Burns & Trauma, 2026, 14, tkag026
https://doi.org/10.1093/burnst/tkag026
Published: 23 March 2026
Review

OXFORD

Large language models in emergency and critical care medicine: a comprehensive review of applications, challenges, and future directions

Jie Yang[1,‡], Suibi Yang[1,‡], Ziyao Shao[2,‡], Tianqi Chen[1], Hongjie Shen[1], Pengmin Zhou[1], Boming Xia[1], Xiong Lei[1], Lihui Wang[3], Dong Xue[2], Shaojiang Zheng[4,*], Yuetian Yu[3,*], Zhongheng Zhang[1,5,6,7,*]

[1]Department of Emergency Medicine, Sir Run Run Shaw Hospital, Zhejiang University School of Medicine, 3 Qingchun East Road, Shangcheng District, Hangzhou 310016, China
[2]Key Laboratory of Smart Manufacturing in Energy Chemical Process, Ministry of Education, East China University of Science and Technology, 130 Meilong Road, Xuhui District, Shanghai 200237, China
[3]Department of Critical Care Medicine, Renji Hospital, Shanghai Jiaotong University School of Medicine, 160 Pujian Road, Pudong New District, Shanghai 200001, China
[4]Key Laboratory of Emergency and Trauma of Ministry of Education, Engineering Research Center for Hainan Biological Sample Resources of Major Diseases, the First Affiliated Hospital, Hainan Medical University, 31 Longhua Road, Longhua District, Haikou 570102, China
[5]Key Laboratory of Precision Medicine in Diagnosis and Monitoring Research of Zhejiang Province, Sir Run Run Shaw Hospital, Zhejiang University School of Medicine, 3 Qingchun East Road, Shangcheng District, Hangzhou 310016, China
[6]School of Medicine, Shaoxing University, 508 Huancheng West Road, Yuecheng District, Shaoxing 312000, China
[7]Longquan Industrial Innovation Research Institute, No. 2 Industrial Road, Longyuan Street, Longquan City, Lishui 323799, China

*Corresponding authors. Zhongheng Zhang, E-mail: zh_zhang1984@zju.edu.cn; Yuetian Yu, E-mail: fishyyt@sina.com; Shaojiang Zheng, E-mail: zshaojiang@muhn.edu.cn
‡Jie Yang, Suibi Yang, and Ziyao Shao contributed equally.

Abstract

Emergency and critical care medicine requires the rapid synthesis of heterogeneous clinical data under extreme time constraints. Early artificial intelligence tools lacked the flexibility to manage real-world patient heterogeneity. Large language models (LLMs) offer a paradigm shift by demonstrating advanced natural language understanding, cross-task generalization, and context-sensitive reasoning, thereby bridging the gap between fragmented algorithms and holistic clinical decision support. The effective deployment of these models is grounded in four methodological pillars: domain adaptation, knowledge integration, multimodal and temporal modeling, and transparency. Domain adaptation and knowledge integration specifically empower the context-sensitive reasoning required for high-stakes intensive care. This theoretical framework enables their application across clinical decision support, documentation optimization, medical education, and clinical research. Integrating continuous physiological waveforms with multi-omics data facilitates dynamic risk stratification for complex conditions like sepsis, while natural language-to-structured query language capabilities accelerate clinical data extraction and quality improvement. The transition of LLMs from experimental settings to routine clinical deployment remains constrained by model hallucinations, multimodal integration barriers, and unresolved ethical governance. Sustainable implementation requires a human-in-the-loop copilot design, rigorous multicenter prospective validation, and transparent regulatory frameworks. Addressing these challenges is essential to ensure that technological innovations safely translate into measurable improvements in patient survival and clinical outcomes.

Keywords Large language models, Emergency and critical care medicine, Clinical decision support, Multimodal integration, Artificial intelligence, Precision medicine

Highlights

- LLMs offer a paradigm shift in emergency and critical care medicine (ECCM) by providing context-sensitive reasoning and cross-task generalization to manage heterogeneous, time-constrained patient data.
- The effective deployment of LLMs in high-acuity clinical environments is grounded in four methodological pillars: domain adaptation, knowledge integration, multimodal and temporal modeling, and transparency.

Received: 6 September 2025. **Revised:** 3 March 2026. **Accepted:** 17 March 2026
© The Author(s) 2026. Published by Oxford University Press
This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial License (https://creativecommons.org/licenses/by-nc/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact reprints@oup.com for reprints and translation rights for reprints. All other permissions can be obtained through our RightsLink service via the Permissions link on the article page on our site—for further information please contact journals.permissions@oup.com.

Downloaded from academic.oup.com/burnstrauma/article/doi/10.1093/burnst/tkag026/8537608 by guest on 12 October 2026

- Current LLM applications in ECCM are clustered into four primary domains: clinical decision support, clinical documentation and administrative workflow optimization, medical education, and clinical research.
- Integrating multimodal and time-dependent data—such as continuous physiological waveforms and multi-omics profiles—enables LLMs to transition from retrospective documentation tools into proactive, dynamic decision-support systems.
- Transitioning LLMs to routine clinical use requires overcoming model hallucinations and ethical hurdles through a human-in-the-loop copilot design, rigorous multicenter validation, and transparent regulatory frameworks.

Background

Recent advances in deep learning have expanded the applications of large language models (LLMs) in medicine [1–3], with growing interest in critical care [4–11]. Emergency and critical care medicine (ECCM) requires managing extreme time constraints, high clinical acuity, and precise risk mitigation [12–16]. Within intensive care units (ICUs) and emergency departments (EDs), clinicians rapidly synthesize heterogeneous, dynamically evolving datasets, including laboratory results [17, 18], imaging studies [19–21], continuous physiological waveforms [22, 23], and unstructured text [24–26]. These data directly guide interventions for life-threatening conditions such as early sepsis recognition [27, 28], ventilator parameter adjustment in acute respiratory distress syndrome (ARDS) [29, 30], and trauma triage [31, 32]. Diagnostic delays or errors in these scenarios cause irreversible patient harm, establishing the demand for timely, integrative clinical reasoning as a longstanding driver of artificial intelligence (AI) applications in acute care [33–35].

Early AI generations in ECCM proved ill-suited for real-world clinical complexities [36]. Rule-based decision-support systems could execute formulaic workflows but lacked the flexibility to account for patient heterogeneity [37, 38]. Subsequent task-specific predictive models, such as sepsis early warning algorithms [39–42], improved performance in narrow use cases yet remained largely siloed [43, 44].

To address these limitations, LLMs bridge the gap between fragmented AI tools and holistic clinical decision support. Pre-trained on extensive biomedical corpora encompassing scientific literature, clinical practice guidelines, structured electronic health records (EHRs), ICU/ED documentation, and text-associated multimodal data, LLMs exhibit core capabilities critical to ECCM. They demonstrate advanced natural language understanding that enables robust summarization [45], clinical reasoning [11], and question-answering [10]. They also support cross-task generalization, allowing a single system to manage diverse workflows from triage and diagnosis to risk prediction [46]. Crucially, they perform context-sensitive reasoning aligned with real-world clinical practice patterns [47, 48]. Domain adaptation strategies further refine model interpretation of specialized terminology [49–52], and retrieval-augmented generation (RAG) grounds outputs in evidence-based guidelines, mitigating hallucinations risks [53, 54].

LLMs also facilitate multimodal and temporal data integration. Beyond text processing, advanced architectures increasingly incorporate numerical data such as laboratory trajectories, image-derived features from chest radiographs and computed tomography, continuous physiological signals, and omics data like transcriptomic profiles in sepsis [55–57]. This integration builds dynamic, patient-centric representations, transitioning LLMs from retrospective documentation tools to prospective decision-support systems. Early applications span triage outcome prediction, anticipatory alerts for patient deterioration, automated real-time clinical documentation, patient communication support, medical education simulation, and accelerated cohort identification for clinical research (Figure 1). Collectively, these functions address the ECCM requirements for timely decision-making and comprehensive patient assessment [58].

Critical challenges remain. Genuine multimodal fusion is largely experimental [59]. Current interpretability and transparency levels cannot support full clinical accountability, and ethical dilemmas remain unresolved [60–63]. Regulatory frameworks ensuring the safe, validated, and equitable deployment of LLMs in ECCM are still in early developmental stages [64–66].

While existing reviews examined LLM applications across various medical specialties, they lack systematic evaluation of the unique ECCM environment. Acute care requires continuous multimodal physiological monitoring and high-stakes clinical decisions for rapidly evolving patient conditions. This review provides a comprehensive synthesis of LLM applications specifically designed for the ECCM context.

Review

Literature search strategy

We conducted a comprehensive literature search to identify studies on LLM applications in ECCM. PubMed, Web of Science, Embase, and Google Scholar databases were searched from January 2023 through December 2025 using combinations of the following terms: “large language model”, “LLM”, “GPT”, “transformer”, “artificial intelligence”, “natural language processing”, “emergency medicine”, “emergency department”, “critical care”, “intensive care unit”, “ICU”, “triage”, “clinical decision support”, “sepsis”, “ARDS”, “trauma”, and “resuscitation”.

Inclusion criteria were defined as follows: first, studies should evaluate LLM applications (including models based on transformer architectures) in emergency or critical care settings; second, publications must appear in peer-reviewed journals; third, studies must provide empirical data or systematic analysis relevant to ECCM practice. Studies employing transformer-based architectures that partially fulfill the relevant application purposes were also considered. Exclusion criteria included editorials and commentaries without original data, conference abstracts without full-text availability, and studies focusing exclusively on specialties outside ECCM. Reference lists of included articles were manually screened to identify additional relevant publications.

This search strategy identified a substantial number of potentially relevant articles. Following application of the inclusion

Downloaded from academic.oup.com/burnstrauma/article/doi/10.1093/burnst/tkag026/8537608 by guest on 12 October 2026

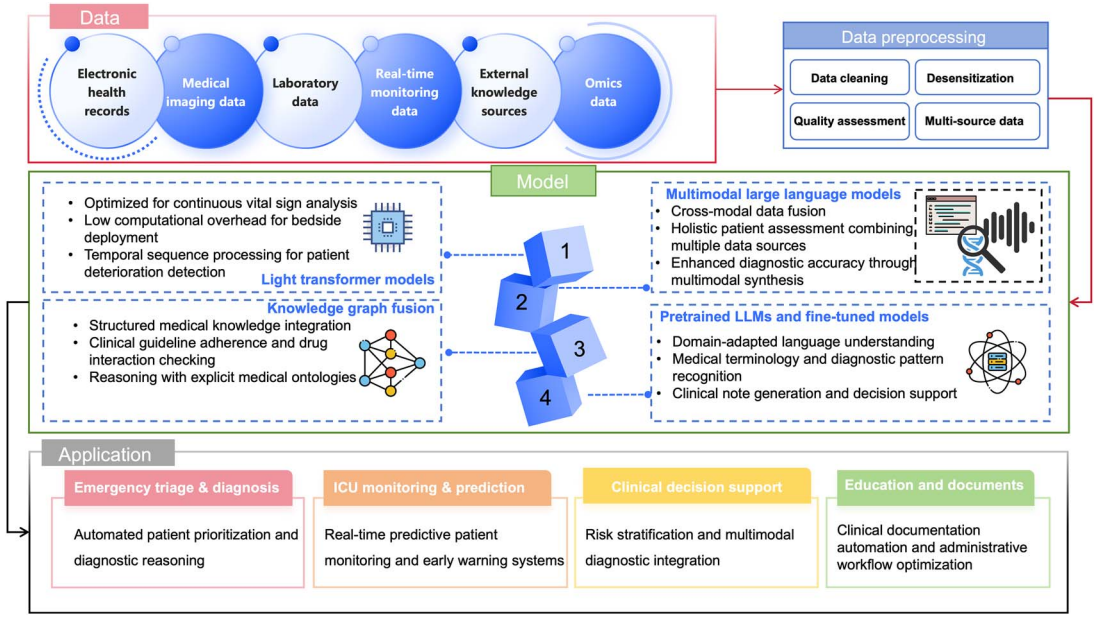


Figure 1 Illustration of a comprehensive framework of large language model (LLM) applications in emergency and critical care, covering the entire workflow from diverse data sources that undergo data preprocessing. It then extends to model development with core components: domain adaptation, knowledge integration, multimodal and temporal modeling, and transparency and interpretability. Additionally, the framework outlines various application scenarios in emergency and critical care: clinical decision support, multimodal integration, documentation and administration, education and training, and research and quality improvement

and exclusion criteria, primary studies meeting all criteria were selected and are summarized in Table 1. The remaining articles informed the background, theoretical framework, and discussion sections of this review.

