System Report for CCL25-Eval Task 8: Structured ICD Coding with LLM-Augmented Learning and Group-specific Classifiers

Bo Wang[1,†], Kaiyuan Zhang[1,†], Chong Feng[1], Ge Shi[2†], Jinhua Ye[3], Jiahao Teng[2], Shouzhen Wang[1], Fanqing Meng[1], Changsen Yuan[1], Yan Zhuang[4]
[1]School of Computer Science, Beijing Institute of Technology, Beijing, China
[2]Faculty of Information Technology, Beijing University of Technology, Beijing, China
[3]Hangzhou JustHealth Technology Co., Ltd
[4]Medical Innovation Research Department of PLA General Hospital, Beijing 100853, China
[5]Southeast Academy of Information Technology, Beijing Institute of Technology, China

Abstract

The International Classification of Diseases (ICD) provides a standardized framework for encoding diagnoses, serving critical roles in clinical scenarios. Automatic ICD coding aims to assign formalized diagnostic codes to medical records for documentation and analysis, which is challenged by an extremely large and imbalanced label space, noisy and heterogeneous clinical text, and the need for interpretability. In this paper, we propose a structured multi-class classification framework that partitions diseases into clinically coherent groups, enabling group-specific data augmentation and supervision. Our method combines input compression with generative and discriminative fine-tuning strategies tailored to primary and secondary diagnoses, respectively. On the CCL2025-Eval Task 8 benchmark for Chinese electronic medical records, our approach ranked first in the final evaluation.

Keywords: ICD Coding , Adversarial training , Multi-label Classification

1 Introduction

With the increasing adoption of electronic medical records in modern healthcare systems, there is a growing need for automated methods to extract structured clinical knowledge from unstructured narratives. One crucial application is the automatic assignment of the International Classification of Diseases (ICD) codes to discharge summaries and other clinical documents. Accurate ICD coding plays a vital role in medical billing, clinical research, and health statistics (Teng et al., 2022), yet manual coding remains labor-intensive, error-prone, and heavily reliant on domain expertise.

To address the inefficiencies of manual ICD coding, recent research has framed the task as a multi-label text classification problem (Huang et al., 2022, Li and Yu, 2020, Mullenbach et al., 2018, Vu et al., 2020, Wang et al., 2022). However, applying this formulation to real-world clinical settings poses several unique challenges (Dong et al., 2022). First, clinical notes are often lengthy and composed of heterogeneous, noisy content, characterized by domain-specific terminology, abbreviations, and inconsistent structure. Second, the relevance of textual information varies significantly across different ICD codes—some rely on explicit symptom descriptions, while others require inference from longitudinal history or diagnostic tests. Third, the label space exhibits complex dependencies, including hierarchical relationships and frequent co-occurrence, and there is a heightened need for interpretability to ensure clinical reliability and trustworthiness. CCL2025-Eval ICD Coding Task introduces a benchmark dataset tailored for the evaluation of automatic ICD coding on de-identified Chinese electronic medical records. The dataset comprises 1,485 EMRs annotated with primary and secondary diagnosis codes, including 5 distinct primary diagnoses and 53 secondary diagnoses, conforming to the ICD-10 standard. Given a clinical document containing both structured and unstructured fields, the task requires accurately extracting diagnostic information and assigning the appropriate primary and secondary ICD codes.

In this work, we focus on the task of automatic ICD code prediction from real-world inpatient medical records, using rich, fine-grained EMR data that includes chief complaints, disease history, diagnostic

† Corresponding Author, † denotes the equal contribution.

findings, and discharge instructions. ICD coding is formulated as a structured multi-class classification task. Diseases are partitioned into clinically meaningful groups, where labels within each group are mutually exclusive while those across groups are compatible. This structure is exploited to guide data augmentation using fine-grained diagnostic information, thereby enhancing generalization under low-resource conditions. Furthermore, we perform data compression to reduce input noise, thereby improving learning efficiency and robustness. Finally, we apply group-specific supervised fine-tuning (SFT), adopting a generative framework for primary diagnoses and a discriminative framework for secondary diagnoses. This tailored supervision strategy, combined with the representational capacity of decoder-only language models, yields substantial performance improvements on the CCL2025-Eval benchmark.

