

Beyond Label Attention: Transparency in Language Models for Automated Medical Coding via Dictionary Learning

John Wu

University of Illinois Urbana-Champaign
johnwu3@illinois.edu

David Wu

Vanderbilt University
David.h.wu@vanderbilt.edu

Jimeng Sun

University of Illinois Urbana-Champaign
jimeng@illinois.edu

Abstract

Medical coding, the translation of unstructured clinical text into standardized medical codes, is a crucial but time-consuming healthcare practice. Though large language models (LLM) could automate the coding process and improve the efficiency of such tasks, *interpretability* remains paramount for maintaining patient trust. Current efforts in interpretability of medical coding applications rely heavily on label attention mechanisms, which often leads to the highlighting of extraneous tokens irrelevant to the ICD code. To facilitate accurate interpretability in medical language models, this paper leverages *dictionary learning* that can efficiently extract sparsely activated representations from dense language model embeddings in superposition. Compared with common label attention mechanisms, our model goes beyond token-level representations by building an interpretable dictionary which enhances the mechanistic-based explanations for each ICD code prediction, even when the highlighted tokens are medically irrelevant. We show that dictionary features can steer model behavior, elucidate the hidden meanings of upwards of 90% of medically irrelevant tokens, and are human interpretable.

1 Introduction

Transparency is a vital factor in healthcare to gain patients' trust, especially when AI models make critical decisions in clinical practice (Rao et al., 2022). One of the essential applications of AI models is to assign International Classification of Diseases (ICD) codes automatically based on the clinical text (we name this task as *medical coding*). These ICD codes categorize patient diagnoses, conditions, and treatments for billing, reporting, and treatment purposes (Hirsch et al., 2016; Johnson et al., 2021). However, assigning ICD codes is complex and requires expertise and time (O'Malley et al., 2005). Recent advancements in medical pre-trained language models (PLMs)

have made it possible to treat medical coding as a high-dimensional multilabel classification challenge (Edin et al., 2023; Huang et al., 2022). These AI models led to significant success in efficient ICD coding (Kaur et al., 2021). However, their transparency remains of great concern (Hakkoum et al., 2022). Therefore, developing automated interpretability methods is crucial to upholding transparency in medical coding processes.

Significant progress has occurred in the field of black-box interpretability, particularly concerning feature attribution, with the emergence of perturbation-based methods such as SHAP (Lundberg and Lee, 2017) and its approximate counterpart LIME (Ribeiro et al., 2016; Moraffah et al., 2020). These techniques, rooted in information and game theory, are recognized for assessing feature relevance in detail by intelligently perturbing and ablating input features (Lundberg and Lee, 2017; Ribeiro et al., 2016). While approximation methods have greatly improved the speed of calculating Shapley values, exact computations remain expensive (Lundberg et al., 2020; Chen et al., 2022; Shrikumar et al., 2019; Mosca et al., 2022). The huge computational cost makes their application impractical towards automated medical ICD coding since clinical notes usually contain thousands of high dimensional token embeddings in a vast multilabel prediction space (Johnson et al., 2023).

As such, we seek a human-interpretable approach that scales efficiently with large datasets, highlights essential features, and offers more comprehensive explanations of PLM predictions. Recent advancements in mechanistic interpretability methods (Cunningham et al., 2023; Räuker et al., 2023) have demonstrated the potential to surpass the computational challenges posed by traditional black-box approaches by elucidating the roles of specific neuron subsets within a network. This level of mechanistic understanding is precious in the medical field, where explaining the significance

of a feature is as crucial as its identification.

In recent years, the attention mechanism (Abnar and Zuidema, 2020; Vaswani et al., 2023; Chefer et al., 2021, 2022) has been heavily used to interpret and explain the behavior of large transformer models. Within the realm of automated medical ICD coding, label attention (LAAT) variants are most prevalent due to their efficiency in handling extensive sequence lengths. They calculate an attention score for each token relative to each label, thereby identifying tokens crucial for ICD code predictions (Mullenbach et al., 2018; Vu et al., 2020). Nevertheless, studies such as (Pandey et al., 2023) have questioned the interpretability and validity of explanations provided by attention mechanisms.

For example, the LAAT mechanism may highlight incoherent or irrelevant tokens, such as stop words for highly medically specific ICD codes. For instance, LAAT attributes the stop token "and" to the medically specific ICD code "998.59 postoperative wound infection" despite conjunctions being irrelevant to medical prognosis, thus undermining the interpretability of LAAT as shown in Figure 1. We attribute this issue to be the result of neuron polysemanticity—where a single neuron responds to diverse, unrelated inputs—complicates direct interpretation (Olah et al., 2020).

Elhage et al. (2022) theorizes polysemanticity to be a form of superposition, occurring when the count of independent data features surpasses layer dimensions, leading to data features being represented by linear combinations of neurons. As a result, individual neurons are often directly uninterpretable (Subramanian et al., 2017; Cunningham et al., 2023). Addressing this, sparse autoencoders (Olshausen and Field, 1997) have been applied to distill these complex dense representations into interpretable, sparse linear combinations, performing what is known as *dictionary learning* (DL). This strategy has effectively decomposed different language model layers, such as neural word embeddings (Subramanian et al., 2017), MLPs (Cunningham et al., 2023; Bricken et al., 2023), and residual connections (Yun et al., 2023a), proving scalable and monosemantic.

Our paper generalizes these sparse coding concepts to better understand the attention mechanism and pre-trained language model (PLM) embeddings by constructing effective dictionaries with LAAT to improve the interpretability of the medical ICD coding task. We summarize our main

contributions:

- Expanding on (Bricken et al., 2023)’s ablation studies, we show that combining learned dictionary features with another mechanistic component (LAAT) (Vu et al., 2020), in our new interpretability framework AutoCodeDL, improves explainability of downstream ICD predictions.
- We build medically relevant dictionaries with sparse autoencoders that can capture medically relevant concepts hidden within superposition.
- We develop new automated proxy metrics for assessing the human understandability of constructed dictionaries and conduct extensive evaluations on large scale clinical text-based corpus.

2 Related Work

2.1 Automated Interpretability in ICD Coding

Alternative automated ICD coding methods, like phrase matching (Cao et al., 2020) and relevant phrase extraction using manually curated knowledge bases (Duque et al., 2021), offer inherent interpretability but fall short in expressive power compared to neural network-based approaches. This discrepancy highlights a persistent tradeoff between interpretability and performance in ICD coding tasks.

Furthermore, the prevailing interpretability method for deep neural models in ICD coding tasks rely on the attention mechanism (Yan et al., 2022). Specifically, the LAAT mechanism projects token embeddings into a label-specific attention space, where each token receives a score indicating its relevance to each ICD prediction. Such attention-based associations between tokens and classes have been employed in various architectures, including convolutional models like CAML (Mullenbach et al., 2018), recurrent neural networks (Vu et al., 2020), and large language models (Huang et al., 2022; Yang et al., 2023). While computationally efficient, it overlooks the potentially richer information hidden within the embedding space, hindering our understanding of automated ICD predictions. Our work builds on top of such past works, and directly interprets the medical PLM embeddings.

2.2 Dictionary Learning

Dictionary learning aims to find a sparse representation of input data in the form of linear combina-

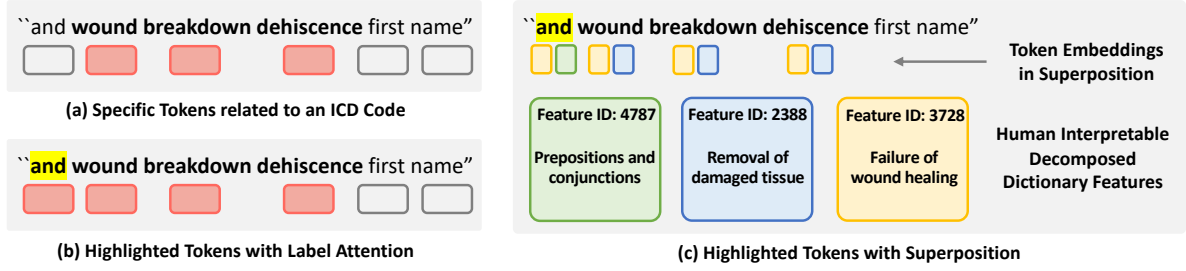


Figure 1: Motivation: LAAT identifies the most relevant tokens for each ICD code (b). Compared to our inspection of which tokens are most relevant to an ICD code (a), we assume "and" is irrelevant to an ICD code prediction. Although it may appear as though "and" is irrelevantly highlighted, taking token embeddings out of superposition allows us to decompose dense token embeddings into more semantically meaningful dictionary features that show that concepts of "failure of wound healing" are embedded within its token embedding (c), thereby giving justification for its highlighting by LAAT for a wound-related ICD code.

tion of basic elements (Olshausen and Field, 1997). Such an approach has been applied across various domains, including word embedding decomposition (Subramanian et al., 2017), interpretation of language model activations (Bricken et al., 2023; Cunningham et al., 2023; Yun et al., 2023a), enhancement of representation learning (Ghosh et al., 2023; Tang et al., 2023), and analysis of time-series data (Xu et al., 2023), highlighting their versatility (Zhang et al., 2015a).

Our research aims to understand if and how dictionary learning improves upon existing interpretability methods like LAAT for predicting medical codes (ICDs) in a highly practical setting. By analyzing diverse medically relevant ICDs, we assess the learned dictionaries’ ability to capture specific and meaningful medical concepts. Unlike previous work requiring extensive human annotation (Bricken et al., 2023; Cunningham et al., 2023; Subramanian et al., 2017), we propose new automated metrics to measure how understandable these dictionaries are due to the cost of expert annotation.

3 Methodology

As depicted in Figure 2, our focus within dictionary learning involves explicitly building dictionaries where relevant tokens and ICD codes are mapped to dictionary features. We examine two sparse autoencoder approaches aimed at creating interpretable representations from dense language model embeddings: one via L_1 minimization (Cunningham et al., 2023; Bricken et al., 2023) and another via SPINE’s loss function (Subramanian et al., 2017). While our discussion primarily centers on the L_1 minimization technique for its widespread application and illustrative clarity regarding dictionary

learning’s objectives in section 3.1, further details on SPINE are provided in the Appendix A.6.

Then, using our trained sparse autoencoder, we perform ablation studies to understand the downstream effects of dictionary features in section 4.1 and map the relevant ICD codes to each dictionary feature as discussed in section 3.2. Finally, we leverage sparse encoding and its ablation techniques in constructing our final dictionary, mapping both relevant tokens and ICD codes to each dictionary feature in section 3.3, which is used in our new proposed method AutoCodeDL in Figure 3.

