Idan and Einav *Critical Care* (2025) 29:238
https://doi.org/10.1186/s13054-025-05479-4

Critical Care

REVIEW

Open Access

Primer on large language models: an educational overview for intensivists

Daphna Idan[1*] and Sharon Einav[2]

Abstract

The integration of artificial intelligence (AI) and machine learning-enabled medical technologies into clinical practice is expanding at an unprecedented pace. Among these, large language models (LLMs) represent a subset of machine learning designed to comprehend linguistic patterns, semantics, and contextual meaning by processing vast amounts of textual data. This educational primer aims to inform intensivists on the foundational concepts of LLMs and how to approach emerging literature in this area. In critical care, LLMs have the potential to enhance various aspects of patient management, from triage and clinical documentation to diagnostic support and prognostic assessment of patient deterioration. They have also demonstrated high appropriateness in addressing critical care-related clinical inquiries and are increasingly recognized for their role in post-ICU rehabilitation and as educational resources for patients' families and caregivers. Despite these promising applications, LLMs still have significant limitations, and integrating LLMs into clinical workflows presents inherent challenges, particularly concerning bias, reliability, and transparency. Given their emerging role as decision-support tools and potential collaborative partners in medicine, LLMs must adhere to rigorous validation and quality assurance standards. As the trajectory toward AI-driven healthcare continues, responsible and evidence-based integration of LLMs into critical care practice is imperative to optimize patient outcomes while ensuring ethical and equitable deployment.

Keywords Large language models, Critical care, Artificial intelligence

Introduction

In 2022, the US Food and Drug Administration (FDA) approved 139 medical devices incorporating artificial intelligence (AI) [1]. By August 2024, this number had surged to 950 AI or machine learning-enabled medical devices [2], highlighting the exponential growth of AI integration into medicine in general and clinical practice in particular [3].

*Correspondence:
Daphna Idan
daphnaid@post.bgu.ac.il
[1]Ben-Gurion Faculty of Health Sciences, Beer-Sheva, Israel
[2]Hebrew University Faculty of Medicine and Regional Medical Director at Maccabi Healthcare and Chief Scientist, Medint Medical Intelligence, Hebrew University, Jerusalem, Israel

Artificial intelligence (AI) is broadly defined as the seeming ability of a device to perform tasks that typically require human cognitive skills, such as reasoning, learning, and decision-making. At its core, AI is intended to embody the capacity to "do the right thing" in a given context by employing logic-based methods efficiently and safely.

Early AI systems were designed to achieve specific outcomes for well-defined tasks and are therefore often referred to as task-specific AIs. Typical examples are the early customer service chatbots, which relied on rigid rules to recognize specific keywords and generate pre-programmed responses to these words. These simple systems relied on algorithms - mathematical constructs first formalized in the ninth century by the Persian mathematician Muhammad ibn Musa al-Khwarizmi. Algorithms

define specific rules for specific outcomes and have been the foundation of decision-making models for centuries [4].

More recently, AI has gained an additional layer of sophistication - the ability to learn from vast datasets (i.e., big data). This progress has given rise to neural networks, which mimic the interconnected pathways of the brain to model complex relationships. Neural networks evaluate multiple possible pathways and select the most probable outcome (probabilistic reasoning). Bayesian statistics comprise a critical component of this process. Bayesian methods may be applied to all types of data, and they are typically used to calculate outcome likelihoods based on prior probabilities and new evidence. The resultant outputs of these methods are ranges of confidence related to the likelihood of the outcome (i.e., confidence intervals) rather than a definitive answer (i.e., cutoff levels). While highly effective, Bayesian approaches to data highlight a fundamental AI challenge - the need to balance statistical precision with uncertainty. Or, in other words, creative problem-solving.

Here enters generative AI, whose emergence marked another milestone in AI development. Generative AI models differ from earlier models in the production of outputs for new inputs, based on learned patterns rather than solely following predefined pathways. Generative AI models introduce an element of creativity, as they generate predictions (often also called insights) by synthesizing prior information that is frequently not overt. For example, AI systems trained on extensive imaging datasets can identify new patterns and diagnose novel cases [5]. Such advancements fuel ongoing debates regarding the limits of AI and the delicate balance between creativity, transparency, and replicability.

Of particular relevance to critical care is the AI field of machine learning (ML) that leverages vast datasets to train algorithms that infer logic and adapt to new scenarios. Using electronic health records (EHRs) as the data source, granular and diverse clinical data are aggregated to enable predictive modeling using ML. As discussed below, these datasets can be leveraged to empower AI systems to anticipate clinical decisions, optimize patient care pathways, and improve medical decision-making. This educational primer aims to teach intensivists the foundational concepts of large language models and how to approach emerging literature in this area.

The building blocks of large language models

Large language models (LLMs) are a type of machine learning specifically designed to understand the relationships between words and phrases [6]. LLMs learn the grammar, semantics, and contextual use of human language by processing vast amounts of data from diverse sources. A key component of this learning involves creating embeddings, mathematical representations of words. These allow the model to group words with similar meaning close together in a kind of language map [7]. For example, the words "heart rate", "blood pressure", "temperature", and "respiratory rate" would be grouped in the map as "vital signs". These relationships help the model understand how words are used in different contexts.

These models originate from natural language processing (NLP), a field that employs computational techniques to represent text using algorithmic structures based on word co-occurrence frequencies within a given dataset [8]. Unlike traditional NLP models, which rely strictly on the data provided within a specific dataset, LLMs can incorporate contextual information and generate responses based on learned patterns. Table 1 presents key terms related to augmented intelligence and large language models.

A key distinction for understanding how LLM algorithms function is the model distinction between words and tokens. Words are linguistic units that convey meaning, while tokens are discrete symbols that comprise or represent these units (e.g., sub-words or even punctuation marks). For instance, a simple word like *"note"* may be represented by a single token, while a longer or compound word such as *"hospitalization"* might be broken into two tokens (e.g., *"hospital"*, *"ization"*). Each token is assigned a probability of appearing next in a sequence based on the tokens that preceded it. The model employs a mechanism called self-attention to evaluate all parts of the input simultaneously, allowing it to determine contextual relationships and generate coherent output. In other words, this mechanism is designed to evaluate the relevance of different parts of the input when generating a response. For example, when a physician prompts an LLM to generate an admission or discharge summary, the model may refer to the patient using both "the patient" and the pronoun "his/her." In the sentence, "The patient was given his medications," the model is able to associate "his" with "the patient" because it considers the entire sentence holistically, not just word by word. This process ultimately results in an output reflecting the most likely and contextually appropriate sequence of tokens [9] - a "predictive language assembly" that enables the model to form medically coherent and grammatically accurate responses (Fig. 1).

