DILA: Dictionary Label Attention for Mechanistic Interpretability in High-dimensional Multi-label Medical Coding Prediction

John Wu JOHNWU3@ILLINOIS.EDU
University of Illinois Urbana-Champaign, USA

David Wu DAVID.H.WU@VANDERBILT.EDU
Vanderbilt University, USA

Jimeng Sun JIMENG@ILLINOIS.EDU
University of Illinois Urbana-Champaign, USA

Abstract

Automated medical coding, a clinical high-dimensional multilabel task, requires explicit interpretability. Existing works often rely on local interpretability methods, failing to provide comprehensive explanations of the overall mechanism behind each label prediction within a multilabel set. We propose a mechanistic interpretability module called DIctionary Label Attention (DILA) that disentangles uninterpretable dense embeddings into a sparse embedding space, where each nonzero element (a dictionary feature) represents a globally learned medical concept. Through human evaluations, we show that our sparse embeddings are more human understandable than its dense counterparts by at least 50 percent. Our automated dictionary feature identification pipeline, leveraging large language models (LLMs), uncovers thousands of learned medical concepts by examining and summarizing the highest activating tokens for each dictionary feature. We represent the relationships between dictionary features and medical codes through a sparse interpretable matrix, enhancing our global understanding of the model's predictions while maintaining competitive performance and scalability without extensive human annotation.

Keywords: Interpretability, Medical Coding

Data and Code Availability MIMIC-III (Johnson et al., 2016) dataset from PhysioNet (Goldberger et al., 2000 (June 13)), processed using split from (Edin et al., 2023). Code in supplemental material.

Institutional Review Board (IRB) No IRB approval needed for expert evaluations.

1. Introduction

Medical coding, which involves assigning potentially tens of thousands of International Classification of Diseases (ICD) codes to lengthy, unstructured clinical notes for categorizing diagnoses and procedures (Hirsch et al., 2016), is a crucial task in the clinical pipeline. This complex and time-consuming task requires explicit medical expertise, making human annotation expensive (O'Malley et al., 2005). While recent advancements in pre-trained language models have significantly improved coding efficiency and accuracy by treating it as a high-dimensional multilabel classification problem (Edin et al., 2023; Huang et al., 2022; Yan et al., 2022), the opaque nature of these black-box models raises concerns about their decision-making processes and potential biases (Hakkoum et al., 2022; Räuker et al., 2023). To uphold transparency and maintain patient trust, these models must be capable of explaining their code predictions, which are crucial for billing, research, and clinical treatment purposes (Rao et al., 2022). This is especially the case where misclassifications can directly impact patient outcomes, underscoring the need for interpretable and transparent AI models (Johnson et al., 2021).

Existing interpretability solutions are inadequate for medical coding. Model-agnostic methods like SHAP (Lundberg and Lee, 2017; Chen, 2021) are computationally infeasible for long clinical notes with large multilabel prediction spaces (Lundberg et al., 2020; Chen et al., 2022). Post-hoc mechanistic interpretability methods, such as label attention mechanisms (Chaudhari et al., 2021; Vu et al., 2020), often fail to provide comprehensive explanations and are generally limited to local attribution (Serrano and Smith, 2019; Zhang et al., 2021). Intrinsically explainable models, like prototype models (Tang et al., 2023a; Ma et al.,

2023), require costly exemplar corpora, which are impractical due to the vast number of ICD codes and privacy concerns (Edin et al., 2023; Hirsch et al., 2016). White-box approaches (Yu et al., 2023) can address neuron polysemanticity (Elhage et al., 2022), but pretraining new models is expensive. Ultimately, such interpretability approaches are either too expensive or not comprehensive enough for clinical applications.

However, studies have shown that increasing sparsity in decision-related layers can enhance interpretability where neurons only activate for specific data features (Wong et al., 2021; Thompson et al., 2024). Inspired by these results, we propose an interpretable DIctionary Label Attention (DILA) module incorporating sparsity via dictionary learning (Olshausen and Field, 1997; Bricken et al., 2023). To enhance the interpretability of the learned sparse dictionary features without needing expert annotations, we leverage medical large language models (LLMs) to automatically interpret our learned sparse abstractions. Using our approach DILA, we show:

- Interpretability: The dictionary label attention layer in DILA is more human-interpretable than its dense counterparts. Learning sparse medical abstractions enables insights into the model's global decision-making process for medical coding.
- Scalability: While imperfect, we demonstrate that medical LLMs can serve as capable domain-specific dictionary feature summarizers and annotators. This enables the scalability of deep auto-interpretability pipelines, overcoming the limitations of prototype design in the medical domain, where expert knowledge is scarce and expensive.
- Performance: Despite the incorporation of sparsity and interpretability, DILA maintains competitive performance with current state-of-the-art black-box baselines on the cleaned MIMIC-III dataset (Edin et al., 2023).

2. Related Work

2.1. Dictionary Learning

Much of our work was directly inspired by the use of dictionary learning to better understand the predictions made by auto-regressive LLMs (Bricken et al., 2023; Cunningham et al., 2023; Yun et al., 2023) in a post-hoc manner. However, we note that the sparse coding problem (Zhang et al., 2015; Olshausen and Field, 1997) is not unique to the language domain and has been applied to improve interpretability in various other modalities such as vision (Yu et al., 2023; Ghosh et al., 2023; Liu et al., 2023) and time-series (Xu et al., 2023; Tang et al., 2023b) tasks. To our knowledge, such an approach has not been directly leveraged to explain deep extreme multi-label prediction settings better, as seen in the automated ICD coding task where the input space consists of thousands of tokens with an equally large prediction space.

2.2. Interpretability in Automated ICD Coding

There exists many attempts in making interpretable automated ICD coding methods (Yan et al., 2022), each with their own trade-offs. Phrase matching (Cao et al., 2020) and 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. Alternative approaches using large language models (LLMs) for evidence generation (Yang et al., 2023) face challenges with hallucination (Zhao et al., 2023; Huang et al., 2023), potentially compromising explanation faithfulness.

Furthermore, the predominant interpretability method for deep neural models in ICD coding tasks is the label attention (LAAT) mechanism (Yan et al., 2022). LAAT projects token embeddings into a label-specific attention space, scoring each token's relevance to ICD predictions. This approach has been incorporated into various architectures, including convolutional (Mullenbach et al., 2018), recurrent (Vu et al., 2020), and pre-trained language models (Huang et al., 2022). While LAAT efficiently highlights locally relevant tokens for each ICD prediction, its nonlinear projections and dense pretrained language model (PLM) embeddings hinder direct interpretation of global prediction mechanisms due to polysemanticity (Olah et al., 2020). These limitations underscore the ongoing trade-off between interpretability and performance in ICD coding tasks.

Our method DILA, in comparison, directly interprets and disentangles concepts learned within the PLM embedding space. It maps global medically relevant concepts in the form of dictionary features to each ICD code while retaining the local interpretability derived from the label attention mechanism.

3. DILA

Overview. DILA consists of three key components: (1) dictionary learning to disentangle dense embeddings into sparse, interpretable dictionary features; (2) a dictionary label attention module that utilizes a sparse, interpretable matrix capturing the global relationships between dictionary features and medical codes to generate each clinical note's label attention matrix, representing the local token-code relationships; and (3) an automated interpretability approach using medical LLMs, as shown in Figure 1.

3.1. Dictionary Learning

Dictionary learning decomposes vectors $\mathbf{x}$ that often represent tokens or words into a sparse linear combination of basis vectors. We formulate this problem using a sparse autoencoder with dense PLM embeddings $\mathbf{x} \in \mathbb{R}^d$, sparse dictionary feature activations $\mathbf{f} \in \mathbb{R}^m$, as well as encoder and decoder weight matrices $\mathbf{W_e} \in \mathbb{R}^{d \times m}$ and $\mathbf{W_d} \in \mathbb{R}^{m \times d}$ with corresponding bias units $\mathbf{b_e} \in \mathbb{R}^m, \mathbf{b_d} \in \mathbb{R}^d$:

$$\bar{\mathbf{x}} = \mathbf{x} - \mathbf{b_d} \tag{1}$$

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

$$\hat{\mathbf{x}} = \mathbf{W_d}\mathbf{f} + \mathbf{b_d} \tag{3}$$

$$\mathcal{L}_{\text{saenc}} = \frac{1}{|X|} \sum_{\mathbf{x} \in X} \|\mathbf{x} - \hat{\mathbf{x}}\|_2^2 + \lambda_{L_1} \|\mathbf{f}\|_1 + \lambda_{L_2} \|\mathbf{f}\|_2^2 \tag{4}$$

The $\lambda_{L_1} \|\mathbf{f}\|_1$ and $\lambda_{L_2} \|\mathbf{f}\|_2^2$ elastic loss terms Zou and Hastie (2005) in Equation (4) enforce sparsity on the encoded dictionary feature activations $\mathbf{f}$, ensuring that only specific $f_i \in \mathbf{f}$ activate (i.e., are nonzero) for specific token embeddings $\mathbf{x}$. This sparsity influences each element f_i to be more interpretable, often corresponding to a specific medical concept. To ground these interpretations within the embedding space, we reconstruct $\mathbf{x}$ using our dictionary embeddings $\mathbf{W_d} = \begin{bmatrix}\mathbf{h_0}, & \mathbf{h_1}, & \dots & ,\mathbf{h_m}\end{bmatrix}^T$, allowing any $\mathbf{x}$ to be represented through a sparse linear combination of $\mathbf{h_i}$ as defined by our sparse autoencoder: $\mathbf{x} \approx \sum_i^m f_i \mathbf{h_i}$.

To construct a human interpretable dictionary, we sample token embeddings from a text corpus and sort for the highest activating tokens for f_i, called a dictionary feature's context, as they describe the underlying meaning of each f_i (see Appendix A.2). For instance, in step 3 of Figure 1, the contexts of a dictionary feature f_i all share the same theme of falls and subdural hematomas, a type of head injury.

3.2. Dictionary Label Attention

We propose a simpler, disentangled version of label attention called dictionary label attention that maps the concepts represented by each of the m dictionary features f_i to their corresponding ICD codes (Figure 1). Given a clinical note's tokenized PLM embeddings $\mathbf{X}_{\text{note}} \in \mathbb{R}^{s \times d}$ of length s, we encode them into disentangled dictionary features $\mathbf{F}_{\text{note}} \in \mathbb{R}^{s \times m}$ using a sparse autoencoder. We initialize the sparse projection matrix $\mathbf{A}_{\mathbf{f}_{\text{icd}}} \in \mathbb{R}^{m \times c}$, which maps the relationship between each of the m dictionary features f_i and c ICD codes, by encoding the tokens in each ICD code's description into their respective dictionary features $\mathbf{F}_{\text{desc}}^{(c)} \in \mathbb{R}^{l \times m}$, where l is the description length. We then average pool these features into $\mathbf{f^{(c)}} \in \mathbb{R}^m$ and perform the following operations:

$$\mathbf{A}_{\mathbf{f}_{\text{icd}}} = \begin{bmatrix}\mathbf{f}^{(1)}, & \mathbf{f}^{(2)}, & \dots & ,\mathbf{f}^{(c)}\end{bmatrix} \in \mathbb{R}^{m \times c} \tag{5}$$

$$\mathbf{A}_{\text{laat}} = \text{softmax}(\mathbf{F}_{\text{note}}\mathbf{A}_{\mathbf{f}_{\text{icd}}}) \in \mathbb{R}^{s \times c} \tag{6}$$

$$\mathbf{X}_{\text{att}} = \mathbf{A}_{\text{laat}}{}^T \mathbf{X}_{\text{note}} \in \mathbb{R}^{c \times d} \tag{7}$$

Once the label aware representation $\mathbf{X}_{\text{att}}$ is computed, it is passed through a decision layer for the final prediction $\hat{\mathbf{y}} \in \mathbb{R}^c$.

Relationship to Dense Label Attention. In contrast to the original label attention mechanism in Vu et al. (2020), we have essentially replaced the original nonlinear projection of $\mathbf{A}_{\text{laat}} = \text{softmax}(\mathbf{ZW_c})$ where $\mathbf{Z} = \tanh(\mathbf{X}_{\text{note}}\mathbf{W_z})$ with a single linear sparse projection matrix $\mathbf{A}_{\mathbf{f}_{\text{icd}}}$, representing each of the c ICD codes as a set of dictionary features. Unlike $\mathbf{Z}$, since $\mathbf{f}$ is always positive, every element's magnitude in $\mathbf{A}_{\mathbf{f}_{\text{icd}}}$ indicates the strength of the overall relationship between a dictionary feature f_i and an ICD code $\hat{\mathbf{y}}_\mathbf{i}$.

