System Report for CCL25-Eval Task 8: Improving ICD Coding with Large Language Models via Disease Entity Recognition

Tengxiao Lv, Juntao Li, Chao Liu, Haobin Yuan, Ling Luo*, Jian Wang, Hongfei Lin
School of Computer Science and Technology, Dalian University of Technology, Dalian, 116024
{tengxiaolv,juntaoli,liuchao2464687308,yhhhb}@mail.dlut.edu.cn
{lingluo,wangjian,hflin}@dlut.edu.cn
*Corresponding author

Abstract

With the widespread adoption of Electronic Medical Records (EMRs), automated coding of the International Classification of Diseases (ICD) has become increasingly essential. However, the complexity of Chinese clinical texts presents significant challenges to traditional methods. To address these issues, CCL25-Eval Task 8 organized the Chinese EMRs ICD Diagnosis Coding Evaluation. This paper presents a method based on Large Language Models (LLMs), which divides the task into primary and other diagnosis coding. For the primary diagnosis, a confidence-guided semantic retrieval strategy is applied, while ensemble learning enhanced with Named Entity Recognition (NER) is used for other diagnoses. The proposed approach achieved 83.42% accuracy on the official test set, ranking second in the evaluation.

Keywords: Chinese Electronic Medical Records , ICD Diagnosis Coding , LLMs

1 Introduction

With an aging population and increased health awareness, healthcare systems face growing pressure. The widespread use of Electronic Medical Records (EMRs) improves service efficiency and quality (Jha et al., 2010). To standardize medical data and enable sharing, the World Health Organization (WTO) introduced the International Classification of Diseases (ICD), which encodes diseases into standardized codes critical for clinical records and health statistics. Manual ICD coding is inefficient and prone to errors (Shi et al., 2017), making automated coding systems essential.

To promote automatic ICD coding from Chinese EMRs, the 2025 China National Conference on Computational Linguistics (CCL25) organized task 8: ICD coding from Chinese EMRs. The task involves predicting diagnosis codes from free-text records. Challenges include the flexible Chinese clinical language and semantic ambiguity from intertwined diagnoses. Traditional methods show low accuracy (Yan et al., 2022).

Large Language Models (LLMs) have advanced natural language understanding and generation, offering new potential for medical text processing (Mann et al., 2020). However, ICD coding remains difficult due to semantic gaps between clinical text and ICD codes (Huang et al., 2022), lack of confidence estimation in outputs, and complex diagnosis dependencies.

We propose an LLM-based approach for automatic ICD coding of Chinese EMRs. In our approach, the task is divided into two subtasks: predicting the primary diagnosis code and predicting other diagnosis codes. For each subtask, we apply targeted training and optimization strategies. Our main contributions are as follows:

(1) We enhance the robustness of primary diagnosis coding by integrating token-level confidence modeling, semantic retrieval, and few-shot re-ranking.

(2) We improve other diagnosis coding by incorporating disease entity information and applying ensemble learning.

(3) We achieve 83.42% accuracy on the official test set and rank second in the public evaluation challenge.

Proceedings of the 24th China National Conference on Computational Linguistics, pages 304-311, Jinan, China, August 11-14, 2025.

2 Related Work

In medical informatics, the WHO's ICD standard converts thousands of diseases into a standardized coding system. However, manual ICD coding of EMRs is inefficient and error-prone, making automated systems necessary (Mahdi et al., 2023). Machine learning has enabled statistical methods like KNN, Naive Bayes, and SVM to automate this process by learning from annotated data (Larkey and Croft, 1996; Koopman et al., 2015; Scheurwegs et al., 2017). Yet, these methods need large datasets and complex feature engineering.

In recent years, deep learning techniques, including convolutional neural networks (CNNs) and recurrent neural networks (RNNs), have been widely adopted for semantic representation and code prediction from clinical texts (Mullenbach et al., 2018; Luo et al., 2021). For instance, Li (Li et al., 2023) proposed a knowledge-enhanced framework combining graph attention networks (GATs) and multi-task learning (MTL). This framework constructs a heterogeneous text graph consisting of concept nodes and document nodes to capture semantic correlations and mitigate the issue of data imbalance. Moreover, the emergence of pre-trained language models (Devlin et al., 2019; Liu et al., 2019) has further advanced the automation of ICD coding. Zhang (Zhang et al., 2020) proposed the BERT-XML model, which integrates BERT pre-training with a multi-label attention mechanism. The model is trained from scratch on EMRs to accommodate domain-specific vocabulary and extend sequence length. This approach effectively addresses the high cost and low efficiency of manual ICD coding, as well as the limitations of traditional methods in processing EMRs texts.

