Application of Conformer Architecture in Clinical Speech Input and Intelligent Medical Record Generation

Xiangkun Zou[1], Lei Wang[1], Jing Sun[1], Shengyu Guo[1], Ning Li[2]
[1]Internet Hospital and Remote Collaboration Office, The First Affiliated Hospital of Hebei North University, Zhangjiakou 075000, Hebei, China
[2]Nursing Department, The First Affiliated Hospital of Hebei North University, Zhangjiakou 075000, Hebei, China
Corresponding author: Lei Wang (e-mail: wljc2025@163.com)

ABSTRACT Accurate clinical speech recognition remains challenging because rapid pronunciation, domain-specific terminology, and background noise often degrade automatic speech recognition and subsequent medical record generation. This study proposes a multi-stage intelligent documentation framework that integrates a 12-layer Conformer architecture, BERT-BiLSTM-CRF semantic modeling, and BART-based structured text generation. The Conformer encoder captures both local acoustic characteristics and long-range contextual dependencies, while the semantic module performs medical entity recognition and normalization to enhance terminology consistency. The extracted information is subsequently incorporated into a BART generator with clinical knowledge prompts to produce standardized SOAP-compliant medical records. Experimental results demonstrate a word error rate of 6.3%, medical term accuracy of 95.8%, low response latency of approximately 940–960 ms, and generation quality approaching physician-written records. Beyond clinical documentation, the proposed framework illustrates the effectiveness of deep time-frequency feature extraction and contextual sequence modeling for complex noisy signals, offering methodological insights for electromagnetic signal interpretation, antenna measurement data processing, and intelligent information extraction in propagation-related applications.

INDEX TERMS clinical speech recognition, conformer architecture, intelligent medical record generation, semantic feature extraction, electromagnetic signal processing

I. INTRODUCTION

CLINICAL speech capture is vital for improving documentation efficiency in smart healthcare [1,2]. However, traditional ASR systems [3,4], trained on general corpora, struggle in clinical settings due to dense terminology, rapid speech, colloquial expressions, noise, and accents [5]. These challenges lead to high error rates in recognizing key medical entities, undermining the accuracy and usability of generated records [6,7]. The high-precision and robust clinical speech processing framework developed in this study possesses significant potential for cross-domain transfer. In the context of intelligent upgrading in the industry, scenarios such as noise control in production environments, verbal instruction delivery for complex processes, and rapid voice input of quality inspection results face similar challenges to the clinical environment—namely, the need to accurately identify technical terms and generate structured records under noise interference. The Conformer architecture employed in this system, with its keen ability to capture local features and its backend semantic understanding and generation process that integrates domain knowledge, provides a feasible technical paradigm for developing intelligent voice interaction and automated report generation systems suitable for workshops. This offers important reference value for promoting the digitalization and intelligentization of the industry.

This paper presents a clinical speech-to-medical-record system based on the Conformer architecture to address inaccurate terminology recognition and semantic distortion in medical ASR. The Conformer-Transducer model serves as the acoustic front-end, leveraging convolutional modules and self-attention with relative position encoding to capture local acoustic patterns and improve discrimination of confusable medical terms. A multi-stage integration pipeline then integrates BERT-BiLSTM-CRF for medical entity recognition and UMLS terminology normalization, while a BART-large model generates SOAP-compliant records using learnable clinical knowledge prompts and template-driven constraints. By converting SOAP structures into sequence constraints, the system ensures compliance with electronic medical record standards. The proposed approach demonstrates Conformer's superiority in local feature sensitivity and contextual modeling, offering a reusable technical paradigm for intelligent clinical documentation.

II. RELATED WORK

Automatic speech recognition technology has evolved over the past few decades from the traditional GMM (Gaussian Mixture Model)-HMM (Hidden Markov Model) system to a deep, multi-stage integration architecture [8]. Early systems [9] relied on a cascade structure of acoustic, pronunciation, and language models, with limited modeling capabilities and inconsistent optimization goals. With the rise of deep neural networks, deep belief networks, and recurrent neural networks [10,11], these models have been widely used for acoustic modeling to improve recognition performance. In recent years, multi-stage integration models based on attention mechanisms have become mainstream, and the Listen-Attend-Spell [12] and Transformer [13] architectures have achieved direct mapping of speech sequences to text. However, pure attention mechanisms exhibit high computational complexity when processing long speech sequences, and their ability to model local time-frequency features is limited, which can easily lead to the misrecognition of short-term phonemes. To address this shortcoming, the Conformer model [14,15] combines the local perception characteristics of convolutional neural networks with the global context modeling advantages of the Transformer. It enhances frame-level feature extraction by utilizing convolutional modules. It uses multihead self-attention of relative position encoding to capture long-distance dependencies, achieving leading performance on benchmark datasets such as LibriSpeech. Conformer-Transducer [16] further supports streaming recognition and expands its application in real-time scenarios.

Some work has adopted dictionary weighting or pronunciation dictionary expansion strategies to enhance the recognition ability of drug names and disease names; however, it is challenging to address unregistered words and pronunciation variations [17]. Another method type utilizes transfer learning [18,19] to fine-tune pre-trained models on medical speech data, achieving progress on multiple datasets. However, the model structure is not optimized for the high term density and short-term acoustic changes of medical speech, resulting in a high error rate in key entity recognition. In recent years, a small number of studies have attempted to apply Transformer [20] or Conformer architectures to process speech, but have not conducted an in-depth analysis of the collaborative mechanism of convolution and attention modules in medical terminology modeling. In addition, most systems only output raw text, lacking semantic understanding and structured processing of recognition results, making it difficult to generate documents that meet electronic medical record standards directly [21]. In the area of intelligent medical record generation [22,23], early systems relied on template filling and rule engines, resulting in limited flexibility. Recent research has employed Seq2Seq (Sequence-to-Sequence) models [24,25] for multi-stage integration generation; however, these systems often rely on human-generated text as input and lack a deep coupling with the speech recognition frontend, resulting in error propagation.

III. METHOD

A. OVERALL SYSTEM ARCHITECTURE

Figure 1 shows the overall system architecture.

Figure 1 illustrates the multi-stage integration workflow of the Conformer architecture-based system for generating intelligent medical records from clinical speech. The system input is the physician's spoken clinical speech. After audio preprocessing, an 80-dimensional Mel-spectrogram is extracted and fed into a Conformer-Transducer model for speech recognition. This model utilizes a 12-layer stack of Conformer blocks, combined with a multi-head self-attention mechanism and depthwise separable convolutions, to jointly model the short-term acoustic features and long-range semantic dependencies of medical terms, and outputs preliminary ASR text.

The semantic understanding layer performs multi-level processing on the ASR [26,27] text: first, a spelling correction module based on Levenshtein distance is used to correct terminology misrecognition; then the BERT-BiLSTM-CRF model is used to perform medical named entity recognition, covering 12 types of clinical entities such as diseases, drugs, and dosages, and standardized mapping is performed through the UMLS and ICD-10 terminology systems to generate structured semantic representations.