Translating words into tokens is called encoding, and translating these tokens back into text is called decoding. Embeddings are a critical component of LLM formulation of the context, nuance, and subtle meanings of words and phrases that undergo decoding and encoding. These embeddings represent each token's place in the map of language space. LLM inputs are first transformed into tokens and embeddings, and after being processed as a

Table 1 Key terms related to artificial intelligence and large language models

Application Programming Interface (API)	A framework that facilitates communication and data exchange between software applications, enabling integration of features and functionalities.
Artificial Intelligence (AI)	Computer systems that are designed to perform tasks that usually require human intelligence. These may be classified into narrow AIs focused on specific tasks like language translation or playing chess and general AIs capable of broader functions like learning, reasoning, and problem-solving.
Artificial Neural Network	Interconnected layers of computational units (neurons) that process information.
Bayesian Statistic	A statistical framework that applies Bayes' theorem to update probabilities based on new evidence. It is particularly suited for decision-making in circumstances of uncertainty, providing a probabilistic approach to data analysis and model inference.
Big Data	Large datasets, distinguished by their size, diversity, and processing speed, that may facilitate advanced data analysis and pattern recognition essential for machine learning and AI applications.
Deep learning	A subset of machine learning that uses artificial neural networks with multiple layers to model complex patterns. The term "deep" refers to the number of layers in the network, with deeper networks enabling the execution of more intricate tasks.
Embedding	A mathematical representation of data (such as words, sentences, or images) in a dense, low-dimensional space. Embeddings capture semantic relationships between items, allowing AI systems to analyze similarity and contextual meaning efficiently.
Encoding and Decoding	Processes in machine learning and AI that transform input data into structured formats (encoding) and convert structured representations back into human-readable formats (decoding).
Generative Artificial Intelligence	A technology that produces content by identifying patterns within large datasets. Depending on the model, outputs may include text, images, music, and more.
Intelligence Augmentation	The ability of computer systems to enhance human capabilities and improve performance rather than merely automating tasks.
Large Language Model	A type of neural network-based AI capable of performing diverse linguistic tasks by analyzing large volumes of text data. LLMs identify relationships between token sequences and compute probabilities, enabling the performance of a variety of tasks such as language translation, summarization, and content generation.
Machine Learning (ML)	A practical subfield of computer science and AI based on statistical models. ML utilizes algorithms that allow systems to learn patterns from data without explicit programming. Through iterative learning from experience, these systems may improve performance over time.
Natural Language Processing	A field that employs computational techniques to represent text using algorithmic structures based on word co-occurrence frequencies within a given dataset.
Retrieval Augmented Generation (RAG)	A framework that enhances LLMs by incorporating relevant and updated data from appropriate sources to produce more informed responses.
Tokens	Discrete units of text (such as words, subwords, or characters) that are used by language models to process and analyze language. Tokens are the building blocks for the computational understanding of text, allowing models to generate or interpret language.

series of layers in the model architecture, the output is created using the reverse process. This multi-layer architecture, known as a transformer encoder-decoder, allows the representations of tokens to be progressively refined. Each successive layer helps the model learn increasingly complex patterns, from basic grammar to more abstract semantic relationships. Statistical strategies are applied in the decoding process when the model selects one token over another based on learned probabilities - the likelihood of each token appearing in a given context based on training data. As a result of this selection process, the model generates text. Figure 2 presents a simplified large language model processing algorithm.

Large language models in critical care

LLMs are still in their infancy concerning use in critical care, with very few clinically validated applications and a glaring lack of scientific consensus on their actual use. Today's truths may be rapidly swept away by exponentially improving technology. Yet, considering how these tools might be integrated into critical care could help frame the essential discussion on their future role in clinical decision-making. LLMs will likely be used across the critical care continuum to assist in triage and documentation, diagnosis and prediction of patient deterioration, and patient management. LLMs have demonstrated a high median score for appropriateness in addressing clinical questions related to critical care [10] and have also been proposed to support patient rehabilitation after discharge from the ICU [11, 12].

As of this writing, at least six additional papers are under review in prepublication databases that propose using LLMs in critical care for treatment planning, patient care management, and prediction of deterioration and mortality.

Triage

Early identification of patients at risk of rapid clinical deterioration can improve triage and enable early response, including ICU admission and implementation of care interventions, which machine learning models aim to predict. Efforts are being made to improve the

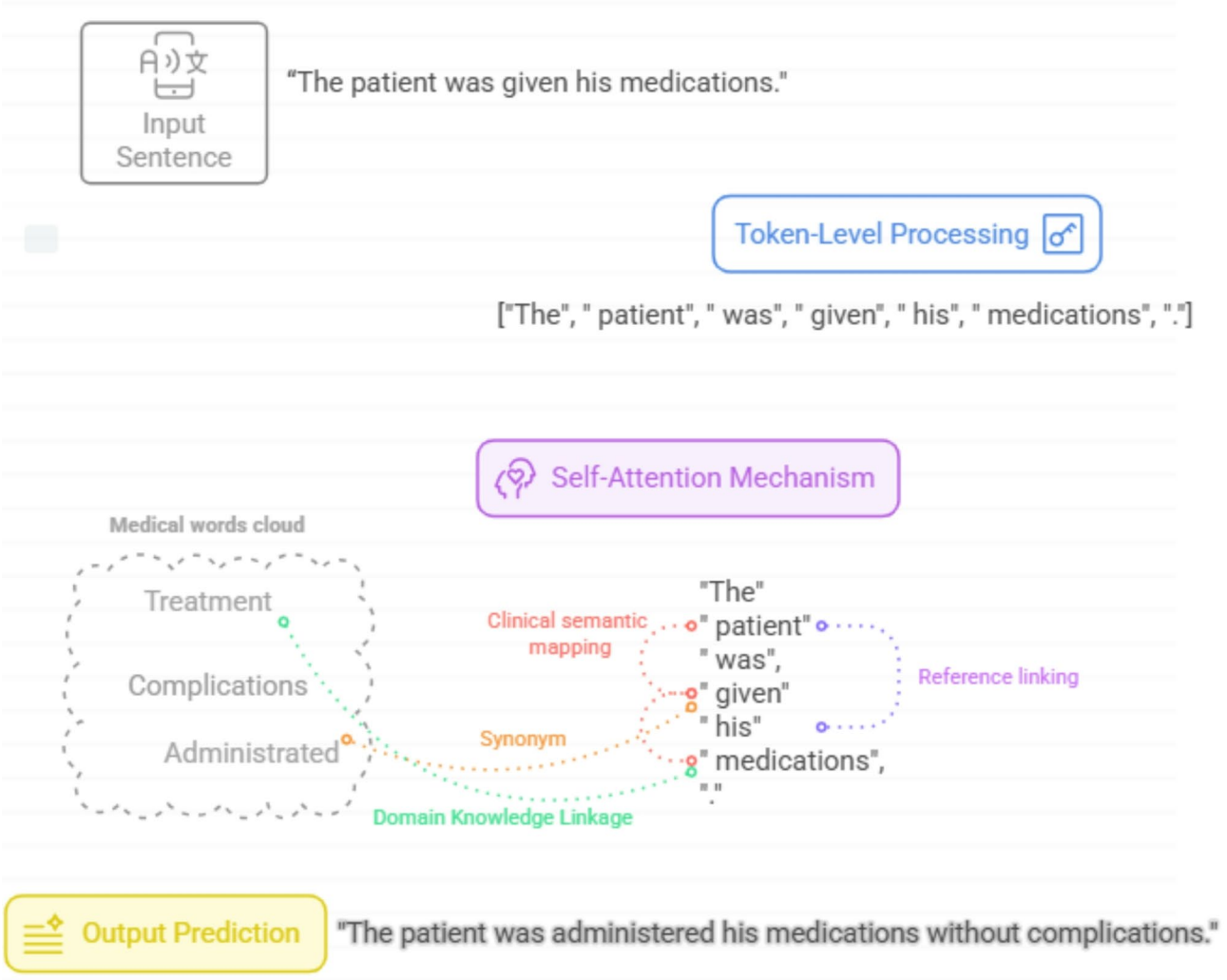

Fig. 1 Token-based processing of a large language model in a Clinical Context

accuracy of ICU admission predictions, including integrating NLP technology to enhance the quality of the data used for model development [13].