2 Related Work

Early approaches to ICD coding predominantly employed rule-based systems, wherein domain experts manually devised regular expressions (Zhou et al., 2020), logical rules (Farkas and Szarvas, 2008), keyword matching (Koopman et al., 2015), or dictionary lookups (Seva et al., 2017) to map clinical text to codes. Despite their interpretability and simplicity, such methods lacked scalability and robustness to the variability and non-standardization of clinical language. Subsequent machine learning methods, including support vector machines, logistic regression, and random forests (Huang et al., 2018, Ruch et al., 2008, Schäfer and Friedrich, 2019, Xu et al., 2019), framed ICD coding as a multi-label text classification problem. These models leveraged features such as TF-IDF and Bag-of-Words but were limited in capturing semantic nuances and long-range dependencies within clinical narratives.

Recent advances have centered on deep neural networks that jointly learn feature representations and perform classification. Convolutional neural networks (CNNs), as utilized in models such as CAML (Mullenbach et al., 2018) and MultiResCNN (Li and Yu, 2020), extract hierarchical features from lengthy clinical notes via multi-scale and dilated convolutions. Recurrent architectures, including Bi-GRU and BiLSTM as employed in LAAT (Vu et al., 2020), model sequential dependencies and contextual information, with bidirectional configurations enhancing performance. Transformer-based models represent the current state of the art. Approaches like PLM-ICD (Huang et al., 2022) leverage pre-trained language models to generate rich contextual embeddings, substantially improving performance on the complex, multi-label classification task inherent to ICD coding.

Building upon the strengths of transformer-based architectures, our approach leverages decoder-only language models fine-tuned via generative and discriminative approaches, preserving the model's capability for long-text understanding while enhancing its effectiveness on the ICD coding task.

3 Methodology

As shown in Figure 1, our framework comprises four stages. We explicitly partition the candidate primary and secondary diseases into distinct groups, wherein diseases within the same group are mutually exclusive and diseases across different groups are compatible. Based on the grouping, we employ prompt-based data augmentation to enhance the quantity and diversity of training samples. Furthermore, we compress the medical records to accommodate the input length constraints of language models while reducing noise. Eventually, we fine-tune models with different architectures for primary and secondary diseases, endowing the system with the capability for ICD coding-based diagnosis.

3.1 Structured Disease Grouping

Based on in-depth clinical observation and analysis, we partition the candidate primary and secondary diseases into distinct groups. Specifically, the five primary diseases are categorized into three groups: (1) Coronary Artery Disease Group, (2) Hypertension Group, (3) Heart Failure Group. This grouping reflects clear clinical distinctions: the first group emphasizes coronary perfusion insufficiency, the second highlights hemodynamic overload due to hypertension, and the third focuses on cardiac decompensation. Each group is associated with different pathological mechanisms and characteristic symptoms. Details of the group assignments are illustrated in Table 1.

Proceedings of the 24th China National Conference on Computational Linguistics, pages 312-321, Jinan, China, August 11-14, 2025.