B. CLINICAL SPEECH RECOGNITION BASED ON CONFORMER-TRANSDUCER

The multi-head self-attention mechanism uses relative position encoding to enhance the model's ability to model the temporal relationships between phonemes in speech sequences. Let the input sequence be $x \in \mathbb{R}^{T\times d}$, where T is the time step and d is the feature dimension. The attention output is:

$$\text{Attention}(Q,K,V) = \text{softmax}\left(\frac{QK^\top + R^\top}{\sqrt{d_k}}\right)V \quad (1)$$

The convolution module uses a depth-wise separable one-dimensional convolution with a kernel size of 15 and a gated linear unit activation function to enhance the ability to extract local speech patterns. Its calculation form is:

$$\text{ConvModule}(X) = (X * W_{\text{conv}}) \otimes \sigma(X * W_{\text{gate}}) \quad (2)$$

In formula (2), $\otimes$ represents element-by-element multiplication, and σ is the sigmoid function, which implements the gating mechanism. The feedforward network adopts a two-layer fully connected structure, with a hidden layer dimension of 2048 and an activation function of Swish, which is in the form of:

$$\text{FFN}(X) = \text{Swish}(XW_1 + b_1)W_2 + b_2 \quad (3)$$

All submodules are trained using residual connections and layer normalization to stabilize the training process:

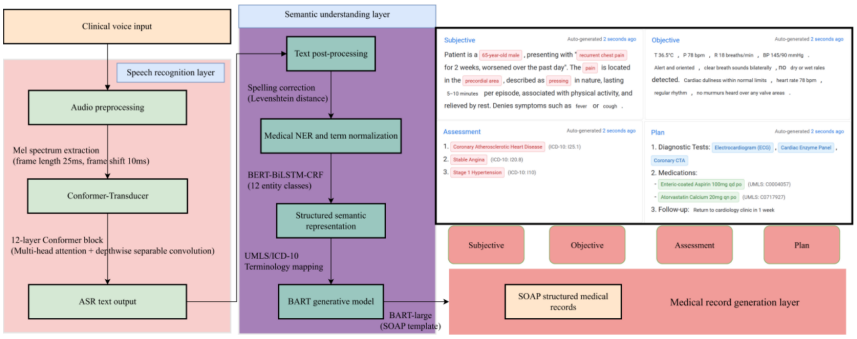

FIGURE 1. Overall architecture of this system

$$X_{\text{out}} = \text{LayerNorm}(X + \text{Dropout}(\text{Sublayer}(X))) \quad (4)$$

The entire model uses the RNN (Recurrent Neural Network)-Transducer loss function to achieve multi-stage integration alignment of the input acoustic frame sequence and the output symbol sequence without forced synchronization. The Transducer loss is defined as:

$$L_{\text{Transducer}} = -\log \sum_{z \in Z^*(x,y)} p(z \mid x) \quad (5)$$

In formula (5), Z^* is the set of all alignment paths corresponding to the target sequence y, and $p(z \mid x)$ is determined by the conditional probability output by the joint network:

$$p(y_u \mid h_t^{\text{enc}}; s_{u-1}^{\text{pred}}) = \text{Softmax}\Big(W_j[h_t^{\text{enc}}; s_{u-1}^{\text{pred}}] + b_j\Big) \quad (6)$$

In formula (6), h_t^{enc} represents the encoder hidden state, and s_{u-1}^{pred} represents the state of the prediction network at the previous moment.

C. MEDICAL TEXT POST-PROCESSING AND SEMANTIC UNDERSTANDING

Medical terminology error correction based on Levenshtein distance. ASR output text often misrecognizes medical terms due to similar pronunciations. To this end, we constructed a medical terminology dictionary, D_{med}, which contains approximately 15,000 standard terms, including ICD-10, generic drug names, and commonly used clinical abbreviations. For each word w in the ASR output that does not match in the dictionary, we calculate its Levenshtein edit distance with all candidate words c in D_{med}:

$$LD(w,c) = \min \begin{cases} LD(w_{1:i-1}, c_{1:j}) + 1, & \text{deletion} \\ LD(w_{1:i}, c_{1:j-1}) + 1, & \text{insertion} \\ LD(w_{1:i-1}, c_{1:j-1}) + I(w_i \neq c_j), & \text{substitution} \end{cases} \quad (7)$$

In formula (7), a candidate set with similar $LD(w,c)$ lengths is selected, and the maximum probability item is selected as the correction result in combination with the n-gram language model, effectively reducing the misrecognition rate of homophones and near-phones.

A BERT-BiLSTM-CRF hybrid architecture is used for fine-grained medical entity recognition [28-29]. The input is the corrected text, which is then segmented by WordPiece and fed into the Chinese clinical pre-trained model BERT-wwm-ext-med to obtain the contextualized word vector representation $H = [h_1, \ldots, h_n] \in \mathbb{R}^{n \times d}$. Then it is fed into the BiLSTM to further extract sequence features:

$$\overrightarrow{h}_t = \text{LSTM}_{\text{forward}}(h_t, \overrightarrow{h}_{t-1}) \quad (8)$$

$$\overleftarrow{h}_t = \text{LSTM}_{\text{backward}}(h_t, \overleftarrow{h}_{t+1}) \quad (9)$$

$$s_t = [\overrightarrow{h}_t; \overleftarrow{h}_t] \quad (10)$$

The label transfer relationship is modeled through the conditional random field layer, and the optimal label sequence is output:

$$y^* = \arg\max_y \sum_{t=1}^{n} \left(A_{y_{t-1}, y_t} + W_{y_t} s_t\right) \quad (11)$$

Distance similarity is edited:

$$S_{\text{edit}}(m,c) = 1 - \frac{LD(m,c)}{\max(|m|, |c|)} \quad (12)$$

Semantic similarity: Using the medical semantic encoder trained with SimCSE (Simple Contrastive Learning of Sentence Embeddings) to calculate the cosine similarity:

$$S_{\text{sem}}(m,c) = \frac{E(m)^\top E(c)}{\|E(m)\| \|E(c)\|} \quad (13)$$

Comprehensive score:

$$S_{\text{final}}(m,c) = \alpha S_{\text{edit}}(m,c) + (1-\alpha) S_{\text{sem}}(m,c) \quad (14)$$

D. INTELLIGENT MEDICAL RECORD GENERATION MODEL BASED ON BART

The model input is an enhanced representation that combines the original speech recognition text with medical semantic annotations. The input sequence is constructed as follows:

$$x_{\text{input}} = [[CLS]] + t_{\text{ASR}} + [SEP] + e_{\text{NER}} + [SEP] \quad (15)$$

In formula (15), t_{ASR} is the text word sequence output by ASR, and e_{NER} is the corresponding NER annotation sequence. The modal information is isolated by special separators, enabling the encoder to jointly model the original speech text and structured semantic features.

