ACE-ICD: Acronym Expansion As Data Augmentation For Automated ICD Coding

Tuan-Dung Le[1,2], Shohreh Haddadan[1], Thanh Q. Thieu[1,2]
[1]Moffitt Cancer Center and Research Institute, [2]University of South Florida
{tuandung.le, shohreh.haddadan, thanh.thieu}@moffitt.org

Abstract

Automatic ICD coding, the task of assigning disease and procedure codes to electronic medical records, is crucial for clinical documentation and billing. While existing methods primarily enhance model understanding of code hierarchies and synonyms, they often overlook the pervasive use of medical acronyms in clinical notes, a key factor in ICD code inference. To address this gap, we propose a novel effective data augmentation technique that leverages large language models to expand medical acronyms, allowing models to be trained on their full form representations. Moreover, we incorporate consistency training to regularize predictions by enforcing agreement between the original and augmented documents. Extensive experiments on the MIMIC-III dataset demonstrate that our approach, **ACE-ICD** establishes new state-of-the-art performance across multiple settings, including common codes, rare codes, and full-code assignments. Our code is publicly available [1].

Discharge Summary		
... history of present illness: ortho hpi: 86m w/ severe b/l oa, admitted to ortho for sequential bilateral tka ... icu hpi: 86 y/o m with pmhx of arthritis, bph & osteoporosis s/p elective right total knee replacement ... past medical history: osteoporosis anemia (family h/o g6pd deficiency) bph osteoarthritis cataracts ... empiric vancomycin and ceftriaxone for possible uti were initiated ...		
Label	**KEPT**	**ACE-ICD**
599.0 urinary tract infection, site not specified	✗	✓
715.36 osteoarthrosis, localized, not specified whether primary or secondary, lower leg	✗	✓
81.54 total knee replacement	✓	✓
285.9 anemia, unspecified	✓	✓
...	...	...

Table 1: Example predictions from the MIMIC-III-full dataset (HADM_ID = 108519) using KEPT(Yang et al., 2022) and our ACE-ICD models.

1 Introduction

Assigning standardized codes based on the International Classification of Diseases (ICD[2]) known as ICD Coding is essential for efficient medical record management, accurate billing processes, and streamlined insurance reimbursements (Park et al., 2000; Sonabend et al., 2020). However, traditional ICD coding relies on manual effort, making it time-intensive and error-prone driving the development of automated coding methods.

Accurate ICD code assignment, whether manual or automated, requires a comprehensive understanding of clinical notes, which include detailed information such as symptoms, diagnoses, and test results. However, healthcare professionals often rely on acronyms and abbreviations to reduce documentation effort (Amosa et al., 2023). This shorthand introduces significant ambiguity, making it difficult for both human coders and language models to interpret the clinical text correctly. To better understand the impact of acronyms on ICD code assignment, we analyze expert-annotated evidence spans from the MDACE dataset (Cheng et al., 2023), which is derived from MIMIC-III (Johnson et al., 2016). Our analysis shows that 22% of the evidence spans either consist entirely of acronyms and abbreviations or include at least one such term, highlighting that acronyms often carry useful information for assigning ICD codes. Previous studies improved model understanding of medical acronyms by leveraging their appearance in code synonyms (Yuan et al., 2022; Yang et al., 2022; Gomes et al., 2024). However, they still fail to predict codes which contain acronyms in their synonym description. Table 1 shows several instances of this shortcoming in previous mod-

[1]https://github.com/LangIntLab/ACE-ICD
[2]who.int/standards/classifications/classification-of-diseases

Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, pages 1650–1662

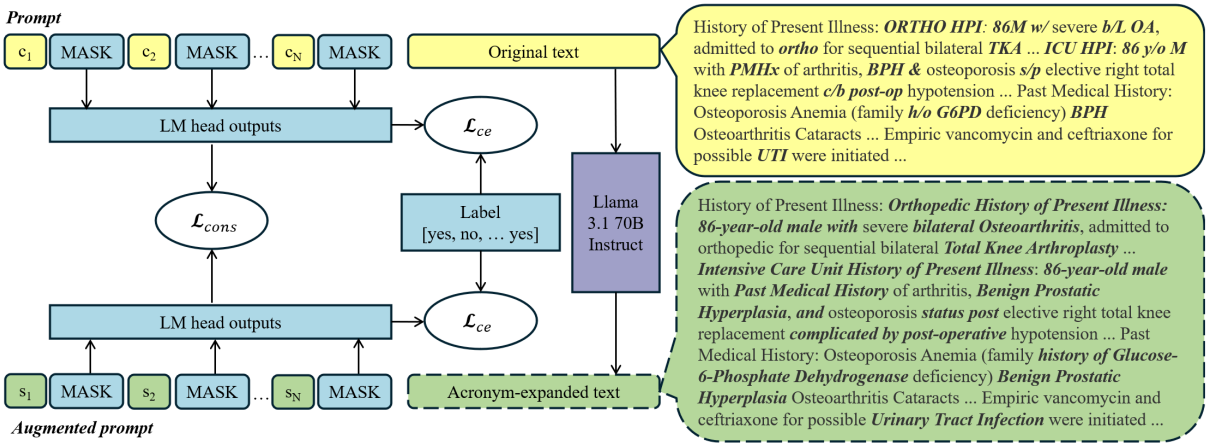

Figure 1: Our training pipeline incorporating acronym-expanded data augmentation and consistency training.

els. These failures underscore the need for new approaches to effectively address medical acronyms to consequently improve the ICD coding system.

In this paper, we propose **AC**ronym **E**xpansion for **ICD** coding (**ACE-ICD**) to investigate the impact of expanding medical acronyms to full terms on coding models. Our system harnesses the strong capability of large language models in disambiguating clinical acronyms (Kugic et al., 2024; Liu et al., 2024), as a data augmentation method that expands acronyms using open-source LLM prompting. We further apply KL divergence consistency regularization to enforce alignment between predictions from the original and augmented examples. Finally, we conduct experiments on three standard ICD coding tasks using the MIMIC-III dataset, demonstrating that our approach outperforms previous state-of-the-art methods across all tasks.

