DOI：10.61189/502047ilpumd

·产学研融合·

基于大语言模型的中文电子病历命名实体识别

程　洁，刘独玉*，陈思旭，钱姝妤
西南民族大学电气工程学院，四川 成都 610041

［摘要］ 命名实体识别（NER）作为自然语言处理（NLP）的核心任务，在电子病历（EMRs）中识别疾病、症状等医学实体，对临床辅助决策和医学知识库构建意义重大。但传统方法依赖大量标注数据与复杂模型，训练及推理成本较高。本文提出一种融合语义检索与提示学习的大语言模型生成式医学NER方法。首先，构建句子级向量数据库，对电子病历进行语义编码以实现可检索表示；然后，基于输入语句进行语义相似度检索，将相似示例动态注入提示模板，引导模型完成实体抽取；最后，通过结构化特殊标记生成实体类型标注结果，实现直接解码输出。实验表明，该方法在自建电子病历数据集和瑞金医院糖尿病数据集上均表现良好，尤其在低资源场景下具备较强的鲁棒性与迁移能力。

［关键词］ 命名实体识别；电子病历；大语言模型
［中图分类号］ TP　391　　［文献标志码］ A

Named entity recognition in chinese electronic medical records based on large language models

CHENG Jie, LIU Duyu*, CHEN Sixu, QIAN Shuyu
Southwest Minzu University, College of Electrical Engineering, Chengdu 610041, Sichuan, China

［**Abstract**］ Named Entity Recognition, as a core task in Natural Language Processing, plays a crucial role in identifying medical entities such as diseases and symptoms in Electronic Medical Records, which is of great significance for clinical decision support and the construction of medical knowledge bases. However, traditional methods rely heavily on large amounts of annotated data and complex models, resulting in high training and inference costs. This paper proposes a generative medical NER method that integrates semantic retrieval and prompt learning with large language models. First, a sentence-level vector database is constructed to semantically encode EMRs for retrievable representations. Then, based on the input sentence, semantic similarity retrieval is performed, and similar examples are dynamically injected into a prompt template to guide the model in entity extraction. Finally, entity type annotation results are generated through structured special markers, enabling direct decoding output. Experimental results demonstrate that the proposed method performs well on both a self-constructed EMR dataset and the Ruijin Hospital diabetes dataset, and exhibits strong robustness and transferability, especially in low-resource scenarios.

［**Key Words**］ named entity recognition; electronic medical records; large language models

NER作为NLP领域的一项基础任务，其主要目标是识别出文本中有特定含义的实体，并对这些实体进行标注[1]。它也是自然语言处理领域中许多研究问题的基础，如关系抽取[2]、事件抽取[3]、知识图谱[4]、机器翻译[5]、问答系统[6-7]等。电子病历作为医疗信息的核心载体，包含大量非结构化数据，蕴含着疾病、症状、检查、治疗等关键医学实体信息。从电子病历中高效、准确地抽取医学实体，不仅是深入分析医疗文本的基础，也是提升临床决策与治疗水平、实现智能医疗辅助的关键步骤。

当前，主流的医学NER方法多基于序列标注框架[8]，如BiLSTM-CRF、BERT-CRF等。这类方法通常依赖复杂的模型结构和大规模高质量标注语料，在处理医学文本时面临诸多挑战，如语言表达复杂、实体边界模糊、标签类别重叠等，使得模型泛化能力不足，迁移困难，且训练与部署成本较高。

近年来，大语言模型（large language models，LLMs）以其卓越的语言理解与生成能力，为医学NER任务提供了新的解决思路。通过合理设计提示（prompt），可以在无需微调参数的前提下引导模型完成实体识别，有效缓解标注资源依赖问题。然而，LLMs的性能高度依赖于提示的上下文质量。传统手工设计的提示模板往往通用性强而针对性弱，难以有效捕捉特定医学语境下的实体语义，从而影

［收稿日期］ 2025-08-14　　［接受日期］ 2025-09-16
［作者简介］ 程洁，硕士研究生. E-mail：19822909065@163.com
*通信作者（Corresponding author）. Tel:15828397145, E-mail: liuduyu10000@163.com

响识别效果的稳定性与准确性。

为此，本文提出一种结合语义检索与提示学习的上下文实体标记生成方法，旨在提升提示内容的相关性和大模型识别能力。该方法通过构建以句子级为单位的向量数据库，对已标注电子病历进行语义编码，实现可检索的语义表示；随后，利用语义检索机制从中动态选取与输入文本最相似的句子作为示例，并将其注入提示模板中，引导大语言模型进行实体抽取；最后，采用结构化标记格式输出实体及其类型，实现统一建模与直接解码。该方法将传统NER任务由“序列标注”转化为“文本生成”，不仅简化了建模流程，也更契合大语言模型的生成式能力。

