DOI: 10.24938/kutfd.1922944 Kırıkkale Üniversitesi Tıp Fakültesi Dergisi 2026;28(2):314-320
Original Article *Özgün Araştırma*

SCENARIO-BASED PREHOSPITAL TRIAGE OF CARBON MONOXIDE POISONING: COMPARING FIVE LARGE LANGUAGE MODELS WITH EMERGENCY MEDICAL SERVICES PERSONNEL IN A PROSPECTIVE STUDY

Karbonmonoksit Zehirlenmesinde Senaryo Tabanlı Hastane Öncesi Triyaj: Beş Büyük Dil Modeli ile Acil Sağlık Hizmetleri Personelinin Prospektif Karşılaştırılması

Vildan ÖZER[1] Özlem BÜLBÜL[2] Efnan BAYRAK ERBÖLÜKBAŞ[3] Serdar KARAKULLUKÇU[4] Aynur ŞAHİN[5]

[1] *Department of Emergency Medicine, Faculty of Medicine, Karadeniz Technical University, TRABZON, TÜRKİYE.*
[2] *Emergency Department, Ministry of Health A. Ilhan Ozdemir State Hospital, GİRESUN, TÜRKİYE*
[3] *Department of Health Services, Republic of Türkiye Ministry of Health, Trabzon Provincial Health Directorate, TRABZON, TÜRKİYE*
[4] *Department of Public Health, Faculty of Medicine, Karadeniz Technical University, TRABZON, TÜRKİYE.*
[5] *Department of Emergency Medicine, Medical Toxicology ICU, Basaksehir Cam and Sakura City Hospital Health Sciences University, İSTANBUL, TÜRKİYE*

ABSTRACT

Objective: Carbon monoxide (CO) poisoning requires rapid identification and timely decisions regarding the need for hyperbaric oxygen therapy (HBOT) to improve clinical outcomes. This study aimed to compare the decision-making performance of emergency medical services (EMS) personnel and large language models (LLMs) in accurately determining the need for HBOT during the prehospital phase of CO poisoning cases, and to explore the potential implementation of LLMs as decision-support tools for patient triage and referral.

Material and Methods: In this prospective, simulation-based diagnostic accuracy study, 128 standardized scenario-based clinical cases (64 requiring HBOT, 64 not requiring HBOT) were developed based on established indications. Sixty EMS personnel and five LLMs (GPT-4o, GPT-4.5, GPT-o3, Gemini 2.5 Pro, and DeepSeek-R1) evaluated each scenario. Their responses were compared with the gold standard answers established by consensus between a medical toxicologist and a hyperbaric medicine specialist.

Results: The overall accuracy of the EMS personnel was 65.1%, whereas the highest accuracy was observed with GPT-o3 (96.9%). All LLMs demonstrated 100% sensitivity, although the specificity varied, ranging from 26.6% (GPT-4.5) to 93.8% (GPT-o3). The accuracy of GPT-o3 (96.9%, p<0.001), GPT-4o (78.9%, p=0.001), and Gemini 2.5 Pro (75.8%, p=0.012) was significantly higher than that of the EMS personnel. GPT-o3 also significantly outperformed all other models (all p<0.001).

Conclusion: Our study suggests that LLMs can support prehospital decision-making, facilitating timely referral to appropriate centers, improving early access to HBOT, and potentially enhancing outcomes. However, these findings are based on standardized simulation scenarios, and real-world clinical validation is necessary before widespread implementation.

Keywords: *Artificial intelligence, carbon monoxide, Emergency medical service, poisoning, triage*

ÖZ

Amaç: Karbonmonoksit (CO) zehirlenmesi, klinik sonuçları iyileştirmek için hızlı tanımlama ve hiperbarik oksijen tedavisi (HBOT) ihtiyacına yönelik zamanında karar verilmesini gerektirir. Bu çalışma, CO zehirlenmesi vakalarında hastane öncesi evrede HBOT ihtiyacının doğru şekilde belirlenmesinde acil sağlık hizmetleri (ASH) personeli ile büyük dil modellerinin (LLM) karar verme performansını karşılaştırmayı ve LLM'lerin hasta triyajı ve sevkinde karar destek aracı olarak kullanım potansiyelini kullanımını araştırmayı amaçlamıştır.

Gereç ve Yöntemler: Bu prospektif ve simülasyon temelli tanısal doğruluk çalışmasında, yerleşik endikasyonlara dayalı olarak 128 standart vaka temelli klinik senaryo (64'ü HBOT gerektiren, 64'ü gerektirmeyen) geliştirildi. Altmış ASH personeli ve beş LLM (GPT-4o, GPT-4.5, GPT-o3, Gemini 2.5 Pro ve DeepSeek-R1) her bir senaryoyu değerlendirdi. Katılımcıların yanıtları, bir tıbbi toksikolog ve bir hiperbarik tıp uzmanı arasındaki fikir birliği ile oluşturulan altın standart cevaplarla karşılaştırıldı.

