OBSERVETRACE: EXPLICIT EVIDENCE STATES FOR PARTIALLY OBSERVED EHR DECISIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

At the moment a renal-risk medication is ordered, no local creatinine can mean never measured, advertised by an external exchange but unavailable, or too old for dosing; a note-derived value can be present yet uncertain. Standard code–value timelines collapse these conditions even though they imply different evidence needs and safety actions. *ObserveTrace* makes the observation process a first-class model interface: activated evidence requirements become canonical, round-trip-parseable events with value, units, time, provenance, and one of five evidence states. The same event object drives a 7B decoder, constrained alerts, frozen safety rules, and reviewer cards. Observability-aware denoising (OAD) reconstructs workflow-conditioned evidence losses, while incremental clinical consistency (ICC) regularizes adjacent-prefix decisions only when a frozen clinical protocol detects no rule or threshold change. Across 847K high-risk orders from 312K patients at seven health systems, ObserveTrace-7B records 3.2 (2.8–3.7) queue-reaching misses per 1,000 eligible orders, versus 3.8 (3.3–4.4) for an information-, backbone-, and training-budget-matched structured long-context model (paired difference -0.6 (-1.0 to -0.2), $P = 0.003$); its false-alert point estimate is also lower, 42.3 versus 43.7. In a fully crossed study, replacing a plain timeline with the complete interface transformation lowers misses from 9.1 to 5.1 before either objective. On 128,640 held-out gate-stable prefix pairs, ICC lowers argmax changes from 0.218 (0.206–0.231) to 0.129 (0.121–0.138), while the action-update rate on 34,920 gate-changing pairs is 0.924 (0.912–0.935). Endpoint concordance is 95.6% with blinded re-adjudication agreement $\kappa = 0.84$; a later 184.4K-order silent study preserves 2.9–3.4 logged queue-reaching misses per 1,000 across Epic, Cerner, and MEDITECH workflows. These results establish explicit evidence state as an effective common boundary for partially observed longitudinal decisions.

1 INTRODUCTION

At a medication decision timestamp, two empty creatinine fields can demand opposite actions. One may denote a test that was never measured; the other may point to an external result that can be acquired before dosing. A recorded value can likewise be too old for the current protocol or uncertain because it was extracted from narrative text. Conventional longitudinal encodings reduce these states to an absent token, a carried-forward value, or a site-specific feature. The distinction then disappears before constrained generation, deterministic rules, and human review can use it.

Missingness-aware sequence models learn that absence is informative through masks, elapsed-time channels, decay, or imputation (Che et al., 2018; Lipton et al., 2016; Baytas et al., 2017). That strategy is effective for prediction, but it leaves the observation process implicit: the model does not receive a persistent identity for *why* required evidence is unavailable. This becomes consequential when several consumers must agree on the same decision-time record. A learned embedding, an alert grammar, a dosing rule, and a reviewer card can otherwise encode different notions of presence, provenance, and freshness.

ObserveTrace makes evidence state the common boundary. A frozen trigger table declares the evidence slots required by each decision. Every activated slot has exactly one canonical witness—a matched source event or a synthetic REQUIRED event—and every witness carries value, unit, anchor, elapsed time, provenance, and one of MEASURED, NOT_MEASURED, EXTERNAL_ONLY, STALE, or

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

LOW_CONFIDENCE. Fixed precedence, escaping, ordering, and parse–serialize equality preserve event identity from model input through constrained alerts, the post-decoding rule layer, and review. The primary contribution is therefore a decision-conditioned observability contract, rather than another latent missingness feature.

Two objectives exploit transitions on that contract. Observability-aware denoising (OAD) draws removal, delay, external-only, staleness, and extraction-confidence corruptions from event-, setting-, and site-conditioned gap distributions fitted on training trajectories. Incremental clinical consistency (ICC) aligns adjacent-prefix recommendations only when a version-frozen protocol detects neither a threshold crossing nor a rule-status change. The strongest learned-channel comparator receives the same candidate events, values, units, time, provenance, and evidence states on the same 7B backbone and 32K context; the separate $2 \times 2 \times 2$ study varies the complete interface transformation, OAD, and ICC against a conventional plain timeline.

The evidence follows this mechanism. Across 847K high-risk orders from 312K patients at seven health systems, ObserveTrace-7B records 3.2 (2.8–3.7) queue-reaching misses per 1,000 eligible orders, compared with 3.8 (3.3–4.4) for the information-, backbone-, and budget-matched structured comparator (paired difference -0.6 (-1.0 to -0.2), $P = 0.003$), while also attaining a lower false-alert point estimate. Replacing the plain timeline with the interface lowers misses from 9.1 to 5.1 before OAD or ICC; targeted field removals identify observability as the largest individual intervention. OAD contributes most under workflow-conditioned evidence loss. On 128,640 held-out gate-stable prefix pairs, ICC lowers argmax changes from 0.218 (0.206–0.231) to 0.129 (0.121–0.138), with a 0.924 (0.912–0.935) action-update rate on 34,920 gate-changing pairs. Endpoint re-adjudication, rule-off tests, every-site paired effects, difficult-cohort analyses, and a later 184.4K-order silent multi-vendor study connect the interface mechanism to persistent operating performance. Together, these results show that preserving the reason evidence is unavailable changes longitudinal decisions while keeping that reason inspectable throughout the decision path.

2 Related Work

Missingness and irregular clinical time. Informative missingness is a defining property of EHR data (Weiskopf & Weng, 2013; Rubin, 1976; Little & Rubin, 2002). GRU-D, T-LSTM, and related models encode masks, time decay, or irregular intervals into latent dynamics (Che et al., 2018; Baytas et al., 2017; Lipton et al., 2016). Those methods define missingness over an existing variable set. ObserveTrace instead begins from decision-triggered evidence requirements: each activated slot must resolve to a source event or an explicit absence event. Its contribution is the persistence of that resolution across learning, constraints, rules, interventions, and review, not the discovery that missingness predicts outcomes.

Structured EHR foundation models and long context. Pretrained timeline models establish the value of temporal structure, scale, and reusable clinical representations (Li et al., 2020; Rasmy et al., 2021; Wornow et al., 2023b;a; Jiang et al., 2023; Kraljević et al., 2024; Pang et al., 2024); recent work directly studies long EHR contexts (Fallahpour et al., 2025; Wornow et al., 2025). The strongest comparator here follows that line and receives the same normalized values, units, time, provenance, five evidence states, candidate universe, 7B backbone, context, and training budget as ObserveTrace-7B, but carries nonvalue information in learned channels. The 3.8-versus-3.2 comparison therefore tests the complete ObserveTrace+OAD+ICC system against an information-matched learned-channel recipe. The crossed plain-sequence study separately measures the complete interface transformation and each objective; it is not treated as a lexicalization-only comparison.

Schemas, denoising, and consistency. OMOP and FHIR standardize clinical data exchange (Observational Health Data Sciences and Informatics, 2019; Ayaz et al., 2021), while graph linearization, table-to-text generation, schema-guided prediction, and constrained decoding make structured objects consumable by language models (Ribeiro et al., 2021; Chen et al., 2020; Li et al., 2021; Rastogi et al., 2020; Lee et al., 2021; Willard & Louf, 2023). ObserveTrace adds a different invariant: required evidence has one decision-time identity before and after serialization, including explicit absence. OAD inherits the robustness principle of denoising objectives (Vincent et al., 2008; Devlin et al., 2019) but estimates corruption support from workflow gap distributions and reconstructs provenance and evidence state. ICC is related to consistency regularization (Tarvainen & Valpola, 2017); unlike

invariance under generic perturbations, it compares real adjacent prefixes and disables its penalty when a frozen clinical protocol changes. These distinctions make observability, corruption, and prefix stability counterfactually testable as named parts of one interface.

Clinical safety systems. Medication decision support combines learned ranking with deterministic contraindication, dosing, and drug–drug interaction checks (Phansalkar et al., 2010; Sittig et al., 2008; Wright et al., 2016). Existing layered systems can pass learned scores and rule inputs through different representations. ObserveTrace makes the alert card and hard-rule evidence address the same round-trip-parseable event fields used by the learner. Retrained field removals then test named parts of that boundary, and rule-off evaluation identifies the learned ordering before deterministic enforcement.

3 APPROACH

ObserveTrace joins four independently testable objects at one event boundary: a source-record-to-evidence-state transformation, OAD for workflow-conditioned evidence loss, ICC for prefix stability relative to a frozen clinical protocol, and an external deterministic rule layer. The boundary enforces slot closure—every activated evidence requirement has exactly one matched or synthetic witness—and consumer consistency: parsing and reserializing a witness preserves the bytes addressed by the learner, output grammar, rules, and evidence card.

Candidate events and required-evidence slots. At prediction time t, every source record available by t is canonicalized exactly once into E_t^{raw}, while a frozen trigger table activates required-evidence slots Z_t. A slot (r, a, W, S) declares a trigger class, required code, lookback window, and source set. The resolver matches it only to records with the declared code, source, and window. A match reuses the raw event and records $(\text{trigger key}, \text{slot key}) \mapsto \text{raw event key}$; it never duplicates the event. An unmatched slot emits one REQUIRED tuple whose versioned slot key contains the required code, whose anchor is the trigger time, and whose unavailable value/unit receives external provenance with EXTERNAL_ONLY or none provenance with NOT_MEASURED. Rules and cards reconstruct the same map from the canonical sequence and table, so the audit map is not a learned input. In the medication study, an order activates the renal, hepatic, allergy, interaction, dose, and formulary inputs used by its prespecified safety rules. No absence is emitted outside an activated slot; trigger definitions, windows, sources, and clinical horizons are frozen before validation.

Evidence-state assignment. Each candidate becomes

$$e_i = (\tau_i, a_i, t_i^{\mathrm{anchor}}, \Delta t_i, v_i, u_i, s_i, o_i).$$

The fields encode event type, clinical code or slot key, anchor, elapsed time, value, normalized unit, provenance, and observability in serialization order. State assignment follows one fixed precedence. If no value is available, the external pointer test above selects EXTERNAL_ONLY or NOT_MEASURED. For an available value, calibrated extraction confidence below 0.85 selects LOW_CONFIDENCE; otherwise age beyond the frozen code–setting horizon selects STALE; all remaining values are MEASURED. Imported external values enter the value-present branch and retain external provenance, allowing an event to be external and stale without conflating provenance with observability. Structured order, laboratory, and medication fields receive confidence 1. Text and reconciliation adapters expose a source probability calibrated by isotonic regression on a held-out subset of the training partition, separately by site and source type with source-level backoff below 500 examples; calibration and the 0.85 cutoff are frozen before validation. Clinical horizons come from the same version-frozen guideline and formulary table used by the rule layer (e.g., 30 days for creatinine in the evaluated dosing rules), rather than from test-time gap statistics.