3.1 Sparse Autoencoders

Following (Bricken et al., 2023)’s approach, let $x \in \mathbb{R}^d$ be the token embedding we wish to interpret, d the dimension size of the token embedding, and m the dimension size of the latent sparse dictionary feature activations $f \in \mathbb{R}^m$ generated by the sparse autoencoder. Our L_1 sparse autoencoder is shown below in equations 1 through 4 where $W_e \in \mathbb{R}^{m \times d}$ is the encoder weight matrix, $b_e \in \mathbb{R}^m$ is the encoder weight bias term, $b_d \in \mathbb{R}^d$ is the decoder bias term, and $W_d \in \mathbb{R}^{d \times m}$ represents the sparse dictionary embeddings:

$$\bar{x} = x - b_d \quad (1)$$

$$f = \text{ReLU}(W_e \bar{x} + b_e) \quad (2)$$

$$\hat{x} = W_d \cdot f + b_d \quad (3)$$

$$\mathcal{L} = \frac{1}{|X|} \sum_{x \in X} \|x - \hat{x}\|_2^2 + \lambda_{L_1} \|f\|_1 \quad (4)$$

The L_1 norm $\|f\|_1$ in the loss function (see eq. 4) enforces the sparse representation of f in training. As a result, only certain elements within f activate for certain types of token embeddings x ,

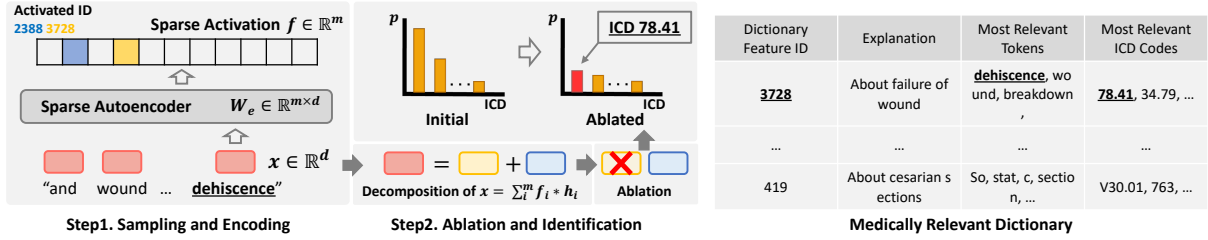


Figure 2: Building a dictionary involves several steps: A sparse autoencoder decomposes each token embedding into a sparse latent space, where each nonzero element represents a dictionary feature ID (step 1). This process enables the creation of mappings between tokens and various dictionary features and ICD codes. In step 2, ICD codes are mapped to dictionary features based on the softmax probabilities of each ICD prediction after dictionary embedding ablations, as detailed in section 3.2. Once a dictionary is constructed, it is utilized to enhance explanations by applying it to highlighted tokens identified by LAAT in Figure 3.

creating a direct mapping between different tokens (words) represented by x and their respective features identified by f_i . For instance, we observe that "depression"-related tokens only activate the dictionary feature f_{5732} .

While there exist many other useful properties of L_1 minimizations (Zhang et al., 2015b), the key idea is that a sparse linear combination of learned dictionary embeddings represents every token embedding. First, let us examine the dictionary embedding matrix W_d as defined below.

$$W_d = [h_0, h_1, \dots, h_m]^T \quad (5)$$

Let $h_i \in \mathbb{R}^d$ be the dictionary embedding associated with a dictionary feature f_i ; we have the following decomposition of any token embedding x into sparse features.

$$x \approx \sum_i^m f_i * h_i \quad (6)$$

Another key intuition behind why these decompositions are directly interpretable is that since certain f_i only activate (i.e., is nonzero) for certain types of tokens, its dictionary embeddings h_i have a defined direction in the PLM embedding space that should directly correspond to some meaningful concept within the original clinical text, thus having downstream implications.

3.2 Mapping Dictionary Features to ICD Codes

To build a medically relevant dictionary, ICD codes should map to their respective meaningful dictionary features. Following the methodology in (Bricken et al., 2023), we ablate features in clinical notes by targeting any activated dictionary feature

$f_i > 0$ in token embedding $x \in \mathbb{R}^d$. For each feature f_i with corresponding feature embedding $h_i \in \mathbb{R}^d$, we define the ablated token embedding as \tilde{x} .

$$\tilde{x} = x - f_i \cdot h_i \quad (7)$$

For any token in a clinical note, we perform token embedding ablations, recalculate the ablated model’s softmax probabilities for all \mathbb{C} classes or ICD codes, and compute the probability differences δ_i .

$$\delta_i = p(x) - p(\tilde{x}), \delta_i \in \mathbb{R}^{\mathbb{C}} \quad (8)$$

Finally, for any given class $c \in \{1, 2, \dots, C\}$, we sort and record the top $\delta_{i,c}$ ’s when ablating each dictionary feature f_i and its embedding h_i , identifying its most relevant ICD codes. Such ablations are later used for evaluating the model explainability of dictionary features and building more human-interpretable medical dictionaries with the sparse autoencoder.

3.3 Building Medically Relevant Dictionaries to Augment ICD Explanations

In essence, the sparse autoencoder contains a dictionary within its latent space. While efficient in time and space complexity, its direct interpretation requires the construction of a more human-interpretable and literal dictionary containing the most relevant tokens and ICD codes for each dictionary feature f_i . Building a dictionary can be summed up into two sorting steps.

Sampling and Encoding. We sample a certain number of clinical notes from our test set. For every token in each clinical note, we encode their PLM embeddings with our sparse autoencoder and retrieve its sparse feature activations f_i . Then, for each dictionary feature i , we sort by each token’s re-

spective f_i and select the top k tokens with the highest feature activations for each dictionary feature. Since there are often compound words, consisting of multiple tokens, we also retrieve any neighboring tokens with nonzero activations.

Ablation and Identification. For every clinical note, we perform ablations for each activated dictionary feature’s embedding h_i , measuring the change in predicted probability for each ICD code. For each dictionary feature, we identify its most relevant classes based on the largest probability drops after ablation. We formalize this process for a single clinical note in the pseudocode shown in algorithm 1 in the Appendix.

Proposed Method of Interpretability. After constructing the dictionary, we can utilize it to query the dictionary features of any PLM token embeddings, enhancing interpretability. Integrated with LAAT in our proposed new method AutoCodeDL, we initially identify the key tokens for each ICD prediction and subsequently match their embeddings to features in our dictionaries, thereby refining explanations (see Figure 3).

4 Interpretability Evaluations

While there does not exist a unified singular definition of neural model interpretability, there is a general consensus that model interpretability methods should be both model explainable and human understandable (Zhang et al., 2021). We define explainability as how much do the dictionary features learned predict the pre-trained language model’s ICD predictions. We define human understandability within the lens of monosemanticity: a dictionary feature is only human interpretable when the tokens that highly activate said dictionary feature are related and its underlying concept can be easily identified.

Sparse Autoencoders. We explore two major dictionary learning (DL) sparse autoencoder approaches, SPINE by (Subramanian et al., 2017) and L_1 described in section 3.

Baselines. Following (Cunningham et al., 2023), we explore four unsupervised baseline encoders that we compare to a true dictionary encoding: Independent Component Analysis (ICA), capable of decomposing word embeddings into semantically meaningful independent components (ICs) (Musil and Mareček, 2022); Principal Component Analysis (PCA), which has been used to analyze the structure of word embeddings (Musil, 2019); an Identity ReLU encoder, which effectively treats

each element in the PLM embedding as its own dictionary feature; and a random encoder.

Dataset and Model. We train on PLM embeddings from a 110M medical RoBERTa model leveraged by the state of the art PLM-ICD coding model (Huang et al., 2022; Lewis et al., 2020) and evaluate our method using the cleaned MIMIC-III dataset, using 38,427 clinical notes for training and 8,750 for evaluation, as detailed by (Edin et al., 2023). Additional details on training are in Appendix A.1.

4.1 Model Explainability

Faithfulness. Inspired by explainability evaluation metrics within the vision domain (Chattopadhyay et al., 2018; Samek et al., 2015), we assess our interpretable dictionary features by removing them (ablation) and measuring their impact on predicted ICD codes. This approach helps us quantify how well our explanations align with the model’s predictions. For example, ablating the "depression" feature should primarily impact related ICD codes like "depressive disorders" while leaving unrelated ones like "postoperative wound infection" relatively unaffected.

To address this, we consider both the *decrease* in the most likely code’s softmax probability after ablation and the sum of absolute changes in softmax probabilities for all other codes. Computing a ratio of these measures (detailed in Table 1) allows us to accurately gauge an interpretability method’s explanatory power. This ablation metric is analogous to Comprehensiveness (Comp) by (Chan et al., 2022; DeYoung et al., 2020), but instead of erasing entire tokens, we ablate specific components of the embedding (see eq. 7).

Setup. We compare our interpretability framework, AutoCodeDL, against three sets of baselines. The first set involves feature ablations using baseline encoders and full token ablations (labeled "token" in Table 1) with LAAT to pre-highlight relevant tokens. The second set excludes LAAT and simply explains only with DL feature ablations across all tokens. The third set excludes LAAT and only utilizes the baseline encoder ablations. For additional details on the ablations for each baseline encoder, please refer to Appendix A.7.

Results. From Table 1, ablations of dictionary features of highlighted tokens with our proposed AutoCodeDL method show minimal impact on other ICD code predictions while retaining large drops in softmax probabilities. As a result, our

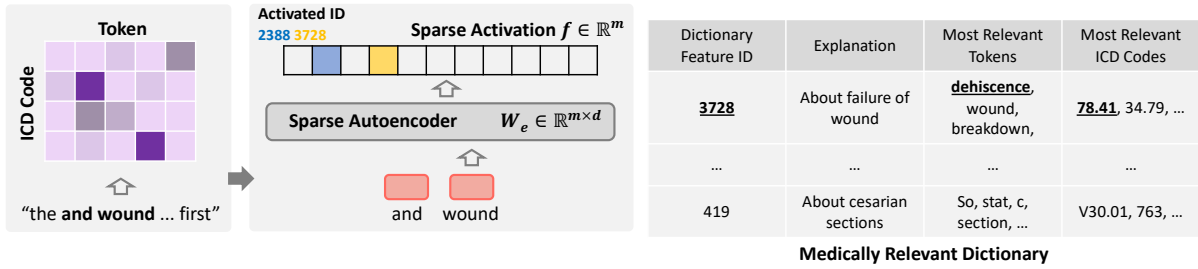


Figure 3: Proposed method for automated ICD interpretability pipeline: **AutoCodeDL**. LAAT identifies the most important words "and wound". Then, the sparse autoencoder queries its most activated dictionary features, returning its respective dictionary feature ids that can be leveraged to further explain the PLM’s predictions and attention highlights.

proposed method outperforms all baselines in our ratio explainability metric. Within our baselines, their ablations varied tremendously. Most notably, ablating ICs from embeddings using ICA had minimal impact on predictions, possibly because ICA approximations identify features with negligible mutual information while principal component ablations were similar to full token ablations.

Addressing Superposition. We decompose token embeddings to understand why label attention highlights specific tokens for ICD codes. But each token can hold multiple meanings, making manual evaluation of its dictionary features laborious. So, we introduce a new metric to assess if our dictionaries can identify hidden meanings even when attention identifies "extraneous" words.