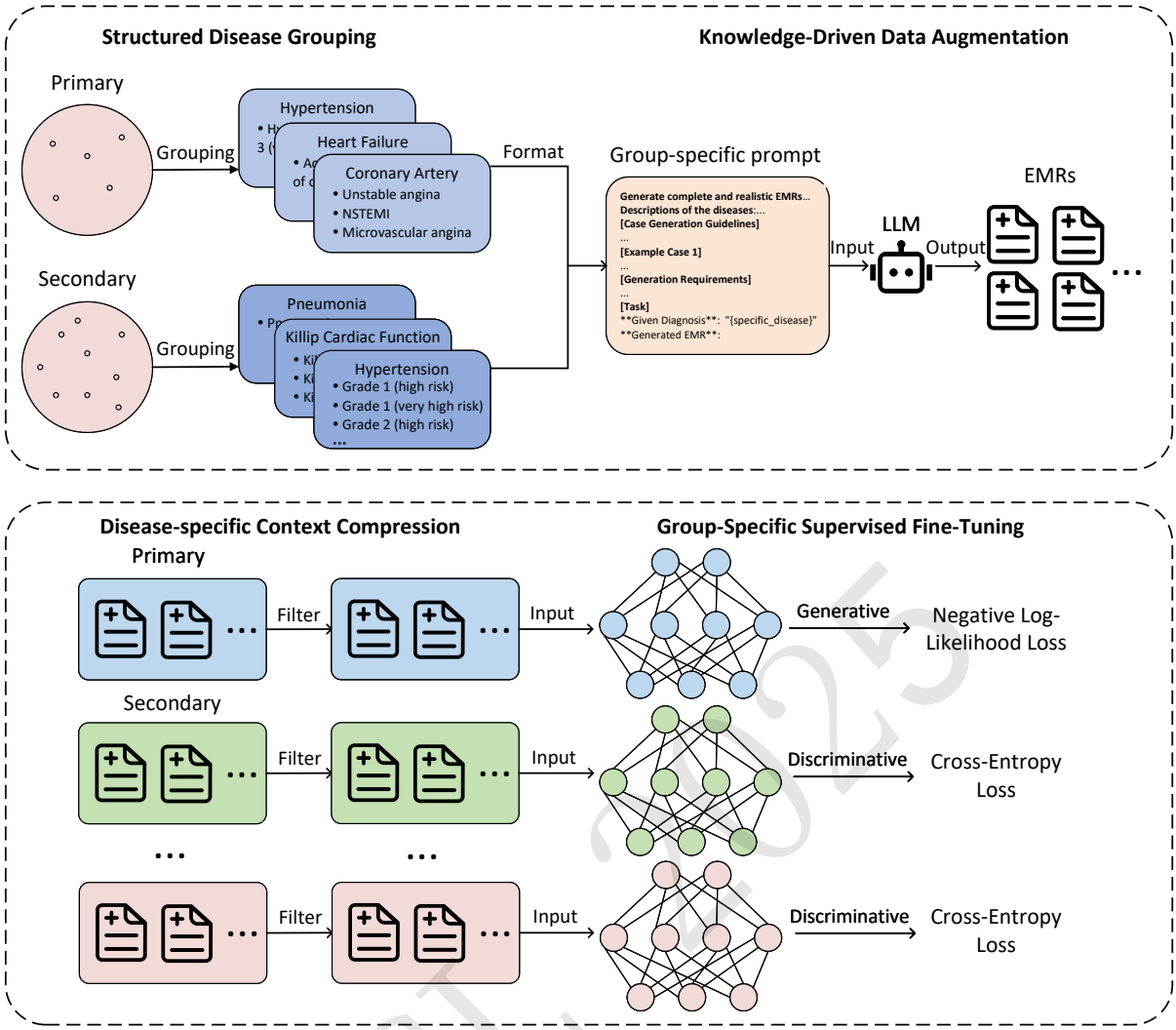

Figure 1: Overall framework of our method. We partition diseases into distinct groups based on clinical knowledge and leverage fine-grained group-level descriptions to generate diverse EMRs, which are then compressed to reduce noise and improve efficiency. Subsequently, we fine-tune group-specific models, employing a generative approach for primary diagnoses and a discriminative strategy for secondary diagnoses.

Group	Diseases	Total
Coronary Artery	Unstable angina, Non-ST-segment elevation myocardial infarction (NSTEMI), Microvascular angina	3
Hypertension	Hypertension Grade 3 (very high risk)	1
Heart Failure	Acute exacerbation of chronic heart failure	1

Table 1: Grouping of primary diseases, including 3 groups.

