Interpretable ICD Code Classification with Faithful Sentence Extraction

Yichen Wang[†*], **Lian Hong**[†*], **Masato Mizogaki**[‡‖*],
Shunnosuke Umeda[†], **Toshimune Kenmotsu**[†], **Akihiro Tamura**[§] and **Daniel Andrade**[†¶]
[†] Hiroshima University, Japan, andrade@hiroshima-u.ac.jp
[‡] Bourbon Corporation, Japan
[§] Doshisha University, Japan

Abstract

Transformer-based models such as PLM-CA achieve strong performance for automatic ICD coding, but their attention weights do not provide faithful explanations of their predictions. This is a major limitation for electronic medical records, where users often need concise and trustworthy evidence for each assigned code. To address this issue, we jointly train a sentence extractor and an ICD code classifier such that predictions are based only on the extracted sentences. As a result, the extracted sentences serve as faithful rationales for each predicted code and substantially reduce the effort required to inspect long medical records. Experiments on MIMIC-III show that our method approaches the performance of a transformer baseline that processes the full record while using only a small fraction of the document.

1 Introduction

The International Classification of Diseases (ICD), defined by the World Health Organization, helps match patients to appropriate treatments.[1] However, the manual assignment of ICD codes is labor-intensive and requires substantial expertise. Consequently, automatic ICD coding from electronic medical records has been the focus of extensive research (Dong et al., 2022; Edin et al., 2023). Recent advances in pretrained transformer-based models have further improved ICD coding performance (Edin et al., 2023; Aden et al., 2024). In particular, Edin et al. (2024) proposed PLM-CA, a model pretrained on medical data that achieves state-of-the-art classification performance.

Unfortunately, attention mechanisms in transformer-based classification models do not provide faithful explanations for ICD code predictions (Bastings and Filippova, 2020). To address this limitation, we jointly train a sentence extractor and an ICD code classifier such that the classifier relies only on the selected sentences. Following Lyu et al. (2024), we refer to such a classifier as faithful.

Our working hypothesis is that a few key sentences, also referred to as rationales (Ehsan et al., 2019), can be sufficient to correctly judge each ICD code of a medical record. As a consequence, we propose the following two step method: (1) Select a subset of relevant sentences $\hat{S}$ from the document. (2) Predict the ICD code with an interpretable classifier using only $\hat{S}$.

With the requirement of an independent sentence selection in step (1), we prove the faithfulness of the classifier, and show that the parameters of the models in step (1) and (2) can be jointly trained (see Section 3). Furthermore, our experiments using the MIMIC-III dataset (Johnson et al., 2016) show that the proposed method can successfully identify a small number of relevant sentences that are critical for the correct ICD code classification (see Section 4).

2 Background and Related Works

ICD code prediction for electronic medical records is often formulated as a (multi-label) classification task (Edin et al., 2023). ICD code prediction is challenging due to the fact that there are several thousand codes, where many are rare and *not* mutually exclusive. As a result, many works have focused on increasing prediction accuracy (Edin et al., 2023; Dong et al., 2022).

[*] Equal contribution.
[‖] Work done at Hiroshima University during bachelor, not associated with Bourbon Corporation.
[¶] Corresponding author.
[1] https://icd.who.int/

Proceedings of the 25th Workshop on Biomedical Language Processing (BioNLP 2026), pages 679–686

Transformer-based models are often among the best performing models (Edin et al., 2023; Ravichandran et al., 2024). In particular, PLM-CA is considered to be one of the best models for ICD code prediction. The PLM-CA model, proposed by Edin et al. (2024), is an improved version of the PLM-ICD model of Huang et al. (2022), with a multi-head attention mechanism for ICD codes and RoBERTa pretrained on medical data (Lewis et al., 2020). Though, it is tempting to use a model's attention-mechanism for producing explanations of the prediction (Ravichandran et al., 2024; Mullenbach et al., 2018), these explanations are not faithful (Bastings and Filippova, 2020; Lyu et al., 2024). In particular, there are no theoretical guarantees that removing any non-highlighted text will not change the outcome of the classifier. For example, Feng et al. (2018) demonstrate that using attention scores for rationale extraction can lead to unintuitive behavior that is difficult to interpret.

