山西大学学报(自然科学版)49(4):547—556,2026
Journal of Shanxi University(Nat. Sci. Ed.)

DOI:10.13451/j.sxu.ns.2025118

大模型的药物不良反应实体识别能力评估

王子天,徐康*,董振江
(南京邮电大学 计算机学院,软件学院,网络空间安全学院,江苏 南京 210023)

摘　要:本研究针对药物不良反应领域的命名实体识别任务,构建了用于药物警戒领域的中文命名实体识别数据集(Chinese Named Entity Recognition Pharmacovigilance Dataset, CNER-PH),并在此基础上对主流命名实体识别(Named Entity Recognition,NER)模型与大语言模型进行了系统评估。方法上,采用多种预训练模型基于Transformer的双向编码器表示(Bidirectional Encoder Representations from Transformers, BERT)、鲁棒优化的BERT预训练方法(Robustly Optimized BERT Pretraining Approach, RoBERTa)与大语言模型(GPT-4o、DeepSeek等),重点分析其在实体类别多样化及边界模糊场景下的识别能力差异。实验结果表明,BERT系列模型在应对复杂实体边界时表现更为稳健,*F*1值整体高于大语言模型。研究结果揭示了不同模型在药物不良反应识别中的特点与不足,为后续方法改进提供了依据。

关键词:大语言模型;药物警戒;命名实体识别;数据集构建
中图分类号:TP391　　**文献标志码**:A　　**文章编号**:0253-2395(2026)04-0547-10

Evaluation of Large Language Models for Adverse Drug Reaction Entity Recognition

Wang Zitian, Xu Kang*, Dong Zhenjiang
(*School of Computer Science, Nanjing University of Posts and Telecommunications, Nanjing* 210023*, China*)

Abstract: This study focused on the named entity recognition task in the field of adverse drug reactions and constructed a Chinese dataset named Chinese named entity recognition pharmacovigilance dataset (CNER-PH). Based on this dataset, mainstream named entity recognition (NER) models and large language models were systematically evaluated. The methods used included several pretrained models such as bidirectional encoder representations from transformers (BERT) series and robustly optimized BERT pretraining approach (RoBERTa), as well as large language models such as GPT-4o and DeepSeek. The analysis emphasized their recognition ability under conditions of diverse entity types and unclear entity boundaries. The experimental results show that BERT series models are more stable in handling complex entity boundaries, and their overall *F*1 scores are higher than those of large language models. The findings reveal the strengths and weaknesses of different models in adverse drug reaction recognition and provide a basis for future methodological improvements.

Key words: large language models; pharmacovigilance; named entity recognition; dataset construction

收稿日期:2025-09-27;**修回日期**:2025-11-13
基金项目:国家自然科学基金(62202240;61872190;62276142);江苏省社会发展项目(BE2023025);国家社会科学基金重点项目(23ATQ009);江苏省社会科学基金(23TQB007);江苏省高校哲学社会科学研究重大项目(2020SJZDA102)
作者简介:王子天(2000—),女,江苏宜兴人,硕士研究生,研究方向为大语言模型应用。E-mail:1023040922@njupt.edu.cn
* **通信作者**:徐康(Xu Kang),E-mail:kxu@njupt.edu.cn
引文格式:王子天,徐康,董振江.大模型的药物不良反应实体识别能力评估[J].山西大学学报(自然科学版),2026,49(4):547-556. DOI:10.13451/j.sxu.ns.2025118.

0 引言

药物不良反应是指由药物相关干预引起的明显有害或不愉快的反应[1]。尽管药物在上市前会经过大量的临床试验,但患者的个体差异,如性别、年龄、怀孕状况、健康状况和种族,仍可能影响药物的治疗效果[2]。因此,在药物使用过程中,可能会出现一些在临床试验中未曾观察到的新反应,这就需要持续监测并及时更新药物不良反应记录。药物不良反应的发生率相对较高,严重时可能导致死亡,所以能够有效识别药物不良反应并采取适当的预防措施至关重要。命名实体识别(Name Entity Recognition,NER)是自然语言处理(Natural language Process,NLP)应用中的基础任务,旨在识别自然语言文本中具有特定含义的实体。在提取药物不良反应信息的任务中,NER广泛应用于识别与不良反应相关的重要实体,通过将NER应用于大量的医学文献、患者报告和社交媒体内容,可以有效地捕捉潜在的不良反应信息,为药物安全监测提供数据支持。

早期的药物安全监测相关数据集研究中,LePendu等[3]将电子健康记录中的非结构化临床文本转化为去标识化的患者特征矩阵,用于识别药物不良反应和药物相互作用风险。Demner-Fushman等[4]与美国食品药品监督管理局(The United States Food and Drug Administration,FDA)合作,构建了结构化产品标签药物不良反应(The Structured Product Labels Adverse Drug Reactions, SPL-ADR-200db)数据库,包含了200种FDA批准药物的标准化不良反应信息,用于支持不良反应提取工具的开发和评估。Sun等[5]引入了药物安全监测数据集药物警戒事件抽取(Pharmacovigilance Event Extraction, PHEE),包含了5 000多个标注事件,用于分析患者的人口统计信息、治疗和副作用,并评估了先进生物医学事件提取方法的有效性和局限性。Sboev等[6]开发了一个俄语药品评论语料库,支持复杂的命名实体识别和共指解析任务。Raithel等[7]创建了第一个用于检测德国患者生成内容中的药物不良反应的数据集,包含4 169篇二进制标记文档。D'Oosterlinck等[8]开发了生物医学药物不良事件抽取(Biomedical Adverse Drug Event Extraction,BioDEX),这是一个大规模的不良药物事件提取数据集,旨在通过自然语言处理改进药物安全监测。Raithel等[9]构建了一个多语言的药物不良反应(Adverse Drug Reaction, ADR)语料库,涵盖了德语、法语和日语中的实体和关系标注,支持多语言医学语言模型的发展。

