RETROSPECTIVE SUPERVISION RESHAPES EHR PATIENT STATES BEYOND FUTURE PREDICTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Electronic health record (EHR) foundation models compress longitudinal records into patient states that are reused across downstream tasks, yet these representations are evaluated mainly by how well they predict future outcomes. This leaves unmeasured how much of the patient's history remains accessible and which patients the representation places near one another. We introduce CHRONOS, a patient-state model trained with prospective and retrospective objectives: predicting future clinical events and reconstructing a structured fingerprint of prior history. On MIMIC-IV, frozen CHRONOS states achieve 0.871 mean AUROC across 16 clinical prediction tasks, compared with 0.862 for the strongest existing model, and 0.701 mean AUROC across eight target-masked phenotype-retrieval cohorts, outperforming the best learned baseline by 0.06. In a matched ablation, retrospective supervision leaves future-task performance essentially unchanged while improving retrieval from 0.651 to 0.697 and making both observed and omitted history more recoverable. These effects replicate across independently trained models and persist under projection and record-footprint controls. Our results show that prospective predictiveness, retrospective accessibility, and patient-similarity geometry capture distinct properties of reusable EHR representations.

1 INTRODUCTION

Electronic health records (EHRs) accumulate a longitudinal account of patient care, including diagnoses, procedures, medications, encounters, and laboratory measurements. EHR foundation models compress this history into a patient state that is reused across downstream tasks such as outcome prediction, patient retrieval, and adaptation with limited labels (Xiao et al., 2018; Rajkomar et al., 2018; Steinberg et al., 2021; Wornow et al., 2023a; Guo et al., 2024; Zhang et al., 2026). Because the same state serves many purposes, pretraining determines what information is preserved and how it is exposed to downstream models.

Current EHR pretraining objectives are diverse. They include masked-event reconstruction (Devlin et al., 2019; Li et al., 2020b; Rasmy et al., 2021; Pang et al., 2021; Yang et al., 2023), next-event and next-visit prediction (Choi et al., 2016b; Steinberg et al., 2021; Pang et al., 2024; Renc et al., 2024), time-to-event likelihoods (Steinberg et al., 2023; Jing et al., 2026), generative modeling (Wang & Sun, 2022; Theodorou et al., 2023), and latent trajectory alignment (Sun et al., 2025; Yang et al., 2026). Yet the resulting representations are still compared predominantly by how well they predict future outcomes (Steinberg et al., 2021; Wornow et al., 2023a; Guo et al., 2024). This implicitly treats prospective performance as a proxy for representation quality.

Prospective prediction alone does not determine what a patient state preserves. A future-prediction objective only constrains information that is useful for the modeled futures; other aspects of the observed history may be discarded or encoded in forms that are difficult for downstream readouts to recover. Nor does predictive performance determine representation geometry: two states can achieve similar probe accuracy while placing different patients near one another because of differences in scale, covariance, and organization of the embedding space. We therefore distinguish three properties of a reusable patient representation: *prospective predictiveness*, how well it supports future outcome prediction; *retrospective accessibility*, how much of the patient's prior history can be recovered from the state; and *patient-similarity geometry*, which patients become neighbors under the representation.

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.

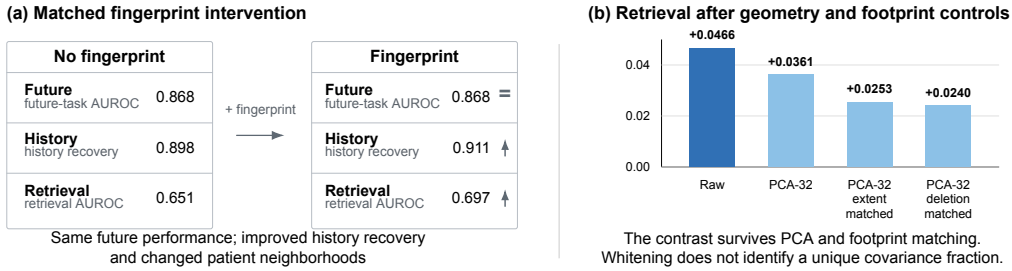


Figure 1: Retrospective supervision changes accessibility and patient-similarity geometry while leaving prospective performance nearly unchanged. (a) Matched models with and without fingerprint supervision. (b) Retrieval contrast survives train-fitted PCA and footprint matching; whitening effects depend on the transformation and do not identify a unique covariance contribution.

Structured EHRs make these properties directly measurable. A frozen state can be probed for whether a clinical code has occurred, how often it occurred, and what its latest value was. Representation geometry can be evaluated through patient retrieval, including a stricter setting in which the codes defining the target phenotype are removed from the input. We call this *target-masked retrieval*. At the same time, EHRs reflect not only underlying clinical state but also the observation process: patients who are sicker or more closely monitored often accumulate more visits, measurements, and coded history (Phelan et al., 2017; Harton et al., 2022). Patient neighborhoods may therefore encode documentation intensity and record extent as well as clinical similarity, making these factors important controls.

We introduce CHRONOS, a patient-state model designed to preserve information in both directions of the record. A block-causal Transformer processes events in the order they became available, with unrestricted attention within groups of simultaneously recorded events and causal attention across groups. Learned slots then pool the encoded sequence into a multi-slot patient state. Training combines two complementary objectives. A prospective marked time-to-event objective predicts which clinical event occurs next, when it occurs, and with what value (Jing et al., 2026). A retrospective objective reconstructs a structured fingerprint of the patient's history, including code presence, frequency, and latest values, as well as information from events preceding the model's input window. Both prediction heads are discarded when the representation is exported.

On MIMIC-IV (Johnson et al., 2023), frozen CHRONOS states achieve a mean AUROC of 0.871 across 16 clinical prediction tasks, compared with 0.862 for ORA (Jing et al., 2026), 0.844 for CEHR-GPT (Pang et al., 2024), and 0.802 for CORE-BEHRT (Odgaard et al., 2024). On eight target-masked phenotype-retrieval cohorts, CHRONOS reaches 0.701 mean AUROC, compared with 0.641 for ORA and 0.678 for a same-window TF-IDF code bag.

More importantly, a matched ablation isolates what changes when retrospective supervision is added while holding the encoder, data, training schedule, and state shape fixed (Figure 1). Future-task performance remains essentially unchanged, while retrieval AUROC increases from 0.651 to 0.697 and historical information becomes more recoverable from the frozen state. The gain extends to events that occurred before the input window, showing that the representation learns to infer omitted history rather than merely re-encode visible tokens. An independently trained model pair reproduces the same dissociation. The retrieval effect persists after matching patients on record footprint and after low-rank projection, although documentation intensity remains strongly encoded in both models. Finally, increasing the number of state slots from one to 16 improves prediction AUROC by 0.013 but retrieval AUROC by 0.035, suggesting that representation capacity can materially alter patient geometry even after predictive performance begins to saturate.

These results show that similar prospective performance can conceal substantial differences in what a patient representation retains and how patients are organized in representation space. Evaluating EHR foundation models therefore requires more than future prediction alone.

2 RELATED WORK

Pretraining objectives for longitudinal EHRs. Longitudinal encoders have predicted upcoming events with recurrent networks, reverse-time attention, hierarchical visit structure, and ontology-

or graph-aware embeddings (Choi et al., 2016b;a; 2018; 2016c; 2017; Shang et al., 2019; Choi et al., 2020; Hur et al., 2024). Transformer pretraining uses masked-event reconstruction (Li et al., 2020b; Rasmy et al., 2021; Pang et al., 2021; Yang et al., 2023; Odgaard et al., 2024) or causal next-event and next-visit prediction (Steinberg et al., 2021; Pang et al., 2024; Renc et al., 2024), with generative (Wang & Sun, 2022; Theodorou et al., 2023) and long-context variants (Li et al., 2023; Fallahpour et al., 2025). Clinically aligned objectives forecast irregular measurements or utilization (Tipirneni & Reddy, 2022; Kumar et al., 2021), scale censored time-to-event pretraining (Steinberg et al., 2023), jointly model event time and value (Jing et al., 2026), or align latent trajectories (Sun et al., 2025; Yang et al., 2026). These objectives motivate strong prospective evaluation, while masked-event objectives also test local reconstruction. CHRONOS adopts ORA's marked likelihood as its prospective term and adds an explicit full-prefix summary objective. Our contribution is the controlled evidence that prospective performance, retrospective accessibility, and patient-similarity geometry can change independently in the exported patient state.

Patient summaries, similarity, and multi-vector states. Deep Patient and later autoencoders reconstruct aggregated clinical features to learn patient vectors (Miotto et al., 2016; Landi et al., 2020), and metric learning has formalized patient similarity under weak supervision (Wang & Sun, 2015). These studies establish history reconstruction and patient similarity as useful learning signals. We study whether explicit retrospective supervision changes the information and geometry of a reusable state when prospective fit is held nearly fixed within a temporal event model. Learned query sets summarize variable-sized inputs in Set Transformer and Perceiver (Lee et al., 2019; Jaegle et al., 2021), and Slot Attention shows that latent slots can specialize (Locatello et al., 2020). CHRONOS uses slot-specific pooling as distributed capacity without assuming object-like or organ-specific disentanglement.

Evaluation and the observation process. Benchmarks evaluate reusable encoders across outcomes, few-shot tasks, temporal shift, and health systems (Rajkomar et al., 2018; Steinberg et al., 2021; Wornow et al., 2023a; Guo et al., 2023; 2024), and recent critiques question what such evaluations establish (Wornow et al., 2023b). Routine EHR data also encode when and how often care was observed: informative presence induces associations through visit frequency and record completeness that adjusting for encounter intensity does not fully remove (Phelan et al., 2017; Harton et al., 2022). Cosine retrieval further depends on embedding anisotropy and covariance, which flows, whitening, and contrastive training can change without altering the encoder (Li et al., 2020a; Huang et al., 2021; Gao et al., 2021). We therefore pair prediction with target-masked retrieval, utilization matching, train-fitted geometry transforms, and observation-process probes.

3 METHODS

3.1 PROSPECTIVE AND RETROSPECTIVE PRESERVATION

Let a patient record contain events $e_i = (c_i, m_i, t_i, a_i, x_i)$, where c_i is a clinical code, m_i is its modality, t_i is its clinical time, a_i is the time at which the event becomes available, and x_i is an optional numeric or categorical value. We order the input by availability time. Events with the same availability time form one block. This distinction prevents a late-entered measurement from becoming visible before it was recorded.

Let u_a denote an anchor's availability cutoff and H_a the complete observed prefix. A sampled window ω contains temporal events W_ω and block-end anchors $\mathcal{A}_\omega$. The actual context at anchor a contains only events in that window available by u_a:

$$\begin{aligned} H_a &= \big(d_0, (e_i : a_i \le u_a)\big), \quad V_{\omega a} = \big(d_0, (e_i \in W_\omega : a_i \le u_a)\big), \\ Z_{\omega a} &= E_\theta(V_{\omega a}) \in \mathbb{R}^{K \times d_s}. \end{aligned} \tag{1}$$

Here d_0 denotes static context and $s_{\omega a} = [z_{\omega a1}^\top, \ldots, z_{\omega aK}^\top]^\top$ is the exported state. The retrospective target is $\Phi(H_a)$ and excludes the window-only summary $\Phi(V_{\omega a})$. For targets $Y_+ = Y^+$ (future events) and $Y_- = \Phi(H)$ (history summary), readout family $\mathcal{G}_j$, and loss ℓ_j, define the operational decoding risk

$$\mathcal{R}_j(E) = \inf_{g \in \mathcal{G}_j} \mathbb{E}_{P_j}[\ell_j(g(E(V)), Y_j)], \qquad j \in \{+, -\}. \tag{2}$$

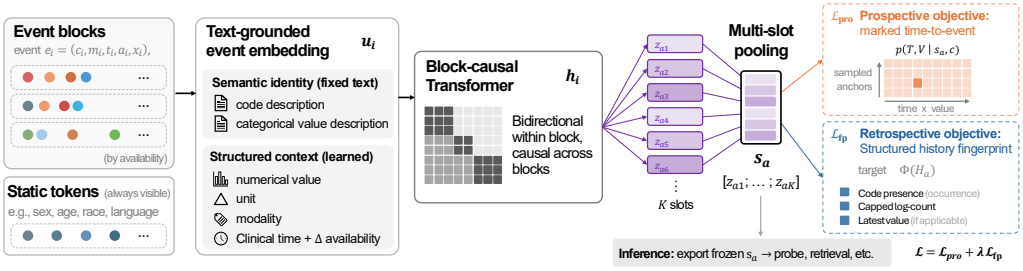


Figure 2: **CHRONOS overview.** Event blocks ordered by availability are encoded by a block-causal Transformer, pooled into a multi-slot patient state, and trained with prospective marked time-to-event prediction and a retrospective structured history fingerprint. At inference, all pretraining heads are removed and the frozen patient state is exported for linear-probing, retrieval, etc.