Training. Training consists of two steps. First, we train a sparse autoencoder on all of the embeddings generated by our PLM within the training set. Then, we initialize our label attention module, and do end-to-end training using a combination of our sparse autoencoder loss defined in equation 4 and the binary cross entropy loss function. We use an additional hyperparameter λ_{saenc} to prevent L_{saenc} from dominating L_{BCE}, giving us our final loss function in equation 8.

$$L = \lambda_{\text{saenc}} L_{\text{saenc}} + L_{\text{BCE}} \tag{8}$$

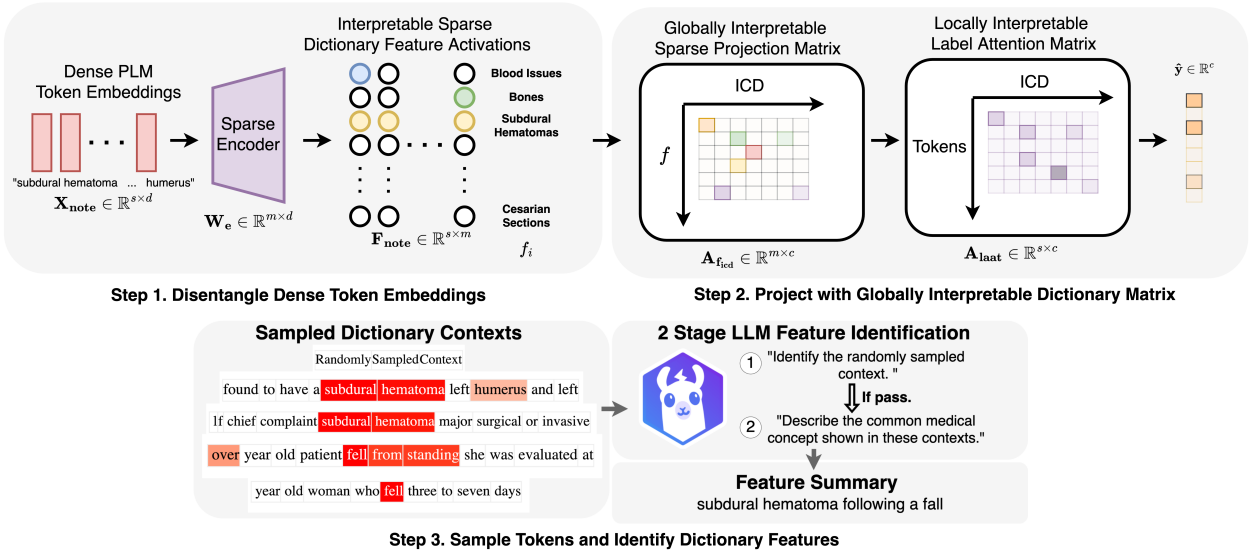

Figure 1: DILA composes of three steps: First, we disentangle each token embedding into its dictionary features. Then, we project each set of dictionary features with our globally interpretable $\mathbf{A}_{\mathbf{f}_{\text{icd}}}$ to generate our local explanation $\mathbf{A}_{\text{laat}}$ for downstream multilabel prediction. Finally, medical LLMs identify the learned dictionary feature to understand the learned relationships in $\mathbf{A}_{\mathbf{f}_{\text{icd}}}$.

3.3. Auto-Interpretability

One major challenge in our framework is that the sampled dictionary contexts and its sparse dictionary feature representations $\mathbf{f}$ are unlabeled. Ideally, human expert annotators Bricken et al. (2023) would inspect the dictionary features' highest activating contexts. However, unlike general language disciplines Zhang et al. (2023), expert medical professional annotators are not as readily available and often lack the time to provide high-quality annotations for each learned dictionary feature. An automated pipeline to identify and quantify the quality of such dictionaries is crucial for scalable interpretability, as medical PLMs learn thousands of domain-specific concepts. Inspired by Subramanian et al. (2017)'s feature identification and Bills et al. (2023)'s LLM neuron interpretability experiments, we develop a two-stage auto-interpretability pipeline shown in Figure 1.

More specifically, to discern whether or not the LLM can identify the dictionary feature faithfully, we perform an identification test that asks the LLM to identify the randomly sampled context that does not activate the dictionary feature representing a specific medical concept out of five total contexts. If the LLM is capable of discerning the outlier context, it implies it can understand that the other contexts all belong to the same underlying medical concept. In practice, we prompt it with only the highlighted tokens (red) to avoid information contamination from the context window as there can often be neighboring medical concepts that may misdirect the LLM. Once determined to be identifiable, we prompt the LLM to summarize the dictionary feature given the original contexts.

4. Experiments

Models. We mainly explore two models, our method DILA, and its nearest dense label attention comparison PLM-ICD (Huang et al., 2022), both of which use the same text medical RoBERTa PLM. While we reuse the weights provided by (Edin et al., 2023) in our interpretability comparison, we also retrain their model using the same training hyperparameters used in DILA to get a more direct comparison in performance in Section 4.3 as we were unable to replicate the reported large batch size of 16 (Edin et al., 2023) due to GPU limitations. We report the training hyperparameters used in the Appendix (A.1).

Annotators. We ask two medical experts for our human evaluations, a clinically licensed physician and a medical scientist trainee with extensive clinical training. For the LLM, we use a state-of-the-art quantized, medically fine-tuned Llama 3 OpenBioLLM 70b model in our auto-interpretability pipeline.

Overview. We explore the interpretability, mechanistic insights, and performance of our proposed method, DILA, for automated ICD coding. Using human evaluations, we assess the human understandability of its learned dictionary features and the efficacy of the auto-interpretability pipeline, and demonstrate its ability to efficiently provide precise global explanations through ablation studies, visualizations, and human predictability experiments. Finally, we evaluate DILA's performance compared to baselines.

4.1. Automated Interpretability of Dictionary Features

Setup. To evaluate our dictionary feature identification method in Figure 1 and quantify the interpretability or human understandability of our trained dictionaries, we conduct a medical expert identification experiment across 100 randomly sampled dictionary features, as described in Section 3.3. An example of the LLM identification prompt is showcased in the Appendix (A.3.1).

As hallucination is always plausible, human medical experts are asked to evaluate the summaries generated by the LLMs of the identified dictionary features. We ask whether they agree with the originally sampled contexts and how confident they are in their responses, from 1, being unsure, to 4, absolute confidence.

Baselines. Furthermore, we compare our dictionary features to dense $\mathbf{Z}$ activations from the label attention mechanism (Section 3.2) in PLM-ICD (Huang et al., 2022) and a random baseline where contexts are randomly sampled from a large medical token corpus from the test set. We run our automated pipeline across all observed features: 6,088 active features in our dictionary $\mathbf{f}$, 768 features in our dense $\mathbf{Z}$, and 1,000 randomly sampled contexts. However, since we felt random context summaries were trivial, we only ask our annotators to evaluate the summaries generated from interpreting the dictionary features $\mathbf{f}$ and its dense label attention counterpart $\mathbf{Z}$.

Identification of Dictionary Features. Table 1 demonstrates improved interpretability by leveraging a sparse embedding $\mathbf{f}$ over the dense embedding $\mathbf{Z}$, with significantly more interpretable $\mathbf{f}$ features (3,524) compared to $\mathbf{Z}$ features (263), suggesting over 3,000 interpretable abstract concepts were hidden in superposition (Elhage et al., 2022). Comparing LLM and human responses using a vector similarity metric reveals that human expert annotations identify more interpretable features and that human alignment of medical LLM feature identifications deteriorates as interpretability declines, suggesting a substantial gap between domain-specific annotators and medical LLMs. Further qualitative examinations reveal that LLMs fail to identify features with intrinsic relationships between contexts not obvious by language, such as linking "banding" and blood loss in Appendix Figure 7. However, given the volume of identified interpretable features, LLMs dramatically reduce the number of human annotations needed for obviously interpretable features.

Quality of LLM Dictionary Feature Summarizations. From Table 2, we see that the majority of the summaries generated by the LLMs are in agreement with our human evaluators. Our qualitative evaluations show that the contexts of summaries rejected by our human annotators for the dictionary features are substantially more coherent than the ones rejected by the ones in the dense neurons $\mathbf{Z}$ as shown in Appendix Section A.4.2. Crucially, many of the dictionary feature summaries that were rejected, were rejected due to their lack of specificity rather than being unrelated to the dictionary contexts or hallucinations, allowing us to better conduct mechanistic interpretability experiments.

4.2. Mechanistic Interpretability

Runtime Performance. One crucial utility of mechanistic interpretability is the ability to generate efficient interpretations of a model's predictions, whether locally or globally. In Table 3, we show that using an off-the-shelf KernelSHAP (Lundberg and Lee, 2017) interpreter to analyze the dense PLM embeddings of a single clinical note is extremely computationally expensive, when compared to its mechanistic counterparts, highlighting SHAP's impractical use in high dimensional multilabel prediction.

Setup. To evaluate the explainability of our $\mathbf{A}_{\mathbf{f}_{\text{icd}}}$ matrix, we compare to the common local attribution approach of token-based attention ($\mathbf{A}_{\text{laat}}$) through an ablation study. For each clinical note, we identify the highest softmax probability ICD code and ablate the weights corresponding to the observed activated dictionary features f_i, and measure the softmax probability drop for the target code and the sum of absolute changes for other codes.

Baselines. To compare against alternative local interpretability approaches, we identify the most relevant tokens for the highest-probability ICD code using $\mathbf{A}_{\text{laat}}$ and perturb the PLM embeddings by ablating, noising, or replacing these tokens with medically irrelevant ones, measuring the same downstream effects.

	Medical Expert 1 (100) ↑	Medical Expert 2 (100) ↑	LLM (100) ↑	LLM (all) ↑	Avg. Cosine Similarity ↑	Avg. Jaccard Similarity ↑	No. of LLM Identified Features ↑
Dict. **f** (DILA)	**0.67 (+55.8%)**	**0.69 (+64.3%)**	**0.59 (+73.5%)**	**0.58 (+70.6%)**	**0.77 (+30.5%)**	**0.62 (+44.2%)**	**3,524 (+1239.9%)**
Dense **Z**	0.43	0.42	0.34	0.34	0.59	0.43	263
Random	0.23	0.19	0.27	0.19	0.15	0.08	193

Table 1: Identification test accuracy comparing human experts and LLMs on interpreting dictionary features (**f**), dense embeddings (**Z**), and random contexts. The results show the superior performance of human experts and the importance of sparse embeddings for interpretability. The relative improvement of Dict. **f** (DILA) over Dense **Z** is shown in parentheses.

	Expert 1, Agreement ↑	Expert 2, Agreement↑	Expert 1, Confidence ↑	Expert 2, Confidence↑
Dict. **f** (DILA)	**0.83 (+5.1%)**	**0.92 (+12%)**	$\mathbf{3.85 \pm 0.41}$	$\mathbf{3.80 \pm 0.45}$
Dense **Z**	0.79	0.82	3.79 ± 0.41	3.44 ± 0.75

Table 2: Human evaluations of LLM summaries of dictionary features. We report the standard deviations of our confidence scores. The relative percentage improvement in LLM summary agreement of Dict. **f** (DILA) over Dense **Z** are shown in parentheses.

	Model-agnostic	Mechanistic (DILA)		
	KernelSHAP	$\mathbf{A}_{\text{laat}}$	$\mathbf{F}_{\text{note}}$	$\mathbf{A}_{\mathbf{f}_{\text{icd}}}$
Time ↓	62m 10.71s	0.04s	0.03s	0.00s

Table 3: Runtime comparison for interpreting a single clinical note using KernelSHAP (Lundberg and Lee, 2017) (model-agnostic) vs. our DILA method. $\mathbf{A}_{\mathbf{f}_{\text{icd}}}$ access is near-instantaneous, demonstrating the efficiency of mechanistic interpretability.

Global vs. Local Explanations. Local explanation methods can identify relevant tokens for a model's predictions but fail to isolate the mechanisms behind individual ICD code predictions in extreme multilabel settings, where medically-specific tokens often relate to multiple ICD codes. Our sparse weight matrix $\mathbf{A}_{\mathbf{f}_{\text{icd}}}$ overcomes this limitation by directly mapping encoded dictionary features **f** to each ICD code prediction. Through weight ablation, we can pinpoint the specific mechanisms driving each code's prediction. Table 4 demonstrates that ablating relevant weights in our global $\mathbf{A}_{\mathbf{f}_{\text{icd}}}$ matrix does not affect the prediction of other ICD codes, unlike token ablation, which impacts multiple codes due to tokens' relationships with various codes.

Visualization. Another crucial utility is the ability to visualize the overall medical concepts that the model has learned to associate with each of the several thousand ICD codes. For instance, we can summarize that the model has learned that hypoglycemia, obesity, pancreatic abnormalities, and diabetes mellitus is highly predictive of diabetes-related ICD codes in Figure 2. Furthermore, we visualize the model's understanding of different medical conditions in space by plotting the UMAP of $\mathbf{A}_{\mathbf{f}_{\text{icd}}}$ as shown in Figure 3.

