From Triage to Discharge: A Survey of NLP Tasks, Methods, and Open Challenges in the Emergency Department

Dipankar Srirag Aditya Joshi Salil S. Kanhere Padmanesan Narasimhan
University of New South Wales, Sydney, Australia
{d.srirag, aditya.joshi, salil.kanhere, padmanesan}@unsw.edu.au

Abstract

Emergency departments (EDs) operate under time pressure, generating multimodal data such as clinical conversations, triage notes, and discharge documents. Recent advances in natural language processing (NLP), particularly pretrained transformers and large language models, have created new opportunities to support language and time-intensive stages of emergency care. Yet existing surveys map clinical NLP across the broader hospital workflow or focus on specific tasks. This survey analyses 46 papers spanning the three phases of ED: triage, diagnosis, and disposition, covering tasks such as triage classification, clinical summarisation, automatic diagnosis, report generation, and discharge documentation. We examine modelling paradigms, evaluation practices, and emerging benchmarks and shared tasks. Across tasks, we identify common trends, including a shift from task-specific neural architectures to pretrained language models, growing interest in interactive clinical systems, and increasing attention to clinically grounded evaluation. Finally, we detail open challenges such as limited generalisability, noisy clinical inputs, and workflow constraints that inform future ED-NLP research.

1 Introduction

The use of natural language processing (NLP) in medical domains has been widely studied (Wu et al., 2020). Modern NLP approaches, particularly following the development of Transformer-based large language models (LLMs), have enabled a range of clinical tasks, including summarisation (Adams et al., 2021), dialogue modelling (Xu et al., 2023), information extraction (Fornasiere et al., 2024), and clinical decision support (Xu et al., 2024b). This survey focuses on a specific clinical context: emergency departments (EDs), where clinicians work under severe time pressure, often within a four-hour window from arrival to disposition (Jones and Schimanski, 2010; Mason et al., 2012). ED care unfolds across three phases: *triage and arrival*, *diagnosis and assessment*, and *patient disposition* (See Appendix A). These phases generate triage notes, chief complaints, patient speech, consultation notes, radiology and laboratory reports, and discharge summaries, comprising both structured and unstructured textual records that cover initial reports, intermediate clinical reasoning, and final decisions. Applying NLP techniques to ED data therefore has the potential to reduce the documentation burden, surface important clinical signals, and support consistent communication with the patient throughout the ED workflow. Appendix B provides a detailed motivation for NLP approaches in the ED.

Survey	Tri.	Dia.	Dis.	Task	Method
Klug et al. (2024)	✕	✓	✓	–	–
Valizadeh and Parde (2022)	✕	✓	✕	✓	–
Stewart et al. (2023)	✓	✕	✕	–	✓
Tyler et al. (2024)	✓	✕	✕	✕	✓
Yao and Yu (2026)	–	–	–	–	✓
Ours	✓	✓	✓	✓	✓

Table 1: Comparison with prior surveys across ED phases (**Tri**age, **Dia**gnosis, **Dis**position) and analytical dimensions (NLP **Task** formulation, **Method** analysis). ✓ = covered; ✕ = not covered; – = partial.

Existing surveys of Artificial Intelligence (AI) and NLP in emergency medicine, as described in Table 1, provide useful overviews of use cases and clinical outcomes. However, they primarily draw on clinical settings and rarely examine methodological aspects relevant to NLP, such as task formulation, modelling assumptions, or evaluation protocols in a systematic way (Kirubarajan et al., 2020; Stewart et al., 2023; Tyler et al., 2024). Surveys within the NLP community often focus on specific tasks or model architectures without assessing their evaluation practices or alignment with real-world workflow constraints (Di Martino and Delmastro, 2022; Valizadeh and Parde,

Proceedings of the 2026 Conference on Empirical Methods in Natural Language Processing, pages 20918–20942
October 24-29, 2026 ©2026 Association for Computational Linguistics

2022; Wang et al., 2025b). The two closest recent surveys share a workflow-level view but differ from ours in scope. Klug et al. (2024) catalogues clinical-NLP applications along the inpatient hospital pathway from admission to discharge, and Yao and Yu (2026) analyse LLM-based multi-agent clinical systems around workflow observables and deployment-readiness framework. Neither is ED-specific, and neither examines how ED-phase constraints shape task formulation and evaluation. This fragmentation leaves open fundamental questions about how ED-oriented NLP tasks are formulated, how modelling paradigms are chosen, and whether evaluation practices adequately reflect operational realities in emergency care.

To address this gap, we survey NLP research applied to ED workflows through a corpus of 46 papers[1]. We analyse how these systems are deployed across ED phases and examine trends in dataset language, input modality, and modelling paradigms. We further synthesise task-level developments across triage classification, clinical interaction summarisation, automatic diagnosis, radiology report generation, disposition prediction, and discharge documentation. Finally, we discuss benchmarking resources and outline open challenges and future research directions for deploying NLP systems in ED settings.

2 Research Trends

Tables 2 and 3 in Appendix D provide a paper-level catalogue of all 46 included papers. We outline the trends across three dimensions: clinical settings, training paradigms, and evaluation paradigms.

2.1 Clinical settings

ED phase. Most work targets *diagnosis and assessment* (35/46), where language data are relatively rich and naturally support tasks such as clinical summarisation, diagnostic reasoning, and report generation. By contrast, triage receives limited attention (5/46), despite being one of the most time-critical stages of ED care. Disposition tasks appear in 6 studies, including disposition prediction and discharge documentation. Research therefore concentrates on diagnostic reasoning and documentation rather than early triage decision support. Applying the ED-relatedness tags, the corpus splits into 10 ED-native, 17 ED-anchored, and 19 ED-adjacent papers. Papers corresponding to triage and disposition are almost entirely ED-native, whereas automatic diagnosis is dominated by ED-adjacent work (16/19). This distribution makes the inclusion boundary visible, as much of what is called ED-NLP in CS venues is methodologically transferable to ED but not evaluated there, which conditions the trend estimates that follow.

Language. English remains the dominant language (32/46) for NLP research in ED, followed by Chinese (13/46). Other languages, such as German and Spanish, appear in a small number of studies (3/46), and the source language is unspecified in 2/46. This imbalance largely reflects the availability of public clinical corpora, especially the MIMIC family of datasets (Johnson et al., 2019, 2023a,b), which have become standard resources for ED-facing clinical NLP. The limited representation of other languages highlights the scarcity of multilingual datasets and evaluation resources.

2.2 Training paradigms

Input modality. Text is the primary input modality (43/46), including triage notes, clinical conversations, physician documentation, and radiology reports. Structured electronic health record variables appear in 22/46 studies, while only 4/46 incorporate medical imaging directly. Despite the availability of heterogeneous multimodal information, most current work remains text-centred.

Training strategies. Supervised methods appear in 13/46 studies, agent-based systems in 10/46, transfer learning in 8/46, knowledge-grounded models in 6/46, pretraining in 4/46, reinforcement learning in 3/46, retrieval-augmented generation in 2/46, and prompt-based methods in only one paper. This distribution suggests a lack of a single dominant strategy beyond conventional supervised and transfer learning approaches. The scarcity of prompt-based methods in the corpus contrasts with their prevalence in shared task submissions (Section 4). Agent-based and retrieval-augmented methods are concentrated in 2024 and 2025, reflecting the recent influence of LLM-based pipelines.

2.3 Evaluation paradigms

Study design. Most papers rely on retrospective records (24/46), simulation, or synthetic data (21/46). Only one study adopts a prospective design (Ip et al., 2024), with data collected alongside live clinical care. As a result, most findings re-

[1] Search strategy and inclusion criteria in Appendix C

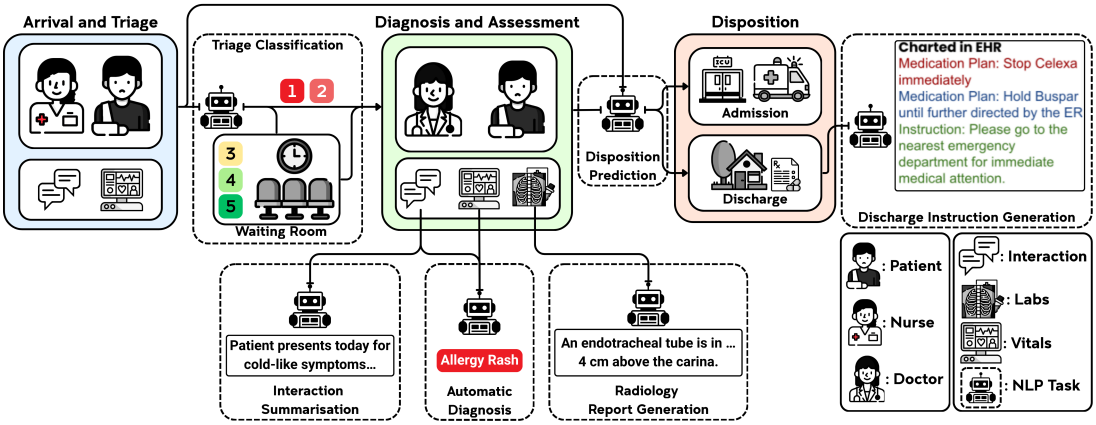

Figure 1: Overview of NLP tasks across the three phases of the ED workflow.

main supported by offline evaluation rather than real-time clinical use.

Clinician evaluation. Only 11/46 papers supplement automated metrics with clinician or human judgement. The remainder rely entirely on automatic evaluation (Lin, 2004), despite the limited alignment between such metrics and clinical utility (Moramarco et al., 2022). This gap is especially important in ED settings due to high cost of error.

Deployment. No paper in our corpus reports clinical deployment within an ED-specific workflow. The closest case, MTDiag (Hou et al., 2023), is deployed in an online medical-consultation platform, which by our scope does not qualify as ED deployment. Real-world ED validation remains the most concrete open challenge for the field.

3 NLP Tasks in the ED

NLP methods in healthcare support several functions that align closely with the three phases of the ED workflow. Figure 1 summarises the main ED-focused tasks across the three ED phases. The following subsections describe each task, the main training strategies (Figure 2), and the metrics used to evaluate them. Appendix E provides a summary of reported performance across the reviewed tasks.

Common formulation. We formalise the ED NLP tasks under a common conditional prediction framework. Each task defines an input space $\mathcal{X}$ and an output space $\mathcal{Y}$, and models the conditional distribution parameterised by θ: $\hat{y} = \arg\max_{y \in \mathcal{Y}} p_\theta(y \mid x)$. Tasks differ in what constitutes x (the information available at the corresponding ED phase) and y (the required output); each subsection below specifies both.

3.1 Triage Classification

Triage classification assigns an urgency or severity category (ETEK, 2024; Gilboy et al., 2005) at *arrival* and *triage*. Under the common formulation, the output space is an ordered set of K acuity levels: $\mathcal{Y} = \{1 \prec 2 \prec \cdots \prec K\}$ with $K = 5$ under most triage scales (Appendix A), used for prioritisation and resource allocation (Stewart et al., 2023; Tyler et al., 2024). The ordering $\prec$ reflects that acuity is ordinal (Level 1 is most urgent), though most work treats $\mathcal{Y}$ as a flat categorical space. Early work used topic modelling on triage notes to reveal symptom clusters and temporal patterns, suggesting that free text can support triage decisions (Kocbek et al., 2014). Existing work utilises several modelling paradigms, which can be grouped into transfer-learning approaches based on pretrained transformers, agent-based, and retrieval-augmented LLM systems.

Transfer learning approaches. These formulate the task as supervised classification using contextual embeddings from pretrained transformers, with input $x = (c, v)$ combining the free-text chief complaint c with vital signs v serialised into textual form (Maschhur et al., 2024; Liu et al., 2025). The resulting representation is then used for triage classification (Maschhur et al., 2024), along with expert knowledge (Liu et al., 2025). Overall, these approaches use pretrained encoders to jointly model short textual descriptions and structured triage variables.

Agent-based approaches. Recent work explores LLMs acting as collaborative agents for triage decision-making, with input $x = (c, I)$ comprising the chief complaint c and an instruction/prompt I. One line of work uses multiple LLM-based medical

experts that iteratively discuss a case and converge on a triage decision through multi-round deliberation (Lu et al., 2024). Another approach extends the input to $x_t = (c, v, h_t, I)$ at turn t, where h_t is the dialogue history, modelling triage as a dynamic process that revises the assigned level as additional patient information becomes available (Zhu et al., 2025c). This formulation explicitly models interacting with patients to gather additional information.