We use the representation contract as operational vocabulary for these two decoding axes. It specifies future and history targets under an exported-state budget Kd_s; claims of zero loss or universal statistical sufficiency fall outside its scope. Patient-similarity geometry is measured separately from these decoding risks. P_+ samples prospective anchors, whereas P_- samples final-window anchors. Appendix E.1 describes the sampler. Appendices F.1, F.3, and F.2 define the scope of this interpretation.

Figure 2 summarizes how CHRONOS instantiates these two preservation axes: availability-ordered event blocks are encoded by a block-causal Transformer, pooled into a multi-slot patient state, and trained with prospective marked time-to-event and retrospective history-fingerprint objectives.

3.2 Event Representation and Temporal Encoder

We combine projected frozen text-description embeddings with value, modality, and time features. A block-causal Transformer allows attention within an availability block and to earlier blocks; later blocks remain masked. Static tokens remain in the initial context. The full feature construction, mask, and temporal conventions are given in Appendices A.4 and E.1.

3.3 Distributed Patient State

Each slot uses an anchor-conditioned query and private key/value projections to pool contextualized tokens in $V_{\omega a}$. The ordered concatenation of these slot vectors is the exported state. We use slot-specific pooling without assumptions of statistical independence or fixed clinical assignments. The model has no cross-slot self-attention or diversity loss. The final model has $K = 16$ slots of width $d_s = 128$; full pooling equations and parameter sharing are in Appendix E.2.

3.4 Prospective Targets: Marked Time-to-Event Prediction

For each supervised code, the prospective target is its first eligible occurrence with clinical time after the anchor availability cutoff u_a. The target pool is scanned independently for each code and selects its first eligible occurrence. This target differs from prediction of the next event overall. The implementation uses $r_a = u_a$ as the clinical prediction origin. Because visibility follows availability time, some events with future clinical time may already be documented at the anchor. Let $\mathcal{V}_c$ be the valid mark classes, P the number of time intervals, and ℓ_{acpv} the head's joint time and mark logits. An unmarked code has one dummy class. Using ORA's joint time and value parameterization (Jing et al., 2026), the grid is

$$D_{ac} = 1 + \sum_{p,v} e^{\ell_{acpv}}, \quad \pi_{ac}(p, v) = e^{\ell_{acpv}}/D_{ac}, \quad \pi_{ac}(\varnothing) = 1/D_{ac},$$
$$\ell^{+}_{ac} = -\log \sum_{y \in \mathcal{A}_{ac}} \pi_{ac}(y). \tag{3}$$

The sum ranges over $p = 1, \dots, P$ and $v \in \mathcal{V}_c$. Here $\varnothing$ denotes no occurrence in the modeled horizon. For an observed first event, $\mathcal{A}_{ac}$ is its joint time and mark cell; after q fully observed event-

free intervals, it contains $\varnothing$ and all cells with $p > q$. Pairs with no event and no completed follow-up interval are uninformative and masked. Target-extension truncation stops follow-up before the earliest omitted event. Appendices E.1 and E.3 specify the target clock, interval boundaries, missing-mark treatment, censoring rule, and exact loss reduction.

3.5 Retrospective targets: structured history fingerprint

At the final anchor, the fingerprint summarizes the complete observed prefix:

$$\begin{aligned} n_{ac} &= \sum_{e_i \in H_a} \mathbf{1}[c_i = c], \quad y^p_{ac} = \mathbf{1}[n_{ac} > 0], \\ y^n_{ac} &= \log(1 + \min(n_{ac}, C)), \quad y^v_{ac} = \nu_{ac}. \end{aligned} \tag{4}$$

$C = 32$ caps counts and ν_{ac} denotes the latest valid numeric-value target by clinical time; $\Phi(H_a)$ collects these targets over supervised codes. Presence uses slot-wise logits, max pooling, and capped-IDF-weighted binary cross-entropy; positive counts use Huber loss with diagnosis and DRG codes excluded; valid latest values use a code-query categorical readout. Static input-only codes are excluded from all targets. Full-prefix supervision combines reconstruction of visible summaries with inference about omitted history. Omitted events remain unavailable at inference. Appendix E.4 specifies the valid-first latest-value selector, code-query readout, IDF weights, Huber convention, and exact reductions.

3.6 Full objective

The full pretraining objective is given by

$$\mathcal{L}_{\text{CHRONOS}} = \mathcal{L}_{\text{pro}} + \lambda_{\text{fp}} \left(\mathcal{L}_{\text{pres}} + \lambda_{\text{count}} \mathcal{L}_{\text{count}} + \lambda_{\text{value}} \mathcal{L}_{\text{latest}} \right). \tag{5}$$

The prospective loss is applied at sampled anchors and the fingerprint loss at the final anchor of each window. This fixed-weight objective is a surrogate for complementary target preservation. At inference, the pretraining heads are removed and the concatenated slot state is exported (see Appendix E.5 for coefficients and aggregation; Appendix F.3 for optimization).

4 Experimental Setup

4.1 Data

We use MIMIC-IV v3.1 (Johnson et al., 2023). The representation includes 22 temporal modalities spanning encounters, diagnoses, procedures, medications, laboratory tests, microbiology, bedside observations, inputs, outputs, and clinical orders, together with input-only sex, race, and language tokens. ICD-9 diagnoses are mapped to three-character ICD-10 categories. The supervision vocabulary contains 8,591 event types. Numeric marks use code-specific bins; categorical result text is input-only in the reported models. Vocabulary thresholds, mappings, value bins, and time bins are fit on the training split.

We use a fixed patient-level split of 111,638 patients for pretraining, 37,213 for validation, and 37,213 for testing. The training split contains 315.8 million events. Events entered after their clinical time remain unavailable until entry, and all prediction snapshots exclude events unavailable at the cutoff. Architecture and hyperparameters were selected on validation data only; the test split was reserved for final frozen-state evaluation.

Reported implementation settings are in Appendix A; Appendix E gives the exact target, readout, censoring, and reduction conventions used for training.

4.2 Baselines

We compare with a paper-aligned ORA model using its last event as the patient vector, causal CEHR-GPT, and bidirectional CORE-BEHRT. All use the same split, cutoff rule, modalities, and sequence limit; we otherwise retain each method's native objective, architecture, serialization, and optimizer.

These native-recipe comparisons contextualize frozen-state performance; differing objectives, architectures, parameter counts, serialization, and value handling preclude isolating the representation contract. Every evaluation freezes the encoder. Appendix A records value-handling and parameter-count differences.

4.3 EVALUATION

Frozen clinical prediction. At each task cutoff, we encode the available record, freeze the state, and fit an L2-regularized logistic regression using only training and validation data for preprocessing and model selection. We report test AUROC on 16 tasks and patient-cluster bootstrap uncertainty.

Nonlearned code-bag control. We build a sparse code control from the same target-masked event window: the cutoff, defining-code deletion before window selection, backfill, and complete-block-aligned suffix match the learned models. It retains counts over the 8,591-code vocabulary and omits numeric values, timing, units, categorical values, and text-derived embeddings. Sublinear TF-IDF uses $1+\log(\text{count})$ to damp repeated documentation and $\log(N/\mathrm{df}_c)$ to downweight common codes. We fit df_c without labels on 56,032 structurally eligible *training* patients. The held-out retrieval pool contributes no representation estimates. Vectors are L_2-normalized and ranked by cosine similarity.

Target-masked patient retrieval. Each positive test patient queries the eligible cohort by cosine similarity, with no fitted model. Before encoding, we delete every diagnosis and medication code that defines that phenotype and backfill the window while holding its label fixed. Auditing this rule against the full text catalog finds drug-name proxies outside the diagnosis and medication modalities, including vancomycin levels reported as laboratory tests, albuterol and vancomycin chart observations, and a vancomycin-resistance screen; widening the deletion set to all 92 such codes across every modality changes mean retrieval AUROC by at most 0.0005 for any arm. We report per-phenotype and mean AUROC with positive-query bootstrap uncertainty.

Retrospective recovery and geometry controls. To measure retrospective preservation directly, we fit identical ridge readouts from frozen states to code presence and capped counts; code eligibility and regularization use training data only, and the pretrained fingerprint head is never reused. We evaluate the full prefix, the visible input window, and codes that occur only in the omitted prefix. Separate ridge readouts recover latest numeric values under a train-only winsorization policy. Because cosine neighborhoods depend on representation covariance, we fit PCA projections, truncated PCA whitening, and ridge-regularized whitening on a fixed label-free sample of training patients, freeze each transform, and apply it to held-out retrieval states.

Observation-process controls. For the matched 10-slot fingerprint pair, we repeat retrieval within strata and against negatives matched jointly on event count, active days, admissions, observation density, age, and sex, all computed after target masking. We also probe nine documentation-intensity and record-size quantities from frozen states. Appendix B gives the complete protocols and supplementary documentation-inflation interventions.

Slot interventions. We remove evidence groups and replace individual slots by their cohort means, measuring effects on the state, prospective loss, probes, and retrieval. Within-row slot permutations test whether effects align consistently with specific slots; leave-one-slot-out refitting tests redundancy. Appendix C gives the full protocol.

5 RESULTS

5.1 PROSPECTIVE PERFORMANCE DOES NOT DETERMINE THE PATIENT STATE

Table 1 isolates retrospective supervision under one encoder, prospective head, split, schedule, and state shape. In the primary seed-0 pair, the resolved configurations differ only in whether the fingerprint objective is enabled, and both arms share the same source tree; the pretraining-only head is discarded, so both models export the same 25.4M-parameter inference architecture. The independently trained seed-1 pair uses matched configurations from different source-tree versions and

Table 1: Exact matched fingerprint intervention (seed 0). Newly fitted readouts measure full-prefix and omitted-prefix code presence and latest-value recovery. Prospective loss uses the pretraining objective. Retrieval uses raw cosine.

Fingerprint	Future probe	Prospective loss ↓	Full history	Omitted history	Latest value R^2	Raw retrieval
No	0.868	0.07535	0.898	0.905	0.158	0.651
Yes	**0.868**	**0.07517**	**0.911**	**0.916**	**0.202**	**0.697**

Table 2: Retrieval gain from fingerprint supervision after train-fitted transformations and record-footprint controls. PCA-32 retains the 32 leading principal components without whitening. Brackets are 95% patient-cluster bootstrap intervals over query patients (1,000 replicates); every interval excludes zero and every row improves on all eight phenotypes. Seed 0 is the exact matched intervention; seed 1 is an independent replication.

Evaluation	Seed 0	Seed 1
Raw, unconditional	0.0466 [0.0446, 0.0486]	0.0510 [0.0489, 0.0530]
PCA-32, unconditional	0.0361 [0.0343, 0.0379]	0.0371 [0.0352, 0.0388]
PCA-32, utilization matched	0.0224 [0.0202, 0.0244]	0.0260 [0.0238, 0.0283]
PCA-32, record extent matched	0.0253 [0.0229, 0.0276]	0.0272 [0.0247, 0.0296]
PCA-32, deletion load matched	0.0240 [0.0215, 0.0265]	0.0258 [0.0228, 0.0287]

serves as replication. Future probes and the prospective objective change little, while direct historical recovery and target-masked retrieval improve.