Group	Diseases	Total
Hypertension	Hypertension Grade 1 (high risk), Hypertension Grade 1 (very high risk), Hypertension Grade 2 (high risk)...	6
Killip Cardiac Function	Killip Class II, Killip Class III, Killip Class IV	3
Complex Premature Beats	Occasional Atrioventricular Premature Contractions, Frequent Premature Contractions	2
Renal Insufficiency	Renal Insufficiency	1
Pneumonia	Pneumonia	1

Table 2: Examples of the grouping of secondary diseases.

For the 53 secondary diseases, we define 12 groups encompassing 30 diseases while treating the remaining 23 diseases as individual groups. Within each group, diseases are mutually exclusive, whereas diseases across different groups are considered compatible. Specifically, mutual exclusivity is defined based on clinical incompatibility—diseases which cannot co-occur. This includes different stages or risk levels of the same condition (e.g., various grades of hypertension or diabetes), or distinct pathological entities affecting the same organ system (e.g., different thyroid disorders or forms of heart failure). Representative examples of the grouping of secondary diseases are shown in Table 2.

Furthermore, we provide a concise description for each disease, highlighting the core distinctions within each group, facilitating fine-grained differentiation and supporting model interpretability.

3.2 Knowledge-Driven Data Augmentation

Due to the limited availability of annotated training data, we adopt a knowledge-driven data augmentation strategy to construct disease-specific synthetic records. Specifically, we design customized prompts for each disease group and utilize a LLM to generate corresponding clinical narratives for target disease. Each generated record comprises six structured fields: *Chief Complaint*, *History of Present Illness*, *Past Medical History*, *Clinical Course*, *Discharge Status*, and *Discharge Instructions*.

Prompt Design As shown in Figure 2, we incorporate comprehensive differential diagnostic descriptions of all diseases within the same disease group into the prompt, including pathology, typical symptoms, auxiliary diagnostic criteria, and treatment protocols. To utilize the reasoning ability of LLM, we explicitly define rules for case generation, which include:

- Disease Classification Logic: Generated cases must adhere to disease-specific diagnostic criteria and exhibit distinguishing features among diseases in the same group.
- Field Consistency Logic: Each of the six fields is required to contain specific, medically coherent content aligned with the clinical description of the disease.

To define the response format and better guide the generation process, we provide a one-shot example of a real clinical case formatted in JSON, covering all six fields. This enables the construction of high-quality, diverse, and medically consistent synthetic records, each tailored to specific disease and enriched with group-level distinguishing characteristics. Specifically, we construct 1,200 samples for each disease via Qwen2.5-7B-Instruct, resulting in a total of 6,000 samples corresponding to primary diseases. For secondary diseases, since not every case includes a confirmed diagnosis for each group, we sample additional records of other diseases to form the training sets.

3.3 Disease-specific Context Compression

To construct the input context for training, we perform disease-specific compression of medical records by retaining only the most informative fields. For **primary diseases**, we concatenate six clinically relevant sections—*Chief Complaint*, *History of Present Illness*, *Past Medical History*, *Clinical Course*, *Discharge Status*, and *Discharge Instructions*—to form the model input. For **secondary diseases**, where

Proceedings of the 24th China National Conference on Computational Linguistics, pages 312-321, Jinan, China, August 11-14, 2025.

