Comparative Evaluation of ChatGPT-4o and DeepSeek-R1 for Emergency Severity Index (ESI) Triage Classification

Ahmad M. Nazzal, MD, PhD

Abstract

Large language models (LLMs) like ChatGPT-4o and DeepSeek-R1 show promise in automating emergency triage, but their alignment with clinical standards remains understudied. This study evaluates both models against a human physician gold standard using the Emergency Severity Index (ESI). ChatGPT-4o demonstrated substantial agreement (Cohen's Kappa = 0.717, 95% CI: 0.56-0.85; 80% absolute agreement), outperforming DeepSeek-R1 (Cohen's Kappa = 0.583, 95% CI: 0.41-0.75; 70% absolute agreement). While both models excelled in high-acuity cases (ESI 1-2), their performance declined for mid-level categories (ESI 3-5), underscoring the risks of automation bias in ambiguous scenarios.

Introduction

Emergency department (ED) triage is a critical process that prioritizes patient care based on the severity of their condition. The Emergency Severity Index (ESI) is a widely used five-level triage system in the United States, designed to categorize patients by acuity and anticipated resource needs [1]. Accurate triage is essential for optimizing resource allocation, reducing wait times, and improving patient outcomes. However, ED overcrowding and staffing shortages have increased pressure on triage nurses, leading to variability in decision-making and potential errors [2].

Recent advances in large language models (LLMs), such as OpenAI's ChatGPT-4o and DeepSeek-R1, offer promising opportunities to assist or automate triage. These models excel in natural language understanding and can process complex clinical narratives to generate actionable insights [3]. Studies have demonstrated their potential in medical tasks, including diagnosis support, patient communication, and clinical documentation [4]. However, their application to ESI triage remains underexplored, particularly regarding alignment with clinical standards and performance across diverse patient scenarios.

While LLMs show promise, their adoption in high-stakes clinical settings requires rigorous evaluation. Prior research has highlighted challenges such as over-reliance on pattern recognition, susceptibility to ambiguous inputs, and limited interpretability [5]. Moreover, LLM performance in triage—where decisions must balance urgency, resource needs, and clinical context—has not been systematically compared. This gap is especially critical for mid-level ESI categories (e.g., ESI 3–4), where subtle clinical cues can significantly impact patient outcomes.

This study evaluates the performance of ChatGPT-4o and DeepSeek-R1 in assigning ESI triage levels to synthetic patient cases. Using a physician-validated dataset, we assess inter-rater agreement, absolute accuracy, and class-wise performance across ESI levels**.** Our findings offer insights into the strengths and limitations of LLMs in emergency triage, informing their potential role as decision-support tools in clinical practice.

Methods

Dataset

A synthetic dataset of 50 patient vignettes was created to represent all five Emergency Severity Index (ESI) levels. Each vignette contained relevant demographic details (e.g., age, sex, medical history) and clinical symptoms (e.g., “crushing chest pain radiating to the left arm, diaphoresis”). A single physician assigned an ESI level to each vignette in accordance with ESI guidelines [6]. The dataset included (i) high-acuity scenarios (ESI 1–2, e.g., cardiac arrest, stroke), (ii) mid-level scenarios (ESI 3–4, e.g., stable abdominal pain), and (iii) non-urgent scenarios (ESI 5, e.g., rash).

Large Language Models (LLMs)

Two LLMs were evaluated: (1) ChatGPT-4o [7] DeepSeek-R1 [8]. Both models were provided identical patient vignettes and instructed to assign the most appropriate ESI level based on their underlying reasoning. Rationale given by each model was recorded.

Evaluation Protocol

All vignettes were introduced to each model using a standardized prompt. The assigned ESI level and rationale were recorded and then compared with the physician-designated gold standard. Dicrepancies were documented to identify over-triaging or under-triaging patterns.

Prompt

```
Assign an Emergency Severity Index (ESI) level (1–5) for the following patient based on U.S. emergency triage guidelines:

Patient Profile: [Insert data]
Symptoms: [Insert data]

Guidelines:
- Level 1: Requires immediate life-saving intervention (e.g., cardiac arrest, severe trauma).
- Level 2: High-risk, unstable, or confused/lethargic/disoriented, or severe pain/distress-
- Level 3: Stable but requiring ≥2 resources (e.g., ECG + labs, CT scan + IV meds).
- Level 4: Stable but requiring 1 resource (e.g., X-ray, sutures).
- Level 5: Stable and requiring no resources (e.g., prescription refill).

High-risk vital signs are as follows:
Age Group High-Risk HR (bpm) High-Risk RR (bpm) SpO2 (%)
< 1 month > 190 > 60 < 92%
1–12 months > 180 > 55 < 92%
1–3 years > 140 > 40 < 92%
3–5 years > 120 > 35 < 92%
5–12 years > 120 > 30 < 92%
12–18 years > 100 > 20 < 92%
> 18 years > 100 > 20 < 92%

Provide the ESI level (1–5) and a brief rationale. Do not use markdown.
```

Figure 1. Emergency Severity Index (ESI) triage prompt for patient assessment based on U.S. emergency triage guidelines. The prompt outlines five ESI levels, ranging from immediate life-saving intervention (Level 1) to stable patients requiring no resources (Level 5). High-risk vital sign thresholds for different age groups are also provided to assist in classification.

