GLoC-EHR: Evidence-Cited Clinical Reasoning over Global Context and Local EHR Events

Anonymous authors
Paper under double-blind review

Abstract

Structured electronic health records (EHRs) contain a patient's clinical trajectory as a sequence of clinical codes. Answering clinical questions from such records requires both the context of the whole trajectory and the specific events that support the answer. We introduce GLoC-EHR, a multimodal language model that reads a contextual encoding of the record through a fixed-size global memory of the trajectory and a local memory of selected events. The model learns to generate hospital-course summaries from the global memory and descriptions of masked concepts from the local memory, aligning both with clinical text. It is then trained to cite evidence before answering, through rationale fine-tuning followed by group relative policy optimization (GRPO) with rewards for correct answers and record-supported evidence. On three MIMIC-IV outcome tasks, GLoC-EHR attains the highest macro AUROC among the compared models when it answers directly, whereas zero-shot LLMs reading the serialized record fall far behind. With evidence-cited reasoning, it stays close to its direct multi-task counterpart in macro AUROC, and the evidence terms of the objective reduce unsupported evidence at a similar macro AUROC. The local memory adds distinct supported findings, particularly under strict matching, without a detectable change in macro AUROC. Without retraining, GLoC-EHR transfers to EHRSHOT on par with EHR-BERT and answers two unseen laboratory questions better than zero-shot prompting of its own backbone.

1 Introduction

Electronic health records (EHRs) capture diagnoses, medication orders, procedures, and measurements over time, providing a rich longitudinal account of a patient's health and care. These records offer both a broader view of the patient's clinical trajectory and specific observations relevant to a clinical question. However, clinically relevant information can be dispersed across lengthy, heterogeneous patient records, making it challenging for clinicians to identify and synthesize information needed to guide patient care (Reichert et al., 2010).

To harness longitudinal EHR data for clinical prediction, EHR-based models have been developed for a wide range of clinical outcome-prediction tasks (Choi et al., 2016; Li et al., 2020b; Rasmy et al., 2021). In particular, pretrained EHR encoders learn contextual representations of medical codes from patient histories. BEHRT adapts masked language modeling to diagnosis sequences with age and visit information (Li et al., 2020b); Med-BERT learns transferable diagnosis representations through large-scale pretraining (Rasmy et al., 2021); and CEHR-BERT incorporates temporal information across multiple structured domains (Pang et al., 2021). These models provide a reusable foundation for prediction and can reduce the need for labeled data. However, their usual downstream interface is a task-specific predictive head: adapting to a new outcome generally requires fitting a new head or fine-tuning the model with corresponding labels (Rasmy et al., 2021; Wornow et al., 2023). Thus, transfer remains dependent on task-specific supervision, even when the underlying representation is reused. Moreover, an outcome score alone does not express the recorded findings that support a clinical inference. Extending these representations to natural-language reasoning therefore requires connecting the encoded record, the clinical question, and the evidence supporting an answer.

One approach to connecting structured EHRs with language models, which encode substantial clinical knowledge (Singhal et al., 2023; Jiang et al., 2023), is to serialize medical codes into textual

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.

descriptions that the models can embed or reason over (Hegselmann et al., 2023; Ben Shoham & Rappoport, 2024; Zhu et al., 2024; Hegselmann et al., 2026; Lin et al., 2025; Liao et al., 2025). Serialization retains an explicit account of the selected events at a growing context cost (Wornow et al., 2025). In our downstream cohorts, serializing even the 24-hour observation window yields an average of 4.7k tokens (90th percentile: 14k), increasing to 12.8k (48k) for the full index admission. Other approaches use intermediate textual summaries (Gao et al., 2026) or fuse language-model embeddings of domain-specific pseudo-notes (Lee et al., 2025). Such compression may omit details needed to substantiate a clinical inference. A distinct approach, Llemr, aligns individually encoded clinical events with a generative language model (Wu et al., 2024), reducing the input length relative to textual serialization. However, the resulting input still scales with the number of events.

Connecting structured EHRs with language models makes clinical information available for reasoning, but does not by itself ensure that the generated reasoning is supported by the patient's record. Recent work, including EHRMIND (Lin et al., 2025) and EHR-R1 (Liao et al., 2025), develops EHR-based reasoning through reinforcement learning with rewards for final-answer correctness, which do not explicitly distinguish supported from unsupported findings.

We therefore introduce GLoC-EHR, an EHR–text multimodal large language model that combines a global–local interface with evidence-aware reward training. A contextual EHR encoder reaches the language model through two routes: a global route, whose compact memory summarizes the observed trajectory, and a local route, whose memory retains selected individual events. Trajectory alignment, which generates the hospital-course summary of an admission, aligns the global route with clinical text, and concept alignment, which recovers descriptions of masked concepts, aligns the local route. Beyond answer correctness, an evidence-matching reward credits findings supported by the patient's record rather than concepts merely associated with the task. We study how the local route and this reward shape the generated reasoning on clinical outcome tasks, with direct prediction as a reference for the information in the representation. Our contributions are threefold:

- A global–local interface whose two routes are aligned with clinical text, giving a language model trajectory context and access to individual events; answering directly, GLoC-EHR attains the highest macro AUROC over three MIMIC-IV tasks (Section 4.2).
- An evidence-aware objective that combines this reward with a sufficiency loss on the cited findings; it reduces unsupported evidence with no detectable change in macro AUROC, which stays close to that of direct prediction (Sections 4.2–4.4).
- An evaluation of cited evidence under stricter matching and cited-slot masking, showing that the local route shapes which findings are cited, with no detectable effect on the answer; without retraining, GLoC-EHR also matches EHR-BERT on EHRSHOT and outperforms its zero-shot backbone on unseen laboratory questions (Sections 4.3 and 4.5).

2 Related Work

Structured EHR representations. RETAIN uses reverse-time attention over visits and variables (Choi et al., 2016), and the BERT-style encoders discussed above learn reusable contextual code representations (Li et al., 2020b; Rasmy et al., 2021; Pang et al., 2021). Autoregressive and time-to-event foundation models are instead pretrained to forecast future events in patient timelines (Steinberg et al., 2021; 2024; Kraljevic et al., 2024; Renc et al., 2024). We build on contextual representations to support language generation grounded in the record.

EHR–language alignment. Clinical text connects structured records with language models in several roles: serialized records read or embedded for prediction (Hegselmann et al., 2023; Ben Shoham & Rappoport, 2024; Hegselmann et al., 2026), post-training for EHR reasoning, including reinforcement learning (Lin et al., 2025; Liao et al., 2025), modality-specialist agents that summarize non-text inputs (Gao et al., 2026), and pseudo-notes embedded by a frozen language model (Lee et al., 2025). Closest to our setting, Llemr connects a frozen clinical event encoder to an LLM through a projection, as multimodal LLMs do for images, time series, and individual health data (Liu et al., 2023; Jin et al., 2024; Belyaeva et al., 2023), and trains it with schema alignment followed by reasoning instruction tuning (Wu et al., 2024). Our interface adds a compressed trajectory memory to a separate memory of selected events, and our reward checks whether the generated findings are supported by the record.

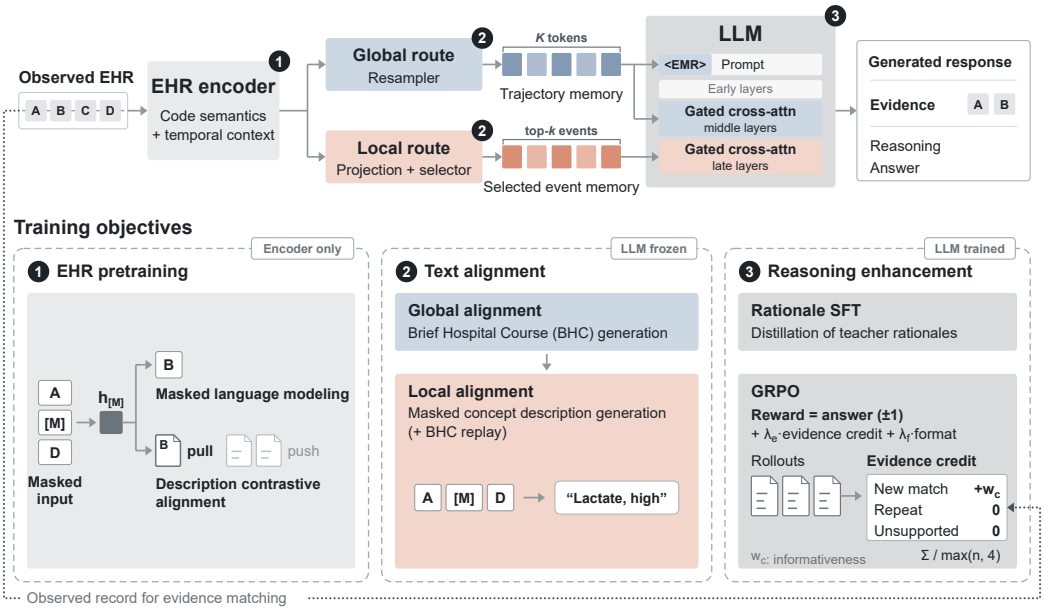

Figure 1: **Overview.** *Top:* An EHR encoder contextualizes the observed events. The global route compresses them into a few tokens that fill the <EMR> slot and are re-read by gated cross-attention in the middle layers; the local route supplies the top-k events through gated cross-attention in the late layers. The LLM generates cited evidence, a rationale, and an answer. *Bottom:* ① encoder pretraining; ② trajectory (global) then concept (local) alignment; ③ SFT on teacher rationales followed by GRPO, whose evidence credit is scored against the patient's record (dotted line).

Evidence support and rationale faithfulness. Attribution metrics check whether generated statements are supported by a source (Rashkin et al., 2023; Gao et al., 2023; Min et al., 2023), but a rationale's consistency with its input differs from its faithfulness to the process that produced an answer (Jacovi & Goldberg, 2020; Turpin et al., 2023). ERASER separates agreement with evidence annotations from behavioral tests of comprehensiveness and sufficiency (DeYoung et al., 2020), and studies of natural language explanations and chain-of-thought use interventions to test whether stated reasons affect predictions (Atanasova et al., 2023; Lanham et al., 2023). Accordingly, our matching reward for group relative policy optimization (GRPO) (Shao et al., 2024) measures support in the record rather than causation, and we evaluate it together with evidence coverage and an intervention that removes cited versus comparable uncited local representations before regenerating the response.

3 METHOD

GLoC-EHR connects an EHR encoder to a language model through two complementary routes: a global route for trajectory-level context and a local route for individual event-level evidence. After aligning both routes with clinical text, we train the model for evidence-grounded reasoning through supervised fine-tuning (SFT) on rationales followed by GRPO with answer-correctness and evidence-support rewards. Figure 1 summarizes the architecture and training procedure.

Problem setup. Given a clinical task instruction q, let $\mathbf{x} = (e_1, \ldots, e_T)$ be the events retained within its observation window. Each event consists of a medical concept c_t and associated temporal and event-type attributes. Given $(\mathbf{x}, q)$, our goal is to generate a response $o = (r, a)$, where r contains cited findings and a rationale grounded in $\mathbf{x}$, and a is the final task answer.

3.1 DESCRIPTION-INFORMED EHR ENCODER

A bidirectional EHR encoder produces contextualized event states $\mathbf{H} = f_{\text{EHR}}(\mathbf{x}) \in \mathbb{R}^{T \times d_E}$. Building on masked language modeling (MLM) (Devlin et al., 2019) pretraining for structured EHRs (Li et al.,

2020b; Rasmy et al., 2021; Pang et al., 2021), it learns these states from the concept, age, date, and type of each event (Appendix A.1). Prior studies show that incorporating textual concept descriptions can improve transfer across institutions and coding systems (Kim et al., 2026; Kirchler et al., 2026). We therefore incorporate description embeddings into the EHR encoder, using the backbone LLM to encode each concept's description rather than introducing a separate text encoder.

Because the backbone LLM is not trained for text embedding, we first assessed how well its representations distinguish concept descriptions (Appendix A.2) and found strong anisotropy in both mean-pooled input-token embeddings and final-layer hidden states (Ethayarajh, 2019). To address this, we applied centering, whitening, and ℓ_2 normalization (Su et al., 2021; Huang et al., 2021) to the LLM's mean-pooled final-layer hidden states. We use the resulting near-isotropic vectors, which showed the clearest concept structure among the evaluated representations, as frozen description embeddings $\mathbf{z}_c$. The final event input embedding combines the projected description embedding with a learned code-ID embedding as a scaled residual, together with temporal and event-type embeddings.

The encoder is pretrained with masked-concept prediction and a contrastive objective that aligns masked event states with their target concept descriptions. We apply code-ID dropout to discourage shortcut learning through exclusive reliance on learned code-ID embeddings and encourage the encoder to use description embeddings. At prediction targets, both the code-ID and description inputs are masked to prevent target-description leakage. The pretrained encoder is then frozen during interface alignment and downstream adaptation. Description processing, embedding equations, pretraining objectives, and sampling settings are specified in Appendix A.1.

3.2 Global and local EHR memories

Given the event states $\mathbf{H}$, we define two complementary routes for injecting EHR information into the LLM. First, the global route summarizes the overall context of the observed trajectory; second, the local route retains individual representations of selected events as event-level evidence.

Global memory. Following latent-query resampling (Jaegle et al., 2021; Alayrac et al., 2022; Li et al., 2023), a learned resampler uses K_g queries $\mathbf{Q}$ to attend to the event sequence and produce a fixed-size memory:

$$\mathbf{M}^{\mathrm{glob}} = f_{\mathrm{res}}(\mathbf{Q}, \mathbf{H}) \in \mathbb{R}^{K_g \times d_L}, \tag{1}$$

where d_L is the LLM's hidden dimension. These global memory tokens are inserted into the prompt as continuous embeddings and also serve as the memory for global cross-attention. The resampler therefore provides a compact interface whose size is independent of the number of retained events.

Local memory. Evidence-grounded reasoning requires linking generated findings to specific recorded concepts and their descriptions. The local route therefore maps each event state to the LLM's hidden dimension with a projection p_ℓ and ranks it with a selector s:

$$\mathbf{c}_t = p_\ell(\mathbf{H}_t), \qquad \mathcal{I} = \mathrm{TopK}_{K_\ell}\{s(\mathbf{c}_t) : t \in \mathcal{V}\}, \qquad \mathbf{M}^{\mathrm{loc}} = (\mathbf{c}_t)_{t\in\mathcal{I}}, \tag{2}$$

where $\mathcal{V}$ is the set of eligible event positions and selection retains up to K_ℓ events. Local vectors are accessed only through cross-attention. The selector does not see the task instruction, so its ranking is shared across tasks. It is first learned during concept alignment (Section 3.3) and then adapted once to rationale generation (Section 3.4); it remains fixed for direct prediction and GRPO.

