Title: Integrating a host transcriptomic biomarker with a large language model for diagnosis of lower respiratory tract infection

Authors: †Hoang Van Phan[1], †Natasha Spottiswoode[1], Emily C. Lydon[1], Victoria T. Chu[2,3], Adolfo Cuesta[1], Alexander D. Kazberouk[4], Natalie L. Richmond[1], Carolyn S. Calfee[5], *Charles R. Langelier[1,3]

†equal contributions
*Corresponding author: chaz.langelier@ucsf.edu

Affiliations:
[1] Department of Medicine, Division of Infectious Diseases, University of California San Francisco
[2] Department of Pediatrics, Division of Infectious Diseases and Global Health, University of California San Francisco
[3] Chan Zuckerberg Biohub San Francisco
[4] Department of Medicine, University of California San Francisco
[5] Department of Medicine, Division of Pulmonary, Critical Care, Allergy and Sleep Medicine, University of California San Francisco

Abstract

Lower respiratory tract infections (LRTIs) are a leading cause of mortality worldwide. Despite this, diagnosing LRTI remains challenging, particularly in the intensive care unit, where non-infectious respiratory conditions can present with similar features. Here, we tested a new method for LRTI diagnosis that combines the transcriptomic biomarker *FABP4* with assessment of text from the electronic medical record (EMR) using the large language model Generative Pre-trained Transformer 4 (GPT-4). We evaluated this methodology in a prospective cohort of critically ill adults with acute respiratory failure, in which we measured pulmonary *FABP4* expression and identified patients with LRTI or non-infectious conditions using retrospective adjudication. A diagnostic classifier combining *FABP4* and GPT-4 achieved an area under the receiver operator curve (AUC) of 0.92 ± 0.06 by five-fold cross validation (CV), outperforming classifiers based on *FABP4* expression alone (AUC 0.83) or GPT-4 alone (AUC 0.84). At the Youden's index within each CV fold, the combined classifier achieved a mean sensitivity of 92% ± 7%, specificity of 90% ± 17% and accuracy of 91% +/- 8%. Taken together, our findings suggest that combining a host transcriptional biomarker with interpretation of EMR data using artificial intelligence is a promising new approach to infectious disease diagnosis.

Brief Communication

Lower respiratory tract infections (LRTIs) are a leading cause of death worldwide, yet remain challenging to diagnose[1]. This is especially true in the intensive care unit (ICU), where non-infectious acute respiratory illnesses can have similar clinical manifestations. Further complicating accurate diagnosis is the inability to identify a causative pathogen in most clinically recognized cases of LRTI[2]. The resulting diagnostic uncertainty drives the overuse of empiric antibiotics, leading to adverse outcomes ranging from *Clostridioides difficile* infection to the development of antimicrobial-resistance[3,4].

Host transcriptomic biomarkers have emerged as a promising approach to LRTI diagnosis that can overcome several limitations of traditional microbiologic tests[5,6]. While best studied in the peripheral blood of patients with mild to moderate infection[5], recent work demonstrates that lower airway transcriptomic signatures can also enable accurate LRTI diagnosis in critically ill patients[6,7]. Pulmonary *FABP4*, for instance, was recently identified as a LRTI diagnostic biomarker in critically ill patients with acute respiratory failure, achieving an area under the receiver operating characteristic curve (AUC) of 0.90 ± 0.07 in children, and 0.85 ± 0.12 in adults[7]. *FABP4,* which encodes a lipid chaperone that modulates inflammatory signaling, is expressed in an alveolar macrophage subpopulation specifically depleted during LRTI[8-10]. Although *FABP4* exhibits respectable diagnostic performance, we recognized that improving it could boost the biomarker's practical clinical utility.

Large language model (LLM)-based artificial intelligence (AI) chatbots such as Generative Pre-trained Transformer 4 (GPT-4)[11] have shown promise in diversity of medical applications. These include image interpretation[12], patient risk stratification[13], and assisting with clinical reasoning[14,15]. Whether the performance of an infectious disease diagnostic biomarker could be augmented by an AI chatbot, however, had not been previously explored.

