Journal of Artificial Intelligence in Medical Applications (JAIMA)
Vol. 1, No. 1, 2025, **pp**. 76–88
ISSN: 3106-1494 | **DOI**: https://doi.org/10.65511/jaima.v1i1.735

Enhancing Medical Department Prediction in Chinese Clinical Records via BERT Fine-Tuning and Rare-Class Augmentation

Mohammed Salah Ibrahim[1, *], Mohammed Al-Jabbar[2], Rida E Moustafa[3]

[1]Department of Artificial Intelligence, College of Computer Science and Information Technology, University of Anbar, Ramadi, Iraq; moh.salah@uoanbar.edu.iq
[2]Computer Department, Applied College, Najran University, Najran, Saudi Arabia; mosalqahtani@nu.edu.sa
[3]dMining Technology, LLC, Bentonville, Arkansas, USA; moustafa@dMining-technology.com

Received: 19/08/2025, Revised: 17/10/2025, Accepted: 27/10/2025, Published: 30/12/2025

ABSTRACT: Accurately predicting the appropriate referral department from clinical records is essential for improving triage efficiency, resource allocation, and overall hospital workflow. In this study, we present an end-to-end pipeline for department prediction using unstructured Chinese electronic health records (EHRs) from the publicly available MedCD dataset. The dataset contains detailed admission notes written in natural Chinese, which often introduce challenges such as class imbalance and linguistic variability. To address these issues, we applied a series of preprocessing steps, including the removal of incomplete records, the consolidation of closely related departments (e.g., orthopedic sub-units), and the exclusion of labels with very low sample frequency. To further mitigate class imbalance, we used nlpcda, a Chinese-specific data augmentation toolkit, which generated additional samples for underrepresented classes through synonym substitution and textual perturbation. The augmented dataset was then balanced using class-weighted loss functions. For classification, we fine-tuned the Chinese-BERT-wwm-ext model on the processed dataset. Evaluation across multiple runs demonstrated clear performance gains, particularly in macro and weighted F1-scores. Overall accuracy reached 91%, while the macro F1-score improved from 0.66 (baseline) to 0.89 after augmentation. Departments with historically limited data—such as Radiology, Emergency Medicine, and Pain Management—showed marked improvement. In addition, we reported Top-3 prediction accuracy to better reflect real-world triage scenarios, where the correct department may reasonably be among the top three recommendations. These findings highlight the value of combining language-specific data augmentation with fine-tuned deep learning models for clinical text classification, especially in low-resource healthcare settings.

Keywords: Electronic Health Records, Chinese BERT, Transformers, Chinese Clinical Text, Chinese BERT with Whole Word Masking, NLP Chinese Data Augmentation, Clinical Department prediction.

1. INTRODUCTION

Efficient triage and patient referral in hospitals depend on accurately identifying the appropriate department for each patient. Traditionally, this decision has relied on structured clinical data and physician expertise. Recent developments in artificial intelligence (AI) techniques, particularly in the area of natural language processing (NLP), have created new avenues for improving and automating referral procedures through the use of unstructured clinical narratives. This is especially relevant for Chinese-language admission notes, which present distinctive linguistic and structural challenges [1]. In the broader clinical NLP domain, transformer-based models have shown strong potential in extracting meaningful information and predicting patient outcomes. For example, ClinicalBERT has achieved notable success in modeling English-language notes for hospital readmission prediction tasks [2], [3]. Within the Chinese medical context, earlier work has primarily focused on entity recognition and classification. A deep learning pipeline for extracting radiological features from free-text reports, for instance, reported strong F1 scores, laying an important foundation for subsequent research [4].

Beyond entity-level tasks, researchers have begun to address classification challenges in Chinese clinical text. A dual-channel attention-based model demonstrated that incorporating both semantic and temporal features can enhance department recommendation accuracy [5]. At the same time, domain-specific data augmentation methods such as the NLP Chinese Data Augmentation (nlpcda) have been developed to enrich Chinese medical text corpora through synonym substitution and paraphrasing—strategies that are particularly valuable for addressing rare classes in imbalanced datasets [6]. Despite these advances, the specific problem of automated department referral prediction using fine-tuned BERT models on Chinese clinical records has received limited attention. Performance remains especially weak for less-represented departments due to data scarcity. To address this gap, our study integrates Chinese-specific data augmentation techniques (nlpcda) [7] with class-balancing strategies and a fine-tuned Chinese BERT with Whole Word Masking , which publicly known as Chinese-BERT-wwm-ext model [8]. This approach yields

substantial improvements in classification performance across both common and rare referral departments. Accordingly, our study is driven by the following research questions:

- Q1: To what extent can a fine-tuned BERT model accurately predict referral departments from unstructured Chinese admission records?
- Q2: How does the use of domain-specific data augmentation influence classification performance, particularly for rare departments?
- Q3: Can combining fine-tuned BERT with class-weighted learning and Top-3 prediction strategies provide practical support for real-world triage systems?

The aim of this study is to develop and evaluate an end-to-end pipeline for automated department referral prediction from Chinese admission records. Specifically, the study seeks to enhance classification accuracy across both common and rare departments by fine-tuning a Chinese BERT model, incorporating domain-specific data augmentation, and applying class-weighted learning strategies. In addition, the study aims to validate the real-world utility of the proposed approach through Top-3 prediction analysis, reflecting practical triage scenarios in hospital settings. The key contributions of this study are threefold:

1. Application of fine-tuned Chinese BERT: We demonstrate the effectiveness of fine-tuning Chinese-BERT-wwm-ext for department classification using the MedCD dataset, a large corpus of unstructured Chinese admission notes.

2. Improved rare-class prediction: By integrating domain-specific data augmentation (*nlpcda*) and class-weighted learning, we substantially enhance model performance for underrepresented departments.

3. Practical validation: Our pipeline achieves high predictive performance (91% accuracy, macro F1-score of 0.89) and demonstrates real-world utility through Top-3 predictions, simulating triage decision support in hospital workflows.

The remainder of this study is as follows. A study history and a list of earlier research are provided in the second part. Next, section 3 reports a thorough explanation of the methodology and procedures employed in our study, along with some specifics on the dataset and evaluation measures. The outcomes of the model's application and the findings of other investigations are then discussed, and results and limitations are provided. The conclusion and future work are finally discussed.

