Canadian Journal of Emergency Medicine
https://doi.org/10.1007/s43678-026-01214-2

ORIGINAL RESEARCH

Skyer: a novel benchmark for evaluating the effectiveness of large language models in emergency department triage

Shadi Nourbakhsh[1]

Received: 22 October 2025 / Accepted: 28 June 2026

Abstract

Objectives Emergency department (ED) overcrowding causes diagnostic challenges, prolonged wait times, and impairs appropriate triage, often due to human error and fatigue. Large language models can assist ED staff in triage, improving patient care by mitigating these problems.

Methods We designed an evaluation method (Skyer benchmark) to assess fifteen large language models, including DeepSeek-R1 (70B, 7B), ChatGPT versions (4, 4.5-preview), Gemini iterations (1.5-pro, 2.0-Pro-experimental, 2.5_03-25, 2.5_05-06), Mistral-7B, Llama-3.3, Gemma models (2-27b-it, 3-12b-it, 3-27b-it), Qwen-2.5, and Phi-4-14B. We assessed the performance of models in ED triage by using 55 realistic clinical pediatric scenarios. The main objective was to evaluate models' triage performance using a weighting system that accounted for the impacts of over-triage and under-triage, in addition to simple accuracy. A secondary objective was to assess models' consistency, by repeating tests across scenarios three times.

Results Our findings indicate that only two models, ChatGPT-4.5-preview and Gemini-2.5_05-06, demonstrated superior and reliable triage performance. ChatGPT-4.5-preview (77% accuracy, mean weight 377.5 out of 550) and Gemini-2.5_05-06 (74% accuracy, mean weight 365/550) significantly outperformed human-triage-experts accuracy (64% accuracy, mean weight 253.5/550) and other models. This difference was statistically significant (p-value < 0.05), with an extremely large effect size (Cohen's $D = 2.18$ and 1.98). Furthermore, they demonstrated sufficient reliability due to acceptable consistency in their triage performance (85% and 82%).

Conclusion Our work establishes Skyer as an evaluation approach of large language models. Skyer selected the best-performing models, which demonstrated significantly higher triage performance than the human experts and showed consistent results. These large language models can streamline the delivery of quality healthcare in overcrowded EDs. Despite promising outcomes, we identified limitations prohibiting these large language models as replacement of human experts. Instead, we demonstrate their potential for a substantial role in assisting the staff in overcrowded EDs.

Keywords Artificial intelligence (AI) · Large language models · Pediatrics emergency department (ED) · Triage · Benchmark

Résumé

Objectifs Le surpeuplement des services d'urgence (SU) entraîne des difficultés diagnostiques, des temps d'attente prolongés et un triage inapproprié, souvent en raison d'erreurs humaines et de la fatigue. Les grands modèles de langage peuvent aider le personnel du service d'urgence dans le triage, améliorant ainsi les soins aux patients en atténuant ces problèmes.

Méthodes Nous avons conçu une méthode d'évaluation (benchmark de Skyer) pour évaluer quinze grands modèles de langages, y compris DeepSeek-R1 (70B, 7B), les versions ChatGPT (4, 4.5-preview), les itérations Gemini (1.5-pro, 2.0-Pro-experimental, 2.5_03-25, 2.5_05-06), Mistral-7B, Llama-3.3, les modèles Gemma (2-27b-it, 3-12b-it, 3-27b-it), Qwen-2,5 et Phi-4-14B. Nous avons évalué la performance des modèles dans le triage ED en utilisant 55 scénarios pédiatriques cliniques réalistes. L'objectif principal était d'évaluer la performance de triage des modèles à l'aide d'un système de pondération qui

✉ Shadi Nourbakhsh
shadi@skyeinsights.net

[1] Skye Insights LTD, Toronto, ON, Canada

Published online: 11 July 2026

tenait compte des impacts du triage excessif et du triage insuffisant, en plus de la simple précision. Un objectif secondaire était d'évaluer la cohérence des modèles, en répétant trois fois les tests entre les scénarios.

Résultats Nos résultats indiquent que seuls deux modèles, ChatGPT-4.5-preview et Gemini-2.5_05-06, ont démontré des performances de triage supérieures et fiables. ChatGPT-4.5-preview (précision de 77%, poids moyen 377,5 sur 550) et Gemini-2.5_05-06 (précision de 74%, poids moyen 365/550) ont significativement surpassé la précision des experts en triage humain (précision de 64%, poids moyen 253,5/550) ainsi que d'autres modèles. Cette différence était statistiquement significative (valeur $p > 0{,}05$), avec une taille d'effet extrêmement grande (D de Cohen = 2,18 et 1,98). De plus, ils ont démontré une fiabilité suffisante en raison d'une cohérence acceptable dans leur performance de triage (85% et 82%).

Conclusion Notre travail établit Skyer comme une approche d'évaluation des grands modèles linguistiques. Skyer a sélectionné les modèles les plus performants, avec une performance de triage significativement supérieure à celle des experts humains. Ces grands modèles de langage peuvent rationaliser la prestation de soins de santé de qualité dans les services d'urgence surpeuplés. Malgré des résultats prometteurs, nous avons identifié des limites interdisant ces grands modèles de langage en remplacement des experts humains. Au lieu de cela, nous démontrons leur potentiel pour un rôle substantiel dans l'assistance au personnel dans les services d'urgence surpeuplés.

