SparseEHR: Scalable Foundation Modeling for Structured EHR via Conditional Computation

Anonymous Authors[1]

Abstract

Structured electronic health records (EHRs) are a natural substrate for healthcare foundation models, but dense transformers remain expensive to scale across heterogeneous code vocabularies, irregular longitudinal records, and very large patient populations. We present **SparseEHR**, a hybrid dense-to-sparse transformer for structured EHR sequences that uses dense warm-start layers followed by mixture-of-experts (MoE) layers with top-2 routing and a shared expert pathway. SparseEHR is pretrained on longitudinal diagnosis and procedure sequences from approximately 50 million de-identified individuals in the OptumLabs Data Warehouse. In strictly zero-shot transfer to MIMIC-IV, without any fine-tuning, SparseEHR achieves 0.463 Recall@10 and 0.551 Recall@20 for next-visit ICD-10 prediction, outperforming recent public baselines. The selected hybrid configuration also reduces active parameters per token from 530M to 470M and training step time from 1.889s to 1.682s relative to an all-MoE variant, showing that conditional computation can improve transfer while lowering per-token compute for structured health data.

1. Introduction

Structured health data lies at the center of clinical machine learning: diagnoses, procedures, medications, labs, and physiologic measurements define longitudinal patient trajectories and support forecasting, risk stratification, and operational decision-making. A key problem is next-event prediction from large longitudinal EHR, where anticipating future diagnoses enables earlier risk stratification, care planning, and population-level resource allocation. However, existing dense models become costly as patient histories, vocabularies, and datasets scale. Recent transformer models for structured EHR, including BEHRT, Med-BERT, CEHR-BERT, and more recent large-scale pretrained models (Tang et al., 2023; Wornow et al., 2024), have improved representation learning and prediction, but scaling dense architectures remains expensive for long sequences and large corpora (Li et al., 2020; Rasmy et al., 2021; Pang et al., 2021; Du et al., 2022; Jiang et al., 2024).

Three properties of structured EHR make this challenge particularly acute. First, timelines are *irregular*: clinically meaningful events are grouped into visits with variable time gaps. Second, the token space is *heterogeneous*: common chronic-disease codes coexist with rare diagnoses, procedures, and workflow markers. Third, deployment is constrained by *compute*: practical models must serve long histories and large cohorts without incurring the full cost of dense billion-parameter models at every token.

Mixture-of-experts (MoE) architectures offer a compelling alternative by decoupling model capacity from per-token compute through conditional routing (Shazeer et al., 2017; Fedus et al., 2022). However, naive sparse routing can be unstable, fragment shared structure, and perform poorly on heterogeneous clinical data. The central question is therefore: *can conditional computation enable a scalable **structured-EHR** foundation model without sacrificing transfer?*

We answer yes with **SparseEHR**, a hybrid transformer for structured EHR sequences. The model uses dense layers to learn shared patient-state representations, followed by sparse experts that specialize on heterogeneous medical code patterns. This paper makes two key contributions: **(i)** strong *zero-shot cross-system transfer* from OptumLabs Data Warehouse (OLDW) to MIMIC-IV, and **(ii)** a clear *efficiency gain* from hybrid conditional computation. Together, these results position SparseEHR as a practical approach for structured health data modeling, where transfer, irregular longitudinal structure, and scalable deployment are critical.

2. SparseEHR for Structured EHR

Input representation. Each patient is represented as a chronological sequence of typed tokens built from demo-

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country.**AUTHORERR: Missing \icmlcorrespondingauthor.**

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

graphics and structured clinical codes. We prefix the sequence with static attributes (e.g., age bucket, sex, region), then append daily bundles of standardized diagnosis and procedure tokens with explicit delimiters. ICD codes are normalized to canonical prefixes and CPT codes to fixed numeric forms, producing a compact structured vocabulary while retaining clinically meaningful granularity.

Task formulation. Given a patient timeline $x = (x_1, \ldots, x_t)$, SparseEHR is trained to predict the next structured token autoregressively:

$$p(x_{t+1} \mid x_{\leq t}) = \mathrm{softmax}(W h_t). \tag{1}$$

At evaluation time, we roll the model forward to generate the next visit bundle token by token and score whether the true ICD-10 codes appear within the top-K predictions. This formulation lets one pretrained model support both representation learning and clinically meaningful next-event forecasting.