2 Methods

Previous methods frame the ICD coding task as a multi-label classification problem, as a single clinical note can contain multiple diagnosis or procedure codes. Given a clinical note t, the task is to assign a binary label $y_i \in \{0, 1\}$ for each ICD code i (where $i = 1, 2, ..., N_c$ and N_c is the total number of ICD codes). A label of 1 denotes relevance, while 0 indicates irrelevance.

We followed the prompt-based fine-tuning approach by (Yang et al., 2022) for the ICD coding task, reformulating the multi-label classification task as a cloze task (Schick and Schütze, 2021; Gao et al., 2021). Specifically, we construct a prompt template by concatenating each ICD code description c_i, appending a [MASK] token after each description, and adding the clinical note t. The prompt P is given by: $P = c_1$ [MASK] c_2 [MASK] ... c_{N_c} [MASK] t. The model is trained via a masked language modeling objective to predict "yes" or "no" in each [MASK] position, corresponding to a label of 1 or 0, respectively. For tasks where N_c is large (e.g., thousands of codes), this approach is typically used as a reranker (Tsai et al., 2021; Yang et al., 2022, 2023b; Kailas et al., 2023), as including all code descriptions in the prompt is infeasible.

2.1 Acronym-expansion as data augmentation

Motivated by capabilities of LLMs in clinical acronym disambiguation (Liu et al., 2024; Kugic et al., 2024), we use the open-source Llama 3 model (Dubey et al., 2024) to generate acronym-expanded version of clinical notes, denoted as t_a, from the original notes t. Due to the considerable length of the MIMIC-III notes, we first split each discharge summary based on the headers identified through automatic section-based segmentation (Lu et al., 2023). We then prompt the instruction-tuned Llama 3 models to generate augmented sections, which are concatenated into the final acronym-expanded note, using the following prompt:

```
<|begin_of_text|>
<|start_header_id|> system <|end_header_id>
You are a helpful assistant. <|eot_id|>
<|start_header_id|> user <|end_header_id|>
Expand all acronyms to their full forms while
preserving all the details in the following paragraph,
do not mention the acronyms again. Paragraph: ___
<|eot_id|>
<|start_header_id|> assistant <|end_header_id|>
Here is the paragraph with all acronyms expanded to
their full forms:
```

2.2 Consistency training

Inspired by (Shen et al., 2020; Wu et al., 2021), we incorporate a consistency loss into the training objective to encourage the model to generate similar predictions for the original clinical note t and its acronym-expanded counterpart t_a. We first construct a second prompt template using the augmented note t_a and a synonym s_i for each ICD code description collected from UMLS from previous studies (Yuan et al., 2022; Gomes et al., 2024): P_a = s_1 [MASK] s_2 [MASK] ... s_N [MASK] t_a. The training objective can be written as the following:

$$\mathcal{L} = \frac{1}{2}(\mathcal{L}_{ce}(P, y) + \mathcal{L}_{ce}(P_a, y)) + \alpha\mathcal{L}_{cons}(P, P_a, y)$$

where $\mathcal{L}_{ce}$ is cross-entropy loss for masked language modeling, and $\mathcal{L}_{cons}$ enforces consistency by minimizing the bidirectional KL-divergence:

$$\mathcal{L}_{cons} = \frac{1}{2}(KL[p(y|P)||p(y|P_a)] + KL[p(y|P_a)||p(y|P)])$$

3 Experiments

3.1 Acronym expansion evaluation

Dataset. Several annotated datasets for medical abbreviations and acronyms have been introduced in prior work (Moon et al., 2012; Rajkomar et al., 2022). To evaluate our zero-shot prompting approach on the MIMIC-III corpus, we utilize the annotated dataset developed by Rajkomar et al. (2022), which we consider most suitable for our task. This dataset was constructed using a reverse-substitution technique applied to MIMIC-III discharge summaries, resulting in a large-scale, high-quality resource for acronym expansion. As our experiments are also based on MIMIC-III, this dataset offers a consistent and reliable benchmark for evaluation.

Metrics. We perform zero-shot prompting experiments using four instruction-tuned variants of the Llama model. We first report detection precision and recall (Rajkomar et al., 2022), which assess the model's ability to identify abbreviations in text, regardless of whether the corresponding expansions are correct. We then compute total accuracy, defined as the proportion of abbreviations in the gold standard that are correctly expanded to their full forms by the model. To assess the quality of expansions, we consider two evaluation settings. Strict accuracy requires an exact string match between the model-generated expansion and the gold-standard full form. However, as illustrated in Table 3, certain expansions may differ lexically yet convey the same meaning. To accommodate such cases, we introduce lenient accuracy, which computes a similarity score based on the normalized inverse edit distance between the generated and reference expansions. Specifically, we normalize the edit distance by the length of the reference and consider expansions with a similarity score of at least 70% as correct. Although some semantically valid expansions may fall below this threshold, we adopt the 70% cutoff to balance recall and precision while avoiding acceptance of clearly incorrect outputs.

Implementation Details. To extract abbreviations and their corresponding expansions from model outputs, we employ the Python difflib library[3], which aligns two sequences.

3.2 ICD coding evaluation

We evaluate our methods on three MIMIC-III tasks, following (Mullenbach et al., 2018) for MIMIC-III-50 and MIMIC-III-full dataset splits and (Yang et al., 2022) for constructing MIMIC-III-rare50, which focuses on the top 50 codes with fewer than 10 occurrences (See Table 2).

Dataset	Train	Dev	Test	N_C
MIMIC-III-50	8066	1573	1729	50
MIMIC-III-full	47723	1631	3372	8922
MIMIC-III-rare50	249	20	142	50

Table 2: Statistics of the MIMIC-III ICD-9 datasets.