Gated integration. The language model attends to each memory through separate residual modules (Alayrac et al., 2022). For route $r \in \{\mathrm{glob}, \mathrm{loc}\}$, a module after decoder block l updates the block's output hidden states $\mathbf{Z}^{(l)}$ as

$$\mathbf{Z}^{(l)} \leftarrow \mathbf{Z}^{(l)} + \sigma(\gamma_l^r)\,\mathrm{XAttn}_l^r\big(\mathbf{Z}^{(l)}, \mathbf{M}^r\big), \tag{3}$$

where XAttn_l^r is cross-attention from $\mathbf{Z}^{(l)}$ to the memory $\mathbf{M}^r$, σ is the logistic sigmoid, and the learned scalar gate γ_l^r controls the contribution of that route. Motivated by layer-dependent cross-modal information flow (Zhang et al., 2025b), we place global modules before local modules, making trajectory context available before the final event-level injections. Memory budgets, projections, layer placements, and attention masks are given in Appendix A.3.

3.3 TEXT ALIGNMENT OF THE TWO ROUTES

We align the global–local interface with clinical text at both the trajectory and concept levels. For trajectory-level alignment, we pair the structured event sequence of a hospital admission with the Brief Hospital Course (BHC) section of its discharge summary, treating them as structured and free-text views of the same admission. For concept-level alignment, we pair individual medical concepts with their textual descriptions. We train the global route first and the local route second, keeping the EHR encoder and base language model frozen throughout alignment.

Trajectory alignment. The global route is trained to generate the paired BHC text from the structured admission trajectory. Response-only next-token prediction updates the resampler and global cross-attention modules, with the local route disabled. BHC text serves as supervision for communicating trajectory-level information; downstream inputs consist of the structured observation window and the task instruction.

Concept alignment. We then mask clinical concepts and train the local route to recover their textual descriptions, interleaving these examples with BHC replay. The selector follows three phases. In *forced selection*, masked event positions are placed in local memory, and direct supervision trains the selector to identify these positions as initial selection targets. In *soft selection*, masked positions are excluded from the candidate set, and the description-generation loss trains the selector through differentiable memory weights to prioritize unmasked events that help reconstruct the target descriptions. Finally, *hard selection* retains the top-k events, matching the bounded local memory used downstream. Masked positions remain excluded, and the projection and selector are frozen while local cross-attention continues to adapt. The full schedule, auxiliary selector losses, and trainable components are detailed in Appendix A.4.

3.4 EVIDENCE-GROUNDED REASONING WITH GRPO

Supervised warm-up precedes reinforcement learning in general and medical reasoning models (Guo et al., 2025; Chen et al., 2024b) and, for EHRs, establishes task competence and preserves structured clinical reasoning (Lin et al., 2025). We therefore first perform SFT on teacher responses containing evidence, reasoning, and an answer. The teacher receives the observed record and gold outcome, so the targets are label-conditioned rationales. Conditioned only on the structured EHR representations and task instruction, the student is trained to generate the corresponding evidence-cited rationale and task answer. SFT also adapts the local route, because events useful for recovering masked concept descriptions need not be those that support a rationale. As in concept alignment, the selector and local projection are trained under soft selection, now with the rationale loss, and then frozen under hard top-K_ℓ selection, which matches the memory used in GRPO and evaluation. Teacher configuration and tuning details appear in Appendix A.6.

Reward design. Answer correctness alone does not distinguish a supported rationale from one that cites findings absent from the patient record. We therefore combine an outcome signal with an explicit evidence-support term:

$$R(o) = r_{\text{ans}} + \lambda_{\text{evid}} r_{\text{evid}} + \lambda_{\text{fmt}} r_{\text{fmt}} - r_{\text{len}}. \tag{4}$$

The answer term rewards a correct outcome, the format term encourages a parseable response, and the length term discourages excessive generation. Invalid answers and statements of the outcome itself (e.g., that the patient died) receive a fixed failure reward below any valid total.

The evidence term scores the bullets listed under the evidence heading of the response; the reasoning prose is not scored. A lexical matcher counts a bullet as *supported* if its content words match the typed description of an event in the patient's input window, by containment or an IDF-weighted overlap of at least 0.55, with the same direction, negation, and family-history attribution (Appendix A.7). Matching covers every event in the input window, not only the K_ℓ events in local memory, so the reward itself does not favor the local route; descriptions and concept IDs are used only for scoring and are never shown to the policy.

Let $f(c)$ be the fraction of training records containing concept c. The i-th of m bullets receives credit $\xi_i = 1 - f(c)$ if it is the first supported citation of c in the response. Unsupported bullets, repeated

citations, and non-clinical metadata such as age or sex receive zero credit. The evidence reward is

$$S = \frac{1}{\max(m, m_0)} \sum_{i=1}^{m} \xi_i, \qquad r_{\text{evid}} = 2S - 1 \quad (m > 0), \tag{5}$$

with $m_0 = 4$ and $r_{\text{evid}} = -1$ when no evidence is listed. Informativeness weighting keeps ubiquitous findings, such as a routine laboratory panel, from earning full credit; repeated and unsupported bullets lower the score through the denominator, and the floor m_0 removes the gain from listing fewer than m_0 findings. Because the term does not reward long lists, we report support precision together with the number of distinct supported findings.

Group relative policy optimization. Starting from supervised initialization, we sample a group of responses to each patient–task prompt and standardize their rewards within the group (Shao et al., 2024). Responses with higher relative reward are reinforced, and a reference-policy KL term regularizes the update. Two auxiliary losses complement the reward. The answer loss is the cross-entropy of the gold answer at the answer position; it supplies a correctness signal even when all responses in a group give the same answer, so that the answer reward does not vary within the group. The evidence-sufficiency loss is the cross-entropy of the gold answer given only the task prompt and the response's evidence bullets, with the reasoning prose removed and both EHR memories switched off, as in select-then-predict rationales (Lei et al., 2016; Jain et al., 2020); it asks whether the cited findings are informative about the outcome, and it is a loss rather than a reward because the reward form favored shorter evidence lists. The encoder and memory constructors remain frozen; the language model and cross-attention interfaces are updated. Reward coefficients, matcher rules, sampling settings, and the precise policy-update implementation are specified in Appendix A.7. To examine whether cited local representations affect the model's answer, we also define a matched regeneration intervention alongside support and coverage measurements (Appendix A.8).

4 Experiments

4.1 Data and tasks

Our main dataset is MIMIC-IV v2.2 (Johnson et al., 2023b), converted to the OMOP Common Data Model (CDM) (Hripcsak et al., 2015), with BHC targets from MIMIC-IV-Ext-BHC (Aali et al., 2025). It comprises the 128,350 patients with at least one admission that has a BHC. Patients are split 7:1:2 into training, validation, and test sets; every stage, from alignment to the downstream tasks, uses this split and trains only on training patients. We predict three common binary outcomes (Rajkomar et al., 2018): in-hospital mortality and long length of stay (more than 7 days) from the first 24 hours of an inpatient admission, and 30-day readmission from the full index admission. EHRSHOT (Wornow et al., 2023) serves as an external validation cohort without retraining. Appendix B gives the cohort flow, task definitions, and sample sizes.

4.2 Outcome-prediction comparison

Table 1 compares binary outcome prediction on the three internal tasks. Prediction serves as an outcome-level control for the proposed EHR–language interface.

Baselines. *Zero-shot LLMs* (Qwen3-1.7B, the backbone of GLoC-EHR, and Llama-3.3-70B-Instruct (Grattafiori et al., 2024)) read the records serialized as text (Hegselmann et al., 2026) and answer with the ClinicalBench instruction (Chen et al., 2024a), directly or after chain-of-thought (CoT) reasoning (Wei et al., 2022). *Task-specific* models are trained per task: count-based gradient-boosted trees (Ke et al., 2017; Chen & Guestrin, 2016), RETAIN (Choi et al., 2016), classifier heads on frozen text embeddings (Zhang et al., 2025a) of the serialized record and/or frozen EHR-BERT vectors (adapted from Hegselmann et al., 2026), fine-tuned EHR-BERT, and GLoC-EHR answering directly (ST). *Multi-task* models are trained once on all three tasks: EHR-BERT (MT) uses task-specific heads on a shared encoder, whereas GLoC-EHR (MT) and GLoC-EHR (Reasoning) receive the task only through the instruction, and the latter writes evidence-cited reasoning before its answer. Table 1 shows LightGBM and the frozen-vector baseline that uses all three feature sets; the others and per-task AUPRC are in Appendix C.

Protocol. Supervised models are selected by validation AUROC (macro AUROC for multi-task models) and retrained with seeds 42–46; we report the test mean and SD, averaging the three tasks within each seed for macro scores. Zero-shot LLMs are evaluated once. GLoC-EHR (Reasoning) is a single GRPO run, for which we report the mean and SD over five sampled chains per case, reflecting decoding rather than training variability. Direct GLoC-EHR variants score only the answer tokens, renormalized over Yes/No (Appendix A.6).

Results. Zero-shot LLMs fall far below every trained model, consistent with ClinicalBench (Chen et al., 2024a), where LLMs did not outperform traditional machine-learning models on mortality, length-of-stay, and readmission prediction; CoT prompting lowers their AUROC further. Among task-specific models, GLoC-EHR (ST) attains the highest macro AUROC and AUPRC (0.879 and 0.549), mainly through readmission, and stays close to fine-tuned EHR-BERT (ST) on mortality and long length of stay. Sharing one model across the three tasks lowers GLoC-EHR's scores slightly, yet GLoC-EHR (MT) still exceeds every baseline in macro AUROC, including EHR-BERT (ST), although its macro AUPRC falls below those of the EHR-BERT models and several tree-based baselines (Table 11). Two single-run controls support the interface (Appendix C.1): training it from random initialization without alignment lowers AUROC on all three tasks, and LoRA (Hu et al., 2022) fine-tuning of the backbone on the serialized record falls below GLoC-EHR (MT). GLoC-EHR (Reasoning) trails GLoC-EHR (MT) by 0.008 in macro AUROC, staying close to EHR-BERT (ST), and by 0.042 in macro AUPRC, with the AUPRC gap concentrated on mortality and long length of stay. Section 4.3 separately evaluates the support and informativeness of the generated evidence.

Table 1: Internal outcome prediction on MIMIC-IV (test): mean ± SD over five seeds, or over five sampled chains of one run for GLoC-EHR (Reasoning). Best in bold, second underlined. C: code counts; T_{4096}: text embedding; E_{CLS}: frozen EHR-BERT vector; / LGBM: LightGBM head.

Method	AUROC ↑			Macro (3 tasks) ↑	
	Mortality (24 h)	Long LOS (24 h)	Readmission (full)	AUROC	AUPRC
Zero-shot LLMs on serialized records					
Qwen3-1.7B (direct)	0.8140	0.6062	0.5183	0.6461	0.2240
Qwen3-1.7B (CoT)	0.5488	0.5363	0.5052	0.5301	0.1631
Llama-3.3-70B (direct)	0.8268	0.6391	0.5269	0.6643	0.2065
Llama-3.3-70B (CoT)	0.6780	0.5193	0.5034	0.5669	0.1644
Task-specific models					
LightGBM (counts)	0.9527 ± 0.0013	0.8624 ± 0.0011	0.7758 ± 0.0028	0.8637 ± 0.0013	0.5299 ± 0.0102
C+T_{4096}+E_{CLS} / LGBM	0.9449 ± 0.0027	0.8589 ± 0.0013	0.7625 ± 0.0068	0.8554 ± 0.0017	0.4940 ± 0.0042
EHR-BERT (ST)	**0.9607** ± 0.0018	<u>0.8701</u> ± 0.0008	0.7817 ± 0.0057	0.8708 ± 0.0020	<u>0.5380</u> ± 0.0034
GLoC-EHR (ST)	0.9561 ± 0.0017	**0.8740** ± 0.0018	**0.8075** ± 0.0015	**0.8792** ± 0.0008	**0.5494** ± 0.0066
Multi-task models					
EHR-BERT (MT)	<u>0.9575</u> ± 0.0017	0.8672 ± 0.0019	0.7865 ± 0.0026	0.8704 ± 0.0013	0.5312 ± 0.0062
GLoC-EHR (MT)	0.9545 ± 0.0013	0.8671 ± 0.0030	0.7976 ± 0.0052	<u>0.8731</u> ± 0.0030	0.5236 ± 0.0088
GLoC-EHR (Reasoning)	0.9459 ± 0.0010	0.8469 ± 0.0004	<u>0.8042</u> ± 0.0018	0.8656 ± 0.0008	0.4819 ± 0.0045

4.3 Evidence grounding

Metrics. We score the evidence bullets of every sampled chain with the matcher used by the reward (Section 3.4; Appendix A.7). *Support precision* is the informativeness-weighted fraction of bullets that cite a distinct supported finding, $\frac{1}{m}\sum_i \xi_i$. The *unsupported rate* is the fraction of bullets that match no event in the input window, and *distinct supported findings* counts the concepts credited in a response. *Bullets-only AUROC* measures how informative the cited evidence is to the model that wrote it: the model answers from the task prompt and its evidence bullets alone, with both EHR memories switched off. Test-set values average the five chains of each case. Over GRPO training, the unsupported rate falls for GLoC-EHR and the global-route-only variant but stays near its SFT level without the evidence terms (Appendix C.3).

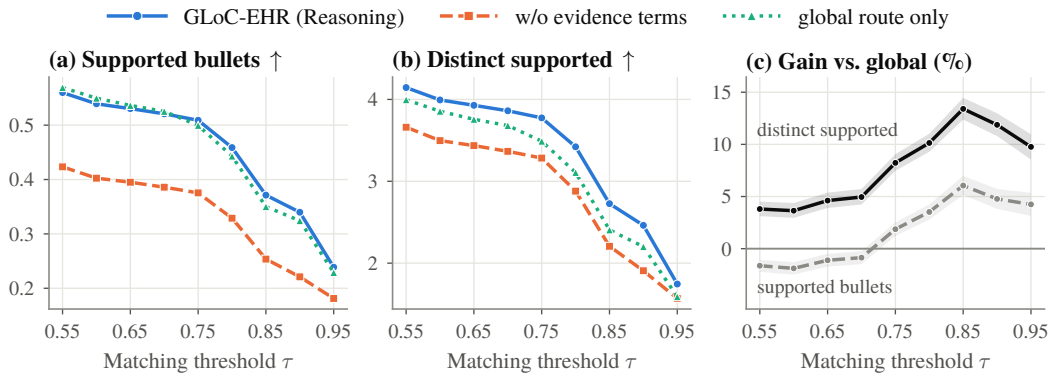

Figure 2: Evidence of the final checkpoints on MIMIC-IV test chains (one per case) as matching becomes stricter: every match must reach an IDF-weighted overlap of at least τ (τ=0.55 is the reward's matcher). (a) Share of supported bullets, (b) distinct supported findings per response, and (c) relative gain of GLoC-EHR (Reasoning) over the global-route-only variant, with 95% paired case-bootstrap intervals.