Setup. We analyze stop words highlighted by label attention to evaluate our dictionary’s ability to explain their relevance to ICD code predictions. From 8,000 test set clinical notes, we extract *all, not just attention-highlighted tokens* and build a dictionary linking tokens and ICD codes to dictionary features (described in Section 3.2). We then focus on stop words deemed highly relevant by the label attention mechanism. For each stop word, we query its relevant dictionary features using our trained sparse autoencoder, and see if the original label ranks among the top 10 classes pre-mapped by the sampled dictionary for each activated feature. Since feature activation magnitudes vary, we define "highly activated" features as those exceeding the 96.5th percentile feature magnitude per token embedding. Additional details are listed in Appendix A.9.

Results. Table 2 presents the proportion of stop word embedding labels correctly identified by our DL framework, as well as the performance of baseline methods. Notably, our DL framework, partic-

ularly the L1 sparse autoencoder variant, achieves an impressive accuracy of 91%, outperforming all baselines. These results underscore the robustness of our DL approach in capturing the hidden medically relevant meanings embedded within superposition in stop words, which are often overlooked in traditional interpretability analyses.

Model Steering. Meaningful dictionary features should effectively steer model behavior by increasing the likelihood of related codes (Templeton et al., 2024). We confirm this in the multilabel setting by "clamping" relevant feature activations, demonstrating that we can drive our model to predict specific subsets of medical codes without additional tokens, modifying model weights, or explicitly training steering vectors (Subramani et al., 2022). This finding may inspire cheaper alternatives for quickly modifying model behavior, especially as ICD coding models become larger.

Setup. To measure the direct impact of each dictionary feature on ICD code predictions, we input pad tokens to generate a blank canvas of PLM embeddings. Instead of ablating decomposed dictionary features (eq. 7), we manually "clamp" or increase each feature’s activation to a large value (50) and reconstruct new embeddings. We then measure the increases in downstream ICD code probabilities, counting the number of ICD codes with a softmax probability increase of 0.5 or more (i.e., a decision flip) and their respective number of dictionary features. To validate the meaningfulness of each clamped dictionary feature, we construct a new dictionary with the top ICD code probabilities increased by each clamped feature and rerun the hidden meaning identification experiment from Table 2.

Results. Table 3 shows that while not all dictionary features are meaningful in steering model

Ablating Dictionary Features of Highlighted Tokens													
Experiment	AutoCodeDL		LAAT + Baselines					DL		Baselines			
	L1	SPINE	ICA	PCA	Identity	Random	Token	L1	SPINE	ICA	PCA	Identity	Random
Top (Comp) \uparrow	0.834	0.862	4.347e-5	0.834	0.575	0.845	0.678	0.878	0.959	5e-3	0.909	0.806	0.967
NT \downarrow	2.568	2.703	8.565e-3	2.628	2.105	387.901	2.640	183.530	15.850	0.064	47.000	758.140	256.136
Ratio \uparrow	0.326	0.319	0.051	0.318	0.273	0.002	0.316	0.005	0.061	0.008	0.019	0.001	0.003

Table 1: Softmax probability changes in downstream ICD predictions resulting from feature ablations (i.e., comprehensiveness). “Top” represents the mean magnitude of softmax drops for the most probable ICD code, while “NT” signifies the sum of absolute softmax probability changes of other ICD codes for each clinical note. The “Ratio” indicates the ratio between these two measures. We bold and distinguish the results obtained using our combined LAAT and dictionary learning framework, and observe that our method has the most precise effect on downstream ICD predictions, suggesting improved explanatory power.

Hidden Medical Meaning Identification Accuracy						
AutoCodeDL		Baselines				
L1	SPINE	ICA	PCA	Identity	Random	
0.91	0.89	0.40	0.53	0.61	0.37	

Table 2: Proportion of stop word embedding labels correctly identified by our AutoCodeDL framework, alongside the baseline methods. Such results showcase that DL is capable of effectively identifying hidden meanings embedded within superposition.

Model Steering Experiment with Dictionary Features						
Metrics	AutoCodeDL		Baselines			
	L1	SPINE	ICA	PCA	Identity	Random
No. Code Flips \uparrow	3449	3681	0	511	1	3681
No. Meaningful f_i \uparrow	928	1353	0	10	1	768
ID Accuracy \uparrow	0.55	0.89	0.29	0.32	0.26	0.26

Table 3: Table 3: Model steering experiment results comparing AutoCodeDL (L1 and SPINE) with baselines. Effectively changing all medical codes and still attaining high identification accuracy through the discovery of clamped classes, AutoCodeDL with SPINE is capable of steering the model in highly interpretable ways.

behavior, the dictionary embeddings can change the predictions of nearly all ICD codes. Moreover, the clamped classes used to generate the dictionary can still recover the hidden medical codes of extraneous stop tokens, suggesting each feature’s explainability. Further verification is provided by a UMAP plot (Figure 4), where the color indicates the maximum increase in probability of its top medical code. The plot reveals clusters of semantically meaningful features with a direct ability to change model predictions of specific subsets of codes, annotated based on their dictionary contexts and relevant medical codes.

4.2 Human Understandability of Dictionaries

Evaluating the human understandability of interpretability methods lacks a clear definition, often

depending on qualitative assessments (Zhang et al., 2021; Alangari et al., 2023). Given the limited availability of clinically licensed physicians, there’s a need for scalable proxy metrics to measure the understandability of learned dictionaries across extensive clinical notes, encompassing thousands of tokens. Through manual examination of various dictionary features, we pinpoint two primary factors that contribute to a dictionary feature’s understandability.

Coherence. A dictionary feature is deemed understandable if its top tokens are semantically related, indicating a clear conceptual identity. Conversely, semantic randomness among tokens complicates the identification of a feature’s underlying concept.

Setup. Advancements in Siamese encoder transformers, like Siamese BERT, have boosted the efficiency and effectiveness of semantic similarity analyses (Reimers and Gurevych, 2019), mimicking human perceptions of text pair similarity. Utilizing Siamese encoder embeddings, we calculate the average cosine similarity among the top k tokens of each dictionary feature to gauge their conceptual relatedness. For more methodological specifics of our Siamese BERT experiment, please refer to Appendix A.10. Although LAAT is part of our method for better explaining downstream ICD predictions, we also treat the LAAT matrix as a “dictionary” of ICD codes, similar to dictionary features learned by sparse autoencoders. While comparing supervised (LAAT) and unsupervised (autoencoders) methods isn’t perfect, LAAT remains the current standard for coherence in this area of automated ICD coding with PLMs.

Results. The coherence of the unsupervised dictionary learning (DL) methods, represented by the DL columns, decreased as the number of top k to-

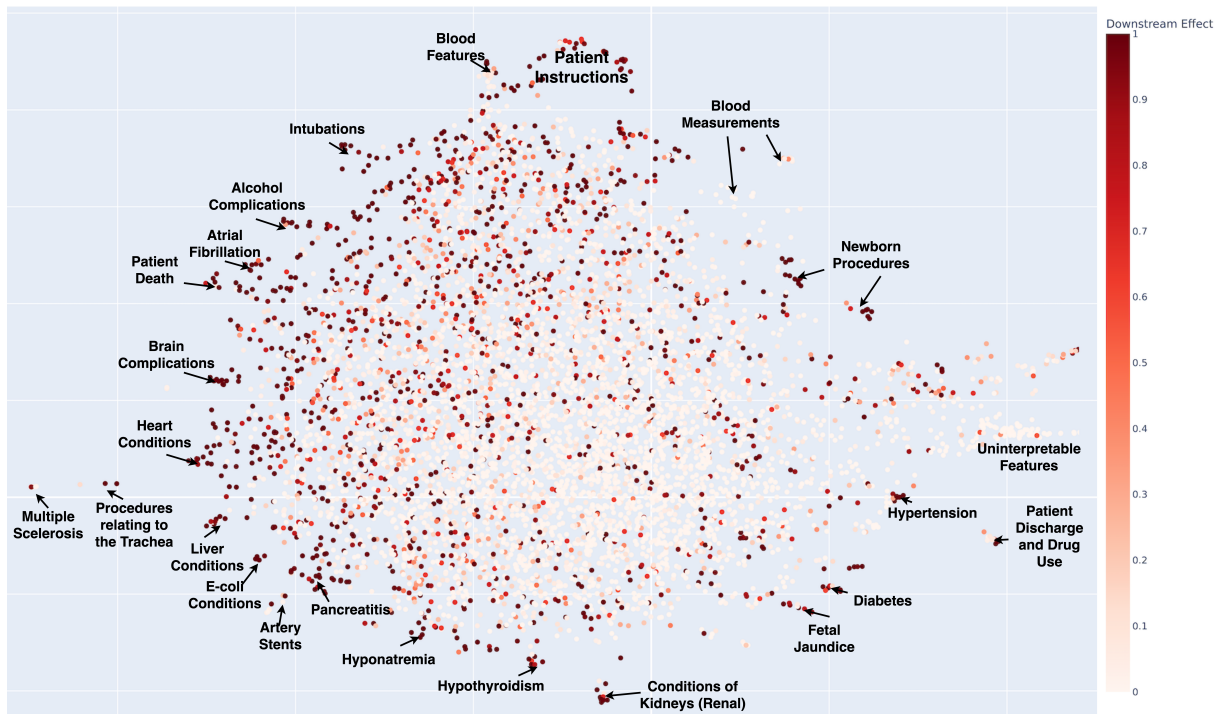


Figure 4: UMAP of SPINE Embeddings: Dictionary features are interpretable in steering model behavior. More darker red colors indicates higher maximum observed probability increase of the top medical code from a feature’s exclusive clamping. Each dot is a dictionary feature embedding projected into 2D.

kens increased. Contrary to expectations from Subramanian et al. (2017), sparse autoencoders with L1 minimization exhibited the highest coherence among unsupervised methods, with the highest average cosine similarity across different k values. Qualitative examples and the relationship between dictionary contexts and coherence are provided in Appendix A.11.1. The most interpretable dictionary features correspond to high cosine similarities, potentially filtering out easy cases for human annotators to focus on more complex features.

Coherence of All Activated Dictionary Features							
$k \uparrow$	DL		Baselines				Supervised
	L1	SPINE	ICA	PCA	Identity	Random	LAAT
2	0.3130	0.3074	0.2138	0.2371	0.2498	0.2591	0.4185
4	0.2981	0.2872	0.2133	0.2294	0.2440	0.2449	0.4083
10	0.2747	0.2684	0.2095	0.2218	0.2384	0.2358	0.3889

Table 4: Average cosine similarity between the top k tokens extracted from each dictionary feature or ICD code, measured from Siamese encoder embeddings. Higher values indicate a stronger thematic connection within the feature or code. The "DL" columns represent our dictionaries constructed, while the remaining columns are baselines.

Distinctiveness. If a dictionary feature has a clear, coherent theme based on its top tokens, unre-

lated tokens should be readily discernible.

