Evaluation and LLM-Guided Learning of ICD Coding Rationales

Mingyang Li[1], Viktor Schlegel[1,2,3], Tingting Mu[1], Wuraola Oyewusi[1], Kai Kang[4], Goran Nenadic[1]

[1]University of Manchester, Department of Computer Science,
[2]Imperial Global Singapore, [3]Imperial College London, Department of Bioengineering,
[4]Shanxi Medical University

{mingyang.li, Tingting.Mu, gnenadic}@manchester.ac.uk, v.schlegel@imperial.ac.uk
wuraola.oyewusi@postgrad.manchester.ac.uk, k.nikey0422@gmail.com

Abstract

ICD coding is the process of mapping unstructured text from Electronic Health Records (EHRs) to standardised codes defined by the International Classification of Diseases (ICD) system. In order to promote trust and transparency, existing explorations on the explainability of ICD coding models primarily rely on attention-based rationales and qualitative assessments conducted by physicians, yet lack a systematic evaluation across diverse types of rationales using consistent criteria and high-quality rationale-annotated datasets specifically designed for the ICD coding task. Moreover, dedicated methods explicitly trained to generate plausible rationales remain scarce. In this work, we present evaluations of the explainability of rationales in ICD coding, focusing on two fundamental dimensions: faithfulness and plausibility—in short how rationales influence model decisions and how convincing humans find them. For plausibility, we construct a novel, multi-granular rationale-annotated ICD coding dataset, based on the MIMIC-IV database and the updated ICD-10 coding system. We conduct a comprehensive evaluation across three types of ICD coding rationales: entity-level mentions automatically constructed via entity linking, LLM-generated rationales, and rationales based on attention scores of ICD coding models. Building upon the strong plausibility exhibited by LLM-generated rationales, we further leverage them as distant supervision signals to develop rationale learning methods. Additionally, by prompting the LLM with few-shot human-annotated examples from our dataset, we achieve notable improvements in the plausibility of rationale generation in both the teacher LLM and the student rationale learning models.

1 Introduction

Clinical coding is the process of translating free-text descriptions in patients' Electronic Health Records (EHRs) into standardized codes, serving a critical role in billing, reimbursement, auditing, and decision support within healthcare systems (Blundell, 2023). In this study, we focus on ICD coding—the document-level assignment of codes from the International Classification of Diseases (ICD) system (Tzitzivacos, 2007), which provides hierarchical alphanumeric identifiers representing medical diagnoses and procedures.

ICD coding relies on manual efforts by trained professionals, which is costly, labour-intensive, and error-prone (Nguyen et al., 2018). To mitigate these challenges, rule-based systems were developed (Pereira et al., 2006; Crammer et al., 2007), followed by machine learning approaches such as Support Vector Machines (SVM) (Lita et al., 2008). With the rise of deep learning, approaches based on like Gated Recurrent Units (GRUs) (Catling et al., 2018) and Convolutional Neural Networks (CNNs) (Karimi et al., 2017) substantially improved coding efficiency and accuracy. More recently, attention-based architectures (Liu et al., 2021; Van Aken et al., 2022; Yuan et al., 2022) such as transformers (Michalopoulos et al., 2022; Yogarajan et al., 2022; Yang et al., 2022) have been adopted, consistently achieving state-of-the-art results on clinical coding benchmarks.

Although these methods have achieved notable success in ICD coding, their inherent lack of explainability poses a major challenge for understanding and interpreting model decisions. This limitation may undermine trust and transparency, which are critical for enabling healthcare professionals and patients to rely on AI-driven recommendations (Amann et al., 2020). To address this issue, researchers have increasingly developed methods that provide reliable explanations, often by extracting short text snippets (*rationales*) using attention mechanisms. For the evaluation of these explanations, some prior studies rely on physicians' assessments, which are based on various evaluation rubrics; while others use the only exist-

Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics
Volume 1: Long Papers, pages 4969–5003

ing rationale-annotated resource, MDACE (Cheng et al., 2023), based on MIMIC-III and thus using the outdated ICD-9 system. Furthermore, MDACE was re-annotated with new codes, resulting in a significant label distribution shift from the original MIMIC-III labels, exacerbating the evaluation of training-based approaches.

These issues reveal three main gaps in this task: *(a)* the absence of a rationale (evidence) dataset specifically designed for the ICD coding task and constructed based on modern resources; *(b)* limited explorations of different types of rationales—as most existing studies focus exclusively on rationales derived from attention scores—and thus a lack of their consistent comparative evaluation; and *(c)* the absence into *rationale learning* approaches given large-scale supervised datasets required for effective training do not exist.

To address this gap, we comprehensively evaluate the quality of rationales in ICD coding from the angles of faithfulness and plausibility—in short, how rationales affect model classificatio ndecisions and how plausible human annotators find them—following Edin et al. (2024). Specifically, we systematically evaluate three types of rationales, which is enabled by a new rationale dataset we introduce in this paper. Finally, we explore two rationale learning approaches and examine the benefits of leveraging few-shot examples from our dataset for rationale generation and rationale learning. Our key contributions[1] are summarised below:

- We construct a new rationale dataset for ICD coding task specifically based on the up-to-date MIMIC-IV benchmark with ICD-10 coding system for plausibility evaluation. This dataset provides richer rationales across multiple levels of granularity.

- We conduct a comprehensive comparison of rationale plausibility across three types: (1) naive entity-level rationales derived from an entity linking dataset; (2) strong LLM-generated rationales, generated by both cloud-based and locally deployed LLMs; and (3) model-generated rationales, including both unsupervised and supervised approaches.

- We investigate two rationale learning approaches—multi-objective learning and a named entity recognition (NER) formulation—supervised by LLM-generated weak rationale labels.

- We demonstrate that leveraging few-shot human-annotated examples from our rationale dataset further improves both rationale generation and rationale learning.

[1]The annotated rationale dataset, the code used in this study, and the Gemini-generated rationale dataset covering 122K MIMIC-IV ICD-10 documents are publicly available at: https://github.com/mingyangligithub/ICD-Coding-Explainability-Evaluation.

2 Related Work

Rationale Snippets. In one of the seminal studies on model explainability in clinical coding, Mullenbach et al. (2018) focus on extracting the most influential text snippets associated with predicted labels based on the importance values (attention weights) to n-grams in the discharge summaries. Lovelace et al. (2020) follow the same idea, but apply attention mechanisms over multiple convolutional filters of different lengths, which allow them to consider variable spans of text. Dong et al. (2021) build a Hierarchical Label-wise Attention Network (HLAN), which has label-wise word-level and sentence-level attention mechanisms, providing more comprehensive explanations for each label by highlighting key words and sentences. Wang et al. (2022) visualise the attention distribution to provide explanations. Similarly, Gao et al. (2024) design heatmap visualisation to help coders better understand the inference logic from the notes.

Evaluation Methods and Rationale Learning. Mullenbach et al. (2018) evaluate the models' effectiveness in identifying highly informative rationales by a physician's assessment. Kim et al. (2022) assess explainability using human-grounded evaluation, where annotators rate each explanation rationale for a predicted code as highly informative, informative, or irrelevant. Van Aken et al. (2022) evaluate the explainability by faithfulness and conduct a manual analysis to judge whether highlighted tokens and prototypical patients aided decision-making. These studies primarily focus on evaluating the attention-based rationales. Edin et al. (2024) additionally assess the explainability of gradient-based and perturbation-based rationales. Furhtermore, both Cheng et al. (2023) and Edin et al. (2024) investigate whether rationales can be learned jointly with the main task of ICD coding (multi-objective), but their approaches are constrained by the small size of the dataset and the limited set of labels that it covers.

In this work, we examine whether other types of rationales, specifically those produced by entity linking and large language models (LLM), can serve as effective explanations. We further propose a new rationale learning approach based on an NER formulation, which surpasses the multi-objective learning approach in generating plausible rationales. Moreover, we examine the effectiveness of LLM-generated labels in addressing the limitations of scarce and biased supervised data.

Data Limitation. To facilitate automated evaluation of rationales Cheng et al. (2023) introduced MDACE, the only existing rationale-annotated ICD coding dataset other than the one we present in this paper. However, MDACE exhibits several limitations that limit its suitability to be the de-facto standard of rationale evaluation of ICD coding models:

Distribution Shift: The MDACE annotations are conducted independently of existing MIMIC-III labels, by coding each chart from scratch[2]. Furthermore, coders used ICD-10 in line with their experience, which were subsequently mapped to ICD-9. This results in a significant code distribution shift from the standard MIMIC-III training set. Specifically, overlap for both Top-50 and all codesis only 37.00% and 14.59% on average per document, respectively. Notably, 40 new ICD-9 are introduced in MDACE, while 725 out of 1,281 codes in the Full setting are completely omitted. As a consequence, the performance of PLM-ICD (Huang et al., 2022) drops from 78.02% Precision@8 on the official MIMIC-III to only 55.34% on MDACE (Appendix D for more details); thus, MDACE significantly underestimates the performance of trained models.

Sparse Rationale Annotations: Most ICD codes have very sparse rationale annotations, often limited to a single instance. Such limited coverage may omit valid generated rationales, which identify other statements that support the classification, such as repeated mentions of the same finding or drug that supports a diagnosis.

Outdated Relevance of MIMIC-III: With the obsoletion of ICD-9, the relevance of MIMIC-III vanes for ICD coding research—recent work uses the more recent MIMIC-IV dataset with ICD-10 annotation.

Taken together, these observations mandate a strong need to develop a modern rationale annotation resource specifically for ICD coding. We heed the call in this paper by introducing a rationale-annotated dataset, based on MIMIC-IV and its ICD-10 code annotations. We demonstrate further utility of this resource by using instances as few-shot examples to prompt LLMs, which benefits both rationale generation and rationale learning.

[2]coders could consult the original ICD-9 codes for reference

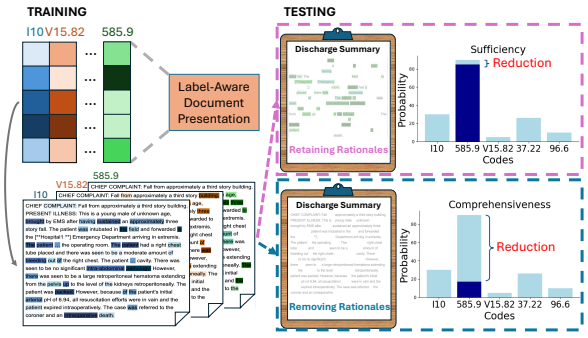

Figure 1: Faithfulness testing workflow. Sufficiency and comprehensiveness are evaluated by retaining or removing rationales from the original documents and using the modified texts as inputs to the trained ICD coding models.

3 Preliminary on Explainable ICD Coding

ICD coding is treated as a multi-label classification task. It assigns ICD codes based on a patient's clinical record, describing their diseases or procedures. We consider clinical documents, each of which is a discharge summary denoted as $\mathbf{x}_i = \{\mathbf{t}_{i,1}, \mathbf{t}_{i,2}, ..., \mathbf{t}_{i,N_t}\}$, consisting of N_t tokens. The ICD coding model computes a label distribution over N_l labels for the input $\mathbf{x}_i$, i.e., $f(\mathbf{x}_i) = \mathbf{p}_i = [p_{i,1}, p_{i,2}, \ldots, p_{i,N_l}]$. The final codes are selected by thresholding the predicted probabilities with $0 < \tau < 1$, i.e., $\hat{y}_{i,l} = \begin{cases} 1, & \text{if } p_{i,l} > \tau \\ 0, & \text{otherwise} \end{cases}$, for $l = 1, 2, \ldots, N_l$.

To enable explainable ICD coding, the prediction models are expected to provide rationale explaining the decision making. The family of attention-driven ICD coding models achieves this by flagging key text rationales influential to the prediction using attention weights. We select three attention-driven ICD coding models: CAML, LAAT and PLM-ICD. They compute an attention weight $\tilde{a}_{i,j,l}$ for each token $\mathbf{t}_{i,j}$ and for each label l. Specifically, CAML (Mullenbach et al., 2018) employs a single-filter CNN to encode the input text and computes the attention weight by $\tilde{a}_{i,j,l} = \text{softmax}(\mathbf{u}_l^T \mathbf{t}_j)$. LAAT

Table 1: Statistics of RD-IV-10 and MDACE. The statistics are computed based on annotations from documents that appear in the ICD coding datasets. A / B: A represents the statistics of the rationale datasets, and B represents the statistics of MIMIC-IV ICD-10 (compared with RD-IV-10) and MIMIC-III ICD-10 (compared with MDACE).

Statistics	RD-IV-10	MDACE
No. documents	150	354
Tokens / doc	1690.63	1837.27
Codes / doc	14.82 / 16.15	11.57 / 17.54
Code overlap (Full / Top-50)	93.15% / 83.88%	37.00% / 14.59%
No. codes	2223 / 2422	4096 / 6208
No. distinct codes	989 / 1044	1195 / 1381
No. annotations	5391	4992
Annotations / doc	35.94	14.10
Tokens / annotation	5.44	2.13

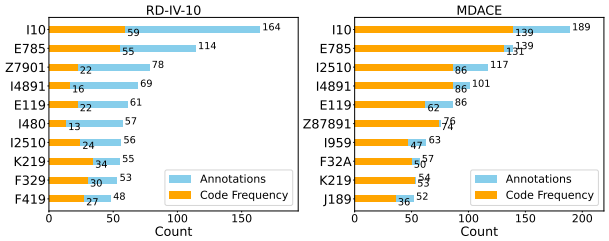

Figure 2: Statistics of code and annotation frequencies for the top 10 codes in RD-IV-10 and MDACE. RD-IV-10 provides richer annotations for each label than MDACE.

(Vu et al., 2020) shares the same underlying framework as CAML, but leverages a BiLSTM to represent the input text $\mathbf{x}_i$ and computes the attention weight by $\tilde{a}_{i,j,l} = \text{softmax}\left(\mathbf{u}_l^T \tanh(\mathbf{W}_j \mathbf{t}_j)\right)$, introducing an additional weight matrix $\mathbf{W}_j$. PLM-ICD (Huang et al., 2022) utilizes a transformer pre-trained on biomedical and clinical texts to encode the input text, and employs the same attention layer as LAAT. The attention weights are then used to compute a label-aware document representation by $\tilde{\mathbf{h}}_{i,l} = \sum_{j=1}^{N_t} \tilde{a}_{i,j,l}\, \mathbf{t}_j$. Together with the label representation vector $\mathbf{z}_l$, $\tilde{\mathbf{h}}_{i,l}$ is used to estimate the label probability by $p_{i,l} = \sigma\left(\mathbf{z}_l^T \tilde{\mathbf{h}}_{i,l}\right)$ through the sigmoid function σ. These models are trained by minimizing a binary cross-entropy loss:

$$\mathcal{L}_{\text{coding}} = -\frac{1}{N_D} \sum_{i=1}^{N_D} \sum_{l=1}^{N_l} \left[y_{i,l} \log \hat{y}_{i,l} + (1 - y_{i,l}) \log(1 - \hat{y}_{i,l}) \right], \quad (1)$$

where N_D denotes the document number. Rationales are then extracted by selecting *top* $p\%$ *tokens* or *top* N *tokens* with the highest attention weights.

4 An Empirical Analysis of ICD Rationales

4.1 Evaluation Metrics

Explainability is typically evaluated from two complementary perspectives (Mendez Guzman et al., 2024): a) Model-centric faithfulness, assessing how accurate the extracted rationales reflect the internal reasoning of the prediction models. b) Human-centric plausibility, measuring how convincing the rationales appear to people.

Faithfulness. We denote the extracted rationale for explaining why input $\mathbf{x}_i$ is predicted to label l by $\hat{\mathbf{r}}_{i,l}$. Its faithfulness can be assessed by two metrics, including **sufficiency** and **comprehensiveness** (DeYoung et al., 2019), quantifying the effect of only retaining or removing rationales, respectively. A rationale $\hat{\mathbf{r}}_{i,l}$ is considered sufficient if it enables a prediction that closely approximates the one produced by using the full input $\mathbf{x}_i$. This motivates the sufficiency metric of $\text{Suff} = P(f(\mathbf{x}_i)) - P(f(\hat{\mathbf{r}}_{i,l}))$, where $P(\cdot)$ denotes a used performance measure for ICD coding, e.g., classification accuracy, precision and recall, etc. Conversely, a highly comprehensive explanation should significantly impact the model prediction when the rationale $\hat{\mathbf{r}}_{i,l}$ is removed. This motivates the comprehensiveness metric that measures the prediction change when the rationale $\mathbf{r}_{i,l}$ is excluded from the input $\mathbf{x}_i$, computed as $\text{Comp} = P(f(\mathbf{x}_i)) - P(f(\mathbf{x}_i \backslash \hat{\mathbf{r}}_{i,l}))$. In general, lower sufficiency and higher comprehensiveness indicate better faithfulness.

Plausibility. A rationale is considered plausible when it highlights text rationales that are perceived by humans (e.g., domain experts) as relevant and appropriate to support model prediction. We assess plausibility by comparing the three types of rationales with human-annotated rationales, using the same matching metrics as in MDACE—exact / position-independent (PI) token matches (TM) and span matches (SM)—as in MDACE. To facilitate this, we construct an rationale dataset as below.

4.2 Rationale Dataset Construction

We introduce a new **r**ationale **d**ataset derived from MIMIC-**IV** and aligned with the ICD-**10** coding system - **RD-IV-10**, annotated by medical professionals. It includes detailed annotations capturing richer rationales supporting each code assignment, such as direct and indirect mentions, medica-

tions, and other pertinent clinical factors. Details of dataset construction, including Data Selection, Annotator, Annotation Guidelines, Annotation Platform, Inter-Annotator Agreement and Details of Data Processing are provided in Appendix C.

As shown in Table 1, the annotations in RD-IV-10 much closely match the original distribution of the ICD coding dataset compared to MDACE—93.15% and 83.88% in the Full and Top-50 settings, respectively, versus 37.00% and 14.59% in MDACE. We also report the specific labels for which no supporting rationale could be identified in Appendix C, offering further insight into the annotation quality of the MIMIC-IV ICD-10 dataset. Furthermore, our dataset provides richer rationale: it includes an average of 35.94 rationale spans for 14.82 labels per document, whereas MDACE contains only 14.10 spans for 11.57 labels. Figure 2 further presents code-level statistics, showing that the number of annotations far exceeds the code frequency in the RD-IV-10 dataset. This confirms our earlier observation that MDACE typically offers only a single piece of supporting rationale per code. In addition, our dataset features more comprehensive annotation formats across multiple levels of granularity, including words, phrases, and both complete and partial sentences. The average length of each rationale span is also greater—5.44 tokens compared to 2.13 tokens in MDACE. A detailed case study comparing the annotation quality of the two datasets is presented in Appendix E.

4.3 Explainability Evaluation

We focus on examining the faithfulness of rationales extracted based on attention weights for CAML, LAAT and PLM–ICD. These are **model-generated rationales** produced with only supervision from ICD coding labels. The overall architecture for testing faithfulness is illustrated in Figure 1. In evaluating plausibility, we also analyze two additional types of rationales. One is **naive entity-level rationales** derived from an existing entity linking dataset – SNOMED CT Entity Linking Challenge dataset (Hardman et al., 2023). This dataset links ICD codes to direct named-entities appeared in text, e.g., entities '*type 2 diabetes*', '*T2DM*', and '*Diabetes II*' are linked to ICD-10 code '*E11.9 – Type 2 diabetes mellitus without complications*'. The other is **strong LLM-generated rationales**. We prompt LLMs to extract rationale spans from patient notes that support specific ICD code assignments, and have observed that 2-Flash performs the best. Given that LLMs occasionally fail to produce spans that exactly match the original text, we design algorithms to align the generated spans back to the original document. Details of our prompt design and span alignment algorithm are provided in Appendix B.

Motivated by the strong plausibility of LLM-generated rationales, we incorporate supervision from rationale labels produced by the best-performing Gemini 2-Flash model, with the methodology detailed in the following section. Figure 3 illustrates representative examples of each rationale type.

5 LLM-Guided Rationale Learning

LLMs have demonstrated strong performance across a variety of tasks in the clinical domain. Also, we have observed from the previous rationale analysis that the LLM-generated rationales align quite well with human-annotated rationales. This motivates us to take advantage of LLMs, i.e., being able to quickly identify rationale with reasonable quality, and design LLM-guided rationale learning approaches. We propose to use rationales produced by prompting LLMs as distant supervision signal, aiming at maximizing simultaneously classification accuracy for ICD coding and plausibility of the model-generated rationales.

