TIMELY Bench: Neuro-Symbolic Evaluation of Temporal Clinical Reasoning in Large Language Models

Anonymous Submission

Abstract

Most clinical LLM benchmarks treat EHR evidence as a static record, although bedside reasoning is made under partial observation. We present TIMELY Bench, an offline neuro-symbolic benchmark for anchor-bounded temporal reasoning over MIMIC-IV intensive-care records. TIMELY Bench fixes an anchor hour, masks explicit post-anchor evidence, serializes visible multimodal EHR context, and scores JSON responses through the Clinical Reasoning Evaluation Suite. Symbolic physiology templates define phases, expected trajectories, and atypical variants for acute kidney injury, delirium, and sepsis. Stroke is evaluated as a separate anchor-bounded neurological task family. In a frozen primary comparison of nine LLM providers across 53,070 prompts per provider and 20 auto-scorable task-dimension pairs, general-purpose models led medical-domain adapted models within this provider set. Difficulty varied by condition, reasoning dimension, and template support.

1 Introduction

ICU records are not produced like a regular machine-learning time series. Measurements appear when care teams decide to measure them, notes summarize what clinicians considered salient, and treatment changes alter the next observations. A clinician therefore does not decide whether kidney injury is progressing, delirium is resolving, or septic shock is emerging from one isolated value. The decision depends on what was visible at a specific time, including measurements, medication changes, procedures, and notes observed so far. Existing ICU benchmarks have made structured time-series prediction reproducible [Harutyunyan *et al.*, 2019; Wang *et al.*, 2020], and longitudinal EHR benchmarks have advanced transfer evaluation [Wornow *et al.*, 2023]. Medical LLM benchmarks test knowledge, reading comprehension, or interactive workflow performance [Jin *et al.*, 2021; Jiang *et al.*, 2025]. These settings do not isolate whether an LLM can interpret a fixed patient information state without using future evidence.

We frame TIMELY Bench as a neuro-symbolic evaluation problem. At anchor hour T, the patient state is $s_T = E[0, T]$, the evidence visible up to that hour. A condition template maps s_T to symbolic targets such as event time, phase, trajectory support, and scoring labels. The LLM produces a JSON answer from the serialized prompt, and the Clinical Reasoning Evaluation Suite (CRES) scores that answer against the rule-derived target. In NILA terms, the symbolic layer fixes what evidence is visible, what event is being queried, and how the answer is judged. This differs from general temporal-reasoning benchmarks, which focus on event ordering or duration over non-clinical evidence [Wang and Zhao, 2024; Fatemi *et al.*, 2025]. It also differs from temporal instruction tuning, where the goal is to improve model training rather than to define an auditable evaluation state [Cui *et al.*, 2025].

TIMELY Bench follows from this evaluation problem. It fixes an anchor-bounded ICU information state, uses symbolic physiology templates to expose phase and trajectory structure for the template-covered conditions, and evaluates nine LLM providers with CRES dimensions that separate temporal localization, threshold reasoning, trend interpretation, diagnostic discrimination, contrastive reasoning, and attribution. The symbolic layer contributes an anchor-bounded evidence state, executable clinical event rules, rule-derived scoring targets, and support strata for error analysis.

2 Benchmark Design

TIMELY Bench is derived from MIMIC-IV ICU records, including clinical notes [Johnson *et al.*, 2023]. Each stay is aligned to a 168-hour grid. For a stay and anchor hour T, all downstream evaluation views use the same visible evidence interval $[0, T]$. The anchor defines the evidence boundary of the task, not just the sampling time. Post-anchor evidence is masked. Text evidence is truncated by note time relative to the anchor, and prompts are serialized from evidence available within $[0, T]$. Each task retains at most two anchors per stay, one early and one late anchor when available, before expansion into task-dimension prompts. This anchor-bounded design controls explicit retrospective leakage and lets the same patient state be evaluated across multiple representations.

The benchmark materializes four synchronized views of the same stay and anchor (Fig. 1). Branch A is a compact structured summary, Branch B1 is an hourly sequence,

