Review

Large Language Models for Electronic Health Records in Emergency Department (ED) and ED-Adjacent Emergency Care—Technical Strategies, Clinical Tasks, and Workflow Integration: Systematic Review

Zhuohao Wu[1,2], BSc; Huilin Yu[2], BSc; Peng Deng[3], MD, PhD; Di Liu[1,2,4], PhD; Yong Zhao[5], PhD; Yu Cao[3*], PhD, MD; Kang Li[1,4*], PhD; Rong Yin[1,2,4*], PhD

[1]West China Biomedical Big Data Center, West China Hospital of Sichuan University, Chengdu, Sichuan, China

[2]Department of Industrial Engineering, Sichuan University, Chengdu, Sichuan, China

[3]Department of Emergency Medicine, West China Hospital of Sichuan University, Chengdu, Sichuan, China

[4]Med-X Center for Informatics, West China Hospital of Sichuan University, Chengdu, Sichuan, China

[5]Department of Computer Science, Sichuan University, Chengdu, Sichuan, China

[*]these authors contributed equally

Corresponding Author:
Rong Yin, PhD
West China Biomedical Big Data Center
West China Hospital of Sichuan University
No.37, Guoxue Lane, Wuhou District
Chengdu, Sichuan, 610207
China
Phone: 86 02885990100268
Email: rong.yin@scupi.cn

Abstract

Background: Large language models (LLMs) are increasingly applied in clinical text processing, but their role in emergency departments (EDs) and ED-adjacent emergency care remains unclear because ED electronic health records (EHRs) are heterogeneous, time-critical, and safety-sensitive.

Objective: This study aims to synthesize peer-reviewed evidence on LLM applications to real-world ED EHR data and ED-adjacent emergency care data, focusing on clinical tasks, technical strategies, evaluation practices, and workflow integration.

Methods: We conducted a PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses)–guided systematic review of PubMed and Web of Science (January 1, 2022, to November 14, 2025). We included peer-reviewed English-language studies applying LLM strategies to ED and ED-adjacent EHR data. We extracted dataset characteristics, model families, task formulations (classification, summarization, and information extraction), adaptation strategies (prompting, fine-tuning, retrieval-augmented generation, and agentic/hybrid workflows), evaluation metrics, and ED workflow phase. Two reviewers also independently completed a design-matched methodological quality and risk-of-bias appraisal.

Results: Thirty-one studies met our inclusion criteria. Many relied on single-center datasets (n=12), relatively small sample sizes (<10,000 encounters; n=20), and English-language unstructured text alone or mixed EHR inputs combining text with structured or semistructured fields. Prompt engineering was the dominant adaptation strategy (n=27), while fine-tuning, retrieval-augmented generation, and hybrid/agentic pipelines were less common. Postencounter studies frequently demonstrated task-specific retrospective technical feasibility, while triage, diagnostic-support, and disposition studies evaluated heterogeneous tasks and outcomes; these nonequivalent measures were not compared as indicators of workflow maturity. Workflow-phase assignment did not imply real-time integration, particularly when model inputs were available only after the relevant clinical decision point. Overall, 22 (71%) studies were classified as technical-performance-only evaluations; clinical utility was empirically assessed in only 3 (9.7%) studies, and prospective deployment in 4 (12.9%). No study directly measured patient outcomes or downstream patient-safety outcomes attributable to LLM use.

XSL•FO
RenderX

Conclusions: Current LLM applications across ED and ED-adjacent emergency care primarily function as documentation, surveillance, and auditing assistants rather than autonomous decision engines. Progress toward safe, deployable systems will require larger multicenter datasets, external and prospective validation, safety- and fairness-centered evaluation, and governance mechanisms that support responsible clinical integration. However, most evidence remains technical or retrospective, while prospective workflow integration, empirically demonstrated clinical utility, patient safety, and patient benefit remain insufficiently evaluated.

(J Med Internet Res 2026;28:e95400) doi: 10.2196/95400

KEYWORDS

EHR; electronic health record; emergency medicine; generative AI; large language model; LLM; systematic review; workflow integration

Introduction

Large language models (LLMs) are a class of generative AI models, typically based on transformer architectures, that learn statistical regularities from large-scale corpora and can generate fluent, context-aware text. Recent model families such as GPT, PaLM, Llama, Gemini, and Claude have demonstrated strong performance across natural-language tasks, including question answering, summarization, and information extraction [1-5]. These capabilities have accelerated interest in clinical applications, especially in text-heavy settings where rapid synthesis of information is essential.

Electronic health records (EHRs) are central to modern diagnosis and treatment but remain difficult to use efficiently because much of the clinically relevant information is embedded in unstructured narratives, fragmented across time, and recorded under severe workflow constraints [6,7]. In emergency departments (EDs), clinicians must make high-stakes decisions with limited time while navigating dense documentation, triage notes, imaging reports, and handover summaries. LLMs may help by automating documentation, extracting key features, and supporting clinical communication and decision support [4,8,9]. In this review, we focused on language foundation models capable of performing generative clinical text tasks, including decoder-only models and encoder-decoder models with text-generation capabilities. Encoder-only BERT-like models used solely for discriminative representation learning and lacking generative capabilities were excluded.

Despite rapid progress, LLM deployment in health care remains constrained by technical and governance challenges. Limited transparency and hallucinations can produce plausible but incorrect statements [7,10]. Training data are often narrow and institution-specific, which may impair generalizability and exacerbate bias [6,11,12]. Privacy, regulatory compliance, and cross-border data transfer further shape feasible deployment options, frequently favoring on-premises or tightly controlled infrastructures [11-13]. These concerns are particularly acute in ED workflows, where errors can propagate quickly and oversight time is limited.

Prior reviews of LLMs in health care have largely focused on general technical frameworks, evaluation approaches, or ethical considerations [2,5,14]. However, a dedicated synthesis mapping LLM applications that operate on real-world ED EHR data and ED-adjacent emergency care data across the broader emergency-care pathway is lacking. Workflow integration refers to the extent to which an LLM-based tool is embedded in an ED or ED-adjacent process with specified users, timing of use, information available at that point, actionable outputs, human oversight, safety mechanisms, and measurable workflow or patient-care outcomes. To address this gap, we conducted a PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses)–guided systematic review of studies applying LLMs to ED EHRs and ED-adjacent emergency care data, summarizing clinical tasks, technical strategies, evaluation practices, and opportunities for safe workflow integration [4,15]. The synthesis examines how methodological quality, the timing of model inputs, and the outcomes actually assessed qualify the evidence for each intended clinical use.

Methods

Search Strategy

We conducted a systematic review following the PRISMA guidelines. The completed PRISMA 2020 checklist is provided in Multimedia Appendix 1. We searched PubMed and Web of Science for articles published from January 1, 2022, to November 14, 2025. Database-specific search strings were developed using the keywords listed in Multimedia Appendix 2. The final search was completed on November 14, 2025. The search yielded 260 records from PubMed and 200 from Web of Science. The search strategy was developed by the review team and was not formally peer-reviewed using the PRESS (Peer Review of Electronic Search Strategies) guideline or validated by an information specialist.

An internal review protocol was developed, but the review was not prospectively registered in PROSPERO (International Prospective Register of Systematic Reviews) or another public registry, and the protocol was not publicly available. The original review synthesized study characteristics, LLM methods, core natural language processing (NLP) tasks, workflow-phase assignments, and reported performance. During peer review, we added the design-matched methodological appraisal, detailed workflow integration and temporal availability coding, clinical/implementation evidence coding, and phase-by-task cross-tabulation to clarify the clinical relevance and limitations of the evidence. These additions were developed after the initial synthesis and were treated as post hoc descriptive assessments.

XSL•FO
RenderX

Inclusion and Exclusion Criteria

Two reviewers (ZW and HY) independently screened all titles and abstracts and then independently assessed potentially eligible full-text reports. Disagreements were resolved through discussion; unresolved disagreements were adjudicated by a third reviewer (RY).

For this review, the eligible data scope comprised (1) real-world ED EHR data and (2) ED-adjacent emergency-care records when both the analyzed records and the intended task were directly linked to the emergency-care pathway. ED-adjacent records included, for example, prehospital emergency medical services (EMS) records used for triage, severity assessment, or early disease identification. General inpatient or outpatient records without a direct link to emergency care were excluded.

After deduplication, records were excluded if they met any of the following criteria: (1) not a peer-reviewed research article (eg, preprints, editorials, comments, reviews, or meta-analyses), (2) non-English, (3) full text not available, (4) did not apply an LLM strategy, (5) used an LLM only as a minor component within a broader non-LLM pipeline, (6) was not conducted in an ED or ED-adjacent emergency care setting, (7) did not analyze ED or ED-adjacent emergency care EHR data, or (8) did not target a clinical task directly linked to the ED or ED-adjacent emergency-care pathway. Eligible studies were peer-reviewed, English-language papers that primarily applied language foundation models with generative text capabilities to tasks based on real-world ED EHR data or ED-adjacent emergency-care records. Consistent with the definition adopted in this review, eligible models included decoder-only models and encoder-decoder models capable of text generation. Studies relying solely on encoder-only BERT (Bidirectional Encoder Representations from Transformers)–like models for discriminative representation learning and lacking generative capabilities were excluded. Eligibility was therefore determined by the generative capability of the model rather than by the downstream task type. Both peer-reviewed journal articles and full papers in peer-reviewed conference proceedings were eligible; preprints and conference abstracts without a full peer-reviewed report were excluded.

Data Extraction and Classification

For each eligible study, we extracted variables across 3 analytic dimensions: data characteristics, modeling strategy, and ED workflow context. Two reviewers (ZW and HY) independently extracted study characteristics, model and task information, and reported performance results using a standardized form. Discrepancies were resolved by discussion, with adjudication by a third reviewer (RY) when necessary.

First, we recorded dataset characteristics, including source type (open clinical database vs institution-specific EHR), center structure (single- vs multicenter), sample size, accessibility, language (English, non-English, or mixed), and input composition (unstructured text alone, mixed EHR input combining text with structured or semistructured fields, or structured EHR items serialized into prompts). In this review, "mixed EHR input" does not denote raw clinical multimodality such as imaging, audio, video, or bedside observations. We also noted whether evaluation involved clinicians (human assessment) or relied on automated metrics and/or comparisons with non-LLM baselines.

Second, we extracted LLM and method information, including the model family (eg, GPT, Llama, Gemini, and Claude), whether multiple LLMs were compared, the presence of non-LLM baselines, the core NLP task formulation (classification, generation/summarization, or information extraction based on a HELM-aligned taxonomy), the adaptation strategy (prompting, fine-tuning, retrieval-augmented generation [RAG], and agentic or hybrid workflows), and all reported performance metrics. Metrics were subsequently mapped to a multidimensional evaluation framework.

Third, we captured ED task characteristics by assigning each study to a clinical workflow location using an adaptation of the input-throughput-output model of emergency care described by Asplin et al [16]. To preserve clinically specific distinctions among LLM use cases, we retained 4 operational subphases within this broader framework: prehospital care and ED triage were classified as input-phase applications because they involve information generated before or at entry into emergency care; initial assessment and diagnostic support were classified as throughput because they occur during ED evaluation and care delivery; patient disposition and handover were classified as output because they concern admission, discharge, transfer, or transfer of responsibility; and postencounter analysis and quality improvement were classified as downstream use of ED information after the index encounter. The grouping of prehospital care and ED triage as input did not imply that EMS and ED triage share the same organizational workflow, documentation system, legal responsibility, or end user group. For each phase, we also summarized the intended end users and use settings, distinguishing clinical support from retrospective analysis or secondary use. Selected study-level coding variables for all 31 studies—including data source and input composition, workflow phase, clinical task, target user, core NLP task, model family, adaptation method, clinical/implementation evidence, temporal availability, and challenge domains—are provided in Multimedia Appendix 3. Workflow phase was assigned according to intended clinical function, whereas temporal availability was coded separately according to whether the reported inputs were available at, later than, or not clearly linked to the intended decision point.

To be specific, in this review, technical feasibility referred to whether an LLM could perform the specified task under the study's evaluation conditions. Workflow integration required evaluation within a prospective clinical data stream, simulated clinician workflow, or active operational process, depending on the reported design. Real-time clinical support was considered supported only when the information used by the model was available at or before the relevant clinical decision point and the study evaluated use within, or directly applicable to, that decision context.

Finally, to distinguish model-level performance from clinical and implementation evidence, each included study was additionally coded across 8 dimensions: clinical utility, clinician acceptance, workflow effect, time saving, patient outcome,

XSL•FO
RenderX

safety outcome, prospective deployment, and technical-performance-only evaluation. For the first 7 dimensions, evidence was classified as "assessed" when the study empirically evaluated the corresponding outcome, "discussed only" when the potential benefit or implication was described without empirical evaluation, and "not assessed" when the dimension was neither empirically evaluated nor substantively discussed. Clinical utility was defined as empirical evidence that LLM use supported or improved the intended clinical task or decision beyond model-performance metrics. Technical-performance-only studies were those whose empirical evidence was limited to model-level measures without user-, workflow-, time-, patient-, safety-, or prospective deployment evidence. The evidence classifications were independently reviewed by 2 investigators using the operational criteria used for the revision-stage coding, with discrepancies reconciled by consensus. The resulting study-level clinical/implementation classifications are provided in Multimedia Appendix 3.

