Synthetic Medical Data and Contrastive Learning for Automatic ICD Coding

Author: Truong Vo, Weiyi Wu
Group: REAL Lab

Outline

Motivation

ICD coding challenges & long-tail problem

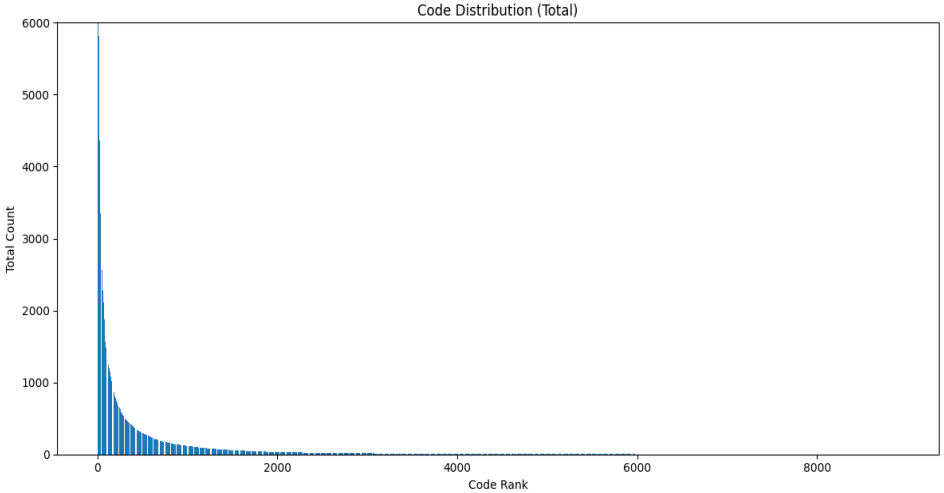
Related Work

Model-centric & data-centric approaches

Proposed Method

Synthetic data & LLM fine-tuning

Motivation & Long-Tail Problem



Motivation:

- Manual ICD coding is labor-intensive and error-prone
- Over 68k diagnostic and 87k procedural codes make coding complex
- Clinical notes vary from <500 to >3,000 words and contain abbreviations, typos & synonyms
- The distribution is extremely skewed: top 10% of codes account for the majority of occurrences

Related Work

01

Data-Centric & Synthetic

Recent works (MedSyn, Falis 2024) generate synthetic notes to address long-tail labels, yet prior methods generate notes for single codes.

02

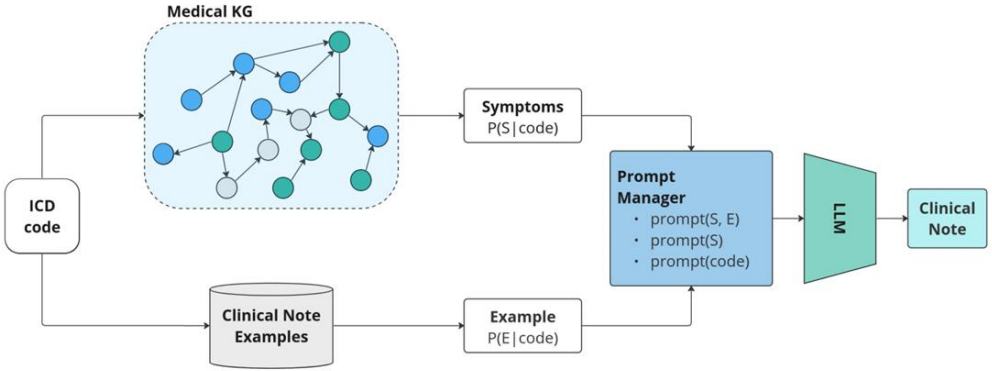
Model-centric Approach

CNNs & RNNs (CAML, MultiResCNN, LAAT) achieve high micro-F1 yet struggle with rare labels.
Frameworks like GKI-ICD combine PLMs with descriptions, synonyms & hierarchies
LLMs is being applied in ICD Coding with strong language understanding but often hallucinate and show low precision.