Named Entity Recognition (NER) is a key Natural Language Processing (NLP) task that identifies entities like names, locations, diseases, and symptoms in text (Xu et al., 2024). In this study, NER helps extract clinical information from EMRs, such as symptoms and disease names, supporting accurate ICD coding by providing semantic context for diagnosis. Meanwhile, large language models (LLMs) (Achiam et al., 2023; Guo et al., 2025) enhance automated coding through prompt engineering, few-shot learning, and domain-specific tuning. For instance, the Taiyi model (Luo et al., 2024), based on the Qwen framework, shows improved performance in tasks like medical QA and NER after fine-tuning on medical datasets.

3 System Overview

We designed an LLM-based ICD diagnosis coding framework for Chinese EMRs, as shown in Figure 1. The framework maps diagnosis codes to standardized disease names, develops fine-tuned models for primary and other diagnoses, enhances primary diagnosis robustness with confidence-based retrieval, and improves other diagnosis understanding with NER information and model ensembling. The predictions are then combined to produce the complete ICD coding results. Details are provided in the following sections.

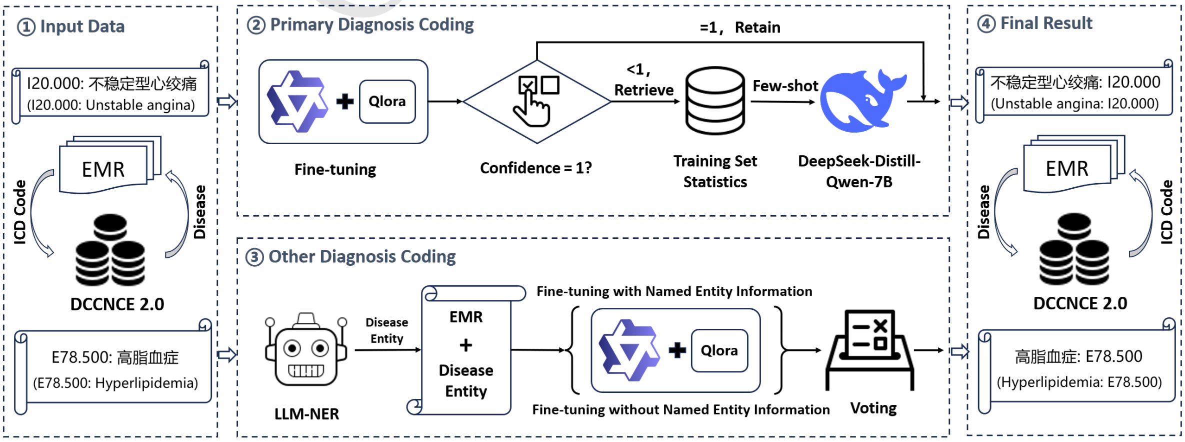
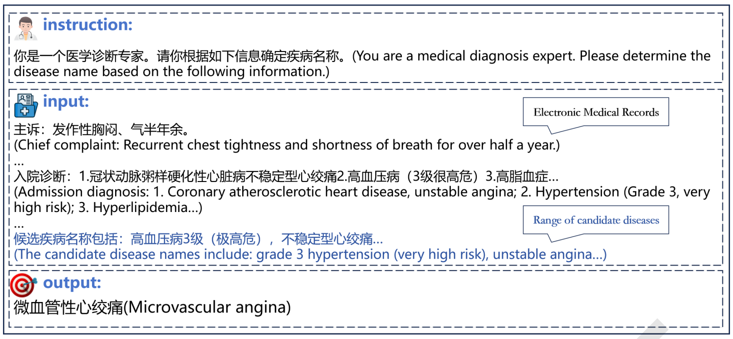


Figure 1: Diagram of ICD diagnosis coding framework for Chinese EMRs based on LLMs

