Journal of Medical Systems (2025) 49:65
https://doi.org/10.1007/s10916-025-02197-w

RESEARCH

Assessment and Integration of Large Language Models for Automated Electronic Health Record Documentation in Emergency Medical Services

Enze Bai[1] · Xiao Luo[2,3] · Zhan Zhang[1] · Kathleen Adelgais[4] · Humaira Ali[7] · Jack Finkelstein[5] · Jared Kutzin[6]

Received: 14 February 2025 / Accepted: 5 May 2025

Abstract

Automating Electronic Health Records (EHR) documentation can significantly reduce the burden on care providers, particularly in emergency care settings where rapid and accurate record-keeping is crucial. A critical aspect of this automation involves using natural language processing (NLP) techniques to convert transcribed conversations into structured EHR fields. For instance, extracting temperature values like "102.4 Fahrenheit" from the transcribed text "His temperature is 39.1, which is 102.4 Fahrenheit." However, traditional rule-based and single-model NLP approaches often struggle with domain-specific medical terminology, contextual ambiguity, and numerical extraction errors. This study investigates the potential of integrating multiple Large Language Models (LLMs) to enhance EMS documentation accuracy. We developed an LLM integration framework and evaluated four state-of-the-art LLMs—Claude 3.5, GPT-4, Gemini, and Mistral—on a dataset comprising transcribed conversations from 40 EMS training simulations. The evaluation focused on precision, recall, and F1 score across zero-shot and few-shot learning scenarios. Results showed that the integrated LLM framework outperformed individual models, achieving overall F1 scores of 0.78 (zero-shot) and 0.81 (few-shot). In addition to quantitative evaluation, a preliminary user study was conducted with domain experts to assess the perceived usefulness and challenges of the integrated framework. The findings suggest that this approach has the potential to reduce documentation effort compared to traditional manual documentation. However, challenges such as misinterpretation of medical context and occasional omissions were noted, highlighting areas for further refinement and future work. This research is the first to systematically explore and evaluate the use of LLMs for real-time EMS EHR documentation. By addressing key challenges in automated transcription and structured data extraction, our work lays a foundation for real-world implementation, improving efficiency and accuracy in emergency medical documentation.

Clinical trial number

Not applicable.

Keywords Large language models · Information extraction · Electronic health records · Emergency medical services

✉ Xiao Luo
xiao.luo@okstate.edu

✉ Zhan Zhang
zzhang@pace.edu

[1] School of Computer Science and Information Systems, Pace University, New York City, NY, USA

[2] Department of Management Science and Information Systems, Oklahoma State University, Stillwater, OK, USA

[3] School of Medicine, Indiana University, Indianapolis, IN, USA

[4] School of Medicine, University of Colorado, Aurora, CO, USA

[5] Interfaith Medical Center, New York City, NY, USA

[6] Mount Sinai Hospital, New York City, NY, USA

[7] Maimonides Medical Center, New York City, NY, USA

Published online: 17 May 2025

Introduction

Using electronic health records (EHRs) to document patient information is a key component of the current healthcare system worldwide. However, research has shown that this practice is time-consuming, leading to increased clinician burnout and reduced time spent on patient care [1, 2]. Studies have found that a typical family physician may dedicate an average of 4.5 h per day to documentation [3]. As such, there has been growing interest in automating EHR documentation to allow healthcare providers to focus more on patient care. This is particularly relevant in fast-paced emergency care settings, such as emergency medical services (EMS) or pre-hospital care [4–6]. Accurate, comprehensive, and timely EMS documentation is crucial for effective communication and the safe transfer of patient care [7]. However, in these settings, care providers often have limited time and capacity to use EHR system to chart patient record, as they are both physically and cognitively preoccupied with intense and dynamic patient care activities [6]. The consequences of inaccurate or incomplete documentation can be significant. Research has shown that failure of documenting essential scene physiology measures and treatments is linked to failure to document essential scene physiology and treatments is associated with inappropriate clinical decisions, delayed interventions, inaccurate billing, and even adverse outcomes [7–9].

Recent studies have focused on developing advanced automatic speech recognition (ASR) systems for real-time documentation during medical interviews in typical clinical office settings [10]. For instance, digital scribes have been utilized to capture physician-patient conversations and generate documentation in the EHR for clinical visits. However, there is limited research on automating documentation in the EMS environment, where noise levels are higher, and data collection presents additional challenges. Our research is motivated by the urgent need to automate EHR documentation in EMS, with a particular focus on minimizing high-risk documentation errors. By doing so, we aim to improve both clinical safety and workflow efficiency in these critical settings.

Automating EHR documentation involves several key components that work together to streamline the documentation process. Perhaps one of the most critical components is leveraging natural language processing (NLP) techniques to transcribe and process spoken or written medical information into structured data. Existing commercial solutions, such as Nuance's DAX (Dragon Ambient eXperience) [11, 12] and Suki AI assistant [13], can transcribe and summarize clinician-patient interactions in real-time with greater efficiency and accuracy, and then integrate the summary directly into the free-form clinical note section within EHR. However, these solutions primarily focus on in-hospital EHR systems and generating summaries of patient encounters, while little attention has been paid to extracting relevant information from transcribed content and mapping it to structured EHR data fields.

A key aspect of developing NLP techniques for automated EHR documentation is ensuring data accuracy and reliability, as automated systems must accurately interpret and extract diverse and complex medical information [14]. However, even advanced NLP systems can still struggle with medical terminology and contextual nuances, leading to potential errors in patient records [15]. With the advancement of large language models (LLMs), these models have been used to process clinical notes and extract clinical information for applications such as disease prediction [16], identifying patients for clinical trials [17], clinical decision support [18], and more. These studies often focus on extracting specific clinical variables in non-emergent scenarios, such as medication details [19], social determinants of health [20], or disease specific variables from clinical notes [21, 22] or guidelines [23]. Although a few research efforts have evaluated the performance of LLM-based approaches in extracting medical information from text for EHR documentation, they are limited to a few elements or specific documentation templates, such as symptom documentation [24] or sleep study reports [25]. To the best of our knowledge, LLMs have not yet been applied to transcribed conversational text for extracting clinical information to populate EHR fields in the EMS domain.

To address the research gap in the EMS EHR documentation, we are conducting a large-scale research project aimed at developing an artificial intelligence (AI) system (see Fig. 1) that can be integrated into the EMS EHR system to automate documentation. As part of this project, we conducted the current study to evaluate different LLMs and the integration of LLMs for extracting key clinical information corresponding to EMS EHR fields from transcribed conversations between EMS providers during pre-hospital care. Specifically, we developed a computational framework for automating EMS EHR documentation by integrating four different LLMs for clinical information extraction and compared them under zero-shot and few-shot settings. Our research is the first to integrate multiple LLMs to extract clinical information from transcribed conversational text for automated EMS EHR documentation. The main contributions of this paper include: (1) systematically assessed the performance and limitations of four different state-of-the-art LLMs for clinical information extraction with regards to the EHR fields in the context of EMS care, demonstrating the feasibility of utilizing LLMs for EMS information extraction from conversational text; (2) developed a novel LLM integration framework to leverage the strengths of various

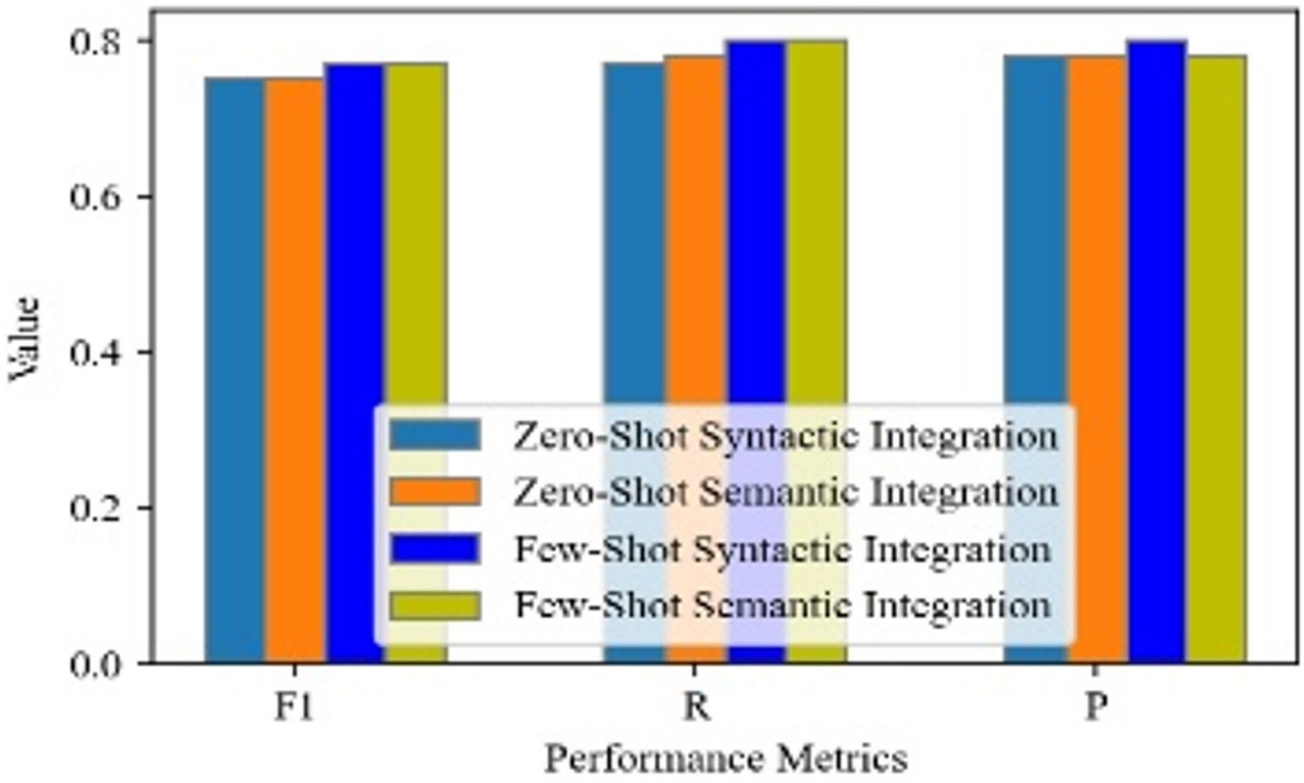

Fig. 1 Overall LLM integration framework for automated EMS EHR documentation

LLMs for automating EMS EHR documentation, highlighting the need for new computational models to capitalize on the benefits of different LLMs; (3) conducted a preliminary user evaluation study to assess the accuracy and perceived usefulness of the integrated LLM framework for extracting and mapping information to EHR fields—a key step in automating EMS EHR documentation; (4) showed the limitations in including all possible EMS EHR fields and discussed challenges and pathway for future real-world implementation.