Theoretical framework of AI architectures in critical care

The deployment of LLMs in emergency and critical care requires a theoretical framework that links advanced computational architectures with the unique demands of high-acuity clinical environments. Unlike routine contexts, critical care involves heterogeneous data streams, rapidly evolving physiologic changes, and urgent decisions that must be timely and evidence based. Drawing upon and synthesizing insights from prior research in medical AI, clinical informatics, and healthcare natural language processing, we identify four methodological pillars that are particularly critical for ECCM applications: domain adaptation, knowledge integration, multimodal and temporal modeling, and transparency and interpretability [67]. These pillars are not novel constructs but rather represent a synthesis of established concepts, reframed and prioritized for the specific demands of acute care environments. Table 2 provides definitions and clinical relevance of each pillar for readers less familiar with LLM terminology. Collectively, these pillars establish a foundation for the safe and effective application of LLMs in critical care. The unique requirements of ECCM that motivate this framework are detailed in Table 3, while accessible definitions of each pillar are provided in Table 2.

Domain adaptation

LLMs trained on general-purpose corpora demonstrate strong linguistic fluency but often lack precision when applied to the specialized terminology, abbreviations, and clinical reasoning conventions of intensive care practice [68, 69]. Domain adaptation through both fine-tuning and instruction tuning is therefore essential. Fine-tuning on ICU and ED notes enables the model to recognize domain-specific terminology while adjusting contextual outputs to critically ill patient populations. Instruction tuning [70] further aligns model behavior with the structured query–response style of clinicians who expect concise, guideline-concordant recommendations. Incorporating structured indices such as Sequential Organ Failure Assessment (SOFA) or Acute Physiology and Chronic Health Evaluation II scores enhances task specificity, enabling reliable risk stratification and prioritization of care. Thus, domain adaptation provides the linguistic accuracy and contextual relevance required for LLMs to function effectively in high-stakes clinical discourse [71–73].

Knowledge integration

While LLMs can generate clinically plausible text, their tendency toward factual errors necessitates grounding outputs in authoritative knowledge sources [74–77]. Knowledge integration can be achieved through several complementary strategies. RAG connects the model to curated repositories [78] including clinical guidelines, institutional protocols, and current literature, thereby ensuring evidence-based recommendations. Embedding structured logic derived from pathways such as advanced life support algorithms or sepsis bundles further constrains model outputs to align with standardized practice. Hybrid symbolic–neural architectures [79–81] combine probabilistic reasoning with explicit clinical rules, such as contraindications to nephrotoxic agents in patients with renal dysfunction, thereby enhancing both flexibility and reliability. By merging tacit statistical inference with codified medical knowledge, LLMs can engage clinicians fluently while adhering to established standards of care [46, 82].

Downloaded from academic.oup.com/burnstrauma/article/doi/10.1093/burnst/tkag026/8537608 by guest on 12 October 2026

Table 1 Summary of studies on large language model (LLM) applications in emergency and critical care medicine

	Author	Title	Application domains	Study focus
1	Liu J, *et al*. 2025	Comparing text-based clinical risk prediction in critical care: a note-specific hierarchical network and large language models	Clinical decision support	Risk stratification (text data-driven clinical risk prediction in critical care)
2	Wu X, *et al*. 2025	A large language model improves clinicians' diagnostic performance in complex critical illness cases	Clinical decision support	Diagnostic support (enhancing clinicians' diagnostic accuracy in complex critical illness)
3	Shi B, *et al*. 2025	Large language models and artificial neural networks for assessing 1-year mortality in patients with myocardial infarction: analysis from the Medical Information Mart for Intensive Care IV (MIMIC-IV) Database	Clinical decision support	Risk stratification (1-year mortality prediction in myocardial infarction patients)
4	Lorenzoni G, *et al*. 2024	Use of a large language model to identify and classify injuries with free-text emergency department data	Clinical decision support	Emergency triage (injury identification and classification via emergency department free-text data)
5	Lafuente C, *et al*. 2025	Evaluating large language models on hospital health data for automated emergency triage	Clinical decision support	Emergency triage (automated emergency triage using hospital health data)
6	Gao H, *et al*. 2025	A large language model based pipeline for extracting information from patient complaint and anamnesis in clinical notes for severity assessment	Clinical decision support	Emergency triage (severity assessment via patient complaint/anamnesis extraction from clinical notes)
7	Goh E, *et al*. 2024	Large language model influence on diagnostic reasoning: a randomized clinical trial	Clinical decision support	Diagnostic support (assessing large language model (LLM) impact on clinician diagnostic reasoning)
8	Fonseca J, *et al*. 2023	Mortality prediction using medical time series on TBI patients	Clinical decision support	Risk stratification (mortality prediction in traumatic brain injury (TBI) patients)
9	Dai L, *et al*. 2024	A clinically actionable and explainable real-time risk assessment framework for stroke-associated pneumonia	Clinical decision support	Risk stratification (real-time risk assessment for stroke-associated pneumonia)
10	Giesa N, *et al*. 2024	Applying a transformer architecture to intraoperative temporal dynamics improves the prediction of postoperative delirium	Clinical decision support	Risk stratification (postoperative delirium prediction using intraoperative temporal dynamics)
11	Sundrani S, *et al*. 2023	Predicting patient decompensation from continuous physiologic monitoring in the emergency department	Clinical decision support	Risk stratification (patient decompensation prediction via continuous physiologic monitoring in emergency department (ED))
12	Suñol F, *et al*. 2025	Leveraging large language models for patient-ventilator asynchrony detection	Clinical decision support	Therapeutic guidance (patient–ventilator asynchrony detection)
13	Antonie NI, *et al*. 2025	The role of ChatGPT and AI chatbots in optimizing antibiotic therapy: a comprehensive narrative review	Clinical decision support	Therapeutic guidance (optimizing antibiotic therapy via artificial intelligence (AI) chatbots)
14	De Vito A, *et al*. 2025	Comparing large language models for antibiotic prescribing in different clinical scenarios: which performs better?	Clinical decision support	Therapeutic guidance (comparing LLM performance in antibiotic prescribing)
15	Wang K, *et al*. 2025	Leveraging large language models for preoperative prevention of cardiopulmonary bypass-associated acute kidney injury	Clinical decision support	Therapeutic guidance (preoperative prediction/prevention of cardiopulmonary bypass (CPB)–associated acute kidney injury (AKI))
16	Sheikh MS, *et al*. 2024	Personalized medicine transformed: ChatGPT's contribution to continuous renal replacement therapy alarm management in intensive care units	Clinical decision support	Therapeutic guidance (continuous renal replacement therapy (CRRT) alarm troubleshooting in intensive care)
17	Williams CYK, *et al*. 2025	Physician- and large language model-generated hospital discharge summaries	Documentation and administration	Documentation automation (hospital discharge summary generation)

(Continued)

Downloaded from academic.oup.com/burnstrauma/article/doi/10.1093/burnst/tkag026/8537608 by guest on 12 October 2026

Table 1 Continued

	Author	Title	Application domains	Study focus
18	Hartman V, *et al*. 2024	Developing and evaluating large language model-generated emergency medicine handoff notes	Documentation and administration	Documentation automation (emergency medicine handoff note generation)
19	Wang H, *et al*. 2024	DRG-LLaMA: tuning LLaMA model to predict diagnosis-related group for hospitalized patients	Documentation and administration	Administrative efficiency (diagnosis-related group (DRG) classification for hospitalized patients)
20	Wan P, *et al*. 2024	Outpatient reception via collaboration between nurses and a large language model: a randomized controlled trial	Documentation and administration	Patient-facing communication (nurse–LLM collaborative outpatient reception)
21	Liu S, *et al*. 2024	Using large language model to guide patients to create efficient and comprehensive clinical care message	Documentation and administration	Patient-facing communication (guiding patients to generate clinical care messages)
22	Madden MG, *et al*. 2023	Assessing the usefulness of a large language model to query and summarize unstructured medical notes in intensive care	Documentation and administration	Documentation automation (querying/summarizing unstructured intensive care unit (ICU) medical notes)
23	Sumner J, *et al*. 2025	Perspectives and experiences with large language models in health care: survey study	Education and training	Healthcare professional training (surveying LLM application experiences in healthcare education)
24	Qiang S, *et al*. 2025	Application of large language models in stroke rehabilitation health education: 2-phase study	Education and training	Patient/professional education (stroke rehabilitation health education via LLM)
25	Li Z, *et al*. 2025	Empowering standardized residency training in China through large language models: problem analysis and solutions	Education and training	Residency training (enhancing standardized residency training via LLM)
26	Ko A, *et al*. 2025	Artificial intelligence-guided bronchoscopy is superior to human expert instruction for the performance of critical-care physicians: a randomized controlled trial	Education and training	Procedural training (AI-guided bronchoscopy operation training for critical-care physicians)
27	Cook DA, *et al*. 2025	Virtual patients using large language models: scalable, contextualized simulation of clinician-patient dialogue with feedback	Education and training	Virtual teaching (scalable clinician–patient dialogue simulation via LLM-powered virtual patients)
28	Mahajan A, *et al*. 2025	Performance of trauma-trained large language models on surgical assessment questions: a new approach in resource identification	Education and training	Knowledge assessment (trauma-related surgical knowledge assessment via specialized LLM)
29	Lee H, 2024	The rise of ChatGPT: exploring its potential in medical education	Education and training	Medical education (exploring ChatGPT's potential in medical education)
30	Cheng Y, *et al*. 2025	A review of ChatGPT in medical education: exploring advantages and limitations	Education and training	Medical education (reviewing ChatGPT's advantages/limitations in medical education)
31	Lee S, *et al*. 2025	CXR-LLaVA: a multimodal large language model for interpreting chest X-ray images	Multimodal integration	Medical imaging (chest X-ray interpretation via multimodal LLM)
32	Chung P, *et al*. 2024	Large language model capabilities in perioperative risk prediction and prognostication	Multimodal integration	Perioperative care (perioperative risk prediction/prognostication via LLM)
33	Theodoris CV, *et al*. 2023	Transfer learning enables predictions in network biology	Multimodal integration	Molecular data integration (network biology prediction via transfer learning)
34	Kwok AJ, *et al*. 2023	Neutrophils and emergency granulopoiesis drive immune suppression and an extreme response endotype during sepsis	Multimodal integration	Molecular data integration (sepsis endotype analysis via immune cell biology)
35	Cao C, *et al*. 2025	Development of prompt templates for large language model-driven screening in systematic reviews	Research and quality improvement	Research support (systematic review screening via LLM prompt templates)
36	Zihang C, *et al*. 2025	Enhanced pre-recruitment framework for clinical trial questionnaires through the integration of large language models and knowledge graphs	Research and quality improvement	Research support (clinical trial pre-recruitment via LLM–knowledge graph integration)