Metrics. We evaluate performance using macro-AUC, micro-AUC, macro-F1, micro-F1, and precision@k (k = 5 for MIMIC-III-50, k = 8 and 15 for MIMIC-III-full). We determine the optimal threshold for micro-F1 on the development set and report test metrics using the best-performing checkpoint. To ensure robustness, we run each experiment with five random seeds and report mean results.

Implementation Details. We initialize our model with two pre-trained language models:

[3] https://docs.python.org/3/library/difflib.html

Acronym	Expanded form by LLM	Full form	Inverse Edit distance
pod#15	post-operative day #15	post-operative day #15	100.0
intrabd	intrabdominal	intra-abdominal	86.67
p/w	presented with	presents with	84.61
dvt	deep vein thrombosis	deep venous thrombosis	81.82
co	complained about	complained of	69.23
angio	angiography	angiogram	66.67
dec	dead	deceased	50.0
x3	three	three times	45.45
p.a.	personal assistant	physician assistant	68.42
pms	previous medical symptoms.	premenstrual syndrome	28.57
vma	visual motor assessment	vanillylmandelic acid	0.0
opa	office of personnel administration	oropharyngeal airway	-25.0

Table 3: Examples of expanded abbreviations, their correct full-forms and the value of the length-normalized inverse edit distance. The threshold is considered 70%, thus the expanded terms with a lower threshold are considered incorrect in our evaluation.

KEPTLongformer[4] and KEPT-PMM3[5]. Following (Yang et al., 2022), we preprocess MIMIC discharge summary by removing de-identification tokens, replacing non-alphanumeric characters (except punctuation) with whitespace, and truncating at 8,192 tokens. If the length exceeded this limit, irrelevant sections were removed prior to truncation to retain the most relevant sections. The MIMIC-III full dataset contains 8,922 ICD codes, making it infeasible to include all descriptions in one prompt. We re-rank the top 300 candidates predicted by MSMN (Yuan et al., 2022) and process 50 candidates at a time, as described in (Yang et al., 2022; Kailas et al., 2023).

To determine the consistency loss weight (α), we perform an ablation study on the MIMIC-III-50 dataset by varying $\alpha \in \{0.02, 0.05, 0.1, 0.2\}$. $\alpha = 0.05$ yields the best performance across all evaluation metrics (see Table 6). This value is applied to all other experiments, as tuning on the larger MIMIC-III-full dataset is computationally expensive. For the top-50 and rare-code datasets, we randomly select 4 synonyms to construct the augmented prompt P_a, following findings from (Yuan et al., 2022; Gomes et al., 2024) that using 4 or 8 synonyms improves ICD coding model performance. For MIMIC-III-full dataset, we use the same code descriptions for both P and P_a, as we observed that incorporating synonyms increases training time and slows convergence on this larger dataset.

[4] https://huggingface.co/whaleloops/keptlongformer

[5] https://huggingface.co/whaleloops/KEPTlongformer-PMM3

All experiments are conducted on a single NVIDIA H100 80GB GPU, with training time and hyperparameters detailed in Appendix A.2. We use the Llama-3.1-70B-Instruct model to perform zero-shot acronym expansion for all ICD coding experiments, except for the ablation study reported in Table 7.

4 Results

4.1 Acronym expansion performance

Our experimental results indicate that the size of the large language model (LLM) used for zero-shot prompting has minimal impact on acronym detection. Detection precision ranges from 93.7% with the smallest model to 96.6% with the largest, while recall increases from 84.5% to 90.5%. In contrast, model size has a substantial effect on acronym expansion accuracy, showing a notable performance improvement of +42 percentage points, from 18.8% with the 1B-parameter model to 60.8% with the 70B-parameter model (Table 7). Evaluating expansions under the lenient accuracy criterion yields additional gains of up to 4%.

Table 3 illustrates example outputs from acronym expansion using the Llama-3.1-70B-Instruct model, along with their corresponding inverse edit distance scores used for lenient accuracy evaluation. As shown in the second section of the table, several expansions remain marked as incorrect even under the lenient 70% similarity threshold, demonstrating our effort to avoid misclassifying incorrect expansions as correct (for example, the first example in the third section). As such, lenient accuracy should be viewed as a conservative,

lower-bound estimate of the model's true acronym expansion performance.

4.2 ICD coding performance

Results show that **ACE-ICD** outperforms previous state-of-the-art methods across all three MIMIC-III datasets (Table 4 and 5), regardless of whether KEPTLongformer or KEPT-PMM3 is used for initialization. From now on, we refer to ACE-ICD (PMM3) as ACE-ICD. To assess statistical significance, we conduct 1,000 rounds of permutation testing comparing the predictions of ACE-ICD and the KEPT baseline. The resulting p-values are all below 0.05 across evaluation metrics and datasets, indicating that our proposed approach yields statistically significant improvements in ICD coding performance.

On the MIMIC-III-50 task (Table 4), ACE-ICD achieves a macro AUC of 94.4 (+0.6), micro AUC of 95.9 (+0.5), macro F1 of 71.6 (+1.2), and micro F1 of 75.0 (+1.0) and precision@5 of 70.0 (+1.1), with values in parentheses indicate improvements over the previous best results. For the MIMIC-III-full task (Table 4), ACE-ICD outperforms prior methods on most metrics except macro F1, achieving a macro F1 of 13.2 (-0.8), micro F1 of 62.7 (+2.0), precision@8 of 79.4 (+1.0), and precision@15 of 63.9 (+0.2). We manage to improve the macro F1 score to 14.3 (+0.3), by applying code-specific threshold optimization (see Appendix A.1). Under the MIMIC-III-rare50 setting (Table 5), ACE-ICD achieves a macro AUC of 91.1 (+2.2), micro AUC of 91.9 (+2.0), macro F1 of 54.0 (+13.7), and micro F1 of 55.8 (+13.2) when fine-tuned from the MIMIC-III-50 checkpoint. Notably, ACE-ICD fine-tuned from KEPT-PMM3 outperforms prior methods initialized from the MIMIC-III-50 checkpoint.