Retrieval-augmented generation (RAG) is a framework that augments a Large Language Model (LLM) with updated data to generate more informed responses. The retrieval model accesses, selects, and prioritizes the most relevant documents and data from appropriate sources, transforms these into an enriched contextual prompt, and invokes the LLM through an application programming interface (API, see Table 1) to generate the response. This type of machine learning is comparable to stock traders' use of publicly available historical financial information and live market data feeds to make decisions. Numerous approaches to integrating LLMs for patient triage into clinical practice are being explored, including a RAG approach. Yazaki et al. used Chat-GPT3.5 with RAG to enhance contextual understanding and achieved a 70% accuracy rate in triaging emergency cases from the Japanese National Examination for Emergency Medical Technicians [14]. A retrospective study used real-world data from seven hospitals to evaluate ChatGPT-4's accuracy in predicting hospital admissions after Department of Emergency Medicine visits. When RAG was incorporated, the prediction accuracy was 81.3% [15].

Documentation

Documentation is essential to any clinical care process, starting from patient intake. LLMs have been proposed for creating clinical notes based on the assumption that such use may reduce physician burnout [16]. In the demanding field of critical care, where clinicians face significant workload pressures, yet daily notetaking must cover most physiological systems, such use may be particularly beneficial.

Madden et al. highlighted the transformative potential of ChatGPT in processing and synthesizing real-time summaries from the daily free-text entries from ICU

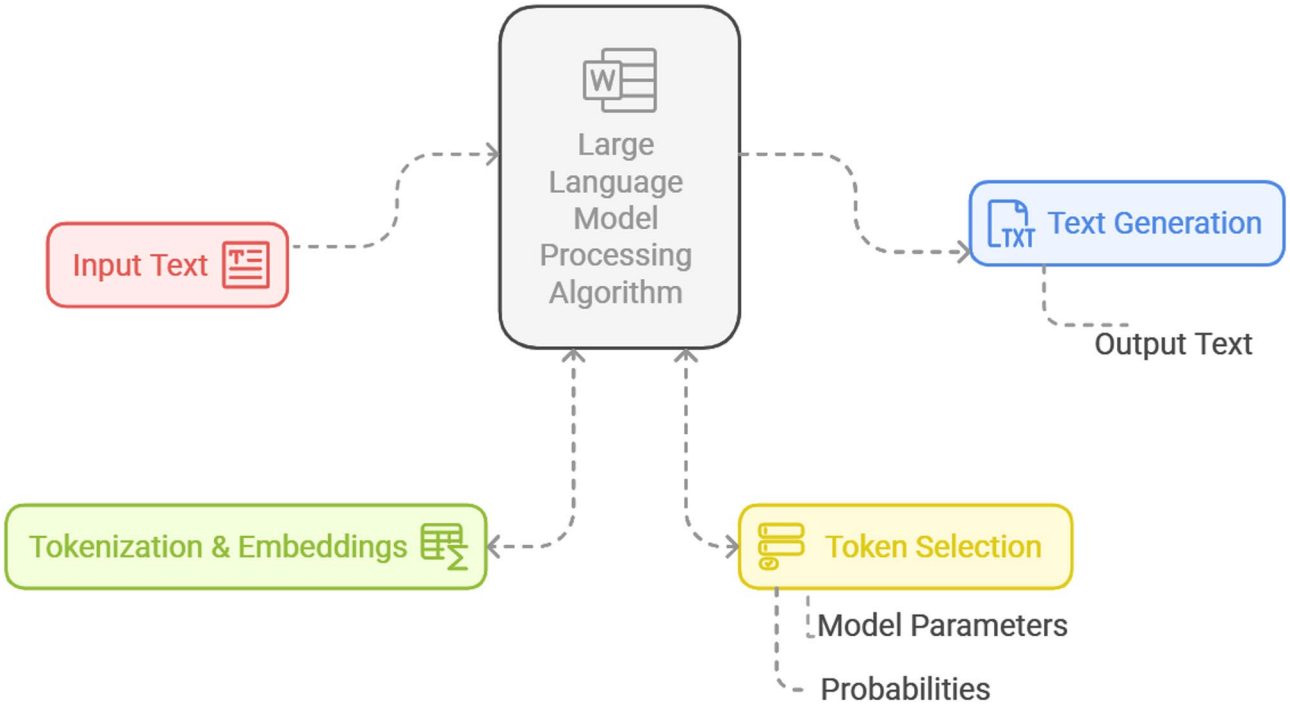

Fig. 2 A simplified large language model processing algorithm

electronic health records [17]. These entries, authored by doctors, nurses, and allied health professionals, often contain critical information but are typically informal, abbreviated, and poorly structured. In their study, ChatGPT-4 generated concise, actionable summaries and responded to queries (e.g., provided timelines of administered medications). A pilot feasibility study also demonstrated the ability of LLMs to generate concise summaries of ICU admissions for discharge documentation [18]. Another single-blind trial found that the quality of discharge letters generated by Chat-GPT4 was comparable to those written by junior clinicians [19].

Z codes (Z55-Z65) are the International Classification of Diseases, Tenth Revision Clinical (ICD-10) Modification diagnosis codes used to document social determinants of health data (e.g., housing, food insecurity, transportation, etc.). Guevara et al. investigated the potential use of LLMs for extracting six social determinants of health categories from narrative text in electronic health records, including employment, housing, transportation, and parental status. The best-performing models accurately identified 95.7% of patients with at least one mention of an SDoH category, compared to just 2.0% identified through structured Z-codes in the electronic health record during the same timeframe [20]. Early identification of poor social determinants of health could be invaluable for the prevention of ICU admissions as well as for planning for post-ICU rehabilitation, particularly in patients with complex medical conditions that are subsequently more likely to require such admission. Another study found that LLMs outperformed human coders in extracting ICD-10 codes from patient notes [21].

Informed consent is another critical domain of documentation. A cross-sectional study of surgical procedures revealed that LLM-based, chatbot-generated presentations of risk, benefit, and possible alternatives to surgery outperformed those presented by surgeons in both composite completeness and accuracy scores, based on expert evaluation and readability assessments, compared to presentations by surgeons. Based on these results, the authors suggested that LLMs be integrated into electronic health records to provide personalized risk and benefit assessments before performing invasive procedures [22].

Medication prescription and clinical documentation share similarities, requiring precise and detailed writing. Given the burden of work imposed on physicians, errors can occur in both processes. The "Healthy Technology Act of 2025," a bill introduced in the 119 th Congress of the US House of Representatives (H.R.238), proposes permitting the use of AI for medication prescribing (albeit with precautions) [23]. This development introduces a new dimension to the potential role of AI, including LLMs, in this domain.

Finally, LLMs are used to summarize doctor-patient conversations during palliative care teleconsultations performed almost similarly in medical conversation summarization. Chat-GPT4 balanced content understanding and preserved structural similarity to the source

somewhat better than other models, suggesting clinicians could use this LLM to generate medical summaries of such meetings. These summaries could then be given to the patient and/or their family, who may need to reflect on the content of the meeting [24].

Diagnostic support

Diagnostics is another field where LLMs can play a role in critical care. A randomized, double-blind crossover study compared the performance of the LLM tool AMIE (Articulate Medical Intelligence Explorer) to that of twenty primary care physicians during text-based consultations modeled after an Objective Structured Clinical Examination (OSCE). The study included 149 clinical vignettes evaluated by specialist physicians and patient actors. AMIE showed greater diagnostic accuracy and superior performance on 28 of the 32 axes assessed by the specialist physicians and 24 of the 26 axes evaluated by the patient actors [25].

Another retrospective cohort study conducted in a 40-bed PICU demonstrated the capability of domain-specific LLMs, such as those trained on specific medical data, to generate differential diagnoses. While their performance was inferior to that of human clinicians in terms of quality, pediatric critical care specialists gave them high evaluation scores [26].

