

一种面向特定信息领域的大模型命名实体识别方法

李永斌^① 刘 棟^① 郑 杰^{*②}

^①(军委政法委员会某部 北京 100010)

^②(中国科学院空天信息创新研究院 北京 100080)

摘要: 在特定信息领域,尤其是开源信息领域,传统模型命名实体识别面临缺乏充足标注数据、难以满足复杂信息抽取任务等困难。该文聚焦开源信息领域,提出一种基于大语言模型的命名实体识别方法,旨在通过大语言模型强大的语义推理能力准确理解复杂的抽取要求,并自动完成抽取任务。通过指令微调和利用检索增强生成将专家知识融入模型,结合问题回归模块,使低参数通用型大模型基座能够快速适应开源信息这一特定领域,形成领域专家模型。实验结果表明,仅需少量的成本,便能构建一个高效的领域专家系统,为开源信息领域的命名实体识别提供了一种更为有效的解决方案。

关键词: 大语言模型; 命名实体识别; 指令微调; 检索增强生成; 知识库

中图分类号: TN91; TP391.1

文献标识码: A

文章编号: 1009-5896(2026)02-0662-11

DOI: 10.11999/JEIT250764

CSTR: 32379.14.JEIT250764

1 引言

在当今信息爆炸的时代,数据呈现出指数级的增长态势,大量的非结构化文本数据蕴含着丰富的有价值信息。命名实体识别(named entity recognition)作为自然语言处理(Natural Language Processing, NLP)领域的关键任务之一,旨在从非结构化文本中识别并提取出特定类型的实体,如人物、组织、地点和时间等,并对其进行分类。

近年来,大语言模型(Large Language Model, LLM)发展迅猛,为命名实体识别方法带来了崭新的研究视角。大语言模型拥有卓越的语义理解能力,它能够深入挖掘大量无标注文本,精准捕捉其中丰富的上下文信息。凭借这种强大的能力,零样本学习在实际应用中的效果得到了显著提升。在开源信息分析领域,借助大模型助力技术人员开展命名实体识别工作,有力支撑情报分析、态势感知和战略决策等重要任务。

1.1 命名实体识别

在传统的命名实体识别任务中,主要采用BiLSTM+CRF^[1]或者其变体,如BERT+BiLSTM+CRF^[2]等方法来进行实体抽取。这些方法通常需要预先确定抽取的类别模板,并且依据该模板人工标注大量的训练数据。然而,这种方式存在明显弊端,一方面,人工标注耗费大量人力成本;另一方面,训练所得模型往往泛化性能欠佳,难以有效适应不同领域的文本特征。

为了降低人工成本,近年来众多研究人员致力

于零样本知识抽取工作,取得了显著进展。例如,Wang等人^[3]以及Yang等人^[4]主要聚焦于对现有数据特征进行统计分析,实现对未标注数据的零样本学习,但该方法在大数据量下处理效率低下。针对上述问题,Lu等人^[5]提出了通用信息抽取统一框架,该框架实现了实体抽取、关系抽取、事件抽取和情感分析等多任务的统一建模,并且不同任务之间展现出良好的迁移和泛化能力。然而,该框架需要预先设置抽取模板,并且对于领域专有名词的抽取效果并不理想。

1.2 大模型

近年来,伴随人工智能技术的迅猛发展,以DeepSeek^[6-9]为典型代表的国产大模型取得了可观的成果。在近期的研究中,Wang等人^[10]将知识抽取任务以自然语言描述性指令的形式进行界定,并在各类知识抽取任务中展开训练,实现对未见标签的任务进行评估。此外,在其他领域,众多学者也开展了大量研究工作。例如,Hu等人^[11]开始探究大模型是否能够从放射报告中提取有价值的信息,Kartchner等人^[12]则对利用大模型进行零样本临床实验数据分析的可行性进行了探索,张国宾等人^[13]使用通用领域大模型进行特定领域信息抽取的初步尝试,皮乾坤等人^[14]借助大模型自动化抽取实体信息,并将其作为标注数据对特定领域的知识抽取模型进行微调,户才顺^[15]使用大模型在审计领域识别文本中的命名实体,胡慧云等人^[16]提出一种基于大语言模型的生成式多模态命名实体识别方法。

