TIMER: TEMPORAL INSTRUCTION MODELING AND EVALUATION FOR LONGITUDINAL CLINICAL RECORDS

Hejie Cui*, **Alyssa Unell***, **Bowen Chen, Jason Alan Fries, Emily Alsentzer, Sanmi Koyejo & Nigam Shah**
Stanford University
Stanford, CA 94305, USA
{hcui,aunell}@stanford.edu

ABSTRACT

Large language models (LLMs) have emerged as promising tools for assisting in medical tasks, yet processing Electronic Health Records (EHRs) presents unique challenges due to their longitudinal nature. While LLMs' capabilities to perform medical tasks continue to improve, their ability to reason over temporal dependencies across multiple patient visits and time frames remains unexplored. We introduce **TIMER** (**T**emporal **I**nstruction **M**odeling and **E**valuation for Longitudinal Clinical **R**ecords), a synthetic data generation framework that incorporates temporal distribution of instructions as a critical dimension in both instruction evaluation and tuning for longitudinal clinical records. We develop TIMER-Bench, the first time-aware benchmark that evaluates temporal reasoning capabilities over longitudinal EHRs, as well as TIMER-Instruct, an instruction-tuning methodology for LLMs to learn reasoning over time. We demonstrate that models fine-tuned with TIMER-Instruct improve performance by 7.3% on human-generated benchmarks and 9.2% on TIMER-Bench, indicating that temporal instruction-tuning improves model performance for reasoning over EHR. Our code is available at TIMER.

1 INTRODUCTION

While many language models now handle context lengths of hundreds of thousands of tokens, their ability to reason across longitudinal documents and follow complex instructions remains limited Li et al. (2024b); Kuratov et al. (2024). This limitation is particularly critical in healthcare, where physicians routinely analyze electronic health records (EHRs) spanning multiple years and thousands of entries Huguet et al. (2020). While biomedical LLMs have shown promising results on well-structured tasks like answering USMLE questions and medical knowledge retrieval Singhal et al. (2023); Lu et al. (2024); Lucas et al. (2024), recent evaluations reveal their significant limitations in processing longitudinal patient information and in making clinical decisions over time Hager et al. (2024); Bedi et al. (2024). The gap between isolated question-answering performance and temporal reasoning ability impacts the practical utility of LLMs in healthcare.

Instruction tuning has proven useful for domain-specific tasks Zhang et al. (2023a), but data curation for instruction tuning over longitudinal clinical records faces significant challenges. The cognitive demands of processing lengthy medical documentation make physician instruction-answer pair generation both time-intensive and difficult to scale Wu et al. (2024b). While synthetic data generation offers a promising solution for data access challenges, the temporal aspects of instruction evaluation and tuning—particularly how information distribution across patient timelines affects model performance—remain poorly understood Scheller (2022). Moreover, deploying language models in high-stakes domains like healthcare requires rigorous evaluation in controlled settings.

To address these challenges, we introduce TIMER, a framework for evaluating and enhancing temporal reasoning capabilities of LLMs using synthetic data generated from longitudinal EHRs. We develop TIMER-Bench, a novel evaluation benchmark that leverages LLM-generated instruction-response pairs with explicit temporal evidence to enable systematic evaluation across different temporal distributions. Our analysis reveals critical limitations in existing LLMs' temporal reasoning capabilities, including poor temporal boundary adherence, inaccurate trend analysis, and chronological confusion. Building on these insights, we propose TIMER-Instruct, a methodology for synthetic temporal instruction tuning. Through careful analysis of synthetic data generation patterns, we identified a "lost-in-the-middle" bias where models overlook mid-timeline events. To address this, we investigated different temporal distributions of instructions: recency-focused, edge-focused, and uniform. Our evaluation demonstrates that while distribution alignment between training and evaluation generally improves performance, uniform distribution exhibits particular strength in mid-timeline reasoning tasks. Models tuned with our synthetic instruction data show promising improvements: 7.3% on MedAlign and 9.2% on TIMER-Bench.

This work advances synthetic data applications in healthcare through three key contributions: (1) identification of temporal distribution as a critical yet overlooked dimension in clinical language model evaluation and instruction-tuning, demonstrating how synthetic data uncovers temporal reasoning limitations, (2) TIMER-Bench, a synthetic benchmark that systematically assesses longitudinal reasoning capabilities across temporal distributions, and (3) TIMER-Instruct, a framework for temporal instruction tuning using synthetic data that achieves state-of-the-art performance on both human-curated and synthetic benchmarks. Our code is available at TIMER. TIMER-Bench will be released under a research data use agreement to support responsible evaluation.

2 Related Work

2.1 Synthetic Data Generation

Instruction tuning has emerged as a useful method for aligning models with user preferences Ouyang et al. (2022); Zhang et al. (2023a). While high-quality instruction-response pairs are essential for this process, their limited availability has led researchers to explore synthetic data generation as a scalable alternative Dubois et al. (2024); Long et al. (2024). Self-instruct demonstrates the feasibility of bootstrapping instruction-following capabilities through LLM-generated data Wang et al. (2023); Zhang et al. (2023b), enabling knowledge distillation from larger models to smaller yet capable ones Shirgaonkar et al. (2024). Recent research has further expanded the field by exploring techniques to enhance the quality and diversity of synthetic datasets Ge et al. (2024); Li et al. (2024a); van Breugel et al. (2024a), achieving promising results on benchmarks. Despite these advancements, challenges remain in ensuring the quality of synthetic data, including detecting and mitigating systematic biases and developing rigorous evaluation protocols van Breugel et al. (2024b).

2.2 Temporal Challenges in Reasoning with EHR

Electronic Health Records (EHRs) serve as digital repositories of patient care, containing structured data, unstructured clinical notes, and temporal data across visits Theodorou et al. (2023). Healthcare providers need to synthesize complex information to track disease progression, treatment responses, and temporal relationships between medical events Carrasco-Ribelles et al. (2023); Allam et al. (2021). However, existing EHR evaluation benchmarks inadequately assess temporal reasoning. MIMIC-derived datasets predominantly focus on single ICU visits averaging 7.2 days Wu et al. (2024b), while physician-curated benchmarks like MedAlign Fleming et al. (2024) show strong recency bias, with 55.3% of instructions concentrated in the last quarter of 10.7-year patient histories. In addition to the lack of temporal evaluation benchmarks, current medical instruction tuning approaches primarily focus on brief instruction responses and simple retrieval tasks Zhang et al. (2023b); Tran et al. (2024); Rohanian et al. (2024). These highlight the need for systematic evaluation and instruction-tuning for models' capabilities across comprehensive patient histories.

3 TIMER

We introduce TIMER, a framework for synthetic data generation which evaluates and enhances temporal reasoning capabilities of LLMs on longitudinal EHRs, as present in Figure 1. TIMER consists of two components: (1) TIMER-Bench, which generates evaluation sets with explicit time evidence integration, and (2) TIMER-Instruct, which improves models' longitudinal reasoning through temporal-aware synthetic instruction tuning.

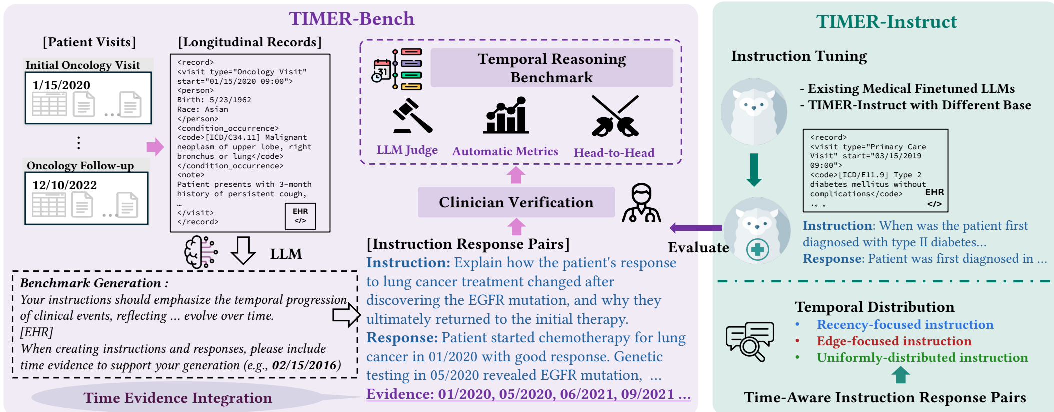

Figure 1: Overview of **TIMER** framework. Left: TIMER-Bench covers questions with explicit temporal evidence across different time periods in patient histories to assess longitudinal EHR reasoning. Right: TIMER-Instruct enables temporal instruction tuning with temporally-diverse examples.