During the generation phase, the model organizes output according to the standard SOAP template [30,31] to ensure that the generated medical records meet the structural requirements of the Electronic Medical Record Application Management Specifications. To enhance the clinical compliance and structural consistency of the generated text, a clinical knowledge prompt mechanism is applied. A learnable text prompt is added to the front end of the input sequence. This prompt is embedded in the input as a task instruction, guiding the model to activate relevant knowledge pathways and improve compliance with medical logic, terminology usage, and format constraints. During the fine-tuning phase, a teacher forcing strategy is adopted, that is, a left-shifted version of the true target sequence is used at the decoder input to maximize the conditional log-likelihood:

$$L_{\text{gen}} = -\sum_{t=1}^{T} \log p(y_t \mid y_{<t}, x_{\text{input}}; \theta) \quad (16)$$

The output of the last layer of the BART [32-33] decoder is linearly projected and calculated with Softmax:

$$p(y_t) = \text{Softmax}(W_V h_t^{\text{dec}}) \quad (17)$$

In formula (17), h_t^{dec} is the decoder's hidden state at step t, and $W_V \in \mathbb{R}^{V \times d}$ is the vocabulary projection matrix.

The interface of the clinical speech-to-intelligent medical record generation system is shown in Figure 2. This clinical knowledge prompting mechanism is essentially a "soft-guided" method based on a continuous vector space. The specific generation and injection logic of the prompt vectors is as follows: First, based on the NER results in the current input text, the system retrieves UMLS Metathesaurus to obtain the SNOMED-CT unique identifiers and semantic types of relevant concepts. Then, these discrete, standardized concepts are encoded into learnable continuous vectors, which are concatenated with the input sequence as prompt prefixes. During model training and inference, these prompt vectors are not immutable hard constraints, but rather dynamically guided by the BART model's cross-attention mechanism to favor activating the semantic space related to the prompt concepts when generating corresponding paragraphs. Specifically, during generation, the model calculates the attention weight between the decoder's hidden state and the prompt vector. This weight implicitly adjusts the probability distribution of the output words, shifting them towards a subset of terms related to the prompt concepts, thus achieving structured and standardized generation while maintaining a certain degree of generation flexibility. This method is optimized through supervised fine-tuning after pre-training, ensuring a balance between the effectiveness of knowledge guidance and the fluency of model generation.

IV. EXPERIMENTAL SETUP

A. DATASET CONSTRUCTION AND DIVISION

The data used in this study are collected through a multi-center collaboration across five Class-A tertiary hospitals. It encompasses physician spoken speech from real-world clinical settings and the corresponding standard electronic medical record text. This study was approved by the Ethics Committee of The First Affiliated Hospital of Hebei North University, and was performed in accordance with the principles of the Declaration of Helsinki. All eligible participants signed an informed consent form. The data totals 1,200 hours of raw speech data, collected using standard hospital-issued directional microphones with a uniform sampling rate of 16 kHz, 16-bit quantization, and monophonic storage. The recordings are collected from eight core clinical departments: Cardiology, Respiratory and Critical Care Medicine, Gastroenterology, General Surgery, Orthopedics, Emergency Medicine, Obstetrics and Gynecology, and Neurology. The speech duration is evenly distributed across departments, ensuring the model's generalizability to diverse terminology and speech patterns.

All speech and text data are de-identified in strict compliance with the Measures for the Ethical Review of Biomedical Research Involving Human Subjects and the Personal Information Protection Law. Direct identifiers such as patient names, ID numbers, and phone numbers are removed, and physician identities are anonymized. Speech engineers and medical quality control personnel then conducted a joint quality screening process, excluding recordings with a signal-to-noise ratio below 5 dB, severe distortion, intermittent speech, or non-monologue speech. Ultimately, 1,000 hours of high-quality, aligned speech-to-text paired data are retained. The dataset is divided into a training set (800 hours), a validation set (100 hours), and a test set (100 hours) in an 8:1:1 ratio based on time independence to ensure objectivity and generalizability of the evaluation results.

The sample information of the dataset is shown in Table 1.

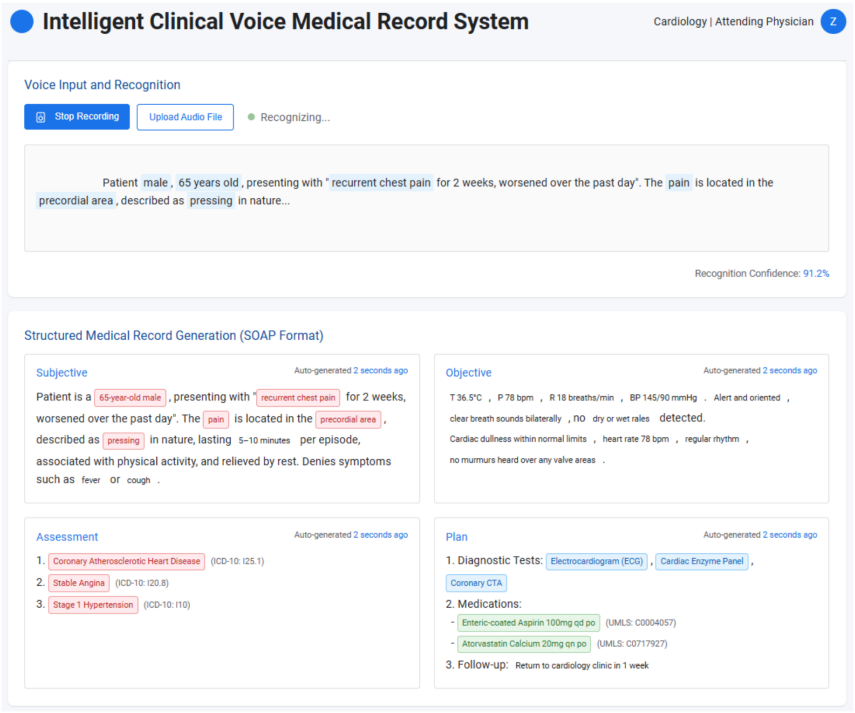

FIGURE 2. System interface

TABLE 1. Distribution of medical record entity data samples

Entity	Training set	Validation set	Test set	Total
Disease	42,150	5,269	5,268	52,687
Symptoms	38,760	4,845	4,842	48,447
Signs	29,840	3,730	3,730	37,300
Medication	33,680	4,210	4,210	42,100
Surgery	18,560	2,320	2,320	23,200
Test	31,240	3,905	3,905	39,050
Anatomic Site	26,720	3,340	3,340	33,400
Time	24,800	3,100	3,100	31,000
Dose	19,360	2,420	2,420	24,200
Frequency	16,480	2,060	2,060	20,600
Route	14,720	1,840	1,840	18,400
Negative	12,480	1,560	1,560	15,600