For retrospective recovery, we fit identical ridge readouts to both frozen states; the pretrained fingerprint head is discarded. Over 2,632 codes selected by training prevalence alone, fingerprint supervision improves full-prefix presence macro AUROC by 0.0123 and 0.0126 in seeds 0 and 1. The corresponding gains are 0.0110 and 0.0112 for micro AUROC, 0.0113 and 0.0103 for Recall@50, and 0.0565 and 0.0500 for count R^2. The effect is also present for codes that occur before the 1,024-event input window and are absent from it: omitted-prefix macro AUROC improves by 0.0116 in both runs, while count R^2 improves by 0.0584 and 0.0521. Thus, the effect cannot be explained only by easier recovery of visible token identities. It reflects statistical inference about omitted history; events absent from the input remain unavailable at inference. Latest-value recovery also improves, with macro R^2 changing from 0.158 to 0.202 and from 0.163 to 0.203. All effects replicate closely in magnitude. The prospective total differs by only 0.00018 and 0.00023, although the fingerprint model has slightly worse event NLL and better value NLL in both runs. The dissociation therefore holds on the prospective pretraining objective as well as downstream probes.

5.2 The retrieval contrast survives footprint control and low-rank projection

Table 2 summarizes the controlled geometry contrasts. In the primary pair, raw-cosine retrieval improves by 0.0466 [0.0446, 0.0486]. The states are strongly low rank: their Shannon effective ranks are 38.7 without the fingerprint and 61.7 with it in 1,280 dimensions, and full-rank whitening amplifies some directions by approximately 10^8. We therefore treat full-rank ZCA as a sensitivity analysis. The fingerprint contrast remains positive over four consecutive train-fitted PCA projections, from 16 to 128 components, in all four matching cohorts and both runs. On pairs matched for target-deletion load, record extent, and truncation, PCA-32 retrieval improves by 0.0240 [0.0215, 0.0265] in seed 0 and 0.0258 [0.0228, 0.0287] in seed 1; all eight phenotypes improve. Whitening attenuates the contrast, but the effect is non-monotonic across truncated and ridge-regularized whitening paths, so the data do not define a unique fraction attributable to covariance. A label-free spectrum transplant reproduces 24% of the raw gap, indicating that variance allocation contributes but does not fully reproduce the effect.

5.3 Observation-process information limits the interpretation of retrieval

The raw fingerprint contrast remains 0.0364 [0.0348, 0.0382] within utilization strata and 0.0285 [0.0266, 0.0305] against utilization-matched negatives. Observation-process variables nevertheless

Table 3: Frozen clinical prediction from exported patient states. AUROC on 16 held-out MIMIC-IV tasks; parentheses give patient-bootstrap standard deviations. The final row is the unweighted task average. Models use their native pretraining recipes, so the comparison provides cross-model context rather than controlled attribution.

Task	TF-IDF bag	CEHR-GPT	CORE-BEHRT	ORA	CHRONOS
General outcomes					
30-day readmission	.667(.004)	.694(.003)	.658(.003)	.704(.003)	**.705(.003)**
ICU LOS $>$ 7 days	.830(.006)	.834(.005)	.779(.007)	.853(.005)	**.854(.005)**
ICU mortality	.870(.006)	.888(.005)	.841(.006)	.912(.004)	**.918(.004)**
In-hospital mortality	.890(.004)	.924(.003)	.880(.004)	.936(.003)	**.940(.002)**
Condition-specific mortality					
Acute kidney injury	.825(.007)	.845(.006)	.805(.007)	.864(.006)	**.877(.005)**
Acute myocardial infarction	.830(.013)	.851(.013)	.820(.014)	.864(.012)	**.875(.011)**
Acute pancreatitis	.885(.024)	.927(.016)	.910(.018)	.931(.019)	.931(.021)
Atrial fibrillation	.835(.008)	.873(.007)	.817(.009)	.888(.006)	**.892(.006)**
ARDS	.717(.015)	.731(.014)	.716(.014)	.768(.013)	**.785(.012)**
Cirrhosis	.822(.021)	.892(.014)	.820(.020)	.909(.013)	**.919(.012)**
Heart failure	.813(.009)	.852(.007)	.807(.009)	.865(.007)	**.876(.007)**
Liver failure	.772(.015)	.799(.014)	.773(.015)	.813(.014)	**.842(.012)**
New GI bleeding	.826(.017)	.865(.012)	.802(.019)	.881(.013)	**.895(.012)**
Septic shock	.743(.011)	.767(.011)	.709(.012)	.793(.011)	**.804(.010)**
Stroke	.852(.012)	.878(.010)	.840(.012)	.905(.009)	**.905(.008)**
Subarachnoid hemorrhage	.887(.027)	.889(.027)	.852(.030)	.904(.027)	**.910(.022)**
Average	.816	.844	.802	.862	**.871**

remain highly decodable: mean R^2 decreases from 0.822 without fingerprint supervision to 0.777 with it. The allocation changes across targets: all six recent-intensity or recency targets decrease, whereas total record size, window length, and truncation increase. Retrieval geometry therefore reflects the joint organization of recorded clinical state and care processes (see Appendix B for documentation-inflation interventions).

5.4 CROSS-MODEL CONTEXT

Having established the controlled dissociation, we ask whether the resulting representation remains competitive under conventional frozen-state evaluation. Table 3 reports all 16 clinical prediction tasks for the final 16-slot CHRONOS model. CHRONOS reaches mean AUROC 0.871, compared with 0.862 for ORA, 0.844 for CEHR-GPT, 0.802 for CORE-BEHRT, and 0.816 for the same-window TF-IDF control. These native-recipe comparisons provide context rather than controlled attribution.

Table 4 reports the complementary target-masked retrieval benchmark across eight phenotype–medication pairs. CHRONOS reaches mean AUROC 0.701, compared with 0.678 for same-window TF-IDF, 0.641 for ORA, 0.611 for CORE-BEHRT, and 0.606 for CEHR-GPT. This raw cross-model advantage depends on the evaluation footprint: after stricter matching on utilization and the deletion/backfill footprint, CHRONOS and TF-IDF are statistically indistinguishable, whereas the within-architecture fingerprint contrast remains positive. We therefore treat Table 4 as an operational retrieval benchmark rather than controlled evidence for the fingerprint intervention.

5.5 STATE CAPACITY SCALES DIFFERENTLY FOR PREDICTION AND RETRIEVAL

Increasing capacity from one to 16 slots raises the four-task probe mean by 0.013 and target-masked retrieval by 0.035, with prediction saturating earlier than retrieval. We selected $K = 16$ on validation data. Because slot count also changes state dimension and parameter count, this sweep does not isolate factorization; a matched-dimensional wide one-slot control nearly matches prediction but trails 10 slots on retrieval. Full results are in Appendix C, Table 9.

Table 4: Target-masked retrieval from exported patient states. AUROC on eight held-out MIMIC-IV phenotype–medication tasks; parentheses give query-bootstrap standard deviations, and the final row is the unweighted phenotype average. Defining diagnosis and medication codes are removed before encoding. Native-recipe comparisons provide operational context; the CHRONOS–TF-IDF difference does not persist under stricter deletion-footprint matching.

Phenotype	TF-IDF bag	CORE-BEHRT	CEHR-GPT	ORA	CHRONOS
Atrial fibrillation → apixaban	.754(.003)	.615(.004)	.694(.003)	.758(.005)	**.795(.003)**
Asthma → albuterol	.609(.002)	.557(.003)	.602(.002)	.549(.002)	**.611(.002)**
Cirrhosis → beta-blocker	.833(.009)	.729(.010)	.710(.010)	.777(.012)	**.843(.009)**
Chronic kidney disease → lisinopril	.698(.003)	.605(.005)	.621(.003)	.656(.004)	**.715(.003)**
Epilepsy → levetiracetam	.582(.004)	.545(.003)	.554(.003)	.566(.004)	**.616(.004)**
Gout → allopurinol	.654(.003)	.636(.006)	.625(.004)	.616(.004)	**.692(.004)**
Iron-deficiency anemia → oral iron	.624(.004)	.552(.004)	.567(.005)	.545(.004)	**.637(.005)**
Sepsis → vancomycin	.668(.008)	.648(.006)	.477(.008)	.657(.010)	**.701(.010)**
Average	.678	.611	.606	.641	**.701**

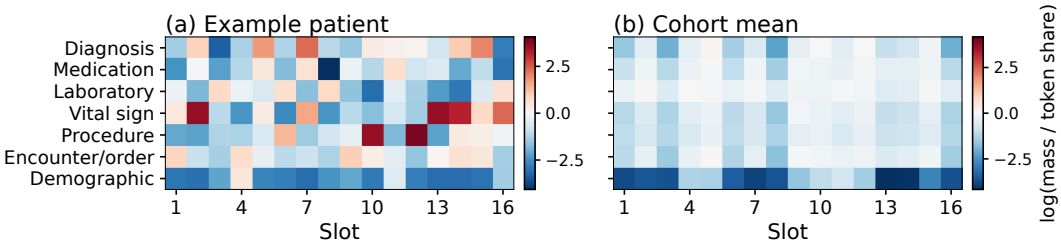


Figure 3: Opportunity-normalized slot routing, $\log[(\text{attention mass} + \epsilon)/(\text{token share} + \epsilon)]$ with $\epsilon = 10^{-3}$: zero denotes attention proportional to token opportunity. (a) One deterministically selected validation patient. (b) Mean over the 3,200-patient cohort. Cells for modalities a patient lacks are excluded (17.6%); panels use separate scales; slots use numeric labels.

5.6 SLOTS LEARN ROUTING PREFERENCES WITH DOWNSTREAM REDUNDANCY

Opportunity-normalized attention shows reproducible slot-by-modality routing asymmetry (Figure 3), although an untrained control indicates that reproducibility alone is partly architectural. Intervention tests do not support fixed semantic slot identities, and downstream readouts remain distributed: removing one slot has little effect after refitting. We therefore interpret the slots as soft routing preferences with redundant downstream capacity; full analyses are in Appendix C.

6 CONCLUSION

Prospective performance does not uniquely characterize an EHR patient state. An exact matched intervention and an independent replication show that structured retrospective supervision leaves future-task probes and prospective pretraining loss nearly unchanged while improving recovery of visible history, omitted-prefix history, and latest values. It also changes target-masked retrieval under record-footprint controls and low-dimensional projection. Whitening attenuates this contrast in a transformation-dependent way, and observation-process information remains strongly decodable, which limits a purely clinical interpretation of the resulting neighborhoods. CHRONOS instantiates prospective and retrospective preservation in a competitive multi-slot model, and slot analyses support distributed, redundant capacity without fixed semantic compartments. Together, these results show why reusable EHR representations require separate evaluation of prospective predictiveness, retrospective accessibility, and patient-similarity geometry.

AI USE STATEMENT

Generative AI tools were used for manuscript editing, literature discovery, and code assistance. All AI-assisted content was reviewed and verified by the authors. The authors take responsibility for the final content, claims, analyses, and artifacts in this work.

ETHICS STATEMENT

This study is a retrospective analysis of MIMIC-IV v3.1, a deidentified dataset available under a PhysioNet data-use agreement; no new patient data were collected.

REPRODUCIBILITY STATEMENT

All experiments use MIMIC-IV v3.1 (Johnson et al., 2023), available to credentialed researchers via PhysioNet. Data processing and the patient-level split are described in Section 4.1 and Appendix A; the exact objective, windowing, pooling, and loss reductions in Appendix E; and all evaluation protocols in Section 4.3 and Appendices A.2, B, and C. Appendix C.3 documents the matched fingerprint intervention and its replication. Code and run configurations will be released upon acceptance.

REFERENCES

Edward Choi, Mohammad Taha Bahadori, Joshua A. Kulas, Andy Schuetz, Walter F. Stewart, and Jimeng Sun. RETAIN: An interpretable predictive model for healthcare using reverse time attention mechanism. In *Advances in Neural Information Processing Systems*, volume 29, pp. 3504–3512, 2016a.

Edward Choi, Mohammad Taha Bahadori, Andy Schuetz, Walter F. Stewart, and Jimeng Sun. Doctor AI: Predicting clinical events via recurrent neural networks. In *Proceedings of Machine Learning for Healthcare*, volume 56 of *Proceedings of Machine Learning Research*, pp. 301–318. PMLR, 2016b.