In the complex diagnostic landscape of the ICU, the perspectives of family members and caregivers are often fraught with misinterpretations and unanswered questions. Scquizzato et al. evaluated the accuracy of ChatGPT in responding to non-professional questions about cardiac arrest. ChatGPT provided highly accurate answers, as assessed by clinicians and researchers specializing in out-of-hospital cardiac arrest, as well as by laypersons. Given the emotionally charged nature of scenarios such as cardiac arrest, which are integral to daily critical care practice, this suggests that leveraging the capabilities of LLMs to help address and clarify clinical situations for families has significant potential [27].

Imaging

The last decades have seen a surge in the use of diagnostic imaging with a related increase in the need for an efficient image interpretation and reporting process. This rising workload has led to concerns regarding decreased efficacy and a higher likelihood of mistakes due to system overload and radiology staff burnout. Radiologists are expected to handle substantial textual information - from diagnostic request forms, medical charts and summaries, information from prior or other examinations, and the most updated medical literature. LLMs may be used to ameliorate this burden if used wisely. While this use may improve the efficiency of radiology services overall, those most likely to benefit are critically ill patients who often require frequent testing and rapid results. Medical imaging of critically ill patients poses unique challenges, including the need to meet stringent time frames and minimize complications stemming from redundant patient transfers. LLMs may be used to improve radiology service efficacy and effectiveness in ways that may be particularly relevant for critically ill patients.

A study comparing human radiologists to Chat-GPT-4 V and Gemini Pro Vision concluded that human radiologists still outperform these LLMs in diagnostic accuracy across various subspecialties (neuroradiology, gastrointestinal, genitourinary) but concluded that LLMs may potentially be used to support clinical decision-making [28]. Another model, CXR-LLaVa, which integrates an LLM with an image encoder, demonstrated 81% diagnostic accuracy in identifying six common clinical conditions from test sets of X-ray images using the Medical Information Mart for Intensive Care (MIMIC) database [29].

These findings have been supported by an additional study that showed that ClotCatcher, a natural language model with data augmentation, can rapidly and accurately identify venous thromboembolism (VTE) from radiology reports. The authors concluded that the model may improve the efficiency and accuracy of incident VTE adjudication in large databases [30].

Monitoring and early prediction of patient deterioration

Critically ill patients may rapidly deteriorate, a situation that requires early diagnosis and effective treatment decisions in complex clinical situations. Additional difficult decisions that typically need to be addressed in the critical care environment are those relating to patient preferences that must be made in conjunction with the families, often on behalf of patients unable to make decisions themselves. Time constraints, cultural reluctance to address end-of-life issues, and clinician burdens may limit the ability to elucidate individual patients' value judgments and preferences. A proof-of-concept study explored the potential of LLMs to integrate patient values into critical care decision-making for incapacitated patients. Automated extractions of the treatment in question were accurate in 88% of scenarios. LLM treatment recommendations were rated by adjudicators with an average Likert score of 3.92 out of 5.00 for being medically plausible and reasonable and 3.58 out of 5.00 for reflecting documented patient values [31].

The possible use of LLMs to predict patient deterioration has also been explored in the context of respiratory failure and support. A machine learning model integrated with natural language processing, ARDSFlag, demonstrated an overall accuracy of 89.0% in identifying ARDS cases [32]. Another small, prospective study found that Chat-GPT4 demonstrated an accuracy comparable to

that of specialized physicians in predicting the need for endotracheal intubation in patients receiving high-flow nasal cannula therapy for 48 h [33]. A third study used natural language processing to identify under-documentation of ARDS in ICU discharge notes [34]. Such use has more than just research implications - it can also serve as an educational tool.

Sepsis remains a leading cause of ICU mortality [35], yet remains a significant diagnostic challenge in the ICU. The SERA algorithm is an AI-enabled tool that uses natural language processing of physicians' clinical notes using structured electronic medical records (EMR) data. SERA had a high predictive accuracy for identifying sepsis 12 h before its onset, with an AUC of 94%. Compared to physician predictions, the SERA algorithm increased early detection of sepsis by as much as 32% while reducing false positives by 17% [36].

Acute kidney injury (AKI) affects 30–57% of critically ill patients and is associated with high morbidity and mortality [37]. Among patients discharged from the ICU with normal renal function after AKI, almost one in three will relapse into renal failure within 5 years [38]. One study evaluated the effectiveness of Chat-GPT4 in teaching patients about AKI and continuous renal replacement therapy (CRRT). The model demonstrated a 97–98% overall accuracy, consistent performance across question types, and no significant differences between AKI and CRRT responses [39].

Management of treatment

Critical care involves providing a broad spectrum of treatment regimens tailored to diverse clinical scenarios. Integrating LLMs into this process may optimize time and efficiency in care delivery. Howard et al. explored whether ChatGPT (version unspecified) may be used to provide antimicrobial treatment recommendations in eight hypothetical infection scenarios. While limitations were noted in addressing complex cases, ChatGPT demonstrated an overall ability to suggest appropriate antimicrobial spectra and regimens for the diagnoses and recognized the implications of clinical responses [40].

Delirium occurs in approximately 30% of ICU patients, with rates rising to 90% among mechanically ventilated patients [41]. Delirium is also associated with increased mortality after ICU admission. One of the studies still in preprint, suggests that DeLLirium - an LLM-based prediction model - achieved better results than other deep-learning models in predicting delirium from electronic health records [42]. Although this tool is primarily intended for critical care research rather than clinical practice, its potential for detecting delirium through conversations with patients or relatives may enable early identification of at-risk individuals. Alternative models may be developed for predicting post-ICU depression among patients and caregivers.

LLMs may also become an essential educational resource for families and caregivers after discharge from the ICU. For example, non-professional caregivers rarely have the training or preparation required for this challenging role. The quality of post-ICU care and the degree of caregiver strain may both be affected by poor preparation. The CaLM (caregiver large language model) has been proposed as a tool for teaching caregivers. The developers of this model aimed to provide caregivers with at least some of the knowledge they require to undertake this challenging role. They showed that by incorporating retrieval-augmented generation (a method used for improving model performance through connection with external knowledge bases), a valuable support tool tailored to specific caregiver scenarios could be created [43].

Patients recovering from ICU admission often require a lengthy and multidisciplinary rehabilitation process. A study that evaluated individualized exercise recommendations generated by an AI chatbot found them to be 41.2% comprehensive and 90.7% accurate. The chatbot could not provide complete and precise recommendations. Still, chatbots are early precursors of the LLMs existing at the time of this writing, and this study represents the potential for supporting rehabilitation efforts through such tools [44].

Figure 3 shows a timeline of patient management in the ICU with the potential application of LLMs at each treatment point.

While the examples presented illustrate potential directions for integrating LLMs into critical care daily practice, these remain exploratory. Figure 4 offers one possible roadmap from model development to clinical integration.

The challenges and limitations of large Language models

Since the introduction of ChatGPT by OpenAI in November 2022, the public adoption of virtual assistants powered by large language models (LLMs) has grown rapidly. The interest in their application in healthcare, including critical care settings, highlights their potential, as shown above. This article summarizes a rapidly evolving technology whose clinical impact remains hypothetical. LLMs are still at the stage of isolated experiments in exploratory studies (for assessment of the studies presented above, refer to Table 2) with limited incorporation of real-world patient care data; a recent systematic review by Bedi et al. found that only 5% of studies use such data [45]. There is no robust clinical validation, and the widespread use of these tools has also brought attention to their limitations and associated challenges [46, 47].