然而,当前命名实体识别研究主要面临以下4个问题:(1)传统实体抽取模型依赖于大量人工标注的训练数据,且训练后难以适应跨领域任务的

需求；(2)基于零样本学习的实体抽取方法在处理专有名词及定制化抽取内容时，识别效果仍不理想；(3)基于通用大模型的实体抽取方法在处理领域专有名词和术语时，抽取效果仍有待提升；(4)缺少特定领域知识库，仅通过数据微调大模型提升有限。

针对上述问题，本文提出一种面向开源信息领域的基于大模型的命名实体识别方法。该方法以DeepSeek大模型为基座，采用LoRA(Low-Rank Adaptation)^[17]微调技术，构建面向开源信息领域的指令微调数据集，使大模型能够充分学习并适应开源信息领域的数据特征。同时，引入检索增强生成(Retrieval-Augmented Generation, RAG)^[18]策略，创新性地针对开源信息领域设计和构建专家知识库，将外部知识库信息与大模型相结合，以增强模型在领域内实体抽取任务中的表现，并有效缓解大模型的“幻觉”现象。通过上述设计，本文方法

旨在提升命名实体识别在开源信息领域的准确性与鲁棒性。

2 基于大模型的命名实体识别方法

本文提出了一种面向开源信息领域的基于大模型的命名实体识别方法。该方法采用DeepSeek作为大模型基座，利用指令微调、RAG检索增强生成以及问题回归，使低参数基座大模型能够快速适配开源信息领域，整体框架如图1所示。

2.1 指令微调

本文面向开源信息领域构建指令微调数据集，使用LoRA微调来指导基础大模型学习用户指令，实现复杂实体抽取的需求。以表1中的样例为例，抽取结果不仅需要包含武器装备名称“尼米兹级航空母舰“杜鲁门”号”，还需挖掘出其对应的状态“紧急维修”和对应的数量“1”。具体执行方法如下。