为验证方法的有效性，本文在两个中文医学实体识别数据集上进行了系统实验，即自建数据集（包括糖尿病和小儿肺炎两个病种的电子病历）和瑞金医院糖尿病数据集。实验结果表明，本文方法在两个数据集上均取得了优异的性能，尤其在低资源条件下表现出显著优势，充分验证了结合语义检索与提示机制的大语言模型生成式方法在中文电子病历信息抽取中的可行性和潜力。

1 研究背景

命名实体概念提出之初，研究多基于规则与字典方法[9]，这类方法在特定场景中有效，但人工依赖高、泛化性差，难适应语言多样性，且在开放领域易出现漏检与误判。因此，引入了机器学习方法，如支持向量机（SVM）、隐马尔可夫模型（HMM）以及条件随机场（CRF）等。Zhou等[10]融合SVM与HMM，在GENETAG数据集上达到了0.83的F1值。Jonnalagadda等[11]与Jiang等[12]基于CRF在I2B2 2010数据集上分别取得了0.82和0.84的F1值。此类方法通过特征建模提升了识别效率，但仍高度依赖人工特征设计，难以捕捉深层语义。

为解决这些问题，深度学习方法应运而生，通过模型自动完成特征提取、训练和预测，常用模型包括卷积神经网络（CNN）、循环神经网络（RNN）、长短期记忆网络（LSTM）和Transformer等。Collobert等[13]提出基于CNN的多层神经网络结构，适用于包括NER在内的多种NLP任务。RNN适用于序列建模，但易出现梯度消失或爆炸。为此，LSTM基于梯度问题进行了改进。Huang等[14]结合BiLSTM和CRF，利用双向状态获取上下文信息。Transformer[15]基于多头自注意力机制，能更好地捕捉长距离依赖关系，是NLP研究中的重要突破，但其效果仍受限于训练语料的规模与质量。

基于迁移学习的思想，预训练语言模型（PLMs）的出现为这一问题提供了解决思路。Peters等[16]提出了ELMo，该模型利用BiLSTM获取上下文表示。谷歌提出的BERT[17]，通过堆叠Transformer与MLM训练构建双向语言模型，被广泛应用并衍生出了许多变体模型。例如，HU等[18]提出BERT-Attention-BiLSTM用于中医短文本匹配。郭等[19]提出RoBERTa-wwm-ext-BiGRU-EGP模型，能够有效提升医学实体解析与文本分析能力。

随着PLMs规模的不断扩大，微调成本急剧上升。提示学习（prompt learning）[20]通过调整下游任务的结构与输入，使其更契合预训练语言模型，从而提升模型参数的利用率与任务性能。在此背景下，GPT-3、ChatGPT、ChatGLM以及Qwen等具备强大语言理解与生成能力的大语言模型（LLMs）相继涌现。基于LLMs的提示学习范式，逐步成为应对下游任务特别是NER任务的一种新趋势。Schick等[21]提出PET，将任务转化为完形填空模板并利用软标签提升低资源条件下的模型性能；GPT-3[22]通过通过任务描述实现少样本高性能；Cui等[23]利用BART构建模板排序促进知识迁移；Ma等[24]提出Template-free Prompt直接预测类别词；Lee等[25]则通过Demonstration-based Learning提升低资源下的模型表现；Shen等[26]统一实体定位与分类并采用动态图匹配提升标签质量。

基于生成式LLMs的NER方法将传统序列标注任务转化为结构化文本生成任务，即以输入文本与提示为条件，生成带有实体标记或结构化信息的输出。该方法在少样本、嵌套实体和跨领域迁移中具备良好的泛化与语义理解能力，尤其适用于中文医学电子病历的复杂实体与语义歧义处理。

2 研究方法

本文针对医学命名实体识别任务，提出了一种结合语义检索与提示学习的大语言模型生成式识别方法。该方法整体流程主要包括以下三个部分。

①实体向量数据库构建：将电子病历中的句子作为基本单元，利用预训练模型进行语义编码，构建可供检索的文本向量集合；②基于语义检索的提示生成：根据输入句子从向量数据库中检索相似样本，并将其作为示例动态注入提示模板中，辅助模型完成实体抽取；③实体标记生成与解码：大语言模型生成嵌入结构化特殊标记的输出句子，直接输出带实体类别的识别结果，便于解析与评估。整体

