EHRADAPT: ADAPTING PRETRAINED LANGUAGE MODELS TO ELECTRONIC HEALTH RECORDS WITH SEMANTIC PRIORS FOR RARE CLINICAL EVENTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Electronic health records (EHRs) encode clinical histories as (time, modality, code) tuples, whereas pretrained language models expect text tokens. Serializing each tuple as text inflates sequence length and repeatedly encodes the same structure. We introduce EHRAdapt, an adapter that maps tuples directly into a frozen language model's embedding space. Modality receives a learned embedding, time gaps enter through learned attention biases, and event codes receive dedicated vectors. Learning these event vectors is the central challenge: clinical vocabularies are long-tailed, leaving rare events with too few observations to estimate reliable representations independently. EHRAdapt therefore represents each event vector as the sum of a *semantic prior* and an *evidence residual*. The prior is a frozen embedding of the event's clinical description, from a biomedical language model trained on clinical ontologies, so it carries clinical knowledge. A shared learned projection maps this embedding into the model's input space, allowing the prior to supply clinical meaning even when observations are scarce. The residual is a learned low-rank, event-specific correction that refines this representation as evidence accumulates. We perform continued pretraining on records from about 4 million patients with three pretrained LLMs as backbones: OLMo2 1B, Llama3.2 1B, and OLMo2 7B. As the LLMs remain frozen, only the adapter is trained, which amounts to 0.1–0.6% of all parameters. The full adapter outperforms all ablations in held-out next-event prediction on every backbone. Removing the semantic pathway hurts rare events most, over ten times more than the most frequent ones, whereas removing the residual hurts overall prediction but improves it for the rarest events. On reportable infectious-disease and syndromic downstream classification tasks, EHRAdapt outperforms text-based LLM and count-based baselines, and both pathways improve rare-disease discrimination. The semantic prior and the evidence residual therefore play complementary roles, and these roles become visible only when results are broken down by event frequency rather than averaged.

1 INTRODUCTION

Pretrained large language models (LLMs) offer an attractive starting point for modeling structured electronic health records (EHRs). Foundation models built specifically for EHRs are trained from scratch on large patient corpora (Wornow et al., 2023; Waxler et al., 2025), and this costly training must be repeated to benefit from newer architectures. Adapting an existing LLM requires less data and computing infrastructure, and newer LLMs can be adopted as they are released. Recent work has therefore adapted pretrained LLMs to clinical prediction from structured EHRs by typically converting each clinical event into text (Ben Shoham and Rappoport, 2024; Hegselmann et al., 2025; Goncalves et al., 2025; Wang et al., 2026). However, standard LLM tokenizers break a clinical event into multiple tokens, increasing the length of patient histories, while information about the event type (e.g., diagnosis, laboratory test, or medication) and its timing is encoded only implicitly.

Clinical events also follow a highly skewed distribution (Wright and Bates, 2010). A small set of routine events occurs very often, while most events are rare. In our training data, the median event type appears about 20 times, whereas the most common appears about 87.6 million times (Figure 4).

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.

Rare events may still carry important diagnostic information, but a model that learns each event only from EHR data sees too few examples of them. The skew also complicates evaluation: the average prediction loss is dominated by common events, so it can hide large effects on rare ones.

In this paper, we introduce EHRAdapt, an adapter that maps structured EHR events into the input space of a frozen pretrained LLM. Each event is represented by a single token, with event type encoded through learned embeddings and temporal information through attention biases. As the LLM backbone remains frozen, EHRAdapt is model-agnostic, and only the adapter is trained, with less than 1% as many parameters as the backbone. To better represent rare events, EHRAdapt combines a semantic prior with an event-specific residual. The semantic prior is a BioLORD embedding of the event's clinical description (Remy et al., 2024), projected into the LLM input space by a learned map shared across events. Because BioLORD is trained on clinical ontologies such as the Unified Medical Language System (UMLS), this prior is informative even for rare events.

We expect the contributions of the semantic prior and event-specific residual to depend on event frequency. Rare events should remain closer to their semantic priors due to limited observations, while common events can adapt more strongly through their learned residuals. This resembles Bayesian learning, where prior information has greater influence under data scarcity, although EHRAdapt neither performs Bayesian inference nor explicitly enforces this behavior.

We pretrain EHRAdapt on records from 4 million Kaiser Permanente Northern California patients collected between 2010 and 2018 (Goncalves et al., 2025). We use three frozen backbones, OLMo 2 1B and 7B (Team OLMo et al., 2025) and Llama 3.2 1B (Grattafiori et al., 2024), and on each one we compare the full model with ablations of the semantic prior, event-specific residual, and learned semantic mapping. Our evaluation asks three questions. First, how does each part affect prediction of the next clinical event, and for which events? To answer this, we report results separately for ten frequency groups, from the rarest events to the most common. Second, do the learned representations improve downstream prediction for patients seen between 2019 and 2022? Third, does the balance between the two parts shift with event frequency, as we expect?

The downstream tasks, reportable infectious disease and syndrome classification, are motivated by biosurveillance, where earlier diagnosis can speed outbreak response (Henning, 2004; Lazarus et al., 2009; Goncalves et al., 2025). Many diseases of surveillance interest are rare (Appendix D.1), making these tasks a natural test of whether the semantic prior helps most where data are scarce. Rare diseases and rare events are related but not identical, so we evaluate both.

Overall, EHRAdapt achieves the lowest test loss across all three backbones, with the semantic prior and event-specific residual playing complementary roles (Section 5). Removing the semantic prior degrades prediction across all frequencies, with the largest effect on rare events. In contrast, removing the residual improves prediction for the rarest events but degrades overall performance, suggesting that residuals primarily benefit common events. Consistently, residuals contribute less to representations of rare events than common events (Section 5.3). On downstream tasks, EHRAdapt outperforms text-based LLM inputs, BioLORD embeddings alone, and a tuned random forest, with both components improving rare-disease prediction (Section 5.2).

2 Related work

Representation learning for structured EHRs has evolved from BERT-style models such as BEHRT, Med-BERT, and CEHR-BERT (Li et al., 2020; Rasmy et al., 2021; Pang et al., 2021) to large-scale foundation models including CLMBR, TransformEHR, MOTOR, and Curiosity (Steinberg et al., 2021; Yang et al., 2023; Steinberg et al., 2024; Waxler et al., 2025). These approaches learn reusable patient representations through pretraining on large longitudinal EHR corpora. A complementary direction adapts existing pretrained LLMs to structured EHR sequences through fine-tuning or prompting, reducing the need for EHR-specific pretraining and facilitating the adoption of newer models (Ben Shoham and Rappoport, 2024; Zhu et al., 2024; Goncalves et al., 2025). However, the long-tailed distribution of clinical events provides limited observations for learning robust representations of rare events. Prior work such as GRAM (Choi et al., 2017) addresses this challenge by leveraging hierarchical relationships among medical concepts, constructing representations by attending over concept ancestors in a medical ontology. In contrast, EHRAdapt can incorporate semantic priors from arbitrary external sources, including biomedical language models and knowledge

graphs, and adaptively balances these priors with representations learned from EHR data according to the available evidence for each clinical event.

Recent approaches use external semantic knowledge to represent clinical events. Biomedical ontologies and representation models such as BioLORD encode semantic relationships among clinical concepts independently of their frequency in a particular EHR corpus (Bodenreider, 2004; Remy et al., 2022; 2024). Most closely related to our work, PORTER (Guo et al., 2026) represents clinical events using natural-language descriptions encoded by a frozen BioLORD encoder, enabling vocabulary-independent representations and improved transfer across coding systems and institutions. Whereas PORTER uses fixed semantic representations primarily for portability, EHRAdapt addresses the complementary problem of data scarcity for rare events through frequency-adaptive integration of semantic priors: rare events rely more on pretrained semantic knowledge, while representations of frequently observed events are increasingly shaped by evidence from the EHR corpus.

A separate line of work adapts LLMs to clinical notes and biomedical literature text. Because these documents are already written in natural language, they can be fed directly to a pretrained language model, and models such as Bio-ClinicalBERT, GatorTron, and NYUTron (Alsentzer et al., 2019; Yang et al., 2022; Jiang et al., 2023) are pretrained or fine-tuned on large collections of notes for tasks ranging from concept extraction to readmission and mortality prediction. Structured EHRs pose a different problem: clinical events are recorded as codes with timestamps rather than text, so they must first be mapped into a form an LLM can process. EHRAdapt focuses on this structured setting and is complementary to models of clinical notes.

3 METHOD

EHRAdapt operates in two stages. First, its parameters are learned through continued pretraining (CPT) on next-event prediction over longitudinal EHR data. Second, the adapted model is reused across downstream tasks. In the simplest setting, it stays frozen and only a task-specific head is trained; alternatively, while not explored in this paper, parameter-efficient methods such as LoRA (Hu et al., 2022) can further adapt it to each task. Figure 1 summarizes the CPT architecture, from structured patient trajectories to next-event prediction. The following subsections describe each component.

3.1 PROBLEM SETUP AND SEQUENCE CONSTRUCTION

A patient history is a sequence of clinical events $(c_1, m_1, t_1), \ldots, (c_L, m_L, t_L)$, where c_i indexes a live event in a vocabulary $\mathcal{V}$ of clinical event types plus special tokens, m_i is the event modality, and t_i is a timestamp. In the continued pretraining phase, the model is trained to predict c_{i+1} given the prefix, so the objective is next-event cross-entropy in nats.

