ORIGINAL ARTICLE

DOI: 10.4274/hamidiyemedj.galenos.2025.60362

Hamidiye Med J 2026;7(1):50-59

Comparison of AI-Based Triage Systems in High-Energy Thoracic Trauma: A Pilot Study Using ChatGPT and Gemini

Yüksek Enerjili Toraks Travmalarında Kılavuz Temelli Klinik Karar Verme Sürecinde ChatGPT-4 ve Gemini 1.5'in Karşılaştırmalı Performansı: 30 Sentetik Olguda Simülasyon Çalışması

Nilay Çavuşoğlu Yalçın[1], Merve Sarı Akyüz[2], Okan Karataş[1], Olgun Keskin[2], Muharrem Özkaya[1]

[1]*University of Health Sciences Türkiye, Antalya Training and Research Hospital, Clinic of Thoracic Surgery, Antalya, Türkiye*

[2]*University of Health Sciences Türkiye, Antalya Training and Research Hospital, Clinic of Pulmonology, Antalya, Türkiye*

ABSTRACT

Background: High-energy thoracic trauma is a leading cause of morbidity and mortality in polytrauma patients. Timely diagnosis and management are crucial for patient outcomes. Artificial intelligence (AI) systems such as ChatGPT and Gemini have demonstrated potential in providing clinical decision support. This study presents a dataset of 30 synthetic thoracic trauma cases and explores the feasibility of using AI tools to assist in triage and early management decisions.

Materials and Methods: Thirty synthetic cases, reflecting real-world patterns of high-energy thoracic trauma, were developed based on established trauma mechanisms, clinical findings, and evidence-based physiological parameters. Each case includes vital signs, injury mechanism, physical findings, and a structured clinical question regarding immediate management. These cases were formatted for AI analysis. Responses from ChatGPT and Gemini will be analyzed for clinical appropriateness, adherence to trauma guidelines (e.g., Advanced Trauma Life Support), and concordance with experts.

Results: Initial simulations (5-case pilot) demonstrated greater adherence to trauma protocols with ChatGPT, particularly in managing tension pneumothorax, tamponade, and flail chest. Gemini responses were more conservative and occasionally delayed critical interventions. A full-scale analysis of all 30 cases is underway, using a scoring system based on accuracy, guideline conformity, and intervention prioritization. Preliminary findings suggest AI tools can assist in high-stakes clinical decision-making, particularly in environments with limited specialist access. ChatGPT's ability to provide structured, guideline-based responses is promising. However, human oversight remains essential. Large-scale validation with real patient data is necessary before clinical deployment.

Conclusion: AI-based models show potential in enhancing early decision-making in high-energy thoracic trauma. While encouraging, these tools require careful integration into clinical workflows and further validation in real-time settings.

Keywords: Thoracic trauma, artificial intelligence, triage, ChatGPT, Gemini, emergency surgery, clinical decision support

ÖZ

Amaç: Yüksek enerjili toraks travması, çoklu travmalı hastalarda önlenebilir ölümlerin önde gelen nedenlerinden biridir. Bu çalışmada, iki gelişmiş büyük dil modelinin (large language model) — ChatGPT-4 (OpenAI) ve Gemini 1.5 (Google DeepMind) — yüksek enerjili toraks travmalarında kılavuzlara uygun klinik karar önerileri üretme yeterliliği değerlendirildi.

Gereç ve Yöntemler: Çeşitli yaralanma mekanizmaları, fizyolojik parametreler ve klinik bulguları içeren 30 sentetik klinik senaryo geliştirildi. Her olgu, klinik özetler ve yönetim sorularını içeren standart istemlerle ChatGPT-4 ve Gemini 1.5 tarafından bağımsız

Address for Correspondence: Nilay Çavuşoğlu Yalçın, University of Health Sciences Türkiye, Antalya Training and Research Hospital, Department of Thoracic Surgery, Antalya, Türkiye

E-mail: drnili@hotmail.com **ORCID ID:** orcid.org/0000-0002-0675-9267

Received: 27.10.2025 **Accepted:** 12.12.2025 **Epub:** 10.02.2026 **Publication Date:** 02.03.2026

Cite this article as: Çavuşoğlu Yalçın N, Sarı Akyüz M, Karataş O, Keskin O, Özkaya M. Comparison of AI-based triage systems in high-energy thoracic trauma: A pilot study using ChatGPT and Gemini. Hamidiye Med J. 2026;7(1):50-59

ÖZ

olarak değerlendirildi. Yanıtlar, model kimliği gizlenen iki uzman travma cerrahı tarafından, İleri Travma Yaşam Desteği ve Doğu Travma Cerrahisi Derneği kılavuzlarına uyuma göre ikili uyum puanlama sistemiyle değerlendirildi. İstatistiksel analizlerde McNemar testi, Cohen'in Kappa katsayısı, ki-kare analizi ve alıcı işletim karakteristiği (AİK) eğrisi kullanıldı.

Bulgular: ChatGPT-4 %83,3 (25/30), Gemini 1.5 ise %60,0 (18/30) oranında kılavuz uyumu sağladı. McNemar testi ChatGPT-4 lehine bir eğilim gösterse de istatistiksel olarak anlamlı değildi ($p = 0,0654$). Modeller arası uyum zayıftı ($\kappa = 0,15$), genel fark anlamlı bulunmadı ($\chi^2 = 2,95$; $p = 0,0856$). Gemini'nin ChatGPT-4'e göre AİK eğrisi altında kalan alanı (AUC) 0,62 idi.