Setup. Inspired by Subramanian et al. (2017), who evaluated the distinctiveness of dictionary features in sparse word embeddings, we investigate by sampling a dictionary feature’s top 4 ($k=4$) activating tokens (and their nearby context windows) and a randomly sampled token outside the feature. An interpretability score is derived by having medical experts (a licensed physician and a medical scientist trainee) identify the randomly sampled token from the set of 5 tokens. The proportion of correctly identified tokens serves as our distinctiveness metric. Due to time constraints, our medical experts evaluated 100 samples each for DL and other baseline encoders. However, given the demonstrated capabilities of state-of-the-art language models in text annotation tasks (Huang et al., 2023; Gilardi et al., 2023) and their extensive medical vocabulary (Bommineni et al., 2023), we utilized the current medically quantized state-of-the-art OpenBioLLM Llama 3 70B model across all dictionary features from the dictionaries constructed in the stop words experiment (Section 3.2).

Results. As shown in Table 5, both sparse autoencoders are more distinctive than their respective unsupervised baselines. Surprisingly, the random and identity encoders are more distinguishable

than their PCA counterparts. We provide examples in Appendix A.11.2 illustrating that the ability to differentiate features can range from exceptionally obvious cases with repeating tokens or differences in specificity to completely uninterpretable cases.

Percentage of Dictionary Features Differentiated							
	DL		Baselines				Supervised
	L1	SPINE	ICA	PCA	Identity	Random	LAAT
No. LLM Id. \uparrow	2828	2713	282	226	297	299	2117
% LLM Id. \uparrow	0.46	0.44	0.37	0.29	0.39	0.39	0.58
% Human Id. (100) \uparrow	0.49	0.56	0.45*	0.41	0.45	0.44	*

Table 5: Percentage of dictionary features successfully distinguished by the quantized OpenBioLLM-70B Llama 3 model and medical experts, determined by correctly identifying the unrelated token from a set of 5 tokens (including the top 4 activating tokens and their context windows) highlighted by our sparse autoencoders. * denotes cases where human evaluations were omitted or reduced (i.e., 40) due to time constraints.

5 Conclusion

This study introduces a novel method that combines dictionary learning with label attention mechanisms to improve the interpretability of medical coding language models. By uncovering interpretable dictionary features from dense language embeddings, the proposed approach offers a dictionary-based rationale for ICD code predictions, addressing the growing demand for transparency in automated healthcare decisions. This work lays the groundwork for future explorations in dictionary learning to enhance healthcare interpretability.

Limitations

We note that there are several key limitations of the dictionary learning applied here. First, as noted in the training details within the appendix, the sparse autoencoders are unable to perfectly reconstruct PLM embeddings, and thus cannot fully capture a model’s downstream performance (Cunningham et al., 2023; Bricken et al., 2023; Templeton et al., 2024). Furthermore, we note as similarly explored by (Cunningham et al., 2023; Bricken et al., 2023; Templeton et al., 2024) that there exist dead or missing features, features that do not activate regardless of token embedding or concepts that exist in the model but are not captured within the dictionary, indicating not all features are useful or interpretable.

In terms of the coherence of highly activating features, we highlight that the sparse dictionaries learned are not as coherent as supervised mappings

of the LAAT, limiting its human understandability. Furthermore, we observe that simply interpreting the PLM embeddings at the last layer, which (Yun et al., 2023b; Cunningham et al., 2023) claim to be the most interpretable, is insufficient for fully capturing and interpreting a PLM’s behavior. Referring to Table 3, the resolution of our dictionary features is lacking in comparison to the granularity of medical codes, where we observe that the total number of highly meaningful features (i.e., those that can change a medical code’s prediction) is smaller than the number of medical codes highly affected, suggesting that the features learned most likely represent higher abstract concepts than highly-specific medical codes.

Recently, (Templeton et al., 2024) has explored scaling up sparse autoencoders to potentially millions of dictionary features for extremely large language models. They show that not all semantic features are learnable, especially in cases without access to large amounts of diverse data (Bricken et al., 2023). While synthetic data generation exists to further augment the diversity of tokens in the dictionary learning corpus, the size of sparse autoencoders must also increase as model complexity increases (Templeton et al., 2024) in order to maintain the granularity of dictionary features. Regardless, such mechanistic explanations are still fairly efficient when compared to black-box alternatives. To the best of our knowledge, its applicability to smaller scale language datasets is unknown.

While our experiments revealed several key benefits of using sparse autoencoders for interpretability, future work offers exciting avenues to push medical PLM interpretability further. Despite SPINE’s (Subramanian et al., 2017) improvements over vanilla L_1 methods, sparse autoencoders’ reconstruction limitations can restrict their ability to capture all information relevant to downstream predictions. Exploring more expressive representation learning techniques like causal and disentanglement methods (Schölkopf et al., 2021; Wang et al., 2023) could hold promise for interpretability gains. Additionally, unraveling the complex relationships within medical PLMs through both automated circuit discovery (Conmy et al., 2023) and dictionary learning approaches (Cunningham et al., 2023) could offer further valuable insights into an ICD code prediction.

Ethics Statement

We note that all clinical notes used from the MIMIC-III dataset (Johnson et al., 2023, 2016) are deidentified and that our method is heavily focused on medical tokens and their conceptual meanings towards ICD predictions. Thus, we follow guidelines laid out by PhysioNet’s MIMIC3 health data license (Johnson et al., 2016) and note our study does not contain any additional patient information that can lead to privacy violations. We note that our expert human annotators have consented and are our collaborators.

References

- Samira Abnar and Willem Zuidema. 2020. [Quantifying attention flow in transformers.](#)
- Nourah Alangari, Mohamed El Bachir Menai, Hassan Mathkour, and Ibrahim Almosallam. 2023. [Exploring evaluation methods for interpretable machine learning: A survey.](#) *Information*, 14(8).
- Felipe Almeida and Geraldo Xexéo. 2023. [Word embeddings: A survey.](#)
- Vikas L Bommineni, Sanaea Bhagwagar, Daniel Balcarcel, Vishal Bommineni, Christos Davazitkos, and Donald Boyer. 2023. [Performance of chatgpt on the mcat: The road to personalized and equitable pre-medical learning.](#) *medRxiv*.
- Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermy, Tom Conerly, Nick Turner, Cem Anil, Carson Denison, Amanda Askell, Robert Lasenby, Yifan Wu, Shauna Kravec, Nicholas Schiefer, Tim Maxwell, Nicholas Joseph, Zac Hatfield-Dodds, Alex Tamkin, Karina Nguyen, Brayden McLean, Josiah E Burke, Tristan Hume, Shan Carter, Tom Henighan, and Christopher Olah. 2023. [Towards monosemanticity: Decomposing language models with dictionary learning.](#) *Transformer Circuits Thread*. <https://transformer-circuits.pub/2023/monosemantic-features/index.html>.
- Pengfei Cao, Chenwei Yan, Xiangling Fu, Yubo Chen, Kang Liu, Jun Zhao, Shengping Liu, and Weifeng Chong. 2020. [Clinical-coder: Assigning interpretable ICD-10 codes to Chinese clinical notes.](#) In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 294–301, Online. Association for Computational Linguistics.
- Chun Sik Chan, Huanqi Kong, and Liang Guanqing. 2022. [A comparative study of faithfulness metrics for model interpretability methods.](#) In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5029–5038, Dublin, Ireland. Association for Computational Linguistics.
- Aditya Chattopadhyay, Anirban Sarkar, Prantik Howlader, and Vineeth N Balasubramanian. 2018. [Grad-CAM++: Generalized gradient-based visual explanations for deep convolutional networks.](#) In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE.
- Hila Chefer, Shir Gur, and Lior Wolf. 2021. [Generic attention-model explainability for interpreting bimodal and encoder-decoder transformers.](#)
- Hila Chefer, Idan Schwartz, and Lior Wolf. 2022. [Optimizing relevance maps of vision transformers improves robustness.](#)
- Hugh Chen, Scott M. Lundberg, and Su-In Lee. 2022. [Explaining a series of models by propagating shapley values.](#) *Nature Communications*, 13(1):4512.
- Arthur Conmy, Augustine N. Mavor-Parker, Aengus Lynch, Stefan Heimersheim, and Adrià Garriga-Alonso. 2023. [Towards automated circuit discovery for mechanistic interpretability.](#)
- Hoagy Cunningham, Aidan Ewart, Logan Riggs, Robert Huben, and Lee Sharkey. 2023. [Sparse autoencoders find highly interpretable features in language models.](#)
- Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C. Wallace. 2020. [Eraser: A benchmark to evaluate rationalized nlp models.](#)
- Andres Duque, Hermenegildo Fabregat, Lourdes Araujo, and Juan Martinez-Romo. 2021. [A keyphrase-based approach for interpretable icd-10 code classification of spanish medical reports.](#) *Artificial Intelligence in Medicine*, 121:102177.
- Joakim Edin, Alexander Junge, Jakob D. Havtorn, Lasse Borgholt, Maria Maistro, Tuukka Ruotsalo, and Lars Maaløe. 2023. [Automated medical coding on mimic-iii and mimic-iv: A critical review and replicability study.](#) In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR ’23*, page 2572–2582, New York, NY, USA. Association for Computing Machinery.
- Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, Roger Grosse, Sam McCandlish, Jared Kaplan, Dario Amodei, Martin Wattenberg, and Christopher Olah. 2022. [Toy models of superposition.](#) *Transformer Circuits Thread*. https://transformer-circuits.pub/2022/toy_model/index.html.
- Subhroshekhar Ghosh, Aaron Y. R. Low, Yong Sheng Soh, Zhuohang Feng, and Brendan K. Y. Tan. 2023. [Dictionary learning under symmetries via group representations.](#)

- Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. [Chatgpt outperforms crowd workers for text-annotation tasks](#). *Proceedings of the National Academy of Sciences*, 120(30).
- Hajar Hakkoum, Ibtissam Abnane, and Ali Idri. 2022. [Interpretability in the medical field: A systematic mapping and review study](#). *Applied Soft Computing*, 117:108391.
- J A Hirsch, G Nicola, G McGinty, R W Liu, R M Barr, M D Chittle, and L Manchikanti. 2016. ICD-10: History and context. *AJNR Am J Neuroradiol*, 37(4):596–599.
- Chao-Wei Huang, Shang-Chi Tsai, and Yun-Nung Chen. 2022. [Plm-icd: Automatic icd coding with pretrained language models](#).
- Fan Huang, Haewoon Kwak, and Jisun An. 2023. [Is chatgpt better than human annotators? potential and limitations of chatgpt in explaining implicit hate speech](#). In *Companion Proceedings of the ACM Web Conference 2023, WWW '23*. ACM.
- Alistair E. W. Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J. Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, Li-wei H. Lehman, Leo A. Celi, and Roger G. Mark. 2023. [Mimic-iv, a freely accessible electronic health record dataset](#). *Scientific Data*, 10(1):1.
- Alistair E.W. Johnson, Tom J. Pollard, Lu Shen, Li-wei H. Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G. Mark. 2016. [Mimic-iii, a freely accessible critical care database](#). *Scientific Data*, 3(1):160035.
- Renee L Johnson, Holly Hedegaard, Emilia S Pasalic, and Pedro D Martinez. 2021. Use of ICD-10-CM coded hospitalisation and emergency department data for injury surveillance. *Inj Prev*, 27(S1):i1–i2.
- Rajvir Kaur, Jeewani Anupama Ginige, and Oliver Obst. 2021. [A systematic literature review of automated icd coding and classification systems using discharge summaries](#).
- Patrick Lewis, Myle Ott, Jingfei Du, and Veselin Stoyanov. 2020. [Pretrained language models for biomedical and clinical tasks: Understanding and extending the state-of-the-art](#). In *Proceedings of the 3rd Clinical Natural Language Processing Workshop*, pages 146–157, Online. Association for Computational Linguistics.
- Scott Lundberg and Su-In Lee. 2017. [A unified approach to interpreting model predictions](#).
- Scott M. Lundberg, Gabriel Erion, Hugh Chen, Alex DeGrave, Jordan M. Prutkin, Bala Nair, Ronit Katz, Jonathan Himmelfarb, Nisha Bansal, and Su-In Lee. 2020. [From local explanations to global understanding with explainable ai for trees](#). *Nature Machine Intelligence*, 2(1):56–67.
- Raha Moraffah, Mansooreh Karami, Ruo Cheng Guo, Adrienne Raglin, and Huan Liu. 2020. [Causal interpretability for machine learning – problems, methods and evaluation](#).
- Edoardo Mosca, Ferenc Szigeti, Stella Tragianni, Daniel Gallagher, and Georg Groh. 2022. [SHAP-based explanation methods: A review for NLP interpretability](#). In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4593–4603, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.
- James Mullenbach, Sarah Wiegrefe, Jon Duke, Jimeng Sun, and Jacob Eisenstein. 2018. [Explainable prediction of medical codes from clinical text](#).
- Tomáš Musil. 2019. [Examining structure of word embeddings with pca](#).
- Tomáš Musil and David Mareček. 2022. [Independent components of word embeddings represent semantic features](#).
- Chris Olah, Nick Cammarata, Ludwig Schubert, Gabriel Goh, Michael Petrov, and Shan Carter. 2020. [Zoom in: An introduction to circuits](#). *Distill*. <https://distill.pub/2020/circuits/zoom-in>.
- Bruno A. Olshausen and David J. Field. 1997. [Sparse coding with an overcomplete basis set: A strategy employed by v1?](#) *Vision Research*, 37(23):3311–3325.
- Kimberly J O'Malley, Karon F Cook, Matt D Price, Kimberly Raiford Wildes, John F Hurdle, and Carol M Ashton. 2005. Measuring diagnoses: ICD code accuracy. *Health Serv Res*, 40(5 Pt 2):1620–1639.
- Lakshmi Narayan Pandey, Rahul Vashisht, and Harish G. Ramaswamy. 2023. [On the interpretability of attention networks](#).
- Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. [GloVe: Global vectors for word representation](#). In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.
- Preethi Rao, Shira H Fischer, Mary E Vaiana, and Erin Audrey Taylor. 2022. Barriers to price and quality transparency in health care markets. *Rand Health Q*, 9(3):1.
- Nils Reimers and Iryna Gurevych. 2019. [Sentence-bert: Sentence embeddings using siamese bert-networks](#).
- Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. ["why should i trust you?": Explaining the predictions of any classifier](#).
- Tilman Räuher, Anson Ho, Stephen Casper, and Dylan Hadfield-Menell. 2023. [Toward transparent ai: A survey on interpreting the inner structures of deep neural networks](#).

- Wojciech Samek, Alexander Binder, Grégoire Montavon, Sebastian Bach, and Klaus-Robert Müller. 2015. [Evaluating the visualization of what a deep neural network has learned.](#)
- Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. 2021. [Towards causal representation learning.](#)
- Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. 2019. [Learning important features through propagating activation differences.](#)
- Nishant Subramani, Nivedita Suresh, and Matthew E. Peters. 2022. [Extracting latent steering vectors from pretrained language models.](#)
- Anant Subramanian, Danish Pruthi, Harsh Jhamtani, Taylor Berg-Kirkpatrick, and Eduard Hovy. 2017. [Spine: Sparse interpretable neural embeddings.](#)
- Yuanbo Tang, Zhiyuan Peng, and Yang Li. 2023. [Explainable trajectory representation through dictionary learning.](#)
- Adly Templeton, Tom Conerly, Jonathan Marcus, Jack Lindsey, Trenton Bricken, Brian Chen, Adam Pearce, Craig Citro, Emmanuel Ameisen, Andy Jones, Hoagy Cunningham, Nicholas L Turner, Callum McDougall, Monte MacDiarmid, C. Daniel Freeman, Theodore R. Sumers, Edward Rees, Joshua Batson, Adam Jermy, Shan Carter, Chris Olah, and Tom Henighan. 2024. [Scaling monosemanticity: Extracting interpretable features from claude 3 sonnet.](#) *Transformer Circuits Thread.*
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2023. [Attention is all you need.](#)
- Thanh Vu, Dat Quoc Nguyen, and Anthony Nguyen. 2020. [A label attention model for icd coding from clinical text.](#) In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-PRICAI-2020*. International Joint Conferences on Artificial Intelligence Organization.
- Xin Wang, Hong Chen, Si’ao Tang, Zihao Wu, and Wenwu Zhu. 2023. [Disentangled representation learning.](#)
- Ruiyu Xu, Chao Wang, Yongxiang Li, and Jianguo Wu. 2023. [Generalized time warping invariant dictionary learning for time series classification and clustering.](#)
- Chenwei Yan, Xiangling Fu, Xien Liu, Yuanqiu Zhang, Yue Gao, Ji Wu, and Qiang Li. 2022. [A survey of automated international classification of diseases coding: development, challenges, and applications.](#) *Intelligent Medicine*, 2(3):161–173.
- Zhichao Yang, Sanjit Singh Batra, Joel Stremmel, and Eran Halperin. 2023. [Surpassing gpt-4 medical coding with a two-stage approach.](#)
- Zeyu Yun, Yubei Chen, Bruno A Olshausen, and Yann LeCun. 2023a. [Transformer visualization via dictionary learning: contextualized embedding as a linear superposition of transformer factors.](#)
- Zeyu Yun, Yubei Chen, Bruno A Olshausen, and Yann LeCun. 2023b. [Transformer visualization via dictionary learning: contextualized embedding as a linear superposition of transformer factors.](#)
- Yu Zhang, Peter Tino, Ales Leonardis, and Ke Tang. 2021. [A survey on neural network interpretability.](#) *IEEE Transactions on Emerging Topics in Computational Intelligence*, 5(5):726–742.
- Zheng Zhang, Yong Xu, Jian Yang, Xuelong Li, and David Zhang. 2015a. [A survey of sparse representation: Algorithms and applications.](#) *IEEE Access*, 3:490–530.
- Zheng Zhang, Yong Xu, Jian Yang, Xuelong Li, and David Zhang. 2015b. [A survey of sparse representation: Algorithms and applications.](#) *IEEE Access*, 3:490–530.

A Appendix

A.1 Sparse Autoencoder Training Details

We train our sparse autoencoders on the PLM activations generated by a 110M medical RoBERTa encoder PLM on the cleaned MIMIC-III train dataset of 38,427 clinical notes and evaluated on their test set of 8,750 clinical notes for a total of 52,712 clinical notes, as detailed by (Edin et al., 2023). We follow the advice of (Bricken et al., 2023) and (Subramanian et al., 2017) in training the L_1 and SPINE autoencoders respectively. Our hyperparameters are shown in Table 6. We also use AdamW as our optimizer. For a fair comparison, we reuse the dictionary feature size m as it has been noted by (Bricken et al., 2023) that larger dictionary feature sizes can potentially increase the resolution of concepts of dictionary features. For instance, dictionary feature may be further decomposed into features of more specific meanings given different token sequence contexts as further discussed by (Bricken et al., 2023). We use PyTorch as our deep learning framework of choice.

Table 6: Sparse Autoencoder Training Details

λ_{L_1}	λ_1	λ_2	m	Batch Size	lr
2e-5	1	1	6,144	8,192	1e-3

We train on randomly sampled embeddings from the training set, and filter out all pad tokens, that are irrelevant to the final prediction, but dominate each

batch due to their use in making GPU inference fast. We note that we cannot get perfect reconstruction loss nor completely match the downstream performance of the original model with our reconstructed embeddings on the test set, but we get very close.

Table 7: Autoencoder Test Loss Metrics

	L1	SPINE	Original
Test Autoencoder Loss	69.46	39.59	N/A
Test F1	0.258	0.260	0.262

A.2 Scalability Discussion

Training these sparse autoencoders takes approximately 6 minutes per epoch on A6000 GPUs. However, the current training process samples new token embeddings for every batch of clinical notes, which is suboptimal. A quick adaptation of existing dataloaders to improve token diversity during training could be beneficial. We find that precomputing and caching PLM embeddings can significantly reduce training time to approximately 15 minutes for 10 epochs, albeit requiring substantial memory (at least 128 GB RAM for caching millions of embeddings). In contrast, decomposing PLM embeddings using our method is extremely fast, on par with LAAT (approximately 0.04 seconds per clinical note). While there is an upfront cost (a couple of hours for sorting and sampling millions of tokens for each dictionary feature), once a dictionary is constructed, interpreting the embedding space of any clinical note is very fast and efficient, which is suitable for this high dimensional multilabel task.

A.3 Build Dictionary

For further clarity, we write up our dictionary construction algorithm here in algorithm 1. In principle, one is just sorting based on encoded activations and ablation softmax drops of different ICD codes for each dictionary feature.

A.4 Additional Baseline Details

There were four main baseline methods that were compared against our exploration of two sparse autoencoders, specifically an ICA encoder, PCA encoder, an identity encoder, and a random encoder. All of their implementations were taken from (Cunningham et al., 2023). The ICA encoder was trained using the FastICA decomposition method from scikit-learn to estimate the activation’s respective independent components that act

Algorithm 1: Build Dictionary

Input: Autoencoder A , feature f_i , tokens x
Output: Dictionary F mapping f_i to tokens x and classes y

```

1  $F \leftarrow \text{dict}$ ; for each token  $x$  do
2    $f \leftarrow A.\text{encode}(x)$ ; for  $f_i$  in  $f$  do
3     if  $f_i > F[i].f_i$  then
4        $F[i].\text{tokens} \leftarrow x$ ;
5        $\delta_i \leftarrow \text{ablation}(f_i)$ ; if  $\delta_i > F[i].\delta$ 
6         then
7            $F[i].\text{classes} \leftarrow \text{drops}(\delta_i)$ ;
7 return  $F$ ;

```

as the encoder weights. However, we note that the training was unstable, most likely due to the same memory and computation limitations that (Cunningham et al., 2023) faced. As a result, we limit the training on only 2,000,000 token embeddings sampled from the training set. The PCA covariance matrix was estimated batchwise with its eigenvectors acting as the encoder weights. For the random and identity encoders, the random encoder was initialized with a normal distribution of mean 0 and variance 1 and the identity encoder has an identity matrix for its encoder weights.

A.5 Label Attention Details

We recognize that we don’t explicitly describe LAAT (Vu et al., 2020; Huang et al., 2022) in detail in the main manuscript. For interested readers, we depict the label attention mechanism in Figure 5.