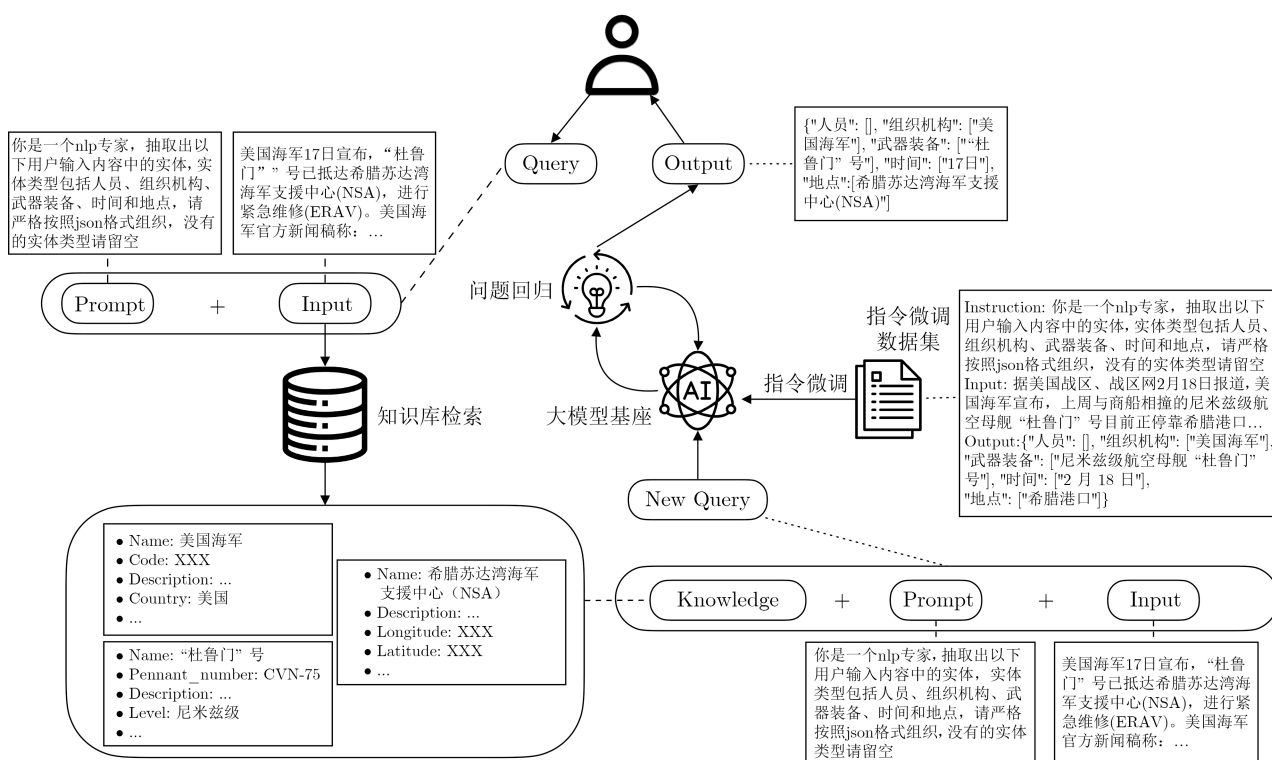


图1 模型框架图

表1 指令数据集样例

样例	
Instruction	你是一个nlp专家，抽取以下用户输入内容中的实体，实体类型包括人员、组织机构、武器装备、时间、地点，其中将文本中关于武器装备对应的状态和数量拼接到武器装备名称后，武器装备的状态和数量通过分隔符**拼接，如果没有则不拼接，同一个武器装备的数量进行合并，数量如果没有说明默认为一个，输出按照json格式，没有的实体类型留空。
Input	2月18日报道，美国海军宣布，上周与商船相撞的尼米兹级航空母舰“杜鲁门”号目前正在停靠希腊港口进行紧急维修。
Output	{ "人员": [], "组织机构": ["美国海军"], "武器装备": ["尼米兹级航空母舰“杜鲁门”号**1**紧急维修"], "时间": ["2月18日"], "地点": ["希腊港口"] }

2.1.1 构建微调数据集

指令微调的核心在于构建指令数据集，本文根据开源信息领域文本的特殊性以及定制化信息抽取需求构建指令数据集，指令数据集样例格式如表1所示。

首先，通过互联网针对性地搜集了一批关于开源信息的新闻文本。为满足构建高质量数据集的需求，本文过滤掉未包含武器装备名称、数量及状态的文本，确保筛选后的文本具备丰富的目标信息。随后，借助信息抽取智能体对筛选后的文本进行批量地实体抽取，将其中时间、地点、组织机构、人员以及武器装备实体提取出来，减少人工标注成本。最后，为构建高质量数据集，对抽取结果开展人工校正工作。针对抽取错误结果和被遗漏实体进行补充完善，并补充武器装备实体的数量和状态信息。

2.1.2 LoRA微调

在完成微调训练数据集的构建后，为使大模型能够更好地适配特定任务与领域需求，需要进行微调训练工作，本文选用LoRA微调方法。具体而言，本文针对模型中的“q_proj”，“k_proj”，“v_proj”，“o_proj”，“gate_proj”，“up_proj”和“down_proj”这几个关键层的参数进行微调。LoRA的微调优化公式为

$$\max_{\Theta} \sum_{(x,y) \in Z} \sum_{t=1}^{|y|} \log_2 (p_{\Phi_0 + \Delta\Phi(\Theta)}(y_t | x, y_{<t})) \quad (1)$$

其中， x 表示输入， y 表示输出， Z 表示微调数据集， Φ 表示模型参数， Θ 表示低秩参数。相较于全量参数 $\Delta\Phi$ 更新，引入参数量更少的 $\Delta\Phi(\Theta)$ 来低秩降维近似，有效减少了可训练参数的数量，降低计算成本的同时，能较好地保持模型性能。

