From Triage to Treatment: Scoping the Role of Large Language Models in Transforming Emergency Medicine

Carl Preiksaitis, Nicholas Ashenburg, Gabrielle Bunney, Andrew Lee Chu, Rana Kabeer, Fran Riley, Ryan Ribeira, Christian Rose

Submitted to: Journal of Medical Internet Research
on: October 19, 2023

Disclaimer: © The authors. All rights reserved. This is a privileged document currently under peer-review/community review. Authors have provided JMIR Publications with an exclusive license to publish this preprint on it's website for review purposes only. While the final peer-reviewed paper may be licensed under a CC BY license on publication, at this stage authors and publisher expressively prohibit redistribution of this draft paper other than for review purposes.

Table of Contents

From Triage to Treatment: Scoping the Role of Large Language Models in Transforming Emergency Medicine

Carl Preiksaitis[1] MD; Nicholas Ashenburg[1] MD; Gabrielle Bunney[1] MD, MBA; Andrew Lee Chu[1] MD; Rana Kabeer[1] MD, MPH; Fran Riley[1] MD, MSE; Ryan Ribeira[1] MD, MPH; Christian Rose[1] MD

[1]Department of Emergency Medicine Stanford University School of Medicine Palo Alto US

Corresponding Author:
Carl Preiksaitis MD
Department of Emergency Medicine
Stanford University School of Medicine
900 Welch Road
Suite 350
Palo Alto
US

Abstract

Background: Artificial intelligence (AI), more specifically large language models (LLMs), hold significant potential to transform the landscape of emergency care delivery. Although enthusiasm for integrating LLMs into emergency medicine (EM) is growing, the existing literature is characterized by a disparate collection of individual studies, conceptual analyses, and preliminary implementations. Given these complexities and gaps in understanding, a cohesive framework is needed to make sense of the existing body of knowledge on the application of LLMs in EM.

Objective: This scoping review sought to map the current literature on the potential uses of LLMs in EM and identify directions for future research.

Methods: Using PRISMA-ScR criteria, we searched Ovid MEDLINE, Embase, Web of Science, and Google Scholar for articles published between January 2018 and August 2023 that discussed LLM use in EM. We excluded other forms of artificial intelligence. Titles and abstracts were screened, and each full text article was independently reviewed by two authors. Data was abstracted independently and 5 authors performed a collaborative quantitative and qualitative synthesis of the data.

Results: Of 1992 identified citations, 42 were included. Studies were predominantly from 2022-2023 and conducted in the USA and China. Four major themes emerged: (1) Clinical decision support, including applications in public health messaging, triage, diagnosis, treatment recommendations, and outcome predictions; (2) Workflow efficiency, through information retrieval and synthesis to reduce physician cognitive load; (3) Risks and ethics, with concerns about model accuracy, transparency, and legal implications noted; and (4) Education and communication, with potential uses in medical training, patient counseling, and knowledge dissemination identified.

Conclusions: This review establishes an initial framework of the capabilities and limitations of LLMs in EM based on reported use cases and identifies key areas for future research. These include prospective validation of proposed applications, developing standards for responsible use, exploring provider and patient perceptions, and fostering physician literacy in artificial intelligence. Thoughtful collaboration and critical evaluation will be essential to safely and effectively integrate LLMs into emergency care.

(JMIR Preprints 19/10/2023:53787)
DOI: https://doi.org/10.2196/preprints.53787

Preprint Settings

No, I do not wish to publish my submitted manuscript as a preprint.

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**

Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain v

Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <a href="http

Original Manuscript

From Triage to Treatment: Scoping the Role of Large Language Models in Transforming Emergency Medicine

ABSTRACT

Background

Artificial intelligence (AI), more specifically large language models (LLMs), hold significant potential to transform the landscape of emergency care delivery. Although enthusiasm for integrating LLMs into emergency medicine (EM) is growing, the existing literature is characterized by a disparate collection of individual studies, conceptual analyses, and preliminary implementations. Given these complexities and gaps in understanding, a cohesive framework is needed to make sense of the existing body of knowledge on the application of LLMs in EM.

Objectives

This scoping review sought to map the current literature on the potential uses of LLMs in EM and identify directions for future research.

Methods

Using PRISMA-ScR criteria, we searched Ovid MEDLINE, Embase, Web of Science, and Google Scholar for articles published between January 2018 and August 2023 that discussed LLM use in EM. We excluded other forms of artificial intelligence. Titles and abstracts were screened, and each full text article was independently reviewed by two authors. Data was abstracted independently and 5 authors performed a collaborative quantitative and qualitative synthesis of the data.

Results

Of 1992 identified citations, 42 were included. Studies were predominantly from 2022-2023 and conducted in the USA and China. Four major themes emerged: (1) Clinical decision support, including applications in public health messaging, triage, diagnosis, treatment recommendations, and outcome predictions; (2) Workflow efficiency, through information retrieval and synthesis to reduce physician cognitive load; (3) Risks and ethics, with concerns about model accuracy, transparency, and legal implications noted; and (4) Education and communication, with potential uses in medical training, patient counseling, and knowledge dissemination identified.

Conclusion

This review establishes an initial framework of the capabilities and limitations of LLMs in EM based on reported use cases and identifies key areas for future research. These include prospective validation of proposed applications, developing standards for responsible use, exploring provider and patient perceptions, and fostering physician literacy in artificial intelligence. Thoughtful collaboration and critical evaluation will be essential to safely and effectively integrate LLMs into emergency care.

INTRODUCTION

Background

Emergency Medicine is at an inflection point. With increasing patient volumes, decreasing staff availability, and rapidly-evolving clinical guidelines, emergency providers are overburdened and burnout is significant [1]. Innovative solutions are needed to continue to provide the high level of care our patients deserve. Artificial intelligence (AI) and, more specifically, large language models (LLMs), stand out as a promising avenue to revolutionize the delivery of emergency care.

A large language model is a deep learning-based artificial neural network trained on vast amounts of textual data enabling them to recognize, translate, predict, or generate text or other content [2]. Characterized by transformer architecture and an ability to encode contextual information using a vast number of parameters, LLMs can utilize knowledge on a large variety of topics. These models can assist in a multitude of tasks, such as real-time data interpretation, augmenting clinical decision-making, and patient engagement in clinical settings. For example, LLMs could sort through electronic health records to flag critical patient history, help clinicians interpret multi-modal diagnostic data, or act as decision support tools in differential diagnosis, enhancing the quality of care while reducing cognitive load and decision fatigue for emergency providers. What is more, these models can generate content, meaning they can create a variety of different text outputs from technical computer code to essays and poetry.

Importance

While interest in applying LLMs to emergency medicine (EM) is gaining momentum, the current body of literature remains a patchwork of isolated studies, theoretical discussions, and small-scale implementations. Moreover, existing research often focuses on specific use-cases like diagnostic assistance or triage prioritization, rather than providing a holistic view of how LLMs can be integrated into the EM workflow. This fragmented landscape makes it challenging for emergency clinicians, who are already burdened by the complexities and pace of their practice, to discern actionable insights or formulate a coherent strategy for adopting these technologies. Despite the promise shown by several models, like ChatGPT-4 or Med PaLM 2, the absence of standardized metrics for evaluating their clinical efficacy, ethical use, and long-term sustainability leaves researchers and clinicians navigating uncharted territory. Consequently, the potential for LLMs to enhance emergency medical care remains largely untapped and poorly understood.

Goals of This Investigation

In light of these complexities and informational disparities, our study undertakes a crucial step to consolidate, assess, and contextualize the fragmented knowledge base surrounding LLMs in EM. Through a scoping review, we aim to establish a foundational understanding of the field's current standing, from technological capabilities to clinical applications and ethical considerations. This synthesis serves a dual purpose: first, to equip emergency providers with a navigable map of existing research and, second, to identify critical gaps and avenues for future inquiry. As EM increasingly embraces technological solutions for its unique challenges, our goal is to provide clarity to the responsible and effective incorporation of LLMs into clinical practice.