架构如图 1 所示。

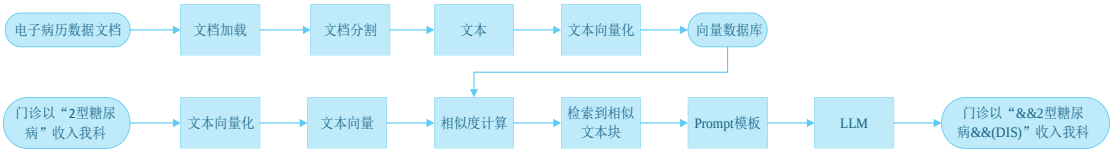

图 1　基于 LLM 的 NER 流程图

2.1　*实体向量数据库构建*　为了为提示增强的生成式医学命名实体识别模型提供可检索的、结构化的医学领域知识支撑，本文设计并构建了一个面向医学电子病历的向量数据库。该向量数据库以电子病历中的句子级文本为基础，通过预训练医学嵌入模型获取其稠密语义表示，为后续提示生成阶段提供高质量、可追溯的示例支撑。向量数据库的构建过程包括以下关键步骤。

2.1.1　*文档加载与分割*　首先加载经过预处理和人工标注后的中文电子病历数据集。原始数据以“content”字段存储完整病历文本。为提高检索粒度与表示一致性，本文采用基于中文句号“。”的切分策略，将每条“content”文本划分为若干语义完整、粒度一致的句子级片段，并以“text”字段形式保存。例如：“text”：“糖尿病史 10 余年，子宫癌术后 8 年，心脏支架术后，无传染病史。”；“text”：“临床表现为反复发作的喘息、胸闷、咳嗽等症状。”该操作确保了每条文本具有明确边界、语义集中，适用于句子级向量建模与相似度检索。

2.1.2　*语义向量编码*　bge-large-zh-v1.5[27] 是由 BAAI（Beijing Academy of Artificial Intelligence）开发的一款基于 Transformer 架构的高性能中文文本嵌入模型，如图 2。该模型采用 1024 维的嵌入空间，能够对句子级文本进行高质量的语义编码，适用于相似度计算与语义检索等任务。其核心技术包括指令微调（instruction fine-tuning）和硬负样本挖掘（hard negative mining），显著提升了向量分布的判别能力与检索效果。同时，bge-large-zh-v1.5 支持最长 8 192 个 token 的长文本输入，兼顾语义精度与处理效率。在 C-MTEB 等中文基准评测中，该模型表现优异，展现出在多语言任务中的广泛适应性与应用潜力。

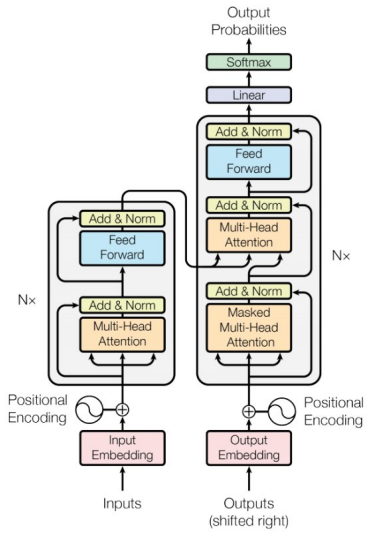

图 2　Transformer 架构图[15]

bge-large-zh-v1.5 的文本表示可形式化包含输入嵌入与位置编码、Transformer 编码和向量归一化三个过程。①输入嵌入与位置编码（embedding + positional encoding）：假设语料库中总共包含 N 个句子，对于其中任一句子 x_j（$j = 1, 2, \cdots, N$）可表示为 $x_j = w_1^j w_2^j \cdots w_n^j$，首先通过嵌入层获得词级输入：

$$E^j = [e_1^j; e_2^j; \cdots; e_n^j] + [p_1^j; p_2^j; \cdots; p_n^j] \quad (2\text{-}1)$$

其中，$e_i^j \in \mathbb{R}^d$ 为 w_i^j 的嵌入向量，$p_i^j \in \mathbb{R}^d$ 为 w_i^j 的位置编码，$d = 1024$ 表示嵌入维度，n 为输入序列的 token 总数，E^j 为 w_i^j 的输入嵌入矩阵，维度为 $\mathbb{R}^{n \times d}$。

②Transformer编码：经过多层Transformer编码器进行上下文建模。

$$H_{(0)}^{j} = E^{j} \tag{2-2}$$

其中，$H_{(0)}^{j}$为初始隐状态，维度为$\mathbb{R}^{n\times d}$。

$$H_{(l)}^{j} = \text{TransformerLayer}^{(l)}(H_{(l-1)}^{j}) \tag{2-3}$$

其中，$l = 1, 2, \cdots, L$，L为Transformer编码器的总层数。

每层包含多头注意力（multi-head attention）和前馈网络（FFN），输出最后一层的隐状态表示为：

$$H_{(L)}^{j} = [h_{[\text{CLS}]}^{j}; h_2^{j}; \cdots; h_n^{j}],\ h_i^{j} \in \mathbb{R}^{d} \tag{2-4}$$