2. RELATED WORKS

The task of predicting medical departments from clinical texts has gained increasing attention as hospitals adopt electronic health records and seek to automate triage and referral processes. In Chinese clinical practice, however, this task presents distinctive challenges: the complexity of medical language, the absence of word segmentation in Chinese text, and the strong imbalance across department labels. Transformer-based models have become the leading approach for clinical text classification. Among them, Chinese-BERT-wwm-ext, introduced by Cui et al., has achieved strong results on biomedical NLP tasks by applying whole-word masking to preserve semantic integrity [8]. This model has since served as a backbone for downstream applications such as disease classification, report summarization, and question answering. For instance, Yao et al. fine-tuned Chinese-BERT on Traditional Chinese Medicine (TCM) records, reporting 89.4% accuracy and a macro-F1 of 88.6% in multiclass diagnosis prediction [9]. These findings demonstrate the ability of transformer models to handle the linguistic complexity and contextual nuances of Chinese clinical documentation.

Several efforts have specifically targeted hospital department recommendation using BERT-based methods. Wang et al. developed a cloud-based intelligent self-diagnosis and department recommendation system (CIDRS), powered by CHMBERT—a Chinese medical BERT trained on a large-scale corpus. Deployed in a containerized cloud environment, CIDRS achieved robust accuracy in real-world outpatient registration, offering practical value for online department selection. However, this work did not explicitly address issues of class imbalance or the poor performance of classifiers on rare departments [10]. To alleviate imbalance-related performance degradation, researchers have proposed data augmentation techniques. Chen and Du introduced the Self-Attentive Adversarial Augmentation Network (SAAN) in combination with a multi-task BERT model, achieving notable improvements in F1 and ROC-AUC on datasets such as CCKS 2017 [11]. Similarly, Chen et al. proposed Controlled Random Replacement (CRR) and Targeted Entity Replacement (TER) to synthetically expand underrepresented classes in Chinese Named Entity Recognition (NER), raising F1 to 83.6% [12]. While effective, these methods often require significant computational resources and extensive domain-specific annotations.

Lightweight augmentation approaches, such as nlpcda, provide a more practical alternative for low-resource domains. Designed specifically for Chinese text, nlpcda applies synonym replacement, random word swaps, and insertions to improve coverage of minority classes at minimal cost [7]. Although widely used in general text classification, its integration with fine-tuned BERT models for department classification in clinical settings has not yet been investigated—a gap addressed in this work. Parallel to augmentation, representation learning continues to evolve. Zhang et al. introduced ChineseBLUE, a domain-adapted model trained on biomedical corpora, which outperformed BERT and RoBERTa on benchmark tasks [13]. Chen et al. further advanced medical entity recognition through a hybrid BERT-BiLSTM-CNN-CRF model, achieving 93–94% F1 across multiple datasets [14]. More recently, large language models (LLMs) such as ChatGPT have been explored for clinical text augmentation. Yuan et al. showed that LLM-generated paraphrases can enhance classification and matching accuracy through contrastive learning, even in low-resource medical contexts [15].

Despite these advances, no prior work has combined synonym-based data augmentation (e.g., nlpcda) with fine-tuned Chinese-BERT for department-level classification of admission records. Furthermore, evaluation using Top-3 prediction—a format that closely reflects real-world triage decision-making—remains underexplored. Our study addresses these gaps by integrating augmentation and class-weighting strategies into a fine-tuned BERT pipeline, resulting in macro-F1 improvements from 0.66 to 0.89 and substantial gains in rare-class performance.

3. THE PROPOSED METHODOLOGY

In our approach, unstructured clinical text is first converted into a tokenized sequence, where each word or character is mapped to an input embedding alongside positional and segment information. These embeddings are then passed into a BERT encoder, which serves as the central representation learner. By processing the entire sequence through multiple self-attention layers, BERT captures both local context and long-range dependencies, producing contextualized embeddings for each token. Among these outputs, the embedding associated with the special classification token [CLS] is treated as the global representation of the input record. This representation is then fed into a fully connected classification layer, followed by a SoftMax function to generate probability distributions across candidate medical departments. During training, the model is optimized using class-weighted loss functions to address imbalances between frequent and rare departments. By leveraging BERT's deep contextual encoding in combination with downstream classification, the methodology enables robust prediction of patient referral departments, ensuring that subtle linguistic cues in clinical notes are effectively captured and translated into accurate triage recommendations. Figure 1 shows the general architecture of the proposed model.