3.2 EVENT REPRESENTATION

For clinical event c, we distinguish the shared event-code vector $\mathbf{e}_c$ from the modality-augmented input $\mathbf{x}_i$ at sequence position i:

$$\begin{aligned} \mathbf{e}_c &= P_{\text{sem}}(\mathbf{z}_c) + \mathbf{r}_c, \qquad \mathbf{r}_c = \mathbf{B}\,\mathbf{a}_c, \quad \mathbf{a}_c \in \mathbb{R}^r, \\ \mathbf{x}_i &= \mathbf{e}_{c_i} + \mathbf{e}_{\text{mod}}(m_i), \end{aligned} \tag{1}$$

where $\mathbf{z}_c \in \mathbb{R}^{d_{\text{sem}}}$ is the frozen BioLORD embedding (Remy et al., 2024) of the event name; $P_{\text{sem}} \in \mathbb{R}^{H \times d_{\text{sem}}}$ is a learned map to the backbone hidden dimension H; $\mathbf{a}_c$ is a learned event-specific code of dimension r; $\mathbf{B} \in \mathbb{R}^{H \times r}$ is a shared projection; and $\mathbf{e}_{\text{mod}}(m_i)$ is a learned modality embedding, from a lookup table with one row per clinical modality and an additional row for special tokens.

Residual rank The rank r controls the capacity of the event-specific correction. Stacking the codes $\mathbf{a}_c$ into $\mathbf{A} \in \mathbb{R}^{|\mathcal{V}| \times r}$, where $|\mathcal{V}|$ is the vocabulary size, gives the residual table $\mathbf{R} = \mathbf{A}\mathbf{B}^\top$ with $\text{rank}(\mathbf{R}) \leq r$. The factorization contains $(V + H)r$ parameters, compared with VH for an unconstrained residual table. For a backbone with $H = 2{,}048$, rank 64 uses approximately 5.15 million residual parameters versus 161 million for an unconstrained table, about 3.2% as many, excluding the semantic projection and other adapter components. Higher ranks relax the constraint,

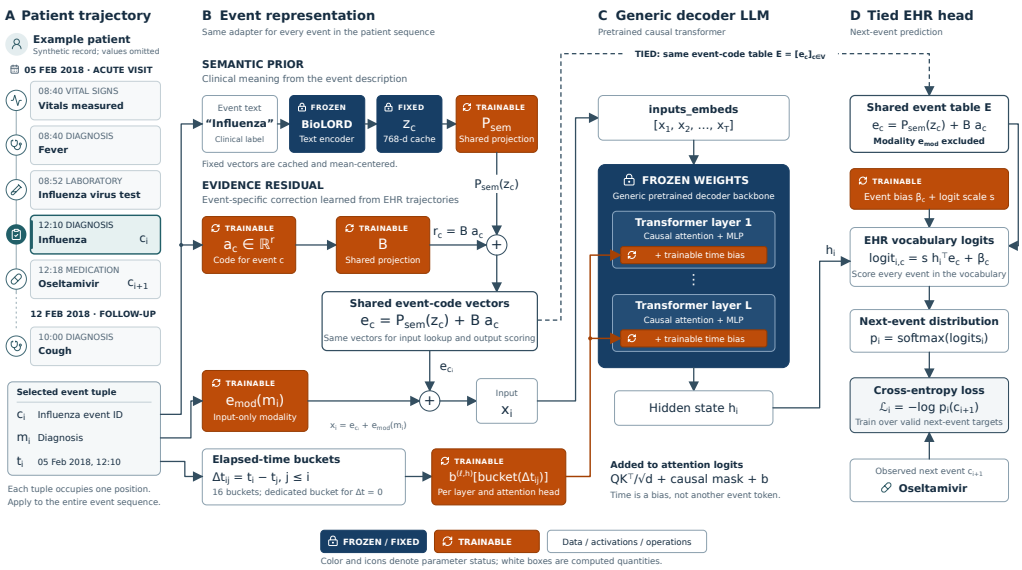

Figure 1: **EHRAdapt CPT architecture.** **(A)** Synthetic patient trajectory of timestamped, typed clinical events. **(B)** Frozen BioLORD encodes an event's clinical description as the semantic vector $\mathbf{z}_c$. A shared learned projection gives the semantic prior $P_{\text{sem}}(\mathbf{z}_c)$, which is added to the low-rank evidence residual $\mathbf{B}\mathbf{a}_c$ to form the event embedding $\mathbf{e}_c$. **(C)** A frozen decoder LLM receives these vectors through `inputs_embeds`. Pairwise elapsed-time buckets index learned attention biases per layer and head. **(D)** The tied output head reuses $\mathbf{e}_c$ without modality, with a learned event bias β_c and logit scale s; CPT minimizes next-event cross-entropy. Dark blue (lock): frozen; orange (update): trainable; white: data or computed.

and any residual table can be expressed once $r \geq H$. However, greater representational capacity does not guarantee better held-out performance, particularly for rare events.

Fixed semantic geometry and the prior interpretation. As an alternative to learning P_{sem}, the projection can be fixed at $P_{\text{sem}} = \alpha Q$, where $Q \in \mathbb{R}^{H \times d_{\text{sem}}}$ satisfies $Q^\top Q = I$ and $\alpha > 0$. Consequently, $\cos(\alpha Q \mathbf{z}_h, \alpha Q \mathbf{z}_l) = \cos(\mathbf{z}_h, \mathbf{z}_l)$, so all angles in the BioLORD space are preserved exactly, and cosine-neighbor rankings are preserved. Combined with the residual, this fixed projection corresponds most closely to a fixed-prior interpretation of Equation 1: the external semantic structure is held constant while the residual absorbs evidence from EHR sequences. Learning P_{sem}, as in the default model, relaxes this constraint and allows the semantic component to adapt to the backbone. We evaluate the fixed projection, with and without the residual, as controls in Section 5.

The decomposition in Equation 1 is central to our analysis, because its two pathways receive evidence differently. The semantic pathway is shared across the whole vocabulary and is therefore optimized predominantly by frequent events, but it reaches rare events through the BioLORD geometry: a rare event's $\mathbf{z}_c$ lies near the $\mathbf{z}$ of semantically related frequent events, so updates to P_{sem} driven by frequent events also move the rare event's representation. The residual pathway instead assigns each event its own learned code $\mathbf{a}_c$, while sharing the projection $\mathbf{B}$.

3.3 Output head, frequency bias, and logit scale

The output head is tied to the event-code representation table $\mathbf{E} = [\mathbf{e}_c]_{c \in \mathcal{V}}$ (Press and Wolf, 2017; Inan et al., 2017). For hidden state $\mathbf{h}_i$,

$$\text{logit}_{i,c} \;=\; s\left(\mathbf{h}_i \cdot \mathbf{e}_c\right) + \beta_c, \qquad s = \exp(\texttt{logit_scale}), \tag{2}$$

where β_c and `logit_scale` are learned and the modality embedding is deliberately excluded from the output representation. With a linear P_{sem} and a projected residual, the unscaled scores before adding β_c factorize as $(\mathbf{h}_i P_{\text{sem}})\mathbf{Z}^\top + (\mathbf{h}_i \mathbf{B})\mathbf{A}^\top$, which avoids materializing the $V \times H$ matrix and avoids constructing a dense event table for every forward pass.

Because the input and output embeddings are tied, each residual $\mathbf{a}_c$ receives gradients from two sources. On the input side, it is updated only when event c appears in a patient history. On the output side, the full softmax scores every event at every position, so $\mathbf{a}_c$ is updated at every step, even when event c is absent. Event frequency therefore determines how often $\mathbf{a}_c$ is updated with event c present, as an input or as the prediction target, but not how often it is updated overall.

3.4 Temporal attention bias

Elapsed time carries clinical information that event order does not, since consecutive events may be minutes or years apart. We therefore add a learned bias to the attention scores. For query position i and key position $j \leq i$, with $\Delta t_{ij} = t_i - t_j$, the score at layer ℓ and head h is

$$\mathbf{q}_i^\top \mathbf{k}_j / \sqrt{d_{\text{head}}} + b^{(\ell,h)}_{\kappa(\Delta t_{ij})}, \tag{3}$$

where κ assigns elapsed time to one of 16 buckets, with bucket 0 for simultaneous events. This follows the learned, bucketed relative-position biases of T5 (Raffel et al., 2020), but uses elapsed time instead of token distance. Biases are initialized to zero, leaving the backbone's attention unchanged at the start of training.

4 Experimental setup

We learn the EHRAdapt parameters by continued pretraining (CPT), then evaluate next-event prediction and transfer to two biosurveillance-related diagnostic prediction tasks. The full adapter and four ablations are trained on three backbones, described below, allowing us to distinguish component effects that recur across models from differences associated with backbone family or size.

4.1 Continued-pretraining data

We use Kaiser Permanente Northern California (KPNC) EHRs from approximately 4 million patients collected between 2010 and 2018. In addition to demographics, the data include longitudinal records of laboratory tests, procedures, medications, encounter types, chief complaints, diagnosis codes, and vital signs, without associated values or outcomes. A deterministic, seeded patient-level split assigns 98%, 1%, and 1% of patients to training, validation, and test sets.

Patient trajectories are partitioned into chunks of 16 to 512 events, with boundaries aligned to timestamp changes where possible so that simultaneous events are not split. A trailing chunk shorter than 16 events is merged into its predecessor, so all events are retained. Each training epoch contains approximately 721 million prediction targets. The vocabulary, frequency strata, and output-bias initialization are derived from the training split only. See Appendix A for additional information.

4.2 Backbones and configurations

We use frozen OLMo-2 1B and 7B backbones (Team OLMo et al., 2025), which are pretrained separately and provide a within-family size comparison, and Llama 3.2 1B (Grattafiori et al., 2024), which provides a cross-family comparison at the same hidden size. We do not include Llama 3.1 8B because Llama 3.2 1B was derived from it by pruning and knowledge distillation, so the pair would confound model size with model derivation. Additional backbone information is in Appendix B.

