UNIVERSITY OF TECHNOLOGY SYDNEY
Faculty of Engineering and Information Technology

Automatic Quantitative Stroke Severity Assessment from Chinese Electronic Health Records based on Advanced Large Language Models

by

Zhanzhong Gu

A Thesis Submitted
in Fulfillment of the
Requirements for the Degree

Doctor of Philosophy

Sydney, Australia

2024

CERTIFICATE OF ORIGINAL AUTHORSHIP

I, Zhanzhong Gu, declare that this thesis, is submitted in fulfilment of the requirements for the award of Doctor of Philosophy, in the School of Electrical and Data Engineering, Faculty of Engineering and Information Technology at the University of Technology Sydney.

This thesis is wholly my own work unless otherwise referenced or acknowledged. In addition, I certify that all information sources and literature used are indicated in the thesis.

This document has not been submitted for qualifications at any other academic institution.

This research is supported by the Australian Government Research Training Program.

Signature: Production Note:
Signature removed prior to publication.

Date: 26 July 2024

ABSTRACT

Stroke is one of the leading causes of death worldwide. Accurate assessment of stroke severity plays a pivotal role in precise diagnosis, development of treatment plans, and efficient allocation of healthcare resources. The assessment of stroke in hospitals is usually conducted manually by clinicians. However, it is labor-intensive, time-consuming, and sometimes unreliable.

With the continuous development of artificial intelligence (AI) techniques in recent years, applying them to automate clinical assessment in EHRs has attracted much interest. In this thesis, we outline an innovation pathway to advance stroke healthcare from ontology construction, clinical named entity recognition (CNER) with pre-training, to LLM-driven automatic quantitative stroke assessment.

The journey begins with the development of "StrokePEO", a stroke clinical ontology co-designed with specialists using advanced natural language processing (NLP) and deep learning techniques. StrokePEO successfully represents clinical terms and relationships in stroke assessment, demonstrating applicability in diverse medical contexts.

Building on this foundation, we develop a deep learning-based framework to automatically assess stroke severity through Chinese CNER and domain-adaptive pre-training. We first construct a new dataset "Chinese Stroke Clinical Records" (CSCR) and pre-train a Chinese clinical embedding "CliRoberta" for CNER. Then, a dictionary-based mapping method is developed to map CNER results into quantitative scores. Comprehensive experiments demonstrate the effectiveness and reliability of the CNER model with our domain-adaptive pre-training. Ultimately, our automatic NIHSS scoring approach achieves excellent inter-rater agreement and intra-class consistency with the ground truth, with significantly improved efficiency.

We further advance toward cutting-edge LLMs with a prompting paradigm "GAPrompt" to empower the generic LLMs to assess diagnostic notes and generate

quantitative evaluation results. GAPrompt assesses the suitability of LLMs for specific tasks through prompting for LLM selection, facilitates their comprehension of task-specific knowledge derived from the constructed knowledge base, enhances the accuracy of knowledge retrieval and demonstration through summary-based generation-augmented retrieval (SGAR), improves LLM inference precision via hierarchical chain-of-thought (HCoT), strengthens generation robustness, and mitigates LLM hallucinations through ensembling. Experimental findings underscore the effectiveness of our approach in empowering LLMs to achieve automated stroke assessment based on EHRs.

Collectively, these works contribute integrative and innovative AI-driven solutions for stroke healthcare, shifting from traditional methods to state-of-the-art LLM techniques, addressing knowledge representation, automated assessment, and quantitative analysis, with broad applications in medical research and practice.

Dissertation directed by A/Prof. Wenjing Jia and Prof Massimo Piccardi
School of Electrical and Data Engineering

Dedication

I would like to dedicate this thesis to my dearest beloved wife, June Cao, and our adorable little angel baby Aaron Gu. I am deeply grateful to my wife for her unwavering support over the past decade and her selfless dedication during my pursuit of a Ph.D. degree. Her boundless encouragement, support, and love have been instrumental in helping me overcome numerous challenges and ultimately complete this PhD thesis. I am also immensely thankful to my mother, FenFen Gao, and my late father, Haijun Gu, for their toil in raising me into adulthood, nurturing a healthy physique, and fostering a sound character. I deeply appreciate their understanding and support in my academic pursuits.

To all of you, my heartfelt love and blessings.

Acknowledgements

First and foremost, I would like to acknowledge my supervisors A/Prof. Wenjing Jia and Prof. Massimo Piccardi for their continuous and endless supervision and encouragements. Furthermore, I would like to thank Prof. Xiangjian He, Prof. Ping Yu, Prof. Yiguang Lin, Dr Gengfa Fang for their huge support and guidance during the early stage of my PhD candidature.

I would like to thank my lab mates, Dr Jiachen Kang, Xiguang Yang, Xudong Song, Jiaoyang Ma, Dr. Xiaochen Fan, Dr. Yue Xi, Dr. Ye Huang, Dr. Qingqing Wang, Dr. Chenpei Xu and Dr Yuanfang Zhang for their assistance and encouragement.

Finally, I am particularly grateful to all people who directly or indirectly provided supports on my research and technical issues, and the authors of the papers that I cited.

Zhanzhong Gu

Sydney, Australia, 2024.

List of Publications

1. **Zhanzhong Gu**, Xiguang Yang, Wenjing Jia, Chengpei Xu, Ping Yu, Xiangjian He, Hongjie Chen and Yiguang Lin, "StrokePEO: Construction of a Clinical Ontology for Physical Examination of Stroke", Proceedings of 2022 9th International Conference on Digital Home (ICDH), 2022.

2. **Zhanzhong Gu**, Xiangjian He, Ping Yu, Wenjing Jia, Yiguang Lin, Gang Peng, Penghui Hu, Shiyan Chen, Hongjie Chen and Yiguang Lin, "Automatic Quantitative Stroke Severity Assessment based on Chinese Clinical Named Entity Recognition with Domain-Adaptive Pre-trained Large Language Model", *under minor revision*, Artificial Intelligence in Medicine, vol 50, 2024.

3. **Zhanzhong Gu**, Wenjing Jia, Ping Yu, "Empowered Large Language Models for Automated Quantitative Assessment of Electronic Health Records through Retrieval-Augmented Generation and Hierarchical Chain-of-Thought", *under review*, Artificial Intelligence in Medicine, 2024.

4. **Zhanzhong Gu**, Xiangjian He, Gengfa Fang, Chengpei Xu, Feng Xia, Wenjing Jia, Millimeter Wave Radar-based Human Activity Recognition for Healthcare Monitoring Robot, IEEE Transactions on Cognitive and Development Systems, *under review*.

Contents

Abstract iii

Dedication v

Acknowledgments vi

List of Publications vii

List of Figures xii

List of Tables xiv

List of Abbreviations xvi

1 Introduction 1

1.1 Automatic Quantitative Stroke Severity Assessment 1

1.2 Key Challenges of Stroke Severity Assessment and Core Scientific Problems 3

1.2.1 Lack of Ontologies for Physical Examination of Stroke 3

1.2.2 Limitation of CNER Models and Pre-trained Embeddings . . 4

1.2.3 Limitations of Automated NIHSS Scoring 5

1.2.4 Applying LLMs in Clinical Domain and Its Limitations 6

1.3 Contributions in This Thesis 7

1.3.1 StrokePEO: Construction of a Clinical Ontology for Physical Examination of Stroke 7

1.3.2 Automatic Quantitative Stroke Severity Assessment based on Chinese Clinical Named Entity Recognition with Domain-Adaptive Pre-trained Large Language Model 8

1.3.3 Empowering LLMs for Automated Quantitative Assessment of EHRs through Retrieval-Augmented Generation and Hierarchical Chain-of-Thought 10

1.4 Organization of This Thesis 11

2 Literature Review 13

2.1 Automatic Quantitative Stroke Assessment 13

2.2 Ontology Construction 14

2.2.1 Existing Stroke Ontologies 14

2.2.2 Text Preprocessing 15

2.2.3 Term and Relationship Extraction 16

2.2.4 Ontology Integration 18

2.3 CNER and Pre-trained Embeddings 18

2.3.1 Existing CNER Models 19

2.3.2 Pre-trained Embeddings 20

2.4 LLMs and Prompting Strategies 21

2.4.1 Large Language Models (LLMs) 21

2.4.2 Retrieval Augmented Generation (RAG) 22

2.4.3 Prompt Engineering 23

3 StrokePEO: Construction of a Clinical Ontology for Physical Examination of Stroke 26

3.1 Background 27

3.2 Methodology 28

3.2.1 Text Preprocessing 29

3.2.2 Ontology Schema Definition 30

3.2.3 Joint Term and Relationship Extraction 31
3.2.4 Term Alignment 32
3.2.5 Ontology Integration 33
3.3 Experiments 34
3.3.1 Dataset 34
3.3.2 Relationship Classification Results 35
3.3.3 Term Extraction Results 36
3.4 Summary 37

4 Automatic Quantitative Stroke Severity Assessment based on Chinese Clinical Named Entity Recognition with Domain-Adaptive Pre-training 39

4.1 Background 40
4.2 The Proposed Method 42
4.2.1 Construction of CSCR Dataset and Entity-to-NIHSS Mapping 42
4.2.2 Chinese CNER with Domain-Specific Pre-trained Embedding . 47
4.2.3 Automated NIHSS Scoring 51
4.3 Experiment Settings and Results 58
4.3.1 Datasets 60
4.3.2 Evaluation Metrics 61
4.3.3 Evaluation of CNER with Pre-trained Embedding 63
4.3.4 Results of Automated NIHSS Scoring 66
4.4 Discussion 71
4.4.1 Principal Findings 71
4.4.2 Advancements and Limitations 72

4.5 Summary 72

5 Empowering LLMs for Automated Clinical Assessment using EHRs 74

5.1 Background 75

5.2 Method 78

5.2.1 Prompt-driven LLM Selection 78

5.2.2 Generation-augmented Knowledge Base Construction 80

5.2.3 Summary-based Generation-augmented Retrieval 82

5.2.4 Hierarchical Chain-of-Thought 83

5.3 Experiment Design 86

5.3.1 Datasets 86

5.3.2 Evaluation Metrics 90

5.4 Results 90

5.4.1 LLM Selection 90

5.4.2 Generation-augmented Retrieval 91

5.4.3 Results of Micro Chain-of-Thought Learning 92

5.4.4 Macro Sequential Chain Results 93

5.4.5 Ablation Studies on the Effectiveness of GAPrompt Components 95

5.5 Summary 97

6 Conclusion and Future Work 98

6.1 Conclusion 98

6.2 Future Work 100

Bibliography 102

List of Figures

3.1 The pipeline of constructing StrokePEO. It illustrates the process from hospital-provided EHRs to an expert-verified ontology using advanced NLP techniques. 29

3.2 The schematic representation of our StrokePEO. 31

4.1 The procedure of the proposed automatic quantitative stroke severity assessment framework. Green color: constructing the CSCR dataset; yellow color: Constructing entity-to-NIHSS mapping dictionary . . . 43

4.2 The architecture of our baseline CNER model with pre-trained embedding. For the convenience of non-Chinese readers, the input sentence in Chinese is accompanied with a word-by-word gloss in English. 48

4.3 The semantic dependency schema and priority order. The schema consists of seven types of entities and twelve relationships. The priority order defines the selection rule during the extraction of relational entity triples. 52

4.4 The algorithm for rational entity triples (RET) extraction. 54

4.5 Examples of RECs constructed from RETs and entities. We show three REC examples in the figure, and each of them is constructed from a different number of RETs and entities. 56

4.6 The procedure of our proposed automated NIHSS scoring pipeline. The left side shows the steps from loading the EHR to capturing the NIHSS scoring result. The right side shows the detailed processing stages including raw EHR tokenization, entities extraction by CNER model with embedding, construction of RETs and RECs, synonyms regulation, dictionary-based mapping, and the calculation of NIHSS score as the final result. 57

4.7 The Algorithm for Entity-to-NIHSS Mapping. 59

4.8 The performance of three CNER models with five different pre-trained embeddings on the CSCR dataset (the top row) and the CCKS2019 dataset (the bottom row). 63

5.1 The architecture of our proposed GAPrompt paradigm. Green color: prompt-driven LLM selection; blue color: generation-augmented knowledge base construction; orange color: summary-based generation-augmented retrieval (SGAR); pink color: hierarchical chain-of-thought (HCoT); purple color: ensembling. 79

5.2 The six prompt templates applied to select the optimal foundation LLM. Six capabilities of LLMs, including Knowledge, Comprehension, Learning, Reasoning, Consistency, and Anti-hallucination, are evaluated using these defined prompts. 81

5.3 The template used by LLMs to generate demonstrations. 82

5.4 The prompt template used for LLMs to generate summarization. . . 83

5.5 The macro sequential chain. The macro sequential chain includes five steps: splitting, translation, retrieval, micro CoT, and ensembling . . 84

5.6 The macro sequential chain. It includes the steps of translation, splitting, retrieval, micro CoT, and ensembling. 86

5.7 The ablation study on the effectiveness of GAPrompt components. . 95

List of Tables

3.1 The statistics of the annotated terms and relationships. 35

3.2 Relationship classification results 36

3.3 Term extraction results 37

4.1 The entities, definitions, count, and percentage of occurrence in the constructed CSCR dataset. 45

4.2 The pre-training corpora of language models. The training tokens are counted by Chinese characters. 49

4.3 The division of two datasets in the experiments. The numbers refer to the count of EHRs. 61

4.4 The performance of the CNER models with pre-trained embeddings on CCKS2019 and CSCR dataset, by F1 score. 64

4.5 The results of the statistical significance test of our pre-trained CliRoberta with four existing pre-trained embeddings. The p_value from paired sample $t-test$ are reported in the table. 65

4.6 Descriptive statistics of NIHSS scoring results by assessors and the ground truth generated by specialists. 67

4.7 Evaluation of NIHSS scoring results by Kappa and ICC metrics for different assessors. The "Time" is the average time taken for scoring one EHR. The "CI95%_low/up" represent the lower and upper limit of the 95% confidence interval, respectively. 69

4.8 The variations in the ICC values between patients with mild and severe stroke severity. The "CI95%_low/up" represent the lower or upper limits of the 95% confidence interval, respectively. 70

5.1 The distribution of the generation-augmented knowledge base. Both the task-specific knowledge and the LLM-generated Demonstrations are reported, along with the count of samples related to each NIHSS component and their corresponding percentages (%). 88

5.2 The distribution of the quantitative assessment dataset. Both micro and macro-level ground truth of each NIHSS component and their corresponding percentage (%) are reported. 89

5.3 The performance of candidate LLMs with six prompting templates, using EM evaluation metrics. 91

5.4 Top-k retrieval accuracy (%) on both task-specific knowledge and the demonstrations. 92

5.5 The micro CoT results (F1 score). 93

5.6 The result of the macro sequential chain. 94

List of Abbreviations

- AI: Artificial intelligence
- BERT: Bidirectional encoder representations from transformers
- BiGRU: Bidirectional gated recurrent unit
- BiLSTM: Bidirectional long short-term memory
- CNER: Clinical named entity recognition
- CNN: Convolutional neural network
- CoT: Chain-of-thought
- CRF: Conditional random field
- CSCR: Chinese stroke clinical records
- EHR: Electronic health record
- GAR: Generation-augmented Retrieval
- GRU: Gated recurrent unit
- ICC: Intraclass correlation coefficient
- KIE: Key inspection entity.
- LLM: Large language model
- LSTM: Long short-term memory
- NIHSS: National institutes of health stroke scale
- NLI: Natural language inference

- NLP: Natural language processing
- NLU: Natural language understanding
- RAG: Retrieval-augmented generation
- REC: Relational entity chain
- RET: Relational entity triple
- SOTA: State-of-the-art

Chapter 1

Introduction

Stroke is a prevalent disease with a significant global impact. Effective assessment of stroke severity is vital for an accurate diagnosis, appropriate treatment, and optimal clinical outcomes. The National Institutes of Health Stroke Scale (NIHSS) [17] is a widely used scale for quantitatively assessing stroke severity. However, the current manual scoring of NIHSS is labor-intensive, time-consuming, and sometimes unreliable. Applying artificial intelligence (AI) techniques to automate the quantitative assessment of stroke on vast amounts of electronic health records (EHRs) has attracted much interest.

This thesis describes our attempts in applying advanced AI techniques for automatic, quantitative stroke severity assessment. Our research pipeline evolves through firstly constructing a stroke ontology, followed by automating the entire NIHSS scoring process on Chinese clinical named entity recognition (CNER) with a domain-adaptive pre-trained CliRoberta embedding, and finally empowering large language models (LLMs) with a novel prompting paradigm.

1.1 Automatic Quantitative Stroke Severity Assessment

Clinically, stroke patients are subjected to specialized tests to characterize the severity of their conditions at admission. When assessing treatment efficacy and care quality, clinicians need to retrospectively analyze patients' previous conditions recorded in EHRs to assess changes in their condition, either improved or deteriorated. Measurement scales (abbreviated as scales) play an important role in this assessment

process. They can be used to score diagnostic records with precise numerical values, intuitively and accurately reflecting the patient's condition in all aspects, and reach a comprehensive understanding of the level of severity.

There are more than 70 scales to evaluate stroke severity from different perspectives [42]. Among them, the NIHSS [17] is a widely accepted, clinically validated assessment tool for clinicians to mark the stroke patients' prognosis and disability level [63]. Especially, it is proven to be reliable and valid for retrospective scoring using data from the existing health records [136]. However, NIHSS scoring takes time and experience. It requires an experienced neurologist to spend at least 5 minutes to complete, and even longer for junior clinicians.

Previous researchers have proposed various automatic algorithms to replace the time-consuming, labor-intensive and unreliable manual assessment method [63,98,144]. However, they either only recognize NIHSS scores that are already recorded and reported in the EHRs, or require external equipment and data to make predictions. To date, there is no existing approach available that achieves automatic, quantitative stroke severity assessment directly from the patients' diagnostic notes in EHRs.

Therefore, it is necessary to develop a novel automatic assessment method using data captured in EHRs to improve the accuracy and efficacy of stroke severity assessment. This motivates us to undertake a series of research to achieve this goal step by step, including constructing a stroke ontology, building a stroke-specific dataset, constructing CNER model with domain-adaptive pre-trained embeddings, developing a dictionary-based automated NIHSS scoring approach, and advancing the foundation LLMs to facilitate the automated quantitative stroke assessment through a novel prompting paradigm.

1.2 Key Challenges of Stroke Severity Assessment and Core Scientific Problems

In this section, we identify the key challenges and core scientific problems on the automatic quantitative stroke severity assessment.

1.2.1 Lack of Ontologies for Physical Examination of Stroke

Clinical ontology is a standardized medical knowledge representation model that facilitates the integration and analysis of a large amount of heterogeneous EHR data. Using ontologies to represent clinical terms can improve data integration to build robust and interoperable medical information systems.

In the medical domain, ontology has successfully supported many important application scenarios, including precision medicine [43, 93], clinical decision support systems [7, 35], recommender systems [55, 89]. Using ontology to represent clinical terms can standardise data and enable data integration. Thus, ontologies have been applied to build robust and interoperable medical information systems, meeting the needs of reusing, sharing, and transmitting medical data, and provide statistical aggregation based on various semantic standards [109].

There are many challenges that are yet to be resolved in clinical ontology research. Among them, insufficient disease coverage, *i.e.*, a lack of high-quality annotated databases for certain diseases, *e.g.*, stroke, remains the biggest obstacle to the advancement of research and applications of clinical ontologies [109]. Literature [6, 16, 41, 68, 84, 125] suggests that none of the publicly available stroke ontologies have modeled the information related to physical examination of stroke.

1.2.2 Limitation of CNER Models and Pre-trained Embeddings

In recent years, there have been major breakthroughs in the application of successful CNER techniques to process natural languages in Chinese EHRs. The achievements include the application of BERT and its variants [32, 80, 121, 122, 145], *i.e.*, BiLSTM-CRF [77, 78, 155, 156, 159], BIGRU-CRF and CNN-LSTM-CRF [58, 100, 152]. However, there are notable gaps in applying high-accurate CNER to EHRs. Training a CNER model requires a considerable amount of labeled data. However, the availability of annotated datasets in the Chinese language is severely limited [22, 47, 150, 151], and there is currently no annotated dataset specifically tailored for stroke assessment. Moreover, the sparsely annotated entity classes in previous studies cannot capture all the valuable information needed for quantitative stroke severity assessment using NIHSS [17].

Pre-trained Embeddings: A word embedding is a vector representation that encodes the meaning of words for text analysis. It has been shown to improve the performance of NLP tasks including CNER [86, 101, 102, 104]. Since 2018, BERT-based embeddings have become the mainstream of text representation because their performance has surpassed the earlier traditional text vector representations [32]. Specifically, for Chinese text representation that plays a critical role in CNER, many Chinese-based BERT and variant models have been developed in recent years. For example, Cui *et al.* [31] proposed several Chinese BERT variants including MacBERT, Chinese ELECTRA, Chinese XLNet, and Chinese-Roberta.

With this increasing interest, research advancements have been made in the Chinese biomedical field that implements BERT-based embeddings, including MC-BERT [151], FT-BERT [77], EMBERT [19], and SMedBERT [154]. However, most of these embeddings are not publicly available, limiting their practical applicability in clinical research. Also, a large amount of training data was crawled from the

Internet, mixed with a lot of general languages rather than pure clinical language, resulting in no guaranteed performance for stroke-specific EHRs [19, 77, 154, 155].

1.2.3 Limitations of Automated NIHSS Scoring

With the advancement of AI techniques, an increasing number of researchers have endeavored to explore its potential in mining valuable insights from the vast realm of clinical EHRs for quantitative stroke research. Zhang *et al.* [63] introduced an automated stroke severity prediction model based on machine learning techniques. Their model takes hospital service parameters as input variables, such as discharge information, length of stay and mortality risk to estimate the overall severity level. Compared with NIHSS scoring that directly examines the diagnostic symptoms exhibited by patients, the estimated overall stroke severity level is not as reliable and accurate as NIHSS scoring that directly from diagnostic symptoms exhibited by patients. Park *et al.* [98] used signals acquired from special sensors to automatically grade the motor level of stroke patients. This method only covers a small portion instead of a full set of the quantified assessment items in NIHSS and therefore cannot provide a comprehensive stroke severity assessment. Another constraint is that it requires the use of additional external equipment.

Recently, Yang *et al.* [144] attempted to develop an automatic approach to identify the NIHSS items and scores reported in EHRs. This method has a restricted application scenario as it necessitates fully recorded NIHSS scores in EHRs, which can only be fulfilled by less than 5% of all clinical stroke EHRs. Therefore, it is useful to further develop automated quantification method for improving frequency and accuracy of stroke severity assessment.

1.2.4 Applying LLMs in Clinical Domain and Its Limitations

Recently, LLMs have demonstrated remarkable ability in natural language understanding (NLU) and natural language inference (NLI) [104, 105]. Unlike traditional approaches, LLMs can comprehend questions and provide answers directly from the given text, without the need for sentence-by-sentence or word-by-word processing and annotation [18]. This surpasses the capabilities of classical machine learning and deep learning methods in complex text-understanding tasks. Therefore, leveraging advanced LLMs to enhance the analysis of EHRs, specifically in the context of quantitative assessment of a patient's condition, holds great promise.

Despite the impressive NLU capabilities of LLMs, their direct applicability in real-world domain-specific scenarios is not without challenges [70, 113]. Most LLMs are trained on general language data and lack proficiency in understanding domain-specific text, such as medical EHRs [127]. Additionally, the very few existing medical domain LLMs are often proprietary and not publicly available [65, 117, 118], and they focus primarily on question-answering tasks while lacking robust quantitative assessment capabilities.

Considering the time and computational resources required to train a domain-specific LLM, it is often impractical for academic researchers and clinical practitioners. Consequently, the most promising approach is to leverage the power of foundation LLMs while enhancing their capability through prompting strategies. The retrieval-augmented generation (RAG) and prompt engineering techniques are thus introduced [73]. RAG is a strategy that combines information retrieval and LLM generation and has been proven to be capable of enhancing the quality and relevance of generated content by incorporating task-specific information retrieved from the external knowledge base [73]. Prompt engineering refers to the process of designing and refining the input queries or instructions given to LLMs, optimizing

the performance of the LLM for a specific task. Previous studies indicate that prompt engineering can significantly enhance the performance of foundation LLMs, particularly in domain-adaptive scenarios [36, 131, 143]. In this study, we propose a prompting paradigm for automated assessment of EHRs based on LLMs. It can automatically assess diagnostic notes in EHRs and provide quantitative assessment results based on the generation-augmented knowledge base.

1.3 Contributions in This Thesis

In this section, we present an overview of the contributions of this thesis.

1.3.1 StrokePEO: Construction of a Clinical Ontology for Physical Examination of Stroke

Clinical ontology serves as a standardized model for representing medical knowledge, facilitating the integration and analysis of diverse EHR data. Utilizing ontologies to depict clinical terms enhances data integration, contributing to the development of robust and interoperable medical information systems. To date, there exists no ontology specifically designed to represent medical knowledge related to the physical examination of stroke. This absence has hindered stroke physicians from fully leveraging clinical information within EHR data to comprehend the health status of stroke patients and devise effective medication and rehabilitation strategies.

In this thesis, we collaboratively design a stroke clinical ontology, "StrokePEO," with two stroke clinical specialists. Leveraging advanced natural language processing and deep learning techniques, we extract terms and their relationships from actual clinical EHRs provided by a tertiary hospital in China. Our experimental results demonstrate that our methods and the output of StrokePEO hold applicability in diverse medical contexts where the extraction of medical knowledge from EHRs is crucial for decision-making.

The contributions of this chapter are as follows.

- We contribute a clinical ontology StrokePEO dedicated to stroke physical examination. StrokePEO provides an essential component for the construction of large stroke knowledge graph, complements the mainstream stroke ontology research and facilitates the development of AI-based diagnosis and recommendation systems.

- We contribute methods and approaches for engaging the domain experts - clinical specialists - into co-designing the ontology StrokePEO, and various advanced natural language processing (NLP) and deep learning techniques to extract the terms and relationships from raw clinical record data to construct the StrokePEO.

- We integrate StrokePEO with globally recognized stroke ontologies, *e.g.* Stroke Ontology (STO) [41] and National Institutes of Health Stroke Scale Ontology (NIHSS) [16].