Sonuç: ChatGPT-4, özellikle anatomik ve radyolojik değerlendirme alanlarında Gemini 1.5'e göre daha yüksek kılavuz uyumu göstermiştir. Ancak, sınırlı model uyumu ve istatistiksel anlamlılığın olmaması, klinik uygulamadan önce hekim denetimi ve ileri doğrulama gerekliliğini vurgulamaktadır. Bulgular, yapay zekâ destekli travma karar destek sistemlerinin potansiyelini göstermekle birlikte, akut bakımda hekim yargısının önemini koruduğunu ortaya koymaktadır.

Anahtar Kelimeler: Torasik travma, yapay zeka, triyaj, ChatGPT, Gemini, acil cerrahi, klinik karar destek sistemi

Introduction

High-energy thoracic trauma represents a critical challenge in emergency medicine, accounting for 20–25% of trauma-related mortality and contributing to an additional 50% of deaths among patients with polytrauma (1). The Golden Hour concept emphasizes that survival outcomes fundamentally depend on rapid recognition, accurate assessment, and timely implementation of evidence-based interventions within the first 60 minutes after injury. Contemporary trauma care relies heavily on standardized protocols, particularly the Advanced Trauma Life Support (ATLS) guidelines developed by the American College of Surgeons Committee on Trauma (2), which provide systematic approaches to primary and secondary survey methodology, diagnostic prioritization, and therapeutic decision-making.

The complexity of thoracic trauma management stems from the anatomical diversity of potential injuries, ranging from simple pneumothorax to life-threatening conditions such as massive hemothorax, cardiac tamponade, and tracheobronchial disruption (3). Each injury pattern requires specific diagnostic approaches and therapeutic interventions, and management decisions often depend on integrating the clinical presentation, physiologic parameters, imaging findings, and mechanisms of injury (4). The time-sensitive nature of these decisions, combined with the high stakes of potential adverse outcomes, creates an environment where clinical decision support tools could provide substantial value.

Artificial intelligence (AI) and machine learning technologies have demonstrated increasing sophistication in medical applications, with large language models (LLMs) showing particular promise in knowledge-based medical tasks. Recent studies (5-7) have evaluated AI performance in medical licensing examinations, diagnostic reasoning, and clinical documentation, with several models achieving performance levels comparable to or exceeding those of human physicians in controlled testing environments (8). However, the translation of these capabilities to acute care scenarios, particularly in trauma medicine where rapid decision-making under uncertainty is paramount, remains largely unexplored (9).

The potential applications of AI in trauma care extend beyond simple knowledge retrieval to include pattern recognition, risk stratification, and decision support in resource-limited environments. Emergency departments and trauma centers increasingly face challenges related to physician workload, diagnostic accuracy under time pressure, and standardization of care across different experience levels. AI-assisted decision support could address these challenges by providing consistent, guideline-based recommendations (10) and maintaining the speed necessary for acute care environments, although notable limitations in accuracy have been documented.

This study aims to evaluate the performance of two leading LLMs, ChatGPT-4 and Gemini 1.5, in providing guideline-concordant recommendations for the management of high-energy thoracic trauma. Through a systematic evaluation of 30 synthetic clinical scenarios designed to reflect authentic trauma presentations, we assess the capacity of these AI systems to support clinical decision-making in acute care settings while identifying specific areas of strengths and limitations that inform future development and implementation strategies.

Materials and Methods

This study employed a retrospective simulation framework involving 30 synthetically generated high-energy thoracic trauma scenarios. Case construction was guided by existing literature on trauma epidemiology, injury biomechanics, and acute clinical management protocols.

Cases were reviewed by two board-certified trauma surgeons to ensure plausibility, clinical relevance, and alignment with real-world trauma profiles.

Two state-of-the-art LLMs were evaluated: ChatGPT-4 (OpenAI) and Gemini 1.5 (Google DeepMind). Each model was independently queried using the 30 cases in a structured prompt format. Each model was queried in a new and isolated session to avoid memory-related carryover effects and minimize prompt-context drift. Prompts included anonymized clinical summaries detailing the mechanism of injury, vital signs, physical findings, and a primary clinical question (e.g., airway management, need for imaging, and need for thoracostomy).

Model outputs were assessed for:

- Clinical appropriateness: Whether the recommendation aligned with accepted trauma protocols.
- Guideline concordance: Based on ATLS, Eastern Association for the Surgery of Trauma (EAST), and trauma surgical best practices.
- Intervention prioritization: Accuracy and urgency of life-saving measures.

Responses were scored dichotomously (1 = guideline-concordant; 0 = discordant or delayed) by an independent panel of trauma specialists blinded to model identity. Statistical analysis included descriptive statistics, McNemar's test, Cohen's Kappa coefficient, chi-square tests, and receiver operating characteristic (ROC) curve analysis to evaluate diagnostic performance and model agreement.

The study was conducted without human subject data and was therefore exempt from institutional review board (IRB) approval.

Inclusion Criteria

A total of 30 synthetic cases were included in this study (Table 1). These cases were designed to emulate real-world presentations of high-energy thoracic trauma, based on established injury patterns observed in clinical practice. Inclusion criteria for case construction were as follows:

- Mechanism of injury consistent with high-energy blunt or penetrating thoracic trauma (e.g., motor vehicle collisions, falls from height, crush injuries).
- Presence of physiologic instability (e.g., hypotension, tachypnea, hypoxia) or pathognomonic thoracic injury signs (e.g., subcutaneous emphysema, flail chest, distended neck veins).
- Availability of clearly defined clinical findings, vital signs, and a specific diagnostic or management question related to acute trauma care.
- Relevance to guideline-based decision-making processes, particularly those addressed by the ATLS and EAST recommendations.