Bulgular: ASH personelinin genel doğruluk oranı %65,1 iken, en yüksek doğruluk oranı GPT-o3 modelinde (%96,9) gözlemlenmiştir. Tüm LLM'ler %100 duyarlılık göstermekle birlikte özgüllük oranları %26,6 (GPT-4.5) ile %93,8 (GPT-o3) arasında değişmiştir. GPT-o3 (%96,9, p<0,001), GPT-4o (%78,9, p=0,001) ve Gemini 2.5 Pro (%75,8, p=0,012) modellerinin doğruluk oranları ASH personeline kıyasla istatistiksel olarak anlamlı düzeyde yüksek bulunmuştur. Ayrıca GPT-o3, diğer tüm modellerden anlamlı derecede daha iyi performans göstermiştir (tümü için p<0,001).

Sonuç: Çalışmamız, LLM'lerin hastane öncesi karar verme süreçlerini destekleyebileceğini, uygun merkezlere zamanında sevki kolaylaştırabileceğini, HBOT'ye erken erişimi artırabileceğini ve potansiyel olarak klinik sonuçları iyileştirebileceğini öne sürmektedir. Bununla birlikte, bu bulgular standart simülasyon senaryolarına dayanmaktadır ve yaygın uygulama öncesinde gerçek dünya klinik doğrulamasının yapılması gerekmektedir.

Anahtar Kelimeler: *Yapay zekâ, karbonmonoksit, acil sağlık hizmeti, zehirlenme, triyaj*

Correspondence/Yazışma Adresi: **Dr. Vildan ÖZER**
Department of Emergency Medicine, Faculty of Medicine, Karadeniz Technical University, TRABZON, TÜRKİYE.
Phone/Tel: +905556216171 **E-mail**/E-posta: dr.vilzan@hotmail.com
Received/Geliş Tarihi: 03.04.2026 **Accepted**/Kabul Tarihi: 04.05.2026

KÜ Tıp Fak Derg 2026;28(2):314-320
Doi: 10.24938/kutfd.1922944

INTRODUCTION

The global cumulative incidence and mortality of carbon monoxide (CO) poisoning are currently estimated at 137 cases and 4.6 deaths per million population, respectively. Although CO-related mortality has declined by approximately 36% over the past 25 years, the incidence of CO poisoning has remained stable during this period.[1] CO poisoning primarily affects the cardiovascular and neurological systems. It is also a well-recognized cause of delayed neuropsychiatric sequelae (DNS), which can emerge days to weeks after exposure and significantly impair long-term neurological function.[2,3]

Hyperbaric oxygen therapy (HBOT) prevents cerebral oxidative damage and late neurological symptoms in moderate-to-severe cases of CO poisoning.[3-6] Studies have demonstrated that early administration of HBOT, particularly within 6 h of exposure, can significantly improve neurocognitive outcomes. Delays beyond 6 h, however, are associated with a worse neurological prognosis.[5,6] As HBOT requires specialized facilities that are not available at all hospitals, the early identification of suitable patients and timely referral to HBO-capable centers during the prehospital phase are critical for optimal outcomes.[7-9]

In recent years, large language models (LLMs) have demonstrated significant potential across various fields of medicine, including clinical decision support, diagnostics, and patient triage.[10,11] In this study, we aimed to compare the decision-making performance of emergency medical services (EMS) personnel and different LLMs in terms of diagnostic accuracy for determining the need for HBOT during the prehospital phase of CO poisoning cases, and to explore the potential of LLMs as decision-support tools for patient triage and referral.

MATERIALS AND METHODS

Study Design and Setting

This prospective simulation-based diagnostic accuracy study, conducted between May 1 and May 30, 2025, compared the performance of EMS personnel and various LLMs in determining the need for HBOT in CO poisoning scenarios during the prehospital phase. This study used standardized, scenario-based simulated cases to evaluate diagnostic decision-making performance. Scenario-based simulated cases related to CO poisoning were developed to address the study objectives. These scenarios were presented to EMS personnel and LLMs (GPT-4o, GPT-4.5, GPT-o3, Gemini 2.5 Pro, and DeepSeek-R1), and their responses were compared with gold standard (GS) answers, established through consensus between a medical toxicologist and a hyperbaric medicine specialist. An overview of the study design, including the evaluated models, clinical scenario structure, comparison groups, and performance metrics, is shown in Figure 1.