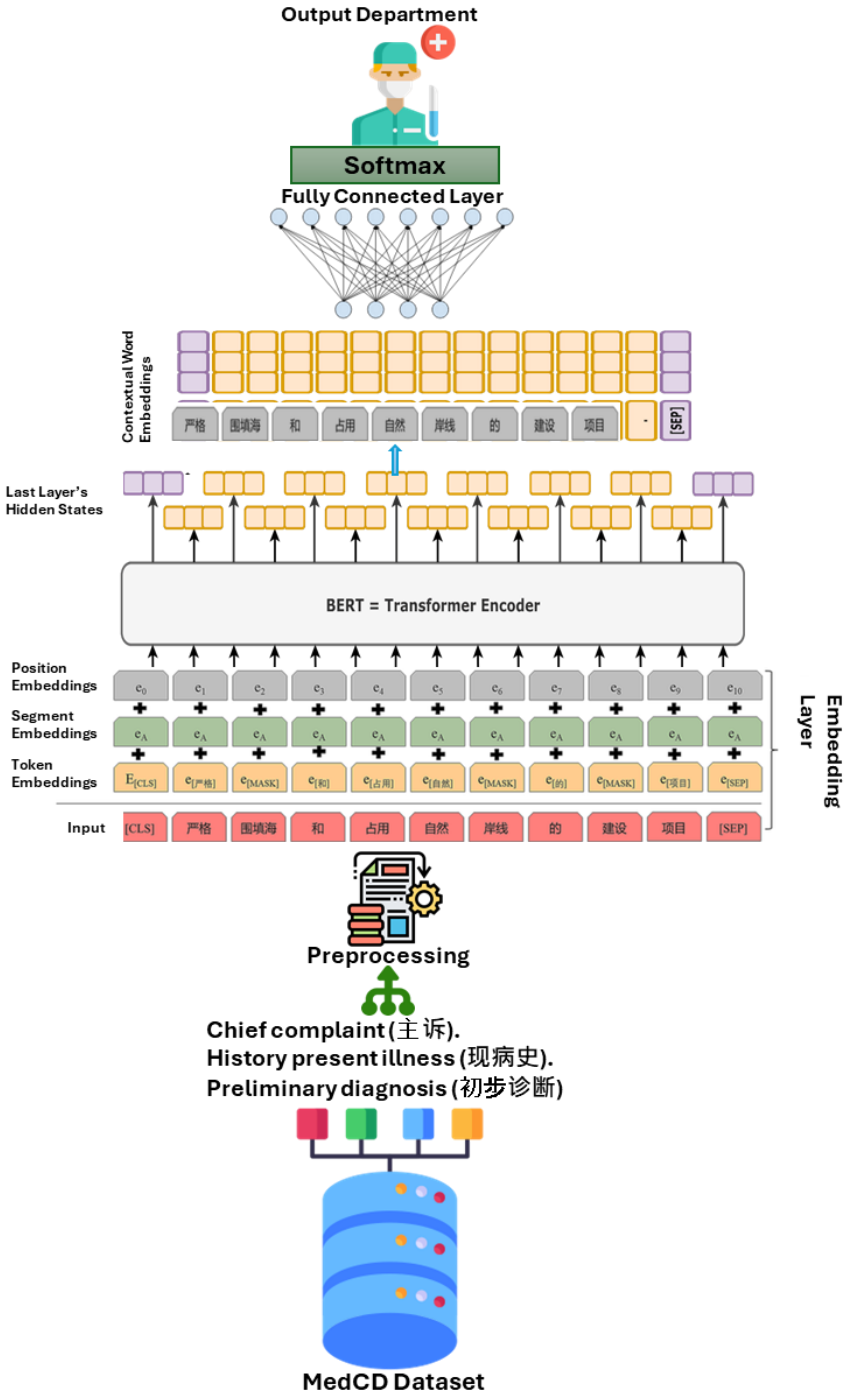

Figure 1. The Proposed Methodology of our proposed system

3.1 Data Source

This study makes use of the publicly available MedCD dataset [16], which contains anonymized Chinese clinical notes spanning different stages of hospital care, including admission records, physician progress notes, and discharge summaries. For the purposes of department prediction, we focus on the admission records, as they provide the richest information for triage: detailed patient complaints, the history of present illness, and the physician's preliminary diagnosis. Each admission record contains more than 30 fields, from which we select a subset most relevant to referral decisions. The ground-truth label for each entry is the department to which the patient was directed, making the dataset well suited for supervised classification tasks.

Specifically, three key text fields are concatenated to form the model input. The chief complaint (主诉) captures the primary symptom or reason for seeking medical care, often recorded in the patient's own words (e.g., "chest pain for two hours") and serving as an immediate cue for triage. The history of present

illness (现病史) provides a narrative of symptom progression, severity, and associated conditions, offering the clinical context necessary to distinguish between otherwise similar complaints (e.g., chest pain with cough and fever suggesting Respiratory Medicine versus chest pain radiating to the arm suggesting Cardiology). The preliminary diagnosis (**初步**诊断) reflects the physician's initial impression, often naming suspected diseases directly and helping to resolve ambiguous cases such as non-specific "weakness." Together, these fields provide a comprehensive input for BERT-based models, which benefit from both the brevity of key symptoms and the depth of contextual narratives. The target output for the task is the "department" field, with Table 1 listing examples of departments along with their English translations.

Table 1 : examples of departments along with their English translations.

Chinese Department Name	English Translation
呼吸与危重症医学科	Respiratory and Critical Care Medicine
消化内科	Gastroenterology
心血管内科	Cardiology (Cardiovascular Medicine)
神经内科	Neurology
神经外科	Neurosurgery
妇科	Gynecology
产科	Obstetrics

3.2 The Preprocessing stage

To prepare the dataset for modeling, we first carried out a series of preprocessing steps aimed at ensuring data quality and consistency. Algorithm 1 shows the main preprocessing steps achieved. Records containing empty or null values in any of the selected text fields or department labels were removed, as incomplete entries can introduce noise into training. We also addressed the issue of extremely small classes: departments represented by fewer than 20 samples were excluded, since such limited data would not provide enough signal for reliable learning or evaluation. This filtering step helped maintain a more balanced dataset while reducing the risk of overfitting to rare, underrepresented categories.

Next, we applied department merging to further minimize label noise. Closely related sub-departments, for example, **骨科(脊柱外科病区)** and **骨科**(创伤骨科病区)—were consolidated under their parent department (**骨科**), which allowed us to simplify the label space and avoid unnecessary fragmentation. After that, the key textual fields—chief complaint, history of present illness, and preliminary diagnosis—were concatenated into a single input string for each record, providing the model with a comprehensive representation of the patient's case. Finally, we applied label encoding to the department field, converting categorical department names into numerical identifiers suitable for classification.

Algorithm 1. Steps of our data preprocessing and data augmentation

Input: MedCD admission records referred to as *A*-
Output: Training data *Atr*, validation data *Aval*
1. combine text
2. $x \leftarrow \text{combine}(A_{chief\_complaint}, A_{history\_of\_present\_illness}, A_{history\_of\_present\_illness}, A_{preliminary\_diagnosis}$
3. $y \leftarrow A_{department}$
4. Remove any record does have x or y
5. Merge subdepartments into one:
6. **e.g. {骨科(创伤骨科病区), 骨科(脊柱外科病区), 运动与关节一/二病区} → 骨科**
7. Remove classes with minor representation
8. remove any class <20 instances
9. Encode departments $y \in \{0,\ldots,|\mathcal{C}|-1\}$
10. Rare class Augmentation:
11. $R \leftarrow \{c \in \mathcal{C}: \text{c count} < 100\}$
12. for each c in R:
13. for each x, y in c:
14. $\{\tilde{x}_1,\tilde{x}_2\ldots,\tilde{x}_k\} \leftarrow \text{Aug}(x; k=3, \rho=0.3)$
15. incorporate $\{(\tilde{x}_j, c)\}$ to *A*.
16. Split 20% for validation and *80%* for training.
17. Tokenization:
18. *Atr*←{(t(x),y)} from training set, *Aval* ←{(t(x),y)} from validation set.