METHODS

We adhered to the Preferred Reporting Items for Systematic reviews and Meta-Analyses extension for Scoping Reviews (PRISMA-ScR) Checklist and used the scoping review methodology proposed by Arksey and O’Malley and furthered by Levac et al [3–5]. This included the following steps: (a) identifying the research question; (b) identifying relevant studies; (c) selecting studies; (d) charting the data; (e) collating, summarizing, and reporting the results, and (f) consultation. Our full review protocol is available online [6].

Identifying the research question

The overall purpose of this review was to map the current literature describing the potential uses of large language models in EM and to identify directions for future research. In order to achieve this goal, we aimed to answer the primary research question: “What are the current and potential uses of LLMs in EM described in the literature?” We chose to explicitly focus on LLMs as this subset of artificial intelligence is rapidly developing and generating significant interest for potential applications.

Identifying Relevant Studies

In August 2023, we searched Ovid MEDLINE, Embase, Web of Science, and Google Scholar for potential citations of interest. We limited our search to articles published after January 2018 as the Bidirectional Encoder Representations from Transformers (BERT) model was introduced this year and considered by many to be the first in the contemporary class of large language models [7]. Our search strategy (Appendix 1), created in consultation with a medical librarian, combined keywords and medical subject headings related to LLMs and EM. We reviewed the bibliographies of identified studies for potential missed articles.

Study Selection

Citations were managed using Covidence online software (Veritas Health Innovation, Melbourne, Australia). Manuscripts were included if they discussed use of a LLM in EM, including applications in the emergency department, prehospital, and peri-admission settings. We also included use cases related to public health, disease monitoring, or disaster preparedness as these are relevant to emergency departments. We excluded studies that used other forms of machine learning or natural language processing that were not LLMs and studies that did not clearly relate to EM. We additionally did not include instances where the sole use of an LLM was in generation of the manuscript without additional commentary.

Two investigators (CP, CR) independently screened 100 abstracts and interrater reliability showed substantial agreement (kappa=0.75). The remaining abstracts were screened by one author (CP), who consulted with a second author as needed for clarification regarding inclusion and exclusion criteria. All manuscripts meeting the initial criteria were independently reviewed in full by two authors (CP and CR). Studies determined to meet the eligibility criteria by both reviewers were included in the analysis. Discrepancies were resolved by consensus and with the addition of a third reviewer (NA) if needed. Our initial search strategy identified 2065 articles, of which 73 were duplicates, resulting in a total of 1992 articles for screening (Figure 1). A total of 1891 articles were excluded based on title or abstract. A total of 101 studies were reviewed in full and 42 manuscripts were found to meet the study inclusion criteria.

<ins>Charting the Data</ins>

Data abstraction was independently conducted using a structured form to capture article details including the author, year of publication, study type, specific study population, study or article location, purpose, and main findings. Data to address our primary research question was iteratively abstracted from the articles as our themes emerged as explained further below.

<ins>Collating, Summarizing and Reporting the Results</ins>

We synthesized and collated the data, performing both a quantitative and qualitative analysis. A descriptive summary of the included studies was created. Braun and Clarke’s methodology was then used to conduct a thematic analysis to address our primary research question [8]. Five authors (CP, CR, AC, NA, RR) independently familiarized themselves with and generated codes for a purposively diverse selection of 10 papers, focusing on content that suggested possible uses for LLMs in EM. The group met to discuss preliminary findings and refine the group's approach. Individuals then independently aggregated codes into themes. These themes were reviewed and refined as a group. Two authors (CP and CR) then reviewed the remaining manuscripts for any additional themes as well as data that supported or contradicted our existing themes. These data were used to refine themes through group discussion. Our analysis included a discussion and emphasis on the implications and future research directions for the field, based on the guidance from Levac et al [4].

Consultation

In order to ensure our review accurately characterized the available knowledge and that our interpretations of it were correct, we consulted with external emergency physicians with topic expertise in artificial intelligence. We incorporated feedback as appropriate. For example, we more completely defined LLMs for clarity and included a table describing common models (Table 3). Our findings and recommendations were endorsed by our consultants.

RESULTS

The majority of identified studies were published in 2023 (28/42, 66.7%). Fourteen studies were conducted in the USA, followed by 6 in China, 4 in Australia, 3 in Taiwan, 2 in Singapore and Korea, and several other individual studies from a variety of countries (Table 1). In terms of study type, 17 papers were methodology studies, 17 were case studies, 6 were commentaries, and there was one each of a case report, qualitative investigation, and retrospective cross-sectional study. Twenty-four of these studies addressed the ED setting specifically, followed by 6 addressing the pre-hospital setting and 6 in other non-ED hospital settings. Three studies were focused on use of LLMs for the general public, 2 studies addressed using them for social media analysis, and one study focused on research applications. Large language models used included versions of GPT (ChatGPT, GPT-4), PaLM (Bard), ELMo, and BERT (BioBERT, ClinicalBERT, DeBERTa) (Table 3).

We identified four major themes in our analysis: (1) Clinical Decision Making and Support, (2) Efficiency, Workflow, and Information Management, (3) Risks, Ethics, and Transparency, and (4) Education and Communication. Major themes, subthemes, and representative quotations are in Table 2.

Theme 1: Clinical Decision Making and Support

The first theme we identified is clinical decision making and support. LLMs have been used or proposed for applications such as providing advice to the public before arrival, aiding in triage as patients arrive to the ED, or augmenting the activities of physicians as they provide care, either through supporting diagnostics or predicting patient resource utilization.

Several applications focused on advising the general public and aiding in symptom checking, self-triage, and occasionally advising first-aid prior to arrival of emergency medical services. These included counseling parents during potential pediatric emergencies, recognizing stroke, or providing advice during potential cardiac arrests [9–11]. Wang, et al. proposed a model that could potentially help patients navigate the complexities of the healthcare system in China and present to the correct medical setting for the care they need [12].

LLMs also have the potential to efficiently screen patients for important outcomes, such as pediatric patients at risk for non-accidental trauma, suicide risk, or COVID infection [13–15]. These can be implemented based on data in the medical record or as clinical data is obtained in real time.

Early identification of patient risks could help physicians more rapidly identify important diagnoses. Several studies discussed implementations of LLMs that work in conjunction with physicians while caring for patients in the ED [16,17]. Brown, et al discuss the potential role for these models to overcome cognitive biases and reduce errors [18]. These models can be used in developing a differential diagnosis, recommending imaging studies, providing treatment recommendations, or interpreting clinical guidelines [19–22].

Several studies centered on predicting outcomes such as presentation to the ED, hospitalization, ICU admission, or in-hospital cardiac arrest [23–26]. Applications of LLMs in the triage process could potentially identify higher risk patients or patients at high risk of certain diagnoses, such as gastrointestinal bleeding [27–30].

Theme 2: Efficiency, Workflow, and Information Management

The second theme identified is information management, workflow, and efficiency. LLMs show great promise in increasing the usability of data available in the electronic health record (EHR). Interactions with the EHR take up a significant amount of physician time and it is often difficult to identify crucial information during critical times [31]. Models can serve a variety of information management functions. They can be used to perform audits for quality improvement purposes, identify potential adverse events, such as drug interactions, anticipate and monitor public health emergencies, and assist with information entry during the clinical encounter [31–39]. Models developed and trained on data from the emergency department can quickly identify similar patient presentations, recognize patterns, and extract important information from unstructured text [35,40,41].