1.3.2 Automatic Quantitative Stroke Severity Assessment based on Chinese Clinical Named Entity Recognition with Domain-Adaptive Pre-trained Large Language Model

In this chapter, we apply various NLP technologies from multiple perspectives attempting to address the challenges. First, to tackle the problem of lacking a labeled stroke-specific data set, we collaborate with stroke clinicians from three top hospitals in China, collect and construct a disease-specific dataset named Chinese Stroke Clinical Records (CSCR). Then, we generate a Chinese clinical word embedding model through domain-adaptive pre-training of the open-source, large amount of clinical EHRs. We validate the high performance of our CNER model through the comprehensive evaluation of its performance against multiple SOTA deep neu-

ral networks. Finally, we develop a dictionary-based NIHSS mapping method to automatically assess stroke severity levels using the learned CNER.

This chapter has made the following three key contributions:

- For clinical stroke research, we construct a CNER dataset named CSCR, based on which fine-grained Chinese clinical entities are annotated. Different from most existing CNER datasets, entities in our CSCR dataset are semantically associated, intensively annotated and expert-verified. It allows excavating as much valuable information as possible from the EHRs for stroke severity assessment.

- We propose a Chinese clinical embedding "CliRoberta" through domain-adaptive pre-training to boost the performance of the CNER model. Experiment results on a public dataset and our CSCR dataset both demonstrate that our pre- trained "CliRoberta" has the best performance compared with the existing embeddings.

- We develop an automatic stroke severity assessment method based on the CNER model trained on the CSCR dataset. Through extracting relational entity triples, developing relational entity chains, and constructing scoring dictionaries and dictionary-based NIHSS mapping, we demonstrate a successful practice of applying NLP techniques for automatic stroke severity assessment. The effectiveness of the proposed method is proved by achieving an excellent reliability of 82.69% Kappa agreement and 0.9907 intra-class consistency coefficient in NIHSS scoring with the golden standard benchmark established by stroke specialists. This has far exceeded the accuracy of less experienced clinicians with significantly reduced task time.

1.3.3 Empowering LLMs for Automated Quantitative Assessment of EHRs through Retrieval-Augmented Generation and Hierarchical Chain-of-Thought

Understanding and extracting valuable information from EHRs holds significant importance in the medical domain, benefiting both clinical practice and medical research. The emerging LLMs have shown promise in natural language understanding (NLU) and inference tasks, making them suitable for automating the often labor-intensive, time-consuming, and tedious analyzing task in EHRs. However, due to the scarcity of publicly available medical LLMs and the complexity of domain-specific fine-tuning, designing appropriate prompting strategies to advance the capacity of foundation LLMs is a promising solution.

This chapter proposes a prompting paradigm for automated analysis of EHRs using foundation LLMs. By leveraging the few-shot in-context learning (ICL) abilities of LLMs, our proposed prompting paradigm enhances the power of foundation LLM through GAR and HCoT prompting, overcoming the limitations of foundation LLM in analyzing domain-specific medical text.

The key contributions of this chapter are as follows:

- We introduce a prompt-driven LLM selection process that effectively and efficiently selects the best foundation LLM for the current task.
- We develop a novel generation-augmented retrieval (GAR) method to dynamically retrieve task-specific knowledge and demonstrations from the self-constructed, generation-augmented knowledge base. Our proposed RAG
- We propose a hierarchical chain-of-thought (HCoT) prompting strategy to integrate the macro sequential chain with the micro chain-of-thought. Experiment results demonstrate the capability of our method to automatically assess EHRs

and generate quantitative results with HCoT prompting.

1.4 Organization of This Thesis

The rest of this thesis is organized as follows:

1. *Chapter 2:* This chapter reviews the related works of automated stroke assessment techniques, including the existing methods for quantitative stroke assessment, ontology construction techniques, current CNER models, pre-trained language models, and state-of-the-art LLMs and prompting strategies.

2. *Chapter 3:* This chapter designs with two stroke clinical specialists a stroke clinical ontology "StrokePEO" using advanced natural language processing and deep learning techniques to extract terms and their relationships from real clinical case records provided by a tertiary hospital in China. We apply the W3C Resource Description Framework (RDF) data model to represent these clinical terms and relationships, and successfully store all case data in a graph database with StrokePEO.

3. *Chapter 4:* This chapter develops an automatic, quantitative stroke severity assessment framework through automating the entire NIHSS scoring process on Chinese clinical EHRs.

4. *Chapter 5:* This chapter develops a prompting paradigm for automated analysis of EHRs using foundation LLMs. We first select LlaMa2-70b as the foundation LLM through a prompt-driven LLM selection process. Subsequently, we develop a novel retrieval-augmented generation (RAG) method to dynamically retrieve task-specific knowledge and demonstrations from the self-constructed, generation-augmented knowledge base. Finally, we propose a hierarchical chain-of-thought (HCoT) prompting strategy to integrate the macro sequential chain with the micro chain-of-thought.

5. *Chapter 6:* A brief summary of the thesis contents and its contributions are given in this chapter. Recommendation for future works is given as well.

Chapter 2

Literature Review

In this chapter, we review the related works of automated stroke assessment with advanced AI techniques. We first discuss the existing methods of automatic quantitative stroke assessment in Section 2.1, which is followed by a detailed review of ontology construction techniques shown in Section 2.2. Then we discuss the state-of-the-art CNER models and existing pre-trained embeddings in Section 2.3. At last, the cutting-edge LLMs and prompting strategies are presented in Section 2.4.

2.1 Automatic Quantitative Stroke Assessment

With the advancement of AI techniques, an increasing number of researchers have endeavored to explore its potential in mining valuable insights from the vast realm of clinical EHRs for quantitative stroke research. Zhang *et al.* [63] introduced an automated stroke severity prediction model based on machine learning techniques. Their model takes hospital service parameters as input variables, such as discharge information, length of stay and mortality risk to estimate the overall severity level. Compared with NIHSS scoring that directly examines the diagnostic symptoms exhibited by patients, this approach is neither reliable nor accurate. Park *et al.* [98] used signals acquired from special sensors to automatically grade the motor level of stroke patients. This method only covers a small portion of the quantified assessment items in NIHSS and therefore cannot provide a comprehensive stroke severity assessment. Also, it requires the use of additional equipment.

Recently, Yang *et al.* [144] attempted to develop an automatic approach to

identify the NIHSS items and scores reported in EHRs. This method has a restricted application scenario as it necessitates fully recorded NIHSS scores in EHRs, which can only be fulfilled by less than 5% of all clinical stroke EHRs. In this study, we address the limitations of previous research by developing an automatic and quantitative stroke severity assessment framework. This framework accurately scores NIHSS directly from diagnostic notes in Chinese EHRs, eliminating the need for pre-existing NIHSS reports or external equipment.

2.2 Ontology Construction

In this section, we summarize the existing stroke ontologies and the technologies for constructing medical ontologies from natural language. The existing research commonly breaks the task of constructing a medical ontology into four key steps, namely text preprocessing, term extraction, relationship extraction, and ontology integration. Below we summarize the existing technologies for each sub-task.

2.2.1 Existing Stroke Ontologies

From the world's largest biomedical ontology portal "BioPortal" [12], we have found two public stroke ontologies. The first is Stroke Ontology (STO) [41]. It has 1,712 classes, 69 instances and 35 properties, covering the knowledge of stroke as suggested by expert review. Currently, it is the largest, most comprehensive and most internationally recognized stroke ontology. The other is NIHSS Ontology [16], which has been linked to STO as a subclass of the "Scales" class. It focuses on quantitative assessment of stroke severity, including 18 classes, 106 instances and 22 properties.

Some academic research on stroke ontology is available. Townsend *et al.* [125] firstly designed a Neural Motor Recovery Ontology "NeuMORE" to represent the stroke patients' neuromotor function recovery status. Teresa *et al.* [84] built a Stroke

Diagnostic Ontology (DStrokeOnto), which contains 456 classes, 77 restrictions and 233 properties. It contributes the formalized medical knowledge for stroke diagnosis. Radhi *et al.* [6] created an ontology to represent knowledge for upper limb stroke rehabilitation in the patient information system. This ontology overcomes the problem of information inconsistency from various assessments. Soonhyun *et al.* [68] proposed a stroke medical ontology based on brain anatomies, lesions and stroke-related disease, aiming to assist the AI-based stroke prediction system.

The literature suggests a lack of effort to construct a comprehensive stroke physical examination ontology. This motivates our research to focus on developing a specific StrokePEO ontology to represent stroke physical examination as a complement to the Stroke ontology research field.

2.2.2 Text Preprocessing

The first step to construct a domain ontology from text is data preprocessing. This can be achieved by applying the common method of natural language processing (NLP) [109] for text parsing. Several successful NLP tools provide mature functions to accomplish these tasks.

The Natural Language Toolkit (NLTK) [13] is an open source platform that provides general text preprocessing capabilities such as sentence segmentation, word tokenization, stemming, part-of-speech (POS) tagging, parsing, and semantic reasoning. FreeLing [96] is another widely used library that supports high-level NLP parsing functions such as word sense disambiguation and semantic role labeling.

Unlike English, Chinese words usually consist of more than two Chinese characters, so special word tokenization methods are required. Jieba [60] is a widely recognized Chinese word tokenization module that provides functions such as word segmentation and part-of-speech tagging. It supports customized dictionaries which is quite helpful for specific domain text processing. HanLP [49] is a multilingual NLP library that

is primarily designed for Chinese text processing. It offers deep parsing functions including semantic dependency parsing, constituency parsing, semantic role labeling and abstract meaning representation (AMR) parsing.

2.2.3 Term and Relationship Extraction

The basic unit of an ontology is often represented in the form of triples, where two associated terms (classes) are described as $<$ term 1, relationship, term 2 $>$. The main task during the construction of a medical ontology is to extract terms and relationships from unstructured data.

2.2.3.1 Term Extraction

In the early years, people used manual extraction to collect relevant terms through experts according to certain rules. However, due to the high cost of manual extraction, automatic term extraction has become a research hotspot, known as "named entity recognition (NER)". The NER task is usually taken as a sequence labeling problem. Hence, classic machine learning methods such as Hidden Markov Model (HMM), Maximum Entropy Markov Model (MEMM) and Conditional Random Field (CRF) are widely applied to the task [132, 139]. With the increase in computing power, deep learning methods attract more and more attention and show their good performance. A bidirectional LSTM with a CRF layer (BiLSTM-CRF) gains much attention by achieving state-of-the-art performance on many CNER datasets [58, 77, 78, 156].

2.2.3.2 Relationship Extraction

Relationship Extraction (RE) is closely related to the NER task, which classifies the relationship between the entities identified in the text. The task is typically formulated into a classification problem that takes a piece of text and two entities in this text as inputs and the possible relation between the entities as output. The existing methods of RE can be roughly divided into two categories, *i.e.*, traditional

methods and neural network approaches. The former is based on feature-based [110] or kernel-based [149] approaches. These models usually spend a lot of time on feature engineering. Neural network methods can extract the relation features without complicated feature engineering. *e.g.*, recurrent capsule network and domain invariant convolutional neural network [112]. However, these methods cannot utilize joint features between entity and relation, resulting in lower generalization performance when compared with joint learning methods.

2.2.3.3 Joint Term and Relationship Extraction

Compared with pipeline methods, joint learning approaches are able to capture the joint features between entities and relations [76]. State-of-the-art joint learning methods can be divided into two categories, *i.e.*, joint tagging methods and parameter-sharing methods. Joint tagging methods transform NER and RE tasks into sequence tagging tasks through a specially designed tagging scheme, *e.g.*, a novel tagging scheme proposed by Zheng *et al.* [160]. Parameter-sharing methods share the feature extraction layer in the models of NER and RE. Compared to joint tagging methods, parameter-sharing methods are able to effectively process multi-map problems. The most commonly shared parameter layer in the medical domain is the Bi-LSTM network [74]. However, compared with the language model, the feature extraction ability of Bi-LSTM is relatively weaker, and the model cannot obtain pre-training knowledge through a large number of unsupervised corpora, which further reduces the robustness of extracted features.

To improve the performance of the BiLSTM-CRF model, word embedding techniques such as Word2Vec [86], GloVe [101], fasttext [14] and BERT [32] were investigated. Among these embedding methods, BERT obtains better word representations compared to the traditional methods. Moreover, domain-specifically fine-tuned embeddings can further improve the performance of medical NER and

RE tasks [142].

2.2.4 Ontology Integration

Ontology integration is the process of organizing the high-level knowledge obtained from different sources and involves data integration, disambiguation, reasoning verification, updating and other steps under the same framework specification.

Ontology integration can be subdivided into intra-class alignment and ontology linkage with other ontologies. Intra-class alignment determines whether classes in multi-source heterogeneous data point refer to the same object in the real world by considering instances and their attribute similarity. Ontology linkage starts from "ontology matching", *i.e.*, matches the semantic similarity of classes in one ontology with those in the other ontologies [116]. As an ontology grows in size and becomes more complex in structure, the classes, attributes, entities and their interrelationships are also taken into consideration. In the medical field, Dieng-Kuntz *et al.* [34] converted medical databases into medical ontology, and then used semi-automatic language tools for semantic extraction from other text corpora, extended and completed ontology building manually, using heuristic rules.

Although there are some meaningful attempts (*e.g.* [21]), it still requires a lot of manual processing to integrate ontologies in the medical field; therefore, further research is required to develop effective technology for efficient ontology integration in this setting.

2.3 CNER and Pre-trained Embeddings

The foundation technique of automated stroke assessment is to extract and map the key terms of symptoms, locations of a clinical presentation, level of severity, *etc.*, to ontology-based entity classes, which is called clinical named entity recognition (CNER). In this section, we summarize the existing CNER models and pre-trained

embeddings that can further improve the performance of CNER.

2.3.1 Existing CNER Models

In recent years, there have been major breakthroughs in the application of successful CNER techniques to process Chinese EHRs. The notable achievements include the application of deep neural networks to efficiently extract biomedical entities from free text with high accuracy. For example, the bidirectional Long Short-Term Memory (LSTM) [53] with a CRF [69] layer (denoted as "BiLSTM-CRF") has achieved state-of-the-art performance on many NER tasks [77,78,156,159]. Some recent works [58, 100, 152] integrated the bidirectional Gated Recurrent Unit (BiGRU) or Convolutional Neural Network (CNN) module with CRF layers, and have achieved competitive results with BiLSTM-CRF. Wan *et al.* [126] proposed an ELMO-ET-CRF model that used fine-tuned domain-specific ELMO as the input, Transformer ET as the encoder, and CRF as the decoder. Their model demonstrates competitive performance with the SOTA results on the CCKS2019 dataset. Zhang *et al.* [157] designed a multi-level representation learning model for CNER, which yielded better performance than the CNN-BiLSTM-CRF models.

The above CNER models are all trained on general medical datasets, and lack of training data specific to the stroke clinical domain, resulting in the inability to accurately identify the unique terms in the stroke clinical assessment. Moreover, the pre-defined entity categories are relatively sparse and cannot contain all valuable information for assessing stroke severity. Therefore, to address the above deficiencies, in this thesis, we first construct a CNER dataset specifically for stroke clinical research with well-defined and densely annotated fine-grained entity types. Then we evaluate SOTA neural network models, *i.e.*, BiLSTM-CRF [77], BIGRU-CRF [100] and CNN-LSTM-CRF [152], and identify the baseline CNER model for the subsequent automatic stroke severity assessment.

2.3.2 Pre-trained Embeddings

An essential approach to improve the performance of CNER models is to embed the EHR text with the embeddings pre-trained on large text corpora. The most popular embeddings can be categorized into three groups: traditional, discriminative and generative embeddings.

The traditional embeddings, in this thesis, serve as a broad category, encompassing all language models that predate the current state-of-the-art discriminative and generative language models. It includes Word2Vec [86], GloVe [101], fasttext [14], *etc.*, generating static word embeddings using statistical, count-based, and prediction-based methods. The discriminative embeddings, represented by BERT [32], ERNIE [121,122], Roberta [31], ELECTRA [25], *etc*, are mostly encoder-based, BERT-style models for identifying masked words within the given text. The generative embeddings are decoder-based, GPT-style models for generating next words beyond the given text [144], including GPT-(1-4) [18, 94, 104, 105], Palm [24], LLaMA [124], *etc.* Considering the application domain of the embeddings and our specific application scenario of high-accuracy CNER in Chinese EHRs, we deem that the encoder-based, discriminative embeddings are the most suitable choice for our task.

Previous studies [19, 77, 106, 151, 154] have suggested that domain-specific embeddings can represent domain terms more accurately than general embeddings. The barrier is the limited amounts of publicly available pre-trained Chinese clinical embeddings and their uncertain performance on new stroke-specific EHRs [19, 77, 154, 155]. However, considering the difficulty in obtaining high-quality domain-specific corpora, they crawl large amounts of low-quality medical data from the web sites as training corpora. Thus another concern is raised: should one prioritize the quantity or quality of the training corpus in domain-adaptive pre-training? In this thesis, we overcome

the shortfall of clinical domain-specific embeddings, and explore the answer to the above concern, by pre-training a Chinese clinical embedding based on the existing SOTA embeddings using specially collected Chinese clinical EHRs.

2.4 LLMs and Prompting Strategies

2.4.1 Large Language Models (LLMs)

Large language models (LLMs) refer to the foundation language models that can understand and generate natural language. They are based on the transformer architecture [137] and pre-trained on a large amount of data, typically containing hundreds or billions of parameters [15]. These include GPT-3.5 [105], GPT-4 [5], Meta's Llama model [124], Google's PaLM model [117], etc.

Many advanced proprietary LLMs have exhibited versatility in handling a wide array of tasks, including those in the field of health and medicine [67, 90, 91]. Furthermore, specific LLMs have been meticulously fine-tuned for medical applications, such as Med-PaLM [117], and Med-PaLM 2 [118]. This dual capability of general applicability and domain-specific refinement underscores the potential of LLMs in health and medicine.

Currently, some open-source LLMs have demonstrated excellent performance even comparable to state-of-the-art (SOTA) proprietary LLMs across various tasks [79, 81]. These models include LlaMa2 [124], BLOOM [138], Falcon [57], Alpaca [123], MedAlpaca [46], and many notable open-source Chinese LLMs, such as Baichuan [10], Qwen [9] and XVERSE [56]. Some LLMs with fewer parameters are specifically fine-tuned on Chinese medical data, such as DoctorGLM [140], and HuatuoGPT [130].

Performance assessments of these models are typically conducted on benchmark datasets with specific tasks, such as MMLU [51], MBPP [8], GSM8K [27], Math [52], to test the model's multilingual knowledge capabilities, translation, mathematical

reasoning, coding, and other capabilities [61, 97]. However, these evaluations may not be adequate for identifying the applicability and performance of LLMs in real-world applications such as clinical assessment using EHRs. To address this, we have designed a set of prompt-driven LLM selection templates to effectively identify the foundation LLMs that align with our specific task requirements (see Section 5.2.1 for more details).

2.4.2 Retrieval Augmented Generation (RAG)

Retrieval augmented generation is a technique aimed at enhancing the performance of LLM generation by incorporating valuable information and demonstrations from an external knowledge base. This external knowledge base can be existing databases and structured resources with domain-specific knowledge [99, 115]. However, building and maintaining a knowledge base suitable for LLMs is a labor-intensive task that demands significant human and time resources. This effort is also susceptible to errors and omissions, which can subsequently impact the effectiveness of the generated content in various tasks [158]. Leveraging the powerful generation capability of LLMs, the technique of generation-augmented self-construction of the external knowledge base has been proposed to address the above challenges and proved to be very efficient and effective [91, 133].

Various retrieval methods can be employed to extract content relevant to the query from the knowledge base. These include classic matching methods such as TF-IDF and BM25 [20], dense representation-based retrieval [40], and other embedding-driven retrieval mechanisms, like KNN [91] and DPR [62]. However, due to the limited information provided in sentence-level queries and the inherent information loss from fixed-sized embeddings in the document-level knowledge base, traditional retrieval methods are prone to fail in precisely identifying the most relevant records [82]. Generative-augmented retrieval (GAR) is introduced to mitigate these limitations by

enhancing the semantics of queries, leading to a substantial improvement in retrieval accuracy [83].

This study employs a generation-augmented approach to construct an external knowledge base for stroke assessment. Leveraging the generative capabilities of LLMs and referencing dataset labels, our approach ensures the efficient generation of a high-quality external knowledge base, validated through a final verification process. Furthermore, to ensure a high retrieval accuracy, we develop an innovative summary-based GAR method to replace the full-text embeddings with LLM-generated summary indexes in which only key terms are extracted and embedded. This effectively enhances retrieval accuracy and the performance of the ultimate task.

2.4.3 Prompt Engineering

Prompt engineering entails the strategic design of effective prompts to guide LLMs in accomplishing downstream tasks. It plays a pivotal role in successful LLM generation. With the rapid development of LLMs, numerous prompting methods have emerged [4, 18, 111, 134, 158]. In this section, we review existing prompting techniques based on their order of effectiveness in three stages: in-context learning (ICL), logical reasoning, and subsequent optimization. Through comparison and discussion, we identify the optimal prompting methods for our tasks.

2.4.3.1 In-context Learning

In-context learning (ICL) is a capability of LLMs to learn and generate responses based on the context of the conversation without fine-tuning. The main prompting methods based on ICL include zero-shot prompting and few-shot prompting.

Zero-shot prompting enhances the use of LLMs by eliminating the need for extensive training data [105]. Instead, it uses carefully crafted prompts to guide the model on new tasks. The model receives a task description in the prompt without

labeled data for specific input-output mappings. It then relies on its pre-existing knowledge to generate predictions based on the given prompt.

Few-shot learning is a key ICL capability of LLMs [18]. It teaches an LLM to learn from only a small number of labeled examples to generate a new, unseen, but similar result. Even a few high-quality examples can significantly improve model performance on complex tasks. However, this approach requires additional tokens to include the examples, which can be a limitation for longer text inputs. Additionally, the selection and composition of prompt examples are crucial as they can significantly influence the model's behavior [111].

2.4.3.2 Logical Reasoning

Numerous studies have shown that breaking down complex tasks into steps and allowing LLMs to perform logical reasoning is an extremely effective prompting method [111, 134, 158]. The most representative of these is the Chain-of-Thought (CoT) prompting [134]. It encourages an LLM to "think step by step", entering a mode of reasoning where it systematically breaks down complex tasks into a sequence of ordered steps. This prompting method has improved the accuracy and coherence of the generated outputs [134, 158], and is entitled CoT to provide a vivid portrayal of the model's sequential thinking process.

Building on the foundation of CoT, numerous logical reasoning prompting methods have emerged, such as automatic chain-of-thought (AutoCoT) [158], Tree-of-Thoughts (ToT) [146], Graph-of-Thoughts (GoT) [148], Thread of Thought (ThoT) [161]. These logical reasoning prompts are suitable for various application scenarios. In this study, considering the sequential nature of EHRs and the challenges posed by long texts, we propose the Hierarchical Chain of Thought (HCoT) method. By decomposing tasks at the paragraph level and applying CoT at the sentence segments level, HCoT significantly improves the performance of LLMs in automatic stroke assessment.

2.4.3.3 Optimization

Due to the inherent uncertainty and potential hallucinations in LLMs when generating responses, their inferences can sometimes produce unexpected biases. To address this issue, researchers have proposed optimization prompting methods, including ReAct prompting [147], Chain-of-Verification (CoVe) [33] and ensembling [4]. Among these, ensembling is one of the most widely used techniques. It combines the outputs of multiple individual models or multiple generations by one LLM with different degrees of randomness to produce a more accurate and reliable result, instead of relying on a single reasoning output [91].

Well-designed prompting strategies have demonstrated comparable or even superior performance than specific fine-tuning methods [91, 119]. However, to date, there is little report on the successful implementation of the emerging prompt engineering techniques in clinical assessment tasks using the EHR data.

To address this methodology gap, we develop a set of generation-augmented prompting strategies and formulate a prompting paradigm entitled "GAPrompt". This paradigm is designed to support foundational generic LLMs in achieving the objectives of specific tasks. These strategies include few-shot prompting, chain-of-thought, and ensembling.

Based on the requirements of our task for automatic quantitative stroke assessment and considering the challenges identified in the literature review, we adopt a progressive approach to achieve our goal. We begin by developing a stroke-specific ontology, followed by utilizing neural network-based CNER models and pre-trained embeddings for entity extraction from EHRs, and completing the NIHSS scoring with a dictionary-mapping algorithm. Finally, leveraging the state-of-the-art LLM techniques, we propose the GAPrompt paradigm to facilitate LLMs in automating stroke assessment based on EHRs. More details are described in the following chapters.

Chapter 3

StrokePEO: Construction of a Clinical Ontology for Physical Examination of Stroke

Clinical ontology is a standardized medical knowledge representation model that facilitates the integration and analysis of a large amount of heterogeneous electronic health record (EHR) data. Using ontologies to represent clinical terms can improve data integration to build robust and interoperable medical information systems. To date, there is no ontology existing to represent the medical knowledge for physical examination of stroke, which has inhibited the stroke physicians to make full use of clinical information captured in EHR data to understand stroke patient's health status and plan effective medication and rehabilitation treatment.

In this chapter, we co-design with two stroke clinical specialists a stroke clinical ontology "StrokePEO" using advanced natural language processing and deep learning techniques to extract terms and their relationships from real clinical case records provided by a tertiary hospital in China. We apply the W3C Resource Description Framework (RDF) data model [3] to represent these clinical terms and relationships, and successfully store all case data in a graph database with StrokePEO. Our experiment results suggest that our methods and the output of StrokePEO can be applied in various medical contexts that require extraction of medical knowledge from free text for decision making. These include, but not limited to, physical assessment, drug and rehabilitation treatment outcome evaluation, medication effect analysis, and patient risk prediction.

3.1 Background

In today's age of information and big data, EHRs are being created and collected at an unprecedented rate in medical setting [43]. As the volume of data has grown exponentially, so has the scope and depth of the stored EHR data, which often include patient demographics, diseases, diagnoses, symptoms, medications, treatments, and other health service data. Therefore, observational electronic health record data is a large treasure chest that waits to be explored and utilized. However, there are also many flaws, such as incomplete, inconsistent or incorrect data, and insufficient data details and missing data in EHR, data from different information systems owned by different healthcare providers can be very different. Therefore, EHR data needs to be handled with special caution to ensure their appropriate use to generate high performing algorithms. It is important to ensure safe and ethical use of EHR that will improve patient safety, healthcare quality and efficiency. To avoid ambiguity and ensure data quality, a standardised data representation that can be recognized by both machines and humans is needed. An ideal data representation needs to standardize knowledge in the relevant health domain and can facilitate analysis and integration of heterogeneous data from diverse data sources. This calls for ontology.

Physical examination is a key step in stroke diagnosis. Through physical examination at admission, doctors can obtain a preliminary understanding of the patient's condition. Based on this, they will further prescribe complex diagnostic tests and treatment plans, *e.g.* medication or rehabilitation treatment. Stroke rehabilitation can reduce or remove the direct pathogenic impact factors for stroke, *e.g.*, ischemia or cerebral hemorrhage. Image tests, such as neuroimaging, are typically aimed at identifying the pathogenic areas but cannot determine whether a patient has recovered from stroke. Only detailed physical examination can provide a comprehensive assessment of the recovery status of various physical functions of the patient.

