# Embedding Electronic Health Records into Large Language Models: Methodologies, Challenges, and a Forward Outlook

**Executive Summary:**

Integrating Electronic Health Records (EHRs) with Large Language Models (LLMs) represents a significant frontier in advancing healthcare analytics, clinical decision support, and personalized medicine.1 This endeavor involves transforming the rich, complex data within EHRs—spanning structured information like medical codes (ICD, LOINC, RxNorm) and lab values, to unstructured clinical notes—into formats that LLMs can effectively process and learn from.4

Key methodologies for this integration include **serializing EHR data** into textual representations, such as structured Markdown or concise summaries, and converting medical codes into human-readable natural language descriptions.7 General-purpose LLMs are increasingly utilized as powerful **EHR encoders**, often matching or exceeding the performance of specialized EHR foundation models, particularly in few-shot learning scenarios.7 Architectural adaptations, such as specific pooling strategies (e.g., last token pooling for decoder-only LLMs) and contrastive learning, are employed to refine embedding quality.7 Furthermore, specialized **medical tokenization** approaches like MedTok, which incorporate textual descriptions and relational context from biomedical ontologies using graph encoders, are proving crucial for enhancing LLM understanding of medical codes.12 For structured tabular EHR data, techniques are advancing beyond simple text serialization to include specialized models like StructLM for structured knowledge grounding, and methods like guided embedding refinement where LLMs generate embeddings based on interpretable clinical attributes.15

Despite these advancements, significant challenges persist. **Data heterogeneity** across EHR systems necessitates robust standardization and interoperability, with standards like HL7/FHIR and OMOP CDM being pivotal.4 **Patient data privacy and security** are paramount, requiring rigorous de-identification methods and the adoption of Privacy-Enhancing Technologies (PETs) such as federated learning and differential privacy.26 Mitigating **algorithmic bias** to ensure fairness and prevent the amplification of health disparities is an ongoing critical concern.37 Enhancing model **interpretability and explainability** through XAI methods like LIME, SHAP, and chain-of-thought reasoning is essential for clinical trust and adoption.42 Effectively managing **temporal dynamics** in longitudinal EHR data often requires specialized architectures like Temporal Dynamic Embedding (TDE) or the TEMPORALCROSS framework.49 Integrating **multimodal EHR data** (text, structured data, imaging) using advanced fusion techniques like those in the EMERGE or ColaCare frameworks is key to a holistic patient understanding.49

Innovations focus on AI-assisted **data harmonization** frameworks (e.g., ehrapy, MedRep) 20, leveraging **knowledge graphs** for enhanced contextual understanding (e.g., DR. KNOWS) 63, ensuring robust **model validation** through guidelines like TRIPOD-LLM 22, and developing strategies for **continuous learning** and adaptation using feedback mechanisms (e.g., EHR-CoAgent, ColaCare).57

The path forward requires navigating an evolving **regulatory landscape** (e.g., FDA guidance) and adhering to strong **ethical principles** (e.g., AMA principles) to ensure responsible deployment.73 Future breakthroughs are anticipated in more specialized medical LLMs, advanced multimodal integration, causal inference capabilities, and enhanced automation of clinical workflows, all necessitating sustained cross-disciplinary collaboration.81

## I. Introduction

### The Imperative of Integrating EHRs with LLMs for Advancing Healthcare

Electronic Health Records (EHRs) have become the cornerstone of modern healthcare, representing a systematized, digital collection of patient and population health information. These records, designed to be shared across diverse healthcare settings, harbor immense potential for driving advancements in clinical prediction, medical research, and the overall quality of patient care.1 The fundamental goal is to leverage the vast datasets contained within EHRs through sophisticated machine learning techniques, thereby enhancing clinical outcomes, optimizing healthcare processes, and potentially reducing associated costs.3

Concurrently, Large Language Models (LLMs) have emerged as a transformative technology, demonstrating remarkable capabilities in understanding, processing, and generating human language.4 Their advanced natural language processing prowess offers an unprecedented opportunity to unlock the rich, often complex, and voluminous information embedded within EHRs, particularly from unstructured textual components like clinical notes, which have traditionally been challenging for automated analysis.6 The process of embedding EHR data into LLMs is a critical technical step that enables these powerful models to ingest, interpret, and ultimately derive meaningful insights from clinical records. This report aims to provide a deep research background on the methodologies employed, the inherent challenges encountered, and the future prospects of this crucial integration.

### Overview of the Report's Scope

This report will conduct a comprehensive exploration of the multifaceted landscape of embedding EHRs into LLMs. The initial sections will establish a foundational understanding of the intrinsic characteristics of EHR data and the architectural underpinnings of LLMs relevant to this integration. Following this, the report will delve into a detailed examination of various methodologies for transforming and embedding diverse EHR data types—spanning structured, semi-structured, and unstructured information—into formats amenable to LLM processing.

A significant portion of this analysis will be dedicated to a critical review of the multifaceted challenges that arise in this domain. These challenges include, but are not limited to, data heterogeneity and the need for robust standardization, the paramount concerns of patient data privacy and security, the pervasive issue of algorithmic bias and the pursuit of fairness, the necessity for model interpretability and explainability in high-stakes clinical decision-making, and the complexities of managing temporal dynamics inherent in longitudinal patient data.

Finally, the report will discuss innovative solutions and emerging best practices designed to address these challenges. It will also touch upon the evolving regulatory environment shaping the deployment of AI in healthcare and outline promising future research avenues that could further advance the field of EHR-LLM integration.

### Transformative Potential vs. Navigational Complexity

The synergy between EHRs and LLMs heralds a new era of possibilities for healthcare. EHRs, as digital repositories, contain a wealth of information including demographics, comprehensive medical histories, medication lists, allergy information, immunization records, laboratory test results, radiological images, vital signs, and billing data.1 LLMs, on the other hand, built upon architectures like the Transformer, excel at discerning intricate patterns and contextual nuances within textual data.4 The integration of these two domains holds the transformative potential to revolutionize healthcare analytics, refine clinical decision support systems, and enable truly personalized medicine.7 For example, LLMs can process unstructured clinical notes to extract critical patient information that might be missed by structured data analysis alone, leading to more accurate risk stratification or earlier disease detection.

However, this promising endeavor is not without its significant hurdles. EHR data is notoriously complex, characterized by its heterogeneity, incompleteness, inherent noise, and the presence of various biases stemming from its primary collection for clinical and administrative purposes rather than research.1 LLMs, despite their power, also come with their own set of limitations, including the potential for generating factually incorrect information (hallucinations), reflecting biases present in their training data, and demanding substantial computational resources.5

Successfully navigating this intricate web of potential and pitfalls is paramount. The process of embedding EHRs into LLMs is therefore not merely a technical conversion but a strategic undertaking that demands meticulous attention to data preparation, sophisticated model adaptation techniques, and robust ethical safeguards. The ultimate success of EHR-LLM integration will depend not only on continued technical innovation in embedding methodologies but also on the concurrent development and implementation of comprehensive frameworks for data governance, stringent ethical oversight, and continuous model validation. These measures are indispensable to ensure patient safety, promote equitable outcomes, and build trust in AI-driven healthcare solutions.

## II. Foundational Understanding: EHRs and LLMs

### A. The Landscape of Electronic Health Records

#### Core characteristics, data types (structured vs. unstructured), and common formats

Electronic Health Records (EHRs) are the digital backbone of modern healthcare, providing a systematized and longitudinal collection of patient health information that can be accessed and shared across different healthcare settings.1 An EHR typically encompasses a wide array of data, including patient demographics (age, gender, contact details, insurance information), detailed medical history (past illnesses, surgeries, family medical history, lifestyle factors), medication and allergy records, immunization status, laboratory test results, radiology images (X-rays, CT scans, MRIs), vital signs (blood pressure, heart rate, temperature), personal statistics (weight, height), and billing information.1

A fundamental distinction within EHR data lies in its structure:

* **Structured Data:** This category includes discrete, well-organized data points that are easily machine-readable and typically stored in relational databases or standardized formats like Comma Separated Values (CSV).10 Examples include patient demographics, coded diagnoses (e.g., using the International Classification of Diseases, ICD-10), medication lists (often coded with RxNorm), laboratory results (which may use Logical Observation Identifiers Names and Codes, LOINC), and procedure codes (e.g., Current Procedural Terminology, CPT).8 Order entry for tests and medications also generates structured data.8
* **Unstructured Data:** This comprises free-form information that does not adhere to a predefined data model, making it richer in context but more challenging for traditional computational analysis. Prime examples include clinical notes (such as progress notes, consultation reports, discharge summaries, and Subjective, Objective, Assessment, and Plan (SOAP) notes), narrative radiology and pathology reports.1 While imaging data like X-rays have associated structured metadata, the images themselves are unstructured visual data.1

Understanding this structured/unstructured dichotomy is critical when considering LLM integration. LLMs are inherently proficient at processing and understanding unstructured text. However, to leverage the valuable information contained in structured EHR fields, specific transformation and embedding strategies are required. The overarching goal is to create a unified, coherent representation of the patient's health status that an LLM can effectively process, drawing insights from both structured codes and narrative text.

#### Inherent complexities: heterogeneity, incompleteness, noise, and the critical role of data standards (HL7/FHIR, OMOP, SNOMED CT, LOINC)

EHR data, despite its richness, is fraught with inherent complexities that pose significant challenges for data analysis and model development. These complexities include:

* **Heterogeneity:** Data often originates from disparate systems with non-standardized formats and varying data exchange protocols, leading to inconsistencies in how information is represented.2 Different institutions may use different local coding systems or versions of standard terminologies.
* **Incompleteness:** Missing data is a pervasive issue in EHRs. Data may be missing completely at random (MCAR), missing at random (MAR), or missing not at random (MNAR), each type requiring different handling strategies.2 For instance, a lab test not being recorded could mean it wasn't performed, the result was normal and thus not explicitly noted, or the data entry was simply missed.
* **Noise and Inaccuracies:** Errors in data entry, inconsistencies in coding practices, and even contradictory information (e.g., measurements reported for deceased patients) contribute to data noise.10 The primary use of EHRs for clinical care and billing, rather than curated research, often means data quality can be variable.10
* **Bias:** EHR data can reflect and perpetuate various forms of bias, including selection bias (non-representative patient samples), filtering bias (inconsistent data inclusion criteria), surveillance bias (differential monitoring frequencies leading to exaggerated associations), and coding biases (inaccuracies in recording or inconsistencies in coding).10

To mitigate these complexities, data standards play a crucial role in promoting interoperability and data quality. Key standards include:

* **Health Level Seven (HL7):** A set of international standards for the exchange, integration, sharing, and retrieval of electronic health information.1
* **Fast Healthcare Interoperability Resources (FHIR):** A newer HL7 standard designed for modern web-based data exchange, promoting easier and faster interoperability, even on mobile devices.8
* **Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM):** A standard that transforms data contained within EHRs and other observational healthcare databases into a common format, vocabulary, and structure, facilitating collaborative research across different datasets.10
* **Systematized Nomenclature of Medicine – Clinical Terms (SNOMED CT):** A comprehensive, multilingual clinical healthcare terminology used for consistent and precise representation of clinical information.15
* **Logical Observation Identifiers Names and Codes (LOINC):** A standard for identifying laboratory tests and observations.15

The "messiness" of real-world clinical data presents both a challenge and an opportunity for LLMs. While LLMs, especially those pre-trained on vast and diverse text corpora, exhibit a degree of robustness to linguistic variation and some types of noise 2, the specific structural and semantic complexities of EHRs can still degrade their performance or lead to skewed interpretations if not carefully managed.22 For example, an LLM might misinterpret an ambiguously written clinical note or draw incorrect inferences from inconsistently coded data. Thus, while LLMs may handle certain linguistic ambiguities better than older NLP models, specialized embedding and fine-tuning strategies that explicitly account for EHR-specific noise, bias, and structural heterogeneity are essential. Standardization efforts through models like OMOP CDM and protocols like FHIR are therefore not just beneficial but become even more critical, as they provide a cleaner, more consistent input format, reducing the inherent noise that LLMs must otherwise contend with. This implies that the journey towards effective EHR-LLM integration is not solely about advancing LLM capabilities but also about concurrently improving the quality and standardization of the underlying EHR data itself.

### B. Large Language Models in the Medical Context

#### Fundamental architecture (Transformer, attention mechanisms) and data processing capabilities

Large Language Models, which are at the forefront of artificial intelligence advancements, are predominantly built upon the Transformer architecture.4 This architecture, a departure from earlier recurrent and convolutional neural network designs, relies heavily on **self-attention mechanisms**. Self-attention allows the model to weigh the importance of different words (or tokens) in an input sequence relative to each other, capturing long-range dependencies and complex contextual relationships within the text.4

Key architectural components of a Transformer include:

* **Input Embeddings:** Input text is first tokenized (broken into smaller units like words or sub-words), and each token is converted into a dense vector representation (embedding).
* **Positional Encoding:** Since self-attention processes tokens in parallel, positional encodings are added to the input embeddings to provide the model with information about the order or position of tokens in the sequence.4
* **Multi-Head Attention:** This mechanism allows the model to jointly attend to information from different representation subspaces at different positions. Essentially, it runs multiple self-attention operations ("heads") in parallel and concatenates their outputs, enabling a richer understanding of the input.4
* **Feed-Forward Networks:** Each attention output passes through a position-wise feed-forward network, which consists of fully connected layers applied independently to each position.
* **Layer Normalization and Residual Connections:** These are used throughout the network to stabilize training and enable deeper architectures.

