基于大语言模型的中文医学命名实体识别

吕腾啸[1], 罗凌[1]†, 吕慧怡[2], 孙媛媛[1], 王健[1], 林鸿飞[1]
[1]大连理工大学，计算机科学与技术学院，大连，116024
[2]大连医科大学附属第二医院，药学部，大连，116023
tengxiaolv@mail.dlut.edu.cn,{lingluo,syuan,wangjian,hflin}@dlut.edu.cn,
dmu_huiyilv@163.com

摘要

从中文文本中准确识别医学命名实体是实现中文医疗信息结构化的关键。传统机器学习方法在面对中文医学实体边界模糊和嵌套结构复杂等问题时效果有限。本文提出一种基于大语言模型的中文医学命名实体识别方法，首先通过任务重构将识别过程转化为文本生成任务，设计了适配的标注策略以统一处理平面与嵌套实体，然后引入实体筛选器过滤错误候选实体，最后通过大语言模型决策进行冲突消解与多模型集成提升系统整体鲁棒性。在CMeEE-V2与CCKS2019两个数据集上实验结果显示，所提方法在识别准确性与鲁棒性方面均达到当前先进水平，F1值分别为0.7785和0.8821。

关键词： 中文医学文本；命名实体识别；大语言模型

Chinese Medical Named Entity Recognition Based on Large Language Models

Tengxiao Lv[1], Ling Luo[1]†, Huiyi Lv[2], Yuanyuan Sun[1], Jian Wang[1], Hongfei Lin[1]
[1]School of Computer Science and Technology,Dalian University of Technology,Dalian,116024
[2]Department of Pharmacy,Second Affiliated Hospital of Dalian Medical University,Dalian,116023
tengxiaolv@mail.dlut.edu.cn,{lingluo,syuan,wangjian,hflin}@dlut.edu.cn,
dmu_huiyilv@163.com

Abstract

Accurate recognition of medical named entities in Chinese text is crucial for structuring clinical information. Traditional machine learning methods often fall short when dealing with vague entity boundaries and complex nested structures. This paper proposes a Chinese medical named entity recognition method based on large language models (LLMs). The task is reformulated as a text generation problem, with a unified annotation strategy designed to handle both flat and nested entities. Then, an entity filter removes wrong candidates, followed by conflict resolution and multi-model integration via LLMs to enhance overall robustness. Experiments on the CMeEE-V2 and CCKS2019 datasets demonstrate that the proposed method achieves competitive performance, with F1 scores of 0.7785 and 0.8821, respectively.

Keywords: Chinese Medical Texts , Named Entity Recognition , Large Language Models

† 通讯作者

1 引言

医学文本（如电子病历、医学书籍、医学文献等）在提升医疗服务效率、减少诊疗错误、辅助临床决策及科研分析等方面发挥着关键作用，已成为现代医疗体系的重要组成部分(杜晋华et al., 2022)。然而，此类文本多以自然语言撰写，结构往往不规范，呈现出半结构化甚至非结构化的特点，限制了其在医疗信息系统中的直接应用。因此，如何应用自然语言处理技术对医学文本进行信息抽取，转化为结构化数据，成为当前研究的热点方向(吴宗友et al., 2021)。不同于英文文本，在中文医学文本中，因汉语缺乏空格分词，命名实体边界识别更具挑战性。同时，医学文本中常包含复杂的嵌套结构，如图1中，在“血管周围单核细胞及浆细胞浸润”这一“临床表现”类型的实体中，还嵌套有“血管周围单核细胞”和“浆细胞”两个“身体”类型的实体，进一步增加了中文医学命名实体识别（Named Entity Recognition, NER）的难度。

医疗设备　身体　临床表现

显微镜下见神经细胞空泡形成，透明变性和染色体分解小神经胶质细胞浸润，血管周围单核细胞及浆细胞浸润。

Figure 1: 中文医学文本数据样例

现存中文医学NER方法主要采用基于预训练模型的序列标注方式实现，但受限于模型表达能力的局限，这些方法在应对语义歧义、实体边界模糊等医学文本特性方面存在明显不足。尤其在同时识别平面实体与嵌套实体的医学文本场景下，通常需要依赖如词对关系建模等复杂的结构设计或标注策略(Li et al., 2022)，增加了系统实现难度。

随着大语言模型（Large Language Models, LLMs）在自然语言处理中的广泛应用，其强大的文本生成和语义理解能力为中文医学NER带来了新思路。然而，现有大多数基于LLMs的通用NER方法，采用的是不包含实体位置的标注策略，即直接生成实体及其对应的类型(Wang et al., 2023)。这种方式虽然简化了任务形式，但在中文医学领域中，因存在大量术语歧义且对信息可靠性要求极高，缺失位置信息将严重影响识别结果的可用性与解释性。

针对上述问题，本文提出一种基于大语言模型的中文医学NER方法。该方法通过设计适配LLMs的标注策略，将中文医学NER从序列标注转化为文本生成任务；并构建筛选器以优化识别结果。在此基础上，进一步探索了多模型集成机制，以提升整体识别性能。本文的主要贡献总结如下：

（1）在中文医学NER上，系统地测试了现存基于LLMs的NER标注策略和评估了多种开源大语言模型（Llama(Grattafiori et al., 2024)、GLM(GLM et al., 2024)、Qwen(Yang et al., 2024)和DeepSeek蒸馏版(Guo et al., 2025)）的表现，并进行了全面的结果错误分析；

（2）提出了一种适配LLMs的中文医学NER方法，设计了能够识别位置的符号标注策略，可以同时识别平面和嵌套实体，并引入正负样本对比微调的筛选器，最后基于大语言模型决策进行了多模型集成，进一步提升实体识别准确性；

（3）在CMeEE-V2与CCKS2019两个中文医学NER数据集上进行了实验验证，结果显示所提出方法能够有效识别平实体、嵌套实体及其文中位置，达到当前先进水平，F1值分别为0.7785和0.8821。

2 相关工作

中文医学命名实体识别在近年来取得了显著进展。早期方法主要基于词典和规则，依赖构建医学专业词典及人工规则模板来进行实体匹配(杨锦锋et al., 2014)，但此类方法存在词典更新滞后、规则通用性差等局限性。为提升识别准确性，统计机器学习方法逐渐被引入中文医学NER任务，包括隐马尔可夫模型（HMM）(Fine et al., 1998)、支持向量机（SVM）(Cortes and Vapnik, 1995)以及条件随机场（CRF）(Lafferty et al., 2001)等。这些方法借助人工特征工程（如词性、偏旁部首等）取得了一定成效(罗凌et al., 2020)。深度学习的发展显著推动了中文医学NER方法的进化。BiLSTM-CRF架构通过结合双向上下文建模能力和标签依赖建模能力，成为早期主流模型(张华丽et al., 2020)；Lattice-LSTM则引入词汇信息以缓解中文分词边界模

糊的问题(Zhang and Yang, 2018)。进一步的研究引入注意力机制(李博et al., 2020)与模型融合策略(许力and 李建华, 2021)，不断提升识别性能。