3.1 Primary Diagnosis Coding Based on Confidence-Guided Retrieval

Before fine-tuning, we standardized diagnostic codes by mapping them to disease names defined in the National Clinical Edition 2.0 of the Disease Classification and Coding (DCCNCE 2.0). Disease names provide richer semantics than abstract codes, helping LLMs understand clinical content. We introduced a structured prompt strategy for fine-tuning with QLoRA, including task instructions, clinical information, and candidate disease names. The model is trained to select the correct disease name for each diagnosis. An example prompt for primary diagnosis fine-tuning is shown in Figure 2. Note that the English in brackets is not part of the input and output; it is the translation of the Chinese.



instruction:

你是一个医学诊断专家。请你根据如下信息确定疾病名称。(You are a medical diagnosis expert. Please determine the disease name based on the following information.)

input:

Electronic Medical Records

主诉：发作性胸闷、气半年余。
(Chief complaint: Recurrent chest tightness and shortness of breath for over half a year.)
...
入院诊断：1.冠状动脉粥样硬化性心脏病不稳定型心绞痛2.高血压病（3级很高危）3.高脂血症...
(Admission diagnosis: 1. Coronary atherosclerotic heart disease, unstable angina; 2. Hypertension (Grade 3, very high risk); 3. Hyperlipidemia...)
...

Range of candidate diseases

候选疾病名称包括：高血压病3级（极高危），不稳定型心绞痛...
(The candidate disease names include: grade 3 hypertension (very high risk), unstable angina...)

output:

微血管性心绞痛(Microvascular angina)

Figure 2: Diagnostic Coding Fine-tuning Prompt Example

During primary diagnosis coding, we use the LLM to generate results and also introduce a confidence scoring mechanism. This mechanism calculates the average prediction probability of each token in the generated sequence to assess the reliability of the model's output. The calculation proceeds as follows:

At the t-th generation step, the model outputs the next token based on the input sequence x and the previously generated tokens $y_{<t}$. The predicted logits vector is:

$$\mathbf{z}^{(t)} = f_\theta(y_{<t}, x) \tag{1}$$

Here, f_θ denotes the conditional probability prediction function of the language model, and $\mathbf{z}^{(t)}$ is the unnormalized score for each token in the vocabulary.

The logits vector $\mathbf{z}^{(t)}$ is converted into a probability distribution via the `softmax` function. The probability of the token $y^{(t)}$ actually generated at step t is defined as the confidence for that step:

$$\text{conf}^{(t)} = \text{softmax}(\mathbf{z}^{(t)})_{y^{(t)}} \tag{2}$$

The average confidence across the entire generated sequence is:

$$\text{Confidence} = \frac{1}{T}\sum_{t=1}^{T} \text{conf}^{(t)} \tag{3}$$

where T denotes the total number of generated tokens. This confidence score reflects the overall certainty of the model's output and can be used for quality control and uncertainty evaluation.

Based on this, we constructed a mapping database of "primary admission diagnosis" and its corresponding primary diagnosis codings to assist in judging and correcting the coding results. The specific process is as follows: If the confidence of the primary diagnosis coding output by the model is 1, the result is directly retained; if the confidence is less than 1, the database is searched for the instance with the closest semantic meaning to the current primary admission diagnosis using cosine similarity. If the model's predicted code is inconsistent with the search result, the current case text and the most similar example are combined into a few-shot input and passed to the DeepSeek-Distill-Qwen-7B model to rejudge the primary diagnosis coding. Finally, all case coding results are integrated to obtain a more accurate prediction of the primary diagnosis coding.

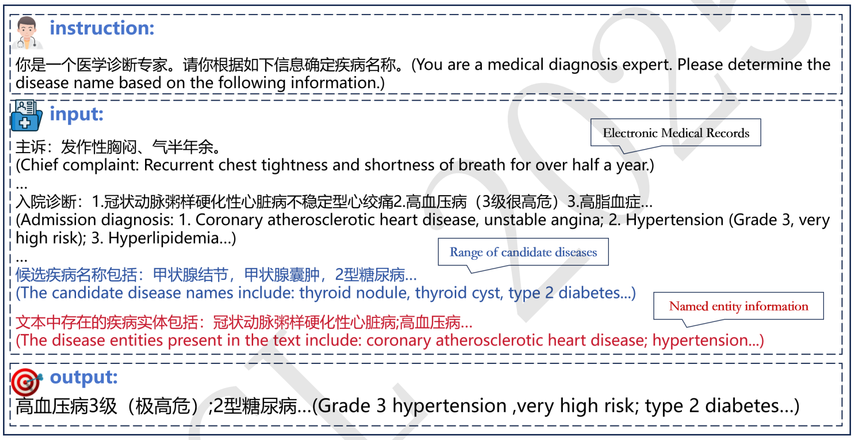
3.2 ICD Coding for Other Diagnoses Based on NER Information