Mots-clés Intelligence artificielle (IA) · Grands modèles de langage · Service des urgences pédiatriques (SU) · Triage · Benchmark

Clinician's capsule

What is known about the topic?
Overcrowded EDs, particularly pediatric EDs, present significant challenges to healthcare systems, resulting in prolonged waiting times, misdiagnoses and improper triage

What did this study ask?
We designed the Skyer benchmark to assess large language models performance as a reliable triage assistant in real-life ED situations

What did this study find?
This study reveals that two of fifteen evaluated models can assist triage-experts performance, reduce errors and improve ED flow

Why does this study matter to clinicians?
Skyer can evaluate current and upcoming models to determine the most effective assistance for ED performance leading to improved patient care

Introduction

Overcrowded emergency departments (EDs) are plagued by prolonged waiting times, staff overfatigue, and an elevated risk of misdiagnosis. Previous research suggests that advanced large language models might help EDs increase diagnosis accuracy [1], prioritize triage, promptly identify critical conditions, and optimize resource allocation [2]. Implementing such models could significantly alleviate these issues and improve the condition of both patients and medical staff, accelerating reliable and efficient care [1]. Nevertheless, previous research relied on virtual or textbook scenarios rather than real-life situations and was mostly focused on adults.

To assess effectiveness of large language models, benchmarking is employed. Benchmarking is the process of subjecting various methods to a set of standard questions to gauge performance and accuracy in a unified way. In our context, it serves as a standardized test for AI, with the results revealing a model's ability to handle diverse information [3]. Benchmarking models against human performance and against one another on clinically relevant tasks enables tracking of their capabilities and progress [4]. Existing AI triage model benchmarks may not be validated for pediatrics due to unique challenges. Pediatric medicine challenges include complex comorbidities, rising emergency admissions, limited access to specialists, differences in disease prevalence, and age-dependent variations compared with adult medicine. These factors threaten the delivery of high quality and timely care. Data privacy for sensitive pediatric medical information is also a key concern [5]. While large language models show significant potential in medical fields, their proficiency in pediatrics requires validation. Improving accuracy in this specialized area necessitates integrating pediatric data into pretraining datasets [6]. Furthermore, patient safety concerns for young individuals demand exceptional accuracy and reliability in any AI-driven triage system. Models trained primarily on adult data or lacking diverse pediatric representation may not perform equitably across all children [7].

Our study evaluated the accuracy and effectiveness of various large language models in triage scoring in realistic pediatric cases. Meanwhile, a significant challenge of using models in EDs is their inconsistent performance. Consistent, trustworthy models should yield similar answers to the same

Table 1 The offline and online large language models evaluated in the study

	Name	Size	Release Date	Creator
Online large language models	ChatGPT-4		March 2023	Open AI
	ChatGPT-4.5-preview		February 2025	Open AI
	Gemini-1.5-pro		May 2024	Google
	Gemini-2.0-Pro experimental		February 2025	Google
	Gemini-2.5_03-25		May 2025	Google
	Gemini-2.5_05-06		June 2025	Google
Offline large language models	Deepseek-R1	7B	January 2025	Deepseek AI developmental
	Deepseek-R1	70B	January 2025	Deepseek AI developmental
	Qwen -2.5	7.6b	September 2024	Alibaba
	Phi-4	14.7b	January 2025	Microsoft
	Llama-3.3	70.6b	December 2024	Meta
	Mistral	7.2b	July 2024	Mistral
	Gemma-2	27b	July 2024	Google
	Gemma-3	12B	March 2025	Google
	Gemma-3	27B	March 2025	Google

prompt across different users and times. A model is typically deemed consistent if 80% of responses to repeated questions meet this standard [8]. Our evaluation focused on comparing the accuracy of weighted triage scores and the consistency of the models. We hypothesized that large language models achieve a consistently superior triage accuracy compared to human experts. It demonstrated their potential to alleviate ED overcrowding.

Methods

Study design and population

To evaluate the performance of the large language models, a high quality set of targeted test scenarios is necessary [9]. Using real data presented several risks, including ethical and legal concerns, data noise due to parent prompts, and bias due to overlap between cases. Public datasets are also not usable, as they could be present in model training sets. Therefore, 55 realistic, synthetic pediatric clinical scenarios (common and severe conditions) were created by an experienced pediatrician. Prompts were solely based on parent descriptions. For a comparative analysis of triage performance, the models were compared against each other, as well as against the assessments of three experienced triage nurses and a physician, all utilizing identical scenario information.

Data collection

Each case was presented individually to the models and human experts, with details mirroring parents' complaints arriving at EDs. They were tasked with assigning a triage score (1—life-threatening, 2—emergent, 3—urgent, 4—less-urgent, 5—non-urgent) using both the American Emergency Severity Index (ESI) [10] and the Canadian Triage and Acuity Scale (CTAS) [11]. This method was used to benchmark fifteen recently developed, highly ranked, efficient large language models listed in Table 1.