生物医学命名实体识别方面,Sarker和Gonzalez[10]采用朴素贝叶斯、支持向量机和最大熵等方法对药物不良反应进行二元分类,通过多语料库训练,他们有效地提高了分类准确率。自2020年以来,基于Transformer的双向编码器表示(Bidirectional Encoder Representations from Transformers,BERT)模型凭借其深度捕捉上下文信息的双向编码能力,显著提升了文本分类[11]、命名实体识别[12]和问答系统[13]等任务的性能。Fan等[14]使用BERT模型从WebMD和Drugs.com的评论中检测和提取药物不良事件(Adverse Drug Events, ADEs),实现了0.94的ADE(Adverse Drug Event,ADE)检测曲线下面积(Area Under the Curve, AUC)和0.97的提取$F1$分数。McMaster等[15]利用预训练的具有解耦注意力的解码增强型BERT(Decoding-enhanced BERT with Disentangled Attention,DeBERTa)、FDA不良事件报告系统(FDA Adverse Event Reporting System,FAERS)模型和微调技术,高效地从出院总结中识别药物不良反应。2023年大语言模型的崛起引领了自然语言处理领域的重大突破,以ChatGPT为代表的模型展示出强大的语言理解能力[16-18]。Juhi等[19]通过对40对药物进行相互作用测试,探索了ChatGPT在预测和解释药物相互作用方面的有效性。

随着药物警戒体系的不断完善,FAERS和EudraVigilance等系统在ADR的监测与分析中发挥了重要作用。然而,现有的公共ADR数据集以英文为主,中文ADR数据集的缺乏限制了在本土语境下对ADR的自动化研究与方法评估。与此同时,中国药品不良反应监测网络自1999年至2023年已累计收集超过2 327.5万份报告,如何借助自然语言处理方法从海量报告中提取关键信息成为迫切需求。

为了评估不同模型在中文药物不良反应识

别任务中的表现,本文构建了用于药物警戒领域的中文命名实体识别数据集(Chinese Named Entity Recognition Pharmacovigilance Dataset,CNER-PH)[①]。该数据集包含2 000条真实的ADR描述文本,涵盖12个命名实体类别,实体标注示例如图1所示。据我们所知,CNER-PH是在中文药物不良反应报告中,实体类型最多样化的命名实体识别数据集。基于该数据集,本文采用预训练模型与大语言模型进行系统评估,考察了不同模型在实体识别精度和边界判定等方面的表现差异,以揭示模型在药物不良反应识别任务中的适用性与局限性。

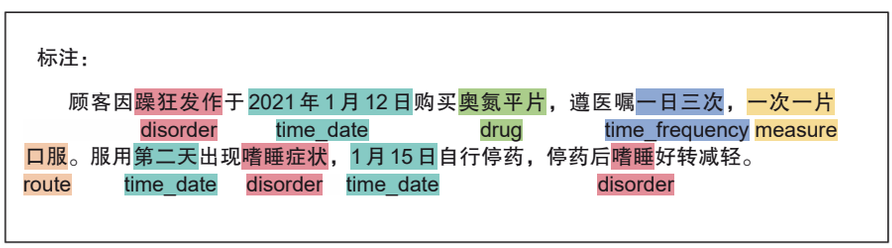


图1 CNER-PH中的实体标注示例

Fig. 1 Example of entity annotation in CNER-PH

1 方法

1.1 中文药物不良反应需求分析

近年来,全球药物不良反应事件的报告数量持续增长,对公共健康、安全用药和医疗决策产生了深远影响。在此背景下,对药物不良反应领域的信息进行研究和分析具有重要的现实意义。本文结合专家意见,收集并整理了中文药物不良反应领域面临的核心需求和问题:

R1:中文药物不良反应报告的处理。药物不良反应信息在全球范围内存在语言和表达方式的差异,与英文等其他语言相比,中文ADR文本的研究相对较少,且其表达方式独特,存在专业医学术语、口语化描述、简写等复杂情况,给实体识别带来了挑战。因此,NER任务需要针对中文ADR报告的语言特点,精准识别关键实体。

R2:医学术语和表达方式的多样性。ADR报告中的实体可能存在多种表达方式,例如,疾病或症状可能以不同名称出现(如"高血压"对比"血压升高"),药物名称可能以通用名、商品名或化学名表述(如"布洛芬"对比"芬必得")。无论其具体表述方式如何,NER任务需要精准识别文本中的所有相关实体。

R3:药物不良反应报告中的复杂实体结构识别。在ADR报告中,部分实体的表达方式并非简单的单个词语,而是以复杂结构出现。例如"低剂量阿莫西林"是完整的Drug实体,而不是单独的"阿莫西林"。