3.3 Data Augmentation and Class Balancing

To mitigate the issue of class imbalance, we employed the NLP Chinese Data Augmentation (*nlpcda)* toolkit [7], which is designed for Chinese text and applies techniques such as synonym replacement, word swapping, and controlled noise injection. Step 10 in Algorithm 1 reports how we accomplished augmentation. Departments with fewer than 100 samples were identified as rare classes, and each record within these categories was augmented threefold, effectively tripling the training data available for underrepresented labels. In addition to augmentation, we incorporated a class-weighted loss function to further stabilize training. Here, class weights were computed inversely proportional to label frequencies, ensuring that minority classes contributed more strongly to the optimization process. This combined strategy allowed the model to achieve better balance across both frequent and rare departments.

3.4 Model Architecture: Chinese-BERT-wwm-ext

The backbone of our approach is the pre-trained Chinese BERT whole world masking, Chinese-BERT-wwm-ext [8], a variant of BERT specifically optimized for Chinese text through whole-word masking during pretraining. At the input layer, clinical records are tokenized using the standard BERT tokenizer, which segments Chinese sequences into subword units suitable for embedding. These token embeddings are then passed through a 12-layer Transformer encoder, each layer comprising 768 hidden units and 12 self-attention heads, allowing the model to capture both local and global contextual dependencies within the text. Algorithm 2 shows the main steps of this model implementation. The input to the model is the training *Atr*, validation data *Aval*, and department class $\mathcal{C}$ collected in the previous step. The output is a pretrained model f_θ tranined on Chinses embeddings. We attach a lightweight supervised head to the base Chinese BERT WWM model to initialize the multi-class classifier. Each tokenized Note (*maxlength = 256) is processed and passed to the model as (input* ids, attention_mask) pairs and hidden vectors are obtained as outputs. From the hidden states, we select the representation of the first token, the noisy speech-based token, H and classify it sequence summary. The first token is subjected to dropout for regularization (p=0.1) and is then sent to a linear projection layer to obtain output class logits $z \in R|C|$ corresponding to every class $Z \in R|C|$ (one for each department). Note that WWM was only a pretraining mask strategy technique and shaped the parameters of the model; during the fine-tuning stage, the model is only given the ids and attention masks associated with the WordPiece tokens; no masking is applied. Each mini-batch appended with class weighted logits corresponding to L_CE (z,y;w) is used to compute weighted cross entropy class L_CE to ensure that minority classes are optimally weighted w to mitigate imbalance. The model $f\theta$, comprising the backbone with dropout, and a linear head, is ready to undergo end to end optimization with Dtr and is subjected to the $Dval$ dataset for evaluation.

The contextualized representation of the sequence is summarized through the [CLS] token, which serves as the pooled embedding. This pooled output is then fed into a fully connected layer that projects the representation onto the space of department classes, producing logits for classification. Formally, the process can be expressed as:

Given a tokenized input sequence $x = (x_1, x_2, \ldots, x_n)$, the model computes embeddings (See Eq. (1):

$$h_0 = Embeddings(x) \quad (1)$$

Then, 12-layer Transformer encoder is applied (Eq (2)):

$$h_i = TransformerLayeri(h_{i-1}), i = 1,\ldots,12 \quad (2)$$

For the classification head, the SoftMax function was used (Eq. (3)):

$$\hat{y} = softmax(W \cdot h + b) \quad (3)$$

Where W is the weight matrix, and b is the bias vector. The loss function is computed using Weighted Cross-Entropy Loss as reported in Eq. @@.

$$\iota = -\sum_{c=1}^{C} w_c \, . y_c \, . \log(\hat{y}_c) \quad (4)$$

Where C is the number of department classes and y_c is the ground-truth label (one-hot encoded). The $\hat{y}_c$ is the predicted probability and w_c is the class weight for class c.

Algorithm 2. Chinese BERT with Whole Masking steps

Input : *Atr*, *Aval*, number of classes |$\mathcal{C}$|, class weights w (balanced), max_length=256
Output: Initialized classifier fθ

1 Load pretrained backbone:
2 f_backbone ← Chinese-BERT-WWM-Ext (hfl/chinese-bert-wwm-ext)
3 // Note: WWM was used in pretraining only; fine-tuning uses the frozen tokenizer outputs.

4 Build classification head:
5 Dropout p_drop = 0.1

 ISSN: 3106-1494 | **DOI**: https://doi.org/10.65511/jaima.v1i1.735

```
6     Linear head: z = W · h_CLS + b, where h_CLS is the [CLS] embedding (hidden size H)
7     f_θ(x) := z  // logits in ℝ^{|𝒞|}

8  Batching & inputs:
9     For a batch B = {(input_ids, attention_mask, y)}:
10        Pass through backbone to get last_hidden_state
11        h_CLS ← last_hidden_state[:, 0, :]       // first token is [CLS]
12        ĥ ← Dropout(h_CLS, p_drop)
13        z ← W · ĥ + b                          // logits

14 Loss:
15    L_CE(z, y; w) = CrossEntropy(z, y; class_weight = w)
16    // w compensates label imbalance

17 Return f_θ
```

After having the pretrained embeddings ready from Chinese-BERT-wwm-ext, we fine-tuned it to be convenient for our work. Algorithm 3 shows the main steps of fine-tuning process. The $f\theta$ for $E = 5$ epochs with batches of size $B = 16$ and AdamW (lr 2 × 2e-5, weight decay = 0.01, no decay on bias and LayerNorm parameters). The learning rate has a linear decay with warm-up (e.g. warm-up ratio 0.06). At the end of every epoch, we assess the model on $Dval$, accumulating logits for macro-F1 (and loss) estimation. The model checkpoints with highest macro-F1 on validation (tie-break with validation loss) is the only one we save, the model is guaranteed to perform well across all classes.

