# Exploring the Application of 7B LLMs for Named Entity Recognition in Chinese Ancient Texts

**Chenrui Zheng[1], Yicheng Zhu[1], Han Bi[2]**
[1]East China Normal University，[2]Shandong Second Medical University
{10194800486,10210110437}@stu.ecnu.edu.cn,
2336007835@qq.com

## Abstract

This paper explores the application of fine-tuning methods based on 7B large language models (LLMs) for named entity recognition (NER) tasks in Chinese ancient texts. Targeting the complex semantics and domain-specific characteristics of ancient texts, particularly in Traditional Chinese Medicine (TCM) texts, we propose a comprehensive fine-tuning and pre-training strategy. By introducing multi-task learning, domain-specific pre-training, and efficient fine-tuning techniques based on LoRA, we achieved significant performance improvements in ancient text NER tasks. Experimental results show that the pre-trained and fine-tuned 7B model achieved an F1 score of 0.93, significantly outperforming general-purpose large language models.

## 1. Introduction

Named Entity Recognition (NER) is a foundational task in natural language processing (NLP) that involves identifying and categorizing entities, such as person names, locations, organizations, and temporal expressions—within unstructured text. Since its inception in the 1990s (Nadeau and Sekine, 2007), NER has evolved significantly, transitioning from rule-based systems to machine learning approaches, and more recently, to deep learning architectures like Bidirectional Long Short-Term Memory networks (Hochreiter and Schmidhuber, 1997) and transformer-based models (Devlin et al., 2019). These advancements have enabled robust performance in modern languages, particularly with the advent of pre-trained language models (e.g., BERT) that capture contextualized representations.

However, applying NER to ancient Chinese texts presents unique challenges. Ancient Chinese, characterized by archaic vocabulary, flexible grammar, and extensive use of homophones, diverges substantially from modern Mandarin. Additionally, historical texts often lack standardized punctuation and contain domain-specific terminology (e.g., official titles in dynastic records or disease names in medical classics), complicating entity boundaries and classification. Furthermore, annotated resources for ancient Chinese are scarce compared to those for contemporary languages, limiting the scalability of data-driven approaches.

## 2. Task Description

In EvaHan2025,[1] our team participates in the open modality track, which allows unrestricted use of external resources, models, and domain-specific knowledge to enhance NER in ancient Chinese texts. The task involves identifying 12 distinct entity categories across three heterogeneous datasets: (1) Shiji（"史记"） (historical records), (2) Twenty-Four Histories ("二十四史"), and (3) Traditional Chinese Medicine Classics ("中医药典籍"). To facilitate our evaluation, we split the dataset into train, development, and test sets using an 8:1:1 ratio. Our objective is to fine-tune the model to achieve the highest possible F1 score, optimizing its performance on the given task.

## 3. Related works

In the domain of NER for ancient Chinese texts, particularly in TCM, the field has witnessed a progression through various methodological approaches. Initially, dictionary-based and rule-based pattern matching methods, such as the maximum matching algorithm (Wang Y et al.,2012), were prevalent. The advent of deep learning ushered in new approaches. For instance, Xie et al.(2022) employed Sikubert and SikuRoBERTa, demonstrating that pre-trained

[1]https://github.com/GoThereGit/EvaHan/tree/main/evahan2025



*Proceedings of the Second Ancient Language Processing Workshop associated with NAACL 2025*, pages 106–111
May 4, 2025. 

models based on ancient texts outperform generic BERT models in these specialized NER tasks.In recent years, Large Language Models (LLMs) have exhibited significant potential in NER tasks. This aligns with contemporary research on LLM applications in NER, such as the work by Raffel et al. (2020), which illustrates that models like GPT and T5, when fine-tuned for NER tasks, can achieve superior performance by leveraging their extensive pre-trained knowledge and contextual understanding.

For NER tasks involving ancient Chinese texts, the flexibility and contextual understanding inherent in LLMs render them especially suitable for addressing the intricacies of these historical documents. He Yuhao's research(2024), for instance, revealed that LLMs outperformed other deep learning models in identifying and extracting entities and relationships from "ZhonghuaYaofang" ( "中华药方"), demonstrating superior performance across precision, recall, and F1-score metrics. This growing body of evidence underscores the rationale behind the present study's objective to further explore and harness the potential of LLMs in processing NER tasks for ancient texts.