In this research, we construct a clinical ontology dedicated to stroke physical examination, called "StrokePEO", which focuses on stroke assessment. Different from the existing ontologies for stroke [16, 41], our source data comes from the real clinical case records of the Third Affiliated Hospital of Sun Yat-sen University, Guangzhou, China. The ontology schema, term classes, relationships, *etc.* are co-designed and validated by two stroke specialists. The terms and relationships in the StrokePEO are represented in the Resource Description Framework (RDF) data model [3]. The annotated dataset is used for training, evaluating and testing of the deep learning-based term relationship extraction methods. Experiments show that our approach can effectively mine clinical terms and relationships critical for stroke physical examination. We conduct ontology integration, including term alignment and linkage with other ontologies, to enhance the robustness, consistency and scalability of the StrokePEO in stroke ontology research.

The rest of this chapter is organized as follows. In Section 3.2 we provide a detailed description of our approach to construct the StrokePEO. Section 3.3 presents the dataset and the experiment results. Finally, the chapter concludes in Section 3.4.

3.2 Methodology

In this section, we illustrate in detail the key steps we take in constructing the StrokePEO. Following the seven-step approach of ontology construction recommended by a Stanford research group [92], we use Protégé [87] to build the StrokePEO *i.e.*, to determine scope, consider reuse, enumerate terms, define classes, define properties, define constraints and create instances.

Figure 3.1 shows the pipeline of our construction of StrokePEO. We apply advanced NLP technologies, including sentence segmentation, word tokenization, named entity recognition and relationship extraction models, to accomplish the tasks at each step. *i.e.*, text preprocessing, ontology schema definition, joint term and

relationship extraction, term alignment and ontology integration.

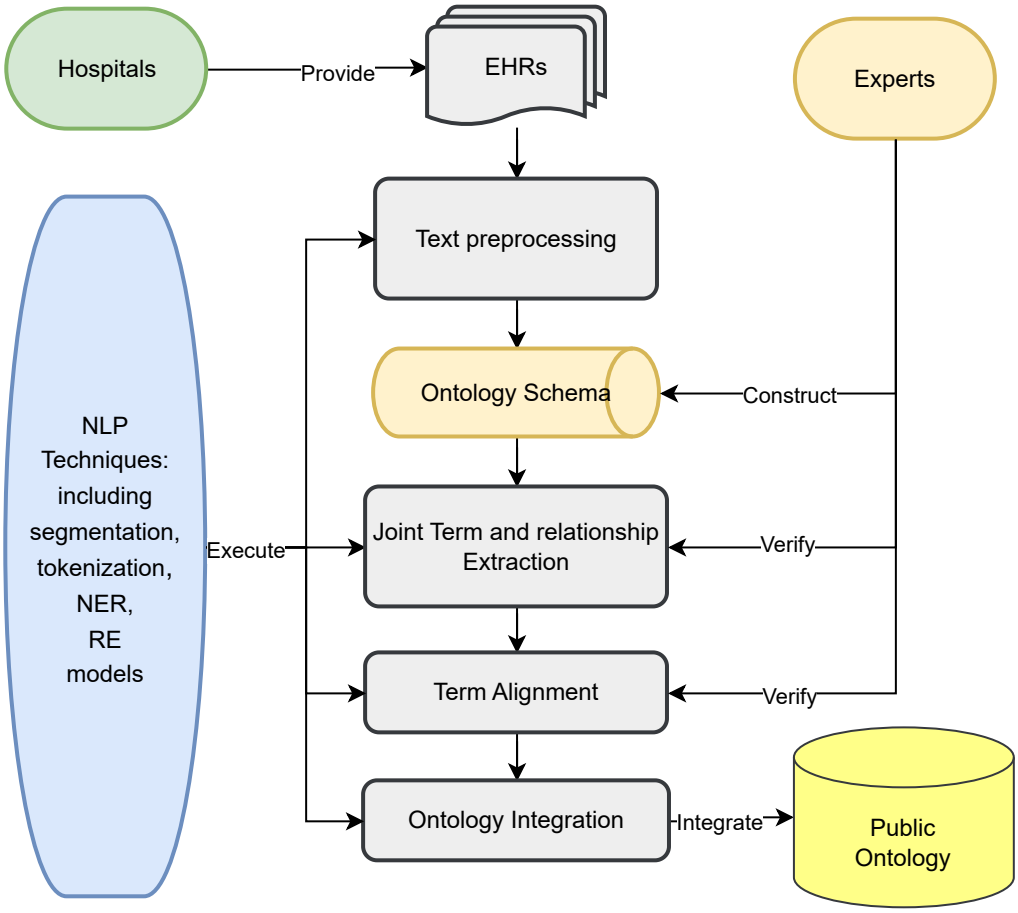

Figure 3.1 : The pipeline of constructing StrokePEO. It illustrates the process from hospital-provided EHRs to an expert-verified ontology using advanced NLP techniques.

3.2.1 Text Preprocessing

We apply a series of NLP techniques to preprocess the raw clinical text data. These include unifying format, removing the staleness, sentence segmentation and word tokenization with POS tagging.

Unlike the common article sentence, the structure of medical record text usually

does not have a complete and standard syntactic structure, but lists multiple subject-predicate phrases in a sentence. This inhibits the effective application of the semantic dependency parsing method to process these clinical records. Therefore, instead of the common practice for English in using "period" as the delimiter, we use "comma" or "semicolon" as the delimiter to divide sentences. The resultant segmented sentences are short in length, but still contain single or multiple terms to form the triples of "subject, predicate and object" (SPO).

To tokenize Chinese words, we adopt the Jieba [60] package with self-defined dictionaries. Two dictionaries are imported to help enhance the accuracy of word tokenization. The first is a dictionary named "THUOCL-medical" [44] produced by Tshinghua University, with medical words and their frequency annotated. The other is annotated by us and reviewed by clinical experts, to handle special terms with POS tagging suitable for stroke physical examination records.

3.2.2 Ontology Schema Definition

Through in-depth analysis of the structure and concepts of stroke terms, we develop a schematic ontology representation and represent the terms using the Resource Description Framework (RDF) data model [3]. Its atomic data format is called RDF triple, which consists of three entities in the form of "subject, predicate, object" to show the semantic statement of "term 1 has relationship with term 2".

Specifically, in our StrokePEO, triples are composed of fine-grained terms and relationships, to express the knowledge as precisely and as accurately as possible. To construct our StrokePEO, we define seven classes of terms, *i.e.*, Anatomy, Inspection, Symptom, Position, Binary, Change and Degree. Twelve relationships are defined among the term classes. The detailed ontology representation schema is shown in Figure 3.2.

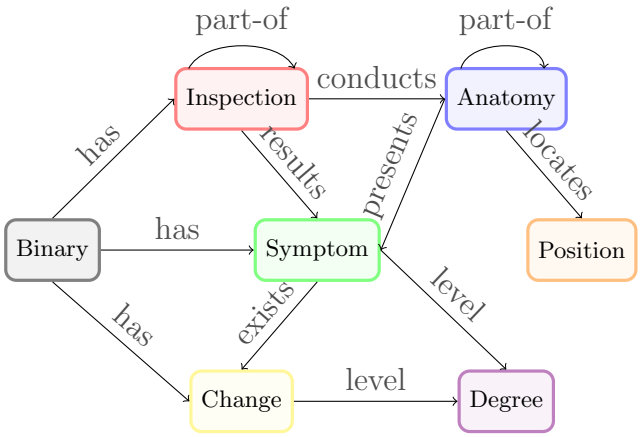

Figure 3.2 : The schematic representation of our StrokePEO.

3.2.3 Joint Term and Relationship Extraction

To construct the StrokePEO, we mine useful terms and their relationships from a large amount of raw clinical EHR text data, using two hot research techniques in the field of text mining, *i.e.*, Named Entity Recognition (NER) and Relation Extraction (RE). With the continuous development of machine learning and deep learning to reach maturity level, many mature NER and RE algorithms are now publicly available.

We co-define with the two stroke specialists in our team the ontological representation of each concept for stroke physical examination. Due to that a sentence usually contains multiple RDF triples, we apply multi-relational classification for model training and prediction. We first classify the relationship constraint. Then we put the same term into different classes in accordance with the relational constraint in a relevant sentence to resolve the ambiguity of semantic relationships expressed by the same term in different context.

We apply the TensorFlow-based Entity and Relation Extraction model [2], a schema-based pipeline entity-relation extraction model. This model has achieved excellent performance comparable to the SOTA model in the “2019 Language and

Intelligence Challenge" [1] and has been widely recognized by high popularity stars in GitHub.

This approach leverages a joint term and relationship extraction model based on Bert-BiLSTM-CRF architecture. Different from other models that first perform NER and then extract relationships. This model addresses term and relationship extraction tasks using a pipelined approach. Initially, a multi-label classification model assesses the relationship types within sentences. Next, both the sentence and potential relationship types serve as input for a sequence labeling model. This model identifies terms (entities) within the sentences. Finally, the predicted relationship is combined with the entity output to form an entity-relationship triples: (term 1, relationship, term 2).

To improve the model accuracy, instead of using the original model, we adopt a more advanced Chinese embedding named Chinese Pre-trained BERT with Whole Word Masking (ROBERTA_wwm_large_ext) [30]. It significantly outperforms the standard BERT embedding for our entity relationship extraction task. We discover as broadly and comprehensively terms as possible, resulting in many terms with similar or even the same semantic meaning.

3.2.4 Term Alignment

The purpose of this step is to unify the synonymous terms into one standardised term to ensure atomicity of the concept classes in the constructed StrokePEO. To improve accuracy, we combine the open source Chinese synonym tool "Synonyms" [128] with the word2vec model [107] to process the clinical data. As both models have fully learned the context information embedded in the adjacent and distant words during training, they can infer, to a large extent, the original meaning of words and their relationships.

We use these two models to obtain the ten most similar terms for each extracted

term, respectively. After filtering out terms with different term classes, the remaining terms are marked as synonyms of the standardised term. Finally, clinical experts are called upon to validate accuracy of the machine-generated thesaurus.

3.2.5 Ontology Integration

We set up the scope of the StrokePEO ontology as the diagnostic physical assessment of stroke patients in clinical setting to address this gap in Chinese stroke ontology. Based on the systematic review of existing ontologies and previous research work, our StrokePEO can be recognized as a complement to the research field of stroke ontology, which can be directly integrated into the current most authoritative Stroke ontology (STO) [41], under the "Stroke-Diagnosis-Evaluation of stroke-Physical Examination" class.

The clinical experts in our team expect the StrokePEO to have the ability to be integrated with other stroke ontologies to meet the needs of real-world applications and research. For example, when there is a clinical requirement to assess the severity of a patient's stroke condition, the clinicians usually use the international standard NIHSS [16]. In order to align with the NIHSS international standard, we integrate the two ontologies, StrokePEO and the NIHSS ontology.

As mentioned in the literature section, there is no fully automatic ontology fusion algorithm in medical domain; therefore, manual fusion has to be conducted in this project. The NIHSS is composed of 11 classes, including consciousness level, eye movement, motor arm and leg, speech, *etc.* It has less classes than our StrokePEO. Thus, with the guidance and quality control of the clinical experts, we manually match the classes of StrokePEO with those in the NIHSS ontology. This integration mainly consists of two tasks, one is to match the "Inspection" class in StrokePEO with the classes in NIHSS, and the other is to match the "Symptom" class in StrokePEO with the value set in NIHSS. After integration, the resulted StrokePEO will afford

people even without professional training to acquire a quantitative assessment score of a patient's stroke condition by observing the patient's clinical manifestations.

3.3 Experiments

To efficiently extract the terms and relationships from the large amount of text data, we apply the advanced deep learning-based techniques to automatically recognize the terms and classify their relationships in each sentence. As supervised learning requires a batch of data annotated with correct labels to train the algorithm, we first introduce our approach to acquire the annotated dataset, and then report the setting and performance of the two algorithms used for term extraction and relationship classification. Finally, we evaluate the quality of the constructed StrokePEO.

3.3.1 Dataset

The study dataset is collected and labeled from the clinical case records of physical examination results for stroke patients from the Third Affiliated Hospital of Sun Yat-sen University, China. The definition of ontology schema, including the classes of terms and relationships are all guided and approved by two stroke experts. The dataset contains 89,351 annotated samples, and are randomly split into training set, evaluation set and test set at a ratio of 4:1:1. Each annotated sample is composed of the raw text and lists of SPO ("subject, predicate, object") triples to show the terms and relationships. As a pipeline model, both term extraction and relationship classification algorithms are trained on the same dataset. Therefore, we add the term type into the SPO triples, indicating the subject type and object type. For example, the sentence "右侧肢体肌力5级" is extracted into three RDF triples, *i.e.*, (subject: "肢体", subject_type: "Anatomy", predicate: "locates", object: "右侧", object_type: "Position"), (subject: "肢体", subject_type: "Anatomy", predicate: "conducts", object: "肌力", object_type: "Inspection"), and (subject: "肌力", subject_type:

Table 3.1 : The statistics of the annotated terms and relationships.

Subject Term Type	Relationship	Object Term Type	Count
Inspection	results	Symptom	85,893
Inspection	conducts	Anatomy	11,968
Anatomy	presents	Symptom	15,888
Anatomy	locates	Position	24,096
Binary	has	Symptom	10,393
Binary	has	Inspection	26,657
Binary	has	Change	386
Symptom	exists	Change	3,219
Symptom	level	Degree	7,187
Change	level	Degree	893

"Inspection", predicate: " results", object: "5级", object_type: "Symptom"). Table 3.1 shows the statistics of the annotated terms and relationships in the dataset.

3.3.2 Relationship Classification Results

We conduct a multi-class classification model to predict the possible relationships in a sentence. The input of this model is raw texts from training samples, which are first tokenized and embedded by a BERT layer. We have found that, replacing the BERT embedding with the ROBERTA embedding [30] has led to much better performance. The embedding sequences are then passed to the multi-class classifier, which outputs the predicted set of possible relationships in the text.

To evaluate the accuracy of relationship classification, we compare the predicted relationships with the golden set. If the predicted relationship matches the golden set, it is marked as "Correct". If the output set is equal to or greater than the golden

Table 3.2 : Relationship classification results

Output Results	Count	Total Numbers	Accuracy (%)
Correct	8125	8228	98.7482
Superset	52	8228	0.6320
Subset	13	8228	0.1580

set, it is marked as a "Superset". Finally, a "Subset" result indicates that the output set contains only a part of the correct relationships in the golden set. Table 3.2 shows the results. As it can be seen that, the classification algorithm can effectively identify all possible relations in sentences with more than 98% accuracy. In a few cases of inaccurate predictions, partially correct relationships can also be identified with a small number of redundant or missing predictions.

3.3.3 Term Extraction Results

We run a sequence labelling model to extract the terms from the input text, *i.e.*, the relationship classification results. First, the model converts a training sample with multiple labels into multiple samples, so that the mapping between the original text and label in each sample is one-to-one relationship. Then the predicted subject and object terms are constrained by the classified relationships to suit their types.

To evaluate the performance of the term extraction algorithm, we calculate the accuracy of the predicated SPO triples. A "correct SPO" indicates that the predicated SPO triples are exactly the same as the golden set regarding the terms, types and relationships. We also report the number of predicted SPO triples and the number of SPO triples in the golden set. Finally, we evaluate the performance of the term extraction model using the common metrics, including Precision (P), Recall

Table 3.3 : Term extraction results

Correct SPO num	14,114	Submitted SPO num	14,567
Golden set SPO num	15,086		
Task	Precision (%)	Recall (%)	F1-score (%)
Term extraction	96.89	93.56	95.19

(R) and F1-score ($F1$), which are defined by:

$$P = TP/(TP + FP)$$

$$R = TP/(TP + FN)$$

$$F1 = (2 \times P \times R)/(P + R),$$

where TP is the number of true positives, FP is the number of false positives, and FN is the number of false negatives.

The detailed results of term extraction are shown in Table 3.3. For all of the evaluation metrics, the larger the values, the better the algorithm performs.

3.4 Summary

For the first time, this study has developed and validated a clinical ontology "StrokePEO" for physical examination of stroke using real clinical case record data. We have applied multiple NLP techniques to preprocess the raw text records and have adopted advanced deep learning techniques to successfully extract the terms and relationships pertaining to the physical examination of stroke. Our method offers a rapid and robust approach for constructing new medical domain ontologies using advanced NLP techniques. Notably, it substantially reduces the labor costs associated with manual construction and can be extended to other diseases. Moreover, our

approach and the resulting StrokePEO ontology provide a useful machine learning model and base for the further development of diverse clinical decision support systems that generate knowledge from rich clinical text. These include, but are not limited to, physical assessment, drug and rehabilitation treatment outcome evaluation, medication effect analysis, and patient risk prediction.

Crucially, the StrokePEO serves as a foundational resource for extracting medical terms, playing a pivotal role in our endeavor for automated stroke assessment. It enhances our deep learning-based clinical named entity recognition (CNER) approach and facilitates the automated scoring of NIHSS, a framework that will be elaborated upon in the next chapter.

Chapter 4

Automatic Quantitative Stroke Severity Assessment based on Chinese Clinical Named Entity Recognition with Domain-Adaptive Pre-training

The previous chapter presents our constructed StrokePEO ontology, which comprehensively defines the clinical terms and relationships related to stroke assessment. This lays the foundation for leveraging advanced AI technologies to achieve automated entity extraction and ultimately automate the assessment process. In this chapter, we develop an automatic, quantitative stroke severity assessment framework through automating the entire NIHSS scoring process on Chinese clinical EHRs. Our approach consists of two major parts: Chinese clinical named entity recognition (CNER) with a domain-adaptive pre-trained embedding and automated NIHSS scoring. To build a high-performing CNER model, we first construct a stroke-specific, densely annotated dataset "Chinese Stroke Clinical Records" (CSCR) from EHRs provided by our partner hospital, based on our constructed StrokePEO ontology (see the previous chapter) that defines semantically related entities for stroke assessment. We then pre-train a Chinese clinical embedding coined "CliRoberta" through domain-adaptive transfer learning and construct a deep learning-based CNER model that can accurately extract entities directly from Chinese EHRs. Finally, an automated, end-to-end NIHSS scoring pipeline is proposed by mapping the extracted entities to relevant NIHSS items and values, to quantitatively assess the stroke severity.

Results obtained on a benchmark dataset CCKS2019 and our newly created CSCR dataset demonstrate the superior performance of our domain-adaptive pre-

trained embedding and the CNER model, compared with the existing benchmark embeddings and CNER models. The high F1 score of 0.990 ensures the reliability of our model in accurately extracting the entities for the subsequent automatic NIHSS scoring. Subsequently, our automated, end-to-end NIHSS scoring approach achieved excellent inter-rater agreement (0.823) and intraclass consistency (0.986) with the ground truth and significantly reduced the processing time from minutes to a few seconds.

Our proposed automatic and quantitative framework for assessing stroke severity demonstrates exceptional accuracy and reliability through directly scoring the NIHSS from diagnostic notes in Chinese clinical EHRs. Moreover, this study also contributes a new clinical dataset, a pre-trained clinical embedding, and an effective deep learning-based CNER model. The deployment of these advanced algorithms can improve the accuracy and efficiency of clinical assessment, and help improve the quality, affordability and productivity of healthcare services.

4.1 Background

With the expanded use of EHRs across healthcare organizations, research interest has grown dramatically in the application of artificial intelligence (AI) and natural language processing (NLP) technologies to automatic disease assessment utilizing the extensive volume of data captured in EHRs [54, 63, 98, 141, 144]. Among them, a key technique is to identify and extract clinical terms from doctors' free-text diagnostic notes, which is known as Clinical Name Entity Recognition (CNER). Many advanced deep learning-based approaches have emerged to tackle the CNER tasks, motivating researchers to apply the state-of-the-art (SOTA) CNER technology to the automatic assessment of diseases.

In terms of stroke assessment, studies have emerged to extract quantitative measurement of stroke severity using CNER and machine learning techniques [63, 98,

144]. However, they either only recognize NIHSS scores that are already recorded reported in the EHRs, or require external equipment and data to make predictions. To date, there is no existing approach available that achieves automatic, quantitative stroke severity assessment directly from the patients' diagnostic notes in EHRs.

These limitations have severely hindered the effective utilization of the vast amount of clinical EHR data for quantitative stroke assessment.

To fill the gap, in this research, we develop an automatic framework for quantitatively assessing stroke severity directly from diagnostic notes in Chinese EHRs, which has the potential to replace the often time-consuming, tedious and unreliable manual assessment widely practiced in clinical settings. Specifically, our first step is to address the insufficiency of a CNER dataset for stroke assessment, where we construct a stroke ontology-based, densely annotated dataset "CSCR", meaning that most words in each sentence in the dataset are annotated with entity labels. We then address the insufficiency of existing language models in representing Chinese clinical EHRs through domain-adaptive pre-training of a clinical domain-specific embedding coined "CliRoberta". Subsequently, inspired by the successful applications of mapping-aided methods in quantitative clinical research [29, 59, 66], we define an entity-to-NIHSS mapping that links the extracted entities to relevant items and values in the NIHSS [17]. Finally, we design and implement an end-to-end pipeline to automatically calculate the NIHSS scores based on the defined mapping. A series of experiments and evaluations conducted on real-world data demonstrate the excellent reliability and superior efficiency of our proposed automatic stroke severity assessment framework.

The rest of this chapter is organized as follows. In Section 4.2, we present the details of our proposed approach, encompassing the construction of datasets and mappings, the development of Chinese CNER with embedding, and the automated

end-to-end NIHSS scoring. The experimental results of our embedding-based CNER and NIHSS scoring are presented in Section 4.3. In Section 4.4, we analyze the key findings, compare our work to prior studies, and discuss its limitations. Finally, Section 4.5 summarizes our work.

4.2 The Proposed Method

Our automatic stroke severity assessment framework comprises two key components: building the embedding-based CNER model, and automated NIHSS scoring. To accomplish these two tasks, we first construct the CSCR dataset and entity-to-NIHSS mapping dictionary. Then, we pre-train a domain-specific embedding "CliRoberta" through domain-adaptive transfer learning on Chinese clinical EHR data, and develop a deep learning-based Chinese CNER model to accurately extract entities. Finally, these entities are mapped to NIHSS scores through an automated, end-to-end entity-to-NIHSS mapping pipeline.

Figure 4.1 illustrates the steps of the proposed automatic stroke severity assessment: (i) constructing the CSCR dataset (in green); (ii) constructing entity-to-NIHSS mapping dictionary (in yellow); (iii) domain-adaptive pre-training of a Chinese clinical embedding (in blue); (iv) generating the CNER model (in red); and (v) automated NIHSS scoring (in purple).

4.2.1 Construction of CSCR Dataset and Entity-to-NIHSS Mapping

Applying NLP techniques for automatic stroke severity assessment requires a disease-specific and densely annotated dataset. To the best of our knowledge, there is no publicly available annotated Chinese EHR dataset for stroke assessment. In this study, we construct a stroke-specific, intensively annotated dataset, CSCR (detailed below), in close collaboration with the stroke specialists from three top hospitals in China, *i.e.*, the Third Affiliated Hospital of Sun Yat-sen University, the First

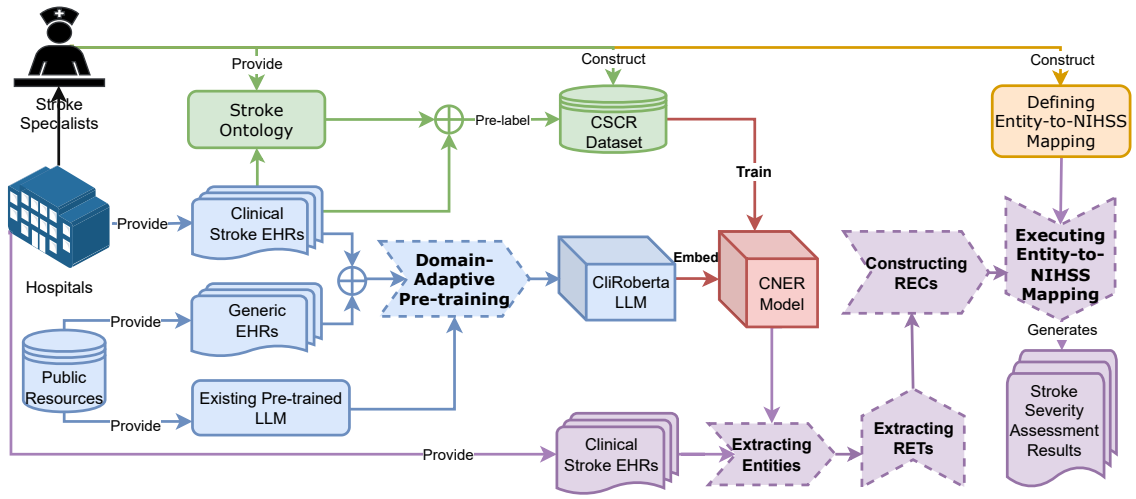

Figure 4.1 : The procedure of the proposed automatic quantitative stroke severity assessment framework. Green color: constructing the CSCR dataset; yellow color: Constructing entity-to-NIHSS mapping dictionary
; blue color: domain-adaptive pre-training of Chinese clinical embedding; red color: generating the CNER model; purple color: automated NIHSS scoring.

Affiliated Hospital of Jinan University, and the First Affiliated Hospital of Fujian Medical University. Subsequently, we establish an entity-to-NIHSS mapping to facilitate the automated NIHSS scoring for quantitative stroke severity assessment.

4.2.1.1 CSCR Dataset Construction

Our CSCR dataset is built from the Chinese clinical stroke EHRs provided by our partner hospital "The Third Affiliated Hospital of Sun Yat-sen University". It consists of de-identified admission and discharge notes of 1,133 patients, including in total 1,067 admission records and 864 discharge records. To construct this dataset, these source EHRs are annotated into clinical terms describing the conditions of stroke patients, referred to as *entities*. Different from the sparsely annotated entities in existing studies [22, 47, 150, 151], we construct a stroke-specific, densely annotated CNER dataset in which most words of each sentence are assigned with entity labels.

Towards this end, we first define seven types of stroke ontology-based entities and then propose a semi-automatic approach to annotate the source Chinese clinical stroke EHRs, which are further validated by clinical specialists, detailed below.

Entity Definition. Based on the stroke ontology [39] and the standardized protocols for NIHSS assessment [17], we identify seven types of semantically related entities, which, when used in combination in the diagnostic notes, can accurately describe stroke symptoms.