This table synthesizes 36 studies on LLM applications in emergency and critical care medicine (ECCM) across five core domains: clinical decision support, documentation and administration, education and training, multimodal integration, and research and quality improvement. Each entry includes study identifier, author/year, title, target domain, and ECCM-specific focus to illustrate LLM research scope in ECCM

Downloaded from academic.oup.com/burnstrauma/article/doi/10.1093/burnst/tkag026/8537608 by guest on 12 October 2026

Table 2 Definitions and clinical relevance of the four methodological pillars for large language model (LLM) deployment in emergency and critical care medicine (ECCM)

Methodological pillar	Definition	Clinical relevance in ECCM	Representative techniques
Domain adaptation	The process of adjusting a general-purpose LLM to perform optimally within a specific clinical domain	Enables recognition of intensive care unit (ICU)/emergency department (ED)–specific terminology, abbreviations, and clinical reasoning patterns	Fine-tuning on clinical notes, instruction tuning, reinforcement learning from human feedback (RLHF)
Knowledge integration	Grounding LLM outputs in external authoritative knowledge sources	Ensures that recommendations align with current clinical guidelines, reduces hallucination risk	Retrieval-augmented generation (RAG), knowledge graph embedding, hybrid symbolic–neural architectures
Multimodal and temporal modeling	Integration of diverse data types (text, images, signals, laboratory values) with temporal sequence modeling	Enables synthesis of chest X-ray findings with ventilator waveforms, laboratory trends, and clinical notes for holistic patient assessment	Vision-language models, temporal transformers, cross-modal attention mechanisms
Transparency and interpretability	The capacity of a model to explain its reasoning process in terms understandable to clinicians	Supports clinical accountability, facilitates error detection, enables appropriate trust calibration	Chain-of-thought prompting, attention visualization, uncertainty quantification, rationale generation

This table provides accessible explanations for readers unfamiliar with technical LLM terminology

Table 3 Emergency and critical care medicine (ECCM)–specific requirements and corresponding large language model (LLM) capabilities

ECCM-specific requirements	Distinction from general medical applications	Required LLM capabilities
Real-time decision support	Response latency must be seconds, not minutes	Low-latency inference, streaming output, edge deployment optimization
Continuous physiological signal integration	Must process electrocardiogram (ECG), arterial waveforms, ventilator data in real time	Temporal sequence modeling, signal-to-text alignment, multimodal fusion
High-stakes early warning	False negatives may result in patient death; false positives cause alert fatigue	Uncertainty quantification, high-sensitivity thresholds, calibrated confidence scores
Rapid clinical deterioration	Patient states evolve within minutes to hours	Dynamic context updating, trajectory prediction, state-transition modeling
Multidisciplinary coordination	Information must be synthesized across multiple specialty teams	Cross-domain knowledge integration, structured handoff generation
Resource-constrained environments	Intensive care units (ICUs)/emergency departments (EDs) often operate at capacity limits	Prioritization algorithms, resource allocation support

This table contrasts the unique demands of emergency and critical care medicine with general medical applications and identifies the LLM architectural features necessary to address each requirement

Multimodal and temporal modeling

The ICU generates vast and rapidly evolving data, including structured laboratory values, continuous physiologic waveforms, unstructured documentation, and bedside imaging [83–87]. To capture this complexity, LLM architectures must evolve beyond text-only paradigms toward multimodal and temporally aware frameworks. Multimodal integration enables concurrent interpretation of textual data, numeric values, temporal laboratory trajectories, and imaging reports, producing a holistic representation of patient status. Temporal modeling, incorporating sequential dependencies and dynamic embeddings, supports prediction of disease trajectories such as impending hemodynamic collapse or respiratory failure hours before onset [88–92]. By synthesizing multimodal and time-dependent signals, LLMs can advance from retrospective documentation tools to proactive clinical decision-support systems that anticipate risk and inform timely interventions [93–95].

Transparency and interpretability

In critical care, where diagnostic and therapeutic decisions carry substantial clinical implications, trust in AI tools depends on their ability to explain reasoning processes. Transparency and interpretability are therefore ethical and practical prerequisites [96]. Beyond providing conclusions, LLMs should deliver explicit, evidence-based rationales that reference pertinent laboratory abnormalities, vital sign trends, or clinical guidelines [97–99]. Techniques such as chain-of-thought reasoning and structured rationale generation can expose these decision pathways to clinicians [100]. Equally important is uncertainty quantification: models must communicate the degree of confidence in

Downloaded from academic.oup.com/burnstrauma/article/doi/10.1093/burnst/tkag026/8537608 by guest on 12 October 2026

predictions and defer to clinician judgment when evidence is insufficient [101]. Prioritizing clarity and evidential grounding not only promotes clinician trust but also addresses regulatory and medico-legal expectations for safe and accountable AI deployment in critical care.

The application of LLMs in ECCM

Current LLM applications in ECCM cluster into four primary domains: clinical decision support, clinical documentation and administrative workflow optimization, medical education and training, and clinical research and quality improvement. Multimodal data integration serves as a cross-cutting methodology enhancing capabilities across all domains rather than a standalone category (Table 1).

Clinical decision support

LLMs function as dynamic aids to frontline decision-making, across triage, diagnosis, risk stratification, therapeutic planning, and multimodal data synthesis [102]. Tolerance for latency varies drastically among these tasks. Triage and sepsis prediction typically allow response times measured in minutes to hours, supported by feasibility studies [42, 103]. Conversely, real-time cardiac arrest detection and ventilator asynchrony identification demand responses within seconds or milliseconds to remain clinically meaningful, as delays directly impact patient outcomes [104, 105]. This distinction dictates LLM architecture choices, such as cloud-based versus edge deployment, determining successful real-world implementation.

Triage and risk stratification Emergency triage is a highly viable LLM application due to its reliance on the rapid synthesis of unstructured clinical information. A systematic evaluation found that Generative Pre-trained Transformer (GPT)-4 achieved 74.7% accuracy in assigning emergency severity index levels, compared to 85.2% for trained nurses [106]. While not matching expert performance, LLMs consistently processed free-text chief complaints and generated recommendations within seconds. Models trained on free-text records successfully predict admission needs and injury classifications across multilingual datasets, supporting equitable triage [107–109]. Hybrid approaches combining LLM reasoning with structured machine learning classifiers outperformed LLM-only systems with area under the curve (AUC) of 0.89 versus 0.82 [108]. In paired comparisons, GPT-4 correctly identified the more acute patient in 88.4% of cases, demonstrating utility during ED surge conditions [110]. Implementation studies demonstrate clinical translation. An LLM-based decision-support system reduced triage-to-physician time by 23% and improved undertriage detection by 15% [111].

Diagnostic support Diagnostic support exhibits contextual variability. A randomized trial found no significant improvement in clinician reasoning when assisted by LLMs despite the model outperforming clinicians on standalone accuracy benchmarks [112]. Conversely, specialized models optimized for high-complexity ICU data, such as DeepSeek-R1, halved response times and enhanced resident diagnostic accuracy [46]. In ED settings, GPT-4 achieved 87.1% diagnostic accuracy compared to residents' 78.3%, demonstrating strong performance in identifying rare conditions [113]. An LLM system developed for syncope recognition achieved 91.2% sensitivity and 88.7% specificity in differentiating cardiac from non-cardiac etiologies [114]. These findings indicate that while generalized diagnostic replacement remains premature, LLMs tailored to intensive care datasets provide tangible value in complex clinical scenarios.

Therapeutic guidance Specific workflows, including antibiotic recommendations, ventilatory support, and continuous renal replacement therapy (CRRT) counseling, benefit from LLM decision aids, though human oversight remains vital to mitigate factual errors [104, 115]. Successful implementations function as copilot models, framing clinical decisions with synthesized evidence while leaving final diagnostic authority for the treating clinician. For antibiotic recommendations, LLMs like ChatGPT-o1 achieve 71.7% accuracy in appropriate drug selection but struggle with difficult-to-treat microbes [116, 117]. A multi-agent LLM system integrating specialized agents achieved 89.4% concordance with infectious disease specialist recommendations for empiric antibiotic selection in sepsis [118]. Regarding organ support therapy, LLMs showed poor direct classification of flow starvation asynchronies (accuracy $\leq$62.7%) but excelled at generating effective neural network code achieving 90.2% accuracy [104]. Multimodal LLM models predict cardiopulmonary bypass-associated acute kidney injury with an AUC of 0.92 [119], and ChatGPT-4 aids CRRT alarm troubleshooting with 90%–94% accuracy [120]. In advanced cardiac life support, LLMs achieved 91.4% accuracy in algorithm selection but provided inconsistent drug dosing recommendations (78.3% accuracy) [121].

Risk stratification and outcome prediction LLMs demonstrate strong risk stratification capabilities through multimodal data integration. The COMPOSER-LLM system achieved an AUC of 0.92 for 6-h sepsis prediction, reducing the time to antibiotic administration by 32% [103]. LLM-generated syndromic classifications from admission notes of 63 704 sepsis patients predicted mortality more accurately than traditional SOFA scores (AUC 0.79 *vs* 0.71) [122]. Transformer architectures robustly predict ARDS, delirium, and stroke-associated complications, often generating warnings days before clinical recognition [118, 123, 124]. Applying transformer architecture to intraoperative temporal dynamics achieved an AUC of 0.84 for postoperative delirium prediction, alerting clinicians 48 h before diagnosis [105]. Domain-specific models outperform general-purpose LLMs in mortality prediction (AUC 0.86 *vs* 0.74) [48]. For cardiac arrest, GPT-4 predicted outcomes with an AUC of 0.81, comparable to validated clinical scores but possessing the advantage of incorporating unstructured notes [125]. Automatic detection of in-hospital cardiac arrest from clinical notes achieved 94.2% sensitivity and 91.8% specificity across five hospitals. Multimodal approaches consistently outperform unimodal systems, with models integrating continuous physiologic monitoring achieving an AUC of 0.91 versus 0.78 for patient decompensation prediction [112].

Multimodal data integration as an enabling technology Multimodal LLMs advance imaging and physiological signal interpretation. The chest X-ray (CXR)–large language and vision assistant (LLaVA) achieved human-comparable lesion detection across 377 110 chest radiographs with 87.3% diagnostic accuracy,

Downloaded from academic.oup.com/burnstrauma/article/doi/10.1093/burnst/tkag026/8537608 by guest on 12 October 2026

outperforming general-purpose LLMs (71.2%) [118, 126]. Electrocardiogram (ECG)–LM achieved 89.7% accuracy in arrhythmia classification while enabling natural language interaction for ECG interpretation. GPT-4 classified ED imaging report actionability with 93.8% accuracy [125]. At the molecular level, foundational models like Geneformer establish frameworks for integrating transcriptomic data with clinical phenotypes [127]. Neutrophil transcriptomic profiles identify sepsis endotypes with 30-day mortality exceeding 40% [125].

