit). The FAPH was calculated by dividing the total number of false alarms by the total number of data points (sum of hourly time points across all patients) in a given cohort.

The COMPOSER model was implemented in TensorFlow. The LLM-based clinical signs or symptom extraction pipeline was implemented using the LangChain framework in Python. The LLM pipeline was run on AWS multi-GPU EC2 instance with NVIDIA A10G GPUs: g5.12xlarge ec2 instance type (cost of $5.672 per hour) for Mixtral 8x7B, g5.2xlarge ec2 instance type (cost of $1.212 per hour) for Llama-3 8B.

2.4. *Prospective deployment:*

COMPOSER-LLM$_{Mixtral}$ and COMPOSER-LLM$_{Llama}$ model were prospectively deployed in silent-mode on a cloud-based platform, as previously described by Boussina et al.[14]. Prospective validation studies are essential in clinical applications of LLMs as retrospective performance may not accurately reflect real-world performance due to factors such as incomplete or missing clinical notes. The real-time platform extracted data at an hourly resolution of all the active patients (across the two Emergency Departments within UCSD Health system) using FHIR APIs with OAuth 2.0 authentication, and passed the input feature set to the COMPOSER-LLM inference engine. The inference engine consisted of COMPOSER microservice and Sepsis likelihood tool microservice hosted within separate EC2 instances. The sepsis risk scores generated by the COMPOSER-LLM pipeline were then written to a flowsheet within the EHR using an HL7v2 outbound message. The flowsheet then triggered a nurse-facing Best Practice Advisory (BPA) that alerted the caregiver that the patient was at risk of developing severe sepsis. As the models were deployed in silent mode, the BPA was not shown to the end-user. The COMPOSER-LLM pipeline was deployed for real-time prediction of sepsis across the two EDs within the UCSD Health system starting from May 1, 2024.

3. Results

Table 1: Comparison of model performance.

		COMPOSER	COMPOSER-LLM$_{Mixtral}$ (Mixtral 8x7B)	COMPOSER-LLM$_{Llama}$ (Llama-3 8B)
Retrospective cohort	**Sensitivity**	72.9%	72.1%	70.3%
	PPV	22.6%	31.9%	32.5%
	F1-Score	34.5%	44.2%	44.4%
	FAPH	0.037	0.021	0.0194
Prospective cohort	**Sensitivity**	70.8%	70.5%	68.7%
	PPV	25.1%	36.3%	36.6%
	F1-Score	37.1%	47.9%	47.7%
	FAPH	0.034	0.020	0.019

The standalone COMPOSER model achieved a sensitivity of 72.9%, positive predictive value (PPV) of 22.6%, F-1 score of 34.5%, and FAPH of 0.037 on the retrospective cohort. In comparison, COMPOSER-LLM$_{Mixtral}$ demonstrated improved performance with a sensitivity of 72.1%, PPV of 31.9%, F-1 score of 44.2%, and FAPH of 0.021. Similarly, COMPOSER-LLM$_{Llama}$ achieved a sensitivity of 70.3%, PPV of 32.5%, F-1 score of 44.1%, and FAPH of 0.0194. Additionally, outputs generated by both the LLMs for the same patient are shown in Table 2.

Within the *prospective cohort*, the performance of COMPOSER-LLM$_{Llama}$ was found to be similar to COMPOSER-LLM$_{Mixtral}$. Specifically, COMPOSER-LLM$_{Llama}$ achieved a sensitivity of 68.7%, PPV of 36.6%, F-1 score of 47.7%, and FAPH of 0.0190 whereas COMPOSER-LLM$_{Mixtral}$ achieved a sensitivity of 70.5%, PPV of 36.3%, F-1 score of 47.9% and FAPH of 0.020.

Table 2: Outputs generated by the LLM for various clinical signs and symptoms for a patient

	COMPOSER-LLM$_{Mixtral}$ (Mixtral 8x7B)	COMPOSER-LLM$_{Llama}$ (Llama-3 8B)
Fever	['Yes', 'Temperature recorded as high as 102.7 degF (4 days ago), and 101degF (today)']	['Yes', 'Temperature of 102.7 degF is greater than or equal to 100.4 degF.']
Hypotension	['No', 'Systolic BP > 100 mmHg, normotensive']	['No', 'BP 122/94 is not less than 90/60.']
Tachypnea	['No', 'Respiratory rate 16-20, within normal limits']	['No', 'Resp rate of 20 is not greater than 20.']
Tachycardia	['No', 'Pulse rate 77-83, within normal limits']	['No', 'Pulse of 83 is not greater than 90.']