R1-R3作为CNER-PH数据集标注工作的基本原则,2.4节将详细介绍实体类别及标注过程。

1.2 任务定义

本文将NER任务理解为序列标注任务,并在CNER-PH数据集上开展了相应的研究(见图2和图3)。

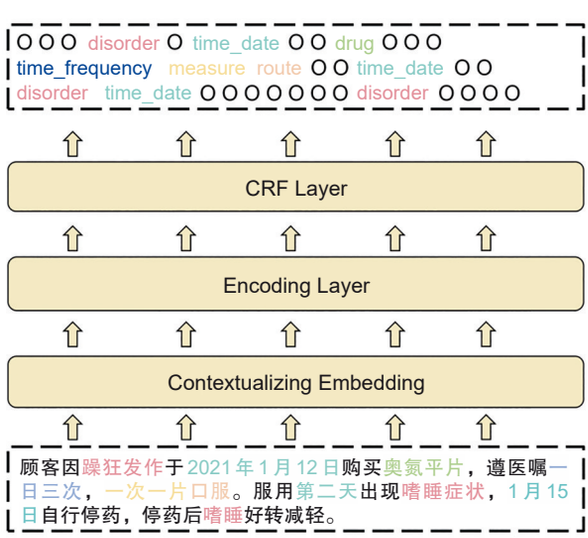


图2 模型结构:基于BiLSTM/BERT的方法

Fig. 2 Model architecture: methods based on BiLSTM/BERT

作为自然语言处理中的基础任务,序列标注旨在准确分类和标注文本序列中的每个元

① 数据集示例https://github.com/SieteSiete-77/CNER-PH,完整数据集请邮件联系1023040922@njupt.edu.cn

素。在命名实体识别的背景下,该任务涉及将文本中的单词或字符与预定义的实体类别关联起来。给定一个句子$S=x1, x2, \cdots, xn$,包含n个标记,本文使用标记级别的标签$y=y1, y2, \cdots, yn$,并应用“BIO”(Begin, Inside, Outside开始、内部、外部)标注方案,为输入句子中的每个标记分配标签。

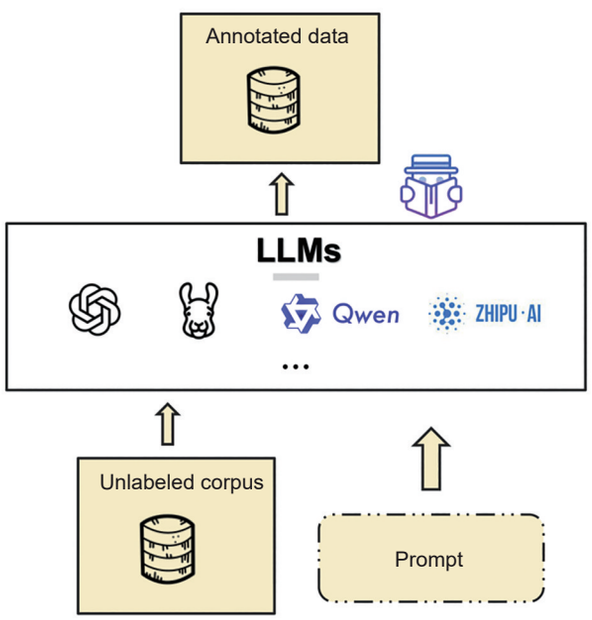


图3 模型结构:基于LLM的方法

Fig. 3 Model architecture: LLM-based approaches

1.3 数据准备

在去除所有可能涉及个人身份的信息后,本文构建了CNER-PH数据集。原始数据来源于中国药品不良反应监测中心归档的ADR报告,这些报告由多家医院、药店及其他医疗相关机构于2021年提交,涵盖了临床用药与药品销售过程中收集到的真实不良反应案例。

本文首先依据以下标准对原始报告进行筛选:1)文本长度不少于15个汉字;2)至少包含一种药物和一种疾病或症状;3)排除格式异常、内容缺失的记录。

在构建数据集时,本文从清洗后的报告中选取了2 000份内容完整、语言规范、实体多样的样本,确保覆盖多种药物类别(如抗生素、降压药、镇痛药)、不同年龄段患者(儿童、成人、老年人),以及多种不良反应类型(皮疹、头晕、呼吸困难等),以提高数据集的代表性和适用性。

1.4 数据标注

CNER-PH数据集由两位医学专家和三位计算机科学研究生标注,研究生在医学专家的指导下,熟悉实体定义、标注规范和边界判定规则。为确保手动标注一致性与可靠性,数据集中的所有样本均由两位标注员独立完成初轮标注。在存在标注意见不一致的情况下,由第三位标注员进行独立复标,随后通过多数投票法确定最终标注结果。针对边界模糊或实体类型争议较大的样本,标注团队会集中讨论,统一标准后再修订标注指南,并对部分已标注样本进行回溯修正,以提升标注一致性,标注流程如图4所示。标注工具选择了Doccano[20]。