Each synthetic case was developed to provide sufficient detail for AI models to interpret the scenario and to propose a prioritized clinical action plan.

Statistical Analysis

Responses from the two AI models were dichotomously scored (1 = guideline-concordant; 0 = discordant or delayed) by an independent panel of trauma specialists blinded to model identity. Descriptive statistics were used to summarize categorical variables as frequencies and percentages.

A paired comparison of binary outcomes was conducted using the McNemar test to detect significant differences in guideline-concordant responses between the two models. Agreement between ChatGPT and Gemini was assessed using Cohen's Kappa coefficient. A chi-square test was performed to examine associations between model performance and case-specific variables. This test was used solely to evaluate differences in paired proportions and was not intended to represent clinical agreement or imply clinical equivalence between the models.

A paired t-test was not applied because the model outputs were binary (0/1), and this test requires continuous, normally distributed data. ROC curves were generated to evaluate the diagnostic performance of each model in predicting expert-aligned responses. In this study, ROC analysis was used as an exploratory model-to-model comparison tool rather than a clinical diagnostic metric, with ChatGPT's outputs serving as the benchmark. The area under the curve (AUC) was calculated to quantify discriminatory power. All analyses were performed using SPSS version 27 (IBM Corp.) and GraphPad Prism version 9.0, with statistical significance set at a p-value < 0.05.

Results

In the context of AI-assisted decision-making in trauma care, concordant response scores refer to the degree of alignment between an AI model's clinical recommendations and established trauma management guidelines, such as ATLS or EAST.

The purpose of concordance scoring is to assess the reliability and clinical appropriateness of AI-generated responses, particularly in high-stakes scenarios such as high-energy thoracic trauma. Each response is evaluated by expert reviewers or automated protocols for its consistency with evidence-based clinical standards.

Concordant response scores provide a quantifiable framework to determine the clinical utility of AI models. In our study, models, such as ChatGPT-4 and Gemini 1.5, were evaluated on 30 synthetic thoracic trauma cases. The scoring system allowed a standardized comparison based on adherence to trauma guidelines, revealing significant differences in model performance (Table 2).

Table 1. Synthetic thoracic trauma cases dataset.

Case ID	Age	Gender	Mechanism of injury	Clinical findings	Vital signs	Primary question
1	72	Male	Assault with blunt object	Subcutaneous emphysema and decreased breath sound on the right	BP: 101/74 mmHg, HR: 122 bpm, RR: 28 bpm, SpO_2: 92%	What should be the immediate clinical approach and need for intervention?
2	54	Female	Assault with blunt object	Massive hemothorax requiring immediate decompression	BP: 80/63 mmHg, HR: 99 bpm, RR: 31 bpm, SpO_2: 86%	What should be the immediate clinical approach and need for intervention?
3	58	Male	High-speed motor vehicle collision	Multiple left rib fractures with minimal hemothorax	BP: 74/77 mmHg, HR: 132 bpm, RR: 19 bpm, SpO_2: 96%	What should be the immediate clinical approach and need for intervention?
4	40	Female	Pedestrian struck by car	Flail chest with bilateral pulmonary contusion	BP: 116/48 mmHg, HR: 97 bpm, RR: 35 bpm, SpO_2: 98%	What should be the immediate clinical approach and need for intervention?
5	63	Male	Assault with blunt object	Multiple left rib fractures with minimal hemothorax	BP: 95/70 mmHg, HR: 140 bpm, RR: 25 bpm, SpO_2: 90%	What should be the immediate clinical approach and need for intervention?
6	64	Female	Pedestrian struck by car	Flail chest with bilateral pulmonary contusion	BP: 91/68 mmHg, HR: 99 bpm, RR: 27 bpm, SpO_2: 97%	What should be the immediate clinical approach and need for intervention?
7	57	Male	High-speed motor vehicle collision	Subcutaneous emphysema and decreased breath sound on the right	BP: 120/66 mmHg, HR: 95 bpm, RR: 20 bpm, SpO_2: 94%	What should be the immediate clinical approach and need for intervention?
8	61	Female	Pedestrian struck by car	Cardiac tamponade signs with hypotension and tachycardia	BP: 80/73 mmHg, HR: 103 bpm, RR: 23 bpm, SpO_2: 95%	What should be the immediate clinical approach and need for intervention?
9	20	Male	Fall from height	Pulmonary contusion with right-sided small pneumothorax	BP: 91/80 mmHg, HR: 123 bpm, RR: 29 bpm, SpO_2: 95%	What should be the immediate clinical approach and need for intervention?
10	34	Female	Pedestrian struck by car	Massive hemothorax requiring immediate decompression	BP: 111/69 mmHg, HR: 123 bpm, RR: 20 bpm, SpO_2: 89%	What should be the immediate clinical approach and need for intervention?
11	50	Female	High-speed motor vehicle collision	Cardiac tamponade signs with hypotension and tachycardia	BP: 107/59 mmHg, HR: 121 bpm, RR: 33 bpm, SpO_2: 88%	What should be the immediate clinical approach and need for intervention?
12	53	Female	Industrial accident	Jugular venous distention and muffled heart sounds	BP: 100/40 mmHg, HR: 119 bpm, RR: 27 bpm, SpO_2: 90%	What should be the immediate clinical approach and need for intervention?
13	75	Male	High-speed motor vehicle collision	Pulmonary contusion with right-sided small pneumothorax	BP: 108/52 mmHg, HR: 97 bpm, RR: 22 bpm, SpO_2: 95%	What should be the immediate clinical approach and need for intervention?
14	69	Female	Fall from height	Cardiac tamponade signs with hypotension and tachycardia	BP: 83/41 mmHg, HR: 91 bpm, RR: 21 bpm, SpO_2: 96%	What should be the immediate clinical approach and need for intervention?
15	52	Female	Pedestrian struck by car	Flail chest with bilateral pulmonary contusion	BP: 86/76 mmHg, HR: 119 bpm, RR: 36 bpm, SpO_2: 89%	What should be the immediate clinical approach and need for intervention?
16	40	Female	Fall from height	Pulmonary contusion with right-sided small pneumothorax	BP: 97/60 mmHg, HR: 106 bpm, RR: 22 bpm, SpO_2: 90%	What should be the immediate clinical approach and need for intervention?
17	31	Male	Motorcycle crash	Tracheobronchial injury suspected with severe air leak	BP: 106/78 mmHg, HR: 134 bpm, RR: 19 bpm, SpO_2: 87%	What should be the immediate clinical approach and need for intervention?

