# Concept-Enhanced Automatic ICD Coding using Large Language Models

**Md Shahrar Fatemi** MFATEMI@CS.STONYBROOK.EDU
*Stony Brook University, USA*

**Zhan Shi** SHZHAN@CS.STONYBROOK.EDU
*Stony Brook University, USA*

**Joel Saltz** JOEL.SALTZ@STONYBROOKMEDICINE.EDU
*Stony Brook University, USA*

**Klaus Mueller** MUELLER@CS.STONYBROOK.EDU
*Stony Brook University, USA*

**Tengfei Ma** TENGFEI.MA@STONYBROOKMEDICINE.EDU
*Stony Brook University, USA*

## Abstract

Automatic ICD coding is a task which assigns disease or procedure codes to clinical notes from patients' electronic health record data. Large language models have been explored for this task, but none of the existing approaches have shown stronger performance than traditional deep learning models due to limited ability to model concepts. Existing methods for ICD coding often utilize the code descriptions or synonyms to enhance performance. In this paper, we propose to use concepts to expand the label space. Utilizing the hierarchy of ICD codes, we construct concepts associated with the codes at different levels, and employ finetuned large language models to obtain concept scores, which are then used for code prediction. Experiments conducted on MIMIC-III-50, and MIMIC-III-rare50 datasets demonstrate that our models achieve excellent performance and largely outperform previous state-of-the-art models. While the current evaluation is constrained in scope and computational tractability, the results provide strong evidence for the potential of concept-driven LLM frameworks to advance automated medical coding.

**Keywords:** ICD Coding, Medical Concepts, Large Language Models (LLMs).

**Data and Code Availability** The experiments in this study were conducted on the publicly available MIMIC-III Clinical Database (Johnson et al., 2016). The code used to run the experiments is available in the GitHub repository [1].

**Institutional Review Board (IRB)** No IRB approval is needed.

## 1. Introduction

Electronic Health Record (EHR) systems have been increasingly employed in US hospitals to maintain and manage patients' data. ICD (International Classification of Diseases) coding is a critical task for EHR management to transform clinical notes into ICD codes to structurally represent appropriate diseases and procedures. The extracted ICD codes can be not only used for insurance companies for reimbursement purposes but also help government agencies to conduct analyses in policy making.

Traditional ICD coding tasks generally rely on a team of experts to manually annotate the medical notes, but they highly rely on expertise and they are also not overly time-efficient. As deep learning has achieved great success in many healthcare problems, it has also been used to automate ICD coding (Mullenbach et al., 2018; Shi et al., 2017; Li and Yu, 2020; Rios and Kavuluru, 2018). Existing approaches often regard this task as a multi-label classification task and utilize some domain knowledge to better represent the codes and enhance the performance, including using code descriptions (Yang et al., 2022),

[1] https://github.com/shahrarfatemi/CEC

code synonyms (Yuan et al., 2022) and code hierarchies (Xie and Xing, 2018; Cao et al., 2020; Lu et al., 2023). However, there remain many challenges, such as rare codes and limited training data. Very recently, large language models have been explored in the ICD coding task (Agrawal et al., 2022; Boyle et al., 2023; Huang et al., 2022; Yang et al., 2023), but surprisingly they do not perform better than the non-pretrained state-of-the-art deep learning models such as MSMN (Yuan et al., 2022). It is claimed that one possible reason is their inefficiency in modeling concept-level information (Dong et al., 2022).

In this paper, we propose the notion of concept-enhanced coding (CEC), which leverages concept-level prediction to improve ICD coding. Concepts are often used in machine learning tasks to enhance interpretability. For example, Concept Bottleneck Models (CBM)(Koh et al., 2020; Oikarinen et al., 2023) construct the prediction of image classification labels from the linear combination of human-annotated concepts. However, these concept-based models focus more on interpretability and they often sacrifice the performance compared to their end-to-end counterparts. In contrast, in the context of ICD coding, we argue that medical concepts associated with ICD codes can not only provide interpretability but also significantly improve the performance of large language models. The concepts for ICD coding are often symptoms of diseases and we extract the concepts which best discriminate different ICD codes (see Figure 1 as an example). Consequently, the concepts for different codes have little overlap, and they are used to expand the label space instead of forming a bottleneck as in CBMs.

We use large language models to first project the medical texts to the concept space and obtain concept scores, and then utilize another large language model to predict the ICD codes. In parallel, considering the inherent hierarchical structure of ICD codes, we generate concepts at different levels. The added concepts from more abstract levels can be used to either enlarge the concept space or to improve the estimation of concept scores. This in turn increases the accuracy of ICD code prediction. While our approach demonstrates strong improvements in predictive accuracy, it requires substantial computational resources to process the expanded concept space and fine-tune large-scale models. Therefore, our current evaluation focuses on the MIMIC-III-50 subset rather than the full dataset, representing a limited but practical scope for experimentation. We evaluate our approach on the MIMIC-III benchmarks using two recent LLM backbones (Llama3-8B (Grattafiori, 2024) and Gemma-7B (GemmaTeam, 2024)). Our models significantly outperform the previous LLM-based models as well as other state-of-the-art (SOTA) models. Specifically, on MIMIC-III-rare-50 our best model achieves over 80% in both Macro F1 and Micro F1, more than double the scores of the next-best model. On MIMIC-III-50, it also exceeds the previous SOTA by over 18% in both metrics.

## 2. Related Work

### 2.1. Pre-LLM Automated ICD Coding Methods

Automatic ICD coding is typically formulated as a multi-label classification problem in many recent deep learning approaches. These methods often utilize CNNs or RNNs to encode clinical texts through attention mechanisms, identifying relevant parts for prediction (Mullenbach et al., 2018; Xie et al., 2019; Li and Yu, 2020; Shi et al., 2017). Subsequent works have increasingly focused on leveraging the structural characteristics of ICD codes and their descriptions for improved representation. For instance, Xie and Xing *et al.*(Xie and Xing, 2018) applied tree-of-sequences LSTM networks to capture the hierarchical relationships among codes, while Vu *et al.*(Vu et al., 2021) proposed a joint hierarchical learning mechanism based on bidirectional LSTM encoder to handle tail codes. Graph Neural Networks (GNNs) have also been explored to enhance modeling capacity (Rios and Kavuluru, 2018). For example, Cao *et al.*(Cao et al., 2020) introduced a co-graph structure that simultaneously considers both code hierarchy and code co-occurrence. Moreover, Yuan *et al.*(Yuan et al., 2022) proposed the Multiple Synonym Matching Network(MSMN) to utilize code synonyms by aligning codes to concepts in UMLS. Building on this, Yang *et al.*(Yang et al., 2022) fine-tuned a pretrained language model with code description prompts and injected knowledge of synonyms, abbreviations, and hierarchy. Following this, Wang *et al.*(Wang et al., 2024a) developed a multi-stage retrieve-and-rerank model that uses external knowledge to retrieve candidate labels and refines them using co-occurrence relationships.