5 Discussion

Effectiveness of our proposed training framework. We evaluate the impact of acronym augmentation and consistency training in ACE-ICD through an ablation study by adding each component separately and assessing performance on the MIMIC-III-50 dataset (Table 6). For data augmentation only, we include the augmented data into the original dataset, doubling the training set size while halving the number of training epochs. For consistency training only, we apply R-Drop (Wu et al., 2021) directly to the original data. Both approaches improve performance over the reproduced KEPT baseline, with consistency training yields better gains. We attribute this to the fact that zero-shot acronym expansion may introduce translation errors, adding noise to the training data. In contrast, R-Drop acts as dropout augmentation and has been shown to enhance ICD coding performance by preventing overfitting (Yuan et al., 2022; Luo et al., 2024). However, the performance further improves when both strategies are applied together.

Impact of acronym expansion quality in improving ICD coding performance. To evaluate how the quality of the acronym expansion process affects ICD coding performance, we conduct an ablation study using Llama 3 models of varying sizes and compare them to a baseline without acronym expansion, as shown in Table 7. We observe that poor expansion quality can hurt performance: the 1B model, with less than 20% accuracy, degrades coding results. Moderate-size models (3B and 8B), achieving 30–50% accuracy, yield only minor gains. In contrast, the 70B model, with over 60% exact match accuracy, provides the most significant performance gains. By observing a few incorrectly expanded acronyms illustrated in Table 3, we hypothesize that expansion errors may have minimal impact if the expanded acronyms are unrelated to any ICD code.

Effect of acronym expansion as data augmentation on different ICD codes. Figure 2 illustrates the F1-score improvement for each ICD code, comparing our ACE-ICD model to the baseline KEPT model in the MIMIC-III-50 settings. The results show that ACE-ICD outperforms the baseline on 39 out of 50 codes, with only marginal decreases observed for the remaining ones. Our method effectively improves the performance of lower-performing codes, including 99.04 (22.8 $\rightarrow$ 29.9), 285.9 (24.5 $\rightarrow$ 35.9), 38.91 (28.0 $\rightarrow$ 43.6), 37.23 (35.2 $\rightarrow$ 51.4), and V15.82 (12.7 $\rightarrow$ 33.4). A Pearson correlation of –0.25 ($p = 0.08$) between absolute F1 improvement and the number of training examples per code suggests that rarer codes tend to benefit more from our method, although the evidence is not statistically significant.

Moreover, although our approach is designed to improve model robustness to the use of acronyms in clinical texts, the impact of acronym expansion on ICD coding performance varies depending on the presence of code-relevant evidence in abbreviated or expanded form across different codes. If several

Methods	MIMIC-III-50					MIMIC-III-full			
	AUC		F1		Pre	F1		Pre	
	Macro	Micro	Macro	Micro	P@5	Macro	Micro	P@8	P@15
MSMN(Yuan et al., 2022)	92.8	94.7	68.3	72.5	68.0	10.3	58.4	75.2	59.9
PLM-ICD(Huang et al., 2022)	-	-	-	-	-	10.4	59.8	77.1	61.3
DiscNet+RE(Zhang et al., 2022)	-	-	-	-	-	**14.0**	58.8	76.5	61.1
KEPT (KL)(Yang et al., 2022) †	92.6	94.8	68.9	72.9	67.3	11.8	59.9	77.1	61.5
CoRelation (Luo et al., 2024)	93.3	95.1	69.3	73.1	68.3	10.2	59.1	76.2	60.7
MSAM(Gomes et al., 2024)	93.7	95.4	70.4	74.0	68.9	-	-	-	-
Extra human annotations									
NoteContrast(Kailas et al., 2023) †	93.8	95.4	69.2	73.6	68.6	11.9	60.7	77.8	62.2
MRR (Wang et al., 2024a)	92.7	94.7	68.7	73.2	68.5	11.4	60.3	77.5	62.3
AKIL (Wang et al., 2024b)	92.8	95.0	69.2	73.4	68.3	11.2	60.5	78.4	63.7
Ours									
KEPT (KL) †∗	93.1	94.9	68.5	72.7	67.6	11.3	60.3	77.5	61.6
KEPT (PMM3) †∗	93.6	95.2	69.6	73.5	68.3	12.9	61.5	78.4	62.7
ACE-ICD (KL) †	**93.9**	**95.6**	**70.9**	**74.5**	**69.2**	11.7	**61.8**	**78.6**	63.0
ACE-ICD (PMM3) †	**94.4**	**95.9**	**71.6**	**75.0**	**70.0**	13.2	**62.7**	**79.4**	**63.9**
GPT4-based									
LLM-codex (Yang et al., 2023a)	92.9	94.8	67.4	71.5	-	-	-	-	-
Multi-Agents (Li et al., 2024)	-	-	**74.8**	58.9	-	-	-	-	-

Table 4: Results on MIMIC-III-50 and MIMIC-III-full datasets, using KEPTLongformer (KL) and KEPT-PMM3 (PMM3). Our reproduced KEPT results (marked with *) closely align with those reported by (Yang et al., 2022). Methods marked with † indicate approaches that re-rank the top 300 predictions from MSMN (Yuan et al., 2022) under the MIMIC-III-full setting.

Methods	Trained from	AUC		F1	
		Macro	Micro	Macro	Micro
MSMN	pre-trained	75.4	77.4	15.3	16.6
KEPT (KL)		79.4	80.7	24.6	23.3
NoteContrast		85.7	86.7	39.0	41.8
ACE-ICD (KL)		**86.8**	**89.1**	37.9	37.7
ACE-ICD (PMM3)		**92.2**	**90.9**	**49.1**	**51.1**
MSMN	MIMIC III-50 checkpoint	59.0	58.9	3.5	5.5
KEPT (KL)		82.3	83.7	29.0	31.4
NoteContrast		88.9	89.9	40.3	42.6
ACE-ICD (KL)		**90.0**	**90.9**	**45.3**	**48.1**
ACE-ICD (PMM3)		**91.1**	**91.9**	**54.0**	**55.8**
GPT4-based					
LLM-codex		82.5	83.2	27.9	30.2
Multi-Agents		-	-	**71.5**	37.6

Table 5: Results on the MIMIC-III-rare50 dataset.