其中，$h_{[\text{CLS}]}^{j}$表示Transformer编码器输出的[CLS] token对应的向量表示。

③向量归一化（CLS token + L2 Normalization）：

$$\vec{x}_j = \frac{h_{[\text{CLS}]}^{(j)}}{\| h_{[\text{CLS}]}^{(j)} \|} \tag{2-5}$$

其中，$\| \cdot \|$表示向量的L2范数。

（3）*向量集合构建* 将语料库中的所有句子完成编码并归一化后，构建向量集合（即向量数据库）$\mathcal{X}$：

$$\mathcal{X} = \{\vec{x}_1, \vec{x}_2, \cdots, \vec{x}_N\} \tag{2-6}$$

该集合仅包含句子级文本向量，不包含实体标注信息，保证了数据库结构的简洁性与通用性。同时，通过句子索引的方式，可在提示生成阶段快速映射与追溯原始样本。

2.2 *基于语义检索的提示生成* 在完成向量数据库的基础上，本文进一步设计了基于语义检索的提示生成机制，旨在充分利用向量数据库中蕴含的医学知识示例，动态增强大语言模型的生成式命名实体识别能力。该模块的核心流程如下。

（1）输入句子编码与相似度计算：本文采用Top_K相似度检索策略，旨在为每一个输入句子构造高质量的语义提示。具体而言，使用嵌入模型bge-large-zh-v1.5（同2.1）将待识别的输入句子q转化为稠密向量表示$\vec{q}$，并与向量集合$\mathcal{X}$中的所有向量进行余弦相似度计算，以检索与其最接近的Top_K个向量所对应句子作为提示示例。

余弦相似度计算：

$$sim(\vec{q}, \vec{x}_j) = \frac{\vec{q} \cdot \vec{x}_j}{\| \vec{q} \| \cdot \| \vec{x}_j \|} \tag{2-7}$$

其中，$\cdot$表示向量点积。显然，$sim(\vec{q}, \vec{x}_j)$的取值范围为[-1, 1]，值越接近1表示两个句子的语义越接近。

$$\text{Top\_K}(q) = \underset{x_j \in \mathcal{X}}{\arg\text{topK}}\ sim(\vec{q}, \vec{x}_j) \tag{2-8}$$

其中，arg topK表示从候选向量集合中，返回与$\vec{q}$相似度最高的前K个向量对应的句子，即Top_K个句子，作为最相关的提示示例。

（2）示例标注重构：检索后，将原始“text”中出现的实体按照特殊标记格式“&&实体&&（实体类型）”替换，从而得到符合提示规范的“输入-输出”示例。所有Top_K示例会拼接并注入到提示模板中的{RAG_text}占位符处，可为大语言模型提供具备相似语义背景和准确标注结构的示例。

（3）提示模板动态填充：设计提示模板如表1所示。该提示模板涵盖了实体抽取规则、实体类型定义、输出格式约束等内容，并支持动态插入检索示例和当前待识别句子{query}，以构建完整的模型输入。

（4）提示引导的生成推理：经过填充后的完整提示将作为输入送入大语言模型，引导其识别当前句子中的医学实体。由于提示中嵌入了结构化规则和真实示例，模型可直接参考标注风格进行仿写，进而提高抽取结果的准确率、实体边界判断的清晰性及类型分辨的一致性。

本节所设计的提示生成机制融合了向量检索与提示模板的优点，具备可扩展、可控、可追溯的特点，为生成式实体识别任务提供了更可靠的输入结构。

2.3 *实体标记生成与解码* 本文采用显式可视的特殊标记方法，即将识别出的实体及其对应类型以“&&实体&&（实体类型）”的格式嵌入原句中，作为模型的直接输出结果，如图3所示。相较于传统BIO（begin-inside-outside）序列标注方式，该标记方法具有明显的优势：BIO标注需要对每个分词位置进行标签预测，解码时再通过后处理算法组合实体，容易导致边界模糊、重叠错误或标签不一致等问题；而本文中的特殊标记方法则能在原文中直接呈现实体及其实体类型，可立即校验，无需复杂的后处理，特殊标记不仅机器可解析，如正则表达式即可抽取，同时也方便人工快速核对，有助于医学专家对生成结果进行验证，提高整体的可追溯性与可解释性。

在解码阶段，大语言模型会根据提示注入的信息以及语义检索返回的相似示例，逐步推理并输出结构化标记的句子。系统后续可通过简单的正则匹配或规则算法，快速解析出实体及其类别。

3 研究实施

本实验的环境配置如表2所示，大语言模型框架是LangChain，BGE（BAAI general embedding）模型是bge-large-zh-v1.5，Transformers版本为4.52.3。