Documentation and administrative workflow optimization

ECCM presents unique documentation challenges suited for LLM applications. These applications encompass real-time trauma resuscitation documentation that captures verbal orders and intervention timestamps, critical care transport handoff notes synthesizing patient status during inter-facility transfers, automated code documentation generating structured records of cardiac arrest events, and shift-to-shift ICU handoffs summarizing overnight events for incoming teams. Documentation automation is a low-risk, high-yield initial application. AI-generated summaries achieve concision and structural quality comparable to human-authored reports. However, insidious errors require human verification, such as "temporal swapping," where the model confuses a patient's medical history with their current presentation [45, 128]. Tools like ICU-GPT improve data accessibility through natural-language querying of hospital databases, lowering technical barriers for clinician-directed analytics [25]. Administrative applications demonstrate efficiency gains. Diagnosis-related group (DRG)–large language model meta AI (LLaMA) successfully classifies billing groups with acceptable accuracy, streamlining reimbursement and resource allocation [129]. Automated pipelines accelerate systematic review screening by up to 90%, reducing workload and expanding clinician involvement in scholarship [127]. Patient-facing communication is an emerging domain. Nurse–LLM hybrid workflows improve patient satisfaction and response efficiency compared with nurse-only systems [130]. Comprehensive LLM Artificial Intelligence Responder (CLAIR), a patient communication assistant, outperformed GPT-4 in relevance and appropriateness of responses to portal messages [131]. Because these applications benefit greatly from human verification, they represent "early adopter" use cases for safe clinical deployment [132].

Medical education and training

Recent empirical studies demonstrate the efficacy of LLMs in ECCM education, primarily through interactive clinical simulations and personalized theoretical tutoring. In clinical training, LLM-driven platforms generate high-fidelity, dynamically evolving patient scenarios that serve as scalable alternatives to standardized patients and physical mannequins. LLMs function as procedural simulators, personalized tutors, and objective evaluators—each addressing key gaps in traditional training paradigms [133–135]. For procedural training, AI-guided bronchoscopy systems enable faster completion and fewer errors than expert-led instruction [136]. Virtual patients powered by LLMs have consistently been rated as authentic, valuable, and cost-effective alternatives to standardized patients [137]. Domain-specific fine-tuning enhances knowledge assessment. For example, Eastern Association of the Surgery of Trauma (EAST)–GPT outperformed GPT-4 in multiple evaluation dimensions, demonstrating the advantage of aligning model training with curricular content [138]. Similarly, LLM-based virtual assistants deliver adaptive, case-based learning while fostering interactive engagement [139]. LLM applications in education represent a low-risk, high-impact domain. They complement traditional mentorship models by handling routine tasks, freeing expert clinicians to focus on high-value interactions, such as nuanced clinical reasoning discussions and professional development mentoring.

Research and quality improvement

LLMs address the longstanding gap between clinicians' medical expertise and limited coding proficiency. Extracting structured EHR data—such as laboratory results, diagnosis codes, or treatment timelines—traditionally required information technology (IT) teams to write structured query language (SQL) queries, delaying studies. LLMs enable natural language-to-SQL conversion: clinicians can simply describe data needs in plain language, and the model generates syntactically correct, clinically relevant SQL queries. This capability eliminates intermediate technical support, placing direct data access into researchers' hands [136, 140]. For quality improvement, automated surveillance systems using LLMs identify near-miss events and documentation errors more effectively than manual chart review, converting retrospective audits into proactive patient-safety monitoring [137, 141]. Accelerating research workflows and embedding real-time error detection aligns directly with institutional goals of advancing patient care.

Challenges and strategic responses

Translating LLMs from experimental applications to routine ECCM deployment faces significant challenges across multiple dimensions. For each challenge, we identify corresponding strategic responses that may enable safe and effective implementation (Table 4).

Reliability and accuracy: domain-specific fine-tuning and validation

While domain-specific models show strong accuracy in targeted tasks [136], general-purpose LLMs remain prone to factual errors and hallucinations [4, 46]. Even when outperforming clinicians in vignette-based diagnostics, their real-world performance is inconsistent. These flaws amplify in high-stakes ECCM settings: false positives risk alert fatigue, while false negatives may delay recognition of clinical deterioration, potentially resulting in patient death [137, 138]. Beyond discrimination metrics like AUC, it is essential to evaluate model calibration, as the tendency of LLMs to be overconfident can result in probability estimates that do not accurately reflect true event rates, thereby compromising their clinical utility. Future validation studies must report calibration metrics, including the Expected Calibration Error, alongside discrimination metrics to ensure that predicted risk scores are reliable for clinical decision-making [142, 143]. Progress demands developing critical-care-specific LLMs trained on ICU and ED datasets encompassing structured variables and clinical outcomes. Models like ICU-GPT demonstrate the performance advantages of domain-specific tuning. Future efforts should establish standardized ECCM benchmarks, scale

Downloaded from academic.oup.com/burnstrauma/article/doi/10.1093/burnst/tkag026/8537608 by guest on 12 October 2026

Table 4 Challenges and strategic responses for large language model implementation in emergency and critical care medicine

Challenge domain	Key issues	Strategic responses
Reliability and accuracy	Hallucination, factual errors, inconsistent real-world performance; false-positive/negative risks	Domain-specific fine-tuning; task-focused benchmarks; prospective multicenter validation
Data integration	Partial multimodal integration; limited physiologic signal and omics incorporation; siloed outputs	Multimodal architectures integrating imaging, waveforms, and omics data; precision medicine approaches
Ethical, legal, and governance	Unclear accountability for artificial intelligence (AI) errors; privacy concerns; regulatory gaps	Explainability features; medico-legal frameworks; regulatory partnerships
Workflow and human factors	AI "silo" outputs; automation bias or disuse; workflow disruption	Human-in-the-loop copilot design; physician verification; intuitive interfaces
Clinician preparedness	Insufficient AI literacy; risk of misunderstanding or misapplying outputs	AI literacy in medical education; simulation-based training curricula

development through federated learning approaches to preserve patient privacy, and mandate prospective multicenter validation before clinical deployment.

Data integration barriers: multimodal and multi-omics architectures

ECCM requires synthesizing highly heterogeneous data streams—structured EHR data, physiologic waveforms, imaging, and omics layers. While multimodal models have emerged, integration remains partial. Most systems merge text and imaging, but seamless incorporation of genomics, proteomics, and metabolomics data remains experimental. Without comprehensive interoperability, models underutilize available patient information [144]. The next generation of clinical LLMs must advance beyond text–image fusion to integrate multi-omics data, including transcriptomics, proteomics, metabolomics, and genomics. Such modeling would enable phenotype-guided rather than one-size-fits-all approaches to patient management, aligning directly with precision medicine paradigms and potentially unlocking novel treatment pathways for syndromes such as sepsis and ARDS [145].

Ethical, legal, and governance issues: transparent governance and regulation

Using LLMs in direct patient care raises critical questions about accountability, privacy, and informed consent. Clear legal frameworks assigning responsibility for misdiagnoses or adverse outcomes arising from LLM-assisted decision support are absent. Data-sharing regulations—particularly regarding genomic, psychiatric, or other sensitive information—pose barriers to adoption and complicate governance [146, 147]. The deployment of LLMs in life-critical ECCM settings raises profound ethical and medico-legal challenges that extend beyond routine clinical AI applications [148]. Moreover, LLMs in such settings require exceptionally rigorous validation—extending beyond benchmark accuracy to robustness across diverse populations and edge cases—yet standardized pathways akin to pharmaceutical trials are lacking. Because errors in ECCM cause irreversible harm rapidly, structured protocols for error disclosure, adverse-event reporting, and continuous monitoring are necessary alongside technical safeguards. To mitigate these risks, a multilayered safeguard framework is proposed, including prospective multicenter trials, human verification for critical decisions, real-time performance monitoring with degradation alerts, established incident reporting, and transparent communication with patients about AI's role in their care. Models must incorporate explainability features—such as highlighting laboratory trends, physiological patterns, or guideline references—that make predictions interpretable to clinicians. Parallel efforts are required to build regulatory structures that address validation standards, liability assignment, and documentation requirements. Academic–industry–regulatory partnerships will be vital to ensure that adoption proceeds with safety, accountability, and equity at the forefront [149].

Workflow and human factors: human-in-the-loop copilot design

LLM effectiveness depends heavily on clinician acceptance and integration into existing workflows [150, 151]. Current systems often create "AI silos," detaching model outputs from bedside tools. Poorly designed interfaces may lead clinicians to either over-rely on outputs or disregard them entirely, undermining both safety and potential benefits. Maintaining human oversight is essential to safeguard reliability and clinical trust. Effective workflows present AI outputs for clinician verification through "copilot" interfaces that summarize evidence while ensuring that final diagnostic authority remains with clinicians. This hybrid design mitigates hallucination risks, preserves efficiency gains, and maintains appropriate human accountability for clinical decisions [101].

Clinician preparedness: education and cultural adoption

Clinician preparedness remains a critical bottleneck, primarily rooted in insufficient AI literacy and inadequate targeted training among frontline providers. This knowledge gap often hinders their ability to comprehend LLM outputs, recognize inherent limitations such as reliance on training data biases and inability to account for rare clinical nuances, and understand contextual constraints like mismatches between model inputs and real-time patient conditions. In turn, it elevates the risk of misunderstanding the validity of LLM recommendations or misapplying outputs to clinical decision. Sustainable adoption requires cultivating

Downloaded from academic.oup.com/burnstrauma/article/doi/10.1093/burnst/tkag026/8537608 by guest on 12 October 2026

an "AI-literate" clinical workforce. Simulation-based training should serve as both a testing ground for LLM applications and a medium to teach clinicians the strengths, weaknesses, and ethical implications of AI. Embedding AI literacy in medical school curricula, residency training programs, and continuing medical education will facilitate sustainable adoption and enhance acceptance across disciplines [136, 137, 139].

Conclusions

LLMs offer transformative potential for critical care medicine through advanced diagnostic support, risk prediction, and multimodal integration. However, their clinical translation is currently hindered by technical limitations such as persistent hallucinations and interpretability deficits, alongside broader regulatory and equity concerns. Notably, the paradox wherein highly capable LLMs fail to consistently enhance human clinician performance underscores the immense complexity of integrating AI into high-acuity workflows. Moving forward, the field must prioritize rigorous scientific validation through multicenter randomized controlled trials and adhere strictly to AI-specific reporting guidelines, such as TRIPOD+AI and TRIPOD-LLM. By combining multidisciplinary collaboration with robust regulatory frameworks, LLMs can be safely harnessed to improve patient outcomes, optimize healthcare efficiency, and ensure medical equity.

