CliCARE: Grounding Large Language Models in Clinical Guidelines for Decision Support over Longitudinal Cancer Electronic Health Records

Dongchen Li[1], Jitao Liang[1], Wei Li[2,3*], Xiaoyu Wang[4], Longbing Cao[5*], Kun Yu[6]

[1] College of Computer Science and Engineering, Northeastern University, Shenyang, China
[2] National Frontiers Science Center for Industrial Intelligence and Systems Optimization, Shenyang, China
[3] Key Laboratory of Intelligent Computing in Medical Image (MIIC), Northeastern University, Shenyang, China
[4] Liaoning Cancer Hospital & Institute, Shenyang, China
[5] Macquarie University, Sydney, Australia
[6] College of Medicine and Biological Information Engineering, Northeastern University, Shenyang, China
2490254@stu.neu.edu.cn, 2472127@stu.neu.edu.cn, liwei@cse.neu.edu.cn, wangxyz007@hotmail.com, longbing.cao@mq.edu.au, yukun@bmie.neu.edu.cn

Abstract

Large Language Models (LLMs) hold significant promise for improving clinical decision support and reducing physician burnout by synthesizing complex, longitudinal cancer Electronic Health Records (EHRs). However, their implementation in this critical field faces three primary challenges: the inability to effectively process the extensive length and fragmented nature of patient records for accurate temporal analysis; a heightened risk of clinical hallucination, as conventional grounding techniques such as Retrieval-Augmented Generation (RAG) do not adequately incorporate process-oriented clinical guidelines; and unreliable evaluation metrics that hinder the validation of AI systems in oncology. To address these issues, we propose CliCARE, a framework for Grounding Large Language Models in **Cli**nical Guidelines for Decision Support over Longitudinal **CA**ncer Electronic Health **RE**cords. The framework operates by transforming unstructured, longitudinal EHRs into patient-specific Temporal Knowledge Graphs (TKGs) to capture long-range dependencies, and then grounding the decision support process by aligning these real-world patient trajectories with a normative guideline knowledge graph. This approach provides oncologists with evidence-grounded decision support by generating a high-fidelity clinical summary and an actionable recommendation. We validated our framework using large-scale, longitudinal data from a private Chinese cancer dataset and the public English MIMIC-IV dataset. In these settings, CliCARE significantly outperforms baselines, including leading long-context LLMs and Knowledge Graph-enhanced RAG methods. The clinical validity of our results is supported by a robust evaluation protocol, which demonstrates a high correlation with assessments made by oncologists.

Code — https://github.com/sakurakawa1/CliCARE

1 Introduction

Large Language Models (LLMs) are emerging as promising tools for clinical decision support, with current research demonstrating their potential to serve as collaborative partners that augment expert workflows, alleviate clinician workloads, and improve decision-making in complex fields such as oncology (Hager et al. 2024; Rajashekar et al. 2024). However, their integration into high-stakes clinical practice is far from straightforward. The reality of clinical oncology involves physicians navigating cognitive burdens from manually integrating fragmented data within multi-year Electronic Health Records (EHRs), a key contributor to professional burnout (Warner et al. 2020; sin 2016). This chaotic, unstructured environment amplifies the critical disparity between LLMs' high performance on standardized benchmarks and their capabilities in the clinic. Indeed, systematic reviews indicate that their performance in cancer decision-making is inconsistent, with critical safety aspects frequently unaddressed (Hao et al. 2025), while other recent studies confirm that even state-of-the-art models struggle to adhere to treatment guidelines or accurately interpret laboratory results (Hager et al. 2024). Therefore, the frontier of this field is not merely the development of more powerful models, but the creation of robust frameworks that ensure these technologies are reliable, safe, and effectively grounded in expert medical knowledge to augment, not supplant, the role of the physician. In practice, augmenting the expert physi-

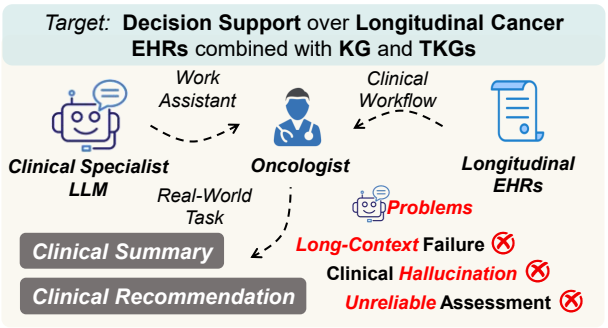

Figure 1: The shared challenges for clinicians and LLM in handling complex longitudinal EHRs.

*Corresponding author.

cian's role means supporting their core clinical workflow: synthesizing a patient's multi-year history into a coherent Clinical Summary, and from that summary, generating an actionable Clinical Recommendation for future treatment.

However, automating this expert workflow with existing LLMs faces three challenges. First, LLMs exhibit an inability to perform effective temporal reasoning over the extensive data records typical of cancer EHRs. Our research addresses a corpus containing large samples of patient records, where a single patient's history can span years, exceed 20,000 tokens, and even include multilingual entries, making brute-force approaches inefficient (Liu et al. 2024). The second challenge is the unacceptable risk of clinical hallucination, which undermines the potential for reliable decision support. Factually incorrect recommendations pose a threat to patient safety, a risk that is exacerbated by the limitations of standard Retrieval-Augmented Generation methods. The retrieval of fragmented text fails to capture the sequential dependencies in a patient's trajectory and cannot effectively bridge the gap with process-oriented clinical guidelines (Li et al. 2024). Finally, the field confronts two interconnected barriers to real-world adoption. The deployment dilemma centers on a trade-off. On one hand, powerful, closed-source models offer state-of-the-art performance but raise significant concerns regarding cost and patient data privacy. On the other hand, open-source alternatives are more efficient and easier to deploy locally, though they often lag in capability. This trade-off is compounded by the significant challenge of reliable evaluation, as the high-stakes nature of clinical content renders conventional automated metrics untrustworthy, thereby hindering reliable progress and diminishing clinical trust (Wang et al. 2023; Zheng et al. 2023).

To address these barriers, we propose CliCARE, a framework for Grounding Large Language Models in **Cli**nical Guidelines for Decision Support over Longitudinal **CA**ncer Electronic Health **RE**cords. CliCARE first tackles long-context temporal analysis by structuring raw EHRs into Temporal Knowledge Graphs (TKGs) to make temporal relationships explicit (Sec 3.1). It then mitigates hallucinations by grounding the model through a deep alignment of patient trajectories with clinical guidelines (Sec 3.2). This representation provides both guideline-grounded data for fine-tuning specialist models and rich context for large generalist models. Finally, we ensure reliable evaluation via an Expert-Validated LLM-as-a-Judge protocol whose ratings correlate with expert judgments (Sec 4.1).