B. MODEL IMPLEMENTATION AND TRAINING CONFIGURATION

The ASR model [34,35] adopts the Conformer-Transducer architecture, the Noam optimizer is used as the optimizer, the initial learning rate is set to 4×10^{-4}, and the learning rate is dynamically adjusted according to $lr = d_{\text{model}}^{-0.5} \cdot \min(step^{-0.5}, step \cdot warmup^{-1.5})$ with the number of training steps. The number of steps warmup is 8,000, the batch size is 32, and the total number of training rounds is 30 epochs. The medical NER model is based on the BERT-BiLSTM-CRF architecture and uses the AdamW optimizer. The model has a learning rate of 3×10^{-5}, a batch size of 16, a maximum sequence length of 512, and is trained for 50 epochs. An early stopping mechanism (patience = 10) is applied to prevent overfitting. The BART medical record generation model also uses the AdamW optimizer, a learning rate of 3×10^{-5}, 20 epochs of training, and a batch size of 16. In the generation phase, beam search decoding is used with a beam width of 5, a length penalty factor of 1.0, and a maximum output length limit of 512 tokens to ensure that the generated medical records are complete and well-structured.

The detailed hardware configuration of this computing cluster is as follows: each computing node is equipped with eight NVIDIA A100 PCIe 80GB GPUs, interconnected via a PCIe 4.0 bus; nodes are connected via a high-speed InfiniBand network (HDR 200Gb/s) to support efficient distributed data-parallel training. All model training is performed in this environment, ensuring computational efficiency and stability for large-scale parameter updates.

C. EVALUATION METRICS

For speech recognition, the Word Error Rate (WER) is used as the core metric to measure the edit distance between the recognized text and the reference text. The calculation formula is:

$$WER = \frac{S + D + I}{N} \times 100\% \tag{18}$$

In formula (18), S is the number of substitution errors, D is the number of deletions, I is the number of insertions, and N is the total number of words in the reference text. MTA is defined as the ratio of the number of key medical terms correctly recognized in the ASR output to the total number of key medical terms in the reference text. It is calculated based on a predefined term dictionary containing 12 core

clinical entities, including diseases, drugs, and examinations. The calculation process involves first performing the same medical named entity recognition (NER) on the ASR output and the reference text. Then, the exact match

Number of Correctly Recognized Medical Terms accuracy is calculated for only the terms in the dictionary:

$$MTA = \frac{\text{Correctly recognized medical terms}}{\text{Total medical terms in reference}} \times 100\%.$$

Compared to the standard NER F1 score, MTA focuses on assessing the complete and accurate transcription of key medical entities in the multi-stage integration speech-to-text process. NER F1 typically assesses the recognition quality of entity boundaries and types in plain text. In the medical named entity recognition task, Precision, Recall, and F1-score are used for evaluation:

$$\text{Precision} = \frac{TP}{TP + FP} \tag{19}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{20}$$

$$F1 = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{21}$$

In the term normalization task, the top-1 accuracy and top-3 accuracy are used to evaluate the normalization effect. In the quality assessment of medical record generation, the automatic evaluation metrics ROUGE-L (Recall-Oriented Understudy for Gisting Evaluation - Longest Common Subsequence), BLEU-4 (Bidirectional Encoder Representations from Transformers), and METEOR (Metric for Evaluation of Translation with Explicit Ordering) are used. The ROUGE-L calculation formula is:

$$ROUGE\text{-}L = \frac{(1 + \beta^2) \cdot LCS(X, Y)}{m \cdot n} \tag{22}$$

In formula (22), LCS(X,Y) is the length of the longest common subsequence between the generated text and the reference text, and m and n are their respective lengths.

A manual evaluation is conducted by 30 attending physicians with more than 5 years of clinical experience who performed double-blind scoring (1–5 points) on the generated medical records. The evaluation criteria included content completeness, medical accuracy, structural standardization, and language readability. For system performance, the P95 value of multi-stage integration response latency is measured to reflect the system's real-time performance. The System Usability Scale (SUS) is used to assess user experience. The SUS consists of 10 questions, each scored on a scale of 1-5. The total SUS score is calculated by subtracting 1 from the score of odd-numbered questions and 5 from the score of even-numbered questions. The sum is then multiplied by 2.5, resulting in a score range of 0-100.

To ensure fairness in the comparison, all baseline models were evaluated on the same dataset split. Whisper-Large and Whisper-Tiny were supervised end-to-end fine-tuned on the 1000-hour Chinese clinical speech training set constructed in this paper using their official pre-trained models, with a greedy decoding strategy. XLS-R (0.3B parameter version) was also fine-tuned using the same training set based on its multilingual pretrained weights. Google USM (General Speech Model) was evaluated based on its publicly available checkpoints pre-trained on medically relevant speech data, without additional fine-tuning, to test its out-of-the-box domain adaptability, with decoding using default beam search (beam size=5). The commercial system Abridge was called through its official API interface, inputting the same test set audio to obtain its generated clinical summary text, representing the current performance level of closed-source systems in the industry. The word error rate (WER) of all models was calculated using the same text normalization process.

V. RESULTS AND ANALYSIS

A. SPEECH RECOGNITION PERFORMANCE ANALYSIS

This paper compares Conformer-Transducer with DeepSpeech 2, Transformer-Transducer, XLS-R (Self-supervised Cross-lingual Speech Representation Learning), Whisper-Tiny, Whisper-Large, and Google USM (Universal Speech Model) to analyze speech recognition performance. The results are shown in Figure 3.

The Conformer-Transducer outperformed mainstream models in clinical speech recognition tasks, achieving a low WER of 6.3% and an MTA of 95.8%, demonstrating its superior acoustic modeling and semantic alignment capabilities in dense medical terminology scenarios. Compared to DeepSpeech 2 (WER 14.9%, MTA 74.2%), this model's advantage stems from its deeply optimized multi-stage integration architecture and precise capture of long-term context, which avoids the vanishing gradient problem of RNNs. Compared to the Transformer-Transducer (9.8%, 85.1%), the Conformer's convolutional modules enhance its ability to extract short-term acoustic features, effectively alleviating the problems of pronunciation compression and connected speech in medical terminology.

The data are grouped and analyzed by department category and doctor title, and the generalization ability of speech recognition is tested using WER. The results are shown in Figure 4.