Here, we investigated whether GPT-4 could improve the LRTI diagnostic performance of *FABP4* in critically ill adults with acute respiratory failure. We hypothesized that incorporating AI

interpretation of electronic medical record (EMR) text data available to clinicians at the time of ICU admission, using a Health Insurance Portability and Accountability Act (HIPAA)-compliant GPT-4 platform, could boost the accuracy of *FABP4* for diagnosing LRTI.

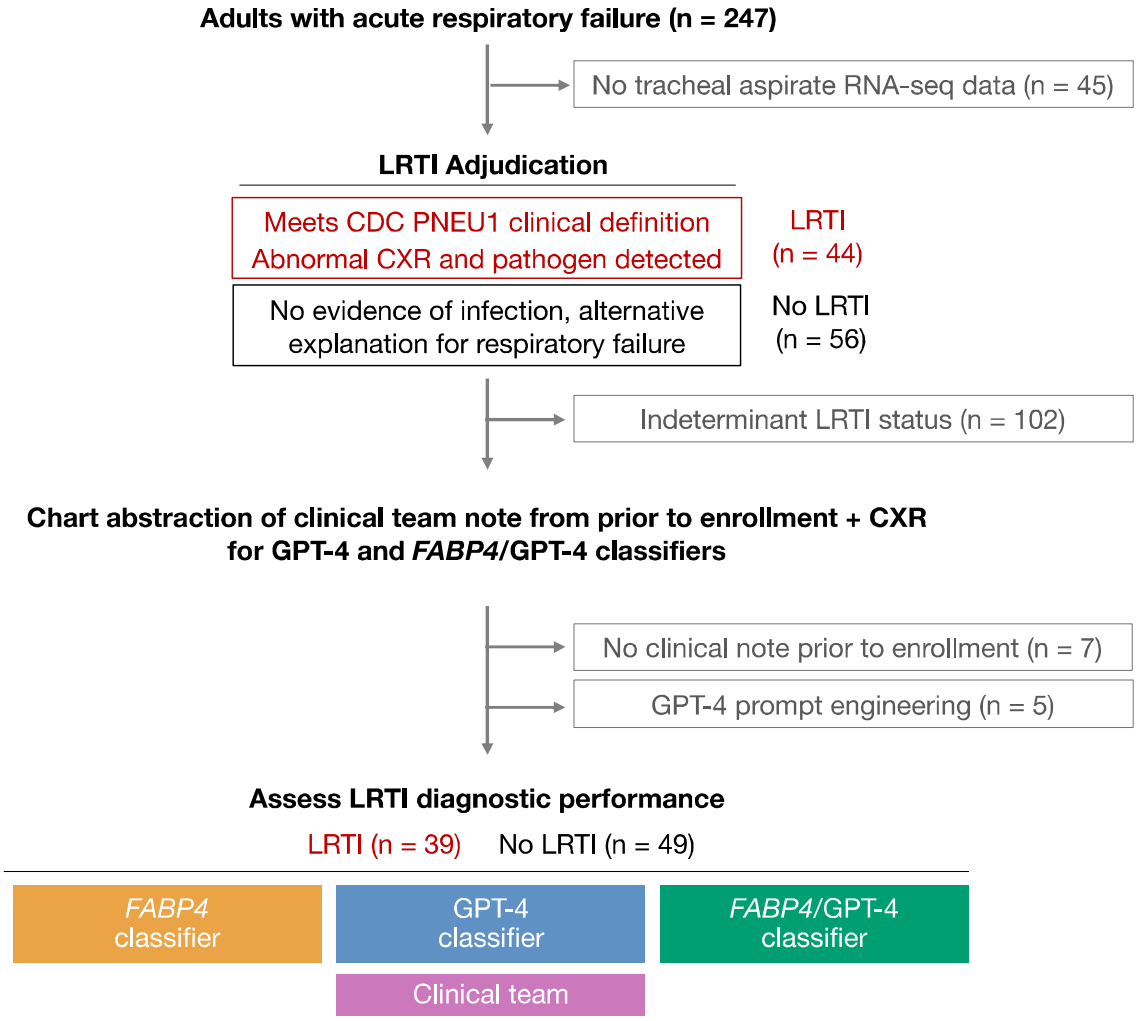


Figure 1. Flow diagram and study overview. Abbreviations: LRTI = lower respiratory tract infection; RNA-seq = RNA sequencing; CXR = chest X ray, *FABP4* = gene encoding fatty acid binding protein 4; CDC = U.S. Centers for Disease Control and Prevention; GPT-4 = Generative Pre-trained Transformer 4.