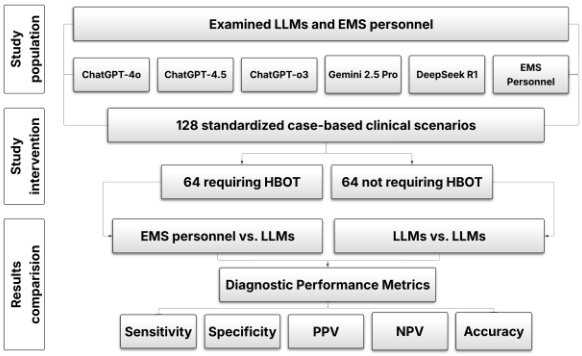

Figure 1: Flowchart of the study

Selection of Participants

The study included EMS personnel actively providing services, as well as several LLMs (GPT-4o, GPT-4.5, GPT-o3, Gemini 2.5 Pro, and DeepSeek-R1). The LLMs were prompted to analyze the clinical scenarios using a standardized case analysis form. In each case, the participants and LLMs were tasked with determining whether the patient required referral to a HBOT-capable center. Participants were tasked with evaluating 128 standardized clinical scenarios to determine the need for HBOT. Detailed information on the participants is provided in the following sections.

EMS Personnel

EMS personnel were recruited on a voluntary basis from paramedics and emergency medical technicians actively serving within the 112 command and control center. EMS personnel included paramedics and emergency medical technicians working as ambulance crew members or emergency call center operators within the 112 command and control center of a city with a population of approximately 820,000. Both ambulance crews and call center operators are authorized to make decisions regarding patient referral to the appropriate medical centers based on the initial scene assessment. Prior to the evaluation, participants were informed verbally about the general nature and purpose of the study. They were not informed of the specific content of the scenarios in advance. Participants were explicitly instructed not to consult any external resources, clinical protocols, reference materials, or artificial intelligence tools during the evaluation. The evaluation was conducted online in a supervised setting, with all participants completing the scenarios simultaneously, thereby preventing any communication or discussion of case content between participants. No formal time limit was imposed, and the estimated completion time was not predetermined. The minimum required sample size for EMS personnel was calculated as 52 using G*Power (two-sided Z-test, EMS accuracy 65%, 20% difference,

KÜ Tıp Fak Derg 2026;28(2):314-320
Doi: 10.24938/kutfd.1922944

α=0.05, power 0.80). All EMS personnel who completed the full set of clinical scenarios were included in the final analysis.

Gold Standard

GS (reference standart) responses regarding HBOT referral in CO poisoning scenarios were determined through consensus between a medical toxicologist and a hyperbaric medicine physician. Discrepancies were resolved by a third medical toxicologist.

Large Language Models

The selected LLMs were chosen to include models developed by different organizations and to allow comparison across commonly used general-purpose language models. All LLMs used in this study (GPT-4o, GPT-4.5 [research preview version], GPT-o3 [OpenAI, L.L.C., San Francisco, CA, USA], Gemini 2.5 Pro [preview version] [Google AI, Mountain View, CA, USA], and DeepSeek-R1 [DeepSeek Inc., Hangzhou, China]) were accessed via their official web-based interfaces, without the use of APIs or third-party platforms. Each scenario was submitted in a separate session to prevent memory effects and contextual carryover. A standardized prompt was used: "I would like to ask your opinion on a prehospital clinical decision. Please read the case below and choose the most appropriate option and briefly explain your reasoning in 1–2 sentences." The prompt used in this study was intentionally designed to be simple and standardized in order to minimize potential bias arising from advanced prompt engineering techniques and to reflect real-world prehospital decision-making conditions. The same prompt was applied across all scenarios and models to ensure comparability. Advanced prompt engineering strategies were not employed. This approach was chosen to evaluate the models' basic clinical reasoning under realistic conditions and to reflect routine use by EMS personnels. The models were expected to indicate whether a referral to an HBOT center was necessary, along with a brief rationale. Sample responses for each model are provided in Appendix 1. All models evaluated the same 128 scenarios under identical conditions.

Scenario-based simulated cases development and presentation

The number of scenarios was determined a priori using MedCalc statistical software (v.23.2.1) (single proportion, 95% CI, expected proportion 80%, CI width 20%), which indicated a minimum of 60 scenarios per group. Accordingly, 128 scenario-based simulated cases (64 requiring and 64 not requiring HBOT) were developed in total, ensuring a balanced binary design. These scenario-based simulated cases were developed using a structured approach informed by established HBOT indications as described in Goldfrank's Toxicologic Emergencies.[12] The scenario set was intentionally balanced to include an equal number of cases requiring and not requiring HBOT in order to enable fair performance comparison across models and participants. All 128 scenarios were constructed de novo by an emergency medicine physician based on current literature and clinical expertise and were independently reviewed by a second emergency medicine physician to ensure clinical accuracy, relevance, and consistency with prehospital practice standards.

Each scenario was designed to represent realistic prehospital conditions and included standardized clinical variables such as patient demographics, medical history, exposure characteristics, vital signs, electrocardiographic findings, physical examination findings, and carboxyhemoglobin (COHb) levels. All scenarios are provided in full in Appendix 2.