From a departmental perspective, Conformer-Transducer achieved the lowest WER across all departments, demonstrating excellent cross-disciplinary generalization. The emergency department had the highest WER (7.2%), due to its rapid speech rate, fragmented expressions, and high background noise, which placed the highest demands on model robustness. Obstetrics and gynecology, cardiology, and gastroenterology performed the best (6.2%, 6.1%, and 6.0%, respectively), likely due to their relatively standardized terminology and moderate speech rates. Among the compared models, DeepSpeech 2 performed the worst across all departments (>14.4%), demonstrating that RNN

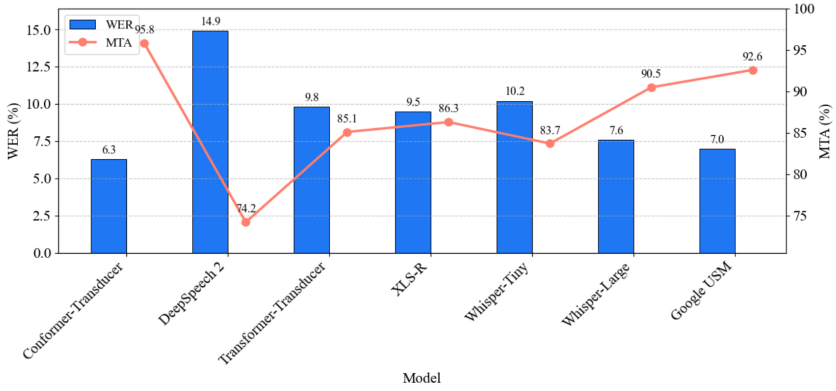

FIGURE 3. Speech recognition performance

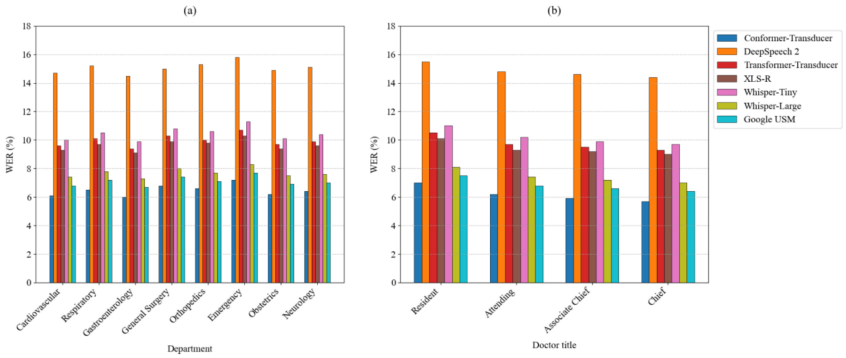

FIGURE 4. Generalization capability of speech recognition: (a) Department category; (b) Doctor title

architectures struggle to adapt to the high variability of medical speech. Whisper-Large and Google USM performed similarly, but Conformer maintained its lead, demonstrating that specialized architectures can outperform general-purpose large models with sufficient fine-tuning. The performance of each model declined consistently in the emergency department and general surgery, reflecting the recognition challenges in surgical and emergency scenarios. Conformer, leveraging its convolutional module's sensitivity to localized acoustic mutations, effectively mitigated recognition bias caused by differences in pronunciation patterns between different departmental terms, validating its robustness in multi-department deployments.

Different signal-to-noise ratio environments are set to analyze the recognition performance of the model in this paper. The results are shown in Table 2.

TABLE 2. Effect of signal-to-noise ratio on recognition performance

SNR (Signal-to-Noise Ratio) (dB)	WER (%)	MTA (%)
-5	15.6	82.3
0	12.1	86.7
5	9.8	89.4
10	7.5	93.1
15	6.6	94.7
20	6.1	95.3
30	6.0	95.6

The Conformer-Transducer demonstrates good noise robustness across varying SNR environments. Recognition performance steadily improves with increasing SNR, maintaining usable accuracy even under low SNR conditions. When the SNR is -5 dB (strong noise), the WER is 15.6% and the MTA is 82.3%, demonstrating that the convolutional module's ability to model local acoustic features effectively suppresses some noise interference. As the SNR increases to 10 dB, the WER drops rapidly to 7.5%, while the MTA reaches 93.1%, approaching the recognition performance achieved in real-world clinical quiet environments.

B. MEDICAL NER AND TERM NORMALIZATION RESULTS

This paper uses BERT-BiLSTM-CRF to automatically identify and classify "entities" from medical texts. The entity recognition performance is shown in Figure 5.

Experimental results show that the BERT-BiLSTM-CRF model used in this paper achieves optimal performance in the medical named entity recognition task, achieving an F1-score of 0.933, outperforming all competing models. Its Precision (94.3%) and Recall (92.4%) are both leading, demonstrating that the model achieves both high accuracy and coverage in entity boundary identification and category classification. In comparison, BioBERT (Biomedical BERT) (F1=0.912) and ClinicalBERT (0.916), while possessing strong semantic representation capabilities based on do-

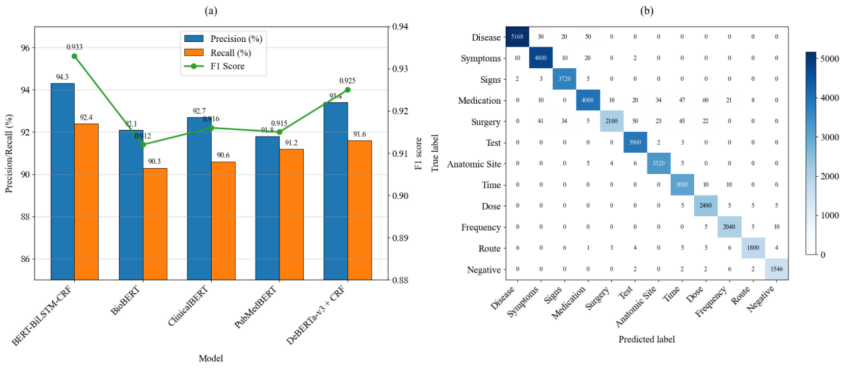

FIGURE 5. Entity recognition performance: (a) Comparison of different models; (b) Confusion matrix of BERT-BiLSTM-CRF

main pre-training, employ a direct Softmax classification head, which can lead to missed recognition or misclassification when handling nested entities and ambiguous terms. PubMedBERT performs similarly (F1=0.915), but lacks sequential dependency modeling and therefore underperforms in modeling the continuity of long entities. DeBERTa-v3 (Decoding-enhanced BERT with disentangled attention)+CRF improves to 0.925 by enhancing the attention mechanism, but still falls short of the proposed model. This suggests that the application of BiLSTM on top of BERT further captures the dynamic nature of word order, while the CRF layer effectively models label transition constraints, suppressing the output of illegal label sequences. This cascaded structure leverages the triplet of pre-trained semantics, sequence context, and label syntax, offering significant advantages in processing complex semantic combinations and contributing to its leading performance.

Terminology normalization maps colloquial expressions to standard medical concepts. The top-1 accuracy reflects the accuracy of the preferred match, while the top-3 accuracy reflects the error tolerance of the candidate. This ensures the correctness of non-standard expressions, improves semantic consistency, and enhances the reliability of subsequent clinical decisions. The results of terminology normalization are shown in Figure 6.