3.1 TIMER-BENCH: EVALUATION BENCHMARK FOR LLM LONGITUDINAL REASONING

Benchmark generation process. We develop an approach for generating benchmarks that integrates time evidence. This method explicitly preserves and utilizes temporal relationships during the creation process, reflecting how clinicians interact with EHRs. As shown in the left panel of Figure 1, we begin by aggregating patient visits and converting them into XML-formatted longitudinal records. These records are then used as input for a language model to generate synthetic instruction-response pairs. We instruct the language model to provide date-time evidence $\mathbf{T}_i = \{T_{i,1}, ..., T_{i,n_i}\}$ for each instruction-response pair (Q_i, A_i) in the benchmark. This evidence connects the instance to related visits in the patient's timeline, resulting in tuples of $(Q_i, A_i, \mathbf{T}_i)$. This temporal grounding enables us to filter benchmark instances by the number of evidence timestamps it generates, so we can ensure that each question evaluates temporal reasoning between visits in a patient record.

We use real patient data from a single medical center to synthetically generate questions, assuring that raw patient data matches the layout of real patients in hospitals. The raw de-identified EHR data utilized in our study has been made accessible through a gated institutional portal on Redivis to promote transparency and responsible research practices. Data access is subject to approval via a user-level Data Use Agreement (DUA) and verification of valid CITI training certification. Institutional patients had previously consented, through an institutional privacy notice, to the research use of their fully de-identified medical records.

Clinical validation. To ensure benchmark quality, we conducted validation with three clinicians who reviewed 100 randomly sampled instruction-response pairs from TIMER-Bench. The clinicians evaluated three key aspects: clinical relevance, which measures alignment with real-world medical scenarios; temporal reasoning complexity, which assesses the depth of temporal synthesis required; and factual accuracy, which verifies perceived medical correctness. The evaluation results included average scores of 95/100 for clinical relevance, 80/100 for temporal reasoning complexity, and 98/100 for factual accuracy. Three clinicians evaluated each question, providing three scores for each axis: clinical relevance, complexity, and accuracy. The results show high inter-rater agreement (86% clinical relevance, 93% accuracy) with low standard deviations (4.32, 1.89 respectively). Complexity scoring, being qualitative, showed more variability but remained significantly above chance (53% observed agreement vs. 12.5% random chance; standard deviation 14.87 of the total assigned score). This validation process confirms that TIMER-Bench maintains high clinical authenticity while effectively testing models' long context temporal reasoning capabilities.

All clinicians completed mandatory CITI training covering the use of de-identified patient data, fully mirroring established ethical standards used in datasets such as MIMIC and PhysioNet. The clinicians voluntarily participate in annotation tasks, without any compensation or other incentives and agree to the public release of annotations. The clinicians interacted exclusively with de-identified data compliant with HIPAA Safe Harbor standards and their involvement was limited to data annotation. As such, their participation does not constitute human subjects research under 45 CFR 46 (the Common Rule).

3.2 TIMER-Instruct: Instruction Tuning for Longitudinal Clinical Records

To improve model performance on temporal reasoning tasks, we develop TIMER-Instruct, a synthetic instruction tuning methodology with time-aware instruction data.

Temporal variability in clinical records. A key challenge in analyzing temporal patterns across patient records is that even with fixed context lengths, the actual time spans can vary significantly due to irregular intervals between clinical events. In clinical records, events like diagnoses, medications, laboratory results, and procedures are distributed irregularly across time, reflecting the reality of patient care where visits may be clustered during acute episodes or spread out during stable periods.

Temporal distribution analysis. We introduce a normalized position metric to account for these varying timescales. For each time evidence $T_j \in \mathbf{T}_i$ of instruction-response pair (Q_i, A_i), we define the relative temporal position P_j as:

$$P_j = \frac{T_j - T_{min}}{T_{max} - T_{min}} \tag{1}$$

where T_j represents an evidence timestamp of this pair, and T_{min} and T_{max} are date time of visits that bound the context window. This normalization enables us to compare temporal patterns across patient records with different absolute time spans while maintaining relative temporal relationships. Using our normalized temporal position metric, we analyze the distribution of temporal evidence in model-generated instructions across patient timelines. Figure 2 reveals a striking "lost-in-the-middle" phenomenon in the default generation pattern.

Such edge-focused distribution indicates that when generating data, LLMs tend to pay more attention to early and recent events in long contexts, while neglecting the intermediate period. These inherent biases in temporal focus during the synthetic data generation process could result in tuned models that overlook important developments in patient timeline journeys.

Instruction tuning with various temporal distribution patterns. Motivated by the temporal biases observed in both model-generated and physician-generated instructions, we explore how temporal distributions in tuning data affect model performance. We construct three instruction-tuning sets from the same set of patient longitudinal records, each reflecting a distinct distribution of instructions' relative temporal position, as demonstrated in Figure 1 right panel:

- *Recency*: a recency-focused set that concentrates instructions in the last quartile of the timeline, similar to the temporal patterns observed in human-annotated datasets
- *Edge*: an edge-focused set which exhibits higher instruction density regarding visits at the context boundaries T_{min} and T_{max} of patient timelines, similar to the natural temporal patterns generated by large language models
- *Uniform*: a uniformly-distributed set that ensures balanced temporal coverage by maintaining consistent instruction density across all relative positions

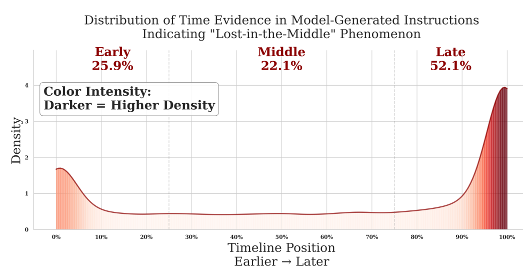

Figure 2: Using our normalized temporal position metric (x-axis: 0% to 100% of the timeline), we find that instructions strongly favor timeline edges, indicating middle periods receive significantly less attention.

By maintaining consistent size and patient timelines across these instruction sets and varying only their temporal distributions over the timeline, we isolate temporal positioning as the key variable of how training data patterns influence models' ability to reason across time points.

4 Experiments

4.1 Experiment Setup

Datasets. TIMER-Bench and TIMER-Instruct utilize patient data from an academic medical center's research data repository, which contains records from its associated healthcare system, includ-

Table 1: Performance (%) of baselines and TIMER-Instruct on MedAlign and TIMER-Bench benchmarks, reported as mean ± std from bootstrap resampling (n=10,000) over test set of 100 samples.

Model Name	MedAlign				TIMER-Bench			
	LLM-as-Judge		Automatic Metrics		LLM-as-Judge		Automatic Metrics	
	Correct	Complete	BERTScore	ROUGE-L	Correct	Complete	BERTScore	ROUGE-L
Existing Medical Finetuned Model								
Meditron-7B*	3.63 ± 2.15	1.32 ± 1.16	60.60 ± 1.20	3.10 ± 0.50	2.99 ± 1.62	1.00 ± 0.87	65.23 ± 1.00	5.39 ± 0.61
MedAlpaca*	12.87 ± 3.63	4.29 ± 2.15	65.90 ± 1.10	4.80 ± 0.90	7.21 ± 2.49	1.49 ± 1.12	72.06 ± 0.74	9.25 ± 0.79
AlpaCare*	27.72 ± 4.95	12.87 ± 3.80	66.50 ± 2.70	11.30 ± 1.20	7.71 ± 2.61	1.24 ± 1.12	75.07 ± 0.10	14.39 ± 0.75
MMed-LLaMA-3-8B*	9.24 ± 3.30	4.29 ± 2.15	65.60 ± 0.80	4.90 ± 0.60	17.66 ± 3.73	6.72 ± 2.36	72.77 ± 0.55	10.80 ± 0.60
PMC-LLaMA-13B*	11.88 ± 3.63	4.62 ± 2.48	65.00 ± 1.10	3.70 ± 0.60	1.24 ± 1.12	0.50 ± 0.62	29.17 ± 2.81	0.77 ± 0.35
MedLM-Large*†	41.30 ± 5.60	20.80 ± 4.60	75.86 ± 0.99	13.78 ± 1.27	22.44 ± 4.11	8.73 ± 2.75	82.19 ± 0.44	22.43 ± 1.16
MedLM-Medium†	50.50 ± 5.60	29.40 ± 5.30	75.61 ± 0.90	13.21 ± 1.34	47.76 ± 4.73	22.64 ± 4.10	83.27 ± 0.47	24.33 ± 1.26
MedInstruct‡	45.90 ± 5.60	27.70 ± 5.10	70.90 ± 0.70	8.70 ± 0.60	59.45 ± 4.73	38.81 ± 4.73	80.14 ± 0.47	18.86 ± 0.73
TIMER-INSTRUCT Tuned Model with Different Base								
Qwen2.5-7B-Instruct	58.42 ± 5.61	41.91 ± 5.61	73.57 ± 0.54	9.51 ± 0.62	67.41±4.48	53.48±4.85	80.81±0.34	18.40±0.64
w/ TIMER-INSTRUCT Tuning	60.40 ± 5.61	43.23 ± 5.61	73.39 ± 0.54	9.36 ± 0.62	69.15±4.48	52.99±4.85	81.53±0.35	19.12±0.66
Llama3.1-8B-Instruct	46.53 ± 5.45	29.70 ± 5.12	70.50 ± 0.70	8.50 ± 0.70	57.96 ± 4.85	34.58 ± 0.46	79.49 ± 0.45	17.78 ± 0.70
w/ TIMER-INSTRUCT Tuning	53.47 ± 5.61	37.29 ± 5.61	76.70 ± 0.80	14.60 ± 1.40	64.68 ± 4.73	46.27 ± 4.85	83.20 ± 0.40	22.60 ± 1.07

*These models have a maximum context length ≤ 8K. We truncated the most recent records to fit within their maximum size.
†MedLM are powered by Med-PaLM 2, which is a medical fine-tuned version of Google PaLM.
‡We instruct-tuned MedInstruct w/ Llama3.1-8B-Instruct as the base model.