Table 1. Continued.

Case ID	Age	Gender	Mechanism of injury	Clinical findings	Vital signs	Primary question
18	66	Male	High-speed motor vehicle collision	Open chest wound with paradoxical movement	BP: 102/58 mmHg, HR: 114 bpm, RR: 34 bpm, SpO_2: 85%	What should be the immediate clinical approach and need for intervention?
19	60	Female	Motorcycle crash	Massive hemothorax requiring immediate decompression	BP: 89/48 mmHg, HR: 101 bpm, RR: 29 bpm, SpO_2: 96%	What should be the immediate clinical approach and need for intervention?
20	19	Male	High-speed motor vehicle collision	Sternal fracture with anterior chest wall deformity	BP: 118/69 mmHg, HR: 134 bpm, RR: 28 bpm, SpO_2: 96%	What should be the immediate clinical approach and need for intervention?
21	23	Male	Assault with blunt object	Flail chest with bilateral pulmonary contusion	BP: 103/44 mmHg, HR: 101 bpm, RR: 34 bpm, SpO_2: 87%	What should be the immediate clinical approach and need for intervention?
22	34	Male	Assault with blunt object	Multiple left rib fractures with minimal hemothorax	BP: 74/44 mmHg, HR: 122 bpm, RR: 27 bpm, SpO_2: 94%	What should be the immediate clinical approach and need for intervention?
23	30	Male	Motorcycle crash	Massive hemothorax requiring immediate decompression	BP: 98/75 mmHg, HR: 101 bpm, RR: 27 bpm, SpO_2: 92%	What should be the immediate clinical approach and need for intervention?
24	32	Female	Industrial accident	Open chest wound with paradoxical movement	BP: 84/80 mmHg, HR: 100 bpm, RR: 30 bpm, SpO_2: 90%	What should be the immediate clinical approach and need for intervention?
25	44	Female	High-speed motor vehicle collision	Open chest wound with paradoxical movement	BP: 117/80 mmHg, HR: 102 bpm, RR: 32 bpm, SpO_2: 98%	What should be the immediate clinical approach and need for intervention?
26	40	Female	Assault with blunt object	Pulmonary contusion with right-sided small pneumothorax	BP: 70/59 mmHg, HR: 98 bpm, RR: 29 bpm, SpO_2: 92%	What should be the immediate clinical approach and need for intervention?
27	21	Male	Fall from height	Tracheobronchial injury suspected with severe air leak	BP: 96/46 mmHg, HR: 117 bpm, RR: 31 bpm, SpO_2: 88%	What should be the immediate clinical approach and need for intervention?
28	41	Female	High-speed motor vehicle collision	Cardiac tamponade signs with hypotension and tachycardia	BP: 94/75 mmHg, HR: 97 bpm, RR: 36 bpm, SpO_2: 90%	What should be the immediate clinical approach and need for intervention?
29	59	Male	Industrial accident	Pulmonary contusion with right-sided small pneumothorax	BP: 90/42 mmHg, HR: 140 bpm, RR: 27 bpm, SpO_2: 95%	What should be the immediate clinical approach and need for intervention?
30	56	Male	Industrial accident	Massive hemothorax requiring immediate decompression	BP: 76/40 mmHg, HR: 128 bpm, RR: 34 bpm, SpO_2: 95%	What should be the immediate clinical approach and need for intervention?

BP, blood pressure; HR, heart rate; RR, respiratory rate.

The implementation of such scoring systems is essential for validating AI tools in critical care environments. While high concordance implies safe and actionable recommendations, discordant outputs underscore the need for clinician oversight and further training of these models.

ChatGPT-4 achieved concordant responses in 25 out of 30 cases (83.3%). Gemini 1.5 achieved concordant responses in 18 out of 30 cases (60.0%; Figure 1).