Using sentences as rationales has also been considered (Herrewijnen et al., 2021; MISAWA et al., 2023), but those methods use information from the entire document for sentence selection, which violates faithfulness, as we describe in Proposition 1. Other methods require manually annotated rationales that are expensive for specialized medical domains (Chan et al., 2022).

Zhou et al. (2021) and Douglas et al. (2025) show that only a small fraction of a discharge summary is relevant for ICD coding. In particular, Douglas et al. (2025) propose a filtering step before applying PLM-CA to retain only clinically relevant text spans. However, this filtering step requires a named entity recognizer and an assertion classifier, both of which rely on fine-grained annotations. To improve interpretability, they also use AttnInGrad (Edin et al., 2024), which still does not guarantee faithfulness.

3 Proposed Method

Our goal is a classification method that bases its prediction result only on a small subset of rationales from a long document. Here, we consider each sentence as a potential rationale.

Since many ICD codes are not mutually exclusive, we consider a multi-label classification model. Let C denote the number of classes (=ICD codes). Assuming a document contains the sentences $\{s_1, s_2, \ldots, s_m\}$, we denote set of their indices by $S := \{1, 2, \ldots, m\}$. Given a constraint $0 < k < m$ on the number of rationales, for each class $c \in \{1, 2, \ldots, C\}$, we proceed as follows:

1. Using some function $h_{\boldsymbol{\eta}}$, we embed each sentence s_i *independently* into a vector representation $\mathbf{x}_i \in \mathbb{R}^H$.

2. We score each sentence for class c using some function $f^c_{\boldsymbol{\theta}} : \mathbb{R}^H \to \mathbb{R}$, and select the sentences with the top-k scores, i.e. $\hat{S}^c := \{i \in S \mid f^c_{\boldsymbol{\theta}}(\mathbf{x}_i) \geq f^c_{\boldsymbol{\theta}}(\mathbf{x}_{(k)})\}$, where

$$f^c_{\boldsymbol{\theta}}(\mathbf{x}_{(1)}) \geq f^c_{\boldsymbol{\theta}}(\mathbf{x}_{(2)}) \ldots \geq f^c_{\boldsymbol{\theta}}(\mathbf{x}_{(k)}) \ldots \geq f^c_{\boldsymbol{\theta}}(\mathbf{x}_{(m)}) .$$

3. We pool the sentence embeddings with the top-k scores and pass the resulting vector again to $f^c_{\boldsymbol{\theta}}$, which return value is interpreted as the logit for class c. That means the probability that the document is assigned ICD code c is given by

$$p(\mathbf{y}_c = 1|S) = \text{sigmoid}(f^c_{\boldsymbol{\theta}}(pool(\hat{S}^c)) , \quad (1)$$

where $\mathbf{y} \in \{0, 1\}^C$ is the class label vector of the corresponding document.

The parameters of the sentence embedder and classifier, namely $\boldsymbol{\eta}$ and $\boldsymbol{\theta}$, are jointly trained.

By construction, the proposed method has the following important property.

Proposition 1 *Removing from the document any sentence from $S \backslash \hat{S}^c$ leaves the classification result for class c unchanged.*

Proposition 1 is a basic requirement for a classifier with faithful explanations. However, transformer-like attention mechanisms do not fulfill this requirement, since the attention score for a sentence (or token) itself depends on all sentences.

Finally, we note that the proposed method (and Proposition 1) have the subtle requirement that $m > k$. Since our focus is on classification of long medical records this is not a limitation. However, we note that as an alternative to taking the top k sentences, we could introduce a threshold τ, and then use all sentences

with $f^c_{\boldsymbol{\theta}}(\mathbf{x}_i) \geq \tau$. This still fulfills Proposition 1, though, this has the disadvantage that for some documents this may lead to a large set $\hat{S}^c$, where many rationales have overlapping information content.

3.1 Joint Sentence Extraction and Classification with LLM

Below, we describe a concrete instance of the proposed method. First, using a large language model (LLM) denoted by $h_{\boldsymbol{\eta}}$, we encode each sentence s_i as a representation $\mathbf{x}_i \in \mathbb{R}^H$, i.e., $\mathbf{x}_i := h_{\boldsymbol{\eta}}(s_i)$. Because of the large number of ICD codes, we use a single LLM $h_{\boldsymbol{\eta}}$ for all classes. Next, let $W \in \mathbb{R}^{C \times H}$ and $\mathbf{b} \in \mathbb{R}^C$ denote the parameters of a linear multi-label classifier over all ICD codes. For each sentence-label pair, we compute

$$f^c_{\boldsymbol{\theta}}(\mathbf{x}_i) = \mathbf{w}_c^\top \mathbf{x}_i + b_c \,,$$

where $\mathbf{w}_c^\top$ is the c-th row of W.