Retrieval-augmented approaches. These augment LLMs with external medical knowledge, with input $x = (c, v, m, R, I)$ where R is retrieved evidence and m captures encounter metadata such as patient demographics. Methods serialise structured ED variables and clinical notes into prompts, then retrieve supporting evidence from sources such as PubMed[2] to guide prediction (Gaber et al., 2025). Experiments suggest that incorporating vital signs and retrieved references can improve performance across several LLM configurations. Comparative evaluations further indicate that general-purpose LLMs can match untrained ED practitioners in vignette-style settings, though they do not yet match professionally trained raters (Masanneck et al., 2024).

Metrics. Triage classification is usually evaluated with standard classification metrics such as accuracy, precision, recall, F1, and AUC (Maschhur et al., 2024; Liu et al., 2025; Zhu et al., 2025c; Gaber et al., 2025). Some work also reports disagreement-based measures such as total discordance, defined as $1 -$ accuracy, to quantify mismatch with clinician-assigned triage levels (Lu et al., 2024). However, most studies treat triage as a standard multi-class classification task and do not account for the ordinality of triage categories. Dataset properties such as long-tailed symptom distributions are also rarely addressed explicitly during model development (Maschhur et al., 2024).

> Evaluations typically use post-hoc triage notes or EHR data that already contain information accumulated during the encounter rather than the limited inputs available at the moment of triage. This assumption inflates apparent accuracy and conflicts with how triage is performed in practice.

[2] https://pubmed.ncbi.nlm.nih.gov/

3.2 Clinical Interaction Summarisation

Clinical interaction summarisation converts multi-speaker doctor-patient conversations into structured clinical notes. Under the common formulation, the output space is a structured note partitioned into S sections: $\mathcal{Y} = \{n = (n^{(1)}, \ldots, n^{(S)})\}$ most commonly in SOAP format (Subjective, Objective, Assessment, Plan; $S = 4$). Generation is typically autoregressive under a sequence-to-sequence model p_θ. Existing work can be grouped into task-specific neural architectures, transfer-learning approaches based on pretrained transformers, domain-adaptive pretraining approaches, and prompt-based methods. These approaches input $x = d = (u_1, \ldots, u_T)$ a multi-turn dialogue where each utterance $u_t = (s_t, w_t)$ pairs a speaker label $s_t \in \{\text{clinician}, \text{patient}\}$ with a token sequence w_t.

Neural architectures. Early work relied on task-specific neural architectures trained directly on annotated dialogue datasets. These methods often decomposed the task into intermediate steps such as utterance classification, clinical entity extraction, and SOAP section prediction (Jeblee et al., 2019; Schloss and Konam, 2020). Other studies explored abstractive summarisation with pointer-generator networks (See et al., 2017) to better model negation and encourage copying of clinically salient content from the source dialogue (Joshi et al., 2020). Pipeline-oriented systems were also developed to combine automatic speech recognition, clinical concept extraction, and rule-based post-processing for structured form filling from emergency medical service intake reports (Rahman et al., 2020). Overall, these methods were highly task-specific and often depended on multi-stage pipelines.

Transfer learning approaches. These formulate summarisation as a transfer-learning problem, with the same input processed through pretrained encoders or encoder-decoder models to support both utterance-level note section assignment and section-conditioned summary generation (Su et al., 2022; Chen et al., 2023c). For example, Krishna et al. (2021) propose a hybrid extractive-abstractive approach that identifies noteworthy utterances using a BERT-LSTM classifier and then generates section-conditioned summaries using a fine-tuned T5 model. Others adapt transformer models to conversational medical data before performing section classification (Chen et al., 2023c) or abstractive summarisation with pointer-generator net-

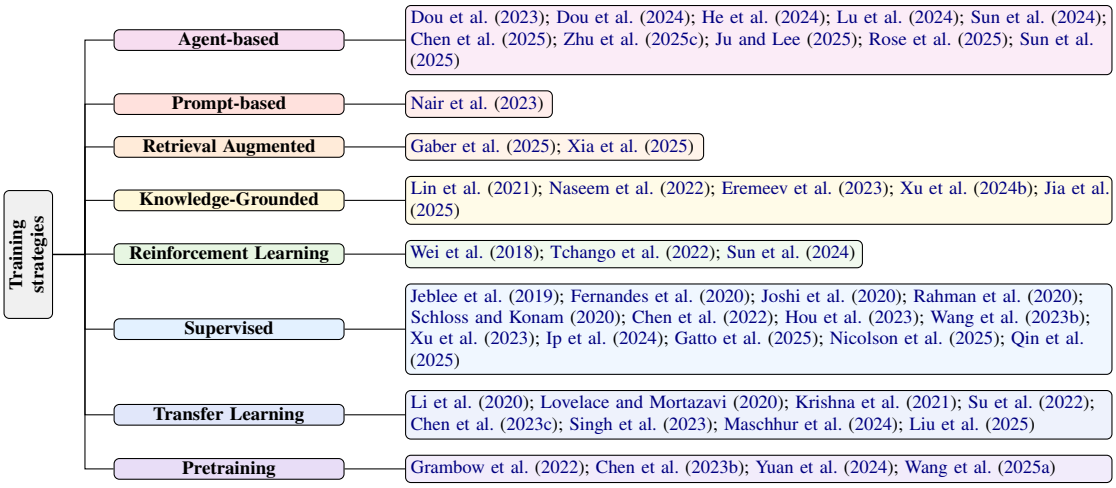

Figure 2: Training strategies used in the included literature.

works (Singh et al., 2023). Compared with earlier systems, these methods rely less on handcrafted pipelines and more on transfer from LLMs.

Pretraining approaches. A related line of work adapts pretrained models to the clinical dialogue domain before task-specific fine-tuning. This includes continued or domain-adaptive pretraining with denoising objectives on conversational or synthetic clinical corpora, followed by fine-tuning for note generation or note section summarisation (Grambow et al., 2022; Chen et al., 2023b). More recent work explores continual pretraining of LLMs on clinical corpora such as MIMIC-IV-Note (Johnson et al., 2023b) before adapting them for summarisation tasks (Yuan et al., 2024). Wang et al. (2025a) further align summarisation quality with clinician expectations through reinforcement learning from human and AI feedback.

Prompt-based approaches. Recent work explores prompt-based approaches using LLMs without task-specific training, extending the input to $x = (d, I)$ with an instruction/prompt I. Nair et al. (2023) formulate the task as a multi-step prompting pipeline using GPT-3 (Brown et al., 2020), first extracting clinical entities and then generating structured summaries conditioned on these entities. Expert evaluation is used to assess the quality and coverage of the summary and clinical concepts.

Metrics. Evaluation typically relies on automated metrics such as ROUGE (Lin, 2004) and BERTScore (Zhang et al., 2020), along with medical concept coverage via tools like QuickUMLS (Grambow et al., 2022; Singh et al., 2023). Some work also reports utterance-level section classification accuracy using metrics such as AUROC (Jeblee et al., 2019; Schloss and Konam, 2020). Savkov et al. (2022) show that grounding evaluations in pre-specified consultation checklists improves the correlation of these metrics with human judgements.

> Approaches assume curated transcripts with clean turn boundaries, correct speaker labels, and complete utterances. Deployed ED summarisation would instead receive noisy ASR output with recognition and diarisation errors, informal patient phrasing, and frequent interruption.

3.3 Automatic Diagnosis

Automatic diagnosis differs from the other tasks in being sequential, the model interleaves symptom inquiries with a terminal diagnosis over multiple turns. At turn t, given dialogue history $h_t = ((u_1, r_1), \ldots, (u_{t-1}, r_{t-1}))$ of prior inquiries u and patient responses r, a policy π_θ selects either an inquiry action or a terminal diagnostic decision: $a_t = \pi_\theta(h_t)$, $a_t \in \mathcal{A}_{\text{inquire}} \cup \mathcal{D}$. $\mathcal{A}_{\text{inquire}}$ is the space of symptom inquiries and $\mathcal{D}$ is the set of possible diagnoses. In the ED, this aligns with *diagnosis and assessment*, where structured history-taking and timely rule-in or rule-out decisions are critical. We focus on approaches trained and evaluated on derived data (dialogues) rather than on single-turn settings with fully observed information (Jin et al., 2021a). Existing work can be grouped into reinforcement learning, supervised, knowledge-grounded, and agent-based approaches. Surveys on medical agents provide complementary

categorisations of this literature (Shi et al., 2024; Wang et al., 2025b). Approaches differ in how π_θ is parameterised and trained.

Reinforcement learning approaches. These formulate diagnosis as sequential decision-making and learn π_θ through interaction rewards. Wei et al. (2018) propose a task-oriented dialogue system that uses a Deep Q-Network to decide whether to inquire about additional symptoms or produce a diagnosis. Later work extends this paradigm with more structured reward design. For example, Tchango et al. (2022) introduce a dual-branch architecture combining an evidence acquisition policy with a disease classifier and reward functions that encourage exploration of differential diagnoses and prioritisation of severe conditions. Despite these advances, such approaches often require carefully engineered reward functions and can suffer from unstable training and limited data efficiency.

Supervised approaches. These reformulate diagnosis as supervised prediction, learning π_θ from labelled interactions using transformer-based architectures (Chen et al., 2022; Wang et al., 2023b; Hou et al., 2023). Some treat diagnosis as symptom-sequence generation (Chen et al., 2022), others use collaborative or multi-task learning to align symptom checking with disease prediction (Wang et al., 2023b; Hou et al., 2023). More recent work models dialogue structure, follow-up questions, and patient responses to improve symptom tracking and downstream reasoning (Xu et al., 2023; Gatto et al., 2025; Qin et al., 2025). However, supervised approaches often assume symptoms in h_t are available in structured form, requiring extraction components to recover them from conversation.

Knowledge-grounded approaches. These augment h_t with structured medical knowledge to improve diagnostic reasoning. Prior work injects medical knowledge graphs into transformer models through knowledge-enriched utterance representations (Naseem et al., 2022), while graph-based methods model symptom-disease relationships with evolving knowledge graphs to guide dialogue generation and entity prediction (Lin et al., 2021). More recent approaches integrate knowledge graphs with LLMs. For example, Jia et al. (2025) propose medIKAL, which combines LLM predictions with knowledge graph retrieval and path-based re-ranking to identify candidate diseases. Other work focuses on improving how LLM reasoning better reflects clinical diagnostic processes by explicitly modelling intermediate reasoning steps during dialogue (Xu et al., 2024b).

Agent-based approaches. These implement π_θ as LLM agents, often decomposing diagnosis into specialised functions such as symptom acquisition, knowledge retrieval, and differential ranking (Rose et al., 2025; Ju and Lee, 2025; Dou et al., 2023). Some approaches inject structured diagnostic protocols via preference learning (Dou et al., 2024) or coordinate multi-agent multidisciplinary consultations (Chen et al., 2025). Prompting-based variants expose intermediate reasoning (He et al., 2024) or dynamic prompt adaptation with retrieved knowledge (Sun et al., 2025).

Metrics. Evaluation of automatic diagnosis systems typically measures diagnostic accuracy and symptom acquisition efficiency. Common metrics include disease prediction accuracy, recall of implicit symptoms, and the number of conversational turns required to reach a diagnosis (Wei et al., 2018; Tchango et al., 2022). Some work also evaluates ranking-based metrics such as Recall@K for differential diagnosis generation (Chen et al., 2022; Wang et al., 2023b). Reinforcement learning approaches additionally report success rate and cumulative reward to measure efficiency. However, few studies evaluate agreement with clinician diagnoses or assess performance in real clinical workflows. Most rely on simulated patients or controlled retrospective settings, raising questions about reliability and privacy when deployed against real ED encounters. Recent work documents systematic biases in LLM-based disease prediction across gender and age subgroups (Zhao et al., 2024b) and broader hallucination risks in clinical LLMs (Zhu et al., 2025b), both of which compound in ED settings.

Sequential formulations optimise for symptom acquisition efficiency, yet real ED diagnosis is bounded by clinician time rather than inquiry count. The turn-count rewards produces models that minimise interaction, and leads to a mismatch neither the training objective nor the evaluation captures.