Hybrid dense–MoE architecture. SparseEHR is a 16-layer transformer with hidden size 2048 and a vocabulary of 20,656 structured code tokens. The first four layers are standard dense transformer blocks; the remaining twelve are MoE blocks with four experts and top-1 routing as well as one general expert per token. For token representation h_i, the MoE output is

$$y_i = \sum_{j \in \mathrm{TopK}(g(h_i),1)} w_{ij}\, \mathrm{Expert}_j(h_i), \tag{2}$$

where $g(h_i)$ is the routing network and w_{ij} are normalized gating weights. We add a load-balancing auxiliary loss to avoid expert collapse and include a shared expert pathway so that ubiquitous clinical patterns do not fragment across specialists. The general expert is responsible of having global view to avoid knowledge sparsity.

Why conditional computation is a natural fit for structured EHR. Structured EHR is not uniformly complex from token to token. Much of a patient timeline is composed of common population-level backbone patterns—recurrent chronic diagnoses, routine outpatient procedures, and visit-structure markers—that benefit from shared dense processing. In contrast, rarer clinical trajectories, specialty-specific procedures, and context-dependent code combinations are both more heterogeneous and more expensive to model with a single shared feed-forward path. A hybrid dense–MoE split therefore matches the data-generating structure of EHR: dense warm-start layers learn stable global patient-state representations, while later experts spend extra capacity on tokens whose clinical context is more specialized. This design is intended not only to save compute, but also to preserve shared structure while allowing selective specialization on the long tail of structured healthcare events.

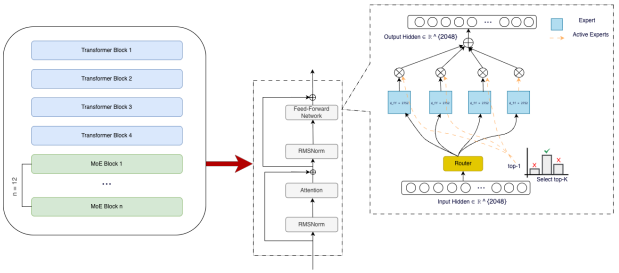

Figure 1. A compact view of the SparseEHR design principle: shared early processing, specialized late processing, and transfer through a common structured tokenization scheme.

Why the hybrid split matters in practice. Dense warm-start layers absorb high-frequency global structure before routing begins, which stabilizes sparse training on irregular, heterogeneous patient histories. The later MoE layers then provide conditional capacity where it matters most: modeling diverse clinical code interactions at scale. The main paper therefore treats the hybrid design as the deployed SparseEHR configuration and reports its efficiency advantage over an all-MoE reference, while the full dense-prefix allocation sweep is deferred to the appendix.

Training stability mechanisms. Conditional computation only helps if routing remains stable on long-tailed clinical data. In addition to the shared expert pathway, SparseEHR uses grouped-query attention, rotary positional embeddings, and an auxiliary load-balancing objective that penalizes extreme expert concentration. Intuitively, the dense prefix handles population-common context while the balance term keeps later specialists from collapsing onto a single expert or ignoring rare but clinically important regions of the code space.

3. Experimental Setup

Pretraining corpus. We pretrain SparseEHR on longitudinal structured records from approximately 50 million de-identified individuals in OLDW (Theel et al., 2015). Training uses an autoregressive next-code objective over chronological diagnosis and procedure sequences. Sequences are truncated or padded to length 2048 and optimized with a causal next-token loss.

Primary transfer evaluation. To isolate cross-system transfer, we evaluate *without fine-tuning* on MIMIC-IV next-visit ICD-10 prediction (Johnson et al., 2023). MIMIC records are converted into the same structured format used for pretraining: admissions are ordered chronologically, diagnosis and procedure codes are standardized, and the resulting timeline is fed directly to SparseEHR. The evaluation is deliberately *code-based rather than note-based*; no clinical notes or MIMIC labels are used for adaptation. We

report Recall@K, which is standard for multi-label diagnosis prediction with large sparse output spaces.

Data harmonization for zero-shot transfer. Because MIMIC-IV contains a mix of coding systems, we apply a fixed preprocessing pipeline before evaluation. Raw patient, admission, diagnosis, and procedure tables are first assembled into temporally ordered patient timelines. Legacy diagnosis codes are then mapped into the ICD-10-oriented token space used by SparseEHR, and procedures are converted to the canonical procedure format used during pretraining. Crucially, this harmonization step is frozen before evaluation and does not use MIMIC labels for tuning. The goal is not to optimize a benchmark-specific adapter, but to test whether a large structured-EHR model can transfer through a common schema alone.