表 1　提示模板设计

你需要对用户输入的句子进行医疗实体抽取。实体抽取的规则有： 1. 实体必须出现在句子中； 2. 实体必须与实体类型相关； 3. 实体和实体之间不能重叠。 抽取的实体类型有： 1. 疾病(DIS)：患者所患的具体疾病或诊断名称。 2. 症状(SYM)：感受到了或表现出的一种令人不快的感觉或表现出的现象，可能是某种潜在疾病或病状的症状。 3. 体征(SIG)：包括体格检查和可观察到的临床异常。 …… 9. 时间(TIM)：时间以及持续时间。包括症状、疾病的具体时间以及持续时间。 10. 治疗手段(PAH)：为了解决疾病或者缓解症状而施加给患者的治疗程序，干预措施。包括药物治疗、手术治疗、医学营养治疗等。 11. 身体部位(BOP)：身体部位或器官。
根据用户输入的句子，识别对应的实体和实体类型。在句子中出现的实体，用"&&实体&&(实体类型)"表示，如：&&支气管肺炎&&(DIS)。 以下是输出示例，不能输出其他对话以及对实体识别任务过程的解释： {RAG_text} 用户输入： {query}

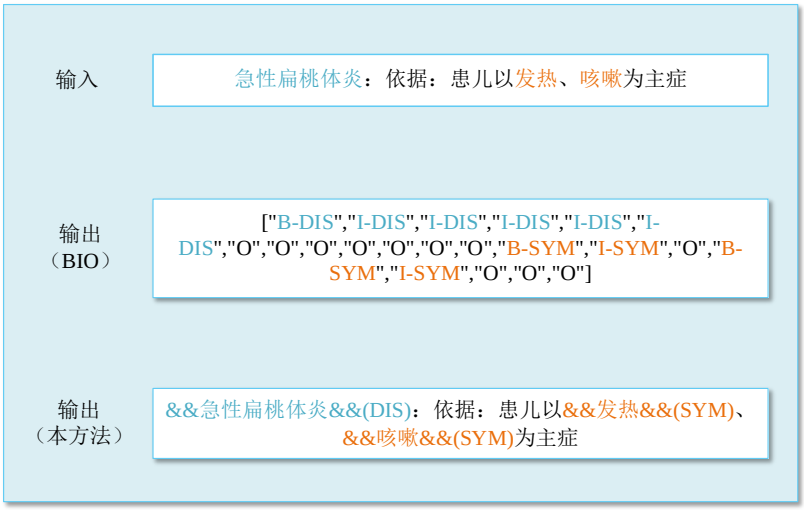

图 3　输出表示

Figure 3　Output representation

表 2　本研究环境配置表

类别	型号
操作系统	Ubuntu 22.04.5 LTS
CPU	Intel(R) Xeon(R) Gold 6330 CPU @ 2.00GHz
显卡	NVIDIA GeForce RTX 4090
CUDA	12.4
显存	24GB
内存	134GB
Python	3.9.18

3.1　*数据集*　本文所使用的医学领域数据集包含自建电子病历数据集和瑞金医院糖尿病数据集。自建电子病历数据集包含四川省某市多家三甲医院真实的 600 份糖尿病电子病历数据和 600 份小儿肺炎电子病历数据。在医学专家的指导下，本文作者所在实验室的两位研究生 A 和 B 采用双盲多轮标注方法标注并构建了自建语料库。

根据病历特点，定义了 11 类医疗实体，实体类

型及其含义如表3所示。将自建数据集拆分为训练集和测试集，其中训练集有748 420个字符，实体数量有107 624个，测试集有93 672个字符，实体数量有13 554个。训练集用于构建实体向量数据库，以支持语义检索与示例选取；测试集则作为实验评估集，用于检验所提出方法在命名实体识别任务中的性能表现。

表3 自建数据集（糖尿病+小儿肺炎）

实体类型	含义	示例
疾病(DIS)	患者所患的具体疾病或诊断名称	糖尿病、小儿肺炎
症状(SYM)	患者自我感受到或主诉的主观不适	头晕、恶心
体征(SIG)	医生客观检查或仪器检测到的体征	心音低钝、呼吸音粗
身体部位(BOP)	涉及人体结构的名称	双下肢、腹部
时间(TIM)	具体时间及持续时间	1年、3天前
程度(DEG)	病情或症状的严重程度	剧烈、轻度
治疗手段(PAH)	施加给患者的治疗程序、干预措施	手术、二级护理
药物(DRU)	具体的药品名称	布洛芬、二甲双胍
药物修饰词(MDM)	用药的频率、剂量以及投药方式	一日两次、口服
医学检查(AUE)	实验室检查和辅助检查	CT、血常规
医学检查结果(RES)	通过医学检查得到的数值或结果	阳性、8.2mmol/L

