Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Implementation of Large Language Models in Electronic Health Records

Maxime Griot

maxime.griot@uclouvain.be

Université Catholique de Louvain

Jean Vanderdonckt

Université Catholique de Louvain

Demet Yuksel

Cliniques universitaires Saint-Luc

Article

Keywords: artificial intelligence, clinical applications, clinical investigations, electronic health records, fhir, generative artificial intelligence, large language models, machine learning, transformers, user-machine interactions

Posted Date: July 4th, 2025

DOI: https://doi.org/10.21203/rs.3.rs-7029913/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Implementation of Large Language Models in Electronic Health Records

Maxime Griot[1,2,3], Jean Vanderdonckt[2], and Demet Yuksel[1,3]
[1]Cliniques universitaires Saint-Luc, 1200 Woluwe-Saint-Lambert, Belgium
[2]Université catholique de Louvain, Louvain Research Institute in Management and Organizations, B-1348 Louvain-la-Neuve, Belgium
[3]Université catholique de Louvain, Institute of NeuroScience, 1200 Woluwe-Saint-Lambert, Belgium

Corresponding author: Maxime Griot (e-mail: maxime.griot@uclouvain.be).

This work was supported by the Fondation Saint-Luc grant number 467E and the Fédération Wallonie-Bruxelles through the Fond Spécial de Recherche of Université catholique de Louvain.

ABSTRACT Electronic Health Records (EHRs) have greatly improved access to clinical documents which also resulted in new challenges for clinicians, with some spending over 50% of their time on non-clinical tasks. While Large Language Models (LLMs) offer promise for reducing this burden, most implementations focus on synthetic benchmarks rather than clinical deployment. This paper presents a secure, fully on-premises, GDPR-compliant LLM chatbot integrated into the Epic EHR system at a European university hospital. The system utilizes Qwen3-235B with Retrieval Augmented Generation (RAG) for contextual awareness across internal records, regional eHealth documents, and medical literature. During a one-month pilot with 28 physicians from nine specialties, over 400 multi-turn conversations were initiated, with 64% of participants using the tool daily. Primary use cases included chart summarization, targeted retrieval, and support for clinical reasoning. Clinicians emphasized the model's utility in rapidly synthesizing dispersed information, while diagnostic use remained limited. Our results demonstrate the technical feasibility and adoption patterns of integrating LLMs into production EHR systems, providing a replicable framework for secure clinical AI deployment.

INDEX TERMS artificial intelligence, clinical applications, clinical investigations, electronic health records, fhir, generative artificial intelligence, large language models, machine learning, transformers, user-machine interactions

I. INTRODUCTION

THE digitization of the medical field brought many benefits, such as better continuity of care, centralized access to data, and larger volumes of data for research [1], [2]. However, these benefits are not without cost; physicians spend more time documenting since the introduction of Electronic Health Records (EHRs), with some spending more than 50% of their time using EHRs [3], [4]. This additional time spent is partially explained by the need to document the same information in different formats. Whereas in the past, clinicians would write a diagnosis in a clinical note, they must now document it in structured fields within EHRs and in the clinical note [5]. This results in either increased documentation time and dissatisfaction or incomplete documentation, often in structured fields. In addition to this double documentation burden, the availability of clinical documents also increases the volume of information available and that physicians have to process when preparing a consultation, on average, 20 minutes per patient for emergency physicians at our hospital. This twofold increase in documentation tasks, either during the preparation phase or writing phase, causes already overworked physicians stress and reduces their ability to see patients [6]. Even though EHR systems offer features to facilitate conversion to structured data, many aspects of optimal EHR usage remain dependent on individual behaviors and cannot be fully controlled.

LLMs have garnered much attention in medicine, with models achieving impressive scores on medical benchmarks [7]. Their impressive ability to process large volumes of text was rapidly identified as a potential solution to handle medical tasks. However, most research focuses on support for clinical reasoning, targeting diagnostics or treatment planning on synthetic benchmarks [8].

While existing research focuses predominantly on clinical accuracy benchmarks using synthetic datasets, there remains a critical gap in understanding how to successfully deploy and integrate LLMs into live clinical environments. The transition from controlled evaluations to production systems presents unique challenges in security, governance, user adoption, and workflow integration that are rarely addressed in the literature. This study addresses these implementation challenges rather than efficacy measurement, focusing on the technical architecture, security framework, and real-world adoption patterns of LLM deployment in clinical practice. Rather than measuring clinical outcomes, we evaluate whether physicians will adopt LLM tools when properly integrated into existing

EHR workflows and characterize how they naturally use such systems in practice.

The main contributions of our work are:

- A comprehensive security and governance framework for clinical AI deployment, including GDPR compliance, on-premises hosting, and institutional oversight protocols
- The description of a complete technical implementation in a live clinical setting, providing a replicable architecture for EHR-integrated LLM systems
- Empirical evidence of physician adoption patterns, with 64% daily usage among 28 clinicians across nine specialties during real-world deployment
- Usage pattern analysis revealing how clinicians naturally interact with LLM assistants, with predominant use for information synthesis (28.5%) and retrieval (28.2%) rather than diagnostic support
- An approach to governance and change management to support innovative and experimental technological advances in clinical settings

II. RELATED WORK

Medicine is a critical field, making it difficult to access real-world clinical data and to perform real-world experiments. This has resulted in a disconnect between the clinical realities and how AI research is conducted. The developed assessment tools have been largely based on multiple choice questions (MCQs) used to evaluate medical students and residents [9]. While the relevance of MCQs to evaluate LLMs is debated [9], [10], they have the benefit of being reproducible, easy to administer, and most importantly available [11]–[13]. While the importance of testing in the real world was emphasized in medical literature, investigations have been lacking, with most real-world evaluations being limited to specific tasks in controlled environments [14]. The two common approaches for real-world clinical evaluation are either based on generating or transforming existing textual data, or generating textual data from speech data, including ambient listening.

A. TEXT-TO-TEXT

Evaluations of large language models (LLMs) in clinical workflows have generally addressed isolated tasks and rarely examined performance under real-world conditions. Decker *et al.* [15] showed that LLM-generated informed-consent documents for common surgical procedures were comparable to, and occasionally surpassed, those written by surgeons. Zaretsky *et al.* [16] likewise demonstrated that LLMs can create patient-friendly discharge summaries of acceptable quality. Beyond educational materials, automatic drafting of replies to patient messages has been piloted in the Epic electronic health record (EHR); however, Tai-Seale *et al.* [17] found no measurable reduction in clinician workload. Baxter *et al.* [18] examined retrospective clinical notes and reported that LLM-generated drafts often required substantial revision, particularly for messages communicating unfavorable information.