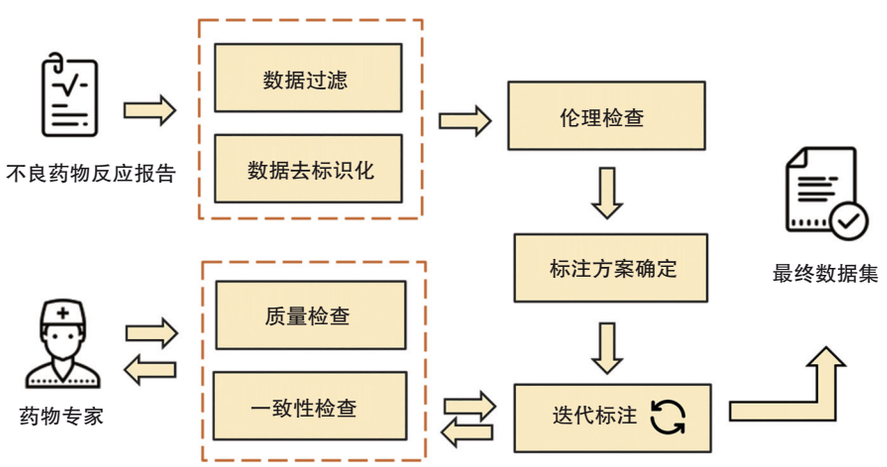


图4 CNER-PH数据集的标注过程

Fig. 4 Annotation process of the CNER-PH dataset

在CNER-PH数据集中,本文系统地标注了12种不同类型的命名实体,每种实体都经过精心选择,涵盖了药物警戒的关键方面。对于每种类型的实体,本文标注了文本中表示实体全名或相关缩写形式的连续片段。这种方法确保了正式术语和常用缩写都被捕捉,尤其适用于医学和药理学领域的文本分析。表1是实体的详细说明。

表1 CNER-PH中的实体及描述

Table 1 Entities and descriptions in CNER-PH

实体名称	实体描述
Drug	包括文本中提到的任何药物名称或成分,如润燥止药胶囊。本文还标注了药物的缩写或通俗名称,如NS表示生理盐水。
Disorder	指与患者健康相关的任何疾病、症状或体征。
Anatomy	指所有器官或解剖部位。需要注意的是,如果一个解剖部位是一个更大实体(如disorder)的一部分,本文优先将其标注为disorder。
Route	表示药物的给药方式,如口服或静脉注射。本文还标注了给药动词的缩写,如ivst表示“即刻静脉注射”。
Function	表示正常的生理功能,如“良好的睡眠”或“正常的心率”。
Test	指所有用于产生医学诊断结果的医疗监测、检查或其他程序。由于不良反应报告内容的灵活性,本文标注了“血压”和“测量血压”作为Test实体。
Measure	包括所有临床相关的测量,通常是药物剂量或检查结果,通常以带有单位的数值形式表示,如1.4 g或100 mL。
Method	表示治疗方法,如“抗感染治疗”。
Time_date	表示具体日期,如“2021年7月7日”。
Time_duration	指时间段,如“三天”。
Time_frequency	表示事件发生的频率,通常与药物服用时间表相关,如“每天三次”。本文还标注了时间频率的缩写,如BID表示“每天两次”。
Time_point	表示具体的时间点,如“下午2:30”。本文还标注了模糊的时间概念,如“那天下午”或“第二天”。

2 实验

本文在CNER-PH数据集上评估了序列标注方法。在本节中,本文将描述实验设计、评估指标和主要结果。

2.1 模型

本文使用了双向长短期记忆网络-条件随机场(Bidirectional Long Short-Term Memory-Conditional Random Field, BiLSTM-CRF)模型[21]和基于BERT[22]的模型进行序列标注。在BiLSTM-CRF模型中,LSTM作为特征提取器,用于建模句子的特征。通过门控机制,它有效地捕捉标记之间的长距离依赖关系,为每个词生成上下文感知的特征表示。本文研究了BERT、RoBERTa和基于纠错式掩码语言模型的BERT(MLM as Correction BERT, MacBERT)①在命名实体识别中的表现。

近年来,大模型在自然语言处理领域表现出显著优势。与传统的命名实体识别方法相比,大模型通过大规模的无监督预训练,能够捕捉语言中的复杂上下文信息。本文评估了16个大模型在NER任务中的表现。本文选取了OpenAI发布的GPT-4o和GPT-3.5-Turbo,GPT-4o在处理非英语文本时比GPT-3.5-Turbo有显著改进,同时更具成本效益。Anthropic的Claude 3.5 Sonnet在细微理解上表现出色,而Meta AI的Llama 3.1-70b、Llama 3.1-405b[23]、Llama 3.3-70b②、Llama-4-Scout-17B③以高效性和可扩展性著称。Google的Gemini-1.5-pro[24]在文本生成和分析方面表现出色。在中文社区中,本文选取了清华大学的GLM-4-plus[25]、DeepSeek团队的DeepSeek-V2.5[26]、DeepSeek-V3[27]、DeepSeek-R1[28],阿里巴巴的Qwen2-72b-Instruct、Qwen2.5-72b-Instruct和Qwen-Max[29],以及01.ai的Yi-large[30]模型。这些模型在中文文本处理任务中表现出了强大的性能。

在使用大模型进行NER任务时,本文采用了少样本、基于提示(prompt)的方法。提示构建方面,本文以结构化标签(例如<drug>…

① https://github.com/ymcui/Chinese-BERT-wwm

② https://huggingface.co/meta-llama/Llama-3.3-70B-Instruct

③ https://huggingface.co/meta-llama/Llama-4-Scout-1

＜/drug＞)明确实体边界与类型,模拟真实标注过程。模型被设定为一位擅长医学文本标注的专家,任务是根据给定的实体类型说明,对输入句子中的实体进行精确标注,并严格输出符合规范的标注结果。本文为模型提供了表1的实体描述,并在提示中加入了3个完整的示例,帮助模型理解任务格式。示例如下:输入:患者因荨麻疹于2017年9月2日入我院门诊就诊,予润燥止痒胶囊2G BID治疗。输出:患者因＜disorder＞荨麻疹＜/disorder＞于＜time_data＞2017年9月2日＜time_data＞入我院门诊就诊,予＜drug＞润燥止痒胶囊＜/drug＞＜measure＞ 2G ＜/measure＞ ＜time_frequency＞BID＜/time_fre-quency＞治疗。提示以结构化说明＋输入输出示例结合的方式,引导模型学习复杂实体边界和类型,有效提升了大模型在少样本条件下的NER表现。