瑞金医院糖尿病数据集来源于中文糖尿病研究领域的权威杂志，共包含363篇文档，约250万字，医疗实体划分为15类，实体类型和实体含义如表4。将数据集拆分为训练集和测试集，其中训练集有662 106个字符，实体数量有48 714个，测试集有257 258个字符，实体数量有19 069个。该数据集的训练集与测试集的作用同自建数据集一样。

表4 瑞金医院糖尿病数据集

实体类型	含义	示例
Disease(DIS)	患者所患的具体疾病或诊断名称	糖尿病、高血压
Reason(REA)	引发疾病或症状的原因、诱因	营养状态差
Symptom(SYM)	临床表现，包括症状和体征	嗜睡、关节痛
Test(TES)	各类医学检查、实验室检查	空腹血糖、TG
Test_Value(TEV)	检查项目对应的数值、结果	23.7kg/m^2、升高
Drug(DRU)	明确的药物名称	双膦酸盐
Frequency(FRE)	药物使用频率	每周1次、餐前
Amount(AMO)	用药剂量	2～5 g、20mg/d
Method(MET)	用药方式	口服、静脉
Treament(TRA)	非药物的治疗方式	降糖治疗、PEI
Operation(OPE)	手术、介入行为	肠道手术
SideEff(SIE)	用药引起的不良反应或副作用	低血糖风险
Anatomy(ANA)	身体或器官部位名称	骨髓干细胞
Level(LEV)	描述病情、症状或指标程度的词语	急性、4期
Duration(DUR)	症状、疾病或用药的持续时间	4周、<10年

3.2 评价指标 采用精确率P、召回率R以及$F1$值这3个指标评估本文方法的性能。

精确率P：模型预测为正例（即正确识别出的实体）中，实际真正正确的比例。

$$P = \frac{TP}{TP + FP} \tag{3-1}$$

召回率R：数据中所有实际的正例（即真实实体）中，被模型正确识别出的比例。

$$R = \frac{TP}{TP + FN} \tag{3-2}$$

其中TP为真正例，表示被正确识别的实体数量，FP为假正例，表示错误识别为实体的数量。

其中FN为假负例，表示模型未能识别出的真实实体数量。

$F1$值：精确率P和召回率R的调和平均。

$$F1 = \frac{2 \times P \times R}{P + R} \quad (3\text{-}3)$$

3.3　实验结果与分析　为验证本文所提出的结合语义检索与提示学习的大语言模型生成式命名实体识别方法的有效性，在自建小儿肺炎与糖尿病混合病种的中文电子病历数据集、瑞金医院糖尿病数据集这两个数据集上进行了实验。比较模型涵盖传统序列标注模型BERT+BiLSTM+CRF、采用API调用的大语言模型GLM-4-Flash（表5简写为GLM-4）和本地部署的开源模型Qwen2-7B-Instruct（表5简写为Qwen2）的多种提示设置（0-shot、few-shot），以及本文提出的基于语义增强的提示生成方法（RAG+few-shot）。实验结果见表5。

表5　实验结果

%

模型	自建数据集			瑞金医院数据集		
	Precision	Recall	F1	Precision	Recall	F1
BERT+BiLSTM+CRF	93.00	92.00	**93.00**	46.00	42.00	**44.00**
GLM-4+0-shot	16.03	6.35	9.10	14.94	9.12	11.33
GLM-4+5-shot	44.22	51.04	47.39	49.30	40.02	44.17
GLM-4+10-shot	48.79	55.63	**51.98**	51.73	40.99	**45.74**
RAG+GLM-4+1-shot	81.98	80.90	81.44	48.22	42.07	44.93
RAG+GLM-4+3-shot	88.27	89.38	88.82	52.62	49.59	51.06
RAG+GLM-4+5-shot	89.86	91.00	**90.42**	55.51	52.43	**53.93**
RAG+Qwen2 +1-shot	84.71	60.25	70.42	28.89	28.58	28.74
RAG+Qwen2 +3-shot	94.61	92.44	93.51	42.70	40.62	41.63
RAG+Qwen2 +5-shot	94.65	94.59	**94.62**	47.37	40.91	**43.90**

从实验结果可见，在自建电子病历数据集上，传统结构化模型BERT+BiLSTM+CRF实现了93.00%的F1值，表现出较强的基线性能。而基于生成式大语言模型的方案在引入语义检索与few-shot提示后进一步提升了识别效果，尤其是RAG+Qwen2-7B-Instruct+5-shot模型达到了94.62%的F1值，全面超越传统方法。这一结果表明：语义检索机制能够有效选取与目标文本语义相关的提示示例，在提升实体识别能力的同时，提高了大模型的泛化性和稳健性。进一步从few-shot配置来看，生成式模型对提示示例极为敏感。例如，GLM-4模型在0-shot下的F1值仅为9.10%，而随着示例数量增加至10-shot，F1值提升至51.98%。在引入语义检索机制后，性能提升更加显著：RAG+GLM-4在5-shot配置下即达90.42%，Qwen2系列即使仅使用3个示例也能实现93.51%的F1值，在5-shot时达到峰值94.62%。