ChatGPT-4 demonstrated a significantly higher concordance with trauma management guidelines,achieving an accuracy of 83.3%, compared to Gemini 1.5's 60.0%. This discrepancy suggests that ChatGPT-4 may offer more reliable clinical decision support in acute trauma scenarios, aligning more consistently with established protocols such as ATLS. These findings underscore the potential of LLMs to augment emergency triage and intervention planning, provided they are appropriately validated and supervised.

Contingency table:

- Both correct: 16
- ChatGPT only correct: 9
- Gemini only correct: 2
- Both incorrect: 3 McNemar test statistic = 2.00, p-value = 0.0654
- Interpretation: A statistically significant result ($p < 0.05$) indicates a difference in accuracy between the two models.

The McNemar test was applied to evaluate whether the difference in accuracy between ChatGPT and Gemini was statistically significant. The test focuses specifically on discordant pairs—instances where one model was correct and the other was not (Table 3). The analysis yielded a test statistic of 2.00 (p = 0.0654).

• Interpretation: While ChatGPT outperformed Gemini in more cases (9 vs. 2), the difference did not reach statistical significance at the conventional threshold (p < 0.05). This suggests a trend toward superior performance by ChatGPT, but the evidence is insufficient to conclusively reject the null hypothesis. Larger datasets or stratified subgroup analyses may be required for more robust inference.

Table 2. Characteristics of concordant response scoring.

Feature	Description
Definition	Binary assessment of whether the AI output adheres to recognized clinical guidelines.
Scoring method	1 = guideline-concordant, 0 = discordant or inappropriate.
Assessment criteria	Timeliness, accuracy, clinical appropriateness, and intervention prioritization.
Review process	Scored by trauma care experts or automated algorithms using decision matrices.
Use case	Evaluating AI reliability in trauma triage and acute management contexts.

AI, artificial intelligence.

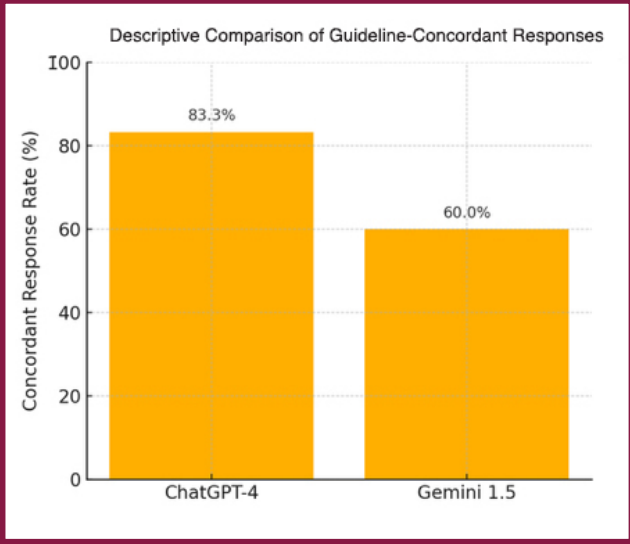

Figure 1. Comparison of guideline-concordant response rates between ChatGPT-4 and Gemini 1.5 across 30 synthetic high-energy thoracic trauma cases.

Cohen's kappa score was 0.15, indicating the level of agreement between the two models' responses (Table 4).

Cohen's Kappa coefficient is a statistical measure used to assess the level of agreement between two raters or models beyond chance. This analysis quantifies how often ChatGPT and Gemini provided the same classification (i.e., guideline-concordant or discordant responses) across 30 trauma cases.

The resulting Cohen's Kappa score of 0.15 indicates only slight agreement between the two AI models. According to the commonly accepted benchmarks (Landis and Koch, 1977), this suggests that the models often diverged in their decisions. Low inter-model agreement may reflect differences in algorithmic reasoning, training data, or interpretation of clinical priorities.

• Implications: While both models aim to mimic clinical reasoning, the modest level of agreement highlights variability in AI-generated recommendations. This variability underscores the necessity for human oversight and systematic validation of AI models before integration into clinical workflows.

Chi-square statistic = 2.95, p-value = 0.0856

• Interpretation: This test assesses whether the overall success rates of the two models differ significantly.

The chi-square test was conducted to determine whether the proportion of guideline-concordant responses differs significantly between ChatGPT-4 and Gemini 1.5 (Table 5). The resulting chi-square statistic was 2.95 (p = 0.0856).

• Interpretation: Although ChatGPT-4 showed a numerically higher rate of concordant responses (83.3%) compared to Gemini 1.5 (60.0%), the observed difference

Table 3. Contingency table of concordant responses between models.

Comparison	Number of cases
Both models correct	16
ChatGPT only correct	9
Gemini only correct	2
Both models incorrect	3

Table 4. Agreement assessment between ChatGPT and Gemini.

Metric	Value
Cohen's Kappa score	0.15
Interpretation	Slight agreement

Table 5. Frequency distribution of guideline-concordant and discordant responses.

Model	Concordant	Discordant
ChatGPT-4	25	5
Gemini 1.5	18	12

did not reach statistical significance at the conventional alpha level ($p < 0.05$). This indicates that while ChatGPT-4 appears more consistent with clinical guidelines, the sample size may be insufficient to definitively establish superiority using this method alone.

- Clinical implication: Chi-square analysis supports preliminary findings but also underscores the importance of larger sample sizes and complementary statistical methods for robust comparative assessments.

The ROC analysis comparing Gemini's concordance with ChatGPT's reference standard yielded an AUC of 0.62.