Model	Macro F1	Micro F1	P@5
KEPT (baseline)	69.6	73.5	68.3
+ data augmentation only	70.0	73.8	68.6
+ consistency training only	71.2	74.7	69.6
ACE-ICD (+ both)	**71.6**	**75.0**	**70.0**
$\alpha = 0.2$	71.0	74.6	69.4
$\alpha = 0.1$	**71.6**	**75.0**	69.7
$\alpha = 0.05$	**71.6**	**75.0**	**70.0**
$\alpha = 0.02$	71.4	74.7	69.4

Table 6: Ablation study on MIMIC-III-50.

full-form expressions are already present in the text, acronym expansion may not be necessary for accurate code prediction (e.g., “tka” and “total knee replacement” as evidence for code 81.54 in the example from Table 1).

A case study is shown in Table 1. Our ACE-ICD demonstrate a better understanding of medical acronyms, correctly predicting codes such as 599.0 and 715.36. In this discharge summary (HADM_ID=108519), the only evidence for inferring code 599.0 is the acronym "uti", as other mentions of "urine" in the text are not relevant. Similarly, code 715.36 can be inferred from "oa" or its synonym "osteoarthritis". Despite the KEPT model being pre-trained to incorporate knowledge from UMLS terms, it still fails to correctly predict these codes, highlighting the effectiveness of our approach in handling medical acronyms.

Comparison with recent methods. Our method outperforms previous state-of-the-art models across all three MIMIC-III datasets, including those using additional human annotations or auxiliary clinical knowledge (Kailas et al., 2023; Wang et al., 2024a,b). Notably, our approach significantly improves over the KEPT baseline using a smaller model, KEPTLongformer (149M parame-

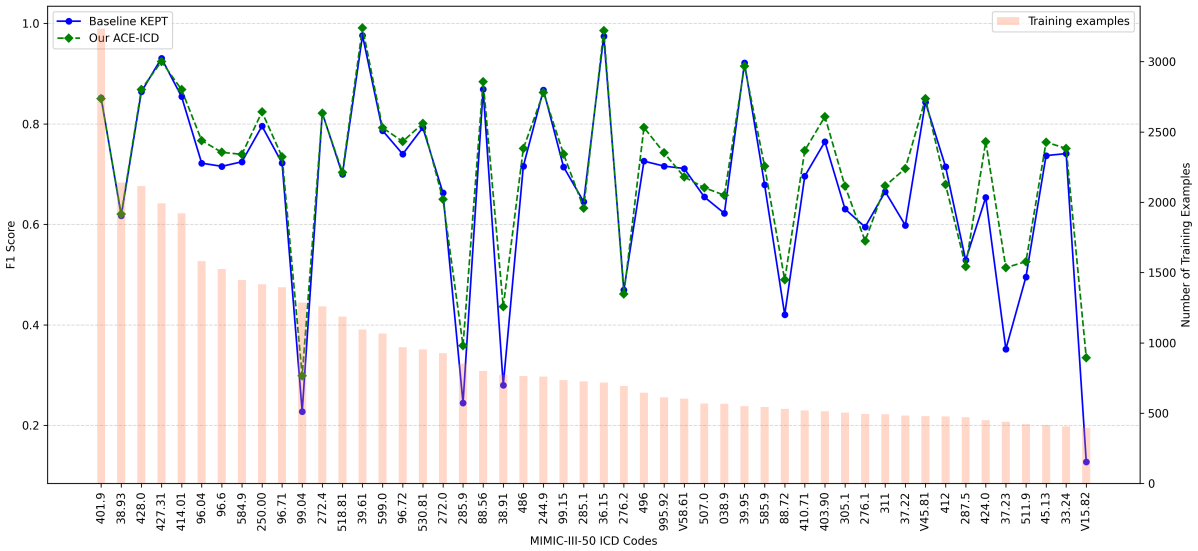

Figure 2: F1 improvement per code in MIMIC-III-50 dataset (sorted by number of training examples in descending order).

Expansion models	Acronym Expansion Performance				ICD coding Performance				
	Detection		Accuracy		AUC		F1		Pre
	Precision	Recall	Strict	Lenient	Macro	Micro	Macro	Micro	P@5
Llama-3.2-1B-Instruct	93.7	84.5	18.8	19.7	94.0	95.6	70.5	73.9	69.2
Llama-3.2-3B-Instruct	95.3	84.3	31.0	33.1	94.2	95.7	71.4	74.7	69.6
Llama-3.1-8B-Instruct	**96.6**	86.9	46.3	49.5	94.1	95.7	71.5	74.8	69.6
Llama-3.1-70B-Instruct	**96.6**	**90.5**	**60.8**	**64.7**	**94.4**	**95.9**	**71.6**	**75.0**	**70.0**
Without acronym expansion (consistency training only)					94.2	95.7	71.2	74.7	69.6

Table 7: Performance of instruction-tuned Llama model variants on acronym expansion and ICD coding. Expansion accuracy is evaluated on the reverse-substituted dataset from Rajkomar et al. (2022), and ICD coding results are reported on MIMIC-III-50.

ters), and even outperforms MSAM(Gomes et al., 2024), built on the larger GatorTron model (345M parameters), under the common code setting.