Baselines and efficiency analysis. For transfer, we compare against representative public baselines spanning structured-EHR pretraining and recent LLM-based diagnosis models: MedBERT, MERA, and NECHO, and recent foundation-model-based EHR approaches (Rasmy et al., 2021; Ma et al., 2025; Koo, 2024; Anonymous, 2025). For efficiency, we compare the SparseEHR main configuration against an all-MoE reference using active parameters per token, time per training step, peak GPU memory, and validation perplexity. The appendix reports the broader dense-prefix ablation that motivated this final choice.

Implementation details. SparseEHR is trained with a causal next-token objective over sequences of maximum length 2048. The attention stack uses grouped-query attention and rotary positional embeddings, while the feed-forward stack follows the dense-to-sparse split described above. For transfer, we evaluate bundle prediction autoregressively: given a patient context, the model generates the next visit token by token and we score whether the ground-truth ICD-10 codes appear in the top-K predictions at each step. This setup keeps the structured prediction task aligned across datasets and avoids any MIMIC-specific calibration.

4. Results

Primary finding. SparseEHR improves zero-shot ICD prediction on MIMIC-IV while reducing active computation relative to an all-MoE design.

Zero-shot transfer to a new health system. SparseEHR achieves 0.463 Recall@10 and 0.551 Recall@20 on MIMIC-IV in a strictly zero-shot setting (Table 1). Relative to the strongest reported public baseline in the draft, MERA with BioMistral-7B, SparseEHR improves Recall@10 by 6.7 points and Recall@20 by 6.0 points. These gains are notable because the model is trained only on OLDW and evaluated on a different health system without task-specific fine-tuning. The result suggests that large-scale structured-EHR pretraining can transfer across institutions when the model architecture respects the sequential and sparse nature of medical code data.

Efficiency from conditional computation. The hybrid design yields a clear systems advantage, even though we keep the main paper focused on transfer. As shown in Appendix Table 3, SparseEHR reduces active parameters per token from $\sim$530M to $\sim$470M and lowers step time from 1.889s to 1.682s, a 10.9% speedup, while leaving peak memory nearly unchanged. The validation perplexity difference is small (28.42 vs. 28.68), indicating that most of the capacity benefit of sparse routing can be retained without paying the full all-expert cost from the start of the stack. The design-selection experiment that motivated this configuration is reported separately in the appendix.

Transfer is not just benchmark adaptation. The zero-shot setup is important substantively, not only procedurally. Many high-performing diagnosis predictors depend on dataset-specific supervision, task-specific prompt engineering, or multimodal components tuned on the target institution. SparseEHR instead transfers through the structured code space itself. Once MIMIC-IV is expressed in the same typed-token format, the pretrained model can forecast the next visit without seeing any MIMIC labels during optimization. This makes the result especially relevant for structured health data research, where schema alignment and transferable representation learning are often more practical than repeated end-to-end retraining at each site.

Supportive diagnostics and interpretation. Qualitative token inspection suggests *partial specialization by EHR function*: some experts absorb diagnosis-backbone tokens, while others focus on outpatient workflow, procedure-linked transitions, or context-specific chronic care. Rare-code prediction remains harder than common-code prediction, but recall still rises with larger K, indicating that sparse experts retain signal for infrequent concepts. Together, these results suggest that conditional computation is not merely a generic scaling trick, but a useful inductive bias for heterogeneous, long-tailed clinical timelines. The appendix reports the fuller routing and ablation analyses.

Secondary in-distribution validation. On the OLDW diabetes-screening endpoint, SparseEHR also outperforms the GPT baselines under the same hard binary decision rule. With $k_{\text{cls}}{=}50$ chosen by validation balanced accuracy, SparseEHR achieves 0.748 accuracy, 0.735 F1, and 0.748 balanced accuracy on the test set, versus 0.600/0.585/0.600 for GPT-5 and 0.530/0.493/0.562 for GPT-4.1 (Appendix Table 6). We treat this as supportive rather than headline evi-

Table 1. Strictly zero-shot transfer from OLDW pretraining to MIMIC-IV next-visit ICD-10 prediction. SparseEHR is evaluated without fine-tuning or MIMIC-specific label adaptation. The efficiency comparison for the selected hybrid design is reported in Appendix Table 3.