Some authors suggest LLMs can improve care throughout the entirety of the EM encounter [15–18].

LLMs could potentially be used as digital adjuncts for clinical decision making, since they could possibly generate differentials, predict final diagnoses, offer interpretations of imaging studies, and suggest treatment plans [15,17,18,42]. They may mitigate human cognitive biases and address human factors (e.g., time constraints, frequent task switching, high cognitive load, constant interruptions, decision fatigue) that predispose emergency physicians to error [18].

The flexibility and versatility of these offer particular benefits to emergency medicine practice. The diverse ways in which these models can aid throughout the entire clinical workflow can help physicians process large quantities of complex clinical data, mitigate cognitive biases, and deliver relevant information in a digestible format [15,17,18,42]. By streamlining these burdensome tasks, LLMs can help improve the efficiency of care for the high volume of patients they routinely see in the ED.

Theme 3: Risks, Transparency, and Ethics

Despite the potential for advancement and improvement in the care that EM physicians can provide through the inclusion of LLMs in practice, there are several issues that limit their implementation into practice at this time.

The most often discussed risk, mentioned in eleven articles, is the reliability of model responses and the potential for erroneous results [15,17,20–22,27,35,38,43–45]. These output errors often result from inaccuracies in the training data, which is most commonly gathered from the internet and unvetted for reliability. Sources of inaccurate responses may be identified by looking at training material, but others due to data noise, mislabeling, or outdated information may be harder to detect [15,21,38,40]. Similarly, biases in training data can be propagated to the model, leading to inaccurate or discriminatory results [17,24,27,46]. In medical applications, the consequences of the errors can be significant, and even small errors could lead to adverse outcomes [17].

Understanding and mitigating errors in large language models is challenging due to issues with transparency and reproducibility of model outputs [18,27,45–47]. Better understanding among clinicians of the algorithms and statistical methods used by LLMs is one suggested way to ensure cautious use [18]. Concentrating on making models more explainable or transparent is another potential approach [46]. However, the degree to which this will be feasible given the complexity of these models remains to be determined.

Patient and data privacy is another clearly articulated risk of using these models in the clinical environment [18,26,27]. There are some proposed methodologies using unsupervised methods that can train models with limited access to sensitive information, however these require further exploration [26]. Patient attitudes and willingness to allow models access to their health information for training and how to address disclosure of this use have not been extensively discussed. Finally, the legal and ethical implications of using LLM output to guide patient care is an often mentioned concern [18,27,45]. How the responsibility for patient care decisions is distributed if LLMs are used to guide clinical decisions has yet to be determined.

Theme 4: Education and Communication

Finally, these models offer several opportunities for education and communication. First, several papers noted that successful integration of LLMs into clinical practice will require physicians to understand the underlying algorithms and statistical methods used by these models [18,45]. There is a need for dedicated educational programs on AI in medicine at all levels of medical education to ensure solutions are developed that align with the clinical environment and address the unique challenges of working with clinical data [13,17,48].

In terms of clinical education, several studies have demonstrated reasonable performance of LLMs on standardized tests in medicine, which could indicate potential for these models to develop study materials [49]. Additionally, these models may be able to help physicians communicate with and educate patients. El Dahdah et al used ChatGPT to answer several common medical questions in easy to understand language, suggesting the ability to enhance physician responses to patient queries [50]. Webb demonstrated the use of ChatGPT to simulate patient conversation and provide feedback to a physician learning how to break bad news [47].

Patient education may be facilitated via these models without physician input as well. As discussed above, several authors described applications designed to educate patients during emergencies prior to arrival in the ED [9–12]. Finally, LLMs can be used to aid in knowledge dissemination. Gottleib et al. and Babl et al. describe potential applications for LLMs in research and scientific writing [51,52]. They highlight potential benefits to individuals who struggle with English or have challenges with writing or knowledge synthesis. Additionally, models may be used to translate scientific articles more rapidly. However, use of these models to generate scientific articles raises concerns regarding the potential for academic dishonesty [51,52].

LIMITATIONS

This scoping review has some limitations worth noting. First, we restricted our search to articles published after 2018, when large language models first emerged. While this captures the current era of LLMs, earlier works relevant to natural language processing in emergency medicine may have been overlooked. Additionally, despite searching four databases and consulting a medical librarian on the search strategy, some pertinent studies may have been missed given the rapidly evolving nature of this research area. However, our review establishes an initial foundation that can be built upon as the field continues to grow.

DISCUSSION

This scoping review found that large language models hold tremendous potential to impact the delivery of emergency care. Although several specific applications and limitations have been reported and suggested in the literature, our analysis identified four major areas of focus for LLMs in emergency medicine: clinical decision support, workflow efficiency, risks/ethics, and education. We propose these topics as a framework for understanding emerging implementations of LLMs and a guide to inform future areas of investigation.

At their core, large language models and their associated natural language processing techniques offer a way to organize and engage with vast amounts of unstructured text data. Depending on how they are trained and used, they can be operationalized to make predictions or identify patterns, which gives rise to the majority of our identified applications. Most commercially-available large language models, like ChatGPT, are trained on massive volumes of text gathered from the internet then optimized for conversational interaction [53]. This ability to access a breadth of general knowledge and resulting wide applicability has contributed to the increased use of LLMs by professionals and the general public across a variety of fields [54]. As these models become more ubiquitous, there is potential for their use across the care continuum. Not only could they support clinical care, but they may offer an opportunity to help advise the general public with medical concerns. Several papers in our review identified the feasibility of using LLMs to provide first aid instructions and offer decision support to potential patients looking to access care [9–11].

Preliminary work suggests that dedicated training can enhance the ability of these models to make triage recommendations, but prospective implementation has not been tested [12]. LLMs certainly could aid patients in self-triage or with basic medical questions, but how this can be effectively and safely implemented needs further exploration, especially with concerns regarding accuracy of outputs. Possibilities to improve outputs include additional dedicated training of models to the medical and emergency settings to improve their reliability and accuracy. These context-specific models could be equipped with information on the local healthcare system to help patients identify available resources, schedule appointments, or activate emergency medical services.

In the emergency department, LLMs can increase workflow efficiency by rapidly synthesizing relevant information from a patient's medical record, structuring and categorizing chief complaint data, and assigning an emergency severity index (ESI) level [27,28,30,41,43,50]. Additionally, quickly accessing data from the medical record could improve efficiency and thoroughness of chart review. A model's ability to identify subtle patterns in data could offer additional diagnostic support by recommending or interpreting laboratory and imaging studies [15,17,18,42]. By facilitating tasks like information retrieval and synthesis, LLMs can reduce this burden for clinicians and minimize errors due to buried or disorganized data, potentially contributing to workflow efficiency. They may also counteract human cognitive biases and fatigue when used to support clinical decisions. Although some studies have demonstrated reasonable accuracy on focused use cases, further validation of any of these applications across diverse settings and patient populations is required. Thoughtful integration of LLMs has the potential to revolutionize emergency medicine by providing clinical decision support, improving situational awareness, and increasing productivity.

However, barriers to seamless implementation exist. As noted by several authors, erroneous outputs remain a concern given dependence on training data [15,17,21,22,26,27,38,43–45]. Information surrounding the most publicly available LLMs today is obscured across three important layers (1) the underlying training data used–commonly reported to be publicly available data on the internet and from third-party licensed datasets, (2) the underlying architecture of the model–whose exact mechanisms are not always easy to discern, and (3) the intricacies of human-led fine tuning–often done at the end of development to provide guardrails for output. These layers of obscurity make it difficult to troubleshoot the cause of any single erroneous output. Issues of privacy and data rights must also be addressed and discussions surrounding information used these models are ongoing [18,27,45]. Additionally, legal and ethical implications of AI-assisted healthcare require further exploration to establish appropriate oversight and accountability structures. Without explainability and transparency, utilization of “black box” LLMs may face clinician resistance.

