Auditing Clinical Concept Fragmentation in Sparse Medical Vision–Language Representations

Anonymous Authors[1]

Abstract

Trustworthy clinical AI requires that model evidence be inspectable at the level of clinically meaningful concepts, not only individual sparse features. Sparse dictionary models expose internal activations, but in medical vision–language models they can split one coherent finding across many atoms, creating a failure mode for clinical auditing. We study this failure mode as *concept fragmentation* and measure it with the Concept Fragmentation Score (CFS), the effective number of sparse features used by each supported clinical concept. We introduce HARP, a hierarchy-aligned, report-guided Poincaré objective that aligns sparse image codes with report-derived UMLS targets at the study level. On held-out MIMIC-CXR, HARP reduces CFS from 76.9 to 22.3 relative to a per-feature Euclidean ontology baseline while preserving reconstruction. The same frozen MIMIC-trained dictionary reduces CFS on CheXpert, NIH ChestX-ray14, and OpenI; linear probes improve on NIH and OpenI and remain comparable on CheXpert. A per-feature Poincaré prototype diagnostic collapses and worsens CFS, showing that ontology metric and supervision granularity must be evaluated together.

1. Introduction

Trustworthy clinical AI requires evidence that can be inspected at the level of clinically meaningful concepts. Sparse dictionary methods offer one route to such audits by decomposing internal states into human-inspectable features, which is especially attractive for medical vision–language models where radiology representations are reused for retrieval, classification, report grounding, and clinical adaptation. Yet radiological findings are not isolated labels: "pleural effusion", "opacity", and "lung abnormality" sit at different levels of a hierarchy of findings, anatomy, and clinical abstractions encoded in resources such as UMLS (Bodenreider, 2004). Sparse dictionary learning can expose activated dimensions (Olshausen & Field, 1997; Bricken et al., 2023; Cunningham et al., 2024; Dunefsky et al., 2024), but standard objectives optimize reconstruction and sparsity rather than clinical concept structure.

This creates a concrete auditing failure mode: a clinically coherent concept can fragment across many sparse features. Post-hoc feature naming may reveal that several atoms relate to the same finding, but it cannot tell whether the model has learned one stable concept, redundant fragments, or dataset-specific correlates. For trustworthy clinical AI, this distinction matters: a reviewer needs to know whether internal evidence for a finding is concentrated enough to inspect, challenge, or follow up with targeted validation.

A natural solution is to align dictionary atoms to ontology nodes. We argue that this feature-level strategy conflates two design choices. First, the ontology is hierarchical, while many alignment losses compare Poincaré coordinates with Euclidean distances, changing both update geometry and semantic neighborhoods. Second, radiology reports supervise studies, not individual dictionary atoms: a report describes a composition of findings rather than a one-to-one assignment between sparse features and concepts. The key audit question is therefore not just whether atoms can be named, but whether the internal sparse representation of a study organizes related clinical concepts compactly enough to support human review.

We propose HARP (**H**ierarchy-**A**ligned **R**eport-guided **P**oincaré sparse representations), a sample-level Poincaré alignment objective for sparse medical vision–language representations. A frozen biomedical vision encoder produces image activations; a TopK sparse dictionary reconstructs those activations; paired reports provide supported UMLS concept sets; and a Karcher mean of fixed UMLS Poincaré embeddings defines one study-level target. HARP aligns sparse image codes to these targets using the same Poincaré distance used to construct them.

Our experiments are organized around falsifiable hypothe-

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

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

ses. H1: sample-level Poincaré alignment reduces concept fragmentation relative to per-feature ontology alignment. H2: the reduction persists when the MIMIC-trained sparse dictionary is frozen and evaluated on external radiology corpora. H3: lower fragmentation does not remove clinically relevant linear signal from sparse codes. H4: applying the Poincaré metric at the wrong granularity, through hard atom-level prototypes, can fail. The results support these bounded claims while also revealing a retrieval–fragmentation trade-off and a missing sample-level Euclidean control.

2. Related Work

Sparse features for clinical model auditing. Sparse autoencoders and transcoders decompose dense activations into sparse feature dictionaries (Olshausen & Field, 1997; Elhage et al., 2022; Bricken et al., 2023; Cunningham et al., 2024; Dunefsky et al., 2024). Recent work improves scaling, sparsity control, and feature recovery with TopK activation, gated variants, and large-model feature analysis (Gao et al., 2024; Rajamanoharan et al., 2024; Templeton et al., 2024). These methods usually recover feature semantics after training by inspecting activations, nearest examples, or downstream associations. HARP changes the training signal and evaluation target: it uses report-derived ontology structure during sparse learning and evaluates whether concepts are concentrated rather than merely nameable.

Hyperbolic and knowledge-grounded medical representations. Hyperbolic spaces embed tree-like structure with low distortion (Nickel & Kiela, 2017; Sala et al., 2018; Ganea et al., 2018; Chami et al., 2019). Medical vision–language models align radiographs and reports with contrastive or local-global objectives (Huang et al., 2021; Boecking et al., 2022; Wang et al., 2022; Zhang et al., 2023a), and knowledge-grounded variants use labels, report graphs, or ontology priors (Jain et al., 2021; Zhang et al., 2023b; Wu et al., 2023). Our focus is different: we shape the sparse dictionary used to inspect a frozen model's internal image representation, rather than training a dense classifier or retrieval model.

Trustworthy clinical AI auditing. Trustworthy AI evaluation emphasizes auditing, failure-mode analysis, evidence standards, and avoiding unintended harms before deployment. In clinical imaging, a model may appear useful at the task level while its internal evidence for a finding is too dispersed to inspect, validate, or challenge. We therefore treat concept fragmentation as an audit failure mode: if a clinically meaningful concept is spread across many sparse atoms, feature-level explanations can become difficult to review even when individual atoms receive plausible names. This paper focuses on whether sparse medical representations are organized compactly enough to support downstream human inspection.