Our contributions are summarized below:

- We introduce CliCARE, an end-to-end framework that grounds LLMs by transforming unstructured clinical text from EHRs into TKGs and aligning them with clinical guidelines. It features an adaptable architecture for both generalist and specialist models.
- We propose a reliable evaluation methodology using an Expert-Validated LLM-as-a-Judge, whose ratings are highly correlated with expert oncologists, addressing the limitations of automated metrics.
- Extensive experiments on diverse datasets show CliCARE significantly outperforms baselines, while ablation studies confirm the contribution of each component.

2 Related Work

2.1 LLMs for Decision Support with Long-Form EHRs

Leveraging LLMs for clinical decision support holds promise for synthesizing complex EHRs. However, the longitudinal, long-form nature of these records presents a barrier. Models often struggle with long-context processing, leading to issues like the lost-in-the-middle problem and performance degradation (Liu et al. 2024; Zhang et al. 2025; Yang et al. 2025b). These challenges are now being systematically evaluated by specialized benchmarks like LongBench and MedOdyssey (Bai et al. 2024; Fan et al. 2024).

To mitigate these issues and unlock the potential of LLMs for decision support, a primary strategy is to transform unstructured data into structured formats. Current approaches range from prompting LLMs on pre-structured data for prediction tasks (Zhu et al. 2024b), or enhancing predictions from structured codes via retrieval augmentation (Xu et al. 2024), to generating Patient Journey Knowledge Graphs (PJKGs) from raw text, though the latter can face reliability challenges (Khatib et al. 2025). The emergence of sophisticated, multi-agent architectures like ColaCare further raises the bar for comprehensive EHR modeling (Wang et al. 2025), highlighting the field's push towards robust, structured solutions for decision support.

2.2 Knowledge Graph-enhanced LLMs and RAG

Augmenting LLMs with external Knowledge Graphs (KGs) is a crucial strategy for mitigating factual errors and hallucinations, which is essential for safety in high-stakes domains such as healthcare (Khan, Wu, and Chen 2024). This approach helps bridge the gap between general-purpose models and specialized clinical knowledge (Yu and McQuade 2025). However, standard RAG often retrieves isolated text snippets, overlooking the relational structures necessary for clinical decision-making (Lewis et al. 2020). Recent efforts seek to address this by focusing retrieval on specific knowledge sources, such as directly incorporating clinical practice guidelines (Oniani et al. 2024) or enhancing multimodal EHR analysis through RAG-driven frameworks like REALM (Zhu et al. 2024a). To more fundamentally resolve knowledge fragmentation, Graph-Aware RAG moves beyond text retrieval to extract structured information, retrieving entire knowledge paths to improve diagnosis prediction (Gao et al. 2025) or pulling relevant subgraphs via frameworks like MedRAG and GNN-RAG (Zhao et al. 2025; Mavromatis and Karypis 2025). An even more advanced frontier moves beyond retrieval to the alignment and fusion of KGs and LLMs at the representation level (Jiang et al. 2024), a synergy that creates a virtuous cycle of grounding and enrichment (Maushagen et al. 2024).

2.3 Assessment of Open-Ended Clinical Generation Tasks

Evaluating open-ended generation from LLMs in high-stakes medical domains presents a challenge. Traditional au-

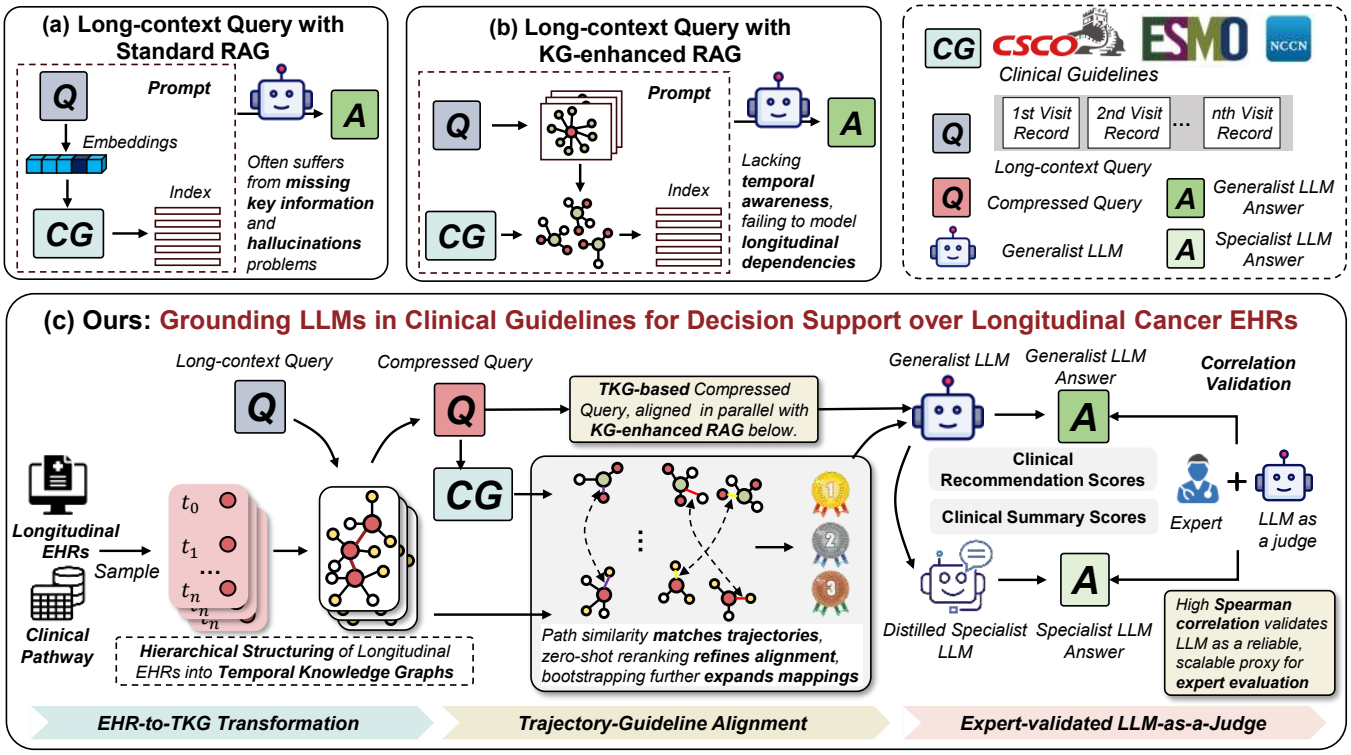

Figure 2: A comparison of RAG approaches for long-form longitudinal clinical tasks. (a) Standard RAG suffers from missing key information and hallucinations. (b) KG-enhanced RAG struggles to model temporal dependencies in patient journeys. (c) In contrast, our CliCARE framework transforms EHRs into Temporal Knowledge Graphs, aligns patient trajectories with guidelines, and generates answers using a distilled specialist model, which are then assessed by our evaluation approach.

