TIMELY-Agent: Privacy-Aware Neuro-Symbolic Orchestration for Temporal Clinical Benchmark Construction

Anonymous Submission

Abstract

Clinical reasoning over electronic health records (EHRs) is temporal, multimodal, and governed by domain knowledge. Yet most evaluations of large language models (LLMs) in healthcare remain centred on static question answering, endpoint prediction, or isolated note understanding. We present **TIMELY-Agent**, a privacy-aware neuro-symbolic framework for constructing multimodal clinical reasoning benchmarks from longitudinal EHR data. The framework separates a *knowledge-facing* layer, where agents use guidelines, literature, terminology resources, and evidence tools to draft disease-specific Condition Graphs and Physiology Templates, from a *patient-facing* layer, where real clinical data are queried only inside local secure environments. We position OMOP-compatible retrieval and FastOMOP/OMCP-style governed interfaces as the standards-aware bridge between symbolic clinical specifications and local EHR databases. TIMELY-Agent therefore treats benchmark construction as an auditable workflow: clinical knowledge is formalised into executable schemas, patient trajectories are assembled into reasoning episodes, and model outputs are evaluated for temporal grounding, trend reasoning, cross-modal consistency, and evidence attribution.

1 Introduction

Large language models and multimodal foundation models have shown strong performance on clinical knowledge tasks and patient timeline modelling [Singhal *et al.*, 2023; Kraljevic *et al.*, 2024; Acosta *et al.*, 2022]. Public critical-care resources such as MIMIC-III and MIMIC-IV have further enabled reproducible research on longitudinal EHR modelling [Johnson *et al.*, 2016; Johnson *et al.*, 2023], including work combining clinical notes with time-series data [Khadanga *et al.*, 2019; King *et al.*, 2023]. However, clinical reasoning is rarely a static text-generation problem. Clinicians interpret laboratory values, vital signs, medications, diagnoses, procedures, and notes as evolving evidence streams. A rising creatinine, for example, is meaningful only in relation to baseline renal function, urine output, nephrotoxic exposure, treatment timing, and the contemporaneous narrative record. Similarly, respiratory deterioration or fluctuating cognition may require reasoning over irregular measurements, delayed documentation, and partially conflicting modalities.

Existing EHR benchmarks have enabled important progress, but many are still endpoint-oriented, focusing on outcomes such as mortality, length of stay, decompensation, or phenotyping [Rajkomar *et al.*, 2018; Harutyunyan *et al.*, 2019]. Such tasks often do not reveal whether a model used the right evidence, aligned the right time window, or relied excessively on one modality, echoing broader concerns about shortcut learning, poor grounding, and limited end-use interpretability in healthcare AI [Ghassemi *et al.*, 2020; Tonekaboni *et al.*, 2019]. In safety-critical clinical settings, this distinction matters: a model may produce a plausible explanation while grounding it in future information, irrelevant notes, or incorrect temporal assumptions.

We propose **TIMELY-Agent**, a neuro-symbolic agentic framework for constructing temporal and multimodal clinical reasoning benchmarks. The contribution is not a new predictive model, but a standards-aware and privacy-aware benchmark construction workflow. Neural agents assist with knowledge synthesis, note abstraction, and task drafting; symbolic structures constrain disease definitions, temporal windows, evidence requirements, and validation rules; local data interfaces enforce privacy and provenance. This makes the benchmark construction process itself inspectable.

2 TIMELY-Agent Framework

TIMELY-Agent is organised around three design principles. First, *clinical grounding*: benchmark instances should reflect clinically meaningful reasoning situations rather than arbitrary data slices. Second, *privacy-aware separation*: public medical knowledge and patient-level data should be handled by different computational layers. Third, *traceable provenance*: every generated benchmark item should be linked back to condition logic, retrieval criteria, timestamps, source modalities, and evidence fragments. Appendix Figure 1 summarises this architecture, governed by OMOP-compatible retrieval, local reasoning-episode construction, and downstream benchmark auditing.