近年来，预训练模型的引入成为中文医学NER领域的重大突破。BERT及其变体通过大规模无监督预训练捕捉语言语义信息，结合领域内微调策略，可显著提升中文医学文本上的实体识别效果(朱岩et al., 2021)。基于BERT的融合模型也逐渐成为主流方法，如结合BiLSTM、CNN、多头注意力机制和CRF的MC_BERT-BiLSTM-CNN-MHA-CRF模型(Chen et al., 2022)。在应对嵌套实体识别方面，W2NER模型(Li et al., 2022)将命名实体识别建模为词-词关系分类，规避了传统序列标注中标签冲突的问题。基于W2NER结构，CNER-GTSS(Zhao et al., 2023)引入U形分割模块，在类图像的特征矩阵中挖掘更丰富的语义表示；CNN-Nested-NER(Yan et al., 2023)则利用卷积神经网络建模评分矩阵中的空间关系；DiFiNet(Cai et al., 2024)进一步构建边界感知的语义区分与过滤网络，有效应对嵌套实体边界识别不敏感的问题。

随着大语言模型的发展，通用模型在自然语言处理任务中展现出强大性能。在医学领域，研究者主要对通用大语言模型进行领域适应性优化。例如，Taiyi在Qwen-7B-base基础上通过两阶段微调提升了其在医疗领域的命名实体识别能力(Luo et al., 2024)。此外，LoRa-LLaMA3模型(张云秋and 殷策, 2024)通过引入大规模中文医学问答语料及数据增强技术，进一步构建出专用于中文医学NER的大语言模型，但主要针对平面实体。

3 大语言模型的中文医学命名实体识别方法

本文构建了一个基于大语言模型的中文医学命名实体识别框架，如图2所示。该框架旨在将传统的命名实体识别任务由序列标注范式转化为更适合LLMs处理的文本生成任务，以实现对平面实体与嵌套实体的统一识别，并精确识别实体的起止位置。

首先，本文对比了四种不同的标注策略，包括JSON标注策略、HTML标注策略、索引标注策略以及符号标注策略，结合指令微调方法进行实验，评估其在中文医学NER任务中的表现差异。通过对比分析，最终确定符号标注策略作为后续建模基础。在此基础上，进一步构建正负样本用于模型微调，设计了实体筛选器，以提升模型对实体边界的识别能力。此外，本文还探索了基于大语言模型的多模型集成策略。通过设计提示方法对不同LLMs模型进行集成，进一步提升整体命名实体识别的性能表现，得到最终的命名实体识别结果。

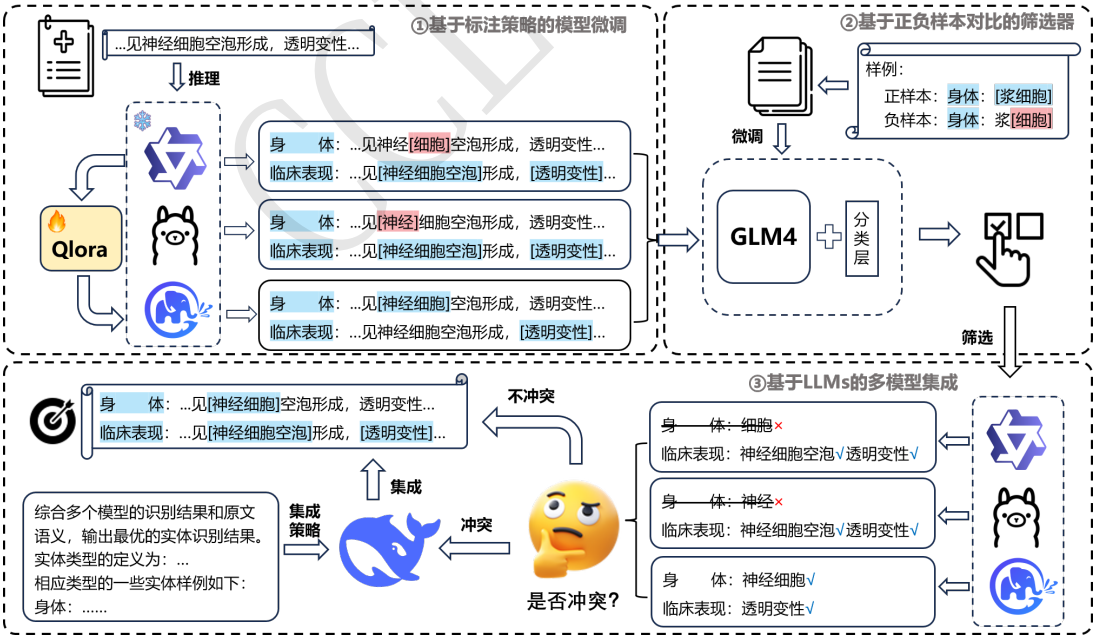

Figure 2: 基于大语言模型的中文医学命名实体识别框架图。蓝色为预测正确，红色为预测错误。

3.1 基于标注策略的模型微调

医学文本常常需要借助上下文来明确实体的真实含义，因此，中文医学NER不仅要求准确识别实体及其类型，还需要精确标注实体在原文中的起始与终止位置。这一点使得输出任务相较于传统命名实体识别更具复杂性。

本文从中文医学NER任务的角度出发，在现存大语言模型NER标注策略的基础上，设计了四种标注策略：JSON标注策略、HTML标注策略、索引标注策略与符号标注策略，相关Prompt示例如图3所示，输入首先设定模型医学命名实体识别专家的角色，其后依次提供原始句子以及各类实体类型的定义；输出则依据所选标注策略生成对应的标注结果。

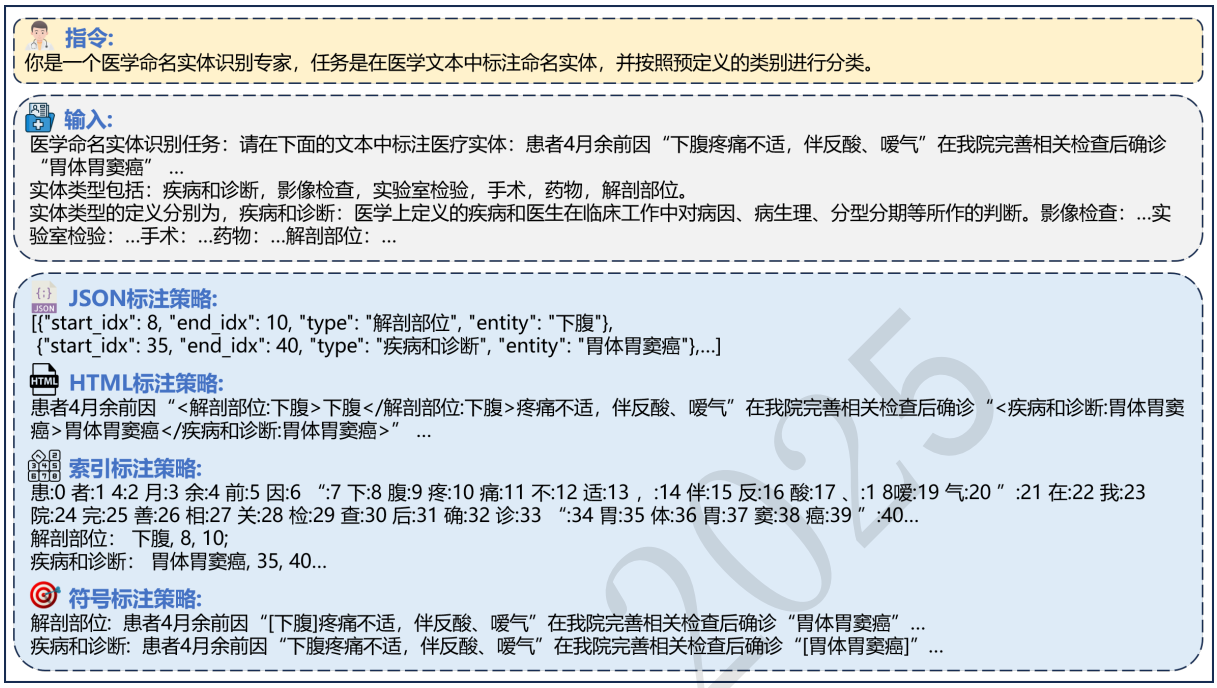

Figure 3: 不同标注策略示例图

JSON标注策略：仿照LoRA-Llama3(张云秋and 殷策, 2024)以结构化的键值对形式输出实体信息，包括实体在原始句子中的起始与结束索引位置（start_idx, end_idx）、实体类型（type）、以及实体文本（entity）。