To support the generation of other diagnosis codes, we trained a large-scale NER model focused on entities of the ”disease and diagnosis” type. The model is based on the Chinese EMRs dataset CCKS2019, with Qwen2.5-7B-Instruct used as the base model.

During inference, the specific entity type to be recognized is clearly specified first. Then, the corresponding raw text sequence is generated. In the output, special symbols are used to mark entity boundaries: the symbol ”[” indicates the start of an entity, and ”]” marks the end. This annotation method processes each entity type separately. It has a clear and structured format, making it easier for the model to learn entity boundary features. This approach transforms the traditional sequence labeling task into a text generation task, which is well-suited to the generalization capabilities of LLMs.

After training the NER model (LLM-NER), we applied it to extract “disease and diagnosis” entities from EMRs texts in the ICD coding task. These recognized entities were then used as supplementary semantic information and injected into the input of the model for other diagnosis coding. The model was further fine-tuned with this enhanced input.

During data preparation, the input was enhanced in two stages. First, a candidate list of possible disease names was added to the original text. Then, the output from the NER results was appended. This strategy helps the model capture richer clinical semantics during training and improves its performance in ICD coding. Figure 3 shows an example prompt for fine-tuning the model for other diagnoses. Note that the English in brackets is not part of the input and output; it is the translation of the Chinese.



instruction:

你是一个医学诊断专家。请你根据如下信息确定疾病名称。(You are a medical diagnosis expert. Please determine the disease name based on the following information.)

input:

主诉：发作性胸闷、气半年余。
(Chief complaint: Recurrent chest tightness and shortness of breath for over half a year.)
...
入院诊断：1.冠状动脉粥样硬化性心脏病不稳定型心绞痛2.高血压病（3级很高危）3.高脂血症...
(Admission diagnosis: 1. Coronary atherosclerotic heart disease, unstable angina; 2. Hypertension (Grade 3, very high risk); 3. Hyperlipidemia...)
...
候选疾病名称包括：甲状腺结节，甲状腺囊肿，2型糖尿病...
(The candidate disease names include: thyroid nodule, thyroid cyst, type 2 diabetes...)

文本中存在的疾病实体包括：冠状动脉粥样硬化性心脏病;高血压病...
(The disease entities present in the text include: coronary atherosclerotic heart disease; hypertension...)

output:

高血压病3级（极高危）;2型糖尿病...(Grade 3 hypertension ,very high risk; type 2 diabetes...)

Figure 3: Example Prompt for Fine-tuning the Other Diagnosis Coding Model

Finally, we ensemble the model fine-tuned with entity information and the model without it. For each model, predictions from eight different inference rounds are aggregated using a hard voting strategy. If a result appears more than seven times, it is considered correct. The final ICD codes for other diagnoses are generated based on the combined voting results from both models.

4 Experimental Results and Analysis

4.1 Experiment Settings and Metrics

The experiments in this study used a Chinese EMRs ICD diagnosis coding evaluation dataset, which contains a total of 1485 samples, divided into training, validation, and test sets with 800, 200, and 485 samples, respectively. The main hyperparameter settings are listed in Table 1, and all experiments were conducted on a single NVIDIA L40 GPU.