Record curation prompt

```
Please generate complete and realistic electronic medical records (EMRs) that
adhere to standard clinical documentation practices.  Each EMR should include the
following six fields...
Given {group_size} diseases:
...
Descriptions of the {group_size} diseases are as follows:
...
# 1. {group[0]}
* Pathology:  Blood pressure overload (Hypertension);
* Symptoms:  Frequent headaches, dizziness...;
* Auxiliary Diagnosis:  Systolic BP ≥ 180 mmHg or Diastolic BP ≥ 110 mmHg...;
* Treatment:  Antihypertensive therapy...

# 2. {group[1]}
...

[Case Generation Guidelines]
1. **Disease Classification Logic**:
   - Clearly reflect the diagnostic criteria for each specified disease...
2. **Field Consistency Logic**:
   - Chief Complaint:  Must reflect core symptoms (e.g...).
   - History of Present Illness:  Provide detailed descriptions expanding on the
chief complaint...
   - Past Medical History:  Include relevant past diseases, prior surgeries...
   - Clinical Course:
     * Include diagnostic findings that confirm the primary diagnosis...
     * Document treatments consistent with clinical guidelines...
   - Discharge Status:  Document improvements in both subjective symptoms and
objective clinical indicators...
   - Discharge Instructions:  Provide follow-up medication plans, lifestyle
recommendations...

[Example Case 1]
**Given Diagnosis**: "Acute Exacerbation of Chronic Heart Failure"
**Generated EMR**:
```

```
{{
  "Chief Complaint": "Recurrent chest tightness for over 10 years...",
  "History of Present Illness": "The patient presented with paroxysma...",
  "Past Medical History": "History of hypertension for over 10 years...",
  "Clinical Course": "Upon admission, comprehensive laboratory...",
  "Discharge Status": "Patient denied chest pain or palpitations...",
  "Discharge Instructions":  "1. Low-salt, low-fat diet, monitor heart rate...",
}}

[Generation Requirements]
Given the **Given Diagnosis**, output a response in JSON format containing the
six fields listed above...

[Task]

**Given Diagnosis**: "{specific_disease}"
**Generated EMR**:
```

Figure 2: The prompt template for generating synthetic EMRs. Each generated record comprises six structured fields: *Chief Complaint*, *History of Present Illness*, *Past Medical History*, *Clinical Course*, *Discharge Status*, and *Discharge Instructions*.

definitive diagnostic evidence may be sparse or inconsistently recorded, we utilize a compressed subset of the context, consisting of *Chief Complaint*, *History of Present Illness*, *Past Medical History*, and *Clinical Course*.

We further tailor the selected sections according to disease-specific characteristics to maximize information relevance and minimize noise. One of the key challenges in long-text understanding lies in handling irrelevant information. To address this, we also employ **Qwen2.5-7B-Instruct** as a feature extractor to identify abnormal clinical indicators across different information points. The model is trained on the same compressed input. Both variants of the context compression strategy are incorporated into our ensemble model library to enrich representation diversity.

3.4 Group-Specific Supervised Fine-Tuning

To leverage the augmented training data effectively, we perform group-specific supervised fine-tuning tailored to the characteristics of primary and secondary diagnoses in ICD coding.

3.4.1 Generative Fine-Tuning for Primary Diseases

For primary diseases, we adopt an instruction-based generative fine-tuning paradigm. Let $\mathbf{x} \in \mathbb{R}^l$ denote the compressed clinical context of a record, and let $\mathbf{c}$ denote the group-specific instruction template. The input to the model is the concatenation:

$$\mathbf{x}_{\text{input}} = \texttt{Concat}(\mathbf{c}, \mathbf{x}), \tag{1}$$

and the label is the corresponding gold disease name $y \in \mathcal{Y}_{\text{prim}}$, where $\mathcal{Y}_{\text{prim}}$ is the set of all primary disease labels. We fine-tune the instruction-following LLM Qwen2.5-7B-Instruct using a negative log-likelihood loss:

$$\mathcal{L}_{\text{gen}} = -\sum_{t=1}^{|y|} \log P(y_t \mid y_{<t}, \mathbf{x}_{\text{input}}), \tag{2}$$

where y_t is the t-th token of the label.

3.4.2 Discriminative Fine-Tuning for Secondary Diseases

In contrast, secondary diseases are handled using a discriminative classification-based paradigm. Given input clinical text $\mathbf{x}$, we first obtain its dense representation using an encoder:

$$\mathbf{h} = \texttt{Encoder}(\mathbf{x}) \in \mathbb{R}^d, \tag{3}$$

where d is the dimensionality of the embedding space. Qwen3-0.6B is used as the encoder backbone for all discriminative fine-tuning tasks.