Data-Centric and Synthetic

MedSyn Framework (Kumichev et al., 2024):

- Integrates LLMs (e.g., GPT-4, LLaMA) with Medical Knowledge Graphs (MKGs).
- Uses structured MKG-derived symptoms and metadata to prompt LLMs.
- Generates clinically coherent discharge notes for rare diseases.
- Achieves **up to 17.8% Macro-F1 gain** on rare codes.

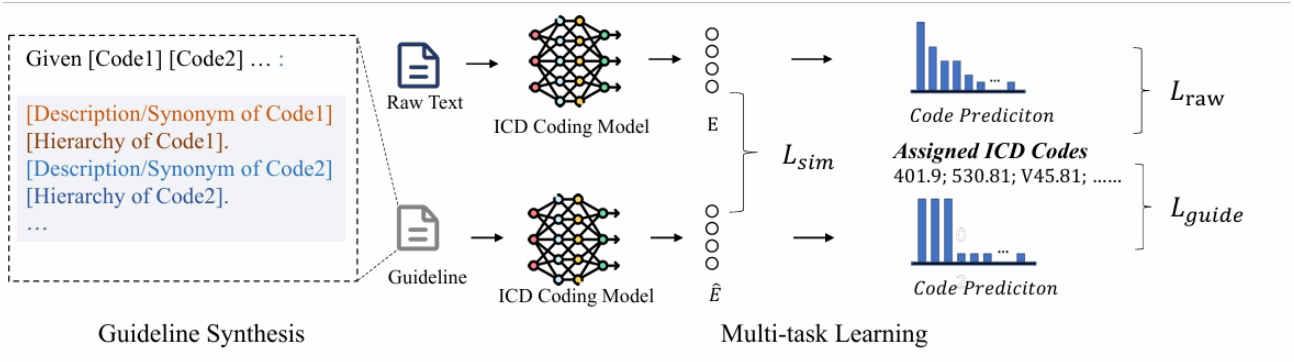


Model-centric Approach

01

Transformer models (PLMs)

CNNs & RNNs (CAML, MultiResCNN, LAAT) achieve high micro-F1 yet struggle with rare labels. SOTA model is GKI-ICD combine PLMs with descriptions, synonyms & hierarchies (**Knowledge-Injection**)