## 4. Methods

In our study, we employed a combination of SikuRoBERTa, BiLSTM, and CRF methodologies to establish a robust baseline for our NLP tasks. After conducting 20 epochs, we evaluated our model on a segmented test dataset and obtained the following results in Table 1, which serve as the foundational benchmark for our experimental analysis.

Table 1: Performance Metrics of SikuRoBERTa+BiLSTM+CRF Model on Segmented Test Dataset after 20 Epochs

| Entity Type | Accuracy | Recall | F1 | Num |
|---|---|---|---|---|
| Symptom (ZS) | 54.34% | 66.20% | 59.68% | 142 |
| Traditional Medicine Disease (ZD) | 65.22% | 68.18% | 66.67% | 66 |
| Syndrome (ZZ) | 65.31% | 69.57% | 67.37% | 46 |
| Acupoint (ZA) | 86.36% | 78.08% | 82.01% | 73 |
| Chinese Formula (ZF) | 84.55% | 90.43% | 87.39% | 115 |
| Time Expression (T) | 88.47% | 83.53% | 85.93% | 340 |
| Official Title (NO) | 77.86% | 79.56% | 78.70% | 137 |
| Geographical Location (NS) | 87.97% | 82.01% | 84.88% | 517 |
| Book Title (NB) | 100% | 80% | 88.89% | 5 |
| Decoction Pieces (ZP) | 92.60% | 93.56% | 93.08% | 388 |
| Country Name (NG) | 90.51% | 96.25% | 93.30% | 347 |
| Person Name (NR) | 93.78% | 93.38% | 93.58% | 1194 |
| Overall | 0.878 | 0.8798 | 0.8789 | 3370 |

These metrics, while commendable, revealed a notable shortcoming when applied to Task C (Precision: 82.29%,Recall: 84.37%,F1 Score: 83.33%), which involves the processing of Traditional Chinese Medicine texts. Specifically, the BERT model's accuracy was significantly lower in this context compared to its performance on other tasks. We hypothesize that this discrepancy stems from BERT's limited exposure to and familiarity with the specialized terminology and nuanced semantics inherent in TCM texts.

To address this limitation, we propose the integration of more semantically aware LLMs that are better equipped to comprehend and process the complex linguistic structures found in TCM texts. By leveraging these advanced models, we aim to enhance the accuracy and effectiveness of our NLP applications in the domain of TCM texts, thereby improving the overall performance and reliability of our system in handling specialized medical texts.

※ Prompt1：

你是一名专注于处理中医领域的文献的专家，你的任务是从我提供的中医文献原文中，直接在原文的基础上标注以下实体:<病名>、<证候>、<方剂>、<饮片>、<症状>、<穴位>。

You are an expert specializing in processing literature from the field of Traditional Chinese Medicine. Your task is to annotate the following entities directly on the original text I provide: <Traditional Medicine Disease> , <Syndrome> , <Chinese Formula>, <Decoction Pieces> , <Symptom> , <Acupoint> .

- input:一男子时疫愈后，遍身发作痒，服补中益气汤而愈。
- input: A man who had recovered from an epidemic disease later developed itching all over his body, took Buzhong Yiqi Tang and recovered.

- output:<实体标注结果>一男子<病名>时疫</病名>愈后，遍身发作痒，服<方剂>补中益气汤</方剂>而愈。</实体标注结果>
- output: <Named entity recognition results>A man who had recovered from <Traditional Medicine Disease>epidemic disease</Traditional Medicine Disease>, later developed itching all over his body, took <Chinese Formula>Buzhong Yiqi Tang</Chinese Formula> and recovered.</Named entity recognition results>

※ Prompt2：

你是一名专注于处理中医领域的文献的专家。你的任务是从我提供的中医文献原文中，直接在原文的基础上标注以下实体:{病名}、{证候}、{方剂}、{药材}、{症状}、{穴位}。

You are an expert specializing in processing literature from the field of Traditional Chinese Medicine. Your task is to annotate the following entities directly on the original text I provide: {Traditional Medicine Disease}、{Syndrome}、{Chinese Formula}、{Chinese Formula}、{Symptom}、{Acupoint}.

- input:一男子时疫愈后，遍身发作痒，服补中益气汤而愈。
- input: A man who had recovered from an epidemic disease later developed itching all over his body, took Buzhong Yiqi Tang and recovered.