Knowledge-facing agents. The first layer defines what should count as a reasoning problem before any patient data

are accessed. A knowledge agent collects and structures disease-specific information from guidelines, review articles, terminology resources, and, where appropriate, medical evidence tools such as OpenEvidence-style systems. Its output is not treated as final clinical truth; instead, it provides a structured draft for expert review.

This scaffold is represented through two artefacts. **Condition Graphs** encode relationships among diagnoses, symptoms, laboratory markers, vital-sign deviations, treatments, contraindications, confounders, and observable events. **Physiology Templates** encode temporal signatures such as onset, deterioration, threshold crossing, fluctuation, treatment response, and recovery. Together, they transform heterogeneous clinical knowledge into partially executable benchmark schemas: anchor events, inclusion and exclusion criteria, concept sets, valid time windows, and evidence requirements. This is the symbolic layer that constrains downstream retrieval and evaluation.

Privacy-aware model separation. TIMELY-Agent deliberately separates knowledge-facing and patient-facing computation. The knowledge-facing layer can use stronger closed-source models because it operates on public or institutionally approved reference material. Its outputs are distilled into symbolic artefacts and passed one way into the local environment. In contrast, the patient-facing layer does not send real patient data to external model providers. It uses locally hosted open-source models, deterministic scripts, and governed database interfaces. This design allows the system to benefit from frontier models for knowledge synthesis while respecting clinical data-governance constraints.

Standards-aware local retrieval. The patient-facing layer requires a controlled interface over heterogeneous EHR data. TIMELY-Agent adopts OMOP CDM as the main analytical representation for local cohort discovery and evidence retrieval [Hripcsak *et al.*, 2015; OHDSI,]. FHIR can support upstream exchange or ingestion [Mandel *et al.*, 2016; HL7 International,], while MEDS can support downstream event-based machine-learning exports [Medical Event Data Standard,]. These standards are not treated as interchangeable: OMOP is used for local retrieval, FHIR for interoperability boundaries, and MEDS for benchmark packaging.

FastOMOP/OMCP provides a natural infrastructure pattern for this layer. OMCP exposes OMOP-compatible databases through a Model Context Protocol interface, allowing agents to perform controlled cohort discovery, terminology lookup, and provenance-aware querying [fastOMOP,]. Recent FastOMOP work further argues that safe agentic real-world evidence generation requires governance, observability, and orchestration layers separated from pluggable agent teams, with deterministic validation at the process boundary [Moeller-Grell *et al.*, 2026]. In TIMELY-Agent, this pattern is used to prevent unconstrained natural-language access to patient data: Condition Graphs and Physiology Templates define what evidence is needed; the governed OMOP-facing layer mediates retrieval; and query outputs retain provenance for later audit.

3 Reasoning Episodes and Evaluation

The core output of TIMELY-Agent is a **reasoning episode**: an anchor-bounded, provenance-preserving slice of a patient trajectory centred on a clinically meaningful question. Each episode contains structured measurements and events around an anchor point, aligned note fragments, the condition schema that motivated retrieval, and metadata describing temporal windows, concept mappings, and source evidence. Unlike full-admission records, episodes are compact enough for error analysis; unlike isolated labels, they preserve the multimodal context required for clinical reasoning.

Constructing episodes is difficult because EHR modalities are asynchronous. Notes may describe events retrospectively or prospectively; laboratory tests and vital signs are irregularly sampled; and treatment changes may alter the interpretation of later observations. TIMELY-Agent therefore uses symbolic templates to condition alignment. A threshold-crossing task may require dense pre-anchor values, while a treatment-response task may require post-intervention trajectories. A note-consistency task may compare narrative statements against nearby structured measurements.