tomated metrics, such as ROUGE and BLEU, are regarded as inadequate because their emphasis on lexical overlap fails to capture essential aspects like clinical validity, factual accuracy, and safety (Wang et al. 2023; Singhal et al. 2023). In response, research has increasingly focused on more nuanced evaluation methods, including dynamic agent-based assessments (Tu et al. 2024; Tadevosyan et al. 2025) and the scalable LLM-as-a-Judge paradigm (Zheng et al. 2023). However, the reliability of LLM judges is compromised by known systematic biases like positional bias and verbosity, raising safety concerns in a field where deep domain expertise is essential (Zheng et al. 2023; Wang et al. 2024). This highlights the urgent need for rigorous methodologies to validate automated judgments against human expert reasoning.

Existing research has treated long-context processing, knowledge grounding, and reliable evaluation as distinct challenges. A research gap exists in developing a solution that simultaneously addresses the long-context limitations in real-world EHRs, provides deep grounding in clinical guidelines that exceeds standard RAG, and guarantees trustworthy assessment. CliCARE is designed to bridge this gap by integrating these capabilities into a unified pipeline.

3 CliCARE

In this section, we present the **CliCARE** framework, as illustrated in Figure 2. CliCARE is designed to systematically analyze long-form, unstructured cancer EHRs to generate a clinically grounded clinical summary and clinical recommendations. A key feature of this design is its extensibility: the guideline knowledge graph can be efficiently updated and expanded to accommodate new clinical evidence and corresponding guidelines.

3.1 EHR-to-TKGs Transformation

The initial stage of CliCARE transforms raw, multi-year EHRs from unstructured text into patient-centric TKGs, effectively addressing the fundamental challenge of long-context temporal reasoning.

Event Extraction. The complete EHR for each patient p can be formalized as a sequence of documents $D_p = (d_{\tau_1}, d_{\tau_2}, \ldots, d_{\tau_n})$ ordered by timestamps $T_p = \{\tau_1, \tau_2, \ldots, \tau_n\}$, where each document d_{τ_i} is an unstructured or semi-structured clinical text at time τ_i. To manage this extensive text sequence, we developed an efficient context processing pipeline, f_{pipeline}, to systematically compress, refine, and structure the raw text before it is input into the final pathway generation model.

$$E_p = f_{\text{pipeline}}(D_p) \tag{1}$$

Here, E_p represents a structured sequence of key clinical events. Specifically, the core of f_{pipeline} is an extractive summarization module based on the Longformer model.

Given the computational cost of processing the entire D_p, we partition the document sequence into the most recent clinical note, d_{τ_n}, and the historical records, $D_p^{\text{hist}} = (d_{\tau_1}, \ldots, d_{\tau_{n-1}})$. We utilize a Longformer model (Beltagy, Peters, and Cohan 2020), $\mathcal{M}_{\text{LF}}$, pre-trained on clinical text, to process the extensive historical records D_p^{hist}:

$$S_p^{\text{hist}} = \mathcal{M}_{\text{LF}}(D_p^{\text{hist}}; \theta_{\text{LF}}) \tag{2}$$

where θ_{LF} are the model parameters and S_p^{hist} is a summary text that includes the most informative sentences from the historical records. This summary effectively functions as the patient's past medical history. The most recent clinical note, d_{τ_n}, is regarded as the history of the present illness. The final structured event sequence E_p is created by concatenating these two components, thus providing a chronologically coherent and condensed patient history. From this combined text, key clinical facts—such as diagnostic confirmations, staging updates, treatment regimens, biomarker trends, and imaging assessments—are identified and organized into discrete events using BERT for information extraction(Yang, Wang, and Li 2021; Huang, Altosaar, and Ranganath 2019).

TKG Instantiation. Extracted event sequences E_p are organized into a patient-centric Clinical TKG, denoted as $G_t = (E_t, R_t, T)$, where E_t is the set of entities, R_t the set of relations, and T the set of timestamps. To enrich the TKG with standardized medical knowledge, we first construct a general, static biomedical knowledge graph $G_B = (\mathcal{E}_B, \mathcal{R}_B)$, where $\mathcal{E}_B$ contains standardized medical concepts and $\mathcal{R}_B$ represents the relations. For each patient, letting $\mathcal{E}_p$ be the set of raw clinical entities extracted from the patient's record, we instantiate a personalized TKG G_t by linking extracted clinical entities from the patient's record to the concepts in G_B. This uses an entity linking function $\phi : \mathcal{E}_p \to \mathcal{E}_B$, which maps textual mentions in the EHR to their corresponding canonical entries in the biomedical ontology. Each entity $e \in \mathcal{E}_t$ is represented as $e = (e_B, \tau, A)$, where $e_B \in \mathcal{E}_B$ is the linked standard entity, $\tau \in T$ is the event timestamp, and A is a set of event-specific attributes.

The TKG employs a hierarchical timestamp granularity by assigning precise timestamps $\tau \in T$ only to macro-level clinical encounters, while linking intra-encounter events through relative temporal relations, thereby mirroring the structure of real-world clinical records designed to capture the dynamic evolution of a patient's disease course.

3.2 Trajectory-Guideline Alignment

To integrate real-world patient data with normative medical knowledge, this stage aligns the descriptive patient TKG with a prescriptive guideline KG through a training-free fusion pipeline, as illustrated in Figure 3.

Knowledge Formalization. Our guideline knowledge graph, G_g, is a normative, static graph formalized as $G_g = (E_g, R_g)$. It is constructed based on authoritative CPGs, where E_g represents abstract medical concepts such as *Cancer*, *ClinicalSituation*, and *Treatment*. The edges, R_g, represent logical and recommendation relationships, collectively forming a graph that represents an idealized clinical workflow. We first perform a basic entity-level alignment between the patient's temporal graph, G_t, and G_g using medical ontologies. Subsequently, the clinical history for each patient p is organized from G_t into a time-ordered sequence of Clinical Events, resulting in a Temporal Trajectory $Tr_p = \langle e_1, e_2, \ldots, e_m \rangle$. The set of all patient trajectories is denoted as $\{Tr_p\}_{p=1}^{P}$. Concurrently, we systematically enumerate all possible normative treatment workflows from G_g to form a set of paths $\{Pa_k\}_{k=1}^{K}$, where each path $Pa_k = \langle s_1, s_2, \ldots, s_l \rangle$ represents a recommended sequence of guideline steps.