The remainder of this paper is structured as follows: Sect. "Related Work" reviews related work on automated EHR documentation, covering methods applied in both EMS and non-EMS domains. Section "Methods" outlines the methodological framework, detailing the process from data collection to the development and evaluation of the computational model. Section "Results" presents the results, while Sect. "Discussion" discusses key findings and study implications. Finally, Sect. "Conclusions" concludes the paper.

Related Work

Automated EHR documentation has become a key area of research, driven by the need to reduce the administrative burden on healthcare professionals and improve the accuracy and efficiency of clinical documentation. Early approaches focused on leveraging NLP to process unstructured clinical text and extract useful information for structured EHR documentation. For instance, Kaufman et al. [26] used MediSapien to extract clinical information from the transcribed documents for EHR documentation and demonstrated that it can enhance the clinical documentation process and end-user experience. Advancements in AI have led to the development of more sophisticated systems, such as ASR tools, that can transcribe physician-patient conversations in real-time and convert them into structured EHR documentation. Xia et al. [27] developed an ASR algorithm to automatically convert speech with accent to text then applied name entity extraction algorithm to extract clinical information from the conversion for EHR documentation. Ahamed et al. [28] compared convolutional neural networks (CNN) with recurrent neural networks (RNN) for speech classification to identify the relevant content in conversation for automatic medical documentation. Finley et al. [29] proposed an architecture involving automatic-speech recognition,

knowledge extraction, and natural language generation. After the conversation was transcribed, a knowledge extraction that utilizes recurrent neural networks and supervised machine learning can be used to extract the structured data for medical documentation. The natural language generation module is to standardize the extracted information for final reports. Maas et al. [30] also used Google cloud speech-to-text API as an ASR tool. Then, a semi-automated approach was used to map the extracted information to ontologies concerning anatomy, symptoms, observation, diagnosis, and treatment for documentation. Woo et al. [31] processed the conversional audio with a deep neural network model based on Speech Enhancement Generative Adversarial Network (SEGAN) to reduce the noise, then used an ASR tool for transcribing. A lexicon-based language model was used to correct the transcribing errors. Finally, the medical information was extracted using MetaMap. Khattak et al. [32] developed a system to first transcribe the conversion into dialogue. Each sentence in the dialogue was then classified as questions, statements, among others. The medical entity identification process is then applied to the statements and other types to extract the medical terms for EHR documentation. Ontologies from BioPortal, Consumer Health Vocabulary, SNOMED-CT, and RxNorm were used for medical entity identification. Text classifiers that use TF-IDF and traditional classification models, e.g., random forest, were built to categorize the extracted medical entities into predefined categories for documentation. Wenceslao et al. [33] used the IBM Watson Speech to Text to transcribe the conversation into text. Then, the cTAKES was used to extract symptoms, diseases, medication, and Procedure from the text.

Despite these advancements, research on automated EHR documentation in pre-hospital settings is still limited. In addition, there is no research investigated into the recent LLMs including GPT 4o to process the transcribed text to extraction information for clinical documentation. This highlights a significant gap in the literature and suggests the need for novel frameworks and computational models tailored for the unique demands of EMS documentation. While automated EHR documentation has made significant strides, there is still much to be done, especially in the context of EMS. Table 1 summarizes the work related to automated EMS EHR documentation in the literature.

Table 1 Overview of the literature regarding automated EMS EHR Documentation

Study	EMS EHR documentation	Traditional-NLP	LLM
Kaufman et al. [26]	No	Yes	No
Xia et al. [27]	No	No	BERT
Ahamed et al. [28]	No	Yes	No
Finley et al. [29]	No	Yes	No
Maas et al. [30]	No	Yes	No
Woo et al. [31]	No	Yes	No
Khattak et al. [32]	No	Yes	No
Wenceslao et al. [33]	No	Yes	No

Methods

Study Framework

In this research, we employed a traditional study framework by first collecting conversational data using different EMS scenarios in a simulation setting. The conversational data is then transcribed into text and annotated to generate ground truth for EHR EMS documentation performance evaluation. Then, we developed a computational framework that utilizes the LLMs to automate the EHR EMS documentation using the transcribed conversational data. Finally, we evaluated the accuracy of the documentation using the traditional name entity recognition (NER) metrics and worked with domain experts on user evaluation towards real-time EHR EMS documentation. Figure 2 shows the overview of the study framework.

Data Collection

As part of a large-scale research project examining EMS workflow [34, 35], we conducted forty high-fidelity EMS simulations in a mobile simulation lab designed to replicate the back of an ambulance. The simulation scenarios varied in clinical conditions and required treatments, including a 15-month-old with seizures, a 1-month-old with hypoglycemia, and a 4-year-old with clonidine ingestion. The simulations used high-fidelity patient mannequins and replicated the dynamic EMS environment, including background noise (e.g., from a distraught parent) and overlapping speech from multiple speakers. Each simulation was conducted by four to six participants, who were full-time, licensed EMS clinicians working for a large fire-based EMS agency. The simulations were recorded in both audio and video formats. The lengths of the video recordings of the simulations varied, ranging from roughly 9 to 14 min.

The conversations among EMS providers were transcribed by professional transcriptionists and reviewed by two medical experts to ensure accuracy. The two medical experts reviewed transcribed conversation independently and annotated the necessary corrections. The inconsistency was solved by discussion to finalize the corrections. The annotators annotated all transcripts according to EMS EHR

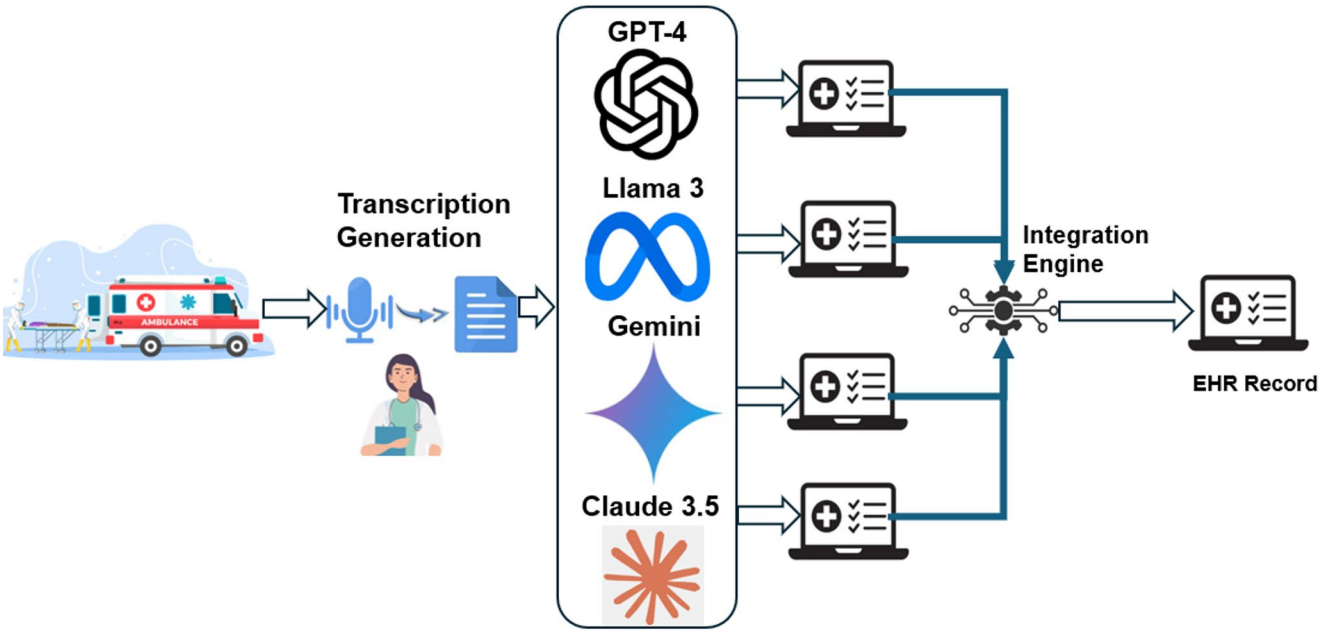

Fig. 2 Study framework

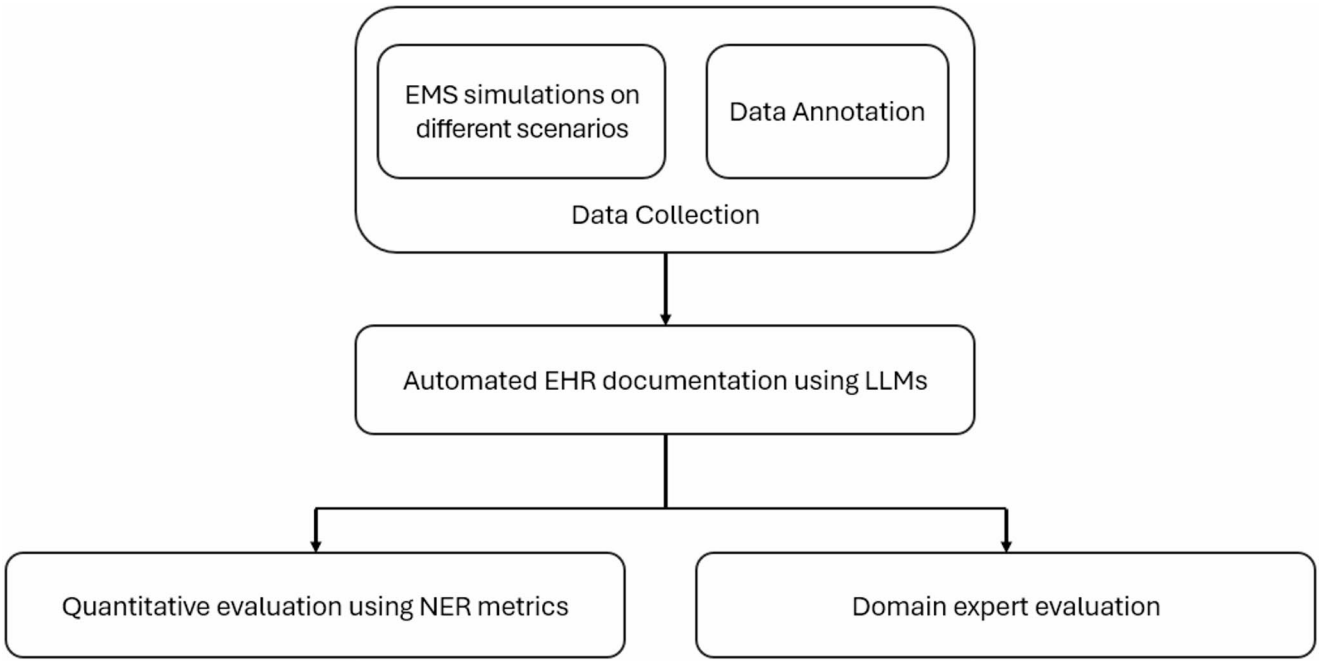

Fig. 3 Snippet of conversation text and data annotation for EHR documentation

data fields. Based on an examination of the structures of two EMS EHR systems, as well as the National Emergency Medical Services Information System (NEMSIS) data structure [36], we included all main EMS EHR fields. Figure 3 shows the snippet of the conversation text with timestamp and annotated data for corresponding EMS EHR data fields, such as Age, Complaint/Symptoms, Mental state, and AVPU.