The evaluation layer turns episodes into diagnostic tasks rather than leaderboard-only labels. TIMELY-Bench tasks will probe temporal grounding, numeric and trend reasoning, threshold and transition detection, note–trajectory consistency, contrastive inference across neighbouring windows, and evidence attribution. The accompanying Clinical Reasoning Evaluation Suite (CRES) checks whether model outputs cite the right evidence, respect temporal order, avoid unsupported causal claims, and remain consistent with structured values. Ongoing prototype work instantiates parts of this pipeline using MIMIC-IV-style episode construction and behavioural verification of LLM-generated clinical reasoning, but these modules are treated as preliminary implementations rather than the main contribution.

4 Neuro-Symbolic Design for Time-Aware Clinical Agents

TIMELY-Agent is neuro-symbolic in three ways. First, clinical knowledge is represented as symbolic artefacts—Condition Graphs, Physiology Templates, concept sets, temporal windows, and validation rules—while neural models assist with knowledge synthesis, note abstraction, and task generation. Second, patient-data access is constrained by standards-aware interfaces and deterministic validation rather than unconstrained natural-language querying. Third, evaluation targets the gap between verbal plausibility and executable reasoning: a model should not receive credit for mentioning a clinically relevant cue unless it uses that cue at the correct time and in the correct evidential context.

This framing is particularly important for non-instantaneous and latency-aware clinical agents. Clinical evidence accumulates over time; some observations arrive late; notes may summarise earlier events; and correct interpretation depends on what information is available at the decision point. TIMELY-Agent can therefore support anchor-bounded evaluation, where post-anchor evidence is masked, delayed, or revealed. This allows us to test whether

an agent reasons prospectively from available information rather than retrospectively exploiting future evidence.

The framework also exposes failure modes that static benchmarks may hide: temporal inversion, over-reliance on one modality, hallucinated evidence, invalid query plans, unsupported causal language, and apparent symbolic understanding without quantitative control. By connecting symbolic disease knowledge to governed local retrieval and auditable reasoning episodes, TIMELY-Agent reframes clinical LLM evaluation as a privacy-aware benchmark construction problem rather than a single-model performance exercise.

References

[Acosta *et al.*, 2022] Julián N. Acosta, Guido J. Falcone, Pranav Rajpurkar, and Eric J. Topol. Multimodal biomedical AI. *Nature Medicine*, 28(9):1773–1784, 2022.

[fastOMOP,] fastOMOP. Omcp: Model context protocol server for the omop common data model.

[Ghassemi *et al.*, 2020] Marzyeh Ghassemi, Tristan Naumann, Peter Schulam, Andrew L. Beam, Irene Y. Chen, and Rajesh Ranganath. A review of challenges and opportunities in machine learning for health. *AMIA Summits on Translational Science Proceedings*, 2020:191–200, 2020.

[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*, 6:96, 2019.

[HL7 International,] HL7 International. FHIR Overview. https://www.hl7.org/fhir/overview.html. Accessed: 2026-05-24.

[Hripcsak *et al.*, 2015] George Hripcsak, Jon D. Duke, Nigam H. Shah, Christian G. Reich, Vojtech Huser, Martijn J. Schuemie, Marc A. Suchard, Rae Woong Park, Ian C. K. Wong, Peter R. Rijnbeek, Johan van der Lei, Nicole Pratt, G. Niklas Norén, Yu-Chuan Li, Paul E. Stang, David Madigan, and Patrick B. Ryan. Observational health data sciences and informatics (OHDSI): Opportunities for observational researchers. *Studies in Health Technology and Informatics*, 216:574–578, 2015.

[Johnson *et al.*, 2016] Alistair E. W. Johnson, Tom J. Pollard, Lu Shen, Li-wei H. Lehman, Mengling Feng, Marzyeh Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G. Mark. MIMIC-III, a freely accessible critical care database. *Scientific Data*, 3:160035, 2016.

[Johnson *et al.*, 2023] Alistair E. W. Johnson, Lucas Bulgarelli, Lu Shen, Alicia Gayles, Ahmad Shammout, Steven Horng, Tom J. Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, et al. MIMIC-IV, a freely accessible electronic health record dataset. *Scientific Data*, 10(1):1, 2023.