Epochs	Dropout	Optimizer	Batch Size	Max Length	Learning Rate
10	0.05	AdamW	2	5120	1e-4

Table 1: Main hyperparameter settings in the experiment

In the Chinese EMRs ICD diagnosis coding task, Acc is used as the evaluation metric. The calculation formula is as follows:

$$\text{Acc} = \frac{1}{N}\sum_{i=1}^{N}\{0.5 \cdot I(y_{\text{main}} = \hat{y}_{\text{main}}) + 0.5 \cdot F1(y_{\text{other}}, \hat{y}_{\text{other}})\}_i \tag{4}$$

Here, $I()$ denotes an indicator function that returns 1 if the condition is met and 0 otherwise; $\hat{y}_{\text{main}}$ and y_{main} represent the predicted and ground-truth labels for the primary diagnosis coding, respectively; $\hat{y}_{\text{other}}$ and y_{other} refer to the predicted and ground-truth sets of other diagnosis codes; N is the number of test samples; and $F1()$ denotes the F1 score.

4.2 Effect of Confidence and Retrieval on Primary Diagnosis

This experiment evaluates how introducing a confidence mechanism and semantic retrieval strategy affects the performance of primary diagnosis coding on the test set. It compares three settings: using model output confidence alone, using semantic similarity retrieval alone, and combining both methods. The experimental results are shown in Figure 4, where the baseline model refers to the result obtained through direct fine-tuning.

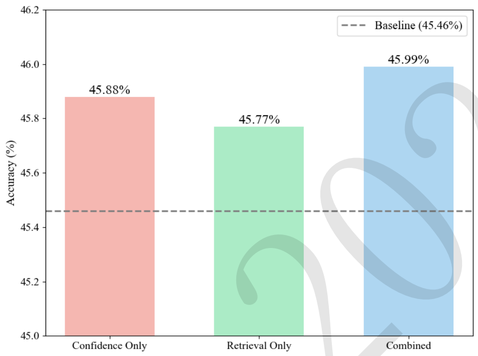


Figure 4: Primary diagnosis boosted by confidence and retrieval on the test set

As shown in Figure 4: (1) Both the confidence mechanism and semantic retrieval strategy individually improve the baseline model's primary diagnosis coding by effectively correcting prediction errors; (2) Combining both strategies achieves the highest accuracy (45.99%), demonstrating their complementary nature; (3) The results highlight that integrating internal generation probabilities with external semantic validation significantly enhances coding accuracy.

4.3 Effect of NER on Other Diagnosis Accuracy

This experiment evaluates the effectiveness of incorporating NER information in improving the performance of the other diagnosis coding task. Three settings are compared: directly fine-tuned models, models fine-tuned with additional NER information, and an ensemble of the two. The results on the validation and test sets are shown in Table 2.

	Direct Fine-tuned	Fine-tuned with NER	Ensemble
Val	36.11	36.75	38.30
Test	35.42	35.86	37.43

Table 2: Comparison of the accuracy of the three methods on the validation set and test set

As shown in Table 2: (1) After incorporating NER information, the model's accuracy on the validation and test sets increased to 36.75% and 35.86%, respectively, indicating that NER information enhances the model's understanding of diagnostic semantics; (2) By integrating the directly fine-tuned model with the NER-enhanced model, the accuracy on the validation and test sets further improved to 38.30% and 37.43%, respectively, demonstrating that model integration effectively improves coding accuracy.

Additionally, we compared two base models—DeepSeek-Distill-Qwen-7B (DS-R1-Qwen-7B) and Qwen2.5-7B—using zero-shot and fine-tuning strategies for NER, and analyzed their impact on other diagnosis coding accuracy. Results are shown in Figure 5.

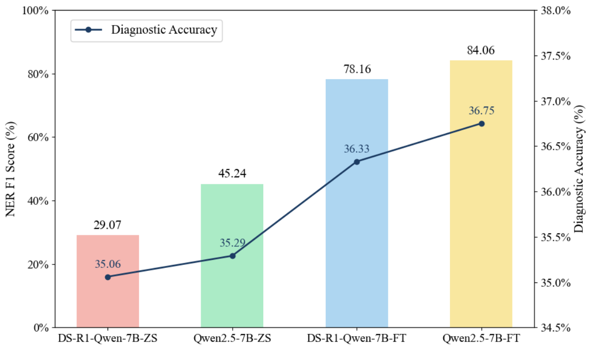


Figure 5: Effect of NER quality on other diagnosis Accuracy on the validation set

As shown in Figure 5: (1) Fine-tuning significantly outperforms zero-shot inference across all base NER models, improving F1 scores and entity boundary/type recognition; (2) Enhanced NER performance increases other diagnosis coding accuracy, confirming high-quality entity information benefits coding tasks; (3) Selecting appropriate base models with targeted fine-tuning effectively improves both NER and ICD coding accuracy.