Held-out evidence. On the test chains (Table 2; brackets: 95% paired case-bootstrap intervals), GRPO lowers the unsupported rate from 0.58 after rationale SFT to 0.44 while writing fewer bullets and citing more distinct supported findings. Without the evidence terms, the unsupported rate stays at the SFT level, 0.137 [0.133, 0.140] above that of GLoC-EHR, and fewer distinct supported findings are cited. The global-route-only variant writes fewer bullets with a slightly higher support precision but cites 0.15 [0.12, 0.18] fewer distinct supported findings per response. Figure 2 re-scores these chains under stricter matching, which requires closer lexical agreement between a bullet and the event it cites. From $\tau = 0.75$ onward GLoC-EHR has a higher share of supported bullets than the global-route-only variant, and at $\tau = 0.85$ it cites 13.4% [12.4, 14.5] more distinct supported findings. The local route thus yields citations that match their events more closely. Zero-shot LLMs that read the serialized record can reach high support by quoting it but answer near chance (Appendix C.2).

Cited-slot masking. To test whether the local route supplies the cited evidence, we mask the local-memory slots of the concepts that a greedy chain cites, regenerate the chain, and compare with masking as many uncited slots of similar informativeness (2,062 test cases whose chains cite local-memory concepts; Appendix A.8). The regenerated rationale drops the masked concepts more often under the cited mask (persistence -0.017 [-0.028, -0.007]), but the answer changes about as much as under the control mask (macro AUROC -0.003 [-0.009, 0.002]; Table 8). The local route thus shapes which findings are cited, with no detectable effect on the answer, consistent with the similar AUROC of the global-route-only variant (Section 4.4).

4.4 Ablations

Table 2 compares GLoC-EHR (Reasoning) with variants that remove the evidence terms (the evidence reward and the evidence-sufficiency loss), use the global route only, or stop after rationale SFT. Stopping after rationale SFT lowers macro AUROC from 0.866 to 0.658, and the SFT model answers almost deterministically (99.8% of its answer probabilities are below 0.02 or above 0.98, vs. 31% after GRPO). Neither removing the evidence terms nor dropping the local route detectably changes macro AUROC (paired bootstrap intervals include zero), but the bullets alone predict the outcome slightly better with the global route only and worse without the evidence terms, a variant that, like rationale SFT, is never trained to answer from its bullets.

4.5 Transferability

Table 3 applies the internally selected checkpoints to EHRSHOT without retraining, after removing events with codes outside the training vocabulary; an EHR-BERT pretrained on code IDs alone isolates the contribution of description embeddings. All three main models lose about 0.1 macro

Table 2: Ablations of GLoC-EHR (Reasoning) on the MIMIC-IV test set. The first row is the mean ± SD over five sampled chains; the other rows score one chain per case. Evidence is matched with the reward's matcher (τ=0.55) and, for distinct supported findings, also at τ=0.85 (Figure 2). Bullets-only AUROC: answers from each model's own bullets with the EHR memories off.

	Macro	Macro	Unsupported	Distinct supported ↑		Bullets-only
Variant	AUROC ↑	AUPRC ↑	rate ↓	τ=0.55	τ=0.85	AUROC ↑
GLoC-EHR (Reasoning)	0.8656 ± 0.0008	0.4819 ± 0.0045	0.440 ± 0.001	4.16 ± 0.01	2.73 ± 0.01	0.758 ± 0.005
one chain	0.8661	0.4810	0.440	4.15	2.73	0.755
w/o evidence terms	0.8665	0.4896	0.577	3.66	2.21	0.707
global route only	0.8641	0.4783	0.431	4.00	2.41	0.767
rationale SFT only	0.6578	0.2013	0.582	3.92	2.34	0.690

Table 3: External validation on EHRSHOT without retraining, with out-of-vocabulary events removed: mean ± SD over five seeds. GLoC-EHR (Reasoning): one sampled chain per case (internal five-chain SD 0.0008), excluding the 2.6% without an answer (0.767 macro AUROC if scored as 0.5). Code IDs only: EHR-BERT pretrained without description embeddings. Best in bold, second underlined.

	AUROC ↑			Macro ↑	
Method	Mortality	Long LOS	Readmission	AUROC	AUPRC
EHR-BERT (MT)	**0.9311** ± 0.0055	0.7531 ± 0.0056	<u>0.6314</u> ± 0.0153	<u>0.7719</u> ± 0.0041	**0.4414** ± 0.0040
code IDs only	<u>0.9303</u> ± 0.0037	0.7525 ± 0.0063	0.5568 ± 0.0328	0.7465 ± 0.0130	0.4167 ± 0.0148
GLoC-EHR (MT)	0.9287 ± 0.0023	<u>0.7584</u> ± 0.0043	**0.6396** ± 0.0131	**0.7756** ± 0.0039	<u>0.4331</u> ± 0.0060
GLoC-EHR (Reasoning)	0.9244	**0.7585**	0.6288	0.7706	0.4222

AUROC relative to their internal test scores, so GLoC-EHR (MT) remains on par with EHR-BERT (MT) and the reasoning variant stays close to both, while EHR-BERT (MT) keeps the highest macro AUPRC. Without description embeddings, EHR-BERT falls further on readmission, the hardest task for every model, which accounts for nearly all of its lower macro AUROC. Asked, without further training, whether a patient's next serum sodium or platelet count will be low, GLoC-EHR (Reasoning) reaches AUROC 0.657 and 0.629 on held-out MIMIC-IV admissions, compared with 0.556 and 0.525 for zero-shot Qwen3-1.7B reading the serialized record (Appendix C.4).

5 Conclusion

GLoC-EHR gives a language model a fixed-size global memory of an EHR trajectory and a bounded local memory of selected events, aligns both with clinical text, and trains evidence-cited reasoning with rewards for correct answers and record-supported evidence. Answering directly, it matches or exceeds fine-tuned EHR encoders in macro AUROC on three MIMIC-IV tasks, far above zero-shot LLMs. Citing evidence first costs little discrimination after GRPO (0.008 macro AUROC below the direct multi-task model, with a larger AUPRC gap), whereas rationale SFT alone reaches only 0.658. At a similar macro AUROC, the evidence terms lower the share of unsupported evidence from 0.58 to 0.44, and the local route adds distinct supported findings and shapes which findings are cited, with no detectable effect on the answer. Without retraining, GLoC-EHR transfers to EHRSHOT on par with EHR-BERT and answers unseen laboratory questions better than its zero-shot backbone.

Limitations and future work. Our evidence matching is lexical: it checks that each cited finding appears in the patient's record, but not whether the finding is clinically correct or whether the model relied on it to answer. Future work includes semantic matching of paraphrased findings, for both the reward and the evaluation, and clinician review of the cited evidence. GLoC-EHR (Reasoning) comes from a single training run on data from one health system; additional runs and cohorts would further test the stability of these effects. We also plan to recalibrate its probabilities for each new cohort: on EHRSHOT, the mean predicted probability of a long stay was 0.57 against a prevalence of 0.31.

AI USE STATEMENT

In this work, we used generative AI tools to assist with translation and to provide feedback on the design of the methodology and experiments. Apart from the teacher rationales used as SFT targets, which are part of our method (Section 3.4), we have not used generative AI tools to generate synthetic data sets, and tasks involving mathematical claims, proofs, or qualitative and thematic data analysis are not applicable to this work. Additionally, we used generative AI tools to draft parts of the paper and edit it for readability, to create and edit analysis and plotting code and figures, and to search for and identify relevant literature. We have reviewed all AI-assisted work: the authors wrote the main content of the paper and revised all AI-drafted text, read and verified every cited work, and checked all AI-assisted code and results. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work uses only de-identified EHR data. MIMIC-IV and MIMIC-IV-Ext-BHC were accessed through PhysioNet under its credentialed data use agreement, and EHRSHOT under its research data use agreement; no new data were collected, and no attempt was made to re-identify patients. All language models, including the teacher and the zero-shot baselines, were run on the authors' servers, so no patient data were sent to external services. GLoC-EHR is a research model and is not intended for clinical decision-making. Its evidence is checked for lexical support in the record, not for clinical correctness; its probabilities are not calibrated on the external cohort (Guo et al., 2017; Van Calster et al., 2019); and it was trained on data from a single health system, whose patient population and documentation practices it may reflect. Because clinical models can degrade under dataset and temporal shift (Finlayson et al., 2021; Guo et al., 2023), clinical use would require prospective validation, recalibration, and evaluation across patient subgroups.

REPRODUCIBILITY STATEMENT

Data. MIMIC-IV v2.2, MIMIC-IV-Ext-BHC, and EHRSHOT are available to credentialed researchers under their data use agreements. Appendix B describes the cohort construction, task definitions, patient-level splits, and sample sizes.

Method. Section 3 describes the architecture and training stages, and Appendix A specifies the encoder, the interface, the alignment schedule, rationale SFT, the GRPO settings, the reward coefficients, and the full evidence matcher. GRPO training of GLoC-EHR (Reasoning) took about 55 hours on four NVIDIA H200 GPUs.

Evaluation. Section 4 and Appendix C specify the metrics, seeds, sampling settings, bootstrap procedures, and the handling of chains without an answer. Main results are reported as means and standard deviations over five seeds or five sampled chains.

Code and model weights. We will release the code upon acceptance. Because the models were trained on credentialed data, we will release the model weights upon acceptance through PhysioNet under the same credentialed-access terms as MIMIC-IV.

REFERENCES

Asad Aali, Dave Van Veen, Yamin Ishraq Arefeen, Jason Hom, Christian Bluethgen, Eduardo Pontes Reis, Sergios Gatidis, Namuun Clifford, Joseph Daws, Arash S. Tehrani, Jangwon Kim, and Akshay S. Chaudhari. A dataset and benchmark for hospital course summarization with adapted large language models. *Journal of the American Medical Informatics Association*, 32(3):470–479, 2025. doi: 10.1093/jamia/ocae312.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob L. Menick, Sebastian Borgeaud, Andy Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikołaj Bińkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karén Simonyan. Flamingo: A visual language model for few-shot learning. In *Advances in Neural Information Processing Systems*, volume 35, pp. 23716–23736, 2022. doi: 10.52202/068431-1723.

Pepa Atanasova, Oana-Maria Camburu, Christina Lioma, Thomas Lukasiewicz, Jakob Grue Simonsen, and Isabelle Augenstein. Faithfulness tests for natural language explanations. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 283–294, 2023. doi: 10.18653/v1/2023.acl-short.25. URL https://aclanthology.org/2023.acl-short.25/.

Parishad BehnamGhader, Vaibhav Adlakha, Marius Mosbach, Dzmitry Bahdanau, Nicolas Chapados, and Siva Reddy. LLM2Vec: Large language models are secretly powerful text encoders. In *First Conference on Language Modeling*, 2024.

Anastasiya Belyaeva, Justin Cosentino, Farhad Hormozdiari, et al. Multimodal LLMs for health grounded in individual-specific data. In *Machine Learning for Multimodal Healthcare Data (ML4MHD 2023)*, Lecture Notes in Computer Science. Springer, 2023. doi: 10.1007/978-3-031-47679-2_7.

Ofir Ben Shoham and Nadav Rappoport. CPLLM: Clinical prediction with large language models. *PLOS Digital Health*, 3(12):e0000680, 2024. doi: 10.1371/journal.pdig.0000680.

Canyu Chen, Jian Yu, Shan Chen, Che Liu, Zhongwei Wan, Danielle Bitterman, Fei Wang, and Kai Shu. ClinicalBench: Can LLMs beat traditional ML models in clinical prediction? *arXiv preprint arXiv:2411.06469*, 2024a.

Junying Chen, Zhenyang Cai, Ke Ji, et al. HuatuoGPT-o1, towards medical complex reasoning with LLMs. *arXiv preprint arXiv:2412.18925*, 2024b.

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

Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pp. 1724–1734, 2014. doi: 10.3115/v1/D14-1179.

Edward Choi, Mohammad Taha Bahadori, Jimeng Sun, Joshua Kulas, Andy Schuetz, and Walter Stewart. RETAIN: An interpretable predictive model for healthcare using reverse time attention mechanism. In *Advances in Neural Information Processing Systems*, volume 29, 2016. URL https://proceedings.neurips.cc/paper/2016/hash/231141b34c82aa95e48810a9d1b33a79-Abstract.html.

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. doi: 10.18653/v1/N19-1423.

Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C. Wallace. ERASER: A benchmark to evaluate rationalized NLP models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4443–4458, 2020. doi: 10.18653/v1/2020.acl-main.408. URL https://aclanthology.org/2020.acl-main.408/.

Bradley Efron and Robert J. Tibshirani. *An Introduction to the Bootstrap*. Chapman & Hall/CRC, 1993.

Kawin Ethayarajh. How contextual are contextualized word representations? Comparing the geometry of BERT, ELMo, and GPT-2 embeddings. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 55–65, 2019.

Samuel G. Finlayson, Adarsh Subbaswamy, Karandeep Singh, John Bowers, Annabel Kupke, Jonathan Zittrain, Isaac S. Kohane, and Suchi Saria. The clinician and dataset shift in artificial intelligence. *New England Journal of Medicine*, 385(3):283–286, 2021. doi: 10.1056/NEJMc2104626.

Jifan Gao, Mahmudur Rahman, John Caskey, Madeline Oguss, Ann O'Rourke, Randall Brown, Anne Stey, Anoop Mayampurath, Matthew M. Churpek, Guanhua Chen, and Majid Afshar. MoMA: A mixture-of-multimodal-agents architecture for enhancing clinical prediction modelling. *npj Digital Medicine*, 9:46, 2026. doi: 10.1038/s41746-025-02219-4.

Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen. Enabling large language models to generate text with citations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 2023. doi: 10.18653/v1/2023.emnlp-main.398.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, et al. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1321–1330, 2017.

Daya Guo, Dejian Yang, Haowei Zhang, et al. DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. *Nature*, 645(8081):633–638, 2025. doi: 10.1038/s41586-025-09422-z.

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*, 13:3767, 2023. doi: 10.1038/s41598-023-30820-8.

Stefan Hegselmann, Alejandro Buendia, Hunter Lang, Monica Agrawal, Xiaoyi Jiang, and David Sontag. TabLLM: Few-shot classification of tabular data with large language models. In *Proceedings of the 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, 2023.

Stefan Hegselmann, Georg von Arnim, Tillmann Rheude, Noel Kronenberg, David Sontag, Gerhard Hindricks, Roland Eils, and Benjamin Wild. Large language models are powerful electronic health record encoders. *npj Digital Medicine*, 9:530, 2026. doi: 10.1038/s41746-026-02915-9.