Transformers can be structured in different ways:

* **Encoder-Decoder Models:** Originally proposed for machine translation, these models have an encoder to process the input sequence and a decoder to generate the output sequence.4
* **Encoder-Only Models:** Examples like BERT (Bidirectional Encoder Representations from Transformers) are designed for language understanding tasks. They process the entire input sequence at once, allowing attention to flow in both directions (bidirectionally).4
* **Decoder-Only Models:** Examples like the GPT (Generative Pre-trained Transformer) series are optimized for language generation tasks. They process input in an auto-regressive manner, predicting the next token based on preceding tokens, typically using masked self-attention to prevent attending to future tokens.4

While LLMs are primarily designed for sequential text data, adaptations are being developed to handle structured data. For instance, specific prompting techniques like "Tabular Chain-of-Thought" (Tabular CoT) guide LLMs to reason over data presented in tables.4 Furthermore, models such as StructLM are being explicitly designed and trained for Structured Knowledge Grounding (SKG), enabling them to interpret and utilize information from tables, graphs, and databases by performing tasks like data-to-text generation, table-based question answering, and SQL generation.24

#### Learning paradigms (pre-training, fine-tuning, prompting) and the impact of scaling laws

The remarkable capabilities of LLMs stem from sophisticated learning paradigms:

* **Pre-training:** This is the initial and most computationally intensive phase. LLMs are pre-trained on massive and diverse unlabeled text corpora (often including books, articles, websites, and code).2 During pre-training, models learn general language representations, grammatical structures, factual knowledge, and some reasoning abilities through self-supervised learning objectives like masked language modeling (predicting masked tokens, as in BERT) or next-token prediction (as in GPT).
* **Fine-tuning:** After pre-training, LLMs can be adapted to specific downstream tasks or domains through fine-tuning on smaller, labeled datasets relevant to the target application.4 For medical applications, this might involve fine-tuning on medical literature, clinical guidelines, or EHR-derived narratives. **Instruction tuning** is a particularly effective fine-tuning method where the model is trained on examples of instructions and their desired outputs, enhancing its ability to follow commands and perform specific tasks.4 Parameter-Efficient Fine-Tuning (PEFT) techniques like LoRA (Low-Rank Adaptation) allow for adapting large models with significantly fewer trainable parameters, reducing computational costs.4
* **Prompting (In-Context Learning):** This involves guiding the LLM's behavior at inference time by carefully crafting the input query or "prompt".4 Different prompting strategies exist:
  + **Zero-shot prompting:** The model performs a task based on instructions alone, without any examples.4
  + **One-shot prompting:** A single example of the task is provided in the prompt.4
  + **Few-shot prompting:** A few examples are included in the prompt to help the model understand the desired output format and task.4
  + **Chain-of-Thought (CoT) prompting:** The model is guided to generate a step-by-step reasoning process before arriving at the final answer, which is particularly effective for complex reasoning tasks.4

The development and performance of LLMs are significantly influenced by **scaling laws**. These are empirical observations demonstrating that model performance (e.g., on benchmark tasks) improves predictably as key factors are increased:

* **Model Size:** The number of parameters in the model. Larger models with more parameters can capture more complex patterns.4
* **Dataset Size:** The quantity of data used for pre-training. Training on larger and more diverse datasets exposes the model to a broader range of language uses and knowledge.4
* **Compute Resources:** The amount of computational power used for training. Training larger models on vast datasets requires substantial compute.4 Scaling laws suggest that as these factors increase, there are often power-law improvements in model capabilities. However, this scaling comes with diminishing returns at very large scales and significantly increased computational costs and environmental impact.4 Scaling laws have also been observed in the context of EHR foundation models, suggesting that similar principles of scaling model size, EHR training datasets, and computational resources can lead to predictable performance improvements in the medical domain.31

The application of LLMs in medicine presents a fundamental tension between leveraging their broad generalization capabilities and the need for deep, domain-specific expertise. General-purpose LLMs, pre-trained on vast, diverse corpora, possess extensive world knowledge and can adapt to new domains with relatively little task-specific data through few-shot learning or prompting.2 This is advantageous for rapidly prototyping applications. However, medical language is highly specialized, filled with jargon, abbreviations, and nuanced contextual meanings that general LLMs may misinterpret.11 Furthermore, the stakes in medical applications are exceptionally high, where factual inaccuracies or "hallucinations" by the LLM can have severe consequences for patient safety.23

To address this, researchers often resort to fine-tuning general LLMs on medical data or developing medical-specific LLMs (e.g., GatorTron 7) through continued pre-training on large biomedical corpora. This specialization aims to instill the necessary domain knowledge and adapt the LLM's language understanding to clinical contexts. Nevertheless, this creates a dilemma: extensive specialization might curtail the model's general reasoning abilities or demand massive, high-quality, curated medical datasets, which are often difficult and expensive to acquire due to privacy and data access restrictions. Conversely, under-specialization risks suboptimal performance and potential safety hazards. Current research often indicates a preference for hybrid approaches, such as using general-purpose LLMs as powerful encoders for EHR data that has been transformed into a textual format, and then fine-tuning these models for specific clinical tasks or using sophisticated prompting strategies.2 This suggests that the optimal path for EHR-LLM integration involves a careful calibration: starting with the robust text understanding capabilities of general pre-trained LLMs but then employing targeted adaptation techniques—be it domain-adaptive pre-training, specialized fine-tuning, or advanced prompt engineering—that are meticulously tailored to the complexities and high-stakes nature of medical data. The specific choice of strategy will invariably depend on the clinical task at hand, the availability and quality of domain-specific data, and the computational resources available.

## III. Bridging the Gap: Methodologies for Embedding EHR Data into LLMs

Integrating EHR data with LLMs requires transforming diverse and complex medical information into numerical representations, or embeddings, that these models can effectively process. This section details the primary methodologies employed for this transformation, covering data serialization, the use of general-purpose LLMs as encoders, strategies for handling structured data, and performance benchmarks.

### A. Transforming EHR Data for LLM Consumption

#### Serialization techniques (e.g., structured Markdown, text summaries)

A foundational step in preparing EHR data for LLM processing is its conversion into a textual format. Given that LLMs are primarily designed to operate on natural language text, structured and semi-structured EHR components must be serialized. One effective method involves converting patient records into **structured Markdown text**.2 This approach preserves a degree of the original data's organization (e.g., using headings for different sections like diagnoses, medications, lab results) while rendering it in a plain text format that LLMs can parse. For instance, a patient's medication list could be formatted as a Markdown list, and lab results as key-value pairs or small tables within the Markdown structure.

Another common strategy is the generation of **concise text summaries** that encapsulate key patient information relevant at the time of a specific prediction or analysis.2 These summaries aim to distill the most pertinent aspects of a patient's record into a narrative or a structured textual representation. The objective of these serialization techniques is to create a textual input that is both comprehensive enough to retain critical clinical details and concise enough to be efficiently processed by LLMs, especially considering their context window limitations.34 This transformation is fundamental because it allows the LLM to apply its powerful semantic understanding capabilities, learned from vast general text corpora, to the clinical domain.

#### Conversion of medical codes (ICD, LOINC, RxNorm) to human-readable natural language descriptors

Structured data within EHRs frequently consists of standardized medical codes, such as ICD codes for diagnoses, LOINC codes for laboratory tests, and RxNorm codes for medications. To make these codes comprehensible to general-purpose LLMs that may not have been specifically pre-trained on medical terminologies, a common and crucial step is their conversion into **human-readable natural language descriptions**.2 For example, the ICD-10 code "I10" would be replaced by its textual description, "Essential (primary) hypertension." Similarly, a LOINC code for a specific lab test would be replaced by the name of that test, and an RxNorm code by the name of the medication, potentially including dosage and form.

This conversion is vital because it allows the LLM to leverage its pre-existing knowledge of natural language to understand the semantic meaning of these clinical concepts, rather than treating the codes as opaque or arbitrary symbols. This step significantly enhances the ability of general-purpose LLMs to interpret structured clinical data and integrate it with information from unstructured notes. For instance, a study comparing LLM classification of EHR terms with clinical experts found high agreement when terms were presented in natural language, demonstrating the LLM's capacity to understand these descriptors.6

### B. LLM-based Embedding Strategies

#### Utilizing general-purpose LLMs (e.g., GTE-Qwen2, LLM2Vec-Llama) as EHR encoders

A promising direction in EHR-LLM integration is the use of general-purpose LLM embedding models to encode the serialized EHR data into high-dimensional vector representations (embeddings).2 Models such as GTE-Qwen2-7B-Instruct and LLM2Vec-Llama3.1-8B-Instruct have been employed for this purpose.2 Typically, the hidden states from one or more layers of these LLMs, generated after processing the textualized EHR input, are used as the final embeddings for downstream tasks like clinical prediction.

The rationale behind this approach is to harness the extensive semantic understanding and generalization capabilities that these LLMs have acquired from pre-training on vast, diverse public corpora of text and code.2 This strategy can potentially bypass the need for large, proprietary medical datasets specifically for training the EHR encoder model itself, making the approach more scalable and adaptable. The pre-trained knowledge allows the LLM to capture nuanced relationships and meanings within the textualized EHR data, even for medical concepts it may not have encountered extensively during its initial training.

#### Architectural adaptations and learning techniques (e.g., contrastive learning, pooling strategies)

To transform general-purpose LLMs, particularly decoder-only models (like GPT variants), into effective embedding generators, specific architectural modifications and learning techniques are often necessary.2 Decoder-only LLMs are primarily optimized for next-token prediction (text generation) and their native hidden states may not directly yield high-quality, discriminative embeddings for entire sequences. A common issue is **anisotropy**, where embeddings tend to occupy a narrow cone in the vector space, limiting their usefulness for similarity comparisons or classification tasks.35

Several strategies are employed to adapt these models:

* **Pooling Strategies:** To derive a single fixed-size embedding for an input sequence from the variable-length hidden states of an LLM, various pooling methods are used. These include simple mean pooling or max pooling over the token embeddings in the last layer. For decoder-only LLMs, "last token pooling" is often considered, as the final token's hidden state has theoretically processed the entire sequence. However, this requires careful intervention, such as using special prompts or fine-tuning to encourage the model to consolidate the sequence's semantics into this last token's representation.35 Other approaches include weighted mean pooling (giving more weight to later tokens in causal models) or partial pooling over specific ranges of hidden states.
* **Attention Mechanism Modifications:** For decoder-only LLMs, the native causal attention (where tokens only attend to previous tokens) can be a limitation for creating holistic embeddings. Some approaches involve converting causal attention to bi-directional attention during a fine-tuning phase, or using dynamic transformation mechanisms that allow the model to switch between attention types.35
* **Additional Projectors:** Linear projection layers can be added after the pooling layer to transform the embeddings, for example, to reduce dimensionality (improving storage and inference efficiency) or to convert dense embeddings into sparse representations (which can be beneficial for tasks like long document retrieval).35
* **Contrastive Learning:** This is a powerful technique used to fine-tune the embedding space. It trains the model to pull embeddings of semantically similar inputs closer together while pushing dissimilar ones further apart.2 For EHR data, this might involve treating different textual representations of the same clinical concept or patient state as positive pairs.
* **Parameter-Efficient Fine-Tuning (PEFT):** Techniques like LoRA (Low-Rank Adaptation) are employed to fine-tune these large embedding models efficiently by updating only a small subset of the model's parameters, significantly reducing computational and memory demands.4

The necessity for these adaptations underscores that deriving high-quality EHR embeddings from general-purpose decoder-only LLMs is a non-trivial engineering challenge. It is not as simple as taking the hidden states of an off-the-shelf model. The choice of the base LLM architecture (encoder-only, decoder-only, or encoder-decoder) significantly influences the ease and effectiveness of deriving embeddings. Encoder-based models like BERT, for example, often use the embedding of a special `` token as a sentence-level representation, a feature not inherently present in decoder-only models. Thus, specialized adaptation and fine-tuning are key to unlocking the full potential of LLMs as EHR encoders.

#### Specialized medical tokenization approaches (e.g., MedTok, BPE, WordPiece)

Tokenization, the process of breaking down input text into smaller units (tokens) that the LLM can process, is a critical preliminary step.27 The quality of tokenization directly impacts the LLM's ability to understand the input and preserve its semantic meaning. Standard tokenization algorithms like Byte Pair Encoding (BPE) or WordPiece, commonly used in general-purpose LLMs, may not be optimal for medical text, particularly for structured medical codes.15 These general tokenizers might treat medical codes as sequences of arbitrary characters or split them into sub-word units that lack clinical meaning, thereby losing critical information embedded in the code's structure or its relation to medical ontologies.15

To address this, specialized medical tokenizers are being developed. **MedTok** is a notable example of a multimodal medical code tokenizer.15 It moves beyond treating codes as mere text strings by incorporating:

1. **Textual Descriptions:** It uses a language model encoder to process the natural language descriptions associated with medical codes.
2. **Relational Context:** It employs a graph encoder to capture the relationships of codes within biomedical ontologies and knowledge graphs (e.g., hierarchical relationships, co-occurrences, drug-treatment associations). MedTok then quantizes information from both these modalities into a unified token space and optimizes these token representations for expressivity, aiming to capture hierarchical relationships and semantic equivalence across different coding systems.

The importance of such specialized tokenizers is underscored by performance improvements. Studies have shown that replacing standard EHR tokenizers with MedTok can lead to significant gains in Area Under the Precision-Recall Curve (AUPRC) across various EHR models and datasets (MIMIC-III, MIMIC-IV, EHRShot), with particularly notable improvements in tasks like drug recommendation.15 This demonstrates that providing LLMs with more semantically rich and contextually aware input tokens, especially for structured coded data, is crucial for enhancing their performance in the medical domain.