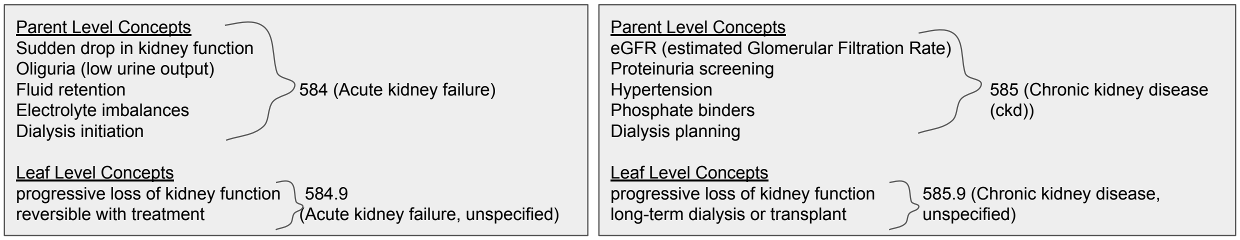

Parent Level Concepts
Sudden drop in kidney function
Oliguria (low urine output)
Fluid retention
Electrolyte imbalances
Dialysis initiation
584 (Acute kidney failure)
Leaf Level Concepts
progressive loss of kidney function
reversible with treatment
584.9
(Acute kidney failure, unspecified)
Parent Level Concepts
eGFR (estimated Glomerular Filtration Rate)
Proteinuria screening
Hypertension
Phosphate binders
Dialysis planning
585 (Chronic kidney disease (ckd))
Leaf Level Concepts
progressive loss of kidney function
long-term dialysis or transplant
585.9 (Chronic kidney disease, unspecified)

Figure 1: Comparative analysis between the concept sets of two ICD Codes: 584.9 and 585.9

### 2.2. Automated ICD Coding with Large Language Models

With the growing success of LLM across various clinical NLP tasks(Agrawal et al., 2022), researchers have started exploring their application to ICD coding. Huang *et al.*(Huang et al., 2022), for instance, used a BERT pretrained on clinical text as an encoder, and incorporated a label-aware attention mechanism to aggregate the representations. Hauksson *et al.* (Hauksson and Einarsson, 2024) examined the fine-tuning of pretrained BERT-based models in low-resource language. Furthermore, Yang *et al.*(Yang et al., 2023) proposed a two-stage model where an LSTM is used to further verify ICD codes predicted by LLMs. Inspired by code hierarchy, Boyle *et al.*(Boyle et al., 2023) prompted the LLMs following a tree-like manner to search the relevant ICD codes, and Wang *et al.*(Wang et al., 2024b) explored the probabilistic label tree decoding. More recent work has directly leveraged large-scale pretrained LLMs such as GPT-3.5, exploring both fine-tuning with clinical data and prompting-based approaches with data augmentation to adapt the model for ICD coding (Nawab et al., 2024; Falis et al., 2024). To improve interpretability and better capture the relationship between clinical features and ICD codes, Wu et al. (Wu et al., 2025) disentangled dense embeddings into a sparse space using dictionary learning, and then leveraged LLMs for dictionary feature identification. In addition, Li et al.(Li et al., 2024) investigated a multi-agent approach that mimics real-world ICD coding workflows by coordinating five specialized LLMs to assign codes in a collaborative manner. However, current LLM-based methods often exhibit lower precision than non-pretrained models such as MSMN, suggesting that their potential in ICD coding remains underexplored. We hypothesize that leveraging clinically meaningful concepts can improve performance, and therefore propose utilizing LLMs to model concept hierarchies rather than directly predicting codes.

## 3. Method

**Problem Definition** The objective is to analyze a clinical note containing medical terminologies, and assign a binary label $y_i \in \{0, 1\}$ to each ICD code in the label set $Y$, where $|Y| = n$. Each candidate code is associated with a short descriptive phrase $d_i$ in plain text. The set of all these code descriptions is denoted as $D$, encompassing all $n_d$ descriptions ($d_i$). In this paper, we use the ICD-9 coding structure[2]. ICD-9 categorizes approximately 13,000 diseases and medical conditions using a hierarchical system.

### 3.1. Overview of our Method

Our proposed method, CEC (Concept-Enhanced Coding) (Figure 2), offers a comprehensive approach to automated ICD coding, leveraging surrounding concepts and their hierarchies. ICD coding draws upon a broad domain of medical knowledge and terminology. This observation motivates the use of a concept-based architecture to perform ICD coding. Alongside this, ICD-9 codes are organized in a complex tree structure, where a major diagnosis or procedure code branches into many specific children codes, which are further subdivided into lower levels. Therefore, determining the leaf-level ICD codes for a medical note requires an understanding of the parent codes at first. The high-level medical concepts present in a note indicate the broader category of parent code it belongs to. Further specifications make it easier to connect the note with the leaf-level codes under the parent category. The example in Figure 1 refers to the concept sets for the ICD-9 codes 584.9 (Acute

[2] http://www.icd9data.com/2015/Volume1/default.htm

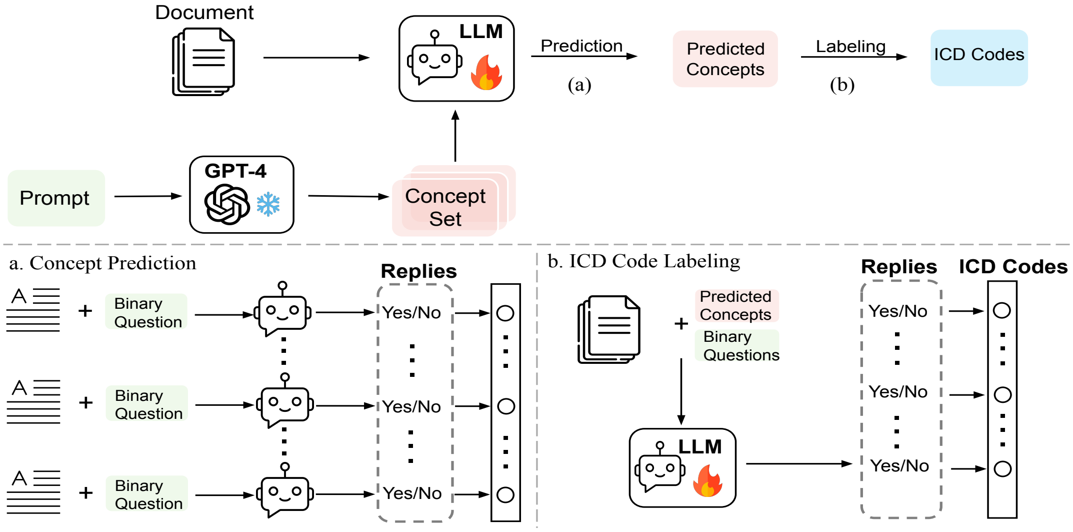


Figure 2: Concept-Enhanced ICD Coding framework. (a) Initial predictions are generated across the concept space. (b) These are then leveraged as intermediate predictions to perform the final labeling task.

kidney failure, unspecified), and 585.9 (Chronic kidney disease, unspecified) respectively. The similarity in their code descriptions is clear; however, the distinction becomes apparent through their parent-level concepts. These concepts highlight the critical difference: one represents a reversible kidney condition, while the other denotes a chronic disease requiring long-term dialysis. To incorporate this idea into ICD coding, we introduce a hierarchical organization of concept sets, consisting of both parent-level, and leaf-level concepts.