Essentially, LAAT computes a cross attention score for each token and ICD code, creating a label attention matrix where each row is an ICD and every column is the token’s attention score with respect to that ICD code.

A.6 Additional SPINE Details

While the L_1 minimization is most commonly used to train sparse autoencoders due to its simplicity of training (Zhang et al., 2015b), we also revisit an alternative sparse formulation SPINE proposed by (Subramanian et al., 2017), specifically designed to decompose neural word embeddings. (Subramanian et al., 2017) showed that they could improve interpretability in GloVe (Pennington et al., 2014) and other forms of neural token embeddings (Almeida and Xexéo, 2023). Their formulation replaces the L_1 regularization loss in favor of using a

Highlighting Tokens with Label Attention Pipeline

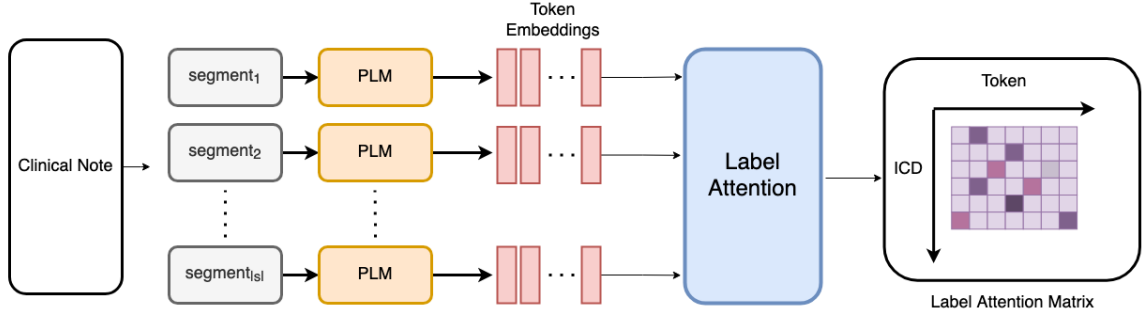


Figure 5: Label Attention identifies the most relevant tokens for each ICD code through a label attention matrix.

combination of average sparsity and a partial sparsity loss function terms. We outline their method in the equations below.

We define our average sparsity term below in equation 9.

$$\mathcal{L}_{\text{asl}} = \sum_i \max(f_i, \rho) \quad (9)$$

Here, ρ is a user-chosen sparsity hyperparameter that prevents the dictionary feature activations f from becoming too large. In practice, one should ideally average each sparse feature activation \bar{f}_i across the entire training set first as outlined by (Subramanian et al., 2017). However, due to memory constraints, we simply average over the batch size.

The partial sparsity term is defined as follows below.

$$\mathcal{L}_{\text{psl}} = \sum_i f_i \cdot (1 - f_i) \quad (10)$$

The general intuition behind the following partial sparsity term is that it further enforces sparsity by penalizing the distribution of f_i 's away from uniformity, which can invariably occur due to the stochastic nature of gradient optimization algorithms. The resulting alternative sparse autoencoder minimization function is shown in equation 11 .

$$\mathcal{L} = \frac{1}{|X|} \sum_{x \in X} \|x - \hat{x}\|_2^2 + \lambda_1 \mathcal{L}_{\text{asl}} + \lambda_2 \mathcal{L}_{\text{psl}} \quad (11)$$

A.7 Baseline Ablation Experiment Details

For every baseline, we iterate through each activated dictionary feature for the tokens, measuring the softmax drop for the selected ICD code

(with the highest softmax probability). This is repeated for each none-DL baseline: In ICA, we ablate embedding independent components. In PCA, we ablate principal components (i.e eigenvectors weighed by their eigenvalues). In the identity encoder, we ablate embedding dimensions. In the random encoder, we essentially add random noise. For the Token baseline, we ablate entire token embeddings.

A.8 Sufficiency Discussion

If comprehensiveness measures the performance drop when ablating key features, sufficiency measures the performance retention when keeping only the key features (Chan et al., 2022). In particular, rather than performing an ablation, we simply set all of the token embeddings to their respective dictionary feature embeddings multiplied by their activations. Clinical coding is high-dimensional, making it computationally expensive to compute different quantiles. Therefore, we only compute metrics for the highly relevant 95% quantile tokens when using LAAT, or for all tokens when not using LAAT.

$$\tilde{x} = f_i \cdot h_i \quad (12)$$

However, when using our specific feature encoders, this metric yields mixed results. While retaining only the relevant token embeddings defined by LAAT preserves the model's class probabilities, we find that a randomly generated embedding from our random encoder remarkably performs similarly or better than the LAAT explanation. This observation raises further questions about the potential inner workings and importance of the embedding space in explaining model behavior. Especially, as the steering experiment is effectively related to the sufficiency metric, but with difference being

f_i 's artificial expansion versus its original encoding. Understanding these challenges in embedding interpretability is an important point of future work.

A.9 Hidden Meaning Stop Words Experiment

We explain more details of our stop words experiments here. To begin, we sample *all* (not just the *attention highlighted*) tokens from 1,600 clinical notes sampled from the test set and then map highly activating tokens to each dictionary feature as well as relevant ICD code predictions through dictionary feature ablations outlined in section 3.2 and 3.3. Then, we collect (12,891) highly relevant stop words (using NLTK) identified by the label attention mechanism, documenting their corresponding labels and PLM embeddings as defined by the label attention matrix. After shuffling each label-embedding pairing, we employ our dictionary to query the stop word's relevant classes via a trained sparse autoencoder, assessing if the original label ranks among the top 10 classes of each of its most highly activated dictionary features. As the magnitudes of different dictionary features varies across decomposition methods and are mostly sparse in sparse autoencoders, we define highly activated features f_i as dictionary features that exceed the 96.5th percentile feature magnitude of encoded from each token embedding. In practice, since the encoded dictionary features for each token by sparse autoencoders are sparse, the highest activated dictionary features are those that are nonzero. The efficacy of dictionary features in clarifying a stop word's significance to an ICD code prediction is quantified by calculating the proportion of shuffled stop word embeddings are correctly contained in these prediction sets generated by their respective dictionaries.

A.10 Additional Siamese BERT Cosine Similarity Experiment Details

We use "all-mpnet-base-v2" from the sentence-transformers package as our Siamese encoder. We show our steps for computing the cosine similarity scores in our evaluation of coherence scores below in algorithm 2. Furthermore, since sparse encoding is a more efficient operation than searching and ablating the most relevant dictionary feature for an ICD prediction, we sample all tokens from every clinical note in the test set, and discern all of the top tokens for each dictionary feature with our sparse autoencoder.

Algorithm 2: Cosine Similarity Score for Dictionary Features

Input: Siamese BERT model M , dictionary feature f_i ;

k number of highly activating tokens in dictionary $T = \{t_{f_i,1}, t_{f_i,2}, \dots, t_{f_i,k}\}$;

Output: Cosine Similarity Score \bar{s}

```

1  $\bar{s} \leftarrow 0$ ;
2 for each  $f_i$  do
3    $t_{pairs} \leftarrow \{(t_{f_i,a}, t_{f_i,b}) \mid a \neq b, t_{f_i,a}, t_{f_i,b} \in T\}$ ;
4   for each  $(t_{f_i,a}, t_{f_i,b})$  in  $t_{pairs}$  do
5      $\hat{s} = \frac{M(t_{f_i,a}) \cdot M(t_{f_i,b})}{\|M(t_{f_i,a})\| \cdot \|M(t_{f_i,b})\|}$ ;
6      $\bar{s} \leftarrow \bar{s} + \frac{\hat{s}}{|t_{pairs}|}$ ;
7  $\bar{s} \leftarrow \frac{\bar{s}}{|f|}$ 

```

A.11 Human Evaluations

We conducted human evaluations with a medical scientist trainee and a licensed physician, specifically the distinctiveness experiment inspired by Subramanian et al. (2017). Below, we showcase examples of incoherent dictionary features (low cosine similarity) and highly coherent features (high cosine similarity). We also provide examples from the human evaluations, including cases where annotators failed to identify the randomly chosen context and cases where it was easy for them to distinguish the random token.

Ablating Dictionary Features of Highlighted Tokens													
Experiment	AutoCodeDL		LAAT + Baselines					DL		Baselines			
	L1	SPINE	ICA	PCA	Identity	Random	Token	L1	SPINE	ICA	PCA	Identity	Random
Suff. ↓	0.143	0.462	0.470	0.424	0.468	-0.029	-0.012	0.137	0.454	0.470	0.112	-0.005	-0.033

Table 8: Softmax probability changes in downstream ICD predictions resulting from sufficiency experiments.

A.11.1 Dictionary Features and Coherence

We provide qualitative examples in Figures 6, 7, and 8 to demonstrate the efficacy of our cosine similarity metric. While an imperfect metric, cosine similarity can effectively discern highly interpretable dictionary features, which typically have repeating tokens. However, many dictionary features are more challenging and may require more complex annotation approaches, such as domain-specific language models, medical experts, or fine-tuning specific Siamese BERT encoders for this task, as the average cosine similarity of the top k tokens is intrinsically low due to the diversity of the text.

Dictionary Feature 1243 - AFib

ades de points **atrial fibrillation** coronary artery disease
gi bleed **atrial fibrillation** non st elevation
moderate aortic stenosis **atrial fibrillation** presented with rest
svt **afib** which was treated
to our hospital **in atrial fibrillation and** hypertensive with a
deconditioning **afib continue coumadin** beta blocker amiodarone
ecg baseline **afib** electrolytes drawn potassium
on patient did **have atrial fibrillation** with date range
with patch angioplasty **atrial fibrillation anticoagulated** history of congestive
her stay had **been afib** indeed interrogation of
mca stroke **atrial fibrillation** status post p
arm embolectomy **atrial fibrillation** insulin dependent diabetes
blood pressure was **in atrial fibrillation** with a heart
lead detected properly **and revealed afib** her ventricular lead
h o cad **a fib on coumadin** htn
high blood pressure **atrial fibrillation** breast cancer leg
home o2 **icdm afib on coumadin** abdominal aortic aneurysm
to treat your **atrial fibrillation** you are now
his paroxysmal atrial **fibrillation** during this admission
u after found **to have new afib** with rv

Figure 6: Example of highly interpretable SPINE feature with high cosine similarity: In this particular case, all activating tokens (red) with their context windows are all atrial fibrillation tokens, giving us a very high cosine similarity (i.e. close to 1).