Ontology supervision as a structured prior. Medical ontologies are imperfect proxies for clinical reasoning, but they provide a useful prior over concept neighborhoods. Our objective does not require each atom to equal a UMLS node. Instead, the ontology target is a study-level summary of the report's supported concepts, and the sparse code is nudged toward that target in the geometry in which the ontology is embedded. This design keeps dictionary atoms free to specialize while still making the representation of a whole study geometrically auditable.

3. Method

3.1. Sparse dictionary and feature-level mismatch

Let $x_i \in \mathbb{R}^m$ be the frozen image representation of the i-th chest radiograph. A TopK sparse dictionary encodes and reconstructs

$$z_i = \mathrm{TopK}(\mathrm{ReLU}(Ex_i + b_E)) \in \mathbb{R}^d_{\geq 0}, \quad \hat{x}_i = Dz_i + b_D, \tag{1}$$

where $D = [w_1, \ldots, w_d]$ contains decoder atoms and TopK keeps the largest k activations. All sparse dictionaries use

$$\mathcal{L}_{\mathrm{rec}} = \frac{1}{n}\sum_i \|x_i - \hat{x}_i\|_2^2, \qquad \mathcal{L}_{\mathrm{sparse}} = \frac{1}{n}\sum_i \|z_i\|_1. \tag{2}$$

With fixed TopK, the sparsity term penalizes activation magnitude rather than cardinality; the same term is used for all comparisons.

The per-feature Euclidean baseline projects atoms to ontology coordinates and assigns each atom to the closest concept under ambient Euclidean distance,

$$b_j = \pi_{\mathrm{atom}}(w_j), \qquad \phi_{\mathrm{Euc}}(j) = \arg\min_{v \in V} \|b_j - h_v\|_2^2, \tag{3}$$

with prototype loss $\mathcal{L}_{\mathrm{atom,Euc}} = d^{-1}\sum_j \|b_j - h_{\phi_{\mathrm{Euc}}(j)}\|_2^2$. This baseline uses ontology embeddings but evaluates them with a distance that is not the ontology metric, and it supervises atoms even though reports provide study-level concept sets.

3.2. Report-derived targets in the Poincaré ball

Let $G = (V, E)$ be a radiology-focused UMLS subgraph and let $h_v \in \mathbb{B}^r_c$ be fixed Poincaré embeddings of concepts. For $u, v \in \mathbb{B}^r_c$,

$$d_c(u, v) = \frac{2}{\sqrt{c}} \operatorname{arctanh}\left(\sqrt{c}\|(-u) \oplus_c v\|_2\right), \tag{4}$$

where $\oplus_c$ is Möbius addition. We clamp $\|h_v\|_2 \leq (1 - \epsilon_{\mathrm{ball}})/\sqrt{c}$ with $\epsilon_{\mathrm{ball}} = 10^{-5}$ for numerical stability.

For each paired report r_i, a concept extractor returns $C_i \subseteq V$. Since reports describe multiple findings, HARP constructs one study-level target by taking a Karcher mean in