Multi-objective Learning One way to embed rational learning into ICD coding is to incorporate another learning objective alongside the primary classification objective of the ICD coding model, minimizing the discrepancy between the model-generated rationales (controlled by attention weights) and the provided rationale labels by LLMs. We define the rationale labels associated with a code label l as $\mathbf{r}_{il}$. To represent these rationales, we construct a binary mask matrix $\mathbf{M}_{i,l}$, where a value of 1 indicates that the corresponding token is part of the rationale, and 0 otherwise. We design the following rationale generation loss by applying binary cross-entropy to the attention weights and rationale masks:

$$\mathcal{L}_{\text{rationale}} = -\frac{1}{N_D}\sum_{i=1}^{N_D}\sum_{l=1}^{N_l}\sum_{j=1}^{N_t}[M_{i,l,j}\log\tilde{a}_{i,j,l} + (1-M_{i,l,j})\log(1-\tilde{a}_{i,j,l})]. \quad (2)$$

The final ICD coding model is trained by minimizing the combined loss of $\mathcal{L}_{\text{coding}} + \mathcal{L}_{\text{rationale}}$.

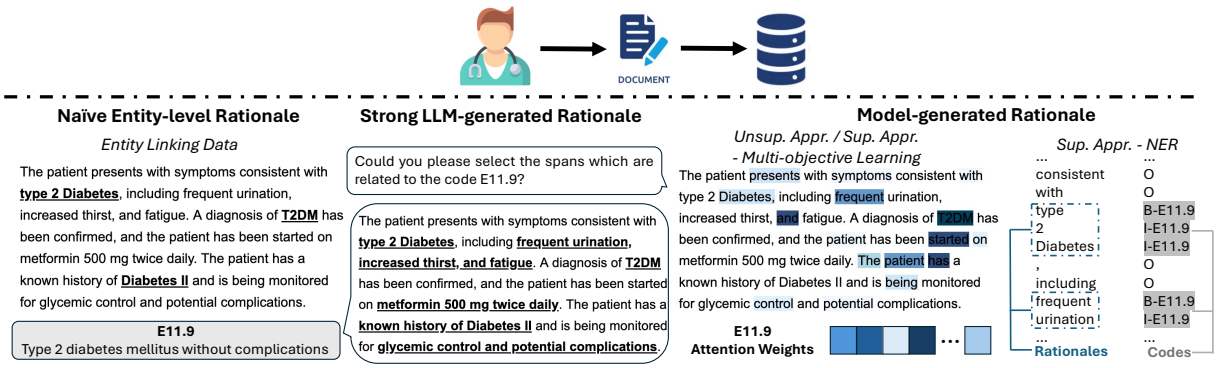

Figure 3: Examples of three types of rationales evaluated for plausibility. *Unsup. / Sup.* denote Unsupervised and Supervised, separately. *Appr.* indicates Approach.

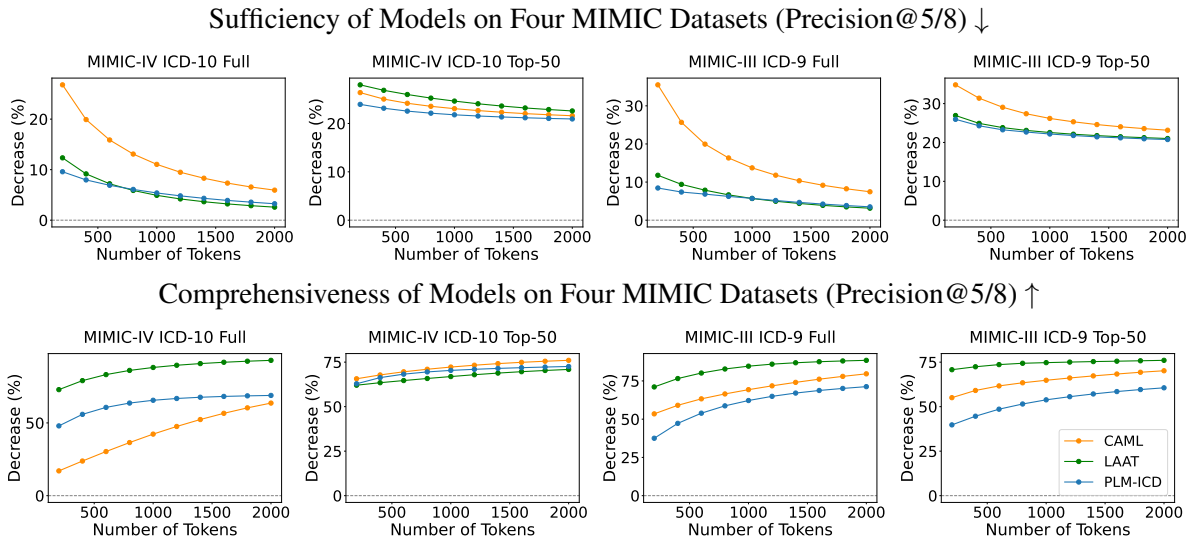

Figure 4: Faithfulness results of ICD coding models on four MIMIC datasets. The y-axis represents the decrease ratio, computed based on Precision@N scores (N = 8 for the Full set and N = 5 for the Top-50 set) as $(S_{orig} - S_{com/suff})/S_{orig} \times 100\%$, where S_{orig} denotes the performance obtained with the original input, and, $S_{com/suff}$ denotes the performance obtained when the input is modified by removing or retaining the rationales. The x-axis denotes the number of most-attended tokens selected. ↑ denotes higher is better; ↓ denotes lower is better.

Learning by NER Formulation An alternative approach to enable both rationale and ICD code learning is to leverage the rationale labels provided by LLMs to train a NER model. Specifically, each rationale is treated as an entity with its corresponding ICD code assigned as its class label, while the span that is not identified as a rationale is assigned a null class. The success of entity recognition contributes to both ICD code classification and rationale extraction. As a result, the ICD coding and rationale learning tasks are neatly converted to one single NER task, solved by following the standard NER training.

Enhanced Supervision by Few-shot Prompting Manual rationale annotation is time-consuming and costly, limiting scalability to large datasets. Although LLMs provide a promising alternative for automatic annotation, they are susceptible to hallucinations and inaccuracies (Li et al., 2023; Ji et al., 2023), particularly in expert domains such as healthcare (Nagar et al., 2024). Prior studies have shown that few-shot prompting, where models are provided with few examples, can substantially improve generation quality (Sivarajkumar et al., 2024). Motivated by this, we further incorporate a small amount of example annotations provided by our constructed rationale dataset into the prompts of Gemini 2-Flash. Details of the prompt design are provided in Appendix B. The rationales generated are then used to supervise the rationale learning.

6 Experiments and Result Analysis

We conduct evaluation using both the MIMIC-III dataset with ICD-9 codes and the more recent

MIMIC-IV benchmark with ICD-10 codes. To assess the ICD coding performance, we use F1, AUC, and Precision@N, following the standard practice. To conduct experiments on plausibility, we use the SNOMED CT Entity Linking Challenge dataset (Hardman et al., 2023) to implement the naive rationale extraction, i.e., entity-level rationales. We compare both cloud-based and locally deployed LLMs, including Gemini 2-Flash, Gemini 1.5-Pro and LLaMA-3.3, to implement the strong rationale LLM-generated rationales. To ensure a fair comparison between the three models of CAML, LAAT and PLM-ICD, we follow the experimental setup of a reproducibility framework Edin et al. (2023). We implement NER models under the default hyperparameter settings in Yang et al. (2020). More details on datasets and implementations are provided in Appendices A and B.

6.1 Comparison of ICD Rationales

6.1.1 Faithfulness of CAML, LAAT and PLM-ICD: *How Explainable Are Rationales to Machines?*

Figure 4 compares the faithfulness of the three ICD coding models with *Top N tokens* rationale selection strategy. Results are reported in terms of precision@5 for the Top-50 datasets and precision@8 for the Full datasets. More results across more metrics with both *Top N tokens* and *Top p% tokens* are reported in Appendix F.

For sufficiency, *PLM-ICD performs comparably to LAAT, and both models outperform CAML* on the MIMIC-IV ICD-10 Full dataset as well as the two MIMIC-III ICD-9 datasets. PLM-ICD achieves the highest sufficiency using only 200 tokens, with a 10% and 25% performance drop on the Top-50 and Full datasets, respectively. These tokens represent just 11–16% of the input text. *For comprehensiveness, removing rationales causes the greatest performance drop in LAAT*, particularly on the MIMIC-IV ICD-10 Full and two MIMIC-III ICD-9 datasets. In contrast, PLM-ICD shows smaller declines, even less than CAML on the ICD-9 datasets, suggesting that its pre-training enhances robustness by effectively extracting relevant information from residual inputs.

6.1.2 Plausibility of Naive, Strong and Model-generated Rationales: *How Explainable Are Rationales to Experts?*

Table 2 summarises the plausibility results across different types of rationales. The results show that model-generated rationales yield very low metric scores, which indicates that they do not align with human explanations. Additional results across a wider range of thresholds are provided in Appendix F. Entity-level rationales rank second, while LLM-generated rationales perform the best, with Gemini 2-Flash achieving the highest scores. An additional case study comparing these rationales with human annotations is provided in Appendix H.

Table 2: Document-level and code-level plausibility results (F1) of naive entity-level rationales, strong LLM-generated rationales, model-generated rationales. 200 denotes selected tokens, "w/o" and "w/" indicate the absence or inclusion of few-shot examples in prompts. All results are reported as percentages.

Model / Code	Settings	Exact SM	PI SM	Exact TM	PI TM
Document-level Evaluation					
Entity-Linking	–	10.3	9.2	6.3	6.1
Gemini 2-Flash	–	**21.6**	**24.1**	**30.1**	**37.3**
Gemini 1.5-Pro	–	13.5	14.6	20.6	26.0
LLaMA-3.3 Ins	–	18.6	21.5	27.8	35.0
LLaMA-3.3 AWQ	–	17.5	20.1	27.2	34.1
CAML	200	0.1	0.2	3.1	6.5
LAAT	200	0.7	0.8	5.0	7.2
PLM-ICD	200	0.5	0.7	4.3	8.8
Code-level Evaluation (Gemini 2-Flash)					
I10	w/o	50.3	63.1	27.0	32.2
	w/	**60.5**	**78.2**	**53.8**	**66.4**
E785	w/o	62.5	76.6	50.2	59.7
	w/	**73.2**	**89.5**	**66.7**	**78.2**
Z7901	w/o	9.0	9.7	35.8	43.3
	w/	**13.0**	**16.8**	**45.8**	**54.8**
I4891	w/o	11.8	16.7	28.0	34.8
	w/	**24.7**	**36.9**	**47.9**	**56.2**
E119	w/o	40.8	48.8	36.2	37.4
	w/	**44.2**	**52.9**	**43.6**	**44.8**

Naive Rationales: *Are Linked Entities Sufficient to Serve as Rationales?* The performance of directly linked entities ranks in the middle among the three types of rationales. However, its actual matching quality is underestimated, as the entity linking and MIMIC-IV ICD-10 dataset use different coding schemes despite referring to the same clinical mentions. For instance, in sample HADM ID: 24813967, all occurrences of '*fall*' are assigned '*R29.6 – Repeated falls*', whereas MIMIC-IV ICD-10 labels the case with '*W01.0XXA – Fall on same level from slipping, tripping and stumbling without subsequent striking against object, initial encounter*'. *In conclusion, directly linked entities can serve as rationales to a certain extent.*

Strong Rationales: *Can LLMs Generate High-Quality Rationales?* *The Gemini 2-Flash model achieves the highest performance among all comparisons.* However, both more cost-effective local LLaMA-3.3 models deliver competitive re-

sults, with only a minor drop observed in the quantized variant AWQ. Importantly, the AWQ requires significantly fewer computational resources-approximately 40 GB of memory, compared to the Instruct model.

Table 3: ICD coding performance of LLM-guided supervised approaches. The experiments are conducted on the Top-50 code settings. All results are reported as percentages.

Model	F1-Mac	F1-Mic	P-Mac	P-Mic	R-Mac	R-Mic
PLM-ICD	68.18	73.40	68.71	73.48	69.38	73.33
Multi-objective	67.93	73.26	67.60	72.66	69.53	73.87
PLM-ICD	61.09	68.18	60.06	67.62	63.90	68.76
NER	53.46	67.75	49.21	60.93	61.79	76.30

6.1.3 Few-shot Prompting: *Does It Improve Plausibility of LLM-Generated Rationales?*

In the code-level evaluation, we conduct experiments on the five most frequent codes in the dataset: I10 (essential (primary) hypertension), E785 (hyperlipidemia, unspecified), Z7901 (long term (current) use of anticoagulants), I4891 (unspecified atrial fibrillation (AFib)), and E119 (type 2 diabetes mellitus without complications). We analyze rationales generated by Gemini 2-Flash with and without few-shot examples in the prompts to examine whether incorporating human-annotated examples can further enhance Gemini's performance. In the few-shot experiments, examples from five test instances per code are included in the prompts, and Gemini then regenerates rationales for the remaining samples, which are then re-evaluated using the same plausibility metrics. *Incorporating examples yields substantial improvements in F1 scores* across all five codes shown in Table 2, with average gains of 39.90%, 48.67%, 50.31%, and 49.01% in Exact Span Match, PI Span Match, Exact Token Match, and PI Token Match, respectively. These results demonstrate that examples from our rationale dataset guide Gemini in generating more plausible rationales. More results for this experiment are provided in Appendix G.

6.2 Results for LLM-Guided Rationale Learning

6.2.1 Standard Prompting: *Does It Improve ICD Coding Performance and Rationale Plausibility?*

Supervised by weak rationale labels generated by Gemini 2-Flash, the multi-objective learning approach–which introduces the additional rationale-targeted objective–does not degrade ICD coding performance (Table 3); instead, it improves plausibility by approximately 1% (F1) across all four metrics (Table 4).

Table 4: Document-level and code-level plausibility results (F1) of LLM-guided supervised approaches. The experiments for multi-objective learning and NER are conducted on different datasets using the Top-50 code settings. 50 denotes selected tokens, "w/o" and "w/" indicate the absence or inclusion of few-shot examples in prompts. **when constructing the training datasets**. All results are reported as percentages.

Model	Settings	Exact SM	PI SM	Exact TM	PI TM
Document-level Evaluation					
PLM-ICD	50	2.7	3.0	9.5	12.2
Multi-objective	50	**3.9**	**4.1**	**10.6**	**13.2**
PLM-ICD	50	4.1	4.3	8.1	12.0
Gemini 2-Flash	–	18.2	23.0	**29.4**	**31.4**
NER	–	**26.5**	**30.6**	21.8	27.0
Code-level Evaluation (NER)					
I10	w/o	55.4	76.9	55.8	75.3
	w/	**62.5**	**85.2**	**64.9**	**86.2**
E785	w/o	67.9	85.5	61.6	75.2
	w/	**70.3**	**87.0**	**63.0**	**76.3**
Z7901	w/o	10.3	**13.8**	**25.5**	38.6
	w/	**10.8**	12.0	22.5	**40.7**
I4891	w/o	22.2	37.7	38.5	48.8
	w/	**26.1**	**43.3**	**40.7**	**54.5**
E119	w/o	**49.4**	**62.0**	**53.2**	**62.4**
	w/	45.8	58.5	49.7	60.2

For the NER-based approach, which adopts a learning formulation distinct from PLM-ICD, there exists a clear trade-off between prediction accuracy and rationale plausibility, e.g., 12.49% lower in coding performance (F1-macro) but on average 363% higher plausibility than PLM-ICD. Notably it achieves the highest span-level plausibility, even surpassing its teacher model, Gemini 2-Flash. This trade-off arises because the NER model is specifically designed for entity recognition. It is trained on rationale–code pairs, focusing on highlighting tokens and assigning labels for each token, rather than to maximise classification accuracy, whereas PLM-ICD is trained on full documents, focusing on optimising document-level classification accuracy. Its competing objectives can reduce its raw predictive performance. However, the NER model benefits from learning entity patterns across all training samples, where these LLM generated training samples can be unstable and may miss some spans. This enables NER to consistently recognize complete instances of relevant entities, and this is why the NER model surpasses its teacher model, Gemini 2-Flash, in plausibility. Moreover, it generates stable outputs for all labels simultane-

ously, whereas Gemini is prompted with one code at a time; providing the entire code set leads to incomplete recognition. However, since the NER model is designed to identify rationales for a set of ICD codes rather than to perform coding directly, it exhibits substantially lower overall coding performance compared to PLM-ICD. Additionally, the NER-based approach offers an alternative, cost-free method for generating rationales. The results presented correspond to the *top 50 tokens*. Additional results under a wider range of experimental settings are provided in Appendix G. Further results for the NER-based approach across different training data sizes are presented in Appendix I.

6.2.2 Few-shot Prompting: *Does It Improve LLM-Guided Rationale Learning?*

This experiment further investigates whether the enhanced rationales generated through few-shot prompting can facilitate the rationale learning process. Specifically, we train separate NER models using Gemini-generated rationales, both with and without few-shot examples in the prompts, for each of the five most frequent codes.

Table 4 shows that the *NER formulation is effective for rationale recognition for specific single code*. It significantly outperforms the teacher models (w/o results in Table 2) across all codes and metrics, achieving average improvements of 28.49%, 45.71%, 37.01%, and 51.21% on the four metrics, respectively. Among all codes, E785 and I10 achieve higher performance, which is attributed to their higher frequencies, providing the model with richer supervision signals. Furthermore, *a model trained on data generated with few-shot examples in the prompts consistently outperforms one trained without such examples in most cases*, with average improvements of 6.30%, 1.74%, 1.19%, and 5.91%. These results further demonstrate that the enhanced rationales generated through few-shot prompting using examples from our dataset further enhance the training of rationale learning models. See Appendix G for complete results across all metrics.

7 Conclusion

In this study, we introduce a rationale dataset specifically designed for ICD coding. We evaluate the faithfulness of coding models and the plausibility of three types of rationales, among which Gemini 2-Flash achieves the best performance. We examine LLM-guided rationale learning approaches, where the NER formulation demonstrates strong potential, as both coding and rationale extraction tasks can be unified under a single NER framework. This approach achieves the highest span-level plausibility. Moreover, incorporating human-annotated examples from our dataset into prompts enhances both rationale generation and rationale learning process.

8 Limitations

While our study contributes to advancing both rationale evaluation and rationale learning in ICD coding area by providing benchmark resource and empirical analyses, it also has several limitations that suggest directions for future research.

First, annotation is a resource-intensive process that demands domain expertise as well as significant time and cost investment. Our dataset currently comprises 150 samples, constrained by budget limitations. With additional funding or institutional support, the scale of annotations could be expanded in future work. With a sufficient number of human-annotated rationale labels, supervision using these labels becomes feasible. This enables a direct comparison between models trained with real labels and those trained with weak labels. Additionally, these 150 samples cover only 989 distinct codes, whereas MIMIC-IV contains 7,942 codes. Increasing the number of samples would improve the diversity of codes.

Second, we evaluate three classic attention-based ICD coding models—CAML, LAAT, and PLM-ICD. There are also many other models that incorporate label-wise attention layers, such as the recent CoRelation (Luo et al., 2024) and MSAM (Gomes et al., 2024). However, for plausibility evaluation, the rationales generated by these attention-based models tend to have low plausibility.

Third, the evaluation of naïve entity-level rationales could be further refined through a more precise alignment of coding schemes.

Finally, our NER models were trained on relatively small datasets of 5,000 randomly selected samples. Future work could extend these experiments to the full datasets.

Acknowledgement

This research is part of the IN-CYPHER programme and is supported by the National Research Foundation, Prime Minister's Office, Sin- gapore under its Campus for Research Excellence and Technological Enterprise (CREATE) programme.

We are grateful for the support provided by Research IT in the form of access to the Computational Shared Facility at The University of Manchester. We also thank the anony- mous ARR reviewers for their feedback that helped us improve the paper further.

References

Julia Amann, Alessandro Blasimme, Effy Vayena, Dietmar Frey, Vince I Madai, and Precise4Q Consortium. 2020. Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. *BMC medical informatics and decision making*, 20:1–9.

James Blundell. 2023. Health information and the importance of clinical coding. *Anaesthesia & Intensive Care Medicine*, 24(2):96–98.

Finneas Catling, Georgios P Spithourakis, and Sebastian Riedel. 2018. Towards automated clinical coding. *International journal of medical informatics*, 120:50–61.