No formal Delphi process was conducted; however, iterative expert review was used to ensure consistency, clarity, and alignment with real-world prehospital decision-making. To minimize the ordering bias, the 128 clinical scenarios were randomized prior to administration using the Fisher–Yates shuffle algorithm implemented in Python.[13] A single randomized sequence was generated and applied uniformly to all EMS personnel and LLMs, thereby maintaining identical presentation conditions across all participants. The participants and models evaluated each scenario to determine whether referral to an HBOT-capable center was necessary.

Data Collection and Statistical Analysis

For each clinical scenario, EMS personnel and LLMs were required to make a referral decision by selecting one of two options: (1) referral to an HBOT-capable center or (2) transfer to the nearest healthcare facility without HBOT capability. Responses were coded in a binary format (Yes/No) to indicate whether referral to an HBOT center was deemed necessary. These responses were compared with the GS answers established by consensus between a medical toxicologist and a hyperbaric medicine physician. The primary outcome measures included sensitivity, specificity, positive predictive value, negative predictive value, and overall accuracy for each participant group relative to the GS. Statistical analyses were conducted using SPSS 27.0 (IBM Corp, Armonk, NY) and MedCalc Statistical Software (v.23.2.1). Descriptive statistics were presented as frequencies and percentages for categorical variables and as means with standard deviations or medians with interquartile ranges for continuous variables, as appropriate. A p-value of <0.05 was considered statistically significant. The study was approved by the Scientific Research Ethics Committee of Karadeniz Technical University (Date: 07/04/2025, Number: 2025/96) and conducted in accordance with the Declaration of Helsinki.

KÜ Tıp Fak Derg 2026;28(2):314-320
Doi: 10.24938/kutfd.1922944

RESULTS

A total of 128 scenario-based clinical cases were developed (64 requiring HBOT and 64 not requiring it). These scenarios were presented to EMS personnel and LLMs. Initially, 93 paramedics participated, but 33 were excluded because of incomplete scenario evaluations. Therefore, 60 EMS personnel were included in the final analysis. The demographic characteristics of the EMS personnel are summarized in Table 1. The diagnostic performance metrics for EMS personnel and LLMs in determining the need for HBOT are shown in Table 2. The overall accuracy for EMS personnel was 65.1%, whereas the highest accuracy was observed with the GPT-o3 model at 96.9%. Although all LLMs demonstrated 100% sensitivity, there were notable differences in specificity. The lowest specificity was observed in GPT-4.5 (26.6%), whereas GPT-o3 achieved the highest specificity at 93.8%. Pairwise comparisons between EMS personnel and LLMs are presented in Table 3, Figure 2 and Figure 3.

Table 1: Demographic characteristics of EMS personnel

Variable	n	(%)
Sex		
Female	42	(70.0)
Male	18	(30.0)
Profession		
Emergency medical technician	33	(55.0)
Paramedic	27	(45.0)
Age Group		
≤ 35 years	26	(43.3)
> 35 years	34	(56.7)
Years of Experience		
< 5 years	8	(13.3)
≥ 5 years	52	(86.7)
Work Unit		
Ambulance crew	39	(65.0)
Command and control center	21	(35.0)

EMS: Emergency medical services

Table 2: Diagnostic performance metrics of EMS personnel and AI models in determining the need for HBOT

Model	Sensitivity (95% CI)		Specificity (95% CI)		PPV (95% CI)		NPV (95% CI)		Accuracy (95% CI)	
EMS personnel	71.6	(70.1–73.0)	58.7	(57.1–60.3)	63.4	(62.4–64.4)	67.4	(66.1–68.6)	65.1	(64.1–66.2)
GPT-o3	100.0	(94.4–100.0)	93.8	(84.8–98.3)	94.1	(86.1–97.6)	100.0	(94.0–100.0)	96.9	(92.2–99.1)
GPT-4o	100.0	(94.4–100.0)	57.8	(44.8–70.1)	70.3	(64.0–76.0)	100.0	(90.5–100.0)	78.9	(70.8–85.6)
Gemini 2.5 Pro	100.0	(94.4–100.0)	51.6	(38.7–64.3)	67.4	(61.6–72.7)	100.0	(89.4–100.0)	75.8	(67.4–82.9)
DeepSeek-R1	100.0	(94.4–100.0)	32.8	(21.6–45.7)	59.8	(55.6–63.9)	100.0	(83.9–100.0)	66.4	(57.5–74.5)
GPT-4.5	100.0	(94.4–100.0)	26.6	(16.3–39.1)	57.7	(54.0–61.2)	100.0	(80.5–100.0)	63.3	(54.3–71.6)

EMS: Emergency medical services, AI: Artificial intelligence, HBOT: Hyperbaric oxygen therapy, PPV: Positive predictive value, NPV: Negative predictive value, CI: Confidence interval