A recent scoping review of clinical summarization studies by Bednarczyk *et al.* [19] identified only retrospective investigations that relied on publicly available datasets, a limitation that restricts analyses to a narrow set of medical specialties. Although study methodologies varied, most focused on output quality and error characterization rather than on effects within clinical workflows. To promote more consistent assessment, Asgari *et al.* [20] introduced an annotation framework that enables fine-grained error reporting and showed that hallucination and omission rates can approach or even fall below those of human experts. Complementing this, MedHELM by Bedi *et al.* [21] offers a comprehensive taxonomy spanning five categories, 22 sub-categories, and 121 clinical text tasks. Their benchmark suite indicates that current LLMs achieve the strongest performance in clinical-note generation and patient communication, findings that align with the individual task studies discussed above. Nevertheless, these benchmarking efforts, like earlier studies, do not quantify the downstream impact on clinical workflows.

B. SPEECH-TO-TEXT

Another branch of real-world evaluation involves generating clinical text from speech data with ambient listening tools that listen to clinician-patient conversations and draft documentation. These systems typically combine automatic speech recognition with LLM-based summarization to produce clinical notes, aiming to relieve physicians from manual charting and allow more focus on the patient. Early results have been encouraging. For instance, Balloch *et al.*demonstrated in a controlled simulation that an AI ''ambient scribe'' could produce higher-quality outpatient documentation (measured by standard scoring) while also shortening visit lengths by about 26% without sacrificing patient interaction time [22]. Clinicians in that study reported a reduced cognitive burden and a better overall experience when the AI handled note-taking. In real clinical settings, preliminary deployments have shown similar promise. A pilot implementation of an ambient AI assistant in primary care found that automatically generated draft notes significantly decreased the time physicians spent on after-hours charting and improved their work satisfaction [23]. On a larger scale, Kaiser Permanente recently rolled out an ambient AI scribe to over 7,000 physicians across 17 medical centers, reporting an estimated 15,700 total hours saved in documentation (about 1 minute less EHR time per patient visit) along with substantially improved physician-patient engagement and job satisfaction [24]. Notably, 84% of surveyed doctors in that deployment felt the AI system allowed them to connect more with patients, and patient feedback was predominantly neutral or positive regarding the tool's presence.

Qualitative evaluations further underscore these benefits. In interviews, physicians have described ambient AI scribes as having a positive impact on their workload, work-life integration, and even patient engagement during visits [25]. In practice, current ambient documentation tools still require clinicians to review and edit the AI-generated notes, and no

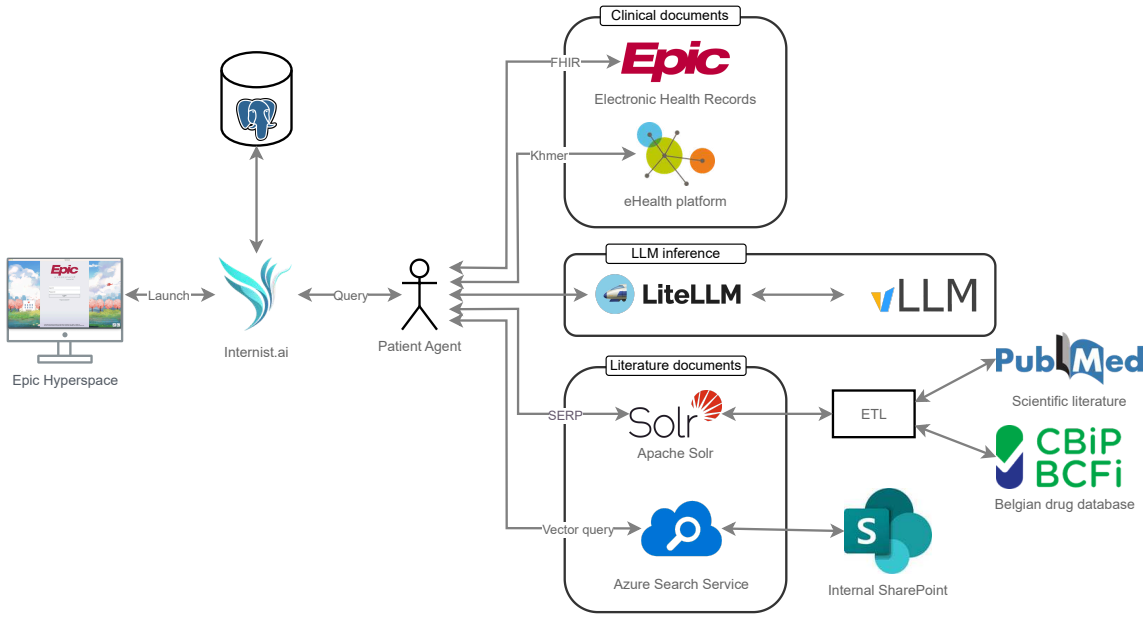


FIGURE 1. Overview of the software architecture and information flow between the components

studies yet quantify their effect on clinical outcomes or error rates. In summary, speech-driven LLM tools offer a compelling solution to documentation burdens and have shown improvements in efficiency and provider satisfaction, but rigorous real-world validation—especially regarding quality assurance and workflow impact—remains an active research area.

III. SYSTEM ARCHITECTURE AND IMPLEMENTATION

The goal of the system is to integrate seamlessly into the day-to-day workflow of physicians. To achieve this goal, we organized workshops with physicians early to teach them the basics of LLMs and discuss the use cases they envisioned [26]. Through this process, we obtained a set of use cases prioritized by the physicians themselves. The most requested use cases were related to improving access to information, be it patient information, scientific literature, or internal procedures.

A chatbot that can retrieve, summarize, and interact with the different sources of information could address most of the use cases proposed by users and serve as a generic entry for current and future use cases. We identified three axes to ensure the tool's usage does not become an additional burden to physicians:

1) The Electronic Health Record must be the initiator and host of the application to reduce context switches and centralize the information in a single application
2) The user interface must be familiar and simple enough to be used without additional training. As users are increasingly used to tools such as ChatGPT, a similar interface appears optimal
3) The integration must not get in the way of existing workflows and should only appear on demand

A. INTEGRATION

Since 2020, the Cliniques universitaires Saint-Luc hospital has utilized the Epic electronic health record (EHR) system, which offers various mechanisms for integrating third-party applications. Among these, we selected web-based integration. Epic supports embedding web applications directly within its user interface and facilitates secure data access through the SMART on FHIR launch protocol.

This protocol enables the EHR to initiate a web application with user-specific context by passing a launch token. The web application can exchange this token for a FHIR access token, which is scoped according to the permissions of the launching user, the application, and the current patient. For instance, when a physician launches the application from within Epic, it is provided a predefined URL and launch context, ensuring that only the necessary and authorized patient data is accessible. This mechanism upholds strict data access control, thereby minimizing the risk of inappropriate data exposure.

This study is designed as a deployment feasibility and adoption study rather than a clinical efficacy trial. Our primary objectives are to demonstrate technical implementation, measure user adoption, and characterize usage patterns. We deliberately did not measure clinical outcomes or time savings, as this would require a controlled study design inappropriate for an initial deployment evaluation.