Data analysis

Our analysis focused on triage accuracy and the impact of incorrect triage scores to identify the top-performing model for ED use. Consistent with prior pediatric triage research [12, 13], triage levels 4 and 5 were merged into a single low-acuity group for the analysis, given that both categories represent stable patients requiring minimal anticipated resource use. Additionally, if a model generates different scores for a case, the higher-priority triage level is selected (e.g., if a case is triaged as 3 or 4, level 3 is assigned).

Inaccurate triage impacts patients' conditions in distinct ways. While assigning a score of 3 instead of 2 to a critical patient might cause non-life-threatening treatment delays, incorrectly assigning a score of 4 significantly risks the patient's life. Therefore, distinguishing between inaccurate scores is crucial. Although over-triage can lead to unnecessary healthcare costs and staff fatigue, it is generally preferred over under-triage for patient well-being [14, 15]. Under-triage can result in disastrous patient outcomes due to increased waiting times and inadequate care.

Conversely, some models might uniformly assign a high triage score (e.g., 2) to all patients, neglecting individual assessment. While this approach minimizes undercare, it can cause staff overfatigue and unacceptably longer waiting

times. Therefore, a system is required to reward accuracy and apply penalties proportional to the degree of impact.

Therefore, in Skyer, we implemented a triage weighting system. This system rewarded accurate triage scores while penalizing both over-triage and under-triage through varying weights. To develop this weighting system, various positive and negative variables were identified through textbook research and prior studies [16, 17]. The models were weighted based on these specific variables:

- Patient outcome/prognosis [−5, +5]
- Patient safety [−3, +3]
- Patient satisfaction/confidence [−2, +2]
- Resource consumption/healthcare utilization [−2, 0]
- Staff fatigue [−2, 0]
- Impact on operational ED flow [−2, 0]

Accurate triage increases expected patient satisfaction (weighted + 2). Under-triage resulting only in delayed treatment is weighted − 1. However, under-triage causing increased anxiety, cost, and reduced ED trust has a worse impact on patient satisfaction, weighted − 2. Expected staff overfatigue from unnecessary care is more detrimental than increased workload from admission instead of observation. These weights facilitated a comparison that not only assessed the correct task but also accounted for the impact of both over-scoring and under-scoring in triage. The model's final triage performance score was determined by the sum of its weighted values across all 55 cases.

Statistical analysis

In Skyer, large language models' triage accuracy was assessed and then re-evaluated using the designed weighting system. Welch's one-sided *t*-test compared models to each other and to human experts. A significance level of 0.05 was applied, and the null hypothesis was rejected when $p < 0.05$, to support the models' superiority. CIs were reported where applicable (Table 3). Focusing solely on triage score accuracy and *p*-value overlooks the critical impacts of under-triage and over-triage. It fails to represent the real ED medical complexities. For instance, one model accurately triaged 60% of cases but severely under-triaged 33% of the remainder. While the accuracy and *p*-value might suggest a good model performance, considering the high under-triage rate indicates the model's suboptimal performance. Therefore, the Skyer weighing system calculated the mean CTAS and ESI weights for each model. These were then compared with the average triage score weights of human experts. We also calculated the effect size (Cohen's *D*), which is valid regardless of the distribution's shape. To further confirm statistical significance, we also performed the Mann–Whitney *U* test. All the above methods yielded the same results. Finally, to evaluate consistency, top-performing models were tested three times on identical cases.

Results

This study evaluated the triage accuracy of fifteen large language models on 55 realistic cases. The models' performance was compared with one another and with human experts (one physician, three triage nurses). Various studies have reported moderate levels of adult triage accuracy among human experts: 66.2% [17], 63% [18], 56.2% [19], 59.3% [20], 62.5% [21], and 46.9% for pediatric [17]. The human-experts' triage accuracy in our study ranged from 56 to 69% (mean: 64%), consistent with previous studies and suitable as a baseline.

Table 2 illustrates the rates of correct triage assignments, over-triage (assigning a higher priority), and under-triage (assigning a lower priority) using ESI and CTAS systems.

The six top-performing models demonstrated a triage accuracy range of 64–80%, over-triage 7–29%, and under-triage 2–24%. This contrasts with triage-experts, who demonstrated a triage accuracy range of 56–69%, over-triage 0–18%, and under-triage 24–44% (Table 2). ChatGPT-4.5-preview demonstrated the highest triage accuracy in both systems, with a mean triage accuracy of 77.5%. Gemini-2.5_05-06 (74%), Deepseek-R1-70b, and Gemini-1.5-pro (72%) also exhibited strong performance. Gemini-2.0-pro-experimental and Gemma-3-27b-it performed well, both at 66%. These models generally demonstrated comparable or superior mean triage accuracy to that of human experts (Fig. 1).

Table 3 clearly illustrates that ChatGPT-4.5-preview and Gemini-2.5_05-06 statistically significantly outperformed human experts in triage performance (p-value < 0.05, CI 24.9, 223.9/12.8, 210.2). While the performance of Deepseek-R1-70b, Gemini-1.5-pro, Gemini-2.0-pro-experimental, and Gemma-3-27b-it practically exceeded that of human experts (large or medium positive effect sizes), these differences were not statistically significant. Other models, however, did not outperform human experts in triage scoring. Inter-rater reliability, among four human raters, showed good reliability (ICC 0.75 to 0.90). Reliability remained good for both CTAS (ICC = 0.86) and ESI (ICC = 0.82), indicating consistent scoring among human experts.