As shown in Table 4.1, these entities are: Inspection (ISP), Symptom (SPT), Position (POS), Binary (Bin), Anatomy (ANT), Change (CHG), and Degree (DEG). Among them, the **ISP** and **SPT** entities, representing the specific inspection item and the corresponding symptom (*i.e.*, the assessing result of the inspection item) of a patient, are the two most prevalent entities in our CSCR dataset, accounting for 38.1% and 30.9% of the annotated entities, respectively. The **POS** and **ANT** entities describe the location of a stroke symptom in a certain body part, accounting for 12.7% and 7.3% of the annotated entities, respectively. The remaining three entities, *i.e.*, **BIN**, **DEG**, and **CHG**, describe the existence and degree of changes of the inspected symptoms.

Entity Annotation. We then develop a semi-automatic approach to annotate the entities in the source EHRs using the BIO (Begin, Inside, Outside) format. Taking the SPT entity as an example, the label "B-SPT" implies the start of a symptom entity, and "I-SPT" marks the character inside the entity. Characters that do not belong to any entities are annotated by "O".

This automatic pre-annotation process proceeds by programmatically automating the annotation of entities to the EHRs, supplemented by the manual revisions of specialists to ensure accuracy and reliability. Various NLP techniques are applied during the automatic pre-annotation, including sentence segmentation, word segmentation,

Table 4.1 : The entities, definitions, count, and percentage of occurrence in the constructed CSCR dataset.

Entity Name	Examples	Definition	Count	Percentage
Inspection (ISP)	“意识” (consciousness), “肌力” (muscle strength)	Specific inspect item	37,964	38.1%
Symptom (SPT)	“流利” (fluent), “麻木” (numb)	Sign of a body condition	30,769	30.9%
Position (POS)	“左” (left), “右” (right)	Location of an anatomy	12,656	12.7%
Binary (BIN)	“有” (has), “无” (no)	Existence or not	7,139	7.1%
Anatomy (ANT)	“眼睛” (eye), “上肢” (arm)	Structure of the body	7,283	7.3%
Degree (DEG)	“明显” (significant), “稍微” (slight)	Level of change	1,993	2.0%
Change (CHG)	“变差” (worsen), “恢复” (recovered)	Treatment effect	1,857	1.9%

and part-of-speech tagging, which are first conducted to preprocess the EHRs, with the support of the Jieba library [120] and Chinese clinical dictionaries [45]. Then, based on the preprocessed EHRs, we revise the segmentation results according to the stroke ontology [39]. Next, we repeat this ontology-based automatic pre-annotation process on all EHRs. Finally, the annotation results are reviewed and verified by the stroke specialists. The CSCR dataset (shown as the green module in Figure 4.1) is thus constructed. In total, there are 347,637 Chinese characters with 99,661 word entities annotated in the CSCR dataset.

Different from the existing CNER datasets [22,47,150,151], this is the first stroke-specific CNER dataset with intensively annotated, semantically related entities for extracting clinical terms from Chinese stroke EHRs. Its more intensive annotation and the semantically associated entities show strong potential in supporting quantitative assessment in clinical research compared with previous works [100, 135].

4.2.1.2 Constructing Entity-to-NIHSS Mapping Dictionary

According to the prescribed guidelines for conducting NIHSS evaluations [17], each score needs to be jointly determined by multiple entities with semantic dependencies. Thus, in order to produce a simple, yet complete mapping to fully represent the semantic relationship and be friendly for query, we design four modules in the mapping, *i.e.*, the Core module M_core, the Categorical module M_cat, the Supplementary module M_sup and the Synonym module M_syn.

Using our trained CNER model and the relational entity triple (RET) extraction method (Algorithm 4.4), we extract the ISP-SPT triples from the stroke clinical EHRs to form the Core module M_core. Each "ISP-SPT" entity pair is associated with its corresponding NIHSS element and a *score*. For the "BIN-SPT" entity pairs, we also provide the reverse score named $score\_r$. For example, NIHSS element 9 aims to assess language ability in understanding and reading, ranging from 0 to 3. "言语" (speech)-"不清" (unclear) scores 1 point, "言语" (speech)-"困难" (hard) scores 2 points, "失语" (aphasia) and "昏迷" (coma) both score 3 points.

The Categorical module M_cat consists of three types of entities, *i.e.*, POS, ANT, and ISP to categorize the core triples to the corresponding NIHSS items. In combination, these three entities can accurately identify the corresponding NIHSS element. For example, "左" (left: POS)-"上肢" (arm: ANT) and "肌力" (inotrope: ISP) belongs to the 5a element of NIHSS, the strength of left upper arm. The POS entity represents location of certain symptom, thus can justify the weighting score.

For example, in NIHSS element 2, “左” (left: POS)-“眼睛” (eye)-“凝视” (gaze) scores 1, while “双侧” (both sides: POS)-“眼睛” (eye)-“凝视” (gaze) scores 2.

The Supplementary module *M_sup* consists of three types of entities, *i.e.*, BIN, CHG and DEG, indicating the existence and degree of changes. The BIN entity denotes existence or not. Existence entities such as “存在” (has) and “发现” (witness) do not change the score of the ISP-SPT REPs. No existence entities such as “不存在” (not exist) and “无” (no), would suggest no symptom, thus impact on the score; therefore, *score_r* in *D_core* needs to be take into consideration in stroke severity assessment. The CHG entities reflect changes in specific symptoms. The higher the NIHSS score, the more serious or positive of the stroke symptom. Positive changes, e.g,“恢复” (recovery), “改善” (improvement), etc. score -1. For negative changes, such as “恶化” (deterioration) and “加重” (aggravation), the corresponding score is 1. The DEG entity is a weight value that helps to precisely score the CHG entity.

The Synonym module *M_syn* is a thesaurus including the synonym pairs of the standard terms for all entities annotated in CSCR [129]. For example, “意识” (consciousness: ISP) is a KIE and it has synonyms, *e.g.* “神志” (sane) and “神智” (sanity). During the evaluation, this mapping is used to map the recognized entity to its corresponding NIHSS score, detailed in Section 4.2.3.

4.2.2 Chinese CNER with Domain-Specific Pre-trained Embedding

To boost the performance of CNER on stroke clinical EHRs, we pre-train a Chinese clinical embedding through domain-adaptive transfer learning on Chinese clinical EHRs. Then, we evaluate the SOTA deep learning models with multiple pre-trained embeddings on a public dataset and the CSCR dataset. The best-performing model that produces the most accurate and reliable CNER results is acquired to extract the relevant entities for the subsequent automatic stroke severity assessment, detailed below.

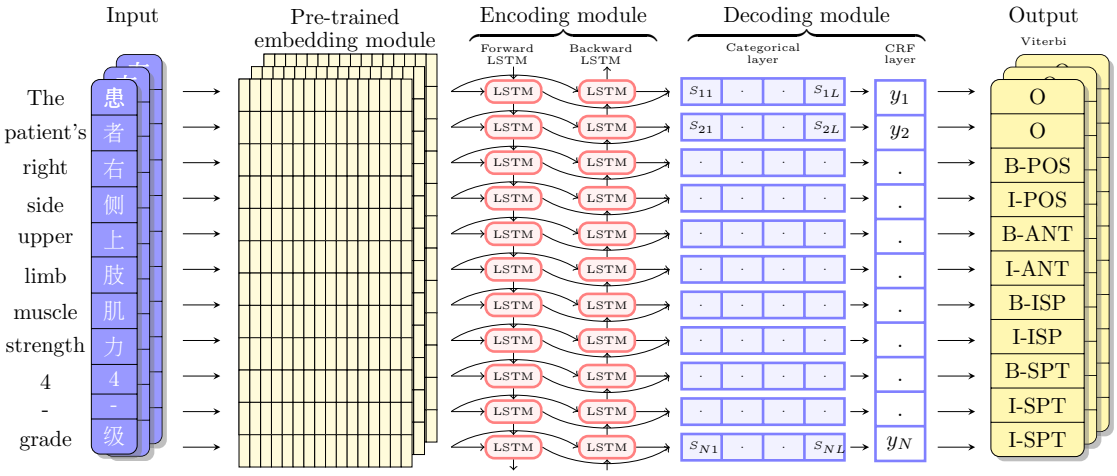

Figure 4.2 : The architecture of our baseline CNER model with pre-trained embedding. For the convenience of non-Chinese readers, the input sentence in Chinese is accompanied with a word-by-word gloss in English.

4.2.2.1 Domain-adaptive Pre-training of Chinese Clinical Embedding

An accurate clinical embedding is critical to high-performing CNER and accurate extraction of entities in EHRs, given that clinical documents typically contain a large number of medical terms that rarely occur and are weakly represented in general language models [19, 77, 106, 151, 154].

Inspired by the previous research that uses domain-adaptive pre-training, a transfer learning mechanism in cross-domain migration [32, 100, 135], we evaluate domain-adaptive pre-training based on the existing SOTA Chinese embeddings and Chinese clinical EHRs and select the Roberta-wwm [31] as the base model in training our clinical embedding. Different from previous research [19, 77, 151] that crawl general medical data from the Internet for pre-training, we use professional clinical EHRs from two data sources: EHRs provided by the partner hospital, and validated EHRs presented at the top Chinese medical conferences [22, 150, 162]. Therefore, it is much more relevant to real-world clinical practice than the other datasets, allowing

us to use a relatively small amount of EHRs to develop a clinical domain-specific embedding. We only use the training set of each dataset for pre-training, so that we could evaluate the test performance accurately on the test set.

Table 4.2 shows the pre-training corpora of our proposed CliRoberta and four baseline language models: word2vec [85], BERT-base [32], Roberta-wwm [31] and MC-BERT [151].

Table 4.2 : The pre-training corpora of language models. The training tokens are counted by Chinese characters.

Embeddings	Corpora Size	Training Tokens	Data Source
Word2vec	10.2MB	1.9M	EHRs from the training set of CCKS and CSCR datasets
BERT-base	-	0.4B	Chinese Wikipedia
ROBERTA-wwm-ext	-	5.4B	Chinese Wikipedia, news, Q&A, medical encyclopedia, etc.
MC-BERT	-	20.1M[1]	Chinese medical corpora, including Q&A, encyclopedia, EHRs, etc.
CliRoberta	72.3MB	8.0M	Selectively collected Chinese clinical EHRs from CBLUE, ChineseBLUE, CCKS, CHIP and CSCR datasets

[1] The training corpora of MC-BERT are reported by sentences rather than tokens [151].

[2] The corpora size with a "-" mark means no exact corpora size was reported.

The Word2vec embedding is derived from the classic Gensim library [108] based on EHRs from the training set of CCKS and CSCR dataset. It exhibits a corpora

size of 10.2MB, trained on 1.9 million tokens sourced from Electronic Health Records (EHRs) within the training sets of both CCKS and CSCR datasets. The BERT-base is the baseline Chinese BERT embedding generated by Google from Chinese Wikipedia [32]. The ROBERTA-wwm-ext is a SOTA variant of BERT with a larger training set size and better representative ability [31]. The reported training tokens for BERT-base and ROBERTA-wwm-ext are 0.4B and 5.4B, respectively, without specifying the exact corpora size *. Both BERT-base and ROBERTA-wwm-ext utilize Chinese Wikipedia as their foundation for pre-training corpora. However, ROBERTA-wwm-ext incorporate extended data, such as medical encyclopedias, news, and question-answering (Q&A) data obtained through web crawling†. The MC-BERT is a Chinese clinical language model by continual pre-training on a large volume of Chinese medical corpora, including biomedical question answering, medical encyclopedia, EHRs, and so on [151]. The training corpora are reported to encompass of 20.1M sentences. Our CliRoberta embedding is obtained through domain-adaptive pre-training based on ROBERTA-wwm-ext, using selectively collected Chinese clinical EHRs from publicly validated datasets, including CBLUE [150], ChineseBLUE [151], CHIP [22], and the training set of CCKS [162] and CSCR datasets. Together it has a size of 72.3MB, with a token count of 8.0M, measured in Chinese characters.

The training corpora are first preprocessed with a tokenizer [31] to fit the input requirements of the base model. We set the chunk size, *i.e.*, the maximum sentence length of the training data, to be 128, and the language model probability to 0.150, allowing up to 15.0% of words to be replaced with the "MASK" token in one sentence. The processed dataset consists of a training set of 120,000 text sequences and a validation set of 13,000, with a token count of 8.0M Chinese characters and a batch

*https://github.com/ymcui/Chinese-BERT-wwm

†https://meta.wikimedia.org/w/index.php?title=List_of_Wikipedias/zh&uselang=zh

size of 32. The maximum number of epochs is set to 100 until early stop criteria are triggered. For the use of the downstream task CNER, we save the embedding layers, forming a Roberta-style embedding, which we named "CliRoberta".

4.2.2.2 CNER with Pre-trained Embeddings

Based on the literature [58, 77, 78, 100, 152, 156, 159], we evaluate three widely-adopted deep learning-based CNER models: BiLSTM-CRF, BiGRU-CRF, and CNN-LSTM-CRF, and compare their performance on two datasets, CCKS2019 and CSCR.

As shown in Figure 4.2, the baseline CNER model consists of an input layer to process raw text sequences of EHRs into the model; a pre-trained embedding module using CliRoberta to represent the words into embedding vectors; an encoding module that consists of both a forward and backward LSTM to encode the semantic association among words in one sentence; a decoding module including a categorical layer to decode the sequence using a softmax loss function, and a CRF layer to regularize the categorization results using a CRF loss [75]; and an output layer to predict the sequence of labels using the Viterbi algorithm [71].

4.2.3 Automated NIHSS Scoring

Thus, with the embedding-based CNER model, all the entities that link to the relevant NIHSS items and their values are extracted. Based on these extracted entities, the automated NIHSS scoring is conducted in an end-to-end pipeline in three steps: extracting relational entity triples (RETs), constructing relational entity chains (RECs), and executing entity-to-NIHSS mapping.

4.2.3.1 Extracting Relational Entity Triples (RETs)

First, semantically associated entities must be paired into RETs according to their relationship or dependency. The existing Chinese dependency parsing methods

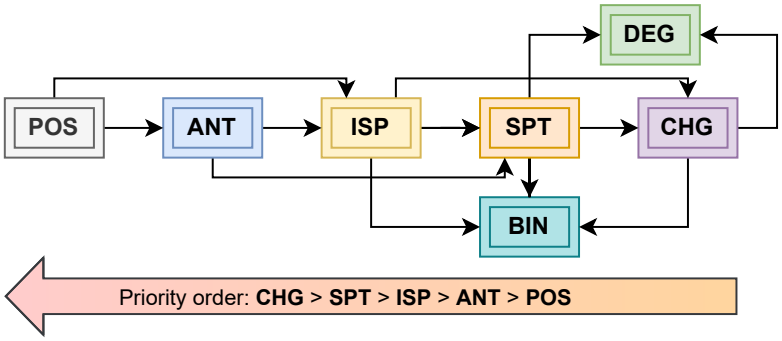

Figure 4.3 : The semantic dependency schema and priority order. The schema consists of seven types of entities and twelve relationships. The priority order defines the selection rule during the extraction of relational entity triples.

such as hanLP [50] and DDParser [153] are mainly based on word segmentation and part-of-speech (POS) tagging, which require a clear "subject-predicate-object" (SPO) structure in a sentence. However, these methods do not work well for the clinical stroke EHRs, in which most clinical terms are not arranged in a clear SPO structure [144]. Therefore, we design a RET extraction algorithm using a schema-constrained fuzzy matching method. The dependency schema includes twelve types of RETs that define the semantic dependencies between seven entity types (see Figure 4.3).

To implement the RET extraction algorithm based on the output sequence of the CNER model (see Figure 4.2), we need to segment the sequence first. Using the "Period" separator, the average length of the resultant sentence is 35 words and 10 entities. This large number of entities in a sentence makes it difficult for the application of the matching rule to accurately extract the RETs. Therefore, we experiment with segmenting the sentences by the "comma" separator, which reduces the segment length and entity counts to an average of 11 words and 3 entities, respectively. More than 90% of the resultant segments have less than 5 entities, and more than 50% of segments contain only 3 entities. Given that the associated entities

in EHRs are often close to each other within one segment, avoiding the long-distance dependency problem, applying the "comma" separator is appropriate for our RET extraction algorithm.

Figure 4.4 shows the RET extraction algorithm in detail. The algorithm iteratively loops through each entity in the segment and combines the current entity and the rest into candidate RETs. The candidate RETs are evaluated against the constraints outlined in the defined schema (as shown in Figure 4.3), where only those that meet the criteria are retained, and those that do not are eliminated. The remaining RETs are evaluated on an individual basis. If the entities are combined in a one-to-one form, it is considered a valid RET. If the entities are in a one-to-many form, a priority rule is applied to determine which RET is considered valid. In many-to-many relationships, a distance rule is applied to ensure one-to-one relationships and avoid cross-combinations. Any entities that have no relationship with other entities are treated as unary "BIN-ISP" RETs by default.

4.2.3.2 Constructing Relational Entity Chains (RECs)

An NIHSS score is not only determined by the ISP-SPT triples showing the inspected symptom, but also the ANT-ISP triples suggesting the location of the symptom, and the SPT-DEG triples measuring the severity degree of the symptom. Therefore, all RETs related to the same inspection item need to be connected into relational entity chains (RECs) and be mapped to the corresponding NIHSS items and values for automatic NIHSS scoring.

We first define the format of RECs as "POS-ANT-BIN & ISP-BIN & SPT-BIN & CHG-DEG" to incorporate all the relevant entities. Matching rules are then applied to generate RECs from RETs. Once a RET's target entity matches another RET's source entity, these two RETs are connected. This matching and connection operation is performed iteratively until all RETs in a sentence have been traversed

Algorithm 1 RET Extraction.

Input: A list of entities $E = [e_1, \ldots, e_i, \ldots, e_n]$, where e_i is the i-th entity in a text segment of n entities.
The defined schema $S = [ret_1, \ldots, ret_j, \ldots, ret_m]$, where ret_j is the j-th RET type in the form of {source entity, predicate, objective entity}, and m is the number of defined RET types;
A priority order list P for entity types.

Output: $RETs$;

$RETs \leftarrow \varnothing$;
for each i, e in enumerate E **do**
  $ret \leftarrow \varnothing$;
  // Count entities with e's entity type in E;
  $c \leftarrow E.count(e.type)$;
  // Combine every entity in E with e;
  $ret \leftarrow [\{e_1, e\}, \ldots, \{e_i, e\}, \ldots, \{e_n, e\}]$;
  // Filter by the schema S;
  $ret \leftarrow [e$ for e in ret if e in $S]$;
  if ret size = 1 **then**
    // One-to-one matching;
    $RETs \leftarrow ret$;
  else if ret size >1 AND c = 1 **then**
    // One-to-many matching, applying priority rule;
    $ret \leftarrow [\{e_k, e\}]$, where e_k has highest priority in P;
    $RETs \leftarrow ret$;
  else
    // Applying many-to-many matching rule;
    calculate the distance of each triple in ret;
    $ret \leftarrow [\{e_d, e\}]$, where e_d is closest to e;
    $RETs \leftarrow ret$;
  end if
  // Extract the unary RETs;
  if RET size = 0 AND e.type = "INSPECTION" **then**
    $ret \leftarrow [\{'has', e\}]$;
    $RETs \leftarrow ret$;
  end if
end for
return $RETs$

Figure 4.4 : The algorithm for rational entity triples (RET) extraction.

and formed a REC.

Figure 4.5 presents examples of RECs constructed from RETs and entities. In the figure, we illustrate examples of three RECs, and each of them is constructed from a different number of RETs and entities. The detailed construction process is shown in Section 4.2.3.2. Initially, following the Algorithm 4.4 for RET extraction and referencing the schema and priorities in Figure 4.3, entities are pairwise combined into various types of RETs. Subsequently, these RETs are connected iteratively in the REC format, forming valid RECs. With this approach, regardless of the entity types and RETs extracted, we have a corresponding method to standardize them into the canonical REC format. Therefore, its output can serve as the basis for subsequent quantitative mapping.

4.2.3.3 Executing Entity-to-NIHSS Mapping

Figure 4.6 illustrates our proposed automated quantitative stroke severity assessment pipeline, demonstrating the NIHSS scoring procedure using a raw EHR example in Chinese accompanied with an English translation. Initially, the EHR is input into our CNER model with pretrained embedding (refer to Section 4.2.2), to identify specific entity types in the text. Following Algorithm 4.4, RETs are extracted (see Section 4.2.3.1). Subsequently, RECs are constructed based on the extracted RETs, as detailed in Section 4.2.3.2. The entity-to-NIHSS mapping (utilizing Algorithm 4.7) is then executed, incorporating synonyms regulation and dictionary-based mapping through four sub-modules (see Section 4.2.1.2). Finally, the scores for each NIHSS item are aggregated to obtain the overall NIHSS score.

Based on the expert-constructed entity-to-NIHSS mapping dictionary in Section 4.2.1.2, we take the valid RECs as input to map the extracted entities to the corresponding NIHSS items and scores. The mapping dictionary has four modules, including M_core, M_cat, M_sup and M_syn, providing specific values of key pa-

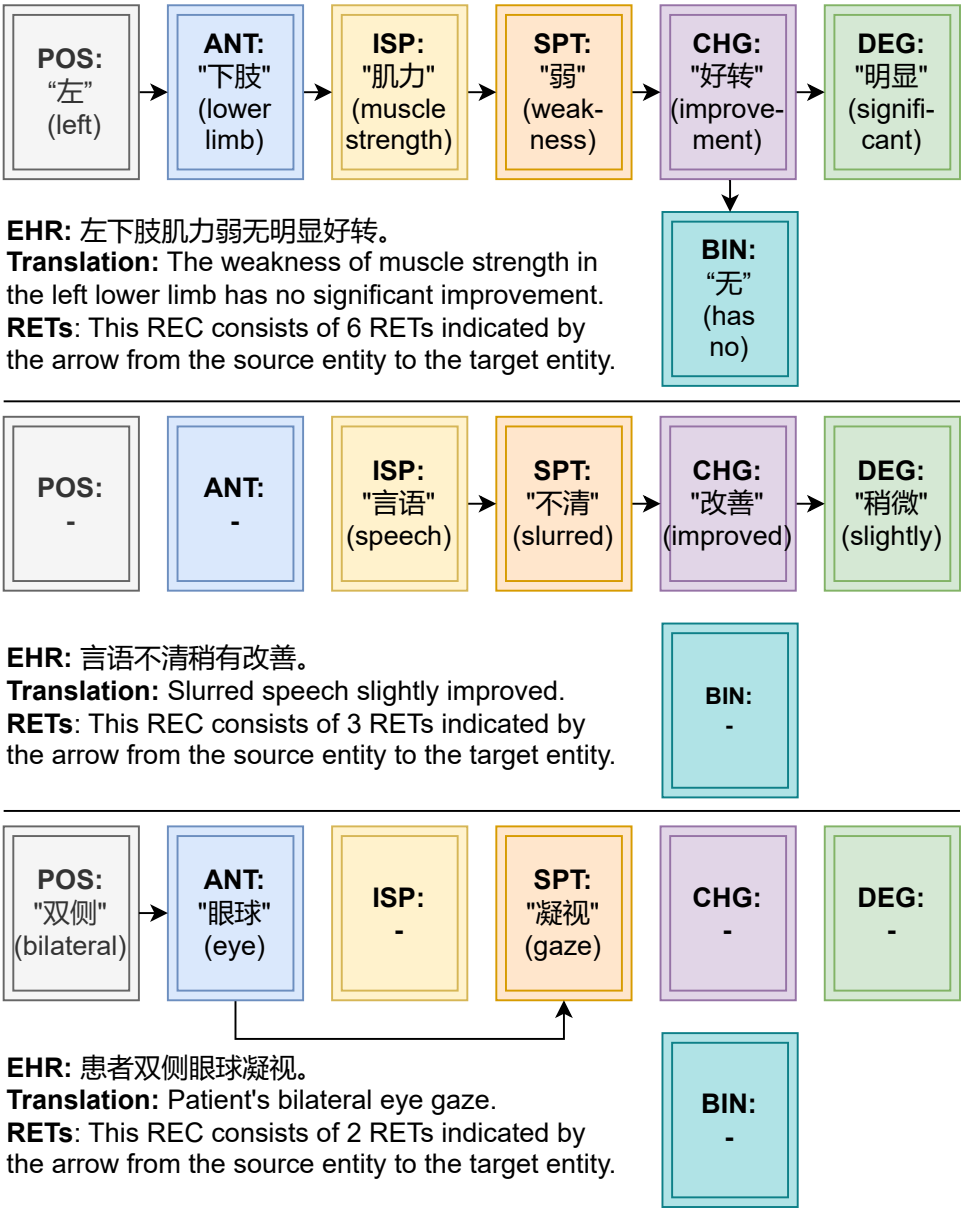

Figure 4.5 : Examples of RECs constructed from RETs and entities. We show three REC examples in the figure, and each of them is constructed from a different number of RETs and entities.

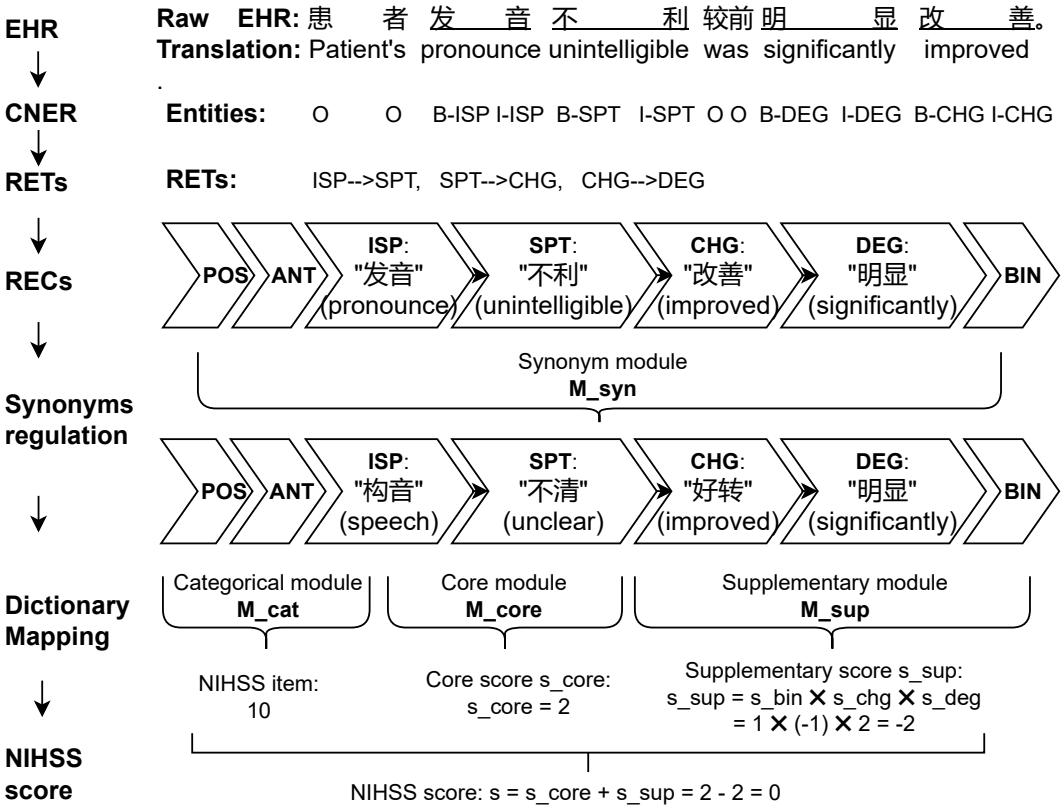

Figure 4.6 : The procedure of our proposed automated NIHSS scoring pipeline. The left side shows the steps from loading the EHR to capturing the NIHSS scoring result. The right side shows the detailed processing stages including raw EHR tokenization, entities extraction by CNER model with embedding, construction of RETs and RECs, synonyms regulation, dictionary-based mapping, and the calculation of NIHSS score as the final result.

rameters for the calculation of NIHSS scores. Figure 4.7 illustrates this detailed mapping process.