1) Infrastructure

LLMs require extensive computing capabilities, especially to support thousands of users in a university hospital. The GenAI

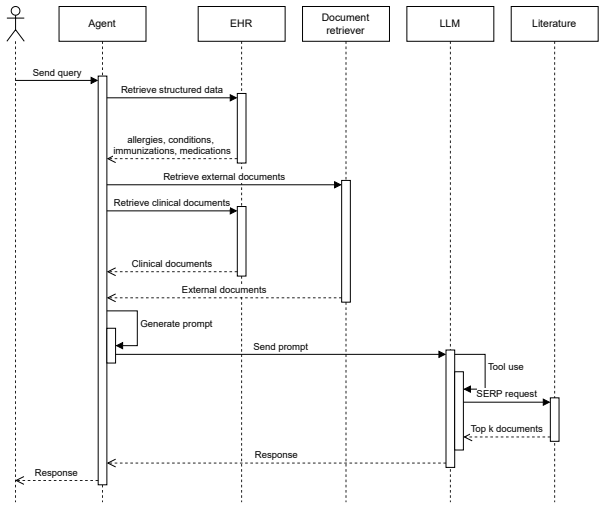


FIGURE 2. **Overview of the response generation process**

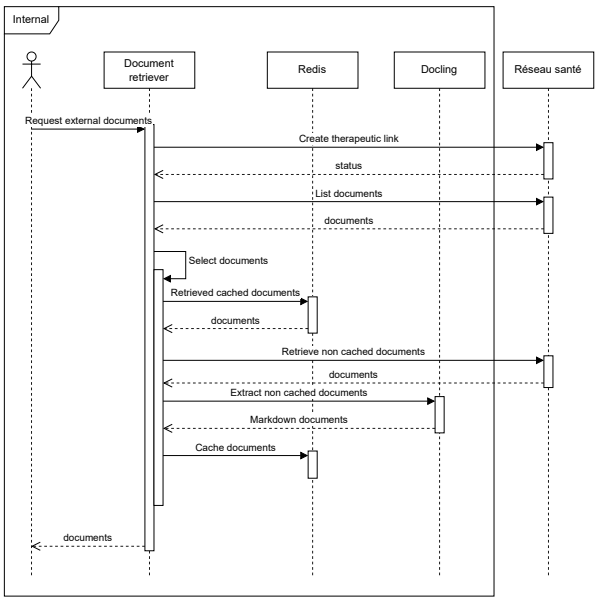


FIGURE 3. **External document retrieval and extraction process**

strategy of our institution is to host medical applications on-site and not to rely on vendor services that do not allow for the level of customization and validation required to make these tools useful and safe in day-to-day workflow.

To support our GenAI efforts, we invested in a server with 8 NVIDIA H200 GPUs, allowing us to host the largest available models, such as DeepSeek V3. In addition to our inference capabilities, this server provides the necessary compute to fine-tune small to medium-sized models on specific tasks.

Our software stack is depicted in Fig. 1 based around vLLM [27], which provides the LLM inference capabilities using the same protocol as OpenAI, allowing us to use existing libraries and methods to interact with the LLM.

2) Cognitive Engine

We deploy Qwen3-235B in FP8 precision [28], enabling efficient inference at scale without compromising model quality. To ensure generated outputs adhere to expected formats, we use vLLM's structured decoding and tool use features. Structured decoding constrains the model to produce outputs that follow predefined schemas, which is critical for downstream systems that rely on predictable, well-formed inputs.

We use LiteLLM to manage access control and system reliability by enforcing usage limits and distributing requests across multiple vLLM instances. This allows us to maintain stable performance and fair resource allocation, even under heavy demand.

B. RETRIEVAL AUGMENTED GENERATION

LLMs are powerful tools, but their effectiveness is limited by the scope of their training data and the content available in their inference context. To improve relevance during inference, it is necessary to inject additional knowledge through Retrieval Augmented Generation (RAG) as depicted in Fig. 2. This can be done through two main strategies: **passive injection**, where a fixed set of knowledge is always included, and **active injection**, where relevant data is retrieved dynamically based on decisions made by the model or the user.

We also distinguish between internal and external knowledge sources. For external sources that pose a risk of data leakage, such as full-text search engines, we apply an Extract-Transform-Load (ETL) process to create a local copy of the data. This ensures the content is properly formatted, indexed in a controlled environment, and filtered to retain only a curated subset aligned with internal requirements.

All documents, regardless of source, are subject to the same security policies. The system operates without internet access, and all documents are either loaded directly onto internal servers or hosted on machines within the internal network. This approach ensures that even if data is inadvertently leaked through user queries, no information is externalized, thereby safeguarding patient data. One exception is the Microsoft SharePoint documents, which are hosted on a private Microsoft Azure cloud instance.

1) Common knowledge

Given the large context windows of current models, we always prepend essential background information to the prompt. Specifically, we include the current date, hospital location, and the provider's name and role. We then add patient data extracted from structured EMR fields via FHIR: demographics (sex, age, encounter type, location) and clinical data (active allergies, chronic conditions, medications, and immunizations). Each clinical datum is accompanied by its recording date, enabling the model to detect and, when necessary, question outdated information.

2) Internal clinical documents

FHIR exposes several document resources stored in the EHR, such as clinical notes and radiology reports. This includes documents imported from our historic EHR. For every user query, we retrieve the complete, chronologically ordered list of signed documents and filter out those addressed to the patient or with low information density. We select up to 20 documents and append their content to the prompt until we reach a 50,000-character budget, remaining safely within the model's context window. For each document, we supply its title together with the signature date.

3) External clinical documents

Belgium's eHealth platform employs a decentralized architecture for sharing healthcare documents. Because our hospital is located in Brussels, we use the *Réseau Santé Bruxellois* (RSB), one of the platform's four regional networks. Through RSB, care providers exchange documents (examination results, clinical notes, correspondence) via web services that implement the Khmer protocol [29]. To ensure we did not overload the eHealth platform, we filtered documents by type and retained only Discharge letters, Consultations, and Imagery reports. After filtering, we selected only the top 10 most recent documents. This decision was made based on the assumption that discharge letters are written to contain all the necessary information for the clinicians to understand the patient's medical history.

Unlike the internal corpus, external files are heterogeneous and often stored as PDFs, whose reliable parsing is non-trivial. We therefore rely on Docling [30], an open-source, fully on-premise tool that converts numerous formats (including PDF) into machine-readable representations such as HTML or Markdown.

Fig. 3 outlines the retrieval pipeline and its caching layer. To reduce Docling's load and improve latency (OCR on large PDFs can exceed a minute), we cache extracted documents for 30 days in Redis. The cache key concatenates the patient's identifier with the document's eHealth identifier, ensuring that a document can be retrieved only when both identifiers are known and preventing access to files deleted from the eHealth network or belonging to other patients.

4) Internal literature

The hospital maintains an internal Microsoft SharePoint in which procedures, guidelines, and other local documentation are stored. This collection is heterogeneous, mostly containing complex PDF and Word files in French. These documents are indexed in an Azure Search Service, which is a vector database used to retrieve documents by passing the embedding vector of the search query.