We studied a recently described cohort of 202 adults with acute respiratory failure[16] who had *FABP4* measured by endotracheal aspirate RNA sequencing within 72 hours of intubation (**Fig. 1, Supp. Table 1**). We evaluated the performance of four different diagnostic approaches (*FABP4* classifier, GPT-4 classifier, integrated *FABP4*/GPT-4 classifier, and admission diagnosis

by the ICU care team) against a gold-standard of retrospective LRTI adjudication performed by ≥2 physicians with access to all EMR data, based on the U.S. Centers for Disease Control and Prevention (CDC) PNEU1 definition[17]. This adjudication process identified 44 patients with LRTI and 56 with no evidence of infection and a clear alternative explanation for respiratory failure (No LRTI group). Patients with indeterminate LRTI status were not further evaluated (**Fig. 1**).

We provided GPT-4 with practical clinical summary information from the EMR that would be available to a treating physician on the day of ICU admission: a chest X-ray (CXR) radiology report and a note written by the medical team from the day prior (**Fig. 1**). Notes and radiology reports from five patients were utilized for GPT-4 prompt engineering and optimization (Methods). Seven patients did not have a clinical note from the day prior to study enrollment (within 72 hours of ICU admission), leaving a total of 39 LRTI and 49 No-LRTI cases available for analysis.

We first compared the accuracy of the medical team's ICU admission diagnosis against the gold-standard retrospective LRTI adjudication described above. We considered antibiotic administration within one day of study enrollment as a proxy for LRTI diagnosis by the medical team, excluding antibiotics given for established non-pulmonary infections or for prophylaxis. The medical team correctly identified 38/39 LRTI cases but unnecessarily administered antibiotics for suspected LRTI in 26/49 No LRTI cases, equating to a sensitivity of 97%, specificity of 53%, and accuracy of 73% (**Fig. 2a, Supp. Table 1**).

We next assessed the diagnostic performance of *FAPB4* expression alone, and found that it achieved an AUC of 0.83 ± 0.04 by five-fold cross validation (**Fig. 2b**). Considering an out-of-fold probability of 50% as LRTI-positive, *FABP4* had a sensitivity of 79%, specificity of 78%, and accuracy of 78% (**Fig. 2b**). We then assessed the performance of GPT-4 alone to diagnose LRTI, repeating in triplicate. Considering ≥1 chatbot diagnosis as positive for LRTI, we found that GPT-4 yielded a sensitivity of 82%, specificity of 73%, and accuracy of 77% (**Fig. 2a**), and achieved an AUC of 0.84 (**Fig. 2b**).

We then combined *FAPB4* and GPT-4 in a logistic regression model, and found that this integrated classifier achieved an AUC of 0.92 ± 0.06 (**Fig. 2b**), outperforming both *FABP4* (P = 0.02, one-sided paired t-test) and GPT-4 alone (P = 0.02, one-sided one-sample t-test). Considering an out-of-fold probability of 50% as LRTI-positive, the integrated *FABP4*/GPT-4 classifier reached a sensitivity of 85%, specificity of 88%, and accuracy of 86% (**Fig. 2a**). Assessment of the integrated classifier's performance at the Youden's index within each CV fold demonstrated a mean sensitivity of 92% ± 7%, specificity of 90% ± 17% and accuracy of 91% +/- 8%. We noted that the *FABP4*, GPT-4, and integrated *FABP4*/GPT-4 classifiers all correctly identified LRTI in the single patient whose LRTI diagnosis was missed by the clinical team, and who ultimately did not survive to hospital discharge (**Fig. 2a**).