Abbreviations

MIMIC-IV: Medical Information Mart for Intensive Care IV; LLM: Large language model; TBI: Traumatic brain injury; ED: Emergency department; AI: Artificial intelligence; CPB: Cardiopulmonary bypass; AKI: Acute kidney injury; CRRT: Continuous renal replacement therapy; DRG: Diagnosis-related group; LLaMA: Large Language Model Meta AI; CLAIR: Comprehensive LLM Artificial Intelligence Responder; ICU: Intensive care unit; ECCM: Emergency and critical care medicine; RLHF: Reinforcement learning from human feedback; RAG: Retrieval-augmented generation; ECE: Expected Calibration Error; ESI: Emergency Severity Index; IT: Information technology; GPT: Generative Pre-trained Transformer; SQL: Structured query language; TRIPOD-LLM: Transparent Reporting of a Multivariable Model for Individual Prognosis or Diagnosis using Large Language Models; CXR: Chest X-ray; LLaVA: Large Language and Vision Assistant; ECG: Electrocardiogram; EAST: Eastern Association of the Surgery of Trauma

Author contributions

Jie Yang (Writing—original draft [equal]), Suibi Yang (Writing—original draft [equal]), Ziyao Shao (Writing—original draft [equal]), Tianqi Chen (Visualization [equal]), Hongjie Shen (Investigation [equal]), Pengmin Zhou (Investigation [equal]), Boming Xia (Investigation [equal]), Xiong Lei (Data curation [equal]), Lihui Wang (Data curation [equal]), Dong Xue (Conceptualization [equal], Supervision [equal]), Shaojiang Zheng (Writing—review & editing [lead]), Yuetian Yu (Conceptualization [equal], Supervision [equal]), and Zhongheng Zhang (Funding acquisition [lead], Resources [lead], Supervision [lead])

Conflicts of interest

None declared.

Funding

This study was supported by funding from the Prevention and Control of Emerging and Major Infectious Diseases–National Science and Technology Major Project (No. 2025ZD01902500, No. 2025ZD01902501, Z.Z.), the China National Key Research and Development Program (No. 2023YFC3603104, Z.Z.), the National Natural Science Foundation of China (82272180, 82472243, Z.Z.), the Fundamental Research Funds for the Central Universities (226-2025-00024, Z.Z.), the Huadong Medicine Joint Funds of the Zhejiang Provincial Natural Science Foundation of China under Grant No. LHDMD24H150001 (Z.Z.), the Key Research & Development Project of Zhejiang Province (2024C03240, Z.Z.), a collaborative scientific project co-established by the Science and Technology Department of the National Administration of Traditional Chinese Medicine and the Zhejiang Provincial Administration of Traditional Chinese Medicine (GZY-ZJ-KJ-24082, Z.Z.), General Health Science and Technology Program of Zhejiang Province (2024KY1099, Z.Z.), the Project of Zhejiang University Longquan Innovation Center (ZJDXLQCXZCJBGS2024016, Z.Z.), the Beijing Municipal Natural Science Foundation (No. 7252298, Z.Z.), Wu Jieping Medical Foundation Special Research Grant (320.6750.2024-23-07, Z.Z.), and the Zhejiang Provincial Science and Technology Program for Disease Control and Prevention (2026JKZ042, Z.Z.).

References

1. Eppler M, Ganjavi C, Ramacciotti LS, Piazza P, Rodler S, Checcucci E, *et al.* Awareness and use of ChatGPT and large language models: a prospective cross-sectional global survey in urology. *Eur Urol* 2024;85:146–53. https://doi.org/10.1016/j.eururo.2023.10.014.
2. Clusmann J, Kolbinger FR, Muti HS, Carrero ZI, Eckardt J-N, Laleh NG, *et al.* The future landscape of large language models in medicine. *Commun Med* 2023;3:141. https://doi.org/10.1038/s43856-023-00370-1.
3. Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language models in medicine. *Nat Med* 2023;29:1930–40. https://doi.org/10.1038/s41591-023-02448-8.
4. Bedi S, Liu Y, Orr-Ewing L, Dash D, Koyejo S, Callahan A, *et al.* Testing and evaluation of health care applications of large language models: a systematic review. *JAMA.* 2025;333:319–28. https://doi.org/10.1001/jama.2024.21700.
5. Bradshaw TJ, Tie X, Warner J, Hu J, Li Q, Li X. Large language models and large multimodal models in medical imaging: a primer for physicians. *J Nucl Med* 2025;66:173–82. https://doi.org/10.2967/jnumed.124.268072.
6. Iqbal U, Tanweer A, Rahmanti AR, Greenfield D, Lee LT-J, Li Y-CJ. Impact of large language model (ChatGPT) in healthcare: an umbrella review and evidence synthesis. *J Biomed Sci* 2025;32:45. https://doi.org/10.1186/s12929-025-01131-z.
7. Johri S, Jeong J, Tran BA, Schlessinger DI, Wongvibulsin S, Barnes LA, *et al.* An evaluation framework for clinical use of large language models in patient interaction tasks.

Downloaded from academic.oup.com/burnstrauma/article/doi/10.1093/burnst/tkag026/8537608 by guest on 12 October 2026

Nat Med 2025;31:77–86. https://doi.org/10.1038/s41591-024-03328-5.
8. Luo X, Rechardt A, Sun G, Nejad KK, Yáñez F, Yilmaz B, *et al.* Large language models surpass human experts in predicting neuroscience results. *Nat Hum Behav* 2025;9:305–15. https://doi.org/10.1038/s41562-024-02046-9.
9. McDuff D, Schaekermann M, Tu T, Palepu A, Wang A, Garrison J, *et al.* Towards accurate differential diagnosis with large language models. *Nature* 2025;642:451–7. https://doi.org/10.1038/s41586-025-08869-4.
10. Singhal K, Tu T, Gottweis J, Sayres R, Wulczyn E, Amin M, *et al.* Toward expert-level medical question answering with large language models. *Nat Med* 2025;31:943–50. https://doi.org/10.1038/s41591-024-03423-7.
11. Tordjman M, Liu Z, Yuce M, Fauveau V, Mei Y, Hadjadj J, *et al.* Comparative benchmarking of the DeepSeek large language model on medical tasks and clinical reasoning. *Nat Med* 2025;31:2550–5. https://doi.org/10.1038/s41591-025-03726-3.
12. Spencer SA, Adipa FE, Baker T, Crawford AM, Dark P, Dula D, *et al.* A health systems approach to critical care delivery in low-resource settings: a narrative review. *Intensive Care Med* 2023;49:772–84. https://doi.org/10.1007/s00134-023-07136-2.
13. Munroe ES, Spicer A, Castellvi-Font A, Zalucky A, Dianti J, Graham Linck E, *et al.* Evidence-based personalised medicine in critical care: a framework for quantifying and applying individualised treatment effects in patients who are critically ill. *Lancet Respir Med* 2025;13:556–68. https://doi.org/10.1016/S2213-2600(25)00054-2.
14. Maslove DM, Tang B, Shankar-Hari M, Lawler PR, Angus DC, Baillie JK, *et al.* Redefining critical illness. *Nat Med* 2022;28:1141–8. https://doi.org/10.1038/s41591-022-01843-x.
15. Bates DW, Syrowatka A. Harnessing AI in sepsis care. *Nat Med* 2022;28:1351–2. https://doi.org/10.1038/s41591-022-01878-0.
16. Bruno RR, Wolff G, Wernly B, Masyuk M, Piayda K, Leaver S, *et al.* Virtual and augmented reality in critical care medicine: the patient's, clinician's, and researcher's perspective. *Crit Care (London, England)* 2022;26:326. https://doi.org/10.1186/s13054-022-04202-x.
17. Rubulotta F, Bahrami S, Marshall DC, Komorowski M. Machine learning tools for acute respiratory distress syndrome detection and prediction. *Crit Care Med* 2024;52:1768–80. https://doi.org/10.1097/CCM.0000000000006390.
18. Douville NJ, Douville CB, Mentz G, Mathis MR, Pancaro C, Tremper KK, *et al.* Clinically applicable approach for predicting mechanical ventilation in patients with COVID-19. *Br J Anaesth* 2021;126:578–89. https://doi.org/10.1016/j.bja.2020.11.034.
19. Meyer H-J, Dermendzhiev T, Hetz M, Osterhoff G, Kleber C, Denecke T, *et al.* Body composition parameters in initial CT imaging of mechanically ventilated trauma patients: single-centre observational study. *J Cachexia Sarcopenia Muscle* 2024;15:2437–46. https://doi.org/10.1002/jcsm.13578.
20. Tsakok MT, Watson RA, Saujani SJ, Kong M, Xie C, Peschl H, *et al.* Reduction in chest CT severity and improved hospital outcomes in SARS-CoV-2 omicron compared with Delta variant infection. *Radiology.* 2023;306:261–9. https://doi.org/10.1148/radiol.220533.
21. Wischmeyer PE, Bear DE, Berger MM, De Waele E, Gunst J, McClave SA, *et al.* Personalized nutrition therapy in critical care: 10 expert recommendations. *Crit Care (London, England)* 2023;27:261. https://doi.org/10.1186/s13054-023-04539-x.
22. Park C, Han C, Jang SK, Kim H, Kim S, Kang BH, *et al.* Development and validation of a machine learning model for early prediction of delirium in intensive care units using continuous physiological data: retrospective study. *J Med Internet Res* 2025;27:e59520. https://doi.org/10.2196/59520.
23. Goodwin AJ, Dixon W, Mazwi M, Hahn CD, Meir T, Goodfellow SD, *et al.* The truth hertz-synchronization of electroencephalogram signals with physiological waveforms recorded in an intensive care unit. *Physiol Meas* 2023;44:085002. https://doi.org/10.1088/1361-6579/ace49e.
24. Cobert J, Mills H, Lee A, Gologorskaya O, Espejo E, Jeon SY, *et al.* Measuring implicit bias in ICU notes using word-embedding neural network models. *Chest* 2024;165:1481–90. https://doi.org/10.1016/j.chest.2023.12.031.
25. Liu J, Nguyen A, Capurro D, Verspoor K. Comparing text-based clinical risk prediction in critical care: a note-specific hierarchical network and large language models. *IEEE J Biomed Health Inform* 2025;29:7657–67. https://doi.org/10.1109/JBHI.2025.3574254.
26. Idan D, Einav S. Primer on large language models: an educational overview for intensivists. *Crit Care (London, England)* 2025;29:238. https://doi.org/10.1186/s13054-025-05479-4.
27. Catenacci V, Sheikh F, Patel K, Fox-Robichaud AE. The prognostic utility of protein C as a biomarker for adult sepsis: a systematic review and meta-analysis. *Crit Care (London, England)* 2022;26:21. https://doi.org/10.1186/s13054-022-03889-2.
28. Adami M-E, Eugen-Olsen J, Giamarellos-Bourboulis EJ. Early sepsis recognition: how difficult can this be? *Crit Care (London, England)* 2024;28:190. https://doi.org/10.1186/s13054-024-04975-3.
29. Barbeta E, Arrieta M, Motos A, Bobi J, Yang H, Yang M, *et al.* A long-lasting porcine model of ARDS caused by pneumonia and ventilator-induced lung injury. *Crit Care (London, England)* 2023;27:239. https://doi.org/10.1186/s13054-023-04512-8.
30. Prkachin Y. The reign of the ventilator: acute respiratory distress syndrome, COVID-19, and technological imperatives in intensive care. *Ann Intern Med* 2021;174:1145–50. https://doi.org/10.7326/M21-0270.
31. El-Menyar A, Naduvilekandy M, Asim M, Rizoli S, Al-Thani H. Machine learning models predict triage levels, massive transfusion protocol activation, and mortality in trauma utilizing patients hemodynamics on admission. *Comput Biol Med* 2024;179:108880. https://doi.org/10.1016/j.compbiomed.2024.108880.
32. Weidman AC, Malakouti S, Salcido DD, Zikmund C, Patel R, Weiss LS, *et al.* A machine learning trauma triage model for critical care transport. *JAMA Netw Open* 2025;8:e259639. https://doi.org/10.1001/jamanetworkopen.2025.9639.