Canonical serialization. A serialized event is shown below; line wrapping is typographic and does not insert a byte.

```
[LAB|code:LOINC%3A2160-0|t_anchor:T0|delta:+3d|val:2.1|unit:mg/dL|
prov:external|obs:stale]
```

Values are converted to canonical units before serialization; unavailable values and units use the reserved token `NA`. Keys have a fixed order, duplicate or unknown keys are rejected, and reserved

delimiters [ ] | : % plus control characters are percent-encoded from UTF-8 bytes. Events are stably sorted by $(t^{\text{anchor}}, \Delta t)$, then event-type priority, code, value/unit, provenance, state, and a SHA-256 digest of canonical event identity. For raw events the identity is the canonical source record; for synthetic events it is the UTF-8 concatenation of trigger key, slot-table version, slot key, and state. The digest is used only to break otherwise exact ties and is not exposed to the learner. Round-trip equality is defined on the canonical tuple: parsing a serialization and serializing it again returns the identical byte string. This contract makes field interventions and evidence cards refer to the same named values used during training.

Backbone and complete objective. ObserveTrace-7B is a 7B decoder-only transformer with a 32K context window, trained for 142B tokens on 256 A100-80GB GPUs for 4.5 days. All decoder-based 7B systems start from the same Llama-2-7B checkpoint (Touvron et al., 2023). The structured long-context control receives the same candidate events, normalized values and units, stable order, time anchors and gaps, provenance, and five evidence states, but supplies nonvalue fields through learned side-channel embeddings. It uses the same three base tasks and omits OAD and ICC, comparing the complete method with an information-, backbone-, context-, token-, and tuning-budget-matched learned-channel recipe. The separate plain input renders observed source type, code, value/unit, and timestamp in source order, without synthetic required-slot events or model-visible canonical gap, provenance, and observability fields. Consequently, the plain-versus-ObserveTrace factor estimates the complete interface transformation rather than lexicalization alone. For clinical record completion (CRC), next-order/care-event prediction (NOCP), and time-to-complication modeling (TCEM), the total loss is

$$\mathcal{L} = \mathcal{L}_{\text{CRC}} + \mathcal{L}_{\text{NOCP}} + 0.5\mathcal{L}_{\text{TCEM}} + 0.7\mathcal{L}_{\text{OAD}} + 0.3\mathcal{L}_{\text{ICC}}.$$

Each term is normalized by its supervised tokens or eligible prefix pairs before weighting. CRC reconstructs code, value bin, unit, provenance, and observability. NOCP uses token cross-entropy for the next order/event and, at a medication trigger, the protocol-assigned action token; Appendix Table S17 defines its safe/unsafe mapping and precedence. The action-token marginal defines p_t. TCEM uses a discrete-time censored hazard loss. All terms update the shared transformer and heads; q_ϕ and the protocol gate receive no gradients. OT-LLM disables OAD and ICC.

Observability-aware denoising. OAD samples $m \sim q_\phi(m \mid \tau, \text{setting}, \text{site})$, where q_ϕ is estimated exclusively from training-set inter-event gaps. The support contains complete event removals, delayed availability, external-only substitution, stale-value substitution, and low-confidence extraction; a batch mask rate is sampled uniformly from 10–50%. For each event type and setting, complete-removal mass is the survival probability of the fitted censored gap model beyond its frozen task or slot horizon; the remaining mass follows the observed transition frequencies. Naturally unavailable events retain their original evidence-state token, whereas training corruption adds a distinct loss-only mask, preventing real absence from being treated as a reconstruction target. For strata with fewer than 500 gaps, estimation backs off from event type×setting×site to event type×setting and then event type, with empirical-Bayes shrinkage at each level. OAD minimizes

$$\mathcal{L}_{\text{OAD}} = \mathbb{E}_{m\sim q_\phi}[\ell_{\text{code}} + \ell_{\text{value}} + \ell_{\text{unit}} + \ell_{\text{provenance}} + \ell_{\text{state}}]$$

over corrupted eligible events. In plain-input cells, the same sampled source-event and value perturbations are applied, but reconstruction is restricted to source-visible fields; provenance and state are neither inputs nor targets. Algorithm S2 gives the estimator and fallback procedure.

Incremental clinical consistency. From training-partition trajectories only, we sample adjacent prefixes $(x_{\le t}, x_{\le t+\delta})$ with $\delta \in \{1, 6, 24, 72\}$ hours at probabilities $(0.35, 0.30, 0.20, 0.15)$. Let p_t and $p_{t+\delta}$ denote distributions over {APPROVE, ALERT, DEFER, ESCALATE}. ICC uses symmetric KL,

$$\mathcal{L}_{\text{ICC}} = \mathbb{E}_{t,\delta}\left[\tfrac{1}{2}\text{KL}(p_t \| p_{t+\delta}) + \tfrac{1}{2}\text{KL}(p_{t+\delta} \| p_t)\right] g_{t,\delta},$$

where $g_{t,\delta} = 1$ only when the newly revealed events neither change rule-layer status nor cross a version-frozen clinical protocol threshold. These protocol thresholds come from guideline and formulary versions fixed before validation and are distinct from the learned model's operating threshold selected on validation miss–alert utility. Setting $g_{t,\delta} = 0$ exempts protocol-changing updates from consistency pressure. The same frozen gate defines the two evaluation strata, so the direct ICC endpoint measures stability relative to this declared clinical protocol rather than an independent judgment of every prefix transition. Algorithm S3 specifies the gate.

External safety floor. After structured decoding, the rule engine evaluates dose bounds, contraindications, allergies, renal/hepatic adjustment, and drug–drug interactions. For every learned system, this engine reads the same frozen canonical source tuple through a model-independent rule API; model-specific input embeddings never alter rule inputs. A triggered rule can upgrade APPROVE to ALERT or ESCALATE and appends the supporting evidence field; it cannot suppress a learned alert. Rule-off experiments bypass this final step, separating learned ordering from deterministic enforcement.

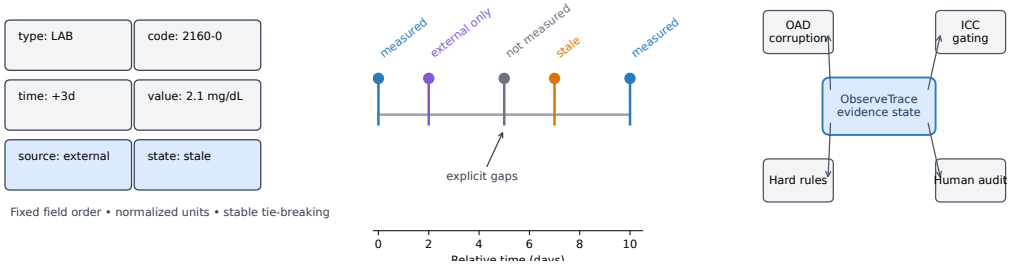

Figure 1: Source records and required-evidence slots become one deterministic event representation. (a) Canonical field order, escaped values, normalized units, and stable tie breaking yield a byte-identical parse–serialize round trip. (b) Anchors and relative gaps preserve irregular time without imputation; activated slots make not_measured and external_only explicit. (c) The serialized fields consumed by ObserveTrace, OAD, and ICC remain available to hard rules and human review.

4 Experimental Setup

Evaluation is organized around four questions: whether the complete system improves the medication miss–alert frontier, which interface component causes the change, whether the targeted robustness and prefix endpoints move as designed, and whether the ordering persists across sites and time. Secondary task and event-stream studies test portability. The full study spans 18 tasks, seven health systems, 3.2M patients, and 41.7M encounters.

Split and cohort construction. The strict 36-month split uses July 1, 2021–June 30, 2023 for training, July 1–December 31, 2023 for validation, and January 1–June 30, 2024 for testing. Sites 1–4 contribute to all three temporal phases; Sites 5–7 are external and contribute test data only. Patients are assigned by first qualifying encounter and appear in one partition; cross-site matches are resolved by privacy-preserving linkage before assignment. Orders, labels, retrieval indexes, q_ϕ, normalization statistics, prompts, and thresholds use only information available at prediction time and in the permitted partition. Repeated orders remain separate decisions, while uncertainty is clustered by patient.

Primary endpoint. Before any model output, a frozen reference protocol labels each qualifying inpatient order safe or unsafe from the full prespecified category definitions and adjudication criteria. It is separate from the runtime rule layer, which implements the executable action-upgrade subset: they share normalized inputs and version provenance, not coverage. Discharge, transfer, or administrative voiding before 60 minutes of observable follow-up is applied before analysis, yielding 847K orders from 312K patients that activate a high-risk category. After decoding and the runtime rules, ALERT, DEFER, and ESCALATE are intervention-positive; APPROVE is negative. A queue-reaching miss is a reference-unsafe order with final APPROVE, no modification or cancellation, and dispensing-queue entry within 60 minutes. A false alert is a reference-safe order with an intervention-positive action. Both rates divide by all 847K eligible decisions, measuring burden per 1,000 opportunities rather than class-conditional sensitivity or specificity. Repeated orders remain separate patient-clustered decisions; an exact re-order after modification starts a linked episode. MAR linkage and blinded pharmacist re-adjudication evaluate the endpoint.

Comparators and selection. Rules, XGBoost (Chen & Guestrin, 2016), GRU-D, a time-aware self-attention model, RAG+LLM (Lewis et al., 2020; Izacard et al., 2022), 7B constrained decoding,

OT-LLM, and the plain counterpart cover deterministic, temporal, retrieval, generation, and interface controls. Motivated by structured EHR and long-context studies (Fallahpour et al., 2025; Wornow et al., 2025), the strongest structured comparator adds learned channels for the same five evidence states, elapsed time, provenance, and normalized units to the shared 7B decoder. Its candidate universe, stable order, parameters, 32K context, 142B tokens, data, base heads, tuning budget, output grammar, and rule layer match ObserveTrace-7B; it omits OAD and ICC. Hyperparameters, retrieval k, prompts, and thresholds are selected once by the common validation miss–alert utility. Main point estimates average three initialization seeds; intervals and paired tests average order-aligned patient-cluster bootstrap replicates across seeds. Appendix Table S1 gives seed rates and Appendix A.1 gives the construction.

Clinical prefix and prospective protocols. The clinical ICC test forms adjacent held-out medication prefixes and stratifies them with the frozen protocol-change gate used during training. Gate-stable flip rate is the fraction whose argmax action differs; gate-changing action-update rate applies the same calculation after a gate change, irrespective of direction. Episode-cluster intervals use the prefix pair as unit. During the later 12-week, three-site prospective phase, model outputs were logged but never shown during care. Existing CPOE is positive when an interruptive alert appears; ObserveTrace-7B uses the same positive/negative action mapping on silent predictions. The prospective ledger applies the same independently assigned reference label and 60-minute queue-reaching definition to both action streams; observed modification or cancellation can therefore affect the operational miss endpoint but not the prediction-time action label. Afterward, 1,800 cases were sampled from the union positive under CPOE or ObserveTrace-7B, stratified by site and reference label, and rendered in both card formats. Two of 18 pharmacists independently reviewed each randomized, blinded pair after a seven-day washout; a third adjudicated helpful-status disagreements. Review time is median (interquartile range), helpful means 4–5 on a five-point scale, and intervals cluster by case and rater.