Altered mental status	['No', "No evidence of altered mental status in the patient's medical note."]	['No', 'No mention of altered mental status in the medical note.']
Elevated inflammatory markers	['Yes', 'Justification: The CBC with diff shows an elevated white blood cell count (16.1) and increased Imm Gran % (1), indicating possible infection or inflammation.']	['Yes', 'Elevated WBC (16.1) and Anion Gap (16) indicating possible inflammatory response.']
Positive blood culture	['No', 'No information provided about blood culture results in the medical note.']	['No', 'No mention of positive blood culture in the medical note.']
Suspicion of bacterial infection	['Yes', 'Elevated WBC, fevers, nausea, vomiting, worsening right flank and right-sided abdominal pain, and recent instillations into nephrostomy tube suggest possibility of infection.']	['Yes', 'Rising WBC of 16.1, worsening renal insufficiency, and fevers, along with a history of ESBL Kleb from a urine culture 3 days ago, suggest suspicion of bacterial infection.']
Organ dysfunction	['Yes', 'Abnormal renal function with history of CKD, abnormal vital signs (hypotension and tachycardia), abnormal white cell count (leukocytosis)']	['No', "The patient's physical exam and diagnostic testing results do not indicate any organ dysfunction. The patient's vital signs are within normal limits, and the physical exam is notable for moderate to significant tenderness in the right side of the abdomen, but no other abnormalities. "]

4. Discussion

The findings from this study highlight the potential of using the new generation of smaller open-source LLMs for enhancing the early sepsis prediction. The results demonstrated that the Llama-3 8B model (COMPOSER-LLM$_{Llama}$) achieved comparable performance to the larger Mixtral 8x7B model (COMPOSER-LLM$_{Mixtral}$), including sensitivity, positive predictive value (PPV), and F-1 score, with slightly fewer false alarms per patient hour (FAPH). When prospectively evaluated, the COMPOSER-LLM$_{Llama}$ pipeline showed similar performance to the COMPOSER-LLM$_{Mixtral}$ pipeline. These outcomes suggest that, for extraction of clinical signs and symptoms from unstructured clinical notes, the Llama-3 generation of smaller language models can perform as effectively and more efficiently than larger models, providing a more efficient and cost-effective solution for real-time clinical decision support systems.

The new generation of smaller LLMs, such as the Llama-3 8B, possess several advantageous properties over older, larger models like the Mixtral 8x7B. These smaller models have been optimized with improved training data, advanced architectural techniques, and enhanced tokenization methods. Despite their reduced parameter size, these advancements allow smaller models to perform at par or even surpass the performance of older, more extensive models[21]. One of the most significant advantages of smaller LLMs is their lower computational resource requirement, making them more accessible and scalable for deployment in resource-constrained environments. The reduction in computational overhead (ec2 instance cost of \$5.672 per hour for Mixtral 8x7B vs \$1.212 per hour for Llama-3 8B) also translates into lower operational costs and faster inference times.

However, this study has several limitations. The sepsis likelihood tool was triggered only after the availability of certain clinical notes ("ED provider note" or "H&P note"), potentially delaying alert generation among patients in the uncertainty interval of 0.5-0.75. However, during the prospective deployment of COMPOSER-LLM, the tool was triggered even if a note was incomplete, as the contextual information within these notes still provided valuable insights. Future research could investigate using LLM-based queries to extract essential patient and provider information, such as suspicion of infection, and explore real-time capture of provider notes through speech recognition and transcription to address issues with missing or incomplete notes. Additionally, while the models were tested on data from two hospitals within a single health system, the generalizability of these findings to other institutions with different patient populations or clinical practices may be limited. Finally, future prospective studies (such as randomized clinical trials) are needed to assess the impact of COMPOSER-LLM on patient care and outcomes.

5. Conclusion

This study demonstrated that a new generation smaller LLM, the Llama-3 8B model (with 8 billion parameters), performed as effectively and more efficiently than an older generation larger LLM, the Mixtral 8x7B model (with 47 billion parameters), for extraction of clinical signs and symptoms from unstructured clinical notes to enable early prediction of sepsis. The results advocate for the potential of smaller models in healthcare, offering a more resource-efficient alternative without compromising accuracy.

Acknowledgments

S.N. is funded by the National Institutes of Health (#R01LM013998, #R35GM143121, #R42AI177108). The opinions or assertions contained herein are the private ones of the author and are not to be construed as official or reflecting the views of the NIH or any other agency of the US Government.

Appendix

Table T1. Patient characteristics of the retrospective and prospective cohorts

	Retrospective cohort		Prospective cohort	
	Septic	*Non-Septic*	*Septic*	*Non-Septic*
# Encounters (%)	215 (16.3%)	1,105	139 (18.4%)	615
Age (in years), median [IQR]	64.7 [51.9 – 75.1]	59.1 [44.9 – 71.2]*	65.1 [50.1 – 75.8]	58.1 [43.3 – 69.6] *
Gender (Male), %	55.8%	50.4%	58.2%	47.7%
Race				
White, %	45.5%	43.9%	50.4%	45.6%
African American, %	11.6%	9.9%	7.2%	7.9%
Asian, %	7.9%	6.5%	4.3%	6.4%
ED Length of Stay (in hours), median [IQR]	24.6 [10.9 – 49.4]	11.8 [6.5 – 29.1] *	22.1 [10.5 – 45.8]	11.9 [6.9 – 30.8] *
CCI, median [IQR]	2 [0 -4]	1 [0 – 3]	2 [0 – 5]	1 [0 -2]
SOFA, median [IQR]	3 [1 – 5]	1 [0 – 2]*	3 [1 – 5]	1 [0 – 2] *
In-hospital mortality, %	8.8%	1.5% *	8.1%	1.2% *
Time from ED triage to onset of sepsis (in hours), median [IQR]	3.2 [1.1 – 8.1]	N/A	2.3 [1.2 – 10.2]	N/A