Model-centric Approach

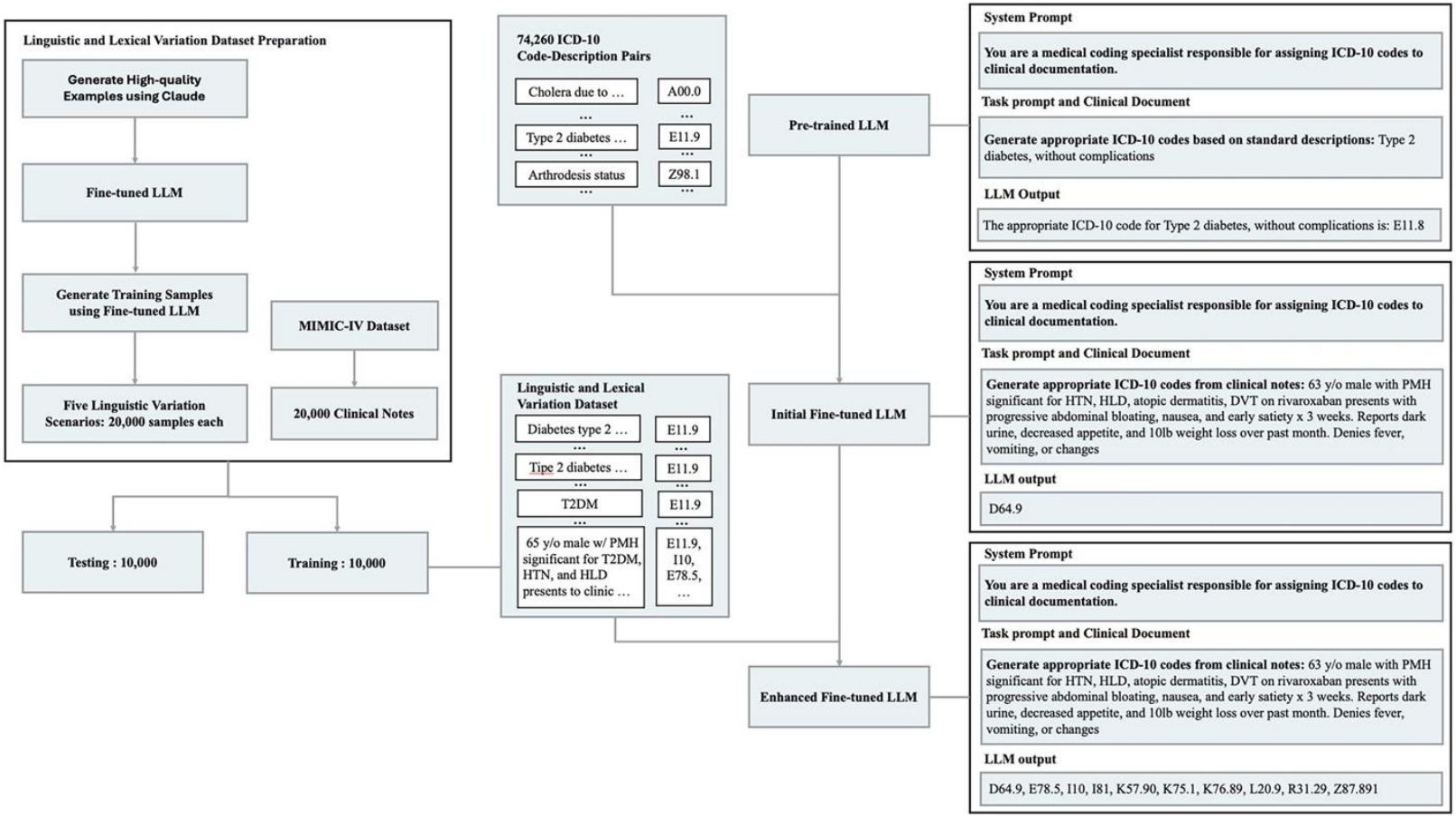
02

Large Language Models (LLMs)

LLMs is being applied in ICD Coding with strong language understanding but often hallucinate and show low precision.

LLM-Only Approaches: Strengths & Pitfalls
GPT-4: Good language understanding but **low precision**, **code overgeneration**
→ 34–46% exact match; <15% agreement with human coders
→ Needs domain adaptation to be effective

Two-stage strategy:
Fine-tune on 74k ICD code-description pairs
Enhance with real-world notes & variability
→ Yields up to 97% exact match, 87% category match