Cross-domain replication and stress tests. Five event-stream domains instantiate the same state-aware serialization principle: OULAD (Kuzilek et al., 2017), Alibaba (Lu et al., 2017), Google cluster traces (Reiss et al., 2012), Retailrocket sessions (Retailrocket, 2015), and MIMIC-IV (Johnson et al., 2023). Matched 350M models are trained separately within each domain on 256 A100-40GB GPUs for 18 hours; Alibaba and MIMIC-IV additionally repeat the comparison at 7B. Domain-native process states occupy a separate state field and are never relabeled as clinical observability. Field-removal variants retrain from the shared initialization and budget. Prespecified clinical stress tests cover workflow-conditioned drop, rule removal, external sites, the six-month held-out test period, at least 10 medications, eGFR below 30, ICU care, age above 75, and worst-site and worst-quarter slices.

5 RESULTS

A lower miss–alert frontier against the strongest matched comparator. ObserveTrace-7B's three-seed mean is 3.2 (2.8–3.7) queue-reaching misses per 1,000 eligible orders, equivalent to 2,710 events at the 847K denominator (Table 1). The information-matched structured comparator records 3.8 (3.3–4.4), 7B constrained decoding 4.2, RAG+LLM 4.8, and rules 9.4. The paired difference from the structured comparator is -0.6 (-1.0 to -0.2) per 1,000 ($P = 0.003$); the point estimates for high-severity misses and false alerts are also lower. Figure 2 shows that the ordering persists as the learned operating threshold moves, rather than arising from one selected point.

The interface, OAD, and ICC move different endpoints. Table 2 crosses the complete interface transformation with both objectives under identical architecture, 32K context, 142B-token budget, split, threshold selection, and rule version. With objectives off, replacing the plain source-order timeline lowers misses from 9.1 to 5.1 per 1,000. On interface input, OAD raises workflow-conditioned-drop retention from 0.73 to 0.83; ICC lowers the 40%-context flip rate from 0.31 to 0.17. Both objectives on plain input reach 6.8 (6.0–7.7) misses, 0.69 retention, and 0.27 flips, versus 3.2, 0.86, and 0.13 for the complete system. Against the information-matched learned-channel comparator, the complete system raises retention from 0.78 (0.75–0.80) to 0.86 and lowers truncation flips from 0.21 (0.19–0.23) to 0.13 while preserving the paired miss advantage. These comparisons attribute the plain-timeline contrast to the full interface transformation and the remaining changes to objective-specific behavior; rule-off results preserve the learned ordering (Appendix Table S15).

Table 1: Medication operating point on 847K eligible high-risk orders from 312K patients. Queue-reaching misses are reference-unsafe orders assigned APPROVE that enter the dispensing queue without modification or cancellation within 60 minutes; false alerts are reference-safe orders assigned an intervention-positive action. Rates use all eligible orders as denominator and report 95% patient-cluster bootstrap intervals ($B = 2{,}000$); they are burden rates, not class-conditional sensitivity or specificity. Thresholds are selected under one validation utility and the rule version is shared.

Metric	Rules	RAG+LLM	Structured long-context	ObserveTrace-7B
Queue-reaching misses / 1K	9.4 (8.6–10.3)	4.8 (4.2–5.5)	3.8 (3.3–4.4)	3.2 (2.8–3.7)
High-severity misses / 1K	3.8	1.9	1.4	1.1
False alerts / 1K	38.7 (36.2–41.3)	46.3 (43.1–49.6)	43.7 (41.0–46.5)	42.3 (39.8–44.9)

Table 2: Crossed interface–objective study. Every cell is trained separately from the shared initialization. Misses use the 847,000-order denominator; retention is performance at 50% workflow-conditioned drop divided by uncorrupted performance; flip rate is the fraction of predictions that change at 40% context. Entries are arithmetic means across three seeds; parenthetical counts are mean-rate-equivalent counts. Intervals quantify patient- or episode-cluster sampling variation, while Appendix Table S1 reports seed variation.

Input	OAD	ICC	Misses / 1K (rate-eq. count)	WCD retention	40% ctx flip
Plain	off	off	9.1 (8.2–10.1) (7708)	0.51	0.47
Plain	on	off	7.4 (6.6–8.3) (6268)	0.66	0.43
Plain	off	on	8.5 (7.6–9.5) (7200)	0.53	0.30
Plain	on	on	6.8 (6.0–7.7) (5760)	0.69	0.27
ObserveTrace	off	off	5.1 (4.5–5.8) (4320)	0.73	0.31
ObserveTrace	on	off	3.9 (3.4–4.5) (3303)	0.83	0.27
ObserveTrace	off	on	4.7 (4.2–5.3) (3981)	0.77	0.17
ObserveTrace	on	on	3.2 (2.8–3.7) (2710)	0.86	0.13

ICC reduces gate-relative prefix changes without depressing the gate-changing update rate. Table 3 evaluates ICC on the medication workflow using the rule-and-threshold protocol declared in Section 4. Among 128,640 eligible adjacent-prefix pairs without a gate-defined decision change, ObserveTrace-7B has a flip rate of 0.129 (0.121–0.138), below its no-ICC ablation, the structured comparator, and RAG+LLM. On 34,920 pairs with a threshold crossing or rule-status change, the action-update rate is 0.924 (0.912–0.935) versus 0.931 (0.920–0.942) without ICC. Their paired difference of -0.007 (-0.018 to 0.004) meets the prespecified -0.02 non-inferiority margin. The measured effect is a 0.089 reduction in gate-stable argmax changes with a non-inferior gate-changing action-update rate. Because the update endpoint counts either direction and uses the same frozen gate that defines ICC eligibility, the result establishes protocol-relative stability rather than independent clinical correctness of each transition.

Table 3: Clinical prefix stability on held-out medication episodes. Flip rate is the fraction of gate-stable pairs whose argmax action changes; action-update rate is the corresponding fraction among pairs with a newly changed rule state or protocol-threshold crossing, irrespective of direction. Intervals are 95% episode-cluster bootstrap intervals. Update non-inferiority was prespecified for the full versus no-ICC comparison; comparator rows report the common flip endpoint.

System	Gate-stable flip rate	Action-update rate
ObserveTrace-7B	0.129 (0.121–0.138)	0.924 (0.912–0.935)
ObserveTrace-7B without ICC	0.218 (0.206–0.231)	0.931 (0.920–0.942)
Structured long-context	0.236 (0.223–0.249)	—
RAG+LLM	0.264 (0.250–0.279)	—

For Figure 2b, threshold-weighted utility at action cutoff p is $U(p) = \mathrm{TP}(p)/N - \mathrm{FP}(p)/N \cdot p/(1-p)$, where reference-unsafe orders are positive and $N = 847$K. The score is $q(x) = 1 - p_\theta(\text{APPROVE} \mid x)$; p is an action cutoff, not a calibrated unsafe-outcome probability. The sweep moves this cutoff while the protocol gate and rule layer remain fixed.

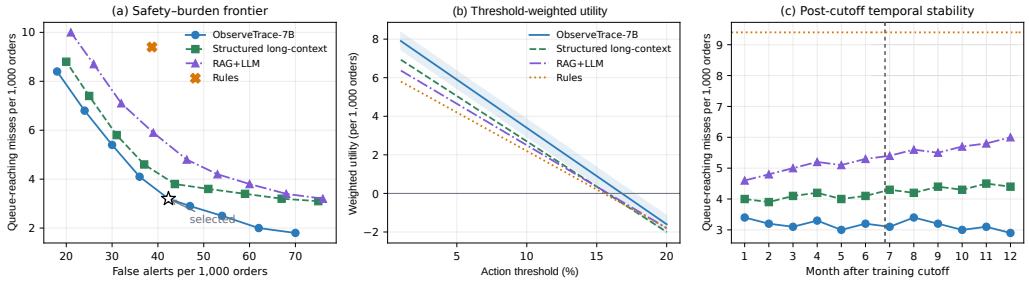

Figure 2: Operating behavior beyond one selected threshold. (a) Queue-reaching misses versus false alerts per 1,000 eligible orders; the marker is ObserveTrace-7B's validation-selected default. (b) Threshold-weighted utility over action cutoffs; the band uses patient-cluster bootstrap. (c) Monthly queue-reaching miss rates for validation months 1–6 and untouched test months 7–12, separated by the vertical marker. Encodings are shared across panels and identified in (a)–(b); site-stratified estimates appear in Appendix Table S3.

The ordering persists under field removal, rule removal, difficult cohorts, and later workflows. Retrained removal of observability raises misses by 141%, temporal gaps by 109%, and canonicalization by 59% (Appendix Table S9), making observability the strongest named-field intervention. With rules disabled, ObserveTrace-7B remains ahead of RAG+LLM (4.7 versus 7.2 per 1,000). Relative miss reductions versus RAG+LLM persist at the worst site (27%), worst quarter (30%), at least 10 medications (23%), and eGFR below 30 (31%) (Appendix Table S21). In the later 184.4K-order silent phase, logged ObserveTrace-7B actions yield 2.9–3.4 queue-reaching misses per 1,000 across Epic, Cerner, and MEDITECH; observed existing-CPOE actions yield 8.1–9.2 under the same operational definition. The separate blinded card study records 87–91% helpful ratings versus 58–64% and shorter review time at every site (Appendix Table S20).

Secondary replications preserve the state-aware ordering. Across 18 clinical tasks, site-macro scores rise from 0.81 to 0.89 at the four internal sites and from 0.77 to 0.86 at the three external sites (Appendix Table S28); pooled task-native deltas favor ObserveTrace for all 18 tasks and average 5.7 percentage points (Appendix Figure S1). In separate OULAD, Alibaba, Google, Retailrocket, and MIMIC-IV experiments, the 350M complete model improves accuracy by 12–17%, reaches 0.859 mean retention at 50% workflow-conditioned drop, and reduces truncation flips by 63–68% relative to plain sequences. Alibaba and MIMIC-IV retain all three advantages at 7B (Appendix Table S27). These studies support portability of deterministic state-aware serialization and the objectives across event vocabularies; the medication study remains the direct test of the five clinical evidence states.

6 Discussion

ObserveTrace changes the unit of longitudinal modeling from a recorded event to a decision-time evidence witness. A required slot receives one canonical value, provenance, time, and observability state that remains addressable through prediction, constrained output, rule evaluation, and review. This shared identity explains why the contribution is stronger than a missingness embedding: the same field can be intervened on, audited, and enforced after the model has consumed it. On the primary operating point, the complete method yields 3.2 queue-reaching misses per 1,000 eligible orders, a paired reduction of 0.6 (0.2–1.0) from the information-matched learned-channel comparator, together with lower high-severity and false-alert point estimates.

The crossed study assigns different empirical roles to the components. The full interface transformation—required-slot closure, canonical ordering, and explicit time, unit, provenance, and observability—provides the largest miss-rate contrast before OAD or ICC. OAD contributes most under workflow-conditioned evidence loss, matching its corruption model. ICC lowers gate-stable argmax changes while meeting the prespecified non-inferiority margin for the gate-changing update rate. Field removals and the rule-off comparison reinforce this account: observability is the largest named-field intervention, and the learned ordering remains when the common deterministic safety floor is bypassed.