Note: The edit distance of this method is Levenshtein, and the edit distance is matched with the SimCSE semantic vector, using a weighted fusion strategy (0.4×edit similarity + 0.6×semantic similarity). The "edit distance + SimCSE semantic matching" method outperforms comparable methods in term normalization tasks, achieving a Top-1 accuracy of 88.9% and a Top-3 accuracy of 93.6%. Traditional methods performed poorly due to their inability to handle synonymous expressions and semantic generalization. MetaMap and ScispaCy rely on rules and shallow semantics, limiting their ability to handle nonstandard abbreviations and contextual ambiguity.

C. QUALITY ASSESSMENT OF INTELLIGENT MEDICAL RECORD GENERATION

The quality of intelligent medical record generation is evaluated using ROUGE-L, BLEU-4, and METEOR. The comparison results of medical record generation quality are shown in Figure 7.

The BART model comprehensively outperformed competing models in the intelligent medical record generation task, achieving a ROUGE-L score of 74.3%, surpassing Seq2Seq (65.4%) and BioGPT (Biomedical Generative Pre-trained Transformer) (70.9%). This demonstrates the highest degree of match between the generated text and the reference medical records in terms of n-gram overlap and longest common subsequence. Based on a bidirectional encoder-autoregressive decoder architecture, BART fully models the contextual semantics of the input speech text and generates structurally coherent SOAP records. The manual evaluation results are shown in Table 3.

TABLE 3. Manual evaluation results

Dimensions	This paper method	Attending physician	t value	p value
Content completeness	4.32 ± 0.58	4.51 ± 0.52	-2.76	0.006
Medical accuracy	4.28 ± 0.61	4.47 ± 0.55	-2.63	0.009
Structural standardization	4.41 ± 0.53	4.49 ± 0.50	-1.87	0.063
Language readability	4.15 ± 0.64	4.38 ± 0.57	-2.94	0.004

This manual assessment was conducted in a double-blind manner by 30 attending physicians with more than 5 years of clinical experience. Each physician assessed 50 records from the system-generated medical records and 50 records from a reference medical record written by another senior physician, for a total of 3000 records assessed (1500 from the system and 1500 from the reference). A 5-point Likert scale was used for scoring, and the final score was the mean and standard deviation of all physicians' scores. Table 3 shows the comparison of scores between the system-generated and reference medical records across various dimensions.

Manual evaluation results showed that the proposed method is close to the writing level of attending physicians in all dimensions, and the differences are clinically acceptable. In terms of content completeness and medical accuracy, the system scored 4.32 and 4.28, respectively. Although slightly lower than manual evaluation ($p < 0.01$), the gap is small, indicating that the system can effectively cover key clinical information and maintain medical cor-

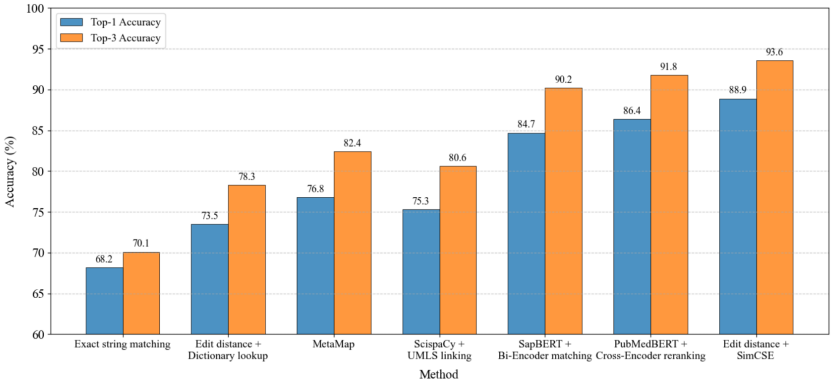

FIGURE 6. Term normalization results

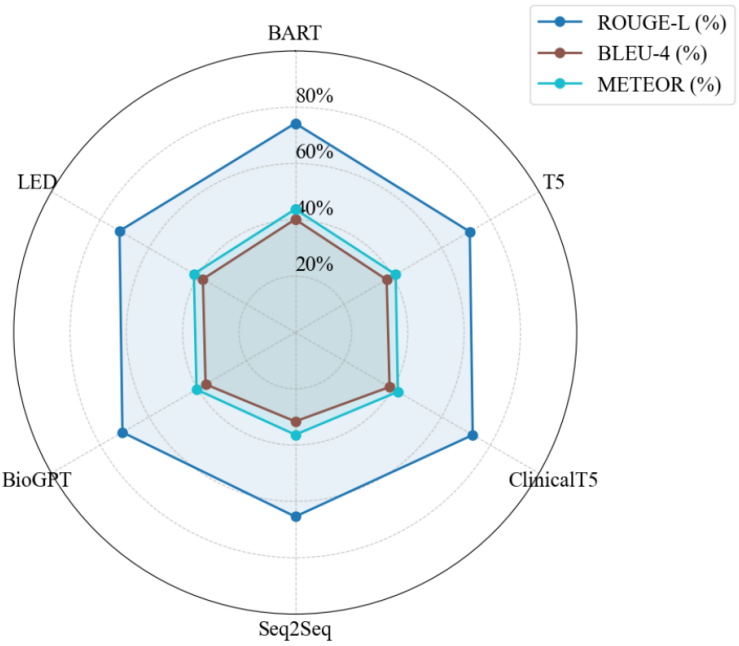

FIGURE 7. Quality of intelligent medical record generation

rectness. There is no significant difference in structural standardization ($p = 0.063$), indicating that the BART model successfully followed the SOAP template and the output format is standardized. The language readability score is relatively low (4.15), mainly due to the presence of a small amount of redundant expressions and terminology stacking in the text generated by the system, which affected reading fluency. Overall, the system has good usability while retaining medical rigor, especially in terms of structured expression, and can serve as an efficient auxiliary tool for clinical documents.

D. SYSTEM RESPONSE AND CLINICAL USABILITY

Abridge is a voice summarization system for doctor-patient conversations. It automatically generates clinical summaries after consultations, supports mobile devices, and emphasizes information extraction and readability. The system proposed in this paper is compared with the Abridge system, and the comparison results are shown in Figure 8.