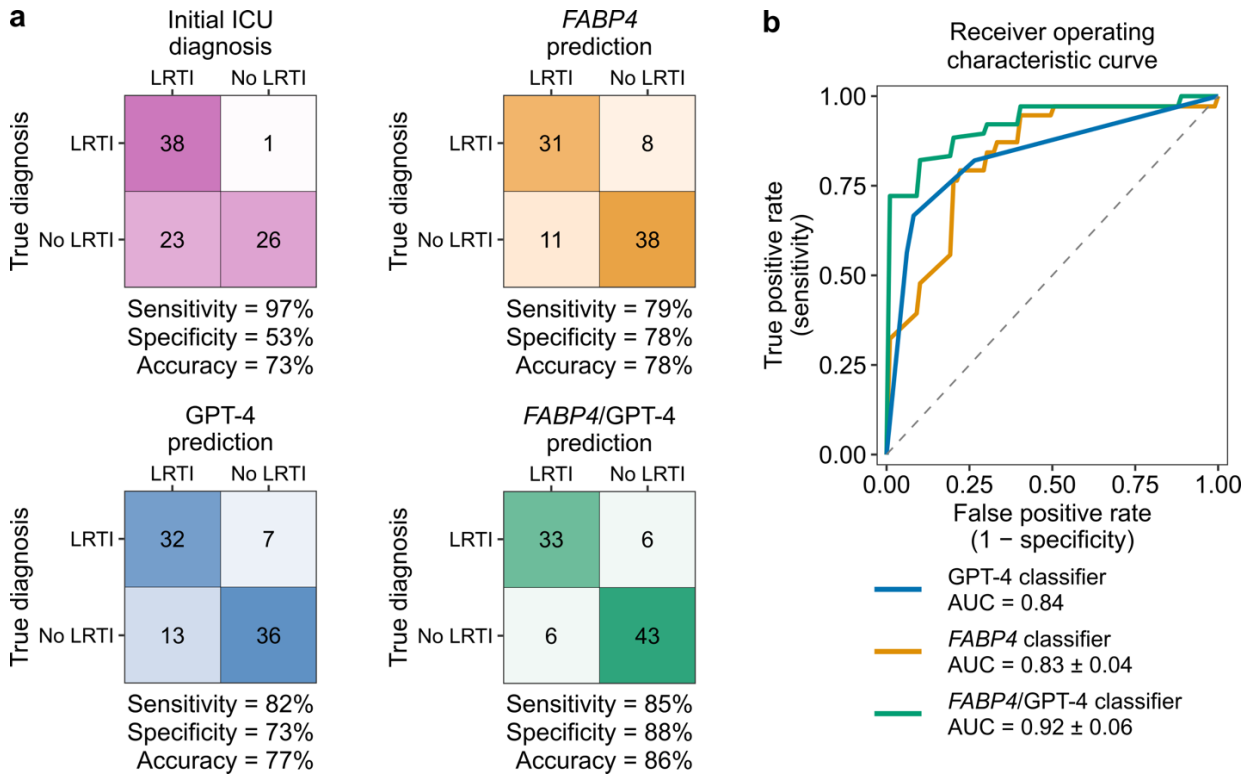


Figure 2. Comparison of LRTI diagnostics from initial ICU diagnosis, *FABP4*-based host classifier, GPT4, and *FABP4*/GPT-4 classifier. a, Heatmaps showing the confusion matrices for initial ICU diagnosis, *FABP4* classifier, GPT-4 classifier, and *FABP4*/GPT-4 integrated classifier. For the *FABP4* and *FAPB4*/GPT-4 classifiers, patients with predicted LRTI probability of 50% or higher are classified as LRTI. **b**, Receiver operating characteristic curves from GPT-4 classifier, *FABP4* classifier, and *FABP4*/GPT-4 integrated classifier. The area under the curves (AUCs) are presented as mean ± standard deviation.