### C. Addressing Structured and Tabular EHR Data

While serializing all EHR data into a single textual stream is a viable strategy, LLMs' inherent proficiency in directly interpreting and utilizing highly structured tabular data remains an area of active research and development.24 Simply converting complex tables into very long text sequences might not be optimal, as LLMs can struggle with extremely long contexts or may fail to capture the inherent relational structure of tabular data effectively.24 Consequently, various techniques beyond simple text serialization are being explored:

* Specialized Models for Structured Data (e.g., StructLM):  
  The StructLM series of models, based on architectures like Code-LLaMA, represents a dedicated effort to enhance LLMs' capabilities for Structured Knowledge Grounding (SKG).24 These models are trained using comprehensive instruction tuning datasets (e.g., 1.1 million examples for StructLM) that specifically target tasks involving structured data. Such tasks include data-to-text generation (summarizing tables in natural language), table-based question answering, engaging in knowledge-grounded conversations that refer to tabular content, fact verification against tables, generating SQL queries to interact with databases, and performing mathematical reasoning over tabular entries. Research with StructLM has indicated that pre-training on code is particularly effective for developing strong SKG reasoning abilities.
* Table-to-Text Conversion and LLM Embeddings for Numerical EHR Data:  
  For numerical EHR data, such as vital signs and laboratory test results, a common approach involves transforming these raw features into structured textual queries that LLMs can process.43 Various formats for this table-to-text conversion are explored, including:
  + **NARRATIVES:** Continuous text descriptions of patient data, aiming for readability similar to clinical notes.
  + **JSON:** Hierarchical structuring of data, facilitating programmatic parsing.
  + **HTML:** Using web-based structures and tags.
  + **MARKDOWN:** Lightweight markup for formatted yet readable plain text. Once the tabular data is textualized, LLM embedding models (e.g., Mistral, Meditron, Llama3-8b) can generate embeddings from these representations, typically using pooling strategies like max or mean pooling. These embeddings then serve as feature inputs for traditional machine learning classifiers, such as XGBoost, for downstream tasks like diagnosis prediction, length of stay estimation, or mortality prediction.43 Studies show that this approach can yield performance comparable to using raw numerical data features directly with ML models for some tasks, although performance gaps may persist, particularly for effectively representing time-varying features.
* Guided Embedding Refinement:  
  This innovative technique proposes using LLMs as auxiliary tools to generate "guided embeddings".45 These embeddings are designed to capture domain-relevant semantic information based on predefined, interpretable attributes. In the EHR context, such attributes could include "severity of disease," "risk of readmission," or "patient adherence." The LLM would analyze the EHR data (potentially already partially processed or textualized) and score the patient or clinical event against these attributes. These guided embeddings can then be combined with base EHR embeddings (derived from other methods) to create "refined EHR embeddings," aiming for both improved performance and enhanced interpretability in clinical prediction tasks.
* Ontology-Enhanced Concept Representations (e.g., MedRep):  
  The MedRep strategy focuses on creating robust, vocabulary-agnostic representations for medical concepts within the OMOP Common Data Model framework.15 It employs a two-pronged approach:
  1. **LLM-Prompted Enrichment:** LLM prompts are used to generate rich, detailed descriptions for OMOP concepts, expanding on their often minimal standard definitions.
  2. **Graph Ontology Enhancement:** These LLM-generated text-based representations are then further enhanced by integrating relational knowledge from the OMOP vocabulary's graph ontology. This is achieved using graph contrastive learning techniques (e.g., the GRACE framework) combined with a "Learning without Forgetting" mechanism to preserve the richness of the initial text-based features while incorporating structural information.

The development of these diverse methods suggests that the future of integrating structured EHR data with LLMs lies in hybrid approaches. Rather than relying on LLMs to passively process long, serialized strings of tabular data, these models are increasingly being positioned as active participants in a more complex data processing and embedding pipeline. This may involve LLMs working in concert with specialized tabular data encoders, graph neural networks for leveraging ontological structures, or other techniques that can explicitly capture the relational nature of structured information. The emphasis is shifting towards creating meaningful, context-rich embeddings that fuse the semantic understanding capabilities of LLMs with the structural insights derived from other specialized methods, ultimately leading to more powerful and nuanced representations of patient data.

### D. Performance Benchmarking

A critical aspect of developing and refining EHR embedding methodologies for LLMs is rigorous performance benchmarking. This involves comparing the efficacy of LLM-based embeddings against established specialized EHR foundation models and traditional machine learning baselines across a range of clinically relevant tasks.

#### Comparative analysis of LLM-based embeddings versus specialized EHR foundation models (e.g., CLMBR-T-Base)

Recent studies have systematically evaluated the performance of general-purpose LLM-based embeddings against models specifically designed and pre-trained for EHR data, such as CLMBR-T-Base (a BERT-like model pre-trained on sequences of medical codes).2 These comparisons often utilize standardized benchmarks like EHRSHOT, which comprises diverse clinical prediction tasks.

The findings from these studies are noteworthy: general-purpose LLM embeddings, derived from models like GTE-Qwen2-7B and LLM2Vec-Llama-3.1-8B after serializing EHR data into text, frequently match or even surpass the performance of specialized EHR foundation models.2 This competitive performance is observed across various task categories, including predicting operational outcomes (e.g., in-hospital mortality), anticipating laboratory test results, assigning new diagnoses, and anticipating findings from chest X-rays.

Particularly compelling is the strong performance of these general-purpose LLM embeddings in **few-shot settings**.2 This indicates that the extensive knowledge and generalization capabilities acquired by LLMs during their pre-training on vast public corpora can be effectively transferred to the medical domain, even with limited task-specific examples. The effectiveness of these LLM-based embeddings often scales with the size of the underlying LLM and the length of the context window it can process, suggesting that larger models with greater capacity can capture more complex patterns from the textualized EHR data.

Furthermore, some research indicates that combining LLM-based embeddings with specialized EHR foundation models can yield even better performance, hinting at complementary strengths.9 For instance, an ensemble or fusion of features from a general LLM encoder and an EHR-specific encoder might capture both broad semantic understanding and domain-specific coding patterns.

The table below, synthesized from recent research 2, provides a comparative overview of different model types on various clinical prediction tasks, typically evaluated using the Area Under the Receiver Operating Characteristic curve (AUROC).

**Table 1: Comparative Performance of LLM-Embedding Models vs. Specialized EHR Foundation Models and Other Baselines on Clinical Prediction Tasks (Illustrative AUROC Scores)**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Model Name** | **Model Type** | **Operational Outcomes** | **Anticipating Lab Test Results** | **Assignment of New Diagnosis** | **Anticipating Chest X-ray Findings** | **Macro Avg. Across Task Groups** | **Source Snippet(s)** |
| GTE-Qwen2-7B | LLM-Embedding Model | 0.771 (.747-.795) | 0.865 (.858-.873) | 0.716 (.675-.757) | 0.666 (.653-.680) | 0.755 (.730-.780) | 2 |
| GTE-Qwen2-1.5B | LLM-Embedding Model | 0.755 (.729-.780) | 0.865 (.859-.872) | 0.700 (.657-.743) | 0.670 (.658-.683) | 0.748 (.722-.774) | 2 |
| LLM2Vec-Llama-3.1-8B | LLM-Embedding Model | 0.753 (.727-.778) | 0.778 (.768-.789) | 0.728 (.682-.774) | 0.678 (.665-.692) | 0.734 (.707-.762) | 2 |
| LLM2Vec-Llama 2 1.3B | LLM-Embedding Model | 0.706 (.680-.731) | 0.649 (.638-.660) | 0.636 (.598-.673) | 0.615 (.603-.627) | 0.651 (.627-.675) | 2 |
| GTE-Qwen2-7B + CLMBR-T-Base | LLM-Embedding Model + EHR Foundation Model | 0.882 (.863-.900) | 0.887 (.881-.893) | 0.725 (.682-.768) | 0.711 (.699-.723) | 0.801 (.777-.826) | 9 |
| CLMBR-T-Base | EHR Foundation Model (Baseline) | 0.824 (.803-.845) | 0.832 (.824-.840) | 0.707 (.667-.746) | 0.713 (.702-.724) | 0.769 (.746-.792) | 9 |
| Counts-based + GBM | Traditional ML (Baseline) | 0.774 (.752-.797) | 0.728 (.716-.741) | 0.719 (.669-.768) | 0.656 (.641-.671) | 0.719 (.691-.748) | 9 |
| DeBERTaV3 large | Encoder Language Model (Baseline) | 0.725 (.699-.752) | 0.712 (.700-.724) | 0.671 (.625-.716) | 0.625 (.612-.639) | 0.683 (.656-.711) | 9 |

*Note: Scores are presented as AUROC with 95% confidence intervals in parentheses, based on data from 9 and 2/.2 Some model versions or specific task results might vary slightly between different pre-print versions or final publications. The table aims to be illustrative of general trends.*

This comparative data is invaluable for researchers and developers. It demonstrates that leveraging general-purpose LLMs for EHR encoding is a highly viable and often superior strategy compared to relying solely on models trained on structured medical codes. The ability of these LLMs to understand natural language descriptions of codes and patient narratives allows them to capture richer semantic context, which translates to improved performance on complex clinical prediction tasks. This informs decisions regarding model selection, resource allocation for fine-tuning, and the overall architectural design for systems integrating EHRs with LLMs.

## IV. Navigating Critical Challenges in EHR-LLM Integration

The integration of Electronic Health Records (EHRs) with Large Language Models (LLMs) is a complex endeavor, marked by a series of critical challenges that span data characteristics, privacy and security imperatives, ethical considerations of bias and fairness, the need for model transparency, and the intricacies of handling dynamic, multimodal patient information. Addressing these challenges comprehensively is essential for the responsible and effective deployment of LLMs in healthcare.

### A. Data Heterogeneity and the Imperative for Standardization and Interoperability

A primary obstacle in leveraging EHR data for LLM applications is its inherent **heterogeneity**.1 EHR systems across different healthcare institutions, and even within the same institution over time, often utilize varied data formats, coding systems (e.g., ICD-9 vs. ICD-10, local codes), and data collection practices. This results in datasets that are difficult to aggregate and compare, posing significant hurdles for training generalizable LLMs and creating uniform embeddings. The lack of consistent data representation can lead to LLMs misinterpreting data or failing to recognize semantically similar concepts presented differently.

Recognizing this challenge, the healthcare informatics community has long championed **data standardization and interoperability** through initiatives like:

* **HL7 Standards (including FHIR):** Health Level Seven (HL7) provides a suite of standards for exchanging clinical and administrative data. Fast Healthcare Interoperability Resources (FHIR) is a particularly important modern HL7 standard that defines data elements (Resources) and an API for data exchange, leveraging web technologies to facilitate easier integration across systems.1 FHIR's modular structure and use of common web standards are intended to simplify the development of applications that can access and use EHR data from diverse systems.
* **OMOP Common Data Model (CDM):** The Observational Medical Outcomes Partnership (OMOP) CDM standardizes the structure and content of observational healthcare databases, including EHRs, by mapping source data to a common set of terminologies (e.g., SNOMED CT for conditions, RxNorm for drugs, LOINC for labs) and a common relational schema.10 This allows for consistent querying and analysis across disparate datasets. Initiatives like OMOP-on-FHIR aim to bridge these two standards, using FHIR for real-time data exchange and transforming this data into the OMOP CDM format for research and analytics.18
* **Controlled Vocabularies (SNOMED CT, LOINC, RxNorm):** These terminologies provide standardized codes and descriptions for clinical concepts, ensuring that diagnoses, laboratory tests, and medications are represented consistently, which is crucial for accurate data interpretation by both humans and machines.15

While these standards provide a foundational layer for harmonization, the process of mapping local data to these standards is itself complex and can involve information loss or the introduction of ambiguities.19 Interestingly, LLMs are now emerging as tools that can **assist in the data harmonization process itself**. Their natural language understanding capabilities can be leveraged to help map local terminologies to standard ontologies, translate between different coding systems, or even validate the quality of existing mappings.21 This creates a potentially synergistic relationship: better-standardized data improves LLM performance, and LLMs, in turn, can contribute to improving data standardization. This suggests that standardization is not merely a static prerequisite for EHR-LLM integration but an ongoing, dynamic process that can be significantly amplified by the capabilities of AI. The future may see LLMs as active participants in data quality improvement pipelines, facilitating more seamless and reliable large-scale clinical data analytics.

### B. Safeguarding Patient Data: Privacy, Security, and Compliance (HIPAA, GDPR)

EHRs are repositories of highly sensitive Protected Health Information (PHI), encompassing intimate details of individuals' health and lives. Consequently, ensuring the **privacy and security** of this data is not just a technical requirement but a profound ethical and legal obligation.1 Strict compliance with data protection regulations such as the Health Insurance Portability and Accountability Act (HIPAA) in the United States and the General Data Protection Regulation (GDPR) in Europe is mandatory when handling EHR data.21

The use of LLMs with EHR data introduces specific privacy risks. LLMs, particularly large generative models, have been shown to "memorize" portions of their training data, potentially leading to the inadvertent leakage of sensitive patient information if trained directly on identifiable EHRs.32 Even if direct identifiers are removed, LLMs might infer sensitive attributes or re-identify individuals through correlational patterns in the data.