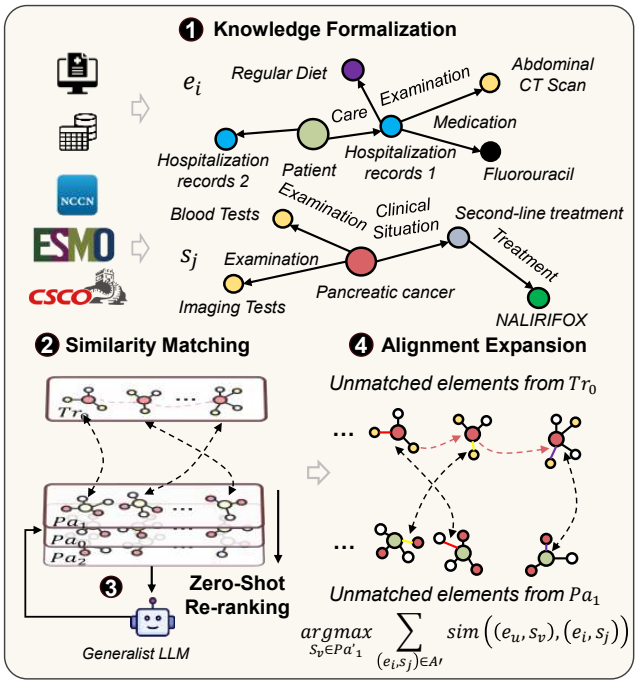

Figure 3: Trajectory-Guideline Alignment workflow. It fuses patient data with guidelines via semantic matching, LLM-based Reranking, and iterative bootstrapping expansion to create a comprehensive, evidence-grounded mapping.

Similarity Matching. We developed a global matching strategy based on deep semantic representations to directly assess the similarity between the entire patient trajectory and each candidate guideline path. Specifically, the matching score for a candidate path $Pa_k = \langle s_1, s_2, \ldots, s_l \rangle$ with the patient trajectory Tr_p is computed as follows. First, for each step s_j in the guideline path, we identify the most semantically similar event in the patient's event trajectory Tr_p. This similarity is calculated using a BERT model, f_{BERT}, pre-trained on biomedical text. Finally, we aggregate the best-match similarities for all steps to derive the total score for the path. Formally, the matching score is defined as:

$$\text{Score}(Tr_p, Pa_k) = \sum_{j=1}^{l} \max_{e_i \in Tr_p} \Big(\cos\_\text{sim}(f_{\text{BERT}}(\text{desc}(s_j)), f_{\text{BERT}}(\text{desc}(e_i))) \Big) \tag{3}$$

where the desc$(\cdot)$ function retrieves the text description of a node, and cos_sim computes the cosine similarity between two vectors. A higher score indicates a better alignment between the normative path Pa_k and the patient's experience. The optimal matching path Pa^* is then determined by selecting the candidate with the highest score:

$$Pa^* = \arg\max_{Pa_k} \text{Score}(Tr_p, Pa_k) \tag{4}$$

This method relies primarily on the deep semantic understanding provided by BERT, transcending simple lexical matching to capture conceptual associations between clinical events and guideline steps, thereby facilitating precise trajectory-path alignment.

LLM-based Reranking. The method described above generates a ranked list of candidate alignment paths for each patient trajectory. However, purely algorithmic matching may fail to capture the complexities of clinical logic. Therefore, we introduce an LLM as a Clinical Reasoner, f_{LLM}, to perform reranking. We provide the LLM with a rich-context prompt in a zero-shot manner, which includes the patient's trajectory Tr_p, the top-N candidate normative paths $\{Pa_1, \ldots, Pa_N\}$, and their corresponding matching scores $\{\text{Score}_1, \ldots, \text{Score}_N\}$. The LLM's task is to evaluate which candidate alignment is the most clinically plausible and to output a reranked list: $\langle Pa'_1, \ldots, Pa'_N \rangle = f_{LLM}(Tr_p, \{\langle Pa_k, \text{Score}_k \rangle\}_{k=1}^{N})$.

Alignment Expansion. To further enhance the coverage and accuracy of our alignments, we introduce an expansion stage inspired by bootstrapping techniques (Sun et al. 2018). After the LLM reranking, the top-ranked alignment path, Pa'_1, and its corresponding aligned node pairs serve as a high-confidence seed set, A'. We then iteratively expand this set. For each unaligned event e_u in the patient trajectory, the framework seeks to identify the best corresponding node $\hat{s}$ from the entire guideline path Pa'_1. The selection process is not based solely on direct similarity; rather, it considers how well the candidate pair (e_u, s_v) coheres with the entire set of existing high-confidence alignments in A'. This is accomplished by selecting the guideline node s_v that maximizes the sum of consistency scores with all established pairs in the seed set. The process is formalized as follows:

$$\hat{s} = \arg\max_{s_v \in Pa'_1} \sum_{(e_i, s_j) \in A'} \text{sim}((e_u, s_v), (e_i, s_j)) \tag{5}$$

where sim is a function that measures the consistency between a candidate pair (e_u, s_v) and an existing seed pair (e_i, s_j). This function utilizes the semantic representations derived from the language model f_{BERT} to compute the similarity between the corresponding nodes within the pairs. A high consistency score indicates that the semantic relationship between the patient event e_u and the candidate guideline node s_v is analogous to the established, high-confidence relationship between the seed event e_i and the guideline node s_j. This approach enables us to leverage established strong associations to infer new alignment relationships, thereby expanding our alignment set A'. After determining the final expanded alignment path, we employ a principled fusion strategy to enrich the guideline knowledge graph G_g with evidence from the patient trajectory Tr_p. Ultimately, this alignment process produces a robust, evidence-fused knowledge representation that serves as a direct, high-quality context for an LLM to generate its final clinical summary and clinical recommendation.

4 Experiments

4.1 Evaluation Method

To assess the quality of generated text, we developed an **Expert-Validated LLM-as-a-Judge** component. This component assesses two primary tasks: retrospective Clinical Summary, referred to as $T_{\mathbf{CS}}$, and prospective Clinical Recommendation, referred to as $T_{\mathbf{CR}}$. Our methodology employs a concise, four-dimensional rubric, co-designed with senior oncologists, to assess *Factual Accuracy*, *Completeness & Thoroughness*, *Clinical Soundness*, and *Actionability & Relevance*. The LLM judge is prompted to assign a score ranging from 1 (poor) to 5 (excellent) for each dimension.