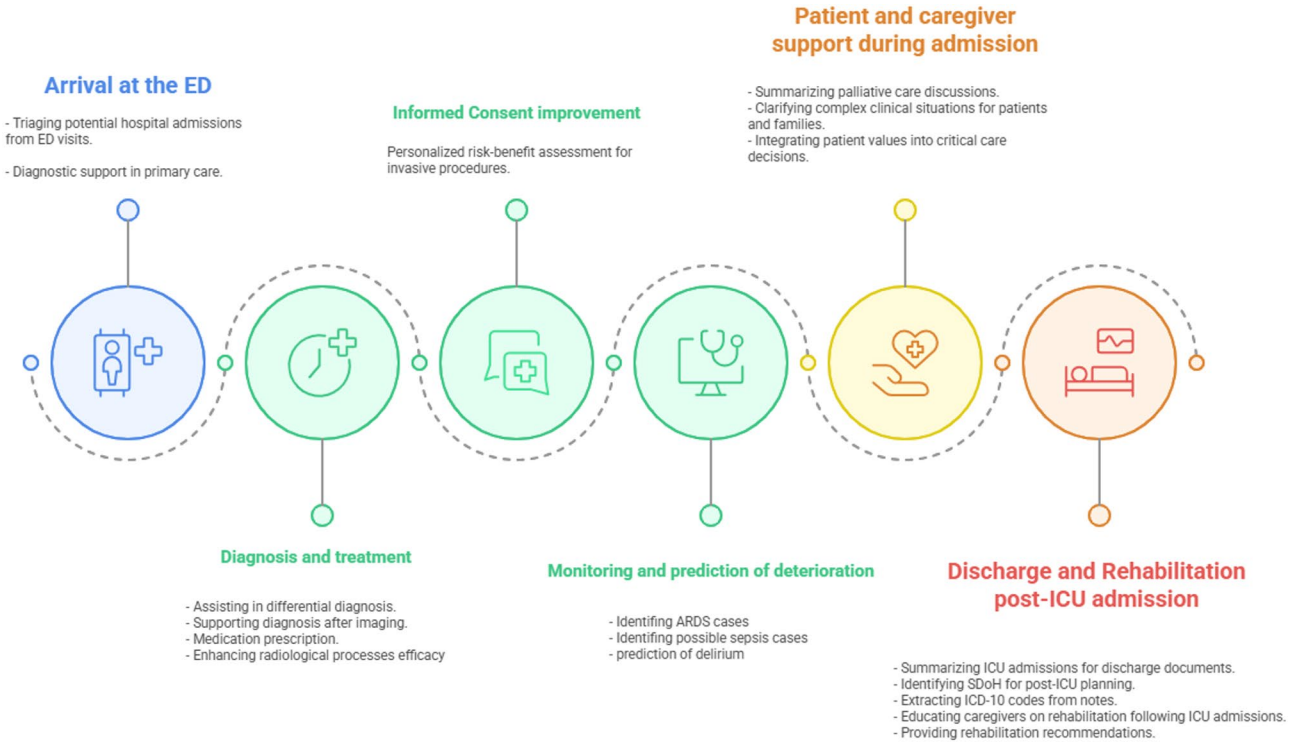

Fig. 3 A timeline of patient management in the ICU with the potential application of LLMs at each treatment point

A critical consideration in incorporating LLMs into clinical practice is their susceptibility to various biases, among them sycophancy bias, which may lead to outputs reinforcing clinicians' preexisting beliefs, potentially increasing errors [48]. Our knowledge of such biases highlights the need to recognize and address how they may influence outputs and the importance of ongoing vigilance when integrating LLM-generated recommendations into clinical decision-making.

Another broader concern regarding the use of AI in general (not limited to LLMs alone) is the phenomenon of overreliance. Clinicians may trust AI-generated diagnoses even when the model produces inaccurate results (Supplementary A) [49]. One study investigated whether providing explanations alongside model-generated diagnoses could help clinicians discern and disregard incorrect outputs. Paradoxically, adding explanations did not improve decision-making accuracy, and reliance on the AI model persisted [50].

Finally, the hurdle of integrating LLMs into clinical workflows remains. These models often function as "black boxes," with limited transparency regarding their internal decision-making processes - a challenge that extends even to their developers and is particularly pronounced among clinicians without even the basic appropriate training. This lack of clarity, coupled with insufficient familiarity among physicians regarding the known capabilities and limitations of these tools, impairs their ability to engage with LLMs in a safe, informed, and clinically meaningful manner. Physicians in family medicine, internal medicine, and emergency medicine exhibited better diagnostic performance on their own compared to when assisted by an LLM. This was assessed based on the accuracy of differential diagnoses, the relevance of supporting and opposing clinical factors, and the appropriateness of the diagnostic evaluation process. The authors interpreted this finding as highlighting "the need for technology and workforce development to realize the potential of physician-artificial intelligence collaboration in clinical practice" [51].

AI tools are rapidly transitioning from simple tools to assistants and potentially even collaborative partners in medicine. They should, therefore, be upheld to similarly rigorous quality assurance standards. The Transparent Reporting of a Multivariable Model for Individual Prognosis Or Diagnosis for LLMs (TRIPOD-LLM) framework has recently been proposed for reporting clinical prediction models developed using large language models [52]. Several critical care leaders have also called for action on AI technologies, emphasizing the need to address technical, ethical, social, and practical issues posed by these tools. Their call highlighted the importance of ensuring that AIs, who may someday be viewed as equal partners to physicians, meet the same ethical and professional standards expected of humans. LLMs must uphold integrity, foster trust in clinical environments, and support