Setup. To assess the interpretability of the global projection matrix $\mathbf{A}_{\mathbf{f}_{\text{icd}}}$, we conduct a predictability experiment. Our medical experts are asked to choose the set of codes that best match a given dictionary feature's LLM summary and its sampled contexts from two sets of codes. They also are asked of their confidence in their choices. Clear and easily understandable dictionary features should represent distinct medical concepts with obvious associated medical codes. The experiment is conducted for 100 randomly sampled dictionary features and repeated using our medical LLM to interpret a larger portion of the matrix.

Baseline. For a direct comparison, we perform the same experiment using the dense $\mathbf{W}_{\mathbf{c}}$ matrix with the identified features of **Z** from Section 4.1. However, as a quick caveat, due to the larger number of identified abstractions of **f** compared to **Z**, $\mathbf{A}_{\mathbf{f}_{\text{icd}}}$ is inherently more informative. Random guessing in this experiment would yield a 50% accuracy.

Human Evaluations. From Table 5, we observe that $\mathbf{A}_{\mathbf{f}_{\text{icd}}}$ is more interpretable than $\mathbf{W}_{\mathbf{c}}$ due to a larger portion of codes being matched when using features from **f** compared to **Z**, but with some interesting caveats that make human predictability challenging. Shown in Appendix Figure 15, many of the incorrect matchings by humans were due to incorrectly learned associations between the identified abstract dictionary features and its corresponding top 5 medical codes despite the **f** generally being interpretable themselves,

	$\mathbf{A}_{\mathbf{f}_{\text{icd}}}$ Weight Ablation	$\mathbf{A}_{\text{laat}}$ Token Ablation	$\mathbf{A}_{\text{laat}}$ Token Noising	$\mathbf{A}_{\text{laat}}$ Token Replacement
Top ICD ↑	0.954 ± 0.1	0.953 ± 0.1	0.204 ± 0.4	0.948 ± 0.2
$\lvert\Delta\rvert$ Other ICD ↓	$\mathbf{0 \pm 0}$	21.7 ± 123.3	372.8 ± 281.4	9.9 ± 5.0

Table 4: Ablating the class-specific weights in $\mathbf{A}_{\mathbf{f}_{\text{icd}}}$ does not affect other classes compared to relevant token perturbations, indicating its disentangled explanation of downstream code predictions.

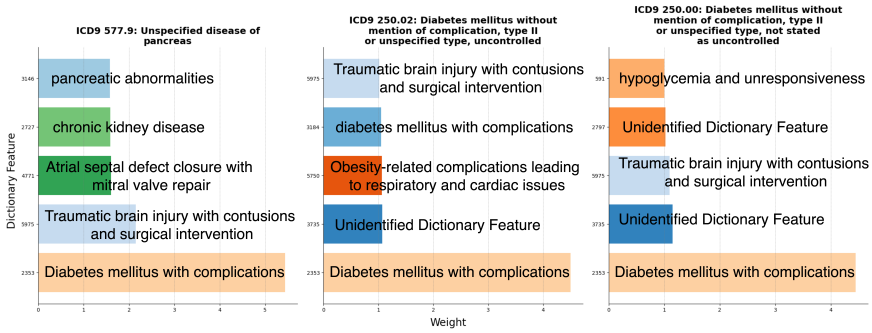

Figure 2: Top 5 Dictionary Features for Diabetes-related ICD Codes.

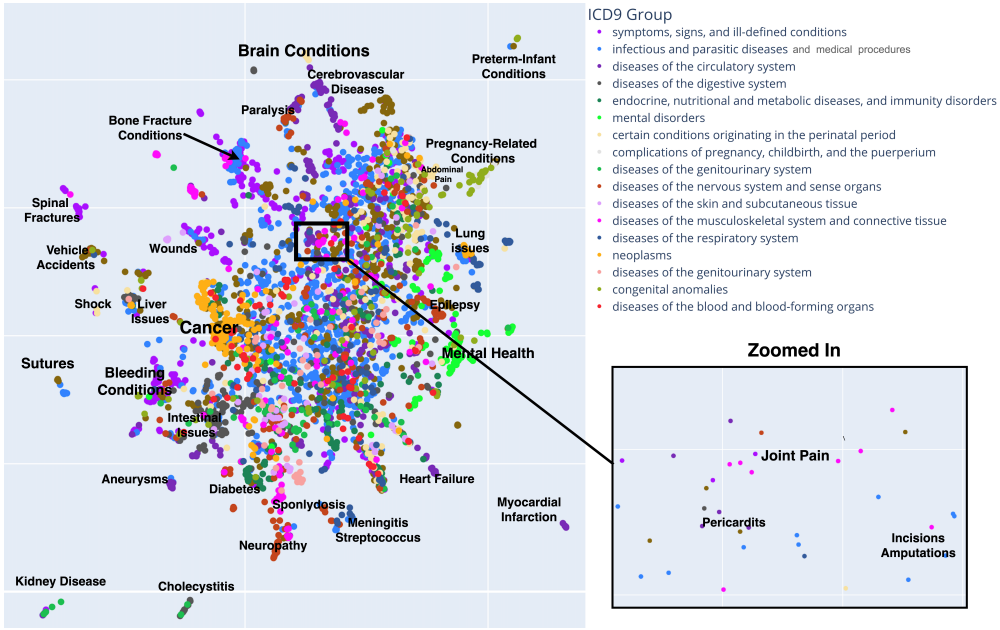

Figure 3: UMAP of $\mathbf{A}_{\mathbf{f}_{\text{icd}}}$ with respect to all medical codes. We observe clusters of medical codes with relative distances that are intuitive. For instance, neuropathy is a common condition associated with diabetes, and vehicle accidents are more closely linked to bone and spinal fractures.

suggesting that $\mathbf{A}_{\mathbf{f}_{\text{icd}}}$ may prove to be useful in identifying incorrectly learned feature predictions.

4.3. Performance

Baselines. The trade-off between interpretability and performance due to reduced expressive power is a critical concern when integrating sparse layers and more interpretable modules (Mansour et al., 2022; Rudin, 2019). However, our findings suggest that this trade-off may not persist. Table 6 compares the performance of automated ICD coding methods such as convolutional and recurrent neural networks, and PLMs for MIMIC-III clean (Edin et al., 2023), including our own reproduction of PLM-ICD (*).

	Medical Expert 1 Matching Accuracy (100) ↑	Medical Expert 2 Matching Accuracy (100) ↑	Medical Expert 1 Confidence (100) ↑	Medical Expert 2 Confidence (100) ↑	LLM Matching Accuracy (100) ↑	LLM Matching Accuracy (all) ↑
$\mathbf{A}_{\mathbf{f}_{\text{icd}}}$	**0.73 (+12.3%)**	**0.69 (+15.0%)**	$\mathbf{3.41 \pm 0.6}$	$\mathbf{2.87 \pm 1.0}$	**0.65 (+6.6%)**	**0.59 (+9.3%)**
$\mathbf{W_c}$	0.65	0.60	3.20 ± 0.57	2.62 ± 0.89	0.61	0.54

Table 5: Human Predictability Experiment with $\mathbf{A}_{\mathbf{f}_{\text{icd}}}$. Our medical experts more confidently and accurately match the corresponding set of ICD codes given a dictionary feature. The relative improvement of $\mathbf{A}_{\mathbf{f}_{\text{icd}}}$ over $\mathbf{W_c}$ is shown in parentheses.

	CNN	Bi-GRU	CAML	MultiResCNN	LAAT	PLM-ICD	PLM-ICD *	DILA * (ours)
Micro F1 ↑	48.0 ± 0.3	49.7 ± 0.4	55.4 ± 0.1	56.4 ± 0.2	57.8 ± 0.2	59.6 ± 0.2	54.6 ± 0.1	54.9 ± 0.2
Macro F1 ↑	9.9 ± 0.4	12.2 ± 0.2	20.4 ± 0.3	22.9 ± 0.6	22.6 ± 0.6	26.6 ± 0.8	26.5 ± 0.3	27.2 ± 0.4
Micro AUC-ROC ↑	97.1 ± 0.0	97.8 ± 0.1	98.2 ± 0.0	98.5 ± 0.0	98.6 ± 0.1	98.9 ± 0.0	97.7 ± 0.0	97.6 ± 0.0
Macro AUC-ROC ↑	88.1 ± 0.2	91.1 ± 0.2	91.4 ± 0.2	93.1 ± 0.3	94.0 ± 0.3	95.9 ± 0.1	92.5 ± 0.0	91.7 ± 0.0

Table 6: Performance comparison of automated ICD coding methods on MIMIC-III clean dataset (Edin et al., 2023). * indicates our training. The average scores are reported along with their standard deviations.

Performance. Our model, DILA, achieves the highest average Macro F1 score and slightly lower Micro F1 score compared to the relevant baselines, indicating better performance on rarer codes but worse performance on edge-cases of common ICD codes. Notably, our reproduction of the previous state-of-the-art baseline, PLM-ICD, yields lower performance than reported, possibly due to memory restrictions limiting batch sizes. As batch size directly affects the optimal performance of ICD coding models (Edin et al., 2023), DILA may still attain better performance with larger batch sizes.

5. Discussion and Conclusion

	Pre-Edit	Post Edit
False Positives	164	158
False Negatives	66	68

Table 7: Results of causal edits of $\mathbf{A}_{\mathbf{f}_{\text{icd}}}$ for ICD 99.20 "Injection of Platelet Inhibitors".

Debugging. Global interpretability allows for quick identification of incorrectly learned mappings between medical concepts and codes by inspecting the sparse mappings in $A_{f_{icd}}$. We have observed numerous improperly learned associations, as visualized in Figures 16, 18, and 17 in the Appendix. For example, the code for "Athlete's Foot" is incorrectly associated with Ovarian cancer in Figure 18, potentially leading to false positives. In a case study, we attempted to remedy the commonly false positive code "Injection of Platelet Inhibitors" by visualizing its top 20 most common abstract dictionary features in Figure 19, revealing incorrect associations like dictionary feature 4443 "Trisomy Disorders" (see Section A.6.4). Ablating the relevant weights in $A_{f_{icd}}$ decreased false positives related to the incorrect dictionary features but slightly increased false negatives (Table 7), suggesting the need to change other weights to better predict true cases, and highlighting the complexities of debugging. An interpretable automated procedure for debugging incorrectly classified ICD codes is an important direction, as over 40% of ICD codes are never predicted correctly Edin et al. (2023).

Improving LLM Annotations. Our current automatic interpretability pipeline with LLMs only uses zero-shot prompting. However, numerous works have improved LLM-assisted annotations Goel et al. (2023) and generation faithfulness such as retrieval augmented generation Lewis et al. (2021). Exploring these ideas could bridge the gap between current state-of-the-art domain-specific LLMs and medical experts for dictionary feature annotation tasks.

Unidentifiable Dictionary Features. Some highly relevant dictionary features may not be identifiable by humans or LLMs, as shown in the Appendix (Figure 17). We investigated a few and showcase their sampled contexts in the Appendix (Table 10). Unidentified features often result from a lack of highly activated contexts within our text corpus or a lack of an explicit coherent medical theme, suggesting the need for larger sampled corpuses in our dictionary construction and potential limitations in our dictionary learning formulation. Other sparse formulations should be explored for optimal interpretable design Rajamanoharan et al. (2024).

Ultimately, our proposed dictionary label attention (DILA) module takes a step towards addressing the need for interpretability in high-dimensional multi-label prediction tasks, particularly in medical coding. By disentangling dense embeddings into a sparse space and leveraging LLMs for automated dictionary feature identification, DILA aims to uncover globally learned medical concepts, provide comprehensive explanations, and facilitate the development of debuggable models. While further research is needed to validate its effectiveness, DILA represents a promising direction in developing more interpretable and transparent models for complex, high-stakes applications, contributing to developing trustworthy AI systems in healthcare and beyond.

References

Vijay Arya, Rachel K. E. Bellamy, Pin-Yu Chen, Amit Dhurandhar, Michael Hind, Samuel C. Hoffman, Stephanie Houde, Q. Vera Liao, Ronny Luss, Aleksandra Mojsilović, Sami Mourad, Pablo Pedemonte, Ramya Raghavendra, John Richards, Prasanna Sattigeri, Karthikeyan Shanmugam, Moninder Singh, Kush R. Varshney, Dennis Wei, and Yunfeng Zhang. One explanation does not fit all: A toolkit and taxonomy of ai explainability techniques, 2019.