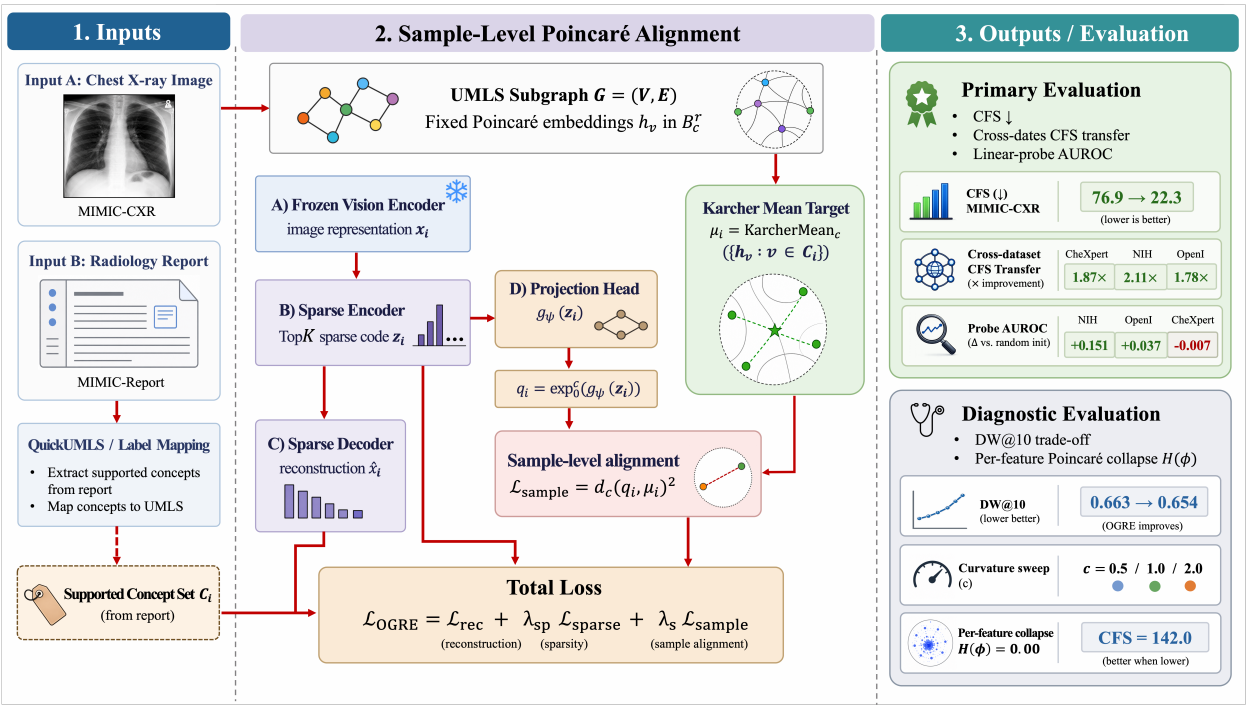

Figure 1. Overview of HARP. A frozen biomedical vision–language encoder produces image representations x_i, which are decomposed into TopK sparse codes z_i and reconstructed as $\hat{x}_i$. The paired report provides a supported concept set C_i; fixed UMLS Poincaré embeddings are summarized by a study-level Karcher target μ_i. HARP maps the sparse code to $q_i = \exp_0^c(g_\psi(z_i))$ and aligns q_i to μ_i with the Poincaré distance. Primary evidence is CFS; DW@10 and per-feature prototype collapse are diagnostics. In the figure, PF-Euc denotes the per-feature Euclidean baseline.

the same Poincaré space:

$$\mu_i = \text{KarcherMean}_c(\{h_v : v \in C_i\}) \tag{5}$$

$$\approx \arg\min_{u \in \mathbb{B}_c^r} \sum_{v \in C_i} d_c(u, h_v)^2. \tag{6}$$

Samples with $C_i = \varnothing$ are masked from the alignment loss.

3.3. Sample-level Poincaré alignment

A projection head $g_\psi : \mathbb{R}^d \to \mathbb{R}^r$ maps sparse codes to the tangent space at the origin and the exponential map sends them to the ball:

$$q_i = \exp_0^c(g_\psi(z_i)) \in \mathbb{B}_c^r. \tag{7}$$

The sample-level alignment loss is

$$\mathcal{L}_{\text{sample}} = \frac{\sum_i m_i d_c(q_i, \mu_i)^2}{\sum_i m_i}, \qquad m_i = \mathbf{1}[C_i \neq \varnothing]. \tag{8}$$

The full objective is

$$\mathcal{L}_{HARP} = \mathcal{L}_{\text{rec}} + \lambda_{\text{sp}} \mathcal{L}_{\text{sparse}} + \lambda_s \mathcal{L}_{\text{sample}}. \tag{9}$$

This objective matches the ontology metric and the report-level supervision unit. The distinction is not only cosmetic:

Proposition 3.1 (Poincaré distance gradient). *For fixed $\mu \in \mathbb{B}_c^r$, the manifold gradient of $d_c(q, \mu)^2$ at q points along the geodesic from q to μ: $\nabla_q^{\mathbb{B}} d_c(q, \mu)^2 = -2 \log_q^c(\mu)$. The Euclidean substitute $\|q - \mu\|_2^2$ follows the ambient chord direction and, under the Poincaré metric tensor, is scaled by λ_q^{-2} with $\lambda_q = 2/(1 - c\|q\|_2^2)$.*

The proof is in Section 5.1. We also evaluate a per-feature Poincaré prototype diagnostic: $a_j = \exp_0^c(\pi_{\text{atom}}(w_j))$, $\phi(j) = \arg\min_v d_c(a_j, h_v)$, and $\mathcal{L}_{\text{proto}} = d^{-1} \sum_j d_c(a_j, h_{\phi(j)})^2$. This variant changes the supervision unit and is not the main method.

3.4. Design choices and expected failure modes

The objective in Equation (9) changes two axes relative to atom-level ontology matching. The first axis is *metric consistency*: Poincaré targets are constructed and optimized with the same geodesic distance. The second axis is *supervision granularity*: report-derived concept sets supervise samples, not individual atoms. These axes are easy to conflate. A method can use hyperbolic coordinates but still optimize ambient Euclidean distances, or it can use the Poincaré metric but impose hard atom-to-concept assignments that do not reflect report supervision.

We use the per-feature Poincaré prototype variant to test this

interaction. If metric choice alone were sufficient, replacing Euclidean prototype distances with d_c should improve concept compactness. If granularity matters, hard prototype assignment can instead concentrate atoms around frequent or central ontology nodes, lowering per-atom distances while making sample-level representations less interpretable. The diagnostic in Table 5 tests exactly this failure mode by measuring assignment entropy $H(\phi)$ and CFS under the same curvature.

A second expected failure mode is a retrieval–fragmentation trade-off. Stronger curvature can pull related report targets together and reduce CFS, but it can also distort nearest-neighbor retrieval at small K. For this reason, DW@10 is reported as a diagnostic rather than the primary success criterion. The primary audit claim is about whether sparse internal evidence for a supported concept becomes less fragmented.

4. Experiments

4.1. Setup and metrics

We evaluate on MIMIC-CXR (Johnson et al., 2019), CheXpert (Irvin et al., 2019), NIH ChestX-ray14 (Wang et al., 2017), and OpenI (Demner-Fushman et al., 2016). MIMIC-CXR concepts are extracted from reports with QuickUMLS (Soldaini & Goharian, 2016). CheXpert and NIH use released pathology label families mapped to supported UMLS concepts; OpenI uses report-linked concept metadata mapped to the same subgraph. Cross-dataset CFS therefore measures fragmentation over each dataset's supported concept or label family, not identical fine-grained concepts at all institutions.

The frozen biomedical vision–language encoder produces $m = 3072$-dimensional patch-pooled features (Zhang et al., 2023a). The UMLS subgraph has $|V| = 2328$ concepts and $|E| = 4396$ *isa/part-of* edges. Poincaré embeddings have $r = 128$ and main curvature $c = 1.0$. Sparse dictionaries use $d = 2048$ atoms and TopK sparsity $k = 32$. We train for 20 epochs with AdamW (Loshchilov & Hutter, 2019), learning rate 10^{-3}, batch size 1024, $\lambda_{\text{sp}} = 10^{-3}$, and $\lambda_s = 10^{-2}$.