Hua Cheng, Rana Jafari, April Russell, Russell Klopfer, Edmond Lu, Benjamin Striner, and Matthew R Gormley. 2023. Mdace: Mimic documents annotated with code evidence. *arXiv preprint arXiv:2307.03859*.

Koby Crammer, Mark Dredze, Kuzman Ganchev, Partha Talukdar, and Steven Carroll. 2007. Automatic code assignment to medical text. In *Biological, translational, and clinical language processing*, pages 129–136.

Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C Wallace. 2019. Eraser: A benchmark to evaluate rationalized nlp models. *arXiv preprint arXiv:1911.03429*.

Hang Dong, Víctor Suárez-Paniagua, William Whiteley, and Honghan Wu. 2021. Explainable automated coding of clinical notes using hierarchical label-wise attention networks and label embedding initialisation. *Journal of biomedical informatics*, 116:103728.

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 D Havtorn, and Tuukka Ruotsalo. 2024. An unsupervised approach to achieve supervised-level explainability in healthcare records. *arXiv preprint arXiv:2406.08958*.

Yue Gao, Yuepeng Chen, Minghao Wang, Jinge Wu, Yunsoo Kim, Kaiyin Zhou, Miao Li, Xien Liu, Xiangling Fu, Ji Wu, and 1 others. 2024. Optimising the paradigms of human ai collaborative clinical coding. *npj Digital Medicine*, 7(1):368.

Gonçalo Gomes, Isabel Coutinho, and Bruno Martins. 2024. Accurate and well-calibrated icd code assignment through attention over diverse label embeddings. *arXiv preprint arXiv:2402.03172*.

Will Hardman, Mark Banks, Rory Davidson, Donna Truran, Nindya Widita Ayuningtyas, Hoa Ngo, Alistair Johnson, and Tom Pollard. 2023. Snomed ct entity linking challenge. *PhysioNet. Version*, 1(0).

Chao-Wei Huang, Shang-Chi Tsai, and Yun-Nung Chen. 2022. Plm-icd: Automatic icd coding with pretrained language models. *arXiv preprint arXiv:2207.05289*.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM computing surveys*, 55(12):1–38.

Alistair Johnson, Lucas Bulgarelli, Tom Pollard, Steven Horng, Leo Anthony Celi, and Roger Mark. 2020. Mimic-iv. *PhysioNet. Available online at: https://physionet. org/content/mimiciv/1.0/(accessed August 23, 2021)*, pages 49–55.

Sarvnaz Karimi, Xiang Dai, Hamed Hassanzadeh, and Anthony Nguyen. 2017. Automatic diagnosis coding of radiology reports: a comparison of deep learning and conventional classification methods. In *BioNLP 2017*, pages 328–332.

Byung-Hak Kim, Zhongfen Deng, Philip S Yu, and Varun Ganapathi. 2022. Can current explainability help provide references in clinical notes to support humans annotate medical codes? *arXiv preprint arXiv:2210.15882*.

Junyi Li, Xiaoxue Cheng, Wayne Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2023. Halueval: A large-scale hallucination evaluation benchmark for large language models. *arXiv preprint arXiv:2305.11747*.

Lucian Vlad Lita, Shipeng Yu, Stefan Niculescu, and Jinbo Bi. 2008. Large scale diagnostic code classification for medical patient records. In *Proceedings of the Third International Joint Conference on Natural Language Processing: Volume-II*.

Yang Liu, Hua Cheng, Russell Klopfer, Matthew R Gormley, and Thomas Schaaf. 2021. Effective convolutional attention network for multi-label clinical document classification. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5941–5953.

Justin Lovelace, Nathan C Hurley, Adrian D Haimovich, and Bobak J Mortazavi. 2020. Dynamically extracting outcome-specific problem lists from clinical notes with guided multi-headed attention. In *Machine Learning for Healthcare Conference*, pages 245–270. PMLR.

Junyu Luo, Xiaochen Wang, Jiaqi Wang, Aofei Chang, Yaqing Wang, and Fenglong Ma. 2024. Corelation: Boosting automatic icd coding through contextualized code relation learning. *arXiv preprint arXiv:2402.15700*.

Erick Mendez Guzman, Viktor Schlegel, and Riza Batista-Navarro. 2024. From outputs to insights: a survey of rationalization approaches for explainable text classification. *Frontiers in Artificial Intelligence*, 7:1363531.

George Michalopoulos, Michal Malyska, Nicola Sahar, Alexander Wong, and Helen Chen. 2022. Icdbigbird: a contextual embedding model for icd code classification. *arXiv preprint arXiv:2204.10408*.

James Mullenbach, Sarah Wiegreffe, Jon Duke, Jimeng Sun, and Jacob Eisenstein. 2018. Explainable prediction of medical codes from clinical text. *arXiv preprint arXiv:1802.05695*.

Aishik Nagar, Yutong Liu, Andy T Liu, Viktor Schlegel, Vijay Prakash Dwivedi, Arun-Kumar Kaliya-Perumal, Guna Pratheep Kalanchiam, Yili Tang, and Robby T Tan. 2024. umedsum: A unified framework for advancing medical abstractive summarization. *arXiv preprint arXiv:2408.12095*.

Anthony N Nguyen, Donna Truran, Madonna Kemp, Bevan Koopman, David Conlan, John O'Dwyer, Ming Zhang, Sarvnaz Karimi, Hamed Hassanzadeh, Michael J Lawley, and 1 others. 2018. Computer-assisted diagnostic coding: effectiveness of an nlp-based approach using snomed ct to icd-10 mappings. In *AMIA Annual Symposium Proceedings*, volume 2018, page 807. American Medical Informatics Association.

Suzanne Pereira, Aurélie Névéol, Philippe Massari, Michel Joubert, and Stefan Darmoni. 2006. Construction of a semi-automated icd-10 coding help system to optimize medical and economic coding. In *MIE*, pages 845–850.

Sonish Sivarajkumar, Mark Kelley, Alyssa Samolyk-Mazzanti, Shyam Visweswaran, and Yanshan Wang. 2024. An empirical evaluation of prompting strategies for large language models in zero-shot clinical natural language processing: algorithm development and validation study. *JMIR Medical Informatics*, 12:e55318.

SNOMED-CT. 2024. Snomed ct u.s. edition release, september 1, 2024. https://www.nlm.nih.gov/healthit/snomedct/us_edition.html.

Dimitri Tzitzivacos. 2007. International classification of diseases 10th edition (icd-10). *CME: Your SA Journal of CPD*, 25(1):8–10.

Betty Van Aken, Jens-Michalis Papaioannou, Marcel G Naik, Georgios Eleftheriadis, Wolfgang Nejdl, Felix A Gers, and Alexander Löser. 2022. This patient looks like that patient: Prototypical networks for interpretable diagnosis prediction from clinical text. *arXiv preprint arXiv:2210.08500*.

Thanh Vu, Dat Quoc Nguyen, and Anthony Nguyen. 2020. A label attention model for icd coding from clinical text. *arXiv preprint arXiv:2007.06351*.

Tao Wang, Linhai Zhang, Chenchen Ye, Junxi Liu, and Deyu Zhou. 2022. A novel framework based on medical concept driven attention for explainable medical code prediction via external knowledge. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1407–1416.

Xi Yang, Jiang Bian, William R Hogan, and Yonghui Wu. 2020. Clinical concept extraction using transformers. *Journal of the American Medical Informatics Association*, 27(12):1935–1942.

Zhichao Yang, Shufan Wang, Bhanu Pratap Singh Rawat, Avijit Mitra, and Hong Yu. 2022. Knowledge injected prompt based fine-tuning for multi-label few-shot icd coding. In *Proceedings of the conference on empirical methods in natural language processing. Conference on empirical methods in natural language processing*, volume 2022, page 1767.

Vithya Yogarajan, Bernhard Pfahringer, Tony Smith, and Jacob Montiel. 2022. Concatenating biomed-transformers to tackle long medical documents and to improve the prediction of tail-end labels. In *International Conference on Artificial Neural Networks*, pages 209–221. Springer.

Zheng Yuan, Chuanqi Tan, and Songfang Huang. 2022. Code synonyms do matter: Multiple synonyms matching network for automatic icd coding. *arXiv preprint arXiv:2203.01515*.

A Additional Information on Datasets and Implementation

A.1 Datasets

MIMIC Datasets The Medical Information Mart for Intensive Care (MIMIC) dataset is a large-scale, de-identified database comprising health records of patients admitted to the emergency department or intensive care units at the Beth Israel Deaconess Medical Center (Johnson et al., 2020). The MIMIC-IV dataset covers over 65,000 ICU admissions and more than 200,000 emergency department visits between 2008 and 2019, coded using ICD-10. The MIMIC-III dataset covers admissions between 2001 and 2012, including 52,723 discharge summaries from 41,126 patients, coded using ICD-9. Table 5 presents the statistics of dataset splits for all datasets used in model training, where "Ra" refers to the subset of data comprising documents with rationale labels generated by Gemini 2-Flash and it is used for training the multi-objective learning model. Top-50 subsets include only the 50 most frequent codes from the respective Full datasets.

Table 5: Dataset Split.

Dataset	Train	Test	Dev
MIMIC-IV ICD-10 Full	88988	19931	13360
MIMIC-IV ICD-10 Top-50	83890	18776	12590
MIMIC-III ICD-9 Full	47719	3372	1631
MIMIC-III ICD-9 Top-50	8066	1729	1573
MIMIC-IV ICD-10 Top-50 Ra	83465	18665	12517

Entity Linking Dataset The SNOMED CT Entity Linking Challenge dataset (Hardman et al., 2023) comprises 272 discharge summaries from MIMIC-IV-Note, annotated with 6,624 unique SNOMED-CT concepts. Among these, 64 documents overlap with the MIMIC-IV ICD-10 dataset. To enable comparison, we align SNOMED CT concepts with ICD-10 codes using established mapping resource (SNOMED-CT, 2024).

Access to MIMIC datasets and entity linking dataset is granted upon completion of human-subjects research training (e.g., CITI program), registration on the PhysioNet platform, acceptance of the dataset's Data Use Agreement, and subsequent retrieval of the data via PhysioNet's web interface or command-line tools.

A.2 Implementation Details

On CAML, LAAT and PLM-ICD To ensure a fair comparison between models, we follow the experimental setup of Edin et al. (2023), which provides a reproducibility framework for state-of-the-art ICD coding models. Details of the key parameter configurations for the three ICD coding models are summarized in Table 5. All models are trained for 20 epochs, although CAML and LAAT typically converge within 10 epochs on MIMIC-IV ICD-10 and both MIMIC-III datasets. The random seed is set to 1337 for all model training.

On Evaluation During the faithfulness testing, if the whole document contains fewer than the threshold N tokens, we include all tokens. To evaluate comprehensiveness, if all tokens are removed, the single token with the lowest attention weight is retained.

On NER We implement the NER models using the default hyperparameter settings provided by Yang et al. (2020), with the parameter configuration summarized in Table 6. All NER models are trained on 5,000 randomly selected samples from the MIMIC-IV Top-50 dataset. Document-level plausibility evaluation is conducted on 139 annotated documents, filtered from an initial set of 150 samples to retain only those associated with the Top-50 codes. For code-level plausibility evaluation, the test set consists of the remaining samples, excluding those 5 samples used as examples in few-shot prompting. The statistics of the original and test sets are shown in Table 7.

B On LLM-Generated Rationales

B.1 Prompting Without Examples

The used prompt without examples follows the format below. In the following, the ***text***, ***code***, and ***description of the code*** are variables that change based on the input note and the target code.

Note Text: ***text*** + Code: ***code***. Description: ***description of the code***. Could you please select the spans (rationales) which are related to the code ***code***? The spans can be words, phrases, or sentences. Only list the exact spans extracted from the 'Note Text', without including their section names. List each span with a number in front. For example: '1. Span1 2. Span2'. Only keep the spans. Do not include any additional responses. Exclude any punctuations at the end of the spans. Keep the spans as what they are in 'Note Text'. Keep the spans as what they are in 'Note Text'. Keep the spans as what they are in the 'Note Text'.

Table 5: Configurations of CAML, LAAT and PLM-ICD.

Parameter	MIMIC-IV ICD-10 Full	MIMIC-IV ICD-10 Top-50	MIMIC-III ICD-9 Full	MIMIC-III ICD-9 Top-50
CAML				
batch size	8	8	8	8
learning rate	5×10^{-3}	5×10^{-3}	10^{-4}	10^{-4}
weight decay	10^{-3}	10^{-3}	-	-
LAAT				
batch size	8	8	8	8
learning rate	10^{-3}	10^{-3}	10^{-3}	10^{-3}
weight decay	10^{-3}	10^{-3}	-	-
PLM-ICD				
batch size	16	16	8	8
learning rate	5×10^{-5}	5×10^{-5}	5×10^{-5}	5×10^{-5}
weight decay	0	0	-	-

Table 6: Configurations of NER models.

Parameter	Value
batch size	8
learning rate	10^{-5}
warmup ratio	0.01
truncation	256
gradient accumulation steps	1
epoch	20
random seed	13

Table 7: Statistics of the test sets used for evaluating code-level NER models.

Code	Test Set / Full Set
I10	55 / 60
E785	50 / 55
Z7901	17 / 22
I4891	11 / 16
E119	17 / 22

B.2 Prompting With Few-Shot Examples

The used prompt with examples follows the format below. In the following,the variable ***examples*** represents annotation examples corresponding to each ***code***.

Note Text: ***text*** + Code: ***code***. Description: ***description of the code***. Could you please select the spans (rationales) which are related to the code ***code***? The spans can be words, phrases, or sentences. Only list the exact spans extracted from the 'Note Text', without including their section names. For example: ***examples***. List each span with a number in front. For example: '1. Span1 2. Span2'. Only keep the spans. Do not include any additional responses. Exclude any punctuations at the end of the spans. Keep the spans as what they are in 'Note Text'. Keep the spans as what they are in 'Note Text'. Keep the spans as what they are in the 'Note Text'.

The examples are drawn from five randomly selected documents containing the given code in the annotation dataset. Table 8 summarizes the samples used for few-shot prompting and the corresponding ***examples*** value for each code.

B.3 LLM Configurations

As the PhysioNet Credentialed Data Use Agreement prohibits sharing MIMIC data with external services such as ChatGPT, we follow PhysioNet's recommendation to use Google Gemini, which does not utilize user prompts or responses for model training. We employ two variants of Gemini 2-Flash and 1.5-Pro. For local deployment, we select LLaMA-3.3, one of the most capable open-weight LLMs available, examining both the 70B Instruct and its quantized variant AWQ. The local LLMs are executed on 4 × and 1 × NVIDIA A100 80GB GPUs, respectively.

For the Gemini variants, we configured both models with a temperature of 0.1 to minimize randomness and encourage more deterministic outputs, as our task requires selecting spans that exactly match those appearing in the input documents. We also set top_p to 0.99.

For the LLaMA variants, we adopted the same parameter settings as Gemini. Additionally, we set max_tokens to 8,000, which exceeds the length of the longest documents in our dataset. This configuration, however, caused the computation for the larger LLaMA model to exceed the default memory capacity of two NVIDIA A100 80GB GPUs.

Table 8: Samples used for few-shot prompting and their annotations.

Code	Description	HADM IDs	Annotations
I10	Essential (primary) hypertension	21893270; 20961577; 20272030; 23048750; 29161744	'HTN', 'Hypertension', 'HYPERTENSION - ESSENTIAL, UNSPEC', 'Essential hypertension', 'hypertension', 'HYPERTENSION'
E785	Hyperlipidemia, unspecified	21893270; 26102343; 20272030; 24014389; 27049443	'HLD', 'Dyslipidemia', 'Hyperlipidemia', 'Hypercholesteremia', 'Dyslipidemia', 'hyperlipidemia'
Z7901	Long term (current) use of anticoagulants	27049443; 29964986; 27021287; 25097869; 29155448	'aspirin', 'Plavix', 'Coumadin', 'Afib on coumadin', "and the decision was made to restart the patient's ASA and Plavix immediately postoperatively", 'Clopidogrel 75 mg PO DAILY', 'Aspirin 325 mg PO DAILY', 'Warfarin', 'on Coumadi', 'Coumadin who presents with dyspnea', 'on coumadin', 'Warfarin held for supratherapeutic INR (INR 3.4)', 'afib on warfarin', 'On coumadin.', 'Warfarin 2 mg PO 1X/WEEK', 'Warfarin 5 mg PO 6X/WEEK', 'Warfarin 2 mg PO 1X/WEEK', 'on high-dose warfarin due to resistance', 'warfarin (5 mg)', 'heparin gtt', 'Aspirin 81 mg PO DAILY', 'Enoxaparin Sodium 140 mg', 'Aspirin 81 mg PO DAILY'
I4891	Unspecified atrial fibrillation	27049443; 24257587; 29964986; 27466246; 29588477	'atrial fibrillation', 'atrial fibrillation', 'Troponinemia', 'Atrial Fibrillation', 'Digoxin', 'afib', 'Afib on coumadin', 'afib s/p', 'Atrial fibrillation', 'Afib s/p ablation', 'Afib'
E119	Type 2 diabetes mellitus without complications	21893270; 27904530; 24257587; 25097869; 22473872	'diabetes', 'DM', 'Diabetes', 'DM2', 'BLOOD Glucose-113', 'Additionally, Diabetes service was consulted for newly found hyperglycemia.', 'insulin dependent DM', 'Diabetes: his hemoglobin A1c was 7.8.', 'type 2 diabetes mellitus'

B.4 Span Alignment for Rationale Generation

LLMs occasionally fail to consistently reproduce spans that exactly match those in the original patient notes, despite our explicit instruction emphasized in the prompt '*Keep the spans as they appear in the 'Note Text'.*' During the prompt engineering, we observe that repeating this instruction three times improve the consistency to some extent, but the generated spans are still not perfectly aligned. To address this issue and enable accurate evaluation against human annotations afterwords, we developed a post-processing method to map the LLM-generated spans back to the original text. Specifically, we first identify all candidate spans. We narrow the search window within the original text and extract all spans that share the same initial and terminal n characters as the generated rationale. We then compute their overlap scores to select the best match.

Overlap Score Calculation Let S_{g} and S_{c} be the generated rationale and candidate target span extracted from the document, respectively. $T_{\mathrm{g}} = \mathrm{Tokenize}(S_{\mathrm{g}})$, and $T_{\mathrm{c}} = \mathrm{Tokenize}(S_{\mathrm{t}})$, where `Tokenize` returns a set of tokens. The overlap score S is calculated by the following equation:

$$S = \frac{T_{\mathrm{g}} \cap T_{\mathrm{c}}}{|T_{\mathrm{g}}|} + \frac{T_{\mathrm{g}} \cap T_{\mathrm{c}}}{|T_{\mathrm{c}}|} \tag{3}$$

We present in Table 5 an example of overlap score calculation for a set of candidate spans corresponding to the generated span '***diagnosis type 2 diabetes***', which shares the same initial character 'd' and terminal character 's' (with a character window size of $n = 1$). The overlap score consists of two components: (1) the overlap ratio with the generated rationale, which captures the degree of alignment with the generated rationale (e.g., '*diagnosis type 2 diabetes*' (1) is preferred over '*diagnosis diabetes*' (0.5)), and (2) the overlap ratio with the candidate span, which reflects the degree of alignment with the candidate span in the original text (e.g., '*diagnosis diabetes*' (1) is preferred over '*diagnosis tuberculosis*' (0.25)).

Rationale Selection We repeat this process across a range of character window sizes and select the candidate with the highest overlap score as the final mapping result. This procedure is applied to all generated spans, and those with an overlap score exceeding 1.7 are retained as supporting rationale for the document. The detailed steps of this mapping process are provided in the accompanying pseudocodes. Statistics for the Gemini 2-Flash–generated dataset across different span alignments, covering 122,004 samples from MIMIC-IV, are provided in Appendix J. Additional results on the plausibility of Gemini 2-Flash–generated rationales across different span alignments are presented in Appendix K. The impact of label selection under different span alignments during rationale learning is presented in Appendix L.

Table 5: Candidates targets of a generated rationale **diagnosis type 2 diabetes**. The decomposition consists of the two components of the overlap score function.

Candidate	Decomposition	Overlap Score
diagnosis	0.25 + 1	1.25
diabetes	0.25 + 1	1.25
diagnosis tuberculosis	0.25 + 0.5	0.75
dyslipidemias	0 + 0	0
diagnosis diabetes	0.5 + 1	1.5
diagnosis type 2 diabetes	1 + 1	2

C More Details on Dataset Construction

Data Selection To build a comprehensive comparison that includes Entity Linking data, we first identify the overlap between the Entity Linking and the MIMIC-IV ICD-10 datasets. There are 64 overlapping samples between the two datasets. We then randomly select an additional 86 samples from the MIMIC-IV ICD-10 test set, resulting in a final dataset of 150 samples.