To address the systematic biases of LLM judges, including positional bias, verbosity, and self-enhancement (Zheng et al. 2023; Wang et al. 2024), and after verifying their presence in our specific context, we implement a robust two-part mitigation protocol. First, to ensure rating stability and reduce model bias, we create a judging ensemble composed of three powerful LLMs, such as GPT-4.1 (Achiam et al. 2023), Claude 4.0 Sonnet (Anthropic 2025), and Gemini 2.5 Pro (Comanici et al. 2025), using their averaged score. Second, to eliminate ordering effects, all items are presented in a randomly shuffled sequence during the evaluation.

We validated the ratings of our LLM judge against those of three experienced oncologists using a subset of data to ensure reliability. We employed Spearman's rank correlation coefficient, denoted as ρ, a non-parametric measure that assesses the monotonic relationship between the LLM's and the experts' rankings. The coefficient is calculated as:

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)} \tag{6}$$

where n is the number of samples and d_i is the difference in ranks for each sample. A high correlation provides evidence that our LLM judge functions as a reliable and scalable proxy for expert assessment. This justifies its application for large-scale evaluations throughout our experiments.

4.2 Datasets

We evaluated our framework using two large-scale clinical datasets. The first dataset is a private Chinese collection, referred to as CancerEHR. It contains longitudinal records for 2,000 patients from Liaoning Cancer Hospital. These records span extensive periods—some exceeding two decades—resulting in inputs of up to 20,000 tokens. The dataset includes a variety of data types, such as physicians' orders, laboratory results, and surgical notes. The second dataset is derived from the publicly available MIMIC-IV dataset (Johnson et al. 2023), filtered to include only patients with cancer-related diagnoses, which we refer to as MIMIC-Cancer. This dataset provides a focus on disease progression similar to CancerEHR; however, the language and data

structure differ, offering a robust test of our method's generalizability. For brevity, in the subsequent implementation and results sections, we will refer to the two datasets as $D_{\mathbf{CEHR}}$ for CancerEHR and $D_{\mathbf{MC}}$ for MIMIC-Cancer. Similarly, the two primary tasks will be abbreviated as $T_{\mathbf{CS}}$ for Clinical Summary and $T_{\mathbf{CR}}$ for Clinical Recommendation.

4.3 Baselines

We compare our proposed CliCARE framework against a variety of robust baseline methods. These include standard RAG pipelines implemented with powerful open-source models such as Mistral-7B and its instruction-tuned variant (Jiang et al. 2023), Qwen3-8B (Yang et al. 2025a), and the domain-specific BioMistral-7B (Labrak et al. 2024). Additionally, we evaluate more advanced KG-enhanced RAG techniques designed for long-context or knowledge-intensive tasks. The selected methods include BriefContext (Zhang et al. 2025), which employs a Map-Reduce strategy, as well as several Graph-Aware RAG frameworks, such as GNN-RAG (Mavromatis and Karypis 2025), KG2RAG (Zhu et al. 2025), and the healthcare-focused MedRAG (Zhao et al. 2025). We evaluate these RAG frameworks using both the open-source models and leading generalist models, including Deepseek-R1 (Guo et al. 2025), Gemini 2.5 Pro, GPT-4.1, and Claude-4.0-Sonnet.

4.4 Implementation Details

In the knowledge graph alignment stage, the threshold is set to 0.7 when using BERT to calculate semantic cosine similarity in the initial step. During the fine-tuning stage, we divided the 2,000-sample dataset into a training set of 1,800 samples and a test set of 200 samples, with 10% of the training data reserved for validation. The key hyperparameters for training include a batch size of 1, a maximum context length of 20,000 tokens, and an initial learning rate of 5×10^{-5} with a cosine scheduler. We utilized BF16 for mixed-precision training, set the maximum output length to 4,096 tokens, and trained for 3 epochs. All experiments were conducted using a configuration of 4 NVIDIA A800 GPUs.

4.5 Experimental Results

High Agreement with Clinician Judgments. Acknowledging the limitations of traditional metrics for clinical tasks, we validated our LLM-as-a-Judge protocol against three experienced oncologists. To ensure a feasible yet representative assessment, we created a validation subset by randomly sampling generated outputs from eight different models. Our protocol minimized bias by evaluating these outputs column-wise and presenting the Clinical Summary and Recommendation tasks together for a comprehensive review. The results demonstrate a strong positive correlation between the automated ratings and those of the experts. Specifically, the Spearman's rank correlation ρ between our LLM judge's scores and the physicians' mean scores was approximately 0.7, confirming that our metric serves as a reliable proxy for human expert judgment.

Method	$D_{\mathbf{CEHR}}$		$D_{\mathbf{MC}}$	
	$T_{\mathbf{CS}}$	$T_{\mathbf{CR}}$	$T_{\mathbf{CS}}$	$T_{\mathbf{CR}}$
Qwen-3-8B				
StandardRAG	1.485	1.527	2.475	2.467
BriefContext	<u>2.681</u>	<u>2.701</u>	<u>2.571</u>	<u>2.497</u>
MedRAG*	2.333	2.366	2.495	2.462
KG2RAG*	2.595	2.558	2.317	2.166
GNN-RAG*	2.508	2.527	2.194	2.182
CliCARE	**3.173**	**3.215**	**2.575**	**2.544**
Gemini 2.5 Pro				
StandardRAG	2.735	2.818	3.563	3.556
BriefContext	<u>4.527</u>	4.468	4.354	4.233
MedRAG*	4.470	<u>4.576</u>	**4.476**	<u>4.323</u>
KG2RAG*	3.845	3.942	3.747	3.797
GNN-RAG*	3.607	3.552	3.683	3.588
CliCARE	**4.976**	**4.965**	<u>4.398</u>	**4.333**

Table 1: CliCARE Outperforms RAG Baselines on Clinical Generation Tasks. Scores are assigned by our Expert-Validated LLM-as-a-Judge. The asterisk (*) denotes KG-enhanced RAG variants.