- output:{实体标注结果}一男子{时疫|病名}愈后，遍身发作痒，服{补中益气汤|方剂}而愈。{实体标注结果}
- output: {Named entity recognition results}A man who had recovered from {epidemic disease|Traditional Medicine Disease}, later developed itching all over his body, took {Buzhong Yiqi Tang|Chinese Formula} and recovered.{Named entity recognition results}

Figure 1 Two distinct prompt formats

### 4.1 Prompt Engineering

To advance our research effectively, the initial step involves the meticulous determination of the prompts to be used with LLMs. This is crucial for structuring both the input and output data in a manner that aligns with our objectives. Drawing upon previous studies, we have meticulously selected two distinct prompt formats that have demonstrated efficacy in similar contexts（Figure 1）.

In our experiments, we utilized the Qwen-Plus[2] model and the Task C test dataset under a 1-shot learning setting to evaluate the performance of the two prompt formats. Our findings revealed that Prompt 2 significantly outperformed Prompt 1 in terms of accuracy（Table 2）. We hypothesize that this is because Prompt 2 provides a more structured and contextually rich input by directly incorporating the original text, which allows the model to better understand the task and generate more accurate outputs. As a result, we decided to adopt Prompt 2 for all subsequent research.

Additionally, we compared the performance of the Qwen-Plus model with the Qwen-7B model. Surprisingly, the Qwen-7B model achieved a higher F1 score than Qwen-Plus on this specific task. We speculate that this may be due to the fact that Qwen-Plus, as a more general-purpose large language model, tends to "overthink" or generalize too much, making it less adaptable to the highly specialized and domain-specific nature of the task at hand. In contrast, the smaller Qwen-7B model, with its more focused architecture, may be better suited for handling the nuances and intricacies of this particular domain. These insights highlight the importance of tailoring both the prompt design and model selection to the specific requirements of

[2] https://github.com/QwenLM/Qwen

the task. Moving forward, we will continue to refine our approach by leveraging Prompt Prompt 2 and exploring the potential of smaller, more specialized models like Qwen-7B for domain-specific NLP tasks.

| Prompt | Precision | Recall | F1 Score |
|---|---|---|---|
| Prompt 1 （Qwen-Plus） | 0.7876 | 0.6729 | 0.717 |
| Prompt 2 （Qwen-Plus） | **0.8945** | **0.8398** | **0.8592** |
| Prompt 1 （Qwen-7B） | 0.8015 | 0.7574 | 0.7719 |

Table 2:difference between Prompt 1 and 2

## 5. Experiments

### 5.1 Data Transformation

To prepare the data for our experiments, we performed a series of preprocessing steps on the raw text data provided by Evahan2025. The original data was in TXT format, and our goal was to transform it into a structured format suitable for training and evaluation. The transformation process involved the following steps:

**5.1.1 Sentence Segmentation**:

We first segmented the text into sentence-level units using punctuation marks such as "。", "！", and "？". This step ensured that each sentence was treated as an independent unit for further processing.

**5.1.2 BEMS to Prompt 2 Conversion**:

The original data was annotated using the BEMS (Begin, End, Middle, Single) tagging scheme, which is commonly used for sequence labeling tasks. We converted these annotations into Prompt 2, a more structured and readable format that aligns with our prompt design. For example:

- Original Text: 一男子时疫愈后,遍身发作痒,服补中益气汤而愈。
- BEMS Tags: ['O', 'O', 'O', 'B-ZD', 'I-ZD', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'B-ZF', 'I-ZF', 'I-ZF', 'I-ZF', 'I-ZF', 'O', 'O', 'O']
- Converted Prompt 2:

{实体标注结果}一男子{时疫|病名}愈后,遍身发作痒,服{补中益气汤|方剂}而愈。{实体标注结果}

This conversion process made the annotations more interpretable and aligned with the input format required by our LLMs.

**5.1.3 Handling Long Sentences**:

After segmentation, we observed that some sentences in the training data were still relatively long. However, given the capability of modern LLMs to handle longer sequences, we decided not to further split these sentences. This approach preserved the contextual integrity of the text while ensuring that the models could still process the data effectively.

By transforming the data into Prompt 2, we created a structured and consistent input format that facilitated better model performance. This preprocessing step was critical for ensuring that the LLMs could accurately interpret and process the specialized terminology and semantic nuances present in the TCM texts.

In the next steps, we will use this transformed dataset to train and evaluate our models, with a focus on improving performance for domain-specific tasks.