The primary rate is an operational burden estimand over all 847K eligible decisions, not a class-conditional sensitivity estimate. Its queue-reaching condition incorporates what happened during the following 60 minutes, so the silent model and active-CPOE comparison can reflect contemporaneous care-process behavior as well as prediction-time actions. MAR concordance and blinded re-adjudication support this endpoint, while the silent phase establishes temporal and vendor persistence under the frozen action mapping rather than an effect of surfaced alerts on patient outcomes. Similarly, ICC establishes argmax stability relative to its declared rule-and-threshold gate; its update rate counts either direction. These qualifications preserve the clinical meaning of the reported comparisons without weakening the interface and objective effects directly measured by the factorial, field, and prefix experiments.

Named evidence states also make residual variation actionable. External-only medications, delayed renal labs, local formulary changes, and high medication counts remain visible as monitorable states rather than unexplained model errors. Every-site paired effects, worst-quarter and difficult-cohort analyses, 18 clinical tasks, and five event-stream replications show that the ordering is not produced by one institution, schema, or vocabulary. The resulting design principle is concrete: define required evidence before interpreting absence, preserve its identity through prediction and enforcement, and train robustness and prefix stability against transitions on that same representation.

7 Conclusion

A blank field is not an evidence state. ObserveTrace turns every activated requirement into a canonical witness whose value, time, provenance, and observability remain shared by learning, constrained output, rules, and review. Across seven health systems, this boundary yields 3.2 queue-reaching misses per 1,000 eligible orders versus 3.8 (3.3–4.4) for the strongest information-matched comparator; the crossed study further isolates a large interface contrast, OAD-specific robustness, and ICC-specific gate-relative stability. Preserving both what was recorded and why required evidence was unavailable makes partially observed longitudinal decisions more reliable and inspectable at every stage.

Ethics Statement

Each participating institution approved the retrospective study through its Institutional Review Board or equivalent ethics committee, with informed-consent waivers where applicable. Analyses used de-identified EHR data under HIPAA, local policy, executed data-use agreements, secure computing, and audited access controls; only aggregate non-identifiable results are reported. The prospective evaluation was silent and involved no patient contact or clinical intervention. MIMIC-IV was accessed under PhysioNet credentialing and its data-use requirements. Evidence-linked outputs received offline human review after silent collection; site-specific threshold monitoring and deterministic contraindication and dosing checks remained active in the evaluation pipeline.

Reproducibility Statement

Section 3 specifies the deterministic ObserveTrace grammar, complete training objective, and corruption and consistency procedures; Section 4 defines split construction, the endpoint, comparator matching, threshold selection, and uncertainty estimators. The appendix provides algorithms, cohort and baseline specifications, hyperparameters, per-seed factorial rates and rate-equivalent counts, prefix-pair construction, stress tests, and full cross-domain mappings. Public-domain evaluations on MIMIC-IV, OULAD, Alibaba, and Google cluster traces provide independently accessible instantiations of the interface and evaluation protocol; partner EHR analyses follow the same documented definitions under institutional privacy and governance requirements.

Manuscript Preparation Disclosure

Generative AI tools assisted the manuscript-preparation workflow, including editorial revision, formatting and consistency review, and reference checking. This assistance informed revisions to the system name, title, abstract, and introduction. This disclosure describes the recorded manuscript-preparation assistance.

REFERENCES

Muhammad Ayaz, Muhammad Faheem Pasha, Muhammad Yousuf Ala, Muhammad Y Alzahrani, Rahmat Budiarto, and Deris Stiawan. The fast health interoperability resources (FHIR) standard: Systematic literature review of implementations, applications, challenges and opportunities. *JMIR Medical Informatics*, 9(7):e21929, 2021. doi: 10.2196/21929. URL https://medinform.jmir.org/2021/7/e21929.

Inci M Baytas, Cao Xiao, Xi Zhang, Fei Wang, Anil K Jain, and Jiayu Zhou. Patient subtyping via time-aware LSTM networks. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 65–74, 2017. doi: 10.1145/3097983.3097997.

Zhengping Che, Sanjay Purushotham, Kyunghyun Cho, David Sontag, and Yan Liu. Recurrent neural networks for multivariate time series with missing values. *Scientific Reports*, 8(1):6085, 2018. doi: 10.1038/s41598-018-24271-9.

Tianqi Chen and Carlos Guestrin. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785–794. ACM, 2016. doi: 10.1145/2939672.2939785. URL https://doi.org/10.1145/2939672.2939785.

Wenhu Chen, Jianshu Chen, Yu Su, Zhiyu Chen, and William Yang Wang. Logical natural language generation from open-domain tables. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 7929–7942, 2020. doi: 10.18653/v1/2020.acl-main.708.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of NAACL-HLT*, pp. 4171–4186, 2019. doi: 10.18653/v1/n19-1423. URL https://aclanthology.org/N19-1423/.

Adibvafa Fallahpour, Mahshid Alinoori, Wenqian Ye, Xu Cao, Arash Afkanpour, and Amrit Krishnan. EHRMamba: Towards generalizable and scalable foundation models for electronic health records. In *Proceedings of the 4th Machine Learning for Health Symposium*, volume 259 of *Proceedings of Machine Learning Research*, pp. 291–307. PMLR, 2025. URL https://proceedings.mlr.press/v259/fallahpour25a.html.

Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. Unsupervised dense information retrieval with contrastive learning. *Transactions on Machine Learning Research*, 2022. URL https://openreview.net/forum?id=jKN1pXi7b0.

Lavender Yao Jiang, Xujin Chris Liu, Nima Pour Nejatian, Mustafa Nasir-Moin, Duo Wang, Anas Abidin, Kevin Eaton, Howard Antony Riina, Ilya Laufer, Paawan Punjabi, Madeline Miceli, Nora C. Kim, Cordelia Orillac, Zane Schnurman, Christopher Livia, Hannah Weiss, David Kurland, Sean Neifert, Yosef Dastagirzada, Douglas Kondziolka, Alexander T. M. Cheung, Grace Yang, Ming Cao, Mona Flores, Anthony B. Costa, Yindalon Aphinyanaphongs, Kyunghyun Cho, and Eric Karl Oermann. Health system-scale language models are all-purpose prediction engines. *Nature*, 619 (7969):357–362, 2023. doi: 10.1038/s41586-023-06160-y.

Alistair EW Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad 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. doi: 10.1038/s41597-022-01899-x.

Željko Kraljević, Daniel Bean, Anthony Shek, Rebecca Bendayan, Harry Hemingway, Joshua Au Yeung, Alexander Deng, Alfred Balston, Jack Ross, Esther Idowu, James 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. doi: 10.1016/S2589-7500(24)00025-6. URL https://doi.org/10.1016/S2589-7500(24)00025-6.

Jakub Kuzilek, Martin Hlosta, and Zdenek Zdrahal. Open university learning analytics dataset. *Scientific Data*, 4(1):170171, 2017. doi: 10.1038/sdata.2017.171.

Chia-Hsuan Lee, Hao Cheng, and Mari Ostendorf. Dialogue state tracking with a language model using schema-driven prompting. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 4937–4949, 2021. doi: 10.18653/v1/2021.emnlp-main.404.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems*, volume 33, pp. 9459–9474, 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html.

Sha Li, Heng Ji, and Jiawei Han. Document-level event argument extraction by conditional generation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 894–908, 2021. doi: 10.18653/v1/2021.naacl-main.69.

Yikuan Li, Shishir Rao, José Roberto Ayala Solares, Abdelaali Hassaine, Rema Ramakrishnan, Dexter Canoy, Yajie Zhu, Kazem Rahimi, and Gholamreza Salimi-Khorshidi. BEHRT: Transformer for electronic health records. *Scientific Reports*, 10(1):7155, 2020. doi: 10.1038/s41598-020-62922-y.

Zachary C Lipton, David C Kale, and Randall Wetzel. Directly modeling missing data in sequences with RNNs: Improved classification of clinical time series. In *Proceedings of the 1st Machine Learning for Healthcare Conference*, volume 56 of *Proceedings of Machine Learning Research*, pp. 253–270, 2016. URL https://proceedings.mlr.press/v56/Lipton16.html.

Roderick JA Little and Donald B Rubin. *Statistical Analysis with Missing Data*. Wiley, New York, 2nd edition, 2002. doi: 10.1002/9781119013563. URL https://doi.org/10.1002/9781119013563.

Chengzhi Lu, Kejiang Ye, Guoyao Xu, Cheng-Zhong Xu, and Tongxin Bai. Imbalance in the cloud: An analysis on alibaba cluster trace. In *2017 IEEE International Conference on Big Data (Big Data)*, pp. 2884–2892. IEEE, 2017. doi: 10.1109/bigdata.2017.8258257.

Observational Health Data Sciences and Informatics. The OMOP common data model. *OHDSI*, 2019. URL https://ohdsi.github.io/CommonDataModel/.

Chao Pang, Xinzhuo Jiang, Nishanth Parameshwar Pavinkurve, Krishna S. Kalluri, Elise L. Minto, Jason Patterson, Linying Zhang, George Hripcsak, Gamze Gürsoy, Noémie Elhadad, and Karthik Natarajan. CEHR-GPT: Generating electronic health records with chronological patient timelines. *arXiv preprint arXiv:2402.04400*, 2024. doi: 10.48550/arXiv.2402.04400. URL https://arxiv.org/abs/2402.04400.

Shobha Phansalkar, Judy Edworthy, Elizabeth Hellier, Diane L Seger, Angela Schedlbauer, Anthony J Avery, and David W Bates. A review of human factors principles for the design and implementation of medication safety alerts in clinical information systems. *Journal of the American Medical Informatics Association*, 17(5):493–501, 2010. doi: 10.1136/jamia.2010.005264.

Laila Rasmy, Yang Xiang, Ziqian Xie, Cui Tao, and Degui Zhi. Med-BERT: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. *NPJ Digital Medicine*, 4(1):86, 2021. doi: 10.1038/s41746-021-00455-y.

Abhinav Rastogi, Xiaoxue Zang, Srinivas Sunkara, Raghav Gupta, and Pranav Khaitan. Towards scalable multi-domain conversational agents: The schema-guided dialogue dataset. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 8689–8696, 2020. doi: 10.1609/aaai.v34i05.6394.

Charles Reiss, Alexey Tumanov, Gregory R Ganger, Randy H Katz, and Michael A Kozuch. Heterogeneity and dynamicity of clouds at scale: Google trace analysis. In *Proceedings of the Third ACM Symposium on Cloud Computing*, pp. 1–13, 2012. doi: 10.1145/2391229.2391236. URL https://doi.org/10.1145/2391229.2391236.

Retailrocket. Retailrocket recommender system dataset. Kaggle, 2015. URL https://www.kaggle.com/datasets/retailrocket/ecommerce-dataset.

Leonardo F R Ribeiro, Martin Schmitt, Hinrich Schütze, and Iryna Gurevych. Investigating pre-trained language models for graph-to-text generation. In *Proceedings of the 3rd Workshop on Natural Language Processing for Conversational AI*, pp. 211–227, 2021. doi: 10.18653/v1/2021.nlp4convai-1.20. URL https://aclanthology.org/2021.nlp4convai-1.20/.