Sara Hooker, Dumitru Erhan, Pieter-Jan Kindermans, and Been Kim. A benchmark for interpretability methods in deep neural networks. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

George Hripcsak, Jon D. Duke, Nigam H. Shah, et al. Observational health data sciences and informatics (OHDSI): Opportunities for observational researchers. *Studies in Health Technology and Informatics*, 216:574–578, 2015.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

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, 2021. doi: 10.18653/v1/2021.findings-emnlp.23. URL https://aclanthology.org/2021.findings-emnlp.23/.

Alon Jacovi and Yoav Goldberg. Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4198–4205, 2020. doi: 10.18653/v1/2020.acl-main.386. URL https://aclanthology.org/2020.acl-main.386/.

Andrew Jaegle, Felix Gimeno, Andy Brock, Oriol Vinyals, Andrew Zisserman, and João Carreira. Perceiver: General perception with iterative attention. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 4651–4664. PMLR, 2021. URL https://proceedings.mlr.press/v139/jaegle21a.html.

Sarthak Jain, Sarah Wiegreffe, Yuval Pinter, and Byron C. Wallace. Learning to faithfully rationalize by construction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4459–4473, 2020. doi: 10.18653/v1/2020.acl-main.409.

Lavender Yao Jiang, Xujin Chris Liu, Nima Pour Nejatian, et al. Health system-scale language models are all-purpose prediction engines. *Nature*, 619(7969):357–362, 2023. doi: 10.1038/s41586-023-06160-y.

Ming Jin, Shiyu Wang, Lintao Ma, et al. Time-LLM: Time series forecasting by reprogramming large language models. In *International Conference on Learning Representations*, 2024.

Alistair Johnson, Tom Pollard, Steven Horng, Leo Anthony Celi, and Roger Mark. MIMIC-IV-Note: Deidentified free-text clinical notes (version 2.2). PhysioNet, 2023a.

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, 2023b. doi: 10.1038/s41597-022-01899-x.

Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. LightGBM: A highly efficient gradient boosting decision tree. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Junmo Kim, Namkyeong Lee, Jiwon Kim, and Kwangsoo Kim. MedRep: Medical concept representations for general electronic health record foundation models. *Journal of the American Medical Informatics Association*, 33(6):1089–1099, 2026. doi: 10.1093/jamia/ocag032.

Matthias Kirchler, Matteo Ferro, Veronica Lorenzini, Robin P van de Water, FinnGen Ganna Andrea 3, Christoph Lippert, and Andrea Ganna. Large language models improve transferability of electronic health record-based predictions across countries and coding systems. *npj Digital Medicine*, 9(1):177, 2026.

Zeljko Kraljevic, Dan Bean, Anthony Shek, et al. Foresight—a generative pretrained transformer for modelling of patient timelines using electronic health records: A retrospective modelling study. *The Lancet Digital Health*, 6(4):e281–e290, 2024.

Tamera Lanham, Anna Chen, Ansh Radhakrishnan, Benoit Steiner, Carson Denison, et al. Measuring faithfulness in chain-of-thought reasoning. *arXiv preprint arXiv:2307.13702*, 2023. doi: 10.48550/arXiv.2307.13702.

Simon A. Lee, Sujay Jain, Alex Chen, Kyoka Ono, Arabdha Biswas, Ákos Rudas, Jennifer Fang, and Jeffrey N. Chiang. Clinical decision support using pseudo-notes from multiple streams of EHR data. *npj Digital Medicine*, 8:394, 2025. doi: 10.1038/s41746-025-01777-x.

Tao Lei, Regina Barzilay, and Tommi Jaakkola. Rationalizing neural predictions. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 107–117, 2016. doi: 10.18653/v1/D16-1011.

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, 2020a. doi: 10.18653/v1/2020.emnlp-main.733.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 19730–19742. PMLR, 2023. URL https://proceedings.mlr.press/v202/li23q.html.

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

Yusheng Liao, Chaoyi Wu, Junwei Liu, Shuyang Jiang, Pengcheng Qiu, Haowen Wang, Yun Yue, Shuai Zhen, Jian Wang, Qianrui Fan, Jinjie Gu, Ya Zhang, Yanfeng Wang, Yu Wang, and Weidi Xie. EHR-R1: A reasoning-enhanced foundational language model for electronic health record analysis. *arXiv preprint arXiv:2510.25628*, 2025. doi: 10.48550/arXiv.2510.25628.

Jiacheng Lin, Zhenbang Wu, and Jimeng Sun. Training LLMs for EHR-based reasoning tasks via reinforcement learning. *arXiv preprint arXiv:2505.24105*, 2025. doi: 10.48550/arXiv.2505.24105.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Wei Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 2023. doi: 10.18653/v1/2023.emnlp-main.741.

Jiaqi Mu and Pramod Viswanath. All-but-the-top: Simple and effective postprocessing for word representations. In *International Conference on Learning Representations*, 2018.

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. URL https://proceedings.mlr.press/v158/pang21a.html.

Qwen Team. Qwen3.5: Towards native multimodal agents. Qwen Blog, 2026. URL https://qwen.ai/blog?id=qwen3.5.

Alvin Rajkomar, Eyal Oren, Kai Chen, Andrew M. Dai, Nissan Hajaj, et al. Scalable and accurate deep learning with electronic health records. *npj Digital Medicine*, 1:18, 2018. doi: 10.1038/s41746-018-0029-1.

Hannah Rashkin, Vitaly Nikolaev, Matthew Lamm, et al. Measuring attribution in natural language generation models. *Computational Linguistics*, 49(4):777–840, 2023.

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*, 4:86, 2021. doi: 10.1038/s41746-021-00455-y.

Daniel Reichert, David Kaufman, Benjamin Bloxham, Herbert Chase, and Noémie Elhadad. Cognitive analysis of the summarization of longitudinal patient records. In *AMIA Annual Symposium Proceedings*, volume 2010, pp. 667–671, 2010. URL https://pubmed.ncbi.nlm.nih.gov/21347062/.

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.

Takaya Saito and Marc Rehmsmeier. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLoS ONE*, 10(3):e0118432, 2015. doi: 10.1371/journal.pone.0118432.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024. doi: 10.48550/arXiv.2402.03300.

Karan Singhal, Shekoofeh Azizi, Tao Tu, et al. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180, 2023. doi: 10.1038/s41586-023-06291-2.

Ethan Steinberg, Ken Jung, Jason A. Fries, Conor K. Corbin, Stephen R. Pfohl, and Nigam H. Shah. Language models are an effective 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 Alan Fries, Yizhe Xu, and Nigam Shah. MOTOR: A time-to-event foundation model for structured medical records. In *International Conference on Learning Representations*, 2024.

Jianlin Su, Jiarun Cao, Weijie Liu, and Yangyiwen Ou. Whitening sentence representations for better semantics and faster retrieval. *arXiv preprint arXiv:2103.15316*, 2021.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel R. Bowman. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Ben Van Calster, David J. McLernon, Maarten van Smeden, Laure Wynants, and Ewout W. Steyerberg. Calibration: The Achilles heel of predictive analytics. *BMC Medicine*, 17:230, 2019. doi: 10.1186/s12916-019-1466-7.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc V. Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

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, 2023. doi: 10.52202/075280-2933.

Michael Wornow, Suhana Bedi, Miguel Angel Fuentes Hernandez, Ethan Steinberg, Jason Alan Fries, Christopher Ré, Sanmi Koyejo, and Nigam Shah. Context clues: Evaluating long context models for clinical prediction tasks on EHR data. In *International Conference on Learning Representations*, 2025.

Zhenbang Wu, Anant Dadu, Michael A. Nalls, Faraz Faghri, and Jimeng Sun. Instruction tuning large language models to understand electronic health records. In *Advances in Neural Information Processing Systems*, volume 37, pp. 54772–54786, 2024. doi: 10.52202/079017-1737.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Yanzhao Zhang, Mingxin Li, Dingkun Long, et al. Qwen3 Embedding: Advancing text embedding and reranking through foundation models. *arXiv preprint arXiv:2506.05176*, 2025a.

Zhi Zhang, Srishti Yadav, Fengze Han, and Ekaterina Shutova. Cross-modal information flow in multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19781–19791, 2025b. doi: 10.1109/CVPR52734.2025.01842.

Yinghao Zhu et al. Prompting large language models for zero-shot clinical prediction with structured longitudinal electronic health record data. *arXiv preprint arXiv:2402.01713*, 2024.

A IMPLEMENTATION AND EVALUATION DETAILS

This appendix records the configuration used in the current implementation. The architectural design and learning objectives are described in Section 3; numerical settings and evaluation procedures are collected here for reproducibility.

A.1 STRUCTURED INPUT AND ENCODER PRETRAINING

Input representation. Each example contains at most $T_{\max} = 2048$ structured events after applying its task-specific observation cutoff and truncation rule. Each event contains a medical concept together with parallel age, date, and event-type attributes. The date field is the rank of the event date within the patient's record, rather than an absolute calendar date. The observed sequence is therefore not the patient's unrestricted longitudinal history. The vocabulary contains 20,561 tokens constructed from the OMOP domains represented in our MIMIC-IV cohort (Johnson et al., 2023b). Measurements carry qualitative or quantile value suffixes, and drug concepts carry exposure-duration suffixes. Cohort construction, task definitions, and the patient-level split shared by all training stages are described in Appendix B.

Description embeddings. For each vocabulary concept c, a short typed description $d(c)$ is formed, such as "condition: Enteropathic arthritis" or "lab: Sodium [Moles/volume] in Serum or Plasma, low". We feed at most 128 description tokens to Qwen3-1.7B (Yang et al., 2025) and mean-pool its final hidden states to obtain a 2,048-dimensional vector $\mathbf{b}_c$. We center, whiten, and ℓ_2-normalize the description vectors:

$$\mathbf{z}_c = \mathrm{norm}_2\Big[(\mathbf{\Sigma} + \epsilon\mathbf{I})^{-1/2}(\mathbf{b}_c - \boldsymbol{\mu})\Big]. \tag{6}$$

The mean and covariance are estimated once over the vocabulary descriptions, which uses no patient data. Whitening follows related normalization for sentence representations (Huang et al., 2021); the comparison that motivated it is reported in Appendix A.2. The description table $\{\mathbf{z}_c\}$ remains frozen during EHR pretraining.

Event embeddings and encoder. The EHR encoder is an eight-layer, 12-head bidirectional Transformer with hidden size $d_E = 768$ and no absolute-position embedding. Let $\mathbf{v}_c$ be a learned code-ID embedding, and let $\mathbf{a}_t$, $\boldsymbol{\delta}_t$, and $\boldsymbol{\tau}_t$ denote learned age, date, and type embeddings. The input at position t is

$$\widetilde{\mathbf{d}}_{c_t} = \mathrm{LN}_d(\mathbf{W}_d\mathbf{z}_{c_t} + \mathbf{b}_d), \tag{7}$$

$$\mathbf{u}_t = \mathrm{LN}\Big(\widetilde{\mathbf{d}}_{c_t} + 0.3\,\mathrm{LN}_c(\mathbf{v}_{c_t}) + \mathbf{a}_t + \boldsymbol{\delta}_t + \boldsymbol{\tau}_t\Big). \tag{8}$$

The description is thus the base representation and the learned code ID is a scaled residual. At eligible non-MLM context positions, the input code ID is replaced by `[UNK]` with probability 0.5, while retaining the correct description-table index. At masked positions, both the code and description indices point to `[MASK]`, preventing target-description leakage.

Sampling and pretraining objective. We sample one view per patient per epoch: a single visit with probability 0.6, the most recent trajectory with probability 0.3, or a random contiguous visit window with probability 0.1. Each view is capped at 2,048 events. We mask 30% of eligible clinical-concept positions and replace every selected position with `[MASK]`, without an 80/10/10 corruption rule. Besides masked concept prediction, each masked contextual state is contrasted against the projected description of its true concept. For M masked positions in a batch, let $\mathbf{h}_i$ be the normalized encoder state and $\widetilde{\mathbf{z}}_i$ the normalized projected target description. The one-directional in-batch contrastive objective is

$$\mathcal{L}_{\mathrm{con}} = -\frac{1}{M}\sum_{i=1}^{M}\log\frac{\exp(\mathbf{h}_i^\top\widetilde{\mathbf{z}}_i/\tau)}{\sum_{j=1}^{M}\exp(\mathbf{h}_i^\top\widetilde{\mathbf{z}}_j/\tau)}, \qquad \tau = 0.1. \tag{9}$$

We optimize

$$\mathcal{L}_{\mathrm{EHR}} = \mathcal{L}_{\mathrm{MLM}} + 0.1\mathcal{L}_{\mathrm{con}}. \tag{10}$$

The resulting contextual sequence is $\mathbf{H} = f_{\mathrm{EHR}}(\mathbf{x}) \in \mathbb{R}^{T\times d_E}$.

A.2 BACKBONE DESCRIPTION EMBEDDINGS

Description-based concept representations improve the transfer of EHR models across vocabularies and institutions, using LLM-generated definitions combined with ontology structure (Kim et al., 2026) or frozen embeddings from a dedicated text-embedding model (Kirchler et al., 2026). GLoC-EHR later maps encoder states into Qwen3-1.7B, so we compute the description embeddings with that backbone. The encoder and the language model then share one semantic reference rather than relying on a third embedding space. Because Qwen3-1.7B is a generative model rather than an embedding model, and decoder-only LLMs typically need adaptation to serve as text encoders (BehnamGhader et al., 2024), we compared several ways of turning a description into a vector before fixing Equation 6 (Table 4).

Setup. All 20,561 vocabulary tokens receive a typed short description; special and metadata tokens receive brief labels such as "masked EHR code". We mean-pool either the input-embedding rows of the description tokens (*input-token mean*) or the final-layer hidden states of a forward pass over at most 128 tokens (*final hidden state*). We then apply no post-processing, removal of the first principal component (Mu & Viswanath, 2018) followed by ℓ_2 normalization, or centering, whitening, and ℓ_2 normalization (Su et al., 2021). Isotropy is summarized by the entropy effective rank of the covariance spectrum, the variance share of the first principal component, and the mean pairwise cosine. Qualifier separation uses the 6,952 tokens that attach a qualifier to a base concept: 1,757 measurement tokens with value bins (e.g., low, high, quartile) and 5,195 drug tokens with short or long exposure. For each, we find its nearest neighbour by cosine among all vocabulary tokens and record whether it is another variant of the same concept or a different concept with the same qualifier.