Algorithm 3. Fine-tuning processing

```
Input : f_θ, Atr, Aval, epochs E=5, batch size B=16, lr=2e-5, weight_decay=0.01
Output: Trained parameters θ*

1  Optimizer:
2     AdamW with (lr, weight_decay); exclude bias/LayerNorm from weight decay
3  LR schedule:
4     Linear decay with warmup ratio r_warmup (e.g., 0.06) or warmup_steps
5  Mixed precision:
6     If GPU supports FP16 → train with autocast (fp16) and gradient scaling
7  Gradient clipping:
8     Clip global norm at 1.0

9  Best checkpoint tracking:
10    metric M := macro-F1 on 𝒟val (or val loss)
11    θ* ← current θ; M* ← -∞

12 For epoch = 1 … E:
13    // ----- Train -----
14    For each batch B in 𝒟tr:
15       z ← f_θ(B)                  // forward pass
16       L ← L_CE(z, y; w)           // weighted CE
17       Backpropagate L
18       Clip gradients; optimizer step; scheduler step; zero grads

19    // ---- Validate (no grad)----
20    Aggregate logits Z_val and labels y_val over 𝒟val
21    Compute macro-F1_val (and loss_val)
22    If macro-F1_val > M*:
23       M* ← macro-F1_val; θ* ← θ      // save best
24 Return θ*
```

3.6 Evaluation Metrics

To assess model performance, we employed a combination of standard and task-specific evaluation metrics. Overall accuracy was reported as a measure of the proportion of correctly classified records. In

addition, we calculated precision, recall, and F1-score at both the per-class and aggregate levels to provide a balanced view of performance across frequent and rare departments. Both macro-averaged and weighted variants of these metrics were considered: macro metrics treat all classes equally, while weighted metrics adjust for class frequency. To further capture clinical applicability, we included Top-3 accuracy, which evaluates whether the correct department appears among the three most likely predictions, a scenario closer to real-world triage support. Finally, confusion matrices were used to visualize class-level misclassifications, with particular attention to rare departments.

1. **Accuracy**

$$Accuracy = \frac{Total\ Number\ of\ Predictions}{Number\ of\ Correct\ Predictions} \quad (5)$$

2. **Precision, Recall, and F1-score**

$$Precision = \frac{TP}{TP+FP} \quad (6)$$

$$Recall = \frac{TP}{TP+FN} \quad (7)$$

$$F1 - score = \frac{2(Precision\ .Recall)}{Precision+Recall} \quad (8)$$

where FP and FN are false positives and false negatives for class.

3. **Macro-Averaged Metrics**

$$Macro - Precision = \frac{1}{C}\sum_{i=1}^{C} Precision_i \quad (9)$$

$$Macro - Recall = \frac{1}{C}\sum_{i=1}^{C} Recall_i \quad (10)$$

$$Macro - F1 - score = \frac{1}{C}\sum_{i=1}^{C} F1_i \quad (11)$$

C is the total number of classes, and i indexes each class.

4. **Weighted F1-score**

$$Weighted\ - F1 - score = \frac{1}{C}\sum_{i=1}^{C} \frac{n_i}{N}.F1_i \quad (12)$$

where $\boldsymbol{n_i}$ is the number of samples in class i.

5. **Top-3 Accuracy**

$$Top - 3\ Accuracy = \frac{1}{C}\sum_{j=1}^{C} 1\{y_j \in Top - 3(\acute{y}_j)\} \quad (13)$$

where y_j is the true label of sample j, $\acute{y}_j$ is the vector of predicted probabilities, and $1\{\cdot\}$is the indicator function.

6. **Confusion Matrix**

$M_{i,j}$=Number of samples with true class i predicted as class j.

- Diagonal entries $M_{i,j}$ represent correct predictions.
- Off-diagonal entries highlight misclassifications between classes.

4. RESULTS AND DISCUSSION

 ISSN: 3106-1494 | **DOI**: https://doi.org/10.65511/jaima.v1i1.735

The results presented in this section provide an in-depth evaluation of the proposed department prediction framework, highlighting its performance across both common and rare clinical categories. We begin by reporting the overall classification metrics, including accuracy, precision, recall, and F1-scores, followed by an analysis of the Top-3 prediction outcomes to better reflect real-world triage scenarios. To gain further insights, confusion matrices are examined to illustrate class-level strengths and weaknesses, with particular emphasis on rare departments that often pose challenges in automated prediction tasks. Finally, the findings are discussed in relation to prior work, emphasizing both the improvements achieved through data augmentation and class-weighted learning, as well as the practical implications of deploying such a system in hospital settings. Our model was run on Kaggle environment. Table 2 shows the training configuration of the applied model.

Table 2. Training configuration

Setting	Type
Pretrained model:	Chinese-bert-wwm-ext
Max token length	256
Optimizer	AdamW with learning rate $2 \times 10 - 5$
Batch size	16
Epochs	5
Hardware	Trained on Kaggle GPU environments with mixed precision (fp16)

To evaluate the proposed framework, we adopted a 5-fold stratified cross-validation strategy, which allowed us to assess performance consistently across all department categories. Three experimental settings were considered. In the first, referred to as the Baseline model, predictions were carried out across all 33 original departments without modification. The second setting, termed Merged–Filtered Classes, involved merging closely related sub-departments and excluding categories with very limited data, thereby reducing label sparsity and noise. In the third experiment, named Augmented–Balanced Classes, departments with fewer than 100 samples were expanded through data augmentation using the *nlpcda*, and class weights were applied to further balance training. The comparative outcomes of these three experiments are summarized in Table 3.

Table 3. Measurements of the different methods applied

Model	Accuracy	Macro F1	Weighted F1
Baseline-AllClas	0.83	0.66	0.83
Merge-Filtered-Clas	0.89	0.77	0.89
Augmented-Balanced	0.91	0.89	0.91

Table 5 illustrates a steady improvement in performance across the three experimental settings. The baseline model, trained on all 33 original departments, achieved reasonable accuracy and weighted F1, yet struggled on macro-F1 due to the sharp imbalance between frequent and rare classes. This discrepancy highlighted how low-frequency categories disproportionately reduced performance despite strong results in well-represented departments. When semantically similar departments were merged and extremely rare ones (fewer than 20 samples) were removed, both accuracy and macro-F1 improved notably. This outcome suggests that simplifying the label space reduced noise and allowed the model to generalize more effectively across the remaining classes.

The best results were obtained in the augmented–balanced setting, where synthetic samples generated with *nlpcda* were used to expand rare categories, and class-weighted training further corrected imbalances. In this case, macro-F1 nearly matched overall accuracy, indicating that rare departments benefited substantially from augmentation. Due to some issues in displaying Chinese letters in many of our experiment figures, we decided to map department names from their original name in Chinese into their match in English by translating them using Google Translator. This mapping was only done for displaying statistics and not during BERT and *nlpcda* process. Figure 2 highlights the performance of the model after applying the *nlpcda*. We can see from this figure how the underrepresented classes such as Emergency Medicine, Pain Management, and Radiology—previously weak performers—showed some of the largest relative gains. Meanwhile, data-rich departments like Cardiology and General Surgery, which were already performing strongly, experienced only modest improvements. This pattern underscores the effectiveness of targeted augmentation in addressing data scarcity, particularly when class imbalance is the dominant limitation.