To annotate the data, two annotators first independently annotated a small set of transcripts ($n=6$) by marking and mapping the content of the transcripts (e.g., words, phrases, etc.) to EHR data fields. Then, we calculated the inter-rater agreement between the two annotators using Cohen's Kappa coefficient on the mapping between content and EHR fields. A Kappa value of 0.80 was achieved, indicating substantial agreement between the two annotators. All

disagreements between the annotators were resolved via discussions. Following this step, the researchers created a data dictionary to standardize the annotation process. One annotator annotated the remaining transcripts using the data dictionary, while another annotator reviewed all the annotations to ensure correctness and consistency across all transcripts. The annotation process was iterative and adaptive to changes (e.g., modifying the data dictionary and revisiting previous annotations for corrections). These transcripts and annotations served as the ground truth for evaluating different LLMs and the proposed LLM integration framework.

Automated EHR Documentation Using Individual LLM

The emergence of LLMs has revolutionized the field of NLP, enabling machines to analyze human language with high accuracy. At the time of this study, based on our literature review, the leading state-of-the-art LLMs included OpenAI's GPT-4o [37], Google's Gemini [38], Meta's LLaMA 3 [39], and Anthropic's Claude 3.5 [40], all of which are designed to push the boundaries of what LLMs can achieve in terms of both performance and efficiency. In this research, we systematically assessed these four LLMs for EMS EHR documentation in two different settings: zero-shot and few-shot learning. In-context few-shot learning offers a key advantage by enabling LLMs to learn from just a few examples given during inference, eliminating the need for retraining. This is especially useful in this study due to the limited availability of labeled data. Providing a handful of examples within the prompt helps the model grasp the specific task requirements, enhancing both its accuracy and ability to adapt to new or varied scenarios. Since in-context few-shot learning does not involve model fine-tuning, it maintains the same response time as zero-shot learning when applied in real-world scenarios.

We conducted the experiments through Application Programming Interfaces (APIs) of the LLMs. These APIs allow users to provide instructions via two roles:

- System: Defines task instructions for the LLM in the desired role. We used the system variable to instruct the model to act as an annotator.
- User: Provides the input transcript text for data annotation.

The user input incorporates a transcript of a simulation recording to be analyzed by an individual LLM to extract data for each EHR field for documentation. The design of the prompt includes directives for the LLM to process the input transcript and deliver the results in a specific format. To implement few-shot learning, we created a dataset that includes two examples of each EMS EHR field to guide the LLM's responses. To standardize the output of the LLM and avoid variations [41], additional instructions, such as the removal of stop words from the output, were provided. The detailed prompt design is included in Tables S1 and S2 in the supplementary material.

Since LLMs occasionally extracted information for categories not included among the predefined EHR fields, we developed a post-processing algorithm to align these extracted categories with the predefined EHR fields. When extracting information in JSON format, the output keys (categories) were first mapped to the predefined fields using strict string matching. For categories that cannot be mapped using strict string matching, we instructed the LLM to perform soft mapping based on the semantic meaning of the categories. Any categories that cannot be mapped to the predefined fields are excluded.

LLM Integration Framework

Based on our previous research [14, 42] and the experimental results of this study, we observed that the performance of each LLM varies. To address this, we developed an LLM integration framework to leverage the strengths of various LLMs in producing the final extraction for each EMS EHR field. Figure 4 shows the algorithm that outlines the steps of the integration framework. Each LLM output (l_0, .., l_n) for a given EHR field was compared pairwise using a proposed similarity function (f), which incorporates both lexical (e.g., Levenshtein distance) and semantic similarity measures (e.g., sentence embeddings). The model output that demonstrated the highest average similarity to all other outputs was selected as the final system-generated result for evaluation, based on the assumption that higher agreement among models indicates greater reliability. If more scenarios are collected and the LLMs can be finetuned using EMS data, this strategy can be refined by incorporating confidence scores from individual LLMs, weighting similarities based on model performance history, or using ensemble voting with threshold mechanisms to filter out low-agreement outputs. Additionally, human-in-the-loop validation could be added for low-consensus cases to ensure robustness and minimize critical errors in clinical contexts.

The proposed similarity measurement considers both syntactic and semantic similarities between the output of the LLMs. The syntactic similarity calculation assumes that the similarity between two pieces of text is proportional to the number of identical words they share, whereas the semantic similarity measures how similar or different two pieces of text are in terms of their meaning and context. There are multiple ways to measure the syntactic [43, 44] and semantic similarity [45–47] between texts. We applied word-based

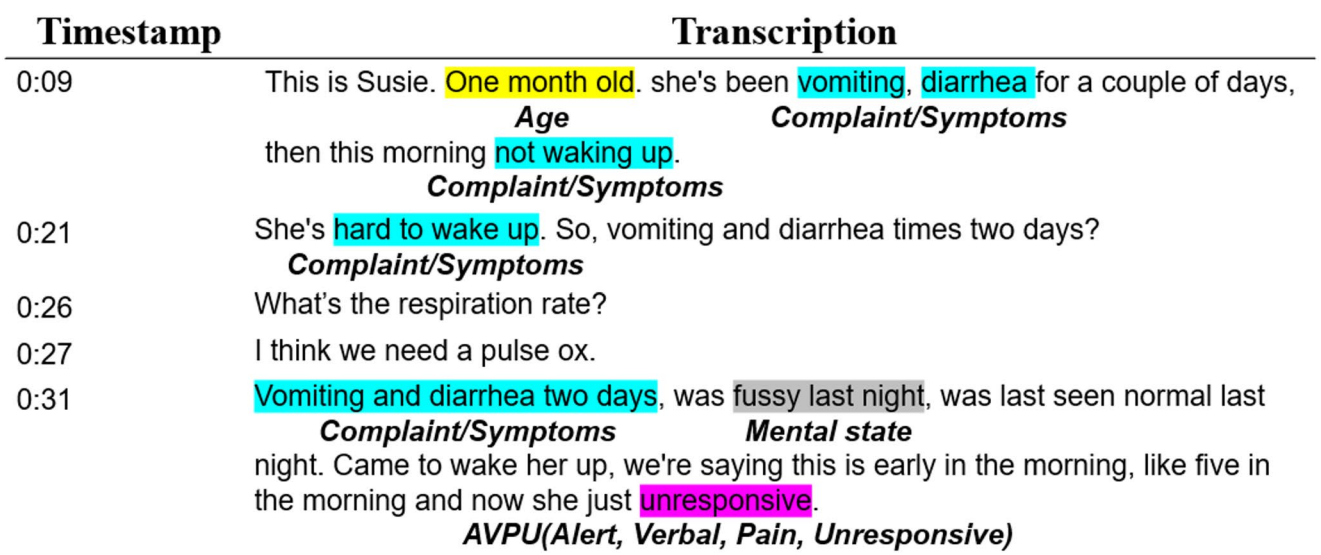

Fig. 4 Algorithm for LLM integration framework

Levenshtein distance similarity [48] to calculate the syntactic similarity between two texts. Let a and b be two texts with m and n words, respectively. The Levenshtein distance $d(m,n)$ is defined as calculation of distance between two texts a and b using Eq. 1, where $d(m,0) = m$; and $d(0,n) = n$ Then, we calculated the normalized Levenshtein distance $normd(i,j)_{(m,n)} = 1 - \frac{d(m,n)}{\max(m,n)}$.

$$d(i,j)_{(m,n)} = min \begin{cases} d(m-1,n)+1 \\ d(m,n-1)+1 \\ d(m-1,n-1)+0, \ if\ i[m] = j[n] \\ d(m-1,n-1)+1, \ if\ i[m] \neq j[n] \end{cases} \quad (1)$$

To calculate the semantic similarity, we first applied sentence transformer [49] to convert candidates and ground truth into embeddings, then, applied cosine similarity to measure the similarity $c(i,j) = \frac{a \cdot b}{||a||||b||}$. The integration of syntactic and semantic similarity was done by a linear combination of both (shown in Eq. 2), which is used to calculate the similarity (f) between the output of the LLMs shown in integration algorithm described in Fig. 3. For the LLM integration, we finetuned the integration ratio α to 0.2 to optimize the performance.

$$f(i,j) = \alpha \times normd(i,j)_{(m,n)} + (1-\alpha) \times c(i,j) \quad (2)$$

LLM Performance Evaluation

To evaluate the performance of the LLMs and the LLM integration framework in extracting clinical information for EMS EHR documentation, we utilized precision (P), recall (R), and F1 as the main metrics [50], which are commonly applied in information extraction tasks and specifically for the EHR documentation in the literature [32, 51]. The metrics evaluate whether the extracted information for the EHR fields is correct, partially correct, or spurious. The details of the calculations are given in Sect. "Related Work" of the supplemental materials.

To conduct a statistical comparison for each medical field between LLM integration and the other models, we first applied Shapiro's test and Levene's test to determine the normality and homogeneity in variety of the data, then the Mann-Whitney U-test was used to calculate the significance value for the comparison. The overall model comparison is based on the statistical comparison of the results gained from different EHR fields.

Preliminary User Evaluation of LLM Integration Framework by Domain Experts

Literature [52, 53] highlights the importance of assessing the performance of systems using LLMs by comparing them to human assessments, particularly in the medical field. To evaluate whether the proposed framework accurately identifies and extracts relevant information and maps it to the appropriate EMS EHR data fields, we invited three domain experts from different EMS agencies in the U.S. One expert has more than 40 years of experience and currently holds the position of EMS director. The second expert has more than 20 years of experience and is currently serving on both state- and city-level emergency medical advisory committees, enacting EMS documentation and care practice policies and procedures. The third expert is an EMS physician who is also involved in quality assurance work on EMS documentation. All of them have extensive, first-hand knowledge of EMS EHR documentation.