Findings. Raw representations are strongly anisotropic, as observed for other pretrained language models (Ethayarajh, 2019; Li et al., 2020a). Input-token means occupy an effective rank of 295 of 2,048 dimensions, and final hidden states collapse further to 70, with a mean pairwise cosine of 0.89. In the raw input space, random qualified concepts have a cosine of 0.61, and 48% of qualified tokens have a different concept with the same qualifier as their nearest neighbour. Whitening restores isotropy (effective rank 1,793 for input-token means and 1,908 for final hidden states) and drives the cosine between unrelated concepts to zero while keeping variants of the same concept close (0.94). Removing only the first principal component leaves the space anisotropic (effective rank 116, random-pair cosine 0.36). Whitened final hidden states give the clearest concept structure: 98.4% of qualified tokens that have a sibling variant find it as their nearest neighbour, compared with 91.1% for whitened and 75.3% for raw input-token means, and the share of same-qualifier neighbours is lowest (31.9%). The forward pass thus composes a multi-token description into a concept-level vector that qualifier words perturb less than a token average does.

The same property means that whitened description embeddings mainly encode concept identity. For "lab: Sodium [Moles/volume] in Serum or Plasma, low", the three nearest neighbours are the high, very low, and unqualified variants of the same test. Value and exposure distinctions are therefore carried chiefly by the learned code-ID residual and the surrounding events (Equation 8), which the contextual encoder combines during pretraining.

A.3 LANGUAGE INTERFACE CONFIGURATION

Backbone and global memory. The language backbone is Qwen3-1.7B, with 28 decoder blocks and hidden size $d_L = 2048$ (Yang et al., 2025). Both routes use the same contextual EHR sequence $\mathbf{H}$. The global resampler has four layers and 16 attention heads, initialized from $K_g = 64$ learned queries. At each layer, these queries cross-attend to all valid encoder states and pass through a feed-forward block. The resulting 64×2048 memory replaces the designated `<EMR>` placeholder with continuous embeddings and also supplies the keys and values of global cross-attention. The inserted embeddings may attend bidirectionally to one another, while ordinary text attention remains causal.

Local projection and selector. The local projection and salience scorer are

$$\mathbf{c}_t = \mathrm{LN}_o(\mathbf{W}_\ell \,\mathrm{LN}_i(\mathbf{H}_t) + \mathbf{b}_\ell)\,, \tag{11}$$

$$s_t = \mathbf{w}_2^\top \,\mathrm{GELU}(\mathbf{W}_1 \,\mathrm{LN}_s(\mathbf{c}_t) + \mathbf{b}_1) + b_2. \tag{12}$$

Table 4: Geometry of Qwen3-1.7B description embeddings for the 20,561 vocabulary concepts (typed short descriptions, mean pooling, 2,048 dimensions). Isotropy: effective rank r_{eff}, variance share of the first principal component, and mean pairwise cosine $\bar{c}$, which is zero by construction after centring. Qualified concepts: the 6,952 concepts with a value or exposure qualifier; $\bar{c}_{\text{same}}$ and $\bar{c}_{\text{rand}}$ are mean cosines between variants of the same concept and between random qualified concepts; NN_c is the share of nearest neighbours that are a variant of the same concept, among the 68.5% of qualified concepts that have one; NN_q is the share whose nearest neighbour is a different concept with the same qualifier. Without the type prefix, the used embeddings give an NN_c of 97.1%, and float32 inference changes NN_c by less than 0.2 points.

		Isotropy			Qualified concepts			
Representation	Post-processing	$r_{\text{eff}} \uparrow$	PC_1 % $\downarrow$	$\bar{c} \downarrow$	$\bar{c}_{\text{same}}$	$\bar{c}_{\text{rand}} \downarrow$	NN_c % $\uparrow$	NN_q % $\downarrow$
Input-token mean	none	295	15.3	0.29	0.97	0.61	75.3	48.4
Input-token mean	whiten + ℓ_2	1,793	0.2	0.00	0.94	0.00	91.1	36.4
Final hidden state	none	70	21.3	0.89	—	—	—	—
Final hidden state	top-PC removal + ℓ_2	116	23.4	0.00	0.99	0.36	87.0	39.5
Final hidden state	whiten + ℓ_2 (used)	1,908	0.3	0.00	0.94	0.00	98.4	31.9

Table 5: Optimization stages. Global and local cross-attention include their scalar gates.

Stage	Supervision	Trainable components
Semantic encoder	masked concepts	EHR encoder and input projections
Global alignment	BHC generation	resampler, global cross-attention
Local alignment	concept descriptions + BHC replay	projection, selector, local cross-attention; local cross-attention only after hard selection
Direct prediction	binary outcome	global and local cross-attention
Rationale SFT, soft	teacher rationale + answer	base LLM, resampler, global and local cross-attention, local projection, selector
Rationale SFT, hard	teacher rationale + answer	base LLM, resampler, global and local cross-attention
GRPO	rule-based reward + answer loss	base LLM, global and local cross-attention

The selector MLP has hidden width 1,024. Hard selection uses $K_\ell = 256$ during concept alignment and $K_\ell = 64$ in every downstream stage; when fewer eligible events are available, all candidates are retained. Local vectors are available through local cross-attention only and are not inserted as ordinary LLM tokens. The selector does not depend on the downstream instruction q, so it produces one ranking shared across tasks. It is learned during concept alignment, adapted during the soft phase of rationale SFT (Appendix A.6), and fixed thereafter.

Cross-attention placement and gates. Each route-specific cross-attention module is applied after its corresponding original decoder block. Using zero-indexed blocks, the global modules are attached at $\mathcal{S}_{\text{glob}} = \{12, 14, 16, 18, 20, 22\}$ and the local modules at $\mathcal{S}_{\text{loc}} = \{23, 25, 27\}$. Each module has one learned scalar gate γ_l^r, initialized to -2, so its initial residual weight is $\sigma(\gamma_l^r) \approx 0.119$. The global interface is therefore available before the final local injections. These depths are an empirical configuration choice.

A.4 Alignment objectives and training schedule

Table 5 specifies trainable components throughout the pipeline. The pretrained EHR encoder and base LLM are frozen in both alignment stages. Direct prediction and rationale SFT are separate downstream branches initialized from the aligned interface.

Global trajectory alignment. The model receives a structured hospital trajectory and generates its corresponding brief hospital course (BHC). We disable the local interface and use the response-only next-token objective

$$\mathcal{L}_{\text{glob}} = -\sum_j \log p_\theta(b_j \mid b_{<j}, q_{\text{BHC}}, \mathbf{M}^{\text{glob}}), \tag{13}$$

updating only the resampler and global cross-attention modules. Discharge text provides an alignment target, not an input to downstream prediction.

Local concept alignment. We replace 5% of eligible concepts by `[MASK]`, up to 32 positions per example, and ask the model to generate descriptions for up to eight masked concepts. Concept-description examples and BHC replay are mixed at rates 0.4 and 0.6, respectively. The EHR encoder, global route, and base LLM remain frozen; the local projection, selector, and local cross-attention are initially trainable. After the forced-selection phase, masked positions are excluded from the selector's candidate set so the local route cannot simply forward the representation at a masked slot.

The three phases are indexed by optimizer updates u:

$$\underbrace{\text{forced masked positions}}_{0\le u<1000} \longrightarrow \underbrace{\text{soft selection, } \tau_s : 1.0 \to 0.3}_{1000\le u<5000} \longrightarrow \underbrace{\text{hard top-256}}_{u\ge 5000}. \tag{14}$$

In the first phase, masked positions are placed in local memory and binary cross-entropy trains the selector to recognize them. In the soft phase, the candidate memory remains full length but each vector is multiplied by $\alpha_t = \mathrm{softmax}(s_t/\tau_s)$, allowing the generation loss to reach the selector. An auxiliary squared budget loss, with weight 0.01, encourages the mean selection probability $\frac{1}{|\mathcal{V}|}\sum_{t\in\mathcal{V}}\sigma(s_t)$ over valid candidates $\mathcal{V}$ to approach 0.1. At the transition to hard selection, the projection and selector are frozen because the top-k indices are non-differentiable; local cross-attention continues to adapt on the same hard route used downstream.

A.5 Route specialization and the local memory budget

Setup. We probe the final Stage-2 checkpoint, the one used by every downstream stage. Its early-stopping copy was selected on BHC loss alone, early in training, before most of the local alignment. Each probe evaluates the held-out alignment split (27,076 admissions of 12,835 validation patients) under hard selection, with both routes open and with the local route, the global route, or both switched off by forcing the corresponding gate logits to -30 ($\sigma \approx 0$). Two validation targets are used: BHC generation, which is the global alignment task, and concept-description generation with fixed masked positions, which are excluded from the local candidates as during training. To choose the downstream budget, the local memory is truncated to $K_\ell \in \{256, 128, 64\}$ at inference without retraining. Switching the local route off, alone or together with the global route, removes the local memory entirely, so those passes do not depend on K_ℓ (Table 6).

The two routes specialize. Removing the local route raises concept-description loss by more than an order of magnitude (0.280 to 3.311, $+1084\%$) but leaves BHC loss unchanged (1.794 to 1.792, -0.1%). Removing the global route roughly doubles both losses (BHC $+78.6\%$, concept description $+92.3\%$), and removing both routes raises concept-description loss to 3.629. The local route is therefore specialized to concept-level recovery, whereas the global route supports both objectives. The pattern is not forced by the objectives: BHC replay during Stage 2 leaves the local route free to contribute to BHC generation, yet BHC loss is unchanged without it. The trajectory summary also does not substitute for access to individual events: with the local route off, the global route alone cannot recover the masked concepts, which motivates keeping the local route for evidence citation.

Choice of K_ℓ. Truncating the local memory from 256 to 128 and 64 events raises concept-description loss by 1.7% and 7.6% with both routes open (0.280, 0.284, and 0.301) and by 2.9% and 10.0% with the global route off, and leaves BHC loss unchanged (1.7937 vs. 1.7942). Even at $K_\ell = 64$, concept-description loss remains an order of magnitude below its value without the local route, so the information the local route supplies is concentrated in the highest-ranked events. The budget also determines whether selection happens at all. Table 7 shows that at $K_\ell = 256$ the local memory would contain every eligible event in 78.9% of 24-h records and 58.5% of readmission records, so the selector would act as an identity on most inputs; the median 24-h record has only 96.5 eligible events. At $K_\ell = 64$, selection is active in 70.4% of 24-h records and 79.7% of readmission records. We therefore align with $K_\ell = 256$, which gives the selector and local cross-attention a broad candidate set during training, and use $K_\ell = 64$ in every downstream stage, where selection remains a real choice at little cost in concept recovery.

Table 6: Route ablation after concept alignment (final Stage-2 checkpoint, hard selection): validation loss with the local and/or global cross-attention gates forced shut, and its change relative to the full model at the same K_ℓ. Passes with the local route off do not read the local memory, so their values are shared across K_ℓ (↑: same as the K_ℓ=256 row); —: not evaluated.

Validation target	K_ℓ	Full	Local off	Global off	Both off
Concept description	256	0.2797	3.3105 (+1083.5%)	0.5378 (+92.3%)	3.6285
	128	0.2844	↑	0.5536 (+94.7%)	↑
	64	0.3010	↑	0.5913 (+96.5%)	↑
BHC generation	256	1.7937	1.7922 (−0.1%)	3.2044 (+78.6%)	—
	64	1.7942	↑	—	—

Table 7: Local budget on the downstream test inputs (6,804 24-h and 7,566 readmission records; mortality and long length of stay share the 24-h inputs): eligible events per record (non-padding positions of the encoder input, truncated at 2,048, a limit that 10.2% of readmission records reach) and the share of records in which top-K_ℓ selection is inactive because the record has at most K_ℓ eligible events. On average, $K_\ell = 256/128/64$ selects 120.1/88.9/56.2 events for the 24-h inputs and 169.1/103.9/58.8 for readmission.

	Eligible events		Selection inactive ($\leq K_\ell$ events)		
Task	Median	p90	$K_\ell = 256$	$K_\ell = 128$	$K_\ell = 64$
Mortality, long LOS (24 h)	96.5	603	78.9%	66.5%	29.6%
Readmission (full)	194	2,048	58.5%	37.8%	20.3%

A.6 Downstream supervision and answer scoring

Direct prediction. The direct branch uses a task instruction and scores the answer slot without generating a rationale. For candidate $a \in \{\texttt{Yes}, \texttt{No}\}$, let $(v_1^a, \ldots, v_{n_a}^a)$ be its answer tokens. We compute

$$s_a = \frac{1}{n_a} \sum_{i=1}^{n_a} \log p_\theta(v_i^a \mid \mathbf{x}, q, v_{<i}^a), \qquad p_\theta(y = 1 \mid \mathbf{x}, q) = \frac{\exp s_{\text{Yes}}}{\exp s_{\text{Yes}} + \exp s_{\text{No}}}. \tag{15}$$

In the multi-task setting, one interface is shared across the three task instructions. Only global and local cross-attention modules are adapted by next-token cross-entropy on the answer continuation and its terminating token. The EHR encoder, resampler, local projection and selector, and base LLM remain frozen. At evaluation, Equation 15 averages answer-token log-likelihoods only, excluding the terminating token. Hard top-64 selection is used during both training and evaluation.

Supervised rationale initialization. The reasoning branch starts from the aligned model rather than the direct checkpoint. Approximately 2.5K examples, balanced across task and label, use rationales generated by Qwen3.5-397B-A17B (Qwen Team, 2026). The teacher receives the gold outcome and a textual rendering of the observed EHR; its outputs are label-conditioned, post-hoc rationales, not independent annotations of clinical causality. After format and content checks, the student is trained to reproduce the rationale and final answer while seeing only the structured EHR interface and task instruction. The EHR encoder remains frozen throughout, while the base LLM, resampler, and both cross-attention interfaces are updated.

SFT runs in two consecutive phases on the same data. In the *soft phase*, the local memory contains all eligible events weighted by $\alpha_t = \mathrm{softmax}(s_t/\tau_s)$ with $\tau_s = 0.3$, as in the soft phase of concept alignment, and the local projection and selector are also trainable, so the rationale loss reaches the event ranking. No selector budget loss is used, keeping the objective a pure rationale cross-entropy. The *hard phase* starts from the best soft-phase checkpoint and restores hard top-64 selection; because top-k indices pass no gradient, the projection and selector are frozen again, and the remaining components continue to adapt to the memory used by GRPO and evaluation. Both phases use learning rate 10^{-5} for the interface and 10^{-6} for the base LLM, weight decay 0.01, a cosine schedule with 5% warmup restarted at the start of each phase, and at most ten epochs with early stopping (patience 2,

minimum improvement 0.005) on the chain cross-entropy of a held-out 10% split. The GRPO policy and its reference are initialized from the best hard-phase checkpoint.