[Khadanga *et al.*, 2019] Swaraj Khadanga, Karan Aggarwal, Shafiq Joty, and Jaideep Srivastava. Using clinical notes with time series data for ICU management. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6432–6437, 2019.

[King *et al.*, 2023] Ryan King, Tianbao Yang, and Bobak Mortazavi. Multimodal pretraining of medical time series and notes. In *Proceedings of Machine Learning for Health*, volume 225 of *Proceedings of Machine Learning Research*, pages 243–257, 2023.

[Kraljevic *et al.*, 2024] Zeljko Kraljevic, Dan Bean, Anthony Shek, Rebecca Bendayan, Harry Hemingway, Joshua Au Yeung, Alexander Deng, Alfred Baston, Jack Ross, Esther Idowu, James T. T. Teo, and Richard J. B. Dobson. Foresight-a generative pretrained transformer for modelling of patient timelines using electronic health records: a retrospective modelling study. *The Lancet Digital Health*, 6(4):e281–e290, 2024.

[Mandel *et al.*, 2016] Joshua C. Mandel, David A. Kreda, Kenneth D. Mandl, Isaac S. Kohane, and Rachel B. Ramoni. SMART on FHIR: a standards-based, interoperable apps platform for electronic health records. *Journal of the American Medical Informatics Association*, 23(5):899–908, 2016.

[Medical Event Data Standard,] Medical Event Data Standard. What is meds?

[Moeller-Grell *et al.*, 2026] Niko Moeller-Grell, Shihao Shenzhang, Zhangshu Joshua Jiang, Richard J. B. Dobson, and Vishnu V. Chandrabalan. FastOMOP: A foundational architecture for reliable agentic real-world evidence generation on OMOP CDM data. *arXiv preprint arXiv:2604.24572*, 2026.

[OHDSI,] OHDSI. OMOP Common Data Model. https://ohdsi.github.io/CommonDataModel/. Accessed: 2026-05-24.

[Rajkomar *et al.*, 2018] Alvin Rajkomar, Eyal Oren, Kai Chen, Andrew M. Dai, Nir Hajaj, Michaela Hardt, Peter J. Liu, Xiaobing Liu, Jared Marcus, Mimi Sun, Patti Sundberg, Hector Yee, Kun Zhang, Gavin E. Duggan, James Irvine, Quoc Le, Kelly Litsch, Andrew Mossin, Jinda Tansuwan, De Wang, James Wexler, James Wilson, Daniel Ludwig, Samuel L. Volchenboum, Kevin Chou, Michael Pearson, Srinivasa Madabushi, Nigam H. Shah, Atul J. Butte, Michael D. Howell, Chunnan Cui, Greg S. Corrado, and Jeff Dean. Scalable and accurate deep learning with electronic health records. *npj Digital Medicine*, 1(1):18, 2018.

[Singhal *et al.*, 2023] Karan Singhal, Shekoofeh Azizi, Tao Tu, Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, et al. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180, 2023.

[Tonekaboni *et al.*, 2019] Sana Tonekaboni, Shalmali Joshi, Melissa D. McCradden, and Anna Goldenberg. What clinicians want: Contextualizing explainable machine learning for clinical end use. In *Proceedings of Machine Learning Research*, volume 106, pages 359–380, 2019.