First, we conduct thesaurus-based entity resolution to regularize all entities in each REC with their standard terms. A query scans through the entities in a REC one by one, and once it finds a match with a term in the thesaurus M_syn, the raw entity will be replaced by the standard term in the M_syn.

We then use the ISP-SPT triple in the REC as a key to query for the core score s_core if the BIN entity exists in the M_core with value 1, and obtain s_core_r if the BIN entity has value 0. Next, we query the BIN, CHG, and DEG entities in the M_sup, to obtain the score of BIN, CHG, and DEG. These three scores are multiplied to generate the supplementary score. The core score and supplementary score are added, multiplied by the POS weight, and mapped to the corresponding NIHSS item by querying the POS, ANT, and ISP entities in M_cat.

After completing the above process for each EHR, we take the maximum score belonging to the same NIHSS item as the patient's resultant NIHSS item score. The addition of these item scores forms the total NIHSS score, indicating the quantified measurement of stroke severity.

4.3 Experiment Settings and Results

To demonstrate the effectiveness of our CNER model with the pre-trained embedding, in this section, we first conduct extensive experiments on a public dataset CCKS2019 [47] and our annotated CSCR dataset to evaluate the performance of the CNER models and embeddings. Five different random seeds are applied when initiating the CNER models in the experiments. Then, to demonstrate the feasibility of our proposed automatic quantitative stroke assessment approach, we evaluate the performance of our proposed approach by comparing its NIHSS scoring results with

Algorithm 2 Entity-to-NIHSS Mapping Algorithm.

Input: A set of RECs that belong to the same assessment, where each REC in the set $C_i = [POS_i, ANT_i, (BIN\&ISP)_i, (BIN\&SPT)_i, DEG_i, (BIN\&CHG)_i]$, for $i = 1, 2, \ldots, n_c$;
A mapping M that consists of four modules, M_core, M_cat, M_sup and M_syn.

Output: Total NIHSS score S

for each i=1 to n_c **do**
  // conduct thesaurus-based entity resolution;
  for each $Entity_i$ in C_i **do**
    if $Entity_i$ in M_syn **then**
      $Entity_i \leftarrow Entity_{standard}$;
    end if
  end for
  // select valid RECs;
  if ISP_i in M_core **then**
    continue;
  else
    break;
  end if
  // get the core score s_{i_core};
  find the $ISP_i - SPT_i$ triple in M_core;
  get the scores $score\_i$ and $score\_r\_i$;
  if $(BIN\&ISP)_i$ or $(BIN\&SPT)_i$ exists **then**
    find BIN_i in M_sup;
    get the score s_{BIN};
    if s_{BIN}=1 **then**
      $s_{i\_core} \leftarrow score\_i$;
    else
      $s_{i\_core} \leftarrow score\_r\_i$
    end if
  end if
  // get the supplementary score s_{i_sup};
  find $BIN\&CHG_i$ and DEG_i in M_sup;
  get the scoress_{BIN}, s_{CHG} and s_{DEG};
  $s_{i\_sup} \leftarrow s_{BIN} * s_{CHG} * s_{DEG}$;
  // sum up the scores;
  $s_i \leftarrow s_{i\_core} + s_{i\_sup}$;
  // categorize the scores to NIHSS items;
  if POS_i-ANT_i in $M\_cat|_{isp=ISP_i}$ **then**
    get the category index k and the weight w_i;
    update $s_i^k \leftarrow s_i w_i$;
  end if
end for
// Take the maximum value for each NIHSS item;
$s^k \leftarrow max_{1 \le i \le n_c} s_i^k$;
// Sum up to get the total NIHSS score;
$S \leftarrow \sum_{k=1a}^{11} s^k$, where k in $[1a, 1b, 1c, 2, 3, 4, 5a, 5b, 6a, 6b, 7, 8, 9, 10, 11]$
return S

Figure 4.7 : The Algorithm for Entity-to-NIHSS Mapping.

the ground truth provided by specialists.

4.3.1 Datasets

CCKS2019. This is a dataset of "Named Entity Recognition for Chinese Electronic Medical Records" [47], which was used in the CNER competition at the CCKS conference in 2019 [26]. It contains 1,000 training EHRs and 379 test EHRs and has six types of entities, *i.e.*, disease and diagnose, imaging examination, laboratory examination, operation, drug, and anatomy.

CSCR. The CSCR dataset (see Section 4.2.1) is curated by our research team. It consists of 1,931 EHR records generated in the clinical process of stroke assessment for two patient groups. We annotate seven types of entities for this data set. We select 1,545 EHRs from one patient group as the training and validation sets, and the rest 386 EHRs from the second patient group as the testing data set.The CSCR dataset is more specific to stroke diseases than CCKS2019, and it annotates most of the valuable entities that correspond to the NIHSS scoring system for stroke assessment. Therefore, this dataset is capable of capturing evidence necessary for the quantitative stroke severity assessment using the NIHSS system.

Table 4.3 shows the division of the two datasets in our experiments (see Section 4.3.1 for more details), where the training and testing EHRs come from different patient groups, with 1545 and 386 EHRs, respectively. During the training stage, the training and validation set is split into 1200 and 345 as training and validation sets. Detailed results are shown in Table 4.3.

Table 4.3 : The division of two datasets in the experiments. The numbers refer to the count of EHRs.

Dataset	Training	Validation	Testing
CCKS2019	600	400	379
CSCR	1200	345	386

4.3.2 Evaluation Metrics

4.3.2.1 CNER Evaluation Metrics

To measure the performance of the CNER models, we adopt the strict evaluation matrices used by the CCKS2019 competition [47], which include precision, recall and F1-score.

We denote the extracted entity set as S and the gold entity set as G. A correct prediction $s_i \in S$ is equal to $g_j \in G$, which means an exact match of the start positions, end positions and categories between the two entity sets S and G. The Precision (P_s), Recall (R_s) and F1-scores ($F1_s$) are defined as:

$$P_s = |S \cap_S G|/|S|, \tag{4.1}$$

$$R_s = |S \cap_S G|/|G|, \tag{4.2}$$

$$F1_s = 2P_sR_s/(P_s + R_s), \tag{4.3}$$

where $\cap_S$ represents the strict intersection of the prediction set S and the gold set G.

4.3.2.2 Kappa Coefficient

Cohen's Kappa coefficient [28] represents the inter-rater agreement between two raters, each assigning items into multiple categories, calculated as:

$$K = (p_o - p_e)/(1 - p_e), \tag{4.4}$$

where p_o represents the degree of agreement between the raters, which is obtained by dividing the observed number of agreements by the total rated records, and p_e represents the probability of each rater randomly rating each category and can be calculated by:

$$p_e = 1/N^2 \sum_k n_{k1} n_{k2}, \tag{4.5}$$

where k is the number of categories, N is the number of rated records, and n_{ki} denotes the times Rater i rated category k.

The Kappa score is a value ranging from -1 to 1. However, in actual applications, they usually fall in the range of $[0, 1]$. A Kappa coefficient < 0.2 indicates very weak agreement; $0.2 \sim 0.4$ indicates weak agreement; $0.4 \sim 0.6$ indicates moderate; $0.6 \sim 0.8$ indicates strong; and $0.8 \sim 1.0$ indicates perfect agreement [71].

4.3.2.3 Intraclass Correlation Coefficient (ICC)

The ICC [11] is one of the reliability coefficient indicators for evaluating the consistency of intraclass measurement. It is equal to the individual's variability divided by the total variability:

$$ICC = \sigma_\alpha^2 / (\sigma_\alpha^2 + \sigma_\varepsilon^2), \tag{4.6}$$

where σ_α^2 is the variance of α_j, an unobserved random effect among values in group j. The σ_ε^2 represents the variance of ε_{ij}, which is an unobserved noise term in the group j and i is the index of observation in the group.

The value of ICC ranges between 0 and 1. According to the guideline given by Koo and Li [64], an ICC below 0.5 indicates poor consistency, an ICC between 0.5 and 0.75 indicates moderate consistency, an ICC between 0.75 and 0.90 indicates good consistency, and an ICC above 0.9 represents excellent consistency.

4.3.3 Evaluation of CNER with Pre-trained Embedding

We compare our pre-trained CliRoberta with four language models: word2vec [108], BERT-base [32], Roberta-wwm [31] and MC-BERT [151] on processing CCKS2019 and CSCR datasets. We apply the widely-used evaluation metrics precision, recall and F1 score to evaluate model performance (see Section 4.3.2.1).

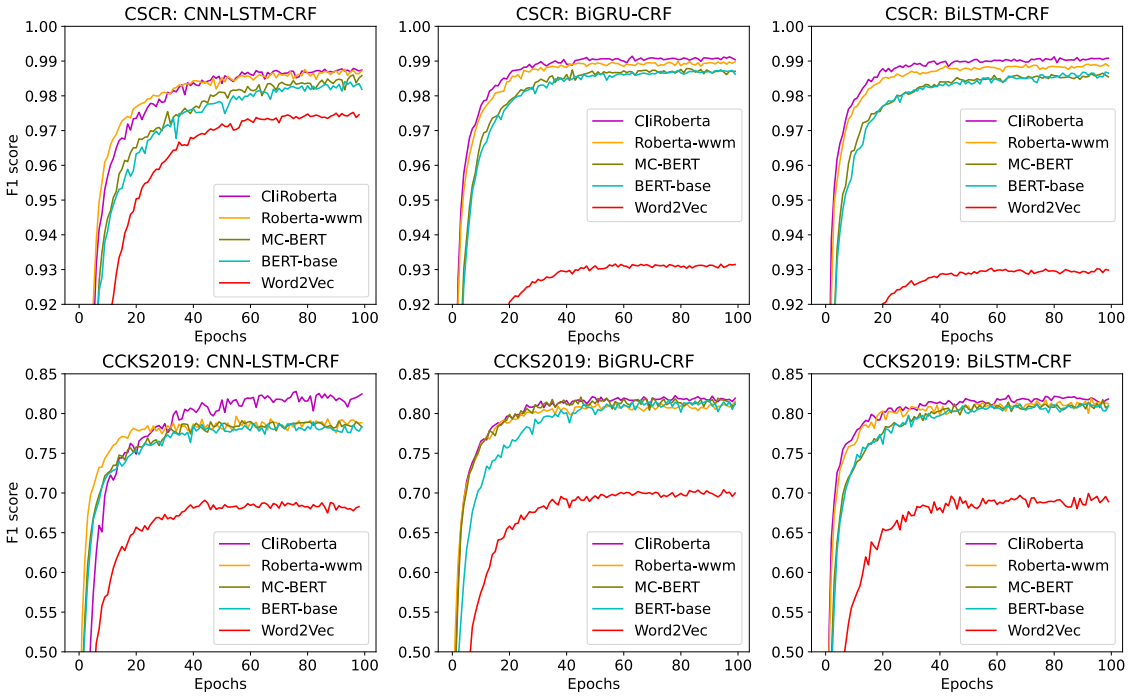

Figure 4.8 : The performance of three CNER models with five different pre-trained embeddings on the CSCR dataset (the top row) and the CCKS2019 dataset (the bottom row).

Table 4.4 shows the detailed experimental results for CNER models with pre-trained embeddings, on two different datasets, CCKS2019 and CSCR. Please refer to Section 4.3.3 for more details on the evaluation and discussion of the results shown in the table.

Figure 4.8 and Table 4.4 show that the four BERT-style embeddings significantly outperform the traditional word2vec model for both CCKS2019 and CSCR datasets.

Table 4.4 : The performance of the CNER models with pre-trained embeddings on CCKS2019 and CSCR dataset, by F1 score.

Dataset	Embeddings	CNN-LSTM-CRF	BiGRU-CRF	BiLSTM-CRF
CCKS2019	Word2Vec	0.696 ± 0.008	0.720 ± 0.007	0.696 ± 0.008
	BERT-base	0.804 ± 0.004	0.805 ± 0.007	0.804 ± 0.005
	Roberta-wwm	0.802 ± 0.005	0.808 ± 0.003	0.804 ± 0.003
	MC-BERT	0.807 ± 0.002	0.808 ± 0.005	0.806 ± 0.003
	CliRoberta	0.805 ± 0.008	0.813 ± 0.005	$\mathbf{0.814 \pm 0.002}$
CSCR	Word2Vec	0.933 ± 0.002	0.942 ± 0.001	0.937 ± 0.003
	BERT-base	0.970 ± 0.001	0.980 ± 0.001	0.978 ± 0.002
	Roberta-wwm	0.974 ± 0.002	0.980 ± 0.001	0.979 ± 0.002
	MC-BERT	0.970 ± 0.002	0.978 ± 0.002	0.974 ± 0.002
	CliRoberta	0.980 ± 0.001	0.988 ± 0.001	$\mathbf{0.988 \pm 0.002}$

Thanks to continual training, MC-BERT and Cliroberta improve over BERT-base and Roberta-wwm on CCKS2019 dataset, exhibiting higher F1 scores in CNER. However, on CSCR dataset, MC-BERT shows slightly worse performance than BERT-base, illustrating an uncertain and unsatisfying performance on our stroke dataset as we stated in Section 4.2.2. This observation motivates us to train our own Chinese clinical embedding through domain-adaptive pre-training to address the limitations of existing works. Table 4.4 demonstrates that our CliRoberta outperforms all pre-trained embeddings on the CCKS2019 and CSCR datasets, despite using a significantly smaller volume of selectively collected clinical corpora than other embeddings (see Table 4.2). To validate the significance of our domain-adaptive pre-training, we adopt the widely used paired sample $t - test$ [37] by comparing the

performance of our CliRoberta with the results of other embeddings one by one, on both CCKS2019 and CSCR datasets. The results are shown in Table 4.5, in which all *p_values* are less than 0.05, proving the significance and persistent improvement of performance by our pre-trained CliRoberta.

From the results of the three widely-used CNER models on both the CCKS2019 and CSCR datasets in Figure 4.8, we can see that the performance of BiGRU-CRF and BiLSTM-CRF is very close, both significantly better than that of CNN-LSTM-CRF. This result is in line with the previous findings [58, 100, 152]. We consider "CliRoberta-BiLSTM-CRF" as our baseline CNER model (see Figure 4.2) because it has achieved the highest F1 scores of 0.817 and 0.991 in CCKS2019 and CSCR datasets, respectively.

Moreover, all the compared models exhibit much higher CNER accuracy on CSCR, since the CSCR dataset is specifically designed stroke-related EHRs and also has a much higher level of annotation density.

Table 4.5 : The results of the statistical significance test of our pre-trained CliRoberta with four existing pre-trained embeddings. The *p_value* from paired sample $t-test$ are reported in the table.

Dataset	CliRoberta *vs* Word2Vec	CliRoberta *vs* BERT-base	CliRoberta *vs* Roberta-wwm	CliRoberta *vs* MC-BERT
CCKS2019	2.565E-14	2.213E-2	1.255E-2	3.083E-2
CSCR	1.227E-18	9.105E-10	1.282E-8	2.541E-11
Overall	1.033E-11	2.681E-06	1.732E-06	2.857E-07

4.3.4 Results of Automated NIHSS Scoring

We conduct automated, end-to-end NIHSS scoring on 33 randomly selected real-world stroke clinical EHRs and compare the results against two comparison groups. The first comparison group is composed of three stroke specialists, all from top tertiary hospitals with extensive clinical experience in stroke and a high level of clinical competence. The second group is made up of three volunteers who have received a brief NIHSS scoring training.

After the three stroke specialists independently perform NIHSS stroke severity assessment on the 33 EHRs, we synthesize their scores based on the voting principle, *i.e.*, the "mode" value with the highest frequency of occurrence is taken as the ground truth item score. The total NIHSS score is the sum of the scores of each NIHSS item. In the case that the three specialists gave three different scores, thus there is no "mode" value, a meeting is called to discuss and review to reach a consensus score.

4.3.4.1 Statistical Evaluation of NIHSS Scoring

Table 4.6 presents the descriptive statistics of the NIHSS scores given by the specialist group, the volunteer group, and our automatic method. It also presents the voted scores of the specialists, *i.e.*, the ground truth. The scores given by the three stroke specialists are relatively close, with a mean value and standard deviation of around 5.500 and 6.300, respectively. The ground truth has a mean value of 5.424 and a standard deviation of 6.383. From the figure, we can also see that the volunteer group gave lower scores for both mean and standard deviation. Among them, there is a gap of 5 points between the maximum score of volunteer 1 and the ground truth.

All of these indicate that people with less experience are prone to missing points in scoring and have difficulty in accurately assessing all NIHSS items. The results produced by our automatic method has a mean score of 5.394, only 0.030 lower than the ground truth, and a standard deviation of 6.123, only 0.260 less than the ground

truth; therefore, the performance of our method is statistically close to that of the ground truth, and much better than that of the volunteer group.

Table 4.6 : Descriptive statistics of NIHSS scoring results by assessors and the ground truth generated by specialists.

Assessor	EHR count	Mean	Std. Dev	Min	Max
Specialist 1	33	5.455	6.099	0	25
Specialist 2	33	5.515	6.350	0	24
Specialist 3	33	5.485	6.491	0	24
Ground Truth	33	5.424	6.384	0	24
Volunteer 1	33	5.212	5.430	0	19
Volunteer 2	33	4.939	5.979	0	25
Volunteer 3	33	5.394	5.932	0	24
Our method	33	5.394	6.123	0	24

4.3.4.2 Evaluation with Kappa and ICC

We further adopt the widely used inter-rater Kappa coefficient [23, 28] (see Section 4.3.2.2) and the Intraclass Correlation Coefficient (ICC) [11] (see Section 4.3.2.3) to assess the level of agreement and consistency of our proposed method with the ground truth. The Kappa coefficient is computed individually between each assessor's result and the ground truth, while each ICC coefficient is calculated within the specialist group, volunteer group, and the group comprising the ground truth with the scoring result of our method, respectively.

As shown in Table 4.7, the Kappa coefficient of the NIHSS scores provided by

the specialists all exceed 0.755 and the ICC coefficient of 0.983, demonstrating perfect agreement and excellent consistency upon high confidence of over 0.600 and 0.910 confidence intervals (CI95%) for the lower and upper limits, respectively. The p-values for all assessors are very low, indicating statistically significant for both Kappa and ICC tests.

These findings suggest that specialists have reached a reliable ground truth. Conversely, the volunteer group shows the lowest Kappa (0.552) and ICC (0.920) with the lowest confidence intervals of 0.374 and 0.730, indicating instability and inaccuracy in scoring by less experienced raters. Our automatic stroke severity assessment method has obtained a Kappa coefficient of 0.823 and an ICC value of 0.986, perfectly agreeing with the ground truth.

Between patients with mild and severe stroke severity, the results shown in Table 4.8 indicate that our model demonstrates good reliability in scoring both mild and severe stroke patients, with ICC values of 0.923 and 0.783, respectively (see Section 4.3.4.3). However, it is noteworthy that the ICC exhibits relatively lower reliability in the severe stroke patients' group, deviating by 0.217 points from the ground truth. This discrepancy is consistent with the majority of errors witnessed by our method. It arises from the compromised mental and physical states of severe stroke patients, making it difficult to fully cooperate in completing all NIHSS assessment items. The poorer quality in describing severe stroke patients' conditions in the EHRs contributes to our method's inability to achieve the same level of reliability observed in the mild stroke patient group.

Furthermore, completing the NIHSS stroke assessment requires experienced specialists an average of 5 to 8 minutes, the volunteers 13 to 17 minutes, and the automatic stroke assessment 0.1 minutes. Overall, the results suggest that our method is a reliable and precise alternative to specialists for assessing stroke severity,

Table 4.7 : Evaluation of NIHSS scoring results by Kappa and ICC metrics for different assessors. The "Time" is the average time taken for scoring one EHR. The "CI95%_low/up" represent the lower and upper limit of the 95% confidence interval, respectively.

Assessor	Time (min)	Kappa				ICC			
		Value	*p*-value	CI95%_low	CI95%_up	Value	*p*-value	CI95%_low	CI95%_up
Specialist 1	5	0.755	3.374E–33	0.600	0.910	0.983	6.647E–26	0.970	0.990
Specialist 2	8	0.823	1.003E–38	0.682	0.964	0.996	3.076E–36	0.990	1.000
Specialist 3	6	0.857	7.960E–40	0.729	0.985	0.999	1.228E–43	1.000	1.000
Volunteer 1	13	0.552	4.093E–30	0.487	0.830	0.920	3.126E–15	0.850	0.960
Volunteer 2	14	0.578	3.684E–21	0.404	0.752	0.953	7.546E–19	0.910	0.980
Volunteer 3	17	0.658	1.601E–22	0.374	0.730	0.970	4.270E–22	0.940	0.990
Our method	0.1	0.823	7.656E–38	0.686	0.959	0.986	2.525E–27	0.970	0.990

and it significantly enhances efficiency.

4.3.4.3 ICC for Mild and Severe Stroke Patients

Table 4.8 presents the ICC values and associated statistical measures for stroke severity assessments conducted by our proposed method against the ground truth. The assessments are categorized into two groups based on the severity of stroke, distinguishing between mild (NIHSS score between 5 and 15) and severe cases (NIHSS score larger than 15) [17].

Table 4.8 : The variations in the ICC values between patients with mild and severe stroke severity. The "CI95%_low/up" represent the lower or upper limits of the 95% confidence interval, respectively.

Assessor	Mild Stroke Severity				Severe Stroke Severity			
	Value	*p*-value	CI95% _low	CI95% _up	Value	*p*-value	CI95% _low	CI95% _up
Our method	0.923	0.001	0.660	0.990	0.783	0.035	0.000	0.990

Notably, our method yielded an ICC of 0.923 for patients with mild stroke severity, with a significant p-value of 0.001 and a 95% CI ranging from 0.660 to 0.990. For patients with severe stroke severity, our method produced an ICC of 0.783, accompanied by a p-value of 0.035 and a 95% CI spanning from 0.000 to 0.990. The results indicate that our model demonstrates good reliability in scoring both mild and severe stroke patients, with ICC values of 0.923 and 0.783, respectively.

However, it is noteworthy that the ICC exhibits relatively lower reliability in the severe stroke patients' group, deviating by 0.217 points from the ground truth. This discrepancy can be attributed to the challenging conditions faced by severe stroke patients, as their compromised mental and physical states make it difficult to fully cooperate in completing all NIHSS assessment items. The poorer quality in describing severe stroke patients' conditions in the EHRs contributes to our method's inability to achieve the same level of reliability observed in the mild stroke patient group.

4.4 Discussion

4.4.1 Principal Findings

In this chapter, we have developed an automatic, quantitative stroke severity assessment framework that has made the following contributions to clinical research and practice:

(1) We constructed a Chinese CNER dataset named CSCR through semi-automatic annotation and expert verification. To the best of our knowledge, this is the first stroke-specific Chinese CNER dataset with densely annotated, semantically related entities, which reliable medical knowledge base can be used in further downstream applications based on Chinese EHRs [100, 135].

(2) We produced a discriminative Chinese clinical embedding named CliRoberta that outperforms the existing general and medical Chinese embeddings [31, 32, 151], in Chinese EHR representation. This, once again, demonstrates that the domain-adaptive pre-trained Chinese clinical embedding is promising in clinical applications such as CNER, with fewer data volumes and superior performance.

(3) Guided and verified by stroke specialists, we defined the entity-to-NIHSS mapping, which supports our algorithm to automatically extract entities and relationships, achieving automated NIHSS scoring. It also supports novice volunteers without prior clinical knowledge to quickly learn and conduct the labor-intensive and time-consuming manual stroke severity assessments.

(4) Finally, we developed an automated, end-to-end NIHSS scoring method based on CNER results, whose effectiveness is proven by its high inter-rater reliability and excellent intraclass consistency compared with the ground truth established by the stroke specialists, both exceeding that of the novice assessors. Furthermore, our method has significantly reduced assessing time from minutes to seconds, thus

improving the efficiency of the assessment. The automated assessment of the NIHSS score is particularly valuable for retrospective studies where information about the NIHSS is missing but EHRs are accessible. In such cases, stroke-related neurological deficits can be extracted and converted into the NIHSS score for analysis.

4.4.2 Advancements and Limitations

For the first time, we have designed a suite of AI-aided, knowledge-based NLP models to automate the clinical stroke severity assessment from Chinese EHRs. Different from previous works [54, 98, 141, 144], our approach extracts intensively annotated, semantically related entities directly from diagnostic notes within Chinese EHRs for automated, end-to-end NIHSS scoring. The ability to incorporate natural language in notes into symptom assessment affords our algorithms higher levels of interpretability, precision, and reliability compared with the previous studies. Furthermore, the whole suite of the AI algorithms and methods designed in this research can be easily replicated or referenced for automated clinical assessments, not only for stroke assessment in particular, but also for clinical assessment in general. It is designed to be applicable to other languages with minimal adjustments, such as replacing the original text with the target language.

A limitation of this research is that the number of test cases is relatively small, with only 33 EHRs. We cannot test the applicability of the created algorithms without trialling to use them on real-world stroke assessment in a clinical setting.