Our method (Figure 2) consists of three key steps: Concept Set Construction, Concept Prediction from a Large Language Model (LLM), and Labeling with ICD Codes. Concept set construction involves creating sets of significant and distinct medical concepts related to ICD codes, and is conducted in an offline preprocessing stage. Concept prediction from the LLM uses structured prompts to assess the relevance of concepts within clinical notes, and we fine-tune the model to address knowledge gaps. Finally, we integrate the predicted concept scores of the leaf and parent concepts into the labeling of clinical notes using another LLM model trained for ICD labeling from concept predictions.

### 3.2. Concept Set Construction

Figure 1 shows that differences in parent-level concepts are crucial for resolving ambiguities in leaf-level concepts, highlighting the need for unique and relevant concepts for accurate medical coding. To generate these concept sets, we employ GPT-4 (ChatGPT) (OpenAI, 2024) to generate candidate concepts by providing a detailed medical coding prompt (Details in Appendix B). The model is asked to list relevant medical concepts for each ICD code and its parent code description. This process helps ensure that the concepts capture adequate information for each ICD code. We then refine the extracted concepts to eliminate redundancy. Specifically, we encode each concept phrase into sentence embeddings using all-MiniLM-L6-v2 (Wang et al., 2020) [3]. Pairwise cosine similarities are computed across all concept embeddings, and concepts exceeding a similarity threshold of $\gamma = 0.83$, determined through empirical examination, are considered duplicates. For such highly similar concepts, we retain a single representative phrase (the first occurrence) and discard the rest, ensuring that the final concept set is distinct and non-

[3] https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2

overlapping. For instance, similar phrases like "low blood pressure" and "minimal blood pressure" are filtered to retain only the most representative terms.

Consequently, we construct distinct sets of concepts $C_y$ for each $y \in Y$ that contain the concepts of the parents and the leaves. For MIMIC-III-50, we identify a set $C = C_1 \cup C_2 \cup \cdots \cup C_n$ of 325 concepts (229 parent-level, and 96 leaf-level). For MIMIC-III-rare-50, the total is 332 concepts (232 parent-level, and 100 leaf-level).

### 3.3. Concept Prediction from LLM

Our goal is to determine the relevance of each concept in the concept set for a given clinical note. We design a structured prompt format to predict concept scores from a clinical note. Specifically, the LLM input is formatted as: [Instruction] + [Clinical Note] + [Question] + [Answer Prompt] (Figure 3). We adopt a QA-style format because framing each code decision as an independent yes/no question reduces ambiguity and simplifies the multi-label coding task. The LLM then answers with "yes" or "no" to each question individually.

The main hindrance in case of the binary prediction from the LLM is due to the lack of knowledge of the LLM on concept specific structure. Hence, the LLMs are fine-tuned under the concept set paradigm. For fine-tuning, we construct training sets by extracting the concepts $c \in C_y$ associated with each true label $y$, treating them as pseudo-positive concepts. Additionally, pseudo-negative concepts are obtained by sampling from concepts linked to negative labels (Details in Appendix C). This process yields pseudo-labeled training data, which may inevitably introduce some noise since concept labels are not manually curated. Our framework, in the subsequent stage, leverages these intermediate concept-level predictions and passes to another LLM fine-tuned on direct label supervision, enabling the model to refine and correct noisy intermediate outputs (if any) before producing the final label predictions.

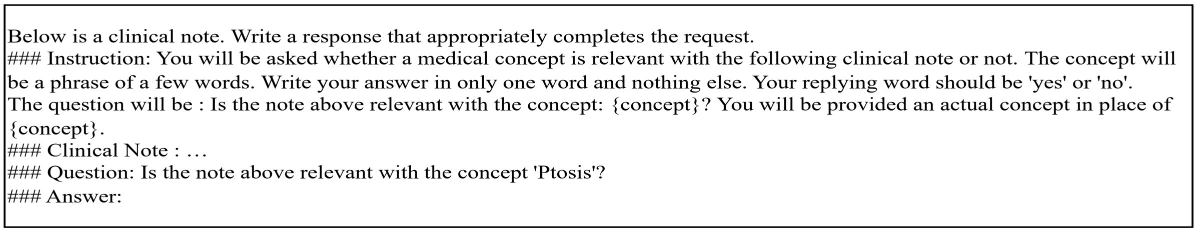
### 3.4. Labeling with ICD Codes

The concepts in the leaf code and its immediate parent code should have direct correlation on the presence of corresponding ICD code in a clinical document. We merge the predictions for all the leaf and parent concepts on a single clinical note together. The set of concepts predicted positively in the previous step is the set of predicted positive concepts ($C'_+$) for that note. The remaining concepts from our concept space are the predicted negative concepts ($C'_-$). The final labeling task consists of using these concept predictions along with the raw clinical note as a prompt to the model. For each code $y \in Y$, we sample $k$ concepts from the combined (both parent, and leaf) concept set $C_y$, and add 2 randomly selected concepts $c_t \in C'_-, c_t \notin C_y$. The predictions ('positive' or 'negative') of these $k + 2$ concepts are included in the prompt to predict the final label of $y$. We prepare this prompt as : [Instruction] + [Clinical Note] + [Concept Predictions] + [Question] + [Answer Prompt] (Figure 4). We repeat this process for all the $n$ codes. The LLM's binary answer gives us the binary label of each $y$, whereas we consider the probit (softmax function applied on the logits of the generated token) as the probability of the presence of $y$ in the note. We keep $k = 3, 4, 5, 6$ to analyze the sensitivity of number of concepts used toward the model's prediction (Table 3).

Similar to the concept prediction task, the model has lack of knowledge regarding the mapping of concept to labels. We have to fine-tune the model to enable it to predict label scores. We create a set of pseudo true concepts for each note from the true labels. For each true label $y$, we consider the concepts in $C_y$ to be pseudo true concepts ($C_+$), and all other concepts in $C \setminus C_y$ to be pseudo false concepts ($C_-$). Then, for each code $l$ in the set of true labels, along with some randomly sampled false labels, we create a prediction dictionary $P$ in the following way.
1) Sample $k = 5$ concepts from $l$'s concept set $C_l$.
2) If $l$ is true, randomly sample $m$ concepts from $C_-$ concepts. If $l$ is false, randomly sample $m$ concepts from $C_+$ concepts.
3) For each concept $c$ in the set constructed in (1), and (2), add "c : Yes/No" in $P$ based on their corresponding label being true/false.
Keeping $m = 2$, we add $P$ in the prompt as the concept predictions, and ask the model to predict the label of $l$. We append the correct answer (Yes/No) in the end during fine-tuning (Details in Appendix C).