Several strategies are employed to mitigate these risks:

* **De-identification:** This is a fundamental step involving the removal or obscuring of explicit PII (e.g., names, addresses, dates of birth, medical record numbers) from EHR text before it is processed by LLMs.51 Common de-identification methods include the HIPAA Safe Harbor method (removing a specific list of 18 identifiers) or expert determination (statistical assessment of re-identification risk). However, de-identification is challenging due to the "semantic gap"—subtle ways in which identity can be inferred through vocabulary mismatches (e.g., "RA" for "rheumatoid arthritis" might be missed), granularity mismatches (e.g., "acute tubular necrosis" implying "acute renal failure"), or implication mismatches (e.g., "amlodipine" implying "hypertension").51 Furthermore, aggressive de-identification can inadvertently remove medically relevant information (e.g., clinical eponyms like "Parkinson's disease") or struggle with out-of-distribution PII not commonly seen.57 Recent approaches like "locally augmented ensembles" aim to improve de-identification accuracy by using institution-specific dictionaries of PII and medical terms to refine the output of ensemble models, showing better performance than LLMs used in isolation for this task.57
* **Privacy-Enhancing Technologies (PETs):** These offer more robust, mathematically grounded privacy protections:
  + **Federated Learning (FL):** Enables collaborative model training across multiple institutions without centralizing raw patient data. Each institution trains a local model, and only aggregated model updates (e.g., gradients or weights) are shared with a central server or peer-to-peer, significantly reducing the risk of raw data exposure.21 A key challenge in FL for healthcare is harmonizing the heterogeneous datasets from different sites.
  + **Differential Privacy (DP):** Provides a formal privacy guarantee by adding carefully calibrated statistical noise to the data, model parameters, or query outputs. This makes it computationally infeasible to determine whether any particular individual's data was included in the computation, thereby protecting against re-identification and inference attacks.49 DPResNet is an example of a deep learning architecture (ResNet) modified to be compatible with DP for medical image classification.50 A trade-off often exists between the strength of DP (lower epsilon values) and model utility.
  + **Secure Multi-Party Computation (SMPC):** Allows multiple parties to jointly compute a function over their private inputs (e.g., aggregating model updates in FL) without revealing those inputs to each other or to a central coordinator.50 This adds another layer of security, especially against an "honest-but-curious" server in FL setups.

The critical nature of patient privacy dictates that it cannot be an afterthought in EHR-LLM system design; it must be a foundational principle. While de-identification is a necessary baseline, its inherent limitations, especially with the inferential power of LLMs, mean that relying on it alone is insufficient. Advanced PETs like FL, DP, and SMPC offer more robust, privacy-by-design solutions. This implies that future architectures for EHR-LLM integration must incorporate these technologies deeply into the data processing, embedding generation, model training, and inference pipelines. A significant research avenue will be the development of embedding techniques that are themselves privacy-preserving—for example, through differentially private embedding algorithms or federated methods for generating embeddings without centralizing sensitive data.

### C. Addressing Bias and Promoting Fairness in Algorithmic Healthcare

EHR data, being a reflection of real-world clinical practice, can unfortunately encapsulate historical and systemic biases related to patient demographics such as race, gender, socioeconomic status, and geographic location.10 These biases can manifest as disparities in diagnosis rates, treatment recommendations, or overall quality of care documented in the records. LLMs trained on such data are at high risk of learning, perpetuating, and even amplifying these biases, potentially leading to inequitable healthcare outcomes, reduced trust in AI systems, and the exacerbation of existing health disparities.5 For example, an LLM might learn to associate certain conditions more strongly with specific demographic groups based on biased historical data, leading to differential diagnostic accuracy or treatment suggestions.

Addressing algorithmic bias is a complex challenge requiring a multifaceted approach:

* **Data-Centric Strategies:**
  + **Dataset Diversification and Augmentation:** Actively working to ensure training datasets are representative of diverse patient populations. This may involve collecting more data from underrepresented groups or using data augmentation techniques to synthetically increase their representation, while being mindful not to introduce new biases.22
  + **Bias Auditing in Data:** Implementing rigorous auditing processes to identify and quantify biases within EHR datasets before they are used for LLM training.
* **Model-Centric Strategies:**
  + **Fairness-Aware Fine-tuning:** Employing fine-tuning techniques that explicitly aim to reduce bias. This can include methods like re-weighting training samples, applying adversarial debiasing, or incorporating fairness constraints into the model's objective function.22 Reinforcement learning with feedback from clinicians or diverse evaluators can also help align model outputs with fairness goals.62
  + **Specialized Prompts and Contextualization:** Using context-specific prompt engineering to guide the LLM towards fairer outputs, and incorporating diverse and international clinical guidelines to ensure broader applicability and reduce regional biases.22
* **Evaluation and Oversight:**
  + **Bias Evaluation Checks:** Implementing specific checks and benchmarks to evaluate LLM outputs for various types of bias across different demographic subgroups.22
  + **Expert Oversight and Clinician Review:** Involving human experts and diverse clinician panels to review LLM outputs for potential biases that automated metrics might miss.22
  + **Retrieval-Augmented Generation (RAG):** Grounding LLM responses in known, accurate, and potentially less biased external data sources (e.g., curated medical knowledge bases, unbiased clinical guidelines) can help mitigate the generation of biased information.22 However, the external knowledge sources themselves must be carefully vetted for bias.

Reporting guidelines like **TRIPOD-LLM** emphasize the importance of transparency regarding data sources and the explicit recognition and addressing of societal biases encoded in clinical models.23 Similarly, the **AMA principles** for augmented intelligence strongly advocate for mitigating bias and promoting health equity in AI systems.55

It is crucial to understand that bias mitigation is not a one-time fix but an active, iterative process. Simply attempting to de-bias a dataset prior to training may be insufficient, as biases can be subtle, deeply embedded, and re-emerge in different contexts or as the model is updated over time. An effective strategy requires continuous monitoring of model performance and outputs for fairness across all relevant demographic groups, coupled with mechanisms for redress if biased outcomes are detected. This underscores the need for a socio-technical approach to fairness in EHR-LLM systems, involving not only algorithmic solutions but also diverse development teams, active community engagement, and ongoing ethical audits throughout the entire lifecycle of the AI model.

### D. Enhancing Trust: Interpretability and Explainability (XAI methods, EHRMind)

The inherent opacity or "black box" nature of many sophisticated machine learning models, including LLMs, poses a significant barrier to their widespread adoption and trust in high-stakes domains like healthcare.5 For clinicians to confidently use LLM-generated insights for patient care, and for patients to accept AI-assisted medical decisions, it is often necessary to understand how these models arrive at their conclusions. **Explainable AI (XAI)** encompasses a range of methods aimed at making AI decisions more transparent and interpretable without unduly compromising performance.

LLMs can play a dual role in XAI: they can be the subject of explanation, and they can also be used to *generate* explanations for their own outputs or the outputs of other complex ML models by transforming them into understandable narratives.64 Key XAI approaches relevant to LLMs include:

* **Intrinsic Interpretability Methods:** These involve designing LLM architectures or training processes that are inherently more understandable:
  + **Chain-of-Thought (CoT) Reasoning:** Prompting or fine-tuning LLMs to generate a step-by-step reasoning process before providing a final answer. This makes the logical flow of the model's "thought process" more explicit and easier to follow.4
  + **Guided Templates:** Providing structured templates that guide the LLM to produce outputs in a logically organized and interpretable format.64
  + **EHRMind Approach:** This method for adapting LLMs to complex clinical reasoning tasks uses a two-stage solution. An initial supervised fine-tuning (SFT) warm-up phase injects domain knowledge and encourages the generation of structured, interpretable outputs. This is followed by Reinforcement Learning with Verifiable Rewards (RLVR) to refine decision-making while maintaining the learned interpretable structure.67
* **Post-hoc Explanation Techniques:** These methods are applied after a model has made a prediction to explain that specific decision:
  + **LIME (Local Interpretable Model-Agnostic Explanations):** Explains individual predictions by learning a simpler, interpretable model (e.g., a linear model) locally around the prediction, identifying key input features that contributed most to the outcome.64
  + **SHAP (SHapley Additive exPlanations):** Uses concepts from cooperative game theory (Shapley values) to assign an importance value to each input feature for a particular prediction, providing both local and global (average feature importance) explanations.64 SHAP can be adapted for LLMs to highlight influential tokens or input segments.
* **Guided Embedding Refinement:** As discussed previously, using LLMs to generate "guided embeddings" based on clinically interpretable attributes (e.g., "disease severity") can inherently make the reasoning of downstream models that use these embeddings more transparent.45

For LLM decisions based on EHR data, these XAI methods can be applied to improve transparency. For example, if an LLM predicts a high risk of readmission for a patient, an XAI technique could highlight the specific elements in the patient's textualized EHR (e.g., recent lab values, specific diagnoses in notes, medication changes) that most influenced this prediction. CoT prompting could make the LLM articulate its reasoning steps: "The patient has a history of heart failure (from structured codes), recent clinical notes mention increased dyspnea (unstructured text), and lab results show elevated BNP (structured data). Therefore, the risk of readmission is high."

However, for explanations to be truly valuable in healthcare, they must be **clinically actionable and relevant**, not just technically sound.64 Generic feature importance scores might not be directly useful to a clinician. Instead, explanations need to be grounded in established medical knowledge and the specific patient's context. For instance, an explanation like "the LLM predicted sepsis because the patient's record contains mentions of fever, tachycardia, and elevated white blood cell count, which are key indicators of SIRS criteria often associated with sepsis" is far more useful than "input tokens 5, 17, and 32 had high attention scores." Approaches like EHRMind, which encourage structured clinical reasoning paths 67, and the use of LLMs to translate complex model outputs into understandable narratives 64, are steps in this direction. The true clinical utility of XAI in EHR-LLM systems will emerge when explanations not only make the AI's process transparent but also help clinicians validate the reasoning, identify potential errors, or even learn new clinical insights that can improve patient care. This necessitates co-designing XAI features with healthcare professionals to ensure they align with clinical workflows and decision-making processes.

### E. Managing Temporal Dynamics, Sparsity, and Noise in Longitudinal EHR Data (e.g., TDE, TEMPORALCROSS)

EHR data is fundamentally **longitudinal**, capturing a patient's health journey as a sequence of events—visits, diagnoses, lab tests, medications—over time.1 This temporal dimension is critical for understanding disease progression, treatment effectiveness, and risk prediction. However, longitudinal EHR data presents unique challenges for modeling:

* **Irregular Sampling:** Clinical events are not recorded at fixed intervals; visits occur as needed, and tests are ordered based on clinical judgment, leading to irregularly spaced time points.2
* **Variable-Length Sequences:** Patient histories vary greatly in length.2
* **Sparsity and Missing Entries:** Not all data types are recorded at every encounter, leading to sparse sequences with many missing values.2
* **Noise:** As with all EHR data, longitudinal records can contain noisy or inaccurate information.2

Standard LLM architectures, primarily designed for contiguous blocks of text, may not natively and optimally handle these temporal irregularities. Simply serializing a long, sparse patient history into a single text string might obscure crucial temporal relationships or exceed the LLM's effective context window, making it difficult for standard attention mechanisms to capture long-range dependencies accurately.70

To address these challenges, specialized techniques for modeling temporal EHR data are being developed:

* **Temporal Dynamic Embedding (TDE):** This approach treats each time-series variable (e.g., a specific lab test) as an embedding vector that evolves over time.69 At each time step, TDE selectively adopts and aggregates only the *observed* variable subsets, thus inherently handling sparsity and irregular sampling without requiring imputation of missing values. It incorporates time embeddings to capture temporal patterns and can use attention-based aggregation to weigh the importance of different observed variables at each point in time. The aggregated local status at each time step is then fed into a recurrent model like a GRU to capture the global temporal trajectory.
* **Temporal Cross-Attention (TEMPORALCROSS) Framework:** Designed for multimodal clinical time series (integrating structured EHR data and unstructured clinical notes), this framework introduces several innovations to create time-aware representations.70 It uses:
  + **Flexible Positional Encoding:** Allows multiple variable tokens recorded at the same timestamp to share the same absolute positional index, and adds relative positional encoding to capture local dependencies within long sequences.
  + **Learnable Time Embeddings:** Techniques like Time2Vec are used to explicitly encode information about the relative time between events, capturing important temporal patterns beyond simple sequence order.
  + **Variable-Specific Encoding:** Captures distinct characteristics and relationships between different temporal variables. A **temporal cross-attention mechanism** then fuses information from the different modalities (structured data and text) over time, allowing the model to learn a joint multimodal temporal representation.
* **TIMER Framework:** This framework focuses on improving and evaluating LLMs' temporal reasoning capabilities in longitudinal EHRs.74 It includes:
  + **TIMER-Bench:** A time-aware benchmark specifically designed to evaluate how well LLMs can reason over temporal dependencies across multiple patient visits and different time frames.
  + **TIMER-Instruct:** An instruction-tuning methodology to train LLMs to better perform temporal reasoning tasks on longitudinal clinical records.

The development of these explicit temporal modeling techniques highlights a crucial understanding: for LLMs to be truly effective with the rich, dynamic information in longitudinal EHRs, they likely need to be augmented with or integrated into architectures specifically designed to represent and reason about time. Patient health is a dynamic process where the sequence, timing, and duration of events are often as diagnostically and prognostically important as the events themselves. Simply serializing temporal data may not be sufficient. The future of LLMs in processing longitudinal EHRs will likely involve hybrid architectures that synergistically combine the natural language understanding strengths of LLMs with specialized temporal modeling components, such as recurrent layers, temporal convolutions, or dedicated temporal attention mechanisms, to accurately capture and interpret patient health trajectories.