To prepare the evaluation data, we randomly selected four transcribed audio recordings and populated the EHR fields after applying our framework in the few-shot learning setting. It is important to note that the primary goal of this study is to evaluate the effectiveness of LLMs in extracting relevant information from conversational text for EHR documentation. The transcribed conversations are segmented into individual sentences and loaded into an annotation tool. Within this tool, extracted data corresponding to specific EHR fields is highlighted and labeled accordingly. For instance, in the sentence "So his temperature is 39.1 which is 102.4 Fahrenheit," the phrase "102.4 Fahrenheit" is marked and tagged as the "temperature" field. Domain experts then assessed the accuracy and relevance of these extractions based on their clinical knowledge and practical experience. These four transcribed recordings cover all EHR fields defined in Table 2. We provided the experts with instructions to evaluate the extracted information for the EHR fields based on the transcribed audio recordings (as sentences). The evaluation tasks include: (1) Evaluating the transcribed and extracted data to the corresponding EHR fields by either correcting it or confirming it as accurate for EHR documentation. If discrepancies arose between the two, the third expert's input were used to resolve them; (2) Categorizing the necessary corrections and noting the types of errors, such as failure to extract relevant information, recording in incorrect EHR fields, or other issues; (3) Completing a survey designed to gauge their perceptions of the real-world applicability of leveraging LLMs to automate the EMS EHR documentation process. The survey questions focused on gathering feedback about whether the automated documentation process could substantially reduce documentation time and burden on EMS providers, seamlessly integrate into existing workflows, and require minimal corrections, among other considerations. The survey questions are included in Sect. "Discussion" of the supplemental material.

Table 2 Statistics and examples of the annotated EMS EHR data

EHR Field Name	Example	Max # in each conversation	Average length (# of words)
Age	1 year old	5	2.47
Airway	clear	2	3.07
Allergies	never had any allergies	2	2.38
AVPU	responds to painful stimulus	7	2.87
B.G.L.	100	4	1.38
B.P.	58 over 36	12	2.61
Capillary Refill Time	6 s	4	2.13
Complaint/Symptoms	vomiting, diarrhea	22	2.88
ECG	190	11	1.73
Electrolytes	potassium	3	1.00
Gender	male	2	1.00
Lung sounds	clear	5	2.35
Medication	D10 20 milliliters	26	2.64
Mental State	Fuzzy	1	1.00
Past medical history	ADHD	15	2.77
Patient Profile	His name is Billy	6	2.90
Pulse	190	7	1.84
Pupils	dilated	4	2.54
RESP	22 to 26	11	2.81
SPO2	100%	8	1.55
Temperature	98.4	4	1.86
Trauma	no signs of trauma	5	3.89
Treatment	rescue breathing	24	2.42

Results

Summary of the Dataset

The transcribed conversation text from the forty high-fidelity EMS simulation videos was used as input for the evaluated LLMs. Each conversation text contains between 744 and 1,957 words (mean: 1,340.5, SD: 297.57). Table 2 provides the statistics and examples of the annotated data for EMS EHR documentation. Some EHR fields, such as "complaint/symptoms", "treatment", and "medication", have high occurrences in each conversation. Additionally, some fields, like "trauma" and "airway", have longer descriptions in terms of word count compared to others.

System Performance Comparison

Tables 3 and 4 present the performance comparison of the four LLMs and our LLM integration framework in zero-shot and few-shot settings, respectively. Details of the hyperparameters for the LLMs are provided in Sect. "Related Work" of the supplemental material.

The results demonstrate that the LLM-integrated model achieved significantly superior performance compared to the other models in multiple fields, including AVPU, Complaint/Symptoms, ECG, Medication, Mental State, Past Medical History, and RESP. Notably, in the Trauma field, the LLM-integrated model outperformed the alternatives only under zero-shot learning conditions, whereas in the Treatments field, its superior performance was observed exclusively under few-shot learning settings. Regarding overall performance, in the zero-shot learning setting, the

Table 3 LLM performance evaluation and comparison (Zero-Shot Learning)

EHR Field Name	GPT-4o			Gemini			LLaMA 3			Claude			LLM Integration		
	F1	*R*	*P*	F1	*R*	*P*	F1	*R*	*P*	F1	*R*	*P*	F1	*R*	*P*
Age	0.83	0.80	0.85	0.95	0.96	0.95	0.88	0.87	0.89	**0.96**	**0.95**	**0.97**	0.90	0.89	0.91
Airway	0.48	0.45	0.53	0.18**	0.17*	0.18**	0.42	0.37	0.48	0.36	0.33	0.41	**0.55**	**0.51**	**0.61**
Allergies	0.66	0.63	0.68	**0.76**	**0.8**	**0.72**	0.45*	0.32**	0.72	0.62	0.63	0.61	0.72	0.73	0.72
AVPU	0.1**	0.08**	0.13**	0.38*	0.38*	0.39**	0.45*	0.41*	0.5*	0.4*	0.35*	0.46*	**0.63**	**0.59**	**0.69**
B.G.L	0.87	0.85	0.9	0.83	0.82	0.83*	0.82	0.8	0.84	0.9	0.87	0.94	**0.91**	**0.88**	**0.94**
B.P	0.76	0.74	0.79	0.68*	0.66*	0.7*	0.72	0.67*	0.76	0.75	0.72	0.79	**0.8**	**0.77**	**0.84**
Capillary Refill Time	0.96	0.98	0.95	0.92	0.98	0.86	0.95	0.97	0.93	**0.98**	**0.98**	**0.98**	0.96	0.98	0.95
Complaint/Symptoms	0.76	0.78	0.75*	0.41**	0.42**	0.4**	0.7**	0.68**	0.72**	0.61**	0.61**	0.61**	**0.82**	**0.83**	**0.81**
ECG (Heart Rate)	0.8	0.78	0.82*	0.73*	0.72	0.74**	0.72*	0.7*	0.75**	0.83	0.83	0.82*	**0.89**	**0.87**	**0.93**
Electrolytes	0.7	0.61	0.81	0.28	0.4	0.21	0.63	0.5	0.85	**0.76**	**0.91**	**0.65**	0.75	0.91	0.64
Gender	**1.0**	**1.0**	**1.0**	0.88*	0.88*	0.88*	**1.0**	**1.0**	**1.0**	**1.0**	**1.0**	**1.0**	**1.0**	**1.0**	**1.0**
Lung Sounds	0.72	0.66	0.8	0.57*	0.52	0.62*	0.72	0.64	0.82	0.66	0.58	0.77	**0.73**	**0.65**	**0.83**
Medication	0.33**	0.31**	0.35**	0.05**	0.05**	0.06**	0.48*	0.42*	0.55	0.36**	0.33**	0.4**	**0.59**	**0.55**	**0.64**
Mental State	0.19**	0.24**	0.16**	0.27*	0.27*	0.26*	0.39	0.4*	0.39	0.51	0.58	0.45	**0.54**	**0.62**	**0.49**
Past Medical History	0.35**	0.32**	0.39**	0.27**	0.25**	0.28**	0.42**	0.35**	0.51*	0.45*	0.43*	0.47**	**0.58**	**0.53**	**0.64**
Patient Profile	0.68	0.68	0.67	0.49**	0.49**	0.48**	0.68*	0.66*	0.7*	0.68*	0.68	0.67*	**0.83**	**0.84**	**0.82**
Pulse	0.59	0.59	0.59	0.61	0.63	0.6*	0.6	0.58	0.63	0.66	0.66	0.66	**0.75**	**0.74**	**0.75**
Pupils	**0.75**	**0.8**	**0.71**	0.65	0.75	0.58	0.71	0.72	0.7	0.68	0.68	0.67	**0.75**	**0.80**	**0.71**
RESP	0.55**	0.51**	0.6*	0.67	0.73	0.62**	0.69	0.63*	0.76	0.59*	0.56*	0.62*	**0.78**	**0.74**	**0.81**
SPO2	0.66	0.63	0.69	0.55*	0.54	0.57*	0.57*	0.53	0.61*	0.68	0.66	0.71	**0.73**	**0.69**	**0.78**
Temperature	0.84	0.85	0.84	0.85	0.86	0.84	0.88*	0.91*	0.84	**0.89**	**0.92**	**0.85**	**0.89**	**0.92**	**0.85**
Trauma	0.77**	0.82*	0.73**	0.86**	0.9**	0.82**	0.63**	0.55**	0.73**	0.86*	0.88	0.84**	**0.87**	**0.91**	**0.84**
Treatments	0.76	0.81	0.71	0.54	0.54	0.53	0.64	0.63	0.64	0.78	0.87	0.71	**0.83**	**0.89**	**0.79**
Macro-Median	0.72	0.68	0.71	0.61**	0.63*	0.60**	0.68**	0.63**	0.72	0.68	0.68	0.67	**0.78**	**0.80**	**0.81**

Note: Bold numbers indicate the highest performance for each corresponding EHR field; ** indicates that LLM Integration is significantly overperforming that model for a $p < 0.01$; * Indicates that LLM Integration is significantly overperforming that model for a $p < 0.05$

LLM-integrated model demonstrated significantly higher recall compared to both Gemini and Llama 3, higher precision compared to Gemini, and superior F1 scores relative to both Gemini and Llama 3. Under the few-shot learning setting, the LLM-integrated model exhibited even greater improvements, significantly outperforming GPT-4o, Gemini, and Llama 3 in both F1 score and precision. Additionally, the results indicate that Claude did not show any significant underperformance when compared to the LLM-integrated model. These results demonstrate that the proposed LLM integration framework effectively leveraged the strengths of different LLMs, resulting in the highest overall performance in both zero-shot and few-shot settings. This approach selects the most semantically aligned and syntactically consistent answers across all LLMs. The integration of LLMs proved robust and adaptable across various tasks, generally surpassing the performance of individual LLMs in most categories. The results show that LLM integration did not outperform in certain areas, such as age, allergies, regardless of whether in zero-shot or few-shot settings. For these fields, our analysis indicates limited consensus among the LLMs, with typically only one or two LLMs significantly outperforming the others.

Comparing the individual LLMs based on their macro-median values, Claude 3.5 achieved a competitive or better a macro-median F1 scores in both zero-shot and few-shot scenarios. Among all LLMs, Gemini performed not as well as others. Comparing the results of zero-shot and few-shot settings, the macro-median shows that the performance of the LLM integration improved overall when few-shot learning is applied, except for a drop in the F1 for past medical history. This might be due to prompt overfitting, where the model's general understanding of fields with multifaceted expressions becomes restricted to the specific examples provided in few-shot learning. [54] Gemini, Llama 3, and Claude 3.5 all experienced performance fluctuations in few-shot learning compared to zero-shot, while GPT-4o generally showed improvement. This could be due to GPT-4o's strong summarization capabilities and balanced generalization ability. Other models may struggle with understanding longer texts and appropriately distributing attention across the prompts.