The major challenge of leveraging the internal documentation is the extraction of documents. PDF files, for instance, are complex to process, especially in non-English languages. The documents also contain many images and tables that are not extracted, further complicating the extraction process. Considering this internal documentation is incomplete and secondary during clinical practice, we did not perform extensive configuration or experiments beyond the standard extraction and vectorization pipeline offered by Microsoft.

5) External literature

Evidence-based medicine is the standard of care, but staying up to date and recalling the details of guidelines is challenging. Physicians often rely on external documents that they review to make optimal decisions for patient care. However, this process is often time-consuming as physicians must find the **relevant documents**, find the **relevant information** in complex documents that can contain hundreds of pages, and often **leave the EHR** to access these documents. The goal of integrating external literature is therefore to address these three concerns to speed up access to the relevant information.

Contrary to the internal literature, which is heterogeneous and difficult to parse, the external documents are homogeneous and structured. This allowed us to use a different approach, not based on vector databases but on classical text search in the content and the metadata. This gives the model the ability to generate different queries and improve the queries based on the result list.

a: Relevant Documents

Identifying relevant documents in the exponentially growing scientific literature is a challenging task in itself, so much so that physicians often rely on private databases such as UpToDate [31] that synthesize medical literature. However, these databases do not allow automatic processing, which implies that alternatives must be found. We therefore rely on the suggestions of physicians made during the workshops [26] and identify subsets of PubMed to reduce the search space and improve the likelihood of finding relevant documents.

In practice, PubMed exposes an FTP service that contains a structured representation of articles and eBooks, which we use to download the selected subset of articles and data relevant to our physicians. The documents are then parsed and transformed into Markdown before being indexed in Apache Solr.

b: Relevant Information

Contrary to internal literature, the well-formed and structured nature of the documents allows for more accurate searches, especially semantic. We therefore opted for full-text semantic search using Apache Solr instead of vector databases. The documents are chunked based on level 1 and 2 titles present in the documents.

The LLM uses a tool call to make a query to Apache Solr with a full-text query, and the documents and their contents are then returned. We make an additional call to the LLM with a list of document names and ask the LLM to pick the top 3 documents, acting as a ReRanker. These 3 documents are then injected as a tool call result.



FIGURE 4. User interface of the chatbot in the EHR

c: Integrated in the EHR

Considering we select the documents and sources in advance, we have a comprehensive list of valid URLs allowed to be opened directly in the EHR, which we whitelisted. As detailed in the User Interface section, the documents used by the LLM are listed and displayed as clickable links to open the source of information in a tab directly in the EHR, ensuring quick access to the original documents.

C. USER INTERFACE

Physicians are increasingly familiar with conversational AI tools such as ChatGPT and Claude, which typically employ a minimalist chat-based interface accompanied by a list of previous interactions. We adopted a similar approach by building upon the open-source OpenWebUI interface [32].

Originally designed for advanced users seeking fine-grained control over language models, OpenWebUI includes numerous configuration options. To streamline the interface for clinical use, we removed most advanced features and retained only the essential interaction loop. In addition, we introduced several enhancements to support day-to-day clinical workflows as shown in Fig. 4, including:

1) **Reasoning control**, which allows users to toggle between a standard and a faster non-reasoning mode;
2) **Source control**, enabling physicians to specify which data sources should contribute to the model's response;
3) **Data retrieval status**, providing visibility into the documents retrieved by the system;
4) **Scoped conversations**, which help maintain contextually relevant and patient-specific interactions.

a: Explainability

In a medical context, ensuring the correctness of information is not just desirable; it is mandatory. To support this, our system emphasizes transparency throughout the retrieval and response generation process. Every answer generated by the model is accompanied by a structured list of the sources consulted, as shown in Fig. 5. These sources include both documents within the EHR and external clinical references. For external documents, the system provides direct hyperlinks that open the relevant material in a dedicated EHR-integrated reading window, allowing physicians to review the original context without leaving the clinical interface. This design not only facilitates rapid verification but also reinforces trust in the AI's recommendations by enabling traceability and auditability of every response.

IV. SECURITY, PRIVACY, AND GOVERNANCE

The deployment of generative AI tools in clinical settings requires not only technical robustness but also a governance model that ensures safety, accountability, and compliance with legal and institutional policies. Our governance framework was developed in collaboration with medical leadership, the legal department, cybersecurity teams, and frontline clinicians to ensure responsible innovation.

A. INSTITUTIONAL OVERSIGHT AND ETHICS

The institutional ethics committee reviewed the project and categorized it as a study of practice rather than a clinical investigation, enabling deployment in a live clinical environment without the need for a formal trial protocol. This classification allowed for iterative development while maintaining alignment with ethical oversight standards.

The hospital's medical leadership was engaged early in the process, and a collective decision by the medical direction authorized the system's deployment in production. The tool was designed to be non-interventional, fully optional, and integrated into existing workflows without altering clinical responsibilities or decision-making.

A key component of governance was the involvement of the hospital's network of physician ''EHR champions,'' representing multiple specialties. These clinicians participated throughout the development lifecycle, from use-case definition to interface testing. Before go-live, the application was demonstrated in a test environment during a dedicated one-hour review session, which served to surface potential limitations and incorporate clinician feedback into final refinements.



FIGURE 5. User interface showing the chatbot's response with collapsed document sources on the left, and the expanded RSB sources on the right showing the title of the documents used and their publish date.

In parallel, the hospital's Data Protection Officer (DPO) was consulted to ensure compliance with privacy and data protection regulations. A Data Protection Impact Assessment (DPIA) was completed before deployment, addressing potential risks associated with the use of sensitive health data and documenting mitigation strategies. The DPIA also confirmed that no patient data would leave the hospital infrastructure, and that the application met the standards of the EU General Data Protection Regulation (GDPR) [33].

Finally, we notified the Belgian Federal Agency for Medicines and Health Products and declared the system as medical device class IIa.

B. ACCESS CONTROL AND DATA MINIMIZATION

Access control is enforced through the SMART on FHIR launch protocol, which delegates user identification and authorization to the EHR system. When a user launches the application, Epic issues a context-specific access token that scopes the data the application can retrieve based on the user's role, the selected patient, and the configured application permissions.

This approach ensures strict adherence to the principle of data minimization: the system can only access information that the user is already authorized to view within the EHR interface. No persistent data storage occurs outside of the application's runtime session, and all access to patient data is ephemeral and auditable.

By leveraging Epic's access controls, the system inherits existing institutional safeguards, such as role-based access restrictions and user audit logging. This design ensures that data retrieval remains compliant with internal policies and external regulations while minimizing the surface area for potential misuse.

C. MODEL GOVERNANCE AND UPDATES

All models deployed in this environment are hosted on-premise and are subject to internal validation procedures. Before any model update or fine-tuning, test cases are evaluated using existing medical benchmarks such as MedQA [11] and internal manual validation on a set of standardized tasks in test environments.

A formal versioning system is maintained to ensure traceability of responses to specific model builds. Changes and new models are rolled out to champions who validate the models in real clinical contexts for a fixed period of time, depending on the magnitude of the change. Based on their feedback, the model is then pushed to all users or rejected.