Method	R@10	R@20
MedBERT	0.259	0.338
MERA (BioMistral-7B)	0.396	0.491
NECHOv3 (GPT-4o-mini)	0.363	—
SparseEHR	**0.463**	**0.551**

dence: it shows that the pretrained structured-EHR model is not only transferable, but also useful on an operational next-visit screen once its ranked code predictions are mapped to the same binary endpoint used by direct generative baselines.

5. Implications for Structured Health Data

Transfer through schema, not site-specific retraining. A practical challenge in structured-health modeling is that many institutions cannot afford to retrain a large model each time the local schema, outcome definition, or patient population changes. The SparseEHR transfer result points to a different strategy: invest in large-scale pretraining over a stable typed-token schema, then adapt new sites primarily through harmonization into that schema rather than through full supervised re-optimization. This is especially attractive for settings where diagnosis and procedure data are abundant but gold labels for a new downstream task are scarce or delayed.

Table 2. Qualitative routing patterns observed in the selected D=4 model. The expert groups summarize the supporting token-inspection analysis from the full draft.

Layer	Representative specialization pattern
4	Earliest routing split between diagnosis-backbone tokens and workflow-linked procedures, demographics, and visit-structure markers.
6	Sharper separation into cardiometabolic follow-up, broad chronic-diagnosis backbone, musculoskeletal/skin care, and mixed screening-workflow context.
11	Shared high-frequency diagnoses remain distributed across experts, while residual experts capture preventive care, procedure-linked transitions, and lower-frequency context.
15	Most mature specialization by EHR function: common diagnosis backbone, visit transitions, chronic ambulatory patterns, and outpatient management/lab-office workflow.

Selective computation matches the structure of EHR. The routing summary in Table 2 suggests that sparse experts in structured EHR do not simply memorize narrow disease labels. Instead, they tend to organize around broader *functions of record generation*: common diagnostic backbone, ambulatory workflow, procedure-linked transitions, and context-specific chronic care. That is a good fit for real EHR, where the same disease token can appear in different operational contexts and where rare but important events are embedded inside common recurrent structure. A dense prefix plus sparse suffix therefore acts as both a systems design and a representational bias: shared early layers capture population-common context, while later experts spend extra capacity where heterogeneity is highest.

6. Discussion and Limitations

This workshop version prioritizes the two clearest signals: transfer and efficiency. Our study is limited to structured diagnosis/procedure sequences plus demographics, and does not yet evaluate multimodal fusion, fairness, calibration, prospective deployment, or broader external validation beyond MIMIC-IV. Formal uncertainty and reliability analysis also remain future work.

The main methodological takeaway is that structured EHR should not require uniform compute per token. Common events benefit from shared representations, while rarer or workflow-specific events can benefit from sparse specialization. SparseEHR suggests that a dense prefix plus expert routing is a better fit to this structure than a uniformly dense backbone. Future work should test whether this extends to medications, labs, time-series signals, and multimodal records (Anonymous, 2024).

7. Conclusion

We presented SparseEHR, a scalable foundation model for structured EHR based on conditional computation. A hybrid dense–MoE transformer pretrained on 50M de-identified longitudinal records transfers zero-shot to MIMIC-IV while reducing active per-token compute relative to an all-MoE design. These results suggest that selective computation is a promising path for population-scale structured health modeling.

References

William Fedus, Barret Zoph, and Noam Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 23(120):1–39, 2022.

Alistair E. W. Johnson, Lucas Bulgarelli, Leo Shen, Abigail Gayles, Ayad Shammout, Steven Horng, Tom J. Pollard, Sophie Hao, Benjamin Moody, Brian Gow, Li-wei H. Lehman, Leo Anthony Celi, and Roger G. Mark. MIMIC-IV, a freely accessible electronic health record dataset. *Scientific Data*, 10(1):1, 2023.

Hyeongu Koo. Next visit diagnosis prediction via medical code-centric multimodal contrastive EHR modelling with hierarchical regularisation. In *Findings of the Association for Computational Linguistics: EACL*, pages 41–55, 2024.