The ROC curve compares the classification performance of Gemini 1.5, using ChatGPT-4's decisions as the reference standard (Figure 2). The area under the ROC curve (AUC) was calculated to be 0.62.

- Interpretation: The AUC reflects the model's ability to distinguish between concordant and discordant classifications. An AUC value of 0.5 indicates no discriminative ability, while a value close to 1.0 indicates excellent performance. The observed AUC of 0.62 suggests a moderate ability of Gemini 1.5 to emulate ChatGPT-4's decision-making pattern, but it indicates limitations in reliability and clinical precision.
- Clinical relevance: Although ROC analysis is typically used in binary classification contexts with a definitive ground truth, its application here illustrates the divergence in decision patterns between AI models. This supports the broader argument that AI model selection and validation must be context-specific, particularly in safety-critical domains such as trauma care.

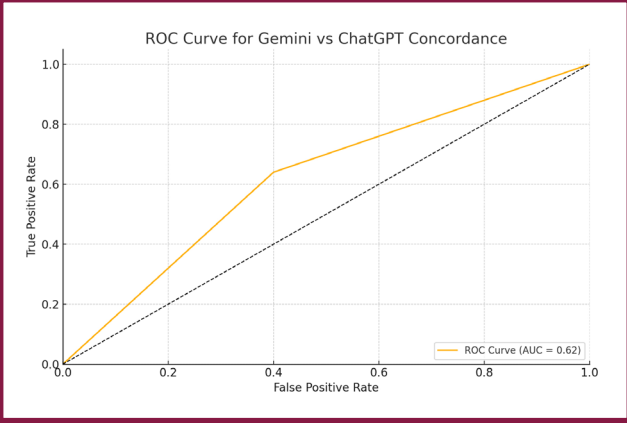

Figure 2. ROC curve showing Gemini 1.5 discriminative performance using ChatGPT-4 as reference.
ROC, receiver operating characteristic.

Discussion

Previous comparative studies evaluating large language models in acute clinical scenarios have reported variability in temporal accuracy and guideline adherence. showing that ChatGPT provided more guideline-consistent and timely interventions in emergency neurosurgical contexts than Gemini. The authors emphasized the significance of timeliness and context awareness in AI-generated clinical recommendations—both of which are critical in trauma scenarios. These findings align with our observations, particularly in high-stakes situations such as the management of tension pneumothorax or cardiac tamponade, where delays can result in catastrophic outcomes.

The findings of this study underscore the potential of AI models in enhancing clinical decision-making in high-stakes environments such as thoracic trauma care. ChatGPT-4 demonstrated superior guideline adherence compared to Gemini 1.5, particularly in cases requiring immediate and protocol-driven interventions. This observation aligns with previous literature, in which ChatGPT has shown relatively high performance in structured clinical assessments (11,12).

Despite promising results, the slight agreement between ChatGPT-4 and Gemini (Cohen's kappa = 0.15) suggests significant variability in how each model interprets and applies clinical guidelines. This phenomenon has been documented in other comparative studies, such as those evaluating AI model performance on neurosurgical board examinations and radiologic decision-making tasks (13,14). The inconsistency between models highlights the influence of underlying architectures, training corpora, and inference mechanisms on the clinical applicability of AI-generated responses.

Concordant response scoring in this context provided a standardized, objective measure of AI performance. This methodology has gained traction in recent AI benchmarking efforts within the healthcare domain (15,16). Our findings contribute to this growing body of literature by applying such metrics in a trauma-specific setting—an area that demands rapid, accurate, and guideline-concordant decisions (Table 6).

While ChatGPT-4 achieved a higher rate of correct responses, the lack of statistical significance in McNemar's test ($p = 0.0654$) and the chi-square test ($p = 0.0856$) underscores the need for cautious interpretation. Larger datasets or prospective validation in real-world clinical encounters may be required to confirm the superiority. Additionally, ROC curve analysis yielded an AUC of 0.62 for Gemini 1.5, indicating limited but detectable discriminative ability when benchmarked against ChatGPT-4. Importantly, the statistical agreement identified by McNemar's test does

not imply clinical equivalence between the models and, therefore, has limited interpretive value in evaluating their real-world performance. Similarly, the ROC results should be interpreted as an exploratory comparison of relative discriminative behavior between models rather than as a measure of clinical diagnostic accuracy.

The modest AUC and observed discordance raise important considerations regarding the use of AI in critical care. As Haemmerli et al. (17) and Aghamaliyev et al. (18) have noted, AI tools must be evaluated not only for accuracy but also for the potential consequences of delayed or inappropriate recommendations. In trauma care—where decisions often have immediate life-or-death consequences—this requirement is paramount.

Table 6. Case-by-case concordance scores.

Case ID	ChatGPT score (guideline-concordant)	Gemini score (guideline-concordant)
1	1	0
2	1	0
3	1	1
4	1	0
5	1	1
6	1	1
7	0	0
8	1	0
9	1	0
10	1	1
11	1	0
12	1	1
13	1	1
14	1	1
15	1	1
16	1	1
17	1	1
18	1	1
19	0	0
20	1	1
21	0	1
22	1	1
23	1	1
24	1	0
25	0	0
26	1	0
27	1	1
28	1	0
29	0	1
30	1	1

Moreover, the pattern of conservative responses observed in Gemini mirrors findings from previous assessments in ophthalmology and otolaryngology, where the model often opted for watchful waiting or further diagnostic evaluation rather than decisive action (19,20). While such caution may be appropriate in certain contexts, it can be detrimental in acute trauma scenarios.