Methodological Quality and Risk-of-Bias Assessment

The included studies evaluated model performance, reference-standard accuracy, comparative workflow effects, and human judgments of generated content or assistant usability. These evaluation structures expose results to different biases; applying one instrument to all studies would leave relevant biases unassessed or require domains to be used outside their intended scope. We therefore selected one primary tool according to the study design and the outcome retained for synthesis. Prediction Model Risk of Bias Assessment Tool for prediction models using regression or artificial intelligence methods (PROBAST+AI) was used when appraisal centered on the predictive or classification performance of a model or pipeline, for which input definition, model selection, separation of development and evaluation data, and performance analysis were central concerns [17]. Quality Assessment of Diagnostic Accuracy Studies 3 (QUADAS-3) was used when the retained outcome was an LLM-based index test's accuracy for a defined target condition against a reference standard, focusing on case spectrum, test and reference-standard procedures, and accuracy analysis [18]. This distinction followed the evaluated comparison, not whether the task was called classification or extraction; study-specific targets and assignments are recorded in Multimedia Appendix 4. The 2 comparisons of clinician writing time with and without LLM assistance required Risk of Bias in Non-Randomized Studies of Interventions (ROBINS-I; 2016) because confounding, intervention order, practice effects, and outcome measurement could affect the estimated workflow benefit [19]. Alberta Quality Assessment Tool: Risk of Bias (AQAT:RoB) was used for human-evaluated generated content and assistant usability, where question and answer selection, reference material, evaluator selection and blinding, and rating methods could influence judgments independently of model accuracy [20].

Two reviewers (ZW and HY) independently completed the applicable signaling questions or support-for-judgment assessments and domain ratings. Disagreements were resolved by discussion or adjudication by a third reviewer (RY). Each study contributed one primary appraisal, with the outcome closest to its stated clinical objective retained when judgments differed across outcomes. Development quality and applicability were distinguished from evaluation risk of bias. For text-labeling models, inputs and reference labels were assessed as predictors and outcomes at the evaluated unit. AQAT:RoB required context-specific interpretation for assistant usability; reference-answer source and selection were assessed when answers formed part of the evaluation, with insufficient reporting rated Unclear. Overall judgments for PROBAST+AI and QUADAS-3 followed their published domain-to-overall rules; ROBINS-I used its ordered risk categories, with the most serious domain limiting the overall judgment. AQAT:RoB leaves the overall threshold to users, so this review defined low as all applicable subdomains low, unclear as at least one unclear and no high subdomain, and high as a material high-risk subdomain. Findings were synthesized within each tool without conversion to a common risk score. Their implications for the conclusions were considered alongside applicability, validation setting, precision, and directness to clinical use. Complete study-to-tool assignments, primary and secondary outcome judgments, domain ratings, item-level assessments, and rationales are provided in Multimedia Appendix 4.

Study-Level Evidence Synthesis

The heterogeneity of clinical tasks, outcomes, and performance metrics precluded quantitative pooling. We synthesized findings by workflow phase, describing the task, model inputs, evaluation setting, and reported performance of the contributing studies. We then summarized evaluation coverage and empirical clinical and implementation outcomes. In interpreting the findings, we considered the relevant study-level appraisal, applicability, validation setting, and availability of inputs at the intended point of use. Performance metrics were retained in their original task context, and risk-of-bias judgments were reported within each tool without conversion to a common score. No formal outcome-level Grading of Recommendations Assessment, Development, and Evaluation (GRADE) assessment was conducted.

To examine whether the distribution of core NLP tasks differed across ED workflow phases, we cross-tabulated the 4 workflow phases against the 3 nonmutually exclusive core NLP task categories. Because a study could contribute to more than 1 task category, percentages within a workflow phase could sum to more than 100%. Study-level assignments and the phase-by-task cross-tabulation are provided in Multimedia Appendix 5.

Results

Overview

Study characteristics describe the data and technical approaches underlying the applications. The methodological appraisal then qualifies the workflow-phase synthesis, which relates task-specific findings to input timing and evaluation setting. Evaluation coverage and clinical/implementation evidence are reported separately to distinguish the dimensions examined from demonstrated effects of model use.

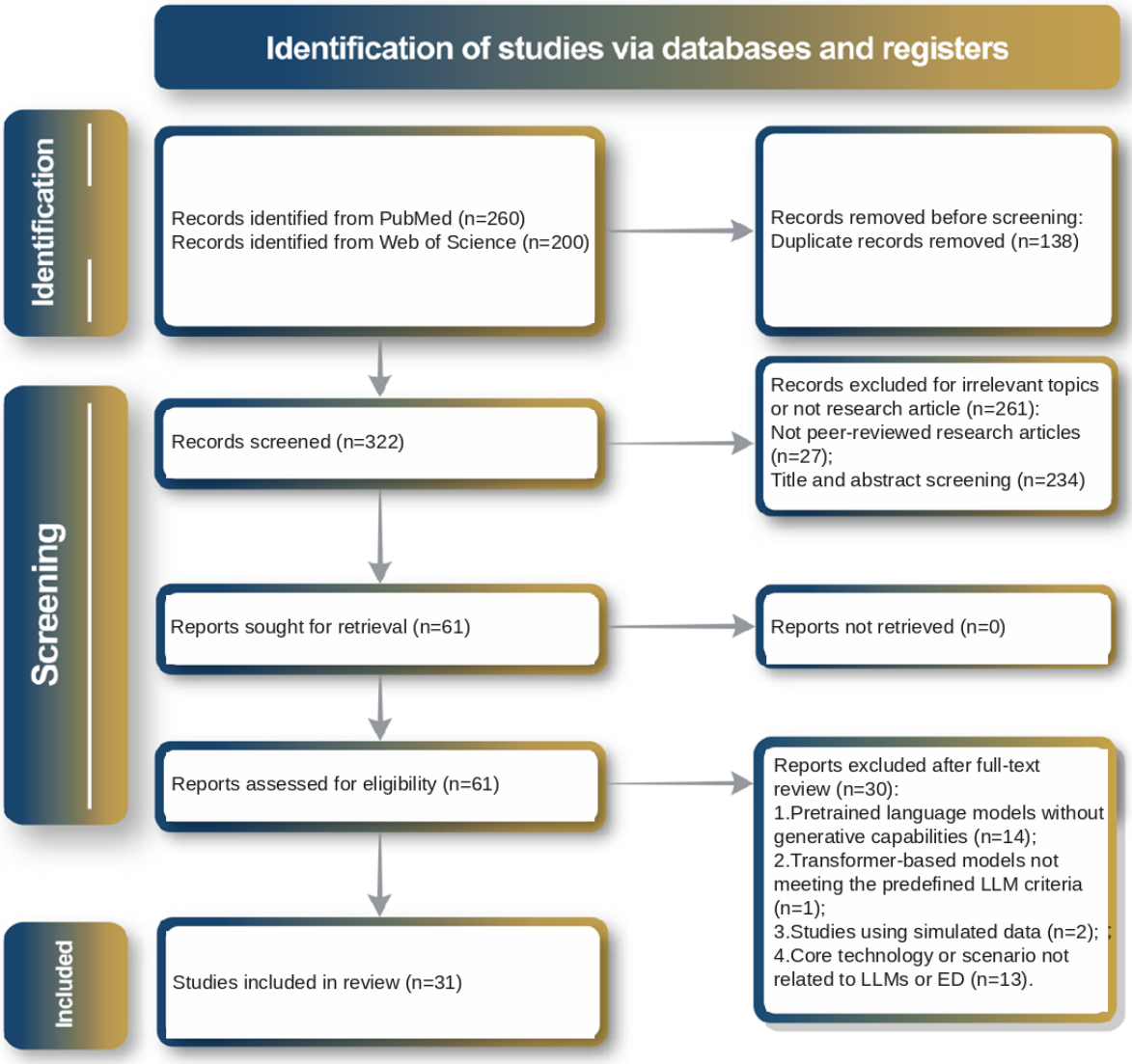
Study Selection

As shown in Figure 1, the initial search results included 460 articles: 260 (56.52%) from PubMed and 200 (43.48%) from

XSL•FO
RenderX

Web of Science. After removing 138 duplicates, 322 articles proceeded to the screening phase. During title and abstract screening, another 261 papers were excluded, including 27 that were not peer-reviewed research articles. The remaining 61 papers were all accessible and written in English. After full-text screening, 31 studies were included according to the predefined inclusion criteria. Of the 30 reports excluded after full-text review (Figure 1), 14 applied pretrained language models (BERT-based models) to analyze EHR data and accomplish clinical tasks in EDs; these models were not included in the definition of LLM in this systematic review.

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) 2020 flow diagram for study identification, screening, eligibility assessment, and inclusion. ED: emergency department; LLM: large language model.



Study Characteristics

Overview

Study characteristics are summarized by data source and setting, model family, NLP task formulation, and adaptation strategy. These characteristics describe the evidence base for the subsequent synthesis by clinical workflow phase; model use and task assignment do not themselves indicate clinical integration.

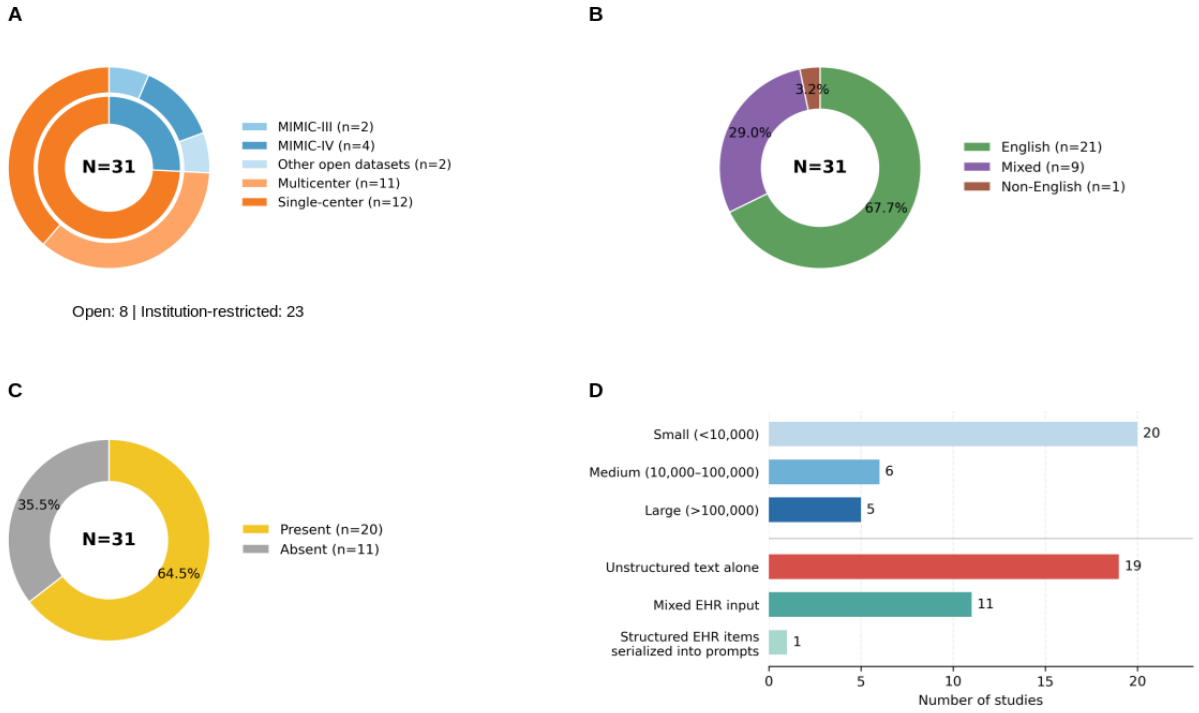
Data Sources and Study Settings

Included studies used datasets ranging from open clinical databases (n=8) to institution-restricted EHR datasets (n=23). Among the open datasets, the Medical Information Mart for Intensive Care (MIMIC) was most commonly used, comprising MIMIC-III (n=2) and MIMIC-IV (n=4). Among the 23 studies using institution-restricted data, 11 used multicenter datasets and 12 used single-center datasets. Most studies used English free text (n=21), while a minority used non-English (n=1) or mixed-language notes (n=9). Given substantial heterogeneity in training corpora and reporting, we categorized dataset scale by sample size as small (<10,000 encounters; n=20), medium

XSL•FO
RenderX

(10,000-100,000; n=6), or large (>100,000; n=5). After study-level source verification, 19 studies used unstructured text alone, 11 used mixed EHR inputs combining text with structured or semistructured fields, and 1 used structured EHR data items serialized into text prompts. More than half of the studies incorporated human evaluation (n=20), whereas the remainder relied on automated metrics and/or comparisons with traditional machine learning baselines (n=11). No included study incorporated raw clinical modalities such as imaging, audio, video, or bedside observations. These characteristics are summarized in Figure 2.

Figure 2. Overview of the reviewed literature: (A) Dataset source and accessibility, (B) Language of clinical records, (C) Human evaluation, and (D) Dataset scale and input composition. EHR: electronic health record; MIMIC-III: Medical Information Mart for Intensive Care.