On each backbone, we compare the full adapter with four configurations that remove or fix components of Equation 1. The *Full model* learns P_{sem} and the rank-64 residual. *Semantic prior only* removes the residual ($r = 0$), so that $\mathbf{e}_c = P_{\text{sem}}(\mathbf{z}_c)$, and *Residual embeddings only* removes the semantic pathway, so that $\mathbf{e}_c = \mathbf{r}_c$. *Frozen isometry + residual* fixes $P_{\text{sem}} = \alpha Q$ as discussed in Section 3.2, while retaining the residual, and *Frozen isometry only* also removes the residual; both retain the BioLORD vectors rather than replacing them with random vectors. In every configuration, the modality embedding, temporal biases, output bias, and logit scale remain learned, so "only" in the configuration names refers to the event-code pathways. On OLMo 1B, we additionally run a within-modality permutation control, which reassigns BioLORD vectors among events of the same modality to break their correspondence to clinical concepts while preserving the vectors and the architecture (Appendix F), and a residual-rank sweep over $r \in \{0, 16, 64, 256, 512\}$ (Appendix G).

Optimization and metrics. EHRAdapt parameters are trained with AdamW, with checkpoints selected by validation cross-entropy (CE) and early stopping. We report test CE globally and within ten deciles of training frequency, from D1 (rarest) to D10 (most frequent). Results are mean ± SD across three seeds, with ablation effects paired within seeds. Training details and additional metrics are given in Appendices C and E.

4.3 BIOSURVEILLANCE TASKS AND COHORTS

Recognizing reportable infections and syndromic illness from clinical histories can support biosurveillance. Following the task designs of Goncalves et al. (2025), we evaluate whether EHR representations distinguish infectious diagnoses and gastrointestinal or respiratory syndromes before the index diagnosis. To be clear, these patient-level tasks assess a component of surveillance, they do not measure outbreak detection or alert lead time. Appendix D provides more information including all class-by-year counts, cumulative training counts, and disease-code definitions.

ICD-based syndromic classification. Syndromic surveillance usually relies on chief complaints (Henning, 2004); following Goncalves et al. (2025), we instead define syndromes by sets of nonspecific ICD codes. The four classes are gastrointestinal (GI) syndrome, respiratory syndrome, other known disease, and matched control, with approximately 118 thousand examples. We report the four-class macro average and the GI and respiratory one-versus-rest scores.

Reportable infectious-disease classification. Classes are defined by ICD codes for specific reportable infections. The 14 classes comprise seven named higher-frequency diseases, six clinically grouped rare-disease classes, and matched control, with approximately 19.1 thousand examples. We report All 14, Diseases 13 (excluding control), Frequent 7, and Rare 6, the principal view when disease labels are scarce. All views average class metrics from the same 14-way model rather than separately trained ones.

For each example, models receive the patient's clinical events from the two months preceding the index date, excluding the index date itself, and predict the example's class, a diagnosis or control label. We use an expanding temporal evaluation: models are trained on 2019 and tested on 2020, trained on 2019–2020 and tested on 2021, and trained on 2019–2021 and tested on 2022.

4.4 DOWNSTREAM MODELS AND BASELINES

We train ℓ_2-regularized logistic regression probes on each CPT model's hidden state at the last event before the index date. The regularization strength $C \in \{0.1, 1, 10\}$ is chosen per representation, task, and training window using patient-grouped validation, and the chosen probe is refit on the full window. To address class imbalance, we weighted each class by the inverse of its effective number of samples, computed from class frequencies in the training set, using $\beta = 0.999$ (Cui et al., 2019). The same protocol applies to the *Full model*, its ablations, and the baselines below.

Baselines. Because EHRAdapt is model-agnostic, our goal is not to find the best backbone or the best task-specific predictor, but to measure the quality of the learned representations for both rare and frequent diseases. We therefore keep all representations frozen, use the same probe for all, and choose baselines that each isolate one aspect of EHRAdapt:

- *Native text:* the same backbones read the patient history written as text, and the probe uses the hidden state of the last token. This holds the backbone fixed and changes only how events enter it.
- *BioLORD:* each event is represented by its frozen BioLORD embedding, and these are averaged into a patient vector, either uniformly (mean) or with weights 2^{-d_i}, where d_i is the event's age in days (decay, a one-day half-life). This tests whether clinical knowledge alone, without sequence modeling or EHR training, is sufficient.
- *Random forest:* trained on counts of clinical events in the same input window, with the number of trees tuned over $\{100, 200, \ldots, 500\}$. Such count-based models are strong baselines for structured EHR prediction (Wornow et al., 2023; Brown et al., 2025).

Medical language models such as GatorTronGPT (Peng et al., 2023) could serve as EHRAdapt backbones, and any of these representations could be further fine-tuned for each task; we leave both backbone comparison and task-specific fine-tuning outside our scope.

5 RESULTS

This section reports next-event prediction results from continued pretraining, downstream classification performance of EHRAdapt and baseline models, and ablations of EHRAdapt's components.

5.1 CONTINUED-PRETRAINING PERFORMANCE

Table 1 shows that EHRAdapt's *Full model* achieves the lowest global test CE among the five configurations on all three backbones. This ordering holds for all paired ablation vs *Full model* comparisons on metrics evaluated. Learning the semantic projection also matters: fixing it (*Frozen isometry*) worsens the global CE and on every data stratum compared with its learned counterpart. See Appendix E for all accuracy results and performance curves for all ablations on each data stratum.

Table 1: Global CPT test CE (nats, lower is better), mean ± sample SD across three seeds.

Configuration	OLMo2-1B	Llama3.2-1B	OLMo2-7B
Full model	$\mathbf{2.847} \pm 0.006$	$\mathbf{2.806} \pm 0.011$	$\mathbf{2.799} \pm 0.003$
Semantic prior only	3.217 ± 0.019	3.129 ± 0.006	3.090 ± 0.008
Residual embeddings only	2.959 ± 0.012	2.937 ± 0.014	2.976 ± 0.050
Frozen isometry + residual	3.158 ± 0.027	2.974 ± 0.015	3.057 ± 0.001
Frozen isometry only	4.689 ± 0.005	4.500 ± 0.017	4.513 ± 0.014

Figure 2 shows how test CE changes for each data stratum when each pathway is removed. Removing the semantic pathway (*Residual embeddings only*) hurts rare events most, and the penalty shrinks steadily as events become more frequent. Removing the residual (*Semantic prior only*) shows the opposite pattern: it improves the rarest events (D1–D3) but hurts events of medium to high frequency. The two curves cross around D5 on all three backbones, and both approach zero in D10, suggesting that abundant data make either pathway nearly sufficient on its own. As aggregate CE is dominated by frequent events, these large effects on rare events have little effect on the total loss.

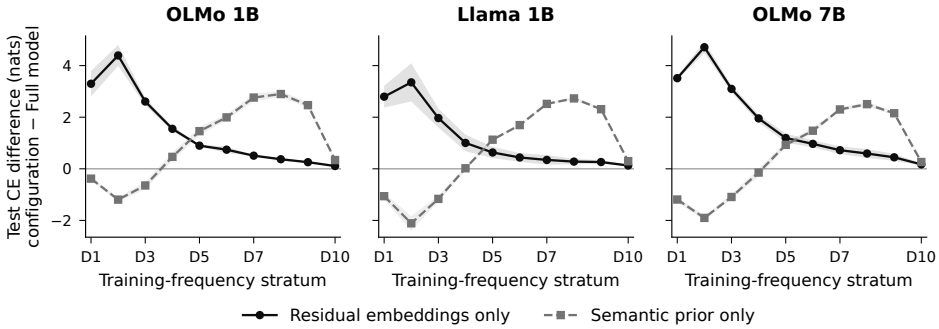

Figure 2: Ablation's test CE minus the *Full model*'s for each training frequency: positive values mean the ablation is worse. *Residual embeddings only* removes the semantic pathway, and *Semantic prior only* removes the residual. D1 is rarest and D10 most frequent. Curves show mean and ±1 sample SD over three seeds.

We also run a permutation control on OLMo2-1B that shuffles BioLORD embeddings within each modality. This keeps the set of vectors and their geometry unchanged but breaks the link between each event and its clinical meaning. The permuted model has higher test CE than the *Full model* globally and in every stratum, indicating that the semantic pathway helps because each event receives its correct clinical description, not merely because of how the vectors are distributed (Appendix F).

Table 2: Downstream AUPRC. Multiclass tasks use macro-AUPRC. Scores are mean ± sample SD over seeds (each seed averaged over 2020–2022).