Edward Choi, Mohammad Taha Bahadori, Elizabeth Searles, Catherine Coffey, Michael Thompson, James Bost, Javier Tejedor-Sojo, and Jimeng Sun. Multi-layer representation learning for medical concepts. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1495–1504, 2016c. doi: 10.1145/2939672.2939823.

Edward Choi, Mohammad Taha Bahadori, Le Song, Walter F. Stewart, and Jimeng Sun. GRAM: Graph-based attention model for healthcare representation learning. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 787–795, 2017. doi: 10.1145/3097983.3098126.

Edward Choi, Cao Xiao, Walter F. Stewart, and Jimeng Sun. MiME: Multilevel medical embedding of electronic health records for predictive healthcare. In *Advances in Neural Information Processing Systems*, volume 31, pp. 4547–4557, 2018.

Edward Choi, Zhen Xu, Yujia Li, Michael W. Dusenberry, Gerardo Flores, Emily Xue, and Andrew M. Dai. Learning the graphical structure of electronic health records with graph convolutional transformer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 606–613, 2020. doi: 10.1609/aaai.v34i01.5400.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 4171–4186, 2019.

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.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. SimCSE: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 6894–6910. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.emnlp-main.552.

Lin Lawrence Guo, Ethan Steinberg, Scott Lanyon Fleming, Jose Posada, Joshua Lemmon, Stephen R. Pfohl, Nigam Shah, Jason Fries, and Lillian Sung. EHR foundation models improve robustness in the presence of temporal distribution shift. *Scientific Reports*, 13(1):3767, 2023. doi: 10.1038/s41598-023-30820-8.

Lin Lawrence Guo, Jason Fries, Ethan Steinberg, Scott Lanyon Fleming, Keith Morse, Catherine Aftandilian, Jose Posada, Nigam Shah, and Lillian Sung. A multi-center study on the adaptability of a shared foundation model for electronic health records. *npj Digital Medicine*, 7:171, 2024. doi: 10.1038/s41746-024-01166-w.

Joanna Harton, Nandita Mitra, and Rebecca A. Hubbard. Informative presence bias in analyses of electronic health records-derived data: A cautionary note. *Journal of the American Medical Informatics Association*, 29(7):1191–1199, 2022. doi: 10.1093/jamia/ocac050.

Junjie Huang, Duyu Tang, Wanjun Zhong, Shuai Lu, Linjun Shou, Ming Gong, Daxin Jiang, and Nan Duan. WhiteningBERT: An easy unsupervised sentence embedding approach. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pp. 238–244. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.findings-emnlp.23.

Kyunghoon Hur, Jungwoo Oh, Junu Kim, Jiyoun Kim, Min Jae Lee, Eunbyeol Cho, Seong-Eun Moon, Young-Hak Kim, Louis Atallah, and Edward Choi. GenHPF: General healthcare predictive framework for multi-task multi-source learning. *IEEE Journal of Biomedical and Health Informatics*, 28(1):502–513, 2024. doi: 10.1109/JBHI.2023.3327951.

Andrew Jaegle, Felix Gimeno, Andrew Brock, Andrew Zisserman, Oriol Vinyals, and Joao Carreira. Perceiver: General perception with iterative attention. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pp. 4651–4664. PMLR, 2021.

Zilin Jing, Vincent Jeanselme, Yuta Kobayashi, Simon A. Lee, Chao Pang, Aparajita Kashyap, Yanwei Li, Xinzhuo Jiang, and Shalmali Joshi. One loss to rule them all: Marked time-to-event for structured EHR foundation models. *arXiv preprint arXiv:2602.00541*, 2026.

Alistair E. W. Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J. Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, Li-wei H. Lehman, Leo A. Celi, and Roger G. Mark. MIMIC-IV, a freely accessible electronic health record dataset. *Scientific Data*, 10(1):1, 2023. doi: 10.1038/s41597-022-01899-x.

Yogesh Kumar, Alexander Ilin, Henri Salo, Sangita Kulathinal, Maarit K. Leinonen, and Pekka Marttinen. SANSformers: Self-supervised forecasting in electronic health records with attention-free models. *arXiv preprint arXiv:2108.13672*, 2021.

Isotta Landi, Benjamin S. Glicksberg, Hao-Chih Lee, Sarah Cherng, Giulia Landi, Matteo Danieletto, Joel T. Dudley, Cesare Furlanello, and Riccardo Miotto. Deep representation learning of electronic health records to unlock patient stratification at scale. *npj Digital Medicine*, 3:96, 2020. doi: 10.1038/s41746-020-0301-z.

Juho Lee, Yoonho Lee, Jungtaek Kim, Adam R. Kosiorek, Seungjin Choi, and Yee Whye Teh. Set transformer: A framework for attention-based permutation-invariant neural networks. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97, pp. 3744–3753. PMLR, 2019.

Bohan Li, Hao Zhou, Junxian He, Mingxuan Wang, Yiming Yang, and Lei Li. On the sentence embeddings from pre-trained language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pp. 9119–9130. Association for Computational Linguistics, 2020a. doi: 10.18653/v1/2020.emnlp-main.733.

Yikuan Li, Shishir Rao, Jose 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, 2020b. doi: 10.1038/s41598-020-62922-y.

Yikuan Li, Mohammad Mamouei, Gholamreza Salimi-Khorshidi, Shishir Rao, Abdelaali Hassaine, Dexter Canoy, Thomas Lukasiewicz, and Kazem Rahimi. Hi-BEHRT: Hierarchical transformer-based model for accurate prediction of clinical events using multimodal longitudinal electronic health records. *IEEE Journal of Biomedical and Health Informatics*, 27(2):1106–1117, 2023. doi: 10.1109/JBHI.2022.3224727.

Francesco Locatello, Dirk Weissenborn, Thomas Unterthiner, Aravindh Mahendran, Georg Heigold, Jakob Uszkoreit, Alexey Dosovitskiy, and Thomas Kipf. Object-centric learning with slot attention. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Riccardo Miotto, Li Li, Brian A. Kidd, and Joel T. Dudley. Deep patient: An unsupervised representation to predict the future of patients from the electronic health records. *Scientific Reports*, 6: 26094, 2016. doi: 10.1038/srep26094.

Mikkel Fruelund Odgaard, Kiril Vadimovic Klein, Martin Sillesen, Sanne Møller Thysen, Espen Jimenez-Solem, and Mads Nielsen. CORE-BEHRT: A carefully optimized and rigorously evaluated BEHRT. In *Proceedings of the 9th Machine Learning for Healthcare Conference*, volume 252 of *Proceedings of Machine Learning Research*. PMLR, 2024.

Chao Pang, Xinzhuo Jiang, Krishna S. Kalluri, Matthew Spotnitz, RuiJun Chen, Adler Perotte, and Karthik Natarajan. CEHR-BERT: Incorporating temporal information from structured EHR data to improve prediction tasks. In *Proceedings of Machine Learning for Health*, volume 158 of *Proceedings of Machine Learning Research*, pp. 239–260. PMLR, 2021.

Chao Pang, Xinzhuo Jiang, Nishanth Parameshwar Pavinkurve, Krishna S. Kalluri, Elise L. Minto, Jason Patterson, Linying Zhang, George Hripcsak, Gamze Gursoy, Noemie Elhadad, and Karthik Natarajan. CEHR-GPT: Generating electronic health records with chronological patient timelines. *arXiv preprint arXiv:2402.04400*, 2024.

Matthew Phelan, Nrupen A. Bhavsar, and Benjamin A. Goldstein. Illustrating informed presence bias in electronic health records data: How patient interactions with a health system can impact inference. *eGEMs*, 5(1):22, 2017. doi: 10.5334/egems.243.

Alvin Rajkomar, Eyal Oren, Kai Chen, Andrew M. Dai, Nissan Hajaj, Michaela Hardt, Peter J. Liu, Xiaobing Liu, Jake Marcus, Mimi Sun, Patrik Sundberg, Hector Yee, Kun Zhang, Yi Zhang, Gerardo Flores, Gavin E. Duggan, Jamie Irvine, Quoc Le, Kurt Litsch, Alexander Mossin, Justin Tansuwan, De Wang, James Wexler, Jimbo Wilson, Dana Ludwig, Samuel L. Volchenboum, Katherine Chou, Michael Pearson, Srinivasan Madabushi, Nigam H. Shah, Atul J. Butte, Michael D. Howell, Claire Cui, Greg S. Corrado, and Jeffrey Dean. Scalable and accurate deep learning with electronic health records. *npj Digital Medicine*, 1(1):18, 2018. doi: 10.1038/s41746-018-0029-1.

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.

Pawel Renc, Yugang Jia, Anthony E. Samir, Jaroslaw Was, Quanzheng Li, David W. Bates, and Arkadiusz Sitek. Zero-shot health trajectory prediction using transformer. *npj Digital Medicine*, 7:256, 2024. doi: 10.1038/s41746-024-01235-0.

Junyuan Shang, Tengfei Ma, Cao Xiao, and Jimeng Sun. Pre-training of graph augmented transformers for medication recommendation. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*, pp. 5953–5959, 2019. doi: 10.24963/ijcai.2019/825.

Ethan Steinberg, Ken Jung, Jason A. Fries, Conor K. Corbin, Stephen R. Pfohl, and Nigam H. Shah. Language models are an effective patient representation learning technique for electronic health record data. *Journal of Biomedical Informatics*, 113:103637, 2021. doi: 10.1016/j.jbi.2020.103637.

Ethan Steinberg, Jason Fries, Yizhe Xu, and Nigam Shah. MOTOR: A time-to-event foundation model for structured medical records. *arXiv preprint arXiv:2301.03150*, 2023.

Minghui Sun, Matthew M. Engelhard, and Benjamin A. Goldstein. Borrowing from the future: Enhancing early risk assessment through contrastive learning. In *Proceedings of the 10th Machine Learning for Healthcare Conference*, volume 298 of *Proceedings of Machine Learning Research*, pp. 1–21. PMLR, 2025.

Brandon Theodorou, Cao Xiao, and Jimeng Sun. Synthesize high-dimensional longitudinal electronic health records via hierarchical autoregressive language model. *Nature Communications*, 14(1):5305, 2023. doi: 10.1038/s41467-023-41093-0.

Sindhu Tipirneni and Chandan K. Reddy. Self-supervised transformer for sparse and irregularly sampled multivariate clinical time-series. *ACM Transactions on Knowledge Discovery from Data*, 16(6):1–17, 2022. doi: 10.1145/3516367.

Fei Wang and Jimeng Sun. PSF: A unified patient similarity evaluation framework through metric learning with weak supervision. *IEEE Journal of Biomedical and Health Informatics*, 19(3): 1053–1060, 2015. doi: 10.1109/JBHI.2015.2425365.

Zifeng Wang and Jimeng Sun. PromptEHR: Conditional electronic healthcare records generation with prompt learning. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 2873–2885, 2022. doi: 10.18653/v1/2022.emnlp-main.185.

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, 2023a.

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*, 6:135, 2023b. doi: 10.1038/s41746-023-00879-8.

Cao Xiao, Edward Choi, and Jimeng Sun. Opportunities and challenges in developing deep learning models using electronic health records data: A systematic review. *Journal of the American Medical Informatics Association*, 25(10):1419–1428, 2018. doi: 10.1093/jamia/ocy068.

Yixuan Yang, Mehak Arora, Ryan Zhang, Baraa Abed, Junseob Kim, Tilendra Choudhary, Md Hassanuzzaman, Kevin Zhu, Ayman Ali, Chengkun Yang, Alasdair Edward Gent, Victor Moas, and Rishikesan Kamaleswaran. Clin-JEPA: A multi-phase co-training framework for joint-embedding predictive pretraining on EHR patient trajectories. *arXiv preprint arXiv:2605.10840*, 2026.

Zhichao Yang, Avijit Mitra, Weisong Liu, Dan Berlowitz, and Hong Yu. TransformEHR: Transformer-based encoder-decoder generative model to enhance prediction of disease outcomes using electronic health records. *Nature Communications*, 14:7857, 2023. doi: 10.1038/s41467-023-43715-z.