A.7 Reward implementation and group relative policy optimization

Sampling and reward terms. Starting from the SFT checkpoint, the current run samples $G = 4$ completions per patient–task prompt, with temperature 0.8, nucleus threshold 0.95, and token-sampling top-$k = 20$. Generation is capped at 1,536 new tokens and stops at the end of the answer. For a completion o_j, the reward is

$$R(o_j) = r_{\text{ans}} + 0.10 r_{\text{fmt}} + 0.15 r_{\text{evid}} - r_{\text{len}}, \tag{16}$$

where $r_{\text{ans}} = 1$ for a correct answer and -1 otherwise. The format reward is $r_{\text{fmt}} = 2q - 1$, where q is the fraction of four structural checks the response passes: a single reasoning block, an evidence heading followed by at least one bullet, a reasoning heading followed by at least one sentence, and exactly one final-answer tag closing the response. q is reduced by 0.2, with a floor of zero, when the response contains an explicit event identifier (`E` followed by three or more digits), which is not part of the student output format. The length penalty is

$$r_{\text{len}} = \min(0.1,\ 0.0005 \max(0, |o_j| - 768)) . \tag{17}$$

A missing or invalid final answer, or a detected reference to future outcome information, overrides Equation 16 with -1.5, below the lowest valid total of -1.35. The leakage check consists of patterns that require outcome language in the past tense or an explicit reference to the label: statements that the patient died, survived, was discharged, or was readmitted; statements that no readmission occurred within 30 days; references to the actual or observed length of stay; and references to a gold or provided answer or label. Prospective statements such as "high risk of death" are not flagged.

Evidence parsing and candidate events. Responses are parsed strictly, without repair. Evidence items are the bullet lines (beginning with "-" or "*") between the heading "Relevant clinical evidence" and the next "Reasoning", "Analysis", or "Uncertainty" heading. An optional role tag (`risk_factor`, `protective_factor`, `counterevidence`, `uncertainty`, `supports_yes`, or `supports_no`) is stripped and not scored. The candidate events are all non-special, non-padding positions of the same truncated input window the model receives ($T_{\max} = 2048$; Appendix A.1), each represented by its typed short description. Candidates therefore include events outside local memory.

Lexical matcher. Text is lower-cased and split into alphanumeric tokens that begin with a letter. Type prefixes ("condition:", "lab:", and so on), single characters, function words, and six generic words (*patient*, *clinical*, *evidence*, *history*, *risk*, *factor*) are removed, which gives a content-token set $T(\cdot)$. Token statistics come from the $N = 20{,}561$ typed descriptions of the model vocabulary. With $\mathrm{df}(t)$ the number of descriptions containing token t, $\mathrm{idf}(t) = \log \frac{N+1}{\mathrm{df}(t)+1} + 1$, and t is an *anchor* if $\mathrm{df}(t) \le 0.02N$ (411 descriptions). The direction words *very*, *high*, *low*, and *normal* are never anchors. Anchors are therefore derived from the vocabulary rather than from an authored list of medical terms; the most frequent content tokens are drug-exposure and dosage words such as *exposure*, *mg*, and *oral*. For a bullet p and a candidate description d, the matcher applies four rules.

1. *Direction.* If p states *very high*, *high*, *very low*, *low*, or *normal*, d must state the same direction. A bullet without a direction may cite a directional event.
2. *Qualifiers.* p and d must assert the same negation qualifiers ("no evidence of", "no history of", "no sign(s) of", "negative for", "deny/denies/denied", "ruled out", "absence of", "not present") and the same family-history attribution. "Without" and "not" are excluded because they act as specifiers in OMOP concept names ("… without hepatic coma", "not otherwise specified"). Across 59,929 logged evidence phrases, "without" caused 30 rejections, all of legitimate citations, whereas the retained qualifiers caused none.
3. *Score.* The IDF-weighted Dice overlap is

$$s(p, d) = \frac{2 \sum_{t \in T(p) \cap T(d)} \mathrm{idf}(t)}{\sum_{t \in T(p)} \mathrm{idf}(t) + \sum_{t \in T(d)} \mathrm{idf}(t)} . \tag{18}$$

4. *Support.* p is supported by d if $T(p) \subseteq T(d)$ or $T(d) \subseteq T(p)$, or if every anchor in $T(p)$ occurs in $T(d)$ and $s(p, d) \geq 0.55$.

Each bullet is assigned the candidate that is largest in the order (supported, s, number of shared anchors); a bullet that no candidate supports is unsupported. The anchor rule is precision-first: "acute on chronic renal failure" cannot cite "acute on chronic heart failure" even though most tokens overlap, because the anchor *renal* is unmatched. The matcher is lexical rather than ontological, which leaves two known failure modes. A paraphrase that introduces a rare word absent from the description, such as "kidney failure" for a renal-failure event, is not matched, so support is conservative for paraphrased findings. Conversely, containment in either direction accepts under-specified bullets: a bare laboratory name such as "Sodium" is supported by the event "Sodium [Moles/volume] in Serum or Plasma, low" and receives that event's credit.

Distinct-concept credit. $f(c)$ is the fraction of GRPO training records, pooled over the three tasks and computed on the untruncated inputs, that contain concept c. The i-th bullet receives $\xi_i = 1 - f(c)$ if it is the first supported citation of c in the response, and $\xi_i = 0$ if it is unsupported, cites a concept already credited in the response, or matches a metadata description (age, sex, gender, visit, or discharge). Metadata is excluded because each individual value, for example one of 126 age buckets, is rare by construction and would otherwise earn nearly full credit. With m bullets and $m_0 = 4$,

$$S = \frac{1}{\max(m, m_0)} \sum_{i=1}^{m} \xi_i, \qquad r_{\text{evid}} = 2S - 1 \quad (m > 0), \qquad r_{\text{evid}} = -1 \quad (m = 0). \tag{19}$$

Two development runs motivated this form. In a run that credited every supported bullet, rollouts with ten or more duplicated bullets averaged a total reward of 0.63, against 0.15 for rollouts with fewer than three, because repeating one supported finding diluted the unsupported bullets. First-citation credit removes this gain. In a run whose denominator was m alone, the mean number of bullets fell from 10.96 to 5.63 over 1,880 steps while weighted precision rose from 0.18 to 0.38. After about 1,100 steps the number of distinct supported findings also fell, from 3.09 to 2.64. The floor m_0 removes the gain from shortening a list below m_0 bullets. Above m_0, dropping a bullet whose credit is below the current average still raises S, so we report distinct supported findings together with precision; in the final run, GRPO shortened responses from 10.1 to 7.8 bullets while distinct supported findings rose (Section 4.3). Credit is the same whether the matched event lies inside or outside local memory; that distinction is logged only for analysis.

Evidence-sufficiency loss. For at most two responses per group, rotated across steps so that every response position is reached, the response is reduced to its evidence heading and bullets, the reasoning prose and answer are removed, and the answer prefix is appended. The EHR is then silenced: the resampled global memory tokens are zeroed and both cross-attention routes are bypassed. The cross-entropy of the gold answer token under this input is added to the loss with weight 0.2 and no class weighting, so its gradient reaches the language model only through the text path. Responses without bullets are skipped because the evidence term already assigns them -1. Only the bullets are shown because, with the full rationale visible, the loss can be minimized by stating the conclusion in the prose. As a reward, $2p_{\text{gold}} - 1$ would be mean-subtracted within each group. On logged rollouts its within-group slope in the number of bullets was -0.024 for negative-label prompts and $+0.029$ for positive-label prompts, so with about 72% negative prompts the reward form would pay for citing less. As a loss it is a proper scoring rule with no group baseline to exploit. The bullets-only answer probability is also logged as a training diagnostic but is not part of the reward.

Policy update. Rewards are standardized within each group and advantages are clipped:

$$A_j = \text{clip}\left(\frac{R(o_j) - \overline{R}}{\text{sd}(R(o_1), \ldots, R(o_G)) + \epsilon}, -5, 5 \right), \tag{20}$$

with zero advantage when the group variance is zero. The GRPO loss combines the clipped policy-ratio surrogate (Shao et al., 2024), a reference-policy KL penalty with coefficient 0.05, the answer loss, and the evidence-sufficiency loss above. The answer loss is the cross-entropy of the gold answer token at the answer position of each sampled response with a valid answer, with coefficient 0.2 and no class weighting. The KL term is omitted for zero-variance groups, whose only remaining

gradient would otherwise pull the policy toward the reference model. In the current configuration there is one inner update, and old-policy log-probabilities are detached current log-probabilities. The importance ratio is therefore one on that update and clipping is inactive. Operationally, this run uses a group-normalized policy-gradient update with KL and auxiliary-loss regularization. The EHR encoder, resampler, local projection, and selector remain frozen; the base LLM and both cross-attention interfaces are updated.

A.8 Evidence intervention and route probes

Matcher support alone does not establish that a cited event affected model behavior (Jacovi & Goldberg, 2020). We define a pre-generation intervention to assess this dependence. For each example, we first decode the full rationale and answer greedily, map citations corresponding to selected local-memory concepts, and record p_{Yes}. We then mask every mapped local slot before generating the first output token and regenerate the complete output. The control intervention masks the same number of uncited selected slots, matched as closely as possible by $1 - f(c)$, because removing inputs without retraining can itself change model behavior (Hooker et al., 2019). We compare three quantities between the cited and control interventions: persistence, the share of the masked concepts that the regenerated rationale still cites; the change in p_{Yes} and in the predicted label relative to the full chain; and AUROC, with the answer read either under the same mask or with the local route switched off, which removes direct access to local memory; masked evidence can then reach the answer only through the regenerated text or the global memory.

Sample. The intervention is applied to the first 500 test cases of each task and to 1,000 further readmission cases, added to test whether a readmission effect seen in the first 500 replicates. Under greedy decoding, 2,280 of the 2,500 chains reach the answer anchor, and 2,062 of them cite at least one concept held in local memory (mortality 453 with 17 positives, long LOS 441 with 152, readmission 1,168 with 75). A chain cites 3.2 such concepts on average, occupying 4.3 of the slots of about 44 distinct concepts in local memory; cited and control concepts have similar informativeness (0.826 vs. 0.833, against 0.715 for all local concepts).

Results. Masking the cited slots lowers persistence by 0.017 [0.007, 0.028] relative to the control mask, most clearly on readmission (Table 8), so the local route contributes to which findings are cited. The answer is no more sensitive to the cited mask than to the control mask. Before extending the readmission sample, we fixed the primary test as the cited-minus-control AUROC on readmission with the local route off, evaluated on the 1,000 added cases, with all 1,500 cases as a secondary test. A difference of -0.010 [-0.027, 0.004] in the first 500 cases did not replicate ($+0.002$ [-0.012, 0.019]; pooled -0.003 [-0.013, 0.007]).

Table 8: Cited-slot masking on MIMIC-IV test cases whose greedy chain cites concepts held in local memory. Persistence: share of the masked concepts (cited in the full chain) that the regenerated rationale still cites. $|\Delta p_{\mathrm{Yes}}|$ and flips: change of the answer probability and of the predicted label relative to the full chain. AUROC: cited minus control, with the answer read under the mask or with the local route switched off. Brackets: two-sided 95% paired bootstrap intervals. AUROC refers to this greedy subset and is not comparable with Table 1.

	Mortality	Long LOS	Readmission	All
Cases (positives)	453 (17)	441 (152)	1,168 (75)	2,062 (244)
Persistence, cited mask	0.874	0.885	0.851	0.863
Persistence, control mask	0.885	0.889	0.875	0.880
Persistence, cited − control	-0.011 $[-0.033, 0.010]$	-0.005 $[-0.026, 0.018]$	-0.024 $[-0.040, -0.009]$	-0.017 $[-0.028, -0.007]$
$\lvert\Delta p_{\mathrm{Yes}}\rvert$, cited / control	0.007 / 0.007	0.021 / 0.021	0.012 / 0.012	0.013 / 0.013
Label flips (%), cited / control	1.1 / 1.3	1.8 / 2.9	0.4 / 0.3	0.9 / 1.1
AUROC, cited − control, masked	-0.001 $[-0.008, 0.004]$	0.001 $[-0.005, 0.008]$	-0.009 $[-0.023, 0.005]$	-0.003 $[-0.009, 0.002]$
AUROC, cited − control, local off	-0.005 $[-0.013, 0.001]$	0.002 $[-0.005, 0.009]$	-0.003 $[-0.014, 0.007]$	-0.002 $[-0.007, 0.002]$

The intervention measures output sensitivity to cited local representations, not causal effects of the corresponding clinical events. Global memory and unmasked local slots remain available, so a small intervention effect can reflect redundant evidence. A cross-attention gate probe also differs from complete route removal: in a global-cross-attention-off probe, the 64 inserted global embeddings remain in the LLM sequence. Such a probe measures the incremental effect of global cross-attention and cannot be described as removing all global information.

B Data and cohort construction

Internal cohort. Our main dataset is MIMIC-IV v2.2 (Johnson et al., 2023b), converted to the OMOP CDM format; it contains 259,770 patients with at least one visit and 431,231 hospital admissions. Discharge summaries come from MIMIC-IV-Note v2.2 (Johnson et al., 2023a), and BHC targets come from MIMIC-IV-Ext-BHC (Aali et al., 2025), which provides the brief hospital course section of 270,033 discharge notes. Figure 3a shows the cohort. Of the patients with at least one visit, 180,733 have a hospital admission, 145,914 have a discharge summary, and 128,350 have at least one admission with an extractable BHC. These 128,350 patients form the cohort. They are split by patient into training (89,845), validation (12,835), and test (25,670) sets, a 70/10/20 split with a fixed seed.

Use of the split. Every stage fits its parameters on training patients only: EHR-BERT pretraining, both alignment stages (189,115 training and 27,076 validation admission–BHC pairs), rationale SFT, GRPO, and all supervised baselines. The teacher rationales for SFT are generated from training records only. Checkpoints are chosen per stage: EHR-BERT pretraining and global alignment by validation loss; the supervised baselines and the direct GLoC-EHR variants by validation AUROC (Section 4.2); rationale SFT by early stopping on a held-out 10% of its training examples (Appendix A.6); and concept alignment and GRPO use their final checkpoints (Appendices A.5 and C.2). The downstream tasks inherit the split, so test records are used neither to fit model parameters nor to select checkpoints. bin thresholds were derived from the full cohort before splitting. Neither uses outcome labels, and refitting the bins on training patients alone would change the value bin of 0.35% of measurement events.

Downstream tasks. Index admissions are inpatient visits, with OMOP visit concept Inpatient Visit or Emergency Room and Inpatient Visit, and every eligible admission of a patient is used. Mortality and long length of stay share one set of admissions: stays of at least 48 hours, excluding patients who died before the input cutoff. The input contains the events recorded in the first 24 hours of the admission, with discharge information removed. Mortality is death during the admission, and long length of stay is a stay longer than 7 days. Thirty-day readmission uses the full index admission with its discharge-disposition tokens removed and excludes admissions that end in death. It is positive when the next inpatient admission begins within 30 days of discharge; a negative additionally requires that the data extend at least 30 days beyond discharge. Table 9 gives the resulting sizes and prevalences.