4.5 Summary

In summary, this chapter designs, implements, and evaluates a suite of AI-based models to automate the clinical assessment task using Chinese EHRs. We effectively applied domain-specific transfer learning to improve the embedding at the pre-training stage and applied deep learning techniques to produce a high-performing CNER

model. Then, we designed a novel entity-to-NIHSS mapping for stroke severity scoring following an end-to-end approach. This represents a novel approach toward a more effective and automatic assessment of stroke severity in an objective way. The reliability and consistency of our automatic stroke severity assessment method have been demonstrated by its comparable performance with the ground truth, and better performance than that of a volunteer group. Furthermore, our method has significantly reduced assessment time from minutes to seconds, thus improving the efficiency of stroke assessment.

At the same time, we also observed that despite the high accuracy of our method, this framework requires close collaboration of multiple steps and the support of expert knowledge for tasks such as dataset construction and dictionary creation. With the current popularity of large language models (LLMs), their emergence has shown promise in natural language understanding (NLU) and inference tasks, making them suitable for automating the often labor-intensive, time-consuming, and tedious analysis tasks in EHRs. To address this, we propose an LLM-based prompting paradigm to achieve automated stroke assessment in a more comprehensive manner, which is detailed in the following chapter.

Chapter 5

Empowering LLMs for Automated Clinical Assessment using EHRs

Despite the solid performance of our proposed framework in the previous chapter that achieves accurate stroke assessment results based on CNER and pre-training, there is still potential for further improvement in both the workload and automation levels. Recently, LLMs have demonstrated significant proficiency in natural language understanding (NLU) and processing, offering promise for automating the typically labour-intensive and time-consuming analytical tasks with EHRs. Despite the active application of LLMs in the healthcare setting, many of the foundation models lack real-world healthcare relevance, applying LLMs to EHRs is still in its early stage. To advance this field, we pioneer a generation-augmented prompting paradigm "GAPrompt" to empower generic LLMs for automated clinical assessment, in particular, quantitative stroke severity assessment in this study, using data extracted from EHRs.

The GAPrompt paradigm comprises five components: (i) selection of LLM driven by prompt, (ii) construction of a knowledge base augmented by generation, (iii) summary-based generation-augmented retrieval (SGAR); (iv) inferencing with a hierarchical chain-of-thought (HCoT), and (v) ensembling of multiple model outputs.

GAPrompt addresses the limitations of generic LLMs in clinical applications in a progressive manner. It efficiently evaluates the applicability of LLMs in specific tasks through LLM selection prompting, enhances their understanding of task-specific knowledge from the constructed knowledge base, improves the accuracy of

knowledge and demonstration retrieval via SGAR, elevates LLM inference precision through HCoT, enhances generation robustness, and reduces hallucinations of LLM via ensembling. Experiment results demonstrate the capability of our method to empower LLMs to automatically assess EHRs and generate quantitative clinical assessment results.

Our study highlights the applicability of enhancing the capabilities of foundation LLMs in medical domain-specific tasks, *i.e.*, automated quantitative analysis of EHRs, addressing the challenges of labor-intensive and often manually conducted quantitative assessment of stroke in clinical practice and research. This approach offers a practical and accessible GAPrompt paradigm for academic researchers and industry practitioners seeking to leverage the power of LLMs in domain-specific applications. Its utility extends beyond the medical domain, applicable to a wide range of fields.

5.1 Background

Hospitals and medical practices around the world have increasingly adopted electronic health record (EHR) systems, resulting in massive amounts of electronic patient data in both structured (*e.g.*, disease codes, medication codes) and unstructured (*i.e.*, clinical narratives such as progress notes) formats. The advancements in AI techniques, including machine learning, deep learning, and natural language processing (NLP), have provided researchers with powerful techniques to automate the methods and process of secondary data analysis to support clinical decisions and research based on these massive amounts of EHR data [54, 63, 95, 141]. Currently, the EHR data analytic methods encounter several significant limitations. These include the requirement for large volumes of labeled datasets for model training, the necessity for entity (health terms) and relationship annotation, labor-intensive preprocessing procedures, and inadequate quantitative assessment capabilities [63, 98, 141].

The recent large language models (LLMs) hold remarkable capability in natural language understanding (NLU) and natural language inference (NLI) [104, 105]. They can comprehend and answer questions directly for a given text, surpassing the classical machine learning and deep learning methods, which require sentence-by-sentence or word-by-word processing and annotation [18]. Therefore, these LLMs are highly promising AI techniques for enhancing EHR analytic technologies to improve the quality and productivity of healthcare services. However, it remains a challenge to directly apply these LLMs in real-world domain-specific tasks [70, 113], because most generic LLMs are trained on general language data and lack domain-specific knowledge [127], while the very few medical domain LLMs are proprietary and not publicly available [65, 117, 118]. Also, there is little report about the application of LLMs in quantitative clinical assessment tasks.

Previous studies have demonstrated that with appropriately designed prompting strategies, generic LLMs can achieve comparable performance to domain-specific LLMs without the time-consuming and costly training or fine-tuning of LLMs [36, 131, 143]. Therefore, we explored the feasibility of applying prompting techniques to enable generic LLMs in completing our clinical assessment task of stroke severity. However, our initial research has found that there are several main challenges of applying generic LLMs directly in implementing automated stroke assessment. These include the evaluation of the applicability of foundation LLMs, the lack of stroke assessment knowledge, the limited context length in processing large EHRs, the inaccuracy of reasoning quantitative assessment results with NIHSS, and the inevitable hallucination during the generation. By leveraging the in-context learning (ICL) ability of LLMs, in this paper, a series of prompting strategies, including prompt-driven LLM selection, generation-augmented knowledge base construction, generation-augmented retrieval (GAR), hierarchical chain-of-thought (HCoT), and an ensembling mechanism, are developed to tackle these issues, empowering LLMs

for automated quantitative clinical assessment from EHRs.

First, with the popularity of LLMs, a plethora of new models are continuously emerging. However, their applicability and performance are not clear in quantitative clinical assessment tasks. Thus, first and foremost, an effective and efficient prompting-based LLM selection approach is needed. Next, to enhance the LLM's knowledge of stroke assessment, retrieval-augmented generation (RAG) [73] is a suitable solution. This process first constructs an external knowledge base compromising stroke assessment guidelines and demonstrations, using the well-established National Institutes of Health Stroke Scale (NIHSS) [17] as the quantitative stroke assessment standard, and generating demonstrations from expert-validated assessment results on a labeled EHRs dataset CSCR [38].It also utilizes a generation-augmented retrieval (GAR) [83] method to accurately extract the corresponding assessment criteria and demonstrations to guide the generation [73], helping to improve the LLM reasoning performance.Subsequently, an HCoT prompting strategy that integrates the document-level macro sequential chain [72] and sentence-level micro chain-of-thought [134], is proven to be effective to overcome the challenges of LLM's limited context length in processing large EHRs, and improve the performance of LLM's inference. Using the popular Langchain library [72], large EHRs are split into short sentences and sequentially processed by LLM inferencing. Meanwhile, the CoT technique [134] is capable of significantly improving the performance of LLM inference through logical solutions provided in the demonstrations. Finally, an ensembling strategy [4] is applied to integrate multiple generation results to control the impact of LLM's hallucination in generation.

In this chapter, integrating the aforementioned promoting strategies, we develop the overarching generation-augmented prompting paradigm named "GAPrompt". This paradigm effectively extends the capability of generic LLMs, thus substantially empowering their proficiency in clinical domain-specific applications, *i.e.*, achieving

automated quantitative stroke severity assessment.

The remainder of this chapter is organized as follows: Section 5.2 details our methods of the GAPrompt paradigm, including the prompt-driven LLM selection, generation-augmented knowledge base construction, summary-based generation-augmented retrieval, hierarchical chain-of-thought, and the ensembling approach. The experiment design and results are presented in Sections 5.3 and 5.4, respectively, followed by a discussion and conclusion in Section 5.5.

5.2 Method

We propose a novel prompting paradigm named GAPrompt that applies retrieval augmented generation to enhance the capabilities of the generic LLMs within our stroke assessment application. GAPrompt specifically addresses the limitations of LLMs including their uncertain applicability, lack of stroke assessment knowledge, limited context length, inaccuracy in quantitative reasoning, and the issue of hallucination.

Our proposed GAPrompt paradigm comprises five process components: (i) prompt-driven LLM selection (in green); (ii) generation-augmented knowledge and demonstrations construction (in blue); (iii) summary-based generation-augmented retrieval (SGAR) (in orange); (iv) hierarchical chain-of-thought (HCoT) (in pink); and (v) ensembling of multiple generations (in purple), as shown in Figure 5.1.

5.2.1 Prompt-driven LLM Selection

To evaluate the applicability of candidate generic LLMs for our specific application scenario, we first devise a prompt-driven LLM selection strategy (see Figure 5.1 in green). In this strategy, we create six prompt templates to evaluate the capabilities of candidate LLMs in the following six aspects: the foundational knowledge required for stroke severity assessment ("Knowledge"), comprehension of stroke-related knowledge

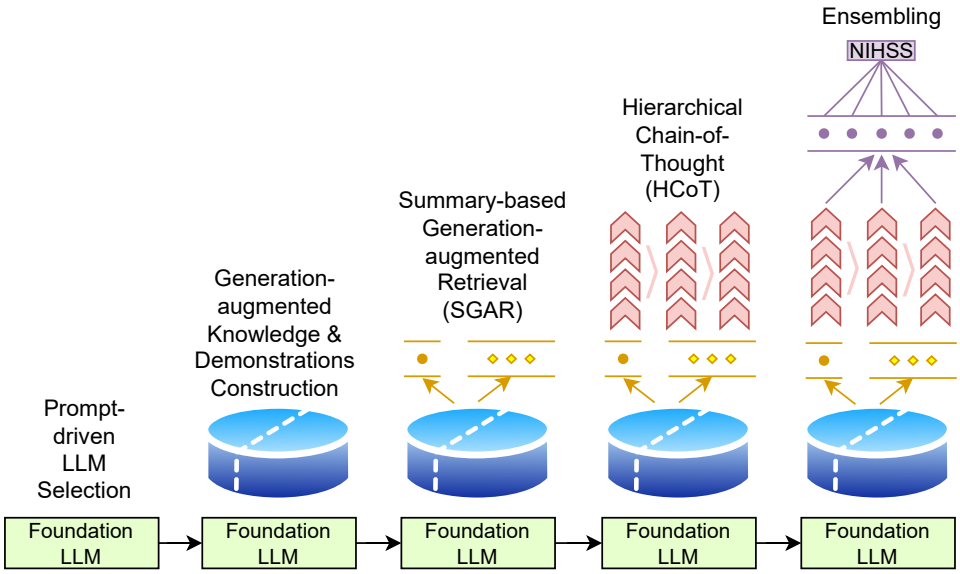

Figure 5.1 : The architecture of our proposed GAPrompt paradigm. Green color: prompt-driven LLM selection; blue color: generation-augmented knowledge base construction; orange color: summary-based generation-augmented retrieval (SGAR); pink color: hierarchical chain-of-thought (HCoT); purple color: ensembling.

and memory capacity ("Understanding"), learning from the few-shot examples about stroke ("Learning"), chain-of-thought (CoT) reasoning ("Reasoning"), ensuring consistency in the generated outputs ("Consistency"), and controlling hallucinations ("Anti-hallucination"). Figure 5.2 presents examples of the detailed format of each prompt template.

The Knowledge prompt, "Tell me the definition of the National Institute of Health Stroke Scale (NIHSS) and its scoring criteria", requires a highly specialized response. It assesses an LLM's foundational knowledge in stroke assessment using NIHSS. In the Understanding prompt, we first present a comprehensive definition of NIHSS along with its scoring criteria, afterwards we pose a similar question to evaluate the LLM's comprehension within the given context. The Learning prompt presents examples in a question-answer format and concludes with a similar question to check

if the LLM can learn from these examples. In the Reasoning prompt, we provide a logical reasoning demonstration in question-answer form, followed by a similar question to assess the LLM's capability to learn logical reasoning from examples. The Consistency prompt repeats a question five times to examine the consistency of the LLM's responses. Finally, in the Hallucination prompt, we pose an initial question and then ask an unrelated one (such as "The patient's speech is unclear. So, what is the patient's muscle strength level on the left leg?") to evaluate the LLM's ability to control hallucinations.

While the above evaluation prompts may not comprehensively assess an LLM's capabilities, they establish a systematic method to assess the performance of generic LLMs in the specific context of stroke severity assessment, and identify the foundation LLM that meets our task requirements. With this process, we have identified the best-performing model LlaMa2-70b from all the models that we evaluated to execute our quantitative stroke assessment tasks.

5.2.2 Generation-augmented Knowledge Base Construction

Two types of external knowledge are required for LLMs to effectively perform the task of quantitative assessment of stroke severity using EHRs: task-specific knowledge and demonstration of the task execution procedure. The former refers to the measurable NIHSS assessment criteria, and the latter are the examples given to the machine for task execution.

Task-specific Knowledge. In our evaluation of LLM performance during the prompt-driven LLM selection process (Section 5.2.1), we have observed that, while LLMs possess a fundamental understanding of stroke assessment, they struggle with consistently identifying assessment items and assigning precise NIHSS scores in reasoning [17]. Therefore, we integrate an explicit NIHSS assessment guideline*

*https://www.ninds.nih.gov/health-information/public-education/know-stroke/health-

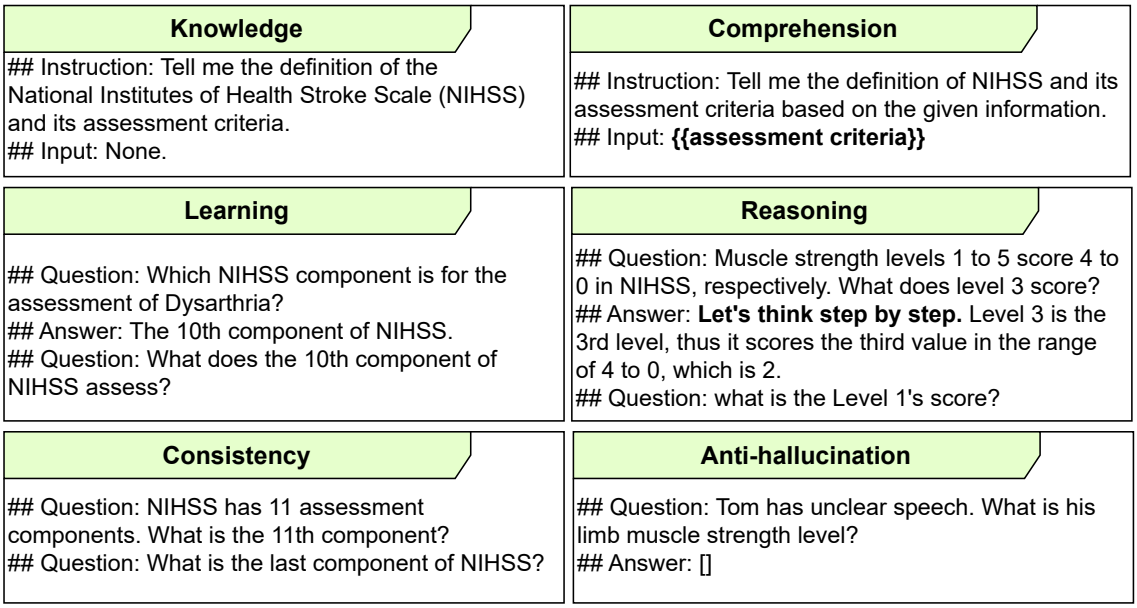

Figure 5.2 : The six prompt templates applied to select the optimal foundation LLM. Six capabilities of LLMs, including Knowledge, Comprehension, Learning, Reasoning, Consistency, and Anti-hallucination, are evaluated using these defined prompts.

as an external task-specific knowledge to support the foundation LLM to improve performance in this task. The NIHSS assessment protocol comprises 11 components, each with distinct assessment objectives and scoring criteria, and varying score ranges. To facilitate this integration, we employ a commonly applied sentence-transformer embedding, "all-mpnet-base-v2" [114], to first convert the assessment criteria of each NIHSS item into fixed-size vectors, and then store these vectors in a database for subsequent queries.

LLM-generated Demonstrations. Previous research works have featured the significance of using demonstrations to improve the performance of LLMs in text generation tasks [118, 134]. They have also explored the potential of substituting manually composed examples with LLM-generated demonstrations. In accordance with the findings that LLMs can automatically generate CoT examples and make

professionals/nih-stroke-scale

corrections based on the given ground truth [91, 143, 158], we introduce the following prompt template, as shown in Figure 5.3, for LLMs to generate demonstrations.

Prompt Template for LLMs to Generate Demonstrations

Context:**{{assessment criteria}}**
Instruction: Please follow the assessment criteria to assess the scores of each NIHSS component from the following report.
{{report}}
Let's think step by step.
1. If the report is not in English, translate it to English first.
2. Determine which components of NIHSS are related to the report, and assess the score.
3. Not mentioned components score 0.
4. Correct the answer according to the ground truth for each component:
{{ground truth}}

Figure 5.3 : The template used by LLMs to generate demonstrations.

5.2.3 Summary-based Generation-augmented Retrieval

Retrieval is a pivotal step in our prompting approach. Previous research has shown that dynamic retrieval, which takes into account the content of each query to accurately retrieve highly relevant demonstrations, significantly improves the overall quality of CoT prompting [91, 158]. Furthermore, the GAR method [83] that uses LLMs to augment the query content has proven effective in enhancing retrieval accuracy. In light of these insights, we propose an innovative summary-based GAR (SGAR) approach that employs LLM-generated summaries to improve retrieval accuracy.

Unlike previous methods that focused solely on enhancing query generation, our approach introduces the concurrent LLM-generated summarization of both the input query and the external knowledge base. This dual summarization approach enables the query to capture essential information at the sentence level, guided by the external knowledge base. Meanwhile, it compresses the information at the document or paragraph level in the knowledge base. This bidirectional compressing process

improves accuracy of matching in retrieval (see Figure 5.1 in orange).

We first define the summarization criteria that focus on retrieving information related to anatomy, inspection, and symptoms, to achieve our research objective of stroke severity assessment based on patients' EHR data (see Figure 5.4). We then instruct the LLM to use these criteria to generate summaries of the raw text. For knowledge records, we apply LLM-based summarization to each paragraph. However, in the case of demonstrations, our focus is solely on summarizing the EHR contained within the question portion, disregarding the answer section.

Prompt Template for LLM-augmented Summarization

Instruction: Please summarize the given text to capture only keywords related to the anatomy, inspection items, and symptoms.
Input: **{{EHR or Knowledge or Demonstration}}**

Figure 5.4 : The prompt template used for LLMs to generate summarization.

After the augmentation process, both the knowledge summary and demonstration summary are embedded with sentence-transformers [114] and then saved as metadata in the vector database corresponding to each record. During the retrieval process, the algorithm searches and matches the summarized query vector with the metadata of the knowledge and demonstrations. Upon a successful match, the raw knowledge and demonstrations are returned, ensuring the preservation of information from the knowledge base without loss.

5.2.4 Hierarchical Chain-of-Thought

To address the limitations of the foundational LLMs that we have encountered, including limited context length and inference error, we introduce two techniques - macro sequential chain and micro CoT, and encapsulate them in our method entitled Hierarchical Chain-of-Thought (HCoT) (see Figure 5.1). The macro sequential chain

breaks down the complex assessment of large EHRs into small, sequential steps, thus helping LLMs to think logically and infer step by step. chain-of-thought has been empirically validated as an effective method for prompting LLMs. It enables systematic reasoning in alignment with the logic flow of the few-shot examples [134, 158].

5.2.4.1 Macro Sequential Chain

Leveraging the Langchain platform [72], we traverse each EHR data through five sequential chains - splitting, translation, retrieval, micro CoT, and summarization (see Figure 5.5). The output of one chain serves as the input for the next chain. Distinct prompt templates are applied at different chains to achieve each one's intended purpose.

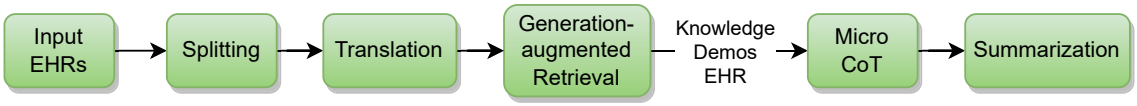

Figure 5.5 : The macro sequential chain. The macro sequential chain includes five steps: splitting, translation, retrieval, micro CoT, and ensembling .

The EHR dataset used in this study, *i.e.*, the CSCR dataset, is provided by our partner hospital in China, thus in Chinese language. First, the Translator translates the EHR dataset into English. Then, the splitter splits the paragraphs in the EHR dataset into short sentences.

Each short sentence is fed into the LLM-augmented Retriever to first summarize the content (as described in Section 5.2.3). Then the compressed content is embedded into vectors. The retriever also retrieves the relevant contextual knowledge and demonstrations from the external knowledge base and stores them in vectors.

The short sentences, contextual knowledge and demonstrations stored in the

vectors are content for prompt templates. The prompt templates are fed into the next chain for Micro CoT learning.

To mitigate the impact of LLM uncertainty and hallucination, we implement an ensembling technique that randomly sets the LLM temperature in accordance with prior research [91]. For each input EHR data, we independently prompt the micro CoT five times. The outputs from these promptings are then subject to a voting process to produce the final output.

5.2.4.2 Micro Chain-of-Thought (CoT)

Micro CoT is the core step of our proposed GAPrompt prompting paradigm (see Figure 5.1 in pink). Figure 5.6 provides a detailed illustration of the micro CoT prompting template. Unlike the existing CoT methods [117, 134] that use a fixed set of examples for few-shot prompting, our sentence-level micro-CoT is underpinned by an external knowledge base in addition to the demonstrations. We incorporate the standard NIHSS assessment criteria into the prompt template as contextual information for information retrieval, addressing inconsistencies caused by LLM's potential uncertainty and hallucination. Given the limited context length of the foundation LLM, we employ a three-shot prompting approach, restricting the number of demonstrations to three instances. Furthermore, our inference process aligns with the hierarchical CoT logic. Therefore, Micro CoT is an important technique to improve the performance of LLMs in analyzing the EHR dataset.

5.2.4.3 Summarization

A voting process selects the optimal score from three demonstrations , which are produced by the Micro CoT reasoning method.

In summary, the GAPrompt machine learning framework starts with identifying the language of the input EHR data. If the data is not in English, it is translated

Micro Chain-of-Thought Template

Context:**{{assessment criteria}}**
Demo: **{{Demostrations}}**
Input: **{{report}}**
Instruction: Please follow the assessment criteria to assess the scores of each NIHSS component from the given report.
Let's think step by step.
1. If the report is not in English, translate it to English first.
2. Determine which components of NIHSS are related to the report, and assess the score.
3. Not mentioned components score 0.

Figure 5.6 : The macro sequential chain. It includes the steps of translation, splitting, retrieval, micro CoT, and ensembling.

into English. Afterwards, the framework classifies the input EHR report according to the relevant NIHSS components. The assignment of the NIHSS score for each assessment item is then carried out.

5.3 Experiment Design

In this section, we first briefly introduce the experiment datasets. These include the prompt-based Q&A templates for LLM selection, the generation-augmented knowledge base for RAG, and the test dataset for stroke severity assessment. This is followed by the metrics for the evaluation of output performance.

5.3.1 Datasets

5.3.1.1 Prompt Templates for LLM Selection

As detailed in Section 5.2.1, we have designed six task-specific prompt templates to evaluate the capabilities of LLMs in our stroke assessment use case. Each prompt template contains a specified query paired with a corresponding ground truth answer. The candidate LLMs are individually loaded and presented with each prompt in sequence. Their performance is assessed by comparing their outputs to the respective

ground truth answers.

5.3.1.2 Generation-augmented Knowledge Base

We refer task-specific knowledge as the detailed definitions and scoring criteria for each of the 11 NIHSS assessment components. A labelled dataset [38] is used to generate task-specific knowledge and demonstrations for the subsequent summary-based generation-augmented retrieval (SGAR), utilizing in-context learning (ICL) capabilities of LLMs. The embeddings of each NIHSS assessment component and its summary are stored in a vector database as metadata (see Figure 5.4).

The LLM-generated demonstrations utilize the original EHR data from the CSCR datasets [38], which contains 1,931 EHRs. These EHRs are split into sentences, each with expert-validated NIHSS scores. After removing duplicate sentences and unrelated items, we are left with 3,314 sentence-level demonstrations. Table 5.1 shows the distribution of the augmented knowledge base for each NIHSS assessment component.

5.3.1.3 Test Dataset for Stroke Severity Assessment

Our test dataset, sourced from our previous study [38], comprises ground-truth stroke assessment scores for 33 patients. Each of these records is processed by the Hierarchical Chain-of-thought (refer to 5.2.4). This starts with Macro Sequential Chain. The resulting prompt templates are inputs of the Micro CoT for sentence-level inferencing, which generates NIHSS assessment items and scores. Consequently, the final test set includes both macro and micro-level ground truth.

Table 5.2 illustrates the distribution of the test dataset, at both micro and macro-levels for all NIHSS components. The micro-level samples represent the results of sentence-level inferencing generated by the micro CoT. These sentences are the product of LLM-augmented Retrieving and LLM-chain for each of the 33 raw EHRs.

Table 5.1 : The distribution of the generation-augmented knowledge base. Both the task-specific knowledge and the LLM-generated Demonstrations are reported, along with the count of samples related to each NIHSS component and their corresponding percentages (%).

NIHSS Component	Task-specific Knowledge		LLM-generated Demonstrations	
	Count	Percentage (%)	Count	Percentage (%)
1a	1	6.67	95	2.87
1b	1	6.67	238	7.18
1c	1	6.67	13	0.39
2	1	6.67	32	0.97
3	1	6.67	30	0.91
4	1	6.67	169	5.10
5a	1	6.67	428	12.91
5b	1	6.67	430	12.97
6a	1	6.67	361	10.89
6b	1	6.67	385	11.62
7	1	6.67	207	6.25
8	1	6.67	226	6.82
9	1	6.67	641	19.34
10	1	6.67	54	1.63
11	1	6.67	5	0.15
Total	15	1	3314	1

Each sample contains an NIHSS assessment component and its score. The macro EHR-level ground truth refers to the original assessment items and scores of the given 33 EHRs. The micro sentence-level texts belonging to the same EHR are summarized by the macro chain of HCoT.

Table 5.2 : The distribution of the quantitative assessment dataset. Both micro and macro-level ground truth of each NIHSS component and their corresponding percentage (%) are reported.