4.4 Comparison with Other Methods

Table 3 presents the top five teams' scores in the public track of this evaluation task. Our team's proposed model achieved excellent performance in the ICD diagnosis coding task, ultimately ranking second. This result fully validates the effectiveness of the coding method based on LLMs combined with confidence-based retrieval and NER enhancement. The approach leverages semantic information and model uncertainty to improve the accuracy of primary diagnosis coding, providing a practical and effective solution for automated EMRs coding.

Rank	1	**2 (Our Team)**	3	4	5
Score	86.72	83.42	82.33	81.59	81.30

Table 3: The official results on the test set

5 Conclusion and Outlook

In this paper, we propose an LLM-based ICD coding method for Chinese EMRs that improves accuracy by integrating confidence-based retrieval mechanisms and NER information. This method achieves good performance in public evaluations, ranking second. However, limitations remain in other diagnosis coding accuracy and in the underutilization of medical entities, such as drugs, in NER. In future work, we will expand entity categories and enhance semantic understanding to advance clinical applications of automated ICD coding.

Acknowledgements

The authors thank the organizers of the CCL25-Eval Task 8. This research was supported by the National Natural Science Foundation of China (No. 62302076, 62276043).

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

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

Ashish K Jha, Catherine M DesRoches, Peter D Kralovec, and Maulik S Joshi. 2010. A progress report on electronic health records in us hospitals. *Health affairs*, 29(10):1951–1957.

Bevan Koopman, Guido Zuccon, Anthony Nguyen, Anton Bergheim, and Narelle Grayson. 2015. Automatic icd-10 classification of cancers from free-text death certificates. *International journal of medical informatics*, 84(11):956–965.

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

Xinhang Li, Xiangyu Zhao, Yong Zhang, and Chunxiao Xing. 2023. Towards automatic icd coding via knowledge enhanced multi-task learning. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pages 1238–1248.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Junyu Luo, Cao Xiao, Lucas Glass, Jimeng Sun, and Fenglong Ma. 2021. Fusion: Towards automated icd coding via feature compression. In *Findings of the association for computational linguistics: ACL-IJCNLP 2021*, pages 2096–2101.

Ling Luo, Jinzhong Ning, Yingwen Zhao, Zhijun Wang, Zeyuan Ding, Peng Chen, Weiru Fu, Qinyu Han, Guangtao Xu, Yunzhi Qiu, et al. 2024. Taiyi: a bilingual fine-tuned large language model for diverse biomedical tasks. *Journal of the American Medical Informatics Association*, 31(9):1865–1874.

Soha Sadat Mahdi, Nikos Deligiannis, and Hichem Sahli. 2023. A review of deep learning methods for automated clinical coding. In *2023 15th International Conference on Computer and Automation Engineering (ICCAE)*, pages 35–39. IEEE.

Ben Mann, N Ryder, M Subbiah, J Kaplan, P Dhariwal, A Neelakantan, P Shyam, G Sastry, A Askell, S Agarwal, et al. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 1:3.

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*.

Elyne Scheurwegs, Boris Cule, Kim Luyckx, Léon Luyten, and Walter Daelemans. 2017. Selecting relevant features from the electronic health record for clinical code prediction. *Journal of biomedical informatics*, 74:92–103.

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*.

Derong Xu, Wei Chen, Wenjun Peng, Chao Zhang, Tong Xu, Xiangyu Zhao, Xian Wu, Yefeng Zheng, Yang Wang, and Enhong Chen. 2024. Large language models for generative information extraction: A survey. *Frontiers of Computer Science*, 18(6):186357.

Chenwei Yan, Xiangling Fu, Xien Liu, Yuanqiu Zhang, Yue Gao, Ji Wu, and Qiang Li. 2022. A survey of automated international classification of diseases coding: development, challenges, and applications. *Intelligent Medicine*, 2(03):161–173.

Zachariah Zhang, Jingshu Liu, and Narges Razavian. 2020. Bert-xml: Large scale automated icd coding using bert pretraining. *arXiv preprint arXiv:2006.03685*.