Linguistic and Lexical Variation Dataset Preparation
Generate High-quality Examples using Claude
Fine-tuned LLM
Generate Training Samples using Fine-tuned LLM
Five Linguistic Variation Scenarios: 20,000 samples each
MIMIC-IV Dataset
20,000 Clinical Notes
Testing : 10,000
Training : 10,000
74,260 ICD-10 Code-Description Pairs
Cholera due to …
A00.0
Type 2 diabetes …
E11.9
Arthrodesis status
Z98.1
Linguistic and Lexical Variation Dataset
Diabetes type 2 …
E11.9
Tipe 2 diabetes …
E11.9
T2DM
E11.9
65 y/o male w/ PMH significant for T2DM, HTN, and HLD presents to clinic …
E11.9, I10, E78.5, …
Pre-trained LLM
Initial Fine-tuned LLM
Enhanced Fine-tuned LLM
System Prompt
You are a medical coding specialist responsible for assigning ICD-10 codes to clinical documentation.
Task prompt and Clinical Document
Generate appropriate ICD-10 codes based on standard descriptions: Type 2 diabetes, without complications
LLM Output
The appropriate ICD-10 code for Type 2 diabetes, without complications is: E11.8
System Prompt
You are a medical coding specialist responsible for assigning ICD-10 codes to clinical documentation.
Task prompt and Clinical Document
Generate appropriate ICD-10 codes from clinical notes: 63 y/o male with PMH significant for HTN, HLD, atopic dermatitis, DVT on rivaroxaban presents with progressive abdominal bloating, nausea, and early satiety x 3 weeks. Reports dark urine, decreased appetite, and 10lb weight loss over past month. Denies fever, vomiting, or changes
LLM output
D64.9
System Prompt
You are a medical coding specialist responsible for assigning ICD-10 codes to clinical documentation.
Task prompt and Clinical Document
Generate appropriate ICD-10 codes from clinical notes: 63 y/o male with PMH significant for HTN, HLD, atopic dermatitis, DVT on rivaroxaban presents with progressive abdominal bloating, nausea, and early satiety x 3 weeks. Reports dark urine, decreased appetite, and 10lb weight loss over past month. Denies fever, vomiting, or changes
LLM output
D64.9, E78.5, I10, I81, K57.90, K75.1, K76.89, L20.9, R31.29, Z87.891

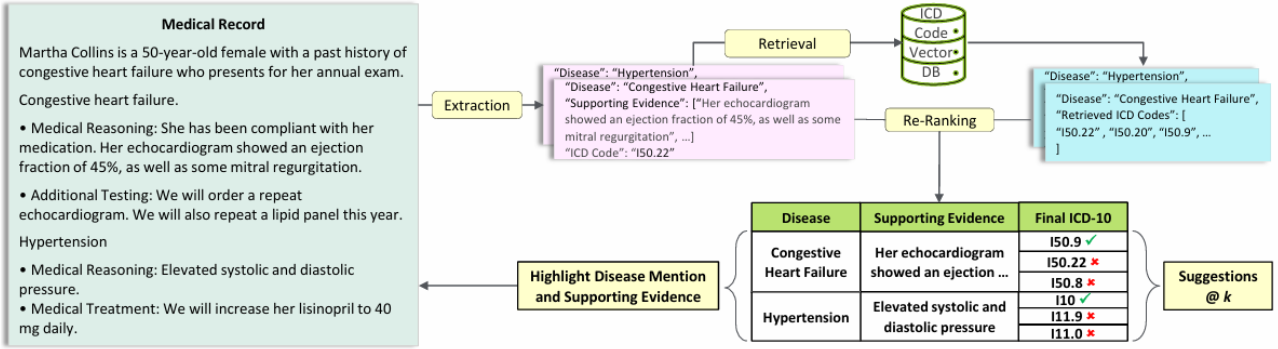
Model-centric Approach

02

Large Language Models (LLMs)

Retrieval-Augmented Generation (RAG) & Modular Architectures

MedCodER: 3-stage system (CoT extraction → ICD retrieval → LLM reranking)
→ Micro-F1 = 0.62; interpretable & accurate



Multi-Agent Coding Systems

MAC-I / MAC-II simulate clinical coding workflows
→ Roles: Patient, Physician, Coder, Reviewer, Adjustor
→ The precision is not verified.