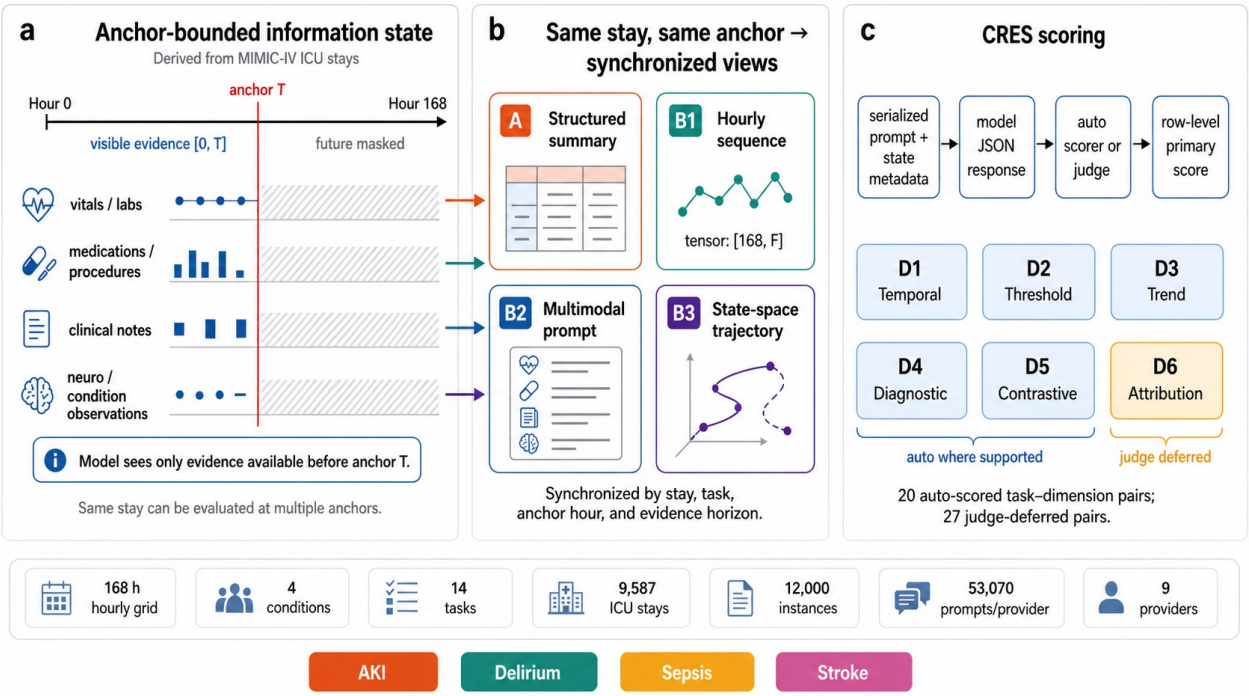

Figure 1: TIMELY Bench fixes the clinical information state at an anchor hour. The symbolic layer defines the anchor, task, template state, and CRES scoring rule, while the LLM sees only serialized visible evidence. The bottom row reports the frozen benchmark scale.

Branch B2 is a prompt-ready multimodal EHR context, and Branch B3 is a state-space trajectory. Here multimodal means structured measurements, medications, procedures, notes, and neurological assessments rather than images or waveforms. The LLM comparison uses Branch B2.

The symbolic layer is condition-aware. AKI tasks use KDIGO Stage 2+ progression and renal-support signals [KDIGO AKI Work Group, 2012]. Delirium tasks use CAM-ICU-derived onset and resolution patterns [Ely *et al.*, 2001]. Sepsis tasks use onset and shock-support physiology consistent with Sepsis-3 reasoning [Singer *et al.*, 2016]. The 14 task definitions cover AKI progression and renal support, delirium onset and resolution, sepsis onset and shock support, and stroke prospective neurological and retrospective extraction tasks. For AKI, delirium, and sepsis, deterministic, hand-specified executable templates define phases, expected marker trajectories, and atypical variants. For example, the AKI template marks a Stage 2+ target when visible creatinine, urine-output, or renal-support evidence satisfies the KDIGO-derived rule at or before the anchor. The first supporting hour becomes the D1 target, and the surrounding renal trajectory determines whether the case is typical, atypical, or low support. These templates are not hidden answers given to the LLM. They define labels, trajectory tiers, and stratification variables used to audit model behavior. Template support is not a data-quality score. It records agreement between the observed marker trajectory and the expected trajectory encoded by the template. A low-support case can still receive task labels when threshold or event criteria are met, but it is flagged as atypical or weakly aligned for stratified interpretation. Stroke is evaluated through neurological deficit and retrospective extraction tasks rather than the same template layer.

CRES decomposes each prompt into one reasoning operation: D1 temporal localization, D2 threshold reasoning, D3 trend interpretation, D4 diagnostic discrimination, D5 contrastive reasoning, and D6 attribution. For example, an AKI-T1 prompt at anchor T can ask when Stage 2+ AKI becomes supported by the visible trajectory (D1), whether a KDIGO threshold is met (D2), or whether renal markers are worsening (D3). The prompt ends with a fixed JSON schema containing reasoning, answer, evidence, and confidence fields. This structure makes free-form model outputs comparable without reducing the task to a single binary label.