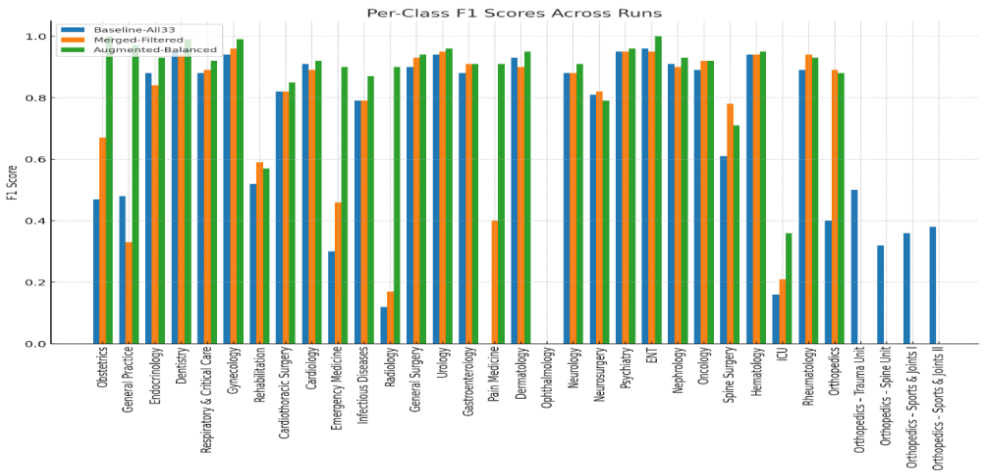

Figure 2. Model performance over different departments and their F1 scores

Figure 3 presents the per-class F1 scores across the three experimental runs, illustrating how performance distribution became progressively more uniform with each intervention. In the baseline run, there was a pronounced gap between the highest- and lowest-performing departments, reflecting both class imbalance and variability in the underlying clinical text. The merged–filtered experiment narrowed this gap by removing extremely rare classes and consolidating overlapping labels, though certain departments still lagged behind. The augmented–balanced setting produced the most consistent results: most departments converged toward high F1 scores (≥0.85), and the overall spread across classes was markedly reduced. This outcome demonstrates that combining class merging with rare-class augmentation and class balancing not only improved overall performance but also enhanced fairness by minimizing disparities between well-represented and underrepresented departments.

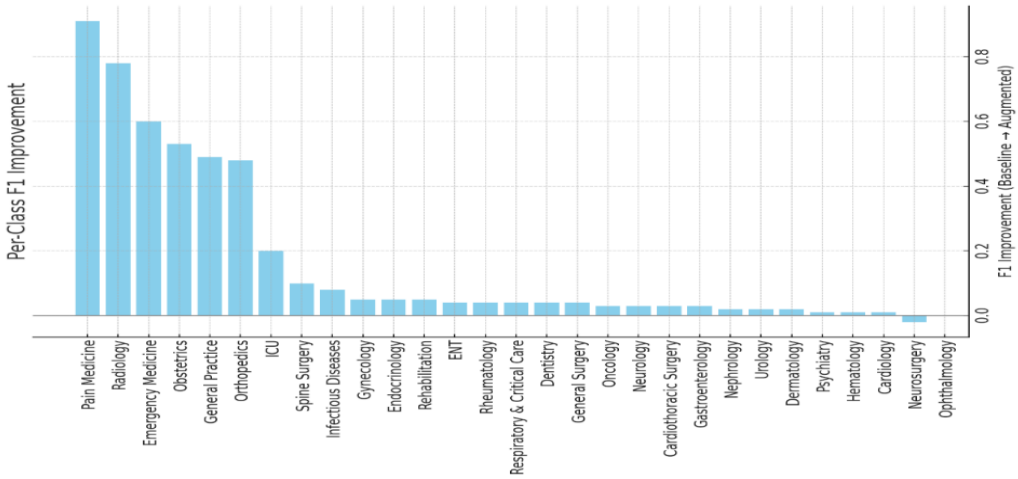

Figure 3. Model performance over different departments and their F1 scores

The general performance of the model implemented is reported in Figure 4 and Figure 5. Figure 4 shows the confusion matrix of the top 10 departments. We chose top 10 departments, to make the confusion matrix clear to readers because 29 departments will make the results of the confusion matrix distorted and not clear. The confusion matrix for Top-10 classes by support reveals significant diagonal concentration which reflects high recall by class on the offered services. General Surgery, Gastroenterology, Orthopedics, Nephrology, Neurology, Urology, and Gynecology have recalls near 0.95–1.00, and only Cardiothoracic Surgery is a bit lower at about 0.89, misclassifying mostly into Cardiology and, less frequently, gastrointestinal categories. The residual confusions are clinically interpretable, for example, Nephrology↔Urology and General Surgery↔Gastroenterology, illustrating overlapping symptom-disease families. These patterns are consistent with aggregate metrics (micro/macro AUCs) and indicate strong discrimination overall with a few semantically related, and therefore, difficult distinguishing between some specialized adjacent fields.