Downloaded from academic.oup.com/burnstrauma/article/doi/10.1093/burnst/tkag026/8537608 by guest on 12 October 2026

33. Bates DW, Levine D, Syrowatka A, Kuznetsova M, Craig KJT, Rui A, *et al.* The potential of artificial intelligence to improve patient safety: a scoping review. *NPJ Digit Med* 2021;4:54. https://doi.org/10.1038/s41746-021-00423-6.
34. Bhasuran B, Schmolly K, Kapoor Y, Jayakumar NL, Doan R, Amin J, *et al.* Reducing diagnostic delays in acute hepatic porphyria using health records data and machine learning. *J Am Med Inform Assoc* 2025;32:63–70. https://doi.org/10.1093/jamia/ocae141.
35. Li L, Peng M, Zou Y, Li Y, Qiao P. The promise and limitations of artificial intelligence in CTPA-based pulmonary embolism detection. *Front Med* 2025;12:1514931. https://doi.org/10.3389/fmed.2025.1514931.
36. Wu Y, Ding X, Wang Y, Ouyang D. Harnessing the power of machine learning into tissue engineering: current progress and future prospects. *Burns Trauma* 2024;12:tkae053. https://doi.org/10.1093/burnst/tkae053.
37. Moreno R, Rhodes A, Piquilloud L, Hernandez G, Takala J, Gershengorn HB, *et al.* The Sequential Organ Failure Assessment (SOFA) Score: has the time come for an update? *Crit Care (London, England)* 2023;27:15. https://doi.org/10.1186/s13054-022-04290-9.
38. Martin-Loeches I, Singer M, Leone M. Sepsis: key insights, future directions, and immediate goals. A review and expert opinion. *Intensive Care Med* 2024;50:2043–9. https://doi.org/10.1007/s00134-024-07694-z.
39. Papareddy P, Lobo TJ, Holub M, Bouma H, Maca J, Strodthoff N, *et al.* Transforming sepsis management: AI-driven innovations in early detection and tailored therapies. *Crit Care (London, England)* 2025;29:366. https://doi.org/10.1186/s13054-025-05588-0.
40. Goh KH, Wang L, Yeow AYK, Poh H, Li K, Yeow JJL, *et al.* Artificial intelligence in sepsis early prediction and diagnosis using unstructured data in healthcare. *Nat Commun* 2021;12:711. https://doi.org/10.1038/s41467-021-20910-4.
41. Jiang S, Dai S, Li Y, Zhou X, Jiang C, Tian C, *et al.* Development and validation of a screening tool for sepsis without laboratory results in the emergency department: a machine learning study. *EClinicalMedicine* 2025;80:103048. https://doi.org/10.1016/j.eclinm.2024.103048.
42. Adams R, Henry KE, Sridharan A, Soleimani H, Zhan A, Rawat N, *et al.* Prospective, multi-site study of patient outcomes after implementation of the TREWS machine learning-based early warning system for sepsis. *Nat Med* 2022;28:1455–60. https://doi.org/10.1038/s41591-022-01894-0.
43. Zhang J, Xie Y, Pang G, Liao Z, Verjans J, Li W, *et al.* Viral pneumonia screening on chest X-rays using confidence-aware anomaly detection. *IEEE Trans Med Imaging* 2021;40:879–90. https://doi.org/10.1109/TMI.2020.3040950.
44. Signoroni A, Savardi M, Benini S, Adami N, Leonardi R, Gibellini P, *et al.* BS-net: learning COVID-19 pneumonia severity on a large chest X-ray dataset. *Med Image Anal* 2021;71:102046. https://doi.org/10.1016/j.media.2021.102046.
45. Williams CYK, Subramanian CR, Ali SS, Apolinario M, Askin E, Barish P, *et al.* Physician- and large language model-generated hospital discharge summaries. *JAMA Intern Med* 2025;185:818–25. https://doi.org/10.1001/jamainternmed.2025.0821.
46. Wu X, Huang Y, He Q. A large language model improves clinicians' diagnostic performance in complex critical illness cases. *Crit Care (London, England)* 2025;29:230. https://doi.org/10.1186/s13054-025-05468-7.
47. Poole S, Sisodia N, Koshal K, Henderson K, Wijangco J, Paredes D, *et al.* Detecting new lesions using a large language model: applications in real-world multiple sclerosis datasets. *Ann Neurol* 2025;98:308–16. https://doi.org/10.1002/ana.27251.
48. Rider NL, Li Y, Chin AT, DiGiacomo DV, Dutmer C, Farmer JR, *et al.* Evaluating large language model performance to support the diagnosis and management of patients with primary immune disorders. *J Allergy Clin Immunol* 2025;156:81–7. https://doi.org/10.1016/j.jaci.2025.02.004.
49. Shi B, Chen L, Pang S, Wang Y, Wang S, Li F, *et al.* Large language models and artificial neural networks for assessing 1-year mortality in patients with myocardial infarction: analysis from the Medical Information Mart for Intensive Care IV (MIMIC-IV) database. *J Med Internet Res* 2025;27:e67253. https://doi.org/10.2196/67253.
50. Akbasli IT, Birbilen AZ, Teksam O. Leveraging large language models to mimic domain expert labeling in unstructured text-based electronic healthcare records in non-English languages. *BMC Med Inform Decis Mak* 2025;25:154. https://doi.org/10.1186/s12911-025-02871-6.
51. Sun C, Teichman K, Zhou Y, Critelli B, Nauheim D, Keir G, *et al.* Generative large language models trained for detecting errors in radiology reports. *Radiology* 2025;315:e242575. https://doi.org/10.1148/radiol.242575.
52. Hsieh C, Moreira C, Nobre IB, Sousa SC, Ouyang C, Brereton M, *et al.* DALL-M: context-aware clinical data augmentation with large language models. *Comput Biol Med* 2025;190:110022. https://doi.org/10.1016/j.compbiomed.2025.110022.
53. Koller P, Clement C, van Eijk A, Seifert R, Zhang J, Prenosil G, *et al.* Optimizing theranostics chatbots with context-augmented large language models. *Theranostics* 2025;15:5693–704. https://doi.org/10.7150/thno.107757.
54. Li H, Huang J, Ji M, Yang Y, An R. Use of retrieval-augmented large language model for COVID-19 fact-checking: development and usability study. *J Med Internet Res* 2025;27:e66098. https://doi.org/10.2196/66098.
55. Renc P, Jia Y, Samir AE, Was J, Li Q, Bates DW, *et al.* Zero shot health trajectory prediction using transformer. *NPJ Digit Med* 2024;7:256. https://doi.org/10.1038/s41746-024-01235-0.
56. Zhou S, Guan C, Deng S, Zhu Y, Yang W, Zhang X, *et al.* A novel sequence-based transformer model architecture for integrating multi-omics data in preterm birth risk prediction. *NPJ Digit Med* 2025;8:536. https://doi.org/10.1038/s41746-025-01942-2.
57. Zeng H, Yin C, Chai C, Wang Y, Dai Q, Sun H. Cancer gene identification through integrating causal prompting large language model with omics data-driven causal inference. *Brief Bioinform* 2025;26:bbaf113. https://doi.org/10.1093/bib/bbaf113.
58. Griot M, Hemptinne C, Vanderdonckt J, Yuksel D. Large language models lack essential metacognition for reliable medical reasoning. *Nat Commun* 2025;16:642. https://doi.org/10.1038/s41467-024-55628-6.

Downloaded from academic.oup.com/burnstrauma/article/doi/10.1093/burnst/tkag026/8537608 by guest on 12 October 2026