### F. Integrating Multimodal EHR Data (Text, Structured Data, Imaging)

EHRs are inherently **multimodal**, containing a rich tapestry of information beyond just textual notes. This includes structured data (coded diagnoses, lab values, medications), unstructured text (clinical narratives), and often imaging data (X-rays, CT scans, MRIs), and potentially even genomic data.1 Effectively integrating these diverse modalities is crucial for obtaining a holistic view of the patient and for developing accurate predictive models and decision support systems. LLMs, and increasingly Multimodal LLMs (MLLMs), are central to this endeavor.

Several frameworks and approaches are emerging to tackle multimodal EHR data integration:

* **Evidence-Based Multimodal Fusion Frameworks:** One such framework, proposed for ICU outcome prediction, involves using distinct encoders for different modalities—for example, ResNet or Transformer-based models for structured EHR data and pre-trained language models (PLMs) for free-text notes.76 The features extracted from each modality are then independently mapped into an "evidence" representation. These pieces of evidence are subsequently combined in an "evidence space," potentially using principles from Dempster-Shafer theory or similar evidential reasoning formalisms, to arrive at a final prediction. This approach aims to reduce false positives and improve prediction reliability.
* **Retrieval-Augmented Generation (RAG) for Multimodal Enhancement (EMERGE Framework):** The EMERGE framework utilizes a RAG-driven approach to enhance predictive modeling with multimodal EHR data (time-series data and clinical notes).71 It employs LLMs to extract relevant medical entities from both time-series data (e.g., using z-score based filtering to identify abnormal values which are then textualized) and clinical notes. These extracted entities are then aligned with a professional knowledge graph (PrimeKG) to ensure consistency and mitigate LLM hallucinations (e.g., by comparing LLM-generated entities with original notes and using embedding similarities for KG alignment). The distilled knowledge, potentially including entity definitions and descriptions, is summarized by a long-context LLM. Finally, an adaptive multimodal fusion network, incorporating a cross-attention mechanism, integrates this summarized knowledge with representations from other modalities to make clinical predictions.
* **Multi-Agent Collaborative Frameworks (ColaCare):** The ColaCare framework exemplifies a multi-agent system for EHR modeling.79 It integrates domain-specific expert models (which process numerical EHR data and provide initial predictions and feature importance) with LLM-driven agents (DoctorAgents and a MetaAgent). The LLM agents engage in a collaborative consultation process, akin to a Multidisciplinary Team (MDT) meeting, to produce reasoning references and decision-making reports. A multimodal fusion network then combines the hidden representations from the expert models with embeddings of the final consensus report from the LLM agents to make a final prediction.
* **Multimodal LLMs (MLLMs) for Time Series Reasoning:** There is a growing recognition that traditional time series analysis often overlooks the rich multimodal context accompanying numerical data. Position papers advocate for the use of MLLMs that can inherently process and reason over time-dependent information that includes not just numerical sequences but also associated textual descriptions, visual data (like medical images linked to specific time points), and even audio signals.78 Such MLLMs could enable deeper temporal and multimodal reasoning, leading to more nuanced understanding of patient trajectories from EHR data.

The evolution of these multimodal fusion techniques indicates a significant trend: a shift away from simple early or late fusion of data streams (e.g., concatenating feature vectors or averaging predictions) towards more sophisticated, semantically rich, and context-aware integration strategies. LLMs are pivotal in this shift, not only by processing the textual components of EHRs but also by facilitating the interpretation and alignment of entities across different modalities (as in EMERGE) or by providing a framework for reasoning over combined information from diverse sources (as in ColaCare). This suggests a future where the goal is not just to combine different data types, but to create a unified, semantically coherent, and dynamic representation of the patient that incorporates information from all available modalities in a contextually aware manner. Achieving this vision will heavily depend on the development of advanced MLLMs or hybrid architectures capable of seamlessly grounding textual information with structured data, interpreting visual and other non-textual data types, and understanding their interplay over time. This also underscores the critical need for large-scale, well-curated, and ethically sourced multimodal EHR datasets to train and rigorously evaluate these advanced models.

## V. Advancing the Frontier: Innovations and Best Practices

As the integration of EHRs with LLMs matures, several innovative approaches and best practices are emerging to address the inherent complexities of medical data and to enhance the capabilities of these AI systems. These advancements span data harmonization, the use of knowledge graphs, robust model validation, and strategies for continuous learning.

### A. Data Harmonization Frameworks and Tools (e.g., ehrapy, MedRep, CDE-Mapper, ontology-LLM alignment)

Data harmonization is a critical precursor to effective EHR-LLM integration, aiming to transform heterogeneous data from diverse sources into a consistent and comparable format. Several frameworks and tools are being developed to facilitate this complex process:

* **ehrapy:** This open-source Python framework is specifically designed for the exploratory analysis of heterogeneous epidemiology and EHR data.10 It provides a unified pipeline that standardizes various analytical steps, from data extraction and quality control to the generation of low-dimensional representations and causal inference. ehrapy leverages the AnnData (annotated data) structure, widely used in omics data analysis, ensuring compatibility and promoting standardized workflows. It supports a range of input formats (e.g., CSV, OMOP, SQL databases) and can handle diverse EHR data types, including demographics, laboratory results, vital signs, diagnoses, medications, unstructured written notes (from which it can extract clinical keywords), and even imaging and omics measurements. Key features include tools for inspecting data quality, detecting and imputing missing values, tracking filtering steps to highlight potential biases, and offering normalization and encoding functions.
* **MedRep:** This strategy focuses on creating vocabulary-agnostic medical concept representations for EHR foundation models, particularly within the OMOP Common Data Model (CDM) framework.15 MedRep addresses the challenge of models encountering unseen medical codes or different concept IDs for semantically similar entities across institutions. It employs a two-step process:
  1. **LLM-Prompted Enrichment:** LLM prompts are used to generate rich, detailed natural language descriptions for OMOP concept names, enhancing their minimal standard definitions.
  2. **Graph Ontology Enhancement:** These LLM-generated text-based representations are then refined by integrating relational knowledge from the OMOP vocabulary's graph ontology using graph contrastive learning (e.g., GRACE framework) and a "Learning without Forgetting" mechanism. This produces robust concept representations that EHR foundation models can directly utilize.
* **CDE-Mapper:** This tool utilizes Retrieval-Augmented Language Models (RALMs) for linking Common Data Elements (CDEs)—standardized questions and allowable answers used in clinical research and healthcare—to controlled terminologies like SNOMED CT, RxNorm, and LOINC.16 Its methodology involves query decomposition (breaking down complex CDEs), knowledge filtering (to refine retrieved candidates), and reranking to identify the most appropriate mappings. It is designed to handle both atomic (single characteristic) and composite (interdependent attributes) CDEs.
* **Ontology- and LLM-based Data Alignment for Federated Learning:** In the context of federated learning (FL), where data from multiple institutions remains decentralized, data harmonization is particularly challenging yet crucial. A proposed two-step strategy involves 21:
  1. **Candidate Generation:** Matching candidates for data elements are generated using either vector-space embeddings of terms or ontology-based converter matching.
  2. **LLM-based Adjudication:** An LLM is then used to evaluate these candidate pairs, accepting or rejecting them based on predefined criteria, thereby harmonizing diverse clinical datasets in a privacy-preserving manner. This approach was applied in a real-world project for semantic mapping of EHR data related to drug exposure during pregnancy.

These innovative frameworks signify an evolution in data harmonization from purely syntactic mapping or rule-based transformations towards more AI-assisted, semantically driven processes. Initial efforts relied on common data models like OMOP and standard ontologies to provide a semantic foundation.15 Current advanced tools are now deeply integrating machine learning and LLMs to automate, scale, and improve the accuracy of this harmonization.10 LLMs, with their nuanced understanding of natural language, can assist in interpreting ambiguous terms found in local EHRs, mapping local codes to standard terminologies, and validating the semantic consistency of mappings. This shift suggests a future where data harmonization is a dynamic, learning-based system, with LLMs and other AI tools continuously refining data mappings and improving data consistency across disparate EHR sources. Such advancements are pivotal for enabling more reliable and scalable large-scale clinical analytics and research.

### B. Leveraging Knowledge Graphs for Enhanced Contextual Understanding (e.g., DR. KNOWS, LLMs for KG completion)

Medical Knowledge Graphs (KGs) are structured representations of medical knowledge, encoding entities (e.g., diseases, drugs, symptoms, genes) and the relationships between them. Integrating KGs with LLMs offers a powerful approach to enhance the contextual understanding and reasoning capabilities of LLMs in the medical domain, particularly when processing EHR data.26

* **DR. KNOWS (Diagnostic Reasoning Knowledge Graph System):** This model exemplifies how KGs can be integrated with LLMs to improve diagnostic predictions from EHR data.83 DR. KNOWS utilizes UMLS (Unified Medical Language System)-based KGs. Its methodology involves:
  1. Extracting UMLS concepts from patient text (e.g., SOAP notes).
  2. Encoding these concepts and their 1-hop subgraphs from the UMLS KG using a Stacked Graph Isomorphism Network (SGIN) to generate node embeddings.
  3. Employing an attention-based path ranker to identify and rank knowledge paths within the KG that are most relevant to the patient's specific clinical context.
  4. Incorporating these retrieved, contextually relevant paths into prompts for an LLM (e.g., T5, ChatGPT), which then makes the diagnostic prediction. This approach demonstrated improved precision and F-scores in diagnosis prediction compared to baselines, and human evaluations suggested that incorporating KG paths led to reasoning more aligned with human expert rationale.
* **LLMs for Knowledge Graph Completion and Augmentation:** Medical KGs, while valuable, are often incomplete due to evolving medical knowledge or limitations in coding systems.26 LLMs, with their ability to process and synthesize information from vast amounts of text (e.g., clinical literature, guidelines), can be used to impute missing relationships or entities in KGs. For instance, LLMs can be queried to suggest potential treatments for diseases not yet explicitly linked in a KG (treatment mapping). However, this application is not without risks: LLMs can generate factually inaccurate or "hallucinated" associations, and their outputs can exhibit instability across different models or prompts. Therefore, using LLMs for KG augmentation requires rigorous validation mechanisms, comparison against established clinical guidelines, and potentially hybrid frameworks that combine generative capabilities with domain-grounded validation.
* **Integrating KGs During LLM Reasoning:** KGs can serve as external, verifiable knowledge sources that LLMs can query or refer to during their reasoning process.29 This can help ground the LLM's outputs in established medical facts, reduce hallucinations, and improve the explainability of its predictions. Retrieval-Augmented Generation (RAG) is a common technique where relevant information from a KG (or other knowledge sources) is retrieved and provided as context to the LLM along with the user's query.

The relationship between KGs and LLMs in the medical domain is increasingly symbiotic. LLMs, pre-trained on vast textual corpora, possess broad implicit knowledge but can suffer from a lack of specific, up-to-date factual grounding and are prone to generating plausible but incorrect statements.5 KGs, on the other hand, provide explicit, structured, and often curated domain knowledge, but they can be incomplete or slow to update.84 By integrating these two technologies, systems can be built where LLMs handle nuanced natural language understanding and flexible reasoning, while KGs provide a robust, verifiable knowledge backbone. This synergy, often termed Neuro-Symbolic AI, is crucial for developing trustworthy and reliable AI systems in medicine. LLMs can ground their reasoning in the factual structures of KGs, leading to more accurate and explainable outputs, while also contributing to the maintenance and enrichment of these KGs, albeit under careful human or automated validation.

### C. Ensuring Robustness: Model Validation, Reproducibility, and Generalizability (e.g., TRIPOD-LLM guidelines)

The high-stakes nature of healthcare necessitates that LLM-based applications processing EHR data are robust, reliable, and generalizable. Rigorous model validation is therefore not just a best practice but an ethical imperative.11 However, evaluating LLMs in the medical context presents unique challenges:

* **Inconsistent Reliability and Accuracy:** Studies have reported inconsistent reliability and accuracy issues with LLMs, including unclear diagnoses or missing essential treatment recommendations, which have severe implications for patient safety.60
* **Lack of Clinical Knowledge:** General-purpose LLMs may lack deep, nuanced clinical knowledge, especially for specialized or rare conditions.60
* **Limitations of Standard Metrics:** While standard machine learning metrics like accuracy, F1-score, and AUROC are commonly used 60, they may not fully capture the critical aspects of LLM performance in clinical settings. For generative LLMs, these metrics do not assess the factual correctness of generated text, the absence of harmful hallucinations, or the clinical relevance and safety of the outputs.23
* **Evaluation Variability:** There is a lack of standardized evaluation frameworks, tools, and datasets, making it difficult to compare results across studies and leading to variability in how LLM performance is assessed.11

To address these issues and promote transparency and rigor in LLM research, reporting guidelines are being developed. The **TRIPOD-LLM statement** is an extension of the Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD) guideline, specifically tailored for studies developing, tuning, prompting, or evaluating LLMs in biomedical applications.23 Key aspects of TRIPOD-LLM include:

* **Comprehensive Checklist:** It provides a detailed checklist of items essential for good reporting, covering title, abstract, introduction, methods (including specifics on LLM version, training data cut-off dates, fine-tuning processes, prompt engineering details), open science practices, patient involvement, results, and discussion.
* **Emphasis on Transparency and Human Oversight:** The guideline stresses the importance of transparently reporting how the LLM-based research was conducted, including measures for human oversight, especially in dataset development and evaluation (e.g., qualifications of assessors, dual annotation).
* **Task-Specific Performance Reporting:** It calls for reporting performance metrics that are relevant to the specific clinical task and context, moving beyond generic NLP metrics.
* **Addressing LLM-Specific Challenges:** It explicitly addresses unique LLM challenges such as hallucinations, omissions, reliability, explainability, reproducibility, privacy, and bias.