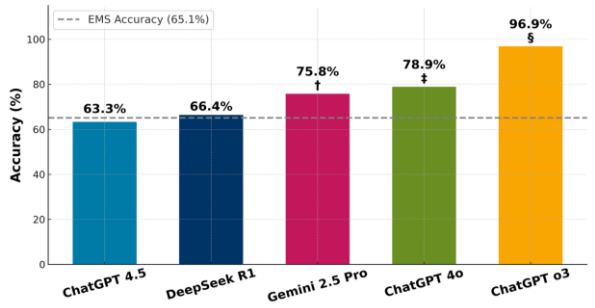

Figure 2: Comparative accuracy analysis between EMS personnel and individual AI models in determining the need for HBOT; symbols indicate statistically significant differences compared to EMS: †p = 0.012; ‡p = 0.001; and §p < 0.001

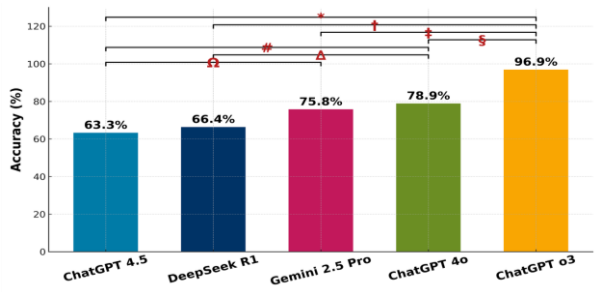

Figure 3: Comparative accuracy analysis between individual AI models in determining the need for HBOT; symbols indicate statistically significant differences between models: *p < 0.001; †p < 0.001; ‡p < 0.001; §p < 0.001; #p = 0.006; Δp = 0.025; and Ωp = 0.030

Table 3: Pairwise accuracy comparisons between EMS personnel and AI models in determining the need for HBOT.

Comparison	Accuracy (%) (Group 1 vs. Group 2)	p
EMS personnel vs. GPT-o3	65.1 vs. 96.9	< 0.001
EMS personnel vs. GPT-4o	65.1 vs. 78.9	0.001
EMS personnel vs. Gemini 2.5 Pro	65.1 vs. 75.8	0.012
EMS personnel vs. DeepSeek-R1	65.1 vs. 66.4	0.766
EMS personnel vs. GPT-4.5	65.1 vs. 63.3	0.661
GPT-o3 vs. GPT-4o	96.9 vs. 78.9	< 0.001
GPT-o3 vs. Gemini 2.5 Pro	96.9 vs. 75.8	< 0.001
GPT-o3 vs. DeepSeek-R1	96.9 vs. 66.4	< 0.001
GPT-o3 vs. GPT-4.5	96.9 vs. 63.3	< 0.001
GPT-4o vs. Gemini 2.5 Pro	78.9 vs. 75.8	0.550
GPT-4o vs. DeepSeek-R1	78.9 vs. 66.4	0.025
GPT-4o vs. GPT-4.5	78.9 vs. 63.3	0.006
Gemini 2.5 Pro vs. DeepSeek-R1	75.8 vs. 66.4	0.098
Gemini 2.5 Pro vs. GPT-4.5	75.8 vs. 63.3	0.030
DeepSeek-R1 vs. GPT-4.5	66.4 vs. 63.3	0.601

Abbreviations: EMS: Emergency medical services, **AI:** Artificial intelligence, **HBOT:** Hyperbaric oxygen therapy.

KÜ Tıp Fak Derg 2026;28(2):314-320
Doi: 10.24938/kutfd.1922944

DISCUSSION

HBOT administered within the first 6 h of CO poisoning significantly improves long-term neurocognitive prognosis and reduces mortality. Conversely, delays in accessing HBOT are associated with worse neurological outcomes.[5,6,14] Therefore, early recognition and timely referral to HBOT-capable centers are important. In this study, EMS personnel made correct decisions in approximately 65% of the scenarios; they failed to identify certain HBOT indications. In contrast, all LLMs demonstrated 100% sensitivity, with specificity varying; GPT-o3 exhibited the highest diagnostic accuracy. Our study represents one of the first investigations into the feasibility of using LLM-supported decision systems for prehospital HBOT referral in CO poisoning. These findings highlight the potential of GPT-o3, in particular, to serve as an effective clinical decision-support tool for the accurate identification of HBOT requirements and ensuring timely referral to appropriate centers during the prehospital phase.