Table 2: Performance analysis of instruction-tuning the same base models with different temporal distribution of instructions: recency-focused ($p_i > 0.75$), edge-focused (higher density at t_{min} and t_{max}), and uniform distribution. We show head-to-head model comparisons, with **Model B** having an aligned temporal distribution with the benchmark, and individual model metrics for each benchmark. Bold numbers indicate the best performance in comparison.

Benchmark	Head-to-Head Comparison: Model A vs. Model B	Win Rate (A / B %)	Tie (%)	LLM-as-Judge Metrics: Model	Correctness	Completeness
MedAlign (Recency) *Human-Annotated, Recency-Focused Distribution*	Edge vs. **Recency**	42.24 / **43.89**	13.86	**Recency**	**55.54**	34.32
	Uniform vs. **Recency**	41.42 / **42.57**	16.01	Edge	53.47	**37.29**
				Uniform	50.83	33.70
TIMER-Bench (Edge) *Model-Generated, Higher Density at T_{min} and T_{max}*	Recency vs. **Edge**	47.89 / **48.76**	3.36	Recency	63.93	40.55
	Uniform vs. **Edge**	47.26 / **47.64**	5.10	**Edge**	64.68	**46.27**
				Uniform	**65.17**	42.54
TIMER-Bench (Uniform) *Model-Generated, Balanced Distribution on Timelines*	Recency vs. **Uniform**	45.16 / **51.01**	3.83	Recency	63.71	39.92
	Edge vs. **Uniform**	47.58 / **48.39**	4.03	Edge	61.69	43.55
				Uniform	**64.52**	**43.55**

ing both adult and children's hospitals. The repository follows the OMOP-CDM (Observational Medical Outcomes Partnership Common Data Model) OHD structure and encompasses 3.67M unique patients with records spanning from 1990 to 2023. We preprocessed patient timelines into chunks that fit the instruction-tuning model's context window. Using Gemini-1.5-Pro Team et al. (2024), we generated 5000 instruction response pairs with temporal evidence supporting the generated answer[1]. For TIMER-Bench, we sampled separate patient timelines with no overlap with the instruction-tuning set. We filtered the benchmark questions by selecting those that have multiple time-stamped pieces of evidence, which require synthesis of the patient record. The prompt for benchmark generation can be found in Appendix A. Three clinicians validated a subset on their clinical relevance, temporal reasoning, and accuracy, which is detailed in Appendix B.

Evaluation metrics. We evaluate models' open-text responses using LLM-Judge that assesses response correctness and completeness[2], which is verified through correlation analysis with clinician evaluation data released alongside the MedAlign benchmark. The human data consists of 9 clinicians that rank models with binary completeness and correctness scores. Our LLM-Judge framework demonstrates strong alignment with human judgment with $|\rho_{corr}|$ = 0.94 for correctness and $|\rho_{corr}|$ = 0.89 for completeness, as shown in Appendix I. We also employ head-to-head comparisons and automated metrics derived from token-level representations, including BertScore Zhang et al. (2019) (using distilbert-based-uncased), ROUGE-L Lin (2004), CHRF Popović (2015), and METEOR Banerjee & Lavie (2005)[3] to provide standard assessment of response quality. All LLM-based evaluations use GPT-4o-mini as the judge.

[1]The prompt for instruction-tuning data generation can be found in Appendix C
[2]Prompts detailed in Appendix H
[3]Results for these additional metrics are in Appendix M.

4.2 Model Evaluation and Analysis

Baselines. We evaluate our approach against several baselines. These include existing medical fine-tuned models from the literature: Meditron-7B Chen et al. (2023) with a 2K context, MedAlpaca Han et al. (2023) with a 2K context, AlpaCare Zhang et al. (2023b) with a 4K context, MMed-LlaMa-3-8B Qiu et al. (2024) with an 8K context, PMC-LlaMa-13B Wu et al. (2024a) with an 8K context, and MedLM-Large Cloud (2024) with an 8K context alongside MedLM-Medium Cloud (2024) with a 16K context. For those models unable to process the entire EHR within their context window, we truncated the most recent K tokens to fit the maximum context size. To identify the effectiveness of conventional medical instruction tuning for long-context reasoning, we include MedInstruct as an additional baseline, applying MedInstruct QA tuning Zhang et al. (2023b) to the long-context capable Llama-3.1-8B-Instruct. We evaluate the role of instruction tuning with TIMER-Instruct by applying it to different base models. Specifically, we tune both Qwen2.5-7B-Instruct Qwen et al. (2025) and Llama3.1-8B-Instruct with TIMER-Instruct, comparing their performance before and after tuning. For further reference, we evaluate several proprietary models (GPT4-32k, GPT-4o, Claude 3.5 Sonnet) and report their performance in Appendix L.

Results. The main results are in Table 1 and head-to-head comparison is shown in Appendix K. Analysis reveals three key findings. First, despite domain-specific training on various medical datasets (e.g., MedQA, medical papers, clinical text), existing medical models with limited context windows struggle with long EHR tasks - even the best performing MedLM-Large achieves only 41.3% correctness on MedAlign. While long-context models like MedLM-Medium and Llama3.1-8B-Instruct perform better, there remains significant room for improvement in temporal reasoning capabilities. Second, simply applying short-form medical instruction tuning, i.e., MedInstruct QA to a long-context model, shows minimal gains over the base Llama3.1-8B-Instruct, showing 1.5% improvement in correctness on TIMER-Bench and even hurts performance on MedAlign, suggesting that traditional medical QA instruction tuning alone is insufficient for complex temporal reasoning in EHRs. Finally, our temporal-aware instruction-tuning approach TIMER-Instruct demonstrates consistent improvements across model architectures. With Llama3.1-8B-Instruct as the base model, TIMER-Instruct improves MedAlign correctness from 46.53% to 53.47% and completeness from 29.70% to 37.29% (average improvement of 7.3%), while improving TIMER-Bench performance from 57.96% to 64.68% in correctness and 34.58% to 46.27% in completeness (average improvement of 9.2%). These improvements generalize to other architectures: for Qwen2.5-7B-Instruct, TIMER-Instruct improves MedAlign correctness from 58.42% to 60.40% and completeness from 41.91% to 43.23%, while similarly enhancing TIMER-Bench performance with correctness and automatic metrics. These gains across architectures demonstrate that our temporal-aware instruction tuning approach robustly enhances temporal reasoning and EHR understanding across models.

4.3 Impact of Temporal Distribution Strategies

Settings. Existing long-context EHR benchmarks primarily focus on questions about the most recent portions of patient timelines, creating an unintended distribution shift in evaluation. We create TIMER-Bench with two controlled temporal distributions: (1) an edge-focused distribution where evaluation instruction-response pairs are randomly sampled from the natural model-generated distribution and (2) a uniform distribution where the evaluation instruction-response pairs are sampled with equal frequency across all patient visits. As shown in Figure 3, these constructed distributions provide complementary evaluation settings to MedAlign, the recency-biased physician-generated longitudinal EHR benchmark, where over half of the instructions focus on visits from the most recent quarter of patient histories.

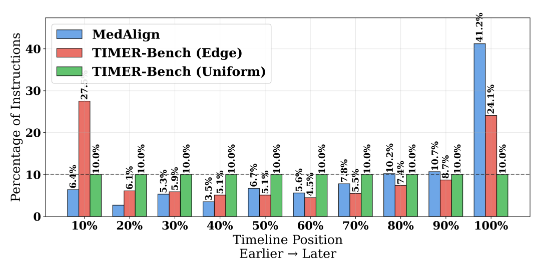

Figure 3: We evaluate on benchmarks with varying temporal distributions: recent-focused, edge-focused, and uniform.