3 Evaluation Protocol

The frozen evaluation sample contains 12,000 benchmark instances from 9,587 condition-positive or at-risk ICU stays. Expanding the task-dimension map yields 53,070 `full_multimodal` prompts per LLM provider because one clinical instance can instantiate multiple CRES prompts. The 14 clinical tasks expand across CRES dimensions into 20 auto-scorable task-dimension pairs plus deferred reasoning pairs. We evaluate nine frozen providers: GPT-5.4, Gemini 3.1 Pro, DeepSeek Chat, Qwen3.5, Gemma4-26B, Aloe-70B, Aloe-7B, Meditron3-8B, and MedGemma-4B. They are grouped as closed general-purpose, open or semi-open general-purpose, and medical-domain adapted models, defined here as models adapted or released for biomedical or clinical use. All outputs were generated from frozen prompts

and canonicalized before scoring. Release manifests record model aliases, decoding settings, retries, invalid-output repair, latency, token totals, source mode, and repair provenance.

Auto-scoring covers 20 supported task-dimension pairs. D1 event-time prompts are scored correct when the response matches the manifest-defined event-time or no-event target within the task-specific tolerance rule. Other binary tasks use binary accuracy, categorical tasks use macro F1, and numeric extraction uses exact match. Provider-level macro primary score is the equal-weight mean over supported task-dimension scores, not a row-weighted average. This choice prevents large task families from dominating the comparison. D6 attribution and other deferred reasoning cases are excluded from the nine-provider primary score because evidence attribution is less reliable under automatic rules. We report frozen descriptive comparisons rather than statistical significance tests. A stratified 500-instance subset is sampled for qualitative response audit [Zheng *et al.*, 2023]. The audit uses four contestants, GPT-5.4, DeepSeek Chat, Aloe-70B, and MedGemma-4B, judged by Claude Opus 4.6 as primary judge with GPT-5.4 and Gemini 3.1 Pro as cross-check judges. It evaluates response plausibility and evidence use under a fixed rubric, not clinical adjudication.

4 Results

Finding	Evidence and interpretation
Primary scope	Nine providers, 53,070 prompts/provider, and 20 auto-scored pairs. D6 and deferred reasoning are sampled for judge audit.
Leading result	Gemini 3.1 Pro scored 0.655. The best medical-domain adapted model, MedGemma-4B, scored 0.535. This compares frozen outputs, not a causal effect of adaptation.
Condition	Condition-level macro means across providers were delirium 0.823, AKI 0.602, sepsis 0.546, and stroke 0.421.
Template	Stay-level support was delirium 96%, sepsis 37%, and AKI 6%. Stroke is excluded from template-support analysis.
Confidence	High-confidence fraction was Gemini 99.2% and GPT-5.4 69.9%. Self-reported confidence is separated from correctness.

Table 1: Compact summary of frozen TIMELY Bench findings. Macro scores are equal-weight means over supported task-dimension scores rather than row-weighted averages.

Table 1 summarizes the frozen results most relevant to NILA. Condition difficulty was not uniform. Delirium had the highest condition-level macro score (0.823), whereas stroke remained the hardest condition family. The delirium result coincided with high template support, while AKI and sepsis had weaker template alignment. A single pooled score would hide this variation.

Given the clinical adaptation of several evaluated models, the tier-level result was counterintuitive. Gemini 3.1 Pro led overall (0.655). The strongest medical-domain adapted model, MedGemma-4B, reached 0.535. Within this frozen provider set, prior medical-domain adaptation did not guarantee stronger performance on anchor-bounded temporal reasoning.

Template support changed how the aggregate scores should be read. Delirium had high template support, sepsis was intermediate, and AKI had low support. Stroke was intentionally excluded from this template-support comparison because it uses a different neurological task design. For the three template-covered conditions, the templates also identified trajectory strata for cautious interpretation.

In the judge audit, mean scores across the three judges were broadly consistent with the auto-scored comparison. GPT-5.4 and DeepSeek Chat received higher qualitative audit scores than Aloe-70B and MedGemma-4B. These scores are not part of the nine-provider primary score.

5 Discussion and Limitations

The template-support results help explain why condition-level scores should not be read as model ability alone. Delirium had high stay-level template support (96%) and the highest condition macro score (0.823). AKI showed the opposite pattern, with only 6% template support and a lower condition macro score (0.602). Sepsis was intermediate on both measures. This suggests that part of the observed difficulty comes from the clinical trajectory itself. In heterogeneous conditions such as AKI, the symbolic layer is less a near-gold-standard event description than a way to mark where the trajectory no longer follows the expected physiology.

The same point also tempers the provider ranking. General-purpose models led the frozen comparison, while the best medical-domain adapted model was MedGemma-4B. We do not interpret this as evidence that medical adaptation is unhelpful. These tasks also test long-context use, instruction following, JSON discipline, and respect for the anchor-time evidence boundary. Provider scores therefore depend on both model behavior and the mix of trajectory types in the benchmark.