Proposed Methods

01

Data Synthesis

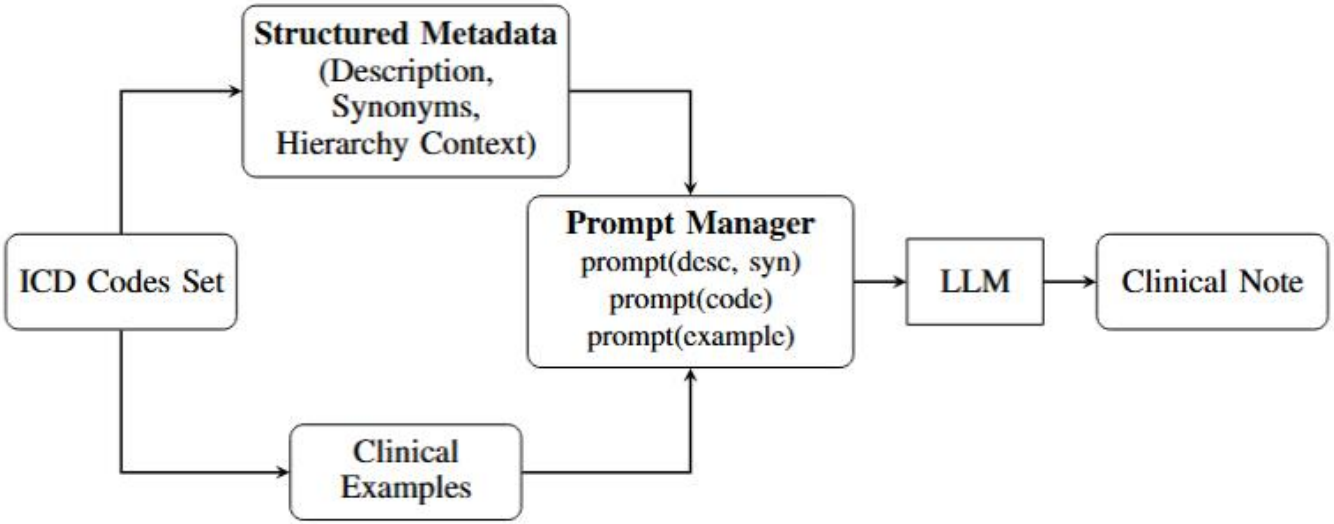
We propose a data-centric framework that systematically generates synthetic discharge summaries for multi-label ICD coding..

02

Contrastive Learning

Synthetic Data Generation Pipeline

Using a knowledge-injected prompting strategy grounded in official ICD descriptions, synonyms, and hierarchy, we augment diverse training medical notes for few-shot codes and create plausible medical notes for zero-shot codes that preserve real-world code co-occurrence patterns



Synthetic Data Generation Pipeline

1. Stratify Codes

- Identify rare and zero-shot codes for augmentation and synthesis

2. Plausible Sets

- Use each rare code as an anchor
- Retrieve realistic co-occurring codes from MIMIC

3. Diversify Notes

- Generate multiple notes per anchor
- Vary code description with synonyms from UMLS

4. Knowledge-Injection

- Insert descriptions, randomly chosen synonyms & hierarchy from dictionary
- Fuse real examples to craft prompts

Example: Suppose we aim to augment the rare code `N18.23` (Chronic kidney disease, stage 3), which is zero-shot. We retrieve a real summary labeled with {`E11.39`, `N18.29`, `I50.19`, `I10`}, and replace `N18.29` with `N18.23`, yielding:

$$\mathcal{C} = \{\texttt{N18.23}, \texttt{E11.39}, \texttt{I50.19}, \texttt{I10}\}$$

This preserves realistic multimorbidity relationships common in chronic kidney disease patients.

Supervised Learning

Stage 1: Description-Aware Pretraining

- Fine-tune the base model on all ICD descriptions
- Ground the LLM in semantic meaning of each code
- Input: code & metadata → Output: official description

Stage 2: Instruction Fine-Tuning

- Supervised learning on 40k real + 40k synthetic notes
- Formulate multi-label coding as an instruction task
- Parameter-efficient LoRA fine-tuning for scalability

Synthetic Generation for Contrastive Learning on Rare Codes

Create Note Pairs:

Way 1: ICD Codes Preserving Generation

Use the same set of target ICD descriptions as those present in the original notes.

e.g.