For concept v, let $S_v = \{i : v \in C_i\}$, and let $V_\tau = \{v : |S_v| \geq \tau\}$ be the supported concepts used for CFS in a given evaluation manifest. We define the normalized activation profile with numerical constant $\epsilon_{\text{num}} = 10^{-12}$:

$$p_v[j] = \frac{\sum_{i \in S_v} z_{ij}}{\sum_{\ell=1}^{d} \sum_{i \in S_v} z_{i\ell} + \epsilon_{\text{num}}}. \tag{10}$$

The Concept Fragmentation Score is the entropy effective number of sparse features,

$$H_v = -\sum_{j=1}^{d} p_v[j] \log(p_v[j] + \epsilon_{\text{num}}), \tag{11}$$

$$\text{CFS}_v = \exp(H_v), \qquad \text{CFS} = \frac{1}{|V_\tau|} \sum_{v \in V_\tau} \text{CFS}_v. \tag{12}$$

Lower CFS means a concept is represented by fewer effective sparse features. CFS equals one when all activation mass for concept v is carried by a single atom, and approaches d when the mass is uniform across the dictionary. It is insensitive to feature ordering and to atom names, which makes it suitable for comparing dictionaries that learn different bases. It is also deliberately conservative: a low value does not prove that the dominant features are clinically correct, while a high value warns that a concept is too distributed to be summarized by a single feature label. We also report DW@10, an ontology-distance-weighted retrieval recall in the Poincaré ball; reconstruction MSE; and macro-AUROC of dataset-specific linear probes trained on frozen sparse codes.

Baselines. We compare three primary dictionary configurations with the same frozen encoder, dictionary size, sparsity budget, optimizer, preprocessing, and validation budget. *Transcoder* is the unregularized sparse dictionary trained with reconstruction and sparsity only. *Per-feature Euclidean* adds atom-level ontology alignment using Euclidean coordinate distance over projected Poincaré embeddings. *HARP* is the sample-level Poincaré objective in Equation (9). Additional post-hoc and supervised controls are reported in Table 6; they test whether ontology behavior can be recovered after training without shaping sparse geometry.

Retrieval diagnostic. For a sample with supported concepts C_i, let $\text{TopK}(i) = \arg\text{topK}_{v \in V}\{-d_c(u_i, h_v)\}$ be the retrieved ontology concepts from a Poincaré representation u_i. We compute

$$\text{DW@K}_i = \frac{1}{|C_i|} \sum_{v^\star \in C_i} \max_{v \in \text{TopK}(i)} \exp(-d_c(h_v, h_{v^\star})). \tag{13}$$

DW@K gives partial credit for retrieving a nearby ontology concept, but it does not measure whether one concept uses many sparse features. This distinction is important in Tables 1 and 5, where fragmentation improves while retrieval remains comparable or decreases.

Statistics and manifests. CFS comparisons use paired Wilcoxon signed-rank tests over concepts (Wilcoxon, 1945); DW@10 comparisons use paired query-level tests; and linear-probe comparisons use macro-AUROC over dataset-specific label families. Holm–Bonferroni correction (Holm, 1979) is applied within the relevant comparison family. All reported values are generated from manifest-level evaluation

Table 1. Held-out MIMIC-CXR evaluation. CFS is the macro-mean effective number of sparse features per supported concept; DW@10 is an ontology-retrieval diagnostic; MSE is reconstruction error. $N = 45{,}351$ studies. * denotes paired Wilcoxon $p < 0.05$ versus the per-feature Euclidean baseline after Holm–Bonferroni correction.

Method	CFS ↓	DW@10 ↑	MSE ↓
Transcoder	327.0	0.529	0.0240
Per-feature Euclidean	76.9	**0.663**	0.0241
HARP	**22.3***	0.654	**0.0240**

counts to avoid mixing raw images, studies, and concept-filtered examples.

Clinical representation audit protocol. For every method we keep the sparse code itself as the object of analysis. CFS is computed from activations before any post-hoc atom naming, probe fitting, or retrieval head. This matters because a feature can retrieve a plausible UMLS neighbor even when a concept's activation mass is distributed over many features. We therefore interpret a method as improving sparse auditability only when three checks are jointly satisfied: reconstruction remains stable, CFS decreases on held-out concepts, and downstream probes do not show a broad loss of clinical label signal. The protocol is intentionally narrower than clinical validation, but it is aligned with the goal of building auditable evidence standards for trustworthy AI systems.

4.2. In-domain fragmentation

Table 1 supports H1. HARP reduces CFS from 76.9 for the per-feature Euclidean baseline to 22.3, a 3.45× reduction in the average effective number of sparse features per concept. Relative to the unregularized Transcoder, CFS decreases from 327.0 to 22.3, a 14.7× reduction. At the concept level, 82% of supported concepts have lower CFS_v under HARP than under the per-feature Euclidean baseline. Reconstruction does not explain the result: MSE remains within 0.4% relative across configurations. DW@10 is comparable but not uniformly improved, with the Euclidean atom baseline highest at 0.663 and HARP at 0.654; we treat retrieval as a diagnostic axis rather than the primary success metric.