Several limits follow from the study design. The benchmark uses one de-identified ICU database and is not a prospective clinical workflow evaluation. Templates cover AKI, delirium, and sepsis more directly than stroke. The rule-derived labels are benchmark reference targets, not clinician-adjudicated diagnoses. The judge audit samples 500 prompt instances and should be treated as qualitative support for the auto-scored results, not as a substitute for expert clinical review. Symbolic definitions of time, thresholds, template support, and visible evidence made the evaluation easier to audit, especially when model outputs were fluent but clinically unsupported.

Ethical Statement

This study uses de-identified MIMIC-IV data accessed through the PhysioNet credentialed data access program and BigQuery under credentialed access. No prospective data collection, patient contact, or intervention was performed. The benchmark is an offline evaluation resource and is not intended for clinical decision support.

References

[Cui *et al.*, 2025] Hejie Cui, Alyssa Unell, Bowen Chen, Jason Alan Fries, Emily Alsentzer, Sanmi Koyejo, and Nigam H. Shah. TIMER: temporal instruction modeling and evaluation for longitudinal clinical records. *npj Digital Medicine*, 8:577, 2025.

[Ely *et al.*, 2001] E. Wesley Ely, Sharon K. Inouye, Gordon R. Bernard, Sharon Gordon, Joseph Francis, Lisa May, Brenda Truman, Theodore Speroff, Shiva Gautam, Richard Margolin, Robert P. Hart, and Robert Dittus. Delirium in mechanically ventilated patients: Validity and reliability of the confusion assessment method for the intensive care unit (CAM-ICU). *JAMA*, 286(21):2703–2710, 2001.

[Fatemi *et al.*, 2025] Bahare Fatemi, Mehran Kazemi, Anton Tsitsulin, Karishma Malkan, Jinyeong Yim, John Palowitch, Sungyong Seo, Jonathan Halcrow, and Bryan Perozzi. Test of time: A benchmark for evaluating LLMs on temporal reasoning. *The Thirteenth International Conference on Learning Representations*, 2025.

[Harutyunyan *et al.*, 2019] Hrayr Harutyunyan, Hrant Khachatrian, David C. Kale, Greg ver Steeg, and Aram Galstyan. Multitask learning and benchmarking with clinical time series data. *Scientific Data*, 6(1):96, 2019.

[Jiang *et al.*, 2025] Yixing Jiang, Kameron C. Black, Gloria Geng, Danny Park, James Zou, Andrew Y. Ng, and Jonathan H. Chen. Medagentbench: A realistic virtual EHR environment to benchmark medical LLM agents. *arXiv preprint arXiv:2501.14654*, 2025.

[Jin *et al.*, 2021] Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14):6421, 2021.

[Johnson *et al.*, 2023] Alistair E. W. Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J. Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, Li-wei H. Lehman, Leo A. Celi, and Roger G. Mark. Mimic-iv, a freely accessible electronic health record dataset. *Scientific Data*, 10(1):1, 2023.

[KDIGO AKI Work Group, 2012] KDIGO AKI Work Group. KDIGO clinical practice guideline for acute kidney injury. *Kidney International Supplements*, 2(1):1–138, 2012.

[Singer *et al.*, 2016] Mervyn Singer, Clifford S. Deutschman, Christopher W. Seymour, Manu Shankar-Hari, Djillali Annane, Michael Bauer, Rinaldo Bellomo, Gordon R. Bernard, Jean-Daniel Chiche, Craig M. Coopersmith, Richard S. Hotchkiss, Mitchell M. Levy, John C. Marshall, Greg S. Martin, Steven M. Opal, Gordon D. Rubenfeld, Tom van der Poll, Jean-Louis Vincent, and Derek C. Angus. The third international consensus definitions for sepsis and septic shock (Sepsis-3). *JAMA*, 315(8):801–810, 2016.

[Wang and Zhao, 2024] Yuqing Wang and Yun Zhao. TRAM: Benchmarking temporal reasoning for large language models. *Findings of the Association for Computational Linguistics: ACL 2024*, pages 6389–6415, 2024.

[Wang *et al.*, 2020] Shirly Wang, Matthew B. A. McDermott, Geeticka Chauhan, Marzyeh Ghassemi, Michael C. Hughes, and Tristan Naumann. MIMIC-Extract: A data extraction, preprocessing, and representation pipeline for MIMIC-III. *Proceedings of the ACM Conference on Health, Inference, and Learning*, 2020.

[Wornow *et al.*, 2023] Michael Wornow, Rahul Thapa, Ethan Steinberg, Jason Fries, and Nigam Shah. EHRSHOT: An EHR benchmark for few-shot evaluation of foundation models. *Advances in Neural Information Processing Systems 36: Datasets and Benchmarks Track*, 2023.

[Zheng *et al.*, 2023] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. *Advances in Neural Information Processing Systems*, 36:46595–46623, 2023.