Steven Bills, Nick Cammarata, Dan Mossing, Henk Tillman, Leo Gao, Gabriel Goh, Ilya Sutskever, Jan Leike, Jeff Wu, and William Saunders. Language models can explain neurons in language models. https://openaipublic.blob.core.windows.net/neuron-explainer/paper/index.html, 2023.

Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, 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. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*, 2023. 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. Clinical-coder: Assigning interpretable ICD-10 codes to Chinese clinical notes. In Asli Celikyilmaz and Tsung-Hsien Wen, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 294–301, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-demos.33. URL https://aclanthology.org/2020.acl-demos.33.

Chun Sik Chan, Huanqi Kong, and Liang Guanqing. A comparative study of faithfulness metrics for model interpretability methods. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5029–5038, Dublin, Ireland, May

2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.345. URL https://aclanthology.org/2022.acl-long.345.

Sneha Chaudhari, Varun Mithal, Gungor Polatkan, and Rohan Ramanath. An attentive survey of attention models, 2021.

Hugh Chen, Scott M. Lundberg, and Su-In Lee. Explaining a series of models by propagating shapley values. *Nature Communications*, 13(1): 4512, Aug 2022. ISSN 2041-1723. doi: 10.1038/s41467-022-31384-3. URL https://doi.org/10.1038/s41467-022-31384-3.

Shikun Chen. Interpretation of multi-label classification models using shapley values, 2021.

Hoagy Cunningham, Aidan Ewart, Logan Riggs, Robert Huben, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models, 2023.

Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C. Wallace. Eraser: A benchmark to evaluate rationalized nlp models, 2020.

Andres Duque, Hermenegildo Fabregat, Lourdes Araujo, and Juan Martinez-Romo. A keyphrase-based approach for interpretable icd-10 code classification of spanish medical reports. *Artificial Intelligence in Medicine*, 121:102177, 2021. ISSN 0933-3657. doi: https://doi.org/10.1016/j.artmed.2021.102177. URL https://www.sciencedirect.com/science/article/pii/S0933365721001706.

Joakim Edin, Alexander Junge, Jakob D. Havtorn, Lasse Borgholt, Maria Maistro, Tuukka Ruotsalo, and Lars Maaløe. 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, 2023. Association for Computing Machinery. ISBN 9781450394086. doi: 10.1145/3539618.3591918. URL https://doi.org/10.1145/3539618.3591918.

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. Toy models of superposition. *Transformer Circuits Thread*, 2022. URL 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. Dictionary learning under symmetries via group representations, 2023.

Akshay Goel, Almog Gueta, Omry Gilon, Chang Liu, Sofia Erell, Lan Huong Nguyen, Xiaohong Hao, Bolous Jaber, Shashir Reddy, Rupesh Kartha, Jean Steiner, Itay Laish, and Amir Feder. Llms accelerate annotation for medical information extraction, 2023.

A. L. Goldberger, L. A. N. Amaral, L. Glass, J. M. Hausdorff, P. Ch. Ivanov, R. G. Mark, J. E. Mietus, G. B. Moody, C.-K. Peng, and H. E. Stanley. PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. *Circulation*, 101(23):e215–e220, 2000 (June 13). Circulation Electronic Pages: http://circ.ahajournals.org/content/101/23/e215.full PMID:1085218; doi: 10.1161/01.CIR.101.23.e215.

Hajar Hakkoum, Ibtissam Abnane, and Ali Idri. Interpretability in the medical field: A systematic mapping and review study. *Applied Soft Computing*, 117:108391, 2022. ISSN 1568-4946. doi: https://doi.org/10.1016/j.asoc.2021.108391. URL https://www.sciencedirect.com/science/article/pii/S1568494621011522.

J A Hirsch, G Nicola, G McGinty, R W Liu, R M Barr, M D Chittle, and L Manchikanti. ICD-10: History and context. *AJNR Am J Neuroradiol*, 37 (4):596–599, January 2016.

Chao-Wei Huang, Shang-Chi Tsai, and Yun-Nung Chen. Plm-icd: Automatic icd coding with pretrained language models, 2022.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions, 2023.

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. Mimic-iii, a freely accessible critical care database. *Scientific Data*, 3(1):160035, May 2016. ISSN 2052-4463. doi: 10.1038/sdata.2016.35. URL https://doi.org/10.1038/sdata.2016.35.

Renee L Johnson, Holly Hedegaard, Emilia S Pasalic, and Pedro D Martinez. Use of ICD-10-CM coded hospitalisation and emergency department data for injury surveillance. *Inj Prev*, 27(S1):i1–i2, March 2021.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive nlp tasks, 2021.

Yingjie Liu, Xuan Liu, Hui Yu, XUAN TANG, and Xian Wei. Learning dictionary for visual attention. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 56589–56601. Curran Associates, Inc., 2023.

Scott Lundberg and Su-In Lee. A unified approach to interpreting model predictions, 2017.

Scott M. Lundberg, Gabriel Erion, Hugh Chen, Alex DeGrave, Jordan M. Prutkin, Bala Nair, Ronit Katz, Jonathan Himmelfarb, Nisha Bansal, and Su-In Lee. From local explanations to global understanding with explainable ai for trees. *Nature Machine Intelligence*, 2(1):56–67, Jan 2020. ISSN 2522-5839. doi: 10.1038/s42256-019-0138-9. URL https://doi.org/10.1038/s42256-019-0138-9.

Chiyu Ma, Brandon Zhao, Chaofan Chen, and Cynthia Rudin. This looks like those: Illuminating prototypical concepts using multiple visualizations, 2023.

Yishay Mansour, Michal Moshkovitz, and Cynthia Rudin. There is no accuracy-interpretability tradeoff in reinforcement learning for mazes, 2022.

James Mullenbach, Sarah Wiegreffe, Jon Duke, Jimeng Sun, and Jacob Eisenstein. Explainable prediction of medical codes from clinical text, 2018.

Chris Olah, Nick Cammarata, Ludwig Schubert, Gabriel Goh, Michael Petrov, and Shan Carter. Zoom in: An introduction to circuits. *Distill*, 2020. doi: 10.23915/distill.00024.001. https://distill.pub/2020/circuits/zoom-in.

Bruno A. Olshausen and David J. Field. Sparse coding with an overcomplete basis set: A strategy employed by v1? *Vision Research*, 37(23):3311–3325, 1997. ISSN 0042-6989. doi: https://doi.org/10.1016/S0042-6989(97)00169-7. URL https://www.sciencedirect.com/science/article/pii/S0042698997001697.

Kimberly J O'Malley, Karon F Cook, Matt D Price, Kimberly Raiford Wildes, John F Hurdle, and Carol M Ashton. Measuring diagnoses: ICD code accuracy. *Health Serv Res*, 40(5 Pt 2):1620–1639, October 2005.

Senthooran Rajamanoharan, Arthur Conmy, Lewis Smith, Tom Lieberum, Vikrant Varma, János Kramár, Rohin Shah, and Neel Nanda. Improving dictionary learning with gated sparse autoencoders, 2024.

Preethi Rao, Shira H Fischer, Mary E Vaiana, and Erin Audrey Taylor. Barriers to price and quality transparency in health care markets. *Rand Health Q*, 9(3):1, June 2022.

Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215, May 2019. ISSN 2522-5839. doi: 10.1038/s42256-019-0048-x. URL https://doi.org/10.1038/s42256-019-0048-x.

Tilman Räuker, Anson Ho, Stephen Casper, and Dylan Hadfield-Menell. Toward transparent ai: A survey on interpreting the inner structures of deep neural networks, 2023.

Sofia Serrano and Noah A. Smith. Is attention interpretable?, 2019.

Anant Subramanian, Danish Pruthi, Harsh Jhamtani, Taylor Berg-Kirkpatrick, and Eduard Hovy. Spine: Sparse interpretable neural embeddings, 2017.

Dennis Tang, Frank Willard, Ronan Tegerdine, Luke Triplett, Jon Donnelly, Luke Moffett, Lesia Semenova, Alina Jade Barnett, Jin Jing, Cynthia Rudin, and Brandon Westover. Protoeegnet: An interpretable approach for detecting interictal epileptiform discharges, 2023a.

Yuanbo Tang, Zhiyuan Peng, and Yang Li. Explainable trajectory representation through dictionary learning, 2023b.

Ryan Thompson, Amir Dezfouli, and Robert Kohn. The contextual lasso: Sparse linear models via deep neural networks, 2024.

Thanh Vu, Dat Quoc Nguyen, and Anthony Nguyen. 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, July 2020. doi: 10.24963/ijcai.2020/461. URL http://dx.doi.org/10.24963/ijcai.2020/461.

Eric Wong, Shibani Santurkar, and Aleksander Madry. Leveraging sparse linear layers for debuggable deep networks, 2021.

Ruiyu Xu, Chao Wang, Yongxiang Li, and Jianguo Wu. Generalized time warping invariant dictionary learning for time series classification and clustering, 2023.

Chenwei Yan, Xiangling Fu, Xien Liu, Yuanqiu Zhang, Yue Gao, Ji Wu, and Qiang Li. A survey of automated international classification of diseases coding: development, challenges, and applications. *Intelligent Medicine*, 2(3):161–173, 2022. ISSN 2667-1026. doi: https://doi.org/10.1016/j.imed.2022.03.003. URL https://www.sciencedirect.com/science/article/pii/S2667102622000092.

Zhichao Yang, Sanjit Singh Batra, Joel Stremmel, and Eran Halperin. Surpassing gpt-4 medical coding with a two-stage approach, 2023.

Yaodong Yu, Sam Buchanan, Druv Pai, Tianzhe Chu, Ziyang Wu, Shengbang Tong, Hao Bai, Yuexiang Zhai, Benjamin D. Haeffele, and Yi Ma. White-box transformers via sparse rate reduction: Compression is all there is?, 2023.

Zeyu Yun, Yubei Chen, Bruno A Olshausen, and Yann LeCun. Transformer visualization via dictionary learning: contextualized embedding as a linear superposition of transformer factors, 2023.

Lining Zhang, Simon Mille, Yufang Hou, Daniel Deutsch, Elizabeth Clark, Yixin Liu, Saad Mahamood, Sebastian Gehrmann, Miruna Clinciu, Khyathi Chandu, and João Sedoc. Needle in a haystack: An analysis of high-agreement workers on mturk for summarization, 2023.

Yu Zhang, Peter Tino, Ales Leonardis, and Ke Tang. A survey on neural network interpretability. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 5(5):726–742, October 2021. ISSN 2471-285X. doi: 10.1109/tetci.2021.3100641. URL http://dx.doi.org/10.1109/TETCI.2021.3100641.

Zheng Zhang, Yong Xu, Jian Yang, Xuelong Li, and David Zhang. A survey of sparse representation: Algorithms and applications. *IEEE Access*, 3:490–530, 2015. ISSN 2169-3536. doi: 10.1109/access.2015.2430359. URL http://dx.doi.org/10.1109/ACCESS.2015.2430359.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. A survey of large language models, 2023.

Hui Zou and Trevor Hastie. Regularization and Variable Selection Via the Elastic Net. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 67(2):301–320, 03 2005. ISSN 1369-7412. doi: 10.1111/j.1467-9868.2005.00503.x. URL https://doi.org/10.1111/j.1467-9868.2005.00503.x.

Appendix A. Appendix

A.1. Training Details

For a fair comparison and to reproduce the PLM-ICD model from (Edin et al., 2023), we followed the hyperparameters swept in their work as closely as possible. Specifically, both models employed the same pre-trained medical RoBERTa encoder architecture. One slight difference in training is that we leverage the PLM already pre-trained on the ICD coding task whereas PLM-ICD is trained from a pre-trained medical RoBERTa model on other medical data. Additionally, we utilized their code to perform the same data splits as reported in their MIMIC-III reproduction, obtained from their repository (Edin et al., 2023). Due to physionet's policies, we are unable to share their training data directly. However, due to GPU memory constraints, we were unable to use the same batch

size, which may have hindered our ability to fully reproduce their PLM-ICD (Huang et al., 2022) performance. Table 8 showcases the hyperparameters used for our methods. It's worth noting that an initial parameter sweep revealed that the performance benefits of PLM label attention models largely stemmed from their batch size. Consequently, we reran our training four times across four randomly selected seeds. We use the optimizer AdamW with its default settings. We also use 1e-6 for our λ_{saenc} hyperparameter. Our binary cross entropy loss function is defined as $L_{\text{BCE}} = -\frac{1}{C}\sum_{c=1}^{C}[y_c \log(\hat{y}_c) + (1-y_c)\log(1-\hat{y}_c)]$. Finally, we also set $m = 6,144$ to stay within GPU memory constraints. In principle, having a higher m would result in more granular and specific dictionary features f learned.