2.2 RAG检索增强生成

本文借助RAG技术，针对已构建的开源信息领域专家知识库展开深度检索，有效规避大模型在处理开源信息相关任务时频繁出现的“幻觉”问题。同时，借助RAG技术与专家知识库，显著增强了大模型在开源信息领域的实体识别能力。具体做法如下。

2.2.1 知识库构建

RAG首先需要构建领域专家知识库。当前主流方法多采用“文本分块-向量嵌入-向量存储”的三段式架构(如Chunk-Embed-Store范式)，虽然具有实施便捷性的优势，但经实证分析发现存在显著缺陷：当处理军事领域文本时，传统分块策略会导致实体边界断裂，且在向量空间中易发生语义漂移现象。

为了解决上述问题，本文设计了3类数据表实现对军事要素的结构化建模：固定目标表、移动目标表和编制序列表。固定目标表定义了“军事基地”“港口”等静态目标的“经纬度”“名称”“描述”等7个核心字段，移动目标建立涵盖“舰船”“航空器”等动态目标的“名称”“别名”“编号”“国家”等14个字段，编制序列构建了“部队”的“名称”“编号”“部署装备”等6个字段。字段详情见表2-表4。

固定目标表“名称、经纬度、描述”等7个核

表2 固定目标知识库数据库表字段信息

序号	字段名称	描述
1	longitude	经度
2	latitude	纬度
3	name_target	固定目标名称
4	name_alias	固定目标别名
5	description	固定目标描述
6	country	国家
7	arm_force	部署部队编号

表3 移动目标知识库数据库表字段信息

序号	字段名称	描述
1	pennant_number	舷号
2	name	移动目标名称
3	name_en	移动目标英文名称
4	name_alias	移动目标别名
5	level	级别
6	country	国家
7	construct_shipyard	造船厂
8	builder	建造商
9	major_experience	主要经历
10	member_of	隶属部队
11	home_port	母港
12	construct_date	建造日期
13	service_date	服役日期
14	retire_date	退役日期

表4 编制序列知识库数据库表字段信息

序号	字段名称	描述
1	country	国家
2	code	编码
3	name	名称
4	description	描述
5	parent	上级部门
6	equipments_names	部署装备名称

心字段，直指军事基地、港口等静态实体的空间唯一性与属性稳定性特征，为实体识别提供“实体核心-关键属性”的强绑定关系：

(1)实体边界清晰化：“名称”作为静态实体的唯一身份标识，与“经纬度”（空间属性）、“隶属单位”（组织属性）等字段直接关联，即使RAG检索到的文本片段存在分块割裂，结构化字段仍能让大模型快速定位“名称+经纬度”对应的完整实体，避免将“某军事基地”与其地理坐标误判为两个独立信息单元。

(2)语义歧义消解：“描述”字段（如“战略轰炸机部署基地”“军民两用港口”）补充实体功能属性，结合其他字段，帮助大模型区分同类型但不同价值的实体（如将“训练机场”与“作战机场”精准识别为两类固定目标），规避因文本语义模糊导致的识别偏差。

移动目标表“名称、别名、编号、国家”等14个字段，适配舰船、航空器等动态实体的流动性与多源表述特性，为实体识别提供跨语境、跨时间的关联依据：

(1)实体唯一性强化：“编号”（如舰船舷号“DDG-1000”、战机编号“82-0001”）作为动态实体的刚性标识，与“国家”字段组合，可直接锁定实体归属与个体身份——即便RAG检索到的文本中实体名称表述不一致（如“F-35A”与“闪电II”），“别名”字段能实现多称谓对齐，避免大模型将同一实体误判为多个不同实体，从根源上缓解语义漂移。

(2)实体属性补全：字段涵盖“型号、载弹量、续航力”等战力属性，与“当前位置、航行状态”等动态属性形成互补，大模型可通过RAG检索的结构化字段，完整识别实体的“身份-状态-战力”全貌（如“美军‘福特’号航母(CVN-78)，当前位于菲律宾海，搭载 F-35C 战机”），避免仅识别实体名称而遗漏核心作战属性。