3.4 Radiology Report Generation

Prior surveys reviewed medical report generation more broadly, including consultations and discharge summaries and their evaluation (Zhou et al.,

2023). In contrast, we focus on *radiology report generation* (RRG) for chest X-rays, emphasising methods and metrics that matter in the ED context. The output space is a structured report with two sections: $\mathcal{Y} = \{r = (r_{\text{find}}, r_{\text{impr}})\}$ where r_{find} is the free-text *Findings* section describing observed abnormalities and r_{impr} is the *Impression* section summarising clinical conclusions. Most models are trained and evaluated on large CXR corpora such as MIMIC-CXR (Johnson et al., 2019). Because many of these examinations occur during ED stays, the task is directly relevant to ED workflows.

Early work adapted image-captioning architectures to radiology with input $x = G$, a chest radiograph. Lovelace and Mortazavi (2020) extract spatial features with a pretrained DenseNet-121 encoder and add an auxiliary objective that enforces agreement between clinical observations extracted from generated and reference reports. More recent work extends the input to $x = (G, \zeta)$ with clinical context ζ (triage vitals, medications, indication, prior imaging). CXRMate-ED (Nicolson et al., 2025) combines chest X-rays with triage vitals and medications, embedding heterogeneous inputs to prompt a Llama-based decoder with a UniFormer image encoder, trained via supervised learning and self-critical sequence training. Retrieval-augmented approaches further extend the input to $x = (G, \zeta, R)$ where R is retrieved evidence. MMed-RAG (Xia et al., 2025) filters retrieved evidence before generation and applies preference-based fine-tuning to align outputs with reference reports. Together, these methods reflect a shift toward context-aware and retrieval-grounded generation to reduce clinically harmful hallucinations.

Metrics. Evaluation of radiology report generation typically relies on automated text generation metrics such as ROUGE, BLEU (Papineni et al., 2002), and CIDEr (Vedantam et al., 2015), along with precision and recall for the extracted clinical concepts. Some recent work also reports diagnostic classification metrics when assessing the clinical correctness of generated reports (Xia et al., 2025). Zhao et al. (2024a) introduce an entity-aware metric that decomposes reports into typed medical entities and compares them via embedding similarity, showing closer alignment with expert preference. The risks of persistent hallucinations mean that factual consistency checks and dedicated error detection remain essential for ED deployment (Min et al., 2022; Rusak et al., 2023).

Current methods populate ζ with structured triage variables (vitals, medications), but radiologists reading ED chest X-rays also draw on the referring clinician's suspected diagnosis, prior imaging, and the specific clinical question being asked. Reports generated without this indication risk being clinically fluent but diagnostically off-target for the question that prompted the imaging.

3.5 Tasks in Patient Disposition

Disposition in the ED spans two language-relevant problems. The first is *disposition prediction*, which forecasts whether a patient will be admitted or discharged and the likely destination, such as a ward or ICU. The second is *discharge instructions*, which translate care plans into patient-facing guidance. Both problems draw on resources available at triage time, such as early triage narratives, chief complaint text, brief histories, clinician notes, vitals and demographics. We structure this section by task as the literature is relatively small.

Disposition Prediction. Disposition prediction aims to identify patients at risk of hospitalisation or clinical deterioration early in the ED encounter. The output is a categorical disposition class $y \in \mathcal{D} = \{\textit{discharge}, \textit{ward}, \textit{ICU}\}$. Early work uses $x = (c, v, m)$ combining routine triage variables with free-text chief complaints, showing that textual features complement structured variables for outcomes such as ICU admission and related critical events (Fernandes et al., 2020). Multimodal variants extend the input to $x = (c, v, m, V)$ with lightweight visual signals V; for example, short triage videos capturing clinical gestalt improve hospitalisation prediction when fused with standard triage variables (Ip et al., 2024). More recent work uses $x = (c, v, m, I)$ with LLM-based approaches that linearise heterogeneous patient data into a single textual prompt: ED-Copilot processes such input via BioGPT (Luo et al., 2022) to predict critical outcomes and recommend informative laboratory tests (Sun et al., 2024), further using reinforcement learning to select laboratory groups that reduce diagnostic delay.

Discharge Instructions Generation. At the point of disposition, language technologies simplify and standardise care plans into patient-facing instructions. The output is a patient-facing instruction sequence $y = \iota$ derived from clinical inputs.

PharmMT (Li et al., 2020) takes $x = p$, a prescription direction, and formulates simplification as neural machine translation from clinical to lay phrasing, reporting substantial gains on automatic metrics together with pharmacist-judged safety and usability at scale. For broader discharge instructions, Eremeev et al. (2023) take $x = (d, K)$ where d is a doctor-patient dialogue and K is injected domain knowledge; this amplifies the use of rare clinical tokens and improves factuality and coherence compared with instruction-tuned baselines.

Metrics. For disposition prediction, common metrics include AUROC, AUPRC, and calibration measures, often with subgroup analyses for higher- and lower-acuity patients (Fernandes et al., 2020; Ip et al., 2024; Sun et al., 2024). For discharge instructions, these are typically complemented by readability assessments and pharmacist or clinician judgements of utility and safety (Li et al., 2020; Liu et al., 2022; Eremeev et al., 2023).

Disposition prediction models optimise for aggregate AUROC, but the clinical cost of errors is asymmetric. Under-triage of an ICU-bound patient carries far greater consequences than over-triage of a ward-bound one. Discharge instruction generation, by contrast, produces the same output ι regardless of patient literacy, language, or cognitive state. Hence, across both tasks, evaluation should remain sensitive to calibration and subgroup performance in the heterogeneous ED population.

4 Benchmarks and Shared Tasks

Early benchmarks for clinical language models relied mainly on exam-style question answering datasets that measured factual recall rather than real clinical decision-making and communication (Jin et al., 2021b, 2019). Recent benchmarks explore open-ended tasks that better reflect clinical workflows, including summarisation, diagnosis, and disposition prediction. ClinicBench (Liu et al., 2024a) represents this shift by aggregating datasets for clinical language understanding, reasoning, and generation. SOTA LLMs report worse performance on open-ended tasks than on exam-style benchmarks. Benchmarks have also expanded beyond text-only evaluation toward multimodal and process-centric settings. MC-BEC (Chen et al., 2023a) integrates structured intake data, radiology reports, vital signs, and physiological waveforms across >100K ED visits for time-sensitive prediction tasks such as decompensation and disposition. Other resources target safety and longitudinal reasoning: MEDEC (Ben Abacha et al., 2025) evaluates detection and correction of clinically significant documentation errors, ACI-Bench (Yim et al., 2023) evaluates dialogue-to-note generation from full doctor-patient conversations. Finally, MedJourney (Wu et al., 2024c), MediQ (Li et al., 2024), and DDxGym (Winter et al., 2024) assess patient-journey and interactive diagnostic reasoning.

Shared tasks provide controlled comparisons under common datasets and protocols. In clinical interaction summarisation, the MEDIQA-Chat 2023 shared task introduced section-wise and full-note generation from doctor-patient conversations (Ben Abacha et al., 2023a). Systems ranged from fine-tuned transformer models with dialogue-aware or retrieval-based components to in-context learning approaches based on LLMs. Results suggest that in-context learning is competitive with fully fine-tuned systems for sectioned note generation (Giorgi et al., 2023; Tang et al., 2023), while the released datasets have also enabled analysis of faithfulness and error patterns in dialogue-to-note generation (Ben Abacha et al., 2023b).

A similar pattern appears in discharge documentation. The BioNLP 2024 Clinical Text Generation challenge introduced the *Discharge Me!* task for generating brief hospital course and discharge instructions from EHR data (Xu et al., 2024a). Submitted systems explored prompt-based LLMs, encoder-decoder fine-tuning, and hybrid pipelines combining structured extraction with controlled generation (Damm et al., 2024; Socrates et al., 2024; Liu et al., 2024b; Wu et al., 2024a). Dynamic input filtering was reported to improve both automatic scores and clinician-rated completeness and correctness (Wu et al., 2024a). Automatic metrics capture only part of the overall quality, making clinician evaluation necessary to assess correctness, completeness, and readability.

5 Conclusion and Future Directions

This survey reviewed recent advances in NLP for the ED workflow. We covered representative modelling approaches and evaluation metrics for key language-intensive tasks in ED settings: triage classification, clinical interaction summarisation,

automatic diagnosis, radiology report generation, disposition prediction, and discharge instruction generation. We also discussed emerging benchmarks and shared tasks that move evaluation beyond exam-style question answering toward more realistic clinical workflows. Across these tasks, several common trends emerge: (i) recent work has shifted from task-specific neural architectures toward pretrained transformer models and LLMs; (ii) evaluation settings are increasingly open-ended, multimodal, and process-oriented, reflecting the complexity of real-world clinical workflows; (iii) many systems are no longer framed as isolated prediction models, but as interactive tools for summarisation, reasoning, and decision support. Despite this progress, our review surfaces several gaps that we believe should shape ED-NLP research.

Data limitations and generalisability. Many studies rely on single-institution datasets (Johnson et al., 2023a) or online health communities (Xu et al., 2019), raising concerns about ecological validity. Documentation style, patient population, case mix, and local workflow vary substantially across institutions, so broader multi-institution evaluation and purpose-built ED corpora outside the MIMIC family are needed for generalisability. Simulation frameworks such as TriageSim (Srirag et al., 2026) generate persona-conditioned triage dialogue and aligned audio from structured EHR seed data. This offers one route to working around regulatory constraints on direct patient recording.

Robustness to real-world clinical data. Many systems assume clean and structured inputs: summarisation models rely on curated transcripts, while triage and diagnosis models assume complete symptom representations. ED data is neither clean nor complete: transcripts contain ASR and diarisation errors (Goss et al., 2016), triage decisions are made under time pressure (Hitchcock et al., 2014), and patient descriptions are often informal or across language barriers (Meeuwesen et al., 2010). Out-of-domain analysis in our corpus (Chen and Hirschberg, 2024) shows substantial degradation from missing information and hallucination rather than format mismatch.

Evaluation and clinical validation. Current evaluation practices are heavily dependent on automated metrics such as accuracy, ROUGE, BLEU, or concept coverage scores (Section 2). While these enable benchmarking, they correlate poorly with clinical usefulness and safety (Moramarco et al., 2022; Savkov et al., 2022). Three concrete shortcomings recur across the corpus. First, almost all results are retrospective or simulated, which limits what can be claimed about real-time clinical use. Second, triage labels are ordinal but are almost always evaluated as flat multi-class classification (Section 3.1), so reported accuracies do not distinguish small misclassifications from clinically dangerous ones. Future works should instead utilise weighted-κ or rank-correlation metrics. Third, clinician evaluation is reported in only 11 of 46 papers and rarely uses pre-specified protocols. Checklist-based protocols (Savkov et al., 2022) measurably improve both inter-rater agreement and the alignment of automatic metrics with human judgement, and should become a default for ED-facing generation tasks. Evaluations are typically performed using data assembled post-encounter rather than the limited inputs available in real time. Recent benchmarks such as EHR2Dial-Triage (Zhu et al., 2026) address this by linking each patient disclosure to its supporting EHR event and the dialogue turn at which it becomes available.

Deployment, privacy and workflow integration. No paper in our corpus reports clinical deployment within an ED-specific workflow (Section 2), making deployment the most concrete open challenge. Many approaches rely on API-based models, conflicting with privacy regulations. Even locally deployable systems must integrate with EHR infrastructure while remaining transparent and accountable. Concerns around privacy (Javed et al., 2025), bias (Zhao et al., 2024b), hallucination (Zhu et al., 2025b), and explainability (Amin et al., 2026) remain major barriers. Closing this gap will require closer collaboration among NLP researchers, clinicians, and healthcare institutions, alongside more realistic benchmarks and clinically grounded evaluation, so that the next generation of ED-NLP systems can move from offline benchmarks to real-time clinical impact.

Limitations

This survey focuses on a defined subset of clinical NLP problems in the ED, specifically tasks aligned with triage, diagnosis and assessment, and disposition. As a result, it does not aim to cover the broader clinical NLP landscape exhaustively, and adjacent problems such as outpatient triage, inpatient documentation, telemedicine consultations,

and post-discharge follow-up fall outside our scope (Appendix C).

The corpus reflects the language and geographic skew of publicly available ED corpora. Our findings about methodological trends and evaluation practices therefore apply most directly to English- and Chinese-language ED-NLP research, and conclusions about generalisability to other languages and care settings should be drawn cautiously. Similarly, our discussion of radiology report generation is restricted to chest X-rays because that is where MIMIC-CXR concentrates the available ED-adjacent data, and our coverage of other imaging modalities is correspondingly limited.