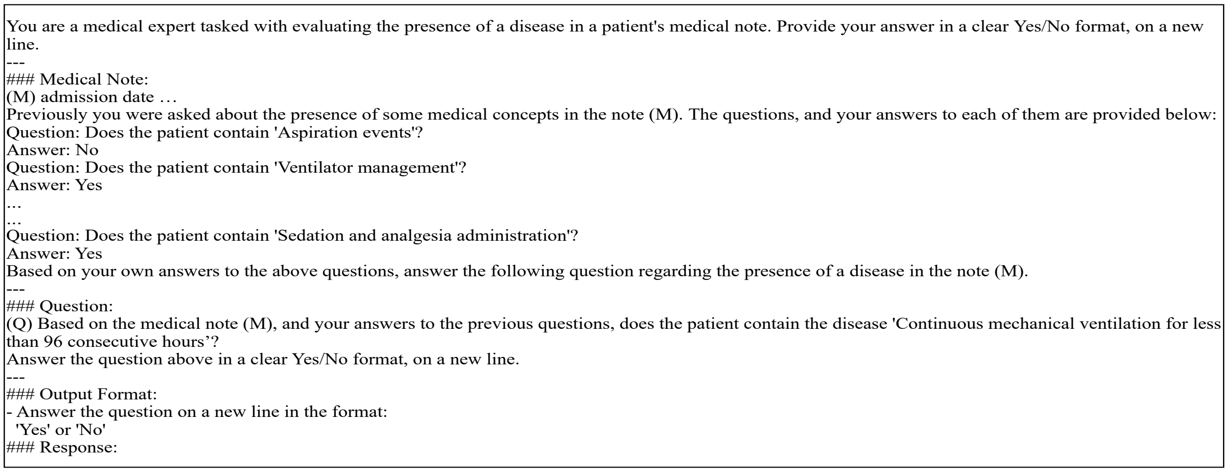
## 4. Experiments

We utilize the publicly available Llama3-8b and Gemma-7b backbones, fine-tuned with LoRA (Hu et al., 2021). Fine-tuning is also applied prior to gen-



Below is a clinical note. Write a response that appropriately completes the request.
### Instruction: You will be asked whether a medical concept is relevant with the following clinical note or not. The concept will be a phrase of a few words. Write your answer in only one word and nothing else. Your replying word should be 'yes' or 'no'.
The question will be : Is the note above relevant with the concept: {concept}? You will be provided an actual concept in place of {concept}.
### Clinical Note : ...
### Question: Is the note above relevant with the concept 'Ptosis'?
### Answer:

Figure 3: Concept prediction prompt to the Large Language Model.



You are a medical expert tasked with evaluating the presence of a disease in a patient's medical note. Provide your answer in a clear Yes/No format, on a new line.
---
### Medical Note:
(M) admission date ...
Previously you were asked about the presence of some medical concepts in the note (M). The questions, and your answers to each of them are provided below:
Question: Does the patient contain 'Aspiration events'?
Answer: No
Question: Does the patient contain 'Ventilator management'?
Answer: Yes
...
...
Question: Does the patient contain 'Sedation and analgesia administration'?
Answer: Yes
Based on your own answers to the above questions, answer the following question regarding the presence of a disease in the note (M).
---
### Question:
(Q) Based on the medical note (M), and your answers to the previous questions, does the patient contain the disease 'Continuous mechanical ventilation for less than 96 consecutive hours'?
Answer the question above in a clear Yes/No format, on a new line.
---
### Output Format:
- Answer the question on a new line in the format:
'Yes' or 'No'
### Response:

Figure 4: ICD Label Prediction Prompt to the Large Language Model using Intermediate Concept Predictions.

erating concept predictions to ensure alignment with our task (Details in Appendix A). For evaluation, we benchmark against recent state-of-the-art models, re-using the standardized scripts from (Yang et al., 2022). The experimental results demonstrate significant improvements in model performance across all evaluation metrics (Table 1).

### 4.1. Dataset

The MIMIC-III dataset (Johnson et al., 2016) is a key factor in healthcare research, offering extensive data on deidentified patients. MIMIC-III-50 contains common medical diagnosis and procedures, while MIMIC-III-rare-50 focuses on less frequent conditions. We use the same data splits as (Mullenbach et al., 2018) and (Vu et al., 2021), and for rare-50, splits from (Yang et al., 2022). MIMIC-III-50 has 8066 training and 1729 test instances, while MIMIC-III-rare-50 has 249 training and 142 test instances.

### 4.2. Discussion

Our results demonstrate that incorporating concept-level reasoning into ICD coding not only bridges the performance gap between LLMs and traditional deep learning models, but also enables LLMs to substantially outperform the state-of-the-art. A key observation is that our method performs similarly well on both the top-50 and rare-50 subsets. On both MIMIC-III-50 and MIMIC-III-rare-50, the concept-enhanced framework achieved consistently higher scores across Macro and Micro F1, with particularly tremendous gains in the rare-code setting. This suggests that concepts provide a critical signal where data scarcity and label imbalance have traditionally hindered performance. Direct code prediction, even with fine-tuning, yielded weak performance, aligning with prior reports that LLMs struggle to map long, noisy notes directly to thousands of candidate codes. By contrast, concept prediction proved to be easier. Once identified, they reliably expand the effective la-

| **Models** | **MIMIC-III-50** | | | | **MIMIC-III-rare-50** | | | |
|---|---|---|---|---|---|---|---|---|
| | AUC | | F1 | | AUC | | F1 | |
| | Macro | Micro | Macro | Micro | Macro | Micro | Macro | Micro |
| CAML (Mullenbach et al., 2018) | 0.875 | 0.909 | 0.532 | 0.614 | 0.574$^*$ | 0.602$^*$ | 0.072$^*$ | 0.083$^*$ |
| HyperCore (Cao et al., 2020) | 0.895 | 0.929 | 0.609 | 0.663 | — | — | — | — |
| LAAT (Vu et al., 2021) | 0.925 | 0.946 | 0.666 | 0.715 | — | — | — | — |
| Joint-LAAT (Vu et al., 2021) | 0.925 | 0.946 | 0.661 | 0.716 | — | — | — | — |
| MSMN (Yuan et al., 2022) | 0.928 | 0.947 | 0.683 | 0.725 | 0.753$^*$ | 0.762$^*$ | 0.171$^*$ | 0.172$^*$ |
| KEPTLongformer (Yang et al., 2022) | 0.926 | 0.947 | 0.689 | 0.729 | 0.827 | 0.833 | 0.304 | 0.326 |
| TwoStage (Nguyen et al., 2023) | 0.926 | 0.945 | 0.689 | 0.718 | — | — | — | — |
| LLM-codex (Yang et al., 2023) | 0.929 | 0.948 | 0.674 | 0.715 | 0.825 | 0.832 | 0.279 | 0.302 |
| Multi-stage Retrieve (Wang et al., 2024a) | 0.927 | 0.947 | 0.687 | 0.732 | — | — | — | — |
| CoRelation (Luo et al., 2024) | 0.933 | 0.951 | 0.693 | 0.731 | — | — | — | — |
| Exploring LLMs MAC (Li et al., 2024) | — | — | 0.748 | 0.589 | — | — | 0.881 | 0.376 |
| CEC - Llama3 | 0.949 | 0.955 | 0.877 | 0.879 | **0.971** | **0.969** | **0.910** | **0.875** |
| CEC - Gemma | **0.997** | **0.997** | **0.908** | **0.919** | 0.946 | 0.937 | 0.651 | 0.640 |

Table 1: Results for MIMIC-III-50 and rare-50 datasets. Baseline results are cited from the original papers, while $^*$ indicates the numbers provided by other papers. — indicates the numbers are not reported.