Model	All 14	Diseases 13	Frequent 7	Rare 6
BioLORD (mean)	0.297	0.254	0.345	0.148
BioLORD (decay)	0.308	0.266	0.356	0.162
OLMo 1B, native text	0.266	0.221	0.303	0.126
Llama 1B, native text	0.261	0.215	0.302	0.113
OLMo 7B, native text	0.255	0.209	0.290	0.114
RF, tuned (count-based)	0.330	0.288	0.373	0.190
EHRAdapt - OLMo 1B	0.366 ± 0.004	0.326 ± 0.005	0.405 ± 0.002	0.233 ± 0.010
EHRAdapt - Llama 1B	$\mathbf{0.380 \pm 0.002}$	$\mathbf{0.341 \pm 0.002}$	$\mathbf{0.421 \pm 0.002}$	$\mathbf{0.247 \pm 0.004}$
EHRAdapt - OLMo 7B	0.378 ± 0.004	0.339 ± 0.004	$\mathbf{0.421 \pm 0.006}$	0.242 ± 0.009
Model	All 4	GI	Respiratory	
BioLORD (mean)	0.651	0.590	0.453	
BioLORD (decay)	0.669	0.605	0.491	
OLMo 1B, native text	0.690	0.586	0.566	
Llama 1B, native text	0.670	0.569	0.530	
OLMo 7B, native text	0.691	0.579	0.572	
RF, tuned (count-based)	0.700	0.667	0.486	
EHRAdapt - OLMo 1B	0.791 ± 0.005	0.681 ± 0.005	0.730 ± 0.009	
EHRAdapt - Llama 1B	0.797 ± 0.003	0.686 ± 0.005	$\mathbf{0.741 \pm 0.004}$	
EHRAdapt - OLMo 7B	$\mathbf{0.798 \pm 0.003}$	$\mathbf{0.688 \pm 0.005}$	0.740 ± 0.008	

5.2 Downstream predictive performance

Table 2 compares the *Full model* with baseline models. Overall, the *Full model* exceed the baselines in all scenarios. Tuned count-based tree models perform well in comparison to LLM-based baselines. CPT's *Full model* outperforms on every reported class view: Rare-6 AUPRC is 0.233–0.247 versus 0.190, and four-class syndromic AUPRC is 0.791–0.798 versus 0.700. The respiratory advantage over RF is larger than the GI advantage (0.244–0.254 versus 0.014–0.021).

The *Full model* outperforms both downstream task averages among the five EHRAdapt configurations, consistent with its global CPT ranking. However, better CPT likelihood does not always transfer: removing the residual improves rare-event likelihood but reduces Rare-6 AUPRC by 0.9–2.1 percentage points. Removing the semantic branch costs 3.6–5.1 Rare-6 points, versus 1.3–2.5 for Frequent 7, while freezing the projection with the residual retained costs 3.2–4.0 Rare-6 points (Table 3). These Rare-6 effects persist across backbones, supporting the contributions of both shared projection learning and event-specific adaptation to rare-disease discrimination.

5.3 Representation Geometry

We examine how EHRAdapt retains the semantic structure supplied by BioLORD while adapting event representations to EHR data. Figure 3 shows how local neighborhood and angular information is preserved at each data stratum, using the OLMo2-1B backbone. The learned semantic projection retains most of BioLORD's nearest neighbors (C) but substantially changes the angles between both neighboring and randomly paired codes (A, B). Thus, learning P_{sem} largely preserves which codes are most similar while changing their similarity values.

Adding the residual further reorganizes local neighborhoods, particularly for frequent events. However, it reduces average angular distortion relative to BioLORD in both local and global comparisons. Preserving similarity values and preserving neighbor rankings thus capture different aspects of the learned representations.

Table 3: *Full model* improvement over ablations (positive favors *Full model*): reduction in CPT test cross-entropy for ΔCE (nats) and gain in AUPRC for downstream columns (percentage points), paired by seed and year. GI/Resp. averages the two class AUPRCs. Full tables in Appendix H.

Configuration	ΔCE	Freq. 7	Rare 6	GI/Resp.
OLMo 1B				
Semantic prior only	0.371	+0.38	+1.91	+0.30
Residual embeddings only	0.112	+1.30	+3.60	+0.66
Frozen isometry + residual	0.312	+2.69	+3.30	+6.69
Frozen isometry only	1.842	+5.95	+8.48	+6.78
Llama 1B				
Semantic prior only	0.323	+0.54	+0.86	+1.82
Residual embeddings only	0.131	+2.16	+4.16	+1.95
Frozen isometry + residual	0.168	+1.43	+3.95	+1.66
Frozen isometry only	1.694	+4.43	+7.64	+3.15
OLMo 7B				
Semantic prior only	0.291	+1.06	+2.08	+0.39
Residual embeddings only	0.177	+2.47	+5.10	+0.72
Frozen isometry + residual	0.258	+1.47	+3.21	+1.61
Frozen isometry only	1.714	+6.52	+9.17	+4.33

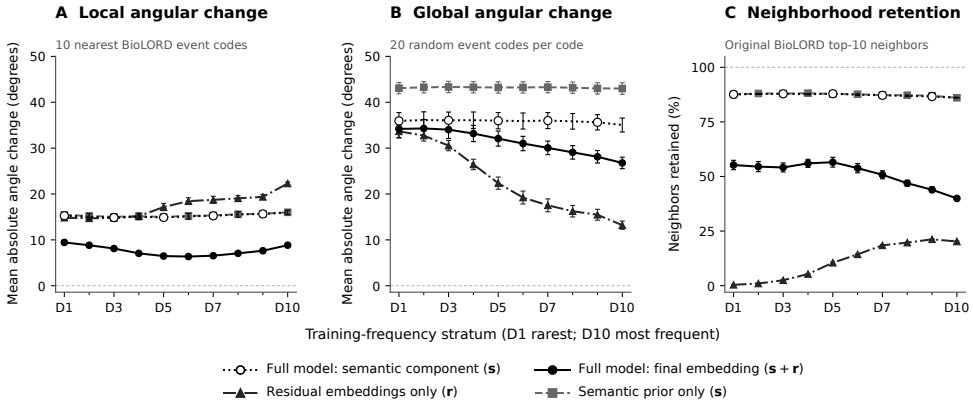

Figure 3: **BioLORD geometry preservation after CPT (OLMo2-1B).** Angular changes relative to frozen, mean-centered BioLORD are measured between each event-code embedding and those of (A) its 10 original nearest neighbors or (B) 20 randomly sampled codes. (C) Percentage of original top-10 neighbors retained. Curves show the *Full model*'s semantic component and final embedding, alongside the final embeddings of separately trained ablations. Each frequency stratum uses 1,000 shared event codes; measurements are averaged equally over codes. Points and error bars show the mean and sample SD across three training seeds.

6 Conclusion

We introduced EHRAdapt, an adapter that maps structured EHR events into frozen pretrained LLMs via a semantic prior and learned residual. Across three backbones, the full model had the best test likelihood. The semantic prior benefited rare events most, while the residual improved overall likelihood but degraded next-event prediction for the rarest events. Both components improved rare-disease classification, showing the need to evaluate across event frequencies and downstream tasks.

Continued pretraining on longitudinal EHR data produced representations that transferred effectively to downstream tasks and outperformed the evaluated baselines, including text-based inputs to the same LLMs. Limitations include the exclusion of laboratory and vital-sign values and demographic information, and evaluation within a single health system. Future work should incorporate these features and assess generalizability across healthcare systems.

Reproducibility statement

The SI specifies data processing, configurations, cohort definitions, tuning, and analysis protocols. Kaiser Permanente records cannot be released. Code and configuration will be available upon paper acceptance.

Ethics statement

Analyses use de-identified EHRs under institutional review and an approved data use agreement in the data holder's secure environment. Outcomes represent coded documentation, not biological onset. Practice and access patterns may affect performance; these results do not support clinical deployment.

The use of large language models

We used LLMs to refine the text of the final manuscript and AI coding assistants to support code development. The authors reviewed all outputs and take full responsibility for the manuscript. All numerical results were produced directly by our experimental pipeline.

References

Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew McDermott. Publicly available clinical BERT embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78, 2019.

Ofir Ben Shoham and Nadav Rappoport. Cpllm: Clinical prediction with large language models. *PLOS Digital Health*, 3(12):e0000680, 2024.

Olivier Bodenreider. The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Research*, 32(suppl 1):D267–D270, 2004.

Katherine E Brown, Chao Yan, Zhuohang Li, Xinmeng Zhang, Benjamin X Collins, You Chen, Ellen Wright Clayton, Murat Kantarcioglu, Yevgeniy Vorobeychik, and Bradley A Malin. Large language models are less effective at clinical prediction tasks than locally trained machine learning models. *Journal of the American Medical Informatics Association*, 32(5):811–822, 2025. doi: 10.1093/jamia/ocaf038. URL https://doi.org/10.1093/jamia/ocaf038.

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

Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. Class-balanced loss based on effective number of samples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9260–9269, 2019.

Andre R. Goncalves, Jose Cadena Pico, Yeping Hu, David Schlessinger, John Greene, Liam O'suilleabhain, Heather Clancy, Michael Vollmer, Vincent Liu, Tom Bates, and Priyadip Ray. AI-enabled diagnostic prediction within electronic health records to enhance biosurveillance and early outbreak detection. *medRxiv*, 2025. doi: 10.1101/2025.05.14.25327606. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC12132134/. Preprint, version 1.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, et al. The Llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

Lin Lawrence Guo, Adam Paul Yan, Emily Vettese, and Lillian Sung. PORTER: Language-grounded event representations for portable structured EHR foundation models, 2026. URL https://arxiv.org/abs/2606.24102. Preprint.

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, 2025. URL https://arxiv.org/abs/2502.17403.

Kelly J. Henning. What is syndromic surveillance? *MMWR Supplements*, 53:7–11, 2004.

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 (ICLR)*, 2022.

Hakan Inan, Khashayar Khosravi, and Richard Socher. Tying word vectors and word classifiers: A loss framework for language modeling. In *International Conference on Learning Representations (ICLR)*, 2017.

Lavender Yao Jiang, Xujin Chris Liu, Nima Pour Nejatian, Mustafa Nasir-Moin, Duo Wang, Anas Abidin, Kevin Eaton, Howard Antony Riina, Ilya Laufer, Paawan Punjabi, 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.