D. COMPLIANCE WITH LEGAL AND REGULATORY FRAMEWORKS

The system was formally declared as an in-house medical device under the European Medical Device Regulation (MDR) [34], following Article 5(5), which allows healthcare institutions to develop and use custom devices internally, provided they meet applicable safety and performance requirements.

In terms of data protection, the system fully complies with the GDPR [33]. No personal data leaves the hospital's infrastructure, and access to patient data is restricted to what is necessary for the user's current clinical context. All access is governed by existing role-based permissions managed within the EHR.

While the European Union Artificial Intelligence Act [35] is not yet fully enacted, we have proactively aligned the development process with its requirements for high-risk AI systems. Specifically, we implemented:

- Documentation of intended use, model behavior, and system limitations;
- Technical documentation of the system;

- Internal traceability and version control for model updates and data pipeline changes;
- User training materials and interface safeguards to prevent over-reliance;
- Organizational policies to support transparency, accountability, and post-deployment monitoring.

This approach ensures that the system is not only compliant under current regulations but also positioned to meet emerging legal standards in AI governance.

E. INCIDENT MANAGEMENT AND USER FEEDBACK

To support safe and responsible deployment, a real-time feedback mechanism was integrated directly into the user interface. Clinicians can rate each response on a scale from 1 to 10, select one or more predefined reasons for their score (e.g., hallucination, irrelevance, omission), and optionally provide free-text comments to elaborate on their evaluation.

Submitted feedback is logged and reviewed by the project team, with reported incidents triaged on a rolling basis and systematically reviewed in monthly meetings. Critical issues—such as incorrect clinical reasoning or information retrieval failures—are prioritized for investigation and resolution.

In addition, a user feedback dashboard aggregates both quantitative metrics (e.g., usage frequency, latency, error rates) and qualitative data (e.g., user comments, flagged examples). This enables continuous monitoring of system performance and user satisfaction.

This physician-in-the-loop feedback loop supports iterative refinement of the system and ensures that future development is guided by clinical needs, usability concerns, and real-world performance.

V. PILOT STUDY AND EVALUATION RESULTS

The one-month pilot study enrolled 28 physicians from nine specialties: oncology, geriatrics, surgery, internal medicine, pediatrics, intensive care, emergency medicine, ophthalmology, and radiotherapy. The intervention focused exclusively on physicians' practice patterns; no patient-level clinical outcomes were collected. During this one-month pilot, we initially only enabled internal clinical document retrieval to allow clinicians to gain experience with the system without the additional complexity of understanding data flow. During the last week of the pilot, the complete system was enabled.

A. ADOPTION

After completing the brief training session, physicians received production access to the model. Over the 30-day observation window, 18 physicians interacted with the system daily, whereas 5 discontinued use after their initial exploration. Adoption remained high despite technical issues during the first week (e.g., a missing launch icon for certain patients and occasional failures to retrieve clinical documents). In total, 482 conversations, 1630 messages were logged, and 47 discrete feedback items were collected: 25 positive and 22 negative.

Radiotherapy clinicians rapidly converged on a set of optimized prompts, which they shared informally within the department to accelerate information retrieval. This behavior suggests that the high engagement observed may be partially attributable to the curiosity and technological enthusiasm of the volunteer cohort.

B. USAGE PATTERNS

Analysis of the first user message and the automatically generated conversation title (without inspecting model responses or subsequent turns to preserve patient anonymity) revealed four predominant usage patterns:

- **Summarization**
- **Information retrieval**
- **Differential diagnosis**
- **Note writing**

Because conversations were multi-turn, a single thread could evolve from one pattern to another. However, we report only the initial message's intent. Only five instances fell outside these four patterns: one asking for a drug reimbursement protocol, another translating a document, and three malformed queries. This concentration of usage in four patterns can be explained by the recent introduction of the tool, which did not allow users enough time to become confident enough with the tool to test different kinds of approaches.

1) Summarization

Summarization was the most common workflow, representing 29.5% of all uses ($n = 142$). Emergency department physicians, in particular, used the model to prepare consultations. The ability to process dozens of documents concurrently enabled rapid assimilation of a patient's history. Only one hallucination was reported: the model indicated that the patient was being considered for palliative care, which was absent from the record. The attending physician noted, however, that this was a true clinical consideration that had not yet been documented. Four omission events were also reported, with only one having details on the omission, which was related to pathology results that were not correctly indexed in our FHIR implementation.

2) Information Retrieval

Targeted information retrieval ranked second in frequency with 29.2% of uses ($n = 141$). Typical queries included requests for results of prior investigations, family history, or medication dosages. Geriatricians additionally retrieved variables needed for risk scores such as HEMORR2HAGES [36]. While the model accurately surfaced the raw data, it misinterpreted certain score components; for example, it failed to recognize that the bleeding-risk element of the HEMORR2HAGES score refers specifically to active bleeding.

3) Differential Diagnosis

The model was also consulted for differential diagnosis in 25.1% of cases ($n = 121$). In one illustrative exchange,

Rating	Reason	Comment
✓	accurate_information	It's excellent that it mentions the patient is Dutch-speaking.
✓	accurate_information	The treating ophthalmologist question needs to be explained more clearly in the prompt.
✓	accurate_information	A slight improvement in the accurate understanding of a medical case: this concerns a patient undergoing hormone therapy for breast cancer, specifically with tamoxifen. Tamoxifen strengthens the bones, thus reducing the risk of osteoporosis, unlike other hormone therapies. Therefore, the following suggestion is not relevant for this type of hormone therapy (it could remain as a general preventive measure, but not in relation to the hormone therapy): Osteoporosis prevention: Measurement of bone mineral density (T-score) before initiating long-term hormone therapy.
✓	accurate_information	-
✓	accurate_information	-
✓	attention_to_detail	-
✓	attention_to_detail	-
✗	factual_errors	the family doctor is missing
✗	factual errors	She received cisplatin
✗	hallucination	"palliative care considered" is not mentioned in clinical notes (that being said, it's true).
✗	misunderstanding	Seems to confuse preventive and therapeutic anticoagulation (different dosages), and does not properly sequence the timeline nor seem to distinguish between ''presence of an embolism/thrombosis'' vs. ''past anticoagulation.'' In general, I think AI will often be used to trace a past treatment/event: perhaps for this type of question, it would be good to redirect the user to Epic tools designed for that purpose? (summary/life chronology)
✗	misunderstanding	The treating ophthalmologist = external ophthalmologist who referred the patient.
✗	misunderstanding	-
✗	omission	The patient's tumor marker has been measured every 3 months for a very long time.
✗	omission	XY is the family doctor
✗	omission	-
✗	omission	The AI seems to have access to imagery reports, but not pathology? It would be very useful!
✗	omission	-
✗	omission	-
✗	other	Answer in Spanish
✗	other	-
✗	other	incomplete information
✗	other	No response after 3 minutes
✓	showcased creativity	Tuberculosis had not been considered. We will exclude it.
✓	thorough_explanation	
✓	thorough_explanation	
✓	-	!!! did not accurately transcribe the surgical protocol described in the letter dated [REDACTED]: ''a recession of the right lateral rectus muscle by [REDACTED] mm and a resection of the right medial rectus muscle by [REDACTED] mm''; instead, it found this information: ''Recurrence after surgery in [REDACTED] (recession and resection of the right lateral rectus).''
✓	-	-
✓	-	-
✓	-	-
✓	-	-
✓	-	-
✓	-	-
✗	-	The system is amazing, but it does not have access to oncologic history
✓	-	-
✓	-	-
✓	-	-
✗	-	Not what I asked
✗	-	-
✓	-	-
✓	-	-
✓	-	-
✗	-	-
✓	-	-
✗	-	-
✓	-	-
✗	-	The conclusion is correct: anticoagulation is needed, and close monitoring is essential due to a high risk of bleeding. However, the details of the hemorrhage score are not entirely accurate. The first H = renal and hepatic insufficiency. E = ethanol (alcohol). M = correct (malignant pathology). O = older age (elderly). R2 = bleeding risk, but this should be assessed based on active bleeding (look for free text in the medical record mentioning this or evidence of red blood cell transfusion - a loss of 2 g/dL acutely is the cut-off to consider). The second H = labile hypertension. Look for blood pressure spikes and episodes of hypotension in the record. A = correct. G = correct. E = excess falls (risk of falling); search for free text in the file. S = correct.