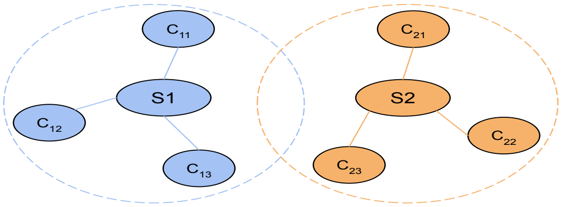


Figure 5: Label space expansion through associated concepts. Each code (e.g., S1, S2) connects to distinct concepts (c11–c13, c21–c23), providing finer granularity and improving discrimination between similar codes.

bel space and guide the model toward the correct codes. This two-stage reasoning process mirrors the workflow of human coders, who first detect medically salient findings before mapping them to ICD codes. The concept sets effectively disambiguated semantically similar codes, cleanly separating those with overlapping descriptions but differing in chronicity or etiology, thereby resolving a common ICD coding error: confusion between near-synonymous leaf codes. Prior ICD coding approaches have explored zero-shot prompting, multi-stage retrieval, or error-specific prompting. They often rely on rigid heuristics or shallow retrieval pipelines that limit robustness. Our method departs from these approaches by integrating concept-aware evidence retrieval and candidate expansion into a unified LLM-based framework, which reduces error propagation and captures implicit signals across sections of clinical notes (Details in Appendix D). This design allows us to improve recall on rare codes while maintaining precision. Collectively, these findings indicate that concepts act as both a performance enhancer and an interpretability mechanism. Unlike prior LLM approaches that often sacrifice precision, our pipeline leverages concept hierarchies to preserve clinical relevance while delivering state-of-the-art accuracy. Beyond ICD coding, this paradigm suggests a promising direction for other clinical NLP tasks: the decomposition of complex label spaces into medically meaningful intermediate reasoning steps.

### 4.3. Ablation Study

We visualize how the concepts can expand the label space of each ICD code in Figure 5. When we look for

| **Models** | **MIMIC-III-50** | | | | **MIMIC-III-rare-50** | | | |
|---|---|---|---|---|---|---|---|---|
| | Llama3 (F1) | | Gemma (F1) | | Llama3 (F1) | | Gemma (F1) | |
| | Macro | Micro | Macro | Micro | Macro | Micro | Macro | Micro |
| Direct coding w/o FT | 0.177 | 0.179 | 0.259 | 0.251 | 0.052 | 0.051 | 0.168 | 0.163 |
| Direct coding w. FT | 0.459 | 0.422 | 0.402 | 0.468 | 0.183 | 0.144 | 0.229 | 0.200 |
| Leaf concepts | 0.832 | 0.773 | 0.794 | 0.803 | 0.774 | 0.693 | 0.784 | 0.767 |
| CEC | 0.877 | 0.879 | 0.908 | 0.919 | 0.910 | 0.875 | 0.651 | 0.640 |

Table 2: Ablation results for MIMIC-III-50 dataset, and MIMIC-III-rare-50 dataset.

| **Number of Concepts** | **MIMIC-III-50** | | | | **MIMIC-III-rare-50** | | | |
|---|---|---|---|---|---|---|---|---|
| | Llama3 (F1) | | Gemma (F1) | | Llama3 (F1) | | Gemma (F1) | |
| | Macro | Micro | Macro | Micro | Macro | Micro | Macro | Micro |
| $C_+ = 6$, $C_- = 2$ | 0.877 | 0.879 | 0.908 | 0.919 | 0.910 | 0.875 | 0.651 | 0.640 |
| $C_+ = 5$, $C_- = 2$ | 0.875 | 0.876 | 0.903 | 0.922 | 0.840 | 0.876 | 0.690 | 0.657 |
| $C_+ = 4$, $C_- = 2$ | 0.882 | 0.874 | 0.898 | 0.917 | 0.825 | 0.792 | 0.668 | 0.657 |
| $C_+ = 3$, $C_- = 2$ | 0.865 | 0.854 | 0.847 | 0.861 | 0.804 | 0.736 | 0.670 | 0.666 |

Table 3: Concept sensitive analysis for MIMIC-III-50 dataset, and MIMIC-III-rare-50 dataset.

concepts, the extracted concepts are relatively unique to each code and we expect them to discriminate different ICD codes; as a result, concepts for different ICD codes have very little overlap. If we consider only the semantic space of a single ICD code, an LLM must project the entire medical note directly into that space for code prediction. By incorporating surrounding concepts, however, any projection into concepts associated with a given code will increase the predicted probability of that code. In addition, if we look into a pair of similar ICD codes, which often causes the errors for previous machine learning methods, the extracted unique concepts can easily help discriminate them. If we compare the results of using concepts, and the direct coding with fine-tuning in Table 2, the concept prediction seems much easier, and highly accurate compared to code prediction, which is understandable because concepts are more straightforward, and often explicitly included in the medical note. In fact, the previously successful synonym-based method (MSMN(Yuan et al., 2022)) can also be seen as a special case of our concept-enhanced method, if we regard synonyms as concepts. We assess the capability of the LLMs by conducting a direct ICD coding task with binary replies for each ICD code's description. Fine-tuning (FT) the LLMs in a few-shot setup improve results but did not surpass our original pipeline. This is consistent with the conclusion from previous LLM papers, and it demonstrates the effectiveness of using concepts. Since concepts are often symptoms included in the text descriptions, predicting concepts is much easier than directly predicting the ICD codes and the concepts can expand the code label space then, and enhance the coding performance. All ablation study results are shown in Table 2. Furthermore, the sensitivity analysis (Table 3) on concept set size shows that performance is stable across different configurations, underscoring the flexibility of the framework and its potential to scale.

## 5. Conclusion

In this paper, we propose a novel framework for automatic ICD coding by leveraging concept-enhanced LLMs. Unlike traditional methods that directly predict ICD codes or rely on code descriptions and synonyms, our approach expands the label space using medical concepts, enhancing the model's understanding of hierarchical code structures. We introduce a two-stage pipeline, where LLMs are fine-tuned to predict the relevance of concepts from clinical notes, followed by code prediction using both concept scores, and the raw note. Our method is highly effective in distinguishing between semantically similar codes by incorporating both parent-level and leaf-level concepts. Experimental results on two benchmark datasets, MIMIC-III-50, and MIMIC-III-rare-50, demonstrate substantial improvements over previous state-of-the-art models. Our approach significantly outperforms both traditional and recent LLM-based methods, particularly in handling rare codes, a challenging aspect in ICD coding tasks.