Finally, this survey synthesises studies that differ substantially in task formulation, dataset design, and evaluation methodology. We report reproduced or self-reported numbers from the original papers where available (Appendix E), but we deliberately avoid head-to-head benchmark rankings across heterogeneous setups. As a result, this work is better suited to identifying broad methodological trends and open challenges than to making strong comparative claims about real-world clinical performance.

Ethical Considerations

The methods discussed in this survey should be treated as supplementary tools for clinicians rather than replacements for clinical expertise. In emergency care, incorrect, incomplete, or poorly calibrated model outputs could contribute to unsafe decisions, and the fluent surface form of LLM-generated text may encourage over-reliance on system recommendations even when the underlying reasoning is unsound. Patient-facing applications, such as discharge instruction generation, require particular caution because errors may directly influence patient understanding and adherence, with disproportionate impact on patients with lower health literacy or limited English proficiency.

Additional risks include privacy concerns arising from API-based deployments of clinical NLP systems, demographic bias in disease prediction and triage models, and limited generalisability across institutions, languages, and workflows. The ED setting compounds these risks because decisions are time-pressured, the patient population is heterogeneous and often vulnerable, and the cost of an under-triage or missed-diagnosis error is high. These concerns reinforce the need for clinician oversight, prospective validation in the target deployment setting, subgroup-aware evaluation, and stronger institutional scrutiny before real-world use.

Acknowledgements

Dipankar Srirag is supported by NHMRC Ideas Grant RG241647, awarded to Padmanesan Narasimhan and Aditya Joshi in 2025. A privacy-preserving AI tool was used to assist with revising portions of the text (sentence structures). All content was manually revised and verified by the authors.

References

Griffin Adams, Emily Alsentzer, Mert Ketenci, Jason Zucker, and Noémie Elhadad. 2021. What's in a summary? laying the groundwork for advances in hospital-course summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4794–4811, Online. Association for Computational Linguistics.

Syed Umar Amin, Mohsen Guizani, and M. Shamim Hossain. 2026. Advances, evaluation, and explainability of large language models in healthcare: A systematic review. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 22(2):1–32.

Asma Ben Abacha, Wen-wai Yim, Griffin Adams, Neal Snider, and Meliha Yetisgen. 2023a. Overview of the MEDIQA-chat 2023 shared tasks on the summarization & generation of doctor-patient conversations. In *Proceedings of the 5th Clinical Natural Language Processing Workshop*, pages 503–513, Toronto, Canada. Association for Computational Linguistics.

Asma Ben Abacha, Wen-wai Yim, Yadan Fan, and Thomas Lin. 2023b. An empirical study of clinical note generation from doctor-patient encounters. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2291–2302, Dubrovnik, Croatia. Association for Computational Linguistics.

Asma Ben Abacha, Wen-wai Yim, Yujuan Fu, Zhaoyi Sun, Meliha Yetisgen, Fei Xia, and Thomas Lin. 2025. MEDEC: A benchmark for medical error detection and correction in clinical notes. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 22539–22550, Vienna, Austria. Association for Computational Linguistics.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child,

Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. *Preprint*, arXiv:2005.14165.

Diane J Chamberlain, Eileen Willis, Robyn Clark, and Genevieve Brideson. 2015. Identification of the severe sepsis patient at triage: a prospective analysis of the australasian triage scale. *Emergency Medicine Journal*, 32(9):690–697.

Emma Chen, Aman Kansal, Julie Chen, Boyang Tom Jin, Julia Rachel Reisler, David A Kim, and Pranav Rajpurkar. 2023a. Multimodal clinical benchmark for emergency care (mc-bec): a comprehensive benchmark for evaluating foundation models in emergency medicine. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

Junying Chen, Dongfang Li, Qingcai Chen, Wenxiu Zhou, and Xin Liu. 2022. Diaformer: Automatic diagnosis via symptoms sequence generation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(4):4432–4440.

Siyuan Chen, Colin A. Grambow, Mojtaba Kadkhodaie Elyaderani, Alireza Sadeghi, Federico Fancellu, and Thomas Schaaf. 2023b. Investigating the utility of synthetic data for doctor-patient conversation summarization. In *INTERSPEECH 2023*, page 2338–2342. ISCA.

Xi Chen, Huahui Yi, Mingke You, WeiZhi Liu, Li Wang, Hairui Li, Xue Zhang, Yingman Guo, Lei Fan, Gang Chen, Qicheng Lao, Weili Fu, Kang Li, and Jian Li. 2025. Enhancing diagnostic capability with multi-agents conversational large language models. *npj Digital Medicine*, 8(1).

Yu-Wen Chen and Julia Hirschberg. 2024. Exploring robustness in doctor-patient conversation summarization: An analysis of out-of-domain SOAP notes. In *Proceedings of the 6th Clinical Natural Language Processing Workshop*, pages 1–9, Mexico City, Mexico. Association for Computational Linguistics.

Zhuohao Chen, Jangwon Kim, Yang Liu, and Shrikanth Narayanan. 2023c. Clinical note section classification on doctor-patient conversations in low-resourced settings. In *Proceedings of the Third Workshop on NLP for Medical Conversations*, pages 1–12, Bali, Indonesia. Association for Computational Linguistics.

Matthew W Cooke and Sarah Jinks. 1999. Does the manchester triage system detect the critically ill? *Emergency Medicine Journal*, 16(3):179–181.

Hendrik Damm, Tabea Margareta Grace Pakull, Bahadır Eryılmaz, Helmut Becker, Ahmad Idrissi-Yaghir, Henning Schäfer, Sergej Schultenkämper, and Christoph M. Friedrich. 2024. WisPerMed at "discharge me!": Advancing text generation in healthcare with large language models, dynamic expert selection, and priming techniques on MIMIC-IV. In *Proceedings of the 23rd Workshop on Biomedical Natural Language Processing*, pages 105–121, Bangkok, Thailand. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Flavio Di Martino and Franca Delmastro. 2022. Explainable ai for clinical and remote health applications: a survey on tabular and time series data. *Artificial Intelligence Review*, 56(6):5261–5315.

Yubo Dong, Hehe Fan, Linchao Zhu, and Yi Yang. 2026. Structured reasoning for LLMs: A unified framework for efficiency and explainability. In *The Fourteenth International Conference on Learning Representations*.

Chengfeng Dou, Zhi Jin, Wenpin Jiao, Haiyan Zhao, Yongqiang Zhao, and Zhengwei Tao. 2023. PlugMed: Improving specificity in patient-centered medical dialogue generation using in-context learning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5050–5066, Singapore. Association for Computational Linguistics.

Chengfeng Dou, Ying Zhang, Zhi Jin, Wenpin Jiao, Haiyan Zhao, Yongqiang Zhao, and Zhengwei Tao. 2024. Integrating physician diagnostic logic into large language models: Preference learning from process feedback. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 2453–2473, Bangkok, Thailand. Association for Computational Linguistics.

Maksim Eremeev, Ilya Valmianski, Xavier Amatriain, and Anitha Kannan. 2023. Injecting knowledge into language generation: a case study in auto-charting after-visit care instructions from medical dialogue. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2373–2390, Toronto, Canada. Association for Computational Linguistics.

Tanguy Espejo, Florian F. Grossmann, Henk B. Riedel, Roland Bingisser, and Christian H. Nickel. 2025. The emergency severity index (esi) version 5: Simulation of predictive validity and triage level distribution. *The Journal of Emergency Medicine*, 78:57–70.

ETEK. 2024. Emergency triage education kit, second edition.

Marta Fernandes, Rúben Mendes, Susana M. Vieira, Francisca Leite, Carlos Palos, Alistair Johnson, Stan Finkelstein, Steven Horng, and Leo Anthony Celi. 2020. Predicting intensive care unit admission

among patients presenting to the emergency department using machine learning and natural language processing. *PLOS ONE*, 15(3):e0229331.

Raffaello Fornasiere, Nicolò Brunello, Vincenzo Scotti, and Mark Carman. 2024. Medical information extraction with large language models. In *Proceedings of the 7th International Conference on Natural Language and Speech Processing (ICNLSP 2024)*, pages 456–466, Trento. Association for Computational Linguistics.

André Freitas, Marco Valentino, and Danilo Silva de Carvalho. 2025. Neuro-symbolic natural language processing. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: Tutorial Abstracts*, pages 14–15, Suzhou, China. Association for Computational Linguistics.

Farieda Gaber, Maqsood Shaik, Fabio Allega, Agnes Julia Bilecz, Felix Busch, Kelsey Goon, Vedran Franke, and Altuna Akalin. 2025. Evaluating large language model workflows in clinical decision support for triage and referral and diagnosis. *npj Digital Medicine*, 8(1).

Joseph Gatto, Parker Seegmiller, Timothy E. Burdick, Inas S. Khayal, Sarah DeLozier, and Sarah M. Preum. 2025. Follow-up question generation for enhanced patient-provider conversations. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 25222–25240, Vienna, Austria. Association for Computational Linguistics.

Nicki Gilboy, Paula Tanabe, Debbie A. Travers, Alexander M. Rosenau, and David R. Eitel. 2005. Emergency severity index, version 4: Implementation handbook.

John Giorgi, Augustin Toma, Ronald Xie, Sondra Chen, Kevin An, Grace Zheng, and Bo Wang. 2023. WangLab at MEDIQA-chat 2023: Clinical note generation from doctor-patient conversations using large language models. In *Proceedings of the 5th Clinical Natural Language Processing Workshop*, pages 323–334, Toronto, Canada. Association for Computational Linguistics.

Foster R. Goss, Li Zhou, and Scott G. Weiner. 2016. Incidence of speech recognition errors in the emergency department. *International Journal of Medical Informatics*, 93:70–73.

Colin Grambow, Longxiang Zhang, and Thomas Schaaf. 2022. In-domain pre-training improves clinical note generation from doctor-patient conversations. In *Proceedings of the First Workshop on Natural Language Generation in Healthcare*, pages 9–22, Waterville, Maine, USA and virtual meeting. Association for Computational Linguistics.

Yuhong He, Yongqi Zhang, Shizhu He, and Jun Wan. 2024. BP4ER: Bootstrap prompting for explicit reasoning in medical dialogue generation. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 2480–2492, Torino, Italia. ELRA and ICCL.

Maree Hitchcock, Brigid Gillespie, Julia Crilly, and Wendy Chaboyer. 2014. Triage: an investigation of the process and potential vulnerabilities. *Journal of Advanced Nursing*, 70(7):1532–1541.

Zhenyu Hou, Yukuo Cen, Ziding Liu, Dongxue Wu, Baoyan Wang, Xuanhe Li, Lei Hong, and Jie Tang. 2023. Mtdiag: An effective multi-task framework for automatic diagnosis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 14241–14248.

Wui Ip, Maria Xenochristou, Elaine Sui, Elyse Ruan, Ryan Ribeira, Debadutta Dash, Malathi Srinivasan, Maja Artandi, Jesutofunmi A. Omiye, Nicholas Scoulios, Hayden L. Hofmann, Ali Mottaghi, Zhenzhen Weng, Abhinav Kumar, Ananya Ganesh, Jason Fries, Serena Yeung-Levy, and Lawrence V. Hofmann. 2024. Hospitalization prediction from the emergency department using computer vision ai with short patient video clips. *npj Digital Medicine*, 7(1).

Haseeb Javed, Farman Ali, Babar Shah, Naqqash Dilshad, and Daehan Kwak. 2025. Mediguard: Protecting sensitive healthcare data with privacy-preserving language models. *IEEE Journal of Biomedical and Health Informatics*, pages 1–14.

Serena Jeblee, Faiza Khan Khattak, Noah Crampton, Muhammad Mamdani, and Frank Rudzicz. 2019. Extracting relevant information from physician-patient dialogues for automated clinical note taking. In *Proceedings of the Tenth International Workshop on Health Text Mining and Information Analysis (LOUHI 2019)*, pages 65–74, Hong Kong. Association for Computational Linguistics.

Mingyi Jia, Junwen Duan, Yan Song, and Jianxin Wang. 2025. medIKAL: Integrating knowledge graphs as assistants of LLMs for enhanced clinical diagnosis on EMRs. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 9278–9298, Abu Dhabi, UAE. Association for Computational Linguistics.

Carlos E Jimenez, John Yang, Alexander Wettig, Shunyu Yao, Kexin Pei, Ofir Press, and Karthik R Narasimhan. 2024. SWE-bench: Can language models resolve real-world github issues? In *The Twelfth International Conference on Learning Representations*.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2021a. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14).