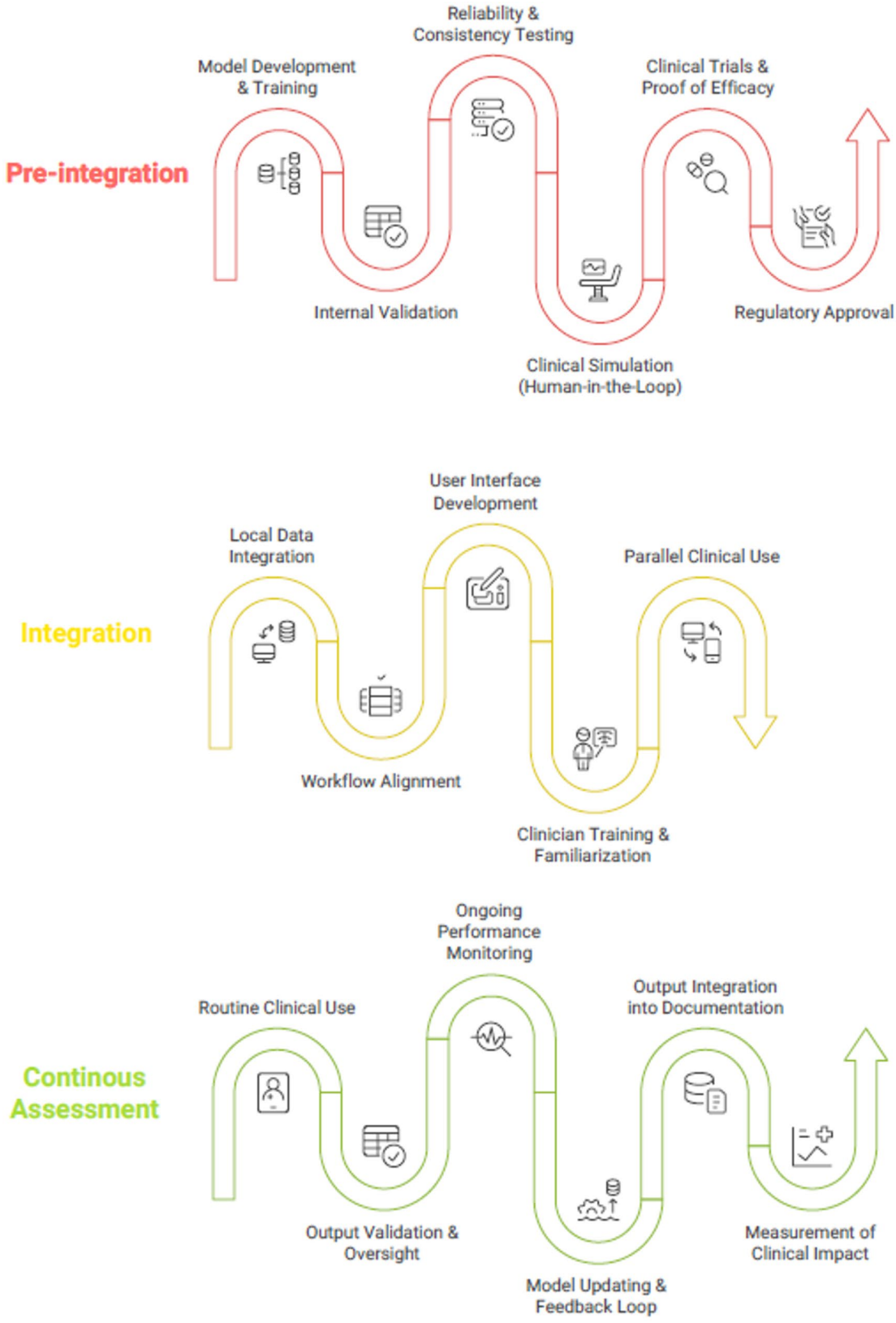

Fig. 4 Proposed LLM implementation pathway

Table 2 Qualitative assessment of LLM studies based on the transparent reporting of a multivariable model for individual prognosis or diagnosis (TRIPOD)-LLM recommendations [52]

Authors	Study Design	Healthcare Context and Intended Use	Source of Data	Objective Evaluation - Metrics and Assessors	Subjective Evaluation - Metrics and Assessors	Performance Comparators (other LLMs, humans, other benchmarks, or standards)	Validation Approach (internal, external, or no formal validation)
Akhondi-Asl et al., 2024 [26]	Single-center retrospective cohort study	Generating differential diagnoses from the admission notes of PICU patients	Admission notes from 10 years period for model development, 130 notes randomly selected for evaluation	None	A 5-point Likert scale of overall quality	Clinicians vs. general LLMs (BioGPT-Large, LLaMa-65B), fine-tuned LLMs (fine-tuned BioGPT-Large, fine-tuned LLaMa-7B)	Internal
Chen et al., 2024 [24]	Pilot study	Summarization of palliative care teleconsultation	Summary of a simulated doctor–patient conversation during teleconsultation	Standardized metrics for precision and similarity to reference text (e.g., ROUGE, BLEU, BERTScore)	None	GPT-3.5, GPT-4, LLaMA-2-7B, LLaMA-2-13B, and LLaMA-2-70B	No formal validation
Contreras et al., 2024 [42] *(preprint)*	Multi-center study	Introducing LLM-based Delirium prediction model in the ICU	eICU Collaborative Research Database, MIMIC-IV and the University of Florida Health's Integrated Data Repository	Standard statistical metrics of AUC	None	None	Internal and external
Glicksberg et al., 2024 [15]	Retrospective study	Predicting the admission of patients arriving at the ED	Electronic health records from seven hospitals	Standard statistical metrics of AUC, AUPRC, and accuracy	None	ML models vs. GPT-4	Internal and external
Guevara et al., 2024 [20]	Comparative study	Identify SDoH in EHRs	Electronic health records	Automated evaluation using macro-F1 score	None	General LLMs (GPT-3.5 and GPT-4) vs. fine tuned LLM (BERT-base and Flan-T5)	Internal and external
Balta et al., 2024 [10]	Cross-sectional comparative study	Evaluation of critical care recommendations	50 Clinical critical care questions synthesized by the authors	Flesch-Kincaid Grade Level (objective readability assessment)	A 5-point Likert scale for appropriateness and consistency	GPT-3.5 vs. GPT-4	No formal validation
Liu et al., 2024 [33]	Prospective multicenter cohort study	Predicting efficacy of high-flow oxygen therapy	Electronic health records of 71 patients	Standard statistical metrics of AUC, sensitivity, specificity, and precision	None	LLMs (GPT-3.5, GPT-4.0) vs. respiratory and critical care specialist physicians and non-specialist physicians	Internal
Madden et al., 2023 [17]	Letter to the editor	Query and summarize medical notes in ICU	None	None	Clinician feedback on the usefulness and clarity of generated summaries	None	No formal validation
Nolan et al., 2024 [31]	Proof-of-cocncept study	Integrate patient values into clinical decision-making processes in critical care for patients who are incapacitated	50 Text-based scenarios of decisionally incapacitated patients	None	A 5-point Likert scales for medical plausibility and alignment with patient values	None	No formal validation
Parmanto et al., 2024 [43]	Exploratory study	Develop a new model for caregivers' questions	Cargiving knowledge bases, including journal articles, care guidelines, and forums	Standardized metrics for precision and similarity to reference text (e.g., ROUGE, BLEU)	None	New LLM (CaLM) and GPT-3.5	Internal

Table 2 (continued)

Authors	Study Design	Healthcare Context and Intended Use	Source of Data	Objective Evaluation - Metrics and Assessors	Subjective Evaluation - Metrics and Assessors	Performance Comparators (other LLMs, humans, other benchmarks, or standards)	Validation Approach (internal, external, or no formal validation)
Sheikh et al., 2024 [39]	Not specified by the authors	Assessing accuracy in responding to patient education questions	89 questions from the Mayo Clinic Handbook for educating patients on AKI and CRRT	None	Subjective accuracy rating	None	Internal
Tommaso et al., 2024 [27]	Not specified by the authors	Address public inquiries related to cardiac arrest and CPR	40 questions	Readability assessed using the Flesch Reading Ease score	A 5-point Likert scales for accuracy, clarity, relevance, comprehensiveness, and overall value	None	Internal
Urquhart et al., 2024 [18]	Pilot study	Synthesize discharge summary of ICU patients	Text from five ICU episodes	None	Subjective evaluation by staff intensivists	ChatGPT, GPT-4 API, and Llama 2	Internal
Yazaki et al., 2024 [14]	Not specified by the authors	Triaging ED patients	100 simulated triage scenarios	Standard statistical metrics of triage accuracy	None	GPT-3.5 with RAG, GPT-3.5 without RAG, and GPT-4 without RAG vs. emergency medical technicians and emergency physicians	Internal
Zaleski et al., 2024 [44]	Mix methods study	Providing individualized exercise recommendations	26 queries on exercise advice	Flesch-Kincaid Grade Level (objective readability assessment)	Subjective assessment of comprehensiveness and factual accuracy	None	Internal

AKI: Acute kidney injury; AUC: Area under the receiver operating characteristic curve; AUPRC: Area under the precision-recall curve; BERT: bert large uncased whole word masking finetuned squad; BERTScore: Bidirectional encoder representations from transformers score; BLEU: Bilingual evaluation understudy; CPR: Cardiopulmonary resuscitation; CRRT: Continous renal replacement therapy; ED: Emergency department; EHRs: Electronic health records; GPT: Generative Pre-trained Transformer; ICD: International classification of diseases; ICU: Intensive care unit; MIMIC-IV: Medical Information Mart for Intensive Care; ML: machine learning; PICU: Pediatrics intensive care unit; RAG: Retrieval augmented generation; ROUGE-L: Recall-oriented understudy for gisting evaluation; SDoH: Social determinants of health; VS: versus

their physician colleagues while maintaining the highest standards of care [53].

Conclusion

The integration of LLMs into critical care is an evolving process that may become transformative in the future. As these models may increasingly permeate various aspects of patient management, it is imperative to avoid over-optimism by emphasising that current results are still far from actual application. That said, if developed and implemented correctly, these models could potentially improve clinical decision-making, alleviate the cognitive and administrative burdens on healthcare professionals, and improve patient and caregiver comprehension of the complexities associated with critical illness during and after hospitalization. The trajectory towards leveraging LLMs for improving patient care and possibly outcomes is increasingly evident, highlighting the need for responsible and evidence-based integration of these tools into critical care practice.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13054-025-05479-4.