TABLE 1. Complete export of the feedback received during this pilot. After giving a thumbs up or down to a response, the users are provided with a list of predetermined reasons to choose from, as well as a free-text form to add a comment or detail the reason if none of the options match the problem. The comments are translated from French.

a physician queried: ''Can you find an explanation for the desaturation?'' The system returned several possibilities and prompted consideration of tuberculosis, an option the team had not initially considered. Overall, diagnostic support was less prevalent than retrieval or clerical assistance, aligning with prior expectations that LLMs have more potential as efficient data retrievers rather than decision makers.

4) Note Writing

Finally, physicians also leveraged the system to draft note sections based on chart data and their shorthand observations in 15.1% of cases ($n = 73$). This behavior is consistent with prior studies underscoring documentation burden in clinical workflows [3]. No feedback was provided for this use case.

C. FEEDBACK

During the study, 815 assistant turns were eligible for rating; users reacted to only 47 of them (5.8 %). Among those 47 ratings, 25 were positive (✓, 53.2 %) and 22 negative (✗, 46.8 %), giving an almost even split.

- *Structured reasons.* Less than half of all ratings (21/47, 44.6 %) carried no predefined reason (''-'' in Tab. 1). Reasons were omitted more often with positive votes (15/25, 60.0 %) than negative ones (6/22, 27.2 %).
- *Free-text comments.* 18 ratings (38.3 %) contained a comment. 13 of these comments accompanied negative votes, underscoring that users tend to justify a down-vote but rarely a thumbs-up.
- *Reason distribution.* When a reason was given, the most frequent negative labels were *Omission* (6/16), *Other* (4/16), *Misunderstanding* (3/16), *Factual error* (2/16), and a single instance of *Hallucination* was recorded. Positive reasons were led by *Accurate information* (5/10), *Attention to detail* (2/10), *Thorough explanation* (2/10), and a single instance of *Creativity*.

The pattern suggests that users intervened primarily when the assistant either omitted clinically relevant context (e.g. pathology reports, prior chemotherapy, explicit HEMORR2HAGES variables) or produced output that was off-topic (wrong language, excessive latency). True content errors, such as hallucinations or factual mistakes, were rare.

Several factors plausibly reduced the feedback rate:

1) **Workflow pressure:** Clinicians in busy wards have limited time; unless a response is conspicuously wrong or extraordinarily helpful, providing a rating competes with patient-care tasks.
2) **Low friction for silence:** The pilot did not gate progress on submitting feedback; the path of least resistance was to say nothing.
3) **Perceived adequacy:** When the assistant's reply was simply ''good enough,'' clinicians had little incentive to spend extra clicks to endorse it. This produced an asymmetry in which users wrote detailed explanations for negative ratings yet stayed silent after positive ones.
4) **Interface cues:** Structured reasons required an extra click and comments an extra text entry; each additional action step measurably decreases response rates.

These observations suggest that future deployments should (i) streamline the rating UI, (ii) actively prompt for brief positive comments, and (iii) log implicit signals (e.g., time-to-next-question) as complementary quality indicators when explicit feedback is scarce.

VI. CONCLUSION AND FUTURE WORK

This work presents a fully on-premises, GDPR-compliant, EHR-embedded LLM-based chatbot deployed in routine clinical care. During a one-month pilot involving 28 physicians from nine specialties, the system processed 404 conversations, with 64% of participants using it daily. The prevailing usage category was related to information retrieval and synthesis, indicating that clinicians perceive the assistant as a documentation aid, rather than a support for clinical reasoning, which accounted for only 26.7% of uses.

The study is limited by its single-center design, short duration, and the absence of patient-level outcome measures. Future work will extend evaluation to multiple sites, include a controlled time-and-motion study to quantify workflow impact, and deploy the tool hospital-wide to reduce selection bias.

By demonstrating that secure, on-premises LLM deployment is feasible and shows early signals of clinical usefulness, this study provides a foundation for AI assistants that complement rather than replace physician expertise.

ACKNOWLEDGMENTS

The authors thank Pavel Jez for assistance with tool setup and deployment, the IT Department for on-premises machine support, and the participating physicians.

COMPETING INTERESTS

All authors declare no financial or non-financial competing interests.

REFERENCES

[1] D. P. Manca, "Do electronic medical records improve quality of care?" *Canadian Family Physician*, vol. 61, no. 10, pp. 846–847, Oct. 2015. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4607324/

[2] D. Capurro, M. Y. PhD, E. van Eaton, R. Black, and P. Tarczy-Hornoch, "Availability of Structured and Unstructured Clinical Data for Comparative Effectiveness Research and Quality Improvement: A Multisite Assessment," *EGEMS*, vol. 2, no. 1, p. 1079, Jul. 2014. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4371483/

[3] Y. Pinevich, K. J. Clark, A. M. Harrison, B. W. Pickering, and V. Herasevich, "Interaction Time with Electronic Health Records: A Systematic Review," *Applied Clinical Informatics*, vol. 12, no. 4, pp. 788–799, Aug. 2021, number: 4. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8387128/

[4] L. A. Baumann, J. Baker, and A. G. Elshaug, "The impact of electronic health record systems on clinical documentation times: A systematic review," *Health Policy*, vol. 122, no. 8, pp. 827–836, Aug. 2018. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0168851018301635

[5] E. Joukes, A. Abu-Hanna, R. Cornet, and N. F. de Keizer, "Time Spent on Dedicated Patient Care and Documentation Tasks Before and After the Introduction of a Structured and Standardized Electronic Health Record," *Applied Clinical Informatics*, vol. 9, no. 1, pp. 46–53, Jan. 2018, number: 1. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5801881/