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2021b. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14):6421.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. 2019. PubMedQA: A dataset for biomedical research question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2567–2577, Hong Kong, China. Association for Computational Linguistics.

Alistair Johnson, Lucas Bulgarelli, Tom Pollard, Leo Anthony Celi, Roger Mark, and Steven Horng. 2023a. Mimic-iv-ed.

Alistair Johnson, Tom Pollard, Steven Horng, Leo Anthony Celi, and Roger Mark. 2023b. Mimic-iv-note: Deidentified free-text clinical notes.

Alistair E. W. Johnson, Tom J. Pollard, Seth J. Berkowitz, Nathaniel R. Greenbaum, Matthew P. Lungren, Chih-ying Deng, Roger G. Mark, and Steven Horng. 2019. Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific Data*, 6(1).

Alistair E.W. Johnson, Tom J. Pollard, Lu Shen, Li-wei H. Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G. Mark. 2016. Mimic-iii, a freely accessible critical care database. *Scientific Data*, 3(1).

Peter Jones and Karen Schimanski. 2010. The four hour target to reduce emergency department 'waiting time': a systematic review of clinical outcomes. *Emergency Medicine Australasia*, 22(5):391–398.

Anirudh Joshi, Namit Katariya, Xavier Amatriain, and Anitha Kannan. 2020. Dr. summarize: Global summarization of medical dialogue by exploiting local structures. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3755–3763, Online. Association for Computational Linguistics.

Chan-Yang Ju and Dong-Ho Lee. 2025. Prediction-augmented generation for automatic diagnosis tasks. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 14225–14246, Vienna, Austria. Association for Computational Linguistics.

Abirami Kirubarajan, Ahmed Taher, Shawn Khan, and Sameer Masood. 2020. Artificial intelligence in emergency medicine: A scoping review. *JACEP Open*, 1(6):1691–1702.

Katrin Klug, Katharina Beckh, Dario Antweiler, Nilesh Chakraborty, Giulia Baldini, Katharina Laue, René Hosch, Felix Nensa, Martin Schuler, and Sven Giesselbach. 2024. From admission to discharge: a systematic review of clinical natural language processing along the patient journey. *BMC Medical Informatics and Decision Making*, 24(1).

Simon Kocbek, Karin Verspoor, and Wray Buntine. 2014. Exploring temporal patterns in emergency department triage notes with topic models. In *Proceedings of the Australasian Language Technology Association Workshop 2014*, pages 113–117, Melbourne, Australia.

Kundan Krishna, Sopan Khosla, Jeffrey Bigham, and Zachary C. Lipton. 2021. Generating SOAP notes from doctor-patient conversations using modular summarization techniques. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4958–4972, Online. Association for Computational Linguistics.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA. Curran Associates Inc.

Jiazhao Li, Corey Lester, Xinyan Zhao, Yuting Ding, Yun Jiang, and V.G.Vinod Vydiswaran. 2020. PharmMT: A neural machine translation approach to simplify prescription directions. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2785–2796, Online. Association for Computational Linguistics.

Shuyue Stella Li, Vidhisha Balachandran, Shangbin Feng, Jonathan S. Ilgen, Emma Pierson, Pang Wei Koh, and Yulia Tsvetkov. 2024. Mediq: Question-asking llms and a benchmark for reliable interactive clinical reasoning. In *Advances in Neural Information Processing Systems*, volume 37, pages 28858–28888. Curran Associates, Inc.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Shuai Lin, Pan Zhou, Xiaodan Liang, Jianheng Tang, Ruihui Zhao, Ziliang Chen, and Liang Lin. 2021. Graph-evolving meta-learning for low-resource medical dialogue generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 13362–13370.

Fenglin Liu, Zheng Li, Hongjian Zhou, Qingyu Yin, Jingfeng Yang, Xianfeng Tang, Chen Luo, Ming Zeng, Haoming Jiang, Yifan Gao, Priyanka Nigam, Sreyashi Nag, Bing Yin, Yining Hua, Xuan Zhou, Omid Rohanian, Anshul Thakur, Lei Clifton, and David A. Clifton. 2024a. Large language models are poor clinical decision-makers: A comprehensive benchmark. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 13696–13710, Miami, Florida, USA. Association for Computational Linguistics.

Fenglin Liu, Bang Yang, Chenyu You, Xian Wu, Shen Ge, Zhangdaihong Liu, Xu Sun, Yang Yang, and David Clifton. 2022. Retrieve, reason, and refine: Generating accurate and faithful patient instructions. In *Advances in Neural Information Processing Systems*, volume 35, pages 18864–18877. Curran Associates, Inc.

Jinghui Liu, Aaron Nicolson, Jason Dowling, Bevan Koopman, and Anthony Nguyen. 2024b. e-health CSIRO at "discharge me!" 2024: Generating discharge summary sections with fine-tuned language models. In *Proceedings of the 23rd Workshop on Biomedical Natural Language Processing*, pages 675–684, Bangkok, Thailand. Association for Computational Linguistics.

Tuo Liu, Yang Gu, Hongyi Chen, Yan Zhang, Leqi Zheng, Xuanqi Huang, Yanjun Xu, Cai Wen, Mansheng Chen, Jiaqi Lin, Dongguo Huang, Feixia Chen, Yulan Zhong, Hui Chen, Yanfeng Guo, Mei Lu, Guangwei Zhang, Hao Wu, Changdong Wang, and 3 others. 2025. A foundational triage system for improving accuracy in moderate acuity level emergency classifications. *Communications Medicine*, 5(1).

Valentin Liévin, Christoffer Egeberg Hother, Andreas Geert Motzfeldt, and Ole Winther. 2023. Can large language models reason about medical questions? *Preprint*, arXiv:2207.08143.

Justin Lovelace and Bobak Mortazavi. 2020. Learning to generate clinically coherent chest X-ray reports. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1235–1243, Online. Association for Computational Linguistics.

Meng Lu, Brandon Ho, Dennis Ren, and Xuan Wang. 2024. TriageAgent: Towards better multi-agents collaborations for large language model-based clinical triage. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 5747–5764, Miami, Florida, USA. Association for Computational Linguistics.

Renqian Luo, Liai Sun, Yingce Xia, Tao Qin, Sheng Zhang, Hoifung Poon, and Tie-Yan Liu. 2022. Biogpt: generative pre-trained transformer for biomedical text generation and mining. *Briefings in Bioinformatics*, 23(6).

Chenyang Lyu, Zefeng Du, Jitao Xu, Yitao Duan, Minghao Wu, Teresa Lynn, Alham Fikri Aji, Derek F. Wong, and Longyue Wang. 2024. A paradigm shift: The future of machine translation lies with large language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 1339–1352, Torino, Italia. ELRA and ICCL.

Lars Masanneck, Linea Schmidt, Antonia Seifert, Tristan Kölsche, Niklas Huntemann, Robin Jansen, Mohammed Mehsin, Michael Bernhard, Sven G Meuth, Lennert Böhm, and Marc Pawlitzki. 2024. Triage performance across large language models, chatgpt, and untrained doctors in emergency medicine: Comparative study. *Journal of Medical Internet Research*, 26:e53297.

Faraz Maschhur, Klaus Netter, Sven Schmeier, Katrin Ostermann, Rimantas Palunis, Tobias Strapatsas, and Roland Roller. 2024. Towards ML-supported triage prediction in real-world emergency room scenarios. In *Proceedings of the 23rd Workshop on Biomedical Natural Language Processing*, pages 559–569, Bangkok, Thailand. Association for Computational Linguistics.

Suzanne Mason, Ellen J Weber, Joanne Coster, Jennifer Freeman, and Thomas Locker. 2012. Time patients spend in the emergency department: England's 4-hour rule—a case of hitting the target but missing the point? *Annals of emergency medicine*, 59(5):341–349.

Ludwien Meeuwesen, Sione Twilt, Jan D. ten Thije, and Hans Harmsen. 2010. "ne diyor?" (what does she say?): Informal interpreting in general practice. *Patient Education and Counseling*, 81(2):198–203.

Dabin Min, Kaeun Kim, Jong Hyuk Lee, Yisak Kim, and Chang Min Park. 2022. RRED : A radiology report error detector based on deep learning framework. In *Proceedings of the 4th Clinical Natural Language Processing Workshop*, pages 41–52, Seattle, WA. Association for Computational Linguistics.

Francesco Moramarco, Alex Papadopoulos Korfiatis, Mark Perera, Damir Juric, Jack Flann, Ehud Reiter, Anya Belz, and Aleksandar Savkov. 2022. Human evaluation and correlation with automatic metrics in consultation note generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5739–5754, Dublin, Ireland. Association for Computational Linguistics.

Varun Nair, Elliot Schumacher, and Anitha Kannan. 2023. Generating medically-accurate summaries of patient-provider dialogue: A multi-stage approach using large language models. In *Proceedings of the 5th Clinical Natural Language Processing Workshop*, pages 200–217, Toronto, Canada. Association for Computational Linguistics.

Usman Naseem, Ajay Bandi, Shaina Raza, Junaid Rashid, and Bharathi Raja Chakravarthi. 2022. Incorporating medical knowledge to transformer-based language models for medical dialogue generation. In *Proceedings of the 21st Workshop on Biomedical Language Processing*, pages 110–115, Dublin, Ireland. Association for Computational Linguistics.

Aaron Nicolson, Shengyao Zhuang, Jason Dowling, and Bevan Koopman. 2025. The impact of auxiliary patient data on automated chest X-ray report generation and how to incorporate it. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 177–203,

Vienna, Austria. Association for Computational Linguistics.

Matthew J Page, Joanne E McKenzie, Patrick M Bossuyt, Isabelle Boutron, Tammy C Hoffmann, Cynthia D Mulrow, Larissa Shamseer, Jennifer M Tetzlaff, Elie A Akl, Sue E Brennan, Roger Chou, Julie Glanville, Jeremy M Grimshaw, Asbjørn Hróbjartsson, Manoj M Lalu, Tianjing Li, Elizabeth W Loder, Evan Mayo-Wilson, Steve McDonald, and 7 others. 2021. The prisma 2020 statement: an updated guideline for reporting systematic reviews. *BMJ*, page n71.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Aske Plaat, Max Van Duijn, Niki Van Stein, Mike Preuss, Peter Van der Putten, and Kees Joost Batenburg. 2025. Agentic large language models, a survey. *Journal of Artificial Intelligence Research*, 84.

Gregory Polyakov, Ilseyar Alimova, Dmitry Abulkhanov, Ivan Sedykh, Andrey Bout, Sergey Nikolenko, and Irina Piontkovskaya. 2025. Tool-Reflection: Improving large language models for real-world API calls with self-generated data. In *Proceedings of the 1st Workshop for Research on Agent Language Models (REALM 2025)*, pages 184–199, Vienna, Austria. Association for Computational Linguistics.

Lang Qin, Yao Zhang, Hongru Liang, Adam Jatowt, and Zhenglu Yang. 2025. Listening to patients: Detecting and mitigating patient misreport in medical dialogue system. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 2650–2664, Vienna, Austria. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(1).

M Arif Rahman, Sarah M. Preum, Ronald Williams, Homa Alemzadeh, and John A. Stankovic. 2020. Grace: Generating summary reports automatically for cognitive assistance in emergency response. volume 34, pages 13356–13362.

Maria C. Raven, Robert A. Lowe, Judith Maselli, and Renee Y. Hsia. 2013. Comparison of presenting complaint vs discharge diagnosis for identifying "nonemergency" emergency department visits. *JAMA*, 309(11):1145.

Daniel Philip Rose, Chia-Chien Hung, Marco Lepri, Israa Alqassem, Kiril Gashteovski, and Carolin Lawrence. 2025. MEDDxAgent: A unified modular agent framework for explainable automatic differential diagnosis. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13803–13826, Vienna, Austria. Association for Computational Linguistics.

Filip Rusak, Bevan Koopman, Nathan J. Brown, Kevin Chu, Jinghui Liu, and Anthony Nguyen. 2023. Catching misdiagnosed limb fractures in the emergency department using cross-institution transfer learning. In *Proceedings of the 21st Annual Workshop of the Australasian Language Technology Association*, pages 78–87, Melbourne, Australia. Association for Computational Linguistics.

Aleksandar Savkov, Francesco Moramarco, Alex Papadopoulos Korfiatis, Mark Perera, Anya Belz, and Ehud Reiter. 2022. Consultation checklists: Standardising the human evaluation of medical note generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 111–120, Abu Dhabi, UAE. Association for Computational Linguistics.