Results. Table 2 analyzes how different temporal distributions in instruction-tuning affect model performance across various evaluation settings. Results show that instruction tuning with the matching distribution performs best on head-to-head comparison metrics for all three benchmark distributions. For MedAlign and TIMER-Bench Edge Cases, instruction tuning with the same temporal distribution as the evaluation data showed better performance than out-of-temporal-distribution tuning data. However, for the uniform evaluation cases, uniform instruction tuning demonstrated a more dramatic advantage, outperforming recency tuning by a margin 3.5 times larger than the performance gaps observed in any other evaluation scenario. These results highlight two key insights: (1) the importance of aligning instruction-tuning temporal distributions with the intended evaluation distribution, and (2) uniformly distributed evaluation sets benefit most from this distributional alignment compared to evaluation sets with inherent temporal biases, illustrating the role of instruction tuning in enhancing model temporal capabilities across a full patient record.

5 Conclusion and Discussion

This work reveals the ability to reason over temporal dependencies as a critical dimension in evaluating and instruction-tuning language models for clinical use. We uncover temporal biases in existing benchmarks that limit our understanding of model capabilities. To address this, we introduce a new temporal benchmark TIMER-Bench, which explicitly includes time evidence and enables controlled evaluation across different temporal distributions. TIMER-Instruct, our method for temporal instruction tuning, shows significant improvements on both physician-generated benchmarks and temporal reasoning benchmarks. While our experiments primarily focus on clinical records, the principles of temporal modeling apply to developing large language models in other fields that require reasoning over documents or sequences of events with complex temporal relationships.

6 Acknowledgments

We would like to thank the members of the Shah Lab and STAIR lab for insightful and helpful conversations regarding this work, as well as clinician annotators Mehr Kashyap, Kameron Black, and Akshay Swaminathan. We would also like to thank the Clinical Excellence Research Center (CERC) at Stanford for their support. SK acknowledges support by NSF 2046795 and 2205329, IES R305C240046, the MacArthur Foundation, Stanford HAI, OpenAI, and Google. NS acknowledges support by the Mark and Debra Leslie Endowment for AI in Healthcare, the Clinical Excellence Research Center at Stanford Medicine, and Technology and Digital Solutions at Stanford Healthcare.

REFERENCES

Observational health data sciences and informatics data standardization. `https://www.ohdsi.org/data-standardization/`.

Ahmed Allam, Stefan Feuerriegel, Michael Rebhan, and Michael Krauthammer. Analyzing patient trajectories with artificial intelligence. *Journal of medical internet research*, 23:e29812, 2021.

Satanjeev Banerjee and Alon Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pp. 65–72, 2005.

Suhana Bedi, Yutong Liu, Lucy Orr-Ewing, Dev Dash, Sanmi Koyejo, Alison Callahan, Jason A Fries, Michael Wornow, Akshay Swaminathan, Lisa Soleymani Lehmann, et al. Testing and evaluation of health care applications of large language models: a systematic review. *JAMA*, 2024.

Lucia A Carrasco-Ribelles, Jose Llanes-Jurado, Carlos Gallego-Moll, Margarita Cabrera-Bean, Mònica Monteagudo-Zaragoza, Concepción Violán, and Edurne Zabaleta-del Olmo. Prediction models using artificial intelligence and longitudinal data from electronic health records: a systematic methodological review. *Journal of the American Medical Informatics Association*, 30: 2072–2082, 2023.

Zeming Chen, Alejandro Hernández Cano, Angelika Romanou, Antoine Bonnet, Kyle Matoba, Francesco Salvi, Matteo Pagliardini, Simin Fan, Andreas Köpf, Amirkeivan Mohtashami, Alexandre Sallinen, Alireza Sakhaeirad, Vinitra Swamy, Igor Krawczuk, Deniz Bayazit, Axel Marmet, Syrielle Montariol, Mary-Anne Hartley, Martin Jaggi, and Antoine Bosselut. Meditron-70b: Scaling medical pretraining for large language models, 2023.

Google Cloud. MedPaLM 2 for Healthcare: Model Reference. `https://cloud.google.com/vertex-ai/generative-ai/docs/model-reference/medlm`, 2024.

Yann Dubois, Chen Xuechen Li, Rohan Taori, Tianyi Zhang, Ishaan Gulrajani, Jimmy Ba, Carlos Guestrin, Percy S Liang, and Tatsunori B Hashimoto. Alpacafarm: A simulation framework for methods that learn from human feedback. *NeurIPS*, 36, 2024.

Scott L Fleming, Alejandro Lozano, William J Haberkorn, Jenelle A Jindal, Eduardo Reis, Rahul Thapa, Louis Blankemeier, Julian Z Genkins, Ethan Steinberg, Ashwin Nayak, et al. Medalign: A clinician-generated dataset for instruction following with electronic medical records. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2024.

Tao Ge, Xin Chan, Xiaoyang Wang, Dian Yu, Haitao Mi, and Dong Yu. Scaling synthetic data creation with 1,000,000,000 personas. *arXiv preprint arXiv:2406.20094*, 2024.

Paul Hager, Friederike Jungmann, Kunal Bhagat, Inga Hubrecht, Manuel Knauer, Jakob Vielhauer, Robert Holland, Rickmer Braren, Marcus Makowski, Georgios Kaisis, et al. Evaluating and mitigating limitations of large language models in clinical decision making. *Nature Medicine*, 30: 2613–2622, 2024.

Tianyu Han, Lisa C Adams, Jens-Michalis Papaioannou, Paul Grundmann, Tom Oberhauser, Alexander Löser, Daniel Truhn, and Keno K Bressem. Medalpaca–an open-source collection of medical conversational ai models and training data. *arXiv preprint arXiv:2304.08247*, 2023.

Nathalie Huguet, Jorge Kaufmann, Jean O'Malley, Heather Angier, Megan Hoopes, Jennifer E DeVoe, and Miguel Marino. Using electronic health records in longitudinal studies: estimating patient attrition. *Medical care*, 58:S46–S52, 2020.

Yuri Kuratov, Aydar Bulatov, Petr Anokhin, Ivan Rodkin, Dmitry Sorokin, Artyom Sorokin, and Mikhail Burtsev. Babilong: Testing the limits of llms with long context reasoning-in-a-haystack. *NeurIPS*, 2024.

Haoran Li, Qingxiu Dong, Zhengyang Tang, Chaojun Wang, Xingxing Zhang, Haoyang Huang, Shaohan Huang, Xiaolong Huang, Zeqiang Huang, Dongdong Zhang, et al. Synthetic data (almost) from scratch: Generalized instruction tuning for language models. *arXiv preprint arXiv:2402.13064*, 2024a.

Tianle Li, Ge Zhang, Quy Duc Do, Xiang Yue, and Wenhu Chen. Long-context llms struggle with long in-context learning. *arXiv preprint arXiv:2404.02060*, 2024b.

Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pp. 74–81, 2004.

Lin Long, Rui Wang, Ruixuan Xiao, Junbo Zhao, Xiao Ding, Gang Chen, and Haobo Wang. On llms-driven synthetic data generation, curation, and evaluation: A survey. *arXiv preprint arXiv:2406.15126*, 2024.

Zhiyong Lu, Yifan Peng, Trevor Cohen, Marzyeh Ghassemi, Chunhua Weng, and Shubo Tian. Large language models in biomedicine and health: current research landscape and future directions. *Journal of the American Medical Informatics Association*, 31:1801–1811, 2024.

Mary M Lucas, Justin Yang, Jon K Pomeroy, and Christopher C Yang. Reasoning with large language models for medical question answering. *Journal of the American Medical Informatics Association*, 31:1964–1975, 2024.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *NeurIPS*, 35:27730–27744, 2022.

Maja Popović. chrf: character n-gram f-score for automatic mt evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pp. 392–395, 2015.

Pengcheng Qiu, Chaoyi Wu, Xiaoman Zhang, Weixiong Lin, Haicheng Wang, Ya Zhang, Yanfeng Wang, and Weidi Xie. Towards building multilingual language model for medicine, 2024.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5: A party of foundation models, 2025.

Omid Rohanian, Mohammadmahdi Nouriborji, Samaneh Kouchaki, Farhad Nooralahzadeh, Lei Clifton, and David A Clifton. Exploring the effectiveness of instruction tuning in biomedical language processing. *Artificial intelligence in medicine*, 158:103007, 2024.

Vibeke Kristine Scheller. Temporal patient trajectories: long stories in short admissions. *Journal of Health Organization and Management*, 36:69–86, 2022.

Anup Shirgaonkar, Nikhil Pandey, Nazmiye Ceren Abay, Tolga Aktas, and Vijay Aski. Knowledge distillation using frontier open-source llms: Generalizability and the role of synthetic data. *arXiv preprint arXiv:2410.18588*, 2024.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. Large language models encode clinical knowledge. *Nature*, 620:172–180, 2023.

Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024.