Table 5 shows the scenario-based performance of applying the LLM integration model using the few-shot learning. We performed both ANOVA and Kruskal-Wallis tests to compare the performance of the LLM-integrated model in data extraction across different medical fields. The results indicate no statistically significant differences in performance among these scenarios.

To demonstrate the distinct impacts of syntactic and semantic similarity on LLM integration, we conducted an ablation study by adjusting the α value to either 1 or 0 to compare the syntactic and semantic integration. Our findings indicate that they yield comparable results, with only minor variations in recall (0.77 vs. 0.78) for zero-shot and precision (0.78 vs. 0.80) for few-shot learning. This shows that the top outputs identified from the LLMs are nearly identical when evaluated using either syntactic or semantic similarity metrics. Figure 5 shows the ablation study result.

LLM Efficiency Comparison

The efficiency of the proposed method for processing the transcribed text towards automated EMS EHR documentation is crucial, given the fast-paced nature of the EMS work environment. Figure 6 compares the average time in seconds required to process a transcribed conversation using different LLMs and the LLM integration approach under zero-shot and few-shot settings, respectively. Google Gemini stands out as the most efficient model, with average processing times of 0.83 s for zero-shot learning and 0.80 s for few-shot learning. GPT-4o and Claude 3.5 exhibit moderate processing times, with GPT-4o averaging 0.74 s in zero-shot and 1.69 s in few-shot, while Claude 3.5 shows times of 1.41 s and 1.48 s, respectively. In contrast, due to the local installation of LLaMA 3, it is the least.

efficient, with significantly longer processing times of 6.15 s in zero-shot and 16.18 s in few-shot scenarios. The LLM integration framework, which runs all models in parallel, shows similar times to LLaMA 3.

Results of Preliminary User Evaluations

Table 6 outlines the necessary corrections to the extracted information and identifies instances of failure to extract relevant information. Most of the demographics and vital signs (B.G.L., B.P., RESP, SpO2, temperature) were correctly documented. However, the extracted information for fields such as airway, ECG, pulse, trauma, medication, and treatment require validation before achieving a level of accuracy sufficient for safe documentation. Some corrections were required where certain extracted information was assigned to incorrect EHR fields. For example, the phrase "good chest rise" was initially categorized by the LLM under the airway field, but domain experts recommended it be assigned under lung sounds. Information related to medication and treatment was the most frequently missed, with 7 and 6 omissions, respectively. Some of these errors and omissions arose from incorrect transcription caused by unclear audio recordings (e.g., Medication D10W being transcribed as V10W), leading to incomplete or incorrect contextual understanding. For medication, treatment, and airway fields, half or one-third of the needed corrections were due to transcription errors. Therefore, one domain

Table 4 LLM performance evaluation and comparison (Few-Shot Learning)

EHR Field Name	GPT-4o			Gemini			LLaMA 3			Claude			LLM Integration		
	F1	*R*	*P*	F1	*R*	*P*	F1	*R*	*P*	F1	*R*	*P*	F1	*R*	*P*
Age	0.90	0.87	0.92	0.89	0.89	0.89	0.87	0.85	0.89	**0.96**	**0.94**	**0.98**	0.93	0.91	0.95
Airway	0.54	0.51	0.57	0.25**	0.26*	0.24**	0.35*	0.32*	0.38*	0.45	0.42	0.49	**0.62**	**0.59**	**0.64**
Allergies	0.63	0.62	0.64	**0.85**	**0.88**	**0.81**	0.62	0.53	0.73	0.62	0.63	0.61	0.75	0.77	0.73
AVPU	0.22**	0.23**	0.22**	0.6	0.61	0.59	0.46**	0.47**	0.45**	0.31**	0.28**	0.34**	**0.73**	**0.73**	**0.73**
B.G.L	0.88	0.86	0.91	0.82*	0.79*	0.84*	0.85	0.83	0.87	0.9	0.87	0.93	**0.91**	**0.88**	**0.95**
B.P	0.69*	0.62**	0.77	0.68*	0.67*	0.7*	0.7*	0.66*	0.75*	0.78	0.73	0.82	**0.81**	**0.77**	**0.84**
Capillary Refill Time	0.96	0.98	0.93	0.85	0.89	0.82	**0.98**	**0.99**	**0.97**	**0.98**	**0.98**	**0.98**	0.96	0.98	0.93
Complaint/symptoms	0.74*	0.77*	0.72**	0.56**	0.57**	0.55**	0.71**	0.73**	0.69**	0.62**	0.63**	0.61**	**0.84**	**0.86**	**0.81**
ECG(Heart Rate)	0.8*	0.82	0.79*	0.76*	0.78	0.74**	0.76	0.76	0.77*	0.73*	0.72*	0.74*	**0.88**	**0.87**	**0.89**
Electrolytes	0.73	0.95	0.59	0.69	0.95	0.54	0.46	0.81	0.32	**0.76**	**0.95**	**0.63**	0.73	0.95	0.59
Gender	**1.0**	**1.0**	**1.0**	0.83*	0.83*	0.83*	**1.0**	**1.0**	**1.0**	**1.0**	**1.0**	**1.0**	**1.0**	**1.0**	**1.0**
Lung Sounds	0.71	0.64	0.81	0.64	0.63	0.66	0.6	0.52	0.71	0.6	0.52	0.71	**0.73**	**0.64**	**0.84**
Medication	0.3**	0.28**	0.32**	0.01**	0.01**	0.01**	0.47	0.43	0.51	0.26**	0.24**	0.3**	**0.54**	**0.51**	**0.58**
Mental State	0.46*	0.53*	0.41**	0.58	0.61*	0.56	0.56*	0.58**	0.55	0.75	0.83	0.67	**0.80**	**0.88**	**0.74**
Past Medical History	0.31*	0.29*	0.34**	0.21**	0.19**	0.24**	0.35	0.3	0.43	0.32*	0.29*	0.36*	**0.44**	**0.39**	**0.51**
Patient Profile	0.68	0.69	0.68	0.2**	0.22**	0.19**	0.62**	0.61**	0.62*	0.72	0.74	0.71	**0.84**	**0.85**	**0.83**
Pulse	0.64	0.63	0.66	0.6	0.6	0.6*	0.62	0.6	0.63*	0.68	0.65	0.71	**0.78**	**0.75**	**0.8**
Pupils	**0.75**	**0.8**	**0.7**	0.62	0.63	0.61	0.68	0.69	0.67	0.67	0.67	0.66	0.72	0.74	0.7
RESP	0.7*	0.66*	0.73*	0.6**	0.62*	0.58**	0.67**	0.65*	0.7**	0.69*	0.68	0.7*	**0.84**	**0.82**	**0.86**
SPO2	0.68	0.68	0.68	0.48*	0.46*	0.5**	0.57*	0.54*	0.6*	0.7	0.65	0.75	**0.74**	**0.71**	**0.78**
Temperature	0.79	0.78	0.81	**0.89**	**0.89**	**0.89**	0.85	0.88	0.82	0.88	0.91	0.84	0.87	0.89	0.86
Trauma	**0.89**	**0.93**	**0.85**	**0.89**	**0.94**	**0.84**	0.83	0.83	0.83	0.86	0.87	0.86	**0.89**	**0.92**	**0.87**
Treatments	0.74**	0.77**	0.71**	0.36**	0.38**	0.34**	0.61**	0.68**	0.55**	0.82	0.85	0.8	**0.85**	**0.88**	**0.82**
Macro-Median	0.71*	0.69	0.71*	0.62**	0.63*	0.6**	0.72**	0.66*	0.69*	0.72	0.72	0.71	**0.81**	**0.85**	**0.82**

Note: Bold numbers indicate the highest performance for each corresponding EHR field; ** indicates that LLM Integration is significantly overperforming that model for a $p < 0.01$; * Indicates that LLM Integration is significantly overperforming that model for a $p < 0.05$

Table 5 Performance comparison on different medical scenarios

EHR Field Name	15-month-old with seizures			1-month-old with hypoglycemia			4-year-old with clonidine ingestion		
	F1	*R*	*P*	F1	*R*	*P*	F1	*R*	*P*
Age	0.89	0.87	0.93	0.94	0.91	0.98	0.94	0.94	0.94
Airway	0.60	0.67	0.55	0.69	0.66	0.75	0.56	0.55	0.62
Allergies	0.75	0.75	0.75	0.75	0.75	0.75	0.70	0.75	0.72
AVPU	0.57	0.55	0.58	0.56	0.53	0.60	0.89	0.93	0.88
B.G.L	0.94	0.93	1.00	0.84	0.84	0.86	0.89	0.87	0.97
B.P	0.84	0.81	0.91	0.74	0.73	0.78	0.78	0.77	0.83
Capillary Refill Time	1.00	1.00	1.00	1.00	1.00	1.00	0.88	0.92	0.86
Complaint/symptoms	0.83	0.86	0.85	0.87	0.90	0.88	0.77	0.82	0.74
ECG (Heart Rate)	0.85	0.85	0.90	0.88	0.89	0.91	0.84	0.85	0.89
Electrolytes	/	/	/	0.63	0.95	0.59	/	/	/
Gender	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Lung Sounds	0.82	0.74	0.99	0.57	0.52	0.71	0.76	0.71	0.86
Medication	0.72	0.68	0.83	0.47	0.49	0.54	0.34	0.35	0.36
Mental State	0.73	0.85	0.75	0.79	0.88	0.77	0.73	1.00	0.62
Past Medical History	0.48	0.46	0.51	0.37	0.38	0.39	0.39	0.35	0.59
Patient Profile	0.83	0.85	0.83	0.77	0.81	0.77	0.88	0.88	0.89
Pulse	0.88	0.85	0.96	0.73	0.73	0.75	0.72	0.71	0.75
Pupils	/	/	/	/	/	/	0.70	0.74	0.70
RESP	0.80	0.76	0.90	0.80	0.80	0.82	0.86	0.89	0.88
SPO2	0.78	0.77	0.90	0.68	0.68	0.76	0.65	0.68	0.71
Temperature	0.97	1.00	0.96	0.57	0.75	0.51	0.79	0.74	0.90
Trauma	0.97	0.94	1.00	0.90	1.00	0.85	0.84	0.91	0.83
Treatments	0.84	0.88	0.86	0.80	0.89	0.80	0.78	0.85	0.79

expert pointed out that the correctness of the transcription is critical for the following information extraction and EHR documentation. This indicates that state-of-the-art LLMs lacking deep EMS domain knowledge cannot match or surpass the performance of human experts. Therefore, training specialized LLMs for EMS documentation is necessary.