CliCARE Significantly Outperforms Baselines. As detailed in Table 1, CliCARE demonstrates a clear performance advantage over a suite of robust baselines, including both context-aware and KG-enhanced RAG methods. The benefits of our framework are most pronounced when paired with a powerful model on complex datasets. With Gemini 2.5 Pro, CliCARE achieves impressive Clinical Summary and Recommendation scores of 4.976 and 4.965, respectively, on the challenging CancerEHR dataset. This performance significantly surpasses that of other methods. For instance, while BriefContext achieves a commendable score of 4.527, it does so through a costly Map-Reduce strategy that involves multiple LLM calls, underscoring the efficiency of CliCARE's approach. Even when utilizing a smaller model like Qwen-3-8B, CliCARE obtains scores of 3.173 and 3.215, substantially outperforming all baselines on the same complex dataset. This success is attributed to CliCARE's TKG transformation, which effectively organizes the chaotic, longitudinal patient records and overcomes the fragmented retrievals that hinder other RAG pipelines.

Structured Knowledge is Key for Complex EHRs. As demonstrated in Table 2, our framework's knowledge structuring offers a significant advantage over standard RAG. The performance uplift is most pronounced on the complex CancerEHR dataset, where nearly all models exhibit substantial gains. Notably, the improvements for Qwen-3-8B and Deepseek-R1 are the largest in their respective groups, with their Clinical Summary scores increasing by +1.688 and +2.279, respectively. This underscores that even advanced models require a coherent structure for effective reasoning on complex records. On the simpler MIMIC-Cancer dataset, while the absolute gains are smaller, CliCARE still delivers a distinct and consistent advantage. For instance, it elevates the score of a strong baseline like GPT-4.1 from 4.419 to 4.737, a gain of +0.318. While the uplift is nearly universal,

Method	Standard RAG				CliCARE			
	D_{CEHR}		D_{MC}		D_{CEHR}		D_{MC}	
	T_{CS}	T_{CR}	T_{CS}	T_{CR}	T_{CS}	T_{CR}	T_{CS}	T_{CR}
Mistral-v0.1-7B	1.120	1.164	2.505	2.505	1.407 (+0.287)	1.526 (+0.362)	2.575 (+0.070)	2.514 (+0.009)
Mistral-Instruct-v0.1-7B	1.054	1.070	2.183	2.115	1.274 (+0.220)	1.355 (+0.285)	2.231 (+0.048)	2.071 (-0.044)
Biomistral-7B	1.161	1.098	**2.785**	**2.698**	1.548 (+0.387)	1.529 (+0.431)	**2.903** (**+0.118**)	**2.742** (+0.044)
Qwen-3-8B	**1.485**	**1.527**	2.475	2.467	**3.173** (**+1.688**)	**3.215** (**+1.688**)	2.575 (+0.100)	2.544 (**+0.077**)
Gemini-2.5-Pro	**2.735**	2.818	3.563	3.556	**4.976** (+2.241)	**4.965** (+2.147)	4.398 (**+0.835**)	4.333 (**+0.777**)
GPT-4.1	2.667	2.873	**4.419**	**4.429**	4.690 (+2.023)	4.703 (+1.830)	**4.737** (+0.318)	**4.676** (+0.247)
Deepseek-R1	2.667	**2.878**	4.016	4.000	4.946 (**+2.279**)	4.935 (**+2.057**)	4.409 (+0.393)	4.319 (+0.319)
Claude-4.0-Sonnet	2.417	2.624	3.898	3.868	3.893 (+1.476)	3.924 (+1.300)	4.183 (+0.285)	4.110 (+0.242)

Table 2: Model performance with standard RAG versus the CliCARE framework. Applying CliCARE provides a substantial performance uplift for most models.

we do note a single case of minor performance degradation, confirming the intricate nature of these tasks.

Method	D_{CEHR}		D_{MC}	
	T_{CS}	T_{CR}	T_{CS}	T_{CR}
CliCARE (Q)	**3.173**	**3.215**	**2.575**	**2.544**
w/o Exp.	3.012 (-)	3.035 (-)	2.075 (-)	2.110 (-)
w/o Rerank	2.857 (-)	2.866 (-)	2.000 (-)	1.962 (-)
w/o Comp.	1.485 (-)	1.527 (-)	2.475 (+)	2.467 (+)
CliCARE (G)	**4.976**	**4.965**	**4.398**	**4.333**
w/o Exp.	4.619 (-)	4.630 (-)	3.737 (-)	3.786 (-)
w/o Rerank	4.542 (-)	4.628 (-)	3.774 (+)	3.824 (+)
w/o Comp.	2.735 (-)	2.818 (-)	3.563 (-)	3.556 (-)

Table 3: Ablation study on CliCARE framework components. Q denotes Qwen-3-8B and G denotes Gemini-2.5-Pro. Exp., Rerank and Comp. signify the removal of Alignment Expansion, LLM-based Reranking, and TKG-based Compression, respectively. The symbols (+)/(-) indicate a performance increase/decrease compared to the row above.

Ablation Study. Our ablation study, with results in Table 3, reveals the nuanced role of each module. This is most evident for the Qwen model on the simpler MIMIC-Cancer dataset; removing TKG-based Compression paradoxically boosts the scores to 2.475 and 2.467. This result is substantially better than when LLM-based Reranking is removed, which causes a drop to 2.000, suggesting aggressive compression can be counterproductive for shorter records. A similar, though less pronounced, positive effect is observed for the Gemini model under the same conditions. Conversely, on the complex CancerEHR dataset, the consistent, significant performance drops from any ablation highlight that the full, integrated CliCARE framework is crucial for achieving optimal performance.

Performance Analysis Based on EHR Length. Further analysis of record length reveals distinct performance patterns, as detailed in Table 4. The smaller model, Qwen-3-8B, performs optimally on Short length records but exhibits a decline in quality when processing the longest records, particularly with the complex CancerEHR data. In contrast, the more powerful Gemini-2.5-Pro model demonstrates strong and consistent performance across all record lengths. Notably, when guided by the CliCARE framework, it achieves its highest scores on the longest record segments for both datasets. This suggests that CliCARE effectively organizes extensive clinical histories and enables advanced models to leverage richer context for enhanced reasoning.

Method		D_{CEHR}		D_{MC}	
		T_{CS}	T_{CR}	T_{CS}	T_{CR}
CliCARE(Qwen-3-8B)					
All	(100%)	3.173	3.215	2.575	2.544
Short	(0˜33%)	3.228	**3.345**	**2.850**	**2.645**
Medium	(33%˜66%)	**3.267**	3.283	2.429	2.450
Long	(66%˜100%)	2.976	2.983	2.460	2.533
CliCARE(Gemini-2.5-Pro)					
All	(100%)	4.976	4.965	4.398	4.333
Short	(0˜33%)	4.982	4.937	4.362	4.311
Medium	(33%˜66%)	4.962	4.976	4.365	4.317
Long	(66%˜100%)	**4.988**	**4.982**	**4.467**	**4.373**

Table 4: Performance analysis by EHR length. Segments are stratified by percentile (0-33%, 33-66%, 66-100%). Average token counts for CancerEHR segments are 4875, 6303, 9411; for MIMIC-Cancer, 4070, 5176, 6463.