Brandon Theodorou, Cao Xiao, and Jimeng Sun. Synthesize high-dimensional longitudinal electronic health records via hierarchical autoregressive language model. *Nature Communications*, 14:5305, 2023.

Hieu Tran, Zhichao Yang, Zonghai Yao, and Hong Yu. Bioinstruct: instruction tuning of large language models for biomedical natural language processing. *Journal of the American Medical Informatics Association*, pp. ocae122, 2024.

Boris van Breugel, Tennison Liu, Dino Oglic, and Mihaela van der Schaar. Synthetic data in biomedicine via generative artificial intelligence. *Nature Reviews Bioengineering*, pp. 1–14, 2024a.

Boris van Breugel, Nabeel Seedat, Fergus Imrie, and Mihaela van der Schaar. Can you rely on your model evaluation? improving model evaluation with synthetic test data. *NeurIPS*, 2024b.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A Smith, Daniel Khashabi, and Hannaneh Hajishirzi. Self-instruct: Aligning language models with self-generated instructions. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pp. 13484–13508, 2023.

Chaoyi Wu, Weixiong Lin, Xiaoman Zhang, Ya Zhang, Weidi Xie, and Yanfeng Wang. Pmc-llama: toward building open-source language models for medicine. *Journal of the American Medical Informatics Association*, pp. ocae045, 2024a.

Zhenbang Wu, Anant Dadu, Michael Nalls, Faraz Faghri, and Jimeng Sun. Instruction tuning large language models to understand electronic health records. In *NeurIPS Datasets and Benchmarks Track*, 2024b. URL `https://openreview.net/forum?id=Dgy5WVgPd2`.

Shengyu Zhang, Linfeng Dong, Xiaoya Li, Sen Zhang, Xiaofei Sun, Shuhe Wang, Jiwei Li, Runyi Hu, Tianwei Zhang, Fei Wu, et al. Instruction tuning for large language models: A survey. *arXiv preprint arXiv:2308.10792*, 2023a.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. *ICLR*, 2019.

Xinlu Zhang, Chenxin Tian, Xianjun Yang, Lichang Chen, Zekun Li, and Linda Ruth Petzold. Alpacare: Instruction-tuned large language models for medical application. *arXiv preprint arXiv:2310.14558*, 2023b.

A Prompt for TIMER-Bench Generation

Prompt Template for TIMER-Bench Generation

```
You are an expert of a practicing medical doctor.

Your task is to interpret the provided electronic health record (EHR) data and generate synthetic instruction-response pairs
based on your medical specialization and expertise. Your instructions should emphasize the temporal progression of clinical
events, reflecting a deep understanding of how medical conditions, treatments, and outcomes evolve over time. Formulate all
instructions from the perspective of a doctor interacting with the EHR and seeking information to reason over a patient's
longitudinal medical history.

[EHR Data Description] Patient EHR data is provided in XML format. This XML document contains comprehensive, timestamped
information covering a patient's medical history, including diagnoses, treatments, medications, test results, and clinical
notes. All clinical events are ordered ascending by time.

[Instruction Response Guidelines] The generated instructions and responses must conform to these guidelines:
Instructions:
- Ground all instructions in the context of the provided EHR and ensure they are relevant to the patient's specific case.
- Prioritize instructions that involve reasoning over the timeline of the patient's care (e.g., comparing test results over
time, evaluating the progression or resolution of a condition, or assessing treatment impacts).
- Highlight the relationships between past events and their implications on current or future clinical decisions.
- Avoid generating generic or irrelevant instructions that do not utilize the temporal nature of the EHR.

Responses:
- Provide accurate, temporally-aware responses that directly address the posed instructions.
- Use the patient's longitudinal medical history to support your responses with precise and relevant details.
- Refer to specific timestamps (dates only, in the format MM/DD/YYYY) to justify responses and highlight temporal reasoning
(e.g., "On 02/15/2016, the patient's lab results showed a significant increase in XYZ levels, indicating...").
- Ensure responses are coherent, structured, and clinically valuable, using specialized terminology as appropriate.

Review the provided EHR and generate five instruction-response pairs that meet all outlined guidelines.
Ensure these pairs reflect comprehensive use of the patient's timeline and involve reasoning over the entire EHR context
window.
Explicitly include evidence for your generation by referencing dates of relevant visits or events in the EHR.

[EHR]
{ehr}

[Output]
```

Figure 4: Prompt template for TIMER-Bench generation.

Figure 4 illustrates the prompt template used to generate TIMER-Bench, which guided Gemini-1.5-Pro to create temporal instruction-response pairs with time evidence from longitudinal records.

B TIMER-Bench Clinician Verification

We ask three clinicians to verify the relevance, quality, and accuracy of a randomly selected subset of the TIMER-Bench instruction-response pairs. We note that the clinicians identified the generated instructions to have high relevance and the responses to be contextually accurate/reasonable, as they were not provided with the patient-protected information to do a complete factual accuracy evaluation. We also note that the majority of questions are considered to be medically complex by the clinicians, thus verifying that the synthetically generated questions are of reasonable complexity to capture model performance. The clinician-verified results are present in Table 3.

Table 3: Clinician Verification

Annotator	**Clinical Relevance**	**Complexity**	**Accuracy**
Annotator 1	97/100	77/100	100/100
Annotator 2	89/100	63/100	96/100
Annotator 3	99/100	99/100	100/100

C Prompt for Instruction-Tuning Data Generation

Figure 5 shows the prompt template used to generate our instruction-tuning dataset, which guides Gemini-1.5-Pro in creating diverse instruction-response pairs that span different parts of patient records.

D Recency bias of MedAlign

MedAlign Fleming et al. (2024) is introduced as the first benchmark created by clinicians covering realistic clinical instructions. However, physician-curated benchmarks show a natural bias toward

Prompt Template for Instruction-Tuning Data Generation

```
You are an expert of a practicing medical doctor.

Your task is to interpret the provided electronic health record (EHR) data and generate synthetic instruction-response pairs
based on your medical specialization and expertise. Formulate all instructions from the perspective of a doctor interacting the
EHR and seeking information from the record.

[EHR Data Description] Patient EHR data is provided in XML format. This XML document contains comprehensive, timestamped
information covering a patient's medical history, including diagnoses, treatments, medications, test results, and clinical
notes. All clinical events are ordered ascending by time.

[Instruction Response Guidelines] The generated instructions and responses must conform to these guidelines:
Instructions:
- Ensure that the instructions are grounded in the context of the provided EHR and are relevant to the patient's specific case
- Formulate questions that are meaningful and valuable for clinicians in their decision-making process
- Avoid generating generic or irrelevant instructions that do not contribute to the understanding of the patient's condition

Responses:
- Provide accurate and informative responses that directly address the posed instructions
- Utilize the expertise and specialized terminology contained within the EHR to deliver precise and targeted answers
- Refrain from using vague or non-specialized language that lacks depth and specificity
- Ensure that the responses are coherent, well-structured, and easy to understand

Review the provided EHR and generate ten instruction-response pairs that conform to all outlined guidelines. Ensure these pairs
cover the entire EHR context window.
When creating instructions and responses, please include evidence to support your generation. Specifically, provide the date of
the relevant visit from the EHR on which the instructions and responses are based, using the format MM/DD/YYYY (e.g.,
02/15/2016). Do not include the time (e.g., 12:00 AM).


[EHR]
{ehr}

[Output]
```

Figure 5: Prompt template for instruction-tuning pairs generation.

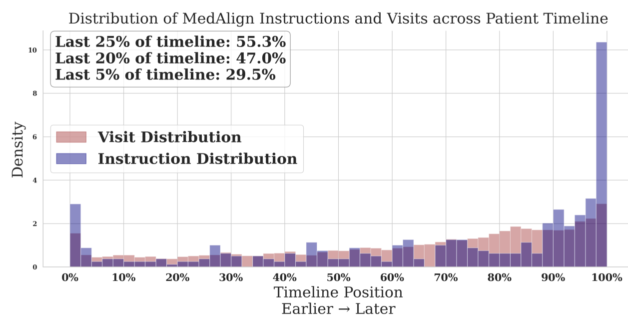

Figure 6: MedAlign instruction benchmark for longitudinal records emphasizes recent portions of each patient's longitudinal record.

recent records, as reviewing long patient histories is cognitively demanding, and creating instruction data manually from scratch is time-consuming. Despite covering a substantial average timespan of 3,895.06 days (approximately 10.7 years), the distribution of clinical instructions is concentrated on recent patient visits. Figure 6 illustrates the distribution of physician-created instruction and the actual patient visits across timelines. The time evidence information is extracted from human-written rationales accompanying the human-written responses in the dataset. It shows that 55.3% of all clinical instructions (n=396) are concentrated in just the last quarter of the patient timeline, with an even more pronounced 47.0% in the last 15% and 29.5% in the final 5%. While this partly reflects clinical practice—where recent summaries capture trajectory information—it may overlook crucial reasoning during earlier periods. Moreover, 71.3% of questions in MedAlign are retrieval-based, focusing on "needle-in-the-haystack" capabilities rather than evaluating the model's ability to synthesize over the full patient timeline.