Our system outperformed Abridge across all physician qualification groups, with a multi-stage integration latency of 940–960 ms, compared to Abridge's 1,280–1,320 ms. This advantage stems from our system's local deployment and lightweight model optimization: the Conformer-Transducer is accelerated through dynamic batching, and the BART generation model uses a caching and pruning strategy, significantly reducing inference time. Abridge, on the other hand, relies on the cloud for voice upload, processing, and return, and is affected by network round-trip latency, server queue scheduling, and multi-tenant resource competition, resulting in slower responses. Although the speaking speed and expression complexity of doctors with different qualifications vary slightly, the latency fluctuation is less than 20 ms, indicating that the system performance is

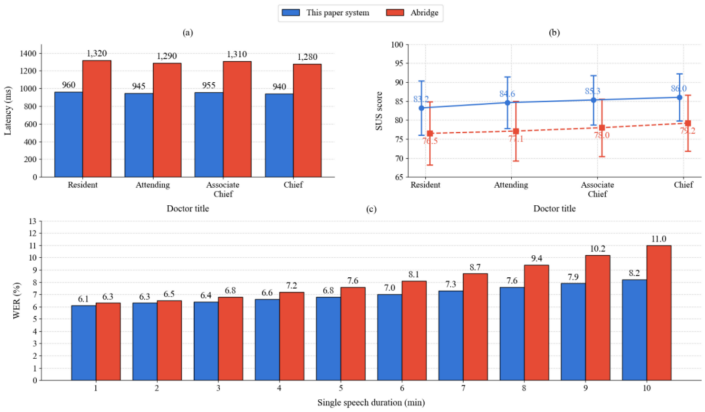

FIGURE 8. System comparison results: (a) Response delay; (b) SUS score; (c) Effect of single speech duration on speech recognition performance

stable and not significantly affected by user characteristics. The low latency feature supports real-time voice feedback, improving the fluency of physician use, and has a clear engineering advantage, especially in time-sensitive scenarios such as emergency rooms.

VI. CONCLUSION

This study presents a clinical speech-to-structured medical record system based on the Conformer architecture. The three-stage pipeline achieves efficient conversion through: 1) the Conformer-Transducer for accurate acoustic modeling of medical terminology, 2) terminology correction and entity recognition using BERT-BiLSTM-CRF for semantic understanding, and 3) the BART generator with SOAP templates and knowledge prompts for standardized record generation. The system achieves a 6.3% WER, accurately extracts 12 clinical entity types, and generates records comparable to physician-level quality. This work establishes a domain adaptation paradigm for medical speech recognition and provides a scalable technical framework for intelligent medical documentation systems.

The multi-stage integrated architecture employed in this study, while achieving modular decoupling and collaborative optimization, also carries the potential risk of error propagation along the processing chain. To mitigate the impact of errors in preceding modules on subsequent tasks, error correction and enhancement mechanisms were introduced at key interfaces during system design: a terminology spelling correction module based on edit distance and a medical dictionary was set up between ASR and NER; between NER and generation, UMLS normalization and semantic similarity matching were used to improve the robustness of entity mapping; finally, the BART generation model, with its powerful context modeling capabilities, was able to recover and generate coherent, structured text from noisy input to a certain extent. Nevertheless, error propagation remains a problem that requires continuous attention and optimization for such serial systems.

AUTHOR CONTRIBUTIONS

Conceptualization – Xiangkun Zou and Lei Wang; methodology – Xiangkun Zou, Lei Wang, Jing Sun, Shengyu Guo and Ning Li; investigation – Xiangkun Zou and Lei Wang; writing-original draft preparation – Xiangkun Zou, Lei Wang, Jing Sun, Shengyu Guo and Ning Li. All authors have read and agreed to the published version of the manuscript.

CONFLICTS OF INTEREST

The authors declare no conflict of interest.

FUNDING

This research was funded by S&T Program of Zhangjiakou. (No.2221151D).

ETHICS APPROVAL AND CONSENT TO PARTICIPATE

This study was approved by the Ethics Committee of The First Affiliated Hospital of Hebei North University, and was performed in accordance with the principles of the Declaration of Helsinki. All eligible participants signed an informed consent form.

ACKNOWLEDGEMENTS

Not applicable.

REFERENCES

[1] Z. T. Guo, W. L. Shi, and T. Yang, "Design and implementation of intelligent voice assistant for clinical diagnosis and treatment of traditional Chinese medicine based on speech recognition," Journal of Medical Informatics, vol. 43, no. 9, pp. 59-62, 2022, doi: 10.3969/j.issn.1673-6036.2022.09.01.

[2] J. R. Liu, Y. Song, and R. Jia, "Identification of protected health information in Chinese clinical texts based on BiLSTM-CRF," Data Analysis and Knowledge Discovery, vol. 4, no. 10, pp. 124-133, 2020, doi: 10.11925/infotech.2096-3467.2020.4.10.124.

[3] A. S. Miner, A. Haque, J. A. Fries, S. L. Fleming, D. E. Wilfley, G. Terence Wilson, et al., "Assessing the accuracy of automatic speech recognition for psychotherapy," NPJ Digital Medicine, vol. 3, no. 1, pp. 82, 2020, doi: 10.1038/s41746-020-0285-8.

[4] Y. Kumar, "A comprehensive analysis of speech recognition systems in healthcare: Current research challenges and future prospects," SN Computer Science, vol. 5, no. 1, pp. 137, 2024, doi: 10.1007/s42979-023-02466-w.

[5] L. Sun, A. N. Wang, Y. M. Song, J. Dong, X. L. Liu, H. Liang, et al., "Application, challenges and prospects of large language models in clinical medicine," Journal of PLA Medical College, vol. 46, no. 1, pp. 50-60, 2025, doi: 10.12435/j.issn.2095-5227.24070201.

[6] M. Jelassi, O. Jemai, and J. Demongeot, "Revolutionizing radiological analysis: The future of French language automatic speech recognition in healthcare," Diagnostics, vol. 14, no. 9, pp. 895, 2024, doi: 10.3390/diagnostics14090895.

[7] B. D. Tran, K. Latif, T. L. Reynolds, and J. Park, ""Mm-hm,""Uh-uh": Are non-lexical conversational sounds deal breakers for the ambient clinical documentation technology? Journal of the American Medical Informatics Association," 2023; 30(4):703-711, doi: 10.1093/jamia/ocad001.

[8] A. Dutta, G. Ashishkumar, and C. V. R. Rao, "Performance analysis of ASR system in hybrid DNN-HMM framework using a PWL euclidean activation function," Frontiers of Computer Science, vol. 15, no. 4, Art. no. 154705, 2021, doi: 10.1007/s11704-020-9419-z.

[9] C. Garcia, D. Nickolai, and L. Jones, "Traditional versus ASR-based pronunciation instruction," Calico Journal, vol. 37, no. 3, pp. 213-232, 2020, doi: 10.1558/cj.40379.

[10] A. M. Deshmukh, "Comparison of hidden markov model and recurrent neural network in automatic speech recognition," European Journal of Engineering and Technology Research, vol. 5, no. 8, pp. 958-965, 2020, doi: 10.24018/eng.2020.5.8.2077.

[11] F. Ye and J. Yang, "A deep neural network model for speaker identification," Applied Sciences, vol. 11, no. 8, pp. 3603, 2021, doi: 10.3390/app11083603.

[12] Z. Ren, N. Yolwas, W. Slamu, R. Cao, and H. Wang, "Improving hybrid ctc/attention architecture for agglutinative language speech recognition," Sensors, vol. 22, no. 19, pp. 7319, 2022, doi: 10.3390/s22197319.