The magnitude is easier to read as an audit workload reduction. Under the per-feature Euclidean baseline, a typical supported concept uses the equivalent of roughly 77 equally active atoms; under HARP, the same measurement falls to about 22 atoms. This does not imply that the representation has become single-feature or fully monosemantic. It means that the set of features an auditor must inspect for a concept is substantially smaller, while still leaving room for legitimate substructure such as anatomical location, severity, co-occurring device findings, or report-style variation. The reduction is therefore a compactness improvement, not a claim that every remaining atom has a one-to-one clinical interpretation.

Table 2. Cross-institutional CFS using frozen MIMIC-CXR-trained sparse representations. Reduction is the ratio of per-feature Euclidean CFS to HARP CFS.

Method	CheXpert	NIH	OpenI	MIMIC-CXR
Transcoder	174.8	174.8	372.2	327.0
Per-feature Euclidean	162.0	174.4	208.0	76.9
HARP	**86.6**	**82.7**	**116.9**	**22.3**
Reduction	1.87×	2.11×	1.78×	3.45×

Table 3. Linear-probe macro-AUROC on dataset-specific labels using frozen MIMIC-CXR-trained sparse codes. Values are mean±std over three probe seeds.

Method	CheXpert	NIH	OpenI
Transcoder	0.557±0.008	0.476±0.011	0.467±0.012
Per-feature Euclidean	**0.589±0.006**	0.517±0.009	0.489±0.013
HARP	0.582±0.007	**0.668±0.010**	**0.526±0.011**

4.3. External transfer under frozen sparse dictionaries

Table 2 supports H2. Without updating the encoder or dictionary, HARP reduces CFS by 1.87× on CheXpert, 2.11× on NIH ChestX-ray14, and 1.78× on OpenI relative to the per-feature Euclidean baseline. The external reductions are smaller than the in-domain MIMIC reduction, as expected from changes in report style, label density, and concept coverage. We therefore do not claim institution-invariant fragmentation removal; the narrower claim is that the frozen MIMIC-trained sparse geometry yields lower fragmentation across three external radiology evaluations.

This transfer test is deliberately strict in one respect and loose in another. It is strict because no external-domain sparse dictionary retraining is allowed; any reduction must come from the geometry learned on MIMIC-CXR. It is loose because each dataset exposes a different vocabulary: CheXpert and NIH provide label families rather than full reports, while OpenI is smaller and report-linked. We therefore avoid averaging raw CFS values across datasets as though they were measured on identical concept sets. The stable conclusion is directional: the same frozen sparse geometry tends to concentrate the concept evidence available in each corpus.

4.4. Downstream separability and diagnostic trade-offs

Table 3 supports H3. Relative to the per-feature Euclidean baseline, HARP improves macro-AUROC by +0.151 on NIH and +0.037 on OpenI; on CheXpert it changes by -0.007, within one seed-level standard deviation. These probes do not establish deployment performance, but they

Table 4. Label families used for macro-AUROC probes. Macro-AUROC is the unweighted average over labels in each row; label-specific claims are not made without corrected per-label tests.

Dataset	Label family	# labels
CheXpert	Competition pathologies	5
NIH ChestX-ray14	Thoracic findings	14
OpenI	Prevalent extracted labels	5

Table 5. Curvature and per-feature diagnostic results on MIMIC-CXR. $H(\phi)$ is atom-to-prototype assignment entropy and applies only to per-feature variants.

Variant	CFS ↓	DW@10 ↑	$H(\phi)$
Per-feature Euclidean	76.9	**0.663**	–
Sample-level, $c = 0.5$	66.4	0.658	–
Sample-level, $c = 1.0$ (HARP)	22.3	0.654	–
Sample-level, $c = 2.0$	**18.5**	0.580	–
Per-feature, d_c, $c = 1.0$	142.0	0.451	0.00

show that lower fragmentation does not merely collapse sparse activations or remove linearly accessible clinical label information.

The probe table is included for a negative reason as much as a positive one. A sparse objective can reduce fragmentation by suppressing rare or difficult features, which would make the representation easier to name but less useful. The macro-AUROC check rules out this simple collapse explanation at the level of dataset label families. It does not rule out subtler failures, such as a loss of rare-disease signal, a shift in calibration, or reliance on spurious acquisition correlates. Those questions require label-specific tests and human-facing audits beyond the scope of this compact submission.

Table 5 supports H4 and clarifies the trade-off. Increasing curvature reduces CFS from 66.4 at $c = 0.5$ to 18.5 at $c = 2.0$, but DW@10 decreases from 0.658 to 0.580. We select $c = 1.0$ because it captures most of the fragmentation reduction while keeping retrieval close to the per-feature Euclidean baseline. Replacing sample-level alignment with hard per-feature Poincaré prototypes collapses all 2048 atoms to a single prototype ($H(\phi) = 0.00$), worsens CFS to 142.0, and drops DW@10 to 0.451. Thus the Poincaré metric is not a drop-in atom-level replacement; metric and supervision granularity interact.

The collapse diagnostic is important for clinical representation auditing because it is a plausible failure of ontology-guided sparsity. If all atoms are pulled toward a frequent or central node, atom-level ontology retrieval can appear superficially coherent while the dictionary no longer separates the mechanisms that explain individual samples. The sample-level objective avoids this hard atom assignment: the dictionary can still allocate different atoms to different visual contexts, but the aggregate code for a study is trained to land near the report-derived concept summary. This is why we treat the diagnostic as evidence against a simplistic "hyperbolic is better" story.

Table 6. Post-hoc and supervised sparse-alignment controls on MIMIC-CXR under a separate post-hoc ontology-retrieval protocol. DW, SH, and SR are retrieval diagnostics; higher is better and not directly comparable to DW@10.