Annotator Two annotators with medical background contributed to this annotation work. Annotator 1 holds a bachelor's degree in medicine, and Annotator 2 holds a master's degree in medicine. Annotator 1 annotated all samples, while Annotator 2 performed a quality check by annotating a subset of codes in 13 samples.

Payment The two annotators were paid £10 per document and an additional £15 for setting up the annotation platform.

Annotation Guidelines In this subsection, we outline the guidelines developed for the annotators.

Title. Identifying Rationales of ICD Codes in Discharge Summaries

Purpose. Explainability is especially critical in the clinical domain, where transparent and well-supported rationales enable healthcare providers and decision-makers to confidently utilize model predictions in patient care. The data you annotate

Algorithm 1 `OverlapScore`: Compute Overlap Score Between Two Spans

Input: Generated span $generated\_span$, Candidate span $candidate\_span$
Output: Overlap score between the two spans

$T_g \leftarrow \texttt{Tokenize}(generated\_span)$
$T_t \leftarrow \texttt{Tokenize}(candidate\_span)$
$I \leftarrow T_g \cap T_t$
$score \leftarrow \frac{|I|}{|T_g|} + \frac{|I|}{|T_t|}$
return $score$
Notes: `Tokenize` *function splits a text span into individual tokens.*

Algorithm 2 `BestCandidate`: Find Best Matching Candidate Span

Input: Generated span $generated\_span$, Document $note$
Output: Best candidate and its overlap score

$best\_candidate \leftarrow$ ”
$max\_score \leftarrow 0$
$candidates \leftarrow [\ ]$
for $n = 7$ **to** 1 **step** -1 **do**
  $span\_start \leftarrow \texttt{Lower}(generated\_span[:n])$
  $span\_end \leftarrow \texttt{Lower}(generated\_span[-n:])$
  for $i = 0$ **to** $\texttt{Len}(note) - n$ **do**
    $window\_start \leftarrow \texttt{Lower}(note[i:i+n])$
    if $window\_start$ == $span\_start$ **then**
      $start\_index \leftarrow i$
      for $j = 0$ **to** $\texttt{Len}(note) - start\_index - n$ **do**
        $end\_index \leftarrow \texttt{Len}(note) - j$
        if $end\_index - n < 0$ **then**
          break
        end if
        $window\_end \leftarrow \texttt{Lower}(note[end\_index - n : end\_index])$
        if $window\_end$ == $span\_end$ **and** $end\_index > start\_index$ **then**
          $text\_candidate \leftarrow note[start\_index : end\_index]$
          $candidates \leftarrow candidates \cup text\_candidate$
        end if
      end for
    end if
  end for
end for
for $item$ **in** $candidates$ **do**
  $score \leftarrow \texttt{OverlapScore}(generated\_span, item)$
  if $score > max\_score$ **then**
    $max\_score \leftarrow score$
    $best\_candidate \leftarrow item$
  end if
end for
return $best\_candidate, max\_score$
Notes: `Lower` *denotes the lowercase conversion function;* `Len` *function returns the length of a string.*

will serve as a gold standard for evaluating the explainability of ICD coding models. By comparing rationales identified by these models with those provided by human annotators, we aim to assess how effectively the models present rationales, particularly in a human-understandable manner.

Task Description. In this task, you will be presented with a patient's discharge summary and its assigned ICD (International Classification of Diseases) codes. Your goal is to identify and highlight text spans that support the given ICD codes as rationales. These highlighted spans may be words, phrases, or sentences (complete or incomplete).

Platform Setup.

- Install Docker following the official installation guide.
- Set up the annotation platform (Doccano).

Annotation Process.

- On the Doccano interface, you will see a patient's discharge summary along with its assigned ICD-10 codes and their descriptions.
- As you review the summary, highlight all text spans that you believe support each label (ICD code).
- When you select a span with your mouse, a selection list will appear. Click the appropriate label to annotate the span with it. The same text span can be annotated with multiple labels.

Completing the Annotation.

- When you would like to save your current annotations, click the 'X' mark on the top left corner to change the status from 'Not Checked' to 'Checked'.
- Once you have finished annotating all the labels for a summary, the sample will be marked as 'Finished'.

Platform We conducted the annotation using Doccano, a free and open-source platform designed to facilitate the creation of labeled datasets for natural language processing tasks. Doccano supports various annotation types, including text classification, sequence labeling (e.g., named entity recognition), and sequence-to-sequence tasks (e.g., machine translation or summarization). For this study, we employ the sequence labeling functionality. In the setup, the 'Allow overlapping spans' and 'Share annotations across all users' options are enabled. Additionally, a separate project is created for each document, each with its own defined label set.

Inter-Annotator Agreement Inter-annotator agreement (IAA) refers to the degree of consistency or reliability among different human annotators who independently annotate the same dataset. It is a critical measure to assess the quality and objectivity of annotated data, especially in tasks involving subjective judgments. We evaluate the annotation quality through a secondary annotator, who annotate the rationales for a subset of codes across 13 samples. To assess agreement, we calculate both span-level and token-level matching scores, including Precision, Recall and F1.

Here we explain how these scores are computed. Let A_2 be the set of tokens annotated by Annotator 2 and A_1 be the set of tokens annotated by Annotator 1. The overlap between these sets is:

$$\text{Overlap} = A_2 \cap A_1. \tag{4}$$

Precision, recall, and F1 scores are computed as follows:

$$\text{Precision} = \frac{|A_2 \cap A_1|}{|A_1|}, \tag{5}$$

$$\text{Recall} = \frac{|A_2 \cap A_1|}{|A_2|}, \tag{6}$$

$$\text{F1} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}. \tag{7}$$

Table 6 presents the results of the inter-annotator agreement analysis. The token-level F1-score (53.32) is substantially higher than the span-level F1-score (31.58), indicating that annotators exhibit greater consistency in identifying relevant content than in determining precise span boundaries. Table 7 provides a subset of detailed annotation matches between the two annotators.

ICD Codes with no supporting evidence During the annotation, we observe that not all codes have their supporting rationales in the text. It is mostly due to initial coding errors made by human annotators during the construction of the MIMIC benchmark. We list all codes with no rationale in Table 8, which also serves the analysis of the quality of the MIMIC-IV ICD10 dataset.

Table 6: Inter-Annotator Agreement

Metric	Value (%)
Span-Level Precision	38.00
Span-Level Recall	30.95
Span-Level F1	31.58
Token-Level Precision	79.84
Token-Level Recall	44.16
Token-Level F1	53.32

Details of Data Processing When evaluating plausibility, the original rationale annotations are preprocessed to align with the input formats of ICD coding models. We apply the same preprocessing procedures used in Edin et al. (2023)'s medical coding reproducibility study. Specifically, CAML and LAAT annotations are cleaned by lowercasing, removing special characters and stray numbers, and trimming extra spaces. For PLM-ICD, this cleaning step is followed by tokenization using the RoBERTa-base-PM-M3-Voc-distill-align-hf tokenizer from Hugging Face. These procedures ensure that the rationales extracted by each model are accurately comparable to our processed gold-standard annotations.

D Code Overlap Analysis

MDACE annotates a subset of MIMIC-III clinical notes with ICD-10 codes independently, then identifies their corresponding rationales. These coding results are then automatically mapped to ICD-9 codes using the General Equivalence Mappings (GEMs). This workflow introduces inconsistencies with the original ICD-9 code distribution in the MIMIC-III dataset commonly used for ICD coding tasks. We analyze the overlap between the two datasets under both the *Full* code setting and the *Top-50* code setting, with average overlaps of 37.00% and 14.59%, respectively. Figure 6 presents the distribution of overlap ratios in terms of number of documents and cumulative proportion. The results indicate that, for approximately 80% of the documents, the code overlap is below 60%. In the Top-50 setting, none of the documents exhibit more than 60% overlap. Figure 7 presents the frequency distribution of the Top-50 ICD-9 codes in both MDACE and the original MIMIC-III ICD-9 dataset. Notably, 6 of the Top-50 codes are entirely missing in MDACE (and 725 codes are missing in the Full-code setting, which comprises 1,281 codes in total). This substantial inconsistency poses challenges in accurately evaluating the explainability of ICD coding models.

In contrast, the annotation workflow of our dataset, RD-IV-10, is conducted by providing annotators with the discharge summaries and their corresponding labels from the MIMIC-IV ICD-10 dataset. The code distribution of our dataset closely matches that of MIMIC-IV ICD-10. As shown in Figure 6, most samples exhibit nearly 100% overlap, indicating a high consistency with the MIMIC-IV ICD-10 distribution. Specifically, the average overlaps reach 93.15% and 83.88% under the Full and Top-50 settings, respectively. Figure 8 further demonstrates the consistency in distribution. The reason they do not reach 100% is that both annotators observed that not all labels have supporting rationales in the text. Details of the codes without supporting rationales are provided in Appendix C.

Additionally, we investigate code switching by evaluating PLM-ICD on both the MDACE dataset and a filtered MIMIC-III ICD-9 test set sharing the same HADM IDs. MDACE samples come from two categories: Inpatient and Profee. The Inpatient category includes 302 samples, while Profee contains 52 samples corresponding to the same discharge summary but assigned with different codes. We combined the codes for samples sharing the same HADM ID. For comparison, we filtered the original MIMIC-III ICD-9 test set to include only samples with HADM IDs matching those in MDACE. We trained PLM-ICD on an extended label space that includes the 40 new ICD-9 codes introduced by MDACE. The testing results differ substantially from the original test set, showing a Precision@8 of 55.34% on MDACE versus 78.02% on the filtered MIMIC-III ICD-9 set, as shown in Table 9.

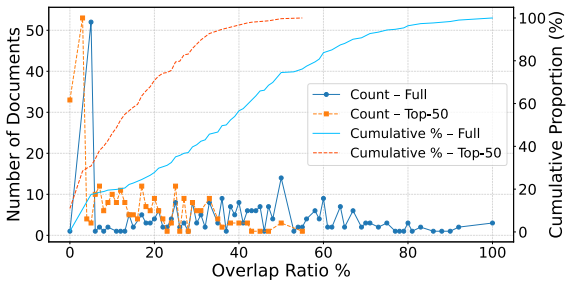

Figure 6: The statistics of overlap between MIMIC-III ICD-9 code set and mapped ICD-9 code set in MDACE.

Table 7: Part of annotation details of two annotators.

HADM ID	Code	Annotator2	Annotator1	Match
29531980	027034Z	'Left heart cath'	'nan'	False
29531980	I10	'hypertension'	'hypertension'	True
29531980	E1140	'type 2 diabetes'	'type 2 diabetes'	True
29531980	N390	'MRSA UTI'	'MRSA UTI '	False
29531980	E1140	'Diabetes'	'Diabetes'	True
29531980	I10	'Hypertension'	'Hypertension'	True
29531980	I10	'Hypertension'	'Hypertension'	True
29531980	E1140	'Diabetes'	'Diabetes'	True
29531980	I10	'hypertension'	'hypertension'	True
29474957	W363XXA	'Trauma'	'nan'	False
29474957	W363XXA	'patient reports that he was blown off by the lid of a highly pressurized natural gas tank in his pickup truck and sustained multiple injuries from the blast'	'he was blown off by the lid of a highly pressurized natural gas tank in his pickup truck and sustained multiple injuries from the blast'	False
29474957	H05221	'Right periorbital soft tissue swelling '	'Right periorbital soft tissue swelling '	True
29474957	S2241XA	'Right-sided rib fractures'	'rib fractures '	False
29474957	S2241XA	'Right-sided rib fractures involving the eighth and ninth ribs which appear comminuted displaced as well as fractured right tenth rib at the costovertebral junction'	'Right-sided rib fractures involving the eighth and ninth ribs which appear comminuted displaced as well as fractured right tenth rib at the costovertebral junction. '	False

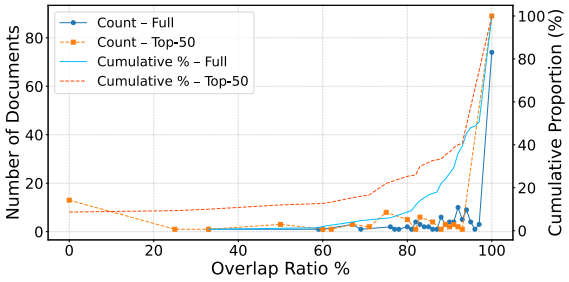

Figure 6: The statistics of overlap between MIMIC-IV ICD-10 code set and ICD-10 code set in RD-IV-10.

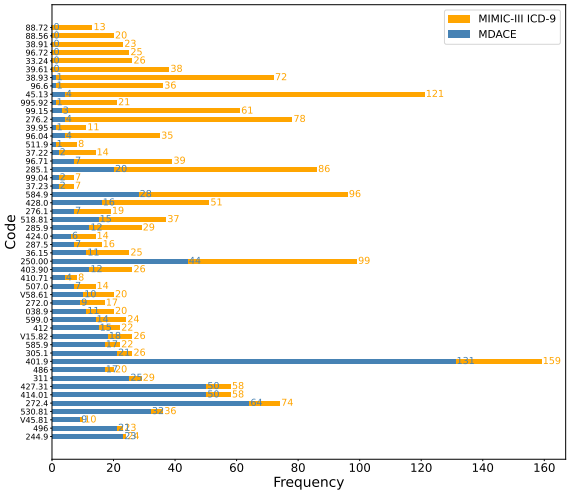

Figure 7: Code distributions in MDACE and MIMIC-III ICD-9. The analysis is based on the Top-50 codes in MIMIC-III ICD-9.

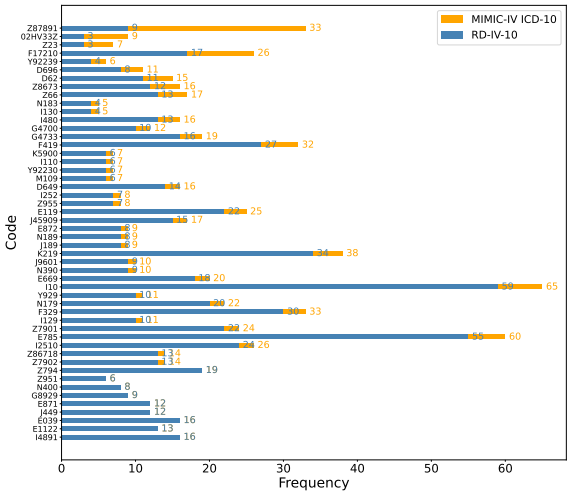

Figure 8: Code distributions in RD-IV-10 and MIMIC-IV ICD-10. The analysis is based on the Top-50 codes in MIMIC-IV ICD-10.

E A Case Study: Comparing Annotation Quality in RD-IV-10 and MDACE

Table 1 indicates that, on average, each ICD code in MDACE is associated with roughly one supporting rationale, given the ratio of 13.83 evidence spans to 11.53 labels per document. In contrast, RD-IV-10 provides substantially richer annotations, averaging 35.94 rationales for 14.82 labels per document. To illustrate this discrepancy, we present a case study shown in Figure 9 comparing both datasets. We randomly select one document annotated with the ICD-10 code **Z79.02 - Long-term use of antithrombotics/antiplatelets** from two datasets. In MDACE, the sole highlighted rationale is '*plavix*

Table 8: Codes without supporting rationales in the document.

HADM ID	ICD-10 Code
21843396	J45909;02HV33Z
29964986	30283B1;Z87891
29918504	F17210
29677969	Z96651;Z22322
28100046	J9811;J95811;Z23
27638102	F17290
27044834	E860
26620438	E8342
25912628	Z87891
24823574	Z87891
23856554	Z87891
23702445	F329
23355051	Z87891
23295582	Z87891;F419
29588477	02HV33Z
22893898	E874;J984;E870;H40052;Y848;Y92230;I9581;05UL0KZ;0NS60ZZ;08N1XZZ;02HV33Z
23051773	I25118;D62;D696;R350;5A1221Z
23618067	Z23
27840655	I5033;I110
22528733	N814
23106502	I082;I671;H409;R6884;Z87891;Y92009
25920183	A549;02HV33Z
26094695	Z66;R627;Z6821
27356906	J9600;Z781;K7290;Z87891;06L34CZ
21222006	Z22321
21318772	I2109;Z23;E861;R740;B1920
21386441	F17210
22257486	G4733
22733522	G4733;E669;Z6833;Z87891
23354056	D696;E860
23694175	K648;R29700
24345583	F17210
24852593	F0390;Z86711
25307585	E881;A630;L732;Z87891
25334768	A419;E46;I272
25510774	F1021
26911900	Z9119
27567712	T859XXA;T81.4XXA;Y92129;Z85820;0FPGX0Z;0F2GX0Z;02HV33Z;0JD80ZZ;0DHA8UZ;3E0H76Z
28716988	Z781
28831703	T8172XA;I808
29520101	Z87891
22032290	K55029;R6521;D62;F05;Z66;F17210
22114206	Z87891
22556702	Z006;K219
22983901	D6832;D6959;Z87891;T45515A
23383624	M25561;Z87820
23869666	E46;Z6829
24309140	B9562;I452;Z8673;Z86711;F17210;0W383ZZ;02H633Z
24352758	D684;N179;Z87891;D6959;Z781;0BC68ZZ;0BC48ZZ;3E0G76Z
24378932	N179;N182
24672299	Z87891
24974242	F17210
24991332	I871;05JY3ZZ
26852604	Z87891
27705753	Z87891;0U20KZ
28206965	I509;F05;BT11YZZ
28731328	Z87891;R0682;F40240
28756843	I959;Z87891
29060004	Z87891
29402217	Z22322

Table 9: Results on MDACE and filtered MIMIC-III ICD-9 test set (same HADM IDs).

Test Set	F1 Macro	F1 Micro	AUC Macro	AUC Micro	Precision@8
MDACE	3.39	50.38	90.82	96.95	55.34
Filtered MIMIC-III ICD-9 Test Set	5.01	59.74	94.78	99.01	78.02

75 mg'. RD-IV-10, however, highlights not only '*Plavix*' but also other directly relevant medications such as '*aspirin*' and '*clopidogrel*', found in multiple locations that MDACE fails to capture (highlighted in grey in the visualization). Additionally, our dataset annotates indirect supportive evidence '*However, patient declined cardiac catheterization given his desire for no invasive procedures*'. This statement reinforces the patient's complex cardiac history and justifies ongoing antiplatelet therapy, despite declining further interventions.

F Complete Results of Faithfulness and Plausibility of CAML, LAAT and PLM-ICD

Results of Faithfulness Tables 14, 15, and 16 summarize the faithfulness results of CAML, LAAT, and PLM-ICD, respectively, using two rationale selection strategies across four MIMIC datasets.

Results of Plausibility Tables 24, 25, and 26 summarize the plausibility results of CAML, LAAT, and PLM-ICD, respectively, using two rationale selection strategies across all threshold settings.

Model-generated Rationales: *Do Tokens with Higher Attention Weights Make Sense?* Figure 10 reports the matching results of *Top N tokens* across all thresholds. Overall, the performances of the three models are comparable. Specifically, PLM-ICD demonstrates better plausibility than LAAT, which in turn outperforms CAML. Span-level matches remain close to zero across all thresholds. Models achieve higher scores at lower thresholds, because shorter spans are selected, which better align with human annotations. For token-level matches, the absolute number of true positives increases at higher thresholds due to the inclusion of more tokens. However, the F1 scores decline because the total number of predicted tokens increases more substantially than the number of true positives. In conclusion, the rationales generated by ICD coding models do not align with human explanations.

G Complete Results of Plausibility

Plausibility Results of Three Types of Rationales Table 10 presents the plausibility results of naive entity-level rationales, strong LLM-generated rationales, model-generated rationales.

Plausibility Results of Top 5 Codes (Rationales Generated by Gemini 2-Flash) Table 11 summarizes the plausibility results of top 5 codes with and without incorporating few-shot human-annotated examples in the prompts. Incorporating few-shot examples substantially improves performance across all five codes and metrics. Notably, the token-level matches for I10 and the span-level matches for I4891 nearly double compared to those generated without few-shot examples.