Donald B Rubin. Inference and missing data. *Biometrika*, 63(3):581–592, 1976. doi: 10.1093/biomet/63.3.581.

Dean F Sittig, Adam Wright, Jerome A Osheroff, Blackford Middleton, Jonathan M Teich, Joan S Ash, Emily Campbell, and David W Bates. Grand challenges in clinical decision support. *Journal of Biomedical Informatics*, 41(2):387–392, 2008. doi: 10.1016/j.jbi.2007.09.003.

Huan Song, Deepta Rajan, Jayaraman J Thiagarajan, and Andreas Spanias. Attend and diagnose: Clinical time series analysis using attention models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, pp. 4091–4098, 2018. doi: 10.1609/aaai.v32i1.11635.

Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *Advances in Neural Information Processing Systems*, volume 30, pp. 1195–1204, 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/hash/68053af2923e00204c3ca7c6a3150cf7-Abstract.html.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023. doi: 10.48550/arXiv.2307.09288.

Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th International Conference on Machine Learning*, pp. 1096–1103. ACM, 2008. doi: 10.1145/1390156.1390294. URL https://doi.org/10.1145/1390156.1390294.

Nicole G Weiskopf and Chunhua Weng. Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research. *Journal of the American Medical Informatics Association*, 20(1):144–151, 2013. doi: 10.1136/amiajnl-2011-000681.

Brandon T Willard and Rémi Louf. Efficient guided generation for large language models. *arXiv preprint arXiv:2307.09702*, 2023. doi: 10.48550/arXiv.2307.09702. URL https://arxiv.org/abs/2307.09702.

Michael Wornow, Rahul Thapa, Ethan Steinberg, Jason A. Fries, and Nigam H. Shah. EHRSHOT: An EHR benchmark for few-shot evaluation of foundation models. In *Advances in Neural Information Processing Systems*, volume 36, pp. 67125–67137, 2023a. URL https://openreview.net/forum?id=CsXC6IcdwI.

Michael Wornow, Yizhe Xu, Rahul Thapa, Birju Patel, Ethan Steinberg, Scott Fleming, Michael A Pfeffer, Jason Fries, and Nigam H Shah. The shaky foundations of large language models and foundation models for electronic health records. *npj Digital Medicine*, 6(1):135, 2023b. doi: 10.1038/s41746-023-00879-8.

Michael Wornow, Suhana Bedi, Miguel Angel Fuentes Hernandez, Ethan Steinberg, Jason Fries, Christopher Re, Sanmi Koyejo, and Nigam Shah. Context clues: Evaluating long context models for clinical prediction tasks on EHR data. In *International Conference on Learning Representations*, 2025. URL https://iclr.cc/virtual/2025/poster/27657.

Adam Wright, Thu-Trang T. Hickman, Dustin McEvoy, Skye Aaron, Angela Ai, Jan Marie Andersen, Salman T. Hussain, Rachel Ramoni, Julie Fiskio, Dean F. Sittig, and David W. Bates. Analysis of clinical decision support system malfunctions: a case series and survey. *Journal of the American Medical Informatics Association*, 23(6):1068–1076, 2016. doi: 10.1093/jamia/ocw005.

A SUPPLEMENTARY EVIDENCE

A.1 MECHANISM AND FACTORIAL CONTROLS

Source-record transformation. Algorithm S1 closes the candidate-universe and state-precedence definitions used by every clinical experiment. Required slots are emitted only after a declared decision trigger; the interface never materializes arbitrary absent diagnoses, tests, or medications. The frozen medication table contains six slot families: renal function for renal adjustment, hepatic function for hepatic adjustment, allergy status for the ordered ingredient and class, active medications for interaction rules, patient factors used by dose bounds, and local formulary status. Each row declares its code set, source adapters, lookback, and staleness horizon. Creatinine uses 30 days in the evaluated dosing rules; other horizons follow the versioned drug- and analyte-specific guideline or formulary row used by the corresponding rule. Site variation enters through the site's frozen formulary version, not through post-test adaptation.

Algorithm S1 Canonical source-record-to-evidence-state transformation

Require: Available source records $R_{\le t}$, prediction trigger r_t, frozen slot table $\mathcal{T}$, horizons H, calibrated source confidences c
Ensure: Canonical event sequence $x_{\le t}$ and deterministic slot audit map A_t
$E \leftarrow$ canonicalize every record in $R_{\le t}$ once, keyed by its source identity
$A_t \leftarrow \varnothing$
$Z \leftarrow \{z : z \text{ is activated by } r_t \text{ in } \mathcal{T}\}$
for all $z \in Z$ **do**
    $M_z \leftarrow$ records in E matching the code, source set, and lookback of z
    if M_z contains a value-bearing record **then**
        $A_t[(\mathrm{key}(r_t), \mathrm{key}(z))] \leftarrow$ latest canonical value-bearing key in M_z
    else if M_z contains an external pointer without an accessible value **then**
        append $(\textsc{required}, \mathrm{key}(z), t, +0\,\mathrm{s}, \mathtt{NA}, \mathtt{NA}, \mathrm{external}, \textsc{external_only})$, keyed by (r_t, z)
    else
        append $(\textsc{required}, \mathrm{key}(z), t, +0\,\mathrm{s}, \mathtt{NA}, \mathtt{NA}, \mathrm{none}, \textsc{not_measured})$, keyed by (r_t, z)
    end if
end for
remove duplicate keys from E
for all $e \in E$ **do**
    if e is value-bearing **then**
        assign LOW_CONFIDENCE if $c_e < 0.85$, else STALE if $t - t_e > H(\mathrm{code}_e, \mathrm{setting}_e)$, else MEASURED
    end if
    normalize code, value, unit, anchor, elapsed time, and provenance for e
    percent-encode reserved UTF-8 bytes and emit fields in the fixed grammar order
end for
compute tie digest from the canonical raw record or synthetic $(r_t, \mathcal{T}\text{-version}, z, o)$
stably sort by the declared fields, then the tie digest
reject invalid keys or escapes and require parse–serialize byte equality
return $(x_{\le t}, A_t)$

Workflow-conditioned corruption. Algorithm S2 estimates OAD corruption from training trajectories only. Hierarchical fallback retains event-type support in sparse site–setting strata, and the fitted distribution is frozen before validation.

Protocol-change gate. Algorithm S3 exempts prefix pairs from ICC whenever newly available evidence changes a frozen rule state or clinical threshold.

Per-seed factorial results. Table S1 gives the three initialization runs underlying the aggregate crossed study. Every cell uses the same 7B backbone, 32K context, 142B-token budget, split, output grammar, validation threshold-selection rule, and rule version.

Algorithm S2 Training-only workflow gap estimator for OAD

Require: Training trajectories $\mathcal{D}_{\text{train}}$, event types $\mathcal{T}$, settings $\mathcal{S}$, sites $\mathcal{H}$, frozen horizons W
Ensure: Frozen corruption distribution q_ϕ

for all $(t, s, h) \in \mathcal{T} \times \mathcal{S} \times \mathcal{H}$ **do**
    collect gaps, right-censoring indicators, and delayed, external-only, stale, and low-confidence transitions
    if the stratum contains at least 500 transitions **then**
        fit empirical transition probabilities and a censored log-normal gap model
    else if (t, s) contains at least 500 transitions **then**
        use the event–setting fit with empirical-Bayes site shrinkage
    else
        use the event-type fit with empirical-Bayes setting shrinkage
    end if
    set complete-removal mass to the fitted gap survival probability $\widehat{S}_{t,s,h}(W_{t,s})$
    allocate remaining mass in proportion to the observed transition frequencies
end for
normalize support; freeze ϕ; sample a 10–50% batch mask rate during training
return $q_\phi(m \mid t, s, h)$

Algorithm S3 Protocol-change detection for ICC

Require: Prefixes $x_{\le t}$ and $x_{\le t+\delta}$, frozen rule layer $\mathcal{R}$, threshold table Θ
Ensure: Gate $g_{t,\delta} \in \{0, 1\}$

compute rule states $r_t = \mathcal{R}(x_{\le t})$ and $r_{t+\delta} = \mathcal{R}(x_{\le t+\delta})$
if $r_t \neq r_{t+\delta}$ **then**
    return 0
end if
for all new laboratory, vital, medication, allergy, diagnosis, and procedure events in $(t, t + \delta]$ **do**
    if the event crosses a decision threshold in Θ or changes contraindication status **then**
        return 0
    end if
end for
return 1

Table S1: Per-seed interface–objective factorial results. Aggregate rates are arithmetic means of the three displayed seeds. Parenthetical counts equal the mean rate times 847,000, rounded to the nearest order; they are rate-equivalent counts, not a pooled fourth run. Patient-cluster intervals quantify sampling variation, while the seed columns expose initialization variation.

Input	OAD	ICC	Seed 1	Seed 2	Seed 3	Aggregate / 1K (rate-eq. count)	Difference vs. full
Plain	off	off	9.0	9.3	9.1	9.1 (8.2–10.1) (7708)	+5.9 (+5.0 to +6.8)
Plain	on	off	7.2	7.6	7.4	7.4 (6.6–8.3) (6268)	+4.2 (+3.4 to +5.0)
Plain	off	on	8.3	8.7	8.5	8.5 (7.6–9.5) (7200)	+5.3 (+4.5 to +6.2)
Plain	on	on	6.6	7.0	6.8	6.8 (6.0–7.7) (5760)	+3.6 (+2.8 to +4.4)
ObserveTrace	off	off	5.0	5.2	5.1	5.1 (4.5–5.8) (4320)	+1.9 (+1.3 to +2.5)
ObserveTrace	on	off	3.8	4.0	3.9	3.9 (3.4–4.5) (3303)	+0.7 (+0.3 to +1.1)
ObserveTrace	off	on	4.6	4.8	4.7	4.7 (4.2–5.3) (3981)	+1.5 (+1.0 to +2.0)
ObserveTrace	on	on	3.1	3.3	3.2	3.2 (2.8–3.7) (2710)	reference

Seed and uncertainty construction. For each of 2,000 replicate indices, patients are sampled with replacement within site and all of their eligible orders are retained. Each seed is evaluated on the same sampled patient clusters; seed-specific rates or paired differences are then averaged with equal weight. Reported point estimates average the three original seed statistics, and 95% intervals are the 2.5th and 97.5th percentiles of the averaged replicate distribution. The two-sided paired P value is the bootstrap tail probability of the averaged, order-aligned difference relative to zero. Prefix and card endpoints use the same construction with linked episode and case–rater clusters, respectively.

Table S2: Alert burden for the crossed factorial cells. Each cell uses its own operating threshold selected under the common validation miss–alert utility.

Input	OAD	ICC	False alerts / 1K
Plain	off	off	51.8
Plain	on	off	49.2
Plain	off	on	48.6
Plain	on	on	46.9
ObserveTrace	off	off	47.8
ObserveTrace	on	off	44.7
ObserveTrace	off	on	45.6
ObserveTrace	on	on	42.3

Per-site comparator effects. The paired advantage over the contemporary structured baseline is present at every held-out site, with the largest absolute rate at Site 5.