2.2 实验设置

1)数据集划分

本文将数据集按照75%/10%/15%的比例划分为1 500个训练集、200个验证集和300个测试集。这三个数据集之间的药物不良反应报告没有重叠。

2)超参数设置

BiLSTM-CRF模型采用单层BiLSTM层,隐藏状态维度为256,词嵌入层维度为300,Dropout率设置为0.3,优化器为Adam。BERT系列模型输出的隐藏层维度设置为768,隐藏层和自注意力机制中的Dropout都设置为0.1,最大输入序列长度为512,优化器为Adam,学习率设置为3×10^{-5}。对于大语言模型,温度超参数通常用于控制模型文本生成的随机性。本文将温度设置为0,以确保模型产生最确定性的结果。

2.3 评估指标

为了评估模型在命名实体识别任务中的表现,本文使用了精确度(Precision,P)、召回率(Recall,R)和$F1$分数作为评估指标。首先,将模型的输出和标签使用特定的标注格式(如BIO)解析为一系列实体,其中实体的提取是基于实体类型以及在序列中的起始和结束位置。对于每个实体类型,本文统计标签中存在的实体、模型预测的实体和正确预测的实体。精确度定义为正确预测的实体数量与模型预测的实体总数之比。召回率定义为正确预测的实体数量与标签中实体总数之比。$F1$分数是精确度和召回率的调和均值,计算公式为:

$$F1=2\times\frac{P\times R}{P+R}。\tag{1}$$

2.4 实验结果

表2展示了在CNER-PH数据集上,使用BiLSTM-CRF模型和基于BERT的模型进行NER任务的表现。通过分析实验结果,本文得出以下结论:① 基于BERT的模型通常在$F1$分数上优于BiLSTM-CRF模型。这是因为BERT通过大规模预训练,能够捕捉到更丰富的上下文语义信息。此外,它的双向Transformer架构使得模型能够更有效地处理长距离依赖和复杂的语境。② 在预训练模型中,chinese-bert-wwm-ext的表现超过了chinese-bert-wwm。这是因为chinese-bert-wwm-ext在预训练过程中使用了更多的数据,本文认为更大的数据集帮助模型学习到了更丰富的语义和语言模式,从而提高了性能。在所有模型中,bert-base-chinese在本数据集上表现最优,表明该数据集的语言模式可能更贴近维基百科语料。③ 在比较不同实体的识别表现时,本文发现所有模型在区分test实体和其他类别时都存在困难,推测这是因为test实体在训练数据中相对较少,其边界较为模糊。

表2 BiLSTM/BERT方法在测试数据集上的结果

Table 2 Results of BiLSTM/BERT methods on the test dataset

模型	精确度/%	召回率/%	$F1$/%
BiLSTM-CRF	60.71	41.58	49.35
BERTbase	87.61	85.71	86.65
BERTwwm	85.48	71.35	77.78
BERTwwm-ext	86.76	80.67	83.61
MacBERTbase	87.22	80.61	83.78
RoBERTawwm-ext	87.34	85.77	86.55

分析表3中的数据,本文可以得出以下结论:① 大型模型通常在命名实体识别任务上的表现比大多数基线模型差。这种差异可能是因为这些模型的预训练语料库中,药物警戒相关内容所占比例较小。当这些模型直接通过

少量示例进行提示,而不是通过训练数据进行微调时,它们获得的领域特定信息较少,从而限制了它们有效理解和处理药物警戒相关概念的能力。② 作为中文语言模型,DeepSeek-V2.5 排名仅次于 Claude-3.5-Sonnet,达到了80.28%的召回率,优于许多其他英语或通用模型。这表明,专门针对中文训练的语言模型可能更能捕捉语言的细微差别,从而在与该语言相关的任务中表现得更为出色。

图 5 展示了不同模型在各个实体类别中的表现差异,所有模型在识别 drug 实体时表现最好,但在识别 function、test 和 anatomy 实体时表现较差。这种差异可能与这些实体的定义和复杂性有关。drug 实体通常具有明确的名称和特征,并且在预训练语料库中出现得较为频繁,因此模型更容易识别。相比之下,function 和 test 实体往往较为抽象,涉及复杂的上下文语义关系,而 anatomy 实体具有模糊的边界,且在预训练数据中较少出现,因此对于模型来说识别起来更具挑战性。尽管 Claude-3.5-Sonnet 在 drug 和 disorder 等类别中取得了相对较高的分数,但在需要细致理解的类别(如 function 和 test)上仍表现不足。同样,像 Qwen-Max 和 Qwen2.5-72b-instruct 这样的模型取得了中等表现,但它们的成绩并未超越较小的预训练模型,表明模型规模并不一定能提高特定精细化提取任务的结果,较大的模型需要更有针对性的预训练或架构调整,才能在多类别信息提取任务中表现出色。