For the pooling function in Equation (1), we use the mean, i.e.,

$$pool(\hat{S}^c) := \frac{1}{|\hat{S}^c|} \sum_{i \in \hat{S}^c} \mathbf{x}_i \,,$$

where $|\hat{S}^c| = k$.

The parameters of the LLM and the classifier, namely $\boldsymbol{\eta}$, W, and $\mathbf{b}$, are trained jointly. We train the model in two stages. In Stage 1, we use all sentences in S to obtain a stable initialization. In Stage 2, we switch to the top-k selection mechanism described above. Since ICD coding is a multi-label task, we optimize binary cross-entropy loss over sigmoid outputs. We denote the resulting method by LLM-SeparateTopK.

3.1.1 Common Sentence Extraction

Note that the method described above extracts k rationales for each ICD code. Alternatively, we also explore a variation where we extract in *total* only k sentences that are used for classification. For that we modify step 2 (Section 3), by determining the set of selected sentences $\hat{S}$ as follows. Define $\hat{S} := \{i \in S \,|\, \alpha_i \geq \alpha_{(k)}\}$, where

$$\alpha_{(1)} \geq \alpha_{(2)} \ldots \geq \alpha_{(k)} \ldots \geq \alpha_{(m)} \,,$$

and α_i is the aggregated positive contribution of sentence i across all labels:

$$\alpha_i := \sum_{c=1}^{C} \mathrm{ReLU}(f^c_{\boldsymbol{\theta}}(\mathbf{x}_i) - b_c) \,,$$

where we subtract the bias term to ensure that each class is given the same weight. We denote the resulting method as LLM-GlobalTopK.

3.2 Bag-of-words Model

Alternatively to the usage of an LLM, we also consider a Bag-of-words (BoW) model for $\mathbf{x}_i$, where we use all unigrams, bi-grams, and tri-grams that are contained in at least 3 sentences. Note that in this case the embedding function h has no trainable parameters. We denote the corresponding methods as BoW-SeparateTopK and BoW-GlobalTopK.

4 Experiments

4.1 Dataset and Preprocessing

For our experiments, we use the discharge summaries of the Beth Israel Deaconess Medical Center collected during 2001 to 2012, which are part of the MIMIC-III dataset (Johnson et al., 2016). For the preprocessing and train/validation/test splits we use MIMIC-III clean as in (Edin et al., 2023).[2] Each document is split into sentences using the NLTK library (Bird et al., 2009) subject to some post-processing described in Appendix A. Note that the resulting median number of sentences per document is 141.

4.2 Baselines

A simple interpretable baseline is a bag-of-words model that uses the whole document (BoW-All). Furthermore, as a rationale extraction baseline, we compare against a method that uses the PLM-CA model to identify important sentences with Integrated Gradients (Sundararajan et al., 2017), and then retrains PLM-CA. We call the resulting method IG-GlobalTopK and provide details in Appendix C. We also compare against methods that simply extract the first k sentences of the document

[2] See https://github.com/JoakimEdin/medical-coding-reproducibility and https://github.com/JoakimEdin/explainable-medical-coding/tree/main/data/splits.