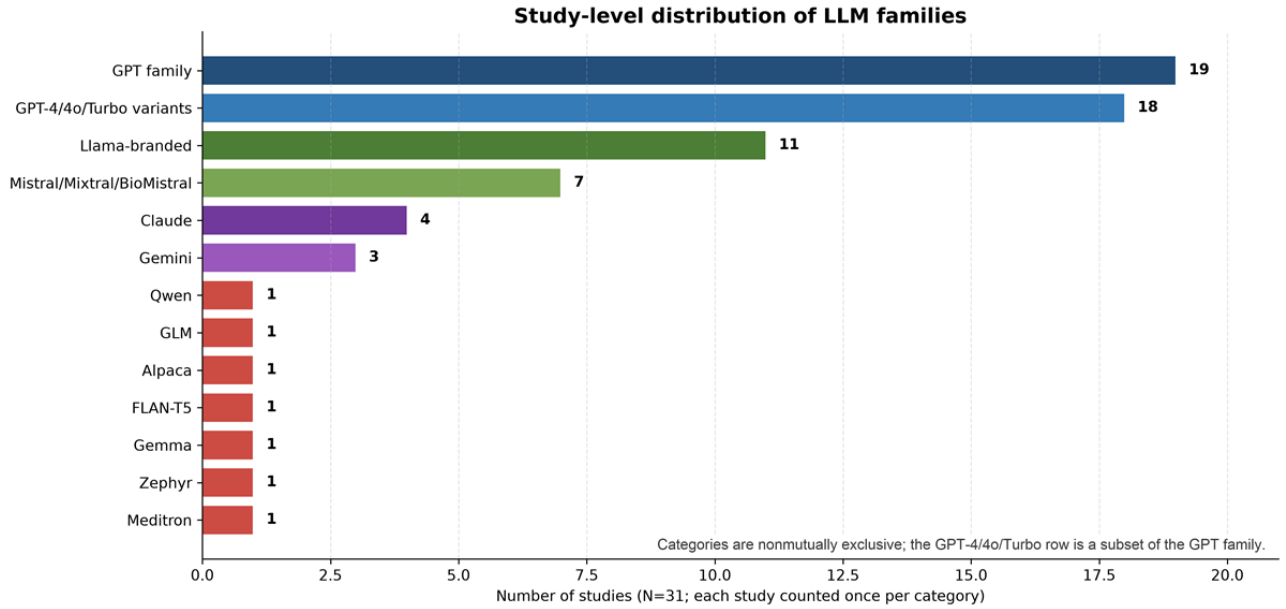
Model Families

In the included studies, various LLMs were used (Figure 3). The GPT series (n=19, including GPT-3, GPT-3.5, GPT-4, and GPT-4o) was the most frequently evaluated [12,21-38]. The Llama family of models (encompassing Llama-2, Llama-3, and their variants) was also widely used, appearing in 11 studies [26,30,33,35,39-45]. Models from the Gemini [28,35,46] and Claude [28,33,35,47] series were evaluated in several comparative studies. Mistral, Mixtral, or BioMistral models appeared in 7 studies. Qwen, GLM, Alpaca, FLAN-T5, Gemma, Zephyr, and Meditron each appeared in 1 study. Full assignments are provided in Multimedia Appendix 3.

Among the 31 studies, 15 used multiple LLMs for comparative analysis [12,21,22,26,28,30,33-36,39,41-43,48], while 16 used a single LLM [23-25,27,29,31,32,37,38,40,44-47,49,50]. The distribution underscored a methodological focus on benchmarking performance across both proprietary and open-source model families within the emergency care domain.

XSL•FO
RenderX

Figure 3. Distribution of the large language models (LLMs) used in the reviewed studies. Each family is counted at most once per study; a study may contribute to more than 1 family.



NLP Task Formulations

This systematic review strictly adhered to the classification of NLP tasks in the holistic evaluation of language models (HELM) framework [51] and, based on the characteristics of the 31 selected clinical studies, summarized the core NLP tasks into the following three categories: (1) text classification; (2) text generation and summarization; and (3) information extraction, which are defined in Table 1. In fact, each selected paper may have multiple NLP tasks, and this review determined the core NLP tasks (which could be more than 1) based on the NLP tasks for which the research assessed performance, while the remaining auxiliary NLP tasks involved in the intermediate steps, such as text reasoning and translation, were not included.

Of the 31 included studies, text classification emerged as the most prevalent task, followed closely by text generation and information extraction (Table 2).

Table 1. NLPa tasks conducted in the corresponding reviewed studies. Task categories were nonmutually exclusive.

NLP[a] task	Definition	Included studies
Text classification	Assigning 1 or more predefined labels to textual or text-serialized EHR[b] inputs [52].	[12,22-24,28,29,31,33,36,37,41,48-50]
Text generation and summarization	Producing new natural-language text or condensing and reorganizing information from an input source while maintaining coherence and, where a source text is available, fidelity to that source [53].	[21,25-27,34,35,38,40,44,46,47]
Information extraction	Detecting and structuring salient entities, relations, or attributes from unstructured text into machine-readable records [54].	[22,25,30,32,39,42,43,45,50]

[a]NLP: natural language processing.

[b]EHR: electronic health record.

XSL•FO
RenderX

Table 2. Distribution of core NLP[a] task assignments across the included studies. A study could contribute to more than 1 core task category; subcategory percentages use the corresponding task-category count as the denominator.

Major NLP domain and application category	Studies, n (%)
Text classification (n=14)	
Diagnosis, phenotyping, and coding	6 (42.9)
Triage and severity assessment	3 (21.4)
Prognosis and disposition prediction	3 (21.4)
Social risk and quality monitoring	2 (14.3)
Text generation and summarization (n=11)	
Documentation generation and summarization	5 (45.5)
Clinical decision and communication	4 (36.4)
Synthetic data and training support	2 (18.2)
Information extraction (n=9)	
Symptom, sign, and injury extraction	3 (33.3)
Specific finding extraction from reports	2 (22.2)
Public health and risk monitoring	2 (22.2)
Privacy protection and deidentification	2 (22.2)

[a]NLP: natural language processing.

Text classification comprised 14 studies (Table 1), primarily operationalizing four families of clinical tasks: (1) outcome and disposition prediction, (2) triage and acuity scoring, (3) diagnosis and phenotype identification, and (4) detection of social risk factors and documentation quality issues. Outcome-related models classified 30-day readmission risk, hospital admission, or discharge versus admission recommendations from ED records [24,36,49]. Triage and acuity applications predicted Emergency Severity Index level, stroke large-vessel-occlusion status, or generic severity scores from prehospital or ED complaints [12,33,48]. Diagnostic and phenotype classifiers targeted specific conditions or symptom clusters, including incidental hepatic steatosis, common ED diseases, mental-health symptoms, *ICD-11* (*International Classification of Diseases, 11th Revision*) diagnosis categories, and psychiatric diagnostic prediction [22,23,28,31,37,50]. Finally, several systems detected incarceration history or stigmatizing language as proxies for social risk and documentation bias [29,41].

Across 11 included studies, text generation and summarization systems served three families of use cases: (1) documentation-support tools automatically produced neurology consultation summaries [46], ED handoff notes [40], longitudinal ED course summaries [27], and discharge reports from existing narrative records or EHR timelines [35,44]; (2) decision-support and communication tools generated GPT-assisted differential diagnosis lists [21,38] and plain-language instructions for actionable incidental radiology findings, aiming to reduce cognitive load and standardize counseling [25], as well as record-retrieval and synthesis support, which was evaluated in TanaGPT [47]; and (3) data-centric applications used local or open-weight models to synthesize radiology reports and emergency dialogues as training data for downstream detection models or benchmarking conversational agents [26,34].

In total, 9 information-extraction studies primarily focused on converting unstructured ED narratives into structured clinical features that could drive downstream surveillance or decision-support models. Symptom- and sign-extraction systems identified sepsis-related manifestations [39], common cardiorespiratory complaints [30], and structured trauma descriptors such as mechanism and injured body region from provider notes [45]. Several papers targeted specific findings in radiology reports, including incidental hepatic steatosis [22] and other actionable incidental findings, which were then linked to alerts or patient-facing instructions [25]. Public-health-oriented pipelines extracted helmet-wearing status in micromobility injuries [32] and vaccine mentions from triage notes to support injury-prevention research and vaccine safety monitoring [42]. Privacy-preserving and psychiatric-record pipelines respectively involved deidentification of protected health information spans before secondary analysis and diagnostic-clue extraction from non-English notes [43,50].

Model Adaptation Strategies

General-purpose LLMs can be adapted to specialized ED tasks using optimization strategies that improve domain alignment, factual grounding, and workflow fit [55-57]. We summarize the main approaches reported in the included studies.

Prompt engineering refers to the systematic design and refinement of input instructions to elicit task-specific behavior from a pretrained model without changing its parameters [55]. In clinical NLP, prompting can simulate expert reasoning and enforce structured outputs [58]. In our review, prompting was the most common strategy (27 studies), ranging from zero-shot prompts (n=22) to few-shot in-context learning (n=9). Six studies explicitly used chain-of-thought prompting to encourage

XSL•FO
RenderX

intermediate reasoning. Nine studies compared multiple prompting strategies and generally reported improvements over zero-shot baselines [22,24,26,31-33,36,41,48].

Fine-tuning updates a model's weights using domain-specific data to improve performance on downstream tasks [56]. Four studies used confirmed parameter-efficient fine-tuning with low-rank adaptation (LoRA) or quantized low-rank adaptation (QLoRA) [33,42,43,45]. Hartman et al [40] and Song et al [44] used supervised or instruction fine-tuning, but the reports did not establish that these were LoRA/QLoRA-based parameter-efficient procedures. One study used a provider-managed standard fine-tuning procedure [37]. Chen et al [41] fine-tuned conventional encoder comparators rather than the evaluated decoder LLMs. These approaches were most often applied to open-weight models in resource-constrained settings. No reviewed study reported targeted fine-tuning of very large, state-of-the-art LLMs on multicenter ED data.

RAG combines an LLM with a retrieval component that supplies relevant external context at inference time (eg, guidelines, similar cases, or prior notes) [57]. Inference- or generation-grounding RAG appeared in 5 studies [24,39,46-48]. Moser et al [34] used RAG only for automated postgeneration evaluation, which was reported separately. Retrieval most commonly drew on semantically similar historical cases or relevant segments from a patient's own record. Retrieval was largely confined to local ED data; no study integrated LLMs with external guideline libraries, curated case repositories, or real-time literature.

Agentic and hybrid workflows decompose complex tasks into multistep pipelines, often combining multiple specialized components. In hybrid systems, LLMs typically handle language-intensive subtasks (eg, extraction or summarization), while conventional models perform prediction subtasks. Two studies implemented agentic designs for documentation and information-access support, and 7 used custom or hybrid pipelines combining LLMs with conventional models, retrieval, task-specific preprocessing, or multistage orchestration [23,24,33,35,39,40,48]. Multiagent systems were implemented in 2 studies, in which LLM agents performed data extraction, summarization, and recommendation refinement [35,47].

Methodological Quality and Risk-of-Bias Findings

All 31 studies underwent a single design-matched primary appraisal [12,21-50]. Of the 18 studies assessed with PROBAST+AI, 11 were unclear overall [23,24,29,30,32,33,39,41,42,48,49], 4 high [25,37,43,46], and 3 low [26,36,45]. Analysis was the most frequent source of concern, with 12 unclear and 2 high judgments; participants and data sources were unclear in 5 studies and high in 2. High judgments reflected selected samples, test-informed model selection, or inconsistent reported metrics (Figure 4A). These concerns limit confidence in the reported discrimination, calibration, and transportability of the models.

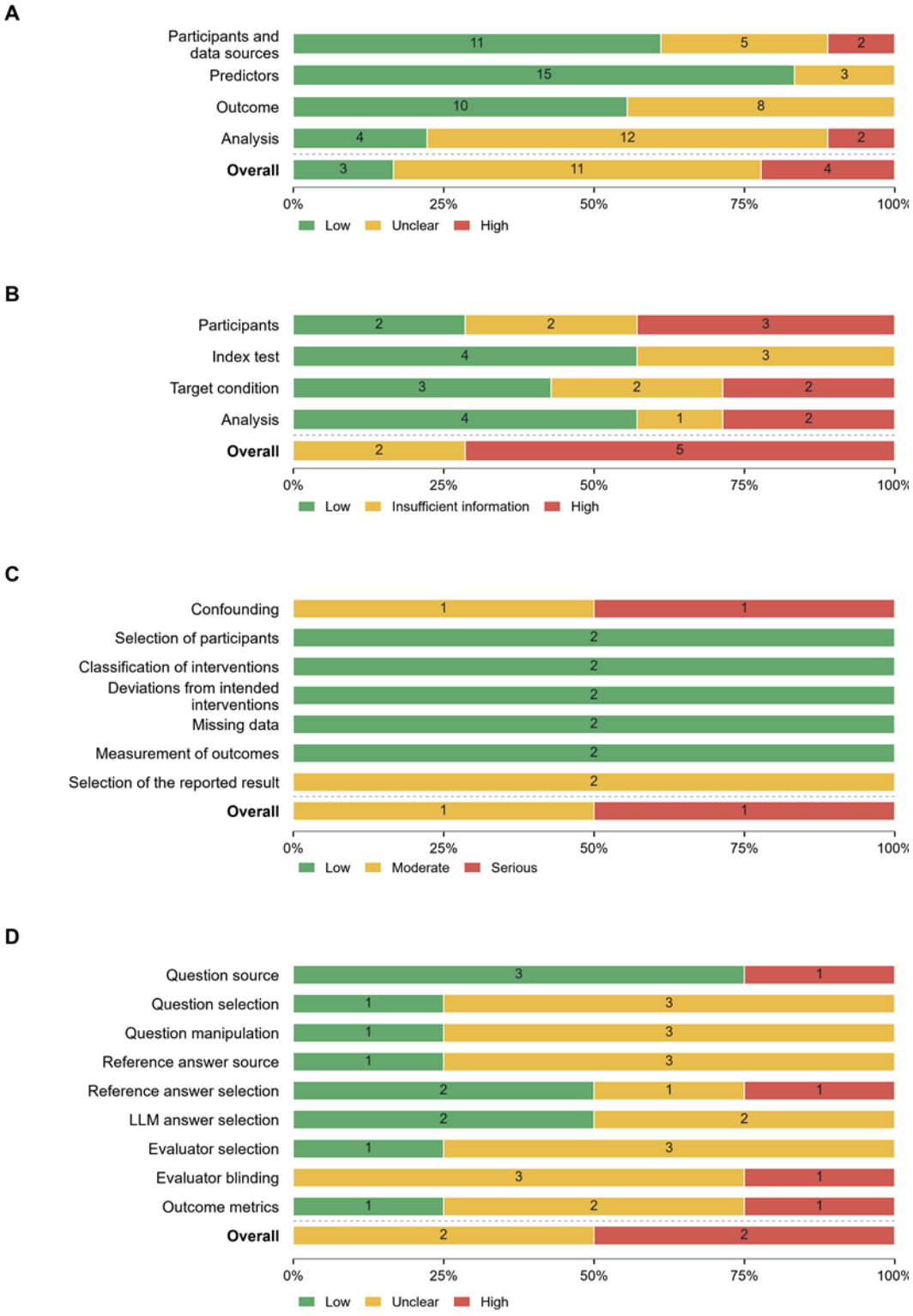
Of the 7 studies assessed with QUADAS-3, 5 were high overall [21,22,28,38,50], and 2 had insufficient information [12,31]. Participant selection was high in 3 studies and insufficiently reported in 2; the target-condition and analysis domains were each high in 2 studies (Figure 4B). Concerns involved selected case spectra, reference-standard independence or target mismatch, and analysis limitations. The reported accuracy estimates therefore remain sensitive to case spectrum, target definition, and reference-standard limitations.