表 3 LLM-based 方法在测试数据集上的结果

Table 3 Results of LLM-based methods on the test dataset

模型	精确度/%	召回率/%	*F*1/%
Claude-3.5-Sonnet	76.96	82.06	79.16
DeepSeek-v2.5	76.60	80.28	78.07
Gemini-1.5-pro	73.33	77.57	75.08
Qwen-Max	74.73	74.92	74.53
DeepSeek-v3	72.41	76.44	74.13
Qwen2.5-72b-instruct	71.44	76.00	73.37
GLM-4-plus	71.36	74.89	72.82
Yi-large	72.93	72.90	72.58
DeepSeek-R1	66.20	73.07	69.11
GPT-4o	65.80	72.28	68.51
Qwen2-72b-instruct	59.98	68.05	63.36
Llama-3.1-405b	60.24	65.09	62.25
Llama-4-Scout-17B-16E-Instruct	57.51	61.72	59.22
GPT-3.5-Turbo	57.19	61.65	58.98
Llama-3.3-70b	54.20	62.04	57.52
Llama-3.1-70b	38.02	38.53	37.90

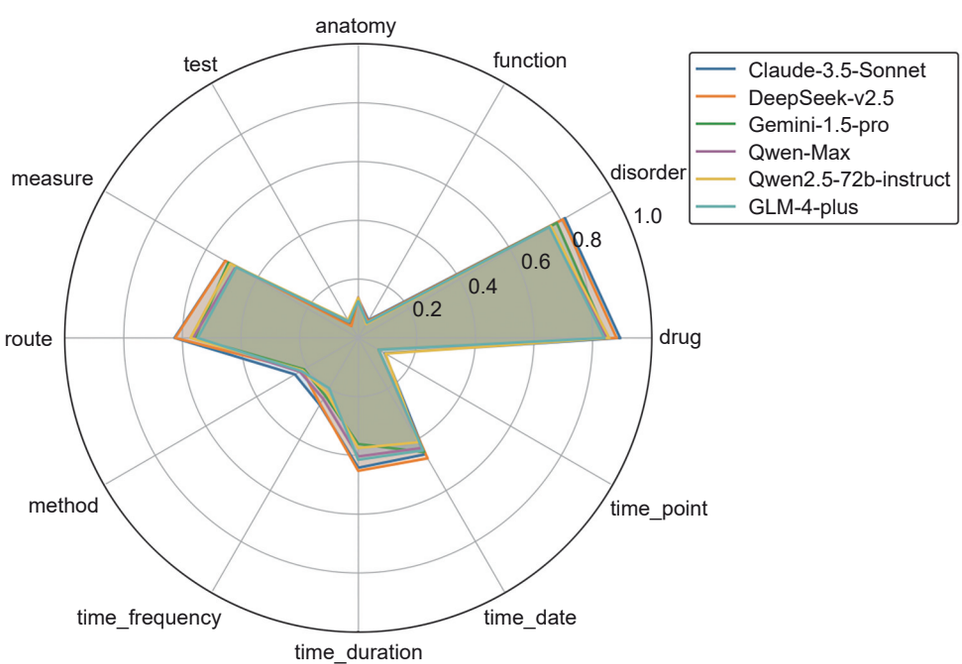


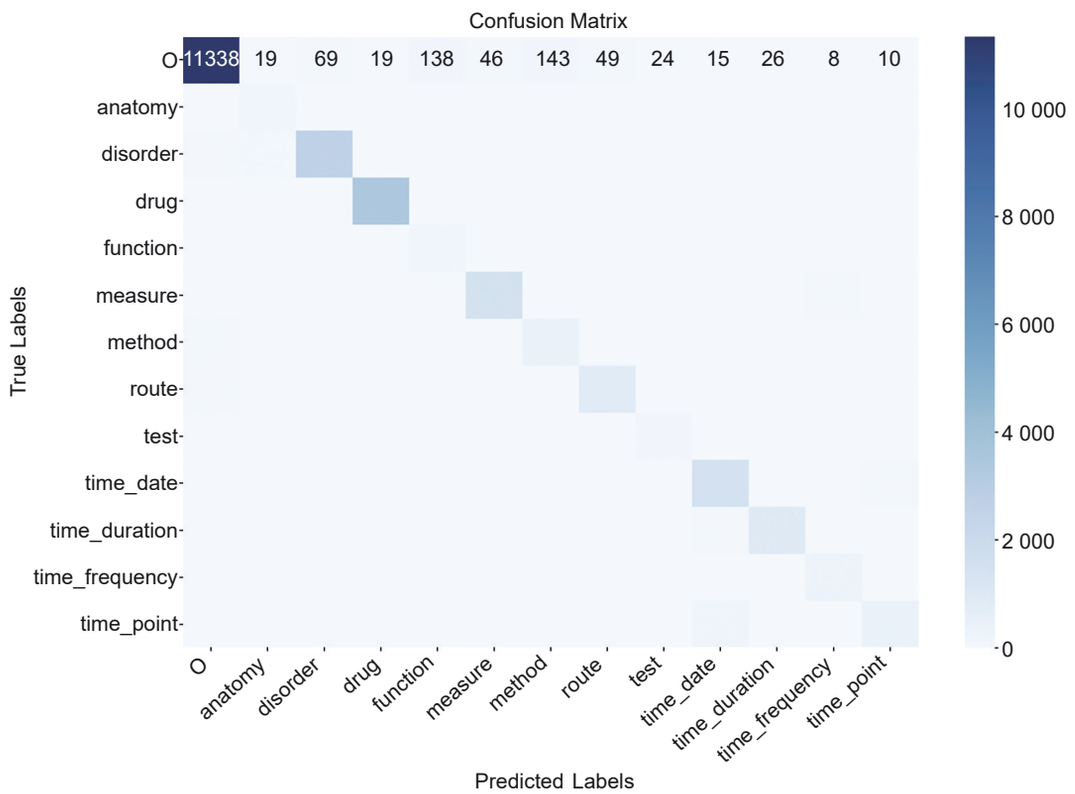
图 5 LLMs 在识别 12 种实体方面的表现

Fig. 5 Performance of LLMs in recognizing 12 entity types

2.5 错误分析

在这项研究中,本文通过绘制混淆矩阵对本实验中表现最佳的大模型进行了深入分析。混淆矩阵提供了模型在不同实体上的预测表现的直观视图,揭示了模型在特定类别上面临的分类困难和偏差。