在瑞金医院糖尿病数据集上，任务复杂度更高。该数据集中包含大量医学英文缩写、嵌套结构等复杂实体类型，导致实体识别任务面临更大挑战。传统结构化模型BERT+BiLSTM+CRF在该数据集上取得了44.00%的F1值。相比之下，GLM-4在无语义增强的5-shot设置下仅达44.17%，而引入语义增强策略后，RAG+GLM-4+5-shot明显提升至53.93%，性能显著提升，验证了语义检索机制在复杂医疗语境中的适用性和有效性。Qwen2系列在该数据集上的表现也有一定改善，RAG+Qwen2-7B-Instruct+5-shot模型实现了43.90%的F1值，达到了和传统模型差不多的效果。

综上所述，实验结果充分验证了本文方法在不同类型医学文本中的广泛适应性与有效性。尤其值得强调的是，自建混合病种数据集在实体类型覆盖全面、语义表达丰富且标注规范等方面具有显著优势，为模型提供了高质量的语义支持和训练基础。正是基于这一高质量标注，生成式大语言模型在该数据集上得以充分学习，在5-shot设置下达到了94.62%的F1值，优于传统结构化方法，表现出更强的识别能力和泛化性能。相比之下，瑞金医院糖尿病数据集由于实体嵌套复杂、缩写频繁等特性，整体识别难度更高，F1值普遍偏低。然而，随着提示示例数量增加，模型表现持续上升，说明其具备良好的学习能力。而在引入语义检索机制后，性能提升尤为明显。这一结果从侧面进一步印证了高质量的标注语料对于医学NER任务至关重要，而自

建数据集的标注质量正是支撑模型取得优异性能的核心保障。

同时,在few-shot低资源配置下,模型也展现了出色的少样本学习能力,即便在示例数量极少的情况下,也能依托语义提示完成精准的实体识别,充分体现了语义增强策略的有效性。因此,本文提出的方法不仅适用于结构复杂、表达多样的多病种电子病历场景,也为后续更大规模的医疗信息抽取任务提供了可扩展、可迁移的解决思路。

4 结 论

本文面向中文电子病历命名实体识别任务,提出了一种结合语义检索机制与提示学习的生成式大语言模型识别方法,以提升模型在低资源、多病种、真实医疗文本环境下的实体识别能力。本文方法通过构建实体向量数据库,从训练集中检索与输入语义相似的句子作为few-shot提示,引导大语言模型在无需参数微调的前提下完成实体识别任务。相比传统依赖大规模标注数据的序列标注方法,本文方法不仅有效缓解了标注资源匮乏的问题,同时充分发挥了生成式语言模型在语义建模与上下文理解方面的优势,适用于实体类型复杂的医学文本。

尽管如此,本研究仍存在一定局限性。一方面,生成式识别方法缺乏位置约束机制,在实体密集或边界模糊的场景中易产生幻觉式输出;另一方面,大语言模型推理资源开销较大,尚难以直接部署于资源受限的实际临床环境。此外,当前模型主要聚焦于实体层级,对于更高阶的结构化信息如事件、因果关系的抽取能力仍有待提升。

未来工作可从以下几个方面展开探索:(1)引入结构控制机制或后验验证模块,以提升输出实体的准确性与一致性;(2)借助参数高效微调方法,推动模型的轻量化与边缘部署;(3)拓展至医学事件、时间表达、因果关系等复杂信息的识别任务,进一步增强模型在医疗信息抽取中的应用广度。

伦理声明 无。

利益冲突 所有作者声明不存在利益冲突。

作者贡献 程洁:论文选题、撰稿;刘独玉:论文修改;陈思旭:论文修改;钱姝好:共同标注数据。

参考文献

[1] 祁鹏年, 廖雨伦, 覃 飙. 基于深度学习的中文命名实体识别研究综述[J]. 小型微型计算机系统, 2023, 44(9): 1857-1868.

[2] YIN L B, MENG X, LI J X, et al. Relation extraction for massive news texts[J]. Comput Mater Continua, 2019, 60(1): 275-285.

[3] MA X, LU Y, LU Y, et al. Biomedical event extraction using a new error detection learning approach based on neural network [J]. Computers, Materials & Continua, 2020, 63(2): 923-941.

[4] ZHOU H, SHEN T, LIU X, et al. Survey of knowledge graph approaches and applications [J]. Journal on Artificial Intelligence, 2020, 2(2): 89-101.