Additionally, ablation study on the impact of concept set size and model configuration highlights the robustness and flexibility of our concept-based architecture. Beyond performance improvements, our framework also contributes to better interpretability in ICD coding by explicitly modeling the reasoning path through clinically meaningful concepts. This makes it not only a high-performing solution but also a more transparent and clinically grounded one. In future work, we aim to enhance concept generation, and prediction efficiency to support even broader clinical applications.

## 6. Limitations

Despite the high accuracy that we obtain in ICD coding, we also identify some limitations. Firstly, our method relies on a good enough LLM (e.g. GPT4) to generate the set of concepts that are highly associated with ICD codes and can also differentiate different codes. Secondly, the off-the-shelf LLMs do not perform well in concept prediction, so the backbone LLMs need to be first fine-tuned for each concept in the concept prediction phase.

Besides, our current architecture introduces significant computational overhead. Due to the large number of concepts, it is not time-efficient, especially when there are thousands of ICD codes. While the proposed design demonstrates strong performance gains, its present computational demands limit our ability to evaluate the method on the MIMIC-III-full dataset. In future work, we plan to investigate more efficient strategies to extend concept-based ICD coding to larger label spaces. Therefore, the current study serves as an initial step toward establishing the feasibility of concept-enhanced coding, with future work aimed at improving computational efficiency and scalability.

## References

Monica Agrawal, Stefan Hegselmann, Hunter Lang, Yoon Kim, and David Sontag. Large language models are few-shot clinical information extractors. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1998–2022, 2022.

Joseph Spartacus Boyle, Antanas Kascenas, Pat Lok, Maria Liakata, and Alison Q O'Neil. Automated clinical coding using off-the-shelf large language models. In *Deep Generative Models for Health Workshop NeurIPS 2023*, 2023.

Pengfei Cao, Yubo Chen, Kang Liu, Jun Zhao, Shengping Liu, and Weifeng Chong. Hypercore: Hyperbolic and co-graph representation for automatic icd coding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3105–3114, 2020.

ContactDoctor. Contactdoctor-biomedical: A high-performance biomedical language model. https://huggingface.co/ContactDoctor/Bio-Medical-Llama-3-8B, 2024.

Tim Dettmers, Mike Lewis, Sam Shleifer, and Luke Zettlemoyer. 8-bit optimizers via block-wise quantization, 2022. URL https://arxiv.org/abs/2110.02861.

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

Matúš Falis, Aryo Pradipta Gema, Hang Dong, Luke Daines, Siddharth Basetti, Michael Holder, Rose S Penfold, Alexandra Birch, and Beatrice Alex. Can gpt-3.5 generate and code discharge summaries? *Journal of the American Medical Informatics Association*, 31(10):2284–2293, 2024.

GemmaTeam. Gemma: Open models based on gemini research and technology, 2024. URL https://arxiv.org/abs/2403.08295.

Aaron Grattafiori. The llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

Haraldur Hauksson and Hafsteinn Einarsson. Applications of bert models towards automation of clinical coding in icelandic. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 1956–1967, 2024.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models, 2021.

Chao-Wei Huang, Shang-Chi Tsai, and Yun-Nung Chen. Plm-icd: Automatic icd coding with pretrained language models. In *Proceedings of the 4th Clinical Natural Language Processing Workshop*, pages 10–20, 2022.

Alistair E.W. Johnson, Tom J. Pollard, Lu Shen, Li-wei H. Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G. Mark. Mimic-iii, a freely accessible critical care database. *Scientific Data*, 3(1):160035, 2016. ISSN 2052-4463. doi: 10.1038/sdata.2016.35. URL https://doi.org/10.1038/sdata.2016.35.

Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. Concept bottleneck models, 2020.

Fei Li and Hong Yu. Icd coding from clinical text using multi-filter residual convolutional neural network. In *proceedings of the AAAI conference on artificial intelligence*, volume 34(5), pages 8180–8187, 2020.

Rumeng Li, Xun Wang, and Hong Yu. Exploring llm multi-agents for icd coding. *arXiv preprint arXiv:2406.15363*, 2024.

Chang Lu, Chandan K Reddy, Ping Wang, and Yue Ning. Towards semi-structured automatic icd coding via tree-based contrastive learning. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

Junyu Luo, Xiaochen Wang, Jiaqi Wang, Aofei Chang, Yaqing Wang, and Fenglong Ma. CoRelation: Boosting automatic ICD coding through contextualized code relation learning. In Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue, editors, *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 3997–4007, Torino, Italia, May 2024.

ELRA and ICCL. URL https://aclanthology.org/2024.lrec-main.355/.

James Mullenbach, Sarah Wiegreffe, Jon Duke, Jimeng Sun, and Jacob Eisenstein. Explainable prediction of medical codes from clinical text. In Marilyn Walker, Heng Ji, and Amanda Stent, editors, *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1101–1111, New Orleans, Louisiana, jun 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1100. URL https://aclanthology.org/N18-1100.

Khalid Nawab, Madalyn Fernbach, Sayuj Atreya, Samina Asfandiyar, Gulalai Khan, Riya Arora, Iqbal Hussain, Shadi Hijjawi, and Richard Schreiber. Fine-tuning for accuracy: evaluation of generative pretrained transformer (gpt) for automatic assignment of international classification of disease (icd) codes to clinical documentation. *Journal of Medical Artificial Intelligence*, 7, 2024.

Thanh-Tung Nguyen, Viktor Schlegel, Abhinav Ramesh Kashyap, and Stefan Winkler. A two-stage decoder for efficient icd coding. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 4658–4665, 2023.

Tuomas Oikarinen, Subhro Das, Lam Nguyen, and Lily Weng. Label-free concept bottleneck models. In *International Conference on Learning Representations*, 2023.

OpenAI. Gpt-4 technical report, 2024. URL https://arxiv.org/abs/2303.08774.

Anthony Rios and Ramakanth Kavuluru. Few-shot and zero-shot multi-label learning for structured label spaces. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing*, volume 2018, page 3132. NIH Public Access, 2018.

S. E. Robertson and S. Walker. Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval. In *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '94, page 232–241, Berlin, Heidelberg, 1994. Springer-Verlag. ISBN 038719889X.

Haoran Shi, Pengtao Xie, Zhiting Hu, Ming Zhang, and Eric P Xing. Towards automated icd coding using deep learning. *arXiv preprint arXiv:1711.04075*, 2017.

Thanh Vu, Dat Quoc Nguyen, and Anthony Nguyen. A label attention model for icd coding from clinical text. In *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*, pages 3335–3341, 2021.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers, 2020. URL https://arxiv.org/abs/2002.10957.