Table S3: Site-level queue-reaching misses per 1,000 eligible orders for the structured long-context comparator and ObserveTrace-7B. Paired differences use patient-cluster bootstrap within site.

Site	Structured long-context	ObserveTrace-7B	Paired difference
1	3.4	2.8	-0.6 (-1.1 to -0.2)
2	3.6	3.0	-0.6 (-1.0 to -0.2)
3	3.8	3.1	-0.7 (-1.2 to -0.2)
4	3.7	3.2	-0.5 (-0.9 to -0.1)
5	4.6	4.1	-0.5 (-1.0 to -0.1)
6	3.5	3.0	-0.5 (-0.9 to -0.1)
7	3.9	3.3	-0.6 (-1.1 to -0.2)

Matched configuration. The modern comparator combines the long-context evaluation regime and structured EHR design choices studied by EHRMamba and Context Clues (Fallahpour et al., 2025; Wornow et al., 2025). It receives the identical candidate universe, normalized values and units, canonical event order, time anchors and gaps, provenance, and five state categories as ObserveTrace-7B. Learned embeddings carry the nonvalue fields into the same 7B decoder. Both systems pass their decoded action to a model-independent rule API that reads the same canonical source tuple, so rule inputs remain identical. The comparator uses the three base tasks without OAD or ICC; the factorial matrix evaluates those objectives separately. The comparison is therefore matched on source information, backbone capacity, 142B-token training budget, tuning budget, and deterministic enforcement, rather than being a representation-only ablation.

All runs use a 2% linear warmup followed by cosine decay, gradient clipping at 1.0, dropout 0.1, bf16 training, activation checkpointing, and validation selection by miss–alert utility. Each system processes 142B training tokens; OAD examples replace rather than add tokens, and ICC uses a matched number of prefix pairs per optimizer step.

Serialization grammar. The fixed surface grammar below supports exact parse–serialize round trips. Fields appear in the listed order; `NA` is the only null token. A value byte outside `SAFE`, including a delimiter, percent sign, or control character, is encoded as `%HH`. For raw events, the SHA-256 tie digest covers the normalized source namespace, table, row key, availability time, code, value, unit, provenance, and source state. For synthetic required-slot events, it covers the trigger key, slot-table version, slot key, and observability state. The digest is removed before model tokenization.

```
EVENT  := "[" TYPE "|" CODE "|" ANCHOR "|" DELTA "|"
          VALUE "|" UNIT "|" PROV "|" OBS "]"
TYPE   := "LAB" | "MED" | "DX" | "PROC" | "VITAL" | "NOTE" |
          "REQUIRED"
CODE   := "code:" TEXT
ANCHOR := "t_anchor:" TEXT
```

Table S4: Input and supervision boundary for the three central controls. "Plain versus ObserveTrace" estimates the complete source-to-interface transformation; "structured versus complete" compares an information-matched learned-channel system with the full method.

Property	Plain timeline	Structured long-context	ObserveTrace-7B
Candidate universe	Observed source events	Observed events plus activated required slots	Same as structured
Value-bearing input	Source type, code, value/unit, timestamp	Normalized value/unit tuple	Same normalized tuple
Required absence	No synthetic event	Learned five-state channel	Named REQUIRED event
Time and provenance	Source timestamp only	Learned anchor, gap, and provenance channels	Named canonical fields
Order and identity	Native source order	Stable canonical order	Stable order plus parse–serialize identity
Base tasks	CRC, NOCP including action, TCEM	CRC, NOCP including action, TCEM	CRC, NOCP including action, TCEM
Transition objectives	Crossed OAD/ICC; source-visible reconstruction only	Neither OAD nor ICC	OAD and ICC over canonical fields
Output and rules	Shared grammar and frozen rule API	Shared grammar and frozen rule API	Shared grammar and frozen rule API

Table S5: Baseline transparency. Learned 7B comparators share data, task supervision, validation tuning, structured output grammar, and the same post-decoding rule layer; the structured long-context row additionally receives the information-matched state, unit, time, and provenance channels described above.

System	Params	Context	Train tokens	Retrieval	Latency	Rule layer
Rules	—	—	—	—	12 ms	Frozen shared
GRU-D (Che et al., 2018)	0.08B	2K	12.4B	—	38 ms	Frozen shared
Time-aware self-attn. (Song et al., 2018)	0.15B	8K	18.6B	—	52 ms	Frozen shared
RAG+LLM	7B	32K	142B	BM25+Contriever, $k = 5$	156 ms	Frozen shared
LLM-7B+Constrained	7B	32K	156B	—	178 ms	Frozen shared
Structured long-context	7B	32K	142B	—	151 ms	Frozen shared
OT-LLM	7B	32K	142B	—	142 ms	Frozen shared
ObserveTrace-7B	7B	32K	142B	—	148 ms	Frozen shared

Table S6: Hyperparameter parity for the 7B factorial and strongest comparator runs. Learning rates and stopping rules are selected on validation data and then fixed across seeds.

System	Initialization	Optimizer	Peak LR	Weight decay	Global batch	Context	Seeds
Plain sequence	Llama-2-7B	AdamW	1.5×10^{-5}	0.1	4M tokens	32K	3
OT-LLM	Llama-2-7B	AdamW	1.5×10^{-5}	0.1	4M tokens	32K	3
Structured long-context	Llama-2-7B	AdamW	1.5×10^{-5}	0.1	4M tokens	32K	3
ObserveTrace-7B	Llama-2-7B	AdamW	1.5×10^{-5}	0.1	4M tokens	32K	3

Table S7: Matched contemporary comparator on the complete safety–robustness–stability endpoint set. Misses and false alerts are per 1,000 high-risk orders; flip rate is measured after truncating context to 40%.

System	Queue-reaching misses	False alerts	WCD retention	Flip rate
Structured long-context	3.8 (3.3–4.4)	43.7 (41.0–46.5)	0.78 (0.75–0.80)	0.21 (0.19–0.23)
ObserveTrace-7B	3.2 (2.8–3.7)	42.3 (39.8–44.9)	0.86	0.13

Table S8: Validation sweep for retrieval, constrained-decoding, and structured long-context comparators. Test results are shown only for the validation-selected configuration of each baseline.

Baseline	Configuration	Retrieval k	Context	Validation misses / 1K	Test misses / 1K
RAG+LLM	structured-v1, greedy	3	16K	5.2	—
RAG+LLM	structured-v2, greedy (selected)	5	32K	4.6	4.8
RAG+LLM	structured-v2, temp. 0.2	10	32K	4.8	—
LLM-7B+Constrained	constrained	—	32K	4.4	—
LLM-7B+Constrained	constrained+SC5 (selected)	—	32K	4.1	4.2
Structured long-context	learned time/missingness	—	16K	4.2	—
Structured long-context	learned time/missingness (selected)	—	32K	3.7	3.8

```
DELTA  := "delta:" SIGNED_DURATION
VALUE  := "val:" (TEXT | "NA")
UNIT   := "unit:" (TEXT | "NA")
PROV   := "prov:" TEXT
OBS    := "obs:" ("measured" | "not_measured" |
          "external_only" | "stale" | "low_confidence")
TEXT   := (SAFE | ESCAPE)+
ESCAPE := "%" HEX HEX
SAFE   := ALPHA | DIGIT | "-" | "_" | "." | "/" | "+"
HEX    := DIGIT | "A" | "B" | "C" | "D" | "E" | "F"
SIGNED_DURATION := ("+" | "-") DIGIT+ DURATION_UNIT
DURATION_UNIT := "s" | "m" | "h" | "d" | "w"
ALPHA  := "A"..."Z" | "a"..."z"
DIGIT  := "0"..."9"
```

The parser rejects missing, repeated, reordered, or unknown fields, malformed escapes, non-canonical numeric forms, and invalid state–value combinations. In particular, NOT_MEASURED and EXTERNAL_ONLY require `val:NA` and `unit:NA`. Imported external values instead use `prov:external` with a value-bearing observability state.

Retrained field interventions. Every named-field row removes that field from the learned model's train, validation, and test serialization and from its corresponding CRC reconstruction target, then retrains from the common initialization and token budget. OAD omits a removed field from its reconstruction sum; ICC sampling and the frozen protocol-change gate remain unchanged. The model-independent rule API and card renderer continue to read the full canonical source tuple, isolating dependence of the learned component. The canonicalization row retains every field and normalized unit but replaces stable ordering and canonical duplicate resolution with native source order and source surface forms. The P column uses patient-cluster paired bootstrap tests with Holm correction across the six named interventions.

Table S9: Retrained ObserveTrace field removals under fixed initialization, architecture, objectives, token budget, tuning budget, and evaluation. WCD denotes retention at 50% workflow-conditioned drop; flip rate is measured after truncating context to 40%.

Variant	Misses / 1K	Change	High severity	WCD retention	Flip rate	P
ObserveTrace-7B	3.2 (2.8–3.7)	—	1.1	0.86	0.13	—
− time anchor	4.8 (4.2–5.5)	+50%	1.7	0.74	0.21	< 0.001
− temporal gap	6.7 (5.9–7.6)	+109%	2.4	0.68	0.28	< 0.001
− units	4.4 (3.8–5.1)	+38%	1.5	0.81	0.16	< 0.001
− provenance	4.1 (3.6–4.7)	+28%	1.4	0.79	0.17	< 0.001
− observability	7.7 (6.8–8.7)	+141%	2.8	0.62	0.31	< 0.001
− canonicalization	5.1 (4.5–5.8)	+59%	1.8	0.76	0.22	< 0.001
Plain sequence	9.1 (8.2–10.1)	+184%	3.4	0.51	0.47	< 0.001

Controlled missingness. OAD trained on workflow-conditioned corruption yields its largest advantage under the corresponding observation process on the 18-task clinical diagnostic.

Table S10: Mean task performance under IID and workflow-conditioned drop (WCD). Entries are unweighted means of the task-specific primary metrics; the 0% columns give uncorrupted performance.

	IID drop				Workflow-conditioned drop			
Model	0%	20%	40%	50%	0%	20%	40%	50%
Without OAD	0.87	0.80	0.69	0.62	0.87	0.74	0.57	0.47
OAD, IID-trained	0.87	0.81	0.71	0.65	0.87	0.76	0.61	0.52
OAD, WCD-trained	0.87	0.82	0.73	0.67	0.87	0.80	0.69	0.62
WCD-trained gain	0.00	+0.02	+0.04	+0.05	0.00	+0.06	+0.12	+0.15

Clinical prefix construction. The held-out test timeline yields 128,640 adjacent gate-stable pairs and 34,920 pairs with a threshold or rule-status change. Pair counts cover seven sites and medication, laboratory, allergy, diagnosis, procedure, and external-reconciliation events. Gate-stable flip rate counts pairs whose argmax action changes; gate-changing action-update rate applies the same calculation after the protocol gate changes, irrespective of direction. Because the same frozen protocol defines ICC eligibility and these strata, "stable" and "changing" refer specifically to that protocol.