Our review reveals several opportunities for future exploration and research. Perhaps the most important is effectively identifying problems which are best solved using LLMs in emergency medicine. Our review outlines several immediate areas of potential exploration including improved communication, translation, and summarization of highly detailed and domain specific knowledge for providers and patients, but further exploration and prospective validation of specific use cases is required. We expect the potential use cases in emergency medicine to grow as LLMs become increasingly complex and develop emergent properties–actions that are not explicitly programmed or anticipated. To bridge the *AI chasm* between innovations in the research realm and widespread adoption, these applications should be identified with significant input from providers in the clinical space who can uniquely identify areas of potential benefit. To accomplish this, a better understanding of the abilities and limitations of LLMs among physicians is needed to optimize their best use and ensure they are effectively implemented, and AI literacy is increasingly described as an essential competency for physicians [55]. We encourage development of curricula and training programs designed for emergency physicians.

Given the black-box nature of LLMs, standardized frameworks and metrics for evaluation that are specific for healthcare use cases are needed to evaluate their performance and implementation effectively. These should include consideration of both the technological abilities and limitations of a model as well as the elements of human interaction that influence use. Implementation and validation of solutions should be done across heterogeneous populations and care environments with special attention to cohorts underrepresented in the training data who could be harmed by model biases. Provider perspectives are essential, but equally important are patient perspectives about the use of LLMs in medicine. Impacts on doctor-patient communication, patient concerns surrounding privacy, and attitudes towards AI-generated recommendations must be further explored. Collaboration between all relevant stakeholders who develop or will be impacted by LLMs for clinical medicine is essential to developing models that can be used effectively, equitably, and safely.

In summary, large language models hold significant potential to impact the delivery of emergency care. This review identified clinical decision support, workflow efficiency, risks/ethics, and education as major areas of focus in the literature and potential avenues for future exploration and research. Collaboration with diverse stakeholders, thoughtful implementation, and critical evaluation appear to be essential processes to ensure the impact of these models on emergency medicine is a positive one.

REFERENCES

1. Petrino R, Riesgo LG-C, Yilmaz B. Burnout in emergency medicine professionals after 2 years of the COVID-19 pandemic: a threat to the healthcare system? Eur J Emerg Med 2022 Aug;29(4):279–284. PMID:35620812

2. Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language models in medicine. Nat Med 2023 Aug;29(8):1930–1940. PMID:37460753

3. Tricco AC, Lillie E, Zarin W, O'Brien KK, Colquhoun H, Levac D, Moher D, Peters MDJ, Horsley T, Weeks L, Hempel S, Akl EA, Chang C, McGowan J, Stewart L, Hartling L, Aldcroft A, Wilson MG, Garritty C, Lewin S, Godfrey CM, Macdonald MT, Langlois EV, Soares-Weiser K, Moriarty J, Clifford T, Tunçalp Ö, Straus SE. PRISMA Extension for Scoping Reviews (PRISMA-ScR): Checklist and Explanation. Ann Intern Med 2018 Oct 2;169(7):467–473. PMID:30178033

4. Levac D, Colquhoun H, O'Brien KK. Scoping studies: advancing the methodology. Implement Sci 2010 Sep 20;5(1):69. doi: 10.1186/1748-5908-5-69

5. Arksey H, O'Malley L. Scoping studies: towards a methodological framework. Int J Soc Res Methodol Routledge; 2005 Feb 1;8(1):19–32. doi: 10.1080/1364557032000119616

6. Preiksaitis C. Protocol for a Scoping Review of the Application of Large Language Models in Emergency Medicine. OSF; 2023 Oct 19; Available from: https://osf.io/tdghu/ [accessed Oct 18, 2023]

7. Devlin J, Chang M-W, Lee K, Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv; 2019. doi: 10.48550/arXiv.1810.04805

8. Braun V, Clarke V. Using thematic analysis in psychology. Qual Res Psychol Routledge; 2006

Jan 1;3(2):77–101. doi: 10.1191/1478088706qp063oa

9. Bushuven S, Bentele M, Bentele S, Gerber B, Bansbach J, Ganter J, Trifunovic-Koenig M, Ranisch R. “ChatGPT, can you help me save my child’s life?”-Diagnostic Accuracy and Supportive Capabilities to lay rescuers by ChatGPT in prehospital Basic Life Support and Paediatric Advanced Life Support cases–an in-silico analysis. 2023; doi: 10.21203/rs.3.rs-2910261/v1

10. Lam WY, Au SCL. Stroke care in the ChatGPT era: Potential use in early symptom recognition. J Acute Dis Medknow; 2023;12(3):129–130. doi: 10.4103/2221-6189.379278

11. Ahn C. Exploring ChatGPT for information of cardiopulmonary resuscitation. Resuscitation Elsevier; 2023;185. doi: 10.1016/j.resuscitation.2023.109729

12. Wang J, Zhang G, Wang W, Zhang K, Sheng Y. Cloud-based intelligent self-diagnosis and department recommendation service using Chinese medical BERT. J CLOUD Comput-Adv Syst Appl 2021;10(1). doi: 10.1186/s13677-020-00218-2

13. Huang D, Cogill S, Hsia R, Yang S, Kim D. Development and external validation of a pretrained deep learning model for the prediction of non-accidental trauma. NPJ Digit Med 2023;6(1). doi: 10.1038/s41746-023-00875-y

14. Pease J, Thompson D, Wright-Berryman J, Campbell M. User Feedback on the Use of a Natural Language Processing Application to Screen for Suicide Risk in the Emergency Department. J Behav Health Serv Res 2023; doi: 10.1007/s11414-023-09831-w

15. Drozdov I, Szubert B, Reda E, Makary P, Forbes D, Chang SL, Ezhil A, Puttagunta S, Hall M, Carlin C, Lowe DJ. Development and prospective validation of COVID-19 chest X-ray screening model for patients attending emergency departments. Sci Rep England;

2021;11(1):20384. doi: 10.1038/s41598-021-99986-3

16. Cheng K, Li Z, Guo Q, Sun Z, Wu H, Li C. Emergency surgery in the era of artificial intelligence: ChatGPT could be the doctor’s right-hand man. Int J Surg Lond Engl 2023;109(6):1816–1818. doi: 10.1097/JS9.0000000000000410

17. Rao A, Pang M, Kim J, Kamineni M, Lie W, Prasad AK, Landman A, Dreyer KJ, Succi MD. Assessing the Utility of ChatGPT Throughout the Entire Clinical Workflow. medRxiv 2023; ((Rao A.; Pang M.; Kim J.; Kamineni M.; Lie W.; Prasad A.K.; Succi M.D., msucci@mgh.harvard.edu) Harvard Medical School, Boston, MA, United States). doi: 10.1101/2023.02.21.23285886

18. Brown C, Nazeer R, Gibbs A, Le Page P, Mitchell AR. Breaking Bias: The Role of Artificial Intelligence in Improving Clinical Decision-Making. Cureus United States; 2023;15(3):e36415. doi: 10.7759/cureus.36415

19. Gupta P, Nayak R, Alazzeh M. The Accuracy of Medical Diagnoses in Emergency Medicine by Modern Artificial Intelligence. Acad Emerg Med 2023;30((Gupta P.) Stanford Emergency Medicine Residency Program, United States):395. doi: 10.1111/acem.14718

20. Barash Y, Klang E, Konen E, Sorin V. ChatGPT-4 Assistance in Optimizing Emergency Department Radiology Referrals and Imaging Selection. J Am Coll Radiol JACR 2023;((Barash Y., yibarash@gmail.com; Sorin V.) Department of Diagnostic Imaging, Chaim Sheba Medical Center, Tel Hashomer, Israel; DeepVision Lab, Chaim Sheba Medical Center, Tel Hashomer, Israel). doi: 10.1016/j.jacr.2023.06.009