Finally, we focused on evaluating the consistency of our top-performing models, a critical issue in healthcare. We conducted three evaluations using the same prompt. Due to unavailability for re-evaluation, Gemini-2.0-pro-experimental was skipped. Only ChatGPT-4.5-preview, Gemini-1.5-pro, and Gemini-2.5_05-06 demonstrated an acceptable triage consistency of over 80% (Table 4).

Table 2 Triage scores of large language models in American system (ESI) and Canadian system (CTAS) ($N=55$)

Triage source	Correct triage *N* (%)	Over-triage *N* (%)	Under-triage *N* (%)
ChatGPT-4.5-preview			
ESI	44 (80)	4 (7)	7 (13)
CTAS	41 (75)	4 (7)	10 (18)
Mean	42.5 (77.5)	4 (7)	8.5 (15.5)
Gemini-2.5_05-06			
ESI	41 (75)	13 (24)	1 (2)
CTAS	40 (73)	12 (21)	3 (6)
Mean	40.5 (74)	12.5 (22)	2 (4)
DeepSeek-R1-70B			
ESI	40(73)	8 (14)	7 (13)
CTAS	39 (71)	8 (14)	8 (14)
Mean	39.5 (72)	8 (14)	7.5 (14)
Gemini-1.5-pro			
ESI	40 (73)	9 (16)	6 (11)
CTAS	38 (69)	9 (16)	8 (14)
Mean	39 (72)	9 (16)	7 (13)
Nurse #3			
CTAS	38 (69)	4 (7)	13 (24)
Nurse #1			
ESI	38 (69)	6 (11)	11 (20)
Gemini-2.0-pro-experimental			
ESI	37 (67)	16 (29)	2 (4)
CTAS	36 (66)	15 (27)	4 (7)
Mean	36.5 (66)	15.5 (28)	3 (6)
Gemma-3-27b-it			
ESI	37 (67)	15 (27)	3 (6)
CTAS	35 (64)	15 (27)	5 (9)
Mean	36 (66)	15 (27)	4 (7)
Mean human expert			
Mean	35.5 (64)	5 (9)	14.5 (27)
Nurse #2			
CTAS	34 (62)	10 (18)	11 (20)
Gemma-3-12b-it			
ESI	34 (62)	20 (36)	1 (2)
CTAS	33 (60)	19 (34)	3 (6)
Mean	33.5 (61)	19.5 (35)	2 (4)
ChatGPT-4			
ESI	34 (62)	4 (7)	17 (31)
CTAS	33 (60)	4 (7)	18 (33)
Mean	33.5 (61)	4 (7)	17.5 (32)
Gemini-2.5_03-25			
ESI	34 (62)	20 (36)	1 (2)
CTAS	32 (58)	20 (36)	3 (6)
Mean	33 (60)	20 (36)	2 (4)
Llama-3.3			
ESI	33 (60)	15 (27)	7 (13)
CTAS	31 (56)	15 (27)	9 (16)
Mean	32 (58)	15 (27)	8 (15)
Gemma-2-27b-it			
ESI	32 (58)	20 (36)	3 (6)
CTAS	30 (55)	20 (36)	5 (9)
Mean	31 (56)	20 (36)	4 (7)
Physician			
ESI	31 (56)	0 (0)	24 (44)
Qwen-2.5			
ESI	30 (55)	22 (40)	3 (6)
CTAS	28 (51)	22 (40)	5 (9)
Mean	29 (53)	22 (40)	4 (7)
Phi-4-14B			
ESI	28 (51)	24 (44)	3 (6)
CTAS	26 (47)	24 (44)	5 (9)
Mean	27 (49)	24 (44)	4 (7)
Mistral-7B			
ESI	26 (47)	26 (47)	3 (6)
CTAS	26 (47)	25 (46)	4 (7)
Mean	26 (47)	25.5 (46.5)	3.5 (6.5)
DeepSeek-R1-7B			
ESI	23 (42)	25 (45)	7 (13)
CTAS	21 (38)	25 (45)	9 (16)
Mean	22 (40)	25 (45)	8 (15)

Discussion

Interpretation of findings

Our Skyer benchmark assessed the reliability of cutting-edge large language models for ED triage and their potential role as ED staff assistants, comparing them against one another and human experts. Overall triage score performance weighted over- and under-triage errors differently due to their significant impact on patient outcomes. Our findings indicate that among various models evaluated for triage using "parent-type prompts," ChatGPT-4.5-preview and Gemini-2.5_05-06 demonstrated statistically significant overperformance. Other over-performing models, including Deepseek-R1-70b, Gemini-1.5-pro, and Gemini-2.0-pro-experimental and at lower rates, Gemma-3-27b-it, showed notable practical improvements over human experts in triage tasks, specifically in reducing both over-triage and under-triage errors. However, these differences were not statistically significant, (Table 3). To ensure consistent and reliable responses in ED settings, our leading models were evaluated twice more, except Gemini-2.0-pro-experimental which was unavailable. The results indicate that ChatGPT-4.5-preview and Gemini-2.5_05-06 demonstrated acceptable consistency