5 Conclusion

We introduced CliCARE, a framework for reliable clinical decision support that transforms cancer EHRs into Temporal Knowledge Graphs and aligns them with clinical guidelines. This approach addresses key challenges in long-context reasoning and hallucination, enabling both small specialist and large generalist models to significantly outperform baselines. We also validated a robust LLM-as-a-Judge protocol that correlates highly with expert oncologist assessments, representing a significant advancement toward deploying trustworthy AI in clinical practice. Future work should focus on generalizing this framework to a wider range of clinical domains.

Acknowledgments

This work was supported by the Natural Science Foundation of Liaoning Province, China (No. 2025-MS-036) and the Doctoral Student Scientific Research Innovation Project of Northeastern University (No. 02190022125036).

References

2016. Allocation of Physician Time in Ambulatory Practice: A Time and Motion Study in 4 Specialties. *Annals of Internal Medicine*, 165.

Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Anthropic. 2025. Introducing Claude 4. https://www.anthropic.com/news/claude-4. Accessed: 2025-11-10.

Bai, Y.; Lv, X.; Zhang, J.; Lyu, H.; Tang, J.; Huang, Z.; Du, Z.; Liu, X.; Zeng, A.; Hou, L.; Dong, Y.; Tang, J.; and Li, J. 2024. LongBench: A Bilingual, Multitask Benchmark for Long Context Understanding. arXiv:2308.14508.

Beltagy, I.; Peters, M. E.; and Cohan, A. 2020. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.

Comanici, G.; Bieber, E.; Schaekermann, M.; Pasupat, I.; Sachdeva, N.; Dhillon, I.; Blistein, M.; Ram, O.; Zhang, D.; Rosen, E.; et al. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

Fan, Y.; Sun, H.; Xue, K.; Zhang, X.; Zhang, S.; and Ruan, T. 2024. Medodyssey: A medical domain benchmark for long context evaluation up to 200k tokens. *arXiv preprint arXiv:2406.15019*.

Gao, Y.; Li, R.; Croxford, E.; Caskey, J.; Patterson, B. W.; Churpek, M.; Miller, T.; Dligach, D.; and Afshar, M. 2025. Leveraging medical knowledge graphs into large language models for diagnosis prediction: Design and application study. *Jmir Ai*, 4: e58670.

Guo, D.; Yang, D.; Zhang, H.; Song, J.; Zhang, R.; Xu, R.; Zhu, Q.; Ma, S.; Wang, P.; Bi, X.; et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Hager, P.; Jungmann, F.; Holland, R.; Bhagat, K.; Hubrecht, I.; Knauer, M.; Vielhauer, J.; Makowski, M.; Braren, R.; Kaissis, G.; and Rueckert, D. 2024. Evaluation and mitigation of the limitations of large language models in clinical decision-making. *Nature Medicine*, 30(9): 2613–2622.

Hao, Y.; Qiu, Z.; Holmes, J.; Löckenhoff, C. E.; Liu, W.; Ghassemi, M.; and Kalantari, S. 2025. Large language model integrations in cancer decision-making: a systematic review and meta-analysis. *npj Digital Medicine*, 8(1): 450.

Huang, K.; Altosaar, J.; and Ranganath, R. 2019. Clinicalbert: Modeling clinical notes and predicting hospital readmission. *arXiv preprint arXiv:1904.05342*.

Jiang, A. Q.; Sablayrolles, A.; Mensch, A.; Bamford, C.; Chaplot, D. S.; de las Casas, D.; Bressand, F.; Lengyel, G.; Lample, G.; Saulnier, L.; et al. 2023. Mistral 7B. *arXiv preprint arXiv:2310.06825*.

Jiang, Z.; Han, X.; Liu, Z.; and Sun, M. 2024. Efficient Knowledge Infusion via KG-LLM Alignment. In *Findings of the Association for Computational Linguistics: ACL 2024*.

Johnson, A. E. W.; Bulgarelli, L.; Shen, L.; Gayles, A.; Shammout, A.; Horng, S.; Pollard, T. J.; Hao, S.; Moody, B.; Gow, B.; wei H. Lehman, L.; Celi, L. A.; and Mark, R. G. 2023. MIMIC-IV, a freely accessible electronic health record dataset. *Scientific Data*, 10(1): 1.

Khan, A.; Wu, T.; and Chen, X. 2024. Data Management Opportunities in Unifying Large Language Models+ Knowledge Graphs. In *Workshops at the 50th International Conference on Very Large Data Bases*. VLDB.org.

Khatib, H. S. A.; Mittal, S.; Rahimi, S.; Marhamati, N.; and Bozorgzad, S. 2025. From Patient Consultations to Graphs: Leveraging LLMs for Patient Journey Knowledge Graph Construction. *arXiv preprint arXiv:2503.16533*.

Labrak, Y.; Vidal, M.; Appert, L.; Ladhari, B.; Bonnet, P.; Ghannay, S.; Gandon, F.; and Nader, F. 2024. BioMistral: A Collection of Open-Source Large Language Models for Medical Domains. *arXiv preprint arXiv:2402.10374*.

Lewis, P.; Perez, E.; Piktus, A.; Petroni, F.; Karpukhin, V.; Goyal, N.; Küttler, H.; Lewis, M.; Yih, W.-t.; Rocktäschel, T.; et al. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems*, volume 33, 9459–9474.

Li, M.-H.; Huang, J. J.; Yeung, J.; Chen, J.; Liu, S.; Zhang, Z.; Lee, E. K.; and Wang, S.-C. 2024. CancerLLM: A large language model in cancer domain. arXiv:2406.10459.

Liu, F.; Li, Z.; Zhou, H.; Yin, Q.; Yang, J.; Tang, X.; Luo, C.; Zeng, M.; Jiang, H.; Gao, Y.; Nigam, P.; Nag, S.; Yin, B.; Hua, Y.; Zhou, X.; Rohanian, O.; Thakur, A.; Clifton, L.; and Clifton, D. A. 2024. Large Language Models Are Poor Clinical Decision-Makers: A Comprehensive Benchmark. In Al-Onaizan, Y.; Bansal, M.; and Chen, Y.-N., eds., *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 13696–13710. Miami, Florida, USA: Association for Computational Linguistics.