Compute Resources. We leverage a compute cluster using A6000 48GB GPUs and note that our maximum batch size was of size 8 due to some clinical notes containing over 4,000 tokens. Training takes approximately a day, but we regret not measuring explicit training times. We recommend having at least 48GB of CPU memory, as depending on the analysis, we store a lot in memory.

Sparsity and Performance. Our additional experiments in Table 9 reveal a trade-off between performance and sparsity.

A.2. Dictionary Contexts Construction

The construction of a human-interpretable dictionary, containing relevant tokens for each f_i, involves two steps. First, we sample a large number of tokens and decompose their embeddings using the sparse autoencoder defined in Section 3.1. For each i in $\mathbf{f}$, we retrieve the sparsely activated nonzero f_i. Next, we sort all tokens by their f_i magnitudes for each f_i to obtain their contexts. To conserve memory, we save the top 10 connected tokens but only use the top 4 connected tokens (i.e., multiple activated tokens within a chunk) for our evaluations. It's worth noting that some dictionary features have very few activating contexts. Algorithm 1 summarizes this procedure. For our dictionary construction, we sample over 8,000 clinical notes from the test set.

We show a resulting output of a pandas dataframe in Figure 4.

Algorithm 1: Build Clinical Token Dictionary

Input : Clinical corpora w/ N tokens: $X = \{x_1, x_2, ..., x_N\}$, where each $x_n \in \mathbb{R}^d$
No. of dictionary features: m
No. of top tokens for each feature: k

Output : Dictionary D of top k tokens for m features

Initialize empty dictionary D

foreach *token* x_n *in clinical corpora* ($n = 1$ *to* N) **do**
  Decompose x_n into sparse combination of m dictionary features: $x_n \approx \sum_{i=1}^{m} f_{i,n} h_i$ where $f_{i,n}$ is scalar magnitude for i-th feature of n-th token
end

foreach *dictionary feature* i ($i = 1$ *to* m) **do**
  Initialize empty list L_i
  foreach *token* x_n ($n = 1$ *to* N) **do**
    Create pair $(x_n, f_{i,n})$
    Add pair to L_i
  end
  Sort L_i in descending order based on $f_{i,n}$ magnitudes
  Select top k pairs from sorted L_i
  Extract only tokens x from k pairs
  Add these k tokens to D under key i
end

return D

A.3. Dictionary Feature Identification

Extended Rationale for Dictionary Feature Identification Approach with LLMs: Initially, we planned to have the LLM simply answer whether or not there was an explicit medical theme within each set of contexts with a yes or no response. However, we observed that, regardless of specificity, the LLMs would always respond affirmatively, stating that the contexts were medically relevant. Consequently, we pursued a different approach originally leveraged in human evaluation experiments by (Subramanian et al., 2017). We noticed that this approach substantially improved the determination of whether a set of dictionary contexts was interpretable, as the LLM was implicitly forced to evaluate whether a common concept existed among the dictionary contexts. To illustrate the dictionary feature identification process, we showcase a couple of LLM prompts and their respective context tokens in the figures below.

Other Minor Identification Experiment Details: We ran our dictionary feature identification process across 6,088 dictionary features. Initially, we eliminated any dictionary feature with fewer than 4 contexts, deeming them unidentifiable due to the lack of contexts. Subsequently, we conducted our simple identification experiment. It's worth noting that, in principle, larger text corpuses could potentially

Table 8: Hyperparameters used in training DILA and PLM-ICD models

	Learning rate	λ_{l1}	λ_{l2}	Batch size	LR scheduler	Epochs	Dropout	Decision Boundary Threshold
DILA *	5e-5	0.0001	0.00001	8	Linear Warmup	20	0.2	0.3
PLM-ICD*	5e-5	N/A	N/A	8	Linear Warmup	20	0.2	0.3
PLM-ICD (Edin et al., 2023)	5e-5	N/A	N/A	16	Linear Warmup	20	0.2	0.3

	string	token_id	context	feature_activation
3122	subdural	37076	to have a| subdural| hematoma left humerus	10.415280
3090	subdural	37076	f chief complaint| subdural| hematoma major surgical	10.374207
3123	hematoma	18801	have a subdural| hematoma| left humerus and	10.311387
3091	hematoma	18801	chief complaint subdural| hematoma| major surgical or	10.185183
3107	fell	11358	year old patient| fell| from standing she	10.151344
3109	standing	12114	patient fell from| standing| she was evaluated	9.239037
3108	from	414	old patient fell| from| standing she was	9.177430
3279	hematoma	18801	discharge diagnosis subdural| hematoma| dementia hip fracture	8.774582
3278	subdural	37076	expired discharge diagnosis| subdural| hematoma dementia hip	8.694788
3128	hip	7238	humerus and left| hip| fracture past medical	7.448319
3233	trauma	5846	admitted to the| trauma| icu head	7.352646
1051	fall	5238	room after a| fall| she was transferred	7.227351
1309	fall	5238	complaint s p| fall| major surgical or	7.029170
1331	fall	5238	who sustained a| fall| at home she	6.987866
1345	hematoma	18801	on chronic subdural| hematoma| with extensive midline	6.875710
1067	hemorrhage	8672	have a subarachnoid| hemorrhage| on ct scan	6.861817
1066	subarachnoid	24685	to have a| subarachnoid| hemorrhage on ct	6.797218
1050	a	261	emergency room after| a| fall she was	6.794368
1344	subdural	37076	acute on chronic| subdural| hematoma with extensive	6.728899
1340	large	1501	to have a| large| acute on chronic	6.643813

Figure 4: Example of token sorting to acquire the necessary token contexts for f_i for dictionary feature 1,871, relating to subdural hemotomas. The far left column indicates the position of the token in the text corpus.

L1	Macro F1	Micro F1
0.1	24.8	55.1
0.01	25.4	54.6
0.001	26.6	55.0
0.0001 (original)	27.2 ± 0.4	54.9 ± 0.2
0.00001	26.8	55.0

Table 9: Increasing the L1 sparsity coefficient reduces model performance. However, we note that the performance drop is not as extreme as once thought.

be beneficial for identifying more dictionary features. However, the majority of dictionary features (5,847 out of 6,088) contained at least 4 contexts.

A.3.1. LLM Identification Prompt

We show the LLM identification prompt in Figures 5 and 6.

A.3.2. Human Identification Examples

We show the human identification experiments in Figures 7 and 8.

A.4. LLM Summarization

We showcase the LLM prompt used, and the human evaluation results below. We showcase the summaries rejected by our medical experts in Figure 11.

A.4.1. LLM Prompt

We showcase the LLM prompt used to summarize dictionary features in Figure 10.

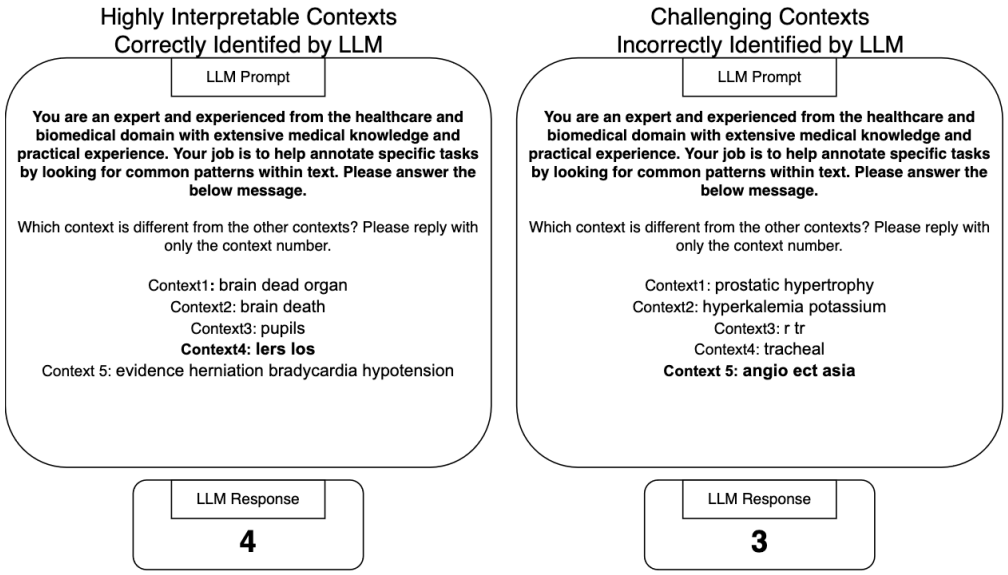

Figure 5: Dictionary Feature f_i Identification Tasks for the LLM. The correct random context is indicated in bold. LLMs struggle to identify the common medical theme in challenging contexts that lack explicit connections (right). Clinical notes often contain abbreviations and shorthand without clear references, making interpretation difficult even for experienced physicians. For example, the abbreviation "rtr" was challenging for our clinical physician to recall (right).

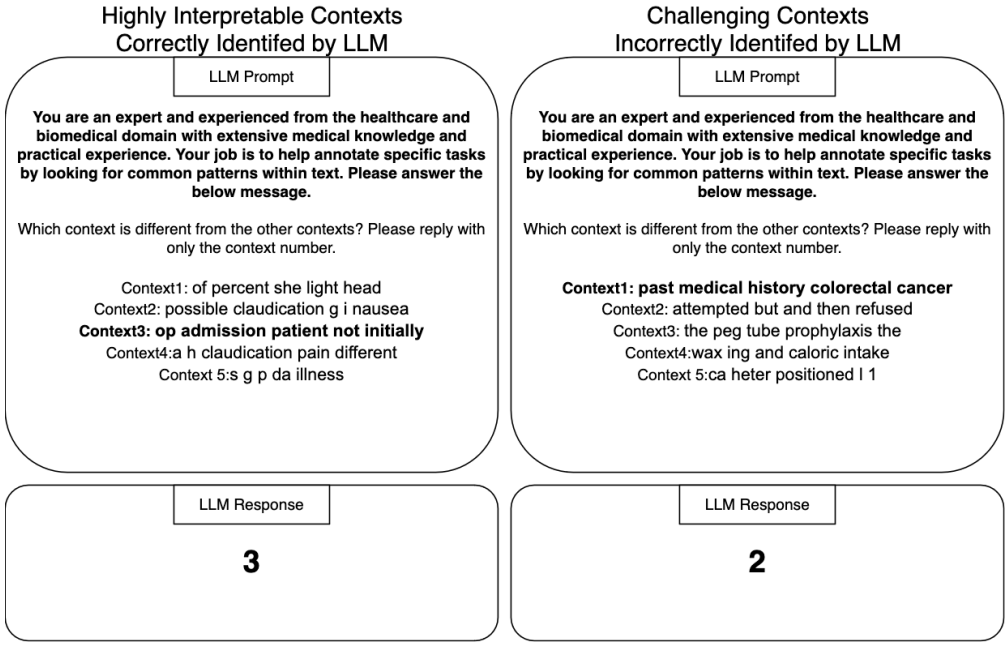

Figure 6: Dense $\mathbf{Z_i}$ Identification Tasks for the Language Model (LLM). The correct random context is indicated in bold. Some dense layer neurons activate for contexts with a common theme, such as claudication and nausea (left). However, many neurons activate for seemingly unrelated contexts (right).

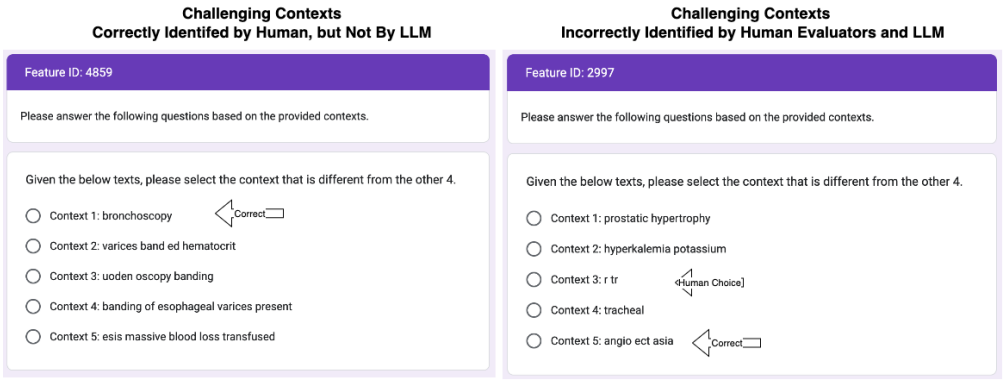

Figure 7: Dictionary Feature f_i Identification Tasks for Humans. The left panel shows a case where the human correctly identified a dictionary feature, but the Language Model (LLM) failed to do so. The LLM incorrectly selected "massive blood loss..." as the random context, despite its relationship to banding. The right panel presents dictionary contexts where both the LLM and humans failed to identify the randomly sampled context, which is understandable given the seemingly unrelated nature of the contexts.