Benjamin Schloss and Sandeep Konam. 2020. Towards an automated soap note: Classifying utterances from medical conversations. In *Proceedings of the 5th Machine Learning for Healthcare Conference*, volume 126 of *Proceedings of Machine Learning Research*, pages 610–631. PMLR.

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1073–1083, Vancouver, Canada. Association for Computational Linguistics.

Jingyuan S. She, Christopher Potts, Samuel R. Bowman, and Atticus Geiger. 2023. ScoNe: Benchmarking negation reasoning in language models with fine-tuning and in-context learning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1803–1821, Toronto, Canada. Association for Computational Linguistics.

Xiaoming Shi, Zeming Liu, Li Du, Yuxuan Wang, Hongru Wang, Yuhang Guo, Tong Ruan, Jie Xu, Xiaofan Zhang, and Shaoting Zhang. 2024. Medical dialogue system: A survey of categories, methods, evaluation and challenges. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 2840–2861, Bangkok, Thailand. Association for Computational Linguistics.

Gagandeep Singh, Yue Pan, Jesus Andres-Ferrer, Miguel Del-Agua, Frank Diehl, Joel Pinto, and Paul Vozila. 2023. Large scale sequence-to-sequence models for clinical note generation from patient-doctor conversations. In *Proceedings of the 5th Clinical*

Natural Language Processing Workshop, pages 138–143, Toronto, Canada. Association for Computational Linguistics.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S. Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, Perry Payne, Martin Seneviratne, Paul Gamble, Chris Kelly, Abubakr Babiker, Nathanael Schärli, Aakanksha Chowdhery, Philip Mansfield, Dina Demner-Fushman, and 13 others. 2023. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180.

Vimig Socrates, Thomas Huang, Xuguang Ai, Soraya Fereydooni, Qingyu Chen, R Andrew Taylor, and David Chartash. 2024. Yale at "discharge me!": Evaluating constrained generation of discharge summaries with unstructured and structured information. In *Proceedings of the 23rd Workshop on Biomedical Natural Language Processing*, pages 724–730, Bangkok, Thailand. Association for Computational Linguistics.

Dipankar Srirag, Quoc Dung Nguyen, Aditya Joshi, Padmanesan Narasimhan, and Salil Kanhere. 2026. Triagesim: A conversational emergency triage simulation framework from structured electronic health records. *Preprint*, arXiv:2603.10035.

Jonathon Stewart, Juan Lu, Adrian Goudie, Glenn Arendts, Shiv Akarsh Meka, Sam Freeman, Katie Walker, Peter Sprivulis, Frank Sanfilippo, Mohammed Bennamoun, and Girish Dwivedi. 2023. Applications of natural language processing at emergency department triage: A narrative review. *PLOS ONE*, 18(12):e0279953–.

Jing Su, Longxiang Zhang, Hamid Reza Hassanzadeh, and Thomas Schaaf. 2022. Extract and abstract with bart for clinical notes from doctor-patient conversations. In *Interspeech 2022*, page 2488–2492. ISCA.

Hongda Sun, Jiaren Peng, Wenzhong Yang, Liang He, Bo Du, and Rui Yan. 2025. Enhancing medical dialogue generation through knowledge refinement and dynamic prompt adjustment. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 25715–25726, Vienna, Austria. Association for Computational Linguistics.

Liwen Sun, Abhineet Agarwal, Aaron Kornblith, Bin Yu, and Chenyan Xiong. 2024. Ed-copilot: reduce emergency department wait time with language model diagnostic assistance. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org.

Xiangru Tang, Andrew Tran, Jeffrey Tan, and Mark Gerstein. 2023. GersteinLab at MEDIQA-chat 2023: Clinical note summarization from doctor-patient conversations through fine-tuning and in-context learning. In *Proceedings of the 5th Clinical Natural Language Processing Workshop*, pages 546–554, Toronto, Canada. Association for Computational Linguistics.

Arsène Fansi Tchango, Rishab Goel, Julien Martel, Zhi Wen, Gaétan Marceau Caron, and Joumana Ghosn. 2022. Towards trustworthy automatic diagnosis systems by emulating doctors' reasoning with deep reinforcement learning. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA. Curran Associates Inc.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, and 49 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

Samantha Tyler, Matthew Olis, Nicole Aust, Love Patel, Leah Simon, Catherine Triantafyllidis, Vijay Patel, Dong Won Lee, Brendan Ginsberg, Hiba Ahmad, and Robin J Jacobs. 2024. Use of artificial intelligence in triage in hospital emergency departments: A scoping review. *Cureus*.

Mina Valizadeh and Natalie Parde. 2022. The AI doctor is in: A survey of task-oriented dialogue systems for healthcare applications. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6638–6660, Dublin, Ireland. Association for Computational Linguistics.

Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. *SuperGLUE: a stickier benchmark for general-purpose language understanding systems*. Curran Associates Inc., Red Hook, NY, USA.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Benyou Wang, Qianqian Xie, Jiahuan Pei, Zhihong Chen, Prayag Tiwari, Zhao Li, and Jie Fu. 2023a. Pre-trained language models in biomedical domain: A systematic survey. *ACM Comput. Surv.*, 56(3).

Hanyin Wang, Chufan Gao, Bolun Liu, Qiping Xu, Guleid Hussein, Mohamad El Labban, Kingsley Iheasirim, Hariprasad Reddy Korsapati, Chuck Outcalt, and Jimeng Sun. 2025a. Towards adapting open-source large language models for expert-level clinical

note generation. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 12084–12117, Vienna, Austria. Association for Computational Linguistics.

Huimin Wang, Wai Chung Kwan, Kam-Fai Wong, and Yefeng Zheng. 2023b. CoAD: Automatic diagnosis through symptom and disease collaborative generation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6348–6361, Toronto, Canada. Association for Computational Linguistics.

Wenxuan Wang, Zizhan Ma, Zheng Wang, Chenghan Wu, Jiaming Ji, Wenting Chen, Xiang Li, and Yixuan Yuan. 2025b. A survey of LLM-based agents in medicine: How far are we from baymax? In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 10345–10359, Vienna, Austria. Association for Computational Linguistics.

Zhongyu Wei, Qianlong Liu, Baolin Peng, Huaixiao Tou, Ting Chen, Xuanjing Huang, Kam-fai Wong, and Xiangying Dai. 2018. Task-oriented dialogue system for automatic diagnosis. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 201–207, Melbourne, Australia. Association for Computational Linguistics.

Benjamin Winter, Alexei Gustavo Figueroa Rosero, Alexander Loeser, Felix Alexander Gers, Nancy Katerina Figueroa Rosero, and Ralf Krestel. 2024. DDxGym: Online transformer policies in a knowledge graph based natural language environment. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 4438–4448, Torino, Italia. ELRA and ICCL.

Haotian Wu, Paul Boulenger, Antonin Faure, Berta Céspedes, Farouk Boukil, Nastasia Morel, Zeming Chen, and Antoine Bosselut. 2024a. EPFL-MAKE at "discharge me!": An LLM system for automatically generating discharge summaries of clinical electronic health record. In *Proceedings of the 23rd Workshop on Biomedical Natural Language Processing*, pages 696–711, Bangkok, Thailand. Association for Computational Linguistics.

Jiageng Wu, Xiaocong Liu, Minghui Li, Wanxin Li, Zichang Su, Shixu Lin, Lucas Garay, Zhiyun Zhang, Yujie Zhang, Qingcheng Zeng, Jie Shen, Changzheng Yuan, and Jie Yang. 2024b. Clinical text datasets for medical artificial intelligence and large language models — a systematic review. *NEJM AI*, 1(6):AIra2400012.

Stephen Wu, Kirk Roberts, Surabhi Datta, Jingcheng Du, Zongcheng Ji, Yuqi Si, Sarvesh Soni, Qiong Wang, Qiang Wei, Yang Xiang, and 1 others. 2020. Deep learning in clinical natural language processing: a methodical review. *Journal of the American Medical Informatics Association*, 27(3):457–470.

Xian Wu, Yutian Zhao, Yunyan Zhang, Jiageng Wu, Zhihong Zhu, Yingying Zhang, Yi Ouyang, Ziheng Zhang, Huimin Wang, Zhenxi Lin, Jie Yang, Shuang Zhao, and Yefeng Zheng. 2024c. Medjourney: Benchmark and evaluation of large language models over patient clinical journey. In *Advances in Neural Information Processing Systems*, volume 37, pages 87621–87646. Curran Associates, Inc.

Peng Xia, Kangyu Zhu, Haoran Li, Tianze Wang, Weijia Shi, Sheng Wang, Linjun Zhang, James Zou, and Huaxiu Yao. 2025. MMed-RAG: Versatile multimodal RAG system for medical vision language models. In *The Thirteenth International Conference on Learning Representations*.

Justin Xu, Zhihong Chen, Andrew Johnston, Louis Blankemeier, Maya Varma, Jason Hom, William J. Collins, Ankit Modi, Robert Lloyd, Benjamin Hopkins, Curtis Langlotz, and Jean-Benoit Delbrouck. 2024a. Overview of the first shared task on clinical text generation: RRG24 and "discharge me!". In *Proceedings of the 23rd Workshop on Biomedical Natural Language Processing*, pages 85–98, Bangkok, Thailand. Association for Computational Linguistics.

Kaishuai Xu, Yi Cheng, Wenjun Hou, Qiaoyu Tan, and Wenjie Li. 2024b. Reasoning like a doctor: Improving medical dialogue systems via diagnostic reasoning process alignment. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 6796–6814, Bangkok, Thailand. Association for Computational Linguistics.

Kaishuai Xu, Wenjun Hou, Yi Cheng, Jian Wang, and Wenjie Li. 2023. Medical dialogue generation via dual flow modeling. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 6771–6784, Toronto, Canada. Association for Computational Linguistics.

Lin Xu, Qixian Zhou, Ke Gong, Xiaodan Liang, Jianheng Tang, and Liang Lin. 2019. End-to-end knowledge-routed relational dialogue system for automatic diagnosis. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):7346–7353.

Zonghai Yao and Hong Yu. 2026. LLM-based multiagent systems for clinical workflows: A survey of AI hospitals. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 45772–45793, San Diego, California, United States. Association for Computational Linguistics.

Wen-wai Yim, Yujuan Fu, Asma Ben Abacha, Neal Snider, Thomas Lin, and Meliha Yetisgen. 2023. Aci-bench: a novel ambient clinical intelligence dataset for benchmarking automatic visit note generation. *Scientific Data*, 10(1).

Dong Yuan, Eti Rastogi, Gautam Naik, Sree Prasanna Rajagopal, Sagar Goyal, Fen Zhao, Bharath Chintagunta, and Jeffrey Ward. 2024. A continued pretrained LLM approach for automatic medical note

generation. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 565–571, Mexico City, Mexico. Association for Computational Linguistics.

Qin Zhang, Shangsi Chen, Dongkuan Xu, Qingqing Cao, Xiaojun Chen, Trevor Cohn, and Meng Fang. 2023. A survey for efficient open domain question answering. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14447–14465, Toronto, Canada. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Weike Zhao, Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2024a. RaTEScore: A metric for radiology report generation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 15004–15019, Miami, Florida, USA. Association for Computational Linguistics.

Yutian Zhao, Huimin Wang, Yuqi Liu, Wu Suhuang, Xian Wu, and Yefeng Zheng. 2024b. Can LLMs replace clinical doctors? exploring bias in disease diagnosis by large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 13914–13935, Miami, Florida, USA. Association for Computational Linguistics.

Yongxin Zhou, Fabien Ringeval, and François Portet. 2023. A survey of evaluation methods of generated medical textual reports. In *Proceedings of the 5th Clinical Natural Language Processing Workshop*, pages 447–459, Toronto, Canada. Association for Computational Linguistics.

Haohao Zhu, Xiaolin Shi, and Jiayu Zhou. 2026. Elicited: Ehr-grounded longitudinal interactive conversations for information-seeking triage evaluation and decision-making. *Preprint*, arXiv:2608.09024.

Rongxin Zhu, Jey Han Lau, and Jianzhong Qi. 2025a. Factual dialogue summarization via learning from large language models. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 4474–4492, Abu Dhabi, UAE. Association for Computational Linguistics.

Zhihong Zhu, Yunyan Zhang, Xianwei Zhuang, Fan Zhang, Zhongwei Wan, Yuyan Chen, QingqingLong QingqingLong, Yefeng Zheng, and Xian Wu. 2025b. Can we trust AI doctors? a survey of medical hallucination in large language and large vision-language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 6748–6769, Vienna, Austria. Association for Computational Linguistics.