Yikuan Li, Shishir Rao, Jose R. A. Solares, Abdelaali Hassaine, Rohan Ramakrishnan, Dexter Canoy, Yajie Zhu, Kazem Rahimi, and Gholamreza Salimi-Khorshidi. BEHRT: Transformer for electronic health records. *Scientific Reports*, 10:7155, 2020.

Ming De Ma, Xin Wang, Yu Xiao, Alessandro Cuturrufo, Vijay S. Nori, Eran Halperin, and William Wang. Memorize and rank: Elevating large language models for clinical diagnosis prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2025.

Chao Pang, Xiaoqian Jiang, K. S. Kalluri, Matthew Spotnitz, Rui Chen, Adler Perotte, and Karthik Natarajan. CEHR-BERT: Incorporating temporal information from structured electronic health records to improve prediction tasks. In *Proceedings of Machine Learning for Health*, volume 158, pages 239–260, 2021.

Laila Rasmy, Yang Xiang, Zhiwen 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.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In *International Conference on Learning Representations*, 2017.

Elitza S. Theel, Ryan D. Johnson, Elizabeth Plumhoff, and Curtis A. Hanson. Use of the Optum Labs Data Warehouse to assess test ordering patterns for diagnosis of *Helicobacter pylori* infection in the United States. *Journal of Clinical Microbiology*, 53(4):1358–1360, 2015.

Du, N., Huang, Y., Dai, A. M., et al. GLaM: Efficient Scaling of Language Models with Mixture-of-Experts. In *Proceedings of ICML*, 2022.

Jiang, A. Q., Sablayrolles, A., Mensch, A., et al. Mixtral of Experts. *arXiv preprint arXiv:2401.04088*, 2024.

Tang, S., Wang, Y., et al. Self-supervised Transformer for Electronic Health Record Representation Learning. *Nature Communications*, 2023.

Wornow, M., Thapa, R., et al. Foresight: Generative Pretraining for Patient Trajectories. *The Lancet Digital Health*, 2024.

Anonymous. CEHR-GPT: Foundation Models for Electronic Health Records. *arXiv preprint*, 2025.

Anonymous. EHRMamba: Scalable State Space Models for Electronic Health Records. *arXiv preprint arXiv:2405.14567*, 2024.

A. Extended Experimental Details

Structured tokenization and sequence construction. Each patient record is serialized as a chronological typed-token timeline. Static demographic attributes are prepended as tokens (for example age bucket, sex, and region), followed by day-level diagnosis and procedure bundles separated by explicit begin/end delimiters. Diagnosis tokens are standardized to a compact ICD-oriented vocabulary and procedure tokens to a canonical procedure vocabulary. This design keeps the token space structured enough for transfer while remaining compatible with autoregressive pretraining.

Zero-shot MIMIC-IV harmonization. The public evaluation pipeline follows four steps. (1) Patient, admission, diagnosis, and procedure tables are joined into temporally ordered timelines. (2) Legacy diagnosis and procedure systems are mapped into the ICD-10-oriented and canonical procedure spaces used by SparseEHR. (3) The resulting sequences are formatted into the same typed-token schema used during OLDW pretraining. (4) Records containing invalid or unresolved tokens are removed before evaluation. Importantly, the conversion pipeline is frozen before testing and does not use MIMIC labels for model selection or adaptation.

Why we use Recall@K. Next-visit diagnosis prediction is a multi-label ranking problem with a large, sparse output space: each visit contains only a few true codes out of many possible codes in the model vocabulary. We therefore report Recall@K, which asks whether the correct code appears in the top-K predictions at each autoregressive step. This metric is more informative than exact-match accuracy for structured clinical prediction because it reflects whether the model surfaces clinically relevant codes near the top of its ranking, even when the first prediction is not identical to the target.

B. Additional Ablation Tables

Table 3. Supporting efficiency comparison for the selected hybrid design used in the main paper. The D=4 dense-prefix model reduces active compute and step time relative to the all-MoE D=0 variant with only a small change in validation perplexity.

Configuration	Active/token	Time/step	Peak mem.	Val PPL
All-MoE (D=0)	530M	1.889	25.7	28.42
Hybrid (D=4)	470M	1.682	25.3	28.68
Relative change	−12.5%	−10.9%	−1.6%	+0.26

Table 4. Dense–MoE allocation ablation from the full draft. The selected D=4 configuration gives the best expert-balance statistic while preserving a strong efficiency/quality trade-off. Lower is better for Val PPL, Time/step, Peak mem., and Expert util. std.