The development and adoption of such guidelines are crucial for improving the quality and trustworthiness of LLM research in healthcare. However, robust validation must extend beyond adherence to reporting standards and technical metrics. It requires a holistic assessment of an LLM's clinical utility and safety. This means evaluating not just its predictive accuracy but also its impact on actual clinical workflows, its ability to support rather than hinder clinician decision-making, its safety profile in real-world scenarios, and its fairness across diverse patient populations. Such comprehensive validation often necessitates human expert evaluation, prospective studies, and carefully designed pilot programs in real clinical environments.32 Establishing best practices for LLM validation in EHR applications will therefore require a paradigm shift towards multi-faceted evaluation frameworks that integrate technical performance with assessments of clinical utility, safety, fairness, and interpretability, ideally supported by standardized benchmarks and the generation of real-world evidence.

### D. Strategies for Continuous Learning and Model Adaptation in Dynamic EHR Environments (e.g., ColaCare, EHR-CoAgent, feedback mechanisms)

Medical knowledge is not static; new research emerges, clinical guidelines evolve, and individual patient conditions change over time. Similarly, EHR data itself is dynamic, with new information being continuously added. LLMs trained on a fixed dataset can therefore become outdated, leading to suboptimal performance or even unsafe recommendations if not adapted to these changes.29 Consequently, strategies for **continuous learning and model adaptation** are vital for the long-term viability and reliability of LLMs in EHR environments.

Several innovative approaches are being explored to enable LLMs to learn and adapt:

* **Multi-Agent Systems with Feedback Loops:**
  + **EHR-CoAgent:** This framework utilizes a **predictor LLM agent** to make disease predictions and generate reasoning, and a **critic LLM agent** to analyze incorrect predictions made by the predictor.79 The critic identifies error patterns and provides instructional feedback. This feedback is then used to refine the prompts given to the predictor agent, enabling the system to learn from its mistakes and improve its reasoning over time.
  + **ColaCare:** This framework employs multiple **DoctorAgents** (each linked to a domain-specific expert model for processing numerical EHR data) and a **MetaAgent** to simulate a Multidisciplinary Team (MDT) consultation.79 The LLM-driven agents engage in an iterative process of review, discussion, and report refinement, effectively allowing the system to converge on a consensus by incorporating diverse perspectives and evidence. This collaborative process inherently involves feedback and adaptation.
* Reinforcement Learning with Human or AI Feedback:  
  Techniques like Reinforcement Learning from Human Feedback (RLHF) or Reinforcement Learning from AI Feedback (RLAIF) are relevant for iteratively improving LLM behavior.62 In a healthcare context, this could involve clinicians providing feedback on the quality, accuracy, and safety of LLM outputs, which is then used to fine-tune the model. The EHRMind approach, for instance, uses Reinforcement Learning with Verifiable Rewards (RLVR) after an initial supervised fine-tuning phase to reinforce outcome correctness and refine the model's decision-making based on verifiable outcomes.67

The core idea behind these continuous learning strategies is to create systems that are not static but can evolve and improve based on new data, feedback, and changing medical knowledge. However, this process must be carefully managed. Blindly updating models based on new data or AI-generated feedback can be risky if the new data contains biases, the feedback is flawed, or if the model experiences "catastrophic forgetting" (losing previously learned knowledge when acquiring new information).

Therefore, continuous learning in EHR-LLM systems necessitates robust **human-AI collaboration** and mechanisms for **dynamic knowledge updating**. Human oversight remains critical to validate the feedback, guide the learning process, and ensure that model updates align with clinical best practices and ethical standards. Furthermore, systems should be designed to integrate new, verified external knowledge (e.g., updated clinical guidelines, new drug approvals) in a controlled manner, perhaps through regularly updated knowledge bases used in conjunction with Retrieval-Augmented Generation (RAG) techniques. The operationalization of such continuously learning EHR-LLM systems in real-world clinical settings will require sophisticated MLOps (Machine Learning Operations) frameworks specifically designed for healthcare. These frameworks must support safe and effective model updating, rigorous version control, continuous performance monitoring, and human-in-the-loop validation processes to ensure ongoing reliability and patient safety.

## VI. Governance, Ethics, and the Regulatory Horizon

The integration of LLMs with EHR data, while promising, operates within a complex ecosystem of governance, ethical considerations, and evolving regulatory frameworks. Ensuring responsible innovation in this domain requires careful navigation of these aspects to maximize benefits while safeguarding patient rights and well-being.

### A. Navigating the Regulatory Landscape: FDA Guidance and International Frameworks

The regulatory landscape for AI and ML in healthcare, particularly for LLMs, is still maturing. In the United States, the **Food and Drug Administration (FDA)** is the primary body overseeing medical devices, and its approach to AI/ML is actively developing.55

* **Software as a Medical Device (SaMD):** AI/ML software intended for medical purposes such as diagnosis, cure, mitigation, treatment, or prevention of disease can be classified as SaMD and thus fall under FDA regulation.55 The FDA generally applies a risk-based approach, where higher-risk applications receive greater scrutiny.
* **Challenges with LLMs:** Traditional medical device regulations were largely designed for deterministic software. The stochastic nature of generative AI like LLMs (where the same input can produce different outputs), their broad general-purpose capabilities, and the fact that many are not initially marketed as medical devices but are being adopted for clinical support, pose unique regulatory challenges.93 For example, it is difficult to define a specific "intended use" for a general-purpose LLM that can be applied to a wide array of tasks.
* **Evolving FDA Stance:** The FDA has issued draft guidance documents and held workshops to gather input on regulating AI in drug development and medical devices.55 While the Pre-Certification (Pre-Cert) program for SaMD, which aimed for a more agile regulatory pathway, has been tabled 55, the agency continues to explore appropriate oversight mechanisms. There is a recognized need for new authorization pathways that might be better suited for "generalized" decision support tools like LLMs.93
* **International Collaboration:** Regulatory bodies are also engaging in international collaboration. For instance, the FDA, Health Canada, and the UK's Medicines and Healthcare products Regulatory Agency (MHRA) have released joint guiding principles for AI transparency, signaling a move towards harmonized approaches.55

The rapid pace of LLM development outstrips the traditional timelines of regulatory framework evolution.4 This creates a period of uncertainty for developers, healthcare providers, and patients. While regulatory bodies are working to adapt, the current environment necessitates proactive engagement from all stakeholders. Reporting guidelines like **TRIPOD-LLM** aim to bridge some of this gap by promoting transparency and standardized reporting in LLM research, which can inform regulatory thinking and build trust.23 The development of agile, adaptive, and appropriate regulatory frameworks that foster innovation while ensuring patient safety and ethical deployment is a critical ongoing task.

### B. Ethical Imperatives and Principles for Responsible Deployment (e.g., AMA Principles)

Beyond formal regulations, the deployment of LLMs with EHR data is governed by a set of profound ethical imperatives. Numerous concerns have been raised, including:

* **Unconsented Data Use:** Using patient data from EHRs to train or operate LLMs without explicit, informed consent raises significant ethical issues.22
* **Lack of Governance and Accountability:** The absence of clear governance structures and lines of accountability for AI-related errors or harms is a major concern, especially in cases of AI-related malpractice.22
* **Bias and Discrimination:** As previously discussed, LLMs can perpetuate and amplify biases present in EHR data, leading to discriminatory outcomes and exacerbating health inequities.22
* **Patient Safety Risks:** Hallucinations, misinformation, and overreliance on LLM outputs can compromise patient safety.22
* **Transparency and Interpretability:** The "black box" nature of LLMs hinders trust and the ability of clinicians to validate recommendations.22

Professional organizations are stepping in to provide ethical guidance. The **American Medical Association (AMA)**, for instance, has proposed principles for "augmented intelligence" in healthcare.55 These principles emphasize:

* **Core Values:** Ethical, equitable, responsible, accurate, transparent, and evidence-based design and deployment.
* **Governance:** Clear national governance policies and mandatory compliance.
* **Bias Mitigation:** A specific focus on mitigating bias and promoting health equity.
* **Risk-Based Oversight:** Scrutiny proportionate to the potential for harm.
* **Human Intervention:** Specified qualified human intervention points in clinical decisions, with physician judgment applied, especially in high-risk scenarios.
* **Liability and Accountability:** Aligning liability with the ability to mitigate risk, with developers of autonomous systems accepting liability for system failures.
* **Data Privacy:** Ensuring individuals know how their data will be used and have rights to opt-out, update, or request deletion.
* **Transparency and Disclosure:** Requirements for developers to disclose detailed information about AI systems to users (physicians), and for the use of AI impacting patient care to be disclosed to patients and documented.

While many high-level ethical principles for AI in healthcare exist, the primary challenge lies in translating these principles into concrete, actionable practices within the development, deployment, and ongoing monitoring of EHR-LLM systems. "Centering equity," for example, requires more than just acknowledging the problem; it demands active measures such as de-biasing data, rigorously testing for fairness across diverse demographic groups, and designing systems that ensure equitable access to the technology's benefits. Similarly, "transparency" necessitates not just making source code available (if applicable) but providing clinically meaningful and validated explanations for LLM outputs.

This translation from principle to practice demands sustained, interdisciplinary collaboration involving AI developers, clinicians, informaticists, ethicists, legal experts, patient advocacy groups, and regulatory bodies. The operationalization of ethical AI will require the development of new tools, metrics, and processes for "ethics by design" and "responsible AI," specifically tailored to the unique context and high stakes of healthcare. This includes frameworks for conducting ethical impact assessments before deployment and establishing mechanisms for continuous ethical auditing throughout the lifecycle of EHR-LLM systems.

## VII. Future Outlook: Charting the Path Forward for EHR-LLM Integration

The integration of Electronic Health Records (EHRs) with Large Language Models (LLMs) is a rapidly evolving field, poised for significant advancements. This section outlines emerging trends, potential breakthroughs, critical research gaps, and the overarching need for cross-disciplinary collaboration to responsibly guide this transformative technology.

### A. Emerging Trends and Potential Breakthroughs in EHR Embedding

The trajectory of EHR-LLM integration points towards increasingly sophisticated and capable systems:

* **More Powerful and Specialized Medical LLMs:** There is a clear trend towards the development of LLMs that are either pre-trained from scratch on massive biomedical and clinical corpora (like GatorTron 7) or are extensively fine-tuned general-purpose LLMs adapted for the nuances of medical language and reasoning.2 These specialized models aim for deeper clinical understanding and improved accuracy on medical tasks.
* **Advanced Multimodal Integration:** The future will likely see more widespread use of Multimodal LLMs (MLLMs) capable of seamlessly integrating and reasoning over the diverse data types found in EHRs, including unstructured text, structured codes, laboratory values, medical images (e.g., X-rays, pathology slides), and potentially genomic data.2 This holistic data fusion is essential for a comprehensive understanding of patient health.
* **Causal Inference Capabilities:** A significant breakthrough would be the development of LLMs that can move beyond identifying correlations in observational EHR data to infering causal relationships.62 While current LLMs primarily learn statistical patterns, integrating causal inference methodologies could revolutionize clinical research, drug discovery, and the development of personalized treatment strategies by helping to understand *why* certain interventions are effective for specific patients.
* **Improved Efficiency and Accessibility:** Research into model compression techniques (e.g., quantization, pruning, knowledge distillation) and more efficient LLM architectures, such as Mixture of Experts (MoE) models 59, will continue. These advancements aim to make powerful EHR-LLMs more computationally tractable and economically viable for deployment in diverse healthcare settings, including those with limited resources.5
* **Enhanced Automation of Clinical Workflows:** LLMs are expected to further streamline and automate clinical documentation, reducing the administrative burden on healthcare professionals.89 Beyond documentation, LLMs could optimize various healthcare service workflows, from patient scheduling and communication to initial triage and post-discharge follow-up.101
* **Personalized Medicine at Scale:** The ability of LLMs to process and synthesize comprehensive embedded EHR data, including individual patient histories, comorbidities, genetic predispositions (if available), and lifestyle factors, will be a key enabler for delivering highly personalized medicine. This includes more accurate patient risk stratification for various conditions, earlier and more precise diagnoses, and tailored treatment plans optimized for individual patient characteristics.13

### B. Identifying Key Unanswered Questions and Critical Research Gaps

Despite rapid progress, several critical research gaps and unanswered questions remain that need to be addressed to fully realize the potential of EHR-LLM integration safely and effectively:

* **Long-Term Reliability and Safety:** How can we ensure the long-term safety, reliability, and robustness of LLMs as medical knowledge continuously evolves, new treatments emerge, and clinical practices change? Mechanisms for dynamic updating and validation in real-world settings are crucial.11
* **Scalable Privacy Preservation:** What are the most effective, scalable, and practically implementable methods for truly privacy-preserving EHR embedding and LLM training and inference, especially when dealing with large, federated datasets? Balancing privacy with data utility remains a core challenge.11
* **Clinically Meaningful Evaluation:** How can we develop standardized, clinically meaningful benchmarks and evaluation metrics for EHR-LLM systems that go beyond traditional technical accuracy (e.g., AUROC, F1-score) to assess their real-world clinical utility, safety, fairness, and impact on patient outcomes?.11
* **Effective Continuous Learning:** What are the optimal strategies for enabling continuous learning and model updating in dynamic EHR environments while preventing issues like catastrophic forgetting (where the model loses previously learned knowledge) or the unintentional introduction of new biases?.29
* **Actionable Explainability:** How can XAI methods be advanced to provide explanations for LLM decisions that are not just technically transparent but genuinely useful, trustworthy, and actionable for clinicians in their daily workflows, enabling them to understand, validate, and appropriately override LLM recommendations?.11
* **Governance and Accountability:** What robust governance models and clear liability frameworks are needed to ensure accountability when EHR-LLM systems are involved in, or influence, clinical decision-making, especially in cases of error or harm?.11
* **Handling Data Scarcity for Rare Diseases:** How can LLMs be effectively applied to EHR data for rare diseases or underrepresented populations where data is inherently scarce, without amplifying biases or making unreliable predictions?
* **User Trust and Adoption:** What factors influence clinician and patient trust in EHR-LLM systems, and how can systems be designed and implemented to foster appropriate levels of trust and encourage adoption while mitigating over-reliance?