Supplementary Material 1

Acknowledgements
We are grateful to Leehee Barak, Medint Data Analyst, for her valuable assistance in verifying the accuracy of the figures and terminology. We also extend our sincere thanks to Prof. Leo Anthony Celi for his insightful contributions to the conceptual development and overall flow of the paper.

Author contributions
D.I. and S.E. were responsible for the conceptualization of the manuscript idea, drafted the main manuscript text, and prepared Figs. 1, 2 and 3. All authors reviewed and approved the final manuscript.

Funding
None.

Data availability
No datasets were generated or analysed during the current study.

Declarations

Competing interests
Daphna Idan is a medical student and data analyst and has no relevant conflict of interests to disclose. Sharon Einav is a Cochrane Editor and is involved in developing an LLM for use by clinicians.

Received: 2 March 2025 / Accepted: 30 May 2025
Published online: 12 June 2025

References

1. Nestor Maslej L, Fattorini R, Perrault V, Parli A, Reuel E, Brynjolfsson J, Etchemendy K, Ligett T, Lyons J, Manyika JC, Niebles Y, Shoham R, Wald, Clark J. The AI Index 2024 Annual Report, AI Index Steering Committee, Institute for Human-Centered AI, Stanford University, Stanford, CA, April 2024.
2. Health C, for D. and R. Artificial Intelligence and Machine Learning (AI/ML)-Enabled Medical Devices. FDA [Internet]. 2021; Available from: https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-aiml-enabled-medical-devices
3. Silcox C, Zimlichmann E, Huber K, Rowen N, Saunders R, McClellan M et al. The potential for artificial intelligence to transform healthcare: perspectives from international health leaders. NPJ Digital Medicine [Internet]. 2024;7(1):1–3. Available from: https://www.nature.com/articles/s41746-024-01097-6
4. Russell SJ, Norvig P. Artificial intelligence: A modern approach, global edition.
5. Rockall AG, Shelmerdine SC, Chen M. AI and ML in radiology: Making progress. Clinical radiology [Internet]. 2023;78(2):81–2. Available from: https://pubmed.ncbi.nlm.nih.gov/36639174/
6. Kasneci E, Sessler K, Küchemann S, Bannert M, Dementieva D, Fischer F et al. ChatGPT for good? On opportunities and challenges of large language models for education. Learning and Individual Differences [Internet]. 2023;103(102274). Available from: https://www.sciencedirect.com/science/article/abs/pii/S1041608023000195
7. Zamani H. W. Bruce Croft. Embedding-based Query Language Models; 2016.
8. Chowdhary KR. Natural Language processing. Fundamentals Artif Intell. 2020;603–49.
9. Avijit Thawani S, Ghanekar, Zhu X, Pujara J. Learn Your Tokens: Word-Pooled Tokenization for Language Modeling [Internet]. OpenReview. 2023. Available from: https://openreview.net/forum?id=O9zrG7NB3X
10. Balta KY, Javidan AP, Walser E, Arntfield R, Prager R. Evaluating the appropriateness, consistency, and readability of ChatGPT in critical care recommendations. J Intensive Care Med. 2024;40(2):184–90.
11. Haw Hwai, Ho YJ, Wang CH, Huang CH. Large Language model application in emergency medicine and critical care. J Formos Med Assoc. 2024. https://doi.org/10.1016/j.jfma.2024.08.032
12. Find and participate in clinical trials and research. studies happening around the world | TrialScreen [Internet]. Trialscreen.org. 2025 [cited 2025 Feb 14]. Available from: https://app.trialscreen.org/trials/assessing-intensive-care-unit-icu-indications-human-vs-chatgpt-4o-predictions-study-nct06726733
13. Fernandes M, Mendes R, Vieira SM, Leite F, Palos C, Johnson A et al. Predicting Intensive Care Unit admission among patients presenting to the emergency department using machine learning and natural language processing. PLoS ONE [Internet]. 2020 Mar 3 [cited 2023 Mar 25];15(3):e0229331. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7053743/
14. Megumi Yazaki, Maki S, Furuya T, Inoue K, Nagai K, Nagashima Y et al. Emergency Patient Triage Improvement through a Retrieval-Augmented Generation Enhanced Large-Scale Language Model. Prehospital Emergency Care [Internet]. 2024;1–7. Available from: https://pubmed.ncbi.nlm.nih.gov/38950135/
15. Glicksberg BS, Timsina P, Patel D, Sawant A, Vaid A, Raut G et al. Evaluating the accuracy of a state-of-the-art large language model for prediction of admissions from the emergency room. Journal of the American Medical Informatics Association: JAMIA [Internet]. 2024;ocae103. Available from: https://pubmed.ncbi.nlm.nih.gov/38771093/
16. Miao J, Charat Thongprayoon WC. Should Artificial Intelligence Be Used for Physician Documentation to Reduce Burnout? Kidney360 [Internet]. 2024 Mar 25 [cited 2024 Aug 18];5(5):765–7. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC11146645/
17. 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(8):1018–20.
18. Urquhart E, Ryan J, Hartigan S, Nita C, Hanley C, Moran P et al. A pilot feasibility study comparing large Language models in extracting key information from ICU patient text records from an Irish population. Intensive Care Med Experimental. 2024;12:17.
19. Yi J, Gill SR, Gui G, Yan D, Ke Y, Ting F, Tan et al. Comparison of the Quality of Discharge Letters Written by Large Language Models and Junior Clinicians: Single-Blinded Study. Journal of Medical Internet Research [Internet]. 2024 Jul 24 [cited 2024 Sep 9];26:e57721–1. Available from: https://www.jmir.org/2024/1/e57721/
20. Guevara M, Chen S, Thomas S, Chaunzwa TL, Franco I, Kann BH et al. Large language models to identify social determinants of health in electronic health records. NPJ Digital Medicine [Internet]. 2024;7(1):1–14. Available from: https://www.nature.com/articles/s41746-023-00970-0
21. Simmons A, Takkavatakarn K, McDougal M, Dilcher B, Pincavitch J, Meadows L et al. Extracting international classification of diseases codes from clinical Documentation using large Language models. Appl Clin Inf. 2024;16(2):337–44.