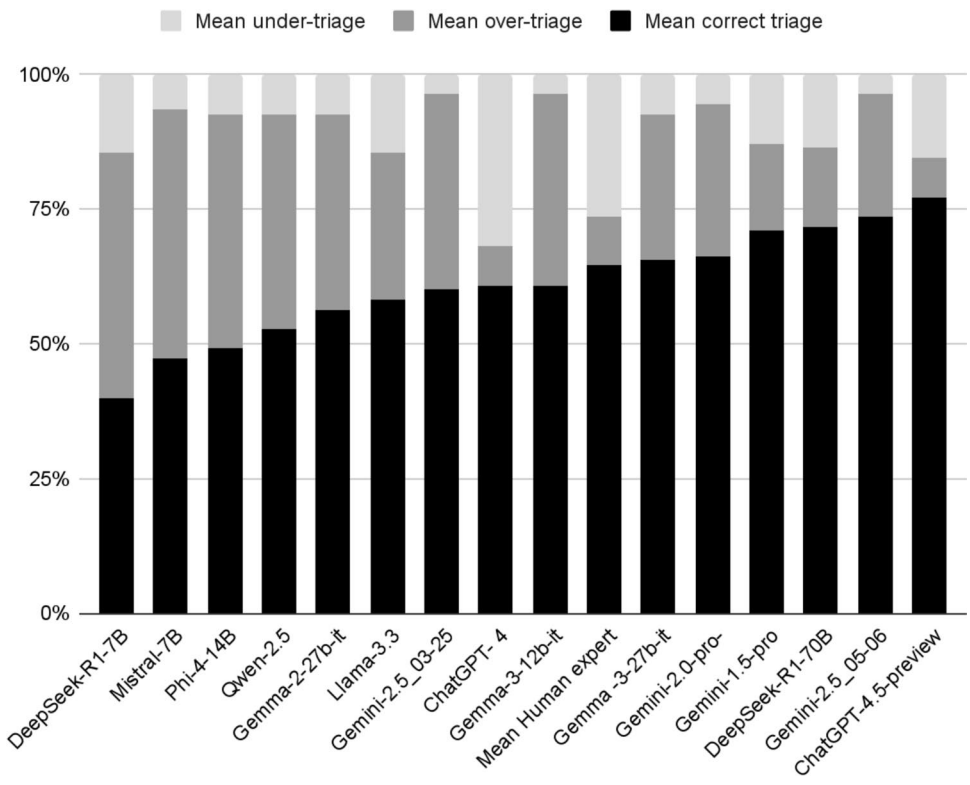

Fig. 1 Mean of ESI and CTAS triage scores of triage sources

across triage tasks, making them the most reliable models. While Gemini-1.5-pro achieved a high weight score and consistency, it lacked statistical significance. Similarly, Gemma-3-27b-it and Deepseek-R1-70b showed promising triage scoring and weighing in the initial assessment, but their inconsistent performance and lack of statistical significance rendered them unreliable for ED triage use (Table 4).

Comparison to previous studies

Previous studies on triage accuracy have yielded varied and even contradictory results. Some research indicates that ChatGPT triage was unreliable for complex patients [22, 23] or ChatGPT-v4 had low accuracy and improved with prompt engineering [24]. Ethical concerns regarding GPT-4o's unreliability were reported [25]. Additionally, ChatGPT and Copilot were found to be inaccurate, though better at identifying high-acuity patients [21]. Conversely, other studies present more optimistic findings. Sorich et al. reported high adult triage performance for GTP-4o, Claude-3.5-Sonnet, and Gemini-1.5-Pro, with urgency overestimation as the main error [4]. Trained GPT-4.0 and Claude-3 Opus showed improved predictive performance and reliability in high triage levels [26]. AI-enabled triage tools achieve substantial agreement with GP urgency assessments [27]. Halwani et al. reported that SVM, RF, and LGBM demonstrated high performance with consistent precision and recall across all CTAS triage levels [28]. Anoka study showed that the KATE triage model improved accurate risk stratification compared to Gemini, Copilot, ChatGPT, and Pi AI [29].

Evaluating the varied results in large language models' triage scoring necessitates a unique triage benchmarking approach. Previous benchmarks focused on Q&A tasks (ClinicBench, Liu et al. [30]) or model-user conversations (Healthbench, Arora et al. [31]). The closest benchmark to Skyer was Kirch et al.'s TRIAGE, which evaluated models' capacity for ethical decision-making during mass casualty incidents [32].

Strengths and limitations

The primary strength of Skyer is its use of pediatric cases that mirror real-life scenarios, with prompts reflecting parents' typical descriptions of their children's symptoms. Its accuracy is evaluated in the most challenging environment, EDs, focusing on complex patient groups: pediatrics. Meanwhile, Skyer employs a weighing system to evaluate the varying impacts of models over- and under-triage on patient outcomes and their consistency.