In contrast, ChatGPT-4's more assertive clinical posture reflects its alignment with structured, guideline-based responses, making it potentially more suitable for integration into triage decision-support tools. Nonetheless, the necessity of clinical oversight cannot be overstated, as even small inaccuracies in high-acuity settings can have cascading effects.

In summary, our findings support the hypothesis that ChatGPT-4 is more consistent with established trauma protocols than Gemini 1.5. However, both models exhibited limitations, particularly in inter-model reliability. These results reinforce the need for hybrid models of care in which AI augments, but does not replace, human clinical judgment.

Future Directions

To further validate and refine the role of AI in trauma care, several avenues of research should be pursued. First, prospective studies using real patient data—preferably in multicenter emergency or trauma settings—are essential to evaluate AI performance under real-world constraints, including time pressure, incomplete data, and diagnostic ambiguity.

Second, future work should aim to develop more nuanced and multidimensional evaluation metrics, incorporating partial correctness, clinical rationale quality, and decision-making efficiency alongside strict guideline adherence. The use of expert panels and Delphi methods may aid in establishing more robust scoring frameworks.

Additionally, future studies will incorporate repeated querying on different days, randomization of case order, and session resets such as clearing cache/cookies to evaluate reproducibility and quantify model-level variability. These methodological refinements will help assess intra-model variability and enhance the robustness of comparative LLM analyses.

Third, expanding the sample size and diversity of trauma mechanisms (e.g., penetrating injuries, blast trauma, pediatric cases) will enhance the generalizability and clinical utility of AI models. Stratified analysis by injury type or severity could yield insights into model-specific strengths and weaknesses.

Fourth, integration of multimodal data—including radiological images, ultrasound findings, and vital sign trends—into model prompts may significantly enhance

diagnostic accuracy and triage appropriateness. Future models capable of processing such heterogeneous inputs should be evaluated.

Lastly, the incorporation of AI into clinical workflows must be accompanied by rigorous usability studies, clinician-AI interaction analyses, and ethical evaluations. Ensuring transparency, interpretability, and accountability in AI-driven recommendations is paramount, especially in high-stakes domains like trauma surgery.

Study Limitations

Despite its structured methodology and expert-reviewed design, this study possesses several limitations. First, the use of synthetic trauma cases, while beneficial for standardization and reproducibility, may not fully replicate the complexity and variability observed in real-world clinical encounters. As such, generalizability to live patient care settings is limited and necessitates cautious interpretation of findings.

Second, the binary concordant scoring system, although practical for assessing guideline adherence, lacks granularity in capturing partial correctness or contextual appropriateness. This may oversimplify the spectrum of clinical acceptability, especially in nuanced cases where multiple valid management strategies may coexist.

Third, the relatively small sample size (n = 30) restricts the statistical power of comparative analyses. While trends favoring ChatGPT-4 were observed, several tests did not reach conventional thresholds of statistical significance, thereby limiting the strength of inferential claims.

Additionally, because each case was tested only once per model, the study does not capture intra-model variability across repeated sessions. This limitation has been added to acknowledge that LLM outputs may vary depending on prompt order, timing, or session resets. Future studies incorporating repeated testing will be necessary to quantify reproducibility.

Furthermore, expert assessment of AI responses, though blinded and based on consensus guidelines, is inherently subjective. Inter-rater variability was not formally quantified, which may affect the consistency of scoring across cases.

Finally, the models were evaluated in a retrospective, asynchronous simulation framework, without real-time interaction, iterative questioning, or multimodal inputs (e.g., imaging, labs), which are crucial elements in actual trauma evaluation. Therefore, performance under dynamic clinical conditions remains untested.

Conclusion

This study demonstrates that LLMs, particularly ChatGPT-4, possess promising capabilities in supporting early decision-making in high-energy thoracic trauma scenarios. ChatGPT-4 outperformed Gemini 1.5 in terms of guideline adherence, intervention prioritization, and overall clinical appropriateness. Although the statistical analyses did not yield definitive significance across all measures, the trend consistently favored ChatGPT-4.

These findings highlight the potential role of AI in supplementing clinical workflows where rapid, evidence-based decisions are critical. In the high-stakes environment of trauma care—where time-sensitive decisions directly influence morbidity and mortality—AI-based support tools may offer tangible improvements in triage, diagnosis, and immediate management, particularly in settings with limited access to experienced trauma surgeons.

Nevertheless, the variability in model performance and low inter-model agreement reinforce the importance of human oversight and rigorous validation. AI models remain prone to hallucinations, contextual misinterpretations, and overly cautious or delayed recommendations, all of which could carry significant clinical risks in acute settings.

Until such systems achieve consistent reliability, AI should be viewed as an adjunct rather than a replacement for expert clinical judgment. Moreover, their utility must be continuously re-evaluated through prospective validation studies, including real patient scenarios, multi-center collaborations, and integration with real-time decision-support systems.

Future research should focus on expanding the sample size, incorporating diverse trauma etiologies, and testing AI models across heterogeneous healthcare environments. The ultimate goal should be the development of interoperable, context-sensitive AI tools that support equitable and high-quality emergency care without compromising patient safety.

Ethics

Ethics Committee Approval: Not required.
Informed Consent: Not required.