HTML标注策略：引导模型使用带有实体类型的类属性标签来标注句子中的命名实体，其中<类型:实体>表示实体的起始位置，</类型:实体>表示实体的终止位置。与BioNER-Llama(Keloth et al., 2024)标签<mask></mask>只能应对单一类型平面实体不同，本文要求起止标签中的实体文本必须完全一致才视为匹配，以此解决多类型嵌套实体场景下的标签匹配问题。

索引标注策略：在输出中首先展示原始句子及其每个字符的索引位置，随后列出识别出的实体、实体类型以及实体在句子中的起止索引位置。这种方式通过显式标注字符位置，有助于模型更清晰地理解和定位实体范围。

符号标注策略：在输出过程中，首先明确指出实体的类型，随后输出原始文本序列。在输出原始文本序列时，通过使用特殊符号“[”和“]”对其中的实体进行标注，其中“[”表示实体的起始位置，“]”表示实体的终止位置。此方法按类型拆解分别进行标注，直观且易于模型学习实体边界，适用于泛化能力强的大语言模型结构。

本文在保持输入一致的前提下，采用QLoRA方法(Dettmers et al., 2023)对大语言模型进行参数高效微调，使其能够适应不同标注策略的输出要求。通过统一Prompt结构构建训练样本，引导模型学习各类标注策略下的输出风格，从而评估不同标注策略在中文医学NER任务中的性能差异。

3.2 基于正负样本对比的筛选器

本文对微调后的LLMs在中文医学NER任务中的错误进行统计与分析，并据此构建实体筛选器。具体而言，错误被划分为两大类：假阳性和假阴性，分别对应模型错误地识别了实体，或未能识别出实际存在的标准实体。在假阳性中，错误源于实体边界或类型的不一致，根据不同的匹配情况，进一步细分为以下几类：（1）位置正确类型错误；（2）位置重叠类型正确；（3）位置重叠类型错误；（4）位置非重叠。假阴性则指模型未预测到的目标实体。

本文在实验中发现，假阳性在整体识别错误中占据较大比例，且约50%的边界识别误差集中在2个token以内。这一现象表明LLMs在实体边界识别方面仍存在显著不足。然而，医学文本对NER的精度要求更高，尤其在临床等高风险应用中容错空间极小。为应对上述问题，本文提出一种基于正负样本对比的筛选器，在不依赖外部知识的前提下，提升模型对实体边界的敏感性和整体识别精度。该筛选器构建于最终采用的符号标注策略之上，其训练数据由正负样本组成，具体构建方式如下：

以训练集中已标注的实体作为正样本，通过对其起始或终止位置进行随机偏移（±1或±2），或更改实体类别的方式生成对应的负样本。最终从中随机采样，构建了包含10,000个正负样本的微调数据集，相关Prompt请参考**附录A**，输入部分首先给出任务要求，其后提供待判断的句子及其中的候选实体；随后给出实体类型及其定义。输出部分需先识别标注类型及实体，再判断该实体是否为正确。

在此数据基础上，本文采用GLM4模型进行微调，在其自注意力模块上引入分类层，用于实现对候选实体的判断，以确定其是否为有效实体。具体过程如下：

设原始文本序列为$\mathbf{x} = [x_1, x_2, \ldots, x_i, \ldots, x_j, \ldots, x_n]$，$i$ 和j 分别为候选实体在原文中的起始和终止位置，满足$0 \leq i \leq j \leq n$。在候选实体前后插入特殊标记符，构造新的输入序列$\mathbf{x}'$，形式如下：

$$\mathbf{x}' = \left[x_{0:(i-1)} \oplus s_1 \oplus x_{i:j} \oplus s_2 \oplus x_{j+1:n}\right] \tag{1}$$

其中，s_1 与s_2 分别表示实体的起始与结束标记符，$\oplus$ 表示序列拼接操作。

构造后的序列$\mathbf{x}'$ 被输入至GLM4模型进行上下文编码，得到上下文隐藏表示序列$\mathbf{h}$：

$$\mathbf{h} = \text{SelfAttention}(t \oplus \mathbf{x}') \tag{2}$$

其中，SelfAttention 表示多层自注意力机制，t 为实体类型表示。

随后，从输出中提取[CLS]位置对应的向量$\mathbf{h}_{[\text{CLS}]}$，并将其输入至分类层，预测该候选实体是否为真实实体。

通过上述方法，本文构建了一个实体筛选器LLM-CLS。结合大语言模型生成的候选实体结果，该筛选器能够过滤边界或类别错误的无效候选，从而提升命名实体识别任务的可靠性，更好地适用于中文医学场景。

3.3 基于LLMs的多模型集成

最近，多个大语言模型的集成方法被广泛用于提升模型性能，尤其在分类、问答等任务中展现出了显著效果(Li et al., 2024)。然而，与分类或问答任务相比，中文医学NER对输出结果的边界和语义一致性要求更高，因此在集成上需要LLMs更加精细化的思考。为此，本文提出了一种基于LLMs的集成方法。从实体边界和类型两个角度，利用LLMs的指令响应能力与上下文建模优势，以实体为最小单元进行冲突消解与集成，具体Prompt参考**附录B**。整体流程如下：

（1）冲突实体检测与定位：对多个模型的预测结果进行逐实体比对，识别出存在差异的冲突实体区域，作为后续集成判断的候选输入；

（2）上下文增强的判别Prompt构造：为提升模型对实体边界的判断能力，针对冲突实体，构建包含原始文本和各模型候选结果的上下文Prompt，引导模型在上下文中对比并感知各实体边界的合理性；

（3）类型定义与代表性实体示例注入：为增强模型对实体类型的判别能力，Prompt中进一步注入实体类型的定义信息。考虑到NER任务更关注实体层面的类型理解，本文基于训练集，采用K-Means聚类为每种类型选取5个代表性中心实体作为示例，引导模型进行语义类比与边界判定；

（4）生成集成输出结果：在上述上下文信息与示例引导下，LLMs根据统一的符号标注策略输出最终的集成识别结果，实现冲突实体的融合与统一。

4 实验结果与分析

4.1 数据集和实验设置

本文采用了CMeEE-V2数据集(Zhang et al., 2021)和CCKS2019数据集(Han et al., 2020)，其统计信息分别见表1和表2。CMeEE-V2为包含嵌套实体识别数据集，CCKS2019则为平面实体识别数据集。由于CMeEE-V2未提供测试集的标准答案且受资源限制，本文随机选取了500条开发集样本作为测试集使用。实验中主要超参数设置如下：batch size设为6，learning rate为1e-4。