E Dataset information of existing EHR instruction benchmarks and TIMER-Bench

Table 4: Dataset information of existing EHR instruction benchmarks and TIMER-Bench.

Dataset	Multi-Visit	Avg Time Span	Raw Record Modality	Curation	Size Scalable	Attribute
MIMIC-Instr (Test)	No	7.2 days	Notes Only	Model-Synthetic	Yes (200)	Instruction + Response
MedAlign	Yes	3895.06 days	Structured Data & Notes	Human-Created	No (303)	Instruction + Response + Rationale
TIMER-Bench	Yes	1294.88 days	Structured Data & Notes	Model-Synthetic	Yes (402)	Instruction + Response + Time Evidence

Table 4 presents a detailed comparison between TIMER-Bench and existing EHR instruction benchmarks. MIMIC-Instr focuses on single-visit scenarios with an average time span of only 7.2 days and relies solely on clinical notes, limiting its ability to capture longitudinal patient histories. While MedAlign incorporates both structured data and clinical notes across multiple visits with the longest average time span (3895.06 days), its manually curated nature restricts scalability with only 303 cases. In contrast, TIMER-Bench bridges these limitations by combining multi-visit support with both structured and unstructured data over a substantial time span of 1294.88 days. Moreover, TIMER-Bench leverages model-synthetic curation to enable scalability while uniquely incorporating temporal evidence in its instruction-response pairs, making it suitable for evaluating temporal reasoning capabilities in clinical settings.

F Data Statistics

Table 5: Lengths of questions and responses in each benchmark and instruction dataset.

Dataset	Count	instructions			Responses		
		Q1	Median	Q3	Q1	Median	Q3
MedAlign	303	8	11	17	14	31	56
TIMER-Bench (Edge)	402	14	18	22	42.25	59	82
TIMER-Bench (Uniform)	248	14	18	21	42.75	61.5	82
TIMER-Instruct (Recency)	5000	51	64	77	97	167	272
TIMER-Instruct (Edge)	5000	52	64	77	100	167	259
TIMER-Instruct (Uniform)	5000	54	65	78	103	168	259

Table 5 shows the length distributions of instructions and responses across evaluation benchmarks, MedAlign and TIMER-Bench, and the three sets of instruction-tuning datasets, TIMER-Instruct.

G Hyper-parameter Tuning

Table 6: Llama-3.1-8B-Instruct hyperparameter grid search.

Name	Values	Best Value
Learning Rate	0, 5e-6, 1e-5, 1e-4	1e-5
Gradient Accumulation	4, 8, 16, 32	16
Weight Decay	0, 1e-2, 1e-3,1e-4	1e-4

Table 6 presents the hyperparameter search space and optimal values for instruction-tuning Llama-3.1-8B-Instruct, where the best performance is achieved with a learning rate of 1e-5, gradient accumulation steps of 16, and weight decay of 1e-4.

H LLM Judge Prompt

Figure 7 shows the prompt template used for the LLM-as-Judge evaluation, which we use to assess model outputs for correctness and completeness according to reference responses.

I The Correlation of LLM-Judge with Human Annotated Rank

To validate our LLM judge, we performed a correlation analysis between the scores from the LLM Judge and the model rankings annotated by clinicians based on the instructions of the MedAlign benchmark, where human evaluators assigned rankings to different models' responses to the same instructions. As shown in Table 7, the LLM-judge metrics demonstrate strong agreement with human judgments, achieving high negative correlations for both correctness ($\rho_{corr} = -0.94$) and completeness ($\rho_{corr} = -0.89$). This inverse relationship aligns with our scoring scheme, where higher LLM scores and lower human ranks indicate better performance. Specifically, the ranking of models by LLM-judge scores consistently matches human preferences—GPT-4 variants are ranked highest, followed by Vicuna models, and then MPT-7B. This strong correlation suggests that our LLM-based evaluation framework provides a reliable proxy for human judgment in assessing temporal reasoning capabilities, enabling us to conduct larger-scale evaluations efficiently.

Prompt Template of LLM Judge

```
You are an expert in electronic health records (EHR) analysis. Your task is to evaluate responses generated by AI models based on given instructions
and EHR. You will assess the quality of the responses based on specific criteria, comparing them to provided reference answers (ground truth). Aim to
be fair and balanced in your evaluation, recognizing both strengths and limitations in the model's response.

EHR Data:
{ehr_data}

Instruction:
{instruction}

Model Response:
{model_response}

Reference Answer:
{reference_answer}

Evaluation Criteria:
1. Correctness (0/1):
- Score 1 if the response is generally accurate and aligns with the key points in the reference answer, even if there are minor discrepancies or
omissions.
- Score 0 only if the response contains significant factual errors or clearly misinterprets the EHR data.
2. Completeness (0/1):
- Score 1 if the response addresses the main aspects of the given instruction and covers the essential points present in the reference answer.
- Score 0 only if the response misses critical information or fails to address the core of the instruction.

Please provide your evaluation in the following JSON format:
{{
      "evaluation": {{
            "correctness": {{
            "score": 0 or 1,
            "explanation": "Your reasoning here, including comparison to reference answer"
            }},
            "completeness": {{
            "score": 0 or 1,
            "explanation": "Your reasoning here, including comparison to reference answer"
            }}
      }},
      "overall_comments": "Brief summary comparing the model's response to the reference answer, highlighting strengths and areas for improvement"
}}
Ensure your output is valid JSON that can be parsed programmatically. Do not include any text outside of the JSON structure.
```

Figure 7: Prompt template for LLM-as-Judge.

Table 7: Correlation between LLM-Judge evaluation and human judgments. Higher LLM scores and lower human ranks indicate better performance. Spearman correlation shows strong agreement between LLM scores and human ranks ($\rho = -0.97$ for average score, $\rho = -0.94$ for correctness, $\rho = -0.89$ for completeness).

Model	**LLM Score** Correctness	Completeness	**Human** Rank↓	**LLM** Rank↓
GPT4-32k	0.419	0.360	2.309	1
GPT4-32k-Multi-Step	0.383	0.365	2.292	2
Vicuna-13B	0.343	0.292	3.259	3
Vicuna-7B	0.318	0.299	3.304	4
MPT-7B-instruct	0.193	0.149	3.688	5

J Head-to-head Comparison Prompt

Prompt Template of Head-to-Head Comparison

```
Please act as an impartial judge and evaluate the quality of the responses provided by two AI assistants to the user question displayed below.
Your evaluation should consider correctness and helpfulness. You will be given a reference answer, assistant A's answer, and assistant B's answer.
Your job is to evaluate which assistant's answer is better. Begin your evaluation by comparing both assistants' answers with the reference answer.
Identify and correct any mistakes. Avoid any position biases and ensure that the order in which the responses were presented does not influence your
decision.
Do not allow the length of the responses to influence your evaluation. Do not favor certain names of the assistants. Be as objective as possible.
After providing your explanation, output your final verdict by strictly following this format: "[[A]]" if assistant A is better, "[[B]]" if assistant
B is better, and "[[C]]" for a tie. You must begin with [[A]] or [[B]] or [[C]].
Assigning "[[C]]" should be a very last resort; used only if you absolutely cannot discern any difference in the quality of the two responses.

Instruction:
{instruction}

Reference:
{ground truth response}

Response 1:
{response of model A}

Response 2:
{response of model B}
```

Figure 8: Prompt template for head-to-head comparison.

Figure 8 displays the prompt template used for the head-to-head comparison between model outputs, where responses from different models are compared to evaluate their relative performance given the instruction and the reference answer.

K Head-to-head Results

Our head-to-head evaluation in Table 8 reveals that TIMER-Instruct consistently outperforms existing medical models across both benchmarks. Against specialized medical models like Meditron-

Table 8: Head-to-head comparison between various models and TIMER-Instruct. Each row shows the winning margin (additional wins by TIMER-Instruct) and tie rates for both benchmarks.

Model	**MedAlign** Win%	**MedAlign** Tie%	**TIMER-Bench** Win%	**TIMER-Bench** Tie%
Existing Medical Finetuned Models				
Meditron-7B	+83.10	2.30	+95.02	0.50
MedAlpaca	+72.80	2.80	+86.41	0.37
AlpaCare	+84.40	1.00	+73.82	0.01
MMed-LlaMa	+80.14	2.60	+81.71	0.87
PMC-LlaMa	+74.40	4.10	+96.14	0.12
MedLM-L	+27.80	6.60	+52.49	1.00
MedLM-M	+39.50	5.90	+20.65	1.99
Models with the Same Base				
MedInstruct	+6.30	6.90	+8.45	1.99
Llama3.1	+23.80	5.60	+17.67	1.99

7B, MedAlpaca, and AlpaCare, TIMER-Instruct achieves substantial winning margins ranging from +72% to +95%. The performance gap is particularly pronounced on the TIMER-Bench dataset, where TIMER-Instruct demonstrates even stronger advantages, notably achieving a +95.02% winning margin against Meditron-7B and +96.14% against PMC-LLaMa. TIMER-Instruct also maintains a clear edge with winning margins of +6.30% to +23.80%. This consistent superiority across different model comparisons demonstrates the effectiveness of our instruction-tuning approach for temporal medical reasoning.