Zihan Zhu, Qiongkai Xu, and Amin Beheshti. 2025c. Heal: Healthcare emergency assistants leveraging large language models. In *Companion Proceedings of the ACM on Web Conference 2025*, WWW '25, page 2959–2962, New York, NY, USA. Association for Computing Machinery.

A Emergency Department Phases

To contextualise the NLP tasks reviewed later, this section outlines the three main phases of the ED workflow: triage and arrival, diagnosis and assessment, and patient disposition. These phases structure how information is collected, documented, and acted upon during an ED encounter, and they motivate the task groupings used in the remainder of the survey.

Triage and Arrival

As shown in Figure 1, the ED workflow begins with patient arrival, whether by self-presentation or emergency services transport. At this stage, a triage nurse rapidly evaluates the patient's condition and assigns an appropriate triage category. While some systems also employ a general practitioner, triage in most EDs is performed exclusively by nurses, who follow structured protocols rather than diagnostic reasoning. This distinction matters for NLP, since tools designed for triage must align with protocol-driven decision rules rather than physician-style assessment. Several triage scales are used internationally, including the Emergency Severity Index (ESI) in the United States (Espejo et al., 2025), the Australasian Triage Scale (ATS) in Australia (Chamberlain et al., 2015), the Manchester Triage Scale in Europe (Cooke and Jinks, 1999), and the CETEC in China (Liu et al., 2025). While these scales share a five-level ordinal structure (with Level 1 denoting the most urgent cases), their underlying criteria differ. For instance, the ESI assigns triage levels based on the anticipated resource requirements, whereas the ATS focuses on the maximum acceptable waiting time before physician assessment. Beyond triage, clinicians generate brief intake notes summarising the chief complaint, relevant history, vital signs, and the assigned category, which form the basis for subsequent decision-making.

Diagnosis and Assessment

Following triage, patients progress to the diagnosis and assessment phase, which forms the core of clinical decision-making in the ED. Physicians combine information from the initial hand-off notes with elicited symptoms, examination findings, and

diagnostic investigations such as laboratory tests or imaging. Documentation at this stage often follows the SOAP structure, consisting of *Subjective* (patient-reported complaints), *Objective* (observed signs and results), *Assessment* (provisional or confirmed diagnosis), and *Plan* (treatment and disposition). Depending on clinical presentation and local workflow, patients may be observed for a short stay before a final disposition decision is made.

Patient Disposition

The final phase of an ED encounter involves disposition, where the clinician formalises the next stage of care. Patients may be discharged with prescriptions and follow-up instructions or admitted to an inpatient ward or intensive care unit. The discharge summary consolidates information about presenting complaints, diagnostic findings, treatment plans, and recommendations for ongoing management. Beyond serving as documentation, these summaries are critical for ensuring continuity of care across providers.

B The Promise of NLP in the ED

Over the years, language technologies have progressed from sparse vector models to large-scale neural sequence models. Early approaches relied on bag-of-words representations, such as TF-IDF and n-gram language models, for tasks such as document retrieval, topic detection, and basic text classification. These were followed by linear and probabilistic models, including logistic regression, maximum entropy models, and conditional random fields, which improved performance on sequence labeling tasks such as part-of-speech tagging and named entity recognition. A major shift came with distributed word representations, in which neural language models and word embeddings replaced sparse features with dense vectors that encode semantic similarity. These representations enabled neural architectures, particularly recurrent networks such as LSTMs and GRUs, to become the dominant paradigm across a wide range of applications, including machine translation, dialogue state tracking, and extractive summarisation. Sequence-to-sequence models with attention further unified many tasks under a common encoder-decoder formulation, where the same architecture could be instantiated for translation, abstractive summarisation, and question answering by changing the training data and objective.

Transformer architectures and large-scale pretraining have since redefined the state of the art for both natural language understanding (NLU) and natural language generation (NLG). Encoder-only models trained with masked language modelling objectives (Devlin et al., 2019) are now routinely fine-tuned on multi-task benchmarks such as GLUE (Wang et al., 2018) and SuperGLUE (Wang et al., 2019) for sentence classification, textual entailment, and natural language inference. Decoder-only (Touvron et al., 2023) and encoder-decoder models (Raffel et al., 2020) trained with causal or span-corruption objectives underpin contemporary systems for open-domain question answering (Zhang et al., 2023), dialogue summarisation (Zhu et al., 2025a), and machine translation (Lyu et al., 2024). These models benefit from training on orders of magnitude more text than earlier systems and from parameter counts that support rich internal representations of context, discourse structure, and world knowledge. LLMs extend this trend by scaling data, model size, and training objectives to the point where a single model can act as a general-purpose language engine. When prompted with natural language instructions and a few examples, LLMs can perform new tasks with little or no task-specific fine-tuning, a capability often described as in-context learning (Brown et al., 2020). They show strong performance on broad multitask benchmarks that cover knowledge-intensive and reasoning-heavy problems (She et al., 2023), as well as specialised benchmarks that evaluate code-generation and software-engineering capabilities, such as program synthesis and issue resolution (Jimenez et al., 2024). In addition, tool-augmented variants integrate retrieval (Lewis et al., 2020), structured reasoning (Dong et al., 2026), and external application programming interfaces (Plaat et al., 2025; Polyakov et al., 2025), enabling them to combine symbolic computation with free-text generation (Freitas et al., 2025).

Building on these advances, a growing body of work has focused on clinical NLP, where models are trained or adapted on electronic health records, clinical notes, radiology reports, and biomedical question answering datasets. Domain-specific pretraining and continued pretraining on de-identified clinical corpora have led to encoders and generative models that better capture medical terminology, abbreviations, and documentation conventions than general-purpose LMs. Evaluations on standard clinical benchmarks for concept extraction, document

classification, natural language inference, and medical question answering consistently show gains when using such specialised models, particularly for tasks that require detailed knowledge of clinical workflows and disease-specific language (Wang et al., 2023a; Liévin et al., 2023; Singhal et al., 2023). From a systems perspective, these models reuse the same architectural components as general-purpose LLMs, i.e., large-scale pretraining, task-specific fine-tuning or instruction tuning, retrieval augmentation, and safety layers, but with training data, ontologies, and evaluation criteria tailored to healthcare (Wang et al., 2023a).

Research on NLP applied to ED clinical data has predominantly focused on predicting outcomes from unstructured free-text triage notes, such as assignment of triage score, need for admission, or critical illness. A recent scoping review of NLP at emergency department triage found that, although NLP models can achieve high predictive accuracy for clinically relevant outcomes and combining free-text with structured data often improves results, the majority of studies exhibited high risk of bias and were retrospective in design. Moreover, only a single study reported actual deployment of an NLP model into clinical practice, highlighting a gap between retrospective performance and real-world impact (Stewart et al., 2023; Wu et al., 2024b). This pattern illustrates both the *promise* and *limitation* of current ED NLP: models can exploit narrative clinical content effectively, yet their validation and integration into clinical workflows remain limited.

Despite widespread adoption of language technologies in other healthcare domains (Valizadeh and Parde, 2022), their direct application to the ED remains limited. Short, data-sparse encounters and the need for rapid documentation constrain methods that depend on longitudinal histories or rich contextual data. Privacy, safety, and workflow integration requirements also make it harder to reuse generic language models without adaptation. These conditions underscore the need for LTs that can operate under time pressure, efficiently integrate multimodal information, and support timely communication between clinicians. Potential applications include triage note classification, automated summarisation of clinician notes, dialogue systems for patient intake, and decision-support tools tailored for real-time environments.

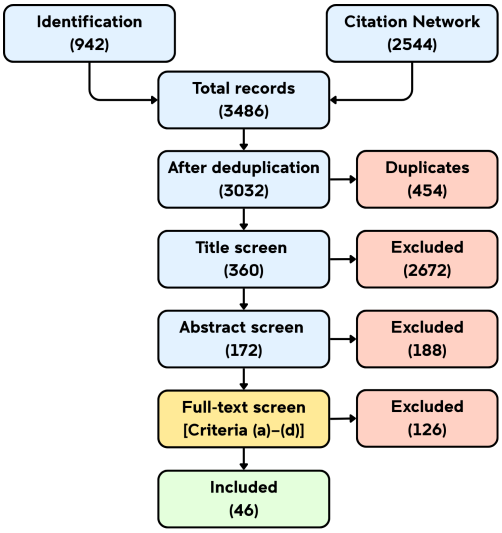

Figure 3: PRISMA-style flow of the search and screening process. Inclusion criteria is described in Section C.

C Scope and Search Criteria

Sources. We searched the following venues in NLP, AI, and health informatics: the ACL Anthology[3], the AAAI Digital Library[4], the NeurIPS proceedings[5], the ACM Digital Library[6], the ISCA Archive[7], the Nature Portfolio[8], and PLOS ONE[9]. The initial search was completed in August 2025; relevant papers published in these venues after this date were retrospectively included up to the submission date.

Search strategy. We queried each source using the terms *emergency department*, *emergency room*, *triage*, *clinical*, *hospital*, *natural language processing*, *NLP*, *artificial intelligence*, *machine learning*, and *clinical*, together with their morphological variants and domain synonyms (e.g., *ED*, *ER*, *A&E*). Terms were combined with AND/OR operators; at least one of *emergency department*, *emergency room*, *triage*, *clinical* or *hospital* was required in all queries. We supplemented keyword search with backward and forward citation snowballing on all retrieved papers.

Inclusion criteria. We follow PRISMA-style reporting (Page et al., 2021) with sequential screening stages. A paper was included if and only if it satisfied all of the following:

[3] https://aclanthology.org
[4] https://aaai.org/aaai-publications/
[5] https://papers.nips.cc
[6] https://dl.acm.org/
[7] https://www.isca-archive.org/
[8] https://www.nature.com/
[9] https://journals.plos.org/plosone/

(a) *ED phase alignment.* The task maps to at least one of the three ED phases defined in Appendix A.
(b) *ED-relevant language resource.* The work uses a language resource that could plausibly arise in an ED setting (e.g., triage notes, chief complaints, symptom descriptions, diagnostic impressions, or patient-clinician interactions) covering presenting complaints or diagnoses representative of ED visits (Raven et al., 2013).
(c) *NLP contribution.* The work proposes or evaluates an NLP method as a primary research contribution, independent of shared-task participation.
(d) *ED-anchored evaluation.* The task or evaluation setting is anchored to ED workflows, evidenced by ED-specific data, ED-specific tasks, or explicit acute-care framing, rather than reliance on general clinical benchmarks or broad inpatient corpora (e.g., MIMIC-III; Johnson et al. 2016) alone.

Screening process. Figure 3 summarises the search and screening flow across three sequential stages. *Stage 1*: title screening to remove clearly off-topic results. *Stage 2*: abstract screening to assess topical fit. *Stage 3*: full-text review against the inclusion criteria below. All three stages were conducted by the lead author. To assess reliability, a second author (an NLP researcher) independently screened a random 20% sample of abstracts reaching Stage 2, yielding a Cohen's κ of 0.80 (rounded to two decimal places), indicating substantial agreement. Disagreements at Stage 2 were resolved by discussion with a third author, an emergency medicine clinician; Stage 3 decisions were made by the lead author. The final corpus comprises 46 peer-reviewed papers that explicitly formulate NLP tasks within emergency department workflows.

D Papers Included in the Survey

Tables 2 and 3 provide a paper-level catalogue of all 46 included papers. We outline the trends across three dimensions: clinical settings, training paradigms, and evaluation paradigms.

E Reported Performances

Tables 4–6 summarise reported performance across the clinical tasks reviewed in this survey. These values should not be interpreted as directly comparable because studies differ substantially in datasets, task formulations, label spaces, and evaluation protocols.

Three broad patterns emerge. First, performance is strongly task dependent: constrained prediction tasks such as triage and disposition prediction generally report stronger headline scores than open-ended generation tasks such as clinical interaction summarisation and radiology report generation. Second, automatic diagnosis shows the greatest evaluation heterogeneity, reflecting the breadth of formulations used in prior work, from disease prediction and symptom acquisition to dialogue generation and multi-step reasoning. Third, newer paradigms such as agent-based, retrieval-augmented, and prompt-based methods are increasingly visible, but their gains remain difficult to interpret consistently because evaluation settings vary widely.