22. Decker H, Trang K, Ramirez J, Colley A, Pierce L, Coleman M et al. Large Language Model – Based Chatbot vs Surgeon-Generated Informed Consent Documentation for Common Procedures. JAMA Network Open [Internet]. 2023 Oct 9 [cited 2023 Nov 29];6(10):e2336997. Available from: https://jamanetwork.com/journals/jamanetworkopen/article-abstract/2810364
23. R-AZ-1 D. Text - H.R.238–119th Congress (2025–2026): To amend the Federal Food, Drug, and Cosmetic Act to clarify that artificial intelligence and machine learning technologies can qualify as a practitioner eligible to prescribe drugs if authorized by the State involved and approved, cleared, or authorized by the Food and Drug Administration, and for other purposes. [Internet]. Congress.gov. 2025. Available from: https://www.congress.gov/bill/119th-congress/house-bill/238/text
24. Chen X, Zhou W, Hoda R, Li A, Bain C, Poon P. Exploring the opportunities of large language models for summarizing palliative care consultations: A pilot comparative study. Digital health [Internet]. 2024;10:20552076241293932. Available from: https://pubmed.ncbi.nlm.nih.gov/39569395/
25. Tu T, Anil Palepu, Schaekermann M, Saab K, Freyberg J, Tanno R, Towards Conversational Diagnostic AI., arXiv et al. (Cornell University). 2024.
26. Yang AA-A, Luchette Y, Burns M, Mehta JP, Alon Geva. Comparing the quality of Domain-Specific versus general Language models for artificial Intelligence-Generated differential diagnoses in PICU patients**. Pediatr Crit Care Med. 2024;25(6):e273–82.
27. Tommaso Scquizzato, Semeraro F, Swindell P, Simpson R, Angelini M, Gazzato A, et al. Testing ChatGPT ability to answer laypeople questions about cardiac arrest and cardiopulmonary resuscitation. Resuscitation. 2024;194:110077–7.
28. Suh PS, Shim WH, Suh CH, Heo H, Park CR, Eom HJ et al. Comparing diagnostic accuracy of radiologists versus GPT-4V and gemini pro vision using image inputs from diagnosis please cases. Radiology. 2024;312(1):e240273.
29. 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. https://doi.org/10.1007/s00330-024-11339-6
30. Wang J, Joao, Gupta S, Upadhyaya P, Lisboa FA, Schobel SA et al. ClotCatcher: a novel natural Language model to accurately adjudicate venous thromboembolism from radiology reports. BMC Med Inf Decis Mak. 2023;23(1):262.
31. Nolan VJ, Balch JA, Baskaran NP, Shickel B, Efron PA, Upchurch GR et al. Incorporating Patient Values in Large Language Model Recommendations for Surrogate and Proxy Decisions. Critical Care Explorations [Internet]. 2024;6(8):e1131–1. Available from: https://pubmed.ncbi.nlm.nih.gov/39132980/
32. Gandomi A, Wu P, Clement DR, Xing J, Aviv R, Federbush M et al. ARDSFlag: an NLP/machine learning algorithm to visualize and detect high-probability ARDS admissions independent of provider recognition and billing codes. BMC medical informatics and decision making [Internet]. 2024 Winter;24(1):195. Available from: https://pubmed.ncbi.nlm.nih.gov/39014417/
33. Liu T, Duan Y, Li Y, Hu Y, Su L, Zhang A. ChatGPT achieves comparable accuracy to specialist physicians in predicting the efficacy of high-flow oxygen therapy. Heliyon [Internet]. 2024;10(11):e31750. Available from: https://www.sciencedirect.com/science/article/pii/S2405844024077818
34. Weissman GE, Harhay MO, Lugo RM, Fuchs BD, Halpern SD, Mikkelsen ME. Natural Language processing to assess Documentation of features of critical illness in discharge documents of acute respiratory distress syndrome survivors. Annals Am Thorac Soc. 2016;13(9):1538–45.
35. Society of Critical Care Medicine. Critical care statistics [Internet]. Society of Critical Care Medicine (SCCM). 2024. Available from: https://www.sccm.org/Communications/Critical-Care-Statistics
36. De Corte T, Van Hoecke S, De Waele J. Artificial intelligence in infection management in the ICU. Crit Care. 2022;26(1):79.
37. Melo F, AF de, Macedo E, Fonseca Bezerra AC, de Melo WAL, Mehta RL, de A Burdmann E et al. G Remuzzi editor 2020 A systematic review and meta-analysis of acute kidney injury in the intensive care units of developed and developing countries. PLoS ONE 15 1 e0226325.
38. Orieux A, Prezelin-Reydit M, Prevel R, Combe C, Gruson D, Boyer A et al. Clinical trajectories and impact of acute kidney disease after acute kidney injury in the intensive care unit: a 5-year single-centre cohort study. Nephrology, dialysis, transplantation: official publication of the European Dialysis and Transplant Association - European Renal Association [Internet]. 2023 Autumn;38(1):167–76. Available from: https://pubmed.ncbi.nlm.nih.gov/35238922/
39. Sheikh MS, Thongprayoon C, Suppadungsuk S, Miao J, Qureshi F, Kashani K et al. Evaluating ChatGPT's Accuracy in Responding to Patient Education Questions on Acute Kidney Injury and Continuous Renal Replacement Therapy. Blood Purification [Internet]. 2024 Apr 26 [cited 2024 Jul 20];1–7. Available from: https://pubmed.ncbi.nlm.nih.gov/38679000/
40. Howard A, Hope W, Gerada A. ChatGPT and antimicrobial advice: the end of the consulting infection doctor? Lancet Infect Dis. 2023;23(4):405–6
41. Miranda F, Gonzalez F, Plana MN, Zamora J, Quinn TJ, Seron P. Confusion Assessment Method for the Intensive Care Unit (CAM-ICU) for the diagnosis of delirium in adults in critical care settings. The Cochrane Database of Systematic Reviews [Internet]. 2023;11(11):CD013126. Available from: https://pubmed.ncbi.nlm.nih.gov/37987526/
42. Contreras M, Kapoor S, Zhang J, Davidson A, Ren Y, Guan Z et al. DeLLiriuM: A large language model for delirium prediction in the ICU using structured EHR [Internet]. arXiv.org. 2024 [cited 2025 Feb 14]. Available from: https://arxiv.org/abs/2410.17363
43. Parmanto B, Aryoyudanta B, Soekinto TW, Setiawan IMA, Wang Y, Hu H et al. A Reliable and Accessible Caregiving Language Model (CaLM) to Support Tools for Caregivers: Development and Evaluation Study. JMIR Formative Research [Internet]. 2024 Jul 31 [cited 2024 Oct 14];8:e54633. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC11325100/
44. Zaleski AL, Berkowsky R, Jean K, Pescatello LS. Comprehensiveness, accuracy, and readability of exercise recommendations provided by an AI-Based chatbot: mixed methods study. JMIR Med Educ. 2024;10:e51308–8.
45. 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(4):319–28.
46. Komorowski M, Del M, Chang AC. How could ChatGPT impact my practice as an intensivist? An overview of potential applications, risks and limitations. Intensive Care Med. 2023;49:844–7.
47. Hager P, Jungmann F, Holland R, Bhagat K, Hubrecht I, Knauer M, et al. Evaluation and mitigation of the limitations of large Language models in clinical decision-making. Nat Med. 2024;30(9):2613–22.
48. Roberts J, Baker M, Andrew J. Artificial intelligence and qualitative research: The promise and perils of large language model (LLM) assistance. Critical Perspectives on Accounting [Internet]. 2024;99:102722. Available from: https://www.sciencedirect.com/science/article/pii/S1045235424000212
49. Buçinca Z, Malaya MB, Gajos KZ. To Trust or to Think: Cognitive Forcing Functions Can Reduce Overreliance on AI in AI-assisted Decision-making. Proceedings of the ACM on Human-Computer Interaction. 2021;5(CSCW1):1–21.
50. Jabbour S, Fouhey D, Shepard S, Valley TS, Kazerooni EA, Banovic N et al. Measuring the Impact of AI in the Diagnosis of Hospitalized Patients: A Randomized Clinical Vignette Survey Study. JAMA [Internet]. 2023;330(23):2275–84. Available from: https://jamanetwork.com/journals/jama/article-abstract/2812908
51. Goh E, Gallo R, Hom J, Strong E, Weng Y, Kerman H, et al. Large Language model influence on diagnostic reasoning. JAMA Netw Open. 2024;7(10):e2440969.
52. Gallifant J, Afshar M, Ameen S, Yindalon Aphinyanaphongs, Chen S, Cacciamani G et al. The TRIPOD-LLM reporting guideline for studies using large language models. Nature Medicine [Internet]. 2025;31. Available from: https://www.nature.com/articles/s41591-024-03425-5
53. Cecconi M, Greco M, Shickel B, Vincent JL, Azra Bihorac. Artificial intelligence in acute medicine: a call to action. Crit Care. 2024;28:258.

Publisher's note