### C. Fostering Cross-Disciplinary Collaboration for Holistic Advancement

Addressing the complex challenges and realizing the full potential of EHR-LLM integration necessitates deep and sustained **cross-disciplinary collaboration**.22 This involves bringing together experts from:

* **AI/Machine Learning:** To develop novel embedding techniques, model architectures, and learning algorithms.
* **Clinical Medicine:** To provide domain expertise, define clinically relevant problems, curate data, and validate model outputs in real-world settings.
* **Healthcare Informatics:** To address issues of data standardization, interoperability, and integration into clinical workflows.
* **Ethics and Law:** To navigate the complex ethical considerations (privacy, bias, fairness, consent) and ensure compliance with legal and regulatory frameworks.
* **Social Sciences:** To understand the societal impact of these technologies and ensure equitable deployment.
* **Patients and Patient Advocacy Groups:** To ensure that patient perspectives, needs, and concerns are central to the development and deployment process.

Insights from other data-intensive fields that are successfully applying LLMs (e.g., finance, law) can offer valuable lessons for healthcare applications. However, it is crucial to recognize the unique safety, ethical, and regulatory demands of the medical domain, where errors can have life-altering consequences. Open-source initiatives, the development of shared benchmark datasets (especially multimodal and longitudinal EHR datasets), and collaborative research platforms can significantly accelerate progress, promote transparency, and foster responsible innovation in this field.

The overall trajectory of EHR-LLM integration points towards a future where AI is not merely a tool for data processing but a collaborative partner in healthcare. Initial AI applications in EHRs often focused on specific predictive tasks using structured data. The advent of LLMs unlocked the vast potential of unstructured text, leading to richer and more nuanced data representations. Current research is pushing the boundaries further, exploring sophisticated multimodal integration, advanced temporal reasoning, the potential for causal inference, and enhanced interpretability. Frameworks like ColaCare 79 and EHR-CoAgent 79, which explicitly model LLMs as collaborative agents or systems incorporating feedback loops, exemplify this trend. The strong emphasis across the research and regulatory landscape on human oversight 23, actionable XAI 64, and robust ethical governance 32 suggests a deliberate move away from the notion of fully autonomous AI making critical clinical decisions. Instead, the vision is one of AI as an intelligent assistant or partner that augments the capabilities of human clinicians, improves efficiency, helps reduce burnout, and ultimately contributes to enhanced patient care through the sophisticated, trustworthy, and collaborative interpretation of complex EHR data.

## VIII. Conclusion and Strategic Recommendations

The integration of Electronic Health Records (EHRs) with Large Language Models (LLMs) through advanced embedding techniques stands as a pivotal development in the pursuit of AI-driven healthcare. This report has traversed the foundational aspects of EHR data and LLM architectures, detailed diverse methodologies for transforming and embedding clinical information, and critically analyzed the array of challenges and innovative solutions shaping this domain.

### Recapitulation of Key Insights on Embedding EHRs into LLMs

The journey to effectively embed EHRs into LLMs is characterized by a compelling juxtaposition of transformative potential and significant complexity. EHRs, as rich, longitudinal, and multimodal repositories of patient information, offer an unparalleled resource for advancing medical knowledge and clinical care. LLMs, with their sophisticated natural language understanding and generation capabilities, provide the means to unlock this potential, particularly from the vast quantities of unstructured text within clinical notes.

Key methodologies involve the **serialization of diverse EHR data types** into textual formats amenable to LLM processing, including the crucial conversion of structured medical codes into human-readable descriptors. General-purpose LLMs are increasingly being repurposed as powerful **EHR encoders**, often outperforming specialized EHR foundation models, especially in few-shot learning scenarios. This leverages their broad semantic understanding, though architectural adaptations and techniques like contrastive learning are often necessary to optimize embedding quality, particularly for decoder-only LLMs. Specialized **medical tokenization** approaches like MedTok further enhance the input quality for LLMs by incorporating domain-specific knowledge.

For **structured and tabular EHR data**, techniques are evolving beyond simple text serialization to include specialized models like StructLM, table-to-text conversions coupled with LLM embeddings, guided embedding refinement, and ontology-enhanced representations like MedRep. These hybrid approaches aim to capture both the semantic content and the inherent relational structure of tabular data.

However, the path is laden with critical challenges. **Data heterogeneity** necessitates robust standardization and interoperability efforts, with standards like FHIR and OMOP CDM playing a vital role, increasingly assisted by LLMs themselves in harmonization tasks. The paramount importance of **patient data privacy and security** demands the integration of de-identification strategies and advanced Privacy-Enhancing Technologies (PETs) like federated learning and differential privacy as foundational design principles. Addressing and mitigating **algorithmic bias** inherited from historical data is an ongoing, iterative process crucial for ensuring fairness and equity. Enhancing **interpretability and explainability** through XAI methods is vital for building clinician trust and ensuring safe deployment. Managing the **temporal dynamics** of longitudinal EHR data requires specialized modeling techniques to capture patient trajectories accurately. Finally, the **multimodal nature** of EHRs calls for advanced fusion strategies to create a holistic patient view.

### Actionable Recommendations for Stakeholders

To navigate this complex landscape and responsibly harness the benefits of EHR-LLM integration, concerted and collaborative efforts are required from all stakeholders:

* **For Researchers:**
  + **Advance Embedding Techniques:** Focus on developing novel embedding methodologies that are robust to EHR data complexities (noise, sparsity, heterogeneity), inherently privacy-preserving (e.g., differentially private embeddings), and highly interpretable.
  + **Develop Specialized Medical LLMs/MLLMs:** Continue research into LLMs and MLLMs specifically pre-trained or extensively fine-tuned on diverse, high-quality, and representative multimodal medical data, including EHRs from varied populations.
  + **Enhance Explainable AI (XAI):** Drive XAI research beyond technical transparency towards methods that provide clinically actionable, validated, and trustworthy explanations for LLM-generated insights from EHR data.
  + **Establish Standardized Benchmarks:** Collaborate to create and adopt comprehensive, standardized benchmarks for evaluating EHR-LLM systems across diverse tasks, modalities (including temporal and imaging data), and patient populations, focusing on clinical utility and safety.
  + **Investigate Causal Inference:** Explore and validate methods for enabling LLMs to perform causal inference from observational EHR data, moving beyond correlational findings.
* **For Developers and Technology Providers:**
  + **Prioritize "Safety and Ethics by Design":** Embed safety, privacy, fairness, and interpretability considerations into the entire lifecycle of EHR-LLM product development, not as afterthoughts.
  + **Build Modular and Adaptable Solutions:** Design EHR-LLM systems that are modular, allowing for easier integration with existing hospital IT infrastructure, and adaptable to evolving medical knowledge and local clinical workflows.
  + **Implement Robust Validation and Continuous Monitoring:** Conduct rigorous internal and external validation before deployment, and establish mechanisms for continuous monitoring of model performance, drift, and potential biases in real-world use.
  + **Ensure Transparency:** Be transparent about the capabilities, limitations, training data characteristics, and potential risks of EHR-LLM systems with healthcare organizations and end-users. Adhere to reporting guidelines like TRIPOD-LLM.
* **For Healthcare Organizations and Clinicians:**
  + **Invest in Data Governance and Quality:** Strengthen efforts in EHR data standardization, quality improvement, and the adoption of common data models like OMOP CDM to create a solid foundation for LLM applications.
  + **Develop Strong AI Governance Frameworks:** Establish clear internal governance policies for the procurement, deployment, and oversight of AI tools, including LLMs, ensuring alignment with ethical principles and regulatory requirements.
  + **Foster AI Literacy:** Invest in training and education programs to enhance AI literacy among clinicians and staff, enabling them to understand, critically evaluate, and effectively use LLM-powered tools.
  + **Conduct Rigorous Pilot Programs:** Implement EHR-LLM solutions through carefully designed pilot programs with thorough real-world evaluation of clinical utility, workflow integration, and impact on patient outcomes before broader rollout.
  + **Maintain Human Oversight:** Ensure that qualified human clinicians remain in control of critical clinical decisions, using LLMs as powerful support tools rather than autonomous decision-makers.
* **For Policymakers and Regulatory Bodies:**
  + **Develop Agile and Adaptive Regulatory Frameworks:** Work collaboratively with stakeholders to create clear, agile, and adaptive regulatory pathways for AI in healthcare that foster innovation while ensuring patient safety, data protection, and ethical conduct.
  + **Promote Data Sharing and Research Collaboration:** Facilitate responsible data sharing for research and development through initiatives that support the creation of large-scale, diverse, and privacy-protected EHR datasets, potentially leveraging federated learning models.
  + **Fund Responsible AI Research:** Increase funding for research into safe, ethical, equitable, and effective AI applications in healthcare, including the development of robust validation methodologies and bias mitigation techniques.
  + **Establish Clear Guidelines for Liability and Accountability:** Develop clear legal and ethical guidelines regarding liability and accountability when AI systems, including LLMs processing EHR data, are involved in clinical decision-making and patient care.

The integration of EHRs with LLMs is not merely a technological advancement but a paradigm shift with the potential to redefine healthcare. By embracing innovation responsibly, fostering collaboration, and prioritizing patient well-being, the medical and AI communities can navigate the challenges and unlock the immense promise of this powerful synergy for a healthier future.