Serialized record length. The token counts in Section 1 serialize the same event sequences that the models receive, without deduplication, using the typed short concept descriptions, and add the direct-answer instruction (about 134 tokens). Tokens are counted with the Qwen3-1.7B tokenizer over all training, validation, and test cases of each task.

Table 9: Downstream sample sizes (admissions) and outcome prevalence. Mortality and long length of stay (LOS) share the same admissions and 24-h inputs. EHRSHOT is used for evaluation only.

		MIMIC-IV			
Task	Input	Training	Validation	Test	EHRSHOT
In-hospital mortality	First 24 h	24,876 (3.19%)	3,549 (3.18%)	6,804 (3.15%)	9,922 (2.75%)
Long LOS (>7 d)	First 24 h	24,876 (37.06%)	3,549 (37.70%)	6,804 (36.67%)	9,922 (31.14%)
30-day readmission	Full admission	27,420 (7.55%)	3,890 (7.46%)	7,566 (6.89%)	12,342 (23.70%)

External cohort. The EHRSHOT release (Wornow et al., 2023) we use contains 6,731 patients in the OMOP CDM; 6,710 have at least one visit and 4,498 have at least one inpatient visit (Figure 3b).

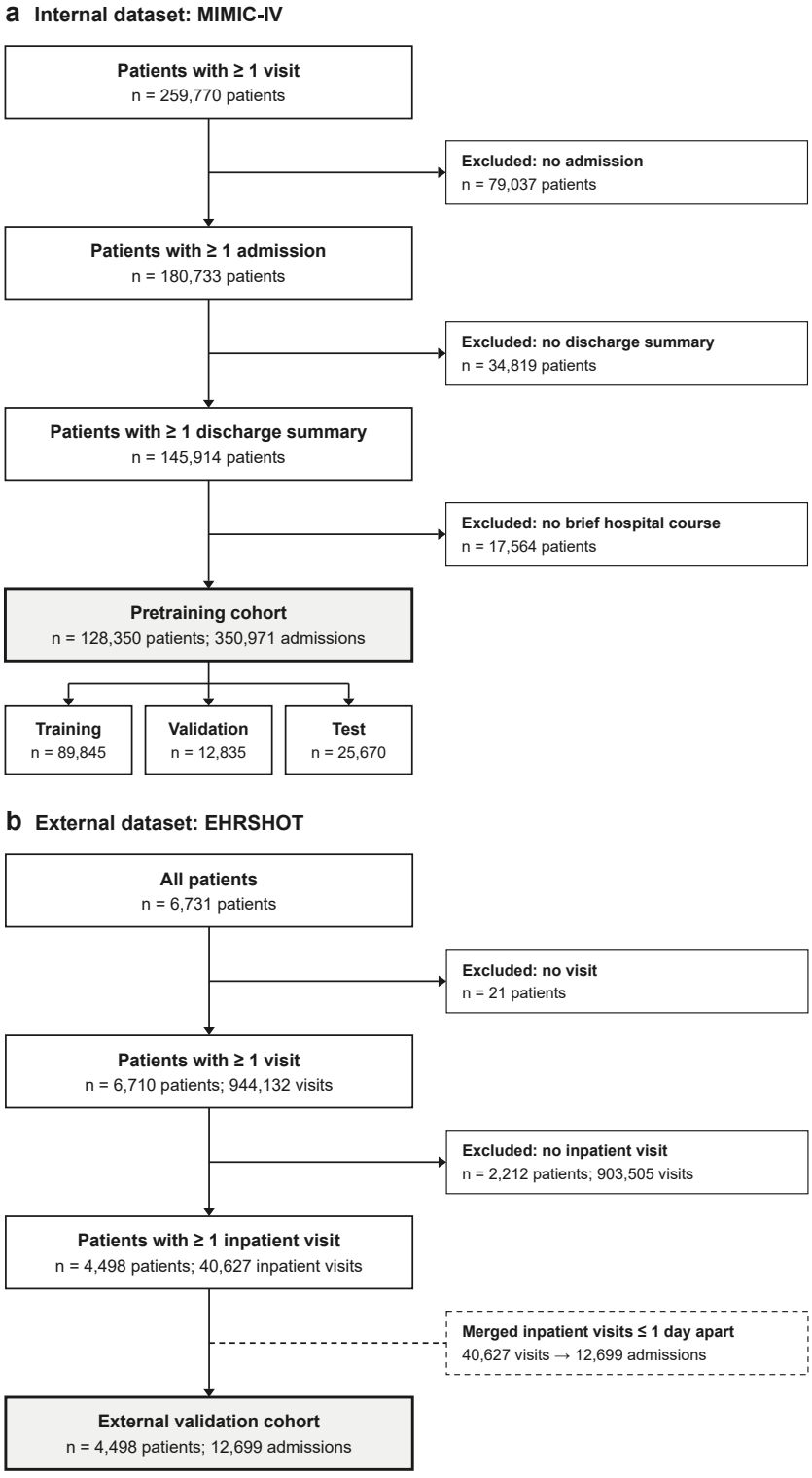

Figure 3: Cohort construction. (a) Internal MIMIC-IV cohort (OMOP CDM format), split by patient. (b) EHRSHOT external cohort; inpatient visits that start within one day of the previous discharge are merged into one admission. Task-level sample sizes are given in Table 9.

EHRSHOT records one hospital stay as several consecutive visits: 60.9% of consecutive inpatient visits start within one day of the previous discharge. For readmission, such visits are merged into one admission (40,627 visits into 12,699 admissions), and the longest fragment serves as the index; without merging, fragments of the same stay would count as readmissions and raise the prevalence to 0.39. Mortality and long length of stay use visit records directly, because the 48-hour minimum already excludes short fragments. The task definitions are otherwise identical to the internal ones, and all models are applied without retraining.

C Additional outcome-prediction details and scores

Tasks and evaluation. All models use the same patient-level MIMIC-IV splits (Appendix B). Mortality and long length of stay are predicted from the first 24 hours of the index admission (test $n = 6{,}804$; prevalence 3.15% and 36.67%), and 30-day readmission from the full index admission (test $n = 7{,}566$; prevalence 6.89%). Supervised configurations are selected by task-specific validation AUROC, or by validation macro AUROC over the three tasks for multi-task models, and the selected configuration is retrained with seeds 42–46. We report test means and sample SDs ($\mathrm{ddof} = 1$). Macro scores average the three tasks within each seed before aggregating across seeds; for task-specific models they summarize separately selected runs rather than a common configuration. Logistic-regression fits are deterministic across seeds and are shown without SD. Bold and underlined values mark the best and second-best distinct four-decimal means within each column; they are descriptive and do not indicate statistical significance.

Zero-shot LLM baselines. Each record is serialized as a newline-separated list of concept names, ordered from most to least recent, de-duplicated, and truncated to 8,192 tokens (Hegselmann et al., 2026). The prompt follows the ClinicalBench instruction (Chen et al., 2024a) ("Answer 1 if yes, 0 if no"). The predicted probability is the next-token probability of the answer at the answer slot, renormalized over the two options. Direct prompts use greedy decoding. For CoT, Llama-3.3-70B-Instruct decodes greedily (at most 1,536 tokens), and Qwen3-1.7B uses thinking mode (T=0.6, top-p 0.95, top-k 20, at most 3,072 tokens). Chains that do not reach an answer are closed with the answer prefix (forced closure); Table 13 reports how often this occurs. The "our prompt" variants replace the ClinicalBench instruction with the Yes/No evidence-then-answer format of GLoC-EHR (Reasoning).

Count-based and sequence baselines. Count features contain 20,567 code counts. LR sweeps $C \in \{10^{-4}, \ldots, 10\}$ with and without positive-class weighting (12 configurations); LightGBM and XGBoost sweep learning rate $\{0.05, 0.1\}$, tree size, and positive-class weighting (8 configurations each) with up to 3,000 boosting rounds and early stopping. GRU (Cho et al., 2014) and RETAIN (Choi et al., 2016) sweep learning rate, hidden size, positive-class weighting, and batch size (24 configurations each) over the same 2,048-event inputs.

Frozen-vector baselines. T_{4096} is the 4,096-dimensional Qwen3-Embedding-8B embedding of the serialized record, E_{CLS} and E_{mean} are 768-dimensional frozen EHR-BERT CLS and mean-pooled vectors, and C denotes code counts. Seven feature sets (E_{CLS}, E_{mean}, T_{4096}, $T_{4096}{+}E_{\mathrm{CLS}}$, $C{+}E_{\mathrm{CLS}}$, $C{+}T_{4096}$, $C{+}T_{4096}{+}E_{\mathrm{CLS}}$) are paired with LR and LightGBM heads, and $C{+}E_{\mathrm{CLS}}$ additionally with XGBoost, giving 15 feature–head combinations with the same sweeps as above. These are local adaptations of the LLM-as-EHR-encoder approach of Hegselmann et al. (2026), not reproductions of its architecture or published results.

Fine-tuned models. EHR-BERT (ST) sweeps learning rate $\{3, 5, 10, 20\} \times 10^{-6}$ and label sampling (natural or power-tempered with $\alpha = 0.5$); EHR-BERT (MT) uses the same grid with task-balanced sampling and one prediction head per task. GLoC-EHR (ST) and (MT) sweep learning rate $\{1, 2, 3, 5, 10\} \times 10^{-4}$ with the same sampling options (task-balanced for MT), use hard top-64 local selection, and score only the answer tokens renormalized over Yes/No (Appendix A.6). All fine-tuned models use early stopping with patience 5.

Tables 10 and 11 report AUROC and AUPRC for every model; AUPRC is more informative than AUROC when positives are rare (Saito & Rehmsmeier, 2015).

Table 10: Full internal AUROC comparison on MIMIC-IV (test): mean ± SD over five seeds, or over five sampled chains of one run for GLoC-EHR (Reasoning); rows without SD are single runs or deterministic (LR). Best in bold, second underlined. ST: one model per task; MT: one model for all three tasks. C: code counts; T_{4096}: text embedding; E_{CLS}/E_{mean}: frozen EHR-BERT vectors; a slash names the classifier head.

Method	Mortality (24 h)	Long LOS (24 h)	Readmission (full)	Macro (3 tasks)
Zero-shot LLMs on serialized records (single run)				
Qwen3-1.7B (direct)	0.8140	0.6062	0.5183	0.6461
Qwen3-1.7B (CoT)	0.5488	0.5363	0.5052	0.5301
Qwen3-1.7B (CoT, our prompt)	0.5203	0.5136	0.4923	0.5087
Llama-3.3-70B (direct)	0.8268	0.6391	0.5269	0.6643
Llama-3.3-70B (CoT)	0.6780	0.5193	0.5034	0.5669
Llama-3.3-70B (CoT, our prompt)	0.6464	0.5004	0.5006	0.5491
Count-based models				
LR	0.9448	0.8463	0.7615	0.8509
LightGBM	0.9527 ± 0.0013	0.8624 ± 0.0011	0.7758 ± 0.0028	0.8637 ± 0.0013
XGBoost	0.9525 ± 0.0009	0.8620 ± 0.0004	0.7733 ± 0.0035	0.8626 ± 0.0010
Sequence models				
GRU	0.9232 ± 0.0023	0.8336 ± 0.0040	0.7400 ± 0.0097	0.8323 ± 0.0037
RETAIN	0.9186 ± 0.0054	0.8341 ± 0.0046	0.7415 ± 0.0045	0.8314 ± 0.0030
Frozen-vector heads				
E_{CLS} / LR	0.9375	0.8397	0.7595	0.8455
E_{CLS} / LGBM	0.9292 ± 0.0023	0.8385 ± 0.0043	0.7481 ± 0.0098	0.8386 ± 0.0042
E_{mean} / LR	0.9285	0.8375	0.7600	0.8420
E_{mean} / LGBM	0.9125 ± 0.0050	0.8389 ± 0.0018	0.7556 ± 0.0034	0.8357 ± 0.0016
T_{4096} / LR	0.9148	0.8175	0.7677	0.8333
T_{4096} / LGBM	0.9062 ± 0.0018	0.8196 ± 0.0006	0.7521 ± 0.0032	0.8260 ± 0.0016
T_{4096}+E_{CLS} / LR	0.9400	0.8440	0.7765	0.8535
T_{4096}+E_{CLS} / LGBM	0.9385 ± 0.0008	0.8493 ± 0.0018	0.7620 ± 0.0043	0.8499 ± 0.0015
C+E_{CLS} / LR	0.9481	0.8523	0.7653	0.8553
C+E_{CLS} / LGBM	0.9497 ± 0.0016	0.8580 ± 0.0010	0.7577 ± 0.0060	0.8552 ± 0.0016
C+E_{CLS} / XGB	0.9520 ± 0.0010	0.8576 ± 0.0006	0.7677 ± 0.0029	0.8591 ± 0.0008
C+T_{4096} / LR	0.9181	0.8225	0.7685	0.8364
C+T_{4096} / LGBM	0.9396 ± 0.0042	0.8504 ± 0.0015	0.7632 ± 0.0039	0.8510 ± 0.0014
C+T_{4096}+E_{CLS} / LR	0.9375	0.8459	0.7764	0.8533
C+T_{4096}+E_{CLS} / LGBM	0.9449 ± 0.0027	0.8589 ± 0.0013	0.7625 ± 0.0068	0.8554 ± 0.0017
Fine-tuned, task-specific (ST)				
EHR-BERT (ST)	**0.9607** ± 0.0018	<u>0.8701</u> ± 0.0008	0.7817 ± 0.0057	0.8708 ± 0.0020
GLoC-EHR (ST)	0.9561 ± 0.0017	**0.8740** ± 0.0018	**0.8075** ± 0.0015	**0.8792** ± 0.0008
Fine-tuned, multi-task (MT)				
EHR-BERT (MT)	<u>0.9575</u> ± 0.0017	0.8672 ± 0.0019	0.7865 ± 0.0026	0.8704 ± 0.0013
GLoC-EHR (MT)	0.9545 ± 0.0013	0.8671 ± 0.0030	0.7976 ± 0.0052	<u>0.8731</u> ± 0.0030
GLoC-EHR (Reasoning)	0.9459 ± 0.0010	0.8469 ± 0.0004	<u>0.8042</u> ± 0.0018	0.8656 ± 0.0008

C.1 Alignment and interface controls

Table 12 reports two single-run controls. *Without alignment*, the GLoC-EHR (MT) architecture is built on the raw backbone with the pretrained EHR encoder but randomly initialized resampler, cross-attention, projection, and selector, and these modules are trained on the three tasks with the backbone frozen. Selection is soft, because a randomly initialized selector cannot be trained under hard top-k selection, and the learning rate is chosen on validation macro AUROC. AUROC is lower than with alignment on all three tasks (macro 0.862 vs. 0.873), with similar AUPRC (0.535 vs. 0.524). The *LoRA* control fine-tunes Qwen3-1.7B (rank 128, $\alpha = 256$) on serialized records to answer Yes or No directly, with one learning rate (10^{-4}) and one seed. Validation macro AUROC is computed every 200 steps on a validation subset, training stops after three evaluations without improvement,

Table 11: Full internal AUPRC comparison on MIMIC-IV (test; prevalence 3.15% / 36.67% / 6.89%): mean ± SD over five seeds, or over five sampled chains of one run for GLoC-EHR (Reasoning); rows without SD are single runs or deterministic (LR). Best in bold, second underlined. ST: one model per task; MT: one model for all three tasks. C: code counts; T_{4096}: text embedding; E_{CLS}/E_{mean}: frozen EHR-BERT vectors; a slash names the classifier head.