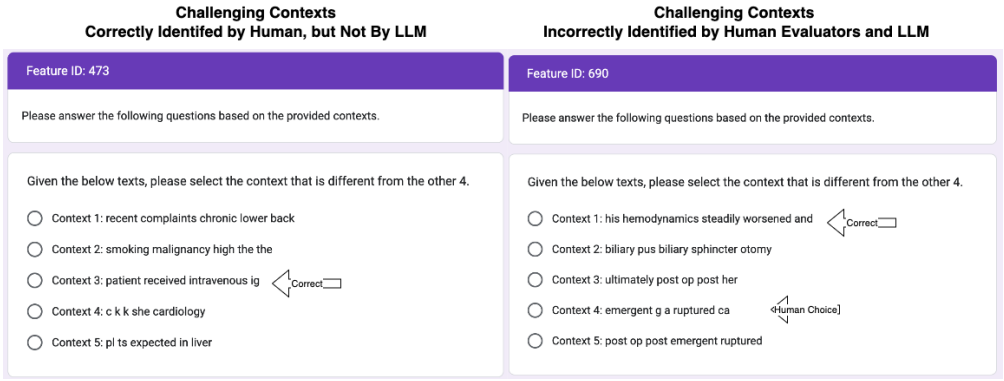

Figure 8: Dense $\mathbf{Z_i}$ Identification Tasks for Humans. The left panel shows an example where humans identified a neuron with hidden underlying relationships, such as the connection between smoking and cardiology, where patients often experience various levels of pain due to complications. Language Models (LLMs) struggle to identify such relationships. The right panel demonstrates a neuron with a diverse set of contexts, making it challenging for both humans and LLMs to identify a common theme.

A.4.2. Medical Expert Rejections of LLM Summary

We showcase only the LLM-generated summaries rejected by our medical expert evaluators, as they are more informative than the accepted summaries. Although the dictionary contexts are highly interpretable and consistent with a specific medical theme, the experts disagreed with certain aspects of the specific LLM summaries. For instance, the term "knee amputation" was deemed insufficiently specific, as "above knee amputation" is the more precise medical condition, differing from "below knee amputations." Other disagreements were related to the LLM's inferences of abbreviations, such as "ig ris," where there was no evidence of "insulin glargine injections" within the context. In such scenarios, dictionary contexts may not have a conclusive summary, requiring further investigation into the clinical notes as well as potentially the need to include some amount of context. Note that, as part of the LLM auto-interpretability evaluation study done in 1, 59 dictionary feature summaries and 34 Dense Z summaries were examined, as shown in Table 2.

On the other hand, from the dense Z contexts that have seemingly passed the LLM identification

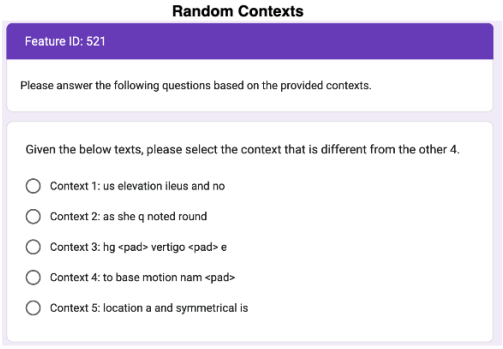

Figure 9: Example of Purely Random Contexts. The image showcases a prompt with truly random tokens, including pad tokens that were not filtered out. A stark reduction in pad tokens activating dense $\mathbf{Z}$ neurons is observed, and pad tokens are essentially never present in the dictionary contexts, suggesting their irrelevance to model predictions.

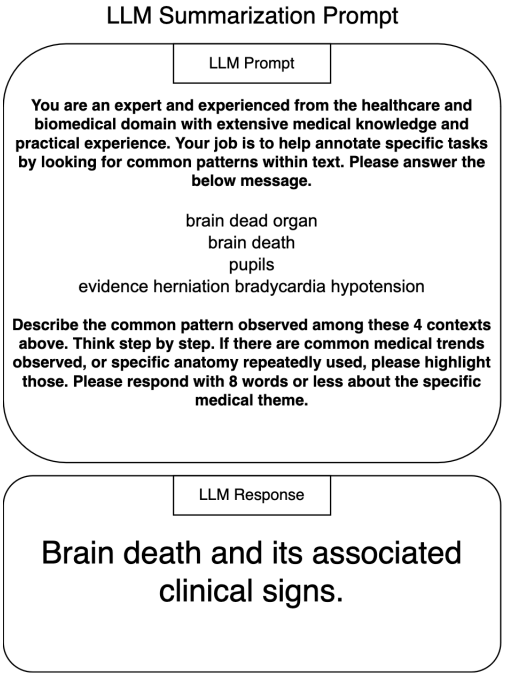

Figure 10: LLM Summarization Prompt for Dictionary Feature Contexts. The image shows the prompt used to summarize the highly connected dictionary contexts identified in the pipeline. The model is limited to 8 words to facilitate processing and extracting the summaries, as allowing an unlimited word count resulted in the inclusion of filler words in the LLM-generated summaries.

test, their summaries were rejected due to the lack of coherence within each context, implying that many of the identified Z features may have been the result of chance by the LLM pipeline.

A.5. Sparse Projection Matrix

To better understand the structure of our sparse matrix, we visualize the first 100 dictionary features and 100 medical codes in our learned $\mathbf{A}_{\mathbf{f}_{\text{icd}}}$ and dense W_c projection matrix in PLM ICD. We note that our $\mathbf{A}_{\mathbf{f}_{\text{icd}}}$ is substantially sparser.

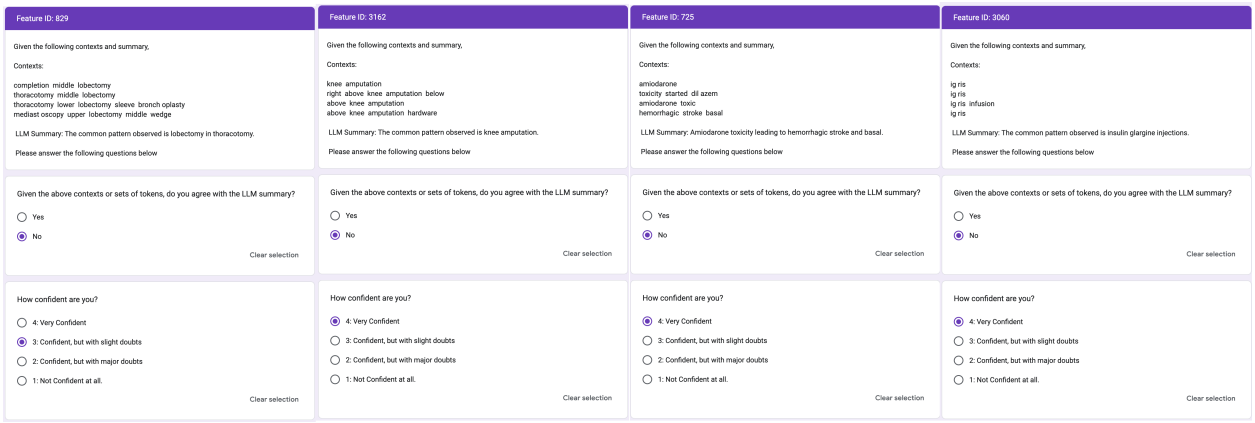

Figure 11: Medical Expert Rejected LLM Summaries of Dictionary Features $\mathbf{f}$. The summaries rejected by the medical experts were generally due to a lack of specificity, such as "knee amputation" not being "above knee amputation" (middle left). Two other summaries (left and middle right) were rejected because the LLM summary was too specific, with unlikely associations such as Amiodarone and strokes or the assumption that thoracotomy was present in all contexts. The far right rejection demonstrates a case of direct hallucination, where the LLM assumed the contexts were related to insulin despite no information hinting at that relationship, even though the contexts shared the same acronym.

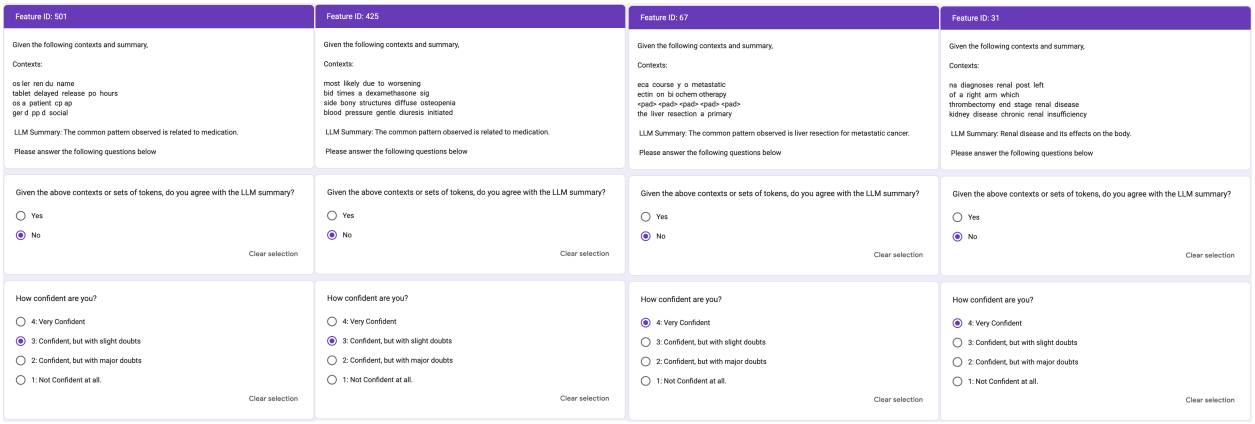

Figure 12: Medical Expert Rejected LLM Summaries of Contexts from Dense Z. Many of the dense Z context summaries were rejected due to their over-generality. For instance, the LLM would frequently generate summaries like "the common pattern observed is related to medication," which is true to some extent as medical text is generally related to medication. However, these LLM summaries (far left, middle left) were deemed uninformative by the medical experts. Other rejections (middle right) were due to the lack of coherence within dense Z contexts, such as the presence of ¡pad¿ tokens and liver-related contexts. Surprisingly, the LLM summary still managed to capture some relevant parts of the context, such as metastatic cancer and liver resection. Finally, the far right example shows a surprisingly consistent medical theme, with renal disease present in three of the four contexts. However, the annotators felt that the summary "effects on the body" was too broad and not specific enough.

We leverage this high sparsity to construct our human interpretability and ablation experiments below, as each abstract dictionary feature is observed to be related to specific medical codes.

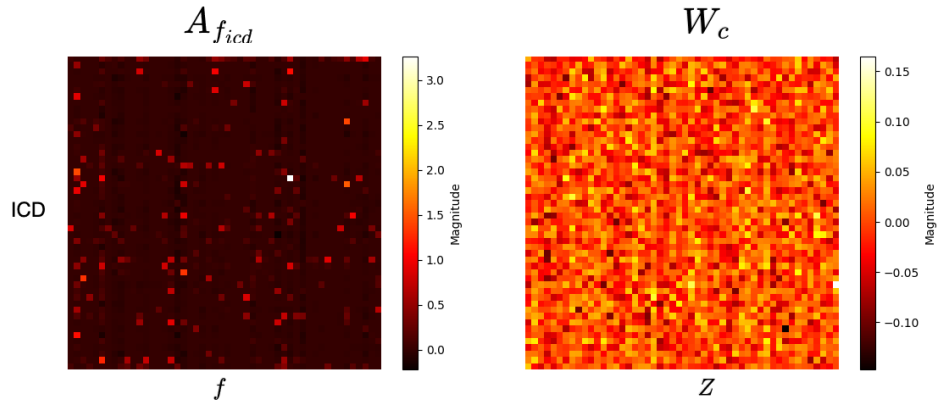

Figure 13: Sparsity comparison between ICD projection matrices $\mathbf{W_c}$ and $\mathbf{A_{f_{icd}}}$ for first 100 dictionary features and medical codes. Visually speaking, it is easy to identify the strong relationships between dictionary features and ICD codes whereas in the dense projection matrix, its weights look almost uninterpretable.

A.5.1. Global vs. Local Interpretability Ablation Experiments

Definition of Highly Relevant Tokens: Since most tokens have very low label attention attribution scores for each class, we define the most relevant tokens as those with attention attribution scores greater than the 95th quantile for a specific class in the label attention matrix.

Weight Ablation Details: Each clinical note can be decomposed into a sparse set of dictionary features. Consequently, we can identify a set of weights corresponding to each nonzero dictionary feature and medical code to be ablated for each clinical note. Since most irrelevant weights are already close to zero, and the relevant weights are positive, ablating them (i.e., setting them to zero) should provide a reasonably close approximation to the optimal explanation. In practice, ablating all the weights for a specific class in our $\mathbf{A_{f_{icd}}}$ matrix does not affect any other class, indicating the potential for precise explanations of a single medical code or class.