如图6所示,某些实体类型,如Anatomy和Disorder,表现出了显著的误分类,特别是Anatomy实体常常被预测为Disorder或Drug。这种混淆反映了大模型在实际应用中面临的挑战,尤其是在医学领域,Anatomy和Disorder有时会有所重叠或相关,此外,Drug和Function之间也存在一定的混淆。另一方面,像度量和时间相关实体出现误分类的频率较低,这表明这些实体类型对于模型来说更容易区分。



注:上面的一行数字表示真实不是实体的内容,被模型分别预测成了哪些标签。11 338是预测正确,后面的19、69、138等是把非实体误判成实体的错误数量。

图6 Claude-3.5-Sonnet在NER任务上的混淆矩阵

Fig. 6 Confusion Matrix of Claude-3.5-Sonnet on the NER Task

3 结论

在快速发展的药物警戒领域,处理包含医学术语、口语化表达和结构化数据的大量复杂文本仍是一个重大挑战,对自动化语言处理提出了严苛考验。为开展相关研究,本文构建了中文药物不良反应命名实体识别数据集CNER-PH,为模型评估提供了实验基础。

基于该数据集,本文系统比较了多种主流NER模型与大语言模型在药物不良反应识别任务中的表现。通过实验与错误分析,揭示了不同模型在面对药物警戒相关命名实体的多样化词汇形式和边界模糊性时的优势与不足。结果表明,即便是表现最优的Claude 3.5 Sonnet,在处理部分复杂案例时依然存在挑战,显示出当前方法在准确性和鲁棒性上的改进空间。

综上所述,本文一方面贡献了一个实体类型多样的中文药物不良反应数据集,另一方面通过方法上的系统对比与误差剖析,为后续研究提供了经验和方向。该工作不仅为自然语言处理领域提供了重要基准,也为药物警戒专家、医学研究人员和公共卫生学者提供了有价

值的参考,有助于推动药物安全监测方法的改进与发展。

目前的数据集主要基于药物不良反应文本报告,忽视了其他模态在药物安全监测中的重要性,未来的研究方向可能包括通过引入视觉数据(如皮肤反应的图像)和临床数据表格来丰富现有的数据集,提高模型处理不同类型数据的能力,捕捉到更全面和复杂的信息。

参考文献:

[1] Edwards I R, Aronson J K. Adverse Drug Reactions: Definitions, Diagnosis, and Management[J]. *Lancet*, 2000, **356**(9237): 1255−1259. DOI: 10.1016/S0140-6736(00)02799-9.

[2] Alomar M J. Factors Affecting the Development of Adverse Drug Reactions (Review Article)[J]. *Saudi Pharm J*, 2014, **22**(2): 83−94. DOI: 10.1016/j.jsps.2013.02.003.

[3] Lependu P, Iyer S V, Bauer-mehren A, *et al*. Pharmacovigilance Using Clinical Notes[J]. *Clin Pharmacol Ther*, 2013, **93**(6): 547−555. DOI: 10.1038/clpt.2013.47.

[4] Demner-fushman D, Shooshan S E, Rodriguez L, *et al*. A Dataset of 200 Structured Product Labels Annotated for Adverse Drug Reactions[J]. *Sci Data*, 2018, **5**: 180001. DOI: 10.1038/sdata.2018.1.

[5] Sun Z Y, Li J Z, Pergola G, *et al*. PHEE: A Dataset for Pharmacovigilance Event Extraction from Text[C]//Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. Stroudsburg, PA, USA: ACL, 2022: 5571−5587. DOI: 10.18653/v1/2022.emnlp−main.376.

[6] Sboev A, Sboeva S, Moloshnikov I, *et al*. Analysis of the Full-size Russian Corpus of Internet Drug Reviews with Complex NER Labeling Using Deep Learning Neural Networks and Language Models[J]. *Applied Sci*, 2022, **12**(1): 491. DOI: 10.3390/app12010491.

[7] Raithel L, Thomas P, Roller R, *et al*. Cross-lingual Approaches for the Detection of Adverse Drug Reactions in German from a Patient's Perspective[C]//Proceedings of the Thirteenth Language Resources and Evaluation Conference. Marseille: European Language Resources Association, 2022: 3637−3649.

[8] D'Oosterlinck K, Remy F, Deleu J, *et al*. BioDEX: Large-scale Biomedical Adverse Drug Event Extraction for Real-world Pharmacovigilance[C]//Findings of the Association for Computational Linguistics: EMNLP 2023. Stroudsburg, PA, USA: ACL, 2023: 13425−13454. DOI: 10.18653/v1/2023.findings−emnlp.896.