NIHSS Component	Micro Level		Macro Level	
	Count	Percentage (%)	Counts	Percentage (%)
1a	35	11.47	30	12.82
1b	29	9.51	20	8.55
1c	3	0.98	3	1.28
2	7	2.30	7	2.99
3	1	0.33	1	0.43
4	31	10.16	25	10.68
5a	10	3.28	10	4.27
5b	29	9.51	27	11.54
6a	11	3.61	10	4.27
6b	28	9.18	27	11.54
7	47	15.41	24	10.26
8	41	13.44	23	9.83
9	24	7.87	19	8.12
10	8	2.62	7	2.99
11	1	0.33	1	0.43
Total	305	100	234	100

5.3.2 Evaluation Metrics

Following prior research [62], we utilize the Exact Match (EM) score to assess the performance of the LLM selection, and we evaluate the retrieval performance using Top-k retrieval accuracy.

Exact Match (EM) is calculated as the proportion of the predicted answer texts that are exactly identical to the ground-truth answer, after string normalization such as article and punctuation removal.

Top-k Retrieval Accuracy is defined as the proportion of questions for which the top-k retrieved records contain at least one correct answer. This metric sets up the upper bound of how many relevant questions are extracted by the retriever.

When evaluating the accuracy of automatic quantitative stroke assessment, unlike the patient-level evaluation in the previous chapter, in this chapter we adopt a fine-grained evaluation method using NIHSS items as groups and segments as units. We use the F1 score as our evaluation metric. The F1 score combines precision (accuracy of positive predictions) and recall (ability to identify actual positive cases) into a single value. It is calculated as the harmonic mean of precision and recall, making it a robust metric for assessing model performance.

5.4 Results

5.4.1 LLM Selection

In Section 5.2.1, we have devised six prompting templates for selecting a candidate pool of LLMs. Table 5.3 shows the results of LLM selection using the EM metric (see Section 5.3.2).

From the table, we can see that LlaMa2-70B and Qwen-72B exhibit superior overall abilities compared to their competitors. Notably, they demonstrate a strong in-context learning (ICL) ability for learning knowledge from the external knowledge

base and understanding the logic from the demonstrations. In contrast, the specifically fine-tuned medical domain LLM HuatuoGPT2 does not perform well. This may be attributed to the nature of our task, which goes beyond a simple medical Q&A task but fully utilizes ICL and CoT to comprehend in-context knowledge and infer from the retrieved knowledge and demonstrations. This test result motivates our selection of the most powerful foundation LLMs, *i.e.*, Llama2-70B, for our task.

Table 5.3 : The performance of candidate LLMs with six prompting templates, using EM evaluation metrics.

LLMs	Knowledge	Understanding	Learning	Reasoning	Consistency	Anti-hallucination	Overall
LlaMa2-70B [124]	0.48	0.73	0.46	0.71	0.89	0.67	**0.66**
Qwen-72B [9]	0.37	0.48	0.37	0.65	0.90	0.66	0.57
Falcon-40B [57]	0.38	0.43	0.35	0.60	0.85	0.70	0.47
Huatuo2-34B [130]	0.40	0.36	0.28	0.56	0.80	0.57	0.50

5.4.2 Generation-augmented Retrieval

Table 5.4 shows the results of our proposed SGAR method using top-k retrieval accuracy on both task-specific knowledge and the demonstrations (see Section 5.3.2).

Table 5.4 : Top-k retrieval accuracy (%) on both task-specific knowledge and the demonstrations.

Method	Knowledge			Demonstrations		
	Top-1	Top-3	Top-5	Top-1	Top-3	Top-5
BM25 [20]	41.18	45.25	48.72	35.26	47.88	54.52
Vdb [62]	56.71	65.51	74.68	48.76	54.64	61.74
BM25-SGAR	70.81	78.39	83.91	81.46	87.64	86.12
Vdb-SGAR	79.64	85.44	**85.84**	89.84	93.81	**95.51**

The experimental results reveal that our retrieval method yields more accurate retrieval outcomes, thereby enhancing subsequent HCoT and achieving more precise inferencing results. Furthermore, it significantly reduces the context length required by LLM. Following retrieval, the context now only includes a single or a few NIHSS components that are most relevant to the query text. In comparison to loading the entire NIHSS assessment guideline, our approach saves up to 90% context occupancy. This efficiency and effectiveness highlight the strength of our retrieval method.

5.4.3 Results of Micro Chain-of-Thought Learning

Table 5.5 shows the performance (F1 score) of LLM inferencing with micro CoT. All the models demonstrate excellent accuracy in the quantitative assessment of sentence-level EHR texts. Four foundation LLMs are tested to conduct a quantitative assessment of EHR sentences and generate the scores for each NIHSS component. The results reported in the table are the ensemble results derived from aggregating the results of five independent inferences. Among these LLMs, LlaMa-70B shows the best performance, slightly surpassing that of Llama2-70B and largely superior to the Falcon-40B and HuatuoGPT2-34B. The result is consistent with the LLM

selection results presented in Section 5.4.1, reaffirming the effectiveness of our simple but efficient prompt-driven LLM selection approach.

Table 5.5 : The micro CoT results (F1 score).

NIHSS Component	**LlaMa2-70B [124]**	**Qwen-72B [9]**	**Falcon-40B [57]**	**HuatuoGPT2-34B [130]**
1a	95.65	94.63	84.62	88.16
1b	95.01	93.60	85.14	84.33
1c	97.56	98.34	85.66	85.00
2	95.42	94.49	81.22	80.88
3	97.25	95.57	83.34	80.43
4	94.80	94.18	79.25	80.61
5a	97.89	96.70	86.15	84.25
5b	96.84	94.44	77.12	81.54
6a	97.67	97.28	80.53	78.27
6b	95.83	94.71	77.84	77.52
7	90.25	86.56	70.53	72.28
8	88.66	86.24	70.11	74.63
9	95.35	93.84	78.62	82.22
10	98.44	97.67	84.51	82.79
11	99.06	98.90	85.12	83.93
overall	**95.71**	94.48	80.65	81.12

5.4.4 Macro Sequential Chain Results

In this section, we evaluate the performance of our macro sequential chain for document-level LLM inferencing. Table 5.6 shows the performance of the macro

sequential chain by four foundation LLMs. The sentence-level LLM inferencing results that belong to the same EHR are comprehensively integrated by the macro sequential chain. Macro results are consistent with the micro CoT results (see Table 5.5), but a bit lower. This is because the cumulative effect of micro-level errors reduces the accuracy of macro results.

Table 5.6 : The result of the macro sequential chain.

NIHSS Component	**Llama2-70b**	**Qwen-72B**	**Falcon-40B**	**HuatuoGPT2-34B**
1a	75.15	70.44	43.12	45.28
1b	60.61	54.36	42.56	38.55
1c	43.55	87.88	40.44	41.23
2	57.04	90.71	47.20	52.65
3	61.88	98.46	51.33	51.44
4	63.84	82.44	52.05	54.64
5a	80.79	89.22	58.60	64.22
5b	89.68	91.02	57.97	71.24
6a	81.80	85.07	50.23	65.72
6b	83.62	83.24	51.54	61.84
7	87.31	85.72	61.35	60.88
8	52.04	63.30	30.51	49.63
9	48.80	56.57	45.54	48.75
10	93.10	94.43	67.88	72.51
11	89.42	96.88	46.40	50.35
overall	**85.42**	81.98	49.78	55.26

5.4.5 Ablation Studies on the Effectiveness of GAPrompt Components

Figure 5.7 presents the results of an ablation study on the effectiveness of GAPrompt components, based on the overall F1 score of the quantitative assessment using Qwen-72B on the test dataset.

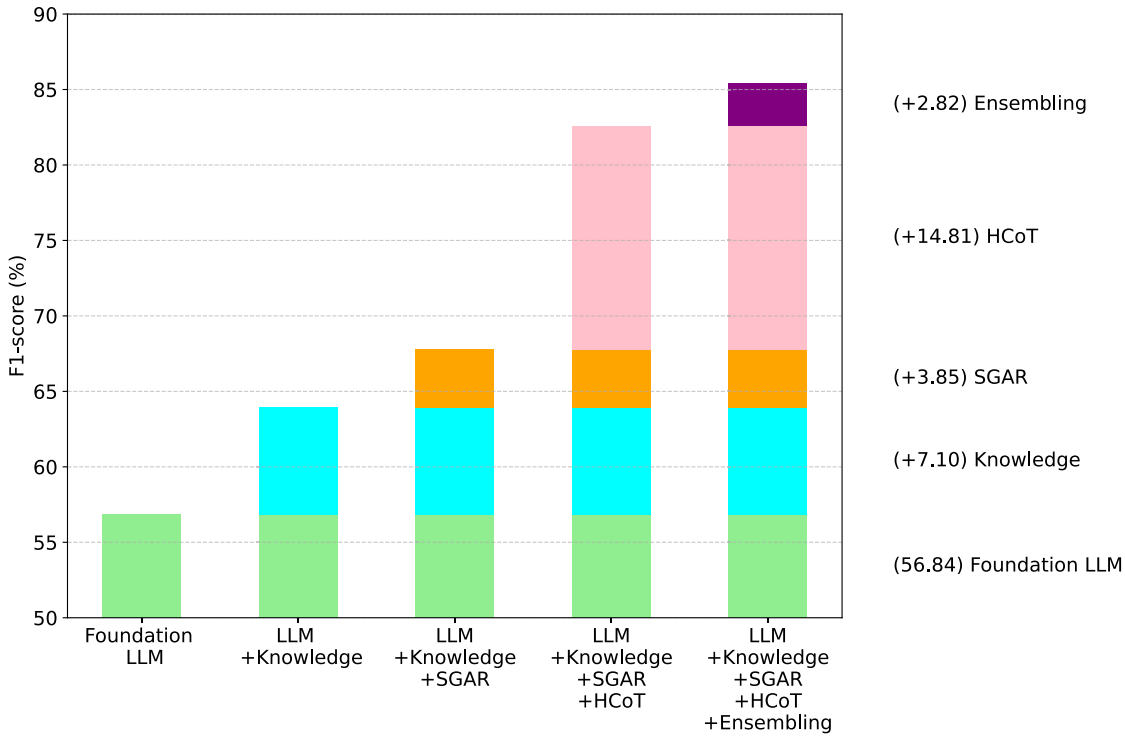

Figure 5.7 : The ablation study on the effectiveness of GAPrompt components.

Our proposed GAPrompt pipeline consists of the generation-augmented knowledge base construction (represented as "Knowledge" in Figure 5.7), the GAR method to retrieve the knowledge and the demonstrations, the HCoT strategy that integrates micro CoT with macro sequential chain, and the ensembling strategy to integrate the inference results from five generations. From the table, we can find how much each component of GAPrompt contributes to the overall results.

The green bar shows the inferencing performance of the foundation LLM, with a moderate F1 score of 56.84%. It is not surprising since the inferencing largely

relies on the basic knowledge of the foundation LLM, which is not fully accurate as described in Section 5.2.1.

The blue bar refers to the performance improvement (+7.10% F1 score) when importing the task-specific knowledge during the LLM inferencing. Two factors lead to this enhancement, the first is the knowledge, *i.e.*, the detailed NIHSS assessment criteria in this study, defines and limits the assessment components of NIHSS. The second is it provides detailed scoring criteria which helps the LLM to better understand the given EHR and conduct NIHSS scoring.

The orange bar indicates the improvement (+3.85% F1 score) when employing our proposed GAR method based on the LLM-generated summary index. This improvement is compared with the performance using the full-text-based retrieval method. Compared with the existing RAG methods [111], our summary-based retrieval indicates higher retrieval accuracy on both knowledge and demonstration retrieval. This is because the LLM-generated summary extracts the most valuable information and excludes noisy information.

The most significant improvement of our GAPrompt falls on the HCoT strategy (+14.81% F1 score, in pink). It indicates that our designed promptings on both the macro sequential chain and the micro CoT contribute the most to empowering the foundation LLM in completing our task. This finding is consistent with the previous works, *i.e.*, CoT [134] and AutoCoT [158], demonstrating that the few-shot step-by-step demonstrations are the most important factor in improving the LLM inference performance. The excellent performance of our method also validates the effectiveness of our HCoT prompting strategy compared to existing reasoning prompting methods [111].

At last, we apply ensembling to further improve the final performance (+2.82% F1 score, in purple) of our proposed GAPrompt paradigm, minimizing the influence

of randomness and hallucination of LLM generation.

Compared with the existing research that only uses RAG to control the hallucination [73, 111], our method applies an ensembling strategy on top of the RAG-based LLM inferencing, providing multiple options and more comprehensive evaluation to better control LLM'S hallucinations.

We set up five different temperatures for the foundation LLM generation, and conduct five independent LLM inferences on each EHR sentence. Finally, the results of the five independent LLM inferences are integrated producing the final inference result with the ensembling mechanism.

5.5 Summary

In conclusion, our study underscores the transformative potential of leveraging foundation LLMs for automating the intricate analysis of EHRs in the medical domain. Focused on stroke as a use case, our prompting paradigm, incorporating LlaMa2-70B and innovative methods including retrieval-augmented generation and hierarchical chain-of-thought, demonstrates the capacity to automatically assess and quantify EHRs.

This approach not only overcomes the challenges of labor-intensive and manually conducted quantitative assessments but also extends its applicability beyond the medical domain. The adaptability of our method positions it as a valuable tool for diverse fields, offering insights and solutions for data-driven analysis in both research and industry applications.

Chapter 6

Conclusion and Future Work

6.1 Conclusion

In this thesis, towards developing advanced AI-based techniques to automate the clinical assessment task using Chinese EHRs, we have developed progressive steps with a range of sophisticated AI-driven approaches to realize this objective.

In Chapter 3, we have constructed a stroke disease-specific ontology "StrokePEO", which is the first ontology that is specifically constructed for the physical examination of stroke. We introduced multiple NLP techniques and deep learning methods to extract the terms and relationships from real clinical EHRs, effectively and efficiently boosting the ontology construction process. With the verification of stroke specialists, our approach and the resulting StrokePEO demonstrate huge potential in supporting the further development of diverse clinical research and practice.

On the basis of this, in Chapter 4, we developed automatic stroke assessment algorithms based on Chinese clinical named entity recognition and domain-adaptive pre-training. Grounded on the US National Institutes of Health Stroke Scale (NIHSS) and clinical practice, we defined an ontology-aided dictionary with seven types of semantically related entities that are combined to describe stroke symptoms. We constructed a labeled dataset "Chinese Stroke Clinical Records" (CSCR) from EHRs of the partner hospital with this dictionary, semi-automatic annotation, NLP techniques and specialist validation. We pre-trained a Chinese clinical word embedding, "CliRoberta", through adaptively transferring the BERT-based embedding, "Roberta-wwm", to the clinical domain using the open-source Chinese EHRs and the EHRs of

our partner hospital, which achieves higher representation accuracy than the existing embeddings. Combining the CSCR and CliRoberta, we created a high-performing Chinese clinical named entity recognition (CNER) model based on BiLSTM-CRF deep neural networks. We defined and implemented an entity-to-NIHSS mapping dictionary and used it in incremental development to automatically generate the stroke assessment score for a patient. With high inter-rater agreement (0.823) and excellent intraclass consistency (0.986) with the ground truth and reduction of processing time to a few seconds, our algorithms demonstrate the value of automatic disease assessment using free-text EHRs.

With the rapid development and outstanding performance of LLMs over the past year, leveraging LLMs to achieve automated stroke assessment based on EHRs showcases a more comprehensive, robust, and general approach compared to previous methods. However, the scarcity of publicly available medical LLMs and the complexity of domain-specific fine-tuning pose challenges. Therefore, designing appropriate prompting strategies to enhance the capabilities of foundation LLMs emerges as a promising solution. In Chapter 5, we propose a novel generation-augmented prompting paradigm called GAPrompt for the automated analysis of EHRs using foundation LLMs. By leveraging the few-shot in-context learning (ICL) abilities of LLMs, our proposed GAPrompt paradigm enhances the power of foundation LLM through a series of prompting strategies, including prompt-driven LLM selection, generation-augmented knowledge base construction, summary-based generation-augmented retrieval, hierarchical chain-of-thought, and ensembling, overcoming the limitations of foundation LLM in analyzing domain-specific EHR text. Our method has demonstrated the capability to automatically assess EHRs and generate quantitative assessment results based on the retrieved assessment criteria and few-shot demonstrations.

All three works completed during my PhD have addressed various challenges in

automating stroke assessment from different perspectives and using various methods, providing advanced AI-based solutions. These works evolve from the NLP and machine learning-based semi-automated ontology construction, to deep learning and pre-trained embeddings-based CNER, from the dictionary-based NIHSS mapping, to leveraging LLM's powerful understanding and reasoning capabilities to achieve more intelligent and comprehensive automated quantitative clinical assessment. These works also represent the development trajectory of advanced AI technologies in clinical research and application, providing significant practical and reference value for both clinical research and practice.

6.2 Future Work

The application of intelligent AI technologies, especially large-scale models, to assist clinical practice is becoming increasingly prominent as a research focus. Our research in this thesis serves as a typical case study, demonstrating the immense potential of LLMs in real-world clinical applications. As a pioneering work in this field, there are still opportunities and challenges for further research.

One future research direction is to extend our proposed methods to multiple levels of clinical practical applications, *i.e.*, automatic stroke severity assessment with various scales and automatic assessments of other diseases. Besides the NIHSS scale, there are other assessment scales that are commonly used by clinicians to assess stroke-induced impairments, *e.g.*, Glasgow Coma Scale (GCS), modified Rankin Scale (mRS), Barthel Index (BI) and quality of life (QOL), *etc.* [48, 88, 103]. Our methods can be modified and fine-tuned to allow automatic assessment using the other scales to obtain more specific quantitative assessment results in a more broad area.

Furthermore, building upon the foundation of this study, exploring further medical applications and research avenues is essential. For example, utilizing LLMs for automatic EHR recording and summarization, converting unstructured EHR data

into structured formats, conducting medicine effectiveness analysis, and developing drug and treatment recommendation systems, etc,.

In conclusion, applying advanced AI technologies to address real-world challenges and promote the practical application of intelligent technologies for the benefit of society is a research focus worth persistently commitment and lifelong dedication.

Bibliography

[1] "2019 language and intelligence challenge," https://www.ccf.org.cn/Chapters/TC/Call_for_Papers/2019-02-22/660510.shtml.

[2] "Entity relation extraction," https://github.com/yuanxiaosc/Entity-Relation-Extraction.git.

[3] "Resource description framework (rdf) model and syntax specification," http://www.w3.org/TR/PR-rdf-syntax, 2017.

[4] H. Abburi, M. Suesserman, N. Pudota, B. Veeramani, E. Bowen, and S. Bhattacharya, "Generative ai text classification using ensemble llm approaches," *arXiv preprint arXiv:2309.07755*, 2023.

[5] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat *et al.*, "Gpt-4 technical report," *arXiv preprint arXiv:2303.08774*, 2023.

[6] R. R. Afandi, "Ontology development in patients information system for stroke rehabilitation," *ICBO*, 2017.

[7] M. Alkahtani, A. Choudhary, A. De, and J. A. Harding, "A decision support system based on ontology and data mining to improve design using warranty data," *Computers & industrial engineering*, vol. 128, pp. 1027–39, 2019.

[8] J. Austin, A. Odena, M. Nye, M. Bosma, H. Michalewski, D. Dohan, E. Jiang, C. Cai, M. Terry, Q. Le *et al.*, "Program synthesis with large language models," *arXiv preprint arXiv:2108.07732*, 2021.

[9] J. Bai, S. Bai, Y. Chu, Z. Cui, K. Dang, X. Deng, Y. Fan, W. Ge, Y. Han, F. Huang, B. Hui, L. Ji, M. Li, J. Lin, R. Lin, D. Liu, G. Liu, C. Lu, K. Lu, J. Ma, R. Men, X. Ren, X. Ren, C. Tan, S. Tan, J. Tu, P. Wang, S. Wang, W. Wang, S. Wu, B. Xu, J. Xu, A. Yang, H. Yang, J. Yang, S. Yang, Y. Yao, B. Yu, H. Yuan, Z. Yuan, J. Zhang, X. Zhang, Y. Zhang, Z. Zhang, C. Zhou, J. Zhou, X. Zhou, and T. Zhu, "Qwen technical report," *arXiv preprint arXiv:2309.16609*, 2023.

[10] Baichuan, "Baichuan 2: Open large-scale language models," *arXiv preprint arXiv:2309.10305*, 2023. [Online]. Available: https://arxiv.org/abs/2309.10305

[11] J. J. Bartko, "The intraclass correlation coefficient as a measure of reliability," *Psychological reports*, vol. 19, no. 1, pp. 3–11, 1966.

[12] BioPortal, https://bioportal.bioontology.org.

[13] S. Bird, E. Loper, and E. Klein, "Natural language processing with python," *O'Reilly Media Inc.*, 2009.

[14] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "Enriching word vectors with subword information," *Transactions of the association for computational linguistics*, vol. 5, pp. 135–146, 2017.

[15] R. Bommasani, D. A. Hudson, E. Adeli, R. Altman, S. Arora, S. von Arx, M. S. Bernstein, J. Bohg, A. Bosselut, E. Brunskill *et al.*, "On the opportunities and risks of foundation models," *arXiv preprint arXiv:2108.07258*, 2021.

[16] T. Brott, J. H. P. Adams, C. P. Olinger, J. R. Marler, W. G. Barsan, J. Biller, J. Spilker, R. Holleran, R. Eberle, and V. Hertzberg, "Measurements of acute cerebral infarction: a clinical examination scale," *Stroke*, vol. 20(7), pp. 864–870, 1989.

[17] T. Brott, J. R. Marler, C. P. Olinger, H. P. Adams Jr, T. Tomsick, W. G. Barsan, J. Biller, R. Eberle, V. Hertzberg, and M. Walker, "Measurements of acute cerebral infarction: lesion size by computed tomography." *Stroke*, vol. 20, no. 7, pp. 871–875, 1989.

[18] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, "Language models are few-shot learners," *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.

[19] Z. Cai, T. Zhang, C. Wang, and X. He, "Embert: A pre-trained language model for chinese medical text mining," in *Asia-Pacific Web (APWeb) and Web-Age Information Management (WAIM) Joint International Conference on Web and Big Data*. Springer, 2021, pp. 242–257.

[20] D. Chen, A. Fisch, J. Weston, and A. Bordes, "Reading Wikipedia to answer open-domain questions," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, R. Barzilay and M.-Y. Kan, Eds. Vancouver, Canada: Association for Computational Linguistics, Jul. 2017, pp. 1870–1879. [Online]. Available: https://aclanthology.org/P17-1171

[21] B. Cheng, J. Zhang, H. Liu, M. Cai, and Y. Wang, "Research on medical knowledge graph for stroke," *Journal of Healthcare Engineering*, 2021.

[22] CHIP, "Evaluation 1: Chinese medical text named entity recognition," http://www.cips-chip.org.cn/2020/eval1, 2020.

[23] H. Chmura Kraemer, V. S. Periyakoil, and A. Noda, "Kappa coefficients in medical research," *Statistics in medicine*, vol. 21, no. 14, pp. 2109–2129, 2002.

[24] A. Chowdhery, S. Narang, J. Devlin, M. Bosma, G. Mishra, A. Roberts, and P. et al., "PaLM: Scaling language modeling with pathways," 2022.

[25] K. Clark, M.-T. Luong, Q. V. Le, and C. D. Manning, "Electra: Pre-training text encoders as discriminators rather than generators," *arXiv preprint arXiv:2003.10555*, 2020.

[26] Y. Cloud, "Yidu-S4K: Yidu cloud structured 4k dataset," http://openkg.cn/dataset/yidu-s4k, 2020.

[27] K. Cobbe, V. Kosaraju, M. Bavarian, M. Chen, H. Jun, L. Kaiser, M. Plappert, J. Tworek, J. Hilton, R. Nakano *et al.*, "Training verifiers to solve math word problems," *arXiv preprint arXiv:2110.14168*, 2021.

[28] J. Cohen, "A coefficient of agreement for nominal scales," *Educational and psychological measurement*, vol. 20, no. 1, pp. 37–46, 1960.

[29] M. Cuadros, N. Pérez, I. Montoya, and A. G. Pablos, "Vicomtech at BARR2: Detecting biomedical abbreviations with ML methods and dictionary-based heuristics." in *IberEval@ SEPLN*, 2018, pp. 322–328.

[30] Y. Cui, W. Che, T. Liu, B. Qin, and Z. Yang, "Pre-training with whole word masking for chinese bert," *IEEE ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3504–3514, 2021.

[31] Y. Cui, W. Che, T. Liu, B. Qin, S. Wang, and G. Hu, "Revisiting pre-trained models for Chinese natural language processing," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, Nov. 2020, pp. 657–668. [Online]. Available: https://www.aclweb.org/anthology/2020.findings-emnlp.58

[32] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of

deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[33] S. Dhuliawala, M. Komeili, J. Xu, R. Raileanu, X. Li, A. Celikyilmaz, and J. Weston, "Chain-of-verification reduces hallucination in large language models," *arXiv preprint arXiv:2309.11495*, 2023.

[34] R. Dieng-Kuntz, D. Minier, and M. Ruzicka, "Building and using a medical ontology for knowledge management and cooperative work in a health care network," *Computers in Biology and Medicine*, vol. 36(8), pp. 871–892, 2006.

[35] P. I. Dissanayake, T. K. Colicchio, and J. Cimino, "Using clinical reasoning ontologies to make smarter clinical decision support systems: a systematic review and data synthesis," *Journal of the American Medical Informatics Association*, vol. 27(1), pp. 159–174, 2020.

[36] M. Fahes, T.-H. Vu, A. Bursuc, P. Pérez, and R. de Charette, "Poda: Prompt-driven zero-shot domain adaptation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 18 623–18 633.

[37] R. A. Fisher, "Statistical methods for research workers," in *Breakthroughs in statistics: Methodology and distribution*. Springer, 1970, pp. 66–70.

[38] Z. GU, H. Chen, P. Yu, X. He, W. Jia, X. Yang, G. Peng, P. Hu, S. Chen, and Y. Lin, "Automatic quantitative stroke severity assessment based on chinese clinical named entity recognition with domain-adaptive pre-trained large language model," *Available at SSRN 4490001*.

[39] Z. Gu, X. Yang, W. Jia, C. Xu, P. Yu, X. He, H. Chen, and Y. Lin, "StrokePEO: Construction of a clinical ontology for physical examination of stroke," in *2022 9th International Conference on Digital Home (ICDH)*. IEEE, 2022, pp. 218–223.

[40] K. Guu, K. Lee, Z. Tung, P. Pasupat, and M. Chang, "Retrieval augmented language model pre-training," in *International conference on machine learning*. PMLR, 2020, pp. 3929–3938.

[41] M. Habibi-Koolaee, L. Shahmoradi, S. R. Niakan Kalhori, H. Ghannadan, and E. Younesi, "Sto: Stroke ontology for accelerating translational stroke research," *Neurology and Therapy*, vol. 10(1), pp. 321–333, 2021.