Method	Mortality (24 h)	Long LOS (24 h)	Readmission (full)	Macro (3 tasks)
Zero-shot LLMs on serialized records (single run)				
Qwen3-1.7B (direct)	0.1203	0.4797	0.0718	0.2240
Qwen3-1.7B (CoT)	0.0347	0.3849	0.0696	0.1631
Qwen3-1.7B (CoT, our prompt)	0.0327	0.3732	0.0679	0.1579
Llama-3.3-70B (direct)	0.0916	0.4552	0.0727	0.2065
Llama-3.3-70B (CoT)	0.0480	0.3759	0.0693	0.1644
Llama-3.3-70B (CoT, our prompt)	0.0439	0.3669	0.0689	0.1599
Count-based models				
LR	0.3969	0.7772	0.3221	0.4987
LightGBM	0.4524 ± 0.0190	0.7966 ± 0.0021	0.3406 ± 0.0096	0.5299 ± 0.0102
XGBoost	<u>0.4612</u> ± 0.0058	0.7980 ± 0.0010	0.3489 ± 0.0048	0.5361 ± 0.0021
Sequence models				
GRU	0.2727 ± 0.0099	0.7512 ± 0.0048	0.2945 ± 0.0110	0.4395 ± 0.0022
RETAIN	0.3377 ± 0.0311	0.7579 ± 0.0064	0.2991 ± 0.0031	0.4649 ± 0.0109
Frozen-vector heads				
E_{CLS} / LR	0.3795	0.7585	0.3157	0.4845
E_{CLS} / LGBM	0.3479 ± 0.0112	0.7526 ± 0.0070	0.3301 ± 0.0165	0.4768 ± 0.0072
E_{mean} / LR	0.3496	0.7604	0.3120	0.4740
E_{mean} / LGBM	0.3112 ± 0.0263	0.7517 ± 0.0039	0.3273 ± 0.0058	0.4634 ± 0.0079
T_{4096} / LR	0.2255	0.7049	0.3056	0.4120
T_{4096} / LGBM	0.2131 ± 0.0079	0.7066 ± 0.0028	0.3132 ± 0.0099	0.4110 ± 0.0033
T_{4096}+E_{CLS} / LR	0.3283	0.7598	0.3148	0.4676
T_{4096}+E_{CLS} / LGBM	0.3416 ± 0.0108	0.7629 ± 0.0021	0.3381 ± 0.0052	0.4809 ± 0.0032
C+E_{CLS} / LR	0.4291	0.7805	0.3339	0.5145
C+E_{CLS} / LGBM	0.4414 ± 0.0128	0.7878 ± 0.0015	0.3479 ± 0.0063	0.5257 ± 0.0039
C+E_{CLS} / XGB	0.4554 ± 0.0102	0.7886 ± 0.0017	0.3485 ± 0.0069	0.5308 ± 0.0048
C+T_{4096} / LR	0.2396	0.7191	0.3060	0.4216
C+T_{4096} / LGBM	0.3513 ± 0.0075	0.7768 ± 0.0015	0.3317 ± 0.0102	0.4866 ± 0.0048
C+T_{4096}+E_{CLS} / LR	0.3078	0.7638	0.3165	0.4627
C+T_{4096}+E_{CLS} / LGBM	0.3667 ± 0.0139	0.7865 ± 0.0020	0.3288 ± 0.0112	0.4940 ± 0.0042
Fine-tuned, task-specific (ST)				
EHR-BERT (ST)	0.4564 ± 0.0064	**0.8072** ± 0.0006	<u>0.3504</u> ± 0.0058	<u>0.5380</u> ± 0.0034
GLoC-EHR (ST)	**0.4667** ± 0.0123	<u>0.8038</u> ± 0.0035	**0.3778** ± 0.0185	**0.5494** ± 0.0066
Fine-tuned, multi-task (MT)				
EHR-BERT (MT)	0.4556 ± 0.0114	0.8009 ± 0.0027	0.3372 ± 0.0063	0.5312 ± 0.0062
GLoC-EHR (MT)	0.4301 ± 0.0110	0.7952 ± 0.0057	0.3455 ± 0.0238	0.5236 ± 0.0088
GLoC-EHR (Reasoning)	0.3638 ± 0.0114	0.7498 ± 0.0011	0.3320 ± 0.0021	0.4819 ± 0.0045

and the step-1,800 checkpoint is evaluated. It reaches a macro AUROC of 0.853, below GLoC-EHR (MT) and the count-based LightGBM (0.864).

C.2 Reasoning evaluation and answer coverage

GLoC-EHR (Reasoning) is the final checkpoint of a single GRPO training run on the three tasks (step 3,216). For each test case we sample five chains (T=0.8, top-p 0.95, top-k 20, at most 1,536 reasoning tokens). The score of a chain is the Yes/No answer probability conditioned on the prompt and the generated reasoning prefix. Each chain index is treated as one sampling run: AUROC and AUPRC are computed separately for each of the five runs, the macro score averages the three tasks within a run, and we report the mean and SD over runs, as for training seeds. A chain that does not

Table 12: Single-run controls on the MIMIC-IV test set: AUROC per task and three-task macro AUROC/AUPRC. GLoC-EHR (MT) is the mean ± SD over five seeds from Table 1. Without alignment: the same architecture trained from randomly initialized interface modules; Qwen3-1.7B + LoRA: the backbone fine-tuned on the serialized record. Both are single runs (Appendix C.1).

	AUROC ↑			Macro ↑	
Model	Mortality	Long LOS	Readmission	AUROC	AUPRC
GLoC-EHR (MT)	0.9545 ± 0.0013	0.8671 ± 0.0030	0.7976 ± 0.0052	0.8731 ± 0.0030	0.5236 ± 0.0088
without alignment	0.9470	0.8578	0.7817	0.8622	0.5354
Qwen3-1.7B + LoRA	0.9414	0.8411	0.7762	0.8529	0.4990

reach the answer anchor is excluded from its own run. Because all runs come from one training run, the SD reflects decoding rather than training variability. Table 13 contrasts anchor coverage with the zero-shot CoT baselines, and Table 14 compares this protocol with scoring chains without an anchor as 0.5 and with averaging the five chains of each case. These outcome metrics do not establish evidence support or reasoning faithfulness, which are evaluated separately. Confidence intervals in the evidence and transfer analyses are two-sided 95% paired case bootstrap intervals (Efron & Tibshirani, 1993): test cases are resampled with replacement, and the compared models are scored on the same resamples. Across the test set, 99.6% of the 105,870 chains reach the answer anchor, and 13 of the 21,174 cases have no chain that does. Scoring chains without an anchor as 0.5 changes the macro AUROC by less than 0.001, and averaging the five chains of each case raises it only slightly, from 0.8656 to 0.8668.

Table 13: Answer-anchor coverage of generated chains (cases reaching a parseable answer; n = 6,804 / 6,804 / 7,566; GLoC-EHR (Reasoning) counts all five chains per case). For the zero-shot baselines, remaining cases are scored by forced closure.

Model	Mortality	Long LOS	Readmission	Overall (%)
GLoC-EHR (Reasoning)	33,966	33,883	37,623	99.62
Qwen3-1.7B (CoT)	6,799	6,799	7,562	99.93
Qwen3-1.7B (CoT, our prompt)	6,380	6,529	7,139	94.68
Llama-3.3-70B (CoT)	6,804	6,804	7,566	100.00
Llama-3.3-70B (CoT, our prompt)	6,804	6,804	7,566	100.00

Table 14: Scoring sensitivities of GLoC-EHR (Reasoning); unweighted three-task macro. Single-chain rows are means over the five chains; n counts cases.

Evaluation variant	n	Macro AUROC	Macro AUPRC
Single chains, valid anchors only (reported)	21,094	0.8656	0.4819
Single chains, no anchor $\mapsto 0.5$	21,174	0.8654	0.4814
Average of five chains, valid anchors	21,161	0.8668	0.4897
Average of five chains, forced closure	21,174	0.8665	0.4897

Evidence of the zero-shot CoT baselines. Table 15 scores the evidence bullets of the "CoT, our prompt" baselines with the reward matcher at τ_0=0.55, as in Table 2, over chains that reach the answer anchor. These models read the record serialized as concept names, with measurements described in plain language (for example, "Lactate – high, above the reference range"), so a bullet can repeat a line of the prompt: 94.1% of the bullets of Llama-3.3-70B and 58.4% of those of Qwen3-1.7B match a recorded concept exactly or by containment. GLoC-EHR (Reasoning) receives the record only through its EHR memories. Qwen3-1.7B, the backbone of GLoC-EHR, reaches a slightly higher support precision than GLoC-EHR (0.452 vs. 0.424) but cites fewer distinct supported findings (3.88 vs. 4.16), and Llama-3.3-70B reaches 0.752; both, however, answer near chance (macro AUROC 0.509 and 0.549). Stricter matching (Figure 2) is not compared across these systems, because the IDF-weighted overlap also penalizes wording that the serialization adds around a concept name, such as "below the reference range", which GLoC-EHR never sees.

Table 15: Evidence of GLoC-EHR (Reasoning) and the zero-shot "CoT, our prompt" baselines on the MIMIC-IV test set, scored with the reward matcher (τ_0=0.55) over chains that reach the answer anchor. GLoC-EHR (Reasoning): mean ± SD over five sampled chains, as in Table 2. The zero-shot models read the serialized record; GLoC-EHR receives it only through its EHR memories. Macro AUROC is taken from Table 10.

Model	Bullets	Support precision ↑	Unsupported rate ↓	Distinct supported ↑	Macro AUROC ↑
GLoC-EHR (Reasoning)	7.83	0.424 ± 0.001	0.440 ± 0.001	4.16 ± 0.01	0.8656 ± 0.0008
Qwen3-1.7B (CoT, our prompt)	6.69	0.452	0.420	3.88	0.5087
Llama-3.3-70B (CoT, our prompt)	6.85	0.752	0.060	6.33	0.5491

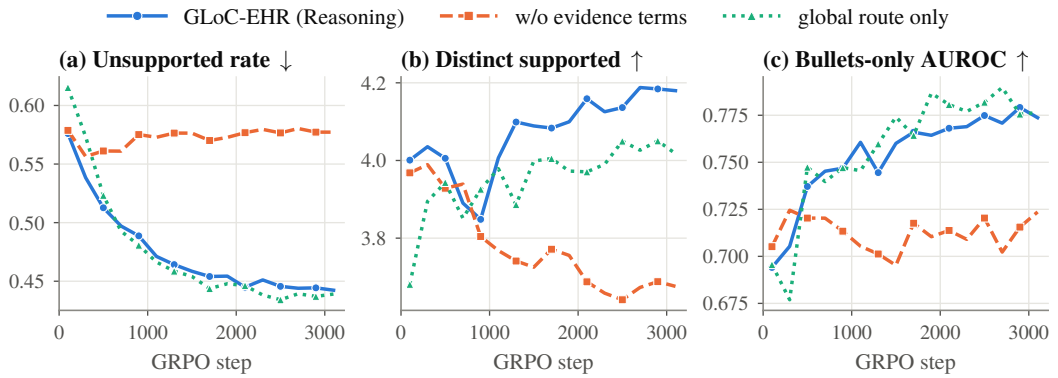

Figure 4: Evidence quality over GRPO training (training rollouts of all ranks, 200-step bins): (a) share of evidence bullets that match no event in the input window, (b) distinct supported findings per response, and (c) AUROC of the answer predicted from the bullets alone with the EHR memories switched off.

C.3 Evidence during GRPO training

Figure 4 tracks the evidence metrics of Section 4.3 over GRPO training for GLoC-EHR (Reasoning) and for the two variants trained without the evidence terms (the evidence reward and the evidence-sufficiency loss) or without the local route, using the training rollouts of all four ranks in 200-step bins. Without the evidence terms, support stays near its SFT level. At the end of training (steps 3,000–3,199), 57.7% of its bullets are unsupported vs. 44.2% for GLoC-EHR, it cites fewer distinct supported findings (3.68 vs. 4.18), and its bullets predict the outcome less well (bullets-only AUROC 0.724 vs. 0.774), at similar balanced accuracy (0.734 vs. 0.740). The global-route-only variant, trained from its own SFT checkpoint, ends with a similar unsupported rate (0.439 vs. 0.442) and bullets-only AUROC (0.776 vs. 0.774) but fewer distinct supported findings (4.02 vs. 4.18). Prompts are sampled with replacement.

C.4 Zero-shot transfer to unseen questions

We ask GLoC-EHR (Reasoning) two laboratory questions that it never saw in training: whether the next serum sodium will be below 135 mmol/L and whether the next platelet count will be below 150 K/μL, given the admission's trajectory up to that measurement. Only the question in the prompt changes. The cases come from MIMIC-IV admissions of held-out patients (none in the training split; 1,871 and 1,894 patients). In each admission, the first measurement below the threshold is taken as a positive and one randomly chosen measurement at or above it as a negative; negatives are matched to positives on sequence length, and 750 positives and 1,500 negatives are drawn per question. Table 16 compares one sampled chain of GLoC-EHR (Reasoning) (chains without an answer, 2.2%, are excluded) with zero-shot Qwen3-1.7B reading the serialized record, prompted as in Table 1. Neither model is trained on these questions. GLoC-EHR (Reasoning) is higher by 0.100 [0.068, 0.131] and 0.104 [0.071, 0.137] AUROC (paired case bootstrap).

Table 16: Zero-shot answers to two laboratory questions absent from training (held-out MIMIC-IV admissions; 750 positives and 1,500 length-matched negatives per question, so AUPRC is 0.33 by chance). GLoC-EHR (Reasoning) scores one sampled chain per case.

	Next sodium $<$ 135 mmol/L		Next platelet count $<$ 150 K/μL	
Method	AUROC	AUPRC	AUROC	AUPRC
GLoC-EHR (Reasoning)	0.657	0.500	0.629	0.499
Qwen3-1.7B, zero-shot	0.556	0.410	0.525	0.344