	Classification				Ranking		
	AUC-ROC		F1		Precision@r		MAP
	Micro	Macro	Micro	Macro	8	15	
Proposed methods							
LLM-SeparateTopK (k = 10)	**98.2**	**94.2**	**51.7**	**18.3**	**65.0**	**50.2**	**55.5**
LLM-GlobalTopK (k = 10)	97.1	89.8	46.0	13.5	59.4	44.6	48.4
BoW-GlobalTopK (k = 10)	94.6	78.1	25.7	2.9	49.9	37.8	39.0
BoW-SeparateTopK (k = 10)	94.0	75.3	38.5	5.8	58.8	44.5	47.2
Baseline methods							
IG-GlobalTopK (k = 10)	96.1	86.1	25.6	1.9	45.5	34.7	35.4
BoW-Head (k = 10)	93.4	74.2	21.1	2.09	45.4	34.4	35.0
BoW-Tail (k = 10)	91.3	66.6	13.4	0.92	36.9	28.5	27.6
BoW-All	94.3	76.1	35.5	5.1	57.6	43.5	46.6
LLM-Head (k = 10)	96.0	86.7	35.6	7.4	45.9	35.0	36.0
LLM-Tail (k = 10)	94.2	80.7	29.1	3.8	36.8	28.6	28.0
LLM-All	95.4	85.0	38.6	10.2	52.2	39.2	41.5
PLM-CA	**98.9**	**95.7**	**59.8**	**29.0**	**72.3**	**56.7**	**65.0**

Table 1: Comparison of baseline and proposed methods for ICD code prediction in terms of micro/macro AUC-ROC, micro/macro F1, MAP, and Precision@r (r = 8, 15). MAP: mean average precision. Best results highlighted in bold, for proposed and baseline methods, respectively.

(i.e., the head sentences) or the last k sentences (i.e., the tail sentences), which we denote as BoW-Head and BoW-Tail, respectively. Analogously, we compare to a method that pools all sentence embeddings (LLM-All), the first k sentence embeddings (LLM-Head), or the last k ones (LLM-Tail). Finally, for reference, we also report the results of the original PLM-CA model, which uses the whole document (Douglas et al., 2025).

4.3 Results

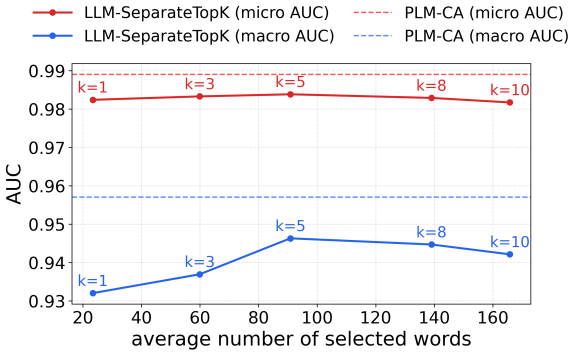

Figure 1: Proposed method's AUC scores for different values of k.

For the proposed methods and corresponding baselines, we set $k = 10$, which reduces most documents by more than 90%. We evaluate all methods on all 3,681 ICD codes occurring in the training and test splits, and report micro/macro AUC-ROC, micro/macro F1, mean average precision (MAP), and Precision@r in

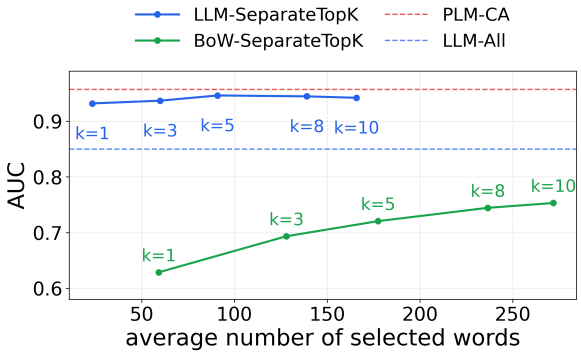

Figure 2: Comparison of different method's macro AUC scores for different values of k.

Table 1. Figure 1 shows the AUC scores for different values of k for the proposed method.

Our results show that the proposed classifier remains accurate while using only a small subset of each document. In particular, for $k = 10$, LLM-SeparateTopK performs better than all baselines in Table 1, including the rationale-based baseline IG-GlobalTopK as well as naive head/tail extraction methods. Although full-document PLM-CA remains a strong reference point, the proposed method often approaches its performance while providing faithful rationale extraction. Table 2 shows one example of the extracted sentences.

Finally, in Figure 2, we also show the macro AUC of the proposed method LLM-SeparateTopK for different number of selected sentences k. Interestingly, we see that the pro-

posed method performs *better* than LLM-All, suggesting that selecting only few sentences has the additional benefit of removing noise.

Rank	Extracted Sentence
1	"discharge diagnosis angioedema"
2	"brief hospital course angioedema followed by dr"
3	"you were admitted to the hospital with throat swelling which was an exacerbation of your chronic angioedema"

Table 2: Top three extracted sentences produced by LLM-SeparateTopK ($k = 10$) for ICD code 995.1 (Angioneurotic edema) from a discharge summary.