The primary limitation is that relying solely on large language models for triage raises ethical concerns. Patients seeking prioritized care might exaggerate symptoms to

Table 3 Statistically comparing triage weighing between large language models with triage-experts

Model	Model scores*	Model mean score	Difference	95% CI for difference	Cohen's *D* (effect size)	*p*-value	Comment
ChatGPT-4.5-preview	ESI: 393, CTAS:362	377.5	+124	24.1, 223.9	2.18 (extreme large positive effect size)	0.03	Statistically significantly outperformed human experts, with a very large effect size (CI excludes 0)
Gemini-2.5_05-06	ESI: 376, CTAS:354	365.0	+111.5	12.8, 210.2	1.98 (extreme large positive effect size)	0.04	Statistically significantly outperformed human experts, with a very large effect size (CI excludes 0)
Deepseek-R1-70B	ESI: 338, CTAS:318	328.0	+74.5	− 24.4, 173.4	1.33 (large positive effect size)	0.10	Higher, stronger overperformance compared to human experts with large effect size, practically but not statistically significant (CI includes 0)
Gemini-1.5-pro	ESI: 348, CTAS:308	328.0	+74.5	− 30.9, 179.9	1.30 (large positive effect size)	0.12	Higher, stronger overperformance compared to human experts with large effect size, practically but not statistically significant (CI includes 0)
Gemini-2.0-pro-experimental	ESI: 325,CTAS:303	314.0	+60.5	− 38.2, 159.2	1.08 (large positive effect size)	0.16	Higher, stronger overperformance compared to human experts with large effect size, practically but not statistically significant (CI includes 0)
Gemma-3-27b-it	ESI: 314, CTAS:283	298.5	+45.0	− 54.9, 144.9	0.79 (medium positive effect size)	0.28	Moderate overperformance compared to human experts with medium effect size, practically but not statistically significant (CI overlapped zero)
Gemma-3-12b-it	ESI: 286, CTAS:264	275.0	+21.5	− 77.2, 120.2	0.38 (small–medium positive effect size)	0.57	No meaningful overperformance, as the scores were nearly identical and the effect sizes were small to medium
Gemini-2.5_03-25	ESI: 291, CTAS:257	274.0	+20.5	− 80.7, 121.7	0.36 (small–medium positive effect size)	0.60	No meaningful overperformance, as the scores were nearly identical and the effect sizes were small to medium
Gemma-2-27b-it	ESI: 256, CTAS:225	240.5	− 13.0	− 112.9, 86.9	− 0.23 (small negative effect size)	0.74	Performance slightly lower than human experts
LLama-3.3	ESI: 255, CTAS:224	239.5	− 14.0	− 113.9, 85.9	− 0.25 (small negative effect size)	0.72	Performance lower than human experts
ChatGPT-4	ESI: 227, CTAS:207	217.0	− 36.5	− 135.4, 62.4	− 0.65 (medium negative effect size)	0.35	Performance lower than human experts
Qwen-2.5	ESI: 221, CTAS:190	205.5	− 0.48	− 147.9, 51.9	− 0.85 (medium–large negative effect size)	0.25	Performance lower than human experts
Phi-4-14b	ESI: 208, CTAS:177	192.5	− 0.61	− 160.9, 38.9	− 1.07 (large negative effect size)	0.164	Performance lower than human experts
Mistral-7b	ESI: 173, CTAS:169	171.0	− 82.5	− 184.6, 19.6	− 1.48 (large negative effect size)	0.083	Performance considerably lower than human experts
DeepSeek-R1-7b	ESI: 124, CTAS:93	108.5	− 145	− 244.9, − 45.1	− 2.55 (extreme large negative effect size)	0.016	Statistically significantly underperformed human experts

Human-expert triage weights: 161, 259, 292, 302, Human experts weights mean: 253.5

Difference mean difference, *DF* degree of freedom, *CI* confidence interval

*The final score for each model was calculated as the sum of its weighted values across all 55 cases, using both triage systems, the first determined using the ESI system, and the second using the CTAS system

Table 4 Consistency of top large language models performances in triage

Large language model	Triage consistency (%)
ChatGPT-4.5-preview	85
Gemini-1.5-pro	82
Gemini-2.5_05-06	80
DeepSeek-R1-70B	60
Gemma-3-27b-it	60

models, without human oversight to verify their claims. This is framed as an ethical challenge rather than a system failure. Another significant concern is overreliance on AI systems for diagnosis and triage, which could lead patients to forgo professional medical care. Our research highlights the necessity of precise descriptions for accurate model diagnoses. We emphasize that non-expert patient interactions with AI may be unreliable, especially given the risk of model hallucinations.

Clinical implications

Overcrowded EDs strain the traditional triage system, leading to human error, fatigue, and long waiting times, which can compromise patient health and care quality. Large language models can potentially assist the ED system. Previous similar studies often relied on textbook or exam questions, which may overlap with the model training datasets and do not accurately reflect real-world scenarios. Meanwhile, their prompts are crucial; for instance, it is unclear whether Sorich et al. models' good performance was responsible for the positive results or their prompt types [4]. Since the prompts cannot be revealed, a unified benchmark is crucial for evaluating the results. Skyer can play an important role in evaluating large language models to identify the best one as an ED assistant. Since relying solely on them for triage raises significant ethical concerns, it can be addressed in two ways:

1. Large language models as an assistant: They can assist triage nurses by rapidly providing triage scores and allow human review if there are any discrepancies.
2. Tiered triage system: Patients with triage scores of 1 and 2 will be sent directly to physicians. Patients with scores 3, 4, or 5 will undergo a second nurse triage to ensure accuracy. This approach minimizes waiting times for critical patients and helps identify and re-triage those attempting to exploit the system.