Recognition of Faithfulness Metrics in Conventional Multiclass Classification: We acknowledge the existence of various other interpretability faithfulness metrics for local attribution methods (Chan et al., 2022), such as comprehensiveness (DeYoung et al., 2020) and monotonicity (Arya et al., 2019). However, due to the large number of tokens (some clinical notes have upwards of almost 6,000 tokens), performing a quantiling removal or token addition scheme is extremely computationally expensive, especially if one were to consider doing so for each code in every multilabel example. As such, we simplified our downstream effects or faithfulness experiment to simply measuring the drop in performance of the most likely ICD code for each clinical note and the absolute change in softmax probabilities of other ICD code predictions.

A.6. Predictability Experiments

In the human predictaiblity experiment, we ask both LLMs and our medical experts to select the best corresponding medical codes given a set of dictionary feature contexts and its LLM summary. We showcase examples of our predictability experiments below.

A.6.1. LLM Prompt

The LLM prompt used for predictability experiments is in Figure 14.

A.6.2. Human Evaluation

Due to being more informative than the positive correct cases (as those dictionary features and codes are more obviously selected for), we showcase the results where the person was unable to select the set of medical codes that the model had highlighted for its downstream predictions. From qualitative inspection with our medical experts, we notice that the medical expert was unable to predict the correct set of medical codes were the direct result of both sets of medical codes being related to the dictionary feature,

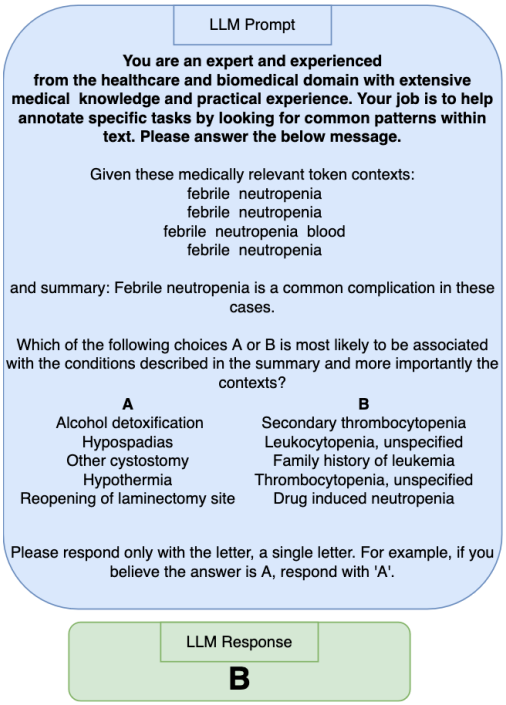

Figure 14: LLM Prompt for Predictability Experiment with $\mathbf{f}$. The LLM is presented with two sets of medical code descriptions related to the contexts and summary it generated, labeled as "A" and "B". The LLM is then prompted to choose the set of medical codes that best describes or relates to the conditions shown in the given contexts and summary.

or the top ICD codes associated with each dictionary feature in our $\mathbf{A}_{\mathbf{f}_{\text{icd}}}$ were incorrectly mapped. For instance, the suture of lacerations of different organs are not directly related to gallstones and double tail stents that prevent biliary obstructions. On the other hand, smoking complications such as Tobacco use are often more likely to cause stomach ulcers than malignant lymphomas. While both can have that as a potential complication, in practice, one is more common over the other, showcasing the wrong weighing of our learned $\mathbf{A}_{\mathbf{f}_{\text{icd}}}$ matrix between different medical conditions and medical codes. Other predictability experiments failed due to the lack of extra information on abbreviations such as "LEVT", which may be a challenge to understand as many of these unstructured clinical notes do not have a specific identification process for these abbreviations. That said, we note that the predictability of these codes is better than random, suggesting that the global dictionary matrix can be leveraged to better understand what our model understands for each set of medical code predictions.

A.6.3. Misaligned Associations Between Learned Medical Concepts and Medical Codes

From qualitative examinations of our human evaluations in Section A.6.2, we notice that the model can often learn unintended associations between medical codes and specific medical conditions. We attempt to better visualize and discern other potential mismappings through heatmap visualizations. For instance, while they can be related, traumatic brain injuries aren't a direct cause of diabetic medical codes or conditions as observed in Figure 17.

A.6.4. Debugging Case Study

Although there are more potentially, we showcase the top 20 different highly relevant dictionary features for the commonly falsely predicted ICD 99.20 code of "Injection of Platelet Inhibitors" in Figure 19. Digging deeper within the top 100 related dictionary features, we observe the following false dictionary features are not related to the medical code:

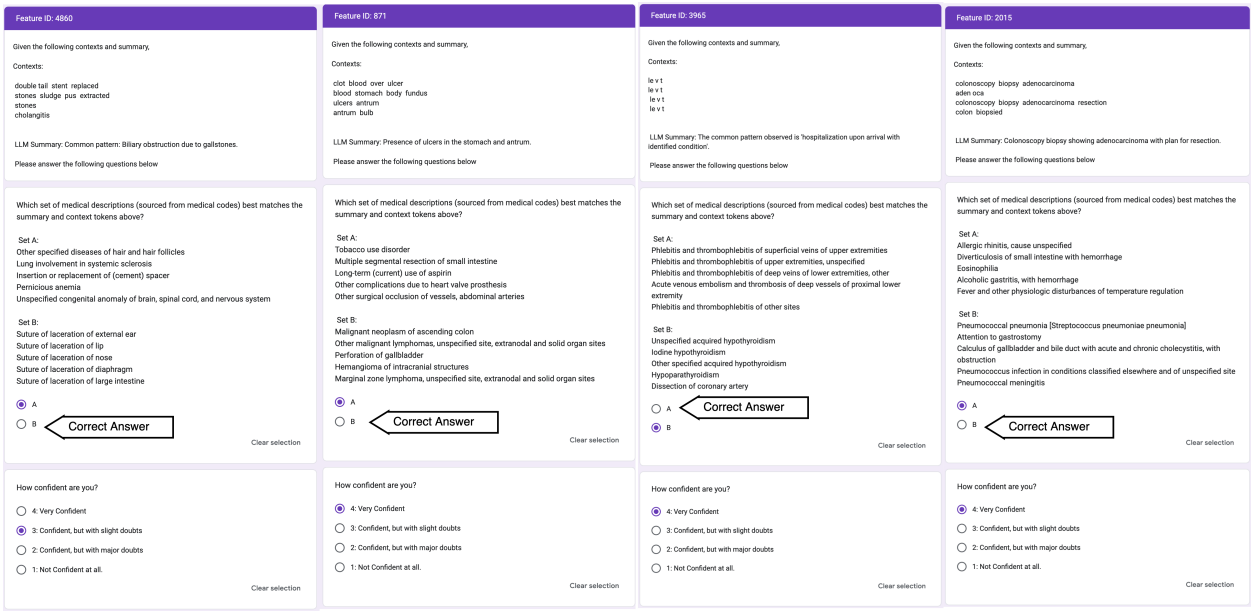

Figure 15: Human Predictability Experiment: Cases with incorrect predictions using dictionary features $\mathbf{f}$. Despite many of the contexts being highly informative, the annotator was often unable to select the set of medical codes corresponding to the top 5 medical codes defined by the $\mathbf{A}_{\mathbf{f}_{\text{icd}}}$ matrix. For example, the left contexts are related to stones, which should have no connection to suture of lacerations, but Set B is the corresponding set of codes observed. In the middle left set of contexts, ulcers are a common complication among tobacco users and smokers, but the model has possibly learned, by association, that they are more indicative of cancer. The "LEVT" contexts and the essentially hallucinated summary were virtually unidentifiable despite being extremely consistent in theme. Finally, both sets of medical codes seem to have very little relation to adenocarcinoma (a form of cancer) in the last set of contexts.

- 5188 Fractures of scapula and glenoid fossa observed repeatedly.
- 4443 Trisomy disorders, likely Down syndrome (Trisomy 21), and associated genetic counseling.
- 345 Lung conditions like bronchiolitis obliterans organizing pneumonia (BOOP) and radiation pneumonitis.
- 1558 Neonatal hyperbilirubinemia, both physiologic and pathologic, and related diagnostic workup.
- 802 Patients with end-stage amyotrophic lateral sclerosis (ALS) who are ventilator-dependent and have complications such as bronchiectasis, pneumonia, and cardiac issues.
- 6069 Patients undergoing total knee replacement surgery, particularly those with complications like respiratory distress, pain management issues, or comorbidities such as schizophrenia.
- 4917 Patients with severe dysphagia and difficulty managing oral secretions, often in the context of advanced illnesses such as cancer or critical conditions.

We ablate these dictionary features to showcase our initial debugging attempts.

A.7. Unidentifiable Dictionary Features

We attempt to investigate some of the highly relevant unidentifiable features in the heatmaps above and observe a couple findings. First, some dictionary features simply lack enough context, and second, some are truly challenging to discern a common pattern, often having a very diverse set of medical conditions that activate a specific dictionary feature.

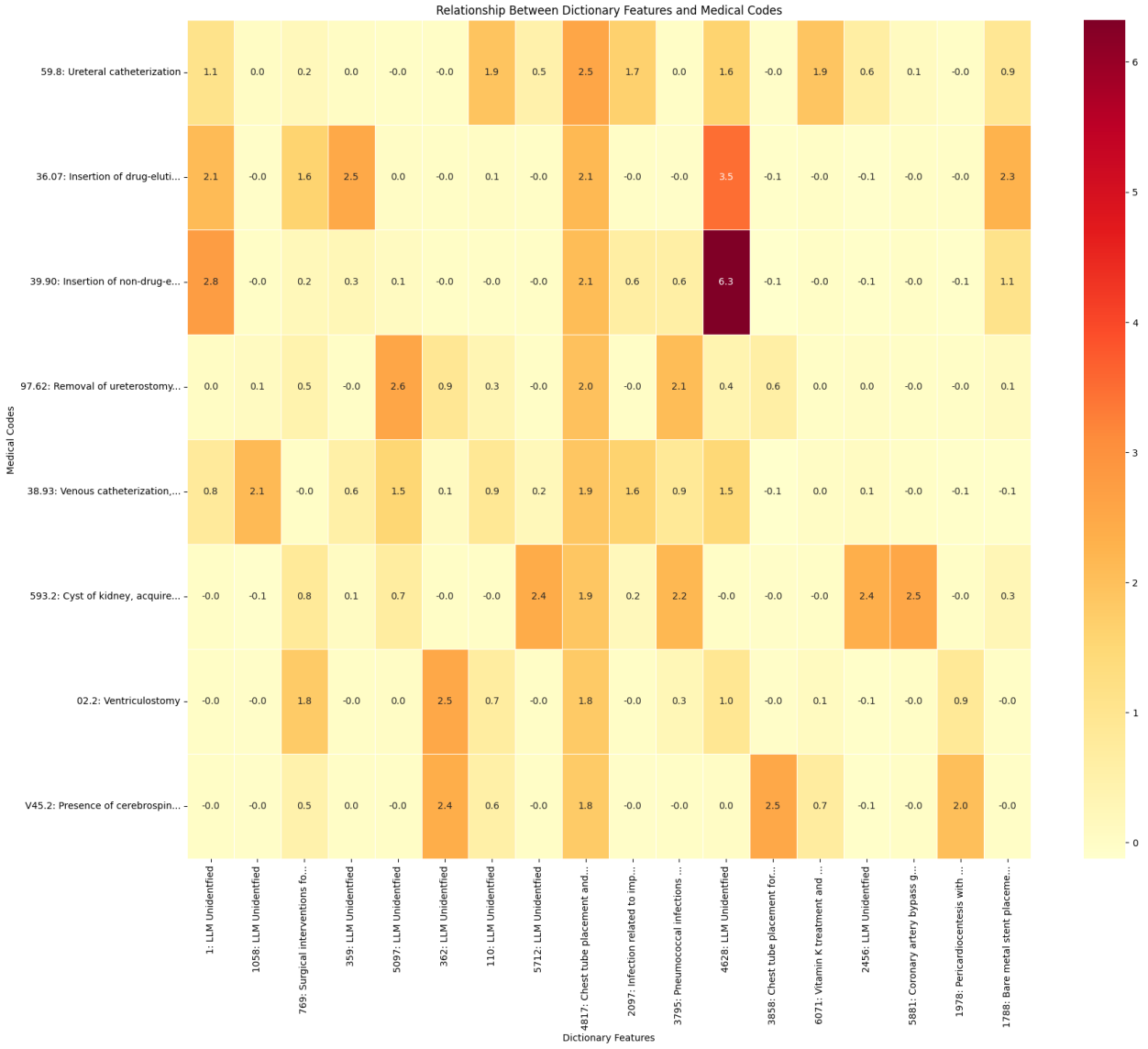

Figure 16: Heat map visualizations of the $\mathbf{A}_{\mathbf{f}_{\text{icd}}}$ matrix for medical codes associated with chest tube-related dictionary features. The x-axis represents the dictionary features, while the y-axis represents the respective medical codes. The intensity of each cell in the heat map indicates the strength of the association between a dictionary feature and a medical code. Dictionary features labeled as "LLM unidentified" denote instances where the LLM pipeline was unable to identify the underlying concept represented by the feature.