5 Conclusions

In this paper, we proposed a provably faithful approach to rationale extraction for ICD code prediction. Experiments on MIMIC-III showed that our method can extract a small number of sentences *without any sentence- or token-level annotations* while achieving competitive coding performance, even compared with methods that process the full discharge summary. Overall, our results suggest that faithful sentence-level rationales are a promising step toward more transparent and clinically useful medical coding systems.

As future work, we plan to evaluate the extracted rationales against human-annotated evidence resources such as MDACE (Cheng et al., 2023), which would also allow us to assess plausibility.

Limitations

The current method judges the importance of each sentence independently of context. While this is an important requirement for the faithfulness, as described in Proposition 1, this might lead to wrongly excluding short sentences that are important when judged in the context of its surrounding sentences.

References

Ilyas Aden, Christopher HT Child, and Constantino Carlos Reyes-Aldasoro. 2024. International classification of diseases prediction from mimiic-iii clinical text using pre-trained clinicalbert and nlp deep learning models achieving state of the art. *Big Data and Cognitive Computing*, 8(5):47.

Jasmijn Bastings and Katja Filippova. 2020. The elephant in the interpretability room: Why use attention as explanation when we have saliency methods? In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 149–155.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit.* " O'Reilly Media, Inc.".

Aaron Chan, Maziar Sanjabi, Lambert Mathias, Liang Tan, Shaoliang Nie, Xiaochang Peng, Xiang Ren, and Hamed Firooz. 2022. Unirex: A unified learning framework for language model rationale extraction. In *International Conference on Machine Learning*, pages 2867–2889. PMLR.

Hua Cheng, Rana Jafari, April Russell, Russell Klopfer, Edmond Lu, Benjamin Striner, and Matthew Gormley. 2023. MDACE: MIMIC documents annotated with code evidence. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7534–7550, Toronto, Canada. Association for Computational Linguistics.

Hang Dong, Matúš Falis, William Whiteley, Beatrice Alex, Joshua Matterson, Shaoxiong Ji, Jiaoyan Chen, and Honghan Wu. 2022. Automated clinical coding: what, why, and where we are? *NPJ digital medicine*, 5(1):159.

James C. Douglas, Yidong Gan, Ben Hachey, and Jonathan K. Kummerfeld. 2025. Less is more: Explainable and efficient ICD code prediction with clinical entities. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 30835–30847, Vienna, Austria. Association for Computational Linguistics.

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*, pages 2572–2582.

Joakim Edin, Maria Maistro, Lars Maaløe, Lasse Borgholt, Jakob Drachmann Havtorn, and Tuukka Ruotsalo. 2024. An unsupervised approach to achieve supervised-level explainability in healthcare records. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 4869–4890, Miami, Florida, USA.

Upol Ehsan, Pradyumna Tambwekar, Larry Chan, Brent Harrison, and Mark O Riedl. 2019. Automated rationale generation: a technique for explainable ai and its effects on human perceptions. In *Proceedings of the 24th international*

conference on intelligent user interfaces, pages 263–274.

Shi Feng, Eric Wallace, Alvin Grissom II, Mohit Iyyer, Pedro Rodriguez, and Jordan Boyd-Graber. 2018. Pathologies of neural models make interpretations difficult. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3719–3728.

Elize Herrewijnen, Dong Nguyen, Jelte Mense, Floris Bex, and 1 others. 2021. Machine-annotated rationales: faithfully explaining text classification. In *Proceedings for the Explainable Agency in AI Workshop at the 35th AAAI Conference on Artificial Intelligence (Washington DC: AAAI Press)*, pages 11–18.

Chao-Wei Huang, Shang-Chi Tsai, and Yun-Nung Chen. 2022. PLM-ICD: Automatic ICD coding with pretrained language models. In *Proceedings of the 4th Clinical Natural Language Processing Workshop*, pages 10–20, Seattle, WA. Association for Computational Linguistics.

Alistair EW 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):1–9.

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.

Qing Lyu, Marianna Apidianaki, and Chris Callison-Burch. 2024. Towards faithful model explanation in NLP: A survey. *Computational Linguistics*, 50(2):657–723.