[42] ——, "STO: Stroke ontology for accelerating translational stroke research," *Neurology and Therapy*, vol. 10, no. 1, pp. 321–333, 2021.

[43] M. A. Haendel, C. G. Chute, and P. N. Robinson, "Classification, ontology, and precision medicine," *New England Journal of Medicine*, vol. 379 (15), pp. 1452–62, 2018.

[44] S. Han, Y. Zhang, Y. Ma, C. Tu, Z. Guo, Z. Liu, and M. Sunc, "Thuocl: Tsinghua open chinese lexicon," 2016.

[45] S. H. Han, Y. Zhang, Y. Ma, C. Tu, Z. Guo, Z. Liu, and M. Sun, "Thuocl: Tsinghua open chinese lexicon," https://github.com/thunlp/THUOCL, 2016.

[46] T. Han, L. C. Adams, J.-M. Papaioannou, P. Grundmann, T. Oberhauser, A. Löser, D. Truhn, and K. K. Bressem, "Medalpaca–an open-source collection of medical conversational ai models and training data," *arXiv preprint arXiv:2304.08247*, 2023.

[47] X. Han, Z. Wang, J. Zhang, Q. Wen, W. Li, B. Tang, Q. Wang, Z. Feng, Y. Zhang, Y. Lu *et al.*, "Overview of the CCKS 2019 knowledge graph evaluation track: entity, relation, event and QA (in chinese)," *arXiv preprint arXiv:2003.03875*, 2020.

[48] J. K. Harrison, K. S. McArthur, and T. J. Quinn, "Assessment scales in stroke:

clinimetric and clinical considerations," *Clinical interventions in aging*, pp. 201–211, 2013.

[49] H. He and J. D. Choi, "The stem cell hypothesis: Dilemma behind multi-task learning with transformer encoders," *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 2021.

[50] ——, "The stem cell hypothesis: Dilemma behind multi-task learning with transformer encoders," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, Nov. 2021, pp. 5555–5577. [Online]. Available: https://aclanthology.org/2021.emnlp-main.451

[51] D. Hendrycks, C. Burns, S. Basart, A. Zou, M. Mazeika, D. Song, and J. Steinhardt, "Measuring massive multitask language understanding," *arXiv preprint arXiv:2009.03300*, 2020.

[52] D. Hendrycks, C. Burns, S. Kadavath, A. Arora, S. Basart, E. Tang, D. Song, and J. Steinhardt, "Measuring mathematical problem solving with the math dataset," *arXiv preprint arXiv:2103.03874*, 2021.

[53] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[54] C. Hong, E. Rush, M. Liu, D. Zhou, J. Sun, A. Sonabend, V. M. Castro, P. Schubert, V. A. Panickan, T. Cai *et al.*, "Clinical knowledge extraction via sparse embedding regression (KESER) with multi-center large scale electronic health record data," *NPJ digital medicine*, vol. 4, no. 1, p. 151, 2021.

[55] I. Huitzil, F. Alegre, and F. Bobillo, "Gimmehop: A recommender system for mobile devices using ontology reasoners and fuzzy logic," *Fuzzy Sets and Systems*, vol. 401, pp. 55–77, 2020.

[56] X. T. Inc., "Xverse," https://github.com/xverse-ai/XVERSE-65B, 2023.

[57] T. I. Institutes, "Falcon," https://huggingface.co/tiiuae/falcon-40b-instruct, 2023.

[58] B. Ji, S. Li, J. Yu, J. Ma, J. Tang, Q. Wu, Y. Tan, H. Liu, and Y. Ji, "Research on Chinese medical named entity recognition based on collaborative cooperation of multiple neural network models," *Journal of Biomedical Informatics*, p. 103395, 2020.

[59] K. Jiang, T. Yang, C. Wu, L. Chen, L. Mao, Y. Wu, L. Deng, and T. Jiang, "Latte: A knowledge-based method to normalize various expressions of laboratory test results in free text of chinese electronic health records," *Journal of Biomedical Informatics*, vol. 102, p. 103372, 2020.

[60] Jieba, "Chinese text segmentation," https://github.com/fxsjy/jieba.

[61] D. Jin, E. Pan, N. Oufattole, W.-H. Weng, H. Fang, and P. Szolovits, "What disease does this patient have? a large-scale open domain question answering dataset from medical exams," *Applied Sciences*, vol. 11, no. 14, p. 6421, 2021.

[62] V. Karpukhin, B. Oğuz, S. Min, P. Lewis, L. Wu, S. Edunov, D. Chen, and W.-t. Yih, "Dense passage retrieval for open-domain question answering," *arXiv preprint arXiv:2004.04906*, 2020.

[63] E. Kogan, K. Twyman, J. Heap, D. Milentijevic, J. H. Lin, and M. Alberts, "Assessing stroke severity using electronic health record data: a machine learning approach," *BMC medical informatics and decision making*, vol. 20, no. 1, pp. 1–8, 2020.

[64] T. K. Koo and M. Y. Li, "A guideline of selecting and reporting intraclass correlation coefficients for reliability research," *Journal of chiropractic medicine*, vol. 15, no. 2, pp. 155–163, 2016.

[65] Z. Kraljevic, A. Shek, D. Bean, R. Bendayan, J. Teo, and R. Dobson, "Medgpt: Medical concept prediction from clinical narratives," *arXiv preprint arXiv:2107.03134*, 2021.

[66] C. S. P. Kumar and L. D. D. Babu, "Evolving dictionary based sentiment scoring framework for patient authored text," *Evolutionary Intelligence*, vol. 14, pp. 657–667, 2021.

[67] T. H. Kung, M. Cheatham, A. Medenilla, C. Sillos, L. De Leon, C. Elepaño, M. Madriaga, R. Aggabao, G. Diaz-Candido, J. Maningo *et al.*, "Performance of chatgpt on usmle: Potential for ai-assisted medical education using large language models," *PLoS digital health*, vol. 2, no. 2, p. e0000198, 2023.

[68] S. Kwon, "Stroke medical ontology for supporting ai-based stroke prediction system using bio-signals," *2021 Twelfth International Conference on Ubiquitous and Future Networks (ICUFN). IEEE*, 2021.

[69] J. Lafferty, A. McCallum, and F. C. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," 2001.

[70] H. Lakkaraju, D. Slack, Y. Chen, C. Tan, and S. Singh, "Rethinking explainability as a dialogue: A practitioner's perspective, 2022," *URL https://arxiv.org/abs/2202.01875*.

[71] J. R. Landis and G. G. Koch, "The measurement of observer agreement for categorical data," *biometrics*, pp. 159–174, 1977.

[72] Langchain-ai, "Langchain," https://github.com/langchain-ai/langchain, 2022.

[73] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel *et al.*, "Retrieval-augmented generation for knowledge-intensive nlp tasks," *Advances in Neural Information Processing Systems*, vol. 33, pp. 9459–9474, 2020.

[74] F. Li, M. Zhang, G. Fu, and D. Ji, "A neural joint model for entity and relation extraction from biomedical text," *BMC bioinformatics*, vol. 18, no. 1, pp. 1–11, 2017.

[75] H. Li, *Statistical Learning Methods*. Qing hua da xue chu ban she, 2012.

[76] Q. Li and H. Ji, "Incremental joint extraction of entity mentions and relations," in *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2014, pp. 402–412.

[77] X. Li, H. Zhang, and X.-H. Zhou, "Chinese clinical named entity recognition with variant neural structures based on bert methods," *Journal of Biomedical Informatics*, p. 103422, 2020.

[78] Y. Li, G. Du, Y. Xiang, S. Li, L. Ma, D. Shao, X. Wang, and H. Chen, "Towards chinese clinical named entity recognition by dynamic embedding using domain-specific knowledge," *Journal of Biomedical Informatics*, p. 103435, 2020.

[79] S. Liu, W. Fang, Y. Lu, Q. Zhang, H. Zhang, and Z. Xie, "Rtlcoder: Outperforming gpt-3.5 in design rtl generation with our open-source dataset and lightweight solution," *arXiv preprint arXiv:2312.08617*, 2023.

[80] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, and et al, "RoBERTa: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.

[81] Z. Liu, Y. Li, P. Shu, A. Zhong, L. Yang, C. Ju, Z. Wu, C. Ma, J. Luo, C. Chen *et al.*, "Radiology-llama2: Best-in-class large language model for radiology," *arXiv preprint arXiv:2309.06419*, 2023.

[82] Y. Luan, J. Eisenstein, K. Toutanova, and M. Collins, "Sparse, dense, and

attentional representations for text retrieval," *Transactions of the Association for Computational Linguistics*, vol. 9, pp. 329–345, 2021.

[83] Y. Mao, P. He, X. Liu, Y. Shen, J. Gao, J. Han, and W. Chen, "Generation-augmented retrieval for open-domain question answering," *arXiv preprint arXiv:2009.08553*, 2020.

[84] P. Marczykowska, T. B. Ciszek, and A. Przelaskowski, "evelopment of diagnostic stroke ontology-preliminary results," *Information Technologies in Biomedicine*, vol. 4, pp. 261–272, 2014.

[85] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.

[86] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," *Advances in neural information processing systems*, vol. 26, 2013.

[87] M. A. Musen, "The protege project: A look back and a look forward," *AI Matters, Association of Computing Machinery Specific Interest Group in Artificial Intelligence*, vol. 1(4), 2015.

[88] R. Nedadur, B. Wang, and W. Tsang, "Artificial intelligence for the echocardiographic assessment of valvular heart disease," *Heart*, vol. 108, no. 20, pp. 1592–1599, 2022.

[89] M. Nilashi, O. Ibrahim, and K. Bagherifard, "A recommender system based on collaborative filtering using ontology and dimensionality reduction techniques," *Expert Systems with Applications*, vol. 92, pp. 507–20, 2018.

[90] H. Nori, N. King, S. M. McKinney, D. Carignan, and E. Horvitz, "Capabilities of gpt-4 on medical challenge problems," *arXiv preprint arXiv:2303.13375*, 2023.

[91] H. Nori, Y. T. Lee, S. Zhang, D. Carignan, R. Edgar, N. Fusi, N. King, J. Larson, Y. Li, W. Liu *et al.*, "Can generalist foundation models outcompete special-purpose tuning? case study in medicine," *arXiv preprint arXiv:2311.16452*, 2023.

[92] N. F. Noy and D. L. Mcguinness, "Ontology development 101: a guide to creating your first ontology," *Stanford Knowledge Systems Laboratory Technical Report KSL-01-05 and Stanford Medical Informatics Technical Report SMI-2001-0880*, 2001.

[93] E. Ong, L. L. Wang, J. Schaub, J. F. O'Toole, B. Steck, A. Z. Rosenberg, F. Dowd, J. Hansen, L. Barisoni, S. Jain, and I. H. de Boer, "Modelling kidney disease using ontology: insights from the kidney precision medicine project," *Nature Reviews Nephrology*, vol. 16(11), pp. 686–96, 2020.

[94] OpenAI, "Gpt-4 technical report," 2023.

[95] T. F. Osborne, Z. P. Veigulis, D. M. Arreola, E. Röösli, and C. M. Curtin, "Automated EHR score to predict COVID-19 outcomes at us department of veterans affairs," *PLOS ONE*, vol. 15, no. 7, pp. 1–7, 07 2020. [Online]. Available: https://doi.org/10.1371/journal.pone.0236554

[96] L. Padro and E. Stanilovsky, "Freeling 3.0: Towards wider multilinguality," *Proceedings of the Language Resources and Evaluation Conference (LREC 2012)*, 2012.

[97] A. Pal, L. K. Umapathi, and M. Sankarasubbu, "Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering," in *Conference on Health, Inference, and Learning*. PMLR, 2022, pp. 248–260.

[98] E. Park, K. Lee, T. Han, H. S. Nam *et al.*, "Automatic grading of stroke symptoms for rapid assessment using optimized machine learning and 4-limb

kinematics: clinical validation study," *Journal of medical Internet research*, vol. 22, no. 9, p. e20641, 2020.

[99] B. Peng, M. Galley, P. He, H. Cheng, Y. Xie, Y. Hu, Q. Huang, L. Liden, Z. Yu, W. Chen *et al.*, "Check your facts and try again: Improving large language models with external knowledge and automated feedback," *arXiv preprint arXiv:2302.12813*, 2023.

[100] D. Peng, Y. Wang, C. Liu, and Z. Chen, "TL-NER: A transfer learning model for chinese named entity recognition," *Information Systems Frontiers*, vol. 22, no. 6, pp. 1291–1304, 2020.

[101] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 2014, pp. 1532–1543.

[102] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer, "Deep contextualized word representations," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, Jun. 2018, pp. 2227–2237. [Online]. Available: https://aclanthology.org/N18-1202

[103] A. Puente-Castro, E. Fernandez-Blanco, A. Pazos, and C. R. Munteanu, "Automatic assessment of Alzheimer's disease diagnosis based on deep learning techniques," *Computers in biology and medicine*, vol. 120, p. 103764, 2020.

[104] A. Radford, K. Narasimhan, T. Salimans, I. Sutskever *et al.*, "Improving language understanding by generative pre-training," *OpenAI*, 2018.

[105] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever *et al.*, "Language models are unsupervised multitask learners," *OpenAI blog*, vol. 1, no. 8, p. 9,

2019.

[106] L. Rasmy, Y. Xiang, Z. Xie, C. Tao, and D. Zhi, "Med-BERT: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction," *NPJ digital medicine*, vol. 4, no. 1, pp. 1–13, 2021.

[107] R. Řehůřek and P. Sojka, "Software framework for topic modelling with large corpora," http://is.muni.cz/publication/884893/en, 2010.

[108] ——, "Software Framework for Topic Modelling with Large Corpora," in *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*. Valletta, Malta: ELRA, May 2010, pp. 45–50, http://is.muni.cz/publication/884893/en.

[109] F. Ren, J. Shen, B. Sun, and J. Zhu, "A review for domain ontology construction from text," *Chinese Journal of Computers*, vol. 3, pp. 654–676, 2019.

[110] B. Rink and S. Harabagiu, "Utd: Classifying semantic relations by combining lexical and semantic resources," in *Proceedings of the 5th International Workshop on Semantic Evaluation*, 2010, pp. 256–259.

[111] P. Sahoo, A. K. Singh, S. Saha, V. Jain, S. Mondal, and A. Chadha, "A systematic survey of prompt engineering in large language models: Techniques and applications," *arXiv preprint arXiv:2402.07927*, 2024.

[112] S. K. Sahu, A. Anand, K. Oruganty, and M. Gattu, "Relation extraction from clinical texts using domain invariant convolutional neural network," *arXiv preprint arXiv:1606.09370*, 2016.

[113] M. Schaekermann, C. J. Cai, A. E. Huang, and R. Sayres, "Expert discussions improve comprehension of difficult cases in medical image assessment," in *Proceedings of the 2020 CHI conference on human factors in computing systems*, 2020, pp. 1–13.

[114] Sentence-transformers, "all-mpnet-base-v2," https://huggingface.co/sentence-transformers/all-mpnet-base-v2, 2021.

[115] Y. Shi, S. Xu, Z. Liu, T. Liu, X. Li, and N. Liu, "Mededit: Model editing for medical question answering with external knowledge bases," *arXiv preprint arXiv:2309.16035*, 2023.

[116] P. Shvaiko and J. Euzenat, "Ontology matching: State of the art and future challenges," *IEEE Trans on Knowledge and Data Engineering*, vol. 25(1), pp. 158–176, 2013.

[117] K. Singhal, S. Azizi, T. Tu, S. S. Mahdavi, J. Wei, H. W. Chung, N. Scales, A. Tanwani, H. Cole-Lewis, S. Pfohl *et al.*, "Large language models encode clinical knowledge," *arXiv preprint arXiv:2212.13138*, 2022.

[118] K. Singhal, T. Tu, J. Gottweis, R. Sayres, E. Wulczyn, L. Hou, K. Clark, S. Pfohl, H. Cole-Lewis, D. Neal *et al.*, "Towards expert-level medical question answering with large language models," *arXiv preprint arXiv:2305.09617*, 2023.

[119] S. Sivarajkumar and Y. Wang, "Healthprompt: A zero-shot learning paradigm for clinical natural language processing," in *AMIA Annual Symposium Proceedings*, vol. 2022. American Medical Informatics Association, 2022, p. 972.

[120] J. Sun, "Jieba Chinese text segmentation: built to be the best python Chinese word segmentation module," https://github.com/fxsjy/jieba, 2020.

[121] Y. Sun, S. Wang, Y. Li, S. Feng, X. Chen, H. Zhang, X. Tian, D. Zhu, H. Tian, and H. Wu, "Ernie: Enhanced representation through knowledge integration," *arXiv preprint arXiv:1904.09223*, 2019.

[122] Y. Sun, S. Wang, Y. Li, S. Feng, H. Tian, H. Wu, and H. Wang, "Ernie 2.0: A continual pre-training framework for language understanding," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, 2020, pp. 8968–8975.

[123] R. Taori, I. Gulrajani, T. Zhang, Y. Dubois, X. Li, C. Guestrin, P. Liang, and T. B. Hashimoto, "Alpaca: A strong, replicable instruction-following model," *Stanford Center for Research on Foundation Models. https://crfm. stanford. edu/2023/03/13/alpaca. html*, vol. 3, no. 6, p. 7, 2023.

[124] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar *et al.*, "Llama: Open and efficient foundation language models," *arXiv preprint arXiv:2302.13971*, 2023.

[125] C. Townsend, J. Huang, D. Dou, H. Liu, L. He, P. Hayes, R. Rudnick, H. Shah, D. Fell, and W. Liu, "Neumore: Ontology in stroke recovery," pp. 821–822, 2011.

[126] Q. Wan, J. Liu, L. Wei, and B. Ji, "A self-attention based neural architecture for chinese medical named entity recognition," *Mathematical Biosciences and Engineering*, vol. 17, no. 4, pp. 3498–3511, 2020.

[127] B. Wang, S. Min, X. Deng, J. Shen, Y. Wu, L. Zettlemoyer, and H. Sun, "Towards understanding chain-of-thought prompting: An empirical study of what matters," *arXiv preprint arXiv:2212.10001*, 2022.

[128] H. L. Wang and H. Y. Xi, "Chinese synonyms for natural language processing and understanding," https://github.com/chatopera/Synonyms, 2017.

[129] ——, "Chinese synonyms tool box," https://github.com/chatopera/Synonyms, 2017.

[130] H. Wang, C. Liu, N. Xi, Z. Qiang, S. Zhao, B. Qin, and T. Liu, "Huatuo: Tuning llama model with chinese medical knowledge," 2023.

[131] J. Wang, E. Shi, S. Yu, Z. Wu, C. Ma, H. Dai, Q. Yang, Y. Kang, J. Wu, H. Hu *et al.*, "Prompt engineering for healthcare: Methodologies and applications," *arXiv preprint arXiv:2304.14670*, 2023.

[132] J. Wang, H. Deng, B. Liu, A. Hu, J. Liang, L. Fan, X. Zheng, T. Wang, and J. Lei, "Systematic evaluation of research progress on natural language processing in medicine over the past 20 years: Bibliometric study on pubmed," *Journal of medical Internet research*, vol. 22, no. 1, p. e16816, 2020.

[133] Y. Wang, Y. Kordi, S. Mishra, A. Liu, N. A. Smith, D. Khashabi, and H. Hajishirzi, "Self-instruct: Aligning language model with self generated instructions," *arXiv preprint arXiv:2212.10560*, 2022.

[134] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou *et al.*, "Chain-of-thought prompting elicits reasoning in large language models," *Advances in Neural Information Processing Systems*, vol. 35, pp. 24 824–24 837, 2022.

[135] G. Wen, H. Chen, H. Li, Y. Hu, Y. Li, and C. Wang, "Cross domains adversarial learning for Chinese named entity recognition for online medical consultation," *Journal of Biomedical Informatics*, vol. 112, p. 103608, 2020.

[136] L. S. Williams, E. Y. Yilmaz, and A. M. Lopez-Yunez, "Retrospective assessment of initial stroke severity with the NIH stroke scale," *Stroke*, vol. 31, no. 4, pp. 858–862, 2000.

[137] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. L. Scao, S. Gugger, M. Drame, Q. Lhoest, and A. M. Rush, "Transformers: State-of-the-art natural language processing," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Online: Association for Computational Linguistics, Oct. 2020, pp. 38–45. [Online]. Available: https://www.aclweb.org/anthology/2020.emnlp-demos.6

[138] B. Workshop, T. L. Scao, A. Fan, C. Akiki, E. Pavlick, S. Ilić, D. Hesslow, R. Castagné, A. S. Luccioni, F. Yvon *et al.*, "Bloom: A 176b-parameter open-access multilingual language model," *arXiv preprint arXiv:2211.05100*, 2022.

[139] S. Wu, K. Roberts, S. Datta, J. Du, Z. Ji, Y. Si, S. Soni, Q. Wang, Q. Wei, Y. Xiang *et al.*, "Deep learning in clinical natural language processing: a methodical review," *Journal of the American Medical Informatics Association*, vol. 27, no. 3, pp. 457–470, 2020.

[140] H. Xiong, S. Wang, Y. Zhu, Z. Zhao, Y. Liu, Q. Wang, and D. Shen, "Doctorglm: Fine-tuning your chinese doctor is not a herculean task," *arXiv preprint arXiv:2304.01097*, 2023.

[141] D. Xu, C. Wang, A. Khan, N. Shang, Z. He, A. Gordon, I. J. Kullo, S. Murphy, Y. Ni, W.-Q. Wei *et al.*, "Quantitative disease risk scores from EHR with applications to clinical risk stratification and genetic studies," *NPJ Digital Medicine*, vol. 4, no. 1, p. 116, 2021.

[142] K. Xue, Y. Zhou, Z. Ma, T. Ruan, H. Zhang, and P. He, "Fine-tuning bert for joint entity and relation extraction in Chinese medical text," in *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE, 2019, pp. 892–897.

[143] C. Yang, X. Wang, Y. Lu, H. Liu, Q. V. Le, D. Zhou, and X. Chen, "Large language models as optimizers," *arXiv preprint arXiv:2309.03409*, 2023.

[144] L. Yang, X. Huang, J. Wang, X. Yang, L. Ding, Z. Li, and J. Li, "Identifying stroke-related quantified evidence from electronic health records in real-world studies," *Artificial Intelligence in Medicine*, vol. 140, p. 102552, 2023.

[145] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. R. Salakhutdinov, and Q. V. Le, "Xlnet: Generalized autoregressive pretraining for language understanding," *Advances in neural information processing systems*, vol. 32, 2019.

[146] S. Yao, D. Yu, J. Zhao, I. Shafran, T. Griffiths, Y. Cao, and K. Narasimhan, "Tree of thoughts: Deliberate problem solving with large language models," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[147] S. Yao, J. Zhao, D. Yu, N. Du, I. Shafran, K. Narasimhan, and Y. Cao, "React: synergizing reasoning and acting in language models (2022)," *arXiv preprint arXiv:2210.03629*, 2023.

[148] Y. Yao, Z. Li, and H. Zhao, "Beyond chain-of-thought, effective graph-of-thought reasoning in language models," *arXiv preprint arXiv:2305.16582*, 2023.

[149] D. Zelenko, C. Aone, and A. Richardella, "Kernel methods for relation extraction," *Journal of machine learning research*, vol. 3, no. Feb, pp. 1083–1106, 2003.

[150] N. Zhang, M. Chen, Z. Bi, X. Liang, L. Li, X. Shang, K. Yin, C. Tan, J. Xu, F. Huang, L. Si, Y. Ni, G. Xie, Z. Sui, B. Chang, H. Zong, Z. Yuan, L. Li, J. Yan, H. Zan, K. Zhang, B. Tang, and Q. Chen, "CBLUE: A Chinese biomedical language understanding evaluation benchmark," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, May 2022, pp. 7888–7915. [Online]. Available: https://aclanthology.org/2022.acl-long.544

[151] N. Zhang, Q. Jia, K. Yin, L. Dong, F. Gao, and N. Hua, "Conceptualized representation learning for Chinese biomedical text mining," *arXiv preprint arXiv:2008.10813*, 2020.

[152] R. Zhang, W. Lu, S. Wang, X. Peng, R. Yu, and Y. Gao, "Chinese clinical named entity recognition based on stacked neural network," *Concurrency and Computation: Practice and Experience*, vol. 33, no. 22, p. e5775, 2021.

[153] S. Zhang, L. Wang, K. Sun, and X. Xiao, "A practical chinese dependency parser based on a large-scale dataset," *arXiv preprint arXiv:2009.00901*, 2020.

[154] T. Zhang, Z. Cai, C. Wang, M. Qiu, B. Yang, and X. He, "SMedBERT: A knowledge-enhanced pre-trained language model with structured semantics for medical text mining," *arXiv preprint arXiv:2108.08983*, 2021.

[155] X. Zhang, Y. Zhang, Q. Zhang, Y. Ren, T. Qiu, J. Ma, and Q. Sun, "Extracting comprehensive clinical information for breast cancer using deep learning methods," *International journal of medical informatics*, vol. 132, p. 103985, 2019.

[156] Z. Zhang, T. Zhou, Y. Zhang, and Y. Pang, "Attention-based deep residual learning network for entity relation extraction in Chinese EMRs," *BMC medical informatics and decision making*, vol. 19, no. 2, p. 55, 2019.

[157] Z. Zhang, L. Zhu, P. Yu *et al.*, "Multi-level representation learning for chinese medical entity recognition: Model development and validation," *JMIR Medical Informatics*, vol. 8, no. 5, p. e17637, 2020.

[158] Z. Zhang, A. Zhang, M. Li, and A. Smola, "Automatic chain of thought prompting in large language models," *arXiv preprint arXiv:2210.03493*, 2022.

[159] Q. Zhao, D. Wang, J. Li, and F. Akhtar, "Exploiting the concept level feature for enhanced name entity recognition in chinese emrs," *The Journal of Supercomputing*, pp. 1–22, 2019.

[160] S. Zheng, F. Wang, H. Bao, Y. Hao, P. Zhou, and B. Xu, "Joint extraction

of entities and relations based on a novel tagging scheme," *arXiv preprint arXiv:1706.05075*, 2017.

[161] Y. Zhou, X. Geng, T. Shen, C. Tao, G. Long, J.-G. Lou, and J. Shen, "Thread of thought unraveling chaotic contexts," *arXiv preprint arXiv:2311.08734*, 2023.

[162] X. Zhu, B. Qin, M. Liu, and L. Qian, *Knowledge Graph and Semantic Computing: Knowledge Computing and Language Understanding: 4th China Conference, CCKS 2019, Hangzhou, China, August 24-27, 2019, Revised Selected Papers*. Springer Nature, 2019, vol. 1134.