Preserving ICD descriptions: [...]

###Editing Rules

- 1. **Identify & preserve** every sentence or phrase that supports the target code(s) according to the provided guideline description. If necessary, revise up to the adjacent 1–2 sentences to ensure coherence and clarity—but do not introduce any findings that would imply unrelated ICD-9 codes.

Way 2: ICD Codes Deletion Generation

Randomly sample one or a few ICD descriptions from the original notes the original notes for targeted deletion.

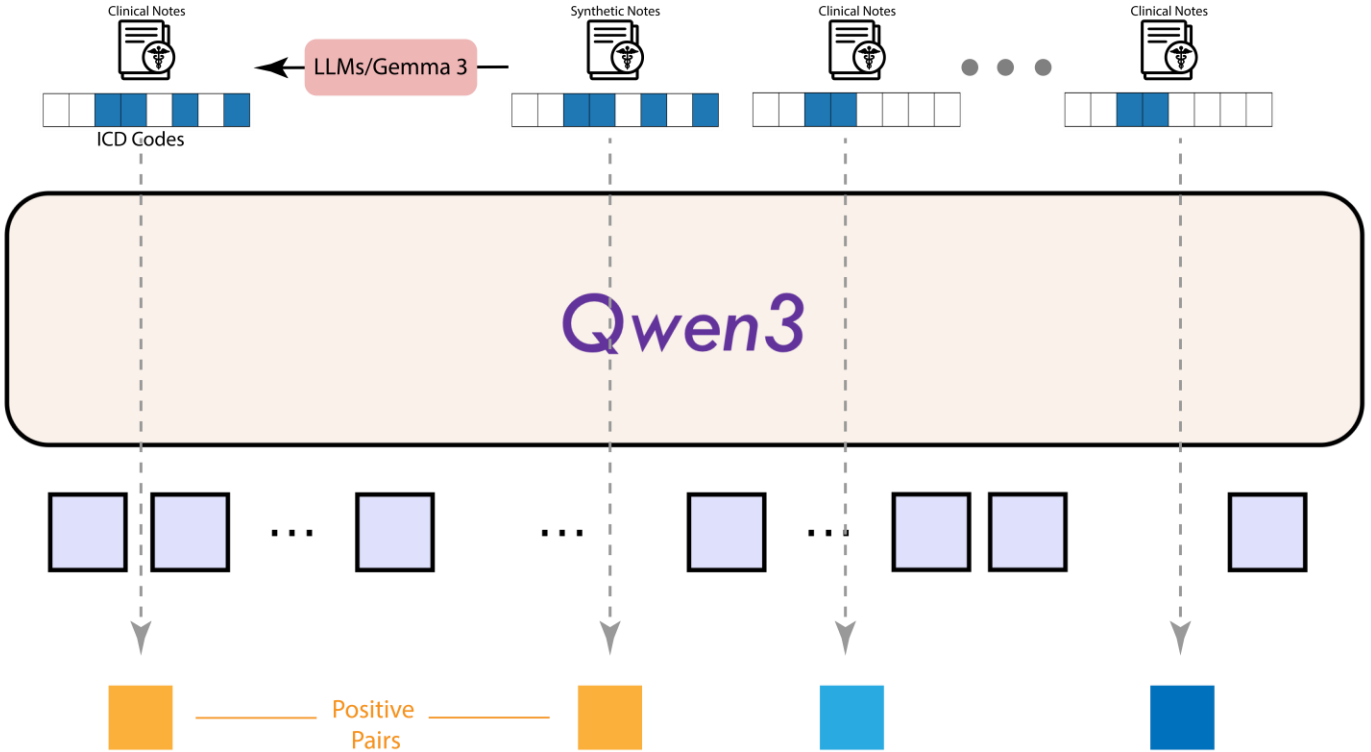
e.g.