Plausibility Results of LLMs-guided Rationale Learning Approaches Table 12 presents the plausibility results of the multi-objective model and its base model, PLM-ICD. We compare their performance across different numbers of selected tokens, ranging from 50 to 200. The multi-objective model consistently outperforms the base model across all settings and metrics, with particularly notable gains in span-level matches. Interestingly, the improvements are more pronounced at lower thresholds, indicating that the model generates more concise rationales, whereas at higher thresholds the performance gap narrows as more tokens are incorporated.

Table 13 presents the plausibility results of the NER model trained on a small dataset of 5,000 randomly selected samples. The results show that the NER model achieves the highest span-level plausibility, even surpassing its teacher model, Gemini 2-Flash. In contrast, PLM-ICD performs the worst. Interestingly, training on the smaller dataset improves plausibility compared to the results in Table 12. We attribute this finding to the influence of rationale proportions on the model's ability to generate plausible rationales.

Plausibility Results of Top 5 Codes (Rationales Generated by NER with Supervision of Rationales Labels Generated by Gemini 2-Flash) Table 14 presents the plausibility results for the top

MDACE HADM-ID: 112338

Admission Date: [**2189-1-29**] Discharge Date: [**2189-2-2**]
Date of Birth: [**2116-1-21**] Sex: F
…
Medications on Admission:
plavix 75mg
omeprazole 20mg [**Hospital1 **]
…
zocor 40mg
aspirin 81mg
percocet
amiodarone 400mg
lopressor 75mg
Discharge Medications:
1. Clopidogrel 75 mg Tablet Sig: One (1) Tablet PO DAILY (Daily). Disp:*60 Tablet(s)* Refills:*0*
2. Omeprazole 20 mg Capsule, Delayed Release(E.C.) Sig: One (1) Capsule, Delayed Release(E.C.) PO BID (2 times a day). Disp:*60 Capsule, Delayed Release(E.C.)(s)* Refills:*0*
3. Docusate Sodium 100 mg Capsule Sig: One (1) Capsule PO BID (2 times a day). Disp:*60 Capsule(s)* Refills:*0*
4. Aspirin 81 mg Tablet, Chewable Sig: One (1) Tablet, Chewable PO DAILY (Daily). Disp:*60 Tablet, Chewable(s)* Refills:*0*
5. Oxycodone-Acetaminophen 5-325 mg Tablet Sig: 1-2 Tablets PO Q6H (every 6 hours) as needed. Disp:*40 Tablet(s)* Refills:*0*
…
[**Name6 (MD) **] [**Name8 (MD) **] MD [**MD Number(2) 173**]
Completed by:[**2189-2-2**]

RD-IV-10 HADM-ID: 27638102

Name: ___ Unit No: ___
Admission Date: ___ Discharge Date: ___
Date of Birth: ___ Sex: M
…
Brief Hospital Course:
Mr. ___ is an ___ yo man with a history of NSTEMI s/p s/p
DES to LAD (___), LCX (___) who presented with right sided
light chest pressure at rest which was similar to his previous
NSTEMIs. Of note, he described holding his aspirin and clopidogrel for a Dermatology procedure.
ACTIVE PROBLEMS
NSTEMI
Community acquired pneumonia
He was noted to have an NSTEMI with troponins peaking at 0.03. His EKG was unremarkable for ischemic change, normal sinus rhythm, TWI in V1, no STE, unchanged from prior. He was placed on a heparin gtt with a plan for cardiac catheterization. However, patient declined cardiac catheterization given his desire for no invasive procedures. He also declines reversal of code status from DNR/DNI for procedures. He is on Plavix, aspirin , carvidilol, atorvastatin 80, and lisinopril-HCTZ at home, and these were continued. Also, we counseled the patient yesterday that he could NOT, under any circumstances, discontinue his DAPT without a cardiologist's permission. …
Medications on Admission:
The Preadmission Medication list is accurate and complete.
1. Amlodipine 10 mg PO DAILY
2. Aspirin 81 mg PO DAILY
3. BuPROPion (Sustained Release) 300 mg PO QAM
4. Clopidogrel 75 mg PO DAILY
5. DiphenhydrAMINE 25 mg PO QHS insomnia
…
Discharge Medications:
1. Amlodipine 10 mg PO DAILY
2. Amphetamine-Dextroamphetamine 10 mg PO TID
3. Aspirin 81 mg PO DAILY
4. Atorvastatin 80 mg PO QPM
5. BuPROPion (Sustained Release) 300 mg PO QAM
6. Clopidogrel 75 mg PO DAILY
7. DiphenhydrAMINE 25 mg PO QHS insomnia
8. Fluticasone Propionate NASAL 2 SPRY NU DAILY PRN
9. Lorazepam 1.5 mg PO QHS insomnia
10. Multivitamins 1 TAB PO DAILY
…
Your ___ team
Followup Instructions:
___

Figure 9: A case study of the annotation quality of RD-IV-10 and MDACE.

Table 10: Plausibility results of naive entity-level rationales, strong LLM-generated rationales, model-generated rationales. Prediction refers to the number of spans or tokens generated by the model, while Accurate denotes the number of spans or tokens matching the human-annotated gold standard. This evaluation is based on 64 documents to enable a comparison of entity linking.

Metric	Model/Dataset	Prdiction	Accurate	TP	FP	FN	Precision (%)	Recall (%)	F1 (%)
Exact SM	Entity-Linking	3546	2260	298	3248	1962	8.4	13.2	10.3
	Gemini 2-Flash	4726	2260	754	3972	1506	16.0	33.4	21.6
	Gemini1.5-Pro	11184	2260	907	10277	1353	8.1	40.1	13.5
	LLaMA-3.3 Ins	5789	2260	747	5042	1513	12.9	33.1	18.6
	LLaMA-3.3 AWQ	6428	2260	761	5667	1499	11.8	33.7	17.5
	CAML	84520	2269	55	84465	2214	0.1	2.4	0.1
	LAAT	91482	2269	330	91152	1939	0.4	14.5	0.7
	PLM-ICD	65639	2269	172	65467	2097	0.3	7.6	0.5
PI SM	Entity-Linking	2751	1762	207	2544	1555	7.5	11.7	9.2
	Gemini-2flash	4664	1762	773	3891	989	16.6	43.9	24.1
	Gemini-1.5pro	10996	1762	932	10064	830	8.5	52.9	14.6
	LLaMA-3.3 Ins	5726	1762	805	4921	957	14.1	45.7	21.5
	LLaMA-3.3 AWQ	6364	1762	816	5548	946	12.8	46.3	20.1
	CAML	77970	2021	88	77882	1933	0.1	4.4	0.2
	LAAT	82542	2021	342	82200	1679	0.4	16.9	0.8
	PLM-ICD	60851	2041	228	60623	1813	0.4	11.2	0.7
Exact TM	Entity-Linking	5629	11422	540	5089	10882	9.6	4.7	6.3
	Gemini-2flash	27639	11422	5881	21758	5541	21.3	51.5	30.1
	Gemini-1.5pro	57708	11422	7109	50599	4313	12.3	62.2	20.6
	LLaMA-3.3 Ins	28800	11422	5588	23212	5834	19.4	48.9	27.8
	LLaMA-3.3 AWQ	32693	11422	6008	26685	5414	18.4	52.6	27.2
	CAML	150968	11428	2493	148475	8935	1.7	21.8	3.1
	LAAT	160732	11428	4266	156466	7162	2.7	37.3	5.0
	PLM-ICD	173633	12517	3975	169658	8542	2.3	31.8	4.3
PI TM	Entity-Linking	4292	8160	381	3911	7779	8.9	4.7	6.1
	Gemini-2flash	18935	8160	5047	13888	3113	26.7	61.9	37.3
	Gemini-1.5pro	37197	8160	5898	31299	2262	15.9	72.3	26.0
	LLaMA-3.3 Ins	20708	8160	5056	15652	3104	24.4	62.0	35.0
	LLaMA-3.3 AWQ	23302	8160	5361	17941	2799	23.0	65.7	34.1
	CAML	114766	9367	4029	110737	5338	3.5	43.0	6.5
	LAAT	121292	9367	4720	116572	4647	3.9	50.4	7.2
	PLM-ICD	120777	10269	5794	114983	4475	4.8	56.4	8.8

Table 11: Plausibility results of top 5 codes. Rationales are generated by Gemini 2-Flash. "w/o" and "w/" indicate the absence or inclusion of few-shot examples in prompts.

Code	Metric	Setting	Prdiction	Accturate	TP	FP	FN	Precision (%)	Recall (%)	F1 (%)	ΔF1
I10	Exact SM	w/o	147	151	75	72	76	51.0	49.7	50.3	+10.2
		w/	107	141	75	32	66	70.1	53.2	60.5	
	PI SM	w/o	140	82	70	70	12	50.0	85.4	63.1	+15.1
		w/	100	79	70	30	9	70.0	88.6	78.2	
	Exact TM	w/o	526	155	92	434	63	17.5	59.4	27.0	+26.8
		w/	182	145	88	94	57	48.4	60.7	53.8	
	PI TM	w/o	411	86	80	331	6	19.5	93.0	32.2	+34.2
		w/	152	83	78	74	5	51.3	94.0	66.4	
E785	Exact SM	w/o	79	97	55	24	42	69.6	56.7	62.5	+10.7
		w/	69	95	60	9	35	87.0	63.2	73.2	
	PI SM	w/o	72	56	49	23	7	68.1	87.5	76.6	+12.9
		w/	60	54	51	9	3	85.0	94.4	89.5	
	Exact TM	w/o	140	107	62	78	45	44.3	57.9	50.2	+16.5
		w/	84	99	61	23	38	72.6	61.6	66.7	
	PI TM	w/o	125	66	57	68	9	45.6	86.4	59.7	+18.5
		w/	75	58	52	23	6	69.3	89.7	78.2	
Z7901	Exact SM	w/o	83	50	6	77	44	7.2	12.0	9.0	+4.0
		w/	102	52	10	92	42	9.8	19.2	13.0	
	PI SM	w/o	82	42	6	76	36	7.3	14.3	9.7	+7.1
		w/	100	43	12	88	31	12.0	27.9	16.8	
	Exact TM	w/o	524	387	163	361	224	31.1	42.1	35.8	+10.0
		w/	645	389	237	408	152	36.7	60.9	45.8	
	PI TM	w/o	354	265	134	220	131	37.9	50.6	43.3	+11.5
		w/	413	266	186	227	80	45.0	69.9	54.8	
I4891	Exact SM	w/o	47	55	6	41	49	12.8	10.9	11.8	+12.9
		w/	42	55	12	30	43	28.6	21.8	24.7	
	PI SM	w/o	45	27	6	39	21	13.3	22.2	16.7	+20.2
		w/	38	27	12	26	15	31.6	44.4	36.9	
	Exact TM	w/o	221	93	44	177	49	19.9	47.3	28.0	+19.9
		w/	124	93	52	72	41	41.9	55.9	47.9	
	PI TM	w/o	164	37	35	129	2	21.3	94.6	34.8	+21.4
		w/	84	37	34	50	3	40.5	91.9	56.2	
E119	Exact SM	w/o	53	50	21	32	29	39.6	42.0	40.8	+3.4
		w/	54	50	23	31	27	42.6	46.0	44.2	
	PI SM	w/o	53	33	21	32	12	39.6	63.6	48.8	+4.1
		w/	54	33	23	31	10	42.6	69.7	52.9	
	Exact TM	w/o	225	79	55	170	24	24.4	69.6	36.2	+7.4
		w/	187	79	58	129	21	31.0	73.4	43.6	
	PI TM	w/o	186	49	44	142	5	23.7	89.8	37.4	+7.4
		w/	152	49	45	107	4	29.6	91.8	44.8	

Table 12: Plausibility results of multi-objective learning approach and the baseline model PLM-ICD. All results are reported as percentages. The experiments are conducted under the Top-50 code settings.

Setting	Model	Exact Span			PI Span			Exact Token			PI Token		
		Pr	Rc	F1	Pr	Rc	F1	Pr	Rc	F1	Pr	Rc	F1
50	PLM-ICD	1.5	14.2	2.7	1.7	17.2	3.0	5.5	35.1	9.5	7.1	43.8	12.2
	Multi-objective	2.1	20.9	3.9	2.2	23.5	4.1	6.1	39.1	10.6	7.6	47.4	13.2
100	PLM-ICD	0.6	9.3	1.1	0.7	11.8	1.3	2.9	36.9	5.5	4.5	51.7	8.3
	Multi-objective	1.0	16.6	1.9	1.1	18.8	2.0	3.2	40.2	5.9	4.7	53.5	8.6
150	PLM-ICD	0.3	7.1	0.6	0.4	10.0	0.8	2.0	37.4	3.8	3.5	56.9	6.6
	Multi-objective	0.7	14.0	1.3	0.8	16.9	1.4	2.1	38.2	3.9	3.5	55.2	6.5
200	PLM-ICD	0.2	6.6	0.5	0.3	9.7	0.7	1.5	36.2	2.8	2.9	59.9	5.6
	Multi-objective	0.5	12.4	1.0	0.6	14.9	1.1	1.6	38.4	3.0	2.9	58.8	5.5

Table 13: Plausibility results for the NER model, the teacher model Gemini 2-Flash, and the baseline model PLM-ICD. All results are reported as percentages. The experiments are conducted under the Top-50 code settings.

Setting	Model	Exact Span			PI Span			Exact Token			PI Token		
		Pr	Rc	F1	Pr	Rc	F1	Pr	Rc	F1	Pr	Rc	F1
-	NER Model	29.3	24.3	26.5	29.0	32.5	30.6	35.1	15.8	21.8	37.1	21.2	27.0
-	Gemini 2-Flash	16.4	20.5	18.2	18.2	31.2	23.0	23.0	40.6	29.4	23.7	46.4	31.4
50	PLM-ICD	2.5	12.8	4.1	2.5	14.8	4.3	4.7	28.2	8.1	7.0	41.6	12.0
100	PLM-ICD	1.5	11.9	2.6	1.6	14.3	2.8	2.7	31.5	4.9	4.7	51.2	8.7
150	PLM-ICD	0.9	9.7	1.7	1.0	12.6	1.9	1.9	33.0	3.5	3.7	56.4	6.9
200	PLM-ICD	0.6	7.5	1.1	0.7	10.0	1.3	1.5	35.1	2.9	3.1	60.5	6.0

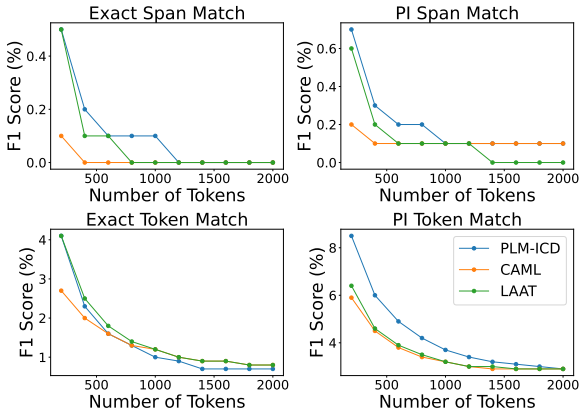

Figure 10: Plausibility results of ICD coding models across all thresholds settings.

5 codes. The rationales are generated using NER models trained on rationales produced by Gemini 2-Flash, with and without the inclusion of few-shot human-annotated examples in the prompts. Models incorporating few-shot examples demonstrate improved performance across most codes and metrics. Furthermore, the supervised model outperforms the unsupervised model for codes I10, E785, and E119, due to the higher frequency of corresponding rationales in the documents.

H A Case Study: Comparative Analysis of Rationales From Four Sources

To examine the rationales produced by different sources in detail, we present a case study comparing Human Annotation, Entity-Linking, Gemini 2-Flash, and PLM-ICD. The focus is on rationales for ICD-10 code N49.2 – Inflammatory disorders of scrotum. We randomly select the document HADM-ID 29964986, to illustrate how each method justifies the assigned code.

Table 11 shows that Entity Linking and Human Annotation partially overlap. Entity Linking highlights only the direct mentions of the code, such as 'SCROTAL' and 'scrotal abscess', whereas Human Annotation additionally mark broader contextual information—the sentence: 'This patient was admitted to the urology service following debridement of scrotal and perineal abscess.' Gemini 2-Flash also performs well in this case. It highlights phrases overlooked by both human annotators and Entity Linking—such as 'scrotal and perineal abscess' and 'right scrotal fluctuance'—and even captures parts of that human-annotated sentence. However, it also returns some incorrect spans, including 'Perineal abscess' and 'Hemiscrotum'. Additionally, we observe that the rationales of PLM-ICD are randomly distributed and lack meaningful explainability; therefore, we do not visualize them in this case study. This case study supports our earlier conclusion that Gemini 2-Flash achieves the highest plausibility, while naive directly linked entities can also serve as rationales to a certain extent. In contrast, the rationales generated by PLM-ICD exhibit the lowest plausibility.

I Trade-off Between ICD Coding Performance and Rationale Plausibility in an NER-Based Approach Across Different Training Data Sizes

We also analyse the trade-off between coding performance (Table 15) and rationale plausibility (Table 16) under different training-data sizes. When the number of training samples is reduced by 80% (retaining only 1000 samples), coding performance drops by about 10% (F1 scores), while plausibility decreases by approximately 10–16%, indicating comparable levels of degradation. Additionally, we observe that both coding performance and plausibility converge as the training size increases - the results with 3,000 and 5,000 samples are highly similar. This suggests that even relatively small training sets are sufficient to achieve high-level plausibility.

RD-IV-10	Entity Linking	Gemini 2-Flash
Name: ___ Unit No: ___ Admission Date: ___ Discharge Date: ___ Date of Birth: ___ Sex: M Service: UROLOGY Allergies: ciprofloxacin Attending: ___. Chief Complaint: SCROTAL/PERINEAL ABSCESS Major Surgical or Invasive Procedure: 1. Exam under anesthesia. 2. Debridement of scrotal and perineal abscess. History of Present Illness: ... chronic indwelling foley; transferred from ___ with right scrotal abscess . Patient presented to his see his urologist (___) ... empiric treatment. When found to have right scrotal fluctuance, there was an ultrasound performed that showed right scrotal abscess with concern for fluid and gas extending into the ... place; ___ wick removed. Hemiscrotum Incision c/d/i w/out ... fluid should be submitted. Brief Hospital Course: This patient was admitted to the urology service following debridement of scrotal and perineal abscess. See operative report for full details. The patient tolerated the procedure well and recovered in the PACU before transfer to the surgical floor. He was admitted on zosyn/clindamycin for antibiotic ... Discharge Diagnosis: PREOPERATIVE DIAGNOSES: 1. Scrotal abscess. 2. Perineal abscess. POSTOPERATIVE DIAGNOSES: 1. Scrotal abscess. 2. Perineal abscess.	Name: ___ Unit No: ___ Admission Date: ___ Discharge Date: ___ Date of Birth: ___ Sex: M Service: UROLOGY Allergies: ciprofloxacin Attending: ___. Chief Complaint: SCROTAL /PERINEAL ABSCESS Major Surgical or Invasive Procedure: 1. Exam under anesthesia. 2. Debridement of scrotal and perineal abscess. History of Present Illness: ... chronic indwelling foley; transferred from ___ with right scrotal abscess . Patient presented to his see his urologist (___) ... empiric treatment. When found to have right scrotal fluctuance, there was an ultrasound performed that showed right scrotal abscess with concern for fluid and gas extending into the ... place; ___ wick removed. Hemiscrotum Incision c/d/i w/out ... fluid should be submitted. Brief Hospital Course: This patient was admitted to the urology service following debridement of scrotal and perineal abscess. See operative report for full details. The patient tolerated the procedure well and recovered in the PACU before transfer to the surgical floor. He was admitted on zosyn/clindamycin for antibiotic ... Discharge Diagnosis: PREOPERATIVE DIAGNOSES: 1. Scrotal abscess . 2. Perineal abscess. POSTOPERATIVE DIAGNOSES: 1. Scrotal abscess. 2. Perineal abscess.	Name: ___ Unit No: ___ Admission Date: ___ Discharge Date: ___ Date of Birth: ___ Sex: M Service: UROLOGY Allergies: ciprofloxacin Attending: ___. Chief Complaint: SCROTAL/PERINEAL ABSCESS Major Surgical or Invasive Procedure: 1. Exam under anesthesia. 2. Debridement of scrotal and perineal abscess . History of Present Illness: ... chronic indwelling foley; transferred from ___ with right scrotal abscess. Patient presented to his see his urologist (___) ... empiric treatment. When found to have right scrotal fluctuance, there was an ultrasound performed that showed right scrotal abscess with concern for fluid and gas extending into the ... place; ___ wick removed. Hemiscrotum Incision c/d/i w/out ... fluid should be submitted. Brief Hospital Course: This patient was admitted to the urology service following debridement of scrotal and perineal abscess. See operative report for full details. The patient tolerated the procedure well and recovered in the PACU before transfer to the surgical floor. He was admitted on zosyn/clindamycin for antibiotic ... Discharge Diagnosis: PREOPERATIVE DIAGNOSES: 1. Scrotal abscess . 2. Perineal abscess . POSTOPERATIVE DIAGNOSES: 1. Scrotal abscess. 2. Perineal abscess.