To gain insight into how GPT-4 makes diagnostic estimations based on limited information, we compared the chatbot against the decision making of three comparison physicians provided identical input. From the same limited EMR data and prompt provided to GPT-4, we asked the comparison physicians to assign a diagnosis of LRTI or no evidence of LRTI for each patient. Considering a threshold of at least one LRTI diagnosis per patient across the three physicians as LRTI-positive, we found a sensitivity of 79%, specificity of 88%, and accuracy of 84% **(Supp. Fig. 1a).** Finally, we sought to identify potential biases in GPT-4 logic by comparing chatbot results to those of the comparison physicians (**Supp. Fig. 1b)**, focusing on cases with two or more discordant LRTI diagnoses. Of the nine patients more frequently diagnosed with LRTI by GPT-4 versus the comparison physicians (**Supp. Fig 1b**), six had clinical notes with no mention of LRTI, but explicit concern for LRTI in the CXR report. This suggested that GPT-4 may have placed more weight on CXR reads relative to physicians. Of the two patients disproportionately diagnosed with LRTI by comparison physicians versus GPT-4 (**Supp. Fig. 1b**), one had a final diagnosis of e-cigarette/vaping associated lung injury (GPT-4 correct) and the other had LRTI attributed to rhinovirus.

Taken together, our findings demonstrate that the combination of a host transcriptomic biomarker with AI assessment of EMR text data can improve LRTI diagnosis in critically ill patients. We found that an integrated *FABP4/GPT-4* classifier achieved higher LRTI diagnostic accuracy than *FABP4* alone, GPT-4 alone, or the treating medical team. Previous studies have found that GPT-4 is influenced by the precise language used in a prompt, leading to a need for prompt engineering[15]. By iterating our prompt on a subset of patients, and through direct comparison to physicians provided identical EMR data, we identified possible blind spots of GPT-4 and gained insights that may help guide future optimization of LLMs for infectious disease diagnosis.

In this critically ill cohort, we found that the initial treating physicians unnecessarily prescribed antibiotics in almost half of patients ultimately found to have non-infectious causes of

acute respiratory failure. Had our integrated classifier results been theoretically available at time of ICU admission, we estimate that inappropriate antibiotic use could have been prevented in 21/23 (91.3%) of No LRTI patients who were unnecessarily treated. Acute respiratory illness is a leading reason for inappropriate antibiotic use[18], and our results suggest a potential role for biomarker/AI classifiers in antimicrobial stewardship, a major goal of the U.S. CDC[19] and the World Health Organization[20].

A primary strength of this study is the novel combination of a host transcriptional biomarker with AI interpretation of EMR text data to advance infectious disease diagnosis. We address one of the most common and challenging diagnostic dilemmas in the ICU, leverage a deeply characterized cohort, and employ a rigorous post-hoc LRTI adjudication approach incorporating multiple physicians. Importantly, clinicians with access to a HIPAA-compliant GPT-4 interface can readily use our prompt without any prior bioinformatics expertise. Weaknesses of this study include a relatively small sample size, restriction to mechanically ventilated patients, and the need for an independent validation cohort. Future work can test whether GPT-4 can improve the marginal performance of widely available clinical biomarkers such as C-reactive protein, assess *FABP4*/GPT-4 classifier performance in a larger independent cohort, and evaluate these methods for the diagnosis of other infectious disease syndromes.

Methods

Patient cohort and adjudication of LRTI

We evaluated patients from a prospective observational cohort study of critically ill adults with acute respiratory failure. Patients were enrolled within 72 hours of intubation under University of California Institutional Review Board protocol #10-02701[16]. Adjudication of LRTI status was performed retrospectively following ICU discharge by ≥2 physicians using all available information in the EMR, and based on the CDC PNEU1[17] criteria including positive microbiologic testing. Patients with a clear alternative reason for their acute respiratory failure besides pulmonary infection were also identified (No LRTI group). Any adjudication discrepancies were resolved by a third physician, and patients with indeterminate LRTI status were excluded.