For each disease group g, we define a classification task with $K_g + 1$ classes, where K_g is the number of diseases in group g, and the additional class represents the absence of any disease from the group. An additional classifier $f_g(\cdot)$ maps the representation to logits, followed by a softmax operation to produce a probability distribution:

$$p_g = \texttt{Softmax}(f_g(\mathbf{h})) \in \mathbb{R}^{K_g+1}. \tag{4}$$

The model is trained using the standard cross-entropy loss:

$$\mathcal{L}_{\text{cls}} = -\log p_g^{(y)}, \tag{5}$$

where $p_g^{(y)}$ denotes the predicted probability corresponding to the ground-truth label y.

3.4.3 Adversarial Training

To improve robustness and generalization, we incorporate adversarial training Liu et al. (2021) and active learning techniques during fine-tuning.

For adversarial training, we perturb the latent representation $\mathbf{h}$ by injecting isotropic Gaussian noise:

$$\tilde{\mathbf{h}} = \mathbf{h} + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(0, \sigma^2 \mathbf{I}), \tag{6}$$

This encourages the model to produce stable predictions under small semantic variations. Additionally, during training, we dynamically identify hard negative samples, those with high predictive uncertainty or elevated training loss, and prioritize them for fine-tuning. This strategy helps the model better discriminate subtle inter-disease differences and handle decision boundary cases. By combining group-specific model architectures and enhancement strategies, our approach enables both precise and robust ICD coding across diverse disease types.

4 Experiments

4.1 Data and Settings

Field	Text
Chief Complaint	Paroxysmal chest tightness and shortness of breath for over six months.
History of Present Illness	The patient experienced chest tightness and gasping following physical activity over six months ago, presenting as a pressure-like sensation...
Past Medical History	History of hypertension for over seven years, with systolic blood pressure peaking at 200 mmHg...
Personal History	Born in the native place, denies long-term residence elsewhere. Denies exposure to epidemic water or endemic areas...
Admission Status	The patient was admitted due to "paroxysmal chest tightness and shortness of breath for over six months." Past history: hypertension for over seven years...
Admission Diagnosis	1. Coronary atherosclerotic heart disease – unstable angina; 2. Hypertension (grade 3, very high risk); 3. Hyperlipidemia; 4. Post-thyroidectomy.
Clinical Course	After admission, relevant laboratory tests were completed. Glycated hemoglobin: 6.90...
Discharge Status	The patient's condition is currently stable, with no episodes of chest pain, no palpitations or chest tightness, and no coughing...
Discharge Instructions	1. Ensure adequate rest, maintain a light diet, engage in appropriate physical activity, avoid overexertion, infections, and emotional agitation...

Table 3: Example of training data.

The gold dataset is divided into training, development, and test sets, containing 600, 200, and 485 samples respectively. Each complete sample consists of 12 structured fields, including *Chief Complaint*, *History of Present Illness*, *Past Medical History*, *Personal History*, *Admission Diagnosis*, *Clinical Course* and so on. A detailed example is provided in Table 3. These fields collectively provide a structured summary of the patient's symptoms, history, diagnosis, and treatment. We employ knowledge-driven data augmentation to expand the dataset, which is subsequently used for supervised fine-tuning of various models.

For the **Qwen2.5-7B-Instruct** model, we perform **LoRA-based fine-tuning** with the following hyperparameters: `lora_rank` = 8, `lora_alpha` = 16, and a learning rate of 1e-5. The training schedule consists of a *linear warm-up* phase followed by *cosine decay*, and the model is fine-tuned for **10 epochs**. For the **Qwen3-0.6B** model, we train **independent models for each disease group**. For each model, we concatenate the outputs of the *last three transformer layers* to construct the feature representation. The classifier consists of a **three-layer linear projection network**, where only the parameters of the linear layers are trainable. Training is conducted with a learning rate of 2e-5 and *weight decay regularization*, over **30 epochs**. The hyperparameters are empirically chosen based on performance on a validation split from our training set.