A TIMELY-Agent Architecture

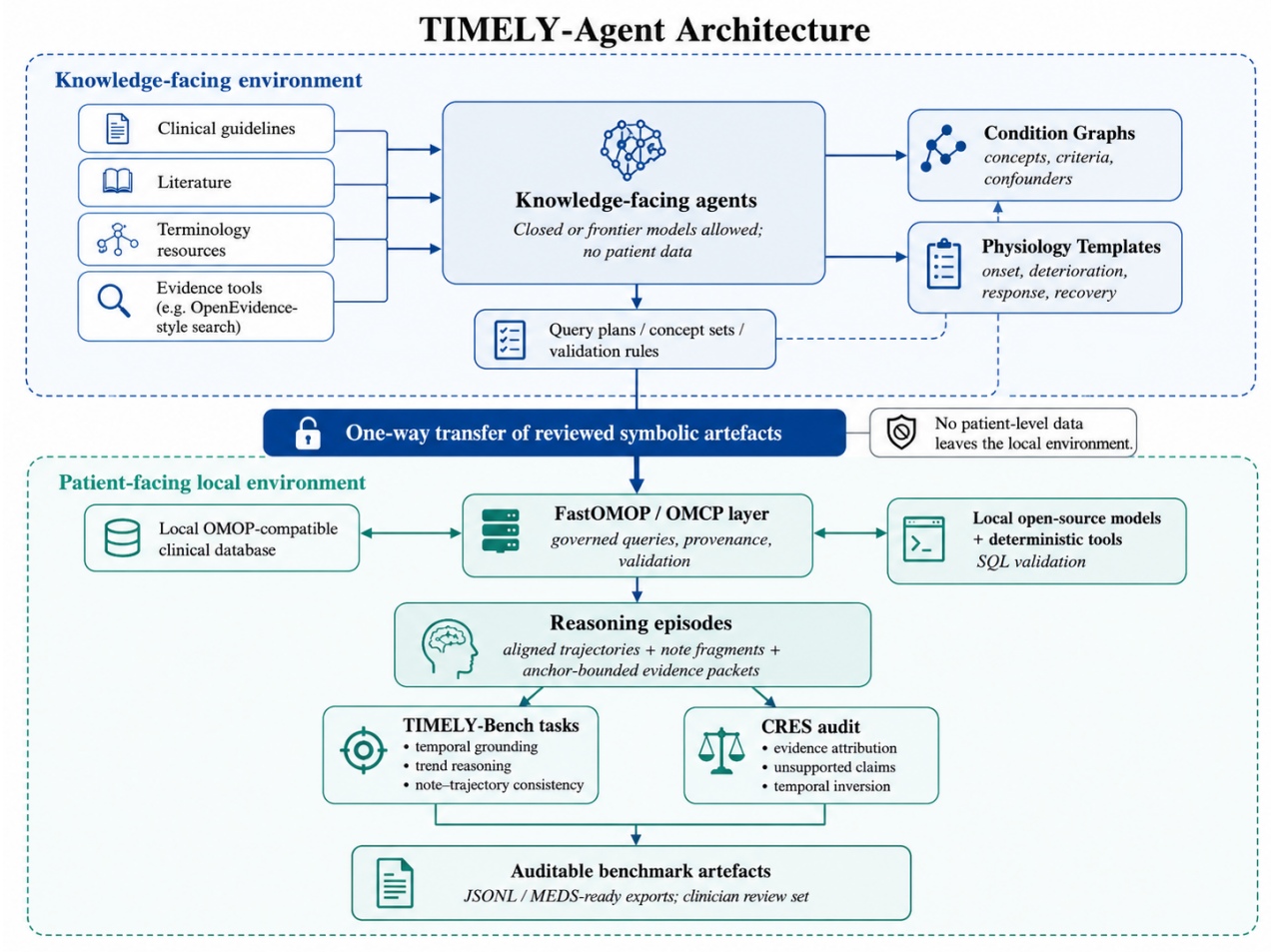

Figure 1: Overview of the TIMELY-Agent architecture. The framework separates a knowledge-facing environment from a patient-facing local environment. Public clinical knowledge sources, including guidelines, literature, terminology resources, and evidence-search tools, are first distilled into reviewed symbolic artefacts such as Condition Graphs, Physiology Templates, query plans, concept sets, and validation rules. These artefacts are transferred one way into the local environment, where patient-level data remain protected. FastOMOP/OMCP-style governed interfaces mediate access to OMOP-compatible clinical databases, while local open-source models and deterministic tools support SQL validation, episode construction, and audit. The resulting reasoning episodes are converted into TIMELY-Bench tasks and CRES audit outputs, producing auditable benchmark artefacts for model evaluation and clinician review.