Method	DW ↑	SH ↑	SR ↑
Transcoder	0.564	0.029	0.423
PostHoc-NN	0.294	0.020	0.001
PostHoc-LinProj	0.296	0.010	0.001
PostHoc-Spectral ($K = 256$)	0.486	0.073	0.042
PostHoc-GraphMatch ($K = 384$)	0.464	0.040	0.022
AlignSAE-style	0.299	**0.218**	0.013
Training-time HARP control	**0.680**	0.141	**0.548**

4.5. Post-hoc and supervised controls

A reviewer concern in sparse model auditing is whether training-time geometry shaping is needed, or whether a vanilla dictionary can be labeled after training. Table 6 reports controls on a fixed Transcoder checkpoint. Naive nearest-neighbor and linear-projection assignment perform poorly under ontology retrieval. Under this post-hoc protocol, structure-aware spectral matching improves DW but remains below training-time hyperbolic shaping on DW and semantic recall. The AlignSAE-style control obtains higher hierarchy precision because it uses flat supervised concept labels, but its DW and semantic recall remain low. These controls support a narrow claim: post-hoc atom labeling partially helps, but it does not explain the training-time geometry behavior. They do not replace the missing same-granularity Euclidean CFS ablation.

4.6. Claim boundaries

Trustworthy clinical AI audits should state what the evidence does and does not support. Table 7 makes the scope explicit. The paper's central claim is not that ontology guidance makes every sparse feature clinically valid. It is that, under one frozen radiology vision–language encoder and supported UMLS concepts, sample-level Poincaré alignment reduces a measurable fragmentation score while preserving reconstruction and linear probe signal.

5. Theory, Reproducibility, and Audit Protocol

5.1. Proof sketch for the Poincaré update

The Poincaré ball $\mathbb{B}_c^r$ has conformal metric tensor $g_q = \lambda_q^2 I$, with $\lambda_q = 2/(1 - c\|q\|_2^2)$. For any smooth scalar objective f, the Riemannian gradient satisfies $\nabla_q^{\mathbb{B}} f = g_q^{-1} \nabla_q f =$

Table 7. Claim-to-evidence map. Each claim has a measured support and a boundary that should not be overextended.

Claim	Evidence	Boundary
HARP reduces concept fragmentation	Table 1: CFS 76.9 to 22.3	Supported MIMIC concepts only
Transfer is directionally preserved	Table 2: lower CFS on three external corpora	Frozen MIMIC dictionary; dataset-specific concept families
Sparse codes keep label signal	Table 3: NIH/OpenI improve, CheXpert comparable	Linear probes only, not deployment
Metric and granularity interact	Table 5: per-feature d_c collapses	Hard nearest prototypes only
Post-hoc labeling is insufficient	Table 6: structured controls remain below training-time control	Retrieval diagnostics, not full causal isolation

Table 8. Sparse dictionary training configuration held fixed across primary methods. The only intended differences are the ontology-alignment objective and its granularity.

Component	Value
Frozen image feature dimension	3072
Dictionary atoms d	2048
TopK sparsity k	32
Training epochs	20
Batch size	1024
Optimizer / learning rate	AdamW / 10^{-3}
Sparse penalty λ_{sp}	10^{-3}
Sample alignment λ_s	10^{-2}

$\lambda_q^{-2}\nabla_q f$. A standard geodesic-distance identity gives

$$\nabla_q^{\mathbb{B}} d_c(q,\mu)^2 = -2\log_q^c(\mu), \tag{14}$$

where $\log_q^c(\mu)$ is tangent to the unique geodesic from q to μ. By contrast, the coordinate loss $\|q-\mu\|_2^2$ has Euclidean gradient $2(q-\mu)$ and follows the ambient chord direction. Interpreting this coordinate loss under the Poincaré metric yields $\lambda_q^{-2}2(q-\mu)$, so Euclidean coordinate matching changes both update direction and scale near the boundary. This is the mathematical reason that metric consistency is an optimization choice, not only a notation change.

5.2. Dictionary training configuration

All primary comparisons freeze the same biomedical vision–language encoder and train sparse dictionaries over the same precomputed activations. Holding the encoder fixed is essential for the representation-auditing question: differences in CFS should reflect the organization of the sparse representation, not changes in the upstream visual model. Holding d and k fixed also prevents a trivial explanation in which one method merely uses fewer active units. The remaining degrees of freedom are the presence of ontology supervision, the geometry used to evaluate it, and whether supervision is applied to atoms or samples.

Table 9. Radiology-focused UMLS subgraph and embedding configuration used by all sparse-dictionary methods.

Property	Value
Concepts $\lvert V\rvert$	2328
Edges $\lvert E\rvert$	4396
isa / *part-of* edges	3247 / 1149
Support threshold τ	50
Embedding dimension r	128
Default curvature c	1.0
Boundary clamp ϵ_{ball}	10^{-5}
Negative samples per edge	50
Optimizer / learning rate	RSGD / 0.05

Table 10. Concept sources used for CFS and probe evaluation. Counts are post-filtered evaluation examples, not raw dataset sizes.

Dataset	Count	Concept source
MIMIC-CXR	45,351	QuickUMLS concepts from reports
CheXpert	188,521	Released pathology labels mapped to UMLS
NIH ChestX-ray14	112,120	Fourteen thoracic labels mapped to UMLS
OpenI	7,470	Report-linked metadata mapped to UMLS

5.3. Ontology construction and embedding

We construct the radiology-focused UMLS subgraph from MIMIC-CXR reports with QuickUMLS. Concepts appearing in at least $\tau = 50$ training reports are retained, UMLS *isa* and *part-of* relations are kept, and immediate parents are added where needed to improve connectivity while preserving hierarchical interpretation. The final graph statistics are shown in Table 9. Poincaré embeddings are trained once on this graph and then frozen for target construction, sample-level alignment, per-feature diagnostics, retrieval, and curvature sweeps. The embedding objective uses positive UMLS edges with negative samples and optimizes Riemannian graph likelihood in the Poincaré ball.