Shotaro MISAWA, Taiki FURUKAWA, Shintaro OYAMA, Ryuji KANO, Hirokazu YARIMIZU, Tomoki TANIGUCHI, Kohei ONODA, Kikue SATO, and Yoshimune SHIRATORI. 2023. Sentence extraction using outcome prediction model trained from clinical data. *Proceedings of the Annual Conference of JSAI*, JSAI2023:3Xin404–3Xin404.

James Mullenbach, Sarah Wiegreffe, Jon Duke, Jimeng Sun, and Jacob Eisenstein. 2018. Explainable prediction of medical codes from clinical text. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1101–1111.

Ajay Madhavan Ravichandran, Julianna Grune, Nils Feldhus, Aljoscha Burchardt, Roland Roller, and Sebastian Möller. 2024. XAI for better exploitation of text in medical decision support. In *Proceedings of the 23rd Workshop on Biomedical Natural Language Processing*, pages 506–513, Bangkok, Thailand. Association for Computational Linguistics.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 3319–3328.

Tong Zhou, Pengfei Cao, Yubo Chen, Kang Liu, Jun Zhao, Kun Niu, Weifeng Chong, and Shengping Liu. 2021. Automatic icd coding via interactive shared representation networks with self-distillation mechanism. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5948–5957, Online.

A Sentence Splitting

First, each document is split into sentences using the sent_tokenize function[3] of the NLTK library (Bird et al., 2009). Afterwards, the following post-processing steps are used to reduce conflation and improve comprehensiveness:

1. If there are two or more line breaks, split the sentence before and after the line break.

2. If there are two or more spaces, split the sentence before and after the spaces.

3. Replace the line break characters with spaces.

4. Remove spaces at the beginning and end of the sentence.

5. If each sentence ends with a ":", join that sentence to the next sentence with a space.

The resulting median number of sentences (and IQR) per document is 141 (102-189). The median number of words (and IQR) in a sentence are 8 (3-13). The median number of words (and IQR) in a document is 1,375 (965-1,900).

[3] https://www.nltk.org/api/nltk.tokenize.sent_tokenize.html

B Training Details

For LLM-SeparateTopK, we trained the model for up to 9 epochs and used early stopping with a patience of 3 epochs according to the validation loss. The learning rates for the pretrained encoder and the linear classifier were set to 3×10^{-6} and 5×10^{-5}, respectively. Owing to the memory requirement of the model, we used a mini-batch size of 1 document and accumulated gradients for 2 steps, yielding an effective batch size of 2. No explicit regularization was used; instead, early stopping was employed to mitigate overfitting. All experiments were run on an NVIDIA Blackwell GPU, and each epoch took around 2 hours.

C Sentence-level Top-k Selection via Integrated Gradients (IG-GlobalTopK)

As a rationale-extraction baseline, we use Integrated Gradients (IG) (Sundararajan et al., 2017) to score sentences and then retrain PLM-CA on the extracted sentences only.

Given a document with S sentences, where each sentence s_i consists of a set of tokens T_i, we first compute the attribution of each token. Let $F : \mathbb{R}^n \to \mathbb{R}$ be the model's loss function (Binary Cross-Entropy), $\mathbf{u} \in \mathbb{R}^n$ be the vector of all tokens in the document, and $\mathbf{u}' \in \mathbb{R}^n$ be a baseline consisting of PAD tokens. The IG score for the j-th token is defined as:

$$\mathrm{IG}_j(\mathbf{u}) = (u_j - u'_j) \int_{\alpha=0}^{1} \frac{\partial F(\mathbf{u}' + \alpha(\mathbf{u} - \mathbf{u}'))}{\partial u_j} d\alpha.$$

To obtain a document-level importance score at the sentence level, we aggregate the token-wise attributions within each sentence by calculating their mean:

$$\mathrm{Score}(s_i) = \frac{1}{|T_i|} \sum_{j \in T_i} \mathrm{IG}_j.$$

Similar to the top-k selection of our proposed method, we then define

$$\hat{S}_{\mathrm{IG}} := \{ i \in S \,|\, \mathrm{Score}(s_i) \geq \mathrm{Score}(s_{(k)}) \},$$

where

$$\mathrm{Score}(s_{(1)}) \geq \mathrm{Score}(s_{(2)}) \geq \ldots \geq \mathrm{Score}(s_{(m)}).$$

We then retrain PLM-CA using only $\hat{S}_{\mathrm{IG}}$, while preserving the original sentence order.