#### Works cited

1. Electronic health record - Wikipedia, accessed June 4, 2025, <https://en.wikipedia.org/wiki/Electronic_health_record>
2. Large Language Models are Powerful Electronic Health Record Encoders - arXiv, accessed June 4, 2025, <https://arxiv.org/html/2502.17403v3>
3. arXiv:2502.17403v2 [cs.LG] 4 Mar 2025, accessed June 4, 2025, [https://arxiv.org/pdf/2502.17403?](https://arxiv.org/pdf/2502.17403)
4. The architecture of language: Understanding the mechanics behind ..., accessed June 4, 2025, <https://www.cambridge.org/core/journals/cambridge-forum-on-ai-law-and-governance/article/architecture-of-language-understanding-the-mechanics-behind-llms/E3DDEFB9C04883733380E04331D6F782>
5. Exploring the Architecture of Large Language Models - Big Data Analytics News, accessed June 4, 2025, <https://bigdataanalyticsnews.com/exploring-architecture-of-large-language-models/>
6. Classifying Unstructured Text in ... - JMIR Medical Informatics, accessed June 4, 2025, <https://medinform.jmir.org/2025/1/e65454>
7. Natural Language Processing for Digital Health in the Era of Large Language Models - PMC, accessed June 4, 2025, <https://pmc.ncbi.nlm.nih.gov/articles/PMC12020548/>
8. What are Electronic Health Records (EHR)? Best Practices | Datavant, accessed June 4, 2025, <https://www.datavant.com/electronic-health-records/electronic-health-records>
9. Large Language Models are Powerful EHR Encoders - arXiv, accessed June 4, 2025, <https://arxiv.org/html/2502.17403v2>
10. Health Record Data Nature Medicine 24 | PDF | Electronic Health ..., accessed June 4, 2025, <https://www.scribd.com/document/863970301/Health-Record-Data-Nature-Medicine-24>
11. Safety challenges of AI in medicine in the era of large language models - arXiv, accessed June 4, 2025, <https://arxiv.org/html/2409.18968v2>
12. Leveraging electronic health record data for endometriosis research - Frontiers, accessed June 4, 2025, <https://www.frontiersin.org/journals/digital-health/articles/10.3389/fdgth.2023.1150687/epub>
13. Personalized medicine and the power of electronic health records ..., accessed June 4, 2025, <https://pmc.ncbi.nlm.nih.gov/articles/PMC6921466/>
14. Using artificial intelligence to develop a measure of orthopaedic ..., accessed June 4, 2025, <https://pmc.ncbi.nlm.nih.gov/articles/PMC12058693/>
15. arXiv:2504.08329v1 [cs.AI] 11 Apr 2025, accessed June 4, 2025, [https://arxiv.org/pdf/2504.08329?](https://arxiv.org/pdf/2504.08329)
16. CDE-Mapper: Using Retrieval-Augmented Language Models for Linking Clinical Data Elements to Controlled Vocabularies - arXiv, accessed June 4, 2025, <https://arxiv.org/html/2505.04365v1>
17. [2402.01711] LLM on FHIR -- Demystifying Health Records - arXiv, accessed June 4, 2025, <https://arxiv.org/abs/2402.01711>
18. Integrating the Clinical Data Through FHIR Bundle to OMOP CDM - PubMed, accessed June 4, 2025, <https://pubmed.ncbi.nlm.nih.gov/40380541/>
19. MENDS-on-FHIR: Leveraging the OMOP common data model and ..., accessed June 4, 2025, <https://pmc.ncbi.nlm.nih.gov/articles/PMC10690355/>
20. arxiv.org, accessed June 4, 2025, <https://arxiv.org/pdf/2504.08329>
21. arxiv.org, accessed June 4, 2025, <https://arxiv.org/pdf/2505.20020>
22. Bridging AI and Healthcare: A Scoping Review of Retrieval-Augmented Generation—Ethics, Bias, Transparency, Improvements, and Applications | medRxiv, accessed June 4, 2025, <https://www.medrxiv.org/content/10.1101/2025.04.01.25325033v1.full-text>
23. The TRIPOD-LLM reporting guideline for studies using large language models, accessed June 4, 2025, <https://www.tripod-statement.org/wp-content/uploads/2025/01/TRIPOD-LLM-Article.pdf>
24. arxiv.org, accessed June 4, 2025, <https://arxiv.org/html/2402.16671v1>
25. openreview.net, accessed June 4, 2025, <https://openreview.net/pdf?id=EKBPn7no4y>
26. arxiv.org, accessed June 4, 2025, <https://arxiv.org/pdf/2503.04748>
27. Large Language Models for Healthcare Text Classification: A Systematic Review - arXiv, accessed June 4, 2025, <https://arxiv.org/html/2503.01159v1>
28. arxiv.org, accessed June 4, 2025, <https://arxiv.org/html/2412.10337v2>
29. Merging Clinical Knowledge into Large Language Models for Medical Research and Applications: A Survey - arXiv, accessed June 4, 2025, <https://arxiv.org/html/2502.20988v1>
30. arxiv.org, accessed June 4, 2025, <https://arxiv.org/pdf/2502.20988>
31. arxiv.org, accessed June 4, 2025, <https://arxiv.org/html/2505.22964v1>
32. Preventing unrestricted and unmonitored AI experimentation in ..., accessed June 4, 2025, <https://pmc.ncbi.nlm.nih.gov/articles/PMC11743128/>
33. arxiv.org, accessed June 4, 2025, <https://arxiv.org/pdf/2502.17403>
34. Shifting Long-Context LLMs Research from Input to Output - arXiv, accessed June 4, 2025, <https://arxiv.org/html/2503.04723>
35. arxiv.org, accessed June 4, 2025, <https://arxiv.org/html/2412.09165v2>
36. arxiv.org, accessed June 4, 2025, <https://arxiv.org/pdf/2412.09165>
37. Evaluating the Effectiveness of Large Language Models in Converting Clinical Data to FHIR Format - Preprints.org, accessed June 4, 2025, <https://www.preprints.org/manuscript/202502.1664/v1>
38. Multimodal Medical Code Tokenizer - arXiv, accessed June 4, 2025, <https://arxiv.org/html/2502.04397v2>
39. Multimodal Medical Code Tokenizer - arXiv, accessed June 4, 2025, <https://www.arxiv.org/pdf/2502.04397>
40. arxiv.org, accessed June 4, 2025, <https://arxiv.org/pdf/2502.04397>
41. A Comprehensive Survey of Synthetic Tabular Data Generation - arXiv, accessed June 4, 2025, <https://arxiv.org/html/2504.16506v2>
42. arxiv.org, accessed June 4, 2025, <https://arxiv.org/pdf/2504.16506>
43. arXiv:2408.11854v2 [cs.CL] 19 Sep 2024, accessed June 4, 2025, [https://arxiv.org/pdf/2408.11854?](https://arxiv.org/pdf/2408.11854)
44. arxiv.org, accessed June 4, 2025, <https://arxiv.org/pdf/2408.11854>
45. arxiv.org, accessed June 4, 2025, <https://arxiv.org/abs/2504.11658>
46. Ontology- and LLM-based Data Harmonization for Federated Learning in Healthcare - arXiv, accessed June 4, 2025, <https://arxiv.org/abs/2505.20020>
47. arxiv.org, accessed June 4, 2025, <https://arxiv.org/html/2505.20020v1>
48. A Case Study Exploring the Current Landscape of Synthetic Medical Record Generation with Commercial LLMs - arXiv, accessed June 4, 2025, <https://arxiv.org/html/2504.14657v1>
49. [2406.15962] Privacy Preserving Machine Learning for Electronic Health Records using Federated Learning and Differential Privacy - arXiv, accessed June 4, 2025, <https://arxiv.org/abs/2406.15962>
50. arxiv.org, accessed June 4, 2025, <https://arxiv.org/abs/2412.00687>
51. arxiv.org, accessed June 4, 2025, <https://arxiv.org/html/2502.06252v1>
52. Privacy-Preserving Machine Learning for Electronic Health Records using Federated Learning and Differential Privacy - arXiv, accessed June 4, 2025, [https://arxiv.org/pdf/2406.15962?](https://arxiv.org/pdf/2406.15962)
53. arxiv.org, accessed June 4, 2025, <https://arxiv.org/abs/2112.04359>
54. TOWARDS ROBUST NATURAL LANGUAGE PROCESSING TO PROMOTE HEALTH EQUITY Keith Harrigian, accessed June 4, 2025, <https://kharrigian.github.io/files/Harrigian_Thesis_Final_2024-08-23.pdf>
55. www.ama-assn.org, accessed June 4, 2025, <https://www.ama-assn.org/system/files/ama-ai-principles.pdf>
56. Clinical insights: A comprehensive review of language models in ..., accessed June 4, 2025, <https://journals.plos.org/digitalhealth/article?id=10.1371/journal.pdig.0000800>
57. www.medrxiv.org, accessed June 4, 2025, <https://www.medrxiv.org/content/10.1101/2024.06.20.24308896v1.full.pdf>
58. arxiv.org, accessed June 4, 2025, <https://arxiv.org/pdf/2406.15962>
59. DeepSeek in Healthcare: A Survey of Capabilities, Risks, and Clinical Applications of Open-Source Large Language Models - arXiv, accessed June 4, 2025, <https://www.arxiv.org/pdf/2506.01257>
60. (PDF) A Systematic Review of Large Language Models in Medical ..., accessed June 4, 2025, <https://www.researchgate.net/publication/386110632_A_Systematic_Review_of_Large_Language_Models_in_Medical_Specialties_Applications_Challenges_and_Future_Directions>
61. arxiv.org, accessed June 4, 2025, <https://arxiv.org/pdf/2502.15871>
62. Medical Hallucination in Foundation Models and Their Impact on Healthcare - medRxiv, accessed June 4, 2025, <https://www.medrxiv.org/content/10.1101/2025.02.28.25323115v1.full-text>
63. Natural Language Processing for Digital Health in the Era of Large Language Models, accessed June 4, 2025, <https://www.researchgate.net/publication/390598834_Natural_Language_Processing_for_Digital_Health_in_the_Era_of_Large_Language_Models>
64. arxiv.org, accessed June 4, 2025, <https://arxiv.org/html/2504.00125v1>
65. www.medrxiv.org, accessed June 4, 2025, <https://www.medrxiv.org/content/10.1101/2024.08.10.24311735v2.full.pdf>
66. arxiv.org, accessed June 4, 2025, <https://arxiv.org/abs/2504.00125>
67. Training LLMs for EHR-Based Reasoning Tasks via Reinforcement Learning - arXiv, accessed June 4, 2025, <https://arxiv.org/html/2505.24105v1>
68. arxiv.org, accessed June 4, 2025, <http://arxiv.org/pdf/2505.24105>
69. arxiv.org, accessed June 4, 2025, <https://arxiv.org/abs/2504.05768>
70. Temporal Cross-Attention for Dynamic Embedding and Tokenization of Multimodal Electronic Health Records - arXiv, accessed June 4, 2025, <https://arxiv.org/html/2403.04012v2>
71. arxiv.org, accessed June 4, 2025, <https://arxiv.org/html/2406.00036v2>
72. arxiv.org, accessed June 4, 2025, <https://arxiv.org/html/2403.04012v1>
73. Explainable Multi-modal Time Series Prediction with LLM-in-the-Loop - arXiv, accessed June 4, 2025, <https://arxiv.org/html/2503.01013v1>
74. TIMER: Temporal Instruction Modeling and Evaluation for Longitudinal Clinical Records, accessed June 4, 2025, <https://arxiv.org/html/2503.04176v1>
75. TEMPORAL CROSS-ATTENTION FOR DYNAMIC EMBEDDING AND TOKENIZATION OF MULTIMODAL ELECTRONIC HEALTH RECORDS - OpenReview, accessed June 4, 2025, <https://openreview.net/pdf?id=zNiIyM3q8v>
76. Evidence-based multimodal fusion on structured EHRs and free-text notes for ICU outcome prediction - arXiv, accessed June 4, 2025, <https://arxiv.org/html/2501.04389v1>
77. The challenge of uncertainty quantification of large language models in medicine - arXiv, accessed June 4, 2025, <https://arxiv.org/html/2504.05278v1>
78. arxiv.org, accessed June 4, 2025, <https://arxiv.org/html/2502.01477v1>
79. LLMs-based Few-Shot Disease Predictions using EHR: A Novel ..., accessed June 4, 2025, <https://pmc.ncbi.nlm.nih.gov/articles/PMC12099430/>
80. ColaCare: Enhancing Electronic Health Record Modeling through Large Language Model-Driven Multi-Agent Collaboration - arXiv, accessed June 4, 2025, <https://arxiv.org/html/2410.02551v2>
81. Edinburgh Research Explorer - ColaCare: Enhancing Electronic ..., accessed June 4, 2025, <https://www.pure.ed.ac.uk/ws/portalfiles/portal/508186863/ColaCare_WANG_DOA20012025_AAM.pdf>
82. Small Models are LLM Knowledge Triggers for Medical Tabular Prediction - arXiv, accessed June 4, 2025, <https://arxiv.org/html/2403.01570v3>
83. Leveraging Medical Knowledge Graphs Into Large ... - JMIR AI, accessed June 4, 2025, <https://ai.jmir.org/2025/1/e58670>
84. Can LLMs Support Medical Knowledge Imputation? An Evaluation-Based Perspective, accessed June 4, 2025, <https://arxiv.org/html/2503.22954v1>
85. [2503.22954] Can LLMs Support Medical Knowledge Imputation? An Evaluation-Based Perspective - arXiv, accessed June 4, 2025, <https://arxiv.org/abs/2503.22954>
86. arXiv:2503.22954v1 [cs.CL] 29 Mar 2025, accessed June 4, 2025, <https://arxiv.org/pdf/2503.22954>
87. arxiv.org, accessed June 4, 2025, <https://arxiv.org/pdf/2503.22954v1>
88. (PDF) A systematic review of large language model (LLM ..., accessed June 4, 2025, <https://www.researchgate.net/publication/389659170_A_systematic_review_of_large_language_model_LLM_evaluations_in_clinical_medicine>
89. Quality assurance informs large-scale use of ambient AI clinical ..., accessed June 4, 2025, <https://permanente.org/quality-assurance-informs-large-scale-use-of-ambient-ai-clinical-documentation/>
90. arXiv:2505.22280v1 [cs.CL] 28 May 2025, accessed June 4, 2025, <https://www.arxiv.org/pdf/2505.22280>
91. Artificial Intelligence for Drug Development - FDA, accessed June 4, 2025, <https://www.fda.gov/about-fda/center-drug-evaluation-and-research-cder/artificial-intelligence-drug-development>
92. FDA's evolving regulatory framework for AI use in drug & device clinical trials and research, accessed June 4, 2025, <https://www.hoganlovells.com/en/publications/fdas-evolving-regulatory-framework-for-ai-use-in-drug-device-clinical-trials-and-research>
93. AI in Health Care and the FDA's Blinspot - Penn LDI, accessed June 4, 2025, <https://ldi.upenn.edu/our-work/research-updates/ai-in-health-care-and-the-fdas-blind-spot/>
94. Navigating US Regulation of Artificial Intelligence in Medicine—A ..., accessed June 4, 2025, <https://pmc.ncbi.nlm.nih.gov/articles/PMC11975648/>
95. Patient Risk Stratification with Time-Varying Parameters: A Multitask ..., accessed June 4, 2025, <https://jmlr.org/papers/v17/15-177.html>
96. (PDF) Large Language Models in Integrative Medicine: Progress ..., accessed June 4, 2025, <https://www.researchgate.net/publication/391868960_Large_Language_Models_in_Integrative_Medicine_Progress_Challenges_and_Opportunities>
97. [2504.10936] Can LLMs Leverage Observational Data? Towards Data-Driven Causal Discovery with LLMs - arXiv, accessed June 4, 2025, <https://arxiv.org/abs/2504.10936>
98. [1611.04660] Causal Inference in Observational Data - arXiv, accessed June 4, 2025, <https://arxiv.org/abs/1611.04660>
99. Beyond Correlation: Towards Causal Large Language Model Agents in Biomedicine - arXiv, accessed June 4, 2025, <https://arxiv.org/html/2505.16982v1>
100. AI-Powered Clinical Documentation and Clinicians' Electronic ..., accessed June 4, 2025, <https://pubmed.ncbi.nlm.nih.gov/39240568/>
101. arXiv:2502.11211v2 [cs.CL] 26 May 2025, accessed June 4, 2025, <https://arxiv.org/pdf/2502.11211>
102. Position: Beyond Assistance – Reimagining LLMs as Ethical and Adaptive Co-Creators in Mental Health Care - arXiv, accessed June 4, 2025, <https://arxiv.org/html/2503.16456v2>