Maushagen, J.; Hext, T.; Meroño-Peñuela, A.; and Pellegrini, T. 2024. Populating CSV Files from Unstructured Text with LLMs for KG Generation with RML. In *SEMANTiCS (Posters, Demos, Workshops & Tutorials)*.

Mavromatis, C.; and Karypis, G. 2025. GNN-RAG: Graph Neural Retrieval for Efficient Large Language Model Reasoning on Knowledge Graphs. In *Findings of the Association for Computational Linguistics: ACL 2025*, 16682–16699.

Oniani, D.; Wu, X.; Visweswaran, S.; Kapoor, S.; Kooragayalu, S.; Polanska, K.; and Wang, Y. 2024. Enhancing large language models for clinical decision support by incorporating clinical practice guidelines. In *2024 IEEE 12th International Conference on Healthcare Informatics (ICHI)*, 694–702. IEEE.

Rajashekar, N. C.; Shin, Y. E.; Pu, Y.; Chung, S.; You, K.; Giuffre, M.; Chan, C. E.; Saarinen, T.; Hsiao, A.; Sekhon, J.; Wong, A. H.; Evans, L. V.; Kizilcec, R. F.; Laine, L.; Mccall, T.; and Shung, D. 2024. Human-Algorithmic Interaction Using a Large Language Model-Augmented Artificial Intelligence Clinical Decision Support System. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI '24. New York, NY, USA: Association for Computing Machinery. ISBN 9798400703300.

Singhal, K.; Azizi, S.; Tu, T.; Mahdavi, S. S.; Wei, J.; Chung, H. W.; Scales, N.; Tanwani, A.; Cole-Lewis, H.; Pfohl, S.; et al. 2023. Large language models encode clinical knowledge. *Nature*, 620(7972): 172–180.

Sun, Z.; Hu, W.; Zhang, Q.; and Qu, Y. 2018. Bootstrapping Entity Alignment with Knowledge Graph Embedding. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, 4396–4402. International Joint Conferences on Artificial Intelligence Organization.

Tadevosyan, M.; Nori, A.; Singh, A.; et al. 2025. Articulate Medical Intelligence Explorer. *Nature*. In Press.

Tu, T.; Azizi, S.; Driess, D.; Schaekermann, M.; Amin, M.; Chang, P.-C.; Carroll, A.; Aha, D.; Corrado, G. S.; Fleet, D. J.; et al. 2024. Towards conversational diagnostic AI. *Nature Medicine*, 1–13.

Wang, L. L.; Otmakhova, Y.; DeYoung, J.; Truong, T. H.; Kuehl, B.; Bransom, E.; and Wallace, B. 2023. Automated Metrics for Medical Multi-Document Summarization Disagree with Human Evaluations. In Rogers, A.; Boyd-Graber, J.; and Okazaki, N., eds., *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 9871–9889. Toronto, Canada: Association for Computational Linguistics.

Wang, P.; Li, L.; Chen, L.; and Lin, B. Y. 2024. Large Language Models are Inconsistent and Biased Evaluators. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 461–472.

Wang, Z.; Zhu, Y.; Zhao, H.; Zheng, X.; Sui, D.; Wang, T.; Tang, W.; Wang, Y.; Harrison, E.; Pan, C.; et al. 2025. Colacare: Enhancing electronic health record modeling through large language model-driven multi-agent collaboration. In *Proceedings of the ACM on Web Conference 2025*, 2250–2261.

Warner, J. L.; Dymshyts, D.; Hripcsak, G.; and Miller, R. S. 2020. Data Fragmentation and the Case for a National Cancer Data Ecosystem: A Report From the American Society of Clinical Oncology. *JCO Clinical Cancer Informatics*, 4: 1022–1031.

Xu, R.; Shi, W.; Yu, Y.; Zhuang, Y.; Jin, B.; Wang, M. D.; Ho, J.; and Yang, C. 2024. RAM-EHR: Retrieval Augmentation Meets Clinical Predictions on Electronic Health Records. In Ku, L.-W.; Martins, A.; and Srikumar, V., eds., *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 754–765. Bangkok, Thailand: Association for Computational Linguistics.

Yang, A.; Li, A.; Yang, B.; Zhang, B.; and .... 2025a. Qwen3 Technical Report. *arXiv preprint arXiv:2505.09388*.

Yang, F.; Wang, X.; and Li, J. 2021. BERT model in Chinese clinical natural language processing: exploration and research. https://github.com/trueto/medbert. Accessed: 2024-07-09.

Yang, Q.; Chen, J.; Sun, Y.; Wang, Y.; and Tan, T. 2025b. Fine-tuning medical language models for enhanced long-contextual understanding and domain expertise.

Yu, H.; and McQuade, F. 2025. Rag-kg-il: A multi-agent hybrid framework for reducing hallucinations and enhancing llm reasoning through rag and incremental knowledge graph learning integration. *arXiv preprint arXiv:2503.13514*.

Zhang, G.; Xu, Z.; Jin, Q.; Chen, F.; Fang, Y.; Liu, Y.; Rousseau, J. F.; Xu, Z.; Lu, Z.; Weng, C.; et al. 2025. Leveraging long context in retrieval augmented language models for medical question answering.

Zhao, X.; Liu, S.; Yang, S. Y.; et al. 2025. Medrag: Enhancing retrieval-augmented generation with knowledge graph-elicited reasoning for healthcare copilot. In *Proceedings of the ACM on Web Conference 2025*, 4442–4457.

Zheng, L.; Chiang, W.-L.; Sheng, Y.; Zhuang, S.; Wu, Z.; Zhuang, Y.; Lin, Z.; Li, Z.; Li, D.; Xing, E. P.; Zhang, H.; Gonzalez, J. E.; and Stoica, I. 2023. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. arXiv:2306.05685.

Zhu, X.; Xie, Y.; Liu, Y.; Li, Y.; and Hu, W. 2025. Knowledge graph-guided retrieval augmented generation. *arXiv preprint arXiv:2502.06864*.

Zhu, Y.; Ren, C.; Xie, S.; Liu, S.; Ji, H.; Wang, Z.; Sun, T.; He, L.; Li, Z.; Zhu, X.; et al. 2024a. Realm: Rag-driven enhancement of multimodal electronic health records analysis via large language models. *arXiv preprint arXiv:2402.07016*.

Zhu, Y.; Wang, Z.; Gao, J.; Tong, Y.; An, J.; Liao, W.; Harrison, E. M.; Ma, L.; and Pan, C. 2024b. Prompting large language models for zero-shot clinical prediction with structured longitudinal electronic health record data. *arXiv preprint arXiv:2402.01713*.