ROBINS-I assessed objective writing time in the 2 nonrandomized documentation studies. Barak-Corren et al [27] were at moderate risk, and Song et al [44] were at serious risk. Confounding was moderate and serious, respectively, and selection of the reported result was moderate in both; the other 5 domains were low (Figure 4C). Fixed intervention order affected both comparisons, with repeated cases creating additional practice and carryover concerns in Song et al [44]. The observed writing-time differences indicate possible efficiency gains but do not establish that the observed gains were attributable to LLM assistance.

Among the 4 AQAT:RoB appraisals, the dialogue-factuality audit by Moser et al [34] and discharge-report evaluation by Trejo Omeñaca et al [35] were high; the handoff-note assessment by Hartman et al [40] and assistant-usability evaluation by Segalini et al [47] were unclear. Evaluator blinding was high in 1 study and unclear in 3, and evaluator selection was unclear in 3 (Figure 4D). Other high judgments involved question-source and outcome alignment in the dialogue audit and selective reference-answer inclusion in the discharge-report evaluation. In Segalini et al [47], reference answer source and reference answer selection were both unclear: physicians reportedly verified answers, but no formal reference standard or selection procedure was described. These findings limit claims about the factuality, completeness, usability, or safety of generated outputs. Complete primary and secondary outcome judgments, domain ratings, item-level responses, and supporting reasons are provided in Multimedia Appendix 4.

XSL•FO
RenderX

Figure 4. Domain-level and overall risk-of-bias judgments: (A) Prediction Model Risk of Bias Assessment Tool for prediction models using regression or artificial intelligence methods (PROBAST+AI, 18 studies), (B) Quality Assessment of Diagnostic Accuracy Studies 3 (QUADAS-3, 7 studies), (C) Risk of Bias in Non-Randomized Studies of Interventions (ROBINS-I, 2 studies), and (D) Alberta Quality Assessment Tool: Risk of Bias (AQAT:RoB; 4 studies). Bars summarize primary appraisal targets; segment labels are study counts. Each study is counted once per domain, using the highest recorded concern across components of the same target.



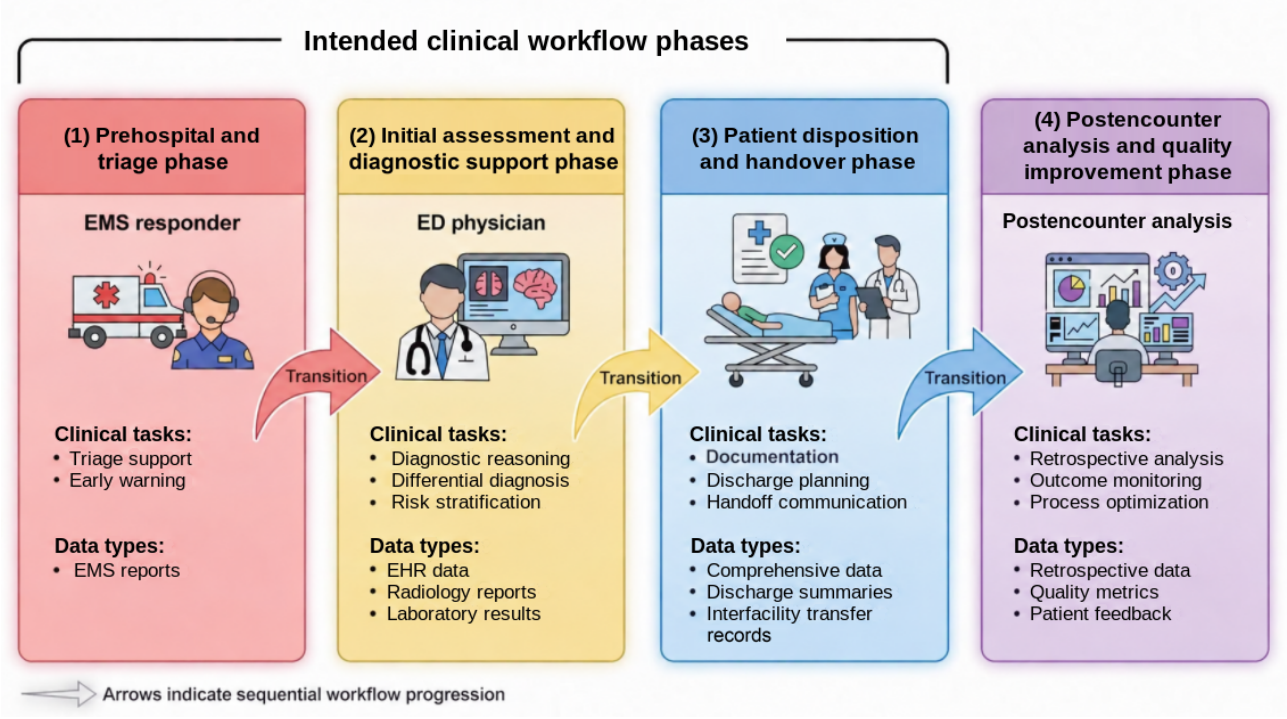
XSL•FO
RenderX

Applications and Performance Across ED Workflow Phases

Overview

The 31 selected studies fall into 4 workflow phases, as shown in the flow diagram in Figure 5. These periods align with the clinical process: (1) prehospital and triage phase; (2) initial assessment and diagnostic support phase; (3) patient disposition and handover phase; and (4) postencounter analysis and quality improvement, including safety-net and problem detection. The first 3 phases were used as an active-care and transition-support taxonomy, but individual studies within them were often retrospective or silent evaluations and should not be interpreted as real-time deployments solely on the basis of their phase assignments. Within each phase, task-specific performance is presented with the relevant evaluation setting, input timing, and appraisal findings. Because the included studies used heterogeneous tasks and nonequivalent metrics (accuracy, F_1-score, area under the receiver operating characteristic curve [AUROC], sensitivity/specificity, cosine similarity, clinician ratings, hallucination rates, and usability scores), values are summarized only within their original task and evaluation context and were not used to rank workflow phases or infer clinical maturity.

Figure 5. Flow diagram of large language model (LLM) use and performance across workflows in emergency departments (EDs). EHR: electronic health record; EMS: emergency medical services.



Prehospital and Triage Phase

The prehospital and triage phase is defined as the initial categorization of patients based on injury or illness severity to prioritize treatment resources [59]. In this phase, study inputs included EMS reports, chief complaints, brief unstructured histories, and mixed EHR inputs available near triage. The included studies (n=4) processed these inputs for triage support and severity assessment [33,48], and early disease identification [12,28]. According to these studies, these LLM applications targeted front-line users, including triage nurses and prehospital EMS clinicians, and were evaluated for triage-related classification, severity assessment, or early disease identification.

For triage support, 1 study showed that 5-level acuity models achieved modest 5-class classification performance, with F_1-scores around 0.45 to 0.50 (the averaging method was not specified), indicating frequent confusion between adjacent urgency levels [33]. Similarly, another study reported that when LLM-derived severity scores were used as additional predictors rather than stand-alone classifiers, early outcome models improved from an AUROC of 0.59 to 0.72 for in-ED death or cardiopulmonary resuscitation [48]. Both severity studies had unclear PROBAST+AI appraisals [33,48].

Disease-focused applications reported higher task-specific values under different targets and metric definitions: gout-flare prediction from chief complaints reached binary F_1-score and accuracy close to 0.9 and 3-class F_1-score above 0.8 [28], while prehospital stroke screening based on EMS texts achieved AUROC values of 0.75 for acute ischemic stroke and 0.71 for large vessel occlusion, with high sensitivity [12]. The disease-focused evaluations had high [28] or insufficient-information [12] QUADAS-3 judgments. Across all 4 studies, inputs were coded as available at or close to the intended decision point, but evaluations remained retrospective/offline and technical-performance-only (Multimedia Appendix 3). Thus, the temporal coding did not indicate evaluated use by triage or EMS clinicians.

Initial Assessment and Diagnostic Support Phase

The initial assessment phase involves the synthesis of subjective history, objective physical examination, and preliminary

XSL•FO
RenderX

diagnostic testing to formulate a working diagnosis and management plan [60]. Studies in this category evaluated LLMs using richer EHR data streams to support diagnostic reasoning, information retrieval, and clinical documentation. Across 8 studies in this phase, LLMs were primarily intended to assist ED physicians and consultants with sepsis feature extraction, diagnostic suggestions, ophthalmic and psychiatric decision support, and clinical documentation rather than autonomous decision-making [21,31,36,38,39,46,47,50].

Three studies evaluated early disease identification [31,39,50]. Early diagnosis models for sepsis in infectious disease and critical care reached a prospective F_1-score of 0.48 with sensitivity of 0.70 [39], while GPT-4 disease classifiers for respiratory disease and septicemia achieved F_1-scores ranging from 0.62 to 0.76 but trailed ClinicalBERT at 0.89 to 0.95 [31]. A 7-billion-parameter psychiatric LLM attained 90% top-2 diagnostic accuracy but showed language-dependent degradation and a 30.2% hallucination rate among extracted clues from Korean notes [50]. Lightweight models achieved extraction performance comparable to larger models for sepsis [39], and domain-specific models outperformed general-purpose ones in precision, although performance remains unstable in non-English contexts [31,50]. The sepsis result came from prospective silent-mode validation, with no live end user exposure; its PROBAST+AI appraisal was unclear [39]. The disease-classifier and psychiatric evaluations had insufficient-information and high QUADAS-3 judgments, respectively [31,50].

Four studies evaluated differential-diagnosis, management, or information-access support [21,36,38,47]. Differential-diagnosis systems showed partial overlap with reference differentials, with partial match rates of 63% to 81% but complete matches below 11% [21], admission and imaging accuracies of 0.29 to 0.74 with systematic overadmission [36], and ophthalmology accuracy of 0.76 to 0.87 depending on documentation quality [38], whereas an ED information-access assistant received a median Usability Metric for User Experience–Lite (UMUX-Lite) score of 75 after adjustment to the System Usability Scale metric [47]. These outcomes measure diagnostic or action agreement and assistant usability, respectively. The differential-diagnosis and ophthalmology evaluations used later/full-encounter information and had high QUADAS-3 judgments [21,38]; the clinical-action agreement evaluation had a low PROBAST+AI judgment [36]. The information-access assistant was evaluated in an operational pilot, but its decision-time boundary was not defined and its usability appraisal was unclear [47].

One study evaluated documentation automation. Documentation-focused neurology summaries reported reference-note performance, with mean Clinical-BioBERT embedding cosine similarity of 0.89 (SD 0.03) to physician-authored reference summaries (semantic/text similarity, not verified clinical correctness) and 78.8% agreement between next-step recommendations and observed admission-vs-discharge outcomes [46]. This retrospective evaluation had an unclear decision-time boundary in the temporal coding and a high PROBAST+AI appraisal [46].

Patient Disposition and Handover Phase

Patient disposition is the critical workflow node determining whether a patient is admitted, discharged, or transferred, and it requires accurate communication of the care trajectory to downstream providers [61]. Seven studies were designed for patient disposition and handover. In this phase, LLMs processed comprehensive data, such as ED course notes and radiology reports, to automate transfer documentation and predict postdischarge outcomes. Across these applications, the intended end user groups included inpatient and specialty teams receiving patients from the ED, for whom model-generated summaries and discharge reports were intended to streamline handover [35,40]; ED physicians, for whom LLMs were intended as decision-support or documentation assistants for admission decisions and discharge instructions [24,25,27,44]; and patients and families, for whom simplified radiology explanations or letters were intended to support understanding of disposition plans [25,27].

Two studies used discharge or ED encounter data for disposition risk prediction: a 30-day readmission model achieved an AUROC of 0.66 and an F_1-score of 0.65, trailing the ClinicalLongformer model despite providing a reasoning process [49], whereas another study integrated GPT-4 and machine learning to develop an admission-prediction model that achieved an AUROC of 0.87 and an accuracy of approximately 83%, which was better than the performance of the machine learning–only comparator [24]. Both used inputs coded as available at or close to their intended prediction point but remained offline technical evaluations; both primary PROBAST+AI appraisals were unclear [24,49].