Deleting ICD descriptions: [...]

Editing Rules

•1. **Identify & delete** every sentence or phrase that supports the target code(s) according to the provided guideline description. If deletion breaks narrative flow, rewrite up to the adjacent 1–2 sentences to restore coherence—but do **not** add any new findings that could trigger other ICD-9 codes.

Contrastive Fine-Tuning of Qwen3-Embedding for Rare-Code Notes



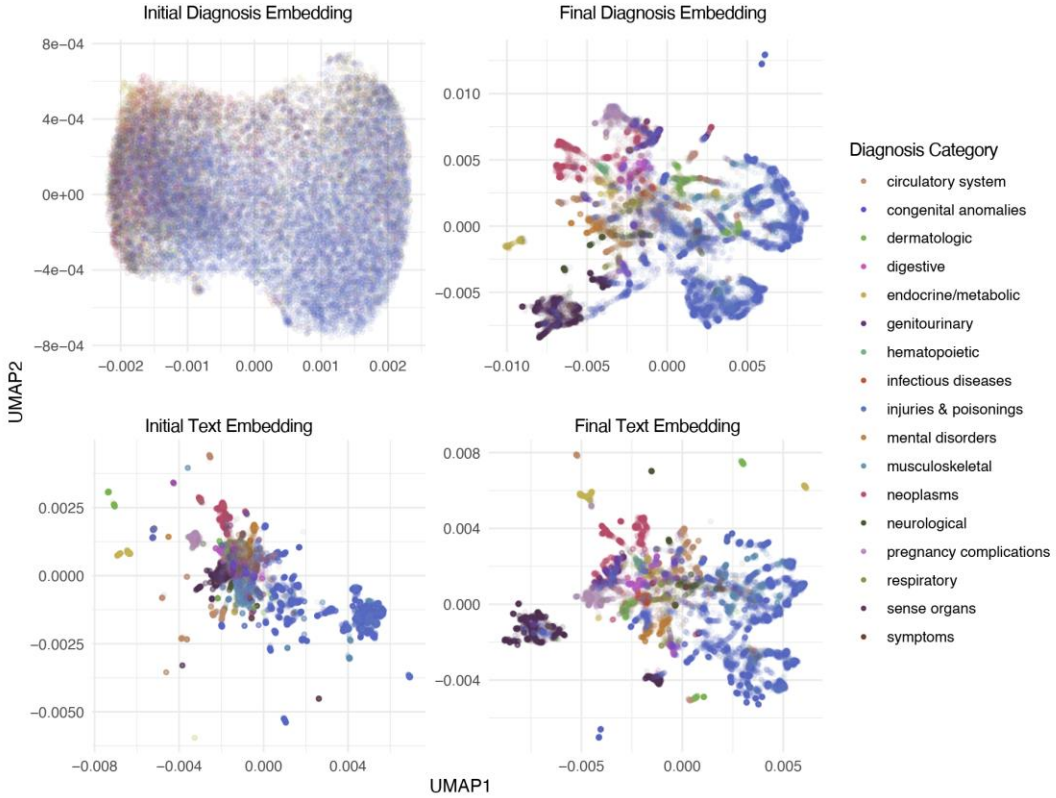
Multi-Label Aware Contrastive Loss

Label Similarity Aware Weighted Contrastive loss:

s_{ip} =Similarity( , )

$$\mathcal{L} = \sum_{i=1}^{N} \frac{-1}{\sum_{p \in P(i)} s_{ip}} \sum_{p \in P(i)} s_{ip} \cdot \log \frac{\exp\left(\frac{z_i \cdot z_p}{\tau}\right)}{\sum_{a \in A(i)} \exp\left(\frac{z_i \cdot z_a}{\tau}\right)}$$

Potential Embedding Visualizations for Rare ICD Codes



(Borrowed from https://proceedings.mlr.press/v225/kailas23a/kailas23a.pdf)