4.2 Evaluation Metrics

To evaluate the performance of the ICD diagnosis code prediction, it adopts an accuracy metric denoted as Acc, which considers both the correctness of the primary diagnosis and the quality of the secondary

Proceedings of the 24th China National Conference on Computational Linguistics, pages 312-321, Jinan, China, August 11-14, 2025.

diagnoses. The metric is defined as:

$$\text{Acc} = \frac{1}{N}\sum_{i=1}^{N}\{0.5 \cdot \mathbb{I}(y_{\text{main}} = \hat{y}_{\text{main}}) + 0.5 \cdot \text{F1}(y_{\text{other}}, \hat{y}_{\text{other}})\}_i \tag{7}$$

where N denotes the total number of test samples. $\hat{y}_{\text{main}}$ and y_{main} represent the predicted and ground truth primary diagnosis codes, respectively, while $\hat{y}_{\text{other}}$ and y_{other} denote the sets of predicted and ground truth secondary diagnosis codes. $\mathbb{I}(\cdot)$ is the indicator function that returns 1 if the condition is true, and 0 otherwise. The $\text{F1}(\cdot)$ function is defined as the harmonic mean of precision and recall:

$$\text{F1}(y, \hat{y}) = 2 \cdot \frac{\text{Precision}(y, \hat{y}) \cdot \text{Recall}(y, \hat{y})}{\text{Precision}(y, \hat{y}) + \text{Recall}(y, \hat{y})} \tag{8}$$

$$\text{Precision}(y, \hat{y}) = \frac{\text{NUM}(y \cap \hat{y})}{\text{NUM}(\hat{y})} \tag{9}$$

$$\text{Recall}(y, \hat{y}) = \frac{\text{NUM}(y \cap \hat{y})}{\text{NUM}(y)} \tag{10}$$

Here, $\text{NUM}(x)$ denotes the cardinality of set x, i.e., the number of elements it contains. This evaluation framework ensures a balanced assessment by accounting for both exact matches on the primary diagnosis and partial matches across the set of secondary diagnoses.

4.3 Main Results

Method	Acc	Acc(primary)	F1(secondary)
Baseline	41.34	-	-
Ours	86.72	96.18	77.27
w/o data_aug	77.41	86.49	68.32

Table 4: Main results of our method and baseline. w/o data_aug represents the SFT manner without data augmentation.

We evaluate the performance of our proposed method on the ICD coding test set and compare it with a baseline model. The results are summarized in Table 4.

Our full method achieves a substantial improvement over the baseline, with an overall accuracy (**Acc**) of 86.72%, compared to 41.34% for the baseline. The primary diagnosis accuracy reaches 96.18%, demonstrating the effectiveness of our grouping-based, data-augmented instruction tuning pipeline. For secondary diagnoses, our method achieves an F1 score of 77.27%, indicating moderate success in handling the complexity and long-tail distribution inherent in these labels.

To further assess the impact of data augmentation, we conduct an ablation study by removing the augmentation strategy from our fine-tuning process (denoted as *w/o data_aug*). The overall accuracy drops to 77.41%, with the primary diagnosis accuracy and secondary diagnosis F1 falling to 86.49% and 68.32%, respectively. These results highlight the critical role of data augmentation in enhancing model generalization and robustness, especially for underrepresented or complex diagnosis cases.

Overall, the results validate the effectiveness of our approach and underline the importance of incorporating data augmentation and task decomposition strategies in supervised fine-tuning for ICD coding.