Extrapolation of clinical team initial LRTI diagnosis

Diagnosis of LRTI by the clinical treatment team at the time of study enrollment (within 72 hours of intubation) was extrapolated based on receipt of antibiotics for empiric treatment of respiratory infection within 1 day of enrollment. Antibiotics administered exclusively for purposes other than empiric LRTI treatment were excluded, including post-operative or peri-operative prophylaxis (n = 8), post-transplant opportunistic infection prophylaxis (n = 4), continued home suppressive therapy (n = 1), antibiotics for a local oropharyngeal infection (n = 2), and treatment of a culture-confirmed non-pulmonary infection (n = 7)**.**

Chart data extraction

For GPT-4 and GPT-4 comparison physician analysis, we extracted from the EMR one clinical note from one day prior to study enrollment and one CXR radiology read from the day of enrollment. Seven patients enrolled on the same day of hospital admission were excluded because no note from the day prior was available (**Fig. 1**). In eight cases, no primary medical

team note was available; therefore a note from consulting intensivists was substituted. In three patients, no note was written on the day prior to enrollment, so a note from two days prior was used instead. In 24 cases, no CXR was performed on the day of enrollment and so the next closest CXR read prior to the date of enrollment was used instead (**Supp. Table 1**).

FABP4 classifier

Host gene counts were obtained from tracheal aspirate RNA-sequencing data using Kallisto as previously described[7]. The gene counts were then analyzed in R v4.3.2. *FABP4* expression was normalized using the varianceStabilizingTransformation function from DESeq2 package (v1.42.1)[21], and used to train a logistic regression classifier. The performance of the *FABP4* classifier was tested using 5-fold cross-validation. For each test fold, the remaining four training folds were filtered to retain only genes with at least 10 counts across 20% of the samples. Each test fold's data was then filtered to retain only these genes from the filtered training folds. The test fold's *FABP4* expression level was normalized using varianceStabilizingTransformation and the dispersions of the training folds, and input to the trained logistic regression classifier to assign LRTI or No LRTI status for each patient in the test fold. The performance and receiver-operating characteristic ROC curve for each of the five folds was evaluated using the package pROC v1.18.5[22]. The mean AUC and standard deviation were calculated from the average AUC derived from each test fold. The sensitivity and specificity at the Youden's index were extracted for each test fold separately using the function coords from the pROC package, and the average and standard deviation were calculated across the 5 cross-validation folds.

GPT-4 input, scoring, and prompt engineering

We used the GPT-4 turbo model with 128k context length and a temperature setting of 0.2, implemented in Versa, a University of California San Francisco (UCSF) chatbot interface

developed through a partnership with partnership with Azure OpenAI, which is Health Insurance Portability and Accountability Act-compliant. For each patient, compiled clinical notes and CXR reads were input into the GPT-4 chatbot interface. Using the same prompt (**Supp. Data 1**), GPT-4 was asked to diagnose LRTI three times for each patient, each time restarting the LLM (i.e. no iterative learning was performed). A per-patient GPT-4 score was calculated based on the total number LRTI-positive diagnoses made by GPT-4.

Prior to testing GPT-4 performance, we carried out prompt engineering. This involved iteratively testing various versions of diagnostic prompts on clinical notes and CXR reads from five randomly selected patients, who were excluded from further analyses. We employed a chain-of-thought prompt strategy[23] that simply involved asking GPT-4 to analyze the note and CXR step-by-step. The prompt engineering and optimization exercise allowed us to realize the need to ask GPT-4 to ignore antibiotic treatment plans in the clinical note to avoid making LRTI diagnoses simply based on documented antibiotic administration. We initially trialed asking GPT-4 to answer in terms of probability that a patient had LRTI. However, we found that GPT-4 frequently answered either 40% or 60% LRTI probability. We then tried asking GPT-4 to choose one of three adjudications: LRTI, no LRTI, or indeterminate LRTI status, and found that GPT-4 favored “indeterminate” in most patient cases. As a result, in our final version of the prompt (**Supp. Data 1**), we asked GPT-4 to choose either LRTI or no LRTI. A deidentified example of GPT-4 LRTI diagnosis output is provided in **Supp. Data 2**.