CMeEE-V2	疾病	临床表现	医疗程序	医疗设备	药物	微生物	身体	科室	医学检验项目
训练集	19371	17118	9688	889	4379	2343	24106	345	4410
测试集	564	489	382	19	108	104	505	7	97

Table 1: CMeEE-V2数据集统计

CCKS2019	疾病和诊断	影像检查	实验室检验	手术	药物	解剖部位
训练集	4212	969	1195	1029	1822	8426
测试集	1323	348	590	162	485	3094

Table 2: CCKS2019数据集统计

4.2 不同标注策略的性能影响

本实验旨在评估四种标注策略在处理嵌套命名实体识别任务（CMeEE-V2）与平面命名实体识别任务（CCKS2019）中的性能表现。所有实验均基于GLM4-9B-Chat模型进行微调，并保持统一的输入设定与训练流程，以公平比较不同标注策略对识别效果的影响。实验结果汇总如表3所示。

	CMeEE-V2			CCKS2019		
	Precision	Recall	F1	Precision	Recall	F1
JSON标注策略	0.7086	0.7726	0.7392	0.8217	0.8358	0.8287
HTML标注策略	0.6921	0.7143	0.7030	0.8405	0.8410	0.8408
索引标注策略	0.7259	0.7754	0.7498	0.8346	0.8412	0.8379
符号标注策略	**0.7366**	**0.7881**	**0.7615**	**0.8674**	**0.8624**	**0.8649**

Table 3: 四种标注策略性能对比

实验结果显示：(1) 符号标注策略在两个数据集上均取得了最优的F1值，在CMeEE-V2数据集上达到0.7615，在CCKS2019上达到0.8649，相比其他三种输出形式均有提升。这表明该标注策略在定位实体边界和识别实体类型方面具有更强的表达能力，尤其适合中文医学文本中复杂结构的实体表示需求；(2) 索引标注策略在CMeEE-V2数据集上表现次优，整体优于JSON和HTML标注策略。其优势可能在于该策略在输出过程中明确标注了实体的起止索引位置，利于模型学习边界信息；(3) JSON与HTML两种结构化标注策略的表现相对较弱。尤其是在CMeEE-V2上HTML标注策略的F1值仅为0.7030，这可能是因为CMeEE-V2中句法结构复杂，嵌套实体较多，HTML标注策略会导致数据结构复杂，不利于生成式模型学习，导致生成时易出现解析误差或实体漏识，具体样例请参考**附录C**；JSON标注策略在CCKS2019数据集上性能较差，观察结果表明，大语言模型难以准确理解并生成代表实体起止位置的数字信息，导致该输出方式在命名实体识别任务中存在较大的性能劣势。

本文进一步探索不同标注策略对推理效率的影响，不同策略所要求的输出结构复杂度和不同，进而导致生成文本长度差异，对推理时间造成显著影响。为保证比较的公平性，本文在基座模型一致（均为GLM4-9B-Chat）和统一运行环境（单张NVIDIA L40 GPU）下，

于CMeEE-V2数据集（平均每句含6个实体）上，对四种标注策略的推理耗时进行了评估。相关推理效率的量化数据如表4。

标注策略	推理时间(s/句)	显存占用(GB)	F1
JSON标注策略	4.38	19.21	0.7392
HTML标注策略	2.71	19.22	0.7030
索引标注策略	4.75	19.24	0.7498
符号标注策略	3.12	19.27	0.7615

Table 4: 不同标注策略的推理效率对比

从实验结果可以观察到：(1)符号标注策略在F1最高的情况下，推理时间与显存占用处于中等水平，体现了精度与效率的良好平衡；(2)不同策略显存使用差异较小，表明效率瓶颈主要来自输出长度而非计算资源。

4.3 不同LLMs的实体识别性能对比

根据上节实验结果，选定性能最佳的符号标注策略进行后续实验，本节实验主要比较4种不同LLMs基座：Qwen2.5-7B-Instruct、Llama-3.1-8B-Instruct、GLM4-9B-Chat和DeepSeek-R1-Distill-Qwen-7B（DS-R1-Qwen-7B）的zero-shot和微调（fine-tuning）效果，相关F1值如图4。

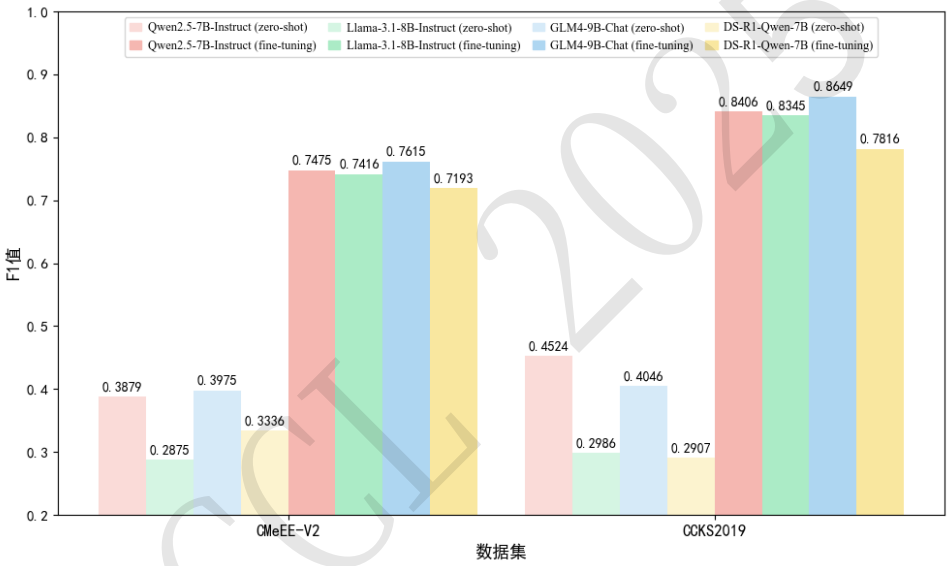

Figure 4: 不同大语言模型基座的微调效果

从图4中可以观察到：(1) 四种LLMs在zero-shot条件下的F1值整体偏低，表明在未经过微调的情况下，LLMs在中文医学NER任务中的能力仍有较大提升空间；(2) 在微调条件下，所有模型的性能均显著提升，其中GLM4-9B-Chat在两个数据集上分别取得了最高F1值，而善于推理的DeepSeek蒸馏模型微调后性能较差，甚至差于原始的Qwen模型；(3) 总体来看，微调显著增强了LLMs在中文医学NER任务中的表现，进一步验证了符号标注策略与指令微调相结合是提升模型实体识别效果的有效途径。

本文进一步对微调后的模型输出结果进行了错误类型统计与边界误差分析。图5展示了性能最好的GLM4-9B-Chat在符号标注策略微调后，在CMeEE-V2与CCKS2019两个数据集上不同类型识别错误的分布情况；图6进一步对实体边界识别错误的token级长度差值范围进行了细化统计。

从图5可见，假阳性在识别错误中占据较大比例，尤其在嵌套实体数据集中达到了57.1%，说明当前模型在识别实体边界与类型时仍存在明显不足。图6显示，大约50%的边界识别误差集中在2个token以内，表明模型倾向于在实体边界附近产生微小偏差，进一步说明实体边界感知能力的提升具有重要意义。