The survey responses show that all three domain experts agreed or strongly agreed that the automated documentation process can significantly reduce documentation time and ease the burden on EMS providers. They also viewed the proposed LLM integration framework as a good starting point, though some corrections may still be necessary. The experts also agreed or strongly agreed that the EHR automated documentation would integrate well into existing EMS workflow, if the need for corrections was minimal. They emphasized that any information generated automatically by the tool should be validated before finalizing the documentation. The experts identified several fields as critical to remain error-free, including medication, past medical history, demographics (age, gender, allergies), vitals (B.G.L., B.P., pulse, RESP, SpO2), and treatment. One expert further noted that the current framework focused only on the main EHR fields and recommended adding fields such as appearance, EtCO2, and medication response. The expert suggested that some existing extractions might be better categorized under these new fields. For example, "cyanotic" should have been placed in a distinct field labeled as appearance. The survey findings also suggest that if this approach is integrated into a full end-to-end system for automated EMS EHR documentation, human oversight should be included to minimize errors and ensure compliance with regulations. This oversight could take place during a staging phase before the final submission into the EHR system, where both the original transcribed conversation and the generated EHR entries are presented for expert review. During this phase, reviewers can validate extracted information, correct inaccuracies, and add any missing data. Additionally, providing access to the original audio recordings is essential, allowing EMS providers to verify the transcription when needed.

Discussion

Opportunities in the USE of LLM for Automating EHR Documentation in Emergency Care Settings

Current manual and semi-automated methods for EMS documentation face several well-documented limitations that hinder clinical efficiency, data quality, and provider satisfaction. Manual documentation, which remains the dominant method in many EMS systems, requires providers to

Algorithm 1: LLM Integration towards EHR Documentation

Data: LLM output $(l_0, \ldots, l_n)$ of each EMS EHR field.
$n \in \{GPT-4, \text{Gemini}, \text{LLaMA 3}, \text{Claude 3}\}$

Result: Final LLM output (l_b)

```
l_b ← null;
i ← 0;
max_sim_score ← −1;
while i ≤ n do
    total_sim_score ← 0;
    for j ∈ {0, ..., n − 1} do
        if i ≠ j then
            total_sim_score ← total_sim_score + f(l_i, l_j);
        end
    end
    if total_sim_score > max_sim_score then
        max_sim_score ← total_sim_score;
        l_b ← l_i;
    end
    i ← i + 1;
end
```

Fig. 5 Ablation study on syntactic and semantic LLM integration

recall and transcribe critical information after the clinical encounter, often under time pressure or cognitive fatigue [6, 55]. This retrospective entry process introduces a high risk of omission [8]. Studies have shown that key clinical fields such as medication administration, patient history, and vital signs are often inconsistently captured, leading to potential gaps in continuity of care and patient safety risks during hospital handoffs [8, 56]. Semi-automated approaches—such as structured electronic templates or basic voice dictation tools—attempt to alleviate some of this burden but remain limited in scope and flexibility [14, 57]. These systems often fail to capture the full richness of EMS narratives or adapt to the variability in patient encounters. Collectively, these limitations underscore the need for more intelligent, adaptive, and context-aware solutions, such as LLM-powered documentation systems, which can better support timely, accurate, and complete EHR generation in emergency medical settings.

This study demonstrates performance improvements through the integration of pre-trained LLMs. In real-world applications, LLM integration can leverage the strengths of different fine-tuned LLMs, such as Me-LLaMA [58]. Given that LLMs can be fine-tuned for various specific domains or objectives, we believe integrating multiple LLMs can optimize performance across groups of EHR fields to achieve overall effectiveness. While this research proposes and tests one method of LLM integration, there may be other advanced approaches worth exploring based on the specific needs of the application. Fine-tuning the LLMs with medical terminology frequently used in EMS could further reduce errors and enhance the accuracy of documenting the critical fields such as medication, past medical history, vitals, and treatment. Deploying LLM models locally can improve

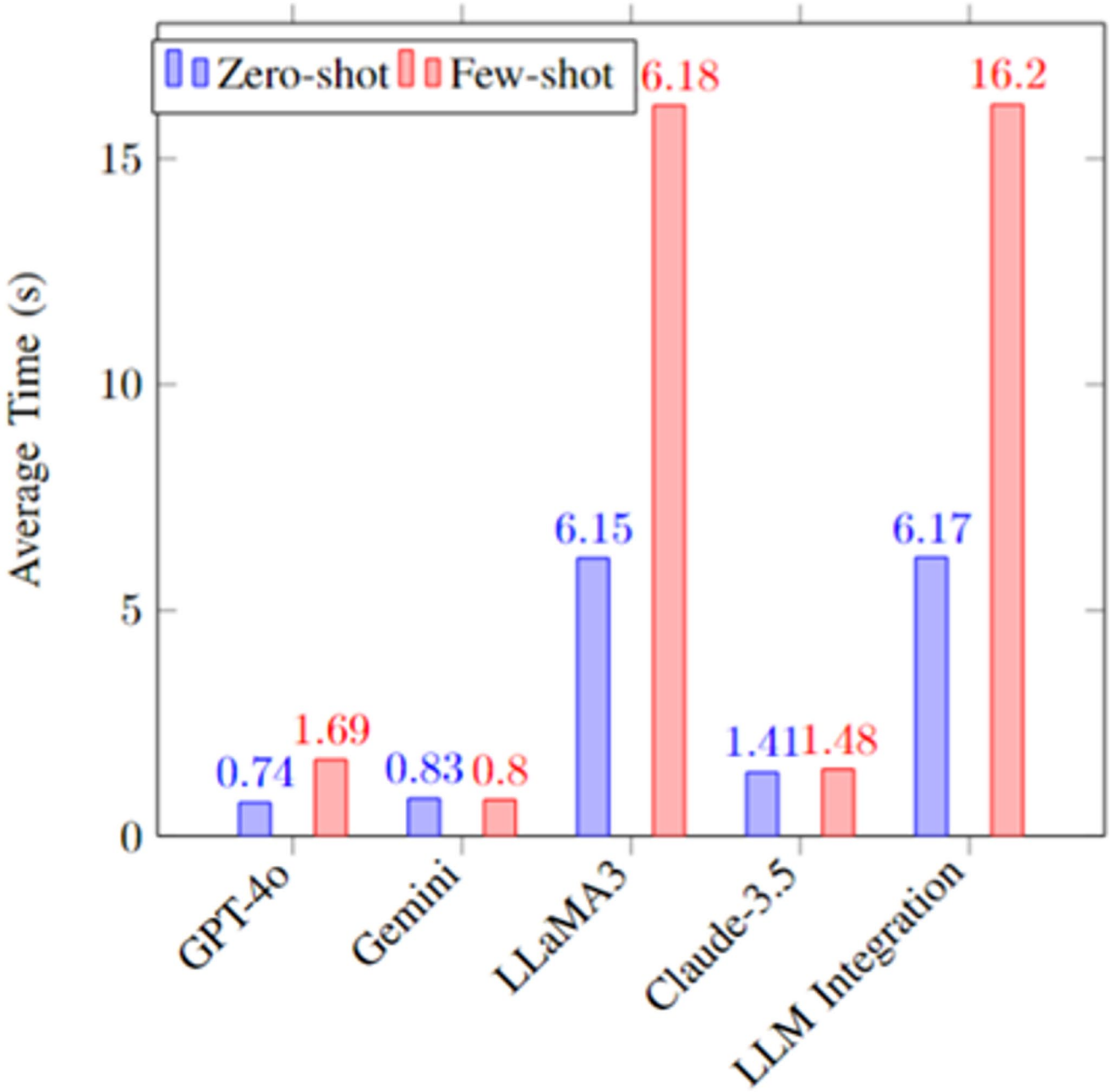

Fig. 6 Efficiency comparison of the LLMs

processing speed, address token limitations, and maintain data privacy—critical considerations in the clinical domain.

Automating EMS EHR documentation not only alleviates the cognitive burden on EMS providers but also enhances the completeness of records, which is critically needed for pediatric EMS cases [59] and in urban fire-based non-transporting EMS agencies [60]. Improved documentation completeness is essential, as it can significantly impact patient outcomes. By providing more accurate and comprehensive EHR fields, the framework can improve continuity of care, as subsequent healthcare providers have access to precise and up-to-date patient information. The future integration of this LLM-based framework into EMS practice represents a significant advancement in the use of AI for prehospital care, potentially setting a new standard for real-time documentation in EMS. In addition, mapping the extracted data to Fast Healthcare Interoperability Resources (FHIR) standard could also improve the generalizability and interoperability of the proposed approach. [61, 62] Our preliminary user evaluation study confirmed that domain experts recognized the great potential of leveraging advanced LLMs to facilitate and potentially automate EMS EHR documentation.

Table 6 Number of corrections needed for Documentation

EHR Field Name	# of occurrence	% of Correction	% of Correction due to Transcription Error	# of Missing
Age	11	2/11(18.2%)	0	1
Airway	4	2/4 (50.0%)	1/2 (50%)	0
Allergies	1	0.0%	0	0
AVPU	5	1/5 (20.0%)	0	0
B.G.L.	7	0.0%	0	1
B.P.	6	0.0%	0	2
Capillary Refill Time	4	0.0%	0	0
Complaint/symptoms	37	6/37(16.2%)	1/6 (16.7%)	3
ECG (Heart Rate)	8	3/8 (37.5%)	0	1
Electrolytes	2	0.0%	0	0
Gender	2	0.0%	0	0
Lung Sounds	1	0.0%	0	1
Medication	26	8/26 (30.8%)	4/8 (50%)	7
Mental State	3	0.0%	0	2
Past Medical History	6	1/6 (16.7%)	0	1
Patient Profile	9	0.0%	0	1
Pulse	2	1/2 (50.0%)	0	1
Pupils	2	0.0%	0	0
RESP	11	0.0%	0	2
SPO2	5	0.0%	0	0
Temperature	6	0.0%	0	0
Trauma	2	1/2 (50.0%)	0	0
Treatments	56	12/56 (21.4%)	4/12 (33%)	6

Challenges and Considerations in the Use of LLM for Automating EHR Documentation in Emergency Care Settings

To fully harness the potential of LLM for automating EHR documentation in real-world emergency care settings, several challenges should be acknowledged and addressed. For example, in our assessment, we found that LLMs can introduce output inconsistency issues by creating additional fields or categories. This could be due to the different training data used to pretrain the LLMs. While OpenAI, Google, and Anthropic have not published much detailed training data information, we hypothesize that GPT-4o and Claude 3.5 have a larger size of training data, which includes common medical domain knowledge. In addition, the LLM seems to suffer from knowledge conflict issue [63]. The LLM models that are pre-trained with a large data might refer to their pre-trained knowledge rather than respecting the information in the prompts. Like other use cases of the LLMs, they struggle to prioritize tasks in long context [64] and face in-context learning failure which is trigger by limitation of the sequence length [65]. On the other hand, LLMs not specifically tailored for the EMS context face challenges in interpreting ambiguous medical abbreviations and are less effective when relevant information is absent from the conversation. These limitations challenge LLM integration, although we believe that LLM integration can be optimized further by considering these limitations. Potential solutions include deploying EMS-specific LLMs trained or fine-tuned on EMS data, developing a human-in-the-loop AI to support continuous performance improvement, and allowing providers to make corrections during the documentation finalization process.