59. Kang B, Fan R, Cui C, Cui Q. Comprehensive prediction and analysis of human protein essentiality based on a pretrained large language model. *Nat Comput Sci* 2025;5:196–206. https://doi.org/10.1038/s43588-024-00733-1.
60. Alber DA, Yang Z, Alyakin A, Yang E, Rai S, Valliani AA, *et al.* Medical large language models are vulnerable to data-poisoning attacks. *Nat Med* 2025;31:618–26. https://doi.org/10.1038/s41591-024-03445-1.
61. Modi ND, Menz BD, Awaty AA, Alex CA, Logan JM, McKinnon RA, *et al.* Assessing the system-instruction vulnerabilities of large language models to malicious conversion into health disinformation Chatbots. *Ann Intern Med* 2025;178:1172–80. https://doi.org/10.7326/ANNALS-24-03933.
62. Degany O, Laros S, Idan D, Einav S. Evaluating the o1 reasoning large language model for cognitive bias: a vignette study. *Crit Care (London, England)* 2025;29:376. https://doi.org/10.1186/s13054-025-05591-5.
63. Cinquin O. Steering veridical large language model analyses by correcting and enriching generated database queries: first steps toward ChatGPT bioinformatics. *Brief Bioinform* 2024;26:bbaf045. https://doi.org/10.1093/bib/bbaf045.
64. Akinci D'Antonoli T, Tejani AS, Khosravi B, Bluethgen C, Busch F, Bressem KK, *et al.* Cybersecurity threats and mitigation strategies for large language models in health care. *Radiol Artif Intell* 2025;7:e240739. https://doi.org/10.1148/ryai.240739.
65. Chow JCL, Li K. Large language models in medical Chatbots: opportunities, challenges, and the need to address AI risks. *Information* 2025;16:549. https://doi.org/10.3390/info16070549.
66. Siddique S, Chow JCL. Machine learning in healthcare communication. *Encyclopedia* 2021;1:220–39. https://doi.org/10.3390/encyclopedia1010021.
67. Zhou H, Liu F, Gu B, Zou X, Huang J, Wu J, *et al.* A survey of large language models in medicine: progress, application, and challenge. arXiv preprint. 2024. https://doi.org/10.48550/arXiv.2311.05112.
68. Workum JD, Volkers BWS, van de Sande D, Arora S, Goeijenbier M, Gommers D, *et al.* Comparative evaluation and performance of large language models on expert level critical care questions: a benchmark study. *Crit Care (London, England)* 2025;29:72. https://doi.org/10.1186/s13054-025-05302-0.
69. Wiest IC, Bhat M, Clusmann J, Schneider CV, Jiang X, Kather JN. Large language models for clinical decision support in gastroenterology and hepatology. *Nat Rev Gastroenterol Hepatol* 2025;22:773–87. https://doi.org/10.1038/s41575-025-01108-1.
70. Nazar W, Nazar G, Kamińska A, Danilowicz-Szymanowicz L. How to design, create, and evaluate an instruction-tuning dataset for large language model training in health care: tutorial from a clinical perspective. *J Med Internet Res* 2025;27:e70481. https://doi.org/10.2196/70481.
71. Dorfner FJ, Dada A, Busch F, Makowski MR, Han T, Truhn D, *et al.* Evaluating the effectiveness of biomedical fine-tuning for large language models on clinical tasks. *J Am Med Inform Assoc* 2025;32:1015–24. https://doi.org/10.1093/jamia/ocaf045.
72. Yang Q, Chen J, Sun Y, Wang Y, Tan T. Fine-tuning medical language models for enhanced long-contextual understanding and domain expertise. *Quant Imaging Med Surg* 2025;15:5450–62. https://doi.org/10.21037/qims-2024-2655.
73. Chen C, Wang X, Guan M, Yue W, Wu Y, Zhou Y, *et al.* Evaluating and improving syndrome differentiation thinking ability in large language models: method development study. *JMIR Med Inform* 2025;13:e75103. https://doi.org/10.2196/75103.
74. Gu Z, Jia W, Piccardi M, Yu P. Empowering large language models for automated clinical assessment with generation-augmented retrieval and hierarchical chain-of-thought. *Artif Intell Med* 2025;162:103078. https://doi.org/10.1016/j.artmed.2025.103078.
75. Isaradech N, Riedel A, Sirikul W, Kreuzthaler M, Schulz S. Zero- and few-shot named entity recognition and text expansion in medication prescriptions using large language models. *Artif Intell Med* 2025;167:103165. https://doi.org/10.1016/j.artmed.2025.103165.
76. Zhong W, Sun M, Yao S, Liu Y, Peng D, Liu Y, *et al.* Enhancing the accuracy of human phenotype ontology identification: comparative evaluation of multimodal large language models. *J Med Internet Res* 2025;27:e73233. https://doi.org/10.2196/73233.
77. Ngoc Nguyen O, Amin D, Bennett J, Hetlevik Ø, Malik S, Tout A, *et al.* GP or ChatGPT? Ability of large language models (LLMs) to support general practitioners when prescribing antibiotics. *J Antimicrob Chemother* 2025;80:1324–30. https://doi.org/10.1093/jac/dkaf077.
78. Gao Y, Xiong Y, Gao X, Jia K, Pan J, Bi Y, *et al.* Retrieval-augmented generation for large language models: a survey. 2024. https://doi.org/10.48550/arXiv.2312.10997.
79. Zhou Z, Yu X, Magoua JJ, Cui J, Luan H, Lin D. Integrating machine learning and a large language model to construct a domain knowledge graph for reducing the risk of fall-from-height accidents. *Accid Anal Prev* 2025;215:108009. https://doi.org/10.1016/j.aap.2025.108009.
80. Li R, Mao S, Zhu C, Yang Y, Tan C, Li L, *et al.* Enhancing pulmonary disease prediction using large language models with feature summarization and hybrid retrieval-augmented generation: multicenter methodological study based on radiology report. *J Med Internet Res* 2025;27:e72638. https://doi.org/10.2196/72638.
81. Feldman MJ, Hoffer EP, Conley JJ, Chang J, Chung JA, Jernigan MC, *et al.* Dedicated AI expert system vs generative AI with large language model for clinical diagnoses. *JAMA Netw Open* 2025;8:e2512994. https://doi.org/10.1001/jamanetworkopen.2025.12994.
82. Yang H, Li J, Zhang C, Sierra AP, Shen B. Large language model-driven knowledge graph construction in sepsis care using multicenter clinical databases: development and usability study. *J Med Internet Res* 2025;27:e65537. https://doi.org/10.2196/65537.
83. Du H, Zhao Y, Zhao J, Xu S, Lin X, Chen Y, *et al.* Advancing real-time infectious disease forecasting using large language models. *Nat Comput Sci* 2025;5:467–80. https://doi.org/10.1038/s43588-025-00798-6.
84. Hou B, Mukherjee P, Batheja V, Wang KC, Summers RM, Lu Z. One year on: assessing progress of multimodal large language model performance on RSNA 2024 case of the day questions. *Radiology* 2025;316:e250617. https://doi.org/10.1148/radiol.250617.

Downloaded from academic.oup.com/burnstrauma/article/doi/10.1093/burnst/tkag026/8537608 by guest on 12 October 2026

85. Suh CH, Suh PS. Evolving multimodal large language models in radiology: a year of diagnostic progress. *Radiology* 2025;316:e252282. https://doi.org/10.1148/radiol.252282.
86. Wang R, Yao Q, Jiang Z, Lai H, He Z, Tao X, *et al.* ECAMP: entity-centered context-aware medical vision language pre-training. *Med Image Anal* 2025;105:103690. https://doi.org/10.1016/j.media.2025.103690.
87. Zhang R, Wang B, Zhang J, Bian Z, Feng C, Ozbay K. When language and vision meet road safety: leveraging multimodal large language models for video-based traffic accident analysis. *Accid Anal Prev* 2025;219:108077. https://doi.org/10.1016/j.aap.2025.108077.
88. Kraljevic Z, Bean D, Shek A, Bendayan R, Hemingway H, Yeung JA, *et al.* Foresight-a generative pretrained transformer for modelling of patient timelines using electronic health records: a retrospective modelling study. *Lancet Digit Health* 2024;6:e281–90. https://doi.org/10.1016/S2589-7500(24)00025-6.
89. Reddy S. Generative AI in healthcare: an implementation science informed translational path on application, integration and governance. *Implement Sci* 2024;19:27. https://doi.org/10.1186/s13012-024-01357-9.
90. Kumari J, Kumar E, Kumar D. A structured analysis to study the role of machine learning and deep learning in the healthcare sector with big data analytics. *Arch Comput Methods Eng* 2023;30:3673–701. https://doi.org/10.1007/s11831-023-09915-y.
91. Wang H, Xiao N, Zhang J, Yang W, Ma Y, Suo Y, *et al.* Static-dynamic coordinated transformer for tumor longitudinal growth prediction. *Comput Biol Med* 2022;148:105922. https://doi.org/10.1016/j.compbiomed.2022.105922.
92. Woldaregay AZ, Årsand E, Walderhaug S, Albers D, Mamykina L, Botsis T, *et al.* Data-driven modeling and prediction of blood glucose dynamics: machine learning applications in type 1 diabetes. *Artif Intell Med* 2019;98:109–34. https://doi.org/10.1016/j.artmed.2019.07.007.
93. Seki T, Kawazoe Y, Ito H, Akagi Y, Takiguchi T, Ohe K. Assessing the performance of zero-shot visual question answering in multimodal large language models for 12-lead ECG image interpretation. *Front Cardiovasc Med* 2025;12:1458289. https://doi.org/10.3389/fcvm.2025.1458289.
94. Yang X, Li T, Su Q, Liu Y, Kang C, Lyu Y, *et al.* Application of large language models in disease diagnosis and treatment. *Chin Med J* 2025;138:130–42. https://doi.org/10.1097/CM9.0000000000003456.
95. Wen J, Huang W, Yan H, Sun J, Dong M, Li C, *et al.* Evaluation of large language models in generating pulmonary nodule follow-up recommendations. *Eur J Radiol Open* 2025;14:100655. https://doi.org/10.1016/j.ejro.2025.100655.
96. Gallifant J, Afshar M, Ameen S, Aphinyanaphongs Y, Chen S, Cacciamani G, *et al.* The TRIPOD-LLM reporting guideline for studies using large language models. *Nat Med* 2025;31:60–9. https://doi.org/10.1038/s41591-024-03425-5.
97. Jin C, Guo S, Zhou S, Guan J. Effective and explainable molecular property prediction by chain-of-thought enabled large language models and multi-modal molecular information fusion. *J Chem Inf Model* 2025;65:5438–55. https://doi.org/10.1021/acs.jcim.5c00577.
98. Huang Z, Zhong S, Zhou P, Gao S, Zitnik M, Lin L. A causality-aware paradigm for evaluating creativity of multimodal large language models. *IEEE Trans Pattern Anal Mach Intell* 2025;47:3830–46. https://doi.org/10.1109/TPAMI.2025.3539433.
99. Yuan W, Chen G, Wang Z, You F. Empowering generalist material intelligence with large language models. *Adv Mater (Deerfield Beach, Fla)* 2025;37:e2502771. https://doi.org/10.1002/adma.202502771.
100. Chen Z, Bie Y, Jin H, Chen H. Large language model with region-guided referring and grounding for CT report generation. *IEEE Trans Med Imaging* 2025;44:3139–50. https://doi.org/10.1109/TMI.2025.3559923.
101. Li H, Fu J-F, Python A. Implementing large language models in health care: clinician-focused review with interactive guideline. *J Med Internet Res* 2025;27:e71916. https://doi.org/10.2196/71916.
102. Li H, Zhen N, Lin S, Li N, Zhang Y, Luo W, *et al.* Deployable machine learning-based decision support system for tracheostomy in acute burn patients. *Burns Trauma* 2025;13:tkaf010. https://doi.org/10.1093/burnst/tkaf010.
103. Shashikumar SP, Mohammadi S, Krishnamoorthy R, Patel A, Wardi G, Ahn JC, *et al.* Development and prospective implementation of a large language model based system for early sepsis prediction. *medRxiv* 2025. https://doi.org/10.1101/2025.03.07.25323589.
104. Suñol F, de Haro C, Santos-Pulpón V, Fernández-Gonzalo S, Blanch L, López-Aguilar J, *et al.* Leveraging large language models for patient-ventilator asynchrony detection. *BMJ Health Care Inform* 2025;32:e101426. https://doi.org/10.1136/bmjhci-2024-101426.
105. Sundrani S, Chen J, Jin BT, Abad ZSH, Rajpurkar P, Kim D. Predicting patient decompensation from continuous physiologic monitoring in the emergency department. *NPJ Digit Med* 2023;6:60. https://doi.org/10.1038/s41746-023-00803-0.
106. Masanneck L, Schmidt L, Seifert A, Kölsche T, Huntemann N, Jansen R, *et al.* Triage performance across large language models, ChatGPT, and untrained doctors in emergency medicine: comparative study. *J Med Internet Res* 2024;26:e53297. https://doi.org/10.2196/53297.
107. Lorenzoni G, Gregori D, Bressan S, Ocagli H, Azzolina D, Da Dalt L, *et al.* Use of a large language model to identify and classify injuries with free-text emergency department data. *JAMA Netw Open* 2024;7:e2413208. https://doi.org/10.1001/jamanetworkopen.2024.13208.
108. Lafuente C, Rahim M. Evaluating large language models on hospital health data for automated emergency triage. *Int J Comput Assist Radiol Surg* 2025;20:1941–52. https://doi.org/10.1007/s11548-025-03475-1.
109. Gao H, Wang K, Yuan Y, Wang Y, Liu Q, Wang Y, *et al.* A large language model based pipeline for extracting information from patient complaint and anamnesis in clinical notes for severity assessment. *Sci Rep* 2025;15:25345. https://doi.org/10.1038/s41598-025-07649-4.
110. Williams CYK, Zack T, Miao BY, Sushil M, Wang M, Kornblith AE, *et al.* Use of a large language model to assess clinical acuity of adults in the emergency department. *JAMA Netw Open* 2024;7:e248895. https://doi.org/10.1001/jamanetworkopen.2024.8895.
111. Niset A, Melot I, Pireau M, Englebert A, Scius N, Flament J, *et al.* Grounded large language models for diagnostic pre-