[6] S. Upadhyay and H.-f. Hu, "A Qualitative Analysis of the Impact of Electronic Health Records (EHR) on Healthcare Quality and Safety: Clinicians' Lived Experiences," *Health Services Insights*, vol. 15, p. 11786329211070722, Mar. 2022. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8902175/

[7] K. Singhal, S. Azizi, T. Tu, S. S. Mahdavi, J. Wei, H. W. Chung, N. Scales, A. Tanwani, H. Cole-Lewis, S. Pfohl, P. Payne, M. Seneviratne, P. Gamble, C. Kelly, A. Babiker, N. Schärli, A. Chowdhery, P. Mansfield, D. Demner-Fushman, B. Agüera y Arcas, D. Webster, G. S. Corrado, Y. Matias, K. Chou, J. Gottweis, N. Tomasev, Y. Liu, A. Rajkomar, J. Barral, C. Semturs, A. Karthikesalingam, and V. Natarajan, "Large language models encode clinical knowledge," *Nature*, vol. 620, no. 7972, pp. 172–180, Aug. 2023, number: 7972 Publisher: Nature Publishing Group. [Online]. Available: https://www.nature.com/articles/s41586-023-06291-2

[8] S. Bedi, Y. Liu, L. Orr-Ewing, D. Dash, S. Koyejo, A. Callahan, J. A. Fries, M. Wornow, A. Swaminathan, L. S. Lehmann, H. J. Hong, M. Kashyap, A. R. Chaurasia, N. R. Shah, K. Singh, T. Tazbaz, A. Milstein, M. A. Pfeffer, and N. H. Shah, "Testing and Evaluation of Health Care Applications of Large Language Models: A Systematic Review," *JAMA*, vol. 333, no. 4, pp. 319–328, Jan. 2025. [Online]. Available: https://doi.org/10.1001/jama.2024.21700

[9] M. Griot, C. Hemptinne, J. Vanderdonckt, and D. Yuksel, "Large Language Models lack essential metacognition for reliable medical reasoning," *Nature Communications*, vol. 16, no. 1, p. 642, Jan. 2025. [Online]. Available: https://doi.org/10.1038/s41467-024-55628-6

[10] R. O. Ness, K. Matton, H. Helm, S. Zhang, J. Bajwa, C. E. Priebe, and E. Horvitz, "MedFuzz: Exploring the Robustness of Large Language Models in Medical Question Answering," Sep. 2024, arXiv:2406.06573 [cs]. [Online]. Available: http://arxiv.org/abs/2406.06573

[11] D. Jin, E. Pan, N. Oufattole, W.-H. Weng, H. Fang, and P. Szolovits, "What Disease does this Patient Have? A Large-scale Open Domain Question Answering Dataset from Medical Exams," Sep. 2020, arXiv:2009.13081 [cs] version: 1. [Online]. Available: http://arxiv.org/abs/2009.13081

[12] A. Pal, L. K. Umapathi, and M. Sankarasubbu, "MedMCQA: A Large-scale Multi-Subject Multi-Choice Dataset for Medical domain Question Answering," in *Proceedings of the Conference on Health, Inference, and Learning*, ser. Proceedings of Machine Learning Research, G. Flores, G. H. Chen, T. Pollard, J. C. Ho, and T. Naumann, Eds., vol. 174. PMLR, Apr. 2022, pp. 248–260. [Online]. Available: https://proceedings.mlr.press/v174/pal22a.html

[13] Q. Jin, B. Dhingra, Z. Liu, W. Cohen, and X. Lu, "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)*. Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 2567–2577. [Online]. Available: https://aclanthology.org/D19-1259

[14] N. H. Shah, D. Entwistle, and M. A. Pfeffer, "Creation and Adoption of Large Language Models in Medicine," *JAMA*, vol. 330, no. 9, pp. 866–869, Sep. 2023. [Online]. Available: https://doi.org/10.1001/jama.2023.14217

[15] H. Decker, K. Trang, J. Ramirez, A. Colley, L. Pierce, M. Coleman, T. Bongiovanni, G. B. Melton, and E. Wick, "Large Language Model-Based Chatbot vs Surgeon-Generated Informed Consent Documentation for Common Procedures," *JAMA Network Open*, vol. 6, no. 10, p. e2336997, Oct. 2023, number: 10. [Online]. Available: https://doi.org/10.1001/jamanetworkopen.2023.36997

[16] J. Zaretsky, J. M. Kim, S. Baskharoun, Y. Zhao, J. Austrian, Y. Aphinyanaphongs, R. Gupta, S. B. Blecker, and J. Feldman, "Generative Artificial Intelligence to Transform Inpatient Discharge Summaries to Patient-Friendly Language and Format," *JAMA Network Open*, vol. 7, no. 3, p. e240357, Mar. 2024, number: 3. [Online]. Available: https://doi.org/10.1001/jamanetworkopen.2024.0357

[17] M. Tai-Seale, S. L. Baxter, F. Vaida, A. Walker, A. M. Sitapati, C. Osborne, J. Diaz, N. Desai, S. Webb, G. Polston, T. Helsten, E. Gross, J. Thackaberry, A. Mandvi, D. Lillie, S. Li, G. Gin, S. Achar, H. Hofflich, C. Sharp, M. Millen, and C. A. Longhurst, "AI-Generated Draft Replies Integrated Into Health Records and Physicians' Electronic Communication," *JAMA Network Open*, vol. 7, no. 4, pp. e246 565–e246 565, Apr. 2024. [Online]. Available: https://doi.org/10.1001/jamanetworkopen.2024.6565

[18] S. L. Baxter, C. A. Longhurst, M. Millen, A. M. Sitapati, and M. Tai-Seale, "Generative artificial intelligence responses to patient messages in the electronic health record: early lessons learned." *JAMIA open*, vol. 7, no. 2, p. ooae028, Jul. 2024, place: United States.