5.4. Dataset-specific concept sources

External datasets differ in label granularity and report availability. We therefore avoid claiming that the same fine-grained report concepts are observed everywhere. Instead, CFS is computed over each dataset's supported concept or label family after mapping to the shared UMLS subgraph. This makes cross-institutional transfer a test of whether the frozen sparse geometry reduces fragmentation under each dataset's available concept vocabulary.

Table 11. Statistical testing plan. CFS and retrieval are paired because the same concepts or queries are evaluated across methods.

Metric	Pairing unit	Test	Correction family
CFS	Concept v	Wilcoxon	Dataset comparisons
DW@10	Query i	Wilcoxon	Dataset comparisons
AUROC	Label	DeLong	Labels within dataset
MSE	Example	Descriptive	Not marked
$H(\phi)$	Atom assignment	Descriptive	Not marked

5.5. Statistical and reproducibility protocol

All numerical claims are generated from one manifest containing method, dataset, split role, evaluation count, metric name, metric direction, value, statistical test, and source artifact. This prevents mixing raw-image counts with study counts or concept-filtered counts. The testing plan is summarized in Table 11. We mark CFS significance only for paired concept-level comparisons and avoid label-specific claims unless per-label DeLong tests (DeLong et al., 1988) are available in the manifest.

5.6. How the method would be used in an audit

The intended use of HARP is not to certify clinical correctness automatically. A representation audit would first identify a supported concept family, compute CFS_v to decide whether the concept is concentrated enough for inspection, and then inspect top-activating images or reports for the dominant sparse features. Low CFS is useful because it narrows the number of features that require human review. High CFS remains informative as a failure signal: it indicates that the concept is too distributed for simple feature-level interpretation and should not be summarized by a single atom label.

A practical representation audit would proceed in three stages. First, compute concept-level CFS on a held-out slice, such as portable frontal studies or a clinically relevant subgroup. Second, rank the dominant atoms for each concept by their aggregate activation mass and inspect top-activating studies, report snippets, and nearest ontology neighbors. Third, treat interventions on these atoms as hypotheses rather than conclusions: ablation or activation patching can test whether the compact features influence downstream behavior, but those causal tests are not included in the present paper. This workflow makes the sparse representation a triage tool for human clinical inspection.

This audit protocol also explains why CFS and DW@10 can disagree. DW@10 asks whether a representation retrieves ontology-near concepts. CFS asks whether the activation mass for one concept is concentrated. A representation can retrieve a nearby concept while still spreading evidence across many atoms, or it can concentrate evidence while losing some nearest-neighbor retrieval fidelity under curvature. We therefore report both, but make CFS the primary metric for the fragmentation claim.

6. Discussion and Limitations

The results support a bounded design principle for trustworthy clinical AI auditing in structured medical domains: ontology guidance is more useful when the alignment metric matches the ontology geometry and the supervision unit matches the data source. HARP reduces concept fragmentation in-domain and across external datasets, while the per-feature Poincaré diagnostic collapses. The method therefore shifts the target from naming isolated sparse atoms to auditing whether concept-level internal evidence is compact enough to inspect.

Several limitations remain. The main causal gap is the missing sample-level Euclidean alignment cell, so our experiments do not fully separate the Poincaré metric from sample-level supervision. Each dictionary configuration is trained once; probe variability reflects probe seeds, not independent dictionary retraining. The study uses one frozen biomedical vision–language encoder and chest radiology data; transfer to other encoders, modalities, and report styles is untested. Concept extraction depends on QuickUMLS and a support-filtered UMLS subgraph, so concepts outside the extracted subgraph are not directly supervised. Finally, we do not claim hallucination detection, radiologist-level feature validation, or clinical deployment readiness. CFS and probes measure representational concentration and linear separability, not clinical correctness.

The most useful next experiment is a factorial ablation that crosses metric choice with supervision granularity: sample-level Euclidean, sample-level Poincaré, atom-level Euclidean, and atom-level Poincaré under matched capacity. A second direction is causal validation, where low-CFS atoms for a concept are ablated or patched to test whether they mediate model behavior on controlled image/report pairs. A third direction is human validation by radiology-trained annotators, because ontology compactness can make an audit tractable but cannot replace expert assessment of whether a sparse feature truly corresponds to a clinical finding.

Responsible Data and Release Statement

We use existing radiology datasets, UMLS resources, and pretrained models under their published access or license terms. We do not redistribute raw images, reports, UMLS content, or patient-level artifacts. If a supplementary artifact is released for review, it will be anonymized and limited to training scripts, evaluation code, configuration files, and aggregate manifests needed to reproduce the reported CFS, retrieval, and probe analyses.

References

Bodenreider, O. The unified medical language system (umls): integrating biomedical terminology. *Nucleic Acids Research*, 32(Database issue):D267–D270, 2004.

Boecking, B., Usuyama, N., Bannur, S., Castro, D. C., Schwaighofer, A., Hyland, S., Wetscherek, M., Naumann, T., Nori, A., Alvarez-Valle, J., Poon, H., and Oktay, O. Making the most of text semantics to improve biomedical vision–language processing. In *European Conference on Computer Vision*, 2022.

Bricken, T., Templeton, A., Batson, J., Chen, B., Jermyn, A., Conerly, T., Turner, N. L., Anil, C., Denison, C., Askell, A., Lasenby, R., Wu, Y., Kravec, S., Schiefer, N., Maxwell, T., Joseph, N., Hatfield-Dodds, Z., Tamkin, A., Nguyen, K., McLean, B., Burke, J. E., Hume, T., Carter, S., Henighan, T., and Olah, C. Towards monosemanticity: Decomposing language models with dictionary learning. Transformer Circuits Thread, 2023.