Five studies evaluated documentation and communication automation in general and pediatric emergency medicine. The systems generated emergency medicine-to-inpatient handoff notes with high lexical and semantic similarity but slightly lower completeness and correctness scores than clinician notes [40], reported shorter writing time and correctness-completeness ratings of 4.2 to 4.5 out of 5 but still showed omissions and isolated confabulations [27,35,44], and identified actionable incidental findings with recall ranging from 95% to 99% and $F1$-score of approximately 0.85 while producing patient instructions rated 4.5 out of 5 [25]. Evaluation settings comprised preimplementation clinician evaluation [25,40], simulated clinician writing tasks [27,44], and a workflow pilot [35]. The writing-time comparisons had moderate and serious ROBINS-I judgments, while the handoff-note and discharge-report human evaluations had unclear and high AQAT:RoB judgments, respectively. These settings and appraisals distinguish favorable output ratings or shorter writing time from a demonstrated effect in routine care.

Postencounter Analysis and Quality Improvement Phase

This phase focuses on retrospective evaluation of care processes and outcomes using EHR-derived performance measures and feedback to benchmark quality and support targeted improvement initiatives [62]. Across the 12 postencounter studies, LLMs primarily served researchers and public health authorities, who leveraged large ED datasets for the following four ED clinical task categories: (1) quality improvement and

XSL•FO
RenderX

safety-net detection, (2) data structuring and research support, (3) text quality and data governance, and (4) epidemiological surveillance and public health.

For quality improvement and safety-net detection, models extracted incidental hepatic steatosis from computed tomography reports with 96.2% to 98.8% accuracy; paired sensitivity/specificity values were 0.994/0.929 for GPT-3.5, 0.966/1.00 for GPT-4, and 0.975/1.00 for GPT-4o [22], while stigma-detection systems identified harmful language with F_1-scores ranging from 0.86 to 0.90 and explicitly evaluated fairness across sex, age, and race [41].

Within data structuring and research support, LLMs converted unstructured text into analyzable labels: mental- and physical-health terms and 61 finer categories (macro-F_1-score 0.93) [23]; synthetic ED dialogues with 87% to 94% medical content accuracy to support future real-time information extraction [34]; large-scale personally identifiable information recognition and replacement with note-level F_1-score of 0.97 [43]; and Turkish pediatric notes classified as respiratory tract infection (RTI) vs non-RTI with 99.88% accuracy, AUROC of 0.97, and F_1-score of 0.97 [37].

A text quality and data governance study used local LLMs to generate synthetic radiology reports that trained downstream fracture-misdiagnosis detectors, achieving an F_1-score around 0.97 while preserving clinically plausible style [26].

Finally, epidemiological surveillance and public health work extracted vaccine names from triage text with an F_1-score of 0.97 [42], incarceration status with F_1-scores of 0.86 for GPT-4 and 0.926 for ClinicalLongformer [29], respiratory symptoms with F_1-scores ranging from 0.91 to 0.94 [30], helmet status with κ up to 1.00 [32], and 5 core injury attributes with accuracy up to 0.972 and macro-F_1-score of 0.883 [45]. These task-specific values describe different targets and constructs and were not compared directly across studies or workflow phases. Eleven of the 12 studies were retrospective/offline evaluations; the vaccine-surveillance study was the exception, with operational deployment [42]. Time-related findings in helmet surveillance and record classification concerned research or data-processing efficiency [32,37]. Low-risk PROBAST+AI appraisals for the synthetic-report and injury-extraction targets [26,45] coexisted with high-risk appraisals for record classification and deidentification [37,43]; postencounter phase assignment was therefore not a proxy for methodological quality.

Task Composition Across Workflow Phases

Task composition differed across workflow phases. Classification was the core task in all prehospital and triage studies (4/4, 100%), whereas generation or summarization was most frequently represented in the patient disposition and handover phase (5/7, 71.4%). Information extraction was most common among postencounter analysis and quality improvement studies (6/12, 50%). When postencounter studies were compared with studies targeting the other workflow phases, information extraction was represented in 50% (6/12) of postencounter studies vs 15.8% (3/19) of non–postencounter studies. In contrast, generation or summarization was represented in 16.7% (2/12) and 47.4% (9/19) of studies, respectively, while classification was represented in 41.7% (5/12) and 47.4% (9/19), respectively. Because core NLP task categories were not mutually exclusive, these distributions were examined descriptively. The cross-tabulation indicates that task composition differed descriptively across workflow phases, limiting direct comparisons of reported performance across workflow phases. The study-level mapping and full cross-tabulation are provided in Multimedia Appendix 5.

Evaluation Dimensions and Reporting Coverage

Figure 6 presents the matrix of NLP tasks and evaluation metrics of the reviewed literature. Building on HELM, which evaluates LLMs across accuracy, calibration, robustness, efficiency, fairness, bias, and toxicity, a multidimensional framework tailored to ED EHR tasks was adopted [51]. In line with recent evaluation work, factuality was added to capture the truthfulness of model outputs, and comprehensiveness was incorporated to describe the completeness of clinical responses, consistent with benchmarks such as TruthfulQA [63]. The category of “efficiency” was broadened to “deployment considerations” (latency, hardware, and cost). Finally, “fairness, bias, and toxicity” were grouped as a single safety dimension, reflecting concerns highlighted by toxicity benchmarks such as RealToxicityPrompts, where “fairness concerns” denotes systematic performance gaps across subgroups and “toxicity” refers to harmful or inappropriate content [64].

Across 31 studies, almost all reported accuracy (30/31, 96.8%), followed by deployment considerations (24/31, 77.4%), factuality (20/31, 64.5%), and comprehensiveness (18/31, 58.1%). Robustness (10/31, 32.3%), calibration and uncertainty (5/31, 16.1%), and fairness/bias/toxicity (5/31, 16.1%) were infrequently assessed.

By core NLP task, information extraction studies (n=9) most often performed multidimensional evaluation, with high attention to factuality and comprehensiveness (each 8/9, 88.9%), robustness (6/9, 66.7%), and deployment (9/9, 100%). Text-generation or summarization studies (n=11) similarly emphasized factuality (8/11, 72.7%), comprehensiveness (7/11, 63.6%), and deployment (10/11, 90.9%). These studies reported accuracy (10/11, 90.9%), calibration and uncertainty (0/11, 0%), and fairness/bias/toxicity (1/11, 9.1%). Classification-focused work (n=14) concentrated on accuracy (14/14, 100%) and deployment (8/14, 57.1%), with limited reporting of robustness (4/14, 28.6%) and fairness/bias/toxicity (4/14, 28.6%).

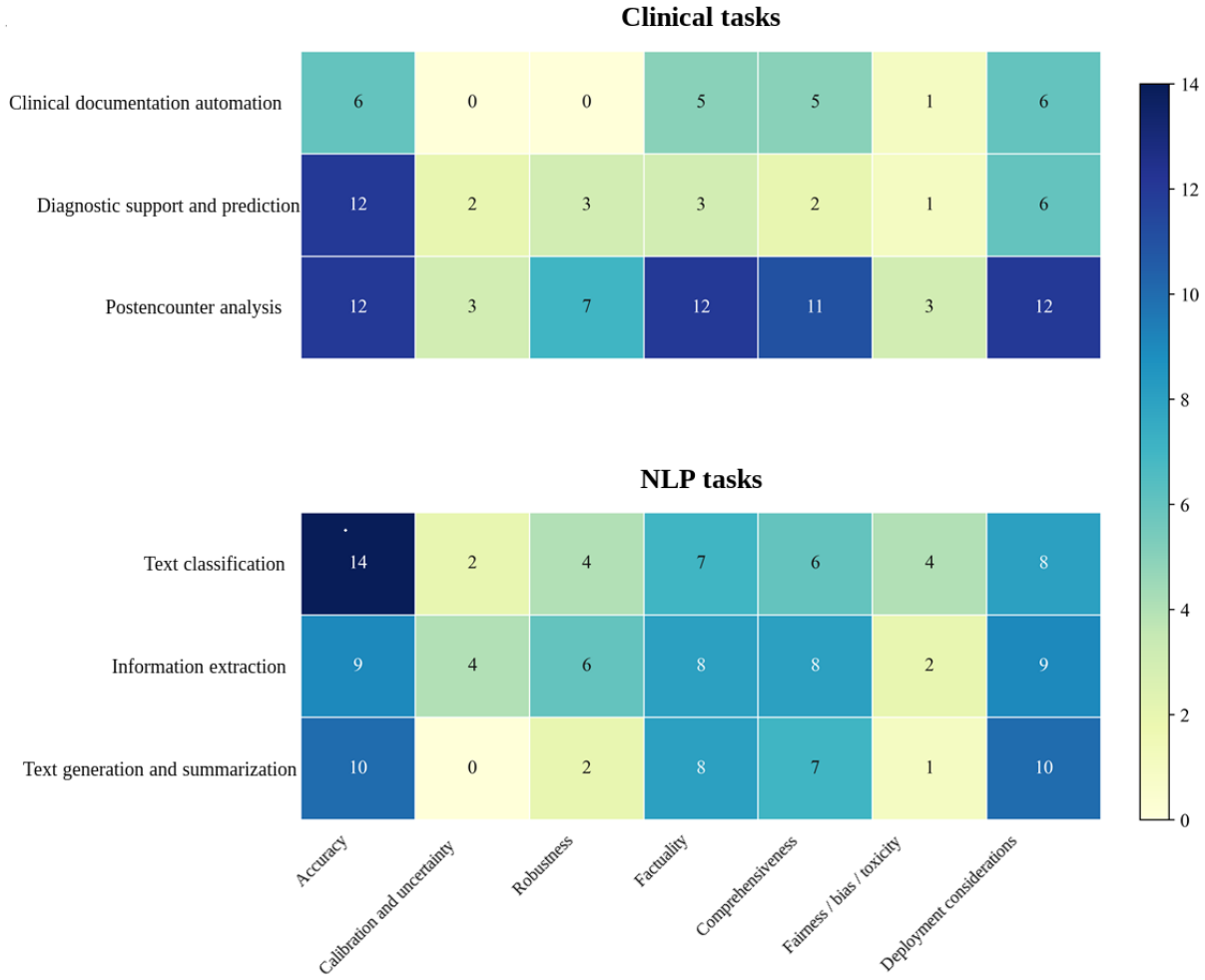
By workflow phase, postencounter analysis and quality improvement studies (n=12) provided a broader range of evaluation dimensions: all reported accuracy, factuality, and deployment (12/12, 100% each), and most addressed comprehensiveness (11/12, 91.7%) and robustness (7/12, 58.3%). In contrast, studies assigned to the 3 non–postencounter workflow phases (n=19) rarely reported robustness (3/19, 15.8%), calibration and uncertainty (2/19, 10.5%), or fairness/bias/toxicity (2/19, 10.5%) and less often assessed factuality (8/19, 42.1%) or comprehensiveness (7/19, 36.8%).

Within the 19 non–postencounter studies, clinical documentation automation studies (n=6) reported factuality and

XSL•FO
RenderX

comprehensiveness (each 5/6, 83.3%), as well as accuracy and deployment (each 6/6, 100%), whereas studies targeting diagnostic support and clinical prediction (n=13) prioritized accuracy (12/13, 92.3%) but far less frequently reported factuality (3/13, 23.1%), comprehensiveness (2/13, 15.4%), or deployment considerations (6/13, 46.2%). These are clinical-task subgroups within the non–postencounter phases; their percentages describe evaluation coverage, not measured clinical effectiveness.

Figure 6. Matrix of natural language processing (NLP) tasks and evaluation metrics of the reviewed literature. Clinical-task labels denote intended uses, not demonstrated real-time deployment.



Workflow Integration and Clinical/Implementation Evidence

Evaluation Settings and Input Timing

Evaluation setting and input timing were assessed separately from intended workflow phase. Of the 31 studies, 21 (67.7%) were retrospective or offline technical evaluations, 4 included retrospective or preimplementation clinician evaluation, and 2 evaluated clinicians in simulated tasks. Only 1 study used prospective silent-mode validation, and 3 reported active operational deployment or workflow pilots. Overall, 27 of 31 studies did not evaluate model use in either a prospective clinical data stream or an active operational workflow.

Temporal availability of input data was another important constraint. Sixteen (51.6%) studies were coded as using information available at or close to the intended decision point, whereas 10 (32.3%) used later or full-encounter information. The decision-time boundary was not applicable or not clearly defined in 3 (9.7%) studies and remained unclear in 2 (6.5%). These classifications were based on information reported in the included publications and extraction fields and should not be interpreted as independent verification that all model inputs were timestamped and available before the relevant clinical decision. Study-level temporal coding for all 31 studies is provided in Multimedia Appendix 3. Studies using discharge summaries, completed radiology reports, full encounter notes, or other downstream documentation may demonstrate retrospective analytic feasibility, but they do not by themselves provide evidence for real-time triage, diagnosis, admission, or disposition support.

Clinical and Implementation Evidence

Clinical and implementation evidence remained substantially less developed than model-level performance evaluation.

XSL•FO
RenderX

Clinical utility was empirically assessed in only 9.7% (3/31) of studies [27,40,44], whereas 24 (77.4%) studies discussed potential clinical utility without directly evaluating it. No study directly measured patient outcomes or downstream patient-safety outcomes attributable to LLM use.