Another key concern is the accuracy of current LLM techniques in transcribing speech and extracting relevant clinical information from fragmented conversations [57]. Establishing clear regulations and guidelines to determine accountability in cases where LLMs inaccurately transcribe, or extract information is crucial [66]. As one of the domain experts pointed out, the accuracy of transcription by AI tools is critical for subsequent information extraction and the automation of EHR documentation. It is important to recognize that accurate clinical information extraction for various EMS EHR fields relies on the high performance of speech recognition and transcription in the EMS domain. Although our transcription was first done by an AI-based transcribing tool then validated by human, some transcription errors persisted, triggering subsequent errors in clinical information extraction. Minimizing transcription errors is therefore essential for automating EMS EHR documentation. Finally, as highlighted by our domain experts in the preliminary evaluation, validating the accuracy of automatically processed and completed information in EHRs is necessary, as EMS providers are liable for the correctness of

patient records (for billing or quality assurance purposes). Therefore, it is essential to emphasize to care providers that human validation remains a critical step in the process of automated clinical documentation. This human oversight helps prevent errors or biases that may arise from the system or the LLMs.

Finally, the use of LLMs for automating documentation introduces critical ethical and legal challenges [14, 66]. Given the highly sensitive nature of EMS medical records, strict adherence to privacy regulations—such as the Health Insurance Portability and Accountability Act (HIPAA) in the United States—is essential. Any future deployment must ensure robust safeguards are in place, including encrypted data transmission and secure storage infrastructures. Accountability and liability also become more complex in the context of AI-generated documentation [67]. Mistakes such as inaccurate or incomplete entries could result in misinformed clinical decisions or even legal disputes. To address this, transparent governance structures must be established, clearly defining the responsibilities of EMS providers, healthcare organizations, and technology vendors. All AI-generated content should be flagged, auditable, and easily verifiable by clinicians to ensure safe and accurate documentation.

Pathway for Real-world Deployment and Implementation of LLM in Emergency Care Settings

Implementing LLMs to automate EMS documentation in the real world is a complex task. It requires addressing not only technical challenges—such as performance and accuracy—but also social and workflow considerations. These include potential resistance to adopting new technologies among providers and the need for alignment with fast-paced clinical workflows [31,57, 66]. To address these implementation challenges, it might be useful to adopt a phased and strategic implementation strategy [11]. The initial phase should involve a limited pilot deployment with a small group of EMS providers, allowing for controlled evaluation and iterative refinement. This pilot phase will enable thorough testing of the system's usability, user acceptance, technical performance, and integration with existing documentation practices [52]. Based on these insights, a broader rollout can be pursued, supported by targeted training sessions, human-in-the-loop validation, and continuous system monitoring [10]. This incremental approach allows for early issue detection, builds provider trust, and ensures safe and effective adoption of LLM technology in EMS documentation workflows.

To maximize the impact of such systems, it is essential to engage key stakeholders from the outset. EMS providers and agencies stand to benefit from reduced documentation burden, improved workflow efficiency, and more accurate records that enhance patient handoffs. Healthcare organizations and hospitals can gain from better continuity of care and reduced risk associated with incomplete or delayed EMS documentation. Technology developers and EHR vendors can apply these findings to design and refine intelligent documentation tools tailored to high-pressure, real-time clinical environments. Researchers and educators may also leverage this work as a foundation for future studies on LLM applications in clinical documentation. To ensure meaningful adoption and future innovation, collaboration among these main stakeholders will be essential.

Lessons Learned and Limitations

Our evaluation revealed several key insights into the performance and limitations of using LLMs for EMS EHR documentation. First, LLM performance varied notably across different EHR fields. Structured data fields such as vital signs were extracted reliably, while narrative-based fields like chief complaint and treatment plan were more susceptible to omissions or inconsistencies. This highlights a need for better handling of complex, context-dependent language within medical narratives.

Another critical lesson was the importance of transcription accuracy. Even minor transcription errors introduced during the upstream process often led to substantial downstream inaccuracies in the generated EHR content. Although the transcriptions used in this study were reviewed by experts, we did not formally quantify transcription accuracy, and small errors may have affected the evaluation results.

In the EMS domain, treatment and medication decisions are typically guided by the patient's condition and state-specific EMS protocols, which vary depending on local policies. However, there is a lack of publicly available EMS data suitable for fine-tuning LLMs for this specialized domain. As a result, even the most advanced LLMs often lack the deep domain knowledge needed to accurately identify certain details, particularly specific treatments or medications—leading to lower performance in these areas.

In practical settings, clinical transcripts frequently include multiple references or questions about specific details, such as a patient's previous medical history, choices concerning medication dosages/routes, and treatment strategies. As the conversation evolves, these mentions might slightly differ. Typically, for EHR documentation, only the final, agreed-upon information is recorded. Our existing approach does not automatically amalgamate these occurrences, which could result in redundant or conflicting records in the EHR unless meticulously handled. This also suggests that future work should focus on developing more sophisticated

methods to consolidate and record only the final, clinically agreed-upon information in the EHR.

Additionally, our current study focused on mapping extracted data to a core set of EMS EHR fields. Fields such as patient appearance—identified as clinically important by domain experts during our user evaluation—were not included but should be considered in future research to expand the coverage and clinical utility of the system.

While our experiments centered on the EMS domain, the proposed computational framework is adaptable to other healthcare contexts, such as office-based clinical visits. However, different specialties often require unique documentation structures and field definitions. As such, information extraction prompts and model configurations need to be adapted accordingly. Fine-tuning the framework with domain-specific LLMs could further enhance performance in these alternative medical settings.

Conclusions

This study provides a comprehensive evaluation of the potential for using LLMs to automate EHR documentation in emergency care settings, such as EMS. By assessing the performance of four advanced LLMs—GPT-4o, Gemini, LLaMA 3, and Claude 3.5—under both zero-shot and few-shot learning paradigms, we identified the strengths and limitations of each model in accurately extracting clinical information from transcribed EMS conversations. We also introduced a novel LLM integration framework that combines the outputs of multiple models to enhance the overall accuracy of data extraction across various EMS EHR fields. Our findings demonstrate that this integrated approach generally outperforms individual models, offering a more robust solution for supporting clinical documentation in real-time, high-stakes environments like EMS. By advancing the development and integration of LLMs in EMS settings, this research paves the way for leveraging advanced AI and LLM techniques to achieve more efficient, accurate, and timely documentation processes, ultimately enhancing the quality of emergency care. The real-world implementation of this automated EMS EHR documentation system will be pursued in future research, where it will be compared against traditional expert-generated documentation.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s10916-025-02197-w.

Author Contributions XL, KA, and ZZ conceptualized and designed the study and secured the funding. KA and ZZ were responsible for simulation data collection. Transcribed data annotation was carried out by EB, XL, and ZZ, whereas EB handled data cleaning. XL and EB were responsible for model implementation and analysis. JF, JK, and HA were experts involved in the user evaluation study. XL, EB, and ZZ ensured the integrity and accuracy of the data analysis. All authors participated in the interpretation and discussion of the findings. XL, EB, and ZZ completed the initial draft, and subsequent versions were reviewed and approved by all authors.

Funding This work was supported in part by funding from the National Science Foundation (Award# 2237097) and National Institute of Health (Award# 1R15LM014556-01).

Data Availability The transcribed simulation data is available based on request, the other data underlying this article are available in the article and in its online supplementary material.

Declarations

Human Ethics and Consent to Participate This study has been approved by the Pace University Institutional Review Board (IRB# 1515261-2). All participants in the simulations agreed that the data could be used for research purposes, provided their identities remain anonymous.

Competing Interests The authors declare no competing interests.

References

1. Pelland KD, Baier RR, Gardner RL. 'It is like texting at the dinner table': a qualitative analysis of the impact of electronic health records on patient–physician interaction in hospitals. BMJ Health & Care Informatics. 2017;24(2)
2. Arndt BG, Beasley JW, Watkinson MD, et al. Tethered to the EHR: primary care physician workload assessment using EHR event log data and time-motion observations. The Annals of Family Medicine. 2017;15(5):419–426.
3. Hudelson C, Gunderson MA, Pestka D, et al. Selection and implementation of virtual scribe solutions to reduce documentation burden: a mixed methods pilot. *AMIA Summits on Translational Science Proceedings*. 2024;2024:230.
4. Pilerot O, Maurin Söderholm H. A conceptual framework for investigating documentary practices in prehospital emergency care. *Proceedings of Tenth International Conference on Conceptions of Library and Information Science*. 2019;4(24)
5. Sarcevic A, Ferraro N. On the use of electronic documentation systems in fast-paced, time-critical medical settings. Interacting with Computers. 2017;29(2):203–219.
6. Zhang Z, Joy K, Harris R, Park SY. Characteristics and challenges of clinical documentation in self-organized fast-paced medical work. *Proceedings of the ACM on Human-Computer Interaction*. 2022;6(CSCW2):1–21.
7. Alatis AS, Monahan BV, Raymond AD, Hudson KB, Vieth JT, Nable JV. Checklists improve EMS documentation: quality improvement in a collegiate-based EMS agency. J Coll Emerg Med Serv. 2020;3(1):16–21.
8. Laudermilch DJ, Schiff MA, Nathens AB, Rosengart MR. Lack of emergency medical services documentation is associated with poor patient outcomes: a validation of audit filters for prehospital trauma care. Journal of the American College of Surgeons. 2010;210(2):220–227.
9. Staff T, Søvik S. A retrospective quality assessment of pre-hospital emergency medical documentation in motor vehicle accidents in south-eastern Norway. Scandinavian journal of trauma, resuscitation and emergency medicine. 2011;19:1–11.