Research implications

Skyer benchmarked recent large language models for their ED triage performance. This establishes the foundation for evaluating models to identify better-performing models as ED assistants for diagnosis and decision-making as well as triage. We are developing an AI-based patient agent that interacts with models, rates them based on expected outcomes, and assists in developing new models.

Conclusion

Our Skyer benchmark used synthetic realistic pediatric ED cases, which presented a greater challenge compared to previous studies. The triage performance was evaluated based on triage accuracy, weights, and consistency. It showed that ChatGPT-4.5-preview and Gemini-2.5-05_06 performed strongly in triage, offering reliable assistance to existing ED triaging systems. We propose that these high-performing models serve as assistive tools in EDs, operating under the careful oversight of human experts. They could offer several benefits, including reduced patient wait times, decreased staff burnout, fewer misdiagnoses, and improved patient outcomes.

Funding The author has no relevant financial or non-financial interests to disclose.

Declarations

Conflict of interest On behalf of all authors, the corresponding author states that there is no conflict of interest.

References

1. Gvajaia N, Kutchava L, Alavidze L, Yandamuri SP, Pestvenidze E, Kupradze V. Artificial intelligence in the medical field: diagnostic capabilities of GPT-4 in comparison with physicians. JMAI. 2024. https://doi.org/10.21037/jmai-24-276.