Metrics

Model performance was quantified using: Cohen’s Kappa (κ): Inter-rater reliability accounting for chance, calculated via the cohen_kappa_score function in scikit-learn. Absolute Agreement: Percentage of exact matches between the model and the gold standard. Class-wise Accuracy: Accuracy stratified by ESI level. Confusion Matrices: Visual distributions of correct and incorrect classifications, highlighting the extent and direction of misclassifications.

Statistical Analysis

Python (v3.9) with scikit-learn (v1.3.0) was used to conduct all analyses. Cohen's Kappa confidence intervals were generated by bootstrapping the data with 1,000 iterations. Descriptive statistics were computed to characterize the degree of agreement and classification trends across ESI levels.

Results

ChatGPT-4o vs. Gold Standard

ChatGPT-4o demonstrated strong alignment with the physician gold standard, achieving: Cohen's Kappa (κ): 0.717 (95% CI: 0.56–0.85), indicating substantial agreement [9] and an absolute agreement of 80.0% (40/50 cases). Class-wise accuracy was ESI 1: 100.0% (2/2 cases), ESI 2: 100.0% (20/20 cases), ESI 3: 36.4% (4/11 cases), ESI 4: 66.7% (4/6 cases), and ESI 5: 90.9% (10/11 cases). ChatGPT-4o excelled in high-acuity cases (ESI 1–2), achieving perfect accuracy. Performance dropped for mid-level categories (ESI 3–4), with frequent over-triaging (e.g., labeling ESI 3 cases as ESI 2). Non-urgent cases (ESI 5) were accurately classified in 90.9% of instances.

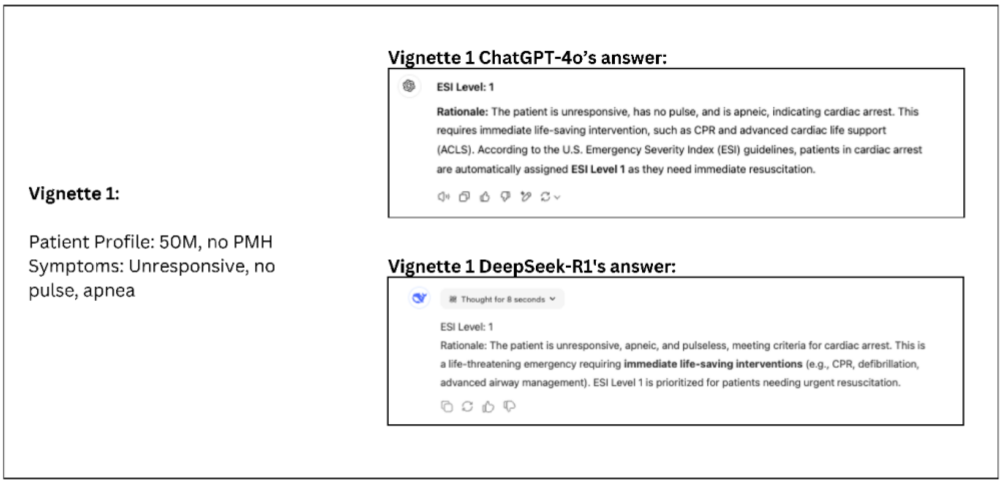

Figure 2. Comparison of responses from ChatGPT-4o and DeepSeek-R1 for an Emergency Severity Index (ESI) triage vignette. The case describes a 50-year-old male with no past medical history, presenting as unresponsive, pulseless, and apneic—consistent with cardiac arrest. Both AI models correctly classify the case as ESI Level 1, requiring immediate life-saving interventions such as cardiopulmonary resuscitation (CPR) and advanced cardiac life support (ACLS).

DeepSeek-R1 vs. Gold Standard

DeepSeek-R1 showed moderate agreement with the gold standard, with Cohen's Kappa (κ): 0.583 (95% CI: 0.41–0.75), indicating moderate agreement [1] and absolute agreement was 70.0% (35/50 cases). Class-wise accuracy was ESI 1: 100.0% (2/2 cases), ESI 2: 95.0% (19/20 cases), ESI 3: 45.5% (5/11 cases), ESI 4: 83.3% (5/6 cases), ESI 5: 36.4% (4/11 cases). DeepSeek-R1 performed well for high-acuity cases (ESI 1–2) but struggled with non-urgent cases (ESI 5), misclassifying 63.6% as ESI 4. Mid-level categories (ESI 3–4) showed mixed performance, with frequent under-triaging (e.g., labeling ESI 3 cases as ESI 2).