Evaluation of intermediate clinical and implementation outcomes was also limited. Clinician acceptance was assessed in 5 (16.1%) studies [27,35,40,44,47] and workflow effects in 3 (9.7%) [27,44,47]. Time-related efficiency was assessed in 5 (16.1%) studies [27,32,37,44,47], but only a subset evaluated clinical task efficiency: 2 studies objectively measured clinician task time, 1 active pilot evaluated perceived time benefits, and 2 studies measured research or data-processing efficiency rather than clinical workflow efficiency. Prospective deployment was evaluated in 4 (12.9%) studies [35,39,42,47], including silent-mode or operational workflow evaluation; however, prospective data flow or pilot implementation did not by itself establish sustained clinical adoption, clinical effectiveness, patient safety, or patient benefit.

Overall, 22 (71%) studies were classified as technical-performance-only evaluations. Evidence beyond technical performance was sparse across clinician-, workflow-, and deployment-related dimensions. Thus, the current evidence base remained dominated by model-level validation rather than demonstrated clinical or implementation effects. Study-level clinical/implementation classifications for all 31 studies are provided in Multimedia Appendix 3.

Reported Implementation Challenges

For the challenge analysis, the included studies were independently coded by 2 reviewers (ZW and HY) using 4 nonmutually exclusive domains: technical adaptability, clinical utility, privacy and regulatory compliance, and human oversight. The operational definitions were informed by previous literature on LLM evaluation and implementation in health care and emergency medicine. Technical adaptability was coded when the reported challenge concerned the ability of an LLM to maintain reliable performance across variations in prompts, input structure, data sources, terminology, or clinical contexts [65]. Clinical utility was coded when the reported challenge concerned whether model outputs were clinically appropriate, valid, relevant, or applicable to the intended clinical task or decision context [66]. Privacy and regulatory compliance was coded when the reported challenge concerned confidentiality, identifiable-data handling, cloud or third-party processing, cross-border data transfer, institutional governance, or compliance with legal or regulatory requirements. Human oversight was coded when a study explicitly described or recommended clinician review, verification, modification, or override of LLM outputs, or retention of final clinical decision authority by clinicians [67]. Clinician involvement solely for annotation, reference-standard development, or study evaluation was not considered operational human oversight. The studies cited in the challenge synthesis below are selected examples rather than exhaustive lists of all studies assigned to each domain. Disagreements between the 2 reviewers were resolved through discussion and, when necessary, adjudication by a third reviewer (RY).

Across the included studies, reported challenges clustered around technical adaptability, clinical utility, privacy and regulatory constraints, and the need for human oversight. Two studies were coded as performance-only or unclassified in the challenge analysis and did not enter any of the 4 challenge domains. These domain counts represent concerns reported by studies, not measured effects of implementation.

Technical adaptability was a recurring barrier (28 studies). Three studies highlighted prompt sensitivity or dependence on prompt design, with performance varying across prompt formulations or prompting strategies [36,41,49]. Two studies reported input-output format mismatches, where models behaved inconsistently even with well-structured inputs, complicating integration into structured ED workflows [25,48]. Two further studies noted that signal quality within EHRs varies widely, making data-source selection and trust calibration nontrivial for reliable extraction [23,42].

Clinical utility-related challenges were identified in 22 studies. These challenges concerned whether model outputs were clinically appropriate, valid, relevant, or applicable to the intended clinical task or decision context. This challenge-domain coding was distinct from the separate clinical/implementation evidence coding reported above. Three studies noted that models may simplify or distort medical terminology, which is risky in scenarios where precise wording is required (eg, discharge documentation and longitudinal summaries) [25,35,50]. Two studies highlighted label or target ambiguity, including actionability criteria and variable-quality clinical targets, which introduces subjectivity and hampers reproducibility [25,40]. Three studies further reported gaps between model outputs and clinician judgment in documentation or recommendation tasks, underscoring the need for careful human review in practice [36,40,44].

Privacy and regulatory compliance were explicitly discussed in 11 studies. Moser et al [34] described privacy and security barriers to acquiring real-world emergency-dialogue data, whereas Song et al [44] noted data-security risks associated with proprietary or externally hosted models. Two studies reported governance constraints favoring local deployment: Hartman et al [40] stated that institutional data were required to remain on premises because of regulatory and legal-governance requirements, while Song et al [44] used a secure on-site deployment partly to avoid domestic restrictions on cross-border data transfer. For evaluation governance, Song et al [44] described their assessment based on completeness, correctness, conciseness, clinical utility, and text similarity as preliminary and called for more comprehensive evaluation frameworks, whereas Hartman et al highlighted unresolved postimplementation governance questions and the need for further safety-focused evaluation [40,44].

Human oversight was consistently framed as essential for clinical use (12 studies). Two described overinterpretation, where models added unsupported conclusions that could mislead clinicians in summarization or diagnostic contexts [25,35]. Two highlighted automation bias: clinicians may over-trust outputs and lower vigilance, potentially reinforcing errors or disparities

XSL•FO
RenderX

[25,41]. Two studies reported confabulations, omissions, or factual errors in generated clinical content [35,44].

Discussion

Principal Findings

Although technical performance was widely evaluated across the included studies, clinical utility was empirically assessed in only 3 studies, and no included study directly measured patient outcomes or downstream patient-safety outcomes attributable to LLM use. This gap indicates that the current evidence base is substantially stronger for model-level technical feasibility than for demonstrated clinical impact.

This systematic review of 31 studies shows that LLM applications to ED EHRs are technically promising but heterogeneous in data, methods, and intended use. From a technical perspective, the literature focused on 3 task families—classification, summarization, and information extraction—with prompt engineering as the dominant adaptation strategy.

Overall, current systems most often function as documentation, surveillance, and auditing assistants rather than as dependable decision engines. Clinical deployment remains constrained by robustness and safety limitations, privacy and regulatory requirements, and the need for explicit human oversight.

Methodological Quality and Strength of the Evidence

Methodological concerns temper the strong task-specific performance reported in these studies. In the PROBAST+AI and QUADAS-3 evaluations, test-informed model selection could inflate performance [43], while enriched case samples limit the transfer of prevalence-dependent measures to routine ED populations [22,25,50]. Unclear and insufficient-information judgments leave these estimates uncertain; they do not demonstrate that bias occurred. The 3 low-risk appraisals support the evaluated report-classification, clinical-action agreement, and injury-extraction targets [26,36,45], but agreement with observed practice does not establish clinical appropriateness or patient benefit. Outcome-specific findings also constrain interpretation: the study by He et al [49] received an unclear judgment for independent-test discrimination, which coexisted with a high-risk assessment of probability estimates or calibration under natural prevalence, limiting conclusions about absolute readmission risk.

The ROBINS-I and AQAT:RoB findings qualify the evidence for documentation support and workflow benefit. Fixed order and, in the study by Song et al [44], repeated cases make it difficult to separate LLM-related time savings from period, practice, or recall effects [27]. Human evaluations were also limited by selection, reference-material, and evaluator concerns [34,35,40,47]. These concerns reduce confidence in claims about generated-content factuality, completeness, and usability. Positive ratings therefore offer preliminary evidence of usefulness but do not establish clinical correctness or a causal improvement in routine care. In particular, usability findings in Segalini et al [47] should not be interpreted as evidence of answer accuracy or measured time savings.

Taken together, the appraisal supports technical feasibility and clinician-reviewed drafting, with limited confidence in claims of clinical effectiveness. More informative studies would use independent test data, representative case spectra, explicit reference standards, and appropriate calibration and uncertainty analyses. Workflow studies need comparisons that address order and carryover effects, along with evaluations of patient and safety outcomes at the intended decision point. These design improvements are necessary to determine whether promising task-level results translate into reliable clinical benefit.

LLM Use and Performance Across Workflows in EDs

While postencounter analysis and quality improvement applications have repeatedly demonstrated task-specific retrospective technical feasibility, applications intended to inform ongoing clinical care require evaluation at the relevant decision point.

In the prehospital and triage phase, inputs were available near the intended decision point, but evaluations remained retrospective [12,28,33,48]. Although these systems conceptually align well with the information needs of triage nurses and EMS clinicians, the evidence base is still thin and heterogeneous, and current models are better viewed as experimental adjuncts than as reliable components of ED triage protocols.

Initial assessment involves high cognitive load and fragmented information. Diagnostic-support studies achieved partial diagnostic agreement, and 1 study reported systematic overadmission [21,36]. The use of later or full-encounter information in some diagnostic evaluations also limits their relevance to early assessment [21,38]. For clinical documentation, high semantic similarity to reference notes does not by itself establish clinical correctness [46]. These findings support further evaluation of LLMs as clinician-reviewed information aids but do not establish their effectiveness as independent decision-makers in this phase.

For patient disposition and handover, the evidence does not yet establish LLMs as reliable independent prognostic models, although they can contribute as feature generators within integrated predictive frameworks [24,49]. In documentation-oriented systems, partial omissions and occasional hallucinations still occurred, so clinicians still need to review outputs carefully [25,27,35,40,44]. Nevertheless, the evidence in this phase supported the feasibility of clinician-reviewed drafting, with preliminary evidence of efficiency gains. Their contribution to high-risk disposition decisions remains uncertain.

Documentation support and decision support were therefore interpreted separately. Drafting handoff notes, discharge summaries, or consultation summaries can be evaluated with completeness, factuality, time saving, and clinician review, whereas triage, diagnostic, admission, or disposition recommendations require calibration, safety evaluation, external validation, and preferably prospective clinical evaluation before they can be considered reliable decision support.

In postencounter applications, performance also varied with task definitions and the handling of contextual cues [29,32].

XSL•FO
RenderX

Offline evaluations can use completed records and allow additional verification time without directly guiding care during the index encounter. These conditions may make postencounter tasks comparatively tractable for retrospective technical evaluation but do not establish lower methodological risk of bias, clinical safety, or greater clinical maturity. Most studies remained focused on data structuring and error labeling, leaving their contribution to root-cause analysis and care pathway optimization uncertain.

The findings in postencounter applications may partly reflect task composition, because extracting information that is already documented in the clinical record is fundamentally different from predicting clinical states or outcomes that are not yet known at the intended decision point. Given the heterogeneity of task formulations, reference standards, and reported performance metrics, the available studies did not permit a valid within-task comparison sufficient to determine whether a phase-related performance difference persisted after accounting for task type.

LLM Technological Strategies for EHRs in EDs

Data characteristics shaped the technological strategies adopted in the included studies. English-language corpora dominate, with only a small subset operating in non-English or mixed-language environments. Prompt-based adaptation is attractive when labeled data are limited, whereas fine-tuning on local documentation requires evaluation of generalizability to other settings.

Against this background, the studies converged on 3 core NLP task types. Classification tasks have been widely explored for tasks such as outcome prediction. Generation and summarization tasks are similarly common, particularly for documentation support and communication. Information-extraction pipelines that transform ED narratives into structured features for surveillance and phenotyping form the third category of actionable outputs. Collectively, these directions were the most frequently studied ways of converting ED free text into task-specific outputs; their nonequivalent metrics do not support a cross-task ranking of performance or maturity.

Optimization strategies for adapting general-purpose LLMs to these tasks were dominated by prompt engineering. Many papers treated zero-shot prompting as a baseline and showed incremental improvements when moving to structured templates or few-shot designs, especially for classification and information-extraction tasks. A smaller subset explicitly implemented chain-of-thought prompting to expose intermediate reasoning steps and in some cases reported modest performance gains. However, the evidence did not establish uniform generalizability.

Fine-tuning was evaluated in fewer studies and sometimes improved performance within individual classification or extraction tasks, but heterogeneous datasets, models, and comparators precluded a general conclusion that fine-tuning outperformed prompting. Moreover, fine-tuning remained limited to a small subset of relatively modest-sized models; none of the reviewed work reported targeted fine-tuning of very large, state-of-the-art LLMs on multicenter ED data. This may reflect constraints in computing resources and governance at the hospital level, but it also means that the potential benefits and risks of large-scale domain adaptation in emergency medicine remain essentially unexplored.

In RAG applications, retrieval was primarily used to surface semantically similar historical cases or relevant segments from a patient's own record, which were then supplied to the model as additional context [24,39,46-48]. These designs were intended to address context-window limitations and improve factual grounding. Individual studies reported task-specific gains, but the available evidence did not establish a general reduction in hallucinations or reliable improvement in clinical correctness. Nevertheless, current RAG applications are narrow in scope: retrieval is almost always confined to local ED databases, and no study integrated LLMs with external guideline libraries, curated case repositories, or real-time literature. The potential of RAG as a bridge between bedside decision-making and broader sources of medical knowledge therefore remains insufficiently evaluated.

Finally, several studies moved beyond single-model setups toward hybrid pipelines and agentic workflows. Several combined LLMs with conventional predictive or encoder-based components, whereas Trejo Omeñaca et al [35] used a multiagent, multistage workflow [23,24,33,39,40,48]. Heterogeneous architectures, tasks, and comparators limit comparisons of these approaches. LLM agents performed data extraction, summarization, and recommendation refinement [35,47]. Within the current evidence base, therefore, agentic architectures are still at a very early stage: they appear conceptually promising for mirroring multistep ED workflows, but concrete implementations and evaluations remain scarce.

Challenges and Opportunities

For prompt sensitivity, several papers point to proposed strategies such as stronger prompt engineering and training-data refinement to better accommodate differences in how inputs are structured [23,46,49]. Similarly, to reduce terminology oversimplification, domain-oriented strategies (eg, clinical fine-tuning on specialized medical corpora) have been suggested so that models handle medical language with greater precision and less loss of nuance [24,36]. Data source selection and the subjectivity embedded in target definitions also require more rigorous methodological treatment [22,33,39,40].

Data privacy/regulatory compliance and human oversight also warrant closer examination. Problems such as cloud-related risk, overinterpretation, automation bias, and confabulation lack widely tested safeguards in the reviewed evidence. The literature describes potential safeguards for institution-restricted or sensitive data [40,44] and potential human-in-the-loop safeguards in which clinicians review and validate outputs before decisions are finalized [25,35,44]. Future work will likely need to reinforce these approaches in real settings or propose alternative mechanisms that can address these risks more reliably.