[19] L. Bednarczyk, D. Reichenpfader, C. Gaudet-Blavignac, A. K. Ette, J. Zaghir, Y. Zheng, A. Bensahla, M. Bjelogrlic, and C. Lovis, "Scientific Evidence for Clinical Text Summarization Using Large Language Models: Scoping Review," *J Med Internet Res 2025;27:e68998 https://www.jmir.org/2025/1/e68998*, May 2025, publisher: Journal of Medical Internet Research. [Online]. Available: https://www.jmir.org/2025/1/e68998

[20] E. Asgari, N. Montaña-Brown, M. Dubois, S. Khalil, J. Balloch, J. A. Yeung, and D. Pimenta, "A framework to assess clinical safety and hallucination rates of LLMs for medical text summarisation," *npj Digital Medicine*, vol. 8, no. 1, pp. 1–15, May 2025, publisher: Nature Publishing Group. [Online]. Available: https://www.nature.com/articles/s41746-025-01670-7

[21] S. Bedi, H. Cui, M. Fuentes, A. Unell, M. Wornow, J. M. Banda, N. Kotecha, T. Keyes, Y. Mai, M. Oez, H. Qiu, S. Jain, L. Schettini, M. Kashyap, J. A. Fries, A. Swaminathan, P. Chung, F. Nateghi, A. Aali, A. Nayak, S. Vedak, S. S. Jain, B. Patel, O. Fayanju, S. Shah, E. Goh, D.-h. Yao, B. Soetikno, E. Reis, S. Gatidis, V. Divi, R. Capasso, R. Saralkar, C.-C. Chiang, J. Jindal, T. Pham, F. Ghoddusi, S. Lin, A. S. Chiou, C. Hong, M. Roy, M. F. Gensheimer, H. Patel, K. Schulman, D. Dash, D. Char, L. Downing, F. Grolleau, K. Black, B. Mieso, A. Zahedivash, W.-w. Yim, H. Sharma, T. Lee, H. Kirsch, J. Lee, N. Ambers, C. Lugtu, A. Sharma, B. Mawji, A. Alekseyev, V. Zhou, V. Kakkar, J. Helzer, A. Revri, Y. Bannett, R. Daneshjou, J. Chen, E. Alsentzer, K. Morse, N. Ravi, N. Aghaeepour, V. Kennedy, A. Chaudhari, T. Wang, S. Koyejo, M. P. Lungren, E. Horvitz, P. Liang, M. Pfeffer, and N. H. Shah, "MedHELM: Holistic Evaluation of Large Language Models for Medical Tasks," Jun. 2025, arXiv:2505.23802 [cs]. [Online]. Available: http://arxiv.org/abs/2505.23802

[22] J. Balloch, S. Sridharan, G. Oldham, J. Wray, P. Gough, R. Robinson, N. J. Sebire, S. Khalil, E. Asgari, C. Tan, A. Taylor, and D. Pimenta, "Use of an ambient artificial intelligence tool to improve quality of clinical documentation," *Future Healthcare Journal*, vol. 11, no. 3, p. 100157,

2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2514664524015479

[23] C. D. Stults, S. Deng, M. C. Martinez, J. Wilcox, N. Szwerinski, K. H. Chen, S. Driscoll, J. Washburn, and V. G. Jones, "Evaluation of an Ambient Artificial Intelligence Documentation Platform for Clinicians." *JAMA network open*, vol. 8, no. 5, p. e258614, May 2025, place: United States.

[24] A. A. Tierney, G. Gayre, B. Hoberman, B. Mattern, M. Ballesca, S. B. Wilson Hannay, K. Castilla, C. S. Lau, P. Kipnis, V. Liu, and K. Lee, "Ambient Artificial Intelligence Scribes: Learnings after 1 Year and over 2.5 Million Uses," *NEJM Catalyst*, vol. 6, no. 5, p. CAT.25.0040, Mar. 2025, publisher: Massachusetts Medical Society. [Online]. Available: https://catalyst.nejm.org/doi/full/10.1056/CAT.25.0040

[25] S. J. Shah, T. Crowell, Y. Jeong, A. Devon-Sand, M. Smith, B. Yang, S. P. Ma, A. S. Liang, C. Delahaie, C. Hsia, T. Shanafelt, M. A. Pfeffer, C. Sharp, S. Lin, and P. Garcia, "Physician Perspectives on Ambient AI Scribes," *JAMA Network Open*, vol. 8, no. 3, p. e251904, Mar. 2025. [Online]. Available: https://doi.org/10.1001/jamanetworkopen.2025.1904

[26] M. Griot, J. Vanderdonckt, D. Yuksel, and C. Hemptinne, "Physician in the Loop Design of Interactive Agents," in *Engineering Interactive Computer Systems. EICS 2024 International Workshops*, L. Zaina, J. C. Campos, D. Spano, K. Luyten, P. Palanque, G. van der Veer, A. Ebert, S. R. Humayoun, and V. Memmesheimer, Eds. Cham: Springer Nature Switzerland, 2025, pp. 94–109.

[27] W. Kwon, Z. Li, S. Zhuang, Y. Sheng, L. Zheng, C. H. Yu, J. E. Gonzalez, H. Zhang, and I. Stoica, "Efficient Memory Management for Large Language Model Serving with PagedAttention," in *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*, 2023.

[28] A. Yang, A. Li, B. Yang, B. Zhang, B. Hui, B. Zheng, B. Yu, C. Gao, C. Huang, C. Lv, C. Zheng, D. Liu, F. Zhou, F. Huang, F. Hu, H. Ge, H. Wei, H. Lin, J. Tang, J. Yang, J. Tu, J. Zhang, J. Yang, J. Yang, J. Zhou, J. Zhou, J. Lin, K. Dang, K. Bao, K. Yang, L. Yu, L. Deng, M. Li, M. Xue, M. Li, P. Zhang, P. Wang, Q. Zhu, R. Men, R. Gao, S. Liu, S. Luo, T. Li, T. Tang, W. Yin, X. Ren, X. Wang, X. Zhang, X. Ren, Y. Fan, Y. Su, Y. Zhang, Y. Zhang, Y. Wan, Y. Liu, Z. Wang, Z. Cui, Z. Zhang, Z. Zhou, and Z. Qiu, "Qwen3 Technical Report," May 2025, arXiv:2505.09388 [cs]. [Online]. Available: http://arxiv.org/abs/2505.09388

[29] eHealth Platform, "eHealth Khmer Standard," 2025. [Online]. Available: https://www.ehealth.fgov.be/standards/kmehr/en

[30] D. S. Team, "Docling Technical Report," Tech. Rep., Aug. 2024, version Number: 1.0.0 _eprint: 2408.09869. [Online]. Available: https://arxiv.org/abs/2408.09869

[31] W. K. N.V., "UpToDate: Industry-leading clinical decision support," 2023. [Online]. Available: https://www.wolterskluwer.com/en/solutions/uptodate

[32] T. J. Baek, N. Vincent, and L. Kim, "Designing an open-source LLM interface and social platforms for collectively driven LLM evaluation and auditing," 2024.

[33] "General Data Protection Regulation (GDPR) – Official Legal Text." [Online]. Available: https://gdpr-info.eu/

[34] "Regulation (EU) 2017/745 of the European Parliament and of the Council of 5 April 2017 on medical devices, amending Directive 2001/83/EC, Regulation (EC) No 178/2002 and Regulation (EC) No 1223/2009 and repealing Council Directives 90/385/EEC and 93/42/EEC (Text with EEA relevance.)," May 2017, pages: 1-175.

[35] "Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act)," *OJ*, Jun. 2024.

[36] B. F. Gage, Y. Yan, P. E. Milligan, A. D. Waterman, R. Culverhouse, M. W. Rich, and M. J. Radford, "Clinical classification schemes for predicting hemorrhage: results from the National Registry of Atrial Fibrillation (NRAF)." *American heart journal*, vol. 151, no. 3, pp. 713–719, Mar. 2006, place: United States.