DeepSeek-R1 performed well for high-acuity cases (ESI 1–2) but struggled with non-urgent cases (ESI 5), misclassifying 63.6% as ESI 4. Mid-level categories (ESI 3–4) showed mixed performance, with frequent under-triaging (e.g., labeling ESI 3 cases as ESI 2).

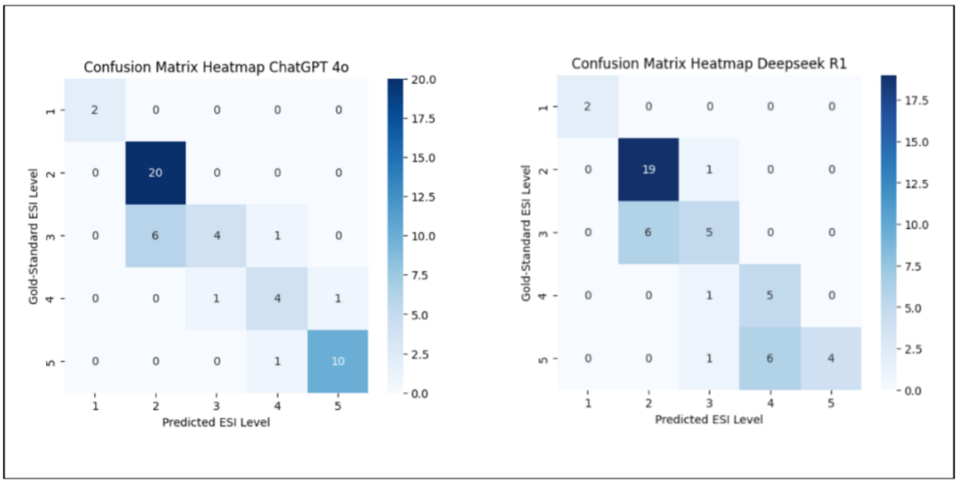

Figure 3. Confusion matrix heatmaps comparing ChatGPT-4o and DeepSeek-R1 in Emergency Severity Index (ESI) level classification. The gold-standard ESI levels (y-axis) are compared to the AI models' predicted ESI levels (x-axis). Darker shades indicate a higher frequency of predictions for a given category. ChatGPT-4o demonstrates high accuracy in ESI Level 2 classification but shows some misclassification in Levels 3 and 4. DeepSeek-R1 exhibits similar performance but with slight variation in misclassification patterns, particularly in Levels 2, 4 and 5.

Comparative Analysis
ChatGPT-4o (κ = 0.717) outperformed DeepSeek-R1 (κ = 0.583), aligning more closely with the gold standard. Both models excelled in high-acuity cases (ESI 1–2) but struggled with mid-level and non-urgent categories. ChatGPT-4o's errors were primarily over-triaging (e.g., labeling ESI 3 as ESI 2), while DeepSeek-R1 exhibited both over- and under-triaging.

Discussion

This study aimed to assess the agreement between these AI models and physician-assigned ESI levels. Our findings indicate that ChatGPT-4o demonstrated substantial agreement with the gold standard (κ = 0.717, 95% CI: 0.56–0.85), outperforming DeepSeek-R1 (κ = 0.583, 95% CI: 0.41–0.75), which showed moderate agreement. Both ChatGPT-4o and DeepSeek-R1 demonstrated strong performance in high-acuity cases (ESI 1–2), achieving near-perfect accuracy. This aligns with prior research suggesting that large language models (LLMs) excel in recognizing critical conditions [4].

However, both models exhibited vulnerabilities in mid-level and non-urgent cases. ChatGPT-4o frequently over-triaged ESI 3 cases (e.g., stable abdominal pain requiring one resource) to ESI 2, potentially leading to unnecessary resource allocation. Conversely, DeepSeek-R1 struggled significantly with non-urgent cases (ESI 5), misclassifying 63.6% as ESI 4. This misclassification pattern could contribute to overcrowding in emergency departments (EDs) by assigning higher urgency than warranted.

One possible explanation for ChatGPT-4o's better performance compared to DeepSeek-R1 is the inherent bias in model fine-tuning based on different healthcare systems. ChatGPT-4o appears to be more aligned with U.S. medical practices and triage protocols, likely due to greater exposure to American healthcare data, guidelines, and clinical decision-making frameworks, including the ESI system. In contrast, DeepSeek-R1 may have been fine-tuned on a different healthcare system, potentially incorporating regional triage practices, alternative classification standards, or differences in clinical workflow. This divergence in fine-tuning introduces a systemic bias in how each model interprets and classifies ESI levels, which could explain why ChatGPT-4o demonstrated higher agreement with the physician gold standard, particularly in high-acuity cases.