Table S11: Direct medication-workflow ICC test. Intervals use 2,000 episode-cluster bootstrap replicates. The gate-changing action-update difference is tested against a prespecified -0.02 non-inferiority margin.

System	Gate-stable flip rate	Gate-changing update rate	Paired update difference
ObserveTrace-7B	0.129 (0.121–0.138)	0.924 (0.912–0.935)	-0.007 (-0.018 to 0.004)
ObserveTrace-7B without ICC	0.218 (0.206–0.231)	0.931 (0.920–0.942)	reference
Structured long-context	0.236 (0.223–0.249)	—	—
RAG+LLM	0.264 (0.250–0.279)	—	—

Table S12: Site distribution of eligible prefix pairs and gate-stable flip rates. Counts sum to the pooled stable and changing denominators.

Site	Gate-stable pairs	Gate-changing pairs	ObserveTrace-7B flip rate	Without ICC
1	21,480	5,840	0.112	0.194
2	19,820	5,360	0.121	0.207
3	18,760	5,040	0.126	0.214
4	17,940	4,720	0.133	0.223
5	16,420	4,310	0.151	0.247
6	17,360	4,850	0.128	0.219
7	16,860	4,800	0.136	0.229

Table S13: Eligible clinical prefix pairs by type of newly available evidence. Counts sum to the pooled gate-stable and gate-changing denominators.

New-evidence event type	Gate-stable pairs	Gate-changing pairs
Medication order or modification	38,120	12,980
Laboratory or vital result	42,680	11,360
Allergy update	9,840	3,420
Diagnosis or procedure	17,200	3,910
External medication reconciliation	20,800	3,250

Rule-layer sensitivity. Rules improve every learned system, while ObserveTrace-7B remains best with the shared rule version enabled or disabled.

Table S14: Prefix-spacing sensitivity. The ICC advantage grows with elapsed context while the gate-changing action-update rate remains stable.

Prefix spacing	ObserveTrace-7B flip rate	Without ICC	Gate-changing update rate
1 hour	0.118	0.205	0.928
6 hours	0.126	0.215	0.926
24 hours	0.137	0.231	0.921
72 hours	0.148	0.246	0.918

Table S15: Rule-layer sensitivity at fixed learned-model thresholds.

System	Rules	Misses / 1K	High severity / 1K	Alerts / 1K
RAG+LLM	off	7.2 (6.4–8.1)	2.8	34.8
RAG+LLM	on	4.8 (4.2–5.5)	1.9	46.3
LLM-7B+Constrained	off	6.3 (5.6–7.1)	2.4	33.9
LLM-7B+Constrained	on	4.2 (3.7–4.8)	1.6	45.2
ObserveTrace-7B	off	4.7 (4.1–5.4)	1.6	31.2
ObserveTrace-7B	on	3.2 (2.8–3.7)	1.1	42.3

A.2 Endpoint Validity and Operating Characteristics

Extended baseline comparison. The full method leads temporal, retrieval, constrained-generation, and structured long-context systems under the same rule layer.

Table S16: Extended medication-safety comparison at each system's operating point selected under the common validation utility.

Baseline	Misses / 1K	High severity	Alerts / 1K	Change vs. ObserveTrace-7B	P
Rules	9.4 (8.6–10.3)	3.8	38.7	+194%	< 0.001
XGBoost + rules	6.7 (6.0–7.5)	2.7	44.2	+109%	< 0.001
GRU-D + rules	5.4 (4.8–6.1)	2.2	43.8	+69%	< 0.001
Time-aware self-attn. + rules	5.1 (4.5–5.8)	2.1	44.1	+59%	< 0.001
RAG+LLM + rules	4.8 (4.2–5.5)	1.9	46.3	+50%	< 0.001
LLM-7B+Constrained + rules	4.2 (3.7–4.8)	1.6	45.2	+31%	< 0.001
Structured long-context + rules	3.8 (3.3–4.4)	1.4	43.7 (41.0–46.5)	+19%	0.003
OT-LLM + rules	5.1 (4.5–5.8)	2.1	47.8	+59%	< 0.001
ObserveTrace-7B+ rules	3.2 (2.8–3.7)	1.1	42.3	—	—

Primary endpoint validity. The reference protocol and runtime rules are separately versioned and frozen. Each site-specific row stores its source guideline or formulary release identifier and date, category, severity, required slots, window, predicate, and action. The snapshot includes rows active no later than June 30, 2023, the day before validation begins; subsequent updates are excluded from validation and test evaluation. The reference object retains every prespecified row and its blinded-adjudication criterion, whereas the runtime engine contains only rows marked as executable action upgrades from available decision-time fields. A reference-unsafe case can therefore remain unflagged by the Rules baseline when no executable row fires. Table S17 fixes the label-to-action mapping used for supervision and evaluation. Shared canonical inputs and source-row provenance prevent version drift without making the runtime subset identical to the reference label. MAR linkage then confirms that dispensing-queue status closely tracks downstream administration across all three prospective sites.

Blinded re-adjudication. Independent pharmacists re-scored 2,400 stratified cases while blinded to system identity. The final row reports the adjudication-corrected miss-rate ratio of ObserveTrace-7B to RAG+LLM after replacing original labels with adjudicated labels.

Table S17: Frozen reference-label and supervised-action protocol. When several rows apply, precedence is ESCALATE>DEFER>ALERT>APPROVE. Runtime rules can upgrade an action only for rows marked executable; reference labels retain the complete prespecified row set.

Reference determination	Action target	Decision-time criterion	Runtime boundary
Reference-safe	APPROVE	No prespecified unsafe or evidence-acquisition row is satisfied	No upgrade
Evidence acquisition required	DEFER	A required slot is unavailable, stale, or low-confidence and its category row blocks approval	Upgrade only if that row is executable
High-severity unsafe	ESCALATE	A satisfied dose, organ-function, allergy, interaction, contraindication, or formulary row is tagged high severity	Upgrade when the corresponding hard row is executable
Other reference-unsafe	ALERT	A remaining prespecified category or blinded-adjudication criterion requires modification or review	Upgrade only for the executable subset

Table S18: Concordance between the operational endpoint and MAR-confirmed administration.

Site	Orders	MAR-confirmed	Concordance	Corrected FN	Corrected FP
Site A	72.4K	68.1K (94.1%)	0.962 (0.958–0.966)	2.1%	1.4%
Site B	64.2K	59.8K (93.1%)	0.954 (0.949–0.959)	2.8%	1.6%
Site C	47.8K	44.2K (92.5%)	0.948 (0.942–0.954)	3.2%	1.9%
Pooled	184.4K	172.1K (93.3%)	0.956 (0.953–0.959)	2.6%	1.6%

Table S19: Blinded endpoint re-adjudication by site and pooled.

Metric	Site A	Site B	Site C	Pooled	κ	95% CI
Original labels confirmed	94.2%	93.1%	92.4%	93.2%	0.84	(0.81–0.87)
Upgraded (FN→TP)	3.1%	4.2%	4.8%	4.0%	0.78	(0.74–0.82)
Downgraded (FP→TN)	2.7%	2.7%	2.8%	2.7%	0.81	(0.77–0.85)
Adjudication-corrected miss-rate ratio vs. RAG+LLM	0.71	0.64	0.68	0.68	—	(0.62–0.74)

A.3 External-Site, Temporal, and Tail-Risk Tests

Silent prospective evaluation. Model outputs were never surfaced during care. The CPOE columns below therefore describe actions from the active care process, while the ObserveTrace-7B columns score silent model detections against the frozen reference endpoint on the same incoming orders. After the observation period, 1,800 cases were sampled from the union of orders positive under CPOE or ObserveTrace-7B, stratified by site and reference label. Each was rendered in both formats, yielding 1,800 within-case comparisons and 3,600 cards. Two of 18 pharmacists independently reviewed each blinded, randomized pair with its formats separated by a seven-day washout; a third adjudicated disagreements in helpful status, prespecified as 4–5 on a five-point scale.

The pooled helpfulness rate is 61.1% (57.8–64.3) for CPOE and 89.0% (86.6–91.2) for ObserveTrace-7B, with inter-rater agreement κ =0.79 (0.75–0.83).

Difficult cohorts and drift. The method leads both RAG+LLM and constrained decoding in every prespecified tail, including the top 1% complexity slice.

Table S20: Twelve-week prospective silent study across 184.4K orders and three EHR vendors. CPOE columns use observed active-workflow actions; ObserveTrace-7B columns use silent logged actions. Both are scored against the same reference label and 60-minute queue-reaching definition. Review time and helpfulness come from the separate blinded offline card review.

	Site A (Epic)		Site B (Cerner)		Site C (MEDITECH)	
Endpoint	CPOE	ObserveTrace-7B	CPOE	ObserveTrace-7B	CPOE	ObserveTrace-7B
Orders (K)	72.4	72.4	64.2	64.2	47.8	47.8
Queue-reaching misses / 1K	8.7 (7.2–10.4)	2.9 (2.1–3.8)	9.2 (7.6–10.9)	3.1 (2.3–4.0)	8.1 (6.5–9.8)	3.4 (2.5–4.4)
High severity / 1K	3.4 (2.6–4.3)	1.0 (0.6–1.5)	3.6 (2.8–4.5)	1.1 (0.7–1.6)	3.1 (2.3–4.0)	1.2 (0.8–1.7)
False alerts / 1K	47.2 (42.1–52.6)	44.8 (39.9–49.9)	49.8 (44.5–55.2)	46.2 (41.2–51.4)	44.6 (39.2–50.2)	43.1 (37.8–48.6)
Review time (s)	52 (38–74)	28 (19–41)	54 (40–76)	29 (20–42)	48 (35–68)	27 (18–39)
Helpful rating	61%	89%	58%	87%	64%	91%

Table S21: Queue-reaching misses per 1,000 eligible high-risk orders in prespecified cohorts and stress slices. Intervals cluster by patient within slice.

Cohort	Orders	RAG+LLM	LLM-7B+Constrained	ObserveTrace-7B	Change vs. RAG
Overall	847K	4.8 (4.2–5.5)	4.2 (3.7–4.8)	3.2 (2.8–3.7)	−33%
$\geq$10 active medications	124K	6.2 (5.4–7.1)	5.4 (4.7–6.2)	4.8 (4.1–5.6)	−23%
$\geq$3 interacting medications	89K	7.1 (6.2–8.1)	6.2 (5.4–7.1)	5.2 (4.5–6.0)	−27%
eGFR $<$ 30	67K	7.4 (6.4–8.5)	6.1 (5.3–7.0)	5.1 (4.4–5.9)	−31%
ICU admission	98K	5.9 (5.1–6.8)	5.1 (4.4–5.9)	4.2 (3.6–4.9)	−29%
Age $>$ 75	156K	6.8 (5.9–7.8)	5.8 (5.0–6.7)	5.1 (4.4–5.9)	−25%
Worst site (Site 5)	112K	5.6 (4.8–6.5)	4.9 (4.2–5.7)	4.1 (3.5–4.8)	−27%
Worst quarter (Q2 2024)	198K	5.4 (4.7–6.2)	4.6 (4.0–5.3)	3.8 (3.3–4.4)	−30%
Top 1% complexity	8.5K	9.2 (7.8–10.8)	7.8 (6.6–9.2)	6.4 (5.4–7.6)	−30%

Assumptions and empirical stress tests. Table S22 connects each operational assumption to its strongest targeted test and evidence location.