[13] L. Tang, "A transformer-based network for speech recognition," International Journal of Speech Technology, vol. 26, no. 2, pp. 531-539, 2023, doi: 10.1007/s10772-023-10034-z.

[14] T. Guo, N. Yolwas, and W. Slamu, "Efficient conformer for agglutinative language ASR model using low-rank approximation and balanced softmax," Applied Sciences, vol. 13, no. 7, pp. 4642, 2023, doi: 10.3390/app13074642.

[15] P. Jiang, W. Pan, J. Zhang, T. Wang, and J. Huang, "A Robust Conformer-Based Speech Recognition Model for Mandarin Air Traffic Control," Computers, Materials & Continua, vol. 77, no. 1, pp. 911, 2023, doi: 10.32604/cmc.2023.041772.

[16] L. Li, Y. Long, D. Xu, and Y. Li, "Boosting Character-based Mandarin ASR via Chinese Pinyin Representation," International Journal of Speech Technology, vol. 26, no. 4, pp. 895-902, 2023, doi: 10.1007/s10772-023-10050-z.

[17] Y. Chen, W. H. Ge, and J. Liao, "Research progress of deep learning in named entity recognition in the medical field," PROGRESS IN PHARMACEUTICAL SCIENCES, vol. 44, no. 1, pp. 28-34, 2020.

[18] W. Zhang, C. Liu, H. B. Fei, W. Li, J. H. Yu, Y. Cao, et al., "Research on speech recognition based on DL-T and transfer learning," Journal of Engineering Science, vol. 43, no. 3, pp. 433-441, 2021.

[19] X. J. Che, H. Xu, M. Y. Pan, and Q. L. Liu, "Two-stage learning algorithm for biomedical named entity recognition," Journal of Jilin University (Engineering Edition), vol. 53, no. 8, pp. 2380-2387, 2023, doi: 10.13229/j.cnki.jdxbgxb.20211156.

[20] R. Mahum, I. Ganiyu, L. Hidri, A. M. El-Sherbeeny, and H. Hassan, "A novel Swin transformer based framework for speech recognition for dysarthria," Scientific Reports, vol. 15, no. 1, Art. no. 20070, 2025, doi: 10.1038/s41598-025-02042-7.

[21] X. Yang, A. Chen, N. PourNejatian, H. C. Shin, K. E. Smith, C. Parisien, et al., "A large language model for electronic health records," NPJ Digital Medicine, vol. 5, no. 1, pp. 194, 2022, doi: 10.1038/s41746-022-00742-2.

[22] J. Z. Yan, Y. X. He, Z. Y. Luo, H. Hu, S. X. Fan, B. Z. Tang, et al., "Potential typical applications and challenges of generative large language models in the medical field," Journal of Medical Informatics, vol. 44, no. 9, pp. 23-31, 2023, doi: 10.3969/j.issn.1673-6036.2023.09.003.

[23] Q. T. Chen, J. W. Ni, J. Xu, X. H. Gao, and L. Z. Xia, "Research on traditional Chinese medicine prescription generation driven by generative artificial intelligence GPT-4," China Pharmacy, vol. 34, no. 23, pp. 2825-2828, 2023, doi: 10.6039/j.issn.1001-0408.2023.23.02.

[24] Z. Y. Yang, S. F. Fang, X. C. Chen, and Z. N. Li, "Traditional Chinese medicine question generation based on multistrategy mechanism and BERT," Journal of Medical Informatics, vol. 43, no. 11, pp. 55-62, 2022, doi: 10.3778/j.issn.1673-9418.2401072.

[25] T. Mairittha, N. Mairittha, and S. Inoue, "Automatic labeled dialogue generation for nursing record systems," Journal of Personalized Medicine, vol. 10, no. 3, pp. 62, 2020, doi: 10.3390/jpm10030062.

[26] J. L. K. E. Fendji, D. C. M. Tala, B. O. Yenke, and M. Atemkeng, "Automatic speech recognition using limited vocabulary: A survey," Applied Artificial Intelligence, vol. 36, no. 1, Art. no. 2095039, 2022, doi: 10.1080/08839514.2022.2095039.

[27] C. Tejedor-Garcia, V. Cardenoso-Payo, and D. Escudero-Mancebo, "Automatic speech recognition (ASR) systems applied to pronunciation assessment of L2 Spanish for Japanese speakers," Applied Sciences, vol. 11, no. 15, pp. 6695, 2021, doi: 10.3390/app11156695.

[28] Q. Qin, S. Zhao, and C. Liu, "A BERT-BiGRU-CRF model for entity recognition of Chinese electronic medical records," Complexity, vol. 2021, Art. no. 6631837, 2021, doi: 10.1155/2021/6631837.

[29] X. Tang, Y. Huang, M. Xia, and C. Long, "A multi-task BERT-BiLSTM-AM-CRF strategy for Chinese named entity recognition," Neural Processing Letters, vol. 55, no. 2, pp. 1209-1229, 2023, doi: 10.1007/s11063-022-10933-3.

[30] C. X. Li, D. Li, X. Ling, X. L. Li, J. F. Tang, Y. Zhao, et al., "Discussion on the content and recording standards of the teaching history of Chinese medicine clinical pharmacists," Chinese Pharmacy, vol. 33, no. 21, pp. 2671-2675, 2023, doi: 10.6039/j.issn.1001-0408.2022.21.20.

[31] J. Y. Zhai, Y. Lu, S. L. Qian, and D. H. Yu, "Course design of clinical diagnosis and treatment thinking for general practice master's students of Tongji University," Chinese General Practice, vol. 26, no. 25, pp. 3202, 2023, doi: 10.12114/j.issn.1007-9572.2022.071.

[32] W. Yim, Y. Fu, A. Ben Abacha, N. Snider, T. Lin, M. Yetisgen, et al., "Aci-bench: A novel ambient clinical intelligence dataset for benchmarking automatic visit note generation," Scientific Data, vol. 10, no. 1, pp. 586, 2023, doi: 10.1038/s41597-023-02487-3.

[33] J. Liang, L. P. Zhang, S. Yan, Y. B. Zhao, and Y. W. Zhang, "Research progress on named entity recognition based on large language model," Journal of Frontiers of Computer Science & Technology, vol. 18, no. 10, pp. 2594, 2024, doi: 10.3778/j.issn.1673-9418.2407038.

[34] A. Dhouib, A. Othman, O. El Ghoul, M. K. Khribi, and A. Al Sinani, "Arabic automatic speech recognition: A systematic literature review," Applied Sciences, vol. 12, no. 17, pp. 8898, 2022, doi: 10.3390/app12178898.

[35] T. T. N. Ngo, H. H. J. Chen, and K. K. W. Lai, "The effectiveness of automatic speech recognition in ESL/EFL pronunciation: A meta-analysis," ReCALL, vol. 36, no. 1, pp. 4-21, 2024, doi: 10.1017/s0958344023000113.