Footnotes

Authorship Contributions

Surgical and Medical Practices: N.Ç.Y., Concept: N.Ç.Y., Design: N.Ç.Y., Data Collection or Processing: N.Ç.Y., M.S.A., O.K., Analysis or Interpretation: N.Ç.Y., O.K., Literature Search: N.Ç.Y., O.K., Writing: N.Ç.Y., M.Ö.

Conflict of Interest: No conflict of interest was declared by the author(s).

Financial Disclosure: The author(s) declared that this study received no financial support.

REFERENCES

1. Battle CE, Hutchings H, Evans PA. Risk factors that predict mortality in patients with blunt chest wall trauma: a systematic review and meta-analysis. Injury. 2012;43:8-17. [Crossref]
2. American College of Surgeons Committee on Trauma. Advanced Trauma Life Support (ATLS) Student Course Manual. 10th ed. Chicago: American College of Surgeons; 2018. [Crossref]
3. Cothren CC, Moore EE. Emergency department thoracotomy for the critically injured patient: objectives, indications, and outcomes. World J Emerg Surg. 2006;1:4. [Crossref]
4. Karmy-Jones R, Jurkovich GJ, Shatz DV, Brundage S, Wall MJ Jr, Engelhardt S, et al. Management of traumatic lung injury: a Western Trauma Association Multicenter review. J Trauma. 2001;51:1049-1053. [Crossref]
5. Gilson A, Safranek CW, Huang T, Socrates V, Chi L, Taylor RA, et al. How does ChatGPT perform on the United States Medical Licensing Examination (USMLE)? The implications of large language models for medical education and knowledge assessment. JMIR Med Educ. 2023;9:e45312. [Crossref]
6. Kung TH, Cheatham M, Medenilla A, Sillos C, De Leon L, Elepaño C, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. PLOS Digit Health. 2023;2:e0000198. [Crossref]
7. Takagi S, Watari T, Erabi A, Sakaguchi K. Performance of GPT-3.5 and GPT-4 on the Japanese medical licensing examination: comparison study. JMIR Med Educ. 2023;9:e48002. [Crossref]
8. Pal A, Sankarasubbu M. Gemini goes to med school: exploring the capabilities of multimodal large language models on medical challenge problems & hallucinations. ArXiv. 2024 Feb 12. [Crossref]
9. Mbakwe AB, Lourentzou I, Celi LA, Mechanic OJ, Dagan A. ChatGPT passing USMLE shines a spotlight on the flaws of medical education. PLOS Digit Health. 2023;2:e0000205. [Crossref]
10. Cadamuro J, Cabitza F, Debeljak Z, De Bruyne S, Frans G, Perez SM, et al. Potentials and pitfalls of ChatGPT and natural-language artificial intelligence models for the understanding of laboratory medicine test results. An assessment by the European Federation of Clinical Chemistry and Laboratory Medicine (EFLM) Working Group on Artificial Intelligence (WG-AI). Clin Chem Lab Med. 2023;61:1158-1166. [Crossref]
11. Mutlucan UO, Zortuk O, Bedel C, Selvi F, Turk CC. Comparison of Chat GPT and Gemini in neurosurgical evaluation questions. Sarcouncil Journal of Medicine and Surgery. 2024;3:10-15. [Crossref]
12. Ali R, Tang OY, Connolly ID, Zadnik Sullivan PL, Shin JH, Fridley JS, et al. Performance of ChatGPT and GPT-4 on neurosurgery written board examinations. Neurosurgery. 2023;93:1353-1365. [Crossref]
13. Hopkins BS, Nguyen VN, Dallas J, Texakalidis P, Yang M, Renn A, et al. ChatGPT versus the neurosurgical written boards: a comparative analysis of artificial intelligence/machine learning performance on neurosurgical board-style questions. J Neurosurg. 2023;139:904-911. [Crossref]
14. Krishna S, Bhambra N, Bleakney R, Bhayana R. Evaluation of reliability, repeatability, robustness, and confidence of GPT-3.5 and GPT-4 on a radiology board-style examination. Radiology. 2024;311:e232715. [Crossref]
15. Meyer A, Soleman A, Riese J, Streichert T. Comparison of ChatGPT, Gemini, and Le Chat with physician interpretations of medical laboratory questions from an online health forum. Clin Chem Lab Med. 2024;62:2425-34. [Crossref]
16. Sau S, George DD, Singh R, Kohli GS, Li A, Jalal MI, et al. Accuracy and quality of ChatGPT-4o and Google Gemini performance on image-based neurosurgery board questions. Neurosurg Rev. 2025;48:320. [Crossref]
17. Haemmerli J, Sveikata L, Nouri A, May A, Egervari K, Freyschlag C, et al. ChatGPT in glioma adjuvant therapy decision making: ready to assume the role of a doctor in the tumour board? BMJ Health Care Inform. 2023;30:e100775. [Crossref]
18. Aghamaliyev U, Karimbayli J, Giessen-Jung C, Matthias I, Unger K, Andrade D, et al. ChatGPT's gastrointestinal tumor board tango: a limping dance partner? Eur J Cancer. 2024;205:114100. [Crossref]
19. Botross M, Mohammadi SO, Montgomery K, Crawford C. Performance of Google's artificial intelligence chatbot "Bard" (now "Gemini") on ophthalmology board exam practice questions. Cureus. 2024;16:e57348. [Crossref]
20. Mete U. Evaluating the performance of ChatGPT, Gemini, and Bing compared with resident surgeons in the otorhinolaryngology in-service training examination. Turk Arch Otorhinolaryngol. 2024;62:48-57. [Crossref]