Figure 11: A case study of annotations of RD-IV-10, Entity Linking and Gemini 2-Flash. The highlights indicate their corresponding annotations.

Table 14: Plausibility results of top 5 codes. Rationales are generated by NER models. "w/o" and "w/" indicate the absence or inclusion of few-shot examples in prompts.

Code	Metric	Setting	Prdiction	Accturate	TP	FP	FN	Precision (%)	Recall (%)	F1 (%)	ΔF1
I4891	Exact SM	w/o	44	55	11	33	44	25.0	20.0	22.2	+3.9
		w/	37	55	12	25	43	32.4	21.8	26.1	
	PI SM	w/o	42	27	13	29	14	31.0	48.1	37.7	+5.6
		w/	33	27	13	20	14	39.4	48.1	43.3	
	Exact TM	w/o	115	93	40	75	53	34.8	43.0	38.5	+2.2
		w/	89	93	37	52	56	41.6	39.8	40.7	
	PI TM	w/o	86	37	30	56	7	34.9	81.1	48.8	+5.7
		w/	62	37	27	35	10	43.5	73.0	54.5	
I10	Exact SM	w/o	80	144	62	18	82	77.5	43.1	55.4	+7.1
		w/	80	144	70	10	74	87.5	48.6	62.5	
	PI SM	w/o	74	82	60	14	22	81.1	73.2	76.9	+8.3
		w/	73	82	66	7	16	90.4	80.5	85.2	
	Exact TM	w/o	93	147	67	26	80	72.0	45.6	55.8	+9.1
		w/	84	147	75	9	72	89.3	51.0	64.9	
	PI TM	w/o	85	85	64	21	21	75.3	75.3	75.3	+10.9
		w/	75	85	69	6	16	92.0	81.2	86.2	
E785	Exact SM	w/o	62	103	56	6	47	90.3	54.4	67.9	+2.4
		w/	62	103	58	4	45	93.5	56.3	70.3	
	PI SM	w/o	56	61	50	6	11	89.3	82.0	85.5	+1.5
		w/	54	61	50	4	11	92.6	82.0	87.0	
	Exact TM	w/o	64	121	57	7	64	89.1	47.1	61.6	+1.4
		w/	63	121	58	5	63	92.1	47.9	63.0	
	PI TM	w/o	57	76	50	7	26	87.7	65.8	75.2	+1.1
		w/	55	76	50	5	26	90.9	65.8	76.3	
E119	Exact SM	w/o	37	52	22	15	30	59.5	42.3	49.4	-3.6
		w/	31	52	19	12	33	61.3	36.5	45.8	
	PI SM	w/o	36	35	22	14	13	61.1	62.9	62.0	-3.5
		w/	30	35	19	11	16	63.3	54.3	58.5	
	Exact TM	w/o	62	96	42	20	54	67.7	43.7	53.2	-3.5
		w/	53	96	37	16	59	69.8	38.5	49.7	
	PI TM	w/o	51	58	34	17	24	66.7	58.6	62.4	-2.2
		w/	45	58	31	14	27	68.9	53.4	60.2	
Z7901	Exact SM	w/o	45	52	5	40	47	11.1	9.6	10.3	+0.5
		w/	59	52	6	53	46	10.2	11.5	10.8	
	PI SM	w/o	44	43	6	38	37	13.6	14.0	13.8	-1.8
		w/	57	43	6	51	37	10.5	14.0	12.0	
	Exact TM	w/o	167	389	71	96	318	42.5	18.3	25.5	-3.0
		w/	190	389	65	125	324	34.2	16.7	22.5	
	PI TM	w/o	138	266	78	60	188	56.5	29.3	38.6	+2.1
		w/	157	266	86	71	180	54.8	32.3	40.7	

J Statistics of Gemini-2 Flash Generated Rationale Dataset Across Different Alignment Settings (Overlap)

We use Gemini 2-Flash to generate rationale labels as weak supervision signals for all 122,004 samples in MIMIC-IV for rationale learning. The statistical analyses under different overlap-score thresholds for the alignments are presented in Table 17. We observe that 93.35% of the spans extracted by the LLM are nearly identical to the original text (an overlap score of 2 indicates an exact match).

K Plausibility Across Different Alignment Settings (Overlap)

We also compare the plausibility results for Gemini 2-Flash across different alignment settings, where the rationales are generated in a separate round. From Table 18, we observe that higher alignment quality consistently yields higher scores across all four metrics. More specifically, including spans with lower overlap scores substantially degrades token-level matching performance, while having only a minor impact on span-level metrics. This is because target spans often cover a wide range of tokens, so adding misaligned spans introduces many irrelevant tokens at the token level, whereas

Table 15: Trade-off with Different Training-Data Sizes – ICD Coding Performance (NER).

Size	F1-Mac	F1-Mic	P-Mac	P-Mic	R-Mac	R-Mic
1000	49.14	60.88	41.69	50.60	62.95	76.39
3000	54.00	68.22	48.76	60.30	64.13	78.54
5000	53.46	67.75	49.21	60.93	61.79	76.30

Table 16: Trade-off with Different Training-Data Sizes – Rationale Plausibility (NER).

Metric	Size	Prediction	Accurate	TP	FP	FN	Precision (%)	Recall (%)	F1 (%)
	1000	1551	1607	358	1193	1249	23.1	22.3	22.7
Exact SM	3000	1420	1607	416	1004	1191	29.3	25.9	27.5
	5000	1332	1607	390	942	1217	29.3	24.3	26.5
	1000	1502	1145	343	1159	802	22.8	30.0	25.9
PI SM	3000	1368	1145	400	968	745	29.2	34.9	31.8
	5000	1283	1145	372	911	773	29.0	32.5	30.6
	1000	2988	5531	803	2185	4728	26.9	14.5	18.9
Exact TM	3000	2713	5531	895	1818	4636	33.0	16.2	21.7
	5000	2493	5531	874	1619	4657	35.1	15.8	21.8
	1000	2684	3891	799	1885	3092	29.8	20.5	24.3
PI TM	3000	2439	3891	852	1587	3039	34.9	21.9	26.9
	5000	2222	3891	825	1397	3066	37.1	21.2	27.0

at the span level the misalignment affects the counts of TP, FP, and FN by only ±1.

L Impact of Label Selection under Different Span Alignment Settings on Rationale Learning

In Tables 19 and 20, we report the downstream learning outcomes under different span alignment settings. We train the NER model using all spans (threshold = 0). We find that lower alignment does not substantially affect ICD coding performance or plausibility scores, because the proportion of training data with low alignment is small, with spans in the range 0.0–1.9 accounting for less than 7% of the dataset.

Table 17: Counts and ratios of rationales across different overlap scores.

Range	Count	Ratio (%)	Range	Count	Ratio (%)
0.0–0.1	174,730	2.82%	1.0–1.1	11,026	0.18%
0.1–0.2	24,021	0.39%	1.1–1.2	2,985	0.05%
0.2–0.3	64,776	1.05%	1.2–1.3	1,637	0.03%
0.3–0.4	63,037	1.02%	1.3–1.4	10,390	0.17%
0.4–0.5	48,663	0.79%	1.4–1.5	1,770	0.03%
0.5–0.6	36,614	0.59%	1.5–1.6	9,493	0.15%
0.6–0.7	19,274	0.31%	1.6–1.7	17,472	0.28%
0.7–0.8	7,397	0.12%	1.7–1.8	17,392	0.28%
0.8–0.9	7,648	0.12%	1.8–1.9	30,288	0.49%
0.9–1.0	4,013	0.06%	1.9–2.0	5,780,713	93.35%
			Total	6,191,350	100%

Table 18: Plausibility results for Gemini 2-Flash across different alignment settings (Overlap).

Metric	Overlap	Prdiction	Accurate	TP	FP	FN	Precision (%)	Recall (%)	F1 (%)
Exact SM	0	3418	2260	711	2707	1549	20.8	31.5	25.0
	0.5	3303	2260	711	2592	1549	21.5	31.5	25.6
	1	3180	2260	710	2470	1550	22.3	31.4	26.1
	1.5	3157	2260	704	2453	1556	22.3	31.2	26.0
	2	3105	2260	700	2405	1560	22.5	31.0	26.1
PI SM	0	3365	1762	730	2635	1032	21.7	41.4	28.5
	0.5	3250	1762	730	2520	1032	22.5	41.4	29.1
	1	3127	1762	729	2398	1033	23.3	41.4	29.8
	1.5	3105	1762	724	2381	1038	23.3	41.1	29.8
	2	3053	1762	720	2333	1042	23.6	40.9	29.9
Exact TM	0	64247	11422	5293	58954	6129	8.2	46.3	14.0
	0.5	53179	11422	5189	47990	6233	9.8	45.4	16.1
	1	20494	11422	4976	15518	6446	24.3	43.6	31.2
	1.5	18151	11422	4911	13240	6511	27.1	43.0	33.2
	2	17723	11422	4867	12856	6555	27.5	42.6	33.4
PI TM	0	36499	8160	4739	31760	3421	13.0	58.1	21.2
	0.5	29972	8160	4617	25355	3543	15.4	56.6	24.2
	1	14343	8160	4383	9960	3777	30.6	53.7	39.0
	1.5	13086	8160	4329	8757	3831	33.1	53.1	40.8
	2	12843	8160	4306	8537	3854	33.5	52.8	41.0

Table 19: Effect of Span Alignment on Rationale Learning (NER) - ICD Coding Performance.

Threshold	F1-Mac	F1-Mic	P-Mac	P-Mic	R-Mac	R-Mic
0	53.35	68.65	49.36	61.12	61.76	78.30
1.7	53.46	67.75	49.21	60.93	61.79	76.30

Table 20: Effect of Span Alignment on Rationale Learning (NER) - Plausibility.

Threshold	Exact Span			PI Span			Exact Token			PI Token		
	Pr	Rc	F1	Pr	Rc	F1	Pr	Rc	F1	Pr	Rc	F1
0	29.5	25.1	27.1	29.8	34.3	31.9	32.2	15.6	21.1	35.3	21.7	26.9
1.7	29.3	24.3	26.5	29.0	32.5	30.6	35.1	15.8	21.8	37.1	21.2	27.0

Table 21: Faithfulness across Top-P and Top-N thresholds for CAML.

Dataset	MIMIC-IV ICD10 Full						MIMIC-IV ICD10 Top-50						MIMIC-III ICD9 Full						MIMIC-III ICD9 Top-50					
Threshold	**F1-Mac**	**F1-Mic**	**AUC-Mac**	**AUC-Mic**	**P@8**	**Retention**	**F1-Mac**	**F1-Mic**	**AUC-Mac**	**AUC-Mic**	**P@5**	**Retention**	**F1-Mac**	**F1-Mic**	**AUC-Mac**	**AUC-Mic**	**P@8**	**Retention**	**F1-Mac**	**F1-Mic**	**AUC-Mac**	**AUC-Mic**	**P@5**	**Retention**
-	15.82	54.87	91.01	98.52	66.09	100.00	65.78	71.23	91.95	94.58	63.53	100.00	9.93	50.84	81.87	97.46	70.81	100.00	55.62	63.98	89.77	92.56	63.32	100.00
Top-p Sufficiency																								
10%	7.72	30.00	79.09	93.61	43.63	12.05	56.90	62.76	86.23	89.17	46.09	15.18	5.05	17.82	67.95	91.84	41.54	10.99	33.45	35.69	80.55	81.50	40.45	15.26
20%	9.15	35.42	83.97	96.15	48.90	21.83	58.97	64.85	88.49	91.28	47.03	24.61	6.03	22.82	71.79	93.81	49.18	20.88	37.31	40.27	83.23	85.25	42.66	24.67
30%	10.12	38.88	86.22	97.11	52.09	31.60	60.10	65.96	89.57	92.30	47.63	34.03	6.61	26.53	74.28	94.95	53.59	30.77	40.55	44.19	85.06	87.68	44.26	34.09
40%	10.83	41.33	87.76	97.61	54.35	41.37	60.87	66.70	90.16	92.92	48.07	43.45	7.02	29.41	75.92	95.66	56.65	40.66	43.01	47.22	86.31	89.15	45.41	43.51
50%	11.36	43.16	88.76	97.91	55.99	51.14	61.48	67.28	90.60	93.36	48.42	52.88	7.33	31.81	76.87	96.18	58.81	50.55	44.88	49.53	87.17	90.09	46.27	52.92
60%	11.81	44.61	89.49	98.13	57.29	60.91	61.97	67.74	90.94	93.69	48.71	62.30	7.56	33.76	78.02	96.55	60.42	60.44	46.24	51.28	87.82	90.77	46.91	62.34
70%	12.19	45.78	90.04	98.27	58.33	70.68	62.38	68.13	91.20	93.95	48.94	71.73	7.74	35.34	78.43	96.73	61.67	70.33	47.26	52.63	88.33	91.29	47.39	71.75
80%	15.14	53.93	90.47	98.38	65.24	80.46	62.73	68.45	91.46	94.17	49.14	81.15	7.88	36.65	78.76	96.85	62.68	80.22	48.03	53.68	88.71	91.65	47.78	81.17
90%	15.39	54.21	90.78	98.46	65.47	90.23	63.02	68.73	91.71	94.37	49.33	90.58	8.01	37.77	79.06	96.96	63.51	90.11	48.62	54.51	89.23	92.11	48.11	90.58
Top-p Comprehensiveness																								
10%	12.62	46.54	89.17	97.92	57.84	90.23	13.72	15.39	71.68	71.27	22.78	90.58	2.80	13.54	75.15	95.37	35.25	90.11	8.62	9.66	75.71	74.72	29.31	90.58
20%	11.97	43.43	87.35	97.17	54.37	80.46	12.73	14.58	67.88	67.80	21.55	81.15	2.59	12.32	72.10	93.60	31.17	80.22	6.90	7.75	66.85	65.97	26.49	81.17
30%	11.33	40.59	85.54	96.34	51.17	70.68	11.97	14.03	65.01	65.33	20.63	71.73	2.45	11.31	69.62	92.06	28.28	70.33	6.20	6.98	60.43	61.35	24.80	71.75
40%	10.73	37.82	84.00	95.44	48.03	60.91	11.29	13.55	62.51	63.21	19.82	62.30	2.36	10.56	67.50	90.71	26.04	60.44	5.84	6.59	56.08	58.58	23.64	62.34
50%	10.11	34.98	82.40	94.31	44.93	51.14	10.74	13.18	60.01	61.34	19.10	52.88	2.29	9.84	65.22	89.41	24.06	50.55	5.77	6.49	53.10	56.58	22.77	52.92
60%	9.46	31.89	80.53	92.84	41.79	41.37	10.29	12.86	57.61	59.61	18.43	43.45	2.20	9.07	63.41	88.11	22.34	40.66	5.77	6.50	50.71	54.94	22.07	43.51
70%	8.77	28.38	78.43	90.86	38.64	31.60	9.87	12.55	55.22	58.08	17.85	34.03	2.13	8.34	61.66	86.76	20.79	30.77	5.80	6.54	48.70	53.44	21.47	34.09
80%	4.16	7.99	75.30	87.81	12.98	21.83	9.52	12.28	53.15	56.62	17.31	24.61	2.00	7.30	59.62	84.65	19.27	20.88	5.90	6.62	47.25	51.91	20.91	24.67
90%	3.17	4.81	70.84	82.12	9.55	12.05	9.23	12.09	51.01	55.13	16.79	15.18	1.76	5.64	56.38	80.38	17.70	10.99	6.03	6.79	46.07	50.32	20.31	15.26
Top-N Sufficiency																								
200	9.13	34.84	81.37	94.70	48.39	16.61	58.68	64.53	87.02	89.94	46.77	19.11	5.78	22.00	70.01	92.67	45.67	14.24	35.27	37.86	81.21	82.40	41.27	18.07
400	10.49	39.54	85.53	96.75	52.93	30.71	60.24	66.10	88.97	91.81	47.62	32.37	6.70	27.69	73.45	94.47	52.63	27.31	39.15	42.49	83.81	85.96	43.44	30.24
600	11.40	42.44	87.53	97.52	55.60	44.48	61.19	67.01	89.91	92.70	48.17	45.39	7.26	31.74	75.62	95.55	56.68	40.06	42.15	46.12	85.34	88.03	44.92	42.10
800	12.03	44.48	88.84	97.91	57.44	57.43	61.87	67.65	90.47	93.26	48.57	57.80	7.61	34.56	77.10	96.14	59.25	52.18	44.37	48.85	86.42	89.31	45.98	53.37
1000	12.51	45.99	89.60	98.14	58.79	68.87	62.36	68.13	90.89	93.67	48.88	68.93	7.86	36.70	77.82	96.54	61.09	63.17	46.04	50.93	87.28	90.25	46.74	63.74
1200	12.90	47.16	90.12	98.28	59.82	78.15	62.76	68.51	91.23	93.97	49.13	78.06	8.06	38.36	78.55	96.79	62.45	72.61	47.28	52.51	87.87	90.88	47.29	72.84
1400	13.22	48.08	90.47	98.37	60.60	85.12	63.10	68.83	91.46	94.17	49.35	84.98	8.22	39.67	79.14	96.95	63.49	80.17	48.26	53.77	88.36	91.37	47.74	80.22
1600	13.48	48.82	90.66	98.42	61.24	90.08	63.38	69.09	91.63	94.31	49.53	89.96	8.35	40.72	79.77	97.08	64.32	85.94	49.01	54.77	88.80	91.74	48.10	85.89
1800	13.70	49.43	90.80	98.46	61.75	93.53	63.61	69.30	91.73	94.40	49.68	93.44	8.45	41.60	80.12	97.18	64.98	90.27	49.62	55.60	89.13	92.04	48.40	90.12
2000	13.88	49.93	90.91	98.48	62.16	95.83	63.81	69.48	91.83	94.47	49.81	95.78	8.54	42.34	80.71	97.27	65.53	93.42	50.13	56.29	89.31	92.22	48.66	93.19
Top-N Comprehensiveness																								
200	11.54	43.65	88.26	97.59	54.84	85.74	13.28	15.04	70.03	69.81	21.82	86.64	2.61	12.73	73.94	94.75	32.95	86.85	8.94	9.69	73.65	72.90	28.46	87.77
400	10.36	38.18	85.78	96.38	50.36	71.64	12.20	14.23	65.25	65.56	20.45	73.38	2.23	9.66	69.99	92.10	28.98	73.79	7.12	7.84	64.24	64.55	25.90	75.60
600	9.01	30.66	83.31	94.65	46.05	57.88	11.31	13.56	61.39	62.41	19.33	60.37	1.72	5.81	65.97	88.96	26.00	61.04	6.47	7.24	57.57	59.46	24.28	63.74
800	7.68	22.08	80.22	91.89	41.97	44.93	10.65	13.11	57.83	59.62	18.42	47.97	1.33	3.46	62.26	85.06	23.73	48.93	6.44	7.24	53.60	56.45	23.17	52.48
1000	6.57	15.14	75.95	87.71	38.13	33.49	10.23	12.83	54.67	57.22	17.64	36.84	1.07	2.13	58.04	78.98	21.75	37.94	6.67	7.49	50.47	53.68	22.27	42.11
1200	5.64	10.44	71.41	82.43	34.66	24.20	10.05	12.74	52.30	55.32	16.98	27.72	0.90	1.48	54.52	71.60	19.98	28.51	7.13	7.97	48.31	51.47	21.46	33.02
1400	4.86	7.46	66.94	76.88	31.51	17.24	10.09	12.79	50.84	54.03	16.41	20.82	0.81	1.15	51.21	63.38	18.37	20.96	7.90	8.81	47.03	49.90	20.72	25.65
1600	4.25	5.59	63.48	71.81	28.69	12.27	10.28	12.99	50.20	53.37	15.95	15.84	0.76	0.99	49.37	56.65	16.90	15.20	8.81	9.77	46.88	49.36	20.05	19.99
1800	3.77	4.37	60.59	67.37	26.21	8.82	10.57	13.26	50.11	53.05	15.56	12.37	0.74	0.91	48.38	50.96	15.60	10.87	9.69	10.70	47.67	49.45	19.42	15.77
2000	3.38	3.54	58.38	63.85	24.04	6.52	10.86	13.54	50.05	52.83	15.24	10.04	0.74	0.86	47.53	45.64	14.42	7.73	10.46	11.51	48.06	49.45	18.89	12.70

Table 22: Faithfulness across Top-P and Top-N thresholds for LAAT.