### 5.2 Model Training and Results

In this section, we detail the model training process and the results obtained from our experiments. The fine-tuning was primarily conducted using the Unsloth framework[3] from, LLaMA-Factory(Zheng Yaowei,2024) and we explored several approaches to optimize the model's performance on the task of TCM text processing. The results can be seen in Table 3.

**5.2.1 Pre-Fine-Tuning Baseline**

Before fine-tuning, we evaluated the baseline performance of the model on the task. This provided a reference point to measure the impact of our subsequent fine-tuning strategies.

**5.2.2 Task C Specific Fine-Tuning**

We fine-tuned the model using the Task C TCM training data with LoRA (Low-Rank Adaptation). Each training sample included three components: Instructions, Input, and Output, following a SFT(supervised fine-tuning) approach. This method allowed the model to learn task-specific patterns and improve its performance on TCM text processing.

**5.2.3 Multi-Task Fine-Tuning (Task A, B, C)**

To further enhance the model's generalization capabilities, we combined the training data from Task A, Task B, and Task C into a single dataset for multi-task fine-tuning. This approach ex-

[3] https://github.com/unslothai/unsloth

posed the model to a more diverse range of materials, which led to a noticeable improvement in accuracy. The results confirmed that providing the model with more varied and extensive training data significantly enhances its performance. As a result, in all subsequent supervised fine-tuning (SFT) stages, we consistently used the combined dataset from all three tasks (A, B, and C) together. This multi-task approach became our standard practice for fine-tuning, leveraging the synergies between the different tasks to improve overall model performance.

#### 5.2.4 Pre-Training with External Data

To further boost the model's performance, we introduced a pre-training phase before fine-tuning. The pre-training data was sourced from the open-source project "殆知阁", [4]and We selected 12.5MB of unannotated classical Chinese text there, including 1/3 historical texts from the Twenty-Four Histories and 2/3 TCM-related texts.

We conducted training on a single NVIDIA RTX 4090 GPU. The learning rate was set to 5e-05 and the train batch size was 1. We first performed unsupervised pre-training using the unsloth framework on the unlabeled domain-specific text for 3 epochs. This was followed by supervised fine-tuning (SFT) using the same hyperparameters as befores(within a combination of Task A, B, C). After pre-training and fine-tuning, the final loss value decreased to around 0.0005. The loss curves for the SFT stage

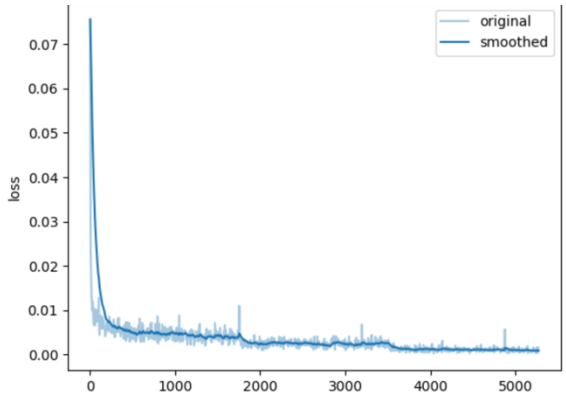


Figure 2 training loss of 7b Fine-tuned After - Multi + Pretrain (Prompt 2)

are shown in Figure 2.

After 3 epochs of pre-training and subsequent fine-tuning, the model **achieved an F1 score of 0.93,** significantly outperforming general-purpose LLMs. This demonstrated the effectiveness of domain-specific pre-training in improving model performance on the target task.

#### 5.2.5 Performance of Smaller Models (3B)

We also experimented with a smaller 3B parameter model using the same training methodology. Surprisingly, this model achieved an F1 score of 0.91, indicating that even smaller models can perform well on specialized tasks when properly trained.

#### 5.2.6 Ensemble Approach with 7B and 3B Models

To further improve accuracy, we implemented an ensemble approach:

- Both the 7B and 3B models generated results independently.
- If their results agreed, the output was considered correct.
- If their results disagreed, we used Qwen-Plus as a teacher model to determine which result was more reliable.