The distribution of results also suggests a clear pattern in training paradigms. Supervised and transfer learning methods remain the most established approaches across tasks, particularly for classification and structured generation settings where labels and evaluation criteria are relatively stable. By contrast, agent-based, prompt-based, and retrieval-augmented methods are more common in tasks that require multi-step reasoning, external knowledge, or flexible generation, such as automatic diagnosis and clinical interaction summarisation. Their growing presence points to methodological expansion rather than convergence, while reinforcement learning remains more specialised and appears mainly in diagnosis settings involving sequential decision-making.

Overall, the table highlights the fragmented state of evaluation in ED-focused NLP. Classification tasks are usually assessed with metrics such as accuracy, AUC, AUROC, and AUPRC, whereas summarisation and generation tasks rely more on ROUGE, BLEU, factuality, and concept-level measures. The table should therefore be read as a compact reference for the range of reported outcomes in the literature rather than as evidence of a single best-performing modelling strategy.

	ED	Phases			Languages			Input modality			Training paradigm								Evaluation paradigm			
Paper	**Tag**	**Triage**	**Diag./Assess.**	**Disposition**	**English**	**Chinese**	**Others**	**Text**	**Struct.**	**Image**	**Pre.**	**TL**	**Sup.**	**RL**	**KG**	**RAG**	**Prompt**	**Agent**	**Retro.**	**Prosp.**	**+Clin.**	**Deploy.**
Wei et al. (2018)	○		✓			✓		✓						✓								
Jeblee et al. (2019)	◐		✓		✓			✓					✓						✓			
Fernandes et al. (2020)	●			✓	✓		✓		✓				✓						✓			
Joshi et al. (2020)	○		✓		✓			✓					✓						✓		✓	
Li et al. (2020)	◐			✓	✓			✓				✓							✓		✓	
Lovelace and Mortazavi (2020)	◐		✓		✓			✓		✓		✓							✓			
Rahman et al. (2020)	●		✓		✓			✓					✓									
Schloss and Konam (2020)	◐		✓		✓			✓					✓						✓			
Krishna et al. (2021)	◐		✓		✓			✓				✓							✓		✓	
Lin et al. (2021)	○		✓			✓			✓						✓							
Tchango et al. (2022)	○		✓		✓				✓					✓								
Grambow et al. (2022)	◐		✓		✓			✓			✓								✓			
Chen et al. (2022)	○		✓			✓		✓	✓				✓									
Liu et al. (2022)	◐			✓	✓			✓											✓		✓	
Naseem et al. (2022)	○		✓		✓			✓	✓						✓						✓	
Su et al. (2022)	◐		✓		✓			✓				✓							✓			
Chen et al. (2023c)	◐		✓		-	-	-	✓				✓										
Chen et al. (2023b)	◐		✓		✓			✓			✓								✓			
Dou et al. (2023)	○		✓		✓	✓		✓	✓									✓			✓	
Eremeev et al. (2023)	◐			✓	-	-	-	✓	✓						✓				✓		✓	
Hou et al. (2023)	○		✓			✓		✓	✓				✓									
Nair et al. (2023)	◐		✓		✓			✓									✓		✓		✓	
Singh et al. (2023)	◐		✓		✓			✓				✓							✓			
Wang et al. (2023b)	○		✓			✓		✓	✓				✓									
Xu et al. (2023)	○		✓			✓		✓					✓								✓	

Table 2: Full catalog of included papers, Part 1 (2018–2023), covering papers from the field's earliest contributions through the emergence of LLM-based approaches. Each paper is annotated across six column groups. The **Tag** column indicates ED-relatedness: ● = ED-native (ED data + ED task), ◐ = ED-anchored (non-ED data + ED-relevant task), ○ = ED-adjacent (telemedicine/online/synthetic data + transferable task). The *Phases* group indicates which stage of the ED workflow the paper addresses: Triage, Diagnosis and Assessment (Diag./Assess.), or Disposition. The *Languages* and *Input modality* groups record the data language and whether the model consumes free text, structured records, or images; a dash in the language columns means the source language is unspecified. The *Training paradigm* group captures the dominant modelling strategy, from classical supervised fine-tuning (Sup.) and transfer learning (TL) through pretraining (Pre.), reinforcement learning (RL), knowledge-grounded generation (KG), retrieval-augmented generation (RAG), prompt-based or in-context learning (Prompt), and agent-based frameworks (Agent). The *Evaluation paradigm* group encodes study rigour: Retro. marks papers that evaluate on real retrospective patient records; Prosp. marks the one prospective study in the corpus; papers with neither mark rely on simulation or synthetic data. +Clin. marks papers whose evaluation includes clinician or human judgment beyond automated metrics; Deploy. marks systems reported as clinically deployed.

	ED	Phases			Languages			Input modality			Training paradigm								Evaluation paradigm			
Paper	**Tag**	**Triage**	**Diag./Assess.**	**Disposition**	**English**	**Chinese**	**Others**	**Text**	**Struct.**	**Image**	**Pre.**	**TL**	**Sup.**	**RL**	**KG**	**RAG**	**Prompt**	**Agent**	**Retro.**	**Prosp.**	**+Clin.**	**Deploy.**
Dou et al. (2024)	○		✓		✓	✓		✓	✓									✓				
He et al. (2024)	○		✓			✓		✓										✓			✓	
Ip et al. (2024)	●			✓	✓		✓			✓			✓							✓		
Lu et al. (2024)	●	✓			✓			✓										✓				
Maschhur et al. (2024)	●	✓					✓		✓			✓							✓			
Sun et al. (2024)	●			✓	✓			✓	✓					✓				✓	✓			
Xu et al. (2024b)	○		✓			✓		✓							✓							
Yuan et al. (2024)	◐		✓		✓			✓			✓								✓			
Chen et al. (2025)	○		✓		✓			✓	✓									✓				
Gaber et al. (2025)	●	✓			✓			✓	✓							✓			✓			
Gatto et al. (2025)	◐		✓		✓			✓	✓				✓						✓			
Zhu et al. (2025c)	●	✓			✓			✓	✓									✓				
Jia et al. (2025)	○		✓			✓		✓	✓						✓				✓			
Ju and Lee (2025)	○		✓			✓		✓	✓									✓				
Liu et al. (2025)	●	✓			✓			✓	✓			✓							✓			
Nicolson et al. (2025)	●		✓		✓			✓	✓				✓						✓			
Qin et al. (2025)	○		✓		✓			✓					✓									
Rose et al. (2025)	○		✓		✓			✓	✓									✓				
Sun et al. (2025)	○		✓		✓	✓		✓	✓									✓				
Wang et al. (2025a)	◐		✓		✓			✓			✓								✓		✓	
Xia et al. (2025)	◐		✓		✓			✓		✓						✓			✓			

Table 3: Full catalog of included papers, Part 2 (2024–2025). Column groups and coding conventions follow Table 2.

Task	Training Paradigm	Paper	Reported performance
Triage classification	Transfer learning	Maschhur et al. (2024)	F1: 0.63
		Liu et al. (2025)	AUC: 0.88
	Agent-based	Lu et al. (2024)	Discordance: 0.19; Acc.: 0.81
		Zhu et al. (2025c)	MAE: 0.27; Acc.: 0.73
	Retrieval-augmented	Gaber et al. (2025)	Acc.: 0.66
Clinical interaction summarisation	Neural architectures	Jeblee et al. (2019)	F1: 0.60
		Schloss and Konam (2020)	F1: 0.44; AUROC: 0.83
		Joshi et al. (2020)	R-L: 0.55
		Rahman et al. (2020)	F1: 0.63
	Transfer learning	Krishna et al. (2021)	R-L: 0.38
		Su et al. (2022)	R-L: 0.34
		Chen et al. (2023c)[a]	Acc. gain: +10.7
		Singh et al. (2023)	R-L: 0.64; Fact-C: 0.69
	Pretraining	Grambow et al. (2022)	R-L: 0.27; UMLS-F1: 0.39
		Chen et al. (2023b)	R-L: 0.27; UMLS-F1: 0.37
		Yuan et al. (2024)[b]	Missed: 4.3; Incorrect: 0.85; Irrelevant: 0.30
		Wang et al. (2025a)	R-L: 0.36; Completeness: 4.3
	Prompt-based	Nair et al. (2023)[c]	GPT-F1: 0.62

Table 4: Reported performance across the reviewed clinical tasks (Part 1 of 3): triage classification and clinical interaction summarisation. See Table 6 for footnotes and abbreviations.

Task	Training Paradigm	Paper	Reported performance
Automatic diagnosis	Reinforcement learning	Wei et al. (2018)	Acc.: 0.65; Turns: 5.11
		Tchango et al. (2022)	Acc.: 99.2; Turns: 5.47
	Supervised	Chen et al. (2022)	Acc.: 0.77; Turns: 14.3
		Wang et al. (2023b)	Acc.: 0.69; Turns: 13.25
		Hou et al. (2023)	Acc.: 81.4; Recall: 87.6; Turns: 14.8
		Xu et al. (2023)	R-1: 0.30; B-1: 0.43; Entity-F1: 0.23
		Gatto et al. (2025)[d]	Questions: 36
		Qin et al. (2025)	R-1: 0.28; B-1: 0.44; Entity-F1: 0.25
	Knowledge-grounded	Naseem et al. (2022)	B-2: 0.15; Human eval.: 4.0
		Lin et al. (2021)[e]	BLEU avg.: 0.36; Entity-F1: 0.47
		Jia et al. (2025)	F1: 0.37
		Xu et al. (2024b)	B-4: 0.21; R-2: 0.13; Entity-F1: 0.24
	Agent-based	Dou et al. (2024)[f]	Symptoms: 0.25; Tests: 0.42; Dx: 0.55
		Chen et al. (2025)	Acc.: 0.34
		Rose et al. (2025)[g]	GTPA@1: 0.72; Avg. rank: 2.2
		Ju and Lee (2025)	Recall@1: 0.75; Recall@3: 0.97
		Dou et al. (2023)	R-L: 0.16; BERTScore: 0.61
		He et al. (2024)	B-4: 0.23; R-2: 0.22
		Sun et al. (2025)	B-4: 0.22; R-2: 0.13; Entity-F1: 0.22

Table 5: Reported performance across the reviewed clinical tasks (Part 2 of 3): automatic diagnosis. See Table 6 for footnotes and abbreviations.

Task	Training Paradigm	Paper	Reported performance
Radiology report generation	Transfer learning	Lovelace and Mortazavi (2020)	B-4: 0.15; CIDEr: 0.31; R-L: 0.32
	Supervised	Nicolson et al. (2025)	R-L: 0.26; B-4: 0.05; BERTScore: 0.25
	Retrieval-augmented	Xia et al. (2025)[h]	R-L: 0.19; BLEU avg.: 0.23; METEOR: 0.27
Disposition prediction	Supervised	Fernandes et al. (2020)	Recall: 0.82; AUROC: 0.91; AUPRC: 0.30
		Ip et al. (2024)	AUROC: 0.71; AUPRC: 0.64
	Agent-based	Sun et al. (2024)	F1: 0.32; AUC: 0.77
Discharge instruction generation	Knowledge-grounded	Eremeev et al. (2023)	BERTScore: 0.32; Concept-F1: 0.76; PPL: 6.96
	Transfer learning	Li et al. (2020)	B-4: 0.60; METEOR: 0.76

Notes. Values are in the original scale used by each paper and are not directly comparable across rows because datasets, label spaces, task formulations, and evaluation protocols differ. Where a paper reported results on multiple datasets, the table reports the average value or the primary result emphasised by the paper. [a] Reports improvement over a baseline rather than an absolute score. [b] Reports the average number of missed, incorrect, and irrelevant facts. [c] Evaluation uses GPT-based scoring. [d] Reports the average number of follow-up questions. [e] BLEU avg. denotes the average of BLEU-1/2/3/4. [f] Reports the probability of selecting the correct symptom, test, or diagnosis category. [g] GTPA@1 denotes the top-1 Ground-Truth Path Accuracy. [h] BLEU avg. denotes the average of BLEU-1/2/4. Abbreviations: Acc. = accuracy; AUC/AUROC = area under the ROC curve; AUPRC = area under the precision-recall curve; MAE = mean absolute error; R-1/R-2/R-L = ROUGE-1/2/L; B-1/B-2/B-4 = BLEU-1/2/4; CIDEr = Consensus-based Image Description Evaluation; PPL = perplexity.

Table 6: Reported performance across the reviewed clinical tasks (Part 3 of 3): radiology report generation, disposition prediction, and discharge instruction generation.