Ross Lazarus, Michael Klompas, Francis X. Campion, Scott J. N. McNabb, Xuanlin Hou, James Daniel, Gillian Haney, Alfred DeMaria, Leslie Lenert, and Richard Platt. Electronic support for public health: Validated case finding and reporting for notifiable diseases using electronic medical data. *Journal of the American Medical Informatics Association*, 16(1):18–24, 2009. doi: 10.1197/jamia.M2848.

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

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 (ML4H)*, volume 158 of *PMLR*, pages 239–260, 2021. URL https://proceedings.mlr.press/v158/pang21a.html.

Cheng Peng, Xi Yang, Aokun Chen, Kaleb E Smith, Nima PourNejatian, Anthony B Costa, Cheryl Martin, Mona G Flores, Ying Zhang, Tanja Magoc, Gloria Lipori, Duane A Mitchell, Naykky S Ospina, Mustafa M Ahmed, William R Hogan, Elizabeth A Shenkman, Yi Guo, Jiang Bian, and Yonghui Wu. A study of generative large language model for medical research and healthcare. *NPJ digital medicine*, 6(1):210, November 2023. doi: 10.1038/s41746-023-00958-w. URL http://doi.org.

Ofir Press and Lior Wolf. Using the output embedding to improve language models. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 157–163, 2017.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020.

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

François Remy, Kris Demuynck, and Thomas Demeester. Biolord: learning ontological representations from definitions for biomedical concepts and their textual descriptions. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 1454–1465, 2022.

François Remy, Kris Demuynck, and Thomas Demeester. BioLORD-2023: semantic textual representations fusing large language models and clinical knowledge graph insights. *Journal of the American Medical Informatics Association*, 31(9):1844–1855, 2024. doi: 10.1093/jamia/ocae029.

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.

Ethan Steinberg, Jason A. Fries, Yizhe Xu, and Nigam H. Shah. MOTOR: A time-to-event foundation model for structured medical records. In *International Conference on Learning Representations (ICLR)*, 2024. URL https://arxiv.org/abs/2301.03150.

Team OLMo, Pete Walsh, Luca Soldaini, Dirk Groeneveld, Kyle Lo, et al. 2 OLMo 2 furious, 2025. URL https://arxiv.org/abs/2501.00656. Shorter version accepted to COLM 2025.

Yuli Wang, Yuwei Dai, Robin Wang, Tej Mehta, Premal Trivedi, Thao Vu, Cheng Ting Lin, Li Yang, Zhicheng Jiao, Ihab Kamel, et al. Integrating large language models for enhanced predictive analytics in healthcare. *npj Digital Medicine*, 2026.

Shane Waxler, Paul Blazek, Davis White, Daniel Sneider, Kevin Chung, Mani Nagarathnam, Patrick Williams, Hank Voeller, Karen Wong, Matthew Swanhorst, Sheng Zhang, Naoto Usuyama, Cliff Wong, Tristan Naumann, Hoifung Poon, Andrew Loza, Daniella Meeker, Seth Hain, and Rahul Shah. Generative medical event models improve with scale, 2025. URL https://arxiv.org/abs/2508.12104.

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 (NeurIPS), Datasets and Benchmarks Track*, 2023. URL https://arxiv.org/abs/2307.02028.

Adam Wright and David W Bates. Distribution of problems, medications and lab results in electronic health records: the pareto principle at work. *Applied clinical informatics*, 1(01):32–37, 2010.

Xi Yang, Aokun Chen, Nima PourNejatian, Hoo Chang Shin, Kaleb E. Smith, Christopher Parisien, Colin Compas, Cheryl Martin, Mona G. Flores, Ying Zhang, et al. A large language model for electronic health records. *npj Digital Medicine*, 5:194, 2022.

Zhichao Yang, Avijit Mitra, Weisong Liu, Dan Berlowitz, and Hong Yu. Transformehr: transformer-based encoder-decoder generative model to enhance prediction of disease outcomes using electronic health records. *Nature communications*, 14(1):7857, 2023.

Yinghao Zhu, Zixiang Wang, Junyi Gao, Yuning Tong, Jingkun An, Weibin Liao, Ewen M Harrison, Liantao Ma, and Chengwei Pan. Prompting large language models for zero-shot clinical prediction with structured longitudinal electronic health record data. *arXiv preprint arXiv:2402.01713*, 2024.

SUPPLEMENTARY INFORMATION

A DATA AND PREPROCESSING

Table 4 summarizes the continued-pretraining cohort and its patient-level splits, and Figure 4 shows the frequency distribution of clinical events in the training split. Frequency strata are derived from training-split counts. The decile boundaries are 1, 1, 2, 5, 9, 20, 47, 120, 379, 2,050, and approximately 87.6M occurrences. The repeated boundary at one reflects tied singleton frequencies, so D1–D10 are not equal-sized bins, and all reported metrics use exact bin memberships. The vocabulary has 78,384 clinical event types and four reserved tokens, for $|\mathcal{V}|$ = 78,388.

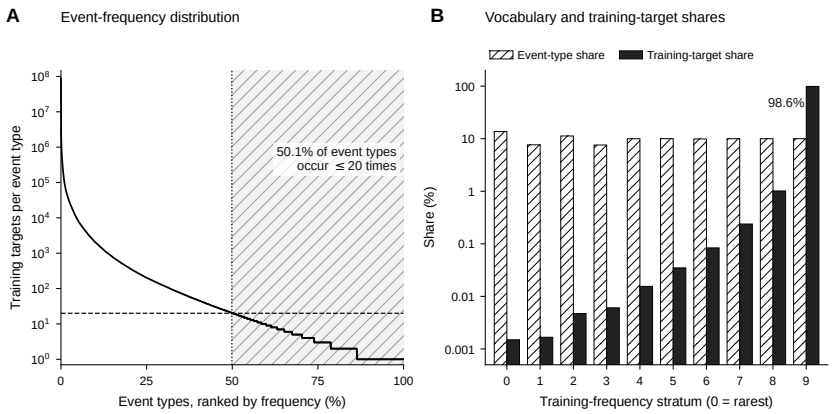

Figure 4: Clinical event frequencies in the KPNC continued-pretraining corpus. Counts refer to the training split and exclude reserved tokens and end-of-sequence targets.

Table 4: Continued-pretraining cohort and splits. Patients with at least one clinical event. Chunking retains all 732M clinical events. Event types by modality: diagnosis 44,079; procedure 28,401; laboratory test 2,853; chief complaint 2,174; medication 817; encounter type 59; vital signs 1.

Split	Fraction	Patients	Chunks	Clinical events
Train	98%	3.95M	4.46M	717M
Validation	1%	40,197	45,484	7.32M
Test	1%	39,734	45,144	7.32M

B BACKBONES AND PARAMETER COUNTS

Table 5 lists the three frozen backbones. The mean norm of each backbone's pretrained token embeddings sets the initialization scale of the semantic projection (Appendix C). Table 6 gives parameter counts: EHRAdapt trains 0.1–0.6% of the parameters of the complete model.

Table 5: Frozen backbones. H: hidden size; Emb. norm: mean norm of the pretrained token embeddings, which sets the initialization scale of P_{sem}.

Backbone	Checkpoint	H	Layers	Heads	KV heads	Emb. norm
OLMo2-1B	`allenai/OLMo-2-0425-1B`	2,048	16	16	16	10.724
Llama3.2-1B	`meta-llama/Llama-3.2-1B`	2,048	16	32	8	0.987
OLMo2-7B	`allenai/OLMo-2-1124-7B`	4,096	32	32	32	7.423

Table 6: Parameter counts for the full EHRAdapt configuration (78,388-event vocabulary, linear semantic projection, rank-64 projected residual, tied prediction head, frozen backbone). The original OLMo2 text models have separate language-model heads, which EHRAdapt replaces; Llama3.2 ties its head to its input embeddings, so its two backbone counts are identical. In panel B, V is the vocabulary size, r the residual rank, H the hidden size, L and A the numbers of layers and attention heads, and K the number of elapsed-time buckets; 768 is the BioLORD embedding dimension.

	Size	OLMo2-1B	Llama3.2-1B	OLMo2-7B
A. Totals				
Original text LM, including LM head		1,484,916,736	1,235,814,400	7,298,617,344
Frozen backbone used by EHRAdapt		1,279,395,840	1,235,814,400	6,887,575,552
Trainable EHRAdapt parameters		6,821,685	6,825,781	8,556,341
Complete EHRAdapt model		1,286,217,525	1,242,640,181	6,896,131,893
Trainable fraction		0.53%	0.55%	0.12%
B. Trainable parameters by component				
Semantic projection	$768H + H$	1,574,912	1,574,912	3,149,824
Event residual table	Vr	5,016,832	5,016,832	5,016,832
Residual projection	rH	131,072	131,072	262,144
Modality embeddings	$8H$	16,384	16,384	32,768
Temporal attention biases	LAK	4,096	8,192	16,384
Prediction bias and logit scale	$V + 1$	78,389	78,389	78,389
Total trainable		6,821,685	6,825,781	8,556,341

C Training configuration

Table 7 lists the training settings. The backbone receives no gradient updates, so only EHRAdapt parameters are optimized. Within each backbone, all configurations use the same settings. Runs use AMD MI300A accelerators under ROCm.

Initialization. The semantic projection P_{sem} is a learned linear map with a scaled semi-orthogonal initialization. The scale is set so that projected semantic vectors match the mean norm of the backbone's pretrained token embeddings (Table 5), so event inputs enter the transformer at a magnitude comparable to its native token inputs. This norm serves only as an initialization target: P_{sem}, $\mathbf{A}$, $\mathbf{B}$, and $\mathbf{e}_{\text{mod}}$ are all trained during CPT. The output bias β_c is initialized to $\log p(c)$ from training-split counts, which lowers the initial CE by 4.899 nats relative to a uniform initialization ($\log V = 11.270$ versus a unigram entropy of 6.371). The logit scale s is calibrated at initialization by one-dimensional minimization on a training-only calibration subset, separately for each configuration, because different initializations produce differently scaled logits.