Overall, LLM applications in ED and ED-adjacent EHR workflows were broad in scope but remained limited in demonstrated workflow integration. The evidence was

XSL•FO
RenderX

concentrated on the technical feasibility of retrospective information extraction, surveillance, and documentation-related tasks. In contrast, substantially fewer studies evaluated LLMs within prospective clinical data streams or active workflows, and evidence for sustained operational adoption, patient safety, or clinical and operational utility remained limited.

Future research should move beyond model-centric evaluation toward prospective, workflow-oriented assessment that considers decision-time information availability, clinician workload, safety, and patient or operational outcomes.

Limitations and Future Studies

The findings of this review should be interpreted in light of several limitations. First, this review was not prospectively registered in PROSPERO or other public registries. Although an internal protocol was developed, the absence of prospective public registration limits independent verification of which analyses were specified before the review began. The methodological appraisal and additional integration, temporal-availability, clinical/implementation, and phase-by-task analyses were developed during peer review, after the included studies had been examined. Knowledge of the evidence may therefore have influenced the choice of coding frameworks and the interpretation of these analyses. Their findings contextualize the original synthesis; they do not confirm prespecified phase-related hypotheses or establish causal explanations for performance differences. Second, the evidence base was assembled from only 2 bibliographic databases (PubMed and Web of Science) over a restricted time window (January 2022 to November 2025). Other papers that meet the review criteria in Embase, Scopus, IEEE Xplore, ACM Digital Library, or other sources might therefore have been missed, and the resulting evidence base may be weighted toward clinically oriented and biomedical publications. Most open-dataset studies used MIMIC-derived data. Because MIMIC is publicly available, some LLMs may have been exposed to similar or overlapping content during pretraining, which could inflate retrospective performance estimates. Few studies explicitly assessed potential data contamination or benchmark exposure. As our research aims focus on the use of LLMs in EHRs and EDs, only studies that applied LLMs to real-world ED EHRs or ED-adjacent emergency-care records were considered; other non-EHR applications were excluded, as were architectures based solely on BERT-like encoder models. As a result, the review does not capture the full spectrum of AI methods in emergency medicine, and some conclusions may not generalize to systems using raw clinical modalities or other non–text-based inputs. Future updates could broaden database coverage and include raw clinical modalities while retaining an explicit ED and ED-adjacent emergency care scope.

The exclusion of preprints may have introduced time-lag bias. Given the rapid pace of development in the LLM field, potentially relevant studies may first appear as preprints and may not have completed peer review by the time of the literature search. Consequently, the review may underrepresent the most recent technical developments and emerging applications. This trade-off reflects our decision to restrict the evidence base to peer-reviewed publications but should be considered when interpreting the completeness and currency of the review.

Moreover, substantial heterogeneity in datasets, tasks, comparators, and outcome metrics precluded formal meta-analysis. The synthesis is therefore descriptive, and apparent performance differences between studies or model families should be interpreted cautiously. Also, some scope and definitions were manually assigned. For instance, the mapping of applications to HELM task categories and ED workflow phases, although guided by predefined definitions, also involved a degree of subjective judgment. Finally, given the rapid evolution of LLMs, newer models, deployment patterns, and regulatory developments emerging after the search date are not reflected in this review.

The study-level appraisal summarized one primary target per study; secondary outcome assessments remain in Multimedia Appendix 4. Applying model-evaluation domains to text labeling and AQAT:RoB to assistant usability required context-specific interpretation. For the usability component, the reference-answer source and selection were assessed, and both were judged Unclear for Segalini et al [47]; this review-specific adaptation has not been separately validated. No formal outcome-level GRADE assessment was conducted. Evidence strength was interpreted by jointly considering tool-specific risk of bias, applicability, validation setting, temporal availability of model inputs, and directness to the intended clinical outcome. These dimensions were considered separately; prospective or external evaluation was not assumed to imply low risk of bias.

Conclusion

This review provides an integrated view of how LLMs are currently being used with ED electronic health records, spanning data characteristics, technical strategies, and clinical workflows. Across 31 studies, current evidence supports selected technical feasibility of LLM applications in ED and ED-adjacent emergency care, particularly for retrospective information extraction, surveillance, auditing, and clinician-reviewed documentation support. Favorable task-specific performance in these settings should not, however, be interpreted as evidence of equivalent clinical maturity or real-time workflow readiness.

When mapped onto the ED care pathway, studies demonstrated task-specific retrospective technical feasibility in postencounter analysis and quality-improvement tasks, including surveillance, phenotyping, documentation quality assessment, and deidentification under offline evaluation conditions. Studies targeting prehospital triage, initial diagnostic assessment, and disposition or handover are fewer in number and address heterogeneous tasks and outcomes, with limited evidence of clinical effectiveness or readiness for real-time decision support.

Taken together, the available evidence does not yet support robust real-time clinical decision support or end-to-end ED workflow integration. Confidence in these findings is limited by methodological concerns and unresolved information in much of the evidence base. Most studies remained retrospective or technically oriented, and empirical evidence of prospective workflow integration, clinical utility, patient safety, and patient benefit was limited or absent. Future research should prioritize

XSL•FO
RenderX

external and prospective validation, decision-time information availability, clinician-centered workflow evaluation, safety assessment, and direct measurement of patient and operational outcomes.

Funding

This study was supported in part by the Natural Science Foundation of Sichuan, China (2025NSFSC1985), and the 1·3·5 Project for Disciplines of Excellence, West China Hospital, Sichuan University (ZYYC21004).

Data Availability

All data generated or analyzed during this review are included in this article and Multimedia Appendices 1-5. No additional individual-participant data were generated.

Generative AI Use

During the preparation of this work, the authors used Grammarly to correct grammatical errors. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the published article.

Authors' Contributions

Conceptualization: YC, KL, RY
Methodology: ZW, HY, DL, YZ, PD
Investigation: ZW, HY, RY
Data curation: ZW, HY, DL
Formal analysis and synthesis: ZW, HY, DL, YZ
Clinical interpretation: PD, YC
Visualization: ZW, HY, DL
Supervision: YC, KL, RY
Project administration: YC, KL, RY
Funding acquisition: KL, RY
Writing - original draft: ZW, HY
Writing - review and editing: ZW, HY, PD, DL, YZ, YC, KL, RY

Conflicts of Interest

None declared.

Multimedia Appendix 1

PRISMA 2020 checklist.
[DOCX File , 22 KB-Multimedia Appendix 1]

Multimedia Appendix 2

Complete database search strategies.
[DOCX File , 14 KB-Multimedia Appendix 2]

Multimedia Appendix 3

Study-level extraction and clinical implementation coding.
[XLSX File (Microsoft Excel File), 17 KB-Multimedia Appendix 3]

Multimedia Appendix 4

Design-matched methodological quality and risk-of-bias assessments.
[XLSX File (Microsoft Excel File), 144 KB-Multimedia Appendix 4]

Multimedia Appendix 5

Workflow phase and core natural language processing (NLP) task cross-tabulations.
[XLSX File (Microsoft Excel File), 14 KB-Multimedia Appendix 5]

References

XSL•FO
RenderX

1. Liu F, Zhou H, Gu B, Zou X, Huang J, Wu J, et al. Application of large language models in medicine. Nat Rev Bioeng. 2025;3(6):445-464. [doi: 10.1038/s44222-025-00279-5]
2. Shool S, Adimi S, Saboori Amleshi R, Bitaraf E, Golpira R, Tara M. A systematic review of large language model (LLM) evaluations in clinical medicine. BMC Med Inform Decis Mak. 2025;25(1):117. [FREE Full text] [doi: 10.1186/s12911-025-02954-4] [Medline: 40055694]
3. Boscardin CK, Gin B, Golde PB, Hauer KE. ChatGPT and generative artificial intelligence for medical education: potential impact and opportunity. Acad Med. 2024;99(1):22-27. [FREE Full text] [doi: 10.1097/ACM.0000000000005439] [Medline: 37651677]
4. Raza MM, Venkatesh KP, Kvedar JC. Generative AI and large language models in health care: pathways to implementation. NPJ Digit Med. 2024;7(1):62. [FREE Full text] [doi: 10.1038/s41746-023-00988-4] [Medline: 38454007]
5. 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):319-328. [doi: 10.1001/jama.2024.21700] [Medline: 39405325]
6. Wornow M, Xu Y, Thapa R, Patel B, Steinberg E, Fleming S, et al. The shaky foundations of large language models and foundation models for electronic health records. NPJ Digit Med. 2023;6(1):135. [FREE Full text] [doi: 10.1038/s41746-023-00879-8] [Medline: 37516790]
7. Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language models in medicine. Nat Med. 2023;29(8):1930-1940. [doi: 10.1038/s41591-023-02448-8] [Medline: 37460753]
8. Singhal K, Azizi S, Tu T, Mahdavi SS, Wei J, Chung HW, et al. Large language models encode clinical knowledge. Nature. 2023;620(7972):172-180. [FREE Full text] [doi: 10.1038/s41586-023-06291-2] [Medline: 37438534]
9. 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(3):943-950. [doi: 10.1038/s41591-024-03423-7] [Medline: 39779926]
10. Clusmann J, Kolbinger FR, Muti HS, Carrero ZI, Eckardt J, Laleh NG, et al. The future landscape of large language models in medicine. Commun Med (Lond). 2023;3(1):141. [FREE Full text] [doi: 10.1038/s43856-023-00370-1] [Medline: 37816837]
11. Meng X, Yan X, Zhang K, Liu D, Cui X, Yang Y, et al. The application of large language models in medicine: a scoping review. iScience. 2024;27(5):109713. [FREE Full text] [doi: 10.1016/j.isci.2024.109713] [Medline: 38746668]
12. Wang X, Ye S, Feng J, Feng K, Yang H, Li H. Performance of ChatGPT on prehospital acute ischemic stroke and large vessel occlusion (LVO) stroke screening. Digit Health. 2024;10:20552076241297127. [FREE Full text] [doi: 10.1177/20552076241297127] [Medline: 39507012]
13. Robertson C, Woods A, Bergstrand K, Findley J, Balser C, Slepian MJ. Diverse patients' attitudes towards artificial intelligence (AI) in diagnosis. PLOS Digit Health. 2023;2(5):e0000237. [FREE Full text] [doi: 10.1371/journal.pdig.0000237] [Medline: 37205713]
14. Zhou S, Xu Z, Zhang M, Xu C, Guo Y, Zhan Z, et al. Large language models for disease diagnosis: a scoping review. NPJ Artif Intell. 2025;1(1):9. [doi: 10.1038/s44387-025-00011-z] [Medline: 40607112]
15. Kirubarajan A, Taher A, Khan S, Masood S. Artificial intelligence in emergency medicine: a scoping review. J Am Coll Emerg Physicians Open. 2020;1(6):1691-1702. [FREE Full text] [doi: 10.1002/emp2.12277] [Medline: 33392578]
16. Asplin BR, Magid DJ, Rhodes KV, Solberg LI, Lurie N, Camargo CA. A conceptual model of emergency department crowding. Ann Emerg Med. 2003;42(2):173-180. [doi: 10.1067/mem.2003.302] [Medline: 12883504]
17. Moons KGM, Damen JAA, Kaul T, Hooft L, Andaur Navarro C, Dhiman P, et al. PROBAST+AI: an updated quality, risk of bias, and applicability assessment tool for prediction models using regression or artificial intelligence methods. BMJ. 2025;388:e082505. [doi: 10.1136/bmj-2024-082505] [Medline: 40127903]
18. Whiting PF, Tomlinson E, Rutjes AW, Davenport CF, Yang B, Westwood ME, et al. QUADAS-3: a revised tool for the quality assessment of diagnostic test accuracy studies. Ann Intern Med. 2026;179(4):548-555. [FREE Full text] [doi: 10.7326/ANNALS-25-02104] [Medline: 41698208]
19. Sterne JA, Hernán MA, Reeves BC, Savović J, Berkman ND, Viswanathan M, et al. ROBINS-I: a tool for assessing risk of bias in non-randomised studies of interventions. BMJ. 2016;355:i4919. [FREE Full text] [doi: 10.1136/bmj.i4919] [Medline: 27733354]
20. Ye C, Mitchell JR, Baumgart DC, Ma Z, Fung AL, Orellana DG, et al. The Alberta quality assessment tool: risk of bias (AQAT:RoB) for the evaluation of medical large language model question-answer studies: development and pilot validation. J Med Internet Res. 2026;28:e87057-e87057. [FREE Full text] [doi: 10.2196/87057] [Medline: 41950508]
21. Shah-Mohammadi F, Finkelstein J. Accuracy evaluation of GPT-assisted differential diagnosis in emergency department. Diagnostics (Basel). 2024;14(16):1779. [FREE Full text] [doi: 10.3390/diagnostics14161779] [Medline: 39202267]
22. Vong T, Rizer N, Jain V, Thompson VL, Dredze M, Klein EY, et al. Automated identification of incidental hepatic steatosis on emergency department imaging using large language models. Hepatol Commun. 2025;9(3). [FREE Full text] [doi: 10.1097/HC9.0000000000000638] [Medline: 39969431]
23. Cardamone NC, Olfson M, Schmutte T, Ungar L, Liu T, Cullen SW, et al. Classifying unstructured text in electronic health records for mental health prediction models: large language model evaluation study. JMIR Med Inform. 2025;13:e65454-e65454. [FREE Full text] [doi: 10.2196/65454] [Medline: 39864953]