21. Altamimi I, Altamimi A, Alhumimidi AS, Altamimi A, Temsah M-H, Altamimi I, Altamimi A, Alhumimidi AS, Altamimi A, Temsah M-H. Snakebite Advice and Counseling From Artificial

Intelligence: An Acute Venomous Snakebite Consultation With ChatGPT. Cureus Cureus; 2023 Jun 13;15(6). doi: 10.7759/cureus.40351

22. Hamed E, Eid A, Alberry M. Exploring ChatGPT's Potential in Facilitating Adaptation of Clinical Guidelines: A Case Study of Diabetic Ketoacidosis Guidelines. Cureus 15(5):e38784. PMID:37303347

23. Chae S, Davoudi A, Song J, Evans L, Hobensack M, Bowles K, McDonald M, Barron Y, Rossetti S, Cato K, Sridharan S, Topaz M. Predicting emergency department visits and hospitalizations for patients with heart failure in home healthcare using a time series risk model. J Am Med Inform Assoc 2023; doi: 10.1093/jamia/ocad129

24. Gebrael G, Sahu KK, Chigarira B, Tripathi N, Mathew Thomas V, Sayegh N, Maughan BL, Agarwal N, Swami U, Li H. Enhancing Triage Efficiency and Accuracy in Emergency Rooms for Patients with Metastatic Prostate Cancer: A Retrospective Analysis of Artificial Intelligence-Assisted Triage Using ChatGPT 4.0. Cancers 2023 Jul 22;15(14):3717. PMID:37509379

25. Tahayori B, Chini-Foroush N, Akhlaghi H. Advanced natural language processing technique to predict patient disposition based on emergency triage notes. Emerg Med Australas 2021;33(3):480–484. doi: 10.1111/1742-6723.13656

26. Chen M, Huang T, Chen T, Boonyarat P, Chang Y. Clinical narrative-aware deep neural network for emergency department critical outcome prediction. J Biomed Inform 2023;138. doi: 10.1016/j.jbi.2023.104284

27. Bhattaram S, Shinde VS, Khumujam PP. ChatGPT: The next-gen tool for triaging? Am J Emerg Med 2023 Jul;69:215–217. PMID:37024324

28. Sarbay İ, Berikol GB, Özturan İU. Performance of emergency triage prediction of an open access

natural language processing based chatbot application (ChatGPT): A preliminary, scenario-based cross-sectional study. Turk J Emerg Med 2023 Sep;23(3):156. doi: 10.4103/tjem.tjem_79_23

29. Shung D, Tsay C, Laine L, Chang D, Li F, Thomas P, Partridge C, Simonov M, Hsiao A, Tay J, Taylor A. Early identification of patients with acute gastrointestinal bleeding using natural language processing and decision rules. J Gastroenterol Hepatol 2021;36(6):1590–1597. doi: 10.1111/jgh.15313

30. Kim D, Oh J, Im H, Yoon M, Park J, Lee J. Automatic Classification of the Korean Triage Acuity Scale in Simulated Emergency Rooms Using Speech Recognition and Natural Language Processing: a Proof of Concept Study. J KOREAN Med Sci 2021;36(27). doi: 10.3346/jkms.2021.36.e175

31. Preiksaitis C, Sinsky CA, Rose C. Chatgpt is not the solution to physicians’ documentation burden. Nat Med Nature Publishing Group US New York; 2023;1–2.

32. Wang H, Yeung W, Ng Q, Tung A, Tay J, Ryanputra D, Ong M, Feng M, Arulanandam S. A Weakly-Supervised Named Entity Recognition Machine Learning Approach for Emergency Medical Services Clinical Audit. Int J Environ Res Public Health 2021;18(15). doi: 10.3390/ijerph18157776

33. Rahman MA, Preum SM, Williams RD, Alemzadeh H, Stankovic J. EMS-BERT: A Pre-Trained Language Representation Model for the Emergency Medical Services (EMS) Domain. 2023;

34. Zhang X, Zhang H, Sheng L, Tian F. DL-PER: Deep Learning Model for Chinese Prehospital Emergency Record Classification. IEEE Access IEEE; 2022;10:64638–64649. doi: 10.1109/ACCESS.2022.3179685

35. Chen Y-P, Chen Y-Y, Lin J-J, Huang C-H, Lai F. Modified Bidirectional Encoder Representations From Transformers Extractive Summarization Model for Hospital Information Systems Based on Character-Level Tokens (AlphaBERT): Development and Performance Evaluation. JMIR Med Inform Canada; 2020;8(4):e17787. doi: 10.2196/17787

36. Gil-Jardine C, Chenais G, Pradeau C, Tentillier E, Revel P, Combes X, Galinski M, Tellier E, Lagarde E. Trends in reasons for emergency calls during the COVID-19 crisis in the department of Gironde, France using artificial neural network for natural language classification. Scand J TRAUMA Resusc Emerg Med 2021;29(1). doi: 10.1186/s13049-021-00862-w

37. Wang T, Lu K, Chow KP, Zhu Q. COVID-19 sensing: negative sentiment analysis on social media in China via BERT model. Ieee Access IEEE; 2020;8:138162–138169. doi: 10.1109/ACCESS.2020.3012595

38. McMaster C, Chan J, Liew DFL, Su E, Frauman AG, Chapman WW, Pires DEV. Developing a deep learning natural language processing algorithm for automated reporting of adverse drug reactions. J Biomed Inform 2023;137((McMaster C., christopher.mcmaster@austin.org.au; Liew D.F.L.; Su E.; Frauman A.G.) Department of Clinical Pharmacology&Therapeutics, Austin Health, Melbourne, VIC, Australia). doi: 10.1016/j.jbi.2022.104265

39. Bradshaw JC. The ChatGPT Era: Artificial Intelligence in Emergency Medicine. Ann Emerg Med 2023;81(6):764–765. doi: 10.1016/j.annemergmed.2023.01.022

40. Chang D, Hong WS, Taylor RA. Generating contextual embeddings for emergency department chief complaints. JAMIA Open 2020;3(2):160–166. doi: 10.1093/jamiaopen/ooaa022

41. Xu B, Gil-Jardiné C, Thiessard F, Tellier E, Avalos M, Lagarde E. Pre-training a neural language model improves the sample efficiency of an emergency room classification model. The AAAI

Press; 2020.

42. Chen H-L, Chen H-H. Have You Chatted Today?-Medical Education Surfing with Artificial Intelligence. J Med Educ □□□□□□□□; 2023;27(1):1–4.