To prevent discrepancies between training and testing, we use the following retraining procedure: we partition the training set into five folds: 4/5 are used for training and 1/5 for IG-based sentence selection, and repeat this 5 times. Finally, we train a new classifier $\mathcal{C}$ on the extracted sentences from the whole training data.

The resulting method, denoted by IG-GlobalTopK, uses at test time classifier $\mathcal{C}$ applied to extracted sentences only. However, note that IG-GlobalTopK is not strictly faithful: a sentence s_i that was *not selected* could have influenced the choice of $\hat{S}_{\mathrm{IG}}$. This in consequence, allows sentence s_i to influence indirectly the final classification result.

D Example of Extracted Sentences

We provide additional examples here, including both successful and failure cases for LLM-SeparateTopK (Tables 3 and 4), as well as a successful case for LLM-GlobalTopK (Table 5).

ICD Code	493.90 long description: Asthma, unspecified type, unspecified
Rank	**Extracted Sentence**
1	"asthma continued flovent prn"
2	"asthma"
3	"fluticasone mcg actuation aerosol sig one puff inhalation hospital1 times a day as needed for shortness of breath or wheezing"

Table 3: An additional successful example of extracted sentences produced by the proposed LLM-SeparateTopK ($k = 10$) for ICD code 493.90 from one discharge summary (admission id = 179767). The selected sentences directly indicate the target respiratory condition and its related medication use.

ICD Code	886.0 long description: Traumatic amputation of other finger(s), complete, without mention of complication
Rank	**Extracted Sentence**
1	"the patient underwent completion amputation of the left ring finger repair of digital ulnar and radial nerve to the middle finger index finger and thumb repair of the fpl ftp of the middle finger and fts ftp of the index finger repair of the epl to the thumb and k wire fixation of the metacarpal of index and phalanx of the first finger"
2	"history of present illness the patient arrived from another hospital in hospital3 to the emergency room with a complete amputation of his thumb index long and ring fingers of his left hand"
3	"discharge diagnosis as described above amputation of thumb index long and ring fingers reattachment of thumb index and long fingers and revision amputation of ring finger"

Table 4: A representative failure example of extracted sentences produced by the proposed LLM-SeparateTopK ($k = 10$) for ICD code 886.0 from one discharge summary (admission id = 174415). Although the selected sentences are clinically plausible and strongly related to traumatic finger injury, the model incorrectly assigns this code.

ICD Code	995.1 and 493.90
Rank	**Extracted Sentence**
1	"asthma"
2	"chronic urticaria and angioedema since c section done on for failure to progress after hour labor"
3	"brief hospital course angioedema followed by dr"
4	"discharge diagnosis angioedema"
5	"asthma continued flovent prn"
6	"physical exam t bp hr rr o2 sat ra wt kg gen nad speaking in full sentences without sob no audible stridor heent moon facies mmm angioedema no edema noted in posterior oropharynx and posterior oropharynx clearly visualized behind tongue neck supple cv rrr s1 split s2 ii vi systolic murmur over lsb lungs cta b l without wheezing rhonchi or rales abd soft obese nt nd normoactive bs ext no le edema wwp neuro aao x skin no rashes noted"
7	"history of present illness yo f with h o idiopathic chronic urticaria and angioedema requiring intubation during previous admission asthma and seasonal allergies who presents with tongue swelling x hrs"
8	"medications on admission epipen prn loratadine mg daily ranitidine mg qam doxepin mg qhs benadryl mg q6h prn calcium mg daily vit d units daily ativan mg qhs flovent mcg puff hospital1 prn for angioedema"
9	"first name4 namepattern1 last name namepattern1 as outpt for idiopathic chronic urticaria and angioedema with multiple admissions in past year for angioedema"
10	"otherwise no history of related urticaria angioedema"

Table 5: An additional successful example of extracted sentences produced by the proposed LLM-GlobalTopK ($k = 10$) for all ICD code from one discharge summary (admission id = 179767). The selected sentences directly indicate the target respiratory condition and its related medication use.