Dataset	MIMIC-IV ICD10 Full						MIMIC-IV ICD10 Top-50						MIMIC-III ICD9 Full						MIMIC-III ICD9 Top-50					
Threshold	**F1-Mac**	**F1-Mic**	**AUC-Mac**	**AUC-Mic**	**P@8**	**Retention**	**F1-Mac**	**F1-Mic**	**AUC-Mac**	**AUC-Mic**	**P@5**	**Retention**	**F1-Mac**	**F1-Mic**	**AUC-Mac**	**AUC-Mic**	**P@8**	**Retention**	**F1-Mac**	**F1-Mic**	**AUC-Mac**	**AUC-Mic**	**P@5**	**Retention**
-	20.88	57.86	95.25	99.02	68.98	100.00	67.50	72.93	92.76	95.22	64.68	100.00	16.89	56.53	89.94	98.59	74.35	100.00	59.51	67.58	90.27	93.06	64.23	100.00
Top-p Sufficiency																								
10%	11.45	44.53	86.80	95.28	58.37	11.27	54.99	62.98	85.39	88.63	46.10	15.52	11.37	46.26	85.83	97.08	64.55	10.21	52.02	58.64	83.99	87.42	46.96	13.64
20%	13.25	47.59	91.60	97.41	60.88	21.13	56.29	64.03	87.06	90.26	46.77	24.90	12.50	48.08	88.33	98.03	66.49	20.19	52.83	60.24	86.34	89.58	48.04	23.24
30%	14.50	49.56	93.32	98.26	62.45	30.99	57.52	65.06	88.26	91.31	47.26	34.29	13.32	49.43	89.27	98.33	67.78	30.17	53.69	61.36	87.34	90.41	48.71	32.83
40%	15.41	50.97	94.22	98.67	63.57	40.85	58.62	65.98	89.18	92.10	47.68	43.68	13.93	50.48	89.71	98.46	68.73	40.14	54.43	62.26	87.90	90.93	49.18	42.43
50%	16.13	52.03	94.77	98.86	64.41	50.71	59.59	66.80	89.89	92.75	48.05	53.07	14.38	51.28	89.82	98.51	69.45	50.12	55.01	62.94	88.42	91.41	49.55	52.02
60%	19.97	56.98	95.00	98.94	68.21	60.56	64.36	70.85	90.47	93.27	50.09	62.45	14.76	51.96	89.88	98.55	70.07	60.09	55.61	63.56	88.72	91.67	49.83	61.62
70%	20.26	57.19	95.14	98.99	68.42	70.42	64.75	71.14	90.98	93.70	50.32	71.84	15.06	52.51	89.99	98.58	70.59	70.07	56.09	64.05	88.95	91.90	50.03	71.21
80%	20.42	57.35	95.21	99.01	68.56	80.28	65.12	71.40	91.42	94.07	50.50	81.23	15.31	52.98	89.97	98.59	71.02	80.05	56.48	64.46	89.16	92.10	50.20	80.81
90%	20.54	57.47	95.24	99.02	68.66	90.14	65.44	71.62	91.83	94.42	50.68	90.61	15.52	53.36	89.94	98.59	71.38	90.02	56.78	64.77	89.45	92.38	50.36	90.40
Top-p Comprehensiveness																								
10%	5.90	14.43	85.66	91.47	21.73	90.14	15.77	20.19	69.27	70.31	25.00	90.61	2.39	7.33	80.53	92.92	23.70	90.02	7.68	7.96	62.84	63.57	19.37	90.41
20%	4.34	10.77	77.83	83.95	17.50	80.28	15.22	19.52	66.17	68.42	24.11	81.23	1.88	5.67	73.83	87.92	19.30	80.05	8.41	8.97	59.67	61.86	18.67	80.81
30%	3.27	7.78	72.15	77.07	14.34	70.42	14.71	18.97	65.25	67.92	23.54	71.84	1.51	4.65	68.47	83.20	16.41	70.07	8.89	9.68	56.78	59.30	17.97	71.21
40%	2.56	5.65	67.33	71.21	11.99	60.56	14.24	18.45	63.84	66.99	23.01	62.45	1.25	3.89	64.77	78.88	14.28	60.09	9.18	10.32	54.80	57.49	17.38	61.62
50%	2.06	4.21	63.59	66.04	10.22	50.70	13.77	17.94	62.40	65.94	22.50	53.07	1.08	3.35	60.95	75.10	12.68	50.12	9.54	11.06	53.46	56.25	16.95	52.02
60%	1.71	3.23	60.37	61.71	8.88	40.85	10.87	15.05	61.12	64.87	19.57	43.68	0.95	2.91	58.32	72.20	11.46	40.14	9.97	11.90	52.96	55.64	16.67	42.43
70%	1.45	2.58	57.71	58.40	7.85	30.99	10.38	14.56	59.83	63.70	19.11	34.29	0.84	2.57	55.99	69.88	10.56	30.17	10.42	12.78	52.18	55.13	16.42	32.83
80%	1.25	2.12	55.03	55.81	7.05	21.13	9.92	14.14	58.43	62.36	18.59	24.90	0.76	2.32	54.76	68.51	9.91	20.19	11.03	13.84	52.02	55.36	16.28	23.24
90%	1.10	1.80	52.68	53.55	6.43	11.27	9.44	13.72	56.58	60.63	18.11	15.52	0.69	2.12	53.34	68.02	9.54	10.21	11.82	14.94	51.16	55.25	16.11	13.64
Top-N Sufficiency																								
200	12.84	46.95	88.98	96.19	60.47	15.89	55.77	63.64	86.06	89.36	46.58	19.41	11.75	46.78	86.56	97.33	65.59	13.54	52.11	59.00	84.54	87.84	46.93	16.57
400	14.54	49.51	92.65	97.88	62.66	30.15	57.42	65.03	87.69	90.96	47.31	32.54	13.01	48.65	88.77	98.15	67.36	26.75	53.35	60.86	86.52	89.81	48.23	28.97
600	15.64	51.16	93.94	98.51	64.01	44.07	58.77	66.18	88.87	92.00	47.87	45.44	13.77	49.93	89.43	98.38	68.50	39.64	54.18	61.93	87.33	90.57	48.93	41.06
800	16.44	52.31	94.58	98.78	64.92	57.14	59.90	67.12	89.85	92.84	48.35	57.75	14.30	50.89	89.81	98.49	69.41	51.86	54.89	62.76	87.86	91.08	49.39	52.55
1000	17.05	53.17	94.91	98.91	65.59	68.68	60.85	67.89	90.64	93.48	48.75	68.84	14.65	51.64	89.88	98.54	70.12	62.96	55.46	63.40	88.28	91.46	49.74	63.13
1200	17.52	53.83	95.08	98.96	66.08	78.02	61.65	68.51	91.22	93.98	49.11	77.97	14.97	52.25	89.98	98.57	70.66	72.47	55.94	63.91	88.74	91.79	50.02	72.37
1400	17.90	54.35	95.16	98.99	66.46	85.01	62.33	69.03	91.67	94.35	49.41	84.89	15.21	52.73	89.98	98.58	71.09	80.07	56.35	64.32	89.05	92.06	50.25	79.90
1600	18.21	54.75	95.21	99.01	66.76	90.01	62.90	69.45	92.01	94.62	49.67	89.89	15.41	53.13	89.97	98.59	71.46	85.87	56.68	64.66	89.36	92.33	50.44	85.69
1800	18.47	55.08	95.23	99.02	67.00	93.48	63.39	69.81	92.27	94.82	49.88	93.38	15.57	53.47	89.95	98.59	71.76	90.23	56.96	64.95	89.60	92.52	50.59	89.99
2000	18.68	55.36	95.24	99.02	67.20	95.79	63.80	70.11	92.44	94.96	50.07	95.73	15.70	53.76	89.98	98.59	72.01	93.39	57.20	65.19	89.83	92.69	50.73	93.11
Top-N Comprehensiveness																								
200	4.72	11.97	81.88	88.43	18.72	85.59	15.43	19.69	67.47	69.24	24.55	86.79	2.18	6.93	78.28	91.80	21.51	86.70	7.80	8.17	59.45	61.31	18.78	87.54
400	3.18	7.68	72.90	78.43	14.45	71.33	14.84	18.98	64.41	67.36	23.64	73.66	1.72	5.39	70.28	85.41	17.43	73.48	8.52	9.12	55.56	58.15	17.72	75.14
600	2.23	5.03	66.68	70.84	11.59	57.42	14.27	18.34	62.68	66.01	22.86	60.75	1.38	4.31	65.42	80.03	14.75	60.60	8.74	9.71	53.09	56.02	16.94	63.05
800	1.68	3.50	61.99	64.58	9.61	44.34	13.68	17.67	60.24	63.85	22.11	48.44	1.13	3.57	61.46	75.33	12.79	48.38	9.05	10.45	51.28	54.31	16.46	51.56
1000	1.32	2.56	58.09	59.34	8.21	32.80	13.20	17.15	57.50	61.26	21.41	37.36	0.94	2.97	57.55	70.94	11.42	37.29	9.65	11.40	50.56	53.52	16.23	40.99
1200	1.08	1.96	55.20	55.40	7.16	23.46	12.88	16.78	54.74	58.60	20.74	28.23	0.79	2.49	54.98	67.07	10.43	27.79	10.42	12.42	49.43	52.42	16.03	31.74
1400	0.92	1.58	53.29	52.91	6.37	16.47	12.73	16.62	52.94	56.87	20.15	21.30	0.69	2.14	52.21	64.16	9.76	20.20	11.30	13.45	48.94	51.89	15.85	24.21
1600	0.81	1.34	51.77	51.57	5.74	11.47	12.77	16.64	52.16	56.09	19.64	16.30	0.62	1.87	50.90	61.69	9.25	14.41	12.20	14.42	48.94	51.60	15.71	18.42
1800	0.72	1.16	51.17	50.82	5.24	8.01	12.90	16.77	51.97	55.86	19.20	12.81	0.56	1.67	50.13	60.20	8.88	10.06	13.03	15.25	49.16	51.61	15.56	14.12
2000	0.66	1.04	50.79	50.41	4.84	5.69	13.09	16.98	52.10	56.01	18.82	10.46	0.52	1.52	49.07	58.83	8.58	6.90	13.76	15.96	49.13	51.34	15.40	11.00

Table 23: Faithfulness across Top-P and Top-N thresholds for PLM-ICD.

Dataset	MIMIC-IV ICD10 Full						MIMIC-IV ICD10 Top-50						MIMIC-III ICD9 Full						MIMIC-III ICD9 Top-50					
Threshold	**F1-Mac**	**F1-Mic**	**AUC-Mac**	**AUC-Mic**	**P@8**	**Retention**	**F1-Mac**	**F1-Mic**	**AUC-Mac**	**AUC-Mic**	**P@5**	**Retention**	**F1-Mac**	**F1-Mic**	**AUC-Mac**	**AUC-Mic**	**P@8**	**Retention**	**F1-Mac**	**F1-Mic**	**AUC-Mac**	**AUC-Mic**	**P@5**	**Retention**
-	19.70	58.32	96.54	99.23	69.79	100.00	68.20	73.42	93.30	95.50	65.35	100.00	16.20	56.33	91.39	98.75	73.10	100.00	65.70	71.18	91.96	94.26	66.65	100.00
Top-p Sufficiency																								
10%	6.44	47.05	85.04	95.21	62.86	11.35	63.49	70.12	89.09	91.99	49.64	14.97	12.34	46.90	79.79	95.24	66.51	10.37	58.83	66.03	86.02	88.14	49.45	12.13
20%	7.63	48.99	90.07	96.94	63.98	21.20	64.37	70.62	91.10	93.70	50.18	24.42	13.74	47.72	84.76	96.75	67.03	20.33	61.03	67.52	88.34	90.80	50.66	21.89
30%	8.44	50.22	92.80	97.79	64.69	31.05	65.07	71.06	92.16	94.56	50.60	33.87	14.47	48.70	87.65	97.73	67.62	30.29	61.95	68.18	89.24	91.75	51.23	31.66
40%	8.99	51.10	94.25	98.27	65.24	40.90	65.56	71.40	92.67	94.97	50.89	43.31	15.00	49.58	89.16	98.19	68.11	40.25	62.59	68.62	90.13	92.55	51.66	41.42
50%	10.98	54.48	95.11	98.55	67.60	50.75	67.42	72.76	92.91	95.19	51.99	52.76	15.36	50.34	89.79	98.42	68.56	50.21	63.05	68.92	90.62	93.16	52.01	51.19
60%	11.40	54.85	95.70	98.76	67.90	60.60	67.54	72.87	93.09	95.33	52.08	62.21	15.54	51.02	90.56	98.59	69.04	60.17	63.37	69.17	90.95	93.51	52.27	60.95
70%	11.90	55.27	96.06	98.92	68.19	70.45	67.66	72.98	93.20	95.42	52.13	71.66	15.66	51.60	90.87	98.67	69.48	70.12	63.62	69.36	91.33	93.84	52.48	70.71
80%	12.62	55.70	96.31	99.05	68.45	80.30	67.76	73.06	93.24	95.46	52.18	81.10	15.76	52.07	91.19	98.72	69.85	80.08	63.85	69.56	91.54	93.95	52.65	80.48
90%	13.60	56.14	96.45	99.16	68.68	90.15	67.84	73.13	93.27	95.48	52.22	90.55	15.82	52.48	91.27	98.73	70.19	90.04	64.01	69.69	91.85	94.15	52.79	90.24
Top-p Comprehensiveness																								
10%	7.30	24.37	91.34	96.93	37.00	90.15	15.25	16.14	73.90	73.50	24.12	90.55	10.40	32.43	87.87	97.42	46.89	90.04	38.46	41.69	83.08	84.77	40.01	90.24
20%	5.36	18.05	83.41	93.87	30.82	80.30	12.16	12.73	66.44	66.82	21.74	81.10	8.47	25.96	80.34	95.11	39.04	80.08	32.28	34.74	78.77	79.37	36.65	80.48
30%	4.26	14.22	76.66	91.42	27.09	70.45	10.16	10.54	61.77	63.63	20.44	71.66	7.01	21.37	73.10	93.16	33.48	70.12	27.32	29.38	75.10	74.74	33.91	70.71
40%	3.53	11.69	70.78	89.94	24.74	60.60	8.70	8.96	58.63	61.90	19.64	62.21	5.99	18.10	66.40	91.71	29.53	60.17	23.54	25.31	72.09	71.52	31.89	60.95
50%	3.02	9.90	65.45	89.12	23.24	50.75	7.58	7.78	57.07	61.26	19.15	52.76	5.23	15.69	62.50	89.90	26.54	50.21	20.60	22.17	69.27	68.95	30.31	51.19
60%	2.64	8.56	60.53	88.62	22.35	40.90	6.71	6.86	56.82	61.15	18.82	43.31	4.64	13.77	59.32	89.89	24.49	40.25	18.33	19.74	67.22	67.14	29.10	41.42
70%	2.35	7.52	56.61	88.29	21.83	31.05	6.01	6.13	57.30	61.09	18.55	33.87	4.18	12.25	58.10	89.60	23.03	30.29	16.48	17.75	65.25	65.53	28.12	31.66
80%	2.12	6.69	53.49	88.08	21.56	21.20	5.44	5.53	57.41	60.70	18.29	24.42	3.79	11.01	55.17	87.56	21.89	20.33	14.92	16.05	63.51	64.04	27.23	21.89
90%	1.94	6.02	51.72	87.96	21.39	11.35	4.97	5.04	56.85	60.06	18.02	14.97	3.47	9.99	55.20	86.69	20.98	10.37	13.60	14.62	60.28	62.37	26.39	12.13
Top-N Sufficiency																								
200	6.68	46.56	82.04	94.81	63.10	13.21	63.51	70.14	88.90	91.86	49.70	16.38	12.20	47.59	80.72	95.53	66.93	11.42	57.96	65.62	84.96	87.67	49.35	12.92
400	7.93	48.63	87.60	96.50	64.23	24.86	64.36	70.64	91.01	93.62	50.22	27.21	13.81	48.75	85.80	97.19	67.70	22.39	60.40	67.27	88.02	90.56	50.46	23.43
600	8.69	49.90	90.89	97.38	64.97	36.39	64.99	71.05	92.05	94.46	50.61	37.96	14.54	49.46	87.83	97.70	68.10	33.25	61.59	68.05	89.17	91.68	51.15	33.89
800	9.39	50.82	92.64	97.92	65.54	47.62	65.45	71.35	92.56	94.88	50.89	48.52	15.00	50.17	88.95	98.15	68.54	43.83	62.20	68.43	89.49	92.17	51.53	44.03
1000	10.07	51.57	93.58	98.26	66.03	58.23	65.82	71.61	92.87	95.14	51.11	58.61	15.34	50.78	89.56	98.34	68.92	53.95	62.67	68.76	90.26	92.90	51.87	53.78
1200	10.77	52.19	94.08	98.46	66.43	67.72	66.12	71.83	93.04	95.28	51.27	67.79	15.59	51.35	90.18	98.52	69.32	63.27	63.00	68.97	90.60	93.26	52.13	62.87
1400	11.48	52.73	94.46	98.63	66.77	75.87	66.35	72.00	93.12	95.36	51.40	75.77	15.76	51.83	90.58	98.60	69.69	71.59	63.30	69.18	90.85	93.47	52.36	71.06
1600	12.17	53.20	94.77	98.75	67.06	82.28	66.55	72.14	93.18	95.41	51.51	82.10	15.88	52.24	90.84	98.66	70.01	78.51	63.53	69.36	91.12	93.71	52.53	78.05
1800	12.81	53.62	95.07	98.86	67.30	87.34	66.72	72.27	93.24	95.45	51.60	87.17	15.97	52.61	91.05	98.70	70.29	84.22	63.73	69.53	91.29	93.85	52.69	83.80
2000	13.40	54.00	95.35	98.94	67.52	91.04	66.85	72.37	93.27	95.48	51.67	90.90	16.03	52.92	91.17	98.72	70.54	88.72	63.91	69.68	91.49	93.95	52.82	88.40
Top-N Comprehensiveness																								
200	7.62	25.40	86.84	96.09	36.33	88.29	15.97	16.63	73.93	73.84	24.28	89.14	10.08	32.48	85.62	96.80	45.67	88.99	40.28	43.28	83.21	84.77	40.15	89.45
400	5.78	19.40	77.83	93.40	30.79	76.64	12.82	13.18	66.56	67.30	22.02	78.31	8.43	26.82	77.48	94.37	38.55	78.02	33.70	36.13	79.05	79.66	36.90	78.94
600	4.63	15.58	71.39	91.60	27.45	65.12	10.76	10.96	61.95	64.05	20.75	67.56	7.14	22.64	71.13	92.77	33.69	67.16	28.81	30.85	75.42	75.48	34.27	68.48
800	3.86	12.96	65.87	90.45	25.38	53.89	9.29	9.38	58.48	62.18	19.93	57.01	6.19	19.50	66.61	91.73	30.19	56.59	25.24	26.91	72.29	72.17	32.33	58.34
1000	3.31	11.06	62.46	89.71	24.04	43.29	8.16	8.20	56.39	61.22	19.37	46.91	5.44	17.09	63.23	91.21	27.64	46.48	22.32	23.82	69.51	69.70	30.77	48.59
1200	2.90	9.63	59.59	89.23	23.18	33.79	7.29	7.30	55.73	60.91	18.97	37.74	4.88	15.17	61.40	90.65	25.64	37.15	20.03	21.40	67.39	67.87	29.59	39.50
1400	2.59	8.51	57.53	88.87	22.59	25.66	6.57	6.56	55.54	60.65	18.64	29.78	4.40	13.61	59.70	90.39	24.08	28.85	18.15	19.40	64.78	66.07	28.57	31.31
1600	2.34	7.61	56.37	88.60	22.19	19.26	5.98	5.96	56.16	60.63	18.37	23.44	4.01	12.32	58.11	89.83	22.84	21.93	16.54	17.70	62.89	64.86	27.67	24.33
1800	2.13	6.88	55.84	88.40	21.92	14.20	5.49	5.46	56.70	60.48	18.14	18.39	3.68	11.23	56.84	89.46	21.86	16.23	15.20	16.28	61.31	63.78	26.92	18.57
2000	1.97	6.27	54.88	88.26	21.72	10.5	5.08	5.04	57.10	60.41	17.93	14.66	3.40	10.30	56.25	89.02	21.01	11.73	14.05	15.04	59.85	63.07	26.27	13.97

Table 24: Plausibility results of CAML across all thresholds.