43. Chang D, Hong WS, Taylor RA. Generating contextual embeddings for emergency department chief complaints. JAMIA Open United States; 2020;3(2):160–166. doi: 10.1093/jamiaopen/ooaa022

44. Chen Y-P, Lo Y-H, Lai F, Huang C-H. Disease concept-embedding based on the self-supervised method for medical information extraction from electronic health records and disease retrieval: Algorithm development and validation study. J Med Internet Res 2021;23(1). doi: 10.2196/25113

45. Okada Y, Mertens M, Liu N, Lam SSW, Ong MEH. AI and machine learning in resuscitation: Ongoing research, new concepts, and key challenges. Resusc Plus 2023 Sep 1;15:100435. doi: 10.1016/j.resplu.2023.100435

46. Fanconi C, van Buchem M, Hernandez-Boussard T. Natural Language Processing Methods to Identify Oncology Patients at High Risk for Acute Care with Clinical Notes. AMIA Summits Transl Sci Proc 2023 Jun 16;2023:138–147. PMID:37350895

47. Webb JJ. Proof of Concept: Using ChatGPT to Teach Emergency Physicians How to Break Bad News. Cureus Cureus; 2023 May 9;15(5). doi: 10.7759/cureus.38755

48. Carl Preiksaitis, Christian Rose. Generative Artificial Intelligence in Medical Education: Opportunities, Challenges, and Future Directions - A Scoping Review. JMIR Med Educ 2023 Sep 28; doi: 10.2196/48785

49. Smith J, Choi PMC, Buntine P. Will code one day run a code? Performance of language models

on ACEM primary examinations and implications. EMA - Emerg Med Australas 2023;((Smith J., jesse.lachlan.smith@gmail.com; Buntine P.) Eastern Health Emergency Medicine Program, Eastern Health, Melbourne, VIC, Australia(Choi P.M.C.; Buntine P.) Department of Neuroscience, Eastern Health, Melbourne, VIC, Australia(Choi P.M.C.) Eastern). doi: 10.1111/1742-6723.14280

50. El Dahdah J, Kassab J, El Helou MC, Gaballa A, Sayles III S, Phelan MP. ChatGPT: A Valuable Tool for Emergency Medical Assistance. Ann Emerg Med Elsevier; 2023;

51. Gottlieb M, Kline JA, Schneider AJ, Coates WC. ChatGPT and conversational artificial intelligence: Friend, foe, or future of research? Am J Emerg Med 2023 May 18;70:81–83. PMID:37229893

52. Babl FE, Babl MP. Generative artificial intelligence: Can ChatGPT write a quality abstract? Emerg Med Australas Wiley Online Library; 2023; doi: 10.1111/1742-6723.14233

53. Introducing ChatGPT. Available from: https://openai.com/blog/chatgpt [accessed Oct 6, 2023]

54. Hu K, Hu K. ChatGPT sets record for fastest-growing user base - analyst note. Reuters 2023 Feb 2; Available from: https://www.reuters.com/technology/chatgpt-sets-record-fastest-growing-user-base-analyst-note-2023-02-01/ [accessed Oct 6, 2023]

55. Boscardin CK, Gin B, Golde PB, Hauer KE. ChatGPT and Generative Artificial Intelligence for Medical Education: Potential Impact and Opportunity. Acad Med :10.1097/ACM.0000000000005439. doi: 10.1097/ACM.0000000000005439

56. Abavisani M, Dadgar F, Keikha M. A commentary on “Emergency surgery in the era of artificial intelligence: ChatGPT could be the doctor’s right-hand man.” Int J Surg Lond Engl 2023; ((Abavisani M.) Student research committee, Mashhad University of Medical Sciences,

Mashhad, Iran). doi: 10.1097/JS9.0000000000000561

57. Chen J, Liu Q, Liu X, Wang Y, Nie H, Xie X. Exploring the Functioning of Online Self-Organizations during Public Health Emergencies: Patterns and Mechanism. Int J Environ Res Public Health 2023;20(5). doi: 10.3390/ijerph20054012

58. Brown TB, Mann B, Ryder N, Subbiah M, Kaplan J, Dhariwal P, Neelakantan A, Shyam P, Sastry G, Askell A, Agarwal S, Herbert-Voss A, Krueger G, Henighan T, Child R, Ramesh A, Ziegler DM, Wu J, Winter C, Hesse C, Chen M, Sigler E, Litwin M, Gray S, Chess B, Clark J, Berner C, McCandlish S, Radford A, Sutskever I, Amodei D. Language Models are Few-Shot Learners. arXiv; 2020. doi: 10.48550/arXiv.2005.14165

59. Schreiner M. GPT-4 architecture, datasets, costs and more leaked. THE DECODER. 2023. Available from: https://the-decoder.com/gpt-4-architecture-datasets-costs-and-more-leaked/ [accessed Oct 12, 2023]

60. Pathways Language Model (PaLM): Scaling to 540 Billion Parameters for Breakthrough Performance. 2022. Available from: https://blog.research.google/2022/04/pathways-language-model-palm-scaling-to.html [accessed Oct 12, 2023]

61. AllenNLP - ELMo — Allen Institute for AI. Available from: https://allenai.org/allennlp/software/elmo [accessed Oct 12, 2023]

62. Open Sourcing BERT: State-of-the-Art Pre-training for Natural Language Processing. 2018. Available from: https://blog.research.google/2018/11/open-sourcing-bert-state-of-art-pre.html [accessed Oct 12, 2023]

TABLES

Table 1: Summary of included studies and identified themes

Author	Country	Study Type	Purpose	Setting/Context	LLM(s) Used	Sample Size	Themes
Xu et al (2020) [41]	France	Methodology	Classification of visits into trauma/non-trauma based on ED notes	ED	GPT-2	161,930 notes	CDMS, EWIM
Wang et al (2020) [37]	China	Retrospective cross-sectional study	Sentiment analysis of social media posts related to COVID	Social Media	BERT	999,978 posts	EWIM
Chen et al (2020) [35]	Taiwan	Methodology	Diagnosis identification from discharge summaries	Inpatient	BERT, BioBERT	258,850 discharge diagnoses	EWIM
Chang et al (2020) [43]	USA	Methodology	Categorize free-text ED chief complaints	ED	BERT, ELMo	2.1 million adult and pediatric ED visits	CDMS, EWIM
Wang et al (2021) [32]	Singapore	Methodology	Summarize EMS reports for clinical audits	EMS/Prehospital	BERT	58,898 ambulance incidents	EWIM
Gil-Jardine et al (2021) [36]	France	Methodology	Classify content of EMS calls during COVID	EMS/Prehospital	GPT-2	888,469 calls (training), 39,907 calls (validation), 254,633 calls (application)	EWIM
Shung et al (2021) [29]	USA	Methodology	Identify patients with GI bleeding from ED triage and ROS data	ED	BERT	7,144 cases	CDMS
Tahayori et al (2021) [25]	Australia	Methodology	Predict patient disposition from ED triage notes	ED	BERT	249,532 ED encounters	CDMS, EWIM
Kim et al (2021) [30]	South Korea	Case study	Assign triage severity to simulated cases	ED	BERT	762 cases	CDMS
Wang et al (2021) [12]	China	Methodology	Predict diagnosis and appropriate hospital team from medical record	Prehospital	BERT, Clinical-BERT	198,000 patient records	EWIM
McMaster et al (2021) [38]	Australia	Methodology	Identify adverse drug events from DC summaries	Inpatient	BERT (Clinical-BERT, DeBERTa)	861 discharge summaries	EWIM
Chen et al (2021) [44]	Taiwan	Methodology	Classify EHR data into disease presentations	ED	BERT	1,040,989 ED visits and 305,897 NHAMCS samples	EWIM
Drozdov et al (2021) [15]	UK	Methodology	Generate annotations for CXRs to train model to identify COVID cases	ED	BERT (to generate image annotations)	214,042 CXRs	CDMS
Zhang et al (2022) [34]	China	Methodology	Classify EMS cases into disease categories	EMS/Pre-Hospital	BERT	3500 records	EWIM
Pease et al (2023) [14]	USA	Qualitative investigation	Determine attitudes of clinicians towards using AI in suicide screening	ED	N/A	3 clinicians	CDMS, RET
Chae et al (2023) [23]	USA	Methodology	Predict ED visits and hospitalizations for patients with heart failure	Pre-hospital (home healthcare)	BERT (Bio-Clinical- BERT)	9362 patients	CDMS, RET
Huang et al (2023) [13]	USA	Methodology	Predict non-accidental trauma	ED	BERT	244,326 trajectories (test), 2,077,852 trajectories (validation)	CDMS
Chen et al (2023) [26]	Taiwan	Methodology	Predict critical outcomes from ED data	ED	BERT (comparator)	171,275 ED visits	CDMS
Smith et al (2023) [49]	Australia	Case study	Determine model performance on EM accreditation exam	ED	GPT-3.5, GPT-4, Bard - PaLM, Bard - PaLM 2, Bing	240 questions	CDMS, RET, EC