 ISSN: 3106-1494 | **DOI**: https://doi.org/10.65511/jaima.v1i1.735

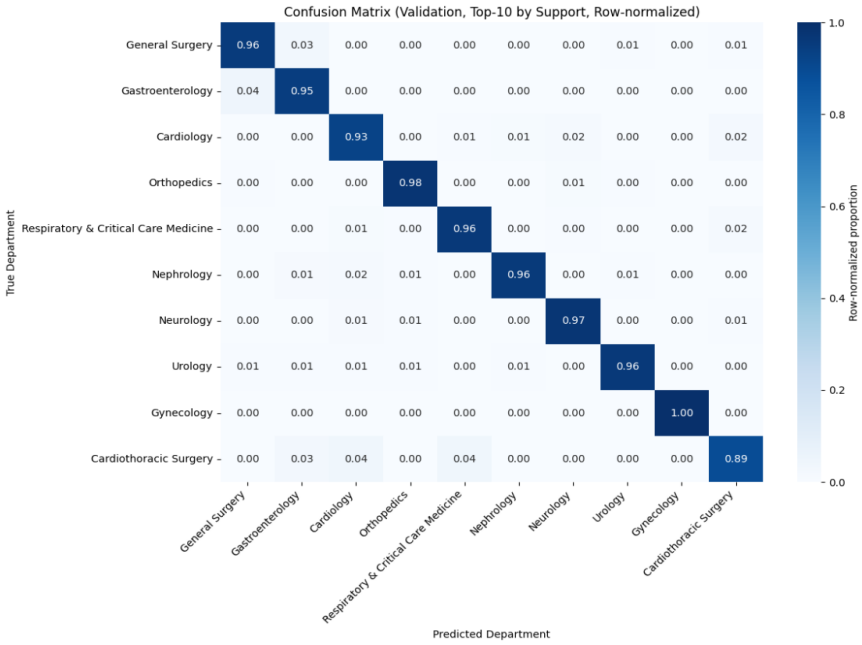

Figure 4. The confusion matrix results for the top 10 departments

The top ten departments ROC curves are displayed in Figure 5 With AUCs ≥ 0.978, all curves go into the upper-left boundary, indicating excellent separability. Gynecology (AUC≈1.000) and Neurology (AUC≈0.998) are the strongest classes. General Surgery and Respiratory & Critical Care Medicine (both ≈0.994), Orthopedics (≈0.992), Gastroenterology (≈0.991), Cardiology and Nephrology (≈0.990), and Urology (≈0.989) are next in line. In the confusion matrix, cardiothoracic surgery is comparatively weaker (AUC≈0.978), which is consistent with its confusions toward cardiology and reflects overlapping terminology (e.g., "chest pain," "valvular," "CABG"). For the majority of classes, the true-positive rate is still high at low false-positive rates, indicating that strong recall can still be achieved with conservative thresholds. All things considered, these ROC profiles demonstrate that the model reliably captures class-specific signals, with residual ambiguity concentrated in semantically related specialties.

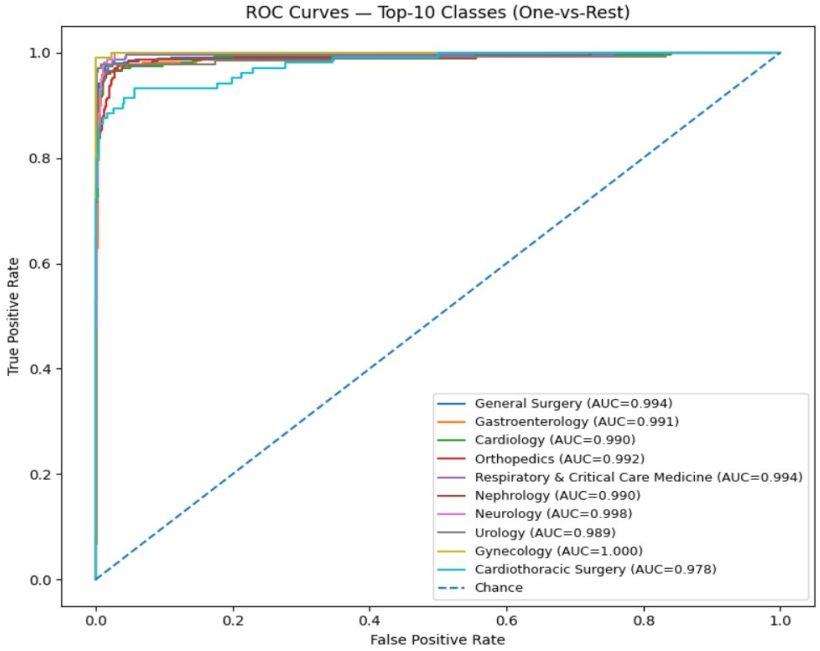

Figure 5. The ROC curves of the top 10 classes

When placed in the context of existing research on department prediction from clinical narratives, our augmented and balanced BERT-based approach delivers competitive, and in many cases superior performance, particularly for rare categories. Table 4 provides a comparison of some studies that are close to

the area of our study. Liu et al. (2022) [4], for instance, employed a BiLSTM-CRF on Chinese outpatient records and achieved a macro-F1 of 0.86, but reported sharp declines in low-support classes. Similarly, Wang et al. (2023) [10] applied CHMBERT, a Chinese BERT model, to emergency triage records, reaching a macro-F1 of 0.87, though disparities between frequent and rare departments remained evident. Chen et al. (2023) [17] used ERNIE, which uses the LLMs to enhance dataset and balance class distribution. Their model achieved 0.88 macro-F1. These approaches continued to struggle with minority class imbalance.

In comparison, our final model—built on Chinese-BERT-wwm-ext with department merging and *nlpcda*-based augmentation—achieved a macro-F1 of 0.89, while maintaining balanced performance across 26 diverse departments. This not only places our results at the upper end of reported outcomes but also highlights a meaningful step forward in addressing the rare-class bottleneck that has consistently limited prior approaches. The reduced variance in per-class F1 scores suggests that our augmentation and balancing strategy effectively improved fairness without diminishing accuracy in well-represented categories, offering a more reliable basis for real-world triage applications.

Table 4. A comparison between our model and the previous studies

Study & Year	Dataset Type	Model	# Depts	Macro-F1	Rare Class Handling
Liu et al. (2022) [4]	Outpatient records (Chinese)	BiLSTM-CRF	20	0.86	None
Wang et al. (2023) [10]	Emergency triage (Chinese)	CHMBERT	22	0.87	Weighted loss only
Chen et al. (2023) [17]	Multi-hospital EHR (Chinese)	ERNIE 3.0	25	0.87	SMOTE for rare classes
Our Work (2025)	MedCD admission + consultation (Chinese)	Chinese-BERT-wwm-ext + nlpcda	26	0.89	nlpcda augmentation + balancing

While the proposed framework achieved strong performance and addressed several key challenges, certain limitations should be acknowledged. These limitations highlight areas where future work may further refine and extend the approach, particularly in terms of dataset coverage, external validation, and real-world applicability. Firstly, the experiments relied exclusively on the MedCD dataset, which, while diverse, may not fully capture the variability of admission records across different hospitals or regions. Although data augmentation improved rare-class performance, classes with extremely few samples were excluded, leaving some specialized departments underrepresented. Moreover, the model was fine-tuned on a single dataset and has not yet been validated by external corpora, raising questions about its robustness across institutions with different documentation practices. Also, the study focused on three primary text fields (chief complaint, history of present illness, and preliminary diagnosis), while other potentially informative fields such as lab results or imaging summaries were not incorporated. In addition to that, Fine-tuning large transformer models like Chinese-BERT-wwm-ext requires significant computational resources, which may limit adoption in low-resource hospital settings.