(a) Results using the Top-p selection strategy.

Metric	Threshold	#Prd	#Act	TP	FP	FN	Pr	Rc	F1
Exact SM	0.1	84520	2269	55	84465	2214	0.1%	2.4%	0.1%
	0.2	146091	2269	44	146047	2225	0.0%	1.9%	0.1%
	0.3	189001	2269	43	188958	2226	0.0%	1.9%	0.0%
	0.4	215458	2269	32	215426	2237	0.0%	1.4%	0.0%
	0.5	225795	2269	29	225766	2240	0.0%	1.3%	0.0%
	0.6	220015	2269	24	219991	2245	0.0%	1.1%	0.0%
	0.7	195599	2269	14	195585	2255	0.0%	0.6%	0.0%
	0.8	154368	2269	8	154360	2261	0.0%	0.4%	0.0%
	0.9	93383	2269	3	93380	2266	0.0%	0.1%	0.0%
PI SM	0.1	77970	2021	88	77882	1933	0.1%	4.4%	0.2%
	0.2	133104	2021	91	133013	1930	0.1%	4.5%	0.1%
	0.3	172526	2021	96	172430	1925	0.1%	4.8%	0.1%
	0.4	198268	2021	85	198183	1936	0.0%	4.2%	0.1%
	0.5	210205	2021	79	210126	1942	0.0%	3.9%	0.1%
	0.6	207520	2021	68	207452	1953	0.0%	3.4%	0.1%
	0.7	187097	2021	54	187043	1967	0.0%	2.7%	0.1%
	0.8	149936	2021	38	149898	1983	0.0%	1.9%	0.1%
	0.9	91959	2021	19	91940	2002	0.0%	0.9%	0.0%
Exact TM	0.1	150968	11428	2493	148475	8935	1.7%	21.8%	3.1%
	0.2	299351	11428	3522	295829	7906	1.2%	30.8%	2.3%
	0.3	445550	11428	4216	441334	7212	0.9%	36.9%	1.8%
	0.4	590911	11428	4660	586251	6768	0.8%	40.8%	1.5%
	0.5	735001	11428	5010	729991	6418	0.7%	43.8%	1.3%
	0.6	876874	11428	5347	871527	6081	0.6%	46.8%	1.2%
	0.7	1016760	11428	5619	1011141	5809	0.6%	49.2%	1.1%
	0.8	1155761	11428	6073	1149688	5355	0.5%	53.1%	1.0%
	0.9	1293460	11428	6065	1287395	5363	0.5%	53.1%	0.9%
PI TM	0.1	114766	9367	4029	110737	5338	3.5%	43.0%	6.5%
	0.2	200255	9367	5190	195065	4177	2.6%	55.4%	5.0%
	0.3	271137	9367	5931	265206	3436	2.2%	63.3%	4.2%
	0.4	331052	9367	6331	324721	3036	1.9%	67.6%	3.7%
	0.5	382967	9367	6692	376275	2675	1.7%	71.4%	3.4%
	0.6	426343	9367	6998	419345	2369	1.6%	74.7%	3.2%
	0.7	462182	9367	7197	454985	2170	1.6%	76.8%	3.1%
	0.8	490358	9367	7423	482935	1944	1.5%	79.2%	3.0%
	0.9	514078	9367	7481	506597	1886	1.5%	79.9%	2.9%

(b) Results using the Top-N selection strategy.

Metric	Threshold	#Prd	#Act	TP	FP	FN	Pr	Rc	F1
Exact SM	200	103366	2269	52	103314	2217	0.1%	2.3%	0.1%
	400	169040	2269	38	169002	2231	0.0%	1.7%	0.0%
	600	204444	2269	31	204413	2238	0.0%	1.4%	0.0%
	800	209526	2269	25	209501	2244	0.0%	1.1%	0.0%
	1000	185153	2269	22	185131	2247	0.0%	1.0%	0.0%
	1200	143952	2269	14	143938	2255	0.0%	0.6%	0.0%
	1400	102270	2269	8	102262	2261	0.0%	0.4%	0.0%
	1600	67889	2269	6	67883	2263	0.0%	0.3%	0.0%
	1800	38832	2269	5	38827	2264	0.0%	0.2%	0.0%
	2000	21988	2269	1	21987	2268	0.0%	0.0%	0.0%
PI SM	200	95197	2021	84	95113	1937	0.1%	4.2%	0.2%
	400	154719	2021	89	154630	1932	0.1%	4.4%	0.1%
	600	188430	2021	85	188345	1936	0.0%	4.2%	0.1%
	800	195157	2021	71	195086	1950	0.0%	3.5%	0.1%
	1000	174043	2021	59	173984	1962	0.0%	2.9%	0.1%
	1200	136261	2021	43	136218	1978	0.0%	2.1%	0.1%
	1400	97396	2021	28	97368	1993	0.0%	1.4%	0.1%
	1600	65016	2021	22	64994	1999	0.0%	1.1%	0.1%
	1800	37116	2021	14	37102	2007	0.0%	0.7%	0.1%
	2000	20846	2021	10	20836	2011	0.0%	0.5%	0.1%
Exact TM	200	194791	11428	2820	191971	8608	1.4%	24.7%	2.7%
	400	385838	11428	3929	381909	7499	1.0%	34.4%	2.0%
	600	573507	11428	4628	568879	6800	0.8%	40.5%	1.6%
	800	758757	11428	5183	753574	6245	0.7%	45.4%	1.3%
	1000	938428	11428	5623	932805	5805	0.6%	49.2%	1.2%
	1200	1088084	11428	5560	1082524	5868	0.5%	48.7%	1.0%
	1400	1206408	11428	5652	1200756	5776	0.5%	49.5%	0.9%
	1600	1284292	11428	5953	1278339	5475	0.5%	52.1%	0.9%
	1800	1337091	11428	5577	1331514	5851	0.4%	48.8%	0.8%
	2000	1365475	11428	5404	1360071	6024	0.4%	47.3%	0.8%
PI TM	200	142740	9367	4470	138270	4897	3.1%	47.7%	5.9%
	400	244099	9367	5692	238407	3675	2.3%	60.8%	4.5%
	600	324089	9367	6374	317715	2993	2.0%	68.0%	3.8%
	800	388125	9367	6818	381307	2549	1.8%	72.8%	3.4%
	1000	439775	9367	7095	432680	2272	1.6%	75.7%	3.2%
	1200	481563	9367	7432	474131	1935	1.5%	79.3%	3.0%
	1400	509233	9367	7615	501618	1752	1.5%	81.3%	2.9%
	1600	530370	9367	7885	522485	1482	1.5%	84.2%	2.9%
	1800	543222	9367	7988	535234	1379	1.5%	85.3%	2.9%
	2000	551266	9367	8080	543186	1287	1.5%	86.3%	2.9%

Table 25: Plausibility results of LAAT across all thresholds.

(a) Results using the Top-p selection strategy.

Metric	Threshold	#Prd	#Act	TP	FP	FN	Pr	Rc	F1
Exact SM	0.1	91482	2269	330	91152	1939	0.4%	14.5%	0.7%
	0.2	152195	2269	183	152012	2086	0.1%	8.1%	0.2%
	0.3	190875	2269	107	190768	2162	0.1%	4.7%	0.1%
	0.4	211973	2269	62	211911	2207	0.0%	2.7%	0.1%
	0.5	217293	2269	40	217253	2229	0.0%	1.8%	0.0%
	0.6	207522	2269	29	207493	2240	0.0%	1.3%	0.0%
	0.7	182928	2269	15	182913	2254	0.0%	0.7%	0.0%
	0.8	143185	2269	3	143182	2266	0.0%	0.1%	0.0%
	0.9	87373	2269	1	87372	2268	0.0%	0.0%	0.0%
PI SM	0.1	82542	2021	342	82200	1679	0.4%	16.9%	0.8%
	0.2	138074	2021	223	137851	1798	0.2%	11.0%	0.3%
	0.3	174708	2021	145	174563	1876	0.1%	7.2%	0.2%
	0.4	196066	2021	102	195964	1919	0.1%	5.0%	0.1%
	0.5	203071	2021	74	202997	1947	0.0%	3.7%	0.1%
	0.6	195899	2021	55	195844	1966	0.0%	2.7%	0.1%
	0.7	174663	2021	41	174622	1980	0.0%	2.0%	0.0%
	0.8	138486	2021	24	138462	1997	0.0%	1.2%	0.0%
	0.9	85688	2021	11	85677	2010	0.0%	0.5%	0.0%
Exact TM	0.1	160732	11428	4266	156466	7162	2.7%	37.3%	5.0%
	0.2	316121	11428	4840	311281	6588	1.5%	42.4%	3.0%
	0.3	467215	11428	5258	461957	6170	1.1%	46.0%	2.2%
	0.4	615384	11428	5468	609916	5960	0.9%	47.8%	1.7%
	0.5	759735	11428	5539	754196	5889	0.7%	48.5%	1.4%
	0.6	901596	11428	5961	895635	5467	0.7%	52.2%	1.3%
	0.7	1040440	11428	6142	1034298	5286	0.6%	53.7%	1.2%
	0.8	1175547	11428	6022	1169525	5406	0.5%	52.7%	1.0%
	0.9	1303770	11428	5812	1297958	5616	0.4%	50.9%	0.9%
PI TM	0.1	121292	9367	4720	116572	4647	3.9%	50.4%	7.2%
	0.2	212105	9367	5732	206373	3635	2.7%	61.2%	5.2%
	0.3	284284	9367	6234	278050	3133	2.2%	66.6%	4.2%
	0.4	344568	9367	6705	337863	2662	1.9%	71.6%	3.8%
	0.5	394117	9367	6979	387138	2388	1.8%	74.5%	3.5%
	0.6	436327	9367	7249	429078	2118	1.7%	77.4%	3.3%
	0.7	470428	9367	7420	463008	1947	1.6%	79.2%	3.1%
	0.8	498098	9367	7545	490553	1822	1.5%	80.5%	3.0%
	0.9	523255	9367	7637	515618	1730	1.5%	81.5%	2.9%

(b) Results using the Top-N selection strategy.

Metric	Threshold	#Prd	#Act	TP	FP	FN	Pr	Rc	F1
Exact SM	200	109729	2269	291	109438	1978	0.3%	12.8%	0.5%
	400	171537	2269	124	171413	2145	0.1%	5.5%	0.1%
	600	200746	2269	75	200671	2194	0.0%	3.3%	0.1%
	800	202268	2269	38	202230	2231	0.0%	1.7%	0.0%
	1000	178051	2269	19	178032	2250	0.0%	0.8%	0.0%
	1200	138551	2269	12	138539	2257	0.0%	0.5%	0.0%
	1400	97825	2269	4	97821	2265	0.0%	0.2%	0.0%
	1600	64925	2269	3	64922	2266	0.0%	0.1%	0.0%
	1800	37745	2269	2	37743	2267	0.0%	0.1%	0.0%
	2000	21659	2269	3	21656	2266	0.0%	0.1%	0.0%
PI SM	200	99384	2021	305	99079	1716	0.3%	15.1%	0.6%
	400	156766	2021	165	156601	1856	0.1%	8.2%	0.2%
	600	185712	2021	120	185592	1901	0.1%	5.9%	0.1%
	800	189014	2021	72	188942	1949	0.0%	3.6%	0.1%
	1000	167363	2021	51	167312	1970	0.0%	2.5%	0.1%
	1200	130794	2021	38	130756	1983	0.0%	1.9%	0.1%
	1400	93001	2021	14	92987	2007	0.0%	0.7%	0.0%
	1600	61976	2021	7	61969	2014	0.0%	0.3%	0.0%
	1800	35988	2021	5	35983	2016	0.0%	0.2%	0.0%
	2000	20502	2021	5	20497	2016	0.0%	0.2%	0.0%
Exact TM	200	206362	11428	4482	201880	6946	2.2%	39.2%	4.1%
	400	404155	11428	5093	399062	6335	1.3%	44.6%	2.5%
	600	596166	11428	5414	590752	6014	0.9%	47.4%	1.8%
	800	782134	11428	5637	776497	5791	0.7%	49.3%	1.4%
	1000	959318	11428	5701	953617	5727	0.6%	49.9%	1.2%
	1200	1104095	11428	5811	1098284	5617	0.5%	50.8%	1.0%
	1400	1217029	11428	5727	1211302	5701	0.5%	50.1%	0.9%
	1600	1294041	11428	5943	1288098	5485	0.5%	52.0%	0.9%
	1800	1343294	11428	5736	1337558	5692	0.4%	50.2%	0.8%
	2000	1368335	11428	5417	1362918	6011	0.4%	47.4%	0.8%
PI TM	200	150423	9367	5104	145319	4263	3.4%	54.5%	6.4%
	400	255097	9367	6039	249058	3328	2.4%	64.5%	4.6%
	600	336224	9367	6659	329565	2708	2.0%	71.1%	3.9%
	800	400263	9367	7100	393163	2267	1.8%	75.8%	3.5%
	1000	452497	9367	7337	445160	2030	1.6%	78.3%	3.2%
	1200	490632	9367	7612	483020	1755	1.6%	81.3%	3.0%
	1400	515576	9367	7784	507792	1583	1.5%	83.1%	3.0%
	1600	535308	9367	7933	527375	1434	1.5%	84.7%	2.9%
	1800	547300	9367	8051	539249	1316	1.5%	86.0%	2.9%
	2000	552280	9367	8042	544238	1325	1.5%	85.9%	2.9%

Table 26: Plausibility results of PLM-ICD across all thresholds.

(a) Results using the Top-p selection strategy.

Metric	Threshold	#Prd	#Act	TP	FP	FN	Pr	Rc	F1
Exact SM	10%	65639	2269	172	65467	2097	0.3%	7.6%	0.5%
	20%	99489	2269	112	99377	2157	0.1%	4.9%	0.2%
	30%	117733	2269	68	117665	2201	0.1%	3.0%	0.1%
	40%	123897	2269	48	123849	2221	0.0%	2.1%	0.1%
	50%	123055	2269	38	123017	2231	0.0%	1.7%	0.1%
	60%	112511	2269	23	112488	2246	0.0%	1.0%	0.0%
	70%	96512	2269	14	96498	2255	0.0%	0.6%	0.0%
	80%	72706	2269	8	72698	2261	0.0%	0.4%	0.0%
	90%	44799	2269	2	44797	2267	0.0%	0.1%	0.0%
PI SM	10%	60851	2041	228	60623	1813	0.4%	11.2%	0.7%
	20%	91601	2041	173	91428	1868	0.2%	8.5%	0.4%
	30%	108523	2041	125	108398	1916	0.1%	6.1%	0.2%
	40%	114615	2041	109	114506	1932	0.1%	5.3%	0.2%
	50%	113966	2041	89	113877	1952	0.1%	4.4%	0.2%
	60%	104917	2041	63	104854	1978	0.1%	3.1%	0.1%
	70%	90569	2041	46	90523	1995	0.1%	2.3%	0.1%
	80%	68893	2041	29	68864	2012	0.0%	1.4%	0.1%
	90%	42952	2041	13	42939	2028	0.0%	0.6%	0.1%
Exact TM	10%	173633	12517	3975	169658	8542	2.3%	31.8%	4.3%
	20%	332771	12517	4194	328577	8323	1.3%	33.5%	2.4%
	30%	486349	12517	4251	482098	8266	0.9%	34.0%	1.7%
	40%	641050	12517	4256	636794	8261	0.7%	34.0%	1.3%
	50%	795367	12517	4168	791199	8349	0.5%	33.3%	1.0%
	60%	949630	12517	4015	945615	8502	0.4%	32.1%	0.8%
	70%	1100798	12517	4189	1096609	8328	0.4%	33.5%	0.8%
	80%	1246480	12517	4106	1242374	8411	0.3%	32.8%	0.7%
	90%	1381492	12517	4208	1377284	8309	0.3%	33.6%	0.6%
PI TM	10%	120777	10269	5794	114983	4475	4.8%	56.4%	8.8%
	20%	203259	10269	6548	196711	3721	3.2%	63.8%	6.1%
	30%	270926	10269	7060	263866	3209	2.6%	68.8%	5.0%
	40%	332829	10269	7461	325368	2808	2.2%	72.7%	4.3%
	50%	388647	10269	7665	380982	2604	2.0%	74.6%	3.8%
	60%	439645	10269	7835	431810	2434	1.8%	76.3%	3.5%
	70%	485691	10269	8082	477609	2187	1.7%	78.7%	3.3%
	80%	526864	10269	8145	518719	2124	1.5%	79.3%	3.0%
	90%	560856	10269	8233	552623	2036	1.5%	80.2%	2.9%

(b) Results using the Top-N selection strategy.

Metric	Threshold	#Prd	#Act	TP	FP	FN	Pr	Rc	F1
Exact SM	200	67058	2269	170	66888	2099	0.3%	7.5%	0.5%
	400	99646	2269	98	99548	2171	0.1%	4.3%	0.2%
	600	115216	2269	68	115148	2201	0.1%	3.0%	0.1%
	800	118029	2269	45	117984	2224	0.0%	2.0%	0.1%
	1000	112128	2269	31	112097	2238	0.0%	1.4%	0.1%
	1200	99278	2269	18	99260	2251	0.0%	0.8%	0.0%
	1400	81722	2269	13	81709	2256	0.0%	0.6%	0.0%
	1600	64482	2269	5	64477	2264	0.0%	0.2%	0.0%
	1800	43416	2269	4	43412	2265	0.0%	0.2%	0.0%
	2000	30369	2269	1	30368	2268	0.0%	0.0%	0.0%
PI SM	200	62197	2041	225	61972	1816	0.4%	11.0%	0.7%
	400	92102	2041	161	91941	1880	0.2%	7.9%	0.3%
	600	106563	2041	125	106438	1916	0.1%	6.1%	0.2%
	800	109437	2041	93	109344	1948	0.1%	4.6%	0.2%
	1000	104052	2041	79	103973	1962	0.1%	3.9%	0.1%
	1200	92471	2041	55	92416	1986	0.1%	2.7%	0.1%
	1400	76305	2041	40	76265	2001	0.1%	2.0%	0.1%
	1600	60330	2041	25	60305	2016	0.0%	1.2%	0.1%
	1800	40620	2041	22	40598	2019	0.1%	1.1%	0.1%
	2000	28513	2041	14	28499	2027	0.0%	0.7%	0.1%
Exact TM	200	182545	12517	3972	178573	8545	2.2%	31.7%	4.1%
	400	348662	12517	4175	344487	8342	1.2%	33.4%	2.3%
	600	509830	12517	4253	505577	8264	0.8%	34.0%	1.6%
	800	672002	12517	4313	667689	8204	0.6%	34.5%	1.3%
	1000	834402	12517	4207	830195	8310	0.5%	33.6%	1.0%
	1200	989036	12517	4262	984774	8255	0.4%	34.0%	0.9%
	1400	1121881	12517	4068	1117813	8449	0.4%	32.5%	0.7%
	1600	1218150	12517	4272	1213878	8245	0.4%	34.1%	0.7%
	1800	1285721	12517	4576	1281145	7941	0.4%	36.6%	0.7%
	2000	1335649	12517	4828	1330821	7689	0.4%	38.6%	0.7%
PI TM	200	126983	10269	5801	121182	4468	4.6%	56.5%	8.5%
	400	212631	10269	6635	205996	3634	3.1%	64.6%	6.0%
	600	283642	10269	7157	276485	3112	2.5%	69.7%	4.9%
	800	346990	10269	7527	339463	2742	2.2%	73.3%	4.2%
	1000	404689	10269	7725	396964	2544	1.9%	75.2%	3.7%
	1200	453426	10269	7908	445518	2361	1.7%	77.0%	3.4%
	1400	491786	10269	7968	483818	2301	1.6%	77.6%	3.2%
	1600	518079	10269	8153	509926	2116	1.6%	79.4%	3.1%
	1800	533761	10269	8111	525650	2158	1.5%	79.0%	3.0%
	2000	547686	10269	8052	539634	2217	1.5%	78.4%	2.9%