* p-value<0.05

CCI = Charlson Comorbidity Index

SOFA = Sequential Organ Failure Assessment score

Table T2. Likelihood values for each of the clinical symptoms conditioned on sepsis

Clinical symptoms	Probability value
Fever	0.9
Hypotension	0.05
Tachypnea	0.7
Tachycardia	0.05
Altered mental status	0.05
Elevated inflammatory markers	0.5
Positive blood culture	0.9
Suspicion of bacterial infection	0.75
Organ dysfunction	0.45

References

1. Rudd, K. E. *et al.* Global, regional, and national sepsis incidence and mortality, 1990–2017: analysis for the Global Burden of Disease Study. *The Lancet* **395**, 200–211 (2020).
2. Singer, M. *et al.* The third international consensus definitions for sepsis and septic shock (Sepsis-3). *J. Am. Med. Assoc.* **315**, 801–810 (2016).
3. Rhee, C. *et al.* Incidence and trends of sepsis in US hospitals using clinical vs claims data, 2009-2014. *J. Am. Med. Assoc.* **318**, 1241–1249 (2017).
4. Ferrer, R. *et al.* Empiric antibiotic treatment reduces mortality in severe sepsis and septic shock from the first hour: results from a guideline-based performance improvement program. *Crit. Care Med.* **42**, 1749–1755 (2014).
5. Liu, V. X. *et al.* The Timing of Early Antibiotics and Hospital Mortality in Sepsis. *Am. J. Respir. Crit. Care Med.* **196**, 856–863 (2017).
6. Peltan, I. D. *et al.* ED Door-to-Antibiotic Time and Long-term Mortality in Sepsis. *Chest* **155**, 938–946 (2019).
7. Islam, K. R. *et al.* Machine learning-based early prediction of sepsis using electronic health records: a systematic review. *J. Clin. Med.* **12**, 5658 (2023).
8. Thirunavukarasu, A. J. *et al.* Large language models in medicine. *Nat. Med.* **29**, 1930–1940 (2023).
9. Guevara, M. *et al.* Large language models to identify social determinants of health in electronic health records. *NPJ Digit. Med.* **7**, 6 (2024).
10. Wang, H., Gao, C., Dantona, C., Hull, B. & Sun, J. DRG-LLaMA: tuning LLaMA model to predict diagnosis-related group for hospitalized patients. *Npj Digit. Med.* **7**, 16 (2024).
11. Tai-Seale, M. *et al.* AI-Generated Draft Replies Integrated Into Health Records and Physicians' Electronic Communication. *JAMA Netw. Open* **7**, e246565–e246565 (2024).
12. Ayers, J. W. *et al.* Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. *JAMA Intern. Med.* **183**, 589–596 (2023).
13. Seymour, C. W. *et al.* Assessment of clinical criteria for sepsis: for the third international consensus definitions for sepsis and septic shock (Sepsis-3). *J. Am. Med. Assoc.* **315**, 762–774 (2016).
14. Boussina, A. *et al.* Impact of a deep learning sepsis prediction model on quality of care and survival. *Npj Digit. Med.* **7**, 14 (2024).
15. Amrollahi, F. *et al.* Inclusion of social determinants of health improves sepsis readmission prediction models. *J. Am. Med. Inform. Assoc.* **29**, 1263–1270 (2022).
16. Shashikumar, S. P., Wardi, G., Malhotra, A. & Nemati, S. Artificial intelligence sepsis prediction algorithm learns to say "I don't know". *NPJ Digit. Med.* **4**, 134 (2021).
17. Lewis, P. *et al.* Retrieval-augmented generation for knowledge-intensive nlp tasks. *Adv. Neural Inf. Process. Syst.* **33**, 9459–9474 (2020).
18. The Llama-3 Herd of Models | Research - AI at Meta. https://ai.meta.com/research/publications/the-llama-3-herd-of-models/.
19. Touvron, H. *et al.* Llama 2: Open foundation and fine-tuned chat models. *ArXiv Prepr. ArXiv230709288* (2023).

20. Jiang, A. Q. *et al.* Mixtral of Experts. Preprint at http://arxiv.org/abs/2401.04088 (2024).

21. Hassid, M., Remez, T., Gehring, J., Schwartz, R. & Adi, Y. The Larger the Better? Improved LLM Code-Generation via Budget Reallocation. *ArXiv Prepr. ArXiv240400725* (2024).