Table S22: Evidence-state assumptions and their targeted empirical tests. Each row identifies the protected object, the trusted transformation, and the comparison that exposes its failure mode.

Protected object	Transformation and trust boundary	Targeted test and comparator	Observed failure and evidence
Unmeasured vs. observed	Triggered slot mapping at model input	10–50% workflow-conditioned drop versus plain sequence	Sparse laboratory workflows reduce retention; Tables 2 and S10
Stale / external evidence	Availability timestamp and provenance in the serializer	Retrained provenance and observability removals	Both removals raise the primary miss rate; Table S9
Prefix recommendation	Frozen rule-and-threshold change gate	Four spacings with and without ICC	Longer gaps increase gate-stable flips; Tables S11–S14
Safety action	Versioned post-decoding rule engine	Rules on/off versus RAG+LLM	Learned ordering persists without rules; Table S15
Site / temporal behavior	Per-system validation-selected thresholds	Seven sites; six validation months for context plus six held-out test months	Site 5 and test-period Q2 2024 define the worst slices; Figure 2 and Table S21
Complex-patient decision	Evidence-linked output under human review	Prespecified polypharmacy and renal subgroups	Older and complex cohorts have higher absolute miss rates; Table S21
Canonical event identity	Unit normalizer, escaping, stable order, and hash tie-break	Retrained unit and canonicalization removals	Unit and tie ambiguity degrade all three endpoints; Table S9

A.4 Cross-Domain Replication and Scaling

Domain mappings. Each benchmark maps domain events into explicit time, provenance, and state fields before model-specific processing. All serialized fields are observed at or before the prediction timestamp. Terminal states used as targets—withdrawal, failure, eviction or rescheduling, conversion, and escalation—are excluded from the input prefix and evaluated only after the declared horizon. The first four domains use domain-native process state; MIMIC-IV uses the clinical observability state.

Table S23: Cross-domain benchmark definitions and ObserveTrace mappings. Process state in the first four domains and clinical observability in MIMIC-IV occupy the same syntactic field but retain their distinct semantics.

Field	OULAD	Alibaba	Google cluster	Retailrocket	MIMIC-IV
Sequences	32.6K students	4.2M tasks	2.4M jobs	8.7M sessions	524K admissions
Events / sequence	847 (312–2104)	24 (8–67)	18 (6–52)	42 (12–128)	156 (48–412)
Time span	9 months	8 days	29 days	6 months	12 months
Event types	click, assess, forum	submit, start, end	submit, schedule, complete	view, cart, purchase	lab, med, vital, note
Provenance	VLE module	machine ID	machine ID	device type	care setting
Observed prefix state	active, inactive	running, pending	queued, running	active, idle	measured, stale, external only
Primary task	dropout	failure	reschedule	conversion	escalation (24 h)

Table S24: Accuracy at 350M parameters across five event-stream domains.

Model	OULAD	Alibaba	Google	Retail	MIMIC-IV
Plain sequence	0.712 (0.698–0.726)	0.684 (0.671–0.697)	0.698 (0.684–0.712)	0.724 (0.711–0.737)	0.756 (0.743–0.769)
ObserveTrace	0.758 (0.744–0.772)	0.731 (0.718–0.744)	0.742 (0.728–0.756)	0.768 (0.755–0.781)	0.792 (0.779–0.805)
ObserveTrace+OAD	0.794 (0.781–0.807)	0.772 (0.759–0.785)	0.781 (0.768–0.794)	0.804 (0.791–0.817)	0.821 (0.808–0.834)
ObserveTrace+OAD+ICC	0.823 (0.810–0.836)	0.801 (0.788–0.814)	0.814 (0.801–0.827)	0.832 (0.819–0.845)	0.847 (0.834–0.860)
Relative improvement	+15.6%	+17.1%	+16.6%	+14.9%	+12.0%

Table S25: Retention at 50% workflow-conditioned drop across domains.

Model	OULAD	Alibaba	Google	Retail	MIMIC-IV	Mean
Plain sequence	0.503	0.430	0.484	0.542	0.578	0.507
ObserveTrace	0.675	0.701	0.688	0.712	0.734	0.702
ObserveTrace+OAD	0.799	0.821	0.812	0.836	0.852	0.824
ObserveTrace+OAD+ICC	0.841	0.858	0.848	0.868	0.881	0.859

Table S26: Prediction flip rate when context is truncated to 40%.

Model	OULAD	Alibaba	Google	Retail	MIMIC-IV	Mean
Plain sequence	0.48	0.52	0.50	0.44	0.41	0.47
ObserveTrace	0.38	0.42	0.40	0.35	0.32	0.37
ObserveTrace+OAD	0.29	0.32	0.30	0.26	0.24	0.28
ObserveTrace+OAD+ICC	0.18	0.19	0.17	0.15	0.13	0.16
Reduction vs. plain	63%	63%	66%	66%	68%	65%

Scale translation. The complete-system advantage persists from 350M through 1.3B to 7B on Alibaba and MIMIC-IV.

A.5 Task-Level Aggregation

Task inventory and units. The 18 tasks are frozen before model comparison and grouped by evaluation object. Tier A uses patient- or encounter-level record tasks: timeline summary, problem-list

Table S27: Accuracy, 50% WCD retention, and 40%-context flip rate across model scales; all structured rows use ObserveTrace+OAD+ICC.

Domain	Model	Accuracy	Retention	Flip rate
Alibaba	Plain 350M	0.684 (0.671–0.697)	0.430	0.52
	ObserveTrace+OAD+ICC 350M	0.801 (0.788–0.814)	0.858	0.19
	Plain 1.3B	0.711 (0.699–0.723)	0.455	0.49
	ObserveTrace+OAD+ICC 1.3B	0.832 (0.820–0.844)	0.872	0.17
	Plain 7B	0.742 (0.730–0.754)	0.472	0.47
	ObserveTrace+OAD+ICC 7B	0.862 (0.851–0.873)	0.885	0.16
MIMIC-IV	Plain 350M	0.756 (0.743–0.769)	0.578	0.41
	ObserveTrace+OAD+ICC 350M	0.847 (0.834–0.860)	0.881	0.13
	Plain 1.3B	0.781 (0.769–0.793)	0.602	0.39
	ObserveTrace+OAD+ICC 1.3B	0.873 (0.861–0.885)	0.892	0.12
	Plain 7B	0.801 (0.789–0.813)	0.621	0.37
	ObserveTrace+OAD+ICC 7B	0.892 (0.881–0.903)	0.903	0.11

completion, medication reconciliation, laboratory trend, eligibility, and adverse-event detection. Tier B uses order-level safety tasks: dose adjustment, contraindication, drug–drug interaction detection, protocol adherence, allergy, and venous-thromboembolism prophylaxis. Tier C uses encounter-level action or trajectory tasks: antibiotic selection, guideline adherence, pathway assignment, escalation, discharge, and care-gap detection. Each task retains its prespecified F1, AUROC, or accuracy and its timestamped eligible cohort; patient- or episode-cluster resampling follows that unit. Figure S1 reports every task separately, so the cross-task mean does not replace task-level effects.

Table S28: Eighteen-task site-macro summary under the strict temporal split. Each native task metric is computed per site, averaged equally within site group, and then averaged across tasks. Figure S1 reports task-level paired differences and intervals.

Model	Tier A	Tier B	Tier C	Overall
Internal sites (Sites 1–4)				
RAG+LLM	0.84	0.79	0.78	0.81
ObserveTrace-7B	0.91	0.88	0.87	0.89
External test-only sites (Sites 5–7)				
RAG+LLM	0.80	0.75	0.74	0.77
ObserveTrace-7B	0.88	0.85	0.84	0.86

All 18 task-level point estimates favor ObserveTrace, and the site panel in Figure S1 shows that the positive mean is not produced by one health system.

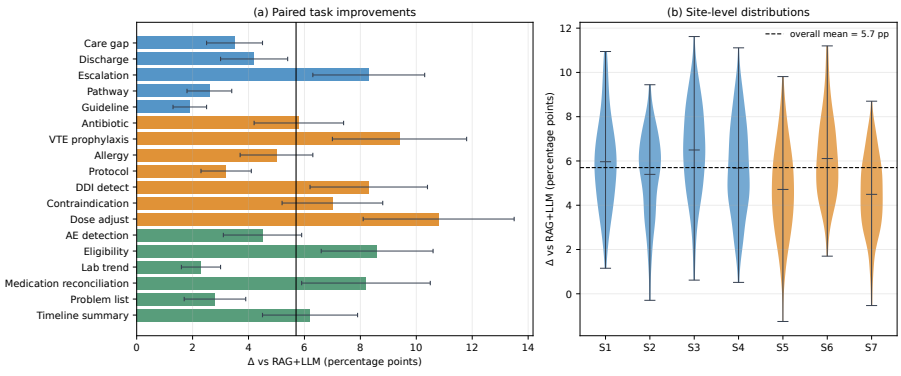

Figure S1: All 18 pooled-across-site task deltas over RAG+LLM. Panel (a) reports each task's recomputed primary metric and 95% cluster-bootstrap interval; panel (b) shows site-level distributions. The 5.7 percentage-point summary is the unweighted mean of the pooled task-level deltas.

Macro-score definition. For task t and site s, $m_{t,s}$ is the prespecified F1, AUROC, or accuracy after orientation so higher is better. Table S28 reports site-macro scores: for site group G, each task score is $m_{t,G} = |G|^{-1} \sum_{s \in G} m_{t,s}$, tier means average their task scores, and the overall statistic is $\bar{M}_G = \frac{1}{18} \sum_{t=1}^{18} m_{t,G}$. Figure S1 instead pools held-out predictions across all seven sites within each task, recomputes its nonlinear metric m_t^{pool}, and reports $\frac{1}{18} \sum_t (m_{t,ObserveTrace}^{\text{pool}} - m_{t,\text{RAG}}^{\text{pool}}) = 5.7$ percentage points. Uncertainty uses patient- or episode-cluster bootstrap within task and site; task-level comparisons apply Benjamini–Hochberg FDR control at $q < 0.05$.

Table S29: Aggregation and uncertainty specification for the 18-task result.

Statistic	Unit	Aggregation	Uncertainty / multiplicity
Site-task score $m_{t,s}$	patient or episode	prespecified F1/AUROC/accuracy within site	2,000 cluster bootstrap replicates
Site-group tier score	site and task	equal-site mean, then unweighted task mean within tier	bootstrap propagated within site and task
Site-group overall $\bar{M}_G$	site and 18 tasks	equal-site mean, then unweighted mean over tasks	bootstrap propagated within site and task
Pooled mean improvement	18 paired tasks	pool predictions across sites, recompute each metric, then average deltas	BH FDR, $q < 0.05$