FABP4/GPT-4 integrated classifier

The integrated classifier’s performance was tested using 5-fold cross-validation. *FABP4* expression level was normalized as above. For each test fold, a logistic regression classifier was trained on the remaining four folds using both the normalized *FABP4* level and the GPT-4 score. The performance and ROC curve for each fold was evaluated using the package pROC[22] v1.18.5. The mean ROC curve (**Fig. 2b**) was calculated from the average of ROC curves, one

from each test fold. The sensitivity and specificity at the Youden's index were extracted for each test fold separately using the function coords from the pROC package, and the average and standard deviation were calculated across the 5 cross-validation folds.

GPT-4 comparison physician control group

We compared LRTI diagnosis by GPT-4 against LRTI diagnosis made by three physicians trained in internal medicine (ADK) or additionally subspecializing in infectious diseases (AC, NLR). The physicians were provided identical information and prompt as GPT-4, and they were asked to assign each patient as either LRTI or No LRTI. The comparison physician group score (0 to 3) was calculated based on the total number of LRTI-positive diagnoses made by the comparison physicians.

Data availability

The gene count data are available at https://github.com/infectiousdisease-langelier-lab/LRTI_FABP4_classifier. Source data are provided in the source data file.

Code availability

The code is available at https://github.com/infectiousdisease-langelier-lab/LRTI_FABP4_GPT4_classifier.

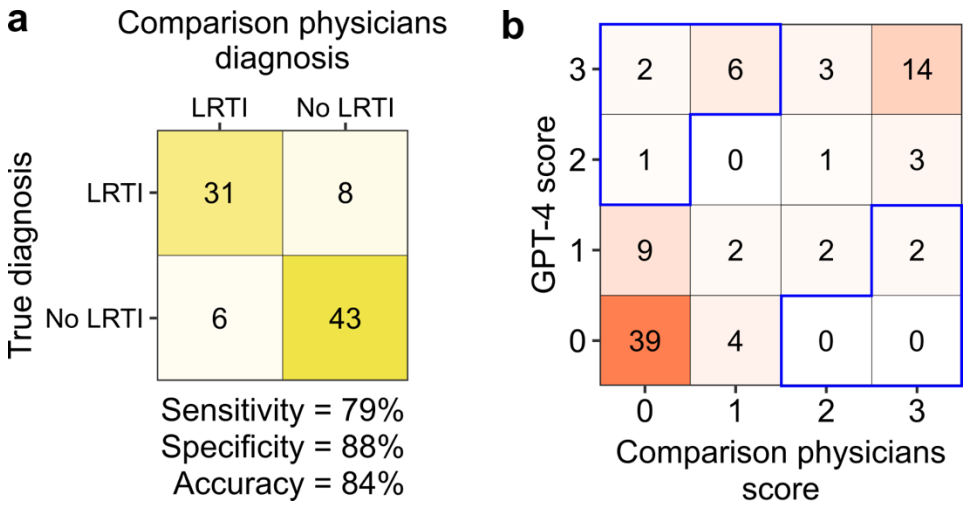
References

1. World Health Organization. Vol. 2024 (2020).
2. Jain, S.*, et al. N Engl J Med* **373**, 415-427 (2015).
3. Langford, B.J.*, et al. Clin Microbiol Infect* **29**, 302-309 (2023).
4. Centers for Disease Control and Prevention. (ed. Centers for Disease Control and Prevention) (https://stacks.cdc.gov/view/cdc/119025, 2022).
5. Tsalik, E.L.*, et al. Sci Transl Med* **8**, 322ra311 (2016).
6. Mick, E.*, et al. J Clin Invest* **133**(2023).
7. Lydon E.C., *et al*. *medRxiv* (2024).
8. Furuhashi, M., *et al*. *Clin Med Insights Cardiol* **8**, 23-33 (2014).
9. Liao, M.*, et al. Nat Med* **26**, 842-844 (2020).
10. Chen, S.T.*, et al. Sci Transl Med* **14**, eabn5168 (2022).
11. OpenAI. Vol. 2024 (2022).
12. Zhou, Y.*, et al. Radiology* **311**, e233270 (2024).
13. Beaulieu-Jones, B.K.*, et al. NPJ Digit Med* **4**, 62 (2021).
14. Maillard, A.*, et al. Clin Infect Dis* **78**, 825-832 (2024).
15. Lee, P., Bubeck, S. & Petro, J. *N Engl J Med* **388**, 2400 (2023).
16. Langelier, C.*, et al. Proc Natl Acad Sci U S A* **115**, E12353-E12362 (2018).
17. Centers for Disease Control and Prevention. (2021).
18. Merenstein, D.J., Barrett, B. & Ebell, M.H. *J Gen Intern Med* (2024).
19. Centers for Disease Control and Prevention. Vol. 2024 (U.S. Department of Health and Human Services, Atlanta: GA, 2019).
20. World Health Organization. Vol. 2024 (2021).
21. Love, M.I., Huber, W. & Anders, S. *Genome Biol* **15**, 550 (2014).
22. Robin, X.*, et al. BMC Bioinformatics* **12**, 77 (2011).
23. Wei, J., *et al. Advances in Neural Information Processing Systems 35 (NeurIPS 2022)* (NeurIPS Proceedings 2022).