CO poisoning affects approximately 50,000 individuals annually in the United States alone, and 15%–40% of those affected develop persistent neurocognitive and psychiatric impairments. In addition, long-term mortality rates are elevated among those exposed to CO.[14-16] Although HBOT has been shown to reduce mortality and DNS while improving neurocognitive outcomes, it remains unavailable in many healthcare facilities.[7-9] Therefore, it is crucial for EMS personnel to promptly identify patients with suspected CO poisoning at the scene and direct them to appropriate centers, thereby facilitating early access to HBOT and improving the clinical outcomes. Considering that the risk of DNS can be minimized if HBOT is initiated within 6 h, accurate prehospital triage and timely referral are critical. Previous studies have demonstrated that prehospital triage by EMS personnel reduces mortality and expedites treatment in time-sensitive emergencies, such as STEMI, trauma, and stroke.[17-19] Moreover, recent studies have highlighted the effectiveness of AI-assisted prehospital triage systems.[20,21] In our study, EMS personnel demonstrated relatively low accuracy (65.1%) in identifying HBOT indications during the prehospital phase. This may be attributed to limited training, lack of medical knowledge, or the missing of critical clinical cues. Additionally, the use of simulated scenarios, rather than real patients may have diminished the realism and clinical judgment of EMS personnel. In contrast, all LLMs achieved 100% sensitivity in detecting the need for HBOT. These findings suggest that LLMs can support EMS personnel in the early prehospital assessment of HBOT requirements and serve as valuable decision-support tools in this context. The need for standardized triage protocols is also evident in the existing literature. For instance, Kijanka et al. proposed a three-step algorithm for the prehospital CO poisoning diagnosis that incorporates exposure history, symptom evaluation, and on-site COHb measurement. The algorithm also recommends teleconsultation with an HBOT center to assess the need for treatment and arrange for transfer.[22] Similarly, Hampson et al. found that on-site COHb measurement reduced the time to HBOT by an average of 1 h.[23] These findings emphasize the importance of rapid on-scene assessment and early referral, suggesting that integrating LLMs into this workflow could provide substantial benefits to the healthcare system. Future studies should evaluate whether LLMs such as GPT-o3 can be integrated into teleconsultation-based triage algorithms to support decision-making when toxicologist access is limited. The present study did not assess integration feasibility or decision-making speed.

In our study, all evaluated LLMs demonstrated 100% sensitivity in determining the need for HBOT in CO poisoning cases; however, their specificity varied widely across models. The ability of LLMs to perform consistent, meticulous, and systematic analyses in structured scenarios may explain this superior performance. Furthermore, although the LLMs used in this study were trained only up to specific knowledge cut-off dates, their capacity for real-time web browsing suggests the potential for continuous updates to medical information, thereby enhancing their decision-support capabilities. These models can retrieve content from specialized academic databases, such as PubMed, enabling them to provide more reliable and scientifically validated responses, particularly in highly specialized domains such as toxicology.[24-33] This functionality may contribute to safer decision-making processes at both the individual and system levels. The 100% sensitivity observed across all LLMs indicates a robust safety net, ensuring that no patient requiring HBOT is overlooked, potentially preventing severe outcomes such as DNS or fetal complications. However, models such as GPT-4.5, which exhibit lower specificity, may lead to unnecessary referrals and inefficient use of healthcare resources. Given the high morbidity and mortality associated with CO poisoning, high sensitivity in identifying true HBOT candidates is crucial. Nonetheless, sufficient specificity is equally important for avoiding overtriage and preserving resource efficiency. Maintaining this balance is essential for optimizing both patient safety and healthcare utilization. Among the evaluated models, GPT-o3 demonstrated the highest diagnostic performance, with 100% sensitivity and 93.8% specificity. This outstanding accuracy is likely attributed to the model's advanced reasoning abilities,

KÜ Tıp Fak Derg 2026;28(2):314-320
Doi: 10.24938/kutfd.1922944

architecture optimized through large-scale reinforcement learning algorithms, and fine-tuning for complex clinical decision-making tasks.[26] GPT-o3's effectiveness stems not only from real-time information access but also from its ability to prioritize and synthesize clinical data, construct logical reasoning chains, and deploy tools strategically at the appropriate decision points. These competencies are particularly valuable in time-sensitive clinical contexts, such as CO poisoning, where context-aware and safe triage decisions are paramount.[26] As GPT-o3 was released in April 2025, this study represents the first known evaluation of its effectiveness in prehospital clinical decision-making and provides preliminary data for future research. Furthermore, while our results highlight the promising potential of LLMs in prehospital triage, it is essential to recognize that simulation-based findings may not fully reflect the complexity and unpredictability of actual clinical settings. Future research should prioritize prospective real-world validation studies to confirm the utility and safety of LLM-assisted decision support in routine prehospital care.