Gupta et al (2023) [19]	USA	Case study	Determine ability for model to correctly diagnose simulated cases	ED	ChatGPT	20 cases	CDMS, RET, EC
Abavisani et al (2023) [56]	Iran	Commentary	Potential uses for model in emergency surgery	Emergency Surgery	ChatGPT	N/A	CDMS, RET
Rahman et al (2023) [33]	USA	Methodology	Identify cases and patterns in unstructured EMS data	EMS/Pre-Hospital	BERT (BioBERT, ClinicalBERT)	40,000 EMS Narratives	EWIM
Lam et al (2023) [10]	China	Case study	Evaluate model response to lay questions regarding stroke	General Public	ChatGPT	3 questions	EC
Bushuven et al (2023) [9]	Germany	Case study	Use of model to advise parents during pediatric emergencies	General Public	Chat GPT, GPT-4	22 cases	CDMS, RET, EC
Ahn (2023) [11]	South Korea	Case study	Use of model to provide lay-person instruction in CPR	General Public	ChatGPT	3 questions	RET, EC
Preiksaitis et al (2023) [31]	USA	Commentary	Potential limitations to using models for clinical charting	General Medicine	ChatGPT	N/A	EWIM, RET
Barash et al (2023) [20]	Israel	Case study	Use of model to aid radiology referral in the ED	ED	GPT-4	40 cases	CDMS, RET
El Dahdah et al (2023) [50]	USA	Case study	Use of model to triage based on chief complaints	ED	ChatGPT	30 questions	CDMS, RET
Gottlieb et al (2023) [51]	USA	Commentary	Discuss advantages/disadvantages to using model in research	ED/Research	ChatGPT	N/A	RET, EC
Babl et al (2023) [52]	Australia	Case study	Determine the ability of model to generate a scientific abstract	Research	ChatGPT	1 abstract	RET, EC
Chen et al (2023) [57]	China	Methodology	Use model to study the functioning of online self-organizations	Social Media	BERT	47,173 users	EWIM
Bradshaw (2023) [39]	USA	Case study	Determine the ability of model to generate discharge instructions	ED	ChatGPT	1 set of discharge instructions	EWIM, EC
Cheng et al (2023) [16]	China	Commentary	Potential uses for model in surgical management	ED	ChatGPT	N/A	CDMS, EWIM
Rao et al (2023) [17]	USA	Case study	Test model performance in several clinical scenarios	General Medicine	ChatGPT	36 Clinical Vignette	EWIM, EC
Brown et al (2023) [18]	Jersey	Case report/Commentary	Discuss possible model uses in supporting decision making/clinical care	ED	ChatGPT	1 case	CDMS, EWIM, RET, EC
Bhattaram et al (2023) [27]	India	Case study	Ability of model to triage clinical scenarios	ED	ChatGPT	5 scenarios	CDMS, RET, EC
Webb (2023) [47]	USA	Case study	Ability of model to be used as a communication skill trainer	ED	ChatGPT-3.5	1 case	RET, EC
Hamed et al (2023) [22]	Qatar	Case study	Ability of model to synthesize clinical practice guidelines for DKA	General Medicine	ChatGPT	3 guidelines	EWIM, RET
Altamimi et al (2023) [21]	Saudi Arabia	Case study	Ability of model to recommend management in snakebites	ED	ChatGPT	9 questions	CDMS, RET
Gebrael et al (2023) [24]	USA	Case study	Predict the disposition of patients with metastatic prostate cancer based on ED documentation	ED	ChatGPT-4	56 patients	CDMS, EWIM, RET
Sarbay et al (2023) [28]	Turkey	Case study	Use of model for patient triage using clinical scenarios	ED	ChatGPT	50 case scenarios	CDMS, EWIM, RET

Okada et al (2023) [45]	Singapore	Commentary	Discuss possible applications for model in resuscitation	ED/ICU	GPT-3, GPT-4	N/A	CDMS, EWIM, RET

CDMS: Clinical Decision Making & Support

EWIM: Efficiency, Workflow & Information Management

RET: Risks, Ethics & Transparency

EC: Education & Communication

Table 2: Major themes identified, associated subthemes, and representative quotations.

Theme 1: Clinical Decision Making and Support	
Subtheme	**Representative Quotation**
Prediction	"Machine-learning and natural language processing can be together applied to the ED triage note to predict patient disposition with a high level of accuracy." [25]
Treatment recommendations	"An under-explored use of AI in medicine is predicting and synthesizing patient diagnoses, treatment plans, and outcomes." [17]
Symptom checking/self-triage	"To our knowledge, this is the first work to investigate the capabilities of ChatGPT and GPT-4 on PALS core cases in the hypothetical scenario that laypersons would use the chatbot for support until EMS arrive." [9]
Classification	"In this proof-of-concept study, we demonstrated the process of developing a reliable NER [named-entity recognition] model that could reliably identify clinical entities from unlabeled paramedic free text reports." [32]
Triage	"...this preliminary study showed the potential of developing an automatic classification system that directly classifies the KTAS [triage] level and symptoms from the conversations between patients and clinicians." [30]
Screening	"We showed that PABLO, a pretrained, domain-adapted outcome forecasting model, can be used to predict both first and recurrent instances of NAT [non-accidental trauma]." [13]
Differential diagnosis building	"These results suggest that ChatGPT has a high level of accuracy in predicting top differential diagnoses in simulated medical cases." [19]
Decision support	"...ChatGPT-4 demonstrates encouraging results as a support tool in the ED. LLMs such as ChatGPT-4 can facilitate appropriate imaging examination selection and improve radiology referral quality." [20]
Clinical augmentation	"AI can serve as an adjunct in clinical decision making throughout the entire clinical workflow, from triage to diagnosis to

	management.” [17]
Theme 2: Efficiency, Workflow, and Information Management	
Unstructured data extraction	“The proposed model will provide a method to further extract the unstructured free-text portions in EHRs to obtain an abundance of health data. As we enter the forefront of the artificial intelligence era, NLP deep-learning models are well under development. In our model, all medical free-text data can be transformed into meaningful embeddings, which will enhance medical studies and strengthen doctors’ capabilities.” [35]
Charting efficiency	“While notes have become more structured and burdensome, the field of data science has rapidly advanced. With such powerful tools available, it seems reasonable to explore their use to automate seemingly mundane tasks such as writing clinical notes. Generative AI models like ChatGPT could be developed to populate notes for patients based on massive amounts of data contained in current EHRs.” [31]
Summarization/Synthesis	“Although ChatGPT demonstrates the potential for the synthesis of clinical guidelines, the presence of multiple recurrent errors and inconsistencies underscores the need for expert human intervention and validation” [22]
Pattern identification	“This embedding system can be used as a disease retrieval model, which encodes queries and finds the most relevant patients and diseases. In the retrieval demonstration, the query subject was a 53-year-old female patient who suffered from abdominal pain in the upper right quarter to right flanks for 3 days and noticed dizziness and tarry stool on the day of the interview. Through the retrieval, we obtained the five most similar patients with similar symptoms that were possibly related to different diseases.” [44]
Workflow efficiency	“Integration of LLMs with existing EHR (with appropriate regulations) could facilitate improved patient outcomes and workflow efficiency.” [17]
Theme 3: Risks, Ethics, and Transparency	
Oversight	“Generally speaking, the Ethics Guideline for Trustworthy AI suggested seven key requirements