Supplementary Materials

	LRTI	No LRTI	*P* value
N	39	49	
Age, years (Median, IQR)	65.0 (51.5 – 74.5)	62.0 (53.0-73.0)	0.12
Female Sex	12 (30.8%)	30 (61.2%)	0.0086
Race			0.76
White	20 (51.3%)	24 (49.0%)	-
Black/African American	4 (10.3%)	5 (10.2%)	-
Asian	8 (20.5%)	9 (18.4%)	-
Native Hawaiian/Pacific Islander	1 (2.6%)	0 (0.0%)	-
Other/Unknown	6 (15.4%)	11 (22.4%)	-
Hispanic ethnicity	5 (12.8%)	11 (22.4%)	0.38
Comorbidities	38 (97.4%)	45 (91.8%)	0.51
Immunosuppressed	9 (23.1%)	6 (12.2%)	0.29
Non-LRTI causes of intubation			-
Surgery		14 (28.6%)	-
Neurologic*		12 (24.5%)	-
Cardiovascular		8 (16.3%)	-
Non-LRTI infection*		5 (10.2%)	-
Other		11 (22.4%)	-
Initial ICU diagnosis			<0.0001
LRTI	38 (97.4%)	26 (51.1%)	-
No LRTI	1 (2.6%)	23 (46.9%)	-
Clinical service writing note			0.18
Medicine	16 (41.0%)	9 (18.4%)	-
Critical Care	5 (12.8%)	7 (14.3%)	-
Neurosurgery	3 (7.7%)	9 (18.4%)	-
Cardiology	3 (7.7%)	6 (12.2%)	-
Liver Transplant	1 (2.6%)	4 (8.2%)	-
Neurology	1 (2.6%)	4 (8.2%)	-
Other	10 (25.6%)	10 (20.4%)	-
Time from note to enrollment			0.05
1 day	36 (92.3%)	49 (100.0%)	-
2 days	3 (7.7%)	0 (0.0%)	-
Time from CXR to enrollment			0.21
0 days	31 (79.5%)	33 (67.3%)	-
1 day	8 (20.5%)	13 (26.5%)	-
2 days	0 (0.0%)	3 (6.1%)	-

Supplementary Table 1. Demographic features of cohort, initial ICU diagnoses, and data collection by LRTI status. *P* values are comparing LRTI to No LRTI patients with chi square tests for categorical variables and Wilcoxon rank-sum test for continuous variables (age). IQR = interquartile range. *One patient was adjudicated as having both neurologic and non-LRTI infection as indications for intubation and is included in both fields.



Supplementary Figure 1. Comparison of GPT-4 performance to physicians provided the same EMR data. a, Heatmap confusion matrix of diagnosis by three GPT-4 comparison physicians who received the same prompt and data as GPT-4. **b**, Comparison of GPT-4 LRTI scores as compared to physicians. In **b,** X-axis depicts the number of times GPT-4 diagnosed LRTI out of 3, Y-axis shows the number of times the physicians called LRTI out of 3. Blue boxes indicate instances in which GPT-4 diagnoses were most discordant with comparison physicians (difference in scores of ≥2).