This study had some limitations. First, the clinical scenarios were standardized and may not fully reflect the complexity and variability encountered in real-world prehospital environments. Second, EMS personnel made decisions based on written case scenarios, which differ from actual field conditions, where visual cues, environmental context, and dynamic interactions contribute to clinical judgment. It should also be acknowledged that this design may inherently favor the LLMs, as written clinical scenarios are the native input modality of these text-based models, whereas EMS personnel typically operate under dynamic field conditions not reproduced here. The observed performance gap may therefore overestimate the real-world advantage of LLMs. Third, the study was limited to a single geographic region, and EMS training and experience may vary across different healthcare systems.Fourth, LLMs evaluated cases based on static text inputs and were not integrated into real-time workflows or equipped with voice input/output capabilities, which could influence their practical utility in the field. Additionally, because LLMs were not trained specifically on toxicology or emergency medicine data, their performance may change as newer model versions are developed or fine-tuned for clinical applications. Moreover, no formal time limit was imposed on participants during the evaluation, which does not fully reflect the time-critical nature of real-world prehospital triage, where decisions must often be made under considerable time pressure. Another important consideration is that all LLMs achieved 100% sensitivity in our analysis.Although striking, this finding may be a consequence of the structured and somewhat simplified nature of the scenario set. The clinical features pointing toward HBOT may have been presented in a clearer and more consistent manner than in real-life prehospital encounters. Finally, our findings (derived from simulation scenarios) may not directly translate to clinical outcomes. In addition, because the scenario set was artificially balanced (1:1 ratio of HBOT required and non-required cases) and may not reflect real-world prevalence, PPV and NPV may not be directly generalizable to real-world clinical settings, as these metrics are influenced by disease prevalence. This should be considered when interpreting the predictive performance of both EMS personnel and LLMs. Therefore, these results should be interpreted with caution, and prospective studies in real-world prehospital environments are required to confirm these findings and guide safe integration into practice. This study demonstrated that LLMs, particularly GPT-o3, outperformed EMS personnel in determining the need for HBOT in standardized CO poisoning scenarios. While all LLMs showed excellent sensitivity, GPT-o3 achieved both high sensitivity and specificity, suggesting its strong potential as a reliable decision-support tool during the prehospital phase. However, this study demonstrated superior diagnostic accuracy only under standardized simulation conditions; real-world integration, decision-making speed, and clinical outcomes were not assessed and require prospective validation.

Conflicts of interest: The authors have no conflict of interest relevant to this article to disclose.

Author Contributions: Concept/Design: VO, OB, EBE, AS; Analysis/Interpretation: VO, OB, SK, AS; Data Collection: VO, OB, EBE, SK, AS; Writer: VO, OB, EBE, SK, AS; Critical Review: VO, OB, EBE, SK, AS; Approver: VO, OB, EBE, SK, AS.

Support and Acknowledgements: We would like to express our sincere gratitude to Dr. Evin Koç, a physician in Underwater and Hyperbaric Medicine, for her valuable contributions in establishing the gold standard decisions regarding the need for hyperbaric oxygen therapy in this study.

Ethics Approval: The study was approved by the Scientific Research Ethics Committee of Karadeniz Technical University (Date: 07/04/2025, Number: 2025/96).