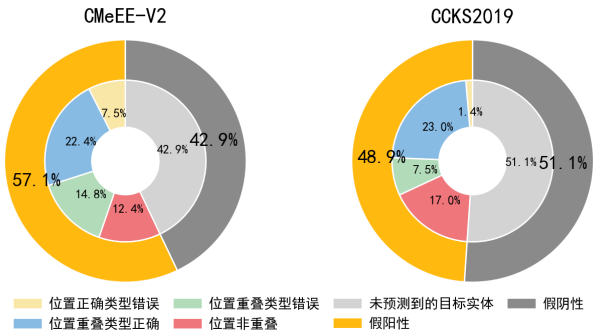

Figure 5: 不同类型识别错误占比

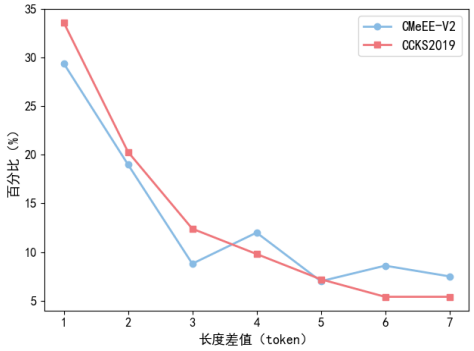

Figure 6: 长度差值占比

4.4 筛选器性能结果

为验证基于正负样本对比思想构建的命名实体识别筛选器在提升模型性能方面的有效性，本文在实验中对比了三种不同的筛选机制在命名实体识别任务中的表现，具体包括：基于微调的BERT分类器（BERT-CLS）、基于大语言模型GLM4-9B-Chat的自回归式筛选器（LLM-CAUSAL），以及本文提出的基于大语言模型GLM4-9B-Chat的分类层筛选器（LLM-CLS）。上述三种机制与基础模型在CMeEE-V2和CCKS2019两个数据集上的性能如表5所示，其中基线模型为GLM4-9B-Chat，采用符号标注策略并完成微调，LLM-CLS在其他模型下的应用效果如**附录D**所示。

	CMeEE-V2			CCKS2019		
	Precision	Recall	F1	Precision	Recall	F1
基线	0.7366	**0.7881**	0.7615	0.8674	**0.8624**	0.8649
+BERT-CLS	0.7427	0.7868	0.7641	0.8698	0.8590	0.8644
+LLM-CAUSAL	0.7492	0.7876	0.7679	0.8842	0.8587	0.8712
+LLM-CLS	**0.7586**	0.7873	**0.7726**	**0.8963**	0.8607	**0.8781**

Table 5: 不同的筛选机制在命名实体识别任务中的表现

从表5中可以观察到：(1) 三种筛选器均在两数据集上提升了基线模型命名实体识别的精确率，验证了引入基于正负样本对比的筛选器在中文医学NER中的有效性；(2) 在两个数据集上，LLM-CLS取得了最优的F1值，分别为相比基线模型提升了约1.1%和1.3%，说明LLM-CLS通过精确建模实体边界与类型一致性，有效减少了假阳性实体的干扰，提升了整体识别质量；(3) 与LLM-CLS相比，LLM-CAUSAL虽然在CMeEE-V2召回率较高，但精确率较低，导致F1值略低，说明LLM-CLS面对复杂嵌套实体能在准确率与召回率之间取得更好的平衡；(4) BERT-CLS筛选器也在CMeEE-V2上获得了轻微提升，但整体优于基线的幅度有限，且在CCKS2019上并没有优势，这说明虽然基于BERT的分类具备一定判别能力，但泛化能力逊于LLMs。

本文进一步以GLM4-9B-Chat用符号标注策略微调，于CMeEE-V2上产生的四类错误类型作为基线，统计在应用本文方法之后，每一类错误数量的变化情况，如表6。

	位置正确类型错误	位置重叠类型正确	位置重叠类型错误	位置非重叠
基线	84	252	166	139
Ours	74(↓12%)	218(↓13%)	157(↓5%)	121(↓13%)

Table 6: 四类实体错误数量在筛选器应用前后的变化统计

从表中结果可以观察到：(1)本文的方法在所有四种错误类型上均实现了错误数量的减少，表明本文筛选器的方法能够有效过滤各类错误；(2)位置重叠类型正确和位置非重叠的改善幅度

最大，均达到了13%，说明本文方法能较好修正边界错误。嵌套实体各错误类型数量变化情况请参考**附录E**。

4.5 多模型集成结果

本实验旨在对比不同多模型集成策略在中文医学NER任务中的性能差异，探索更适用于中文医学NER任务的集成方法。对比所用的单模型包括Qwen2.5+LLM-CLS、Llama3.1+LLM-CLS和GLM4+LLM-CLS，这些模型均基于符号标注策略进行微调，并引入LLM-CLS筛选机制。为评估不同LLMs集成策略的有效性，本文选取了两个具有代表性的LLMs进行对比实验：一个是DeepSeek的蒸馏版本长思考模型DeepSeek-R1-Distill-Qwen-7B（DS-R1-Qwen-7B），该模型通过知识蒸馏技术在保持长思考能力的同时显著降低了推理开销；另一个是非蒸馏版本模型Qwen2.5-7B-Instruct（Qwen2.5-7B）。各模型及其集成策略的性能如表7所示。

	CMeEE-V2			CCKS2019		
	Precision	Recall	F1	Precision	Recall	F1
Qwen2.5+LLM-CLS	0.7536	0.7662	0.7598	0.8717	0.8317	0.8512
Llama3.1+LLM-CLS	0.7456	0.7613	0.7534	0.8625	0.8265	0.8441
GLM4+LLM-CLS	0.7586	0.7873	0.7726	0.8963	0.8607	0.8781
硬投票集成	**0.7987**	0.7446	0.7707	**0.8989**	0.8522	0.8749
Qwen2.5-7B集成	0.7336	**0.8219**	0.7753	0.8768	**0.8836**	0.8802
DS-R1-Qwen-7B集成	0.7595	0.7984	**0.7785**	0.8894	0.8748	**0.8821**

Table 7: 不同多模型集成策略在中文医学NER任务中的性能差异

从表7实验结果可以看出：(1) 相较于最好的基线结果GLM4+LLM-CLS，采用硬投票的集成策略虽然将精确率提升了4.0%，但召回率下降4.3%，导致F1值反而下降约0.19%。这一现象表明基于规则的多数表决机制可能过于保守，容易忽略低置信度但真实存在的实体，适用于对精确率要求更高的场景；(2) 基于LLMs的集成方法在保持较高精确率的同时提升了召回率，从而带来整体F1的提升。例如，DeepSeek-R1-Distill-Qwen-7B相对GLM4+LLM-CLS在CMeEE-V2数据集F1值提升了0.59%，并在精确率与召回率之间取得了良好平衡，还在CCKS2019数据集表现出较强的泛化能力；(3) 不同的大语言模型在集成策略上也呈现出差异性：Qwen2.5-7B更倾向于保留更多候选实体，在CMeEE-V2数据集其召回率相对GLM4+LLM-CLS提升3.5%，说明其实体覆盖能力较强，适用于对召回要求较高的任务；而DeepSeek模型能够通过思考合理选择正确实体，在模型集成上更具优势。