Andrew Zhang, Tong Ding, Sophia J. Wagner, Caiwei Tian, Ming Y. Lu, Rowland Pettit, Joshua E. Lewis, Alexandre Misrahi, Dandan Mo, Long Phi Le, and Faisal Mahmood. A multimodal and temporal foundation model for virtual patient representations at healthcare system scale, 2026. URL https://arxiv.org/abs/2604.18570.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. Qwen3 Embedding: Advancing text embedding and reranking through foundation models. *arXiv preprint arXiv:2506.05176*, 2025.

A Implementation and Evaluation Details

Inputs and batching. We encode catalog descriptions once with Qwen3-Embedding-8B (Zhang et al., 2025) and keep these embeddings fixed. Numeric inputs contain $\operatorname{asinh}(x_i)$, a value-present flag, a code-specific bin embedding, and a unit embedding. Additive time features contain daily, weekly, monthly, quarterly, and yearly sine and cosine phases, normalized age, log absolute time, and the log gap from the preceding availability block. Windows contain at most 1,024 temporal events, end at complete blocks, and contain up to 128 block-end anchors.

Architecture and objectives. The Transformer has six layers, width 512, eight attention heads, and feed-forward width 2,048. Age rotary embeddings use fractional days. The final model has 16 slots of width 128, a 2,048-dimensional exported state, and 46.9M parameters including pretraining heads. The prospective head uses eight intervals over five years, with code-specific quantile boundaries and shared fallbacks. The fingerprint count cap is 32; $\lambda_{\text{count}} = 0.02$, $\lambda_{\text{value}} = 0.05$, and $\lambda_{\text{fp}} = 0.3$ were set by short gradient-balance runs completed before downstream evaluation.

Optimization. We train for 10,000 optimizer steps in bfloat16 with an effective batch of 224 windows, AdamW at learning rate 10^{-4} and weight decay 10^{-4}, 1,000 warmup steps, and cosine decay to 10% of the peak. We evaluate on a fixed 3,200-patient validation subset every 1,000 steps and select the lowest total pretraining loss; no downstream label selects checkpoints or tunes encoders.

A.1 Baseline details

ORA uses a causal Transformer, age rotary embeddings, a last-event patient vector, and the joint grid in Equation 3. CEHR-GPT uses next-token, sub-time-token, and time-to-visit losses and exports its final probe token. CORE-BEHRT uses bidirectional masked-event prediction and exports its classification token; its adapter expands the diagnosis and medication input of the original experiments to all temporal modalities. All models use the same split, cutoff rule, modalities, and 1,024-position limit. CEHR-GPT and CORE-BEHRT retain their native serialization and omit numeric and categorical values under those recipes. Approximate parameter counts are 108M (CEHR-GPT), 22.5M (CORE-BEHRT), 27.9M (ORA), and 46.9M (CHRONOS). The comparison evaluates frozen states; the original CORE-BEHRT protocol fine-tunes its encoder.

A.2 Evaluation protocol

For prediction, each snapshot is the largest suffix of complete availability blocks within 1,024 events. Feature standardization and L2 penalty selection use training and validation folds, capped deterministically at 60,000 and 20,000 samples; the complete test fold is scored once. Already visible outcomes are excluded. The 16 tasks comprise 30-day readmission, ICU length of stay above seven days, ICU and in-hospital mortality, and 12 condition-specific mortality outcomes. Retrieval uses one end-of-record snapshot for 18,568 eligible test patients. Positives contain a diagnosis followed by its associated medication in the original window; all candidates contain some diagnosis followed by a later medication. Each positive queries all other candidates. Defining diagnosis, prescription, and administration codes are deleted before encoding, after which the window is backfilled while the original label remains fixed.

A.3 Cross-model results

Clinical prediction. Table 3 reports the full task-level results. CHRONOS has a higher point estimate than ORA on 15 of 16 tasks; six paired patient-bootstrap intervals exclude zero in favor of CHRONOS, and none excludes zero against it. The only point-estimate loss, on acute-pancreatitis mortality, is below 0.0001 AUROC.

The comparison tests whether task labels are linearly accessible in frozen exported states without downstream encoder fine-tuning. The larger gap to CORE-BEHRT should additionally be read with its all-modality input adaptation and still-improving validation loss in mind; its masked-token loss had not converged at the 150-epoch compute cap.

Target-masked retrieval. Table 4 evaluates frozen-state geometry without a fitted readout. The final 16-slot CHRONOS model reaches 0.701 mean AUROC, versus 0.641 for ORA, 0.606 for CEHR-GPT, 0.611 for CORE-BEHRT, and 0.678 for a same-window sublinear TF-IDF bag. CHRONOS also improves mAP (0.0519 versus 0.0460) and Precision@10 (0.1066 versus 0.0842), each on seven of eight phenotypes (Appendix B.2). Direct defining codes are removed before window selection. The benchmark measures retrieval from indirect context and can still use correlated code co-occurrence.

Target masking leaves correlated diagnoses, treatments, utilization, and demographics, so retrieval reflects indirect context and does not establish inference of an unobserved diagnosis. Under stricter matching on utilization, complete masked-prefix record size, input truncation, and the deletion and backfill footprint, CHRONOS and TF-IDF are statistically indistinguishable (difference 0.0012, 95% CI $[-0.0016, 0.0040]$). We therefore interpret Table 4 as evidence of operational competitiveness, with a cross-model advantage that depends on the evaluation footprint. The matched within-architecture fingerprint contrast remains positive under the same control (0.0251, 95% CI $[0.0230, 0.0275]$; Appendix B.2). CHRONOS and the same-window sparse control also retrieve substantially different patients: their top-10 neighbor sets overlap by about 10%.

A.4 EVENT FEATURES AND TEMPORAL ENCODING

Event input. Each code and categorical value has a short text description. A fixed text encoder maps these descriptions into a semantic input space, and learned projections map them into the model space. Thus, the input path is text-grounded, while the prospective and fingerprint targets use discrete vocabulary identifiers. The input for event i is

$$\begin{aligned} u_i = \mathrm{LN}\big(&W_c E(c_i) + W_x E(x_i^{\mathrm{text}}) + f_{\mathrm{num}}(x_i) + E_{\mathrm{bin}}(x_i) \\ &+ E_{\mathrm{unit}}(x_i) + E_{\mathrm{mod}}(m_i) + f_{\mathrm{time}}(t_i, \Delta a_i)\big). \end{aligned} \tag{6}$$

Numeric values contribute continuous, missingness, code-specific bin, and unit features. The time representation combines absolute patient time with the gap from the previous availability block. The gap is shared by all events in the same block and is therefore independent of their internal order.

Sex, race, and language are encoded as static tokens and placed in an always-visible initial block. We do not include these tokens in the supervision vocabulary. Static attributes therefore provide context but cannot become trivial prediction targets.

Block-causal Transformer. A Transformer encodes the event sequence. A token in block b can attend to all real tokens in blocks $b' \leq b$. Attention is bidirectional inside one block and causal across blocks. This rule removes the arbitrary ordering of events that become available together.

We use rotary position embeddings based on patient age together with the additive time features in Equation 6. Rotary age gives each attention layer a relative temporal signal, while the additive path exposes absolute time patterns and the gap from the preceding block.

B REPRESENTATION AUDITS

The primary attribution uses the exact matched seed-0 10-slot pair: resolved configurations differ only in whether the fingerprint is enabled, and both arms share source tree `5fabd53e`; all 123 shared tensors have identical shapes, and the parameter difference is exactly the pretraining-only fingerprint head. The seed-1 arms have matched configurations and different source-tree versions, so they provide independent replication. All models use identical evaluation rows. Geometry and retrospective-recovery analyses test whether the raw-cosine change survives record-footprint controls and whether each part of the retrospective target is directly recoverable.

B.1 TRAIN-FITTED GEOMETRY TRANSFORMATIONS

For each model, PCA projection, truncated PCA whitening, and ridge-regularized whitening are fitted on the same label-free sample of 20,000 training-fold masked snapshots. The transforms are frozen and applied unchanged to test states. Query bootstraps are conditional on each trained run and runs are never pooled. The full grid contains PCA dimensions 8, 16, 32, 64, and 128 and seven ridge strengths. No transform is selected using retrieval labels.

Table 5: Target-masked retrieval under complementary ranking metrics. Both representations receive the same complete-block-aligned suffix of at most 1,024 events.

Representation	AUROC	mAP	P@10	R@50
Windowed sublinear TF-IDF	0.6778	0.0460	0.0842	0.0174
CHRONOS ($K = 16$)	**0.7012**	**0.0519**	**0.1066**	**0.0190**

Table 6: Direct recovery of the observed history from matched 10-slot states. Seeds are reported separately. Higher values are better for every metric.

Metric	Seed 0 no fingerprint	Seed 0 fingerprint	Difference	Seed 1 no fingerprint	Seed 1 fingerprint	Difference
Full presence macro AUROC	0.8984	**0.9107**	$+0.0123$	0.8979	**0.9105**	$+0.0126$
Visible presence macro AUROC	0.9139	**0.9266**	$+0.0127$	0.9132	**0.9254**	$+0.0122$
Omitted presence macro AUROC	0.9045	**0.9161**	$+0.0116$	0.9049	**0.9165**	$+0.0116$
Full Recall@50	0.2037	**0.2151**	$+0.0113$	0.2044	**0.2147**	$+0.0103$
Full count R^2	-0.0491	**0.0074**	$+0.0565$	-0.0473	**0.0027**	$+0.0500$
Omitted count R^2	0.1698	**0.2283**	$+0.0584$	0.1716	**0.2236**	$+0.0521$
Latest-value macro R^2	0.1576	**0.2018**	$+0.0442$	0.1633	**0.2034**	$+0.0401$

The seed-0 states have Shannon effective ranks of 38.7 without the fingerprint and 61.7 with it; the corresponding values are 40.5 and 60.0 in seed 1. Full-rank covariance condition numbers exceed 10^{15}, and unregularized whitening amplifies some directions by approximately 10^8. Full-rank ZCA is therefore a sensitivity analysis. The contrast remains positive on all eight phenotypes over PCA dimensions 16 through 128 in every cohort and both runs. Truncated and ridge whitening attenuate the contrast non-monotonically, so no unique fraction can be assigned to covariance. In a label-free diagnostic, transplanting the fingerprint model's eigenvalue spectrum into the no-fingerprint states reproduces 24% of the raw gap while retaining the original principal directions. Variance allocation therefore contributes to the raw-cosine effect without fully reproducing it.

B.2 RETRIEVAL RANKING AND MASKING AUDITS

Relative to the same-window TF-IDF control, the paired gains in mAP, Precision@10, and Recall@50 are 0.0059, 0.0223, and 0.0017; each is positive on seven of eight phenotypes and its query-bootstrap interval excludes zero. The representations also induce different neighborhoods: they share only 10.0% of their top-10 neighbors (rank-biased overlap 0.32).

The original 55-code deletion set missed drug-name proxies in other modalities. Expanding it to 92 codes across the full catalog changes no arm by more than 0.0005 mean AUROC. Deletion nevertheless leaves a cohort-dependent footprint: positive patients lose from 38 to 61 events on average, compared with about 19 for negatives. In a sensitivity analysis matching negatives on utilization, complete masked-prefix size, input truncation, deleted events and blocks, and backfill depth, CHRONOS and windowed TF-IDF are tied (difference 0.0012, 95% CI $[-0.0016, 0.0040]$), whereas the matched fingerprint contrast remains positive (0.0251, 95% CI $[0.0230, 0.0275]$, eight of eight phenotypes). The first comparison therefore describes operational retrieval on the unconditioned cohort, and its advantage depends on the evaluation footprint; the within-architecture attribution is more robust.

B.3 DIRECT RETROSPECTIVE RECOVERABILITY

We fit identical multi-output ridge readouts on frozen states. Codes are selected solely by at least 1% training prevalence within each target set. Targets separately summarize the visible input window, codes appearing only in the omitted prefix, and the full history to the cutoff. Regularization is selected within training data and evaluation uses 37,213 untouched test patients; the pretrained fingerprint head is discarded.

The omitted target includes a code only when it occurs before the visible window and is absent from that window. It therefore excludes direct token re-reading. The same-sized presence gains on visible and omitted history show that retrospective supervision also improves inference about history displaced by the context limit. Latest values are winsorized at training-fold 1st and 99th percentiles;

Table 7: Observation-process R^2 for the matched 10-slot models.