Xindi Wang, Robert E Mercer, and Frank Rudzicz. Multi-stage retrieve and re-rank model for automatic medical coding recommendation. *arXiv preprint arXiv:2405.19093*, 2024a.

Zeqiang Wang, Yuqi Wang, Haiyang Zhang, Wei Wang, Jun Qi, Jianjun Chen, Nishanth Sastry, Jon Johnson, and Suparna De. Icdxml: enhancing icd coding with probabilistic label trees and dynamic semantic representations. *Scientific Reports*, 14(1): 18319, 2024b.

John Wu, David Wu, and Jimeng Sun. Dila: Dictionary label attention for mechanistic interpretability in high-dimensional multi-label medical coding prediction. In *Machine Learning for Health (ML4H)*, pages 1014–1038. PMLR, 2025.

Pengtao Xie and Eric Xing. A neural architecture for automated icd coding. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1066–1076, 2018.

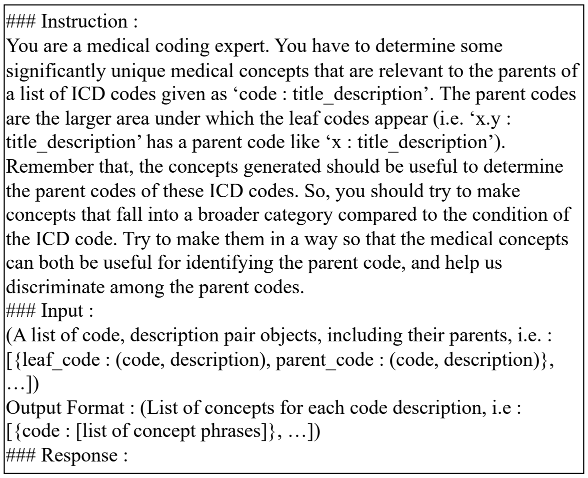
Xiancheng Xie, Yun Xiong, Philip S Yu, and Yangyong Zhu. Ehr coding with multi-scale feature attention and structured knowledge graph propagation. In *Proceedings of the 28th ACM international conference on information and knowledge management*, pages 649–658, 2019.

Zhichao Yang, Shufan Wang, Bhanu Pratap Singh Rawat, Avijit Mitra, and Hong Yu. 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. NIH Public Access, 2022.

Zhichao Yang, Sanjit Singh Batra, Joel Stremmel, and Eran Halperin. Surpassing gpt-4 medical coding with a two-stage approach. *arXiv preprint arXiv:2311.13735*, 2023.

Zheng Yuan, Chuanqi Tan, and Songfang Huang. Code synonyms do matter: Multiple synonyms matching network for automatic icd coding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 808–814, 2022.



### Instruction :
You are a medical coding expert. You have to determine some significantly unique medical concepts that are relevant to the parents of a list of ICD codes given as ‘code : title_description’. The parent codes are the larger area under which the leaf codes appear (i.e. ‘x.y : title_description’ has a parent code like ‘x : title_description’). Remember that, the concepts generated should be useful to determine the parent codes of these ICD codes. So, you should try to make concepts that fall into a broader category compared to the condition of the ICD code. Try to make them in a way so that the medical concepts can both be useful for identifying the parent code, and help us discriminate among the parent codes.
### Input :
(A list of code, description pair objects, including their parents, i.e. : [{leaf_code : (code, description), parent_code : (code, description)}, …])
Output Format : (List of concepts for each code description, i.e : [{code : [list of concept phrases]}, …])
### Response :

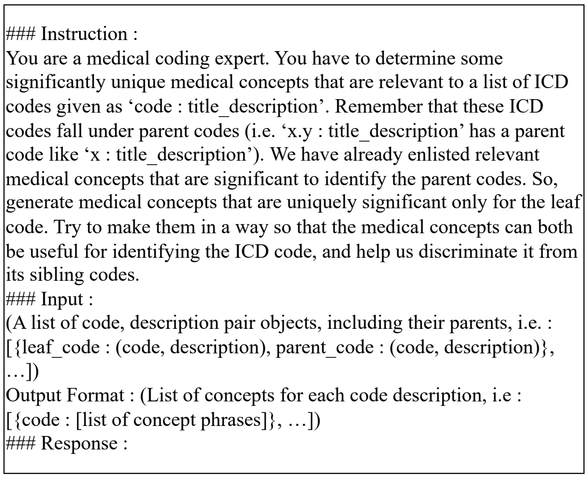
Figure 6: Prompt used to create parent-level concepts using GPT-4

## Appendix A. LoRA Fine-tuning Configuration

We apply LoRA fine-tuning with rank $r = 32$, scaling factor $\alpha = 32$, dropout $= 0.05$, and gradient accumulation steps $= 4$. Training is performed for a maximum of 100 steps, which corresponds to approximately 400 binary question–answer pairs used for weight updates. For comparison, the test set (142 notes in the rare-50 dataset) contains over 47,000 binary questions. For concept prediction, the Gemma model is fine-tuned for 500 steps. For the concept-to-label prediction model, we adopt a higher LoRA scaling factor ($\alpha = 128$) and restrict adaptation to the attention projection layers (k_proj, v_proj). This fine-tuning is run for 2000 training steps using the Bio-Medical-Llama-3-8B (ContactDoctor, 2024) backbone, with paged AdamW (8-bit) optimizer (Dettmers et al., 2022) and a gradient accumulation steps of 4. The Gemma model is also trained in a similar setup.

## Appendix B. Concept Set Construction Prompts

Figure 6 and figure 7 are the prompts we use to construct the concepts for each of the icd code, including their parent levels. At each iteration, 5 to 7 codes are provided.



### Instruction :
You are a medical coding expert. You have to determine some significantly unique medical concepts that are relevant to a list of ICD codes given as ‘code : title_description’. Remember that these ICD codes fall under parent codes (i.e. ‘x.y : title_description’ has a parent code like ‘x : title_description’). We have already enlisted relevant medical concepts that are significant to identify the parent codes. So, generate medical concepts that are uniquely significant only for the leaf code. Try to make them in a way so that the medical concepts can both be useful for identifying the ICD code, and help us discriminate it from its sibling codes.
### Input :
(A list of code, description pair objects, including their parents, i.e. : [{leaf_code : (code, description), parent_code : (code, description)}, …])
Output Format : (List of concepts for each code description, i.e : [{code : [list of concept phrases]}, …])
### Response :

Figure 7: Prompt used to create leaf-level concepts using GPT-4

## Appendix C. Constructing Concept Based Train Set

To construct the concept-based training set, we first extract pseudo-positive concepts for each true ICD label by mapping codes to their associated concept sets. We then generate pseudo-negative examples by sampling concepts from unrelated labels. This produces a weakly supervised binary classification dataset of concepts, which is used to fine-tune the LLM for concept prediction. A training set for the final prediction model is constructed in a similar manner. The complete process is elaborated in Figure 8.