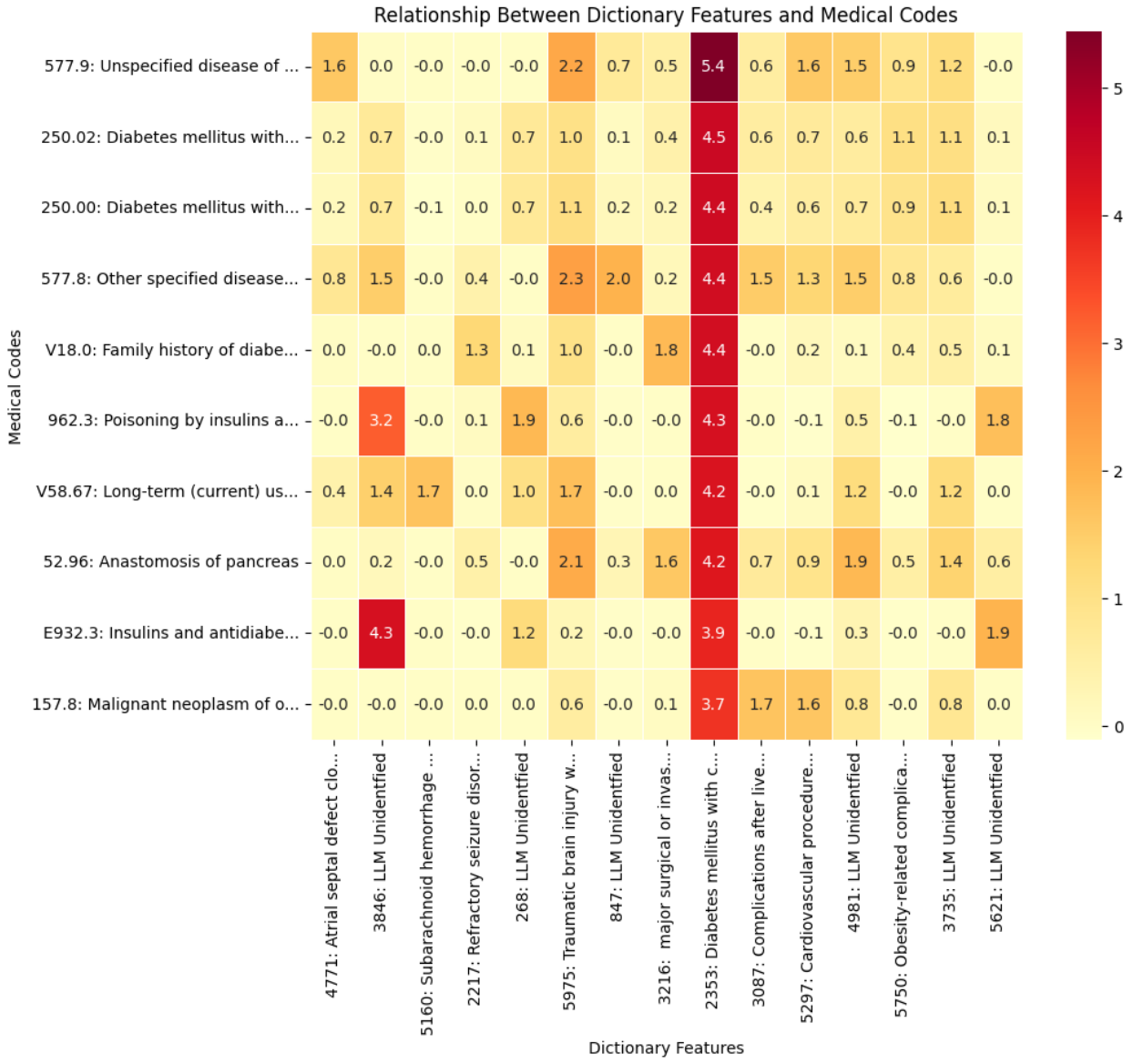

Figure 17: Heat map visualizations of the $\mathbf{A}_{\mathbf{f}_{\text{icd}}}$ matrix for medical codes associated with diabetes-related dictionary features. The x-axis represents the dictionary features, while the y-axis represents the respective medical codes. The intensity of each cell in the heat map indicates the strength of the association between a dictionary feature and a medical code. Dictionary features labeled as "LLM unidentified" denote instances where the LLM pipeline was unable to identify the underlying concept represented by the feature.

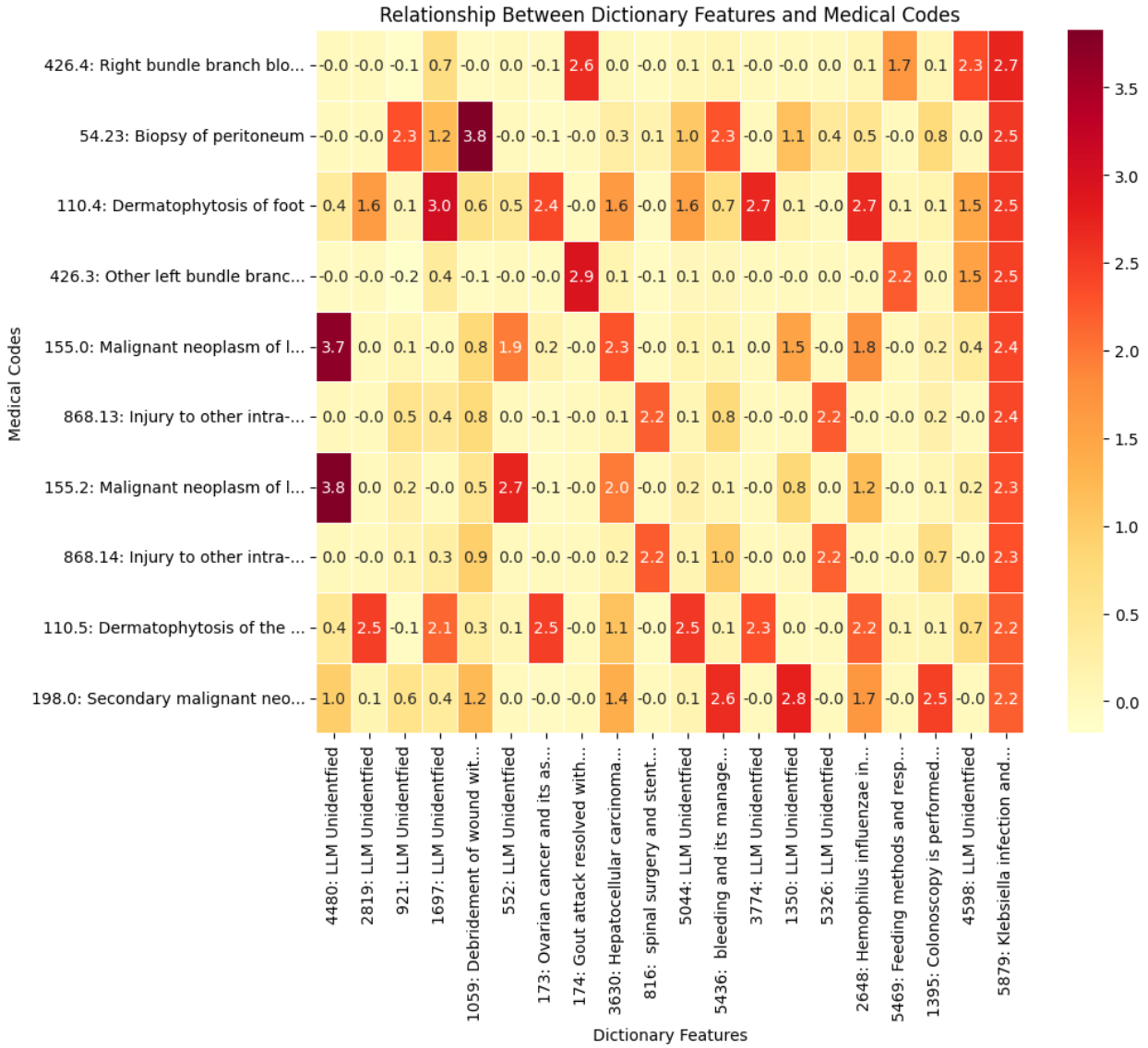

Figure 18: Heat map visualizations of the $\mathbf{A}_{\mathbf{f}_{\text{icd}}}$ matrix for medical codes associated with sepsis-related dictionary features. The x-axis represents the dictionary features, while the y-axis represents the respective medical codes. The intensity of each cell in the heat map indicates the strength of the association between a dictionary feature and a medical code. Dictionary features labeled as "LLM unidentified" denote instances where the LLM pipeline was unable to identify the underlying concept represented by the feature.

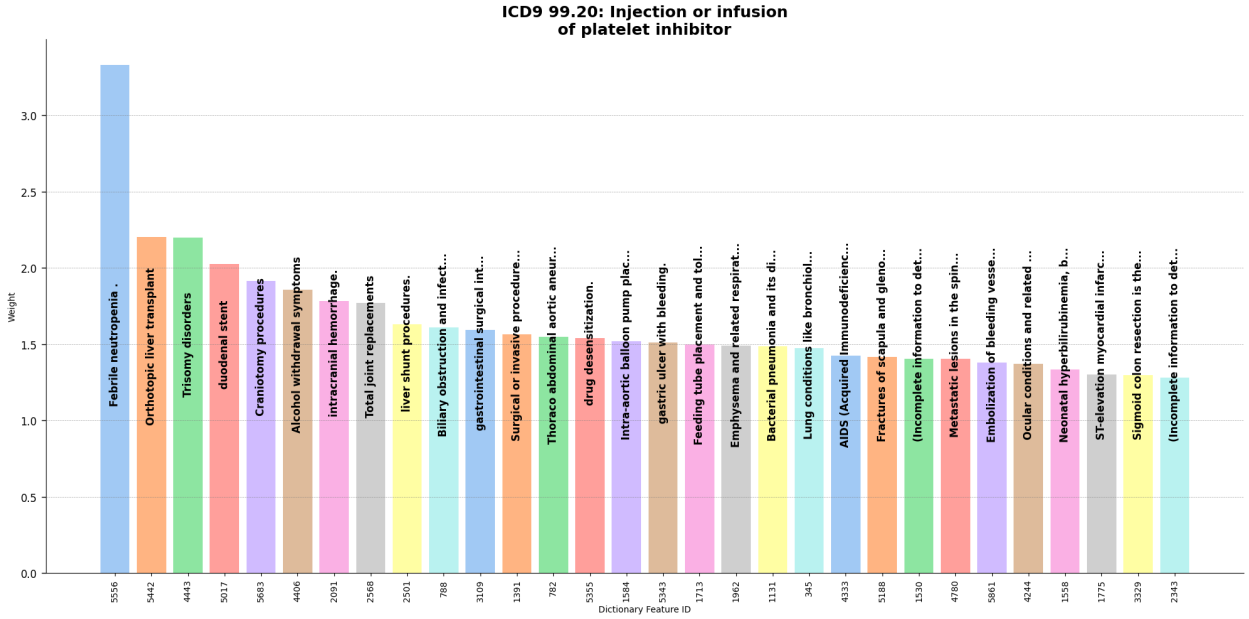

Figure 19: Top 20 dictionary features associated with ICD 99.20 Injection or infusion of platelet inhibitors. We observe many unrelated dictionary features. Note that we also manually identify some of the previously LLM unidentified contexts.

Highly Activating Tokens	Dictionary Feature f_i
ograft	3846
Surgical procedure anterior pelvic ring external fixator posterior ring fixation, Pneum cephalus brain, External fixator posterior ring fixation sacro iliac screw, Temporal, Pneum cephalus, Posterior ring fixation sacro iliac screw suprapubic catheter placement present, Allergic anaphylaxis asthmaticus steroid, Steroids, Removal of external fixator leg, sarc flare	362
Smoker fasci	1
main, e coli, c, cardiac catheterization, va ci, pulmonary, e coli, defibrillator, catheterization	3774
i ab p, liver, i ab p, idiopathic, balloon	552
axilla, v c,v c, line, fistulas, m v c, phal,infectious,port ath, perianal fistulas	1350

Table 10: Examples of Unidentified Dictionary Features. Many lack the number of contexts, or some have very divergent highly activating contexts like e coli and cardiac catheterization.