Target	No fingerprint	Fingerprint	Difference
Events, trailing 30 days	0.771	0.664	-0.108
Active days, trailing 30 days	0.664	0.588	-0.076
Encounters, trailing 30 days	0.769	0.694	-0.075
Laboratory rate	0.774	0.681	-0.092
Medication rate	0.789	0.710	-0.079
Time since encounter	0.903	0.845	-0.058
Total record events	0.881	0.918	$+0.038$
Visible window length	0.972	0.987	$+0.015$
Window truncated	0.879	0.908	$+0.028$
Mean	0.822	0.777	-0.045

385 variables satisfy the train-only support and spread criteria. The result concerns statistically inferable omitted history, since events absent from the input remain unavailable to the encoder.

B.4 Prospective objective and neighborhood readout

On 150 identical validation batches, the prospective total changes from 0.07535 to 0.07517 in seed 0 and from 0.07535 to 0.07512 in seed 1. The hazard term changes by $+0.00009$ and -0.00002. Event NLL is 0.0033 worse in both runs, while value NLL improves by 0.0356 and 0.0266 and value accuracy improves by 0.0193 and 0.0138. Retrospective supervision therefore leaves total prospective fit essentially unchanged while modestly improving its value component.

We also evaluate a nonparametric downstream readout on the 16 prospective tasks. Neighbors come from training subjects, $k \in \{1, 5, 10, 25, 50, 100, 200\}$ is selected on validation, and test is evaluated once. Fingerprint supervision changes mean kNN AUROC from 0.8444 to 0.8468 in seed 0 and from 0.8451 to 0.8488 in seed 1, with improvements on 10 and 13 of 16 tasks. The effect is small, but exceeds the corresponding change under linear probes and agrees with a change in neighborhood geometry.

B.5 Observation-process decodability

We fit ridge probes to one end-of-record state per patient, selecting regularization on validation and evaluating once on the complete test fold. Table 7 reports the clean fingerprint attribution. The unused no-prior-encounter target has zero variance and is excluded.

The fingerprint reduces every recent-intensity or recency readout but increases all three record-extent readouts. This pattern shows a reallocation of observation-process information across targets.

B.6 Documentation-inflation intervention

Laboratory and medication events are duplicated in their original availability blocks. The primary cohort retains every original event after $\times 4$ inflation, isolating redundant documentation from context loss; it contains 2,724 retrieval patients and 39,115 probe snapshots and is consequently selected for shorter records. Because fingerprint counts can change under duplication, this tests functional stability under redundant documentation. Invariance of every training target lies outside its scope. A repeated-forward control gives relative L2 near 10^{-5} and zero probe-label flips.

Fingerprint supervision increases raw coordinate movement by about 10% relative to its control. Behavioral stability remains comparable, and both models change under the intervention. In a secondary fixed-budget intervention, duplicates evict older history; retrieval AUROC then falls by 0.0172 with the fingerprint versus 0.0013 without it. That comparison mixes redundant documentation with genuine information loss and therefore serves as a sensitivity analysis.

Table 8: Effect of content-preserving $\times 4$ documentation inflation on matched 10-slot models. Lower is better for drift and flip rate; higher is better for neighbor overlap.

Statistic	No fingerprint	Fingerprint	Difference
Relative L2 drift	0.2494	0.2732	$+0.0238$
Neighbor Jaccard@10	0.5460	0.5869	$+0.0409$
Probe-label flip rate	0.0199	0.0189	-0.0010
Mean probe AUROC change	-0.0080	-0.0082	-0.0002

Table 9: State-capacity scaling at fixed slot width. Probe AUROC averages four common tasks; retrieval averages eight target-masked phenotypes. The wide one-slot control matches the 10-slot state dimension. $K = 16$ was selected using validation data. Parameter counts include pretraining heads.

Slots	State dim.	Params. (M)	$100\times$ NLL $\downarrow$	Probe $\uparrow$	Retrieval $\uparrow$
1	128	39.0	6.857	0.841	0.666
2	256	39.6	6.789	0.846	0.681
4	512	40.6	6.735	0.851	0.684
8	1024	42.7	6.682	0.853	0.692
10	1280	43.8	6.662	0.854	0.697
16	2048	46.9	6.639	0.854	**0.701**
32	4096	55.3	**6.589**	0.855	0.700
1 (wide)	1280	59.6	6.591	**0.855**	0.690

C Slot Intervention Details

C.1 State capacity

We vary the number of slots at fixed width 128 while retaining the full pretraining objective. Table 9 reports the four prediction tasks available for every slot-sweep run and eight target-masked retrieval cohorts. At $K = 32$, an additional 8.4M parameters improve the four-task probe mean by only 0.001 over $K = 16$ and do not improve retrieval. Parameter counts include pretraining heads that are removed at inference; the sweep does not measure inference efficiency.

C.2 Routing and readout interventions

We use a fixed, label-free cohort of 3,200 validation patients. Evidence interventions remove one clinical group and subtract the effect of deleting the same number of events from matched recency bands. Readout interventions replace one slot by its cohort mean and measure prospective loss, fixed probes, and retrieval. Row-normalized group by slot and task by slot matrices are compared with nulls that permute slot identities independently within each row. We also refit each probe without one slot. Attention probabilities were verified by reconstructing the pooling-module output.

Table 10: Slot alignment under slot interventions. Row-normalized NMI is compared with a null that independently permutes slot identities within each row. No matrix exceeds its null.

Intervention matrix	NMI	Null mean	p
Evidence group $\times$ slot	0.126	0.122	0.290
Evidence recency $\times$ slot	0.406	0.406	0.520
Prospective readout $\times$ slot	0.130	0.119	0.126
Prediction task $\times$ slot	0.132	0.161	1.000
Retrieval phenotype $\times$ slot	0.252	0.317	1.000

The trained cohort profile reproduces across halves at $r = 0.998$, against a within-patient slot-permutation null of -0.009 ± 0.113 ($p = 0.005$). Within one patient, slots can attend to different events. Across patients the largest cohort-mean slot by modality deviation in raw attention mass is

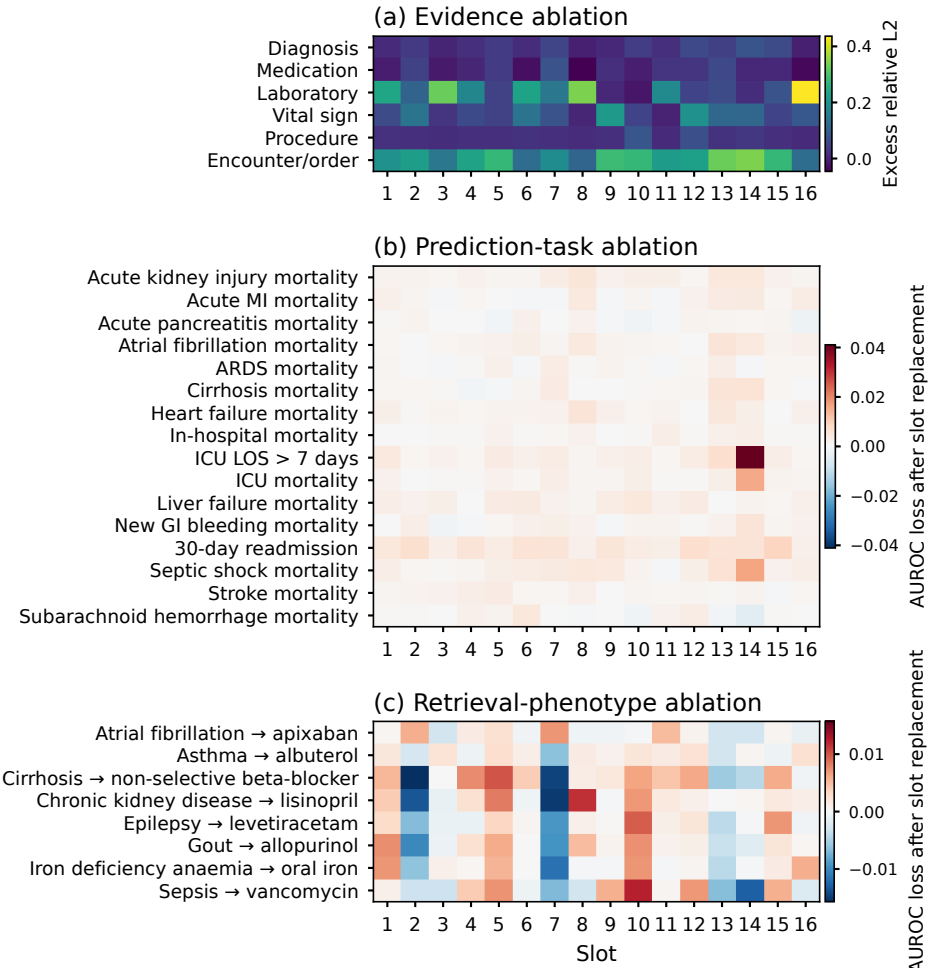


Figure 4: Slot intervention analysis for the final 16-slot model. (a) Change in each slot after removing one evidence group, relative to recency-matched random deletion. Panels (b) and (c) show AUROC lost when each slot is replaced by its cohort mean for frozen prediction probes and target-masked retrieval. Effects are distributed across slots and rows do not align with stable slot identities under the permutation tests in Table 10.

0.092; after normalizing attention mass by each group's token share the cohort profile is reproducible ($r = 0.998$) but still assigns no modality to any slot. An untrained initialization control on the same patients reaches $r = 0.977$, explains 3.4% of patient-level slot-by-modality routing variation, and has a maximum normalized slot-by-modality deviation of 0.14. The trained model explains 27.7% of this variation and has a maximum normalized deviation of 1.96. Thus, split-half reproducibility alone is insufficient evidence of learned routing. None of five intervention matrices exceeds its within-row permutation null (p from 0.126 to 1.000), and tasks use from 6.3 to 14.7 effective slots. Mean leave-one-slot-out AUROC loss falls from 0.0019 with the original probe to 0.0003 after refitting, which recovers 86.4% of the loss. No slot is individually load-bearing.

C.3 Fingerprint replication

The seed-0 10-slot pair is the exact controlled intervention: both arms use source tree `5fabd53e`, and resolved configurations differ only in fingerprint enablement and run name. Seed 1 replicates the effects with matched configurations, identical rows and targets, and separate reporting, but its arms originate from source trees `dcbeca04` and `d425c122`. Seed 1 therefore provides independent replication without the exact source-tree control of seed 0. All four checkpoints select step 10,000. The retrospective probes use identical rows, targets, and selected ridge penalty. Seed 1 includes inert

dynamics and variance-floor configuration fields shared by both arms; a 200-step replay verified unchanged training records under these fields. The final 16-slot model and most slot-count runs use one pretraining seed.

D LIMITATIONS AND BROADER IMPACT

We study one deidentified US hospital dataset, structured events, and fixed task families; external-system transfer and clinical notes remain untested. Frozen probes measure accessibility; fine-tuning ceilings remain unknown. Native baselines differ in tokenization, parameter count, value handling, and training budget. Utilization matching includes measured covariates. Matching on the footprint created by target deletion removes the raw cross-model retrieval advantage over TF-IDF, while the within-architecture fingerprint effect remains positive. A target-free anti-collapse control could not be trained under the exact source tree of the primary pair, so generic variance regularization remains an unresolved alternative mechanism for part of the geometry change. The content-preserving duplication cohort favors short records. Fingerprint supervision increases record-size decodability and is more vulnerable when redundant events displace genuine history from a fixed context. The final model and most slot-count runs use one seed; the exact fingerprint intervention uses seed 0 and the second pair is an independent cross-tree replication. Slot scaling changes dimension and parameters, only partly addressed by the wide one-slot control.

Clinical representations can reproduce inequities in documentation, access, treatment, and outcome. Race and language provide input context but are not targets, and we make no fairness claim. The model is retrospective and not intended for direct clinical decisions; deployment requires prospective validation, calibration, subgroup and privacy review, and human oversight.

E EXACT IMPLEMENTATION SPECIFICATION OF THE EXPORTED-STATE OBJECTIVE

This appendix records the implementation used by the reported CHRONOS models. It fixes the target clocks, window and censoring rules, pooling parameter sharing, mark treatment, and loss reductions that are abbreviated in the main text.