[9] Raithel L, Yeh H S, Yada S, *et al*. A Dataset for Pharmacovigilance in German, French, and Japanese: Annotating Adverse Drug Reactions across Languages[C]//Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC−COLING 2024). Torino: ELRA and ICCL, 2024: 395−414.

[10] Sarker A, Gonzalez G. Portable Automatic Text Classification for Adverse Drug Reaction Detection *via* Multi-corpus Training[J]. *J Biomed Inform*, 2015, **53**: 196−207. DOI: 10.1016/j.jbi.2014.11.002.

[11] Garg S, Ramakrishnan G. BAE: BERT-based Adversarial Examples for Text Classification[C]//Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). Stroudsburg, PA, USA: ACL, 2020: 6174−6181. DOI: 10.18653/v1/2020.emnlp−main.498.

[12] Jia C, Shi Y F, Yang Q R, *et al*. Entity Enhanced BERT Pre-training for Chinese NER[C]//Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). Stroudsburg, PA, USA: ACL, 2020: 6384−6396. DOI: 10.18653/v1/2020.emnlp−main.518.

[13] Zheng W F, Lu S Y, Cai Z H, *et al*. PAL-BERT: an Improved Question Answering Model[J]. *Comput Model Eng Sci*, 2024, **139**(3): 2729−2745. DOI: 10.32604/cmes.2023.046692.

[14] Fan B, Fan W G, Smith C, *et al*. Adverse Drug Event Detection and Extraction from Open Data: A Deep Learning Approach[J]. *Inf Process Manag*, 2020, **57**(1): 102131. DOI: 10.1016/j.ipm.2019.102131.

[15] Mcmaster C, Chan J L, Liew D F L, *et al*. Developing a Deep Learning Natural Language Processing Algorithm for Automated Reporting of Adverse Drug Reactions[J]. *J Biomed Inform*, 2023, **137**: 104265. DOI: 10.1016/j.jbi.2022.104265.

[16] Chen Q J, Sun H T, Liu H Y, *et al*. An Extensive Benchmark Study on Biomedical Text Generation and Mining with ChatGPT[J]. *Bioinformatics*, 2023, **39**(9): btad557. DOI: 10.1093/bioinformatics/btad557.

[17] Tian S B, Jin Q, Yeganova L, *et al*. Opportunities and Challenges for ChatGPT and Large Language Models in Biomedicine and Health[J]. *Brief Bioinform*, 2023, **25**: bbad493. DOI: 10.1093/bib/bbad493.

[18] Gao Z X, Li L Y, Ma S Y, *et al*. Examining the Potential of ChatGPT on Biomedical Information Retrieval: Fact-checking Drug-disease Associations[J]. *Ann Biomed Eng*, 2024, **52**(8): 1919−1927. DOI: 10.1007/s10439-

023-03385-w.

[19] Juhi A, Pipil N, Santra S, *et al*. The Capability of ChatGPT in Predicting and Explaining Common Drug-drug Interactions[J]. *Cureus*, 2023, **15**(3): e36272. DOI: 10.7759/cureus.36272.

[20] Nakayama H, Kubo T, Kamura J, *et al*. doccano: Text Annotation Tool for Human[EB/OL]. (2018-09-26)[2025-08-24]. https://github.com/doccano/doccano.

[21] Huang Z H, Xu W, Yu K. Bidirectional LSTM-CRF Models for Sequence Tagging[PP/OL]. (2015-08-09)[2025-08-24]. https://doi.org/10.48550/arxiv.1508.01991.

[22] Devlin J, Chang M W, Lee K, *et al*. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding[C]//Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Minneapolis: Association for Computational Linguistics, 2019: 4171-4186. DOI: 10.18653/v1/N19-1423.

[23] Dubey A, Jauhri A, Pandey A, *et al*. *The Llama* 3 Herd of Models[PP/OL].(2024-07-31)[2025-08-24]. https://doi.org/10.48550/arxiv.2407.21783.

[24] Gemini Team. *Gemini* 1.5: Unlocking Multimodal Understanding Across Millions of Tokens of Context[PP/OL]. (2024-03-08) [2025-08-24]. https://doi. org/10.48550/arXiv.2403.05530.

[25] The Glm Team. ChatGLM: A Family of Large Language Models from GLM-130B to GLM-4 All Tools[PP/OL]. (2024-06-18) [2025-08-24]. https://doi. org/10.48550/arxiv.2406.12793.

[26] Liu A X, Feng B, Wang B, *et al*. DeepSeek-V2: A Strong, Economical, and Efficient Mixture-of-Experts Language Model[PP/OL]. (2024-05-07) [2025-08-24]. https://doi.org/10.48550/arxiv.2405.04434.

[27] Liu A X, Feng B, Xue B, *et al*. DeepSeek-V3 Technical Report[PP/OL]. (2024-12-27) [2025-08-24]. https://doi.org/10.48550/arxiv.2412.19437.

[28] Guo D Y, Yang D J, Zhang H W, *et al*. *DeepSeek-R*1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning[PP/OL]. (2025-01-22) [2025-08-24]. https://doi.org/10.48550/arxiv.2501.12948.

[29] Yang A, Yang B, Hui B, *et al*. *Qwen*2 Technical Report [PP/OL]. (2024-07-15) [2025-08-24]. https://doi. org/10.48550/arxiv.2407.10671.

[30] Young A, Chen B, Li C, *et al*. Yi: Open Foundation Models by 01.AI[PP/OL]. (2024-03-07)[2025-08-24]. https://doi.org/10.48550/arxiv.2403.04652.