Clinical Implications
ChatGPT-4o achieved a substantial agreement with human rater. Its high agreement with the gold standard suggests that it could assist clinicians in streamlining patient prioritization, especially for high-risk cases. In contrast, DeepSeek-R1's moderate agreement frequent misclassification of non-urgent cases suggest limitations. For reference the inter-rater reliability between nurses' and physicians' ESI rating is near perfect ($\kappa = 0.94$) [10].

Limitations
Several factors may impact the generalizability of these findings: The analysis was conducted on synthetic data, which may not fully capture real-world complexities such as ambiguous presentations or incomplete patient histories. The gold standard was determined by a single expert, whereas inter-rater variability in ESI assignments is well-documented [11,12]. Future studies should incorporate multiple expert annotations to better reflect real-world triage conditions.

Future Directions
Future research needs to focus on enhancing model reliability and applicability by validating performance on real-world emergency department data to assess robustness across diverse clinical presentations. Moreover, future studies should focus on refining model training to reduce systematic biases, particularly in mid-level and non-urgent case classification, to optimize resource allocation in EDs.

References

1. Gilboy, N., Tanabe, P., Travers, D., & Rosenau, A. M. (2011). Emergency Severity Index (ESI): A triage tool for emergency care. AHRQ Publication.
2. Hoot, N. R., & Aronsky, D. (2008). Systematic review of emergency department crowding: Causes, effects, and solutions. Annals of Emergency
3. Riedemann, L., Labonne, M. & Gilbert, S. The path forward for large language models in medicine is open. npj Digit. Med. 7, 339 (2024). https://doi.org/10.1038/s41746-024-01344-w
4. Levine DM, Tuwani R, Kompa B, et al. The diagnostic and triage accuracy of the GPT-3 artificial intelligence model: an observational study. *The Lancet Digital Health*. 2024;6(8):e555-e561. doi:10.1016/S2589-7500(24)00097-9
5. Busch, F., Hoffmann, L., Rueger, C. et al. Current applications and challenges in large language models for patient care: a systematic review. Commun Med 5, 26 (2025). https://doi.org/10.1038/s43856-024-00717-2
6. Gilboy N, Tanabe T, Travers D, Rosenau AM (2020). Emergency Severity Index (ESI): A Triage Tool for Emergency Department Care, Version 4. 2020 Edition (PDF). Emergency Nurses Association.

7. OpenAI. GPT-4 Technical Report. Published online March 27, 2023. https://arxiv.org/abs/2303.08774
8. DeepSeek-AI. DeepSeek-R1 Preprint. Published online January 22, 2025. https://doi.org/10.48550/arXiv.2501.12948
9. Cohen J. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement.* 1960;20(1):37-46. doi:10.1177/001316446002000104
10. Esmailian M, Zamani M, Azadi F, Ghasemi F. Inter-Rater Agreement of Emergency Nurses and Physicians in Emergency Severity Index (ESI) Triage. *Emerg (Tehran)*. 2014;2(4):158-161.
11. Buschhorn HM, Strout TD, Sholl JM, Baumann MR. Emergency medical services triage using the Emergency Severity Index: is it reliable and valid? *Journal of Emergency Nursing.* 2013;39(5):e55-e63. doi:10.1016/j.jen.2011.11.003
12. Pourasghar F, Daemi A, Tabrizi JS, Ala A. Inter-rater reliability of triages performed by the electronic triage system. *Bulletin of Emergency Trauma.* 2015;3(4):134-137.
13. Landis, J. R., & Koch, G. G. (1977). The Measurement of Observer Agreement for Categorical Data. *Biometrics*, *33*(1), 159–174. https://doi.org/10.2307/2529310
14. Fleiss, J. L. (1981). Balanced incomplete block designs for inter-rater reliability studies. *Applied Psychological Measurement, 5*(1), 105–112. https://doi.org/10.1177/014662168100500115

Table 1. Performance comparison of ChatGPT-4o and DeepSeek-R1 in Emergency Severity Index (ESI) level classification. Metrics include Cohen's κ coefficient with 95% confidence intervals, absolute agreement percentage, and class-wise accuracy for each ESI level.

Metric	ChatGPT-4o	DeepSeek-R1
Cohen's κ (95% CI)	0.717 (0.56–0.85)	0.583 (0.41–0.75)
Absolute Agreement	80.0%	70.0%
Class-wise Accuracy		
ESI 1	100.0%	100.0%
ESI 2	100.0%	95.0%
ESI 3	36.4%	45.5%
ESI 4	66.7%	83.3%
ESI 5	90.9%	36.4%