E.1 AVAILABLE HISTORY, SAMPLED CONTEXT, AND TARGET CLOCKS

An event is $e_i = (c_i, m_i, t_i, a_i, x_i)$, with clinical time t_i and availability time a_i. The store is ordered by availability; events with equal availability form one block. Let d_0 denote the input-only static context. At anchor cutoff u_a, the complete observed prefix is

$$H_a = \big(d_0, (e_i : a_i \le u_a)\big). \tag{7}$$

Training samples a temporal window W_ω whose end is an availability-block end. Its start is the earliest boundary of the largest complete-block suffix within the 1,024-event context budget. The sole exception is an availability block larger than the entire budget; that block is retained by cutting inside it and the sample is flagged. For records longer than the budget, the window end is sampled uniformly from eligible complete-block ends. Intermediate anchors are sampled without replacement from block ends with sufficient preceding context, and the final window position is always an anchor. The context at anchor a is

$$V_{\omega a} = \big(d_0, (e_i \in W_\omega : a_i \le u_a)\big), \qquad Z_{\omega a} = E_\theta(V_{\omega a}) \in \mathbb{R}^{K \times d_s}. \tag{8}$$

The exported state is the ordered concatenation

$$s_{\omega a} = \mathrm{rowvec}(Z_{\omega a}) = [z_{\omega a1}^\top, \ldots, z_{\omega aK}^\top]^\top. \tag{9}$$

At evaluation, a snapshot uses the largest complete-block suffix within the same context budget. Training-window resampling is disabled.

Temporal visibility. For real tokens i, j, let $b(i)$ be the availability-block index, with static context in an initial block. The encoder mask is

$$M_{ij} = \begin{cases} 0, & b(j) \le b(i), \\ -\infty, & b(j) > b(i), \end{cases} \qquad \mathrm{Attn}(Q, K, V) = \mathrm{softmax}\left(\frac{QK^\top}{\sqrt{d_h}} + M\right) V. \tag{10}$$

Padding keys are masked. Tokens attend bidirectionally within one availability block and causally across blocks.

Prospective target clock and complete follow-up. The clinical prediction origin is exactly the anchor availability time: $r_a = u_a$. The prospective pool contains the sampled window and the next at most 4,096 events in availability order. For code c, the target is the pool occurrence minimizing

$$(t_i - u_a,\ \mathrm{poolpos}(i)) \quad \text{subject to} \quad c_i = c, \quad t_i > u_a. \tag{11}$$

Thus the delay is strictly positive. Equal-clinical-time ties select the earliest event in deterministic pool order. Availability is not an additional target predicate: a future-dated event already available by u_a remains eligible, whereas a delayed historical event with $t_i \le u_a$ is excluded even if it becomes available later.

If the pool reaches the end of the stored record, its complete-follow-up endpoint C_ω is the maximum clinical time in that record. Otherwise, the extension is shortened to the last complete availability block and

$$C_\omega = \min_{i \in \text{omitted}} t_i - 1 \text{ minute}. \tag{12}$$

Consequently, no omitted event has clinical time at or before C_ω. The usable horizon is $h_a = \min\{T_{\max}, \max(0, C_\omega - u_a)\}$, with $T_{\max} = 5$ years. An occurrence is observed only when its delay is at most h_a. Anchors with $h_a = 0$ contribute no prospective targets. "Complete" here refers to the constructed recorded trajectory through C_ω. Unrecorded biological events beyond database follow-up are outside this definition.

E.2 SLOT-SPECIFIC POOLING AND PARAMETER SHARING

Let $h_i \in \mathbb{R}^d$ be contextualized token vectors. The anchor feature τ_a is produced by the same additive absolute-time encoder used for events, evaluated at u_a with a zero previous-block gap. In the final model each slot is one attention head. With affine maps written without their biases for clarity,

$$q_{ak} = q_k + A_k \tau_a, \tag{13}$$

$$\alpha_{aki} = \frac{\exp\big(\mathrm{LN}_q(q_{ak})^\top W_k^K \,\mathrm{LN}_h(h_i)/\sqrt{d_s}\big)}{\sum_{j\in\mathcal{I}_{\omega a}} \exp\big(\mathrm{LN}_q(q_{ak})^\top W_k^K \,\mathrm{LN}_h(h_j)/\sqrt{d_s}\big)}, \tag{14}$$

$$\widetilde{z}_{ak} = q_{ak} + W^O \sum_{i\in\mathcal{I}_{\omega a}} \alpha_{aki} W_k^V \,\mathrm{LN}_h(h_i), \tag{15}$$

$$z_{ak} = \mathrm{LN}_{\mathrm{out}}(\widetilde{z}_{ak} + f(\mathrm{LN}_f(\widetilde{z}_{ak}))). \tag{16}$$

W_k^K and W_k^V are slot-specific blocks of the key and value projections, and A_k is the corresponding block of the anchor-time projection. The query and token normalizations, output projection W^O, residual MLP f, residual normalization, and output normalization are shared across slots and act independently on each slot vector; none mixes slots. Static demographic tokens belong to $\mathcal{I}_{\omega a}$ and are therefore pooled directly. The reported final model has no separate static output slot, no cross-slot self-attention, and no diversity regularizer. It exports $K = 16$ slots of width $d_s = 128$.

E.3 MARKED TIME-TO-EVENT LIKELIHOOD AND CENSORING

For code c, let $0 = b_{c0} < b_{c1} < \cdots < b_{cP} = T_{\max}$ be right interval boundaries. A delay Δ belongs to interval p exactly when $b_{c,p-1} < \Delta \le b_{cp}$; an event on a boundary stays in the interval ending at that boundary. The reported models use $P = 8$. Boundaries are training-set quantiles of first-occurrence delays when a code has at least 100 observations and global training-set quantiles otherwise.

The final prospective readout uses the ORA paper-grid head. The code-query-over-slots head is absent. One affine map of the complete concatenated state produces a feature u_{ap} for every interval. For a code with numeric marks,

$$\ell_{acpv} = u_{ap}^\top w_{cv} + b_{cv}, \qquad v \in \mathcal{V}_c, \tag{17}$$

and invalid value classes are masked. For an unmarked code the output bank produces one logit ℓ_{acp} per interval, equivalent to a singleton dummy mark. Output weights and biases are code-specific but shared over intervals; interval dependence enters through u_{ap}. Numeric mark classes are fitted on training events using per-code equal-frequency boundaries, with at most five classes and a minimum support of 100. The reported configuration creates no categorical-text mark classes: categorical result text remains an input feature, while codes without supported numeric classes use the unmarked bank.

Let $s_{acp} = \log \sum_{v\in\mathcal{V}_c} \exp(\ell_{acpv})$ for a numeric code and $s_{acp} = \ell_{acp}$ otherwise. With the residual no-event logit fixed to zero,

$$D_{ac} = 1 + \sum_{p=1}^{P} e^{s_{acp}}, \qquad \pi_{ac}(p, v) = \frac{e^{\ell_{acpv}}}{D_{ac}}, \qquad \pi_{ac}(\varnothing) = \frac{1}{D_{ac}}. \tag{18}$$

For a known numeric mark observed in interval $p^\star$, the loss is $-\log \pi_{ac}(p^\star, v^\star)$. If the event time is observed but its numeric mark is missing, only the interval marginal is scored, $-\log \sum_v \pi_{ac}(p^\star, v)$; no missing-value class is added. Unmarked codes use the singleton interval probability.

For a censored pair, define the number of fully observed intervals

$$q_{ac} = \#\{p : b_{cp} \le h_a\}. \tag{19}$$

The admissible mass is the residual plus intervals strictly after the completed intervals,

$$\ell_{ac}^{\mathrm{cens}} = -\log\left[\pi_{ac}(\varnothing) + \sum_{p=q_{ac}+1}^{P} \sum_{v\in\mathcal{V}_c} \pi_{ac}(p, v)\right]. \tag{20}$$

A horizon exactly on a boundary therefore completes that interval. A censored pair with $q_{ac} = 0$ has zero loss and is excluded from the denominator; an observed first-interval event remains informative.

Let $\mathcal{I}_{\text{lik}}$ be all observed pairs and censored pairs with $q_{ac} > 0$ in the effective optimizer batch. The implementation computes

$$\mathcal{L}_{\text{pro}} = \frac{\sum_{(a,c)\in\mathcal{I}_{\text{lik}}} \ell_{ac}^{\text{time}}}{|\mathcal{I}_{\text{lik}}|} + \frac{\sum_{(a,c)\in\mathcal{I}_{\text{mark}}} \ell_{ac}^{\text{mark}}}{|\mathcal{I}_{\text{lik}}|}, \tag{21}$$

where $\mathcal{I}_{\text{mark}}$ contains observed events with a valid numeric class. The two terms together equal the joint grid-cell likelihood when the mark is known; omitting the second term marginalizes a missing mark. A zero likelihood-cell count yields zero contribution. Gradient accumulation weights each microbatch by its component denominator. Equation (21) is therefore the mean over the union. Microbatch means are never averaged.

E.4 Structured history fingerprint

The fingerprint is evaluated only at the final window anchor. Its target prefix contains every temporal event from the start of the stored record through the complete final availability block, including events before the model's sampled input window. Static input-only codes are absent from the supervision vocabulary. For each supervised code,

$$n_{ac} = \sum_{e_i \in H_a} \mathbf{1}\{c_i = c\}, \qquad y_{ac}^{p} = \mathbf{1}\{n_{ac} > 0\}, \tag{22}$$

$$y_{ac}^{n} = \log\bigl(1 + \min(n_{ac}, 32)\bigr). \tag{23}$$

The latest-value selector first removes occurrences without a valid fitted numeric class, then selects the remaining occurrence with greatest clinical time. Clinical-time ties select the later position in deterministic availability/store order. If no valid occurrence remains, the value target is masked. Thus the target is the latest valid recorded numeric value. It can differ from the value on the code's final occurrence and from the patient's current physiological value.

Presence. A shared MLP produces code logits from every normalized dynamic slot, and logits are max-pooled across slots:

$$r_{ack} = g_c^{p}(\text{LN}(z_{ak})), \qquad r_{ac} = \max_k r_{ack}, \qquad \ell_{ac}^{p} = \text{BCEWithLogits}(r_{ac}, y_{ac}^{p}). \tag{24}$$

Let $\mathcal{C}$ be the supervision vocabulary, with nonempty modality slices $\{\mathcal{C}_m : m \in \mathcal{M}\}$. Let N_{tr} be the number of training patients and df_c the training-patient frequency. The code weight is

$$\widetilde{\omega}_c = \min\left\{\log \frac{N_{\text{tr}}}{\max(\text{df}_c, 1)}, 8\right\}, \qquad \omega_c = \frac{\widetilde{\omega}_c}{|\mathcal{C}|^{-1}\sum_{j\in\mathcal{C}} \widetilde{\omega}_j}. \tag{25}$$

For a batch of B windows,

$$\mathcal{L}_{\text{pres}} = \frac{1}{|\mathcal{M}|} \sum_{m\in\mathcal{M}} \frac{1}{B|\mathcal{C}_m|} \sum_{a=1}^{B} \sum_{c\in\mathcal{C}_m} \omega_c \ell_{ac}^{p}. \tag{26}$$

The denominator counts windows and codes within each modality; IDF weights are excluded. Modalities then receive equal weight.

Count. The count head is a two-layer MLP on the concatenated dynamic slots and emits an unconstrained real prediction $\widehat{y}_{ac}^{n}$. Diagnosis and diagnosis-related-group codes are excluded, and the loss is a hurdle term evaluated only for present eligible codes. With residual $r = \widehat{y}_{ac}^{n} - y_{ac}^{n}$,

$$\rho_1(r) = \begin{cases} \frac{1}{2}r^2, & |r| \le 1, \\ |r| - \frac{1}{2}, & |r| > 1, \end{cases} \qquad \mathcal{L}_{\text{count}} = \frac{\sum_{a,c} m_{ac}^{n} \rho_1(r_{ac})}{\max(1, \sum_{a,c} m_{ac}^{n})}, \tag{27}$$

where $m_{ac}^{n} = 1$ exactly for present, count-eligible codes. This is PyTorch Huber loss with $\delta = 1$; scaled smooth-L_1 is not used.