BioLORD's embedding processing. Each event's BioLORD embedding is obtained by mean pooling the token embeddings of its description and L2-normalizing the result. These embeddings are then centered by subtracting their mean over all clinical event types, without whitening.

D Downstream cohorts

D.1 Class counts

Tables 8 and 9 give class counts by year, and Table 10 gives the training and test sizes of each evaluation window. All positive cases are retained. Controls are capped at one demographically matched control per case, so the control-to-case ratio can fall below 1:1. Annual columns are disjoint sets of examples, and each training window accumulates all earlier years.

D.2 Class definitions

Table 11 lists the ICD codes that define each class. In syndromic classification, the Known class contains patients with a case diagnosis outside the two syndromic code lists, sampled at random with a fixed seed. Control patients are matched to patients in the disease classes by demographics

Table 7: Training configuration. Settings in the upper block are shared by all backbones.

Setting	OLMo2 1B	Llama3.2 1B	OLMo2 7B
Backbone optimization	Frozen		
Optimizer	AdamW		
Precision; attention	BF16; `sdpa`		
Effective batch	128 chunks (no gradient accumulation)		
Learning rate: output/temporal	$1e{-4}$ for vocabulary bias, logit scale, and temporal-attention bias		
Learning rate: event encoder	$3e{-4}$ for semantic projection, residual parameters, and modality embeddings		
Weight decay	0.1 on semantic and residual projection weights; 0 otherwise		
Learning-rate schedule	1% warmup, constant, final 10% decay to $0.1\times$		
Gradient clipping	Global norm 1.0		
Validation interval	5,000 steps		
Checkpoint selection	Target-weighted global validation CE		
Early stopping	Patience 3; $\Delta_{\min} = 0.005$ nats		
Validation CE ceiling	6.371 nats (unigram entropy)		
Length-bucket boundaries	32, 64, 128, 256, 384, 512, 527		
Training seeds; split seed	101, 202, 303; 17		
Maximum valid-target budget	4.326×10^9		
Maximum data passes	6		
Scheduler horizon (steps)	213,449		
Steps per epoch	$\sim$35.6K		
Nodes / GPU ranks	4 / 16	4 / 16	8 / 32
Micro-batch per rank	8	8	4

Table 8: Syndromic classification case counts per year. Each test year uses all preceding years for training.

Class	2019	2020	2021	2022	Total
Gastrointestinal	1,420	742	808	1,150	4,120
Respiratory	9,474	6,943	3,765	6,899	27,081
Known	10,894	7,685	4,573	8,049	31,201
Control	19,447	13,524	8,261	14,376	55,608
All positive cases	21,788	15,370	9,146	16,098	62,402
Total	**41,235**	**28,894**	**17,407**	**30,474**	**118,010**

and geographic location. The sizes of the Known and Control classes were set to keep the classes balanced.

E Additional CPT results

All results in this section use held-out test patients and, for each run, the checkpoint selected by global validation CE; values are means and sample SDs over three seeds. Table 12 gives global CE and top-1 and top-10 accuracy, and Figure 5 shows each metric by frequency stratum for every configuration and backbone.

Frequency strata are deciles of event types ranked by training-split frequency (Appendix A). Each stratum's CE is averaged over its test targets. Ablation effects are the ablation's CE minus the Full model's CE within the same seed. SDs reflect variability across training seeds, not uncertainty from sampling test patients.

F Permutation control

On OLMo2-1B, we compare the *Full model* with the same model after shuffling BioLORD vectors among events of the same modality. The control keeps the set of vectors in each modality and the architecture unchanged, but breaks the correspondence between each event and its clinical description.

Table 9: Infectious-disease classification case counts per year. Each test year uses all preceding years for training. Rare-6 diseases are identified by †.

Class	2019	2020	2021	2022	Total
Campylobacteriosis	825	448	479	634	2,386
Salmonellosis (nontyphoidal)	590	357	298	481	1,726
Chickenpox (varicella)	386	255	178	218	1,037
Shigellosis	419	180	246	360	1,205
Pertussis	486	97	28	30	641
Invasive *H. influenzae*	190	118	63	107	478
Viral meningitis	126	93	103	93	415
†Rare vector/travel systemic	142	66	60	80	348
†Rare enteric diarrheal	99	81	70	71	321
†Rare vaccine preventable respiratory/neurologic	129	65	59	45	298
†Rare environmental food/water	79	65	73	67	284
†Hepatitis A	80	71	51	52	254
†Rare zoonotic/tick borne high consequence	59	51	59	49	218
Control	3,582	1,903	1,756	2,277	9,518
All positive cases	3,610	1,947	1,767	2,287	9,611
Total	**7,192**	**3,850**	**3,523**	**4,564**	**19,129**

Table 10: Full-refit training and held-out annual test counts for syndromic and infectious disease tasks.

	Test 2020		Test 2021		Test 2022	
Task	Train	Test	Train	Test	Train	Test
Syndromic	41,235	28,894	70,129	17,407	87,536	30,474
Infectious	7,192	3,850	11,042	3,523	14,565	4,564

Unlike removing the semantic pathway, it keeps the projection parameters. Results use held-out test CE at validation-selected checkpoints (Figure 6).

Global test CE is 2.847 ± 0.006 for the Full model and 2.888 ± 0.014 for the permuted model, a paired reduction of 0.042 ± 0.020 nats. The Full model is better in every stratum and every seed. Its advantage is much larger in D1 and D2 (2.406 ± 0.098 and 3.569 ± 0.266 nats) than in D10 (0.039 ± 0.019), with the largest gap in D2. The most frequent stratum holds approximately 98.60% of test targets, which is why these large effects on rare events carry little weight in the global average.

This comparison supports a role for correct semantic correspondence beyond the distribution of the vectors. It is limited to OLMo2-1B at rank 64. Selected checkpoints differ between conditions (*Full model*: 130K, 115K, and 150K steps; permuted: 135K, 160K, and 180K), so the effect reflects the shared selection rule rather than equal training exposure. Three seeds do not measure uncertainty over possible permutations.

G Residual rank

On OLMo2-1B, we keep the semantic pathway and compare residual ranks 0, 16, 64, 256, and 512, with three seeds each (Figure 7). This sweep characterizes a capacity trade-off on validation data and is reported separately from the held-out test results.

Global validation CE decreases from 3.224 at rank 0 to 2.709 at rank 512, with the same ordering in each seed. Rank 64 uses 6.82M trainable parameters versus 42.9M at rank 512, about 16% as many, while capturing 72% of the global improvement from rank 0 to rank 512. Increasing rank therefore improves aggregate fit, with diminishing gains per added parameter. The stratified curves qualify this pattern: larger ranks improve many middle and frequent strata, but the two rarest strata do not improve consistently with rank, and rank 0 remains competitive in the extreme tail despite its much

Table 11: ICD codes defining the classes in the two downstream tasks. Ranges are inclusive.

Class	ICD-9-CM	ICD-10-CM
Syndromic classification		
Gastrointestinal	–	A02.9, A03.0–A03.3, A03.8–A03.9, A04.5, A05.1, A05.3, A05.5, A32.11, A32.12, A32.7, A32.89, A32.9, B96.21
Respiratory	460, 461.9, 465.8, 465.9, 466.0, 486, 487.0, 487.1, 487.8, 488.01, 488.02, 488.09, 488.11, 488.12, 488.81, 488.82, 488.89, 490	J00, J01.90, J06.9, J09.X1–J09.X3, J09.X9, J10.00, J10.01, J10.08, J10.1, J10.2, J10.81, J10.89, J11.00, J11.08, J11.1, J11.2, J11.89, J12.89, J12.9, J18.1, J18.8, J18.9, J20.9, J40
Known	Any case diagnosis code outside the two syndromic code lists	
Control	Matched patients (same demographics and geographic location) with none of the codes above	
Infectious-disease classification		
Campylobacteriosis	008.43	A04.5
Salmonellosis (nontyphoidal)	003.0, 003.1, 003.20, 003.23, 003.29, 003.9	A02.0, A02.1, A02.20, A02.22–A02.25, A02.29, A02.8, A02.9
Chickenpox (varicella)	052.0–052.2, 052.7–052.9	B01.0, B01.11, B01.2, B01.81, B01.89, B01.9
Shigellosis	004.0–004.3, 004.8–004.9	A03.0–A03.3, A03.8–A03.9
Pertussis	033.0, 033.1, 033.9	A37.00, A37.01, A37.10, A37.11, A37.80, A37.90, A37.91
Invasive *H. influenzae*	041.5, 482.2	A49.2, B96.3, J14, J20.1
Viral meningitis	047.0, 047.1, 047.8, 047.9	A87.0, A87.2, A87.8, A87.9
Rare vector/travel systemic	061, 084.0, 084.1, 084.4–084.6, 084.9, 066.40–066.42, 066.49, 065.4, 062.0, 062.2, 062.3, 062.9	A90, A91, B50.0, B50.8, B50.9, B51.9, B52.0, B52.8, B52.9, B53.8, B54, A92.0, A92.30–A92.32, A92.39, A92.5, A83.0, A83.2, A83.3, A83.5, A83.6, A83.8, A95.1
Rare enteric diarrheal	283.11, 041.41–041.43, 002.0, 007.4, 005.4, 005.81, 001.0, 001.9, 008.44, 002.1–002.3, 002.9	D59.3, D59.30, D59.31, B96.21–B96.23, A01.00, A01.02, A01.03, A01.09, A07.2, A05.3, A05.5, A00.0, A00.9, A04.6, A01.1, A01.3, A01.4
Rare vaccine preventable respiratory/neurologic	045.00, 045.01, 045.10, 045.90, 046.11, 046.19, 037, 055.79, 055.8, 055.9, 036.0–036.2, 036.42, 036.81, 036.82, 036.89, 036.9, 032.0, 032.3, 032.81, 032.83–032.85, 032.89, 032.9	B06.02, B06.89, B06.9, P35.0, Z20.4, A80.30, A80.39, A80.9, B26.0, B26.84, B26.9, A81.00, A81.01, A81.09, A33–A35, B05.2, B05.89, B05.9, A39.0, A39.1, A39.2, A39.4, A39.81, A39.83, A39.9, A36.0, A36.2, A36.3, A36.84, A36.89, A36.9
Rare environmental food/water	040.89, 482.84, 124, 005.1, 027.0	A48.1, A48.2, B75, A05.1, A32.11, A32.12, A32.7, A32.89, A32.9, T61.01XA, T61.11XA, T61.14XA
Hepatitis A	070.0, 070.1	B15.0, B15.9
Rare zoonotic/tick borne high consequence	083.0, 087.0, 087.1, 087.9, 023.0, 023.8, 023.9, 088.82, 065.3, 066.1, 082.9, 083.2, 083.8, 083.9, 071, 082.40, 082.41, 082.49, 021.0, 021.9, 020.2, 020.8, 082.0, 022.0, 022.3, 073.0, 073.8, 073.9	A78, A68.0, A68.1, A68.9, A23.0, A23.1, A23.3, A23.8, A23.9, B60.00, B60.02, B60.10, B60.11, B60.13, B60.19, B60.8, A75.2, A75.3, A77.8, A77.9, A79.1, A79.81, A79.89, A79.9, A82.0, A82.1, A82.9, A77.40, A77.41, A77.49, A21.2, A21.7, A21.8, A21.9, A20.0, A20.7, A20.8, A20.9, A77.0, A98.4, A98.5, A22.7, A70
Control	Matched patients (same demographics and geographic location) with none of the codes above	