在多模型集成中需要考虑到不同模型的结构和推理特点，尤其是模型是否具备“长思考能力”。因此，本文进一步对Prompt进行差异化设计的探索，具体体现在是否引导模型进行分步推理。本文选取了两种具代表性的模型：Qwen2.5-7B-Instruct与DeepSeek-R1-Distill-Qwen-7B（长思考模型），设计了两种Prompt模板进行对比：(1)直接推理标准Prompt：包含角色设定、原文上下文、实体类型定义及典型示例；(2)分步推理Prompt：在标准Prompt基础上，加入“请逐步推理：1. 定位冲突实体的上下文范围；2. 根据实体类型定义判断边界合理性；3. 对比各模型预测，选择最符合医学语义的结果。”以引导模型逐步做出判断。在CMeEE-V2数据集上的实验结果如表8。

模型	Prompt策略	Precision	Recall	F1
Qwen2.5-7B-Instruct	直接推理	0.7336	0.8219	0.7753
	分步推理	0.7191	0.8336	0.7721
DeepSeek-R1-Distill-Qwen-7B	直接推理	0.7595	0.7984	0.7785
	分步推理	0.7496	0.7974	0.7727

Table 8: 基于模型是否具备长思考能力的多模型集成prompt对比

从实验结果可以看出：(1)分步推理未提升性能，两类模型在逐步推理Prompt下性能相对于直接推理下降，说明额外的推理步骤可能引入冗余噪声，而非增强决策精度，并且这对长思

考模型影响更大；(2)当前实验支持统一Prompt的合理性，集成性能更多依赖模型语义理解能力而非Prompt形式。

4.6 与其他现存方法性能对比

如表9所示，本文方法（Ours）和集成前性能最好的单模型GLM4+LLM-CLS（Ours w/o ensemble）在CMeEE-V2和CCKS2019两个数据集上与多种先进的命名实体识别模型进行了性能对比，以验证其有效性。由于部分模型未开源，导致部分实验结果仅在单一数据集上可得。

- BERT-CRF：采用BERT模型生成字符表征，条件随机场（CRF）层进行序列标注。
- ELMo-lattice-LSTM-CRF(Li et al., 2020)：提出结合上下文化字符表示（ELMo）与格点结构的长短期记忆网络（lattice-LSTM），用于构建中文医学NER模型。
- MC_BERT-BiLSTM-CNN-MHA-CRF(Chen et al., 2022)：在医学预训练模型MC_BERT基础上，结合BiLSTM、CNN、MHA与CRF层进行特征建模与解码。
- W2NER(Li et al., 2022)：提出将统一命名实体识别建模为词对关系分类的方法。
- CNN-Nested-NER(Yan et al., 2023)：用CNN建模span特征矩阵的空间关系，提升嵌套命名实体识别性能。
- CNER-GTSS(Zhao et al., 2023)：基于网格标注和语义分割，将字符对关系分类转化为像素级掩码预测问题。
- DiFiNet(Cai et al., 2024)：构建了一种边界感知的语义差异和过滤网络，通过引入自适应语义差异模块和边界过滤模块，提升嵌套命名实体识别的性能。
- LoRa-Llama3(张云秋and 殷策, 2024)：在原始数据基础上通过引入额外的中文医学问答语料与数据增强手段将数据规模扩充十倍，并对Llama3模型进行LoRA微调。
- DeepSeek-R1、DeepSeek-R1-Distill-Qwen-14B（DS-R1-Qwen-14B）(Guo et al., 2025)：直接调用模型采用符号标注策略进行zero-shot命名实体识别。

	CMeEE-V2			CCKS2019		
	Precision	Recall	F1	Precision	Recall	F1
BERT-CRF	0.7155	0.7232	0.7193	0.8249	0.7991	0.8118
ELMo-lattice-LSTM-CRF*	-	-	-	0.8469	0.8535	0.8502
MC_BERT-BiLSTM-CNN-MHA-CRF*	-	-	-	0.8490	**0.8767**	0.8627
W2NER	0.7345	0.7503	0.7423	0.8560	0.8051	0.8297
CNN-Nested-NER	0.7239	0.7723	0.7473	0.8279	0.8387	0.8333
CNER-GTSS*	**0.7659**	0.7464	0.7560	-	-	-
DiFiNet	0.7237	0.7749	0.7485	0.8169	0.8444	0.8304
LoRa-Llama3*	-	-	-	0.8889	0.8660	0.8773
DeepSeek-R1	0.4918	0.5556	0.5218	0.4676	0.5924	0.5226
DS-R1-Qwen-14B	0.4281	0.4189	0.4235	0.4701	0.4288	0.4485
Ours w/o ensemble	0.7586	0.7873	0.7726	**0.8963**	0.8607	0.8781
Ours	0.7595	**0.7984**	**0.7785**	0.8894	0.8748	**0.8821**

Table 9: 与其他方法性能对比。*表示结果引用于原论文，加粗为最高值，下划线为次高值。

从表9结果中可以观察到：(1) 在CMeEE-V2和CCKS2019两个数据集上，本文提出的基于大语言模型的中文医学NER方法均获得了最高的F1值，分别达到0.7785和0.8821，相较目前已有的先进嵌套实体识别方法CNER-GTSS在CMeEE-V2上提升了2.3%，相较于先进的LoRa-Llama3在CCKS2019上提升了0.5%，表明本文方法在不同类型命名实体识别任务中的有效性；(2) 在CMeEE-V2包含嵌套实体数据集上，本文方法不仅在F1值上优于当前多个主流嵌

套NER模型，如CNN-Nested-NER、DiFiNet等，在召回率方面表现最为突出，相比DiFiNet提升了2.4%，体现出其对复杂实体结构的识别能力；(3) 在CCKS2019平面实体识别数据集中，本文方法在保持较高精确率的同时兼顾了较强的召回能力，整体F1值超过LoRa-Llama3等模型，说明其在中文医学NER任务中具备良好的泛化能力；(4) 对比直接调用DeepSeek-R1等通用大语言模型的中文医学NER能力，本文通过设计任务特定的标注策略微调、正负样本筛选器和多模型集成策略，有效弥补了大语言模型在中文医学NER任务中知识不足、边界不清等短板，使得经过指令微调的大语言模型在特定领域任务中展现出先进性能。

4.7 未预测到的目标实体案例分析

为进一步探索可能的改进方向，本文对未预测到的目标实体案例进行分析。发现部分错误并非源于常规的边界模糊，而是由于模型在处理跨短句或跨结构片段的语义整合能力不足，难以准确识别出长距离依赖下的复杂医学实体，如下表10所示，可以看出LLMs在医学NER中在处理跨短语、长距离修饰结构时难以建立稳定语义联结，未来计划引入医学知识图谱进行语义联结，通过在推理过程中辅助模型对接图谱中已有实体，提升其对复杂医学命名实体描述整体性的建模能力。

实体类别	目标实体	原因分析
医疗程序	入院后连续3天动态检测	实体内部包含时间修饰和动宾组合，语义线索分布在多个短语中，模型未能整合为一个统一动作。
临床表现	胸部X线片改变出现较早，可见大小不等的片状阴影，或融合成大病灶	实体为复合描述，语义分散且涉及多个病理特征，模型往往分裂识别成多个局部片段。