Dictionary Feature 264 - Gastric Emptying, Pneumothorax

but declined on the **patient underwent gastric emptying study** which revealed slightly
delayed **gastric emptying** with **time** of
the left thumb secondary **paronychia** diabetes hypertension discharge
right pneumothorax on **post op day** three she was
htn **ex** hypothyroidism sj
approximately **minutes** on the **patient continued to improve** she was toler
gastric emptying with **time** of approximately **minutes**
chest x ray **post chest tube removal** revealed small right pneumothorax on **post op**
pna diabetes **type** bladder diverticulum cop
ating **food** and **fluids** by **mouth** and
was tolerating **food** and **fluids** by
p r n **lactulose** cc po
mycotic aneurysm **rupture** status post embolization
and **fluids** by **mouth** and she had
removed per protocol **chest x ray post**
mediastinal hematoma **r 4th rib fracture** discharge condition stable
time of approximately **minutes** on the **patient**
complete heart block **osteomyelitis** of the left thumb secondary
evening of postoperative **day** eight the patient
with hx **breast and ovarian** cancer and known brain
et **tube ng tube** and right i

Figure 7: Example of interpretable SPINE feature with low cosine similarity: In this case, the activating tokens (red) are diverse, containing various concepts such as food, fluids, gastric emptying, pneumothorax, and related terms. Our clamping experiments show that this dictionary feature is predictive of the code "acidosis," a common complication potentially leading to gastric emptying. For such cases, the cosine similarity metric is not informative, as these coherent features often have lower cosine similarities (closer to 0) despite sharing a common theme.

Dictionary Feature 129 - ???

medical center hospital discharge diagnosis hypotension
 with complaints of **sbc** and dizziness since
 free and was **treated** with aspirin pla
 hospital ward name **vidu oral surgery consulted** for infected tooth upper
 cholecystitis alcohol withdrawal **stable** anemia discharge condition
 renal lac with **hematoma** left iliac doctor
 for infected tooth upper **quadrant bridge consented and taken to the or for** exploration and
 removal of infected foreign body
 following new medications **plavix** 75mg daily
 of infected foreign body upper quadrant **bridge** and **s** and
 mr she had **a bms** placed in the
 manage her daily **activities** with a little
 ed was admitted **to** hospital last weekend
 carcinoma of the **lvu** va fibromyalgia h
 images reviewed of **regional** lv systolic
 and extraction of **infected** foreign body and
 diagnoses status post **colonic** perforation repair doctor last
 head lac per **i** eomi
 date of birth **sex** f service medicine
quadrant bridge and **s** and teeth and **roof**
 lv systolic **dysfunction** is new cardiac

Figure 8: **Example of an uninterpretable SPINE feature with low cosine similarity:** Various tokens are highlighted (red) without an obvious cohesive theme. As expected, these tokens result in very small cosine similarity measurements.

A.11.2 Human Distinctiveness Case Studies

We perform the word intrusion experiment from Subramanian et al. (2017) for the sparse autoencoder features and other baselines. Figures 9, 10, 11, and 12 showcase the expected unrelated "random context" in gold and the expert annotators' choices in red or blue for L_1 features. Qualitative examinations reveal varying levels of distinctiveness. Highly coherent features made it easy to distinguish the outlier token(s), but more complicated cases involving the level of abstraction of specific token contexts (Figure 10) were harder to discern, especially when the resolution of the token mattered. Other features were shown to be highly uninterpretable (Figures 11, 12).

Feature ID: 531

Please answer the following questions below using your best judgement. We have highlighted the most relevant tokens or words and their contexts around them.

Choice 0
 increased **dose** of **metropolol porcine avr** stable does not

Choice 1
2 las s p bovine avr htn

Choice 2
 Unrelated Intruder Context
 nephrectomy and left **laparoscopic para aortic** lymph
 Annotator 1 Selection
 Annotator 2 Selection

Choice 3
 present illness 76 **s p tissue avr** pfo closure c b d

Choice 4
 to get records **porcine avr** cardiomyopathy with l

Figure 9: **Example of an interpretable L_1 feature.** All highly activating tokens are related to heart conditions whereas the token "laparoscopic", an operation performed on the abdomen was clearly an outlier. We also observe the repeating abbreviations "avr".

Feature ID: 4101

Please answer the following questions below using your best judgement. We have highlighted the most relevant tokens or words and their contexts around them.

Choice 0
htn **ca** hypothyroidism sj

Choice 1
Unrelated Intruder Context
Annotator 1 Selection **Annotator 2 Selection**
colonic **poly** hypercholesterolemia **gallbladder poly** prostatic **hypertrophy** benign

Choice 2
developed in leak x3 consistent w demand

Choice 3
last name stt **g** daily daily clonidine

Choice 4
ux amputation also **developed in leak x**

Figure 10: **Example of a less interpretable L_1 feature differentiated by both annotators:** While both annotators acknowledged the lack of an explicit medical theme, they selected the set of tokens discussing a specific piece of anatomy, suggesting that the feature's interpretation activates a more abstract concept.

Feature ID: 26

Please answer the following questions below using your best judgement. We have highlighted the most relevant tokens or words and their contexts around them.

Choice 0
seizure disorder on **tegreto** admitted status post

Choice 1
metformin coronary artery **disease** continued home aspirin

Choice 2
coronary artery disease **ca** complete heart block

Choice 3
Annotator 2 Selection
adenitis with **a large retained duct stone ultimately ll** became **clear she**

Choice 4
Annotator 1 Selection
Unrelated Intruder Context
strep infection **inf** elbow toxic shock

Figure 11: **Example of an uninterpretable L_1 feature where only one annotator identified the random context:** According to our experts, in such cases where no discernible underlying theme exists, selecting the random context is essentially by chance.

Feature ID: 121

Please answer the following questions below using your best judgement. We have highlighted the most relevant tokens or words and their contexts around them.

Choice 0
mrsa abscesses intravenous drug use v

Choice 1
improved to after procedure creatinine peak to

Choice 2
plaza starting tuesday and be dialyzed
Unrelated Intruder Context

Choice 3
at dinner post op she was started
Annotator 1 Selection
Annotator 2 Selection

Choice 4
extended right colectomy right pelvic peritoneal implant excision partial wedge gastrectomy

Figure 12: Uninterpretable L_1 feature neither annotator could differentiate: Both annotators agreed that this dictionary feature lacked a distinctive theme.

Feature ID: 531

Please answer the following questions below using your best judgement. We have highlighted the most relevant tokens or words and their contexts around them.

Choice 0
increased dose of met rop lol porcine avj stable does not

Choice 1
2 as s p bovine avj htn

Choice 2
nephrectomy and left laparoscopic para aortic lymph

Choice 3
present illness 76 f s p tissue avj pfo closure c b d

Choice 4
to get records porcine avj cardiomyopathy with l

Which set of contexts is different than the others?

- Choice 0: increased dose of met rop lol porcine av r stable does not
- Choice 1: 2 as s p bovine av r h t n
- Choice 2: nephrectomy and left laparoscopic para aortic lymph
- Choice 3: present illness 76 f s p tissue av r pf o closure c b d
- Choice 4: to get records porcine av r cardiomyopathy with l

Figure 13: Example of human evaluation form interface.

L1 Identification Survey Form

This survey form is part of an interpretability research project. Your decision to complete this form is voluntary. The only information that will be recorded are your responses, which may later be submitted or published at public scientific venues. Clicking the final SUBMIT button indicates that you agree to this response form voluntarily.

Figure 14: Example of human evaluation form instructions.

A.12 Initial Sparse Autoencoder Experiments

We trained new sparse autoencoders for this revision, as we felt the previous ones were ill-trained due to not filtering out pads during the training process, leading to many irrelevant medical features. Surprisingly, despite finding that only approximately 500 features accounted for many of the ablation downstream results, their feature ablations still highly affected downstream performance, as shown in Table 9. Note that these experiments measured the top ground truth medical code rather than the top predicted medical code, hence the discrepancy in the probability drops. However, the order of these results remains the same. We also report the previous results, such as coherence and similarity, in Section A.12.2. We were surprised to find the relative order of baselines in terms of explainability performance remained the same in this revision.

A.12.1 Other Initial Ablation Experiments

We perform additional validation experiments to showcase that sparse autoencoder (dictionary learning) do outperform their none-sparse baselines in terms of model explainability. For reference, we have investigated the downstream effects of dictionary learning through a total of four ablation experiments (Figure 15). Our analysis centered on the ICD code most likely to be predicted for each note, determined by softmax probabilities, and employed two token ablation benchmarks: (A) complete ablation of all highlighted tokens and (C) random ablation of half the top highlighted tokens. In parallel, dictionary features underwent ablation in two forms: (B) solely ablation of the paramount dictionary feature for all highlighted tokens or (D) ablating half of the tokens and ablating the most significant dictionary feature of the remaining highlighted tokens. For reference, tokens surpassing the 95th percentile in attention scores for their respective ICD code in the label attention matrix were considered "highlighted". We note that we have only shown experimental results for ablation types (A) and (B) in section 4.1.

To investigate if such ablations are perfectly additive, we perform further experiments (C) and (D) where we only ablate half of the relevant tokens and observe the overall changes from ablating the dictionary features of the unablated other half of tokens in Table 11.

We observe that while ablating such dictionary

Downstream Ablation Experiments

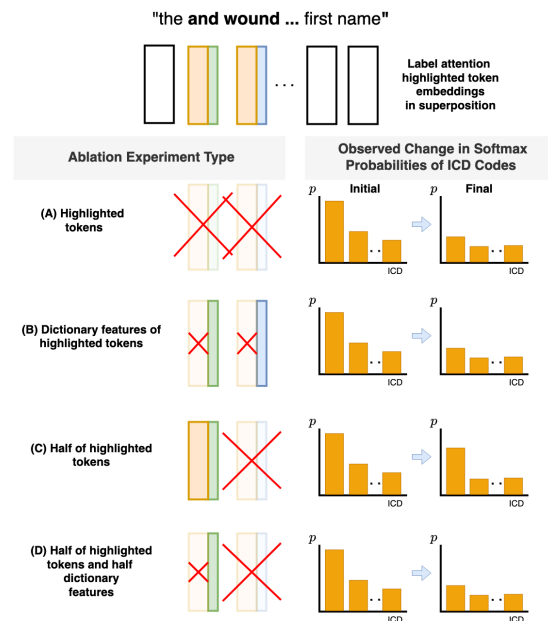


Figure 15: Ablation experiments on clinical notes using a label attention (LAAT) mechanism to highlight relevant tokens for ICD coding: (A) Complete ablation of all highlighted tokens, or ablation of only the most relevant dictionary features of each token embedding (B), and (C) random ablation of half the highlighted tokens compared to (D) random ablation of half the highlighted tokens and the most relevant dictionary features of other half of token embeddings.

features do indeed drop the softmax probabilities of the most likely ICD code, they do not sum to the original softmax drop observed in the ablation experiment in section 4.1, and thus there is some missing information that is not entirely encompassed by the dictionary feature ablations. That being said, DL still provides the best model explanation for downstream ICD predictions in this scenario.

A.12.2 Initial Human Understandability Evaluations

Coherence of top 500 activating dictionary features. Utilizing Siamese encoder embeddings, we calculate the average cosine similarity among the top k tokens of the top 500 dictionary feature to gauge their conceptual relatedness. For more methodological specifics of our Siamese BERT experiment, please refer to Appendix A.10. While not using the same sparse autoencoders, these results are still useful from at least a reproducibility standpoint.