Even though, GPT-4-based approaches show promising results in ICD coding (Yang et al., 2023a; Li et al., 2024), they still underperform compared to fine-tuned encoder-based models, particularly in terms of micro-F1. The multi-agent framework proposed by Li et al. (2024) achieves a higher macro-F1, which suggests better handling of rare codes due to GPT-4's extensive medical knowledge, but it exhibits lower micro-F1, reflecting challenges in accurately predicting frequent codes.(See tables 4 and 5). Additionally, these methods require repeatedly sending clinical notes to third-party services to access proprietary LLMs, raising concerns around privacy, cost, and scalability. In contrast, our method delivers strong performance on both frequent and rare codes through a lightweight, targeted data augmentation strategy that uses open-source LLMs locally in a one-time preprocessing step, preserving privacy and efficiency.

6 Related works

6.1 Automatic ICD coding

Automatic ICD coding is a multi-label classification task that assigns diagnosis and procedure codes to clinical notes (Perotte et al., 2014; Nguyen et al., 2023a). Early approaches use CNNs (Mullenbach et al., 2018; Xie et al., 2019), LSTMs (Vu et al., 2020; Nguyen et al., 2023b), and Transformers (Huang et al., 2022) to encode clinical notes, while incorporating label attention mechanisms to capture relationships between the notes and ICD codes. Recent studies further improve code representation by integrating multiple code synonyms (Yuan et al., 2022; Gomes et al., 2024) or code relation graph learning (Luo et al., 2024). Contrastive learning has also been applied to improve model capabilities, either between medical entities in UMLS

(Yang et al., 2022) or between clinical notes and ICD codes (Kailas et al., 2023). Several studies enhance ICD coding performance by leveraging the discourse structure of clinical notes, such as using section type embeddings (Zhang et al., 2022) or contrastive pre-training between sections (Lu et al., 2023).

While most methods rely solely on the provided clinical notes and code descriptions, some studies focuses on enhancing ICD coding performance using extra human annotations or data augmentation. Kailas et al. (2023) pre-train a model on temporal sequences of diagnostic codes using proprietary data from a large patient cohort, where clinical notes are paired with ICD-10 codes, to provide a strong initialization for finetuning coding model. Wang et al. (2024a,b) incorporate auxiliary information such as diagnosis-related group (DRG) codes, current procedural terminology (CPT) codes, and prescribed medications to improve performance. Lu et al. (2023) introduce masked section training with small ratio as a data augmentation strategy, following contrastive pre-training between note's sections to boost model performance. Falis et al. (2022) propose ontology-guided synonym augmentation and sibling-code replacement to generate silver training examples. However, their method requires a pretrained named entity recognition and linking system to identify code-relevant text spans.

LLMs have demonstrated remarkable capabilities across various general-domain tasks and have recently been explored for ICD coding (Boyle et al., 2023; Soroush et al., 2024). Yang et al. (2023a) introduced LLM-Codex, which generates ICD codes and evidence with GPT-4, followed by LSTM-based verification. Li et al. (2024) use GPT-4 to convert discharge summaries into Subjective, Objective, Assessment, and Plan (SOAP) format, allowing multiple agents to perform ICD coding via predefined workflows. While promising, these approaches still lag behind fully fine-tuned non-LLM models on MIMIC-III common and rare datasets.

To the best of our knowledge, no prior work has explored augmenting data with acronym expansions and incorporating them via consistency training to enhance ICD coding performance. Moreover unlike previous approaches, our approach does not rely on additional annotations or specialized pre-training. Instead, given the evidence of the zero-shot capabilities of general-purpose LLMs to expand medical acronyms, our approach provides a simple yet effective data augmentation strategy to improve ICD coding performance.

6.2 Clinical acronyms disambiguation

Accurately disambiguating clinical acronyms and abbreviations enhances automated clinical note processing which include medical information retrieval and analysis. Several studies focus on training deep learning models with large amounts of annotated data, including word-embeddings (Jaber and Martínez, 2021; Wu et al., 2015), convolutional neural networks (CNNs) (Skreta et al., 2021), fine-tuning transformer-based models such as BioBERT (Li et al., 2024) and BlueBERT (Hosseini et al., 2024), and fine-tuned encoder-decoder architectures like T5 (Rajkomar et al., 2022). To address the scarcity of annotated data and the data-hungry nature of deep learning models, prior work has explored various data augmentation strategies, including reverse substitution (Rajkomar et al., 2022; Liu et al., 2024; Skreta et al., 2021), UMLS-based similar concept retrieval (Skreta et al., 2021), integration of clinical note metadata (Kugic et al., 2024), and the use of generative clinical models (Hosseini et al., 2024). The potential of LLMs in medical context understanding, coupled with their reduced reliance on large annotated datasets, has driven research toward zero-shot and few-shot acronym disambiguation in clinical text requiring less training cost and effort. Kugic et al. (2024), Liu et al. (2024) and Hosseini et al. (2024) evaluate various LLMs on the CASI dataset (Moon et al., 2012), showing that LLM-based prompting achieves performance comparable to supervised models, even in zero-shot settings.

7 Conclusion

In this paper, we introduce ACE-ICD, a system that advances ICD coding performance by utilizing acronym expansion as an innovative data augmentation technique. Furthermore, we incorporate consistency training, a regularization strategy that enforces alignment between original and augmented documents to enhance model predictions. Our approach also outperforms studies which rely on external annotations or proprietary resources. Our extensive experiments reveal that the combination of LLM-based acronym expansion and consistency training elevates ICD coding accuracy, outperforming existing methods and establishing new state-of-the-art benchmarks across various settings.

Limitations

Our data augmentation approach relies on zero-shot prompting to disambiguate medical acronyms in clinical notes, making its effectiveness dependent on the performance of the selected LLMs. We chose Llama 3.1 70B as it was the best-performing open-source model available at the time of our experiments and aligned with our computational resources. More advanced LLMs or prompting techniques could potentially reduce translation errors and generate higher-quality augmented data.

Our work uses KEPT as the base method, but we argue that acronym expansion as a data augmentation technique, combined with consistency training, can benefit other existing ICD coding systems. Additional experiments are needed to thoroughly assess the effectiveness of our proposed strategy across various models.

8 Ethics Statement

This work uses the publicly available MIMIC-III clinical dataset, which contains de-identified patient information in compliance with HIPAA standards. Access to the dataset requires completion of a data use agreement and training in responsible research conduct. Acronym expansion was performed using open-source LLMs on a secure local cluster, and no patient data were transmitted to any third-party services. Our method is intended to support clinical NLP research and is not designed for direct clinical deployment without expert oversight. We do not anticipate any ethical concerns associated with this study.

References

Temitope Ibrahim Amosa, Lila Iznita Bt Izhar, Patrick Sebastian, Idris B Ismail, Oladimeji Ibrahim, and Shehu Lukman Ayinla. 2023. Clinical errors from acronym use in electronic health record: A review of nlp-based disambiguation techniques. *IEEE Access*, 11:59297–59316.

Iz Beltagy, Matthew E Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.

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

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

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Matúš Falis, Hang Dong, Alexandra Birch, and Beatrice Alex. 2022. Horses to zebras: Ontology-guided data augmentation and synthesis for ICD-9 coding. In *Proceedings of the 21st Workshop on Biomedical Language Processing*, pages 389–401, Dublin, Ireland. Association for Computational Linguistics.

Tianyu Gao, Adam Fisch, and Danqi Chen. 2021. Making pre-trained language models better few-shot learners. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3816–3830, Online. Association for Computational Linguistics.

Goncalo Gomes, Isabel Coutinho, and Bruno Martins. 2024. Accurate and well-calibrated ICD code assignment through attention over diverse label embeddings. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2302–2315, St. Julian's, Malta. Association for Computational Linguistics.

Manda Hosseini, Mandana Hosseini, and Reza Javidan. 2024. Leveraging large language models for clinical abbreviation disambiguation. *Journal of medical systems*, 48(1):27.

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

Areej Jaber and Paloma Martínez. 2021. Disambiguating clinical abbreviations using pre-trained word embeddings. In *Healthinf*, pages 501–508.

Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. 2016. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9.

Prajwal Kailas, Max Homilius, Rahul C. Deo, and Calum A. MacRae. 2023. Notecontrast: Contrastive language-diagnostic pretraining for medical text. In *Proceedings of the 3rd Machine Learning for Health Symposium*, volume 225 of *Proceedings of Machine Learning Research*, pages 201–216. PMLR.

Amila Kugic, Stefan Schulz, and Markus Kreuzthaler. 2024. Disambiguation of acronyms in clinical narratives with large language models. *Journal of the American Medical Informatics Association*, 31(9):2040–2046.

Patrick Lewis, Myle Ott, Jingfei Du, and Veselin Stoyanov. 2020. Pretrained language models for biomedical and clinical tasks: Understanding and extending the state-of-the-art. In *Proceedings of the 3rd Clinical Natural Language Processing Workshop*, pages 146–157, Online. Association for Computational Linguistics.

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

Yikuan Li, Ramsey M Wehbe, Faraz S Ahmad, Hanyin Wang, and Yuan Luo. 2023. A comparative study of pretrained language models for long clinical text. *Journal of the American Medical Informatics Association*, 30(2):340–347.

Ying Liu, Genevieve B Melton, and Rui Zhang. 2024. Exploring large language models for acronym, symbol sense disambiguation, and semantic similarity and relatedness assessment. *AMIA Summits on Translational Science Proceedings*, 2024:324.

Chang Lu, Chandan Reddy, Ping Wang, and Yue Ning. 2023. Towards semi-structured automatic icd coding via tree-based contrastive learning. *Advances in Neural Information Processing Systems*, 36:68300–68315.

Junyu Luo, Xiaochen Wang, Jiaqi Wang, Aofei Chang, Yaqing Wang, and Fenglong Ma. 2024. CoRelation: Boosting automatic ICD coding through contextualized code relation learning. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 3997–4007, Torino, Italia. ELRA and ICCL.

Sungrim Moon, Serguei Pakhomov, and Genevieve Melton. 2012. Clinical abbreviation sense inventory.

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

Thanh-Tung Nguyen, Viktor Schlegel, Abhinav Kashyap, Stefan Winkler, Shao-Syuan Huang, Jie-Jyun Liu, and Chih-Jen Lin. 2023a. Mimic-iv-icd: A new benchmark for extreme multilabel classification. *arXiv preprint arXiv:2304.13998*.

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

Jong-Ku Park, Ki-Soon Kim, Tae-Yong Lee, Kang-Sook Lee, Duk-Hee Lee, Sun-Hee Lee, Sun-Ha Jee, Il Suh, Kwang-Wook Koh, So-Yeon Ryu, et al. 2000. The accuracy of icd codes for cerebrovascular diseases in medical insurance claims. *Journal of Preventive Medicine and Public Health*, 33(1):76–82.

Adler Perotte, Rimma Pivovarov, Karthik Natarajan, Nicole Weiskopf, Frank Wood, and Noémie Elhadad. 2014. Diagnosis code assignment: models and evaluation metrics. *Journal of the American Medical Informatics Association*, 21(2):231–237.

Alvin Rajkomar, Eric Loreaux, Yuchen Liu, Jonas Kemp, Benny Li, Ming-Jun Chen, Yi Zhang, Afroz Mohiuddin, and Juraj Gottweis. 2022. Deciphering clinical abbreviations with a privacy protecting machine learning system. *Nature Communications*, 13(1):7456.

Timo Schick and Hinrich Schütze. 2021. Exploiting cloze-questions for few-shot text classification and natural language inference. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 255–269, Online. Association for Computational Linguistics.

Dinghan Shen, Mingzhi Zheng, Yelong Shen, Yanru Qu, and Weizhu Chen. 2020. A simple but tough-to-beat data augmentation approach for natural language understanding and generation. *arXiv preprint arXiv:2009.13818*.

Marta Skreta, Aryan Arbabi, Jixuan Wang, Erik Drysdale, Jacob Kelly, Devin Singh, and Michael Brudno. 2021. Automatically disambiguating medical acronyms with ontology-aware deep learning. *Nature communications*, 12(1):5319.

Aaron Sonabend, Winston Cai, Yuri Ahuja, Ashwin Ananthakrishnan, Zongqi Xia, Sheng Yu, and Chuan Hong. 2020. Automated icd coding via unsupervised knowledge integration (unite). *International journal of medical informatics*, 139:104135.

Ali Soroush, Benjamin S Glicksberg, Eyal Zimlichman, Yiftach Barash, Robert Freeman, Alexander W Charney, Girish N Nadkarni, and Eyal Klang. 2024. Large language models are poor medical coders—benchmarking of medical code querying. *NEJM AI*, 1(5):AIdbp2300040.

Shang-Chi Tsai, Chao-Wei Huang, and Yun-Nung Chen. 2021. Modeling diagnostic label correlation for automatic ICD coding. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human*

Language Technologies, pages 4043–4052, Online. Association for Computational Linguistics.

Thanh Vu, Dat Quoc Nguyen, and Anthony Nguyen. 2020. A label attention model for icd coding from clinical text. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pages 3335–3341. International Joint Conferences on Artificial Intelligence Organization. Main track.

Xindi Wang, Robert Mercer, and Frank Rudzicz. 2024a. Multi-stage retrieve and re-rank model for automatic medical coding recommendation. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4881–4891, Mexico City, Mexico. Association for Computational Linguistics.

Xindi Wang, Robert E. Mercer, and Frank Rudzicz. 2024b. Auxiliary knowledge-induced learning for automatic multi-label medical document classification. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 2006–2016, Torino, Italia. ELRA and ICCL.

Lijun Wu, Juntao Li, Yue Wang, Qi Meng, Tao Qin, Wei Chen, Min Zhang, Tie-Yan Liu, et al. 2021. R-drop: Regularized dropout for neural networks. *Advances in Neural Information Processing Systems*, 34:10890–10905.

Yonghui Wu, Jun Xu, Yaoyun Zhang, and Hua Xu. 2015. Clinical abbreviation disambiguation using neural word embeddings. In *Proceedings of BioNLP 15*, pages 171–176.

Xiancheng Xie, Yun Xiong, Philip S. Yu, and Yangyong Zhu. 2019. 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*, CIKM '19, page 649–658, New York, NY, USA. Association for Computing Machinery.

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

Zhichao Yang, Sunjae Kwon, Zonghai Yao, and Hong Yu. 2023b. Multi-label few-shot icd coding as autoregressive generation with prompt. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 5366–5374.

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. *Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing*, 2022:1767–1781.

Methods	Dev F1		Test F1	
	Macro	Micro	Macro	Micro
ACE-ICD (PMM3)				
w/ single threshold	10.5	62.9	13.2	**62.7**
w/ code-specific thresholds	14.9	66.0	**14.3**	**61.5**
DiscNet+RE	-	-	14.0	58.8

Table 8: Results of different threshold optimization approaches on MIMIC-III-full dataset.

Zheng Yuan, Chuanqi Tan, and Songfang Huang. 2022. 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, Dublin, Ireland. Association for Computational Linguistics.

Shurui Zhang, Bozheng Zhang, Fuxin Zhang, Bo Sang, and Wanchun Yang. 2022. Automatic ICD coding exploiting discourse structure and reconciled code embeddings. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 2883–2891, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

A Appendix

A.1 Threshold Optimization

All results presented in the main paper are obtained using a single threshold for all ICD codes, optimized for the micro F1. An alternative approach is to determine a specific threshold for each ICD code. This method effectively lowers the threshold for ICD codes with less training data compared to a single-threshold approach, leading to an improvement in the macro F1 and surpassing (Zhang et al., 2022) on the MIMIC-III-full dataset. However, this strategy tends to overfit the development set, increasing the performance gap between the development and test sets.

A.2 More implementation details

Pretrained models. We initialize our model with two pretrained variants provided by (Yang et al., 2022). KEPTLongformer is based on Clinical Longformer (Li et al., 2023), while KEPT-PMM3 builds upon RoBERTa-base-PM-M3-Voc-distill (Lewis et al., 2020), a distilled variant of RoBERTa-large pre-trained on PubMed, PMC, and MIMIC-III corpus. These models adapt the Longformer sparse attention mechanism (Beltagy et al., 2020) to handle longer sequences and incorporate medical knowledge through contrastive learning.

Training Details. Table 9 summarizes the training hyperparameters for the three MIMIC-III

datasets. Training ACE-ICD (PMM3) for 8 epochs on a single NVIDIA H100 GPU takes approximately 25 minutes for MIMIC-III-rare, 7 hours for MIMIC-III-50, and 5 days for MIMIC-III-full.

Inference. Evaluating 1,573 examples from the development set of MIMIC-III-50 takes approximately 1 minute and 36 seconds on a single H100 GPU, achieving a throughput of around 16 examples per second. For MIMIC-III-full, the model requires six runs to re-rank 300 candidates, resulting in a throughput of 2.67 examples per second.

Configuration	**MIMIC-III-50**	**MIMIC-III-rare50**	**MIMIC-III-full**
global attention on	code descriptions + masks	code descriptions + masks	code descriptions + masks
global attention stride	1	1	3
synonyms in prompt	yes	yes	no
max length	8192	8192	8192
num epochs	8	8	4
batch size	1	1	1
gradient accumulation steps	1	1	6
learning rate	1.5e-5	1.5e-5	1.5e-5
learning rate scheduler	cosine	cosine	cosine
max grad norm	1	1	1
warm up ratio	0	0	0.1
AdamW epsilon	1e-6	1e-6	1e-7
AdamW betas	(0.9, 0.999)	(0.9, 0.999)	(0.9, 0.999)
weight decay	0.01	0.01	1e-4

Table 9: Training hyperparameters used in our experiments for the three ICD coding tasks on the MIMIC-III dataset.