[5] QIU J, LIU Y, CHAI Y H, et al. Dependency-based local attention approach to neural machine translation [J]. Comput Mater Continua, 2019, 59(2): 547-562.

[6] SHARMA Y, GUPTA S. Deep learning approaches for question answering system [J]. Procedia Comput Sci, 2018, 132: 785-794.

[7] DOU Z Y, WANG X, SHI S M, et al. Exploiting deep representations for natural language processing [J]. Neurocomputing, 2020, 386: 1-7.

[8] LEE J, YOON W, KIM S, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining[J]. Bioinformatics, 2020, 36(4): 1234-1240.

[9] 隋 臣. 基于深度学习的中文命名实体识别研究[D]. 杭州: 浙江大学, 2018.

[10] ZHOU G D, SHEN D, ZHANG J, et al. Recognition of protein/gene names from text using an ensemble of classifiers[J]. BMC Bioinformatics, 2005, 6(Suppl 1): S7.

[11] JONNALAGADDA S, COHEN T, WU S, et al. Enhancing clinical concept extraction with distributional semantics[J]. J Biomed Inform, 2012, 45(1): 129-140.

[12] JIANG M, CHEN Y K, LIU M, et al. A study of machine-learning-based approaches to extract clinical entities and their assertions from discharge summaries [J]. J Am Med Inform Assoc, 2011, 18(5): 601-606.

[13] COLLOBERT R, WESTON J, BOTTOU L, et al. Natural language processing (almost) from scratch [J]. J. Mach. Learn. Res., 2011, 12: 2493-2537.

[14] HUANG Z, XU W, YU K. Bidirectional LSTM-CRF models for sequence tagging[J]. Arxiv, 2015.

[15] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[A]. arXiv, 2023.

[16] PETERS M, NEUMANN M, IYYER M, et al. Deep contextualized word representations [C]//Proceedings of the 2018 Conference of the North American Chapter Ofthe Association for Computational Linguistics: Human LanguageTechnologies, Volume 1 (Long Papers). New Orleans, Louisiana. Stroudsburg, PA, USAACL, 2018: 2227-2237.

[17] DEVLIN J, CHANG M W, LEE K, et al. BERT: pre-training of deep bidirectional transformers for language understanding [A]. arXiv, 2019.

[18] HU X S, ZHANG H J, SUN Y J. Chinese medical short text

matching model based on fine-tuning BERT-attention-BiLSTM [C]//2023 IEEE/ACIS 23rd International Conference on Computer and Information Science (ICIS). June 23-25, 2023, Wuxi, China. IEEE, 2023: 91-96.

[19] 郭振华，宋　波. 基于RBBEGP的中文电子病历命名实体识别研究[J]. 电脑知识与技术，2024，20(16)：6-10.

[20] LIU P F, YUAN W Z, FU J L, et al. Pre-train, prompt, and predict: a systematic survey of prompting methods in natural language processing[J]. ACM Comput Surv, 2023, 55(9): 1-35.

[21] SCHICK T, SCHÜTZE H. Exploiting cloze-questions for few-shot text classification and natural language inference [C]// Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume. Online. Stroudsburg, PA, USAACL, 2021: 255-269.

[22] BROWN T B, MANN B, RYDER N, et al. Language models are few-shot learners[C] //Proceedings of the 34th International Conference on Neural Information Processing Systems. Red Hook, NY, USA: Curran Associates Inc. , 2020: 1877-1901.

[23] CUI L Y, WU Y, LIU J, et al. Template-based named entity recognition using BART [C]//Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021. Online. Stroudsburg, PA, USAACL, 2021: 1835-1845.

[24] MA R T, ZHOU X, GUI T, et al. Template-free prompt tuning for few-shot NER [C]//Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Seattle, USA. Stroudsburg, PA, USAACL, 2022: 5721-5732.

[25] LEE D H, KADAKIA A, TAN K M, et al. Good examples make A faster learner: simple demonstration-based learning for low-resource NER[C]//Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Dublin, Ireland. Stroudsburg, PA, USAACL, 2022: 2687-2700.

[26] SHEN Y L, TAN Z Q, WU S H, et al. PromptNER: prompt locating and typing for named entity recognition [C]// Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Toronto, Canada. Stroudsburg, PA, USAACL, 2023: 12492-12507.

[27] XIAO S T, LIU Z, ZHANG P T, et al. C-pack: packed resources for general Chinese embeddings [C]//Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval. Washington DC USA. ACM, 2024: 641-649.

引用本文

程　洁，刘独玉，陈思旭，等. 基于大语言模型的中文电子病历命名实体识别[J]. 元宇宙医学，2025，2(3)：39-47.

CHENG J，LIU D Y，CHEN S X，et al. Named entity recognition in chinese electronic medical records based on large language models[J]. Metaverse Med，2025，2(3)：39-47.