L PROPRIETARY MODEL PERFORMANCE

Table 9: Proprietary model performance (%) on MedAlign and LongEHR-Bench: reference as cap.

Model	**Open Source**	**MedAlign** LLM-as-a-Judge Correct.↑	**MedAlign** LLM-as-a-Judge Complete.↑	**MedAlign** NLP Metrics BERTScore↑	**MedAlign** NLP Metrics ROUGE-L↑	**LongEHR-Bench** LLM-as-a-Judge Correct.↑	**LongEHR-Bench** LLM-as-a-Judge Complete.↑	**LongEHR-Bench** NLP Metrics BERTScore↑	**LongEHR-Bench** NLP Metrics ROUGE-L↑
Proprietary Long Context Model Performance									
GPT4-32k	✗	62.73±5.61	39.60±5.61	78.12±0.69	15.77±1.33	72.89±4.35	58.46±4.85	81.79±0.45	22.19±0.97
GPT-4o	✗	63.37±5.45	45.21±5.61	76.87±0.71	14.23±1.21	85.32±3.48	70.65±4.35	83.51±0.34	22.60±0.89
Claude 3.5 Sonnet	✗	68.98±5.28	57.76±5.45	77.17±0.55	12.21±0.86	89.55±2.99	88.31±3.11	81.41±0.29	16.91±0.57

For reference purposes only, we present the performance of several proprietary large language models on both MedAlign and LongEHR-Bench datasets in Table 9. It is important to note that direct comparisons between these models and open-source alternatives would not be appropriate due to several factors: (1) the proprietary models have significantly larger model sizes and more extensive pretraining data, (2) their architectures and training procedures are not publicly available, and (3) their training costs are orders of magnitude higher than open-source alternatives. We include these results primarily to establish approximate performance ceilings for these benchmarks and to provide reference points for future research. The performance metrics are reported with standard deviations to account for the inherent variability in model outputs.

M ADDITIONAL METRICS RESULTS

Table 10 presents additional automatic evaluation metrics (METEOR, CHRF, and Google BLEU) for our model and baselines across MedAlign and TIMER-Bench benchmarks, where TIMER-Instruct consistently improves performance across all metrics compared to the base model and medical QA fine-tuning.

N ADDITIONAL TIMER-BENCH RESULTS

Table 11 shows model performance on the uniformly distributed version of TIMER-Bench, where TIMER-Instruct demonstrates consistent advantages over both medical-finetuned baselines and general long-context models, achieving the best performance across all metrics.

Table 10: Additional automatic metrics results (%) of baseline models and TIMER-Instruct, reported as mean ± standard deviation from bootstrap resampling (n=10,000) over the test set with 100 samples.

Model	MedAlign			TIMER-Bench (Edge)			TIMER-Bench (Uniform)		
	METEOR	CHRF	Google BLEU	METEOR	CHRF	Google BLEU	METEOR	CHRF	Google BLEU
Existing Medical Finetuned Model									
Meditron-7B*	6.22±0.98	11.10±0.80	1.19±0.20	9.20±1.14	16.64±0.88	3.40±0.31	8.23±1.11	15.75±1.07	3.24±0.41
MedAlpaca*	7.65±1.04	11.30±1.18	2.28±0.60	12.81±1.36	18.99±0.98	5.52±0.50	12.13±1.36	18.50±12.54	5.47±0.67
AlpaCare*	17.69±1.13	20.39±1.51	5.46±0.70	20.41±1.35	25.43±1.06	7.93±0.57	20.08±1.35	25.04±1.27	7.55±0.65
MMed-LLaMA-3-8B*	10.16±1.00	13.21±0.92	1.76±0.30	19.99±1.24	24.71±0.91	5.69±0.41	19.31±1.20	24.36±1.13	5.49±0.51
PMC-LLaMA-13B*	6.66±0.43	8.84±1.09	1.70±0.30	1.04±0.36	1.34±0.52	0.51±0.22	0.66±0.33	0.87±0.46	0.33±0.19
MedLM-Large*†	19.45±1.66	21.99±1.456	6.66±0.78	27.73±1.55	32.08±1.15	14.24±0.94	34.68±1.55	31.15±1.401	13.50±1.12
MedLM-Medium†	17.84±1.73	20.51±1.59	6.60±0.85	34.68±1.44	38.48±1.20	17.57±1.07	35.18±1.33	38.23±1.41	16.75±1.10
MedInstruct‡	19.30±1.20	20.66±1.11	3.60±0.40	35.18±1.33	38.11±0.91	11.89±0.68	37.76±1.32	37.30±1.15	11.73±0.68
TIMER-INSTRUCT Tuned Model with Different Base									
Qwen2.5-7B-Instruct	21.43±1.15	22.23±1.05	4.11±0.38	36.24±0.94	38.96±0.79	11.76±0.54	35.51±1.22	38.45±1.01	11.62±0.71
w/ TIMER-INSTRUCT Tuning	21.64±1.13	21.73±1.08	4.06±0.39	37.38±0.98	39.84±0.84	12.61±0.58	36.35±1.24	39.32±1.07	12.53±0.74
Llama3.1-8B-Instruct	17.89±1.02	19.63±1.12	3.26±0.40	33.84±1.02	36.77±0.91	10.80±0.57	33.64±1.44	36.41±1.18	10.76±0.59
w/ TIMER-INSTRUCT Tuning	22.91±1.04	25.43±1.57	7.88±1.00	37.76±1.32	41.17±1.06	16.82±0.96	37.10±1.10	40.19±1.29	16.23±1.10

*These models have a maximum context length $\leq$ 8K. We truncated the most recent records to fit within their maximum size.
†MedLM are powered by Med-PaLM 2, which is a medical fine-tuned version of Google PaLM.
‡We instruct-tuned MedInstruct w/ Llama3.1-8B-Instruct as the base model.

Table 11: Performance (%) of baseline models and TIMER-Instruct on TIMER-Bench (Uniform), reported as mean ± standard deviation from bootstrap resampling (n=10,000) over the test set with 100 samples.

Model	TIMER-Bench (Uniform)			
	LLM Judge Metrics		NLP Metrics	
	Correct	Complete	BERTScore	ROUGE-L
Medical Finetuned Model				
Meditron-7B*	3.23±2.22	0.40±0.60	64.38±1.38	4.91±0.75
MedAlpaca*	6.85±3.02	1.61±1.41	71.49±0.97	8.82±1.03
AlpaCare*	7.26±3.23	1.21±1.41	75.20±1.21	13.89±0.83
MMed-LLaMA-3-8B*	17.74±4.64	6.85±3.02	72.84±0.71	10.37±0.75
PMC-LLaMA-13B*	0.81±1.01	0.40±0.60	29.01±3.44	0.44±0.29
MedLM-Large*†	19.35±4.84	6.05±3.02	81.97±0.56	21.24±1.25
MedLM-Medium†	48.79±6.05	24.60±5.44	82.99±0.59	23.18±1.41
MedInstruct‡	58.47±6.05	36.69±6.05	79.93±0.63	18.37±0.94
TIMER-INSTRUCT Tuned Model with Different Base				
Qwen2.5-7B-Instruct	64.52±6.05	50.81±6.05	80.65±0.46	17.84±0.77
w/ TIMER-INSTRUCT Tuning	66.13±5.85	47.58±6.25	81.44±0.45	18.55±0.85
Llama3.1-8B-Instruct	58.47±6.05	33.47±6.05	79.30±0.59	17.62±0.95
w/ TIMER-INSTRUCT Tuning	64.52±6.05	43.55±6.05	84.09±0.56	24.16±1.27

*Models with context length $\leq$ 8K; most recent records truncated to fit.
†MedLM: medical fine-tuned version of Google PaLM (Med-PaLM 2).
‡Instruct-tuned using Llama3.1-8B-Instruct as base model.