10. Falcetta FS, De Almeida FK, Lemos JCS, Goldim JR, Da Costa CA. Automatic documentation of professional health interactions: A systematic review. Artificial Intelligence in Medicine. 2023;137:102487.
11. Haberle T, Cleveland C, Snow GL, et al. The impact of nuance DAX ambient listening AI documentation: a cohort study. Journal of the American Medical Informatics Association. 2024;31(4):975–979.
12. Webpage title: Automatically document care with DAX™ Copilot. Accessed: 05/08, 2025. https://www.nuance.com/asset/en_us/collateral/healthcare/data-sheet/ds-ambient-clinical-intelligence-en-us.pdf
13. Webpage title: AI S. Enterprise-grade AI assistant Accessed 11/17, 2024. https://www.suki.ai/suki-assistant/
14. Willis M, Jarrahi MH. Automating documentation: a critical perspective into the role of artificial intelligence in clinical documentation. Springer; 2019:200–209.
15. Wu S, Roberts K, Datta S, et al. Deep learning in clinical natural language processing: a methodical review. Journal of the American Medical Informatics Association. 2020;27(3):457–470.
16. Goel A, Gueta A, Gilon O, et al. Llms accelerate annotation for medical information extraction. Proceedings of Machine Learning Research. 2023;225:82–100..
17. Nievas M, Basu A, Wang Y, Singh H. Distilling large language models for matching patients to clinical trials. Journal of the American Medical Informatics Association. 2024; 31(9), 1953-1963,.
18. Benary M, Wang XD, Schmidt M, et al. Leveraging large language models for decision support in personalized oncology. JAMA Network Open. 2023;6(11):e2343689-e2343689.
19. Chen A, Yu Z, Yang X, Guo Y, Bian J, Wu Y. Contextualized medication information extraction using transformer-based deep learning architectures. Journal of biomedical informatics. 2023;142:104370.
20. Consoli B, Wu X, Wang S, et al. SDoH-GPT: Using Large Language Models to Extract Social Determinants of Health (SDoH). *arXiv preprint arXiv:240717126*. 2024;
21. Andrew JJ, Vincent M, Burgun A, Garcelon N. Evaluating LLMs for Temporal Entity Extraction from Pediatric Clinical Text in Rare Diseases Context. Proceedings of the First Workshop on Patient-Oriented Language Processing (CL4Health)@ LREC-COLING, 2024; pp 145-152..
22. Sivarajkumar S, Tam TYC, Mohammad HA, et al. Extraction of sleep information from clinical notes of Alzheimer's disease patients using natural language processing. Journal of the American Medical Informatics Association. 2024; 31(10), 2217-2227.
23. Luo X, Tahabi FM, Marc T, Haunert LA, Storey S. Zero-shot learning to extract assessment criteria and medical services from the preventive healthcare guidelines using large language models. Journal of the American Medical Informatics Association. 2024; 31(8), 1743-1753.
24. Koleck TA, Dreisbach C, Bourne PE, Bakken S. Natural language processing of symptoms documented in free-text narratives of electronic health records: a systematic review. Journal of the American Medical Informatics Association. 2019;26(4):364–379.
25. Hsu E, Malagaris I, Kuo Y-F, Sultana R, Roberts K. Deep learning-based NLP data pipeline for EHR-scanned document information extraction. JAMIA open. 2022;5(2):ooac045.
26. Kaufman D, Sheehan B, Stetson P, et al. Natural language processing-enabled and conventional data capture methods for input to electronic health records: a comparative usability study. JMIR Med Inform. 2016; 4 (4): e35. https://doi.org/10.2196/medinform. 5544. 2016.
27. Xia X, Ma Y, Luo Y, Lu J. An online intelligent electronic medical record system via speech recognition. International Journal of Distributed Sensor Networks. 2022;18(11):15501329221134479.
28. Ahamed S, Weiler G, Boden K, et al. Deep neural network driven speech classification for relevance detection in automatic medical documentation. *Public Health and Informatics*. IOS Press; 2021:63–67.
29. Finley GP, Edwards E, Robinson A, et al. An Automated Assistant for Medical Scribes. Proceedings of conference INTERSPEECH, 2018; pp. 3212–3213.
30. Maas L, Kisjes A, Hashemi I, et al. Automated Medical Reporting: From Multimodal Inputs to Medical Reports through Knowledge Graphs. 2021:509–514.
31. Woo M, Mishra P, Lin J, et al. Complete and resilient documentation for operational medical environments leveraging mobile hands-free technology in a systems approach: Experimental study. JMIR mHealth and uHealth. 2021;9(10):e32301.
32. Khattak FK, Jeblee S, Crampton N, Mamdani M, Rudzicz F. AutoScribe: extracting clinically pertinent information from patient-clinician dialogues. *MEDINFO 2019: Health and Well-being e-Networks for All*. L. Ohno-Machado and B. Séroussi (Eds.),IOS Press; 2019:1512–1513.
33. Wenceslao SJMC, Estuar MRJE. Using cTAKES to build a simple speech transcriber plugin for an EMR. Proceedings of the 3rd International Conference on Medical and Health Informatics. 2019; pp. 78-86
34. Kothari K, Zuger C, Desai N, et al. Effect of repetitive simulation training on emergency medical services team performance in simulated pediatric medical emergencies. AEM Education and Training. 2021;5(3):e10537.
35. Zhang Z, Joy K, Upadhyayula P, Ozkaynak M, Harris R, Adelgais K. Data work and decision making in emergency medical services: a distributed cognition perspective. *Proceedings of the ACM on Human-Computer Interaction*. 2021;5(CSCW2):1–32.
36. V3 Data Dictionaries & XSD - NEMSIS https://nemsis.org/technical-resources/version-3/version-3-data-dictionaries/
37. Achiam J, Adler S, Agarwal S, et al. Gpt-4 technical report. *arXiv preprint arXiv:230308774*. 2023;
38. Team G, Anil R, Borgeaud S, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:231211805*. 2023;
39. Dubey A, Jauhri A, Pandey A, et al. The llama 3 herd of models. *arXiv preprint arXiv:240721783*. 2024;
40. Introducing the next generation of Claude — anthropic.com. https://www.anthropic.com/news/claude-3-family
41. Li Y, Ramprasad R, Zhang C. A Simple but Effective Approach to Improve Structured Language Model Output for Information Extraction. *arXiv preprint arXiv:240213364*. 2024;
42. Bhate NJ, Mittal A, He Z, Luo X. Zero-shot learning with minimum instruction to extract social determinants and family history from clinical notes using GPT model. IEEE; 2023:1476–1480.
43. Gali N, Mariescu-Istodor R, Hostettler D, Fränti P. Framework for syntactic string similarity measures. Expert Systems with Applications. 2019;129:169–185.
44. Ferreira R, Lins RD, Simske SJ, Freitas F, Riss M. Assessing sentence similarity through lexical, syntactic and semantic analysis. Computer Speech & Language. 2016;39:1–28.
45. Luo X, Shah S. Concept embedding-based weighting scheme for biomedical text clustering and visualization. Springer; 2018:8.
46. Nguyen HT, Duong PH, Cambria E. Learning short-text semantic similarity with word embeddings and external knowledge sources. Knowledge-Based Systems. 2019;182:104842.
47. Blagec K, Xu H, Agibetov A, Samwald M. Neural sentence embedding models for semantic similarity estimation in the biomedical domain. BMC bioinformatics. 2019;20:1–10.
48. Berger B, Waterman MS, Yu YW. Levenshtein distance, sequence comparison and biological database search. IEEE transactions on information theory. 2020;67(6):3287–3294.

49. Cer D, Yang, Y., Kong, S.Y., Hua, N., Limtiaco, N., John, R.S., Constant, N., Guajardo-Cespedes, M., Yuan, S., Tar, C. and Sung, Y.H. Universal sentence encoder.
50. Chinchor N, Sundheim BM. MUC-5 evaluation metrics. 1993:
51. Luo X, Zhou L, Adelgais K, Zhang Z. Assessing the Effectiveness of Automatic Speech Recognition Technology in Emergency Medicine Settings: A Comparative Study of Four AI-powered Engines. Journal of Healthcare Informatics Research. 2025:1–19.
52. Mustafa A, Naseem U, Azghadi MR. Large language models vs human for classifying clinical documents. International Journal of Medical Informatics. 2025:Volume 195,105800.
53. Singhal K, Tu T, Gottweis J, et al. Toward expert-level medical question answering with large language models. Nature Medicine. 2025: volume 31:1–8.
54. Billa JG, Oh M, Du L. Supervisory Prompt Training. *arXiv preprint arXiv:240318051*. 2024;
55. Lubin JS, Shah A. An incomplete medical record: transfer of care from emergency medical services to the emergency department. Cureus. 2022;14(2): e22446
56. Holzman TG. Computer-human interface solutions for emergency medical care. interactions. 1999;6(3):13–24.
57. Quiroz JC, Laranjo L, Kocaballi AB, Berkovsky S, Rezazadegan D, Coiera E. Challenges of developing a digital scribe to reduce clinical documentation burden. NPJ digital medicine. 2019;2(1):114.
58. Xie Q, Chen Q, Chen A, et al. Me llama: Foundation large language models for medical applications. *arXiv preprint arXiv:240212749*. 2024;
59. Cercone A, Ramgopal S, Martin-Gill C. Completeness of pediatric versus adult patient assessment documentation in the National Emergency Medical Services Information System. Prehospital Emergency Care. 2024;28(2):243–252.
60. Allgood RA, Faris GW, Supples M, Lardaro T, Crowe RP. Results of a quality improvement initiative to increase the completion rate of electronic health records for patient encounters at a large urban fire-based non-transporting EMS agency. Prehospital emergency care. 2024;28(5):696–702.
61. Kiourtis A, Mavrogiorgou A, Menychtas A, Maglogiannis I, Kyriazis D. Structurally mapping healthcare data to HL7 FHIR through ontology alignment. Journal of medical systems. 2019;43:1–13.
62. Kouremenou E, Kiourtis A, Kyriazis D. A data modeling process for achieving interoperability. Springer; 2023:711–719.
63. Shi D, Jin R, Shen T, Dong W, Wu X, Xiong D. IRCAN: Mitigating Knowledge Conflicts in LLM Generation via Identifying and Reweighting Context-Aware Neurons. *arXiv preprint arXiv:240618406*. 2024;
64. Jin H, Han X, Yang J, et al. LLM Maybe LongLM: SelfExtend LLM Context Window Without Tuning.
65. Li T, Zhang G, Do QD, Yue X, Chen W. Long-context llms struggle with long in-context learning. *arXiv preprint arXiv:240402060*. 2024;
66. Ong JCL, Chang SY-H, William W, et al. Ethical and regulatory challenges of large language models in medicine. The Lancet Digital Health. 2024;6(6):e428-e432.
67. Gerke S, Minssen T, Cohen G. Ethical and legal challenges of artificial intelligence-driven healthcare. Editor(s): Adam Bohr, Kaveh Memarzadeh, *Artificial intelligence in healthcare*. Elsevier; 2020:295–336.