2. Kehinde O. A leveraging machine learning for predictive models in healthcare to enhance patient outcome management. Int Res J Mod Eng Technol Sci. 2025. https://doi.org/10.56726/IRJMETS66198.
3. Chiang W-L, Zheng L, Sheng Y, Angelopoulos AN, Li T, Li D, et al. An open platform for evaluating LLMs by human preference. 2024. https://doi.org/10.48550/arXiv.2403.04132
4. Sorich MJ, Mangoni AA, Bacchi S, Menz BD, Hopkins AM. The triage and diagnostic accuracy of frontier large language models: updated comparison to physician performance. J Med Internet Res. 2024;26:e67409. https://doi.org/10.2196/67409.
5. Smith ME, Zalesky C, Lee S, Gottlieb M, Adhikari S, Goebel M, et al. Artificial intelligence in emergency medicine: a primer for the nonexpert. J Am Coll Emerg Physicians Open. 2025;6(2):100051. https://doi.org/10.1016/j.acepjo.2025.100051.
6. Mondillo G, Perrotta A, Frattolillo V, Colosimo S, Indolfi C, del Giudice MM, et al. Large language models performance on pediatrics question: a new challenge. J Med Artif Intell. 2025. https://doi.org/10.21037/jmai-24-174.
7. Alsabri M, Aderinto N, Mourid MR, Laique F, Zhang S, Shaban NS, et al. Artificial intelligence for pediatric emergency medicine. J Med Surg Public Health. 2024;3:100137. https://doi.org/10.1016/j.glmedi.2024.100137.
8. Patwardhan A, Vaidya V, Kundu A. Automated consistency analysis of LLMs. arXiv:2502.07036v1 [cs.CR]. 2025. https://doi.org/10.48550/arXiv.2502.07036
9. Liang P, Bommasani R, Lee T, Tsipras D, Soylu D, Yasunaga M, Zhang Y, Narayanan D, Wu Y, Kumar A, Newman B. Holistic evaluation of language models. arXiv Preprint. arXiv:2211.09110. 2022.
10. Emergency Severity Index Handbook. Fifth edition. 2023. https://media.emscimprovement.center/documents/Emergency_Severity_Index_Handbook.pdf.
11. The Canadian Triage and Acuity Scale Combined Adult/ Pediatric Educational Program, Participant's, Manual version 2.5b. 2013. https://ctas-phctas.ca/wp-content/uploads/2018/05/participant_manual_v2.5b_november_2013_0.pdf.
12. Feral-Pierssens A-L, Gaboury I, Carbonnier C, Breton M. Redirection of low-acuity emergency department patients to nearby medical clinics using an electronic medical support system: effects on emergency department performance indicators. BMC Emerg Med. 2024;24:166. https://doi.org/10.1186/s12873-024-01080-0.
13. Berkowitz D, Morrison S, Shaukat H, Button K, Stevenson M, LaViolette D, et al. Under-triage: a new trigger to drive quality improvement in the emergency department. Pediatr Qual Saf. 2022;7(4):e581. https://doi.org/10.1097/pq9.0000000000000581.
14. Hewes HA, Christensen M, Taillac PP, et al. Consequences of pediatric undertriage and overtriage in a statewide trauma system. J Trauma Acute Care Surg. 2017;83:662–7. https://doi.org/10.1097/TA.0000000000001560.
15. Loftus TJ, Ruppert MM, Shickel B, Ozrazgat-Baslanti T, Balch JA, Hu D, et al. Overtriage, undertriage and value of care after major surgery: an automated, explainable deep learning-enabled classification system. J Am Coll Surg. 2023;236(2):279–91. https://doi.org/10.1097/XCS.0000000000000471.
16. Huabbangyang T, Rojsaengroeng R, Tiyawat G, Silakoon A, Vanichkulbodee A, Sri-on J, et al. Associated factors of under and over-triage based on the emergency severity index; a retrospective cross-sectional study. Arch Acad Emerg Med. 2023;11(1):e57. https://doi.org/10.22037/aaem.v11i1.2076.
17. Mistry B, De Ramirez SS, Kelen G, Schmitz PSK, Balhara KS, Levin S, et al. Accuracy and reliability of emergency department triage using the emergency severity index: an international multicenter assessment. Ann Emerg Med. 2018;71(5):581-587.e3. https://doi.org/10.1016/j.annemergmed.2017.09.036.
18. Tam HL, Chung SF, Lou CK. A review of triage accuracy and future direction. BMC Emerg Med. 2018;18:58. https://doi.org/10.1186/s12873-018-0215-0.
19. Chen SS, Chen JC, Ng CJ, Chen PL, Lee PH, Chang WY. Factors that influence the accuracy of triage nurses' judgement in emergency departments. Emerg Med J. 2010;27(6):451–5. https://doi.org/10.1136/emj.2008.059311.
20. Cetin SB, Eray O, Cebeci F, Coskun M, Gozkaya M. Factors affecting the accuracy of nurse triage in tertiary care emergency departments. Turk J Emerg Med. 2020;20(4):163–7. https://doi.org/10.4103/2452-2473.297462.
21. Arslan B, Nuhoglu C, Satici MO, Altinbilek E. Evaluating LLM-based generative AI tools in emergency triage: a comparative study of ChatGPT Plus, Copilot Pro, and triage nurses. Am J Emerg Med. 2025;89:174–81. https://doi.org/10.1016/j.ajem.2024.12.024.
22. Zaboli A, Brigo F, Brigiari G, Massar M, Parodi M, Pfeifer N, et al. Chat-GPT in triage: still far from surpassing human expertise—an observational study. Am J Emerg Med. 2025;92:165–71. https://doi.org/10.1016/j.ajem.2025.03.028.
23. Pasli S, Yadigaroğlu M, Kirimli EN, Beşer MF, Unutmaz İ, Ayhan AÖ, et al. ChatGPT-supported patient triage with voice commands in the emergency department: a prospective multicenter study. Am J Emerg Med. 2025;94:63–70. https://doi.org/10.1016/j.ajem.2025.04.040.
24. Wang C, Wang F, Li S, Ren Q-W, Tan X, Fu Y, et al. Patient triage and guidance in emergency departments using large language models: multimetric study. J Med Internet Res. 2025. https://doi.org/10.2196/71613.
25. Braam E. Exploring AI-assisted triage: comparing classical machine learning and GPT-4o for clinical decision making. Utrecht University Students Theses. 2025. https://studenttheses.uu.nl/handle/20.500.12932/48926.
26. Ho B, Lu M, Wang X, Butler R, Park J, Ren D. Evaluation of generative artificial intelligence models in predicting pediatric emergency severity index levels. Pediatr Emerg Care. 2025. https://doi.org/10.1097/PEC.0000000000003315.
27. Altalib S, Riboli-Sasco E, Al Ammouri M, Gibson H, Leung K, Painter A, et al. Benchmarking artificial intelligence vs general practitioners decision-making in same-day appointments triage: a mixed-methods study in UK primary care. medRxiv. 2025. https://doi.org/10.1101/2025.06.11.25329441.
28. Halwani MA, Merdad G, Almasre M, Doman G, AlSharif S, Alshiakh SM, et al. Predicting triage of pediatric patients in the emergency department using machine learning approach. Int J Emerg Med. 2025;18:article number 51. https://doi.org/10.1186/s12245-025-00861-z.
29. Anoka WC. A new approach: evaluating the effectiveness of artificial intelligence developments for trauma triage in the emergency department (2025). Spring Honors Capstone Projects. 1. 2025. https://mavmatrix.uta.edu/honors_spring2025/1.
30. Liu F, Li Z, Zhou H, Yin Q, Yang J, Tang X, et al. Large language models are poor clinical decision-makers: a comprehensive benchmark. arXiv. 2024. https://doi.org/10.48550/arXiv.2405.00716.
31. Arora RK, Wei J, Hicks RS, Bowman P, Quiñonero-Candela J, Tsimpourlas F, et al. HealthBench: evaluating large language models towards improved human health. arXiv. 2025. https://doi.org/10.48550/arXiv.2505.08775.
32. Kirch NM, Hebenstreit K, Samwald M. TRIAGE: ethical benchmarking of AI models through mass casualty simulations. arXiv:2410.18991. Submitted on 10 Oct 2024 (v1), last revised 4 Nov 2024 (this version, v2) https://doi.org/10.48550/arXiv.2410.18991.