Table 10: 未预测到的目标实体案例展示和原因分析

5 结论与展望

本文针对中文医学命名实体识别中实体边界模糊与嵌套结构复杂的问题，提出了一个基于大语言模型的生成式识别框架。在CMeEE-V2与CCKS2019两个数据集上的实验结果表明，符号标注策略在实体边界定位和类型识别方面效果最优；引入的筛选器减少了假阳性实体的干扰；多模型集成策略则进一步提升了整体识别性能。最终所提出方法在两个数据集上均取得了当前先进水平。

但本文方法推理效率仍有优化空间，对于未能预测出的实体仍有改进余地。未来工作将着重于提升模型的推理速度与部署效率，同时引入中文医学知识图谱，提升对实体隐含语义的理解与泛化能力，从而更好地服务于真实临床场景中的医学信息抽取任务。

致谢

本研究得到国家自然科学基金项目（62302076，62276043）、国家卫生健康委医院管理研究所医院药学高质量发展研究项目（NIHAYSZX2525）资助。

参考文献

Yuxiang Cai, Qiao Liu, Yanglei Gan, Run Lin, Changlin Li, Xueyi Liu, Da Luo, and JiayeYang JiayeYang. 2024. Difinet: Boundary-aware semantic differentiation and filtration network for nested named entity recognition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6455–6471.

Peng Chen, Meng Zhang, Xiaosheng Yu, and Songpu Li. 2022. Named entity recognition of chinese electronic medical records based on a hybrid neural network and medical mc-bert. *BMC medical informatics and decision making*, 22(1):315.

Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. *Machine learning*, 20:273–297.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. *Advances in neural information processing systems*, 36:10088–10115.

Shai Fine, Yoram Singer, and Naftali Tishby. 1998. The hierarchical hidden markov model: Analysis and applications. *Machine learning*, 32:41–62.

Team GLM, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Dan Zhang, Diego Rojas, Guanyu Feng, Hanlin Zhao, et al. 2024. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *arXiv preprint arXiv:2406.12793*.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

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

Xianpei Han, Zhichun Wang, Jiangtao Zhang, Qinghua Wen, Wenqi Li, Buzhou Tang, Qi Wang, Zhifan Feng, Yang Zhang, Yajuan Lu, et al. 2020. Overview of the ccks 2019 knowledge graph evaluation track: entity, relation, event and qa. *arXiv preprint arXiv:2003.03875*.

Vipina K Keloth, Yan Hu, Qianqian Xie, Xueqing Peng, Yan Wang, Andrew Zheng, Melih Selek, Kalpana Raja, Chih Hsuan Wei, Qiao Jin, et al. 2024. Advancing entity recognition in biomedicine via instruction tuning of large language models. *Bioinformatics*, 40(4):btae163.

John Lafferty, Andrew McCallum, Fernando Pereira, et al. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Icml*, volume 1, page 3. Williamstown, MA.

Yongbin Li, Xiaohua Wang, Linhu Hui, Liping Zou, Hongjin Li, Luo Xu, Weihai Liu, et al. 2020. Chinese clinical named entity recognition in electronic medical records: development of a lattice long short-term memory model with contextualized character representations. *JMIR Medical Informatics*, 8(9):e19848.

Jingye Li, Hao Fei, Jiang Liu, Shengqiong Wu, Meishan Zhang, Chong Teng, Donghong Ji, and Fei Li. 2022. Unified named entity recognition as word-word relation classification. In *proceedings of the AAAI conference on artificial intelligence*, volume 36, pages 10965–10973.

Junyou Li, Qin Zhang, Yangbin Yu, Qiang Fu, and Deheng Ye. 2024. More agents is all you need. *arXiv preprint arXiv:2402.05120*.

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

Xiao Wang, Weikang Zhou, Can Zu, Han Xia, Tianze Chen, Yuansen Zhang, Rui Zheng, Junjie Ye, Qi Zhang, Tao Gui, et al. 2023. Instructuie: Multi-task instruction tuning for unified information extraction. *arXiv preprint arXiv:2304.08085*.

Hang Yan, Yu Sun, Xiaonan Li, and Xipeng Qiu. 2023. An embarrassingly easy but strong baseline for nested named entity recognition. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1442–1452.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. 2024. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*.

Yue Zhang and Jie Yang. 2018. Chinese ner using lattice lstm. *arXiv preprint arXiv:1805.02023*.

Ningyu Zhang, Mosha Chen, Zhen Bi, Xiaozhuan Liang, Lei Li, Xin Shang, Kangping Yin, Chuanqi Tan, Jian Xu, Fei Huang, et al. 2021. Cblue: A chinese biomedical language understanding evaluation benchmark. *arXiv preprint arXiv:2106.08087*.

Xuedong Zhao, Zhiliang Shi, Yan Xiang, and Ying Ren. 2023. Chinese named entity recognition based on grid tagging and semantic segmentation. In *2023 IEEE 9th International Conference on Cloud Computing and Intelligent Systems (CCIS)*, pages 289–294. IEEE.

第二十四届中国计算语言学大会论文集，第102页-第116页，济南，中国，2025年8月11日至14日。

吴宗友, 白昆龙, 杨林蕊, 王仪琦, and 田英杰. 2021. 电子病历文本挖掘研究综述. 计算机研究与发展, 58(03):513–527.

张云秋and 殷策. 2024. 基于大模型的中文电子病历实体自动识别研究. 数据分析与知识发现, pages 1–18.

张华丽, 康晓东, 李博, 王亚鸽, 刘汉卿, and 白放. 2020. 结合注意力机制的bi-lstm-crf 中文电子病历命名实体识别. 计算机应用, 40(S1):98–102.

朱岩, 张利, and 王煜. 2021. 基于roberta-wwm 的中文电子病历命名实体识别. 计算机与现代化, (02):51.

李博, 康晓东, 张华丽, 王亚鸽, 陈亚媛, and 白放. 2020. 采用transformer-crf 的中文电子病历命名实体识别. 计算机工程与应用, 56(5):153–159.

杜晋华, 尹浩, and 冯嵩. 2022. 中文电子病历命名实体识别的研究与进展. 电子学报, 50(12):3030–3053.

杨锦锋, 于秋滨, 关毅, and 蒋志鹏. 2014. 电子病历命名实体识别和实体关系抽取研究综述. 自动化学报, 40(08):1537–1562.

罗凌, 杨志豪, 宋雅文, 李楠, and 林鸿飞. 2020. 基于笔画elmo和多任务学习的中文电子病历命名实体识别研究. 计算机学报, 43(10):1943–1957.

许力and 李建华. 2021. 基于bert 和bilstm-crf 的生物医学命名实体识别. 计算机工程与科学, 43(10):1873.