O CASE STUDIES

As shown in Table 12 and Table 13, Qualitative analysis of model outputs reveals key differences in temporal reasoning capabilities between the baseline model and TIMER-Instruct model. The baseline model exhibits several characteristic failure modes: (1) recall interference, where it retrieves temporally irrelevant information from the context, e.g., listing unrelated historical visits in Case 1, (2) temporal ordering errors, where it misattributes dates and event sequences, e.g., incorrect x-ray dates in Case 3, and (3) repetitive generation patterns when handling long temporal sequences, e.g., redundant phrases about medication weaning in Case 2. In contrast, TIMER-Instruct demonstrates improved temporal processing capabilities: it successfully filters relevant temporal information from the context window, maintains chronological consistency across multiple time points, and generates non-repetitive responses that accurately capture temporal relationships. When presented with complex queries requiring multi-step temporal reasoning. e.g., treatment progression in Case 4, TIMER-Instruct shows robust performance in establishing causal relationships and maintaining contextual relevance throughout the generated response. These improvements suggest that our temporal-aware instruction tuning effectively enhances the model's ability to process and reason about time-dependent information in long-context scenarios.

Table 12: Case studies (Part 1): Qualitative comparison of model responses between base Llama-3.1-8B-Instruct and TIMER-Instruct-tuned model on TIMER-Bench. Key differences in response are highlighted in purple, with analytical insights in pink.

Clinical Query	**Llama-3.1-8B-Instruct**	**w/ TIMER-Instruct**
Case 1: Timeline Analysis Query: The patient was seen on 01/16/2013 for hip and shoulder pain. Provide a timeline of the events leading to this visit, including relevant symptoms and their duration.	**Response:** "The patient was seen on 01/16/2013 for hip and shoulder pain. The timeline of events leading to this visit is as follows:- The patient was seen on 12/31/2011 for an ear lesion, which was described as a bump in the ear that was bleeding 2 days ago and had some crust over it.- The patient was seen on 06/26/2012 for a follow-up visit, where she presented with bumps spreading on her face, which were later diagnosed as molluscum contagiosum infection.- The patient was seen on 11/20/2012 for a hospital visit, but the note does not mention any symptoms related to hip or shoulder pain.- The patient was seen on 11/27/2012 for a hospital visit, but the note does not mention any symptoms related to hip or shoulder pain.- The patient was seen on 01/09/2013 for a telephone encounter note, where her mother reported that the patient was experiencing asthma symptoms ..." *Analysis: The model response inaccurately represents the timeline of events. It fails to mention the specific injury to the hip during PE that occurred a month prior to the visit on 01/16/2013. Additionally, it does not accurately capture the duration and worsening of the shoulder pain and chest pain symptoms, which were reported as occurring over a couple of months leading up to the visit. The response gets distracted by visits in the patient record that do not correspond with the given instruction.*	**Response:** "The patient presented to the clinic on 01/16/2013 with complaints of hip pain and pinching pain in her left shoulder with running. The hip pain had started a month prior to the visit, after she injured her hip in PE doing an obstacle course. The pain initially ignored, but worsened over time. The shoulder pain had been bothering her for about 2 months, specifically when she ran. She also experienced chest pain during exercise, which was alleviated with the use of albuterol. Additionally, she reported shortness of breath. The patient had previously experienced asthma symptoms, for which she had used inhalers in the past. She had also been prescribed Singular as a child. The patient was advised to rest for 2 weeks, use ibuprofen as needed, and ..." *Analysis: The model accurately captures the core facts regarding the patient's hip and shoulder pain, including the timeline of events leading to the visit on 01/16/2013. It correctly identifies the injury during PE, the worsening of hip pain, and the shoulder pain experienced during running. The mention of asthma symptoms and the alleviation of chest pain with albuterol is also correctly noted.*
Case 2: Medication Review Query: The patient is concerned about his current narcotic use. What is his current regimen, and has he expressed concerns about this previously?	**Response:** "The patient is currently taking methadone 60 mg per day, and narco 2 to 3 mg q.h.s. He has expressed a concern about his narcotic use, stating that he has a fear of being on narcotics for the rest of his life and would like to be weaned off of his methadone. He has previously tried to take his medication differently ..." *Analysis: The model incorrectly states the patient's current narcotic regimen as Methadone 60 mg per day and Narco 2-3 mg q.h.s. However, the reference answer specifies that the patient takes Methadone 60 mg daily, Narco 2-3 mg at bedtime, and Motrin 800 mg as needed. Additionally, the model response contains repetitive phrases about the patient's attempts to wean off Methadone.*	**Response:** "The patient is currently taking methadone 60 mg per day, narco 2 to 3 mg q.h.s., and Motrin 800 mg. He has been on a narcotic for some time. In the past, he had tried fentanyl patches, Vicodin, and morphine. He has expressed a concern about being on narcotics for the rest of his life, and would like to be weaned off of his methadone. He has also stated that ..." *Analysis:The model accurately states the patient's current medication regimen, including Methadone 60 mg daily, Narco 2-3 mg at bedtime, and Motrin 800 mg. It correctly notes the patient's concerns about long-term narcotic use and his desire to be weaned off Methadone, aligning with the reference answer.*

Table 13: Case studies (Part 2): Qualitative comparison of model responses between base Llama-3.1-8B-Instruct and TIMER-Instruct-tuned model on TIMER-Bench. Key differences in response are highlighted in purple, with analytical insights in pink.

Clinical Query	Llama-3.1-8B-Instruct	w/ TIMER-Instruct
Case 3: X-ray Findings Over Time Query: The patient had a left knee injury in July 2022. Summarize the findings from his x-rays over time.	**Response:** "The patient had a left knee injury in July 2022. Over time, the x-rays showed the following findings: 1. Initially, there was a subtle irregularity along the proximal tibial metaphysis on lateral view X-ray on July 6, 2022, which raised concern for an occult proximal tibia fracture. 2. Repeat x-rays on July 25, 2022, did not show any obvious bony abnormality, but there was soft tissue swelling in the area of tenderness. 3. ... This confirmed the initial suspicion of a nondisplaced fracture of the proximal tibia. Answered correctly Answered correctly Answered correctly Answered correctly Answered correctly" *Analysis: The model response contains factual inaccuracies regarding the timeline of x-ray findings. It incorrectly states that the initial x-ray was on July 6, 2022, while the correct date is July 18, 2022. Additionally, it misrepresents the findings on July 25, 2022, as showing soft tissue swelling, which was not mentioned in the reference answer.*	**Response:** "1. Initial x-rays on 07/18/2022 showed a subtle irregularity along the proximal tibial metaphysis on lateral view, raising concern for an occult proximal tibia fracture. 2. Repeat x-rays on 07/25/2022 showed no obvious bony abnormality, but there was soft tissue swelling in the area of tenderness. 3... 4. No further x-rays were mentioned in the provided EHR for the left knee after 08/22/2022. 5. The patient was advised to limit his weightbearing activity based on his symptoms, and follow-up was recommended in 2 weeks if he was not comfortably walking without a limp. 6. The patient was also advised to avoid high-risk activities such as going to the playground, bouncy houses, or trampolines. 7. Follow-up x-rays were recommended in 3 weeks if the patient had not returned to full running" *Analysis: The model accurately describes the findings from the x-rays over time, including the initial irregularity, the repeat x-ray findings, and the final assessment of the healing nondisplaced fracture. The temporal relationships are correctly represented, and the dates mentioned align with the EHR data.*
Case 4: Injury Follow-up Query: The patient was seen on 02/13/2017 for a left ankle sprain. Summarize the treatment plan and outcome at his follow-up appointments.	**Response:** "At the follow-up visit on 01/23/2017, the patient was placed in a lace-up ankle brace for the next week. After a week of full-time wear, the brace was to be taken off only to shower and sleep. The patient was to wean out of the brace at home and discontinue it if he remained pain-free. He was advised to avoid physical education for the next week and was told that if the limp persisted or he continued to favor the ankle after 3 weeks, he should return to the clinic. At the follow-up visit on 02/13/2017, the patient's limp had improved, and he had full ankle range of motion without pain. He was no longer tender over the ATFL, CFL, or deltoid ligament. The plan was to refer him for physical therapy if the pain or limp persisted. The patient was advised to continue using the ankle brace as needed ..." *Analysis: The model response is ordered in such a way that it states that the patient was seen on 01/23/2017 for follow-up after it discusses the appointment on 02/13/2017, creating an incorrect temporal flow of events.*	**Response:** "At the follow-up visit on 01/23/2017, the patient was placed in a lace-up ankle brace for the next week. After a week of full-time wear, the brace was to be taken off only to shower and sleep. The patient was to wean out of the brace at home and discontinue it if he remained pain-free. He was advised to avoid physical education for the next week and was told that if the limp persisted or he continued to favor the ankle after 3 weeks, he should return to the clinic. At the follow-up visit on 02/13/2017, the patient's limp had improved, and he had full ankle range of motion without pain. He was no longer tender over the ATFL, CFL, or deltoid ligament. The plan was to refer him for physical therapy if the pain or limp persisted ..." *Analysis: The model accurately describes the treatment plan and outcomes for the patient's follow-up appointments, including the transition from a short leg walking cast to a lace-up ankle brace and the patient's improvement in symptoms. Its ordering of events is temporally grounded and clear, aligning with user preference.*