D	Total params	Active/token	Val PPL	Time/step	Peak mem.	Expert util. std
0	942M	530M	28.42	1.889	25.7	0.148
2	875M	500M	28.17	1.794	27.5	0.134
4	807M	470M	28.68	1.682	25.3	0.120
8	672M	415M	27.88	1.442	21.4	0.128

Table 5. Validation sensitivity of the diabetes decision threshold k_{cls}. The selected value k_{cls}=50 maximizes balanced accuracy on the held-out validation set.

k_{cls}	Precision	Recall	Accuracy	F1	Bal. Acc.
1	1.000	0.009	0.508	0.017	0.504
3	1.000	0.085	0.546	0.156	0.542
5	1.000	0.136	0.571	0.239	0.568
10	0.946	0.297	0.643	0.452	0.640
20	0.907	0.415	0.689	0.570	0.687
50	0.771	0.700	0.746	0.734	0.746
100	0.667	0.898	0.727	0.765	0.728

Table 6. Test-set diabetes prediction under a shared hard binary endpoint. SparseEHR uses k_{cls}=50, selected on validation balanced accuracy.

Method	Acc.	Prec.	Rec.	F1	Bal. Acc.
GPT-4.1	0.530	0.680	0.386	0.493	0.562
GPT-5	0.600	0.590	0.580	0.585	0.600
SparseEHR (k_{cls}=50)	0.748	0.769	0.703	0.735	0.748

Table 7. Rare-code versus common-code recall on the diabetes subset for the selected model. Rare-code prediction remains harder, but recall increases steadily with larger K, indicating useful long-tail retrieval.

K	Overall R@K	Rare-code R@K	Common-code R@K
1	0.1996	0.1027	0.2027
3	0.3096	0.2169	0.3035
5	0.3812	0.2769	0.3742
10	0.4720	0.3260	0.4890
20	0.5546	0.3761	0.5955
50	0.6903	0.4812	0.7454
100	0.7707	0.5684	0.8369

C. Qualitative Routing Summary

Table 8 provides an expanded version of the routed-token analysis summarized in Table 2, offering more detailed descriptions and additional examples of expert behavior across layers. The goal is not to claim that each expert maps to a single disease family. Rather, the evidence supports *partial specialization by EHR function*: some experts emphasize common diagnosis backbone patterns, while others focus more strongly on workflow, procedure transitions, or context-specific chronic care.

Table 8. Expanded routed-token analysis with additional examples and finer-grained patterns for four MoE layers in the selected D=4 model.

Layer	Summary
4	The first MoE layer after the dense warm start splits common diagnosis-sequence backbone tokens from procedure-heavy and workflow-linked context such as lab procedures, office-visit patterns, demographics, and visit markers.
6	Expert roles become sharper: one expert emphasizes cardiometabolic follow-up, another remains a general chronic-diagnosis backbone, a third leans toward musculoskeletal and skin-related chronic care, and a fourth captures mixed screening and outpatient workflow.
11	Routing shows both sharing and specialization. Common chronic-diagnosis tokens remain distributed across multiple experts, while other experts emphasize procedure-linked transitions, preventive care patterns, and lower-frequency contextual content.
15	The final MoE layer shows the most mature specialization by EHR function: a common diagnosis backbone, visit-transition patterns, chronic ambulatory care, and outpatient management/lab-office workflow.

Interpretation. This pattern is important for structured-health modeling because it suggests that strong load balancing does not destroy clinically meaningful specialization. Instead, the model shares ubiquitous high-frequency tokens across experts while still reserving capacity for lower-frequency workflow and chronic-care contexts. That is a plausible and desirable compromise for large structured EHR corpora, where common recurring codes and rare contextual events coexist in the same patient trajectory.

D. Appendix Discussion

The appendix results reinforce the main paper's central claim. First, the dense–MoE split matters: too little dense warm-start leaves routing less stable, but too much dense prefix removes expert capacity. Second, zero-shot transfer is compatible with a fully code-based evaluation protocol, which is appealing for structured-data settings where notes or local labels may be unavailable. Third, the model's long-tail behavior is imperfect but non-trivial: rare-code recall remains below common-code recall at every cutoff, yet it grows substantially as K increases, showing that the sparse architecture does surface useful tail information. Together, these details support the workshop version's argument that conditional computation is a practical scaling principle for foundation models over structured health data.