We initially envisioned that both 3B and 7B models could achieve certain accuracy levels. They're like students working on the same NER task - if they give the same answer, there's a higher probability that their results are correct. If they disagree, someone needs to judge who's right and who's wrong. While general large language models might overthink, they could be quite effective at determining which NER result is correct, so we combined these methods together.This ensemble method achieved a final F1 score of 0.9277. We hypothesize that integrating additional reasoning models, such as DeepSeek-R1, could further enhance performance. This represents a promising direction for future innovation.

| Model | Precision | Recall | F1 |
|---|---|---|---|
| 7b Fine-tuned Before (Prompt 1 ) | 0.8015 | 0.7574 | 0.7719 |
| 7b Fine-tuned (Prompt 1 ) | 0.8199 | 0.7811 | 0.7956 |
| 7b Fine-tuned | 0.8051 | 0.8415 | 0.8142 |
| **7b Fine-tuned - Multi** | 0.8805 | 0.897 | **0.8863** |
| **7b Fine-tuned - Multi + Pretrain** | **0.9297** | **0.9346** | **0.9302** |
| 3b Fine-tuned - Multi + Pretrain | 0.9137 | 0.9173 | 0.9147 |
| 7b+3b+Teacher | 0.9296 | 0.9268 | 0.9277 |

Table 3: Results of the experiments [Prompt 2 used unless otherwise specified]

[4]https://github.com/garychowcmu/daizhigev20

## 6. Conclusion

In this study, we found that domain-specific pre-training and multi-task fine-tuning significantly improved model performance on specialized tasks like TCM text processing. Interestingly, smaller models (e.g., 3B) were able to achieve competitive results when trained with the right methods, showing that model size is not always the limiting factor. Additionally, we found that ensemble methods, combined with teacher models, further enhanced accuracy and reliability. Future work will explore integrating more advanced reasoning models, such as DeepSeek-R1, to push the performance limits of domain-specific NLP tasks. These results highlight the importance of tailored training strategies and the potential of smaller, specialized models to achieve state-of-the-art results in niche domains.

## 7. Limitation

This study has several limitations. **The choice of Qwen-Plus** as the teacher model, while effective, was not extensively compared with alternatives, potentially impacting results. The potential of more advanced inference models remains unexplored. **The quality of data annotations**, crucial in specialized fields like Traditional Chinese Medicine, was not discussed, which could affect result reliability. Additionally, **the evaluation relied primarily on F1 scores**, overlooking other important metrics such as model robustness, inference speed, and performance in resource-constrained environments. These factors collectively suggest areas for future research and improvement in the current approach.

## Reference


Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. In *Proceedings of the 37th International Conference on Machine Learning*, pages 4171–4186. https://doi.org/10.48550/arXiv.1910.10683.

David Nadeau, Satoshi Sekine, 2007. A survey of named entity recognition and classification. Lingvisticae Investigationes, 30(1), 3-26. https://doi.org/10.1075/li.30.1.03nad.

He Yuhao, Li Ming, Luo Xiaolan, Liu Lili, Yang Qi, Zhu Bangxian, Lyu Yuhan, 2024. Research on entity and relation extraction from traditional Chinese medicine knowledge graphs based on GPTs. *Shanghai Journal of Traditional Chinese Medicine*, 58(8), 1-6.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Computing Research Repository, *Computing Research Repository*,arXiv:1810.04805. V2.https://arxiv.org/abs/1810.04805.

Pan Liu, Yanming Guo, Fenglei Wang, Guohui Li, 2022. Chinese named entity recognition: The state of the art. *Neurocomputing*, 473, 37-53. https://doi.org/10.1016/j.neucom.2021.10.101.

Sepp Hochreiter, Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation*, 9(8), 1735-1780.https://doi.org/10.1162/neco.1997.9.8.1735.

Xie Jing, Liu Jiangfeng, Wang Dongbo, 2022. Study on named entity recognition of Traditional Chinese Medicine classics: Taking SikuBERT pre-training model enhanced by the Flat-lattice Transformer for example. *Library Forum*, 42(10), 51-60.

Yaqiang Wang, Zhonghua Yu, Yongguang Jiang, Yongchao Liu, Li Chen, Yiguang Liu, 2012. A framework and its empirical study of automatic diagnosis of Traditional Chinese Medicine utilizing raw free-text clinical records. *Journal of Biomedical Informatics*, 45(2), 210-223.https://doi.org/10.1016/j.jbi.2011.10.003.

Zheng Yaowei, Zhang Richong, Zhang Junhao, Ye Yanhan, Luo Zheyan, 2024. LlamaFactory: Unified Efficient Fine-Tuning of 100+ Language Models. *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics,*V3: 400-410.https://doi.org/10.18653/v1/2024.acl-demos.38.