worse global CE. Extra residual capacity therefore does not guarantee better prediction across the whole vocabulary.

H Downstream results

This section gives the full downstream results summarized in Table 3: paired component effects with SDs (Table 13) and absolute scores for every configuration (Tables 14 and 15).

Class views. All 14, Diseases 13, Frequent 7, and Rare 6 average one-versus-rest class metrics from the same 14-way model. The views change which classes enter the macro average, not which patients are evaluated; in particular, Rare 6 is neither a separately trained classifier nor an evaluation restricted to rare-disease patients. Diseases 13 equals $(7\,\mathrm{AUPRC}_7 + 6\,\mathrm{AUPRC}_6)/13$. The GI and respiratory scores are one-versus-rest scores from the four-way syndromic model, and GI/Resp. in Table 3 averages the two.

Averaging and uncertainty. For each seed, metrics are averaged with equal weight over 2020, 2021, and 2022, and we report the mean and sample SD of these seed averages. Paired effects subtract ablation scores from *Full model* scores within each seed and year before the same aggregation. SDs describe variability across training seeds, not confidence intervals for new patients, and no significance tests are computed. Baselines do not depend on CPT seeds and are run once, so they have no SD.

Table 12: Global held-out CPT test metrics for **Full model** and ablations. Mean ± sample SD over three seeds per configuration

Configuration	CE	Top-1 (%)	Top-10 (%)
OLMo2-1B			
Full model	2.847 ± 0.006	40.201 ± 0.151	75.433 ± 0.077
Semantic prior only	3.217 ± 0.019	36.653 ± 0.269	71.378 ± 0.246
Residual embeddings only	2.959 ± 0.012	39.431 ± 0.126	73.649 ± 0.154
Frozen isometry + residual	3.158 ± 0.027	36.568 ± 0.412	70.913 ± 0.419
Frozen isometry only	4.689 ± 0.005	23.342 ± 0.083	50.138 ± 0.108
Llama3.2-1B			
Full model	2.806 ± 0.011	40.857 ± 0.118	75.984 ± 0.168
Semantic prior only	3.129 ± 0.006	37.948 ± 0.164	72.589 ± 0.068
Residual embeddings only	2.937 ± 0.014	39.713 ± 0.208	73.955 ± 0.145
Frozen isometry + residual	2.974 ± 0.015	39.236 ± 0.220	73.477 ± 0.206
Frozen isometry only	4.500 ± 0.017	25.939 ± 0.176	53.331 ± 0.257
OLMo2-7B			
Full model	2.799 ± 0.003	40.941 ± 0.049	76.108 ± 0.037
Semantic prior only	3.090 ± 0.008	38.401 ± 0.073	73.151 ± 0.129
Residual embeddings only	2.976 ± 0.050	39.267 ± 0.675	73.375 ± 0.771
Frozen isometry + residual	3.057 ± 0.001	38.054 ± 0.009	72.342 ± 0.038
Frozen isometry only	4.513 ± 0.014	24.967 ± 0.200	52.699 ± 0.246

Interpreting component effects. Comparing the *Full model* with *Residual embeddings only* removes shared semantic information together with the projection parameters; comparing with *Frozen isometry + residual* keeps the descriptions and residual but removes projection learning; and comparing with *Semantic prior only* removes the event-specific residual. The permutation control was run for CPT only (Appendix F), so the downstream results do not isolate the effect of correct event-to-description correspondence. Component effects also need not add up. For Rare 6, the residual's benefit under a learned projection minus its benefit under a fixed projection is -3.27, -2.82, and -3.88 AUPRC points for OLMo2-1B, Llama3.2-1B, and OLMo2-7B: the residual helps less when the projection is learned, so the two components partly compensate for each other.

Table 13: Paired downstream advantages of *Full model* over ablations in AUPRC percentage points (mean ± sample SD of three seed-level differences).

Configuration	Frequent 7	Rare 6	GI/Resp.
OLMo2-1B			
Semantic prior only	$+0.38 \pm 0.16$	$+1.91 \pm 1.14$	$+0.30 \pm 0.53$
Residual embeddings only	$+1.30 \pm 0.84$	$+3.60 \pm 1.18$	$+0.66 \pm 0.79$
Frozen isometry + residual	$+2.69 \pm 2.01$	$+3.30 \pm 2.70$	$+6.69 \pm 2.03$
Frozen isometry only	$+5.95 \pm 0.22$	$+8.48 \pm 1.35$	$+6.78 \pm 2.33$
Llama3.2-1B			
Semantic prior only	$+0.54 \pm 0.39$	$+0.86 \pm 1.02$	$+1.82 \pm 0.71$
Residual embeddings only	$+2.16 \pm 0.75$	$+4.16 \pm 0.93$	$+1.95 \pm 2.05$
Frozen isometry + residual	$+1.43 \pm 0.19$	$+3.95 \pm 0.45$	$+1.66 \pm 1.13$
Frozen isometry only	$+4.43 \pm 0.58$	$+7.64 \pm 0.34$	$+3.15 \pm 0.79$
OLMo2-7B			
Semantic prior only	$+1.06 \pm 0.49$	$+2.08 \pm 0.55$	$+0.39 \pm 0.62$
Residual embeddings only	$+2.47 \pm 0.63$	$+5.10 \pm 1.26$	$+0.72 \pm 0.80$
Frozen isometry + residual	$+1.47 \pm 0.50$	$+3.21 \pm 0.62$	$+1.61 \pm 1.46$
Frozen isometry only	$+6.52 \pm 0.72$	$+9.17 \pm 0.54$	$+4.33 \pm 0.47$

Table 14: Infectious-disease classification: absolute macro-AUPRC values for the *Full model* and its ablations. Values are mean ± sample SD across three seed-level annual averages.

Configuration	All 14	Diseases 13	Frequent 7	Rare 6
OLMo2-1B				
Full model	0.366 ± 0.004	0.326 ± 0.005	0.405 ± 0.002	0.233 ± 0.010
Semantic prior only	0.355 ± 0.002	0.315 ± 0.003	0.402 ± 0.001	0.214 ± 0.005
Residual embeddings only	0.344 ± 0.005	0.302 ± 0.006	0.392 ± 0.007	0.197 ± 0.004
Frozen isometry + residual	0.338 ± 0.020	0.296 ± 0.021	0.379 ± 0.019	0.200 ± 0.024
Frozen isometry only	0.298 ± 0.001	0.255 ± 0.001	0.346 ± 0.001	0.148 ± 0.004
Llama3.2-1B				
Full model	0.380 ± 0.002	0.341 ± 0.002	0.421 ± 0.002	0.247 ± 0.004
Semantic prior only	0.373 ± 0.004	0.334 ± 0.005	0.416 ± 0.002	0.239 ± 0.012
Residual embeddings only	0.351 ± 0.010	0.310 ± 0.010	0.400 ± 0.008	0.206 ± 0.013
Frozen isometry + residual	0.355 ± 0.001	0.315 ± 0.001	0.407 ± 0.003	0.208 ± 0.005
Frozen isometry only	0.324 ± 0.003	0.282 ± 0.003	0.377 ± 0.006	0.171 ± 0.000
OLMo2-7B				
Full model	0.378 ± 0.004	0.339 ± 0.004	0.421 ± 0.006	0.242 ± 0.009
Semantic prior only	0.363 ± 0.003	0.323 ± 0.003	0.410 ± 0.001	0.222 ± 0.006
Residual embeddings only	0.343 ± 0.001	0.302 ± 0.001	0.396 ± 0.001	0.191 ± 0.003
Frozen isometry + residual	0.356 ± 0.003	0.316 ± 0.004	0.406 ± 0.002	0.210 ± 0.005
Frozen isometry only	0.304 ± 0.003	0.261 ± 0.002	0.356 ± 0.003	0.151 ± 0.006

Table 15: Syndromic classification: absolute macro-AUPRC (multiclass) and AUPRC (GI and Respiratory) values for the *Full model* and its ablations. Values are mean ± sample SD across three seed-level annual averages.