XSL•FO
RenderX

24. 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. J Am Med Inform Assoc. 2024;31(9):1921-1928. [doi: 10.1093/jamia/ocae103] [Medline: 38771093]
25. Woo KC, Simon G, Akindutire O, Aphinyanaphongs Y, Austrian J, Kim J, et al. Evaluation of GPT-4 ability to identify and generate patient instructions for actionable incidental radiology findings. J Am Med Inform Assoc. 2024;31(9):1983-1993. [FREE Full text] [doi: 10.1093/jamia/ocae117] [Medline: 38778578]
26. Liu J, Koopman B, Brown NJ, Chu K, Nguyen A. Generating synthetic clinical text with local large language models to identify misdiagnosed limb fractures in radiology reports. Artif Intell Med. 2025;159:103027. [FREE Full text] [doi: 10.1016/j.artmed.2024.103027] [Medline: 39580897]
27. Barak-Corren Y, Wolf R, Rozenblum R, Creedon JK, Lipsett SC, Lyons TW, et al. Harnessing the power of generative AI for clinical summaries: perspectives from emergency physicians. Ann Emerg Med. 2024;84(2):128-138. [doi: 10.1016/j.annemergmed.2024.01.039] [Medline: 38483426]
28. Nipu AS, Islam KMS, Madiraju P. How reliable AI chatbots are for disease prediction from patient complaints? IEEE Int Conf Inf Reuse Integr Data Sci (IRI). 2024:210-215. [doi: 10.1109/IRI62200.2024.00052]
29. Huang T, Socrates V, Gilson A, Safranek C, Chi L, Wang EA, et al. Identifying incarceration status in the electronic health record using large language models in emergency department settings. J Clin Transl Sci. 2024;8(1):e53. [FREE Full text] [doi: 10.1017/cts.2024.496] [Medline: 38544748]
30. McMurry AJ, Phelan D, Dixon BE, Geva A, Gottlieb D, Jones JR, et al. Large language model symptom identification from clinical text: multicenter study. J Med Internet Res. 2025;27:e72984-e72984. [FREE Full text] [doi: 10.2196/72984] [Medline: 40743494]
31. Cui W, Finkelstein J. Leveraging LLMs for early diagnosis in the emergency department: comparing ClinicalBERT and GPT-4. Stud Health Technol Inform. 2025;329:1840-1841. [doi: 10.3233/SHTI251241] [Medline: 40776258]
32. Burford KG, Itzkowitz NG, Ortega AG, Teitler JO, Rundle AG. Use of generative AI to identify helmet status among patients with micromobility-related injuries from unstructured clinical notes. JAMA Netw Open. 2024;7(8):e2425981. [FREE Full text] [doi: 10.1001/jamanetworkopen.2024.25981] [Medline: 39136946]
33. Lafuente C, Rahim M. Evaluating large language models on hospital health data for automated emergency triage. Int J Comput Assist Radiol Surg. 2025;20(9):1941-1952. [doi: 10.1007/s11548-025-03475-1] [Medline: 40668511]
34. Moser D, Bender M, Sariyar M. Generating synthetic healthcare dialogues in emergency medicine using large language models. Stud Health Technol Inform. 2024;321:235-239. [doi: 10.3233/SHTI241099] [Medline: 39575815]
35. Trejo Omeñaca A, Llargués Rocabruna E, Sloan J, Catta-Preta M, Ferrer i Picó J, Alfaro Alvarez JC, et al. Leave as fast as you can: using generative AI to automate and accelerate hospital discharge reports. Computers (Basel). 2025;14(6):210. [doi: 10.3390/computers14060210]
36. Williams CYK, Miao BY, Kornblith AE, Butte AJ. Evaluating the use of large language models to provide clinical recommendations in the emergency department. Nat Commun. 2024;15(1):8236. [FREE Full text] [doi: 10.1038/s41467-024-52415-1] [Medline: 39379357]
37. 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(1):154. [FREE Full text] [doi: 10.1186/s12911-025-02871-6] [Medline: 40165165]
38. Lavista Ferres JM, Feng S, Kim M, Popovici N, Lee L, Moore K, et al. Table-based language models for ophthalmology assessment in the emergency department. Graefes Arch Clin Exp Ophthalmol. 2026;264(1):279-284. [doi: 10.1007/s00417-025-06972-w] [Medline: 41081893]
39. Shashikumar SP, Nemati S. A prospective comparison of large language models for early prediction of sepsis. Pac Symp Biocomput. 2025;30:109-120. [FREE Full text] [doi: 10.1142/9789819807024_0009] [Medline: 39670365]
40. 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(12):e2448723. [FREE Full text] [doi: 10.1001/jamanetworkopen.2024.48723] [Medline: 39625719]
41. Chen H, Alfred M, Cohen E. Efficient detection of stigmatizing language in electronic health records via in-context learning: comparative analysis and validation study. JMIR Med Inform. 2025;13:e68955. [FREE Full text] [doi: 10.2196/68955] [Medline: 40825541]
42. Khademi S, Black J, Palmer C, Javed M, Clothier H, Buttery J, et al. Enhancing vaccine safety surveillance: extracting vaccine mentions from emergency department triage notes using fine-tuned Large language models. Stud Health Technol Inform. 2025;329:1170-1174. [doi: 10.3233/SHTI251023] [Medline: 40776041]
43. Dorémus O, Russon D, Contrand B, Guerra-Adames A, Avalos-Fernandez M, Gil-Jardiné C, et al. Harnessing moderate-sized language models for reliable patient data deidentification in emergency department records: algorithm development, validation, and implementation study. JMIR AI. 2025;4:e57828-e57828. [FREE Full text] [doi: 10.2196/57828] [Medline: 40605780]
44. Song JW, Park J, Kim JH, You SC. Large language model assistant for emergency department discharge documentation. JAMA Netw Open. 2025;8(10):e2538427. [FREE Full text] [doi: 10.1001/jamanetworkopen.2025.38427] [Medline: 41118162]

XSL•FO
RenderX

45. Choi DH, Kim Y, Choi SW, Kim KH, Choi Y, Shin SD. Using large language models to extract core injury information from emergency department notes. J Korean Med Sci. 2024;39(46):e291. [FREE Full text] [doi: 10.3346/jkms.2024.39.e291] [Medline: 39623965]
46. Gorenshtein A, Perek S, Vaisbuch Y, Shelly S. AI-generated neurology consultation summaries improve efficiency and reduce documentation burden in the emergency department. Sci Rep. 2025;15(1):38868. [FREE Full text] [doi: 10.1038/s41598-025-22769-7] [Medline: 41198773]
47. Segalini A, Castro C, Lerner J, Horn M, Simon M, Rubin L, et al. TanaGPT: how can I help you? A pilot testing of a GenAI assistant in a healthcare workflow. Stud Health Technol Inform. 2025;329:118-122. [doi: 10.3233/SHTI250813] [Medline: 40775831]
48. 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(1):25345. [FREE Full text] [doi: 10.1038/s41598-025-07649-4] [Medline: 40659653]
49. He Z, Li H, Tian S, Wen J, Yuan G, Li A. A comparative study of structured and narrative EHR data for 30-day readmission risk assessment. Electronics. 2025;14(20):4033. [doi: 10.3390/electronics14204033]
50. Kim M, Hwang G, Chang J, Chang S, Roh HW, Park RW. Performance of open-source large language models in psychiatry: usability study through comparative analysis of non-English records and English translations. J Med Internet Res. 2025;27:e69857. [FREE Full text] [doi: 10.2196/69857] [Medline: 40825309]
51. Liang P, Bommasani R, Lee T, Tsipras D, Soylu D, Yasunaga M, et al. Holistic evaluation of language models. ArXiv. Preprint posted online on November 16, 2022. [doi: 10.48550/arXiv.2211.09110]
52. Sebastiani F. Machine learning in automated text categorization. ACM Comput Surv. 2002;34(1):1-47. [doi: 10.1145/505282.505283]
53. Nenkova A. Automatic Summarization. Found Trends Inf Retr. 2011;5(2):103-233. [doi: 10.1561/1500000015]
54. Sarawagi S. Information extraction. Found Trends Databases. 2008;3(1):261-377. [doi: 10.1561/1900000003]
55. Liu P, Yuan W, Fu J, Jiang Z, Hayashi H, Neubig G. Pre-train, prompt, and predict: a systematic survey of prompting methods in natural language processing. ACM Comput Surv. 2023;55(9):1-35. [doi: 10.1145/3560815]
56. Howard J, Ruder S. Universal language model fine-tuning for text classification. 2018. Presented at: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers); July 15-20, 2018; Melbourne, Australia. [doi: 10.18653/v1/P18-1031]
57. Lewis P, Perez E, Piktus A, Petroni F, Karpukhin V, Goyal N, et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. 2020. Presented at: Advances in Neural Information Processing Systems 33 (NeurIPS 2020); December 6-12, 2020; Vancouver, Canada. URL: https://proceedings.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html
58. Wei J, Wang X, Schuurmans D, Bosma M, Ichter B, Xia F, et al. Chain-of-thought prompting elicits reasoning in large language models. Adv Neural Inf Process Syst. 2022;35:24824-24837. [doi: 10.52202/068431-1800]
59. Ibrahim BE. Sudanese emergency departments: a study to identify the barriers to a well-functioning triage. BMC Emerg Med. 2022;22(1):22. [FREE Full text] [doi: 10.1186/s12873-022-00580-1] [Medline: 35135475]
60. Committee on Diagnostic Error in Health Care, Board on Health Care Services, Institute of Medicine. Balogh EP, Miller BT, Ball JR, editors. Improving Diagnosis in Health Care. Washington, District of Columbia. National Academies Press; 2015.
61. Rutkowski RA, Scheer E, Carlson C, Parks R, Pulia MS, Patterson BW, et al. A scoping review of work system elements that influence emergency department disposition decision-making. Hum Factors Healthc. 2023;4:100059. [doi: 10.1016/j.hfh.2023.100059]
62. Deakyne Davies SJ, Grundmeier RW, Campos DA, Hayes KL, Bell J, Alessandrini EA, et al. The pediatric emergency care applied research network registry: a multicenter electronic health record registry of pediatric emergency care. Appl Clin Inform. 2018;9(2):366-376. [FREE Full text] [doi: 10.1055/s-0038-1651496] [Medline: 29791930]
63. Lin S, Hilton J, Evans O. TruthfulQA: measuring how models mimic human falsehoods. ArXiv. Preprint posted online on September 8, 2021. [doi: 10.48550/arXiv.2109.07958]
64. Gehman S, Gururangan S, Sap M, Choi Y, Smith NA. RealToxicityPrompts: evaluating neural toxic degeneration in language models. ArXiv. Preprint posted online on September 24, 2020. [doi: 10.18653/v1/2020.findings-emnlp.301]
65. Wang L, Wan Z, Ni C, Song Q, Li Y, Clayton E, et al. Applications and concerns of ChatGPT and other conversational large language models in health care: systematic review. J Med Internet Res. 2024;26:e22769. [FREE Full text] [doi: 10.2196/22769] [Medline: 39509695]
66. Seo J, Choi D, Kim T, Cha WC, Kim M, Yoo H, et al. Evaluation framework of large language models in medical documentation: development and usability study. J Med Internet Res. 2024;26:e58329. [FREE Full text] [doi: 10.2196/58329] [Medline: 39566044]
67. Amiot F, Potier B. Artificial intelligence (AI) and emergency medicine: balancing opportunities and challenges. JMIR Med Inform. 2025;13:e70903-e70903. [FREE Full text] [doi: 10.2196/70903] [Medline: 40802997]

XSL•FO
RenderX

Abbreviations

AQAT: RoB: Alberta Quality Assessment Tool: Risk of Bias
AUROC: area under the receiver operating characteristic curve
BERT: Bidirectional Encoder Representations from Transformers
ED: emergency department
EHR: electronic health record
EMS: emergency medical services
GRADE: Grading of Recommendations Assessment, Development, and Evaluation
HELM: holistic evaluation of language models
ICD-11: International Classification of Diseases, 11th Revision
LLM: large language model
LoRA: low-rank adaptation
MIMIC-III: Medical Information Mart for Intensive Care
NLP: natural language processing
PRESS: Peer Review of Electronic Search Strategies
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PROBAST+AI: Prediction Model Risk of Bias Assessment Tool for prediction models using regression or artificial intelligence methods
PROSPERO: International Prospective Register of Systematic Reviews
QLoRA: quantized low-rank adaptation
QUADAS-3: Quality Assessment of Diagnostic Accuracy Studies 3
RAG: retrieval-augmented generation
ROBINS-I: Risk of Bias in Non-Randomized Studies of Interventions
RTI: respiratory tract infection
UMUX-Lite: Usability Metric for User Experience–Lite

Edited by Y Hong; submitted 16.Mar.2026; peer-reviewed by J Bienzeisler, H Sabbineni, P Katyal; comments to author 30.Jul.2026; revised version received 22.Sep.2026; accepted 29.Oct.2026; published 23.Nov.2026

Please cite as:
Wu Z, Yu H, Deng P, Liu D, Zhao Y, Cao Y, Li K, Yin R
Large Language Models for Electronic Health Records in Emergency Department (ED) and ED-Adjacent Emergency Care—Technical Strategies, Clinical Tasks, and Workflow Integration: Systematic Review
J Med Internet Res 2026;28:e95400
URL: https://www.jmir.org/2026/1/e95400
doi: 10.2196/95400
PMID:

©Zhuohao Wu, Huilin Yu, Peng Deng, Di Liu, Yong Zhao, Yu Cao, Kang Li, Rong Yin. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 23.Nov.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX