How to leverage large language models for automatic ICD coding

Anonymous EACL submission

Abstract

ICD coding, which indicates assigning appropriate ICD codes to clinical notes, is imperative for various healthcare circumstances such as health expense claims, insurance claims, and disease research. However, clinical notes contain numerous non-grammatical expressions, abbreviations, professional terms, and synonyms, rendering them notably noisy compared to general documents. Additionally, ICD coding also presents challenges such as a broad label space and a long-tail problem. While Large Language Models (LLMs) possess exceptional ability for natural language comprehension and thus hold potential for high-quality ICD coding, fine-tuning considering the unique properties of clinical notes and ICD codes is requisite. In this research, we propose a novel fine-tuning framework for LLMs toward automatic ICD coding. Our framework includes additional structures of label attention mechanism, note-relevant knowledge injection based on medical expressions, and knowledge-driven sampling for input clinical notes to navigate the input token limitations of LLMs. Our experiments on the MIMIC-III-50 dataset demonstrate that our framework achieves higher scores across both micro and macro measurements compared to the vanilla fine-tuning framework, with notably enhanced performance improvements observed in encoder-decoder models.

1 Introduction

The International Classification of Disease (ICD) is a global healthcare classification system established by the World Health Organization (WHO) (Shull, 2019). Assigning ICD codes is crucial because the assigned codes are utilized for various purposes including health expense claims, insurance claims, and disease research. ICD coding by humans is heavily dependent on clinical knowledge, and it is labor-intensive and time-consuming, rendering the outcome susceptible to human errors (Adams et al., 2002). For that reason, there has been an ongoing need for automatic ICD coding.

The ICD coding task has two main challenges to be addressed. First, clinical notes are noisy and vary in length. They contain synonyms and abbreviations of clinical terminologies which may vary by region, institution, and individual. The clinical notes also include many fragmented sentences without proper grammatical structure. Furthermore, they vary widely in length depending on the patient's medical history. For instance, the Medical Information Mart for Intensive Care III (MIMIC-III) (Johnson et al., 2016) dataset, a commonly used medical database, contains clinical notes that range in length from less than 500 words to over 3000 words. These could be substantive challenges for both humans and machines in interpreting clinical notes and assigning ICD codes accordingly (Yu et al., 2002; Zhou et al., 2021). Second, ICD coding considers a broad label space with a long-tail problem. In the MIMIC-III dataset, the top 10% of all ICD codes account for 85% of all code occurrences, while about 22% of codes appear no more than twice (Zhou et al., 2021). Even among the top 50 most frequent codes, the most frequent code appears about 3,200 times, and the least frequent code appears about 500 times. This extremely unequal distribution of appearances makes it difficult to develop a reliable ICD code classifier (Japkowicz and Stephen, 2002; Buda et al., 2018).

In recent years, Large Language Models (LLMs) have significantly enhanced the ability of machines to understand and generate natural language (Ouyang et al., 2022; Nori et al., 2023; Howard and Ruder, 2018). However, the direct adoption of LLMs in the medical domain encompasses risks due to the relatively insufficient medical domain data during the training of the LLMs. The shortage of medical domain knowledge often leads to generating erroneous responses to questions that require medical expertise (Gilson et al., 2023). In our exploration, OpenAI's GPT-4 (OpenAI, 2023) and Meta-AI's LLaMA (Touvron et al., 2023a) frequently fail to provide the correct description for ICD-9 codes. For example, when we requested the description of ICD-9 code 36.15 to the models, GPT-4 answered 'Insertion of drug-eluting coronary artery stent', and LLaMA answered 'Acute myocarditis'. Both answers are entirely irrelevant to the true description, 'Single internal mammary-coronary artery bypass'. These results highlight the insufficient training of the current LLMs with regard to the medical domain. Therefore, additional fine-tuning of LLMs for ICD coding is required to utilize LLMs for automatic ICD coding.

In this paper, we propose a novel fine-tuning framework for automatic ICD coding based on clinical notes, including three elements. First, we enhance the encoding performance of the LLMs by integrating a label attention mechanism (Vu et al., 2021), which has demonstrated efficacy for multi-class multi-label tasks. Second, we implement a note-relevant medical knowledge injection mechanism to supplement the LLMs with additional information pertaining to the medical expressions, abbreviations, and various synonyms present in clinical notes. Finally, we apply knowledge-based sampling to the clinical note input to ensure that the LLMs verify as much important information as possible within limited input.

2 Related works

Research on machine learning-based automatic ICD coding began in the 1990s. Larkey and Croft (1996) proposed an ICD code classifier using traditional machine learning algorithms such as the K-nearest neighbor, relevance feedback, and Bayesian independence. With the rise of deep learning, Mullenbach et al. (2018) introduced CAML, which employs convolutional neural networks (CNNs) and a label-wise attention mechanism. Xie et al. (2019) also utilized the densely connected CNNs and multi-scale feature attention to enhance the efficacy of feature extraction. Li and Yu (2020) and Ji et al. (2020) adopted residual connections and dilated convolutions to CNNs for automatic ICD coding, respectively. Recurrent neural network (RNN)-based automatic ICD coding has also been actively studied. Shi et al. (2017) and Xie and Xing (2018) attempted the automatic ICD coding using the attentive long short term memory (LSTM), and tree-of-sequences LSTM network, respectively. Vu et al. (2021) designed a hierarchical classifier utilizing LSTM and label attention mechanism and achieved significant performance improvement. Nonetheless, these methods showed the limited capability of interpreting medical notes composed of diverse and noisy text.

The development of LLMs has driven dramatic performance improvements across numerous natural language processing tasks. Google's Text-to-Text Transfer Transformer (T5) transposes a broad range of natural language processing tasks into a text-to-text format (Raffel et al., 2020). Subsequent to its success, OpenAI introduced ChatGPT (ope) and GPT-4 (OpenAI, 2023), demonstrating innovative performances. Furthermore, Meta-AI has introduced the open-source LLMs, LLaMA (Touvron et al., 2023a) and LLaMA2 (Touvron et al., 2023b), leading the development of subsequent models, such as Alpaca (Taori et al., 2023) and Vicuna (Zheng et al., 2023). Leveraging LLMs for the medical domain, ClinicalT5 fine-tuned T5 for the MIMIC-III dataset and achieved higher performance than T5 on several medical benchmark datasets. ChatDoctor, a fine-tuned LLaMA based on 100K patient-physician conversations collected from online medical consultation websites (Yunxiang et al., 2023), performed similar to or better than ChatGPT for a variety of medical queries. Medalpaca recorded high scores on the United States Medical Licensing Examination (USMLE) by fine-tuning LLaMA for self-collected medical datasets (Han et al., 2023). PMC-LLaMA, a fine-tuned LLaMA using a knowledge injection dataset constructed from 4.8M academic papers and 30k medical books and a medical-specific instruction tuning dataset comprising 202M tokens, demonstrated top-tier performance in the Medical QA task (Wu et al., 2023). Nevertheless, there has been no exploration into fine-tuning LLMs for classifying ICD codes from complex and noisy clinical notes. To the best of our knowledge, this study is the first attempt to find an optimal way for fine-tuning LLMs toward automatic ICD coding.

3 Methods

We propose a fine-tuning framework toward the automatic ICD coding for two types of LLMs, the encoder-decoder models (e.g. T5) and the decoder-only models (e.g. LLaMA) which is illustrated in Fig. 1. Our framework contains a label attention mechanism, note-relevant knowledge injec-

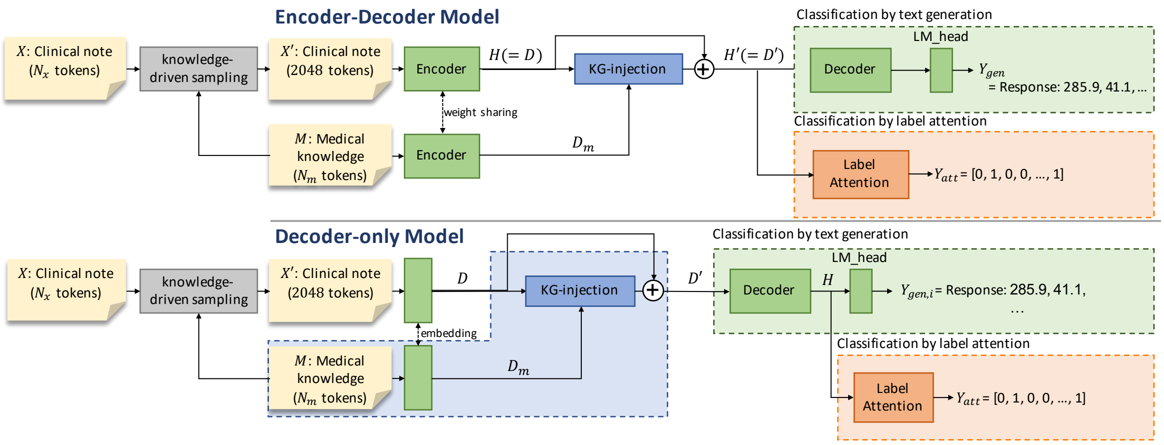

Figure 1: Structural outline of our proposed framework with label attention, note-relevant knowledge injection, and knowledge-driven sampling for encoder-decoder and decoder-only models. The blue box in the decoder-only model is not adopted for our final results because the module degrades the fine-tuning performance.

tion (KG-injection), and knowledge-driven sampling. The model is fine-tuned to predict true assigned codes C from the entire codes $C_{total} = \{c_1, c_2, ..., c_{N_C}\}$ based on a clinical note input X with the prefix prompt (detailed in Appendix A). An objective function L_{gen} to train the LLMs for generating proper ICD codes is defined as a cross-entropy function between the assigned codes C and generated output text Y_{gen}.

3.1 Label attention for LLM-based ICD coding

In order to encourage feature extraction for multi-label classification, we integrate the label attention mechanism (Vu et al., 2021) with LLMs to efficiently solve the multi-label binary classification for a broad label space. The input of the label attention layer is defined as the output of the encoder and decoder for the encoder-decoder and decoder-only models, respectively, as shown in Fig. 1. Given the number of tokens in the input text N_x and the dimension of the hidden state d_h, the input $H \in \mathbb{R}^{N_x \times d_h}$ for the label attention layer is defined as:

$$H = \mathcal{F}(X),$$
$$\mathcal{F} = \begin{cases} \text{encoder, if encoder-decoder model} \\ \text{decoder, if decoder-only model.} \end{cases} \tag{1}$$

Then, the output Y_{att} indicating the possibilities to be assigned codes C_{total} is defined as:

$$Z = \tanh(HW) \tag{2}$$

$$V = \text{softmax}(UZ^{\top})H \tag{3}$$

$$Y_{att} = \text{fcn}(V) \tag{4}$$

where $W \in \mathbb{R}^{d_h \times d_a}$ and $U \in \mathbb{R}^{N_c \times d_a}$ are trainable weight matrices, and d_a is the pre-defined dimension of hidden space. fcn represents fully connected layers to classify the label domain feature $V \in \mathbb{R}^{N_c \times d_h}$ to output possibility $Y_{att} \in \mathbb{R}^{N_c \times 1}$. Consequently, the objective function L_{att} is defined as a cross-entropy function between Y_{att} and C_{att}, the latter being the binary labels indicating whether each code is in the C. The final objective function L_{total} is defined as the summation of L_{gen} and L_{att}.

The final ICD code prediction Y for the clinical note X is obtained as

$$Y = \lambda Y_{gen} + (1 - \lambda)Y_{att}, \tag{5}$$

where the weight value λ is determined depending on the classification performance for put-aside validation data. As Y_{gen} contains multiple codes, we elect to use a binary weight for the ensemble rather than to extract assigning possibilities for each code.

3.2 Note-relevant medical knowledge injection

To enhance LLMs' understanding of various professional terms, abbreviations, and synonyms in clinical notes, we propose a KG-injection with knowledge data for ICD-9 codes which we built using ChatGPT. The details of the knowledge data are in Appendix B. Given the knowledge data M, latent features D and D_m for the clinical notes and knowledge data are obtained as follows:

Input	Output	T5-base				LLaMA-7B			
		F1		Accuracy		F1		Accuracy	
		macro	micro	macro	micro	macro	micro	macro	micro
clinical note	assigned code	**51.09**	**57.4**	**36.31**	**40.25**	**45.24**	**52.53**	**29.95**	**35.62**
clinical note	assigned code + description	48.69	55.64	34.11	38.54	36.51	46.14	23.13	29.99
clinical note + total code	assigned code	50.99	57.14	36.32	40.00	35.33	43.51	20.87	27.80
clinical note + total code + description	assigned code	50.62	56.30	35.69	39.18	36.01	44.27	21.59	28.43

Table 1: Results of T5-base and LLaMA-7B fine-tuned by the MIMIC-III-50 dataset, employing different input-output formats.

$$D = \mathcal{G}(X) \text{ and } D_m = \mathcal{G}(M), \text{where}$$
$$\mathcal{G} = \begin{cases} \text{encoder, if encoder-decoder model} \\ \text{embedding, if decoder-only model.} \end{cases} \tag{6}$$

Given N_m that denotes the number of tokens in the knowledge data, the attention matrix A that represents the attention between the clinical note input and knowledge data is derived using the following equations.

$$Z_m = D_m W \tag{7}$$

$$A = \text{softmax}(D Z_m^T) \tag{8}$$

Then, the attention-applied feature D' is obtained by

$$D' = D + A Z_m. \tag{9}$$

3.3 Knowledge-driven sampling for clinical notes

Input sequences to LLMs have a limited length because of resource constraints, which inevitably results in information loss for long clinical notes. When truncating the MIMIC-III discharge summaries to a limited sequence length, 2048 tokens in our experiments, 44.66% of the total tokens are eliminated, and in the case of the lengthiest discharge summary, 92.44% of the total tokens are eliminated. LAAT (Vu et al., 2021), an LSTM-based automatic ICD coding method, scored 66.6 in macro F1 and 71.5 in micro F1 when truncating inputs to 4000 tokens. However, when the input text is truncated to 2048 tokens, the scores diminished to 48.80 in macro F1 and 58.75 in micro F1 in our experiments. Considering the substantial information loss, it is imperative to strategically include important information from long documents for LLMs. We proposed a knowledge-driven sampling approach to select meaningful parts from the clinical notes.

Clinical notes usually can be divided into several sections. Given the tokens T_i in i-th section of the discharge summary, the number of tokens associated with the assigned code C is defined as:

$$N_{T_i} = \sum_{w \in T_i} \mathbb{1}(w \in \bigcup_{c_j \in C} M_j), \tag{10}$$

where M_j denotes a subset of the knowledge data M associated with the code $c_j \in C$. Sections are primarily selected based on N_{T_i}, and subsequently chosen and sorted according to the importance ratio p, which is defined as:

$$p_{T_i} = \frac{N_{T_i}}{|T_i|}. \tag{11}$$

After selecting sections, paragraphs within each section are ordered according to the paragraph-level importance ratio that is defined in the same manner with p_{T_i}.

4 Experiments

4.1 MIMIC-III-50 dataset

We used the discharge summaries and manually annotated ICD-9 codes from the MIMIC-III dataset to validate the proposed framework as in previous ICD coding studies. We followed the data processing of CAML (Mullenbach et al., 2018) and employed the MIMIC-III-50 dataset for experiments, a subset associated with the top 50 most frequently occurring codes (Mullenbach et al., 2018). This subset encompasses 11,368 discharge summaries, of which 8,066 samples were utilized for training, 1,573 for validation, and 1,729 for testing. We investigated our proposed approach using macro and micro F1-scores along with macro and micro accuracy.

4.2 Training details

Four NVIDIA V100 GPUs were used for the training and testing. We applied a full parameter fine-tuning for the T5 and ClinicalT5 models, while the decoder-based models with 7B parameters were fine-tuned using Low-Rank Adaptation (LoRA) (Hu et al., 2021) with (8, 16) coefficients due to the hardware limitation. The length of input tokens was limited to 2,048 throughout all experiments. The AdamW optimizer was utilized for training, and learning rates of 1e-4 and 3e-4 were applied to encoder-decoder models and decoder-only models, respectively. We employed the base T5 and ClinicalT5 models with 220m parameters, while 7B models were adopted for LLaMA, LLaMA2, Alpaca, Vicuna, MedAlpaca, and PMC-LLaMA. Fine-tuning T5 and ClinicalT5 required 8 GPU days, while the others required 28 GPU days. Because of the long GPU days for training, we report experimental results based on a single run of training.

4.3 Experimented LLMs

4.3.1 T5

T5 (Raffel et al., 2020) is a large-scale transformer-based language model developed by Google Research. The model conceptualized all NLP tasks as a "text-to-text transformation" problem, facilitating a consistent framework for numerous NLP tasks.

4.3.2 ClinicalT5

ClinicalT5 (Lu et al., 2022) is a model derived from T5, fine-tuned on the MIMIC-III dataset.

4.3.3 LLaMA

LLaMA (Touvron et al., 2023a) is an open-source large language model trained on trillions of tokens from publicly available datasets without instruction tuning.

4.3.4 LLaMA-2

LLaMA-2 (Touvron et al., 2023b) is an advanced version of LLaMA. It applied a grouped query attention mechanism and was trained on a dataset 40% larger than LLaMA.

4.3.5 PMC-LLaMA

PMC-LLaMA (Wu et al., 2023) is a model derived from LLaMA, fine-tuned on 4.8M medical documents for knowledge injection and 202M tokens for medical-specific instruction tuning.

4.3.6 Alpaca

Alpaca (Taori et al., 2023) is a fine-tuned version of Meta-AI's LLaMA-7B. The model was trained on 52K instruction-following demonstrations generated in the style of self-instruct using GPT-3.5.

4.3.7 Med-Alpaca

Med-Alpaca (Han et al., 2023) is a medical-specific version of Alpaca. The model is fine-tuned on medical domain datasets incorporating many medical question-answer pairs.

4.3.8 Vicuna

Vicuna (Zheng et al., 2023) is a chatbot model trained on LLaMA using a dialogue corpus collected from ShareGPT (sha).

4.4 Results

4.4.1 Evaluation on input-output formats

We investigated the ICD coding performance depending on input-output formats. The following four input-output formats were examined.

1. Input: X, Output: C
2. Input: X, Output: C with description
3. Input: X + C_{total}, Output: C
4. Input: X + C_{total} with description, Output: C

Table 1 shows the results of T5-base and LLaMA-7B models fine-tuned on the MIMIC-III-50 dataset employing the aforementioned input-output formats. The first format consistently exhibits superior performance, which probably indicates that incorporating additional information into the input and output reduces the portion of the clinical note information, and subsequently degrades the ICD coding performance. Based on these results, all subsequent experimental results adhered to the first format.

4.4.2 Comparison between vanilla and proposed fine-tuning frameworks

Table 2 describes the ICD coding performance of the LLMs after fine-tuning using the MIMIC-III-50 dataset. The encoder-decoder models achieved

	Baseline fine-tuning				Fine-tuning with the proposed framework			
	F1		Accuracy		F1		Accuracy	
model	macro	micro	macro	micro	macro	micro	macro	micro
T5-base	51.09	57.40	36.31	40.25	56.01	64.14	41.27	47.21
ClinicalT5	**56.59***	**63.07***	**40.05***	**46.06***	**58.88***	**65.27***	**43.72***	**48.44***
LLaMA	45.24	52.53	29.95	35.62	47.86	54.84	32.73	37.84
LLaMA-2	**49.60**	**56.11**	31.20	**39.00**	**49.90**	**57.05**	32.49	**39.98**
PMC-LLaMA	45.45	52.30	30.73	35.41	47.47	53.70	32.63	36.70
Alpaca	44.05	50.41	28.74	33.70	46.18	53.83	31.18	36.82
Med-Alpaca	43.76	50.87	28.92	34.11	46.21	52.29	31.74	35.40
Vicuna	48.24	54.93	**33.23**	37.87	48.61	55.09	**33.35**	38.02

Table 2: Comparison of fine-tuning results using vanilla and the proposed frameworks on the MIMIC-III-50 dataset. The bold numbers denote the best performance within each architecture type, and "*" denotes the overall best performance.

higher scores compared to the decoder-only models. Among the decoder-only models, LLaMA2, the most recently introduced model, showed the best performance. This indicates the correlation between natural language understanding capability and ICD coding performance. ClinicalT5 outperformed other models including T5, likely attributable to its prior fine-tuning on the MIMIC-III dataset. However, PMC-LLaMA and MedAlpaca did not evidently demonstrate performance enhancement compared with their baseline models, i.e., LLaMA and Alpaca. This indicates the significant divergence between clinical notes and other medical documents, demonstrating the necessity for fine-tuning specifically geared toward ICD coding.

The column on Fine-tuning with the proposed framework in Table 2 presents the results of our proposed framework on automatic ICD coding. Deriving from the results of ablation studies introduced in section 4.4.3, our fine-tuning framework for T5 and ClinicalT5 integrates the label attention mechanism, KG-injection employing medical expressions, and knowledge-driven sampling for input clinical notes. For the decoder-only models, the label attention mechanism and knowledge-driven sampling excluding KG-injection were applied considering the structural limitations mentioned in section 4.4.3. As shown in Table 2, our proposed framework exhibited enhanced performance across all models compared to the baseline. The most significant performance improvement was observed in T5-base, with score increments of 4.92, 6.74, 4.96, and 6.96 in macro F1, micro F1, macro accuracy, and micro accuracy, respectively, compared to the baseline. The disparity of the performance improvements between encoder-decoder and decoder-only models signifies that our framework elicits more performance enhancements in the former, due to the enhanced efficiency attained when label attention and KG-injection are implemented within the latent space. Same as the baseline fine-tuning result, ClinicalT5 showed the highest performance with fine-tuning using our proposed framework and LLaMA2 showed the highest performance among the decoder-only models.

4.4.3 Ablation studies

We demonstrate the effectiveness of the proposed framework by conducting ablation studies. Table 3 summarizes the results of ablation studies, where the "Classification", "KG-injection", and "Note sampling" sections present the ICD coding performance with regard to label attention, KG-injection, and knowledge-driven sampling, respectively.

Label attention Where *gen (base)* denotes the model fine-tuned with the vanilla fine-tuning framework, *att* in the "Classification" section indicates the fine-tuned model only based on the classification using label attention. *cmp-gen* and *cmp-att* refer to the ICD coding result from the text generation and label attention classification, respectively, where the models were fine-tuned using both the label attention and text generation losses.

The results show that simply applying the label attention classifier does not improve the ICD coding performance (*gen(base)* vs. *att*), which indicates the inefficacy of removing the text generation part of LLMs because the models were pre-trained for the text generation. In contrast, the

		T5-base				LLaMA-7B			
		F1		Accuracy		F1		Accuracy	
Method	Setting	macro	micro	macro	micro	macro	micro	macro	micro
Baseline	gen (base)	50.31	57.84	36.31	40.25	45.24	52.53	29.95	35.62
Classification	att	47.68	55.43	34.17	38.34	39.43	51.32	26.87	34.52
	cmp-gen	51.09	57.40	35.37	40.69	**47.05**	**53.44**	**32.06**	**36.47**
	cmp-att	**52.02**	**60.11**	**37.10**	**42.97**	39.02	51.55	25.79	34.73
KG-injection	code description	53.47	59.36	38.79	42.21	39.31	46.34	25.11	30.16
	medical expressions	**55.11**	**60.94**	**39.82**	**43.82**	**40.77**	**47.21**	**25.85**	**30.90**
Note sampling	section level	**55.27**	59.28	39.17	41.22	45.30	52.07	30.35	35.20
	+paragraph level	55.15	**61.10**	**40.00**	**43.99**	**47.00**	**54.39**	**32.21**	**37.35**

Table 3: Comparison of each component in the proposed method for T5-base and LLaMA-7B model on the MIMIC-III-50 dataset. '+paragraph level' denotes the paragraph-level sampling following the section-level sampling.

cmp-att results of the T5-base and *cmp-gen* results of the LLaMA-7B are better than the *gen (base)* results, respectively. This indicates that integrating the label attention with the text generation effectively enhances the understanding of clinical notes for ICD coding. The different performance superiority between *cmp-gen* and *cmp-att* toward T5-base and LLaMA-7B can be considered due to the structural distinctions of the models, i.e., encoder-decoder and decoder-only. In the encoder-decoder models, the encoding and decoding processes are structurally separated, allowing label attention to directly influence feature encoding. In contrast, the decoder-only models merge encoding and decoding, thus label attention only exerts an indirect effect on feature encoding.

Note-relevant knowledge injection We compared two types of knowledge data: detailed code descriptions and medical expressions (see the details in Appendix B). Applying KG-injection yielded performance improvement for the T5-base model. Specifically, fine-tuning using KG-injection with medical expressions achieved additional score gains of 4.80, 3.10, 3.51, and 3.57 for the macro F1, micro F1, macro accuracy, and micro accuracy, respectively, compared with the baseline results. The substantial enhancement in the macro scores particularly indicates that KG-injection improves the performance of codes with lower occurrence frequencies.

In contrast, applying KG-injection to the LLaMA-7B precipitated the performance decline. This can be attributed to the structural differences between encoder-decoder and decoder-only models. In the encoder-decoder models, KG-injection is applied in a well-reduced latent space following the encoder, whereas, in the decoder-only models, KG-injection is applied in a broader space close to the observation space following the embedding layers. Consequently, KG-injection is likely to be inefficient in the decoder-only models.

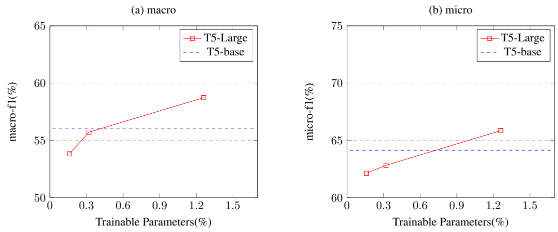

Figure 2: Performance comparison of full fine-tuned T5-base and LoRA fine-tuned T5-large depending on the ratio of trainable parameters for T5-large LoRA fine-tuning.

Knowledge-driven sampling We examined the section-level and paragraph-level knowledge-driven samplings. The proposed knowledge-driven sampling usually constructs input sequences by selecting up to three of the most important sections from the input clinical note (see the details in Appendix C). Regardless of the sampling level, the proposed sampling approach achieved superior scores in all models and metrics compared to the baseline fine-tuning with the slight superiority of paragraph-level sorting combined with section-level sorting. Compared to the baseline fine-tuning, the combination of section-level and paragraph-level sorting shows the performance improvements in macro F1, micro F1, macro accuracy, and micro accuracy were 4.84, 3.26, 3.69, and 3.74 in T5-base, and 1.76, 1.86, 2.26, and 1.73 in LLaMA-7B, respectively. These results also indicate that the knowledge data we built exhibits a correlation with important information within clinical notes.

4.4.4 Explanation for the Performance of Decoder-Only Models

The decoder-only models have demonstrated superior performance in general natural language processing tasks over the encoder-decoder models, owing to the larger number of trainable parameters. However, in our experimental results, the decoder-only models exhibited relatively diminished performance compared to the encoder-decoder models, attributed to the lower number of parameters employed in fine-tuning the decoder-only models. The fine-tuning process for ICD coding with the decoder-only model demands significant resources due to lengthy text and numerous trainable parameters. Constrained by hardware resources, we applied LoRA fine-tuning to the decoder-only model with $r = 8$ and $\alpha = 16$ employing a mere 0.03% of trainable parameters of the 7B models. Those are significantly insufficient for optimally fine-tuning the entire model. Figure 2 illustrates the results of applying LoRA fine-tuning with different coefficients to T5-large with 770M parameters. While the performance excels over the total fine-tuning of the T5-base when the amount of trainable parameters is 1.26% of the total parameters, it recedes below the total fine-tuning of the T5-base at 0.32% and 0.16%. Consequently, a substantial performance enhancement is anticipated in the training of the decoder-only model when either an increment in LoRA coefficients or fine-tuning across all parameters is implemented.

5 Conclusion

In this study, we propose a novel fine-tuning framework for LLMs toward automatic ICD coding. To enhance the performance of multi-class multi-label classification, we adopted a classifier applying a label attention mechanism as an additional classifier. Furthermore, to amplify the capability of understanding diverse medical expressions, abbreviations, and synonyms in clinical notes, we applied KG-injection based on the knowledge data composed of medical expressions. Finally, to overcome the input length limitations of LLMs, we applied knowledge-driven sampling to the input notes grounded on the medical expressions. In experiments across various LLMs, our method demonstrated improved performance compared to the conventional fine-tuning method. Notably, our proposed fine-tuning framework exhibited heightened efficacy in encoder-decoder models, which possess the structure enabling stable application of the label attention mechanism and KG-injection in the latent space.

6 Limitations

Our main limitations come from the restricted resources for experiments. First, the proposed approach was evaluated with the absence of experiments on the MIMIC-III full dataset. Experiments for the MIMIC-III full dataset, which possesses about six times more training samples and over 160 times wider label space than the MIMIC-III-50 dataset, were unfeasible with confined resources. Instead, we exclusively executed experiments using several LLMs and diverse experiment settings with the MIMIC-III-50 dataset. Therefore, we expect that the performance improvement in ICD coding afforded by our proposed framework will be equivalently achieved for the MIMIC-III full dataset.

Second, the fine-tuning of the decoder-only models was conducted by LoRA fine-tuning. Given the extensive trainable parameters of the decoder-only models, we adopted fine-tuning using LoRA exclusively for the 7B models, which probably restricted the potential of the models. Since our proposed fine-tuning framework is not confined to model size, we anticipate it will demonstrate further performance improvement with the full fine-tuning of those models.

Although the proposed fine-tuning approach significantly improves the ICD coding performance of LLMs, the performance is far from practical. Furthermore, because of sequence length constraints, it falls short of other recent methods that don't utilize LLMs. However, we believe that the LLM-based approach is a promising way to solve the challenges in the ICD coding task by leveraging the LLMs' capability of understanding natural language. We hope this study inspires further research to bridge the gap from the general LLMs to practical medical applications.

References

Introducing ChatGPT — openai.com. https://openai.com/blog/chatgpt/. [Accessed 10-10-2023].

ShareGPT: Share your wildest ChatGPT conversations with one click. — sharegpt.com. https://sharegpt.com. [Accessed 12-10-2023].

Diane L Adams, Helen Norman, and Valentine J Burroughs. 2002. Addressing medical coding and billing part ii: a strategy for achieving compliance. a risk management approach for reducing coding and billing errors. *Journal of the National Medical Association*, 94(6):430.

Mateusz Buda, Atsuto Maki, and Maciej A Mazurowski. 2018. A systematic study of the class imbalance problem in convolutional neural networks. *Neural networks*, 106:249–259.

Aidan Gilson, Conrad W Safranek, Thomas Huang, Vimig Socrates, Ling Chi, Richard Andrew Taylor, David Chartash, et al. 2023. How does chatgpt perform on the united states medical licensing examination? the implications of large language models for medical education and knowledge assessment. *JMIR Medical Education*, 9(1):e45312.

Tianyu Han, Lisa C Adams, Jens-Michalis Papaioannou, Paul Grundmann, Tom Oberhauser, Alexander Löser, Daniel Truhn, and Keno K Bressem. 2023. Medalpaca–an open-source collection of medical conversational ai models and training data. *arXiv preprint arXiv:2304.08247*.

Jeremy Howard and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification. *arXiv preprint arXiv:1801.06146*.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Nathalie Japkowicz and Shaju Stephen. 2002. The class imbalance problem: A systematic study. *Intelligent data analysis*, 6(5):429–449.

Shaoxiong Ji, Erik Cambria, and Pekka Marttinen. 2020. Dilated convolutional attention network for medical code assignment from clinical text. *arXiv preprint arXiv:2009.14578*.

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.

Leah S Larkey and W Bruce Croft. 1996. Larkeyandcroft1996. In *Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 289–297.

Fei Li and Hong Yu. 2020. Icd coding from clinical text using multi-filter residual convolutional neural network. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8180–8187.

Qiuhao Lu, Dejing Dou, and Thien Nguyen. 2022. Clinicalt5: A generative language model for clinical text. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 5436–5443.

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

Harsha Nori, Nicholas King, Scott Mayer McKinney, Dean Carignan, and Eric Horvitz. 2023. Capabilities of gpt-4 on medical challenge problems. *arXiv preprint arXiv:2303.13375*.

OpenAI. 2023. Gpt-4 technical report.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551.

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

Jessica Germaine Shull. 2019. Digital health and the state of interoperable electronic health records. *JMIR medical informatics*, 7(4):e12712.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B Hashimoto. 2023. Stanford alpaca: An instruction-following llama model.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023a. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023b. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Thanh Vu, Dat Quoc Nguyen, and Anthony Nguyen. 2021. A label attention model for icd coding from clinical text. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, IJCAI'20.

Chaoyi Wu, Weixiong Lin, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2023. Pmc-llama: Towards building open-source language models for medicine.

Pengtao Xie and Eric Xing. 2018. 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.

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*, pages 649–658.

Hong Yu, George Hripcsak, and Carol Friedman. 2002. Mapping abbreviations to full forms in biomedical articles. *Journal of the American Medical Informatics Association*, 9(3):262–272.

Li Yunxiang, Li Zihan, Zhang Kai, Dan Ruilong, and Zhang You. 2023. Chatdoctor: A medical chat model fine-tuned on llama model using medical domain knowledge. *arXiv preprint arXiv:2303.14070*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *arXiv preprint arXiv:2306.05685*.

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

A Prompts and LLM Response

Figure 3 shows the example of the prompt used in the LLMs' fine-tuning process and the required response for ICD coding.

B Data for knowledge injection

B.1 Detailed code description

The detailed code descriptions, which are used as additional knowledge data in our experiments, contain more detailed information than the official ICD-9 code description. We obtained the data using GPT-3.5-turbo (ope). Figure 4 displays the example of the prompt for obtaining detailed code descriptions using GPT-3.5-turbo and its corresponding response.

B.2 Medical expressions related to ICD-9 codes

The knowledge data consists of all medical expressions related to the ICD code, including medical terms, abbreviations, and synonyms. We obtained the medical expressions pertaining to MIMIC-III-50 ICD codes from GPT-3.5-turbo (ope). In all requests, GPT-3.5-turbo provided 30 medical terms, abbreviations, and synonyms each. Figure 5 presents the prompt and associated response for acquiring medical expression data.

C Results of the knowledge-driven sampling

Table 4 shows the top three sections selected by the knowledge-driven sampling to the MIMIC-III-50 training dataset. The 'Hospital course', 'History of present illness', and 'Pertinent result' sections were identified as the most frequently sampled sections. However, the percentage of sampled sections for each was less than 1%, indicating that the important sections are likely to be different across note samples.

	Hospital course	History present illness	Pertinent result	others
N	58.97	27.69	24.71	less then 8
$p[\%]$	0.12	0.12	0.06	less then 0.03

Table 4: Statistical result of knowledge-driven sampling. N and $p[\%]$ denote the number and the percentage of samples which select the section in the sampling process.

LLM

Q. Below is an instruction that describes a task, paired with an input that provides further context. Write a response that appropriately completes the request.

The following note is the discharge summary of a patient.
Provide all ICD-9 codes applicable to the patient.
Admission Date: XXXX-XX-XX Discharge Date: XXXX-XX-XX

Date of Birth: XXXX-X-XX Sex: M

Service: MEDICINE

Allergies: Penicillins / Shellfish Derived

Attending: First Name3 LF 1253

Chief Complaint: sob, confusion

Major Surgical or Invasive Procedure: Intubation, BAL

History of Present Illness:
56 yom nasal polyps, alcohol abuse, depression, anaphylactic rxn to shellfish and pcn, …

A. Response: 276.1, 285.9, 287.5, 305.1, 486, 507.0, 518.81, 33.24, 38.91, 38.93, 96.6,9 6.72

Figure 3: Example of the prompt and corresponding response for LLMs' fine-tuning

GPT3.5

Q. Provide a description of the given ICD-9 code in one paragraph.
ICD-9 code : 285.9 Anemia, unspecified

A. Anemia, unspecified refers to a general diagnosis of anemia without specifying the underlying cause or type. Anemia is a condition in which there is a deficiency of red blood cells or hemoglobin in the blood, leading to reduced oxygen-carrying capacity and causing symptoms such as fatigue, weakness, and pallor.

Q. Provide a description of the given ICD-9 code in one paragraph.
ICD-9 code : 585.9 Chronic kidney disease, unspecified

A. Chronic kidney disease, unspecified is a code for a progressive and long-term loss of kidney function without specifying the stage or underlying cause. Chronic kidney disease can result from various conditions, such as diabetes, high blood pressure, or autoimmune diseases, and can ultimately lead to kidney failure if not properly managed.

Q. Provide a description of the given ICD-9 code in one paragraph.
ICD-9 code : 38.93 Venous catheterization, not elsewhere classified

A. Venous catheterization, not elsewhere classified, refers to the insertion of a catheter into a vein for various purposes, such as administering medications, monitoring blood pressure, or drawing blood samples. The specific vein used and the purpose of the procedure are not specified in this code.

Figure 4: Examples of the prompts provided to GPT3.5 and their corresponding responses for ICD code description

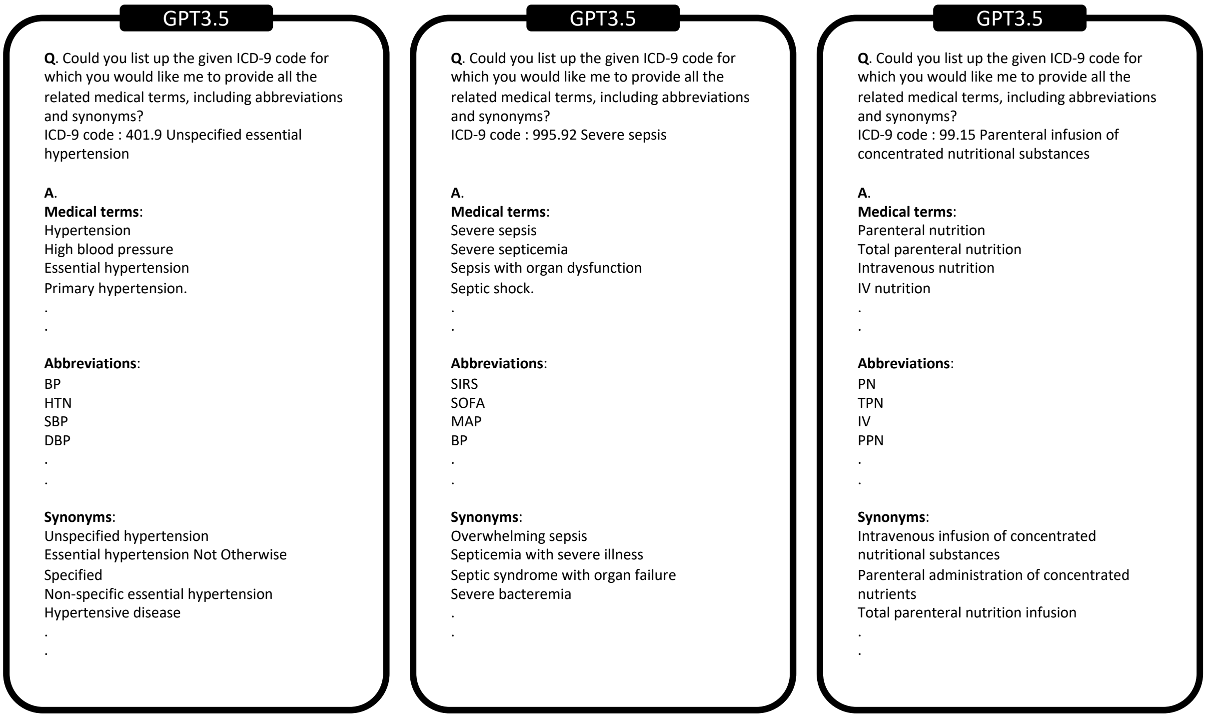

Figure 5: Examples of the prompts provided to GPT3.5 and their corresponding responses for medical terms related to given ICD code