5 Conclusion

In this paper, we propose a structured framework for ICD coding based on clinically informed disease grouping, LLM-enhanced data augmentation, and group-specific model optimization. By partitioning

diseases into mutually exclusive groups, the framework ensures clinically coherent co-occurrence patterns. Building upon this structure, we utilize fine-grained diagnostic descriptions and inter-disease discriminative information to guide data augmentation, thereby improving both the quantity and diversity of training data. To further enhance model performance, we apply input compression to reduce noise and improve learning efficiency. We then perform group-specific supervised fine-tuning, employing a generative approach for primary diagnoses and a discriminative strategy for secondary diagnoses, improving both robustness and generalization. Experimental results on the CCL2025-Eval Task 8 benchmark demonstrate substantial gains in primary diagnosis accuracy and secondary diagnosis F1 score, confirming the effectiveness of our approach.

Future work may focus on addressing the challenges posed by the large, long-tailed ICD label space and the heterogeneous nature of clinical text. Incorporating strategies such as rule-based methods and structured information extraction may further enhance the performance and interpretability of automatic ICD coding systems.

6 Acknowledgement

We sincerely thank the organizers of the shared task and reviewers for their helpful feedback. This work was supported by the Natural Science Foundation of Beijing No.L232119.

References

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

Richárd Farkas and György Szarvas. Automatic construction of rule-based icd-9-cm coding systems. In *BMC bioinformatics*, volume 9, pages 1–9. Springer, 2008.

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

Mengxing Huang, Huirui Han, Hao Wang, Lefei Li, Yu Zhang, and Uzair Aslam Bhatti. A clinical decision support framework for heterogeneous data sources. *IEEE journal of biomedical and health informatics*, 22(6):1824–1833, 2018.

Bevan Koopman, Sarvnaz Karimi, Anthony Nguyen, Rhydwyn McGuire, David Muscatello, Madonna Kemp, Donna Truran, Ming Zhang, and Sarah Thackway. Automatic classification of diseases from free-text death certificates for real-time surveillance. *BMC medical informatics and decision making*, 15:1–10, 2015.

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

Xiao Liu, Ge Shi, Bo Wang, Changsen Yuan, Heyan Huang, Chong Feng, and Lifang Wu. Bit-event at nlpcc-2021 task 3: subevent identification via adversarial training. In *CCF International Conference on Natural Language Processing and Chinese Computing*, pages 400–411. Springer, 2021.

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

Patrick Ruch, Julien Gobeill, Imad Tbahriti, and Antoine Geissbuehler. From episodes of care to diagnosis codes: automatic text categorization for medico-economic encoding. In *AMIA Annual Symposium Proceedings*, volume 2008, page 636, 2008.

Henning Schäfer and Christoph M Friedrich. Umls mapping and word embeddings for icd code assignment using the mimic-iii intensive care database. In *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 6089–6092. IEEE, 2019.

Jurica Seva, Madeleine Kittner, Roland Roller, and Ulf Leser. Multi-lingual icd-10 coding using a hybrid rule-based and supervised classification approach at clef ehealth 2017. In *CLEF (Working Notes)*, 2017.

Fei Teng, Yiming Liu, Tianrui Li, Yi Zhang, Shuangqing Li, and Yue Zhao. A review on deep neural networks for icd coding. *IEEE Transactions on Knowledge and Data Engineering*, 35(5):4357–4375, 2022.

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

Bo Wang, Yi-Fan Lu, Xiaochi Wei, Xiao Liu, Ge Shi, Changsen Yuan, Heyan Huang, Chong Feng, and Xianling Mao. Bit-wow at nlpcc-2022 task5 track1: Hierarchical multi-label classification via label-aware graph convolutional network. In *CCF International Conference on Natural Language Processing and Chinese Computing*, pages 192–203. Springer, 2022.

Keyang Xu, Mike Lam, Jingzhi Pang, Xin Gao, Charlotte Band, Piyush Mathur, Frank Papay, Ashish K Khanna, Jacek B Cywinski, Kamal Maheshwari, et al. Multimodal machine learning for automated icd coding. In *Machine learning for healthcare conference*, pages 197–215. PMLR, 2019.

Lingling Zhou, Cheng Cheng, Dong Ou, and Hao Huang. Construction of a semi-automatic icd-10 coding system. *BMC medical informatics and decision making*, 20:1–12, 2020.