Latest value. The value head uses learned code embeddings e_c. It maps e_c to a query, maps every normalized slot to a key, and attends over the unnormalized slot values:

$$\beta_{ack} = \mathrm{softmax}_k\left(\frac{g_q(e_c)^\top g_k(\mathrm{LN}(z_{ak}))}{\sqrt{r}}\right), \qquad \bar{z}_{ac} = \sum_k \beta_{ack} z_{ak}, \qquad o_{ac} = g_v([\bar{z}_{ac}; e_c]). \tag{28}$$

Logits outside the code's fitted class count are set to $-\infty$. $\mathcal{L}_{\mathrm{latest}}$ is cross-entropy averaged over all cells with a valid latest-value class. If that set is empty, the component contributes zero. The count and latest-value losses do not use IDF weights. Under gradient accumulation, presence, count, and latest-value microbatch means are weighted respectively by window count, eligible positive-count cells, and valid latest-value cells, reproducing each mean over the effective optimizer batch.

Combining these exact reductions,

$$\mathcal{L}_{\mathrm{fp}} = \mathcal{L}_{\mathrm{pres}} + \lambda_{\mathrm{count}}\mathcal{L}_{\mathrm{count}} + \lambda_{\mathrm{value}}\mathcal{L}_{\mathrm{latest}}. \tag{29}$$

E.5 Total objective

The prospective term is evaluated at sampled block-end anchors, whereas the fingerprint is evaluated at one final anchor per window. The reported coefficients give

$$\boxed{\mathcal{L}_{\mathrm{CHRONOS}} = \mathcal{L}_{\mathrm{pro}} + 0.30\mathcal{L}_{\mathrm{pres}} + 0.006\mathcal{L}_{\mathrm{count}} + 0.015\mathcal{L}_{\mathrm{latest}}.} \tag{30}$$

These coefficients are fixed. Measured fractions of gradient magnitude do not define them. At inference all pretraining heads are removed and the ordered concatenation of the slots is exported.

F OPERATIONAL PRESERVATION AND LIMITS OF INTERPRETATION

Status of these arguments. The following elementary, self-contained arguments clarify the representation contract. Their scope excludes new learning-theoretic guarantees and proofs of the reported empirical improvements. The actual method uses restricted readouts, finite data, mixed losses, and numerical optimization. Each argument below states the idealizations under which it applies.

F.1 A TARGET-RELATIVE REPRESENTATION CONTRACT

A representation contract specifies target families, readout families, losses, input information, and a state budget. For target type $j \in \{+, -\}$, define

$$\mathcal{R}_j(E) = \inf_{g \in \mathcal{G}_j} \mathbb{E}_{P_j}\big[\ell_j(g(E(V)), Y_j)\big], \qquad Y_+ = Y^+, \quad Y_- = \Phi(H). \tag{31}$$

P_+ and P_- retain the prospective versus final-anchor sampling distinction. A decoder-relative, budgeted specification is the risk pair $(\mathcal{R}_+(E), \mathcal{R}_-(E))$ for $E \in \mathcal{E}_B$, where $\mathcal{E}_B$ is the chosen encoder class with exported dimension at most B.

Raw loss alone cannot measure universal sufficiency: a stochastic target has irreducible uncertainty. Define a visible-input Bayes reference

$$\mathcal{R}^*_{j,V} = \inf_{h \text{ measurable}} \mathbb{E}_{P_j}[\ell_j(h(V), Y_j)], \qquad \Delta_j(E) = \mathcal{R}_j(E) - \mathcal{R}^*_{j,V} \geq 0. \tag{32}$$

The inequality holds because every state readout is also a function of V. In principle, an operational tolerance contract could require $\Delta_j(E) \leq \epsilon_j$. The present study does not report prespecified tolerances or estimate these Bayes references; the contract therefore serves as a design objective and does not certify these inequalities.

For fitted finite-sample heads, the attained empirical risk is not the infimum over all population readouts. Pretraining heads and downstream linear probes are also different readout families. These distinctions prevent "trained to predict" from silently becoming "proven sufficient for all downstream tasks."

F.2 LOG-LOSS DECOMPOSITION AND THE LIMIT IMPOSED BY HIDDEN HISTORY

Assume Y is finite-valued, $S = E(V)$ is deterministic at evaluation, all relevant entropies are finite, and $g(\cdot \mid S)$ is a probability readout. Then

$$\boxed{\mathbb{E}[-\log g(Y \mid S)] - H(Y \mid V) = I(Y; V \mid S) + \mathbb{E}_S \,\mathrm{KL}\big(P(Y \mid S) \,\|\, g(\cdot \mid S)\big).} \tag{33}$$

Proof. Add and subtract $-\log P(Y \mid S)$ inside the expectation:

$$\mathbb{E}[-\log g(Y \mid S)] = H(Y \mid S) + \mathbb{E}_S \,\mathrm{KL}(P(Y \mid S) \| g(\cdot \mid S)).$$

Because S is a deterministic function of V, $H(Y \mid V, S) = H(Y \mid V)$. Hence

$$H(Y \mid S) - H(Y \mid V) = I(Y; V \mid S),$$

which yields Equation (33). Both right-hand terms are nonnegative. With an unrestricted optimal readout, the KL term is zero. This separates information discarded by the state from decoder approximation error. □

For a fingerprint target $Y = \Phi(H)$, the value of $H(Y \mid V)$ can be positive because the input omits part of the history. If two full histories have identical visible contexts but differ in an old code, every deterministic encoder gives them the same state. No readout can perfectly reconstruct both different labels from that state. The retrospective objective can encourage learning of *predictable* full-prefix summaries; it cannot create patient-specific knowledge absent from the input.

Scope for CHRONOS. The identity applies directly to a finite categorical target with ordinary log loss, including an individual presence target. For latest-value labels, the target distribution must be conditioned on the relevant supervision mask. For a complete prospective joint time and mark outcome it is an idealized reference. It is not, without further assumptions, an identity for the censored observed-data loss, the Huber count loss, or the whole weighted CHRONOS objective. Moreover, accurate marginal readouts of many fingerprint coordinates do not establish preservation of their full joint distribution. This limits the meaning of preservation; it does not establish lossless reconstruction or universal downstream sufficiency.

F.3 SCOPE OF THE WEIGHTED SCALARIZATION

Let $\eta = (\theta, \phi_+, \phi_-)$ contain encoder and both readout parameter sets, and let $\eta \in \mathcal{H}_B$ denote the chosen capacity-constrained model family. One way to motivate a tradeoff is

$$\min_{\eta \in \mathcal{H}_B} \mathcal{R}_+(\eta) \quad \text{subject to} \quad \mathcal{R}_-(\eta) \le \epsilon_-. \tag{34}$$

Its Lagrangian is

$$\mathcal{L}_{\mathrm{Lag}}(\eta, \lambda) = \mathcal{R}_+(\eta) + \lambda\big(\mathcal{R}_-(\eta) - \epsilon_-\big), \qquad \lambda \ge 0. \tag{35}$$

For a *fixed* λ, dropping the constant $-\lambda\epsilon_-$ leaves the same weighted-sum form as CHRONOS. This algebra does not identify a useful ϵ_-, establish constraint satisfaction, imply strong duality, or show that neural-network training solves Equation (34). The method does not optimize a dual variable.

The exact statement available. Suppose $\lambda > 0$ and η^* is a global minimizer of $F(\eta) + \lambda G(\eta)$. Then no feasible η has both $F(\eta) \le F(\eta^*)$ and $G(\eta) \le G(\eta^*)$, with at least one strict inequality. Otherwise its weighted sum would be strictly smaller, a contradiction. Thus a global weighted-sum optimum is Pareto-efficient for these two risks. It also minimizes F subject to $G \le G(\eta^*)$: any feasible point with smaller F would again have smaller weighted sum. □

This is an *ex post* threshold statement at a global optimum. It provides no guarantee for an arbitrary chosen threshold or an optimizer's numerical checkpoint. A nonconvex attainable-risk set can also contain Pareto points that positive linear scalarization does not recover. These facts justify the phrase "a fixed-weight surrogate for the preservation tradeoff." A claim that the loss is mathematically inevitable would exceed the result.

F.4 LINEAR ACCESSIBILITY NEED NOT DETERMINE COSINE GEOMETRY

For invertible $A \in \mathbb{R}^{B\times B}$, set $S' = AS$. An affine readout $w^\top S + b$ is reproduced exactly by $w' = A^{-\top} w$:

$$(w')^\top S' + b = w^\top S + b. \tag{36}$$

Therefore the class of unregularized affine prediction functions is unchanged. However,

$$\cos(As_i, As_j) = \frac{s_i^\top A^\top A s_j}{\sqrt{s_i^\top A^\top A s_i}\sqrt{s_j^\top A^\top A s_j}}, \tag{37}$$

which generally differs from the original cosine. Scaled orthogonal transformations preserve all cosines; a general invertible map does not.

A direct example makes the limitation of decodability explicit. Let $Y \in \{0, 1\}$ be a target label, $N \in \{-1, 1\}$ an unrelated attribute, and

$$s(Y, N) = (2Y - 1, MN), \qquad M > 1. \tag{38}$$

The first coordinate perfectly linearly separates Y. Yet same-label points with opposite N have cosine $(1-M^2)/(1+M^2)$, whereas opposite-label points with equal N have cosine $(M^2-1)/(1+M^2)$. As M grows, the former approaches -1 and the latter $+1$. Recoverability alone therefore does not force label-aligned retrieval geometry.

Important qualifications. Our probe evaluation uses feature preprocessing and L_2-regularized logistic regression. Those fitted procedures need not be invariant under general A: the transformed coefficient penalty is $\|A^{-\top} w\|_2^2$ and can differ from $\|w\|_2^2$. Likewise, the slot-wise max readout family need not be closed under arbitrary mixing of all slots. This argument establishes a conceptual separation. It does not establish exact invariance of the entire CHRONOS training objective or of its empirical probes. Nor does it prove that fingerprint supervision must improve retrieval.

F.5 COORDINATE DRIFT AND FIXED-READOUT STABILITY

Let a documentation transformation T induce $s' = s + \delta$. For a fixed linear logit $r = w^\top s + b$,

$$r' - r = w^\top \delta, \qquad |r' - r| \le \|w\|_2 \|\delta\|_2. \tag{39}$$

There is no converse lower bound in general: if $\delta \perp w$, the logit is unchanged even for large state displacement. For a logistic prediction $p = \sigma(r)$, $\sup_x \sigma'(x) = 1/4$ gives

$$|p' - p| \leq \tfrac{1}{4}|w^\top \delta| \leq \tfrac{1}{4}\|w\|_2\|\delta\|_2. \tag{40}$$

For a collection of fixed linear probes arranged as rows of W, any $\delta \in \ker W$ leaves all their logits unchanged. This demonstrates why latent drift is not synonymous with functional degradation. It does not imply that the observed perturbations actually lie in that nullspace. AUROC depends on between-patient rankings, so coordinate norms alone cannot determine its change.

F.6 Why count preservation conflicts with strict duplication invariance

Suppose a documentation transformation repeats a counted event $r > 1$ times without changing its recorded clinical values. Its count target changes by

$$\Delta y^n = \log\big(1 + \min(rn, C)\big) - \log\big(1 + \min(n, C)\big). \tag{41}$$

This is positive whenever multiplication changes the capped count. If a representation were exactly invariant, $E(T(H)) = E(H)$, a deterministic decoder would output the same count prediction for both records. It therefore could not exactly reconstruct both distinct count targets. Strict state invariance and exact preservation of this multiplicity-sensitive fingerprint cannot both hold on such pairs. □

Presence and latest-value targets can remain unchanged by identical duplication while counts change. Thus "content-preserving duplication" in the clinical sense is not necessarily *fingerprint-target-preserving*. The coherent robustness claim concerns stability of specified downstream functions. Universal invariance of every coordinate lies outside its scope. When duplication evicts genuine history under a fixed context budget, it additionally changes available information; we treat that intervention as a sensitivity analysis.