6 附录

A.筛选器微调Prompt构建

图7和图8展示了用于训练实体筛选器的Prompt构建方法。通过对标准实体样本进行边界或类型扰动生成负样本，结合原始正样本形成训练集。

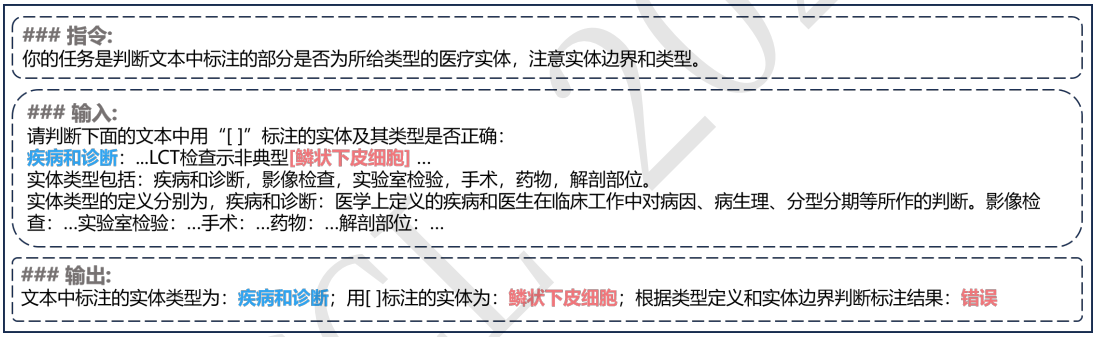

Figure 7: 筛选器中的负样本。蓝色为正确，红色为错误。

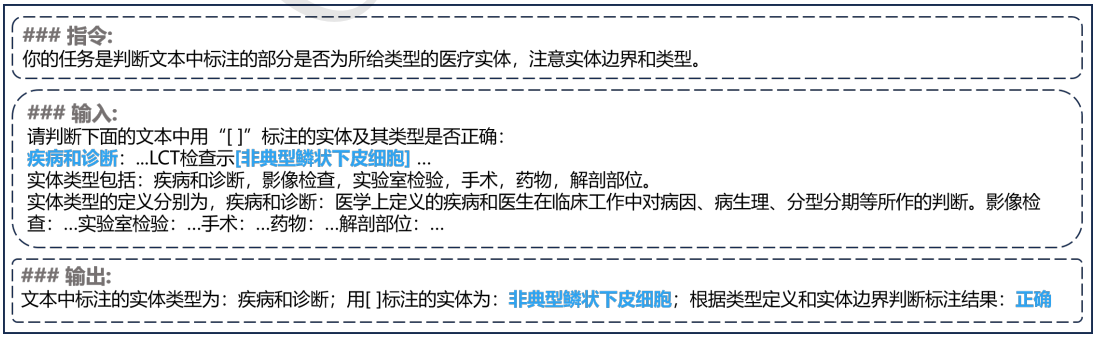

Figure 8: 筛选器中的正样本。蓝色为正确，红色为错误。

B.多模型集成Prompt构建

本文构建的多模型集成Prompt如图9，通过引导LLMs综合判断冲突实体的边界与类型信息，生成最终的识别结果。

指令:
你是一个专业的命名实体识别仲裁专家，需要综合多个模型的识别结果和原文语义，输出最优的实体识别结果。

输入:
原文：...09-26，全腹CT示：胃窦富血供肿瘤...
模型1：解剖部位：...09-26，全腹CT示：[胃窦]富血供肿瘤...
模型2：解剖部位：... 09-26，全腹CT示：胃窦富血供肿瘤...
模型3：解剖部位：... 09-26，全腹CT示：胃窦富血供肿瘤...
对于多模型预测结果冲突的实体，根据实体类型定义和样例决定是否正确。
实体类型的定义为，解剖部位：指疾病、症状和体征发生的人体解剖学部位。
相应类型的一些实体样例如下，可以帮你进一步鉴别模型的识别结果：解剖部位：胸椎、腹、心、肝、肠.....
注意按照原本模型输出方式输出最终结果。

输出:
解剖部位：...09-26，全腹CT示：[胃窦]富血供肿瘤...

Figure 9: 多模型集成Prompt

C.HTML和符号标注策略识别嵌套实体对比样例

图10通过样例对比了HTML与符号标注策略在嵌套实体识别中的表现，可以看出HTML标注结构相对复杂，不利于自回归模型有效学习与生成。

输入文本:
...见神经细胞空泡形成，透明变性和染色体分解，小神经胶质细胞浸润，血管周围单核细胞及浆细胞浸润...

HTML输出范式:
...见<身体:神经细胞><临床表现:神经细胞空泡>神经细胞</身体:神经细胞>空泡</临床表现:神经细胞空泡>形成，<临床表现:透明变性>透明变性</临床表现:透明变性>和<身体:染色体><临床表现:染色体分解>染色体</身体:染色体>分解</临床表现:染色体分解> <身体:小神经胶质细胞><临床表现:小神经胶质细胞浸润>小神经胶质细胞</身体:小神经胶质细胞>浸润</临床表现:小神经胶质细胞浸润>，<身体:血管周围单核细胞><临床表现:血管周围单核细胞及浆细胞浸润>血管周围单核细胞</身体:血管周围单核细胞>及<身体:浆细胞>浆细胞</身体:浆细胞>浸润</临床表现:血管周围单核细胞及浆细胞浸润>...

标注输出范式:
身体: ...见[神经细胞]空泡形成，透明变性和[染色体]分解[小神经胶质细胞]浸润，[血管周围单核细胞]及[浆细胞]浸润...
临床表现: ...见[神经细胞空泡]形成，[透明变性]和[染色体分解][小神经胶质细胞浸润]，[血管周围单核细胞及浆细胞浸润]...

Figure 10: HTML和符号标注策略识别嵌套实体对比

D.LLM-CLS在不同基座模型下的应用效果

本文进一步分析LLM-CLS在不同基座模型下的应用效果如表11所示，其中所有基线模型均基于符号标注策略进行微调。从中可以看出，引入LLM-CLS后Llama-3.1-8B-Instruct（Llama3.1）和Qwen2.5-7B-Instruct（Qwen2.5）的命名实体识别性能均实现了稳定提升，验证了本文所提出方法在不同基座模型下的通用性。

基座	方法	CMeEE-V2			CCKS2019		
		Precision	Recall	F1	Precision	Recall	F1
Llama3.1	基线	0.7212	**0.7631**	0.7416	0.8387	**0.8302**	0.8345
	+LLM-CLS	**0.7456**	0.7613	**0.7534**	**0.8625**	0.8265	**0.8441**
Qwen2.5	基线	0.7282	**0.7679**	0.7475	0.8468	**0.8344**	0.8406
	+LLM-CLS	**0.7536**	0.7662	**0.7598**	**0.8717**	0.8317	**0.8512**

Table 11: 筛选器对其他基座模型的应用效果。加粗表示当前基座本列最高值。

E.嵌套实体各类错误数量变化情况

针对嵌套实体识别效果，本文进一步对CMeEE-V2数据集中嵌套实体的错误类型数量分布进行了细化统计与评估。将GLM4-9B-Chat在符号标注策略微调后所得结果作为基线，与应用筛选和集成方法后的情况作对比。鉴于嵌套实体的结构特性，本文在错误分类中排除了“位置非重叠”这一错误类别（该类错误源于完全错误的冗余实体，与嵌套实体无关），并将嵌套实体的

错误类型划分为以下三类：(1)位置正确类型错误；(2)位置重叠类型正确；(3)位置重叠类型错误。嵌套实体各类型错误数量对比如表12，从表中结果可以看出，本文方法对嵌套实体各类错误数量全面降低，表明也能较为有效提升嵌套实体的识别。

	位置正确类型错误	位置重叠类型正确	位置重叠类型错误
基线	19	80	20
Ours	17(↓11%)	71(↓11%)	18(↓10%)

Table 12: 嵌套实体不同错误数量变化情况统计