Results. Overall, the coherence of the unsuper-

Ablating Dictionary Features of Highlighted Tokens													
Experiment	AutoCodeDL		LAAT + Baselines					DL		Baselines			
	L1	SPINE	ICA	PCA	Identity	Random	Token	L1	SPINE	ICA	PCA	Identity	Random
Top	0.685	0.686	6.169e-5	0.678	0.465	0.709	0.678	0.906	0.929	0.001	0.884	0.744	0.939
NGT	4.847	4.666	5.624	4.926	5.195	618.140	4.925	51.647	39.096	5.614	36.403	493.625	370.035
Ratio	0.141	0.147	1.100e-5	0.138	0.090	0.001	0.138	0.0176	0.0238	1.344e-4	0.024	0.002	0.003

Table 9: Softmax probability changes in downstream ICD predictions resulting from ablation experiments. 'Top' represents the magnitude of softmax drops for the most probable ground truth ICD code, while 'NGT' signifies the sum of absolute softmax probability changes of non-ground truth ICD codes for each clinical note. The 'Ratio' indicates the ratio between these two measures. We bold and distinguish the results obtained using our combined LAAT and dictionary learning framework, and observe that our method has the most precise effect on downstream ICD predictions, suggesting improved explanatory power.

Hidden Medical Meaning Identification Accuracy					
AutoCodeDL		Baselines			
L1	SPINE	ICA	PCA	Identity	Random
0.794	0.864	0.351	0.388	0.327	0.297

Table 10: Proportion of stop word embedding labels correctly identified by our AutoCodeDL framework using previous sparse autoencoders with using 99th percentile activated dictionary features, alongside the baseline methods. Such results showcase that DL is capable of effectively identifying hidden meanings embedded within superposition of stop words.

Experiment	(C) DL		(D) Baselines				
	L1	SPINE	ICA	PCA	Identity	Random	Tokens
Top	0.506	0.506	0.0519	0.205	0.068	0.743	0.0519
NGT	5.146	5.112	5.294	5.164	5.281	11.876	5.296
Ratio	0.098	0.099	0.0098	0.040	0.013	0.063	0.001

Table 11: Softmax probability changes in downstream ICD predictions resulting from ablation experiments illustrated in Figure 15. 'Top' represents the magnitude of softmax drops for the most probable ground truth ICD code, while 'NGT' signifies the sum of absolute softmax probability changes of non-ground truth ICD codes for each clinical note. The 'Ratio' indicates the ratio between these two measures. We highlight and distinguish the results obtained using sparse autoencoders.

vised dictionary learning methods, represented by the DL columns in the table, decreased as the top k tokens considered increased. Contrary to expectations outlined by (Subramanian et al., 2017), sparse autoencoders trained through L_1 minimization exhibited the highest coherence among all methods, as indicated by the highest average cosine similarity values across different values of k . This suggests that the highest activating dictionary features learned through L_1 minimization are more semantically consistent and conceptually coherent compared to other unsupervised methods.

Surprisingly, the coherence of dictionary fea-

tures learned from the top independent components in ICA surpassed that of SPINE, despite ICA's lack of impact on downstream ICD coding performance. This unexpected finding highlights the complex interplay between feature extraction methods and semantic coherence.

As expected, the supervised LAAT method achieved the highest coherence, reflecting its use of labeled data to guide the learning process that results in more tokens closely aligned with the semantics of ICD codes.

k	Coherence of Top 500 Activated Dictionary Features						
	DL		Baselines				Supervised
	L1	SPINE	ICA	PCA	Identity	Random	LAAT
2	0.344	0.283	0.312	0.244	0.269	0.271	0.692
4	0.331	0.278	0.292	0.234	0.242	0.251	0.678
10	0.303	0.260	0.280	0.228	0.234	0.238	0.637

Table 12: Average cosine similarity between the top k tokens extracted from each of the top 500 dictionary feature or ICD code, measured from Siamese encoder embeddings. Higher values indicate a stronger thematic connection within the feature or code. The "DL" columns represents our dictionaries constructed, while the remaining columns are baselines.

Dictionary ICD Overlap. The descriptions of a dictionary feature's most pertinent ICD codes should logically coincide with its most highly activating tokens. Each ICD code comes with a description enriched with medically relevant information, suggesting that a dictionary feature that accurately encapsulates a medical concept will exhibit an overlap between the descriptions of its relevant ICD codes and its activating tokens.

Setup. To assess this, we extract descriptions for each dictionary feature's relevant ICD codes, removing stop words, to compile a list of medically significant tokens. We calculate the overlap between these tokens and each medically relevant

feature’s top activating tokens, using the proportion of overlapping tokens as an additional measure of a dictionary feature’s understandability. We define a feature to be medically relevant when its ablation results in at least a 10% softmax probability drop of any ICD code.

Results. Comparing our dictionary features’ top tokens to those identified by specialized "label attention" (trained to map tokens to ICD codes), we find significant overlap for L1 and SPINE (Table 14). This overlap surpasses baselines, suggesting our unsupervised methods effectively learn medically relevant concepts. Notably, ICA features show minimal overlap. Overall, our dictionaries exhibit strong agreement with medically significant concepts, boosting our model’s interpretability.

Percentage of Dictionary Features GPT3.5 Differentiated						
DL		Baselines				Supervised
L1	SPINE	ICA	PCA	Identity	Random	LAAT
0.352	0.420	0.416	0.320	0.356	0.328	0.540

Table 13: Percentage of the 500 randomly sampled dictionary features successfully distinguished by GPT3.5 Turbo, determined by selecting the unrelated token from a set of four tokens.

Dictionary Overlap with ICD Descriptions						
DL		Baselines				Supervised
L1	SPINE	ICA*	PCA	Identity	Random	LAAT
0.117	0.086	0.000	0.010	0.024	0.003	0.198

Table 14: Proportion of highly activating tokens in dictionary features that overlap with ICD9 descriptions. LAAT represents an upper bound in overlap where we treat each ICD row in the label attention matrix as its own dictionary feature. *In general, the ablation of the independent components in ICA have little effect on downstream ICD predictions, hence its features have no overlap with ICD descriptions in this experiment.

A.13 Initial Examples of Dictionary Features

We manually inspect several highly interpretable dictionary features as a showcase of the potential of these interpretable representations. We show some examples in Figure 16 and 17, relating to depression, cesarian sections, and failure of wound healing.

Dictionary Feature about Cesarean Sections and Pregnancy

Feature ID

	string	token_id	context	feature_activation
9048	so	1091	fetal distress and so the baby was	0.491685
9056	stat	27999	2 ni by stat c section the	0.491327
9058	section	3812	by stat c section the initial ap	0.490882
9057	c	273	ni by stat c section the initial	0.490573
9054	ni	23795	was name2 ni by stat c	0.490307
9050	baby	16485	and so the baby was name2	0.489676
9000	sex	2068	date of birth sex f service neon	0.489209
9052	name	9614	the baby was name 2 ni by	0.488167
9062	g	74	the initial ap g ars were and	0.488026
9070	unit	4035	admitted to the unit of note on	0.486727
9063	ars	3729	initial ap g ars were and was	0.485995
9061	ap	519	section the initial ap g ars were	0.485971
9025	g	326	a year old g 1p 01	0.483746
9013	product	3695	is a kilogram product of a full	0.482721
9020	digits	27643	term pregnancy year digits to a year	0.482082
9027	p	83	old g 1p 01 mother the	0.481495
9074	prenatal	11691	of note on prenatal screens mother was	0.481474
9018	pregnancy	3707	a full term pregnancy year digits to	0.481424
8992		0	job number admission date discharge	0.481391

Dictionary Feature about Depression

Feature ID

	string	token_id	context	feature_activation
2420	depression	3518	s p hanging depression anoxic brain injury	0.439489
2302	depression	3518	past medical history depression social history non	0.438565
20064	depression	3518	n hypothyroid hypercalcemia depression syncope cardiogenic physical	0.419704
2303	social	2526	medical history depression social history noncont	0.409141
8747	depression	3518	history hypertension hypercholesterolemia depression history of stroke	0.392201
8474	depression	3518	medical history anxiety depression social history neg	0.368707
12538	psychiatric	7007	itative care including psychiatric rehabilitation and physical	0.362005
8475	social	2526	history anxiety depression social history neg tob	0.354052
5426	psychiatric	7007	be transferred to psychiatric inpatient unit first	0.353859
5401	psychiatric	7007	will require inpatient psychiatric treatment after her	0.353481
5390	psychiatry	22455	her multiple wounds psychiatry evaluated the patient	0.351970
5400	inpatient	11779	she will require inpatient psychiatric treatment after	0.351756
5427	inpatient	11779	transferred to psychiatric inpatient unit first name	0.347477
10689	psychiatry	22455	was seen by psychiatry on because of	0.340783

Figure 16: Examples of dictionary features. We showcase the most relevant tokens for dictionary feature about cesarian sections and depression as well as their clinical note contexts. We note that these sorted pandas dataframes are from an earlier sparse autoencoder in the previous revision of the paper. However, we felt that they were still worth showcasing.

Dictionary Feature Mapping of Cesarean Sections and Pregnancy to ICD Codes

	most_relevant_tokens	feature_activation	most_relevant_ICD	confidence
0	so	0.491685	V30.01	0.463092
1	stat	0.491327	763.0	0.217211
2	section	0.490882	V31.00	0.210847
3	c	0.490573	762.6	0.195970
4	ni	0.490307	36.01	0.193572
5	baby	0.489676	749.00	0.181664
6	sex	0.489209	766.1	0.168426
7	name	0.488167	770.5	0.164539
8	g	0.488026	764.97	0.157080
9	unit	0.486727	764.96	0.149489
10	ars	0.485995	767.19	0.144933

Dictionary Feature Mapping of Depression to ICD Codes

	most_relevant_tokens	feature_activation	most_relevant_ICD	confidence
0	depression	0.439489	311	0.736157
1	depression	0.438565	296.30	0.725788
2	depression	0.419704	296.20	0.704730
3	social	0.409141	296.50	0.553644
4	depression	0.392201	296.33	0.536124
5	depression	0.368707	300.9	0.516653
6	psychiatric	0.362005	V62.84	0.468462
7	social	0.354052	296.23	0.458238
8	psychiatric	0.353859	296.24	0.446205
9	psychiatric	0.353481	300.4	0.404587
10	psychiatry	0.351970	V17.0	0.387943

Dictionary Feature Mapping of Failure of Wound Healing

	most_relevant_tokens	feature_activation	most_relevant_ICD	confidence
0	dehiscence	0.276791	78.41	0.400468
1	and	0.276086	34.79	0.391046
2	first	0.275124	77.61	0.333510
3	wound	0.274717	998.31	0.328890
4	sex	0.274496	34.01	0.308916
5	breakdown	0.274263	34.03	0.293953
6		0.273906	34.1	0.276632
7	service	0.273818	519.2	0.265560
8	name	0.272039	998.32	0.255004
9	summary	0.271937	730.28	0.248923
10	of	0.270835	998.3	0.244022

Figure 17: Examples of dictionary features. We showcase the most relevant tokens and ICD codes for each respective interpretable dictionary feature. We note that these sorted pandas dataframes are from an earlier sparse autoencoder trained in the previous revision of the paper. However, we felt that they were still worth showcasing to prove a point. They contain various medical codes that are directly related to each token.