Chami, I., Ying, R., Ré, C., and Leskovec, J. Hyperbolic graph neural networks. In *Advances in Neural Information Processing Systems*, 2019.

Cunningham, H., Ewart, A., Riggs, L., Huben, R., and Sharkey, L. Sparse autoencoders find highly interpretable features in language models, 2024.

DeLong, E. R., DeLong, D. M., and Clarke-Pearson, D. L. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*, 44(3):837–845, 1988.

Demner-Fushman, D., Kohli, M. D., Rosenman, M. B., Shooshan, S. E., Rodriguez, L., Antani, S., Thoma, G. R., and McDonald, C. J. Preparing a collection of radiology examinations for distribution and retrieval. *Journal of the American Medical Informatics Association*, 23(2): 304–310, 2016.

Dunefsky, J., Chlenski, P., and Nanda, N. Transcoders find interpretable llm feature circuits, 2024.

Elhage, N., Hume, T., Olsson, C., Schiefer, N., Henighan, T., Kravec, S., Hatfield-Dodds, Z., Lasenby, R., Drain, D., Chen, C., Grosse, R., McCandlish, S., Kaplan, J., Amodei, D., and Olah, C. Toy models of superposition. Transformer Circuits Thread, 2022.

Ganea, O.-E., Bécigneul, G., and Hofmann, T. Hyperbolic neural networks. In *Advances in Neural Information Processing Systems*, 2018.

Gao, L., la Tour, T. D., Tillman, H., Goh, G., Troll, R., Radford, A., Sutskever, I., Leike, J., and Wu, J. Scaling and evaluating sparse autoencoders, 2024.

Holm, S. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979.

Huang, S.-C., Shen, L., Lungren, M. P., and Yeung, S. Gloria: A multimodal global-local representation learning framework for label-efficient medical image recognition. In *IEEE/CVF International Conference on Computer Vision*, 2021.

Irvin, J., Rajpurkar, P., Ko, M., Yu, Y., Ciurea-Ilcus, S., Chute, C., Marklund, H., Haghgoo, B., Ball, R., Shpanskaya, K., Seekins, J., Mong, D. A., Halabi, S. S., Sandberg, J. K., Jones, R., Larson, D. B., Langlotz, C. P., Patel, B. N., Lungren, M. P., and Ng, A. Y. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 590–597, 2019.

Jain, S., Agrawal, A., Saporta, A., Truong, S., Duong, D. N., Bui, T., Chambon, P., Zhang, Y., Lungren, M. P., Ng, A. Y., Langlotz, C. P., and Rajpurkar, P. Radgraph: Extracting clinical entities and relations from radiology reports. In *NeurIPS Datasets and Benchmarks Track*, 2021.

Johnson, A. E. W., Pollard, T. J., Berkowitz, S. J., Greenbaum, N. R., Lungren, M. P., Deng, C.-y., Mark, R. G., and Horng, S. Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific Data*, 6:317, 2019.

Loshchilov, I. and Hutter, F. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

Nickel, M. and Kiela, D. Poincaré embeddings for learning hierarchical representations. In *Advances in Neural Information Processing Systems*, 2017.

Olshausen, B. A. and Field, D. J. Sparse coding with an overcomplete basis set: a strategy employed by v1? *Vision Research*, 37(23):3311–3325, 1997.

Rajamanoharan, S., Conmy, A., Smith, L., Lieberum, T., Varma, V., Kramár, J., Shah, R., and Nanda, N. Improving dictionary learning with gated sparse autoencoders, 2024.

Sala, F., De Sa, C., Gu, A., and Ré, C. Representation tradeoffs for hyperbolic embeddings. In *International Conference on Machine Learning*, 2018.

Soldaini, L. and Goharian, N. Quickumls: A fast, unsupervised approach for medical concept extraction. In *ACM SIGIR Workshop on Medical Information Retrieval*, 2016.

Templeton, A., Conerly, T., Marcus, J., Lindsey, J., Bricken, T., Chen, B., Pearce, A., Citro, C., Ameisen, E., Jermyn,

A., Anil, C., Denison, C., Askell, A., Lasenby, R., Wu, Y., Kravec, S., Schiefer, N., Maxwell, T., Joseph, N., Tamkin, A., Nguyen, K., McLean, B., Burke, J. E., Hume, T., Carter, S., Henighan, T., and Olah, C. Scaling monosemanticity: Extracting interpretable features from claude 3 sonnet. Transformer Circuits Thread, 2024.

Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., and Summers, R. M. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2097–2106, 2017.

Wang, Z., Wu, Z., Agarwal, D., and Sun, J. Medclip: Contrastive learning from unpaired medical images and text, 2022.

Wilcoxon, F. Individual comparisons by ranking methods. *Biometrics Bulletin*, 1(6):80–83, 1945.

Wu, C., Zhang, X., Zhang, Y., Wang, Y., and Xie, W. Medklip: Medical knowledge enhanced language-image pretraining for x-ray diagnosis. In *IEEE/CVF International Conference on Computer Vision*, 2023.

Zhang, S., Xu, Y., Usuyama, N., Bagga, J., Tinn, R., Preston, S., Rao, R., Wei, M., Valluri, N., Wong, C., Lungren, M. P., Naumann, T., and Poon, H. Large-scale domain-specific pretraining for biomedical vision–language processing, 2023a.

Zhang, X., Wu, C., Zhang, Y., Xie, W., and Wang, Y. Knowledge-enhanced visual–language pre-training on chest radiology images. *Nature Communications*, 2023b.