Downloaded from academic.oup.com/burnstrauma/article/doi/10.1093/burnst/tkag026/8537608 by guest on 12 October 2026

diction in real-world emergency department settings. *JAMIA Open* 2025;8:ooaf119. https://doi.org/10.1093/jamiaopen/ooaf119.
112. Goh E, Gallo R, Hom J, Strong E, Weng Y, Kerman H, *et al.* Large language model influence on diagnostic reasoning: a randomized clinical trial. *JAMA Netw Open* 2024;7:e2440969. https://doi.org/10.1001/jamanetworkopen.2024.40969.
113. Hoppe JM, Auer MK, Strüven A, Massberg S, Stremmel C. ChatGPT with GPT-4 outperforms emergency department physicians in diagnostic accuracy: retrospective analysis. *J Med Internet Res* 2024;26:e56110. https://doi.org/10.2196/56110.
114. Levra AG, Gatti M, Mene R, Shiffer D, Costantino G, Solbiati M, *et al.* A large language model-based clinical decision support system for syncope recognition in the emergency department: a framework for clinical workflow integration. *Eur J Intern Med* 2025;131:113–20. https://doi.org/10.1016/j.ejim.2024.09.017.
115. Giesa N, Sekutowicz M, Rubarth K, Spies CD, Balzer F, Haufe S, *et al.* Applying a transformer architecture to intraoperative temporal dynamics improves the prediction of postoperative delirium. *Commun Med* 2024;4:251. https://doi.org/10.1038/s43856-024-00681-x.
116. Antonie NI, Gheorghe G, Ionescu VA, Tiucă LC, Diaconu CC. The role of ChatGPT and AI Chatbots in optimizing antibiotic therapy: a comprehensive narrative review. *Antibiotics (Basel)* 2025;14:60. https://doi.org/10.3390/antibiotics14010060.
117. De Vito A, Geremia N, Bavaro DF, Seo SK, Laracy J, Mazzitelli M, *et al.* Comparing large language models for antibiotic prescribing in different clinical scenarios: which performs better? *Clin Microbiol Infect* 2025;31:1336–42. https://doi.org/10.1016/j.cmi.2025.03.002.
118. Iapascurta V, Fiodorov I, Belii A, Bostan V. Multi-agent approach for sepsis management. *Healthc Inform Res* 2025;31:209–14. https://doi.org/10.4258/hir.2025.31.2.209.
119. Wang K, Lin L, Zheng R, Nan S, Lu X, Duan H. Leveraging large language models for preoperative prevention of cardiopulmonary bypass-associated acute kidney injury. *Ren Fail* 2025;47:2509786. https://doi.org/10.1080/0886022x.2025.2509786.
120. Sheikh MS, Thongprayoon C, Qureshi F, Suppadungsuk S, Kashani KB, Miao J, *et al.* Personalized medicine transformed: ChatGPT's contribution to continuous renal replacement therapy alarm management in intensive care units. *J Pers Med* 2024;14:233. https://doi.org/10.3390/jpm14030233.
121. Pham C, Govender R, Tehami S, Chavez S, Adepoju OE, Liaw W. ChatGPT's performance in cardiac arrest and bradycardia simulations using the American Heart Association's advanced cardiovascular life support guidelines: exploratory study. *J Med Internet Res* 2024;26:e55037. https://doi.org/10.2196/55037.
122. Pak TR, Kanjilal S, McKenna CS, Hoffner-Heinike A, Rhee C, Klompas M. Syndromic analysis of sepsis cohorts using large language models. *JAMA Netw Open* 2025;8:e2539267. https://doi.org/10.1001/jamanetworkopen.2025.39267.
123. Fonseca J, Liu X, Oliveira HP, Pereira T. Mortality prediction using medical time series on TBI patients. *Comput Methods Prog Biomed* 2023;242:107806. https://doi.org/10.1016/j.cmpb.2023.107806.
124. Dai L, Yang X, Li H, Zhao X, Lin L, Jiang Y, *et al.* A clinically actionable and explainable real-time risk assessment framework for stroke-associated pneumonia. *Artif Intell Med* 2024;149:102772. https://doi.org/10.1016/j.artmed.2024.102772.
125. Barash Y, Klang E, Konen E, Sorin V. ChatGPT-4 assistance in optimizing emergency department radiology referrals and imaging selection. *J Am Coll Radiol* 2023;20:998–1003. https://doi.org/10.1016/j.jacr.2023.06.009.
126. Lee S, Youn J, Kim H, Kim M, Yoon SH. CXR-LLaVA: a multimodal large language model for interpreting chest X-ray images. *Eur Radiol* 2025;35:4374–86. https://doi.org/10.1007/s00330-024-11339-6.
127. Cao C, Sang J, Arora R, Chen D, Kloosterman R, Cecere M, *et al.* Development of prompt templates for large language model-driven screening in systematic reviews. *Ann Intern Med* 2025;178:389–401. https://doi.org/10.7326/ANNALS-24-02189.
128. Hartman V, Zhang X, Poddar R, McCarty M, Fortenko A, Sholle E, *et al.* Developing and evaluating large language model-generated emergency medicine handoff notes. *JAMA Netw Open* 2024;7:e2448723. https://doi.org/10.1001/jamanetworkopen.2024.48723.
129. Wang H, Gao C, Dantona C, Hull B, Sun J. DRG-LLaMA: tuning LLaMA model to predict diagnosis-related group for hospitalized patients. *NPJ Digit Med* 2024;7:16. https://doi.org/10.1038/s41746-023-00989-3.
130. Wan P, Huang Z, Tang W, Nie Y, Pei D, Deng S, *et al.* Outpatient reception via collaboration between nurses and a large language model: a randomized controlled trial. *Nat Med* 2024;30:2878–85. https://doi.org/10.1038/s41591-024-03148-7.
131. Liu S, Wright AP, McCoy AB, Huang SS, Genkins JZ, Peterson JF, *et al.* Using large language model to guide patients to create efficient and comprehensive clinical care message. *J Am Med Inform Assoc* 2024;31:1665–70. https://doi.org/10.1093/jamia/ocae142.
132. Madden MG, McNicholas BA, Laffey JG. Assessing the usefulness of a large language model to query and summarize unstructured medical notes in intensive care. *Intensive Care Med* 2023;49:1018–20. https://doi.org/10.1007/s00134-023-07128-2.
133. Sumner J, Wang Y, Tan SY, Chew EHH, Wenjun YA. Perspectives and experiences with large language models in health care: survey study. *J Med Internet Res* 2025;27:e67383. https://doi.org/10.2196/67383.
134. Qiang S, Zhang H, Liao Y, Zhang Y, Gu Y, Wang Y, *et al.* Application of large language models in stroke rehabilitation health education: 2-phase study. *J Med Internet Res* 2025;27:e73226. https://doi.org/10.2196/73226.
135. Li Z, Fu Q, Zhao Y, Ren W. Empowering standardized residency training in China through large language models: problem analysis and solutions. *Ann Med* 2025; 57:2516695. https://doi.org/10.1080/07853890.2025.2516695.

Downloaded from academic.oup.com/burnstrauma/article/doi/10.1093/burnst/tkag026/8537608 by guest on 12 October 2026

136. Ko A, Km C, W D, G V, Hs A, K S, *et al.* Artificial intelligence-guided bronchoscopy is superior to human expert instruction for the performance of critical-care physicians: a randomized controlled trial. *Crit Care Med* 2025;53:e1105–15. https://doi.org/10.1097/CCM.0000000000006629.
137. Cook DA, Overgaard J, Pankratz VS, Del Fiol G, Aakre CA. Virtual patients using large language models: scalable, contextualized simulation of clinician-patient dialogue with feedback. *J Med Internet Res* 2025;27:e68486. https://doi.org/10.2196/68486.
138. Mahajan A, Tran A, Tseng ES, Como JJ, El-Hayek KM, Ladha P, *et al.* Performance of trauma-trained large language models on surgical assessment questions: a new approach in resource identification. *Surgery* 2025;179:108793. https://doi.org/10.1016/j.surg.2024.08.026.
139. Lee H. The rise of ChatGPT: exploring its potential in medical education. *Anat Sci Educ* 2024;17:926–31. https://doi.org/10.1002/ase.2270.
140. Cheng Y, Zhu L. A review of ChatGPT in medical education: exploring advantages and limitations. *Int J Surg (London, England)* 2025;111:4586–602. https://doi.org/10.1097/JS9.0000000000002505.
141. Zihang C, Liang L, Qianmin S, Gaoyi C, Jihan H, Ying L. Enhanced pre-recruitment framework for clinical trial questionnaires through the integration of large language models and knowledge graphs. *Sci Rep* 2025;15:27398. https://doi.org/10.1038/s41598-025-11876-0.
142. Van Calster B, McLernon DJ, van Smeden M, Wynants L, Steyerberg EW. Calibration: the Achilles heel of predictive analytics. *BMC Med* 2019;17:230. https://doi.org/10.1186/s12916-019-1466-7.
143. Alarid-Escudero F, MacLehose RF, Peralta Y, Kuntz KM, Enns EA. Nonidentifiability in model calibration and implications for medical decision making. *Med Decis Mak* 2018;38:810–21. https://doi.org/10.1177/0272989x18792283.
144. Fahrner LJ, Chen E, Topol E, Rajpurkar P. The generative era of medical AI. *Cell* 2025;188:3648–60. https://doi.org/10.1016/j.cell.2025.05.018.
145. Bhavani SV, Carey KA, Gilbert ER, Afshar M, Verhoef PA, Churpek MM. Identifying novel sepsis subphenotypes using temperature trajectories. *Am J Respir Crit Care Med* 2019;200:327–35. https://doi.org/10.1164/rccm.201806-1197OC.
146. Woo BFY, Song J, Middleton E, Fijačko N, Cato K. Teaching critical thinking in the age of AI: safeguarding clinical reasoning in healthcare documentation. *Int Nurs Rev* 2025;72:e70102. https://doi.org/10.1111/inr.70102.
147. Cheung V, Maier M, Lieder F. Large language models show amplified cognitive biases in moral decision-making. *Proc Natl Acad Sci USA* 2025;122:e2412015122. https://doi.org/10.1073/pnas.2412015122.
148. Chow JCL, Li K. Ethical considerations in human-centered AI: advancing oncology chatbots through large language models. *JMIR Bioinform Biotechnol* 2024;5:e64406. https://doi.org/10.2196/64406.
149. Karyukin V, Mutanov G, Mamykova Z, Nassimova G, Torekul S, Sundetova Z, *et al.* On the development of an information system for monitoring user opinion and its role for the public. *J Big Data* 2022;9:110. https://doi.org/10.1186/s40537-022-00660-w.
150. Lin C, Kuo C-F. Roles and potential of large language models in healthcare: a comprehensive review. *Biomed J* 2025;48:100868. https://doi.org/10.1016/j.bj.2025.100868.
151. Choudhury A, Chaudhry Z. Large language models and user trust: focus on healthcare. *J Med Internet Res* 2024;26:e56764. https://doi.org/10.2196/56764.

Downloaded from academic.oup.com/burnstrauma/article/doi/10.1093/burnst/tkag026/8537608 by guest on 12 October 2026

© The Author(s) 2026. Published by Oxford University Press This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial License (https://creativecommons.org/licenses/by-nc/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact reprints@oup.com for reprints and translation rights for reprints. All other permissions can be obtained through our RightsLink service via the Permissions link on the article page on our site—for further information please contact journals.permissions@oup.com.
Burns & Trauma, 2026, 14, tkag026
https://doi.org/10.1093/burnst/tkag026
Review