## Appendix D. Coverage of Evidence Signals from the Medical Notes using Concepts

We assess the coverage of our extracted concepts against clinical notes by segmenting each note into overlapping sentence windows and computing BM25 similarity (Robertson and Walker, 1994) between segments and concept phrases. This analysis shows that nearly 99% of the notes contain at least one segment aligned with a concept from the corresponding set, indicating strong overlap between the concept space and the evidence expressed in clinical text. This supports the relevance of the constructed concept sets as faithful intermediates for ICD coding. An example is explained in Figure 9.

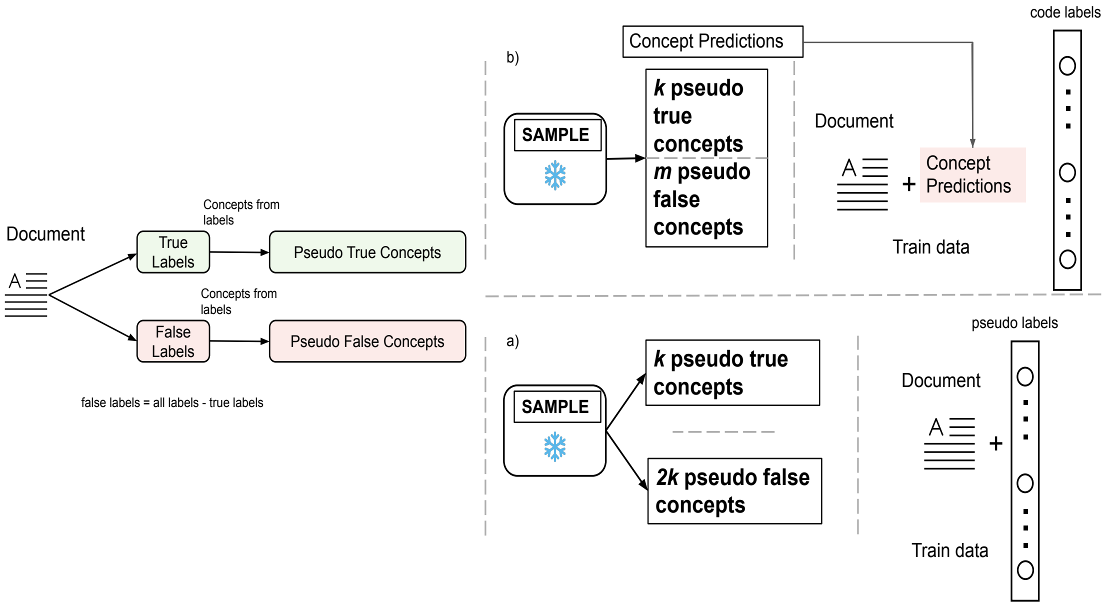


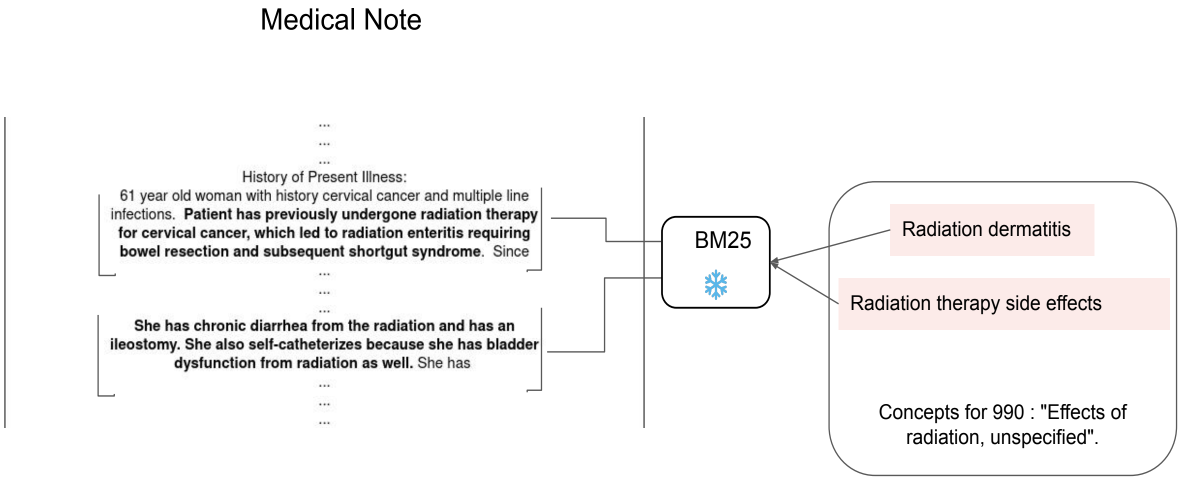
Figure 8: Construction of a weakly supervised binary classification dataset of concepts. (a) Concept-based training data formation using pseudo labels. (b) Pseudo labels are used as concept predictions in creating the training data for the final model.

## Appendix E. Section-Aware Segmentation for Long Text Documents

Clinical notes often exceed the context length of current LLMs, making it difficult to process long documents effectively. We have further experimented with segmenting notes into essential medically relevant sections (e.g., discharge diagnosis, history of present illness, past medical history) and consolidating them into a single document, achieving comparable accuracy without relying on the full note.
Repeating the process from Lu *et al.*(Lu et al., 2023), we implement the DF-IAPF (Document Frequency–Inverse Average Phrase Frequency) algorithm to identify section titles. This method works by scoring n-gram phrases based on how often they appear across documents (high document frequency) but how rarely they repeat within a single note (low average phrase frequency). Phrases with high DF-IAPF scores, such as “history of present illness” or “brief hospital course”, are strong candidates for section titles. This not only emphasizes clinically important content, but also ensures that even lengthy narratives can be handled within the model’s input constraints. This segmentation thus provides both task relevance, and scalability for large clinical texts.

## Appendix F. AUC Scores of Ablation Results

The AUC scores of ablation results using Llama-3-8b model are provided in Table 4.



Medical Note

...
...
...
History of Present Illness:
61 year old woman with history cervical cancer and multiple line infections. **Patient has previously undergone radiation therapy for cervical cancer, which led to radiation enteritis requiring bowel resection and subsequent shortgut syndrome.** Since
...
...
...
**She has chronic diarrhea from the radiation and has an ileostomy. She also self-catheterizes because she has bladder dysfunction from radiation as well.** She has
...
...
...

BM25
Radiation dermatitis
Radiation therapy side effects
Concepts for 990 : "Effects of radiation, unspecified".

Figure 9: Concepts creating evidence signals from the medical note in detecting ICD code labels

| **Models** | **MIMIC-III-50** | | **MIMIC-III-rare-50** | |
|---|---|---|---|---|
| | Macro | Micro | Macro | Micro |
| Direct coding w/o FT | 0.528 | 0.527 | 0.516 | 0.511 |
| Direct coding w. FT | 0.651 | 0.638 | 0.652 | 0.642 |
| Leaf concepts | 0.922 | 0.921 | 0.835 | 0.808 |
| CEC | 0.949 | 0.955 | 0.971 | 0.969 |

Table 4: AUC scores for MIMIC-III-50 dataset, and MIMIC-III-rare-50 dataset using Llama-3-8b.