KÜ Tıp Fak Derg 2026;28(2):314-320
Doi: 10.24938/kutfd.1922944

REFERENCES

1. Mattiuzzi C, Lippi G. Worldwide epidemiology of carbon monoxide poisoning. *Hum Exp Toxicol*. 2020;39(4):387-392.
2. Liao SC, Shao SC, Yang KJ, Yang CC. Real-world effectiveness of hyperbaric oxygen therapy for delayed neuropsychiatric sequelae after carbon monoxide poisoning. *Sci Rep*. 2021;11(1):19212.
3. Arslan A. Hyperbaric oxygen therapy in carbon monoxide poisoning in pregnancy: Maternal and fetal outcome. *Am J Emerg Med*. 2021;43:41-45.
4. Weaver LK, Hopkins RO, Chan KJ, et al. Hyperbaric oxygen for acute carbon monoxide poisoning. *N Engl J Med*. 2002;347(14):1057-1067.
5. Lee Y, Cha YS, Kim SH, Kim H. Effect of hyperbaric oxygen therapy initiation time in acute carbon monoxide poisoning. *Crit Care Med*. 2021;49(10):e910-e919.
6. Jia Y, Han N, Guo H, et al. Hyperbaric oxygen therapy for acute carbon monoxide poisoning patients with coma onset. *Eur J Med Res*. 2025;30(1):125.
7. Talbot Z, Lee A, Boet S. Hyperbaric medicine in Canadian undergraduate medical school curriculum. *Diving Hyperb Med*. 2023;53(2):138-141.
8. Wang SJ, Kang P. Distribution of Hyperbaric Oxygen Chambers for Noxious Gas Disaster. *Prehosp Disaster Med*. 2023;38(1):127-128.
9. NHS England. Reviewing Hyperbaric Oxygen Services: Consultation Guide. Accessed date:10 March 2026. Available from: https://www.england.nhs.uk/long-read/ reviewing-hyperbaric-oxygen-services-consultation-guide/.
10. Piliuk K, Tomforde S. Artificial Intelligence in Emergency Medicine. A Systematic Literature Review. *Int J Med Inform*. 2023;180:105274.
11. Paslı S, Şahin AS, Beşer MF, Topçuoğlu H, Yadigaroğlu M, İmamoğlu M. Assessing the precision of artificial intelligence in emergency department triage decisions: Insights from a study with ChatGPT. *Am J Emerg Med*. 2024;78:170-175.
12. Tomaszewski C. Carbon monoxide. In: Nelson LS, Howland MA, Lewin NA, Smith SW, Goldfrank LR, Hoffman RS, eds. Goldfrank's Toxicologic Emergencies. 11th ed. New York. McGraw-Hill, 2019:1663-1675.
13. Fisher RA, Yates F. Statistical Tables for Biological, Agricultural and Medical Research. 6th ed. London. Oliver and Boyd, 1963.
14. Rose JJ, Nouraie M, Gauthier MC, et al. Clinical outcomes and mortality impact of hyperbaric oxygen therapy in patients with carbon monoxide poisoning. Crit Care Med. 2018;46(7):e649-655.
15. Rose JJ, Wang L, Xu Q, et al. Carbon monoxide poisoning: pathogenesis, management, and future directions of therapy. Am J Respir Crit Care Med. 2017;195(5):596-606.
16. Weaver L, Hopkins R, Churchill S, Deru K. Neurological outcomes 6 years after acute carbon monoxide poisoning. Undersea Hyperb Med. 2008;35(4):241-323.
17. Haas B, Stukel TA, Gomez D, et al. The mortality benefit of direct trauma center transport in a regional trauma system: A population-based analysis. J Trauma Acute Care Surg. 2012;72(6):1510-1517.
18. Prabhakaran S, O'Neill K, Stein-Spencer L, Walter J, Alberts MJ. Prehospital triage to primary stroke centers and rate of stroke thrombolysis. JAMA Neurol. 2013;70(9):1126-1132.
19. Savage ML, Hay K, Murdoch DJ, et al. Clinical outcomes in pre-hospital activation and direct cardiac catheterisation laboratory transfer of STEMI for primary PCI. Heart Lung Circ. 2022;31(7):974-984.
20. Kang DY, Cho KJ, Kwon O, et al. Artificial intelligence algorithm to predict the need for critical care in prehospital emergency medical services. Scand J Trauma Resusc Emerg Med. 2020;28:1-8.
21. Blomberg SN, Folke F, Ersbøll AK, et al. Machine learning as a supportive tool to recognize cardiac arrest in emergency calls. Resuscitation. 2019;138:322-329.
22. Kijanka R, Kawecki M, Białoń P, Szlagor M, Dudek M, Bobiński Three-step analysis of symptoms of carbon monoxide poisoning in pre-hospital medical response teams. Emerg Med Serv. 2022;9(1):48-54.
23. Hampson NB, Piantadosi CA, Thom SR, Weaver LK. Practice recommendations in the diagnosis, management, and prevention of carbon monoxide poisoning. Am J Respir Crit Care Med. 2012;186(11):1095-1101.
24. Uehara O, Morikawa T, Harada F, et al. Performance of ChatGPT-3.5 and ChatGPT-4o in the Japanese National Dental Examination. J Dent Educ. 2025;89(4):459-466.
25. OpenAI. Introducing ChatGPT. Accessed date: 31 May 2025. Available from: https://openai.com/index/hello-gpt-4o/.
26. OpenAI. Introducing OpenAI o3 and o4 mini. Accessed date: 31 May 2025. Available from: https://openai.com/ index/introducing-o3-and-o4-mini/.
27. OpenAI. Introducing GPT-4.5. Accessed date: 31 May 2025. Available from: https://openai.com/index/ introducing-gpt-4-5/.
28. Balel Y, Zogo A, Yıldız S, Tanyeri H. Can ChatGPT-4o provide new systematic review ideas to oral and maxillofacial surgeons? J Stomatol Oral Maxillofac Surg. 2024:101979.
29. Is EE, Menekseoglu AK. Comparative performance of artificial intelligence models in rheumatology board-level questions: Evaluating Google Gemini and ChatGPT-4o. Clin Rheumatol. 2024;43(11):3507-3513.
30. Caspi R, Karp PD. An evaluation of ChatGPT and Bard (Gemini) in the context of biological knowledge retrieval. Access Microbiol. 2024;6(6):000790. v3.
31. Team G, Anil R, Borgeaud S, et al. Gemini: A family of highly capable multimodal models. arXiv preprint arXiv:231211805. 2023.
32. AI G. Model Information 2.5 Pro Model Card. Accessed date: 31 May 2025. Available from: https://deepmind.google/t echnologies/gemini/pro/.
33. DeepSeek I. DeepSeek-R1 Release. Accessed date: 31 May 2025. Available from: https://api-docs.deepseek.com/news/ news250120.