	including human agency and oversight, technical robustness and safety, privacy and data governance, transparency, diversity, nondiscrimination and fairness, environmental and societal well-being, and accountability.” [45]
Fairness	“[Use of LLMs] could also increase equity by assisting researchers with disabilities such as dyslexia.” [51]
Ethical and legal responsibilities	“Legal and ethical implications are associated with using AI in clinical practice, particularly regarding privacy and informed consent issues.” [18]
Reliance on input data	“...data quality can affect the performance of LLMs and NLP techniques applied to the task of extracting and summarizing clinical guidelines.” [22]
Overreliance	“Overreliance on AI systems and the assumption that they are infallible or less fallible than human judgment–automation bias–can lead to errors.” [18]
Explainability/Transparency	“Creating a clinician-interpretable risk prediction model is essential for clinical adoption and implementation of models because it builds trust in decisionmakers, enables error identification and correction in the model, and facilitates integration into clinical workflows.” [23]
Bias propagation	“A risk of bias is possible if the initial training data is not representative of the study population. There is a possibility of compounding of bias and error, leading to incorrect assessment.” [27]
Human bias reduction	“AI tools can offer a near real-time interpretation of medical imaging and clinical decision support and may identify latent patterns that may not be evident to clinicians. While humans are prone to cognitive biases, such as prejudice or fatigue, which can hinder their decision-making process, AI can mitigate these biases and improve accuracy in patient care.” [18]
Accuracy	“LLMs may not be exposed to the broader range of literature (particularly if studies are located behind paywalls), which may limit the comprehensiveness or accuracy of the data.” [51]
Theme 4: Education and Communication	

Clinician education	“While LLM performance in medical examinations may initially seem to be little more than a novelty, their ability to generate coherent and well-explained content hints at other potential uses. As a medical education tool they could potentially help generate practice questions, design mock examinations or provide additional explanations for complex concepts.” [49]
Communication	“Although in its infancy, AI chatbot use has the potential to disrupt how we teach medical students and graduate medical residents communication skills in outpatient and hospital settings.” [47]
Content generation	“ChatGPT or similar programmes, with careful review of the product by authors, may become a valuable scientific writing tool.” [52]
Research assistance	“Conversational AI has some clear benefits and disadvantages. As the technology further evolves, it is incumbent on the scientific community to determine how best to incorporate LLMs into the research and publication process with attention to scientific integrity, adherence to ethical principles, and existing copyright laws.” [51]

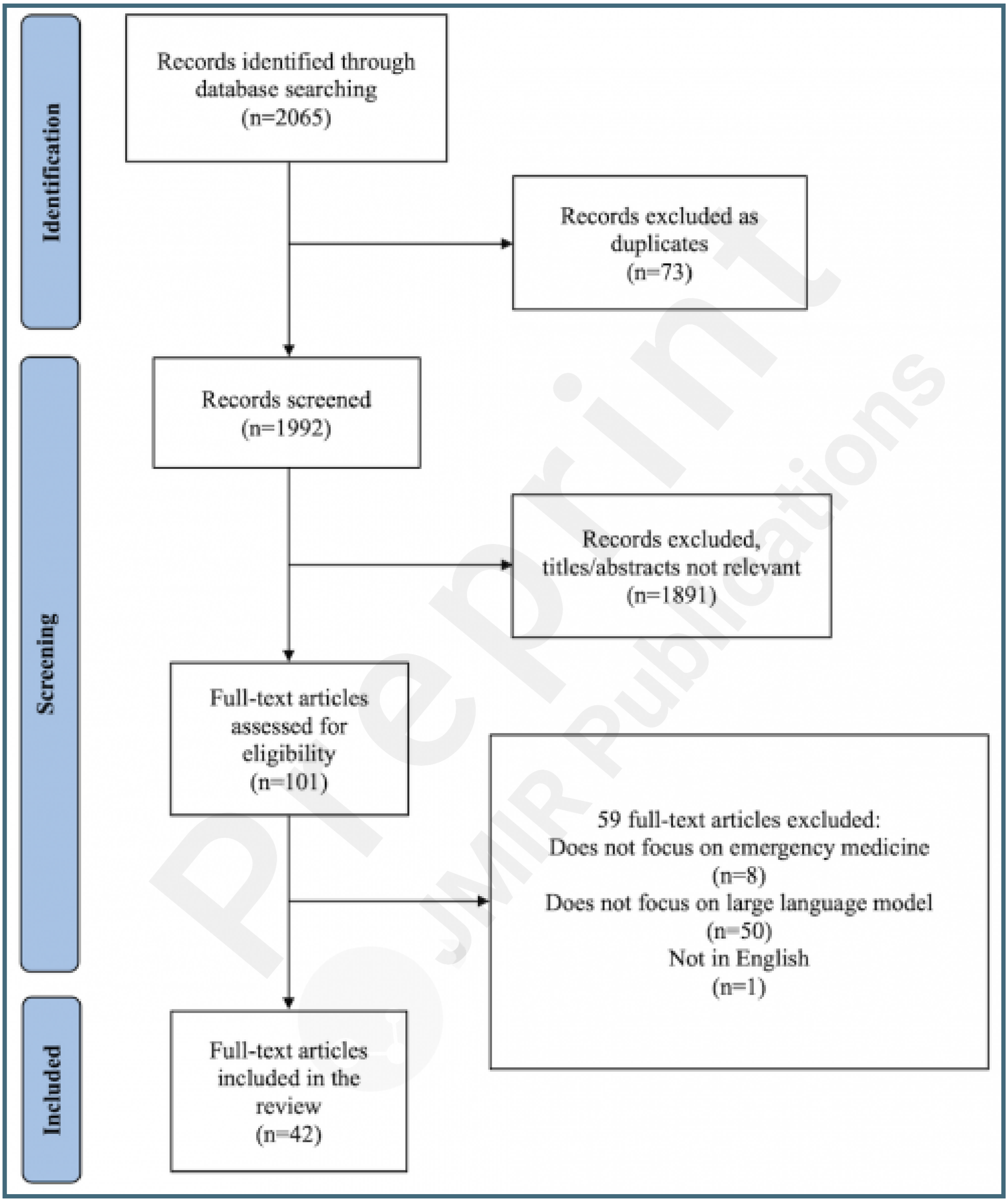
Table 3: Large language models reported in the identified literature.

Model	**Interface**	**Model Size (Parameters)**	**Developer**	**Year Released**
GPT-3.5-turbo	ChatGPT	175B [58]	OpenAI	2022
GPT-4	ChatGPT	~1.8T (estimated) [59]	OpenAI	2023
PaLM	Bard	540B [60]	Google AI	2023
ELMo	Full Model Available	93.6B [61]	Allen Institute for AI	2018
BERT	Full Model Available	110M/340M [62]	Google	2018

Supplementary Files

Figures

PRISMA (Preferred Reporting Items in Systematic Reviews and Meta-Analyses) flow diagram of search and screening for large language models in emergency medicine.


Identification
Records identified through database searching (n=2065)
Records excluded as duplicates (n=73)
Screening
Records screened (n=1992)
Records excluded, titles/abstracts not relevant (n=1891)
Full-text articles assessed for eligibility (n=101)
59 full-text articles excluded:
Does not focus on emergency medicine (n=8)
Does not focus on large language model (n=50)
Not in English (n=1)
Included
Full-text articles included in the review (n=42)

Multimedia Appendixes

Literature review search strategy.
URL: http://asset.jmir.pub/assets/96392fe3fa3d79f92fe3bfe413d0af6b.docx

CONSORT (or other) checklists

PRISMA-ScR checklist.
URL: http://asset.jmir.pub/assets/9f023be752cbebda49aef53a27412434.pdf