Configuration	All 4	GI	Respiratory
OLMo2-1B			
Full model	0.791 ± 0.005	0.681 ± 0.005	0.730 ± 0.009
Semantic prior only	0.787 ± 0.003	0.679 ± 0.008	0.726 ± 0.004
Residual embeddings only	0.786 ± 0.000	0.670 ± 0.003	0.727 ± 0.001
Frozen isometry + residual	0.742 ± 0.022	0.649 ± 0.008	0.628 ± 0.046
Frozen isometry only	0.742 ± 0.013	0.596 ± 0.020	0.679 ± 0.020
Llama3.2-1B			
Full model	0.797 ± 0.003	0.686 ± 0.005	0.741 ± 0.004
Semantic prior only	0.783 ± 0.003	0.677 ± 0.006	0.713 ± 0.006
Residual embeddings only	0.783 ± 0.017	0.665 ± 0.018	0.723 ± 0.031
Frozen isometry + residual	0.783 ± 0.010	0.670 ± 0.006	0.723 ± 0.021
Frozen isometry only	0.771 ± 0.009	0.630 ± 0.010	0.733 ± 0.013
OLMo2-7B			
Full model	0.798 ± 0.003	0.688 ± 0.005	0.740 ± 0.008
Semantic prior only	0.794 ± 0.006	0.683 ± 0.015	0.738 ± 0.007
Residual embeddings only	0.793 ± 0.001	0.677 ± 0.001	0.737 ± 0.005
Frozen isometry + residual	0.786 ± 0.007	0.677 ± 0.008	0.720 ± 0.012
Frozen isometry only	0.765 ± 0.005	0.627 ± 0.008	0.715 ± 0.005

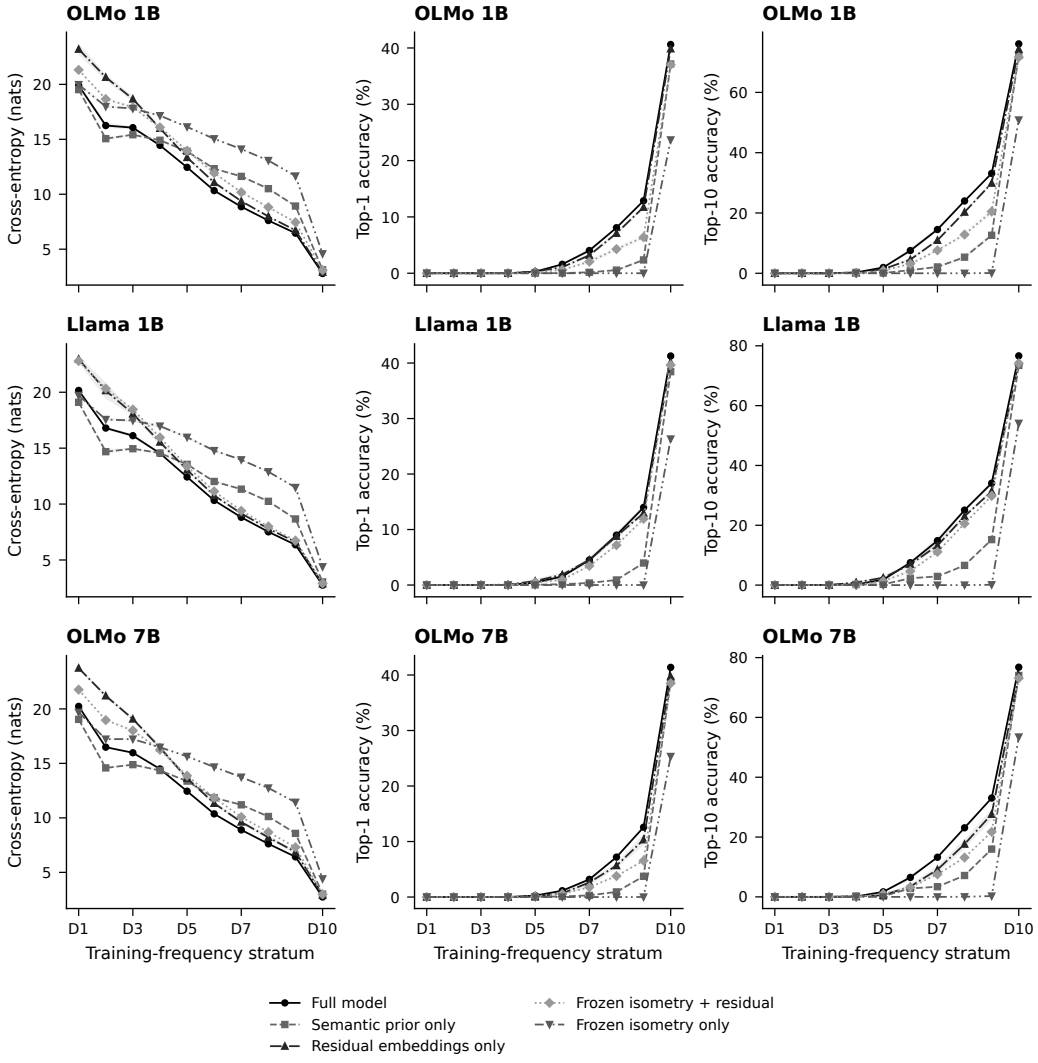

Figure 5: **Absolute CPT test performance by training frequency.** Rows correspond to OLMo2-1B, Llama3.2-1B, and OLMo2-7B; columns show CE, top-1, and top-10 accuracy. D1 contains the rarest event types, D10 the most frequent. Lines and bands show the mean and one sample SD across three seeds. All metrics use the same validation-selected checkpoint within each run.

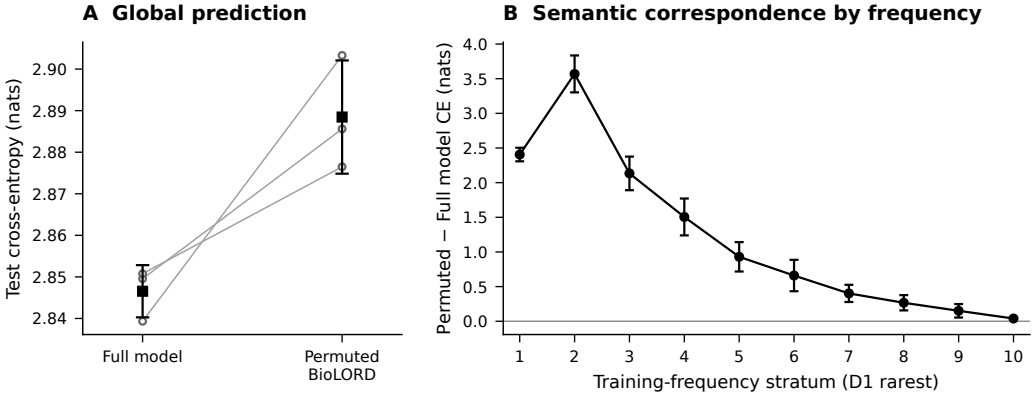

Figure 6: **Meaningful BioLORD assignments improve OLMo2-1B test CE.** (A) Global CE: gray lines connect matched training seeds; black squares and error bars show condition means and sample SDs. (B) Permuted minus Full-model CE by training-frequency stratum. Points and error bars show means and sample SDs of the three paired differences. Positive values favor Full model. D1 is rarest and D10 most frequent.

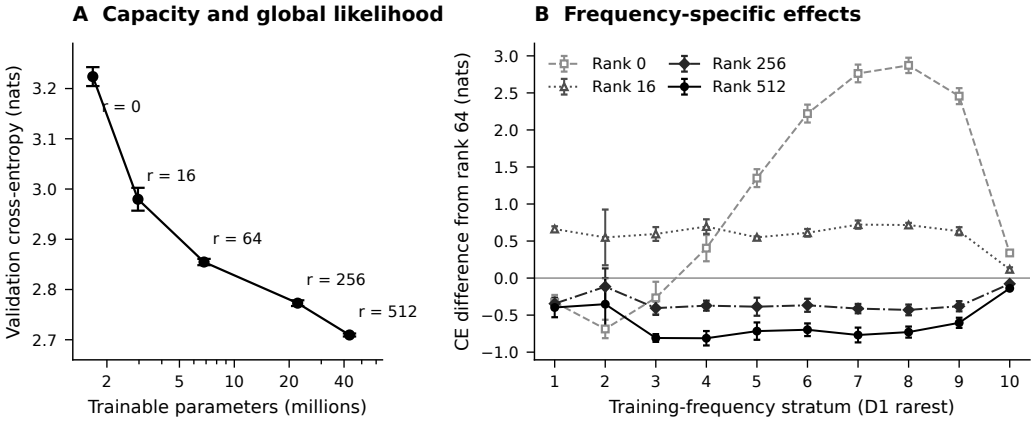

Figure 7: **OLMo2-1B residual capacity and validation likelihood.** (A) Global validation CE versus total trainable adaptation parameters on a logarithmic parameter axis. Error bars are sample SD over three seeds. (B) Per-stratum CE minus rank-64 CE, paired by seed; error bars show SDs of paired differences. Negative values improve on rank 64. D1 is rarest and D10 most frequent.