5. CONCLUSION

In this work, we presented an end-to-end pipeline for predicting medical departments from Chinese admission records using the MedCD dataset. The study focused on three key strategies: first, simplifying the label space by merging closely related departments to reduce ambiguity; second, addressing class imbalance through *nlpcda*-based augmentation targeted at rare categories; and third, fine-tuning the Chinese-BERT-wwm-ext model to capture the linguistic and semantic subtleties of Chinese medical narratives. Together, these steps raised macro-F1 from 0.66 in the baseline run to 0.89 in the final augmented model, while narrowing the performance gap between common and rare departments. This improvement demonstrates that carefully designed augmentation can strengthen minority class performance without compromising results on well-represented classes. For future work could explore the integration of multimodal inputs—for example, combining admission text with laboratory results or imaging reports such as chest X-rays—to enrich the predictive signal. Another promising direction lies in developing cross-lingual transfer models capable of handling both Chinese and other languages, such as Arabic or English. Such extensions would not only broaden the scope of department prediction but also support deployment in multilingual healthcare systems, ultimately making automated triage assistance more accessible across diverse clinical environments.

AUTHORS CONTRIBUATIONS

All authors involved in this paper contributed equally

COFLICTS OF INTERESTS

There is no conflict of interest to our knowledge between our work and any other work already done

DATA AVAILABILITY STATEMENTS

The data used is available on: https://ieee-dataport.org/documents/medcd-medical-clinical-dataset

REFERENCES

[1] C. Shei, *Understanding the Chinese Language: A Comprehensive Linguistic Introduction*. London: Routledge, 2014. doi: 10.4324/9781315767222.

[2] R. Garriga, T. S. Buda, J. Guerreiro, J. Omaña Iglesias, I. Estella Aguerri, and A. Matić, "Combining clinical notes with structured electronic health records enhances the prediction of mental health crises," *Cell Rep Med*, vol. 4, no. 11, p. 101260, Oct. 2023, doi: 10.1016/j.xcrm.2023.101260.

[3] K. Huang, J. Altosaar, and R. Ranganath, "ClinicalBERT: Modeling Clinical Notes and Predicting Hospital Readmission," Nov. 29, 2020, *arXiv*: arXiv:1904.05342. doi: 10.48550/arXiv.1904.05342.

[4] H. Liu *et al.*, "A Natural Language Processing Pipeline of Chinese Free-text Radiology Reports for Liver Cancer Diagnosis," *IEEE Access*, vol. 8, pp. 159110–159119, 2020, doi: 10.1109/ACCESS.2020.3020138.

[5] X. Li, Y. Zhang, J. Jin, F. Sun, N. Li, and S. Liang, "A model of integrating convolution and BiGRU dual-channel mechanism for Chinese medical text classifications," *PLOS ONE*, vol. 18, no. 3, p. e0282824, Mar. 2023, doi: 10.1371/journal.pone.0282824.

[6] S. Zhang, T. Kang, X. Zhang, D. Wen, N. Elhadad, and J. Lei, "Speculation Detection for Chinese Clinical Notes: Impacts of Word Segmentation and Embedding Models," *J Biomed Inform*, vol. 60, pp. 334–341, Apr. 2016, doi: 10.1016/j.jbi.2016.02.011.

[7] X. Long, Z. Zhang, and 425776024, "nlpcda/setup.py at master · 425776024/nlpcda," GitHub. Accessed: Aug. 13, 2025. [Online]. Available: https://github.com/425776024/nlpcda/blob/master/setup.py

[8] Y. Cui, W. Che, T. Liu, B. Qin, S. Wang, and G. Hu, "Revisiting Pre-Trained Models for Chinese Natural Language Processing," in *Findings of the Association for Computational Linguistics: EMNLP 2020*, 2020, pp. 657–668. doi: 10.18653/v1/2020.findings-emnlp.58.

[9] Y. Kang *et al.*, "Deep learning-based classification of traditional Chinese medicine: a novel approach," *Quant Imaging Med Surg*, vol. 15, no. 8, pp. 7483–7496, Aug. 2025, doi: 10.21037/qims-24-1354.

[10] J. Wang, G. Zhang, W. Wang, K. Zhang, and Y. Sheng, "Cloud-based intelligent self-diagnosis and department recommendation service using Chinese medical BERT," *J Cloud Comp*, vol. 10, no. 1, pp. 1–12, Dec. 2021, doi: 10.1186/s13677-020-00218-2.

[11] X. Chen and Y. Du, "Enhancing medical text classification with GAN-based data augmentation and multi-task learning in BERT," *Sci Rep*, vol. 15, no. 1, p. 13854, Apr. 2025, doi: 10.1038/s41598-025-98281-9.

[12] H. Chen, L. Dan, Y. Lu, M. Chen, and J. Zhang, "An improved data augmentation approach and its application in medical named entity recognition," *BMC Med Inform Decis Mak*, vol. 24, no. 1, pp. 1–13, Dec. 2024, doi: 10.1186/s12911-024-02624-x.

[13] N. Zhang, Q. Jia, K. Yin, L. Dong, F. Gao, and N. Hua, "Conceptualized Representation Learning for Chinese Biomedical Text Mining," Aug. 25, 2020, *arXiv*: arXiv:2008.10813. doi: 10.48550/arXiv.2008.10813.

[14] P. Chen, M. Zhang, X. Yu, and S. Li, "Named entity recognition of Chinese electronic medical records based on a hybrid neural network and medical MC-BERT," *BMC Med Inform Decis Mak*, vol. 22, no. 1, pp. 1–13, Dec. 2022, doi: 10.1186/s12911-022-02059-2.

[15] J. Yuan, R. Tang, X. Jiang, and X. Hu, "Large Language Models for Healthcare Data Augmentation: An Example on Patient-Trial Matching," *AMIA Annu Symp Proc*, vol. 2023, pp. 1324–1333, Jan. 2024.

[16] Y. C. Chen, "MedCD: A Medical Clinical Dataset." IEEE DataPort. doi: 10.21227/KH7N-8N28.

[17] H. Chen, Y. Zhang, Y. Jiang, and R. Duan, "Adaptive Hierarchical Text Classification Using ERNIE and Dynamic Threshold Pruning," *IEEE Access*, vol. 12, pp. 193641–193652, 2024, doi: 10.1109/ACCESS.2024.3519954.