编制序列表“名称、编号、部署装备”等6个字段，紧扣部队等编制实体的组织层级与战力构成特性，为实体识别提供“单元-装备-职能”的逻辑链条：

(1)实体类型精准定位：“名称、编号”明确部队在指挥体系中的身份（如“第101空降师”），“隶属层级”“任务类型”字段进一步界定实体职能属性（如“两栖作战部队”“防空部队”），帮助大模型避免将“后勤保障旅”与“作战旅”混淆，提升组织类实体的识别精度。

(2)实体关系显性化：“部署装备”字段直接

关联移动目标表的动态实体（如“装备F-22A的航空兵部队”），使大模型通过RAG检索即可建立“部队编制→装备构成→战力水平”的关联认知，不仅能识别“某部队”这一核心实体，还能完整捕捉其附属装备实体及两者的作战协同关系，突破传统文本分块下“组织与装备信息割裂”的识别局限。

2.2.2 知识检索

构建专家知识库后，需要对库中知识进行检索。目前，很多知识库检索研究工作普遍采用向量检索方法。然而，向量检索方式存在一系列问题：(1)在对文本进行预处理时极易导致语义断裂，破坏原本完整的语义结构。(2)将文本切块后进行向量嵌入操作容易引入噪声，干扰向量的准确性，进而降低检索的质量。(3)向量相似度计算无法充分考虑文本的语义深度、上下文关联以及领域知识，这使得检索结果可能与用户期望的相关信息存在偏差。

鉴于向量检索存在上述诸多问题，针对已构建的开源信息领域专家知识库，本文采用基于倒排索引的方式进行检索。倒排索引能够直接建立从词汇到文档的映射关系，避免了文本切块带来的语义断裂问题，同时减少向量嵌入过程中引入噪声的风险，为开源信息的高效检索提供更可靠的解决方案。

(1)构建索引：具体而言，为了充分发挥已构建的结构化知识库的价值，本文制定了一套针对结构化数据库字段的处理策略。从结构化数据库中挑选若干具有代表性和关联性的字段，以此组建关键词集合。以移动目标表(表3)为例，选取其中的“name”（名称）、“name_alias”（别名）、“construct_shipyard”（建造船厂）等字段，基于这些字段所包含的内容，构建倒排索引。

倒排索引通过预构建“关键词-实体记录”的映射关系，将海量数据扫描转化为关键词精准定位，快速召回含查询相关信息的候选实体记录。其必要性体现在3方面：一是破解海量开源数据检索效率瓶颈，将线性扫描的高延迟转化为关键词映射的快速匹配，满足大模型实时响应需求；二是充分释放结构化字段价值，使“名称”“别名”等强关联属性成为精准检索入口，避免结构化数据优势流失；三是为模糊匹配提供落地基础，先通过关键词圈定候选集再执行正则匹配，大幅降低海量数据模糊匹配的计算成本。

(2)模糊匹配：当用户输入待查询的文本内容时，系统运用分词器对文本进行分词操作。然而，开源信息领域存在大量命名规则复杂、名称较长的武器装备实体，分词器容易出现误切的情况，从而

导致无法检索到正确的内容。为了有效解决这一问题,在进行全文检索时,引入正则表达式进行模糊匹配,即使在分词结果不准确的情况下,也能尽可能地找到相关的信息。例如,“火星-15洲际弹道导弹”会被分词器切分为“火星”、“-”、“15”、“洲际”和“弹道导弹”,无法直接检索,利用正则表达式“ \wedge 火星.*弹道导弹\$”或者“ \wedge .*弹道导弹\$”等进行匹配,能够有效扩大匹配范围,提高检索的召回率。

(3)相似度排序:在完成检索操作并获得查询结果后,为了能够为用户提供最相关的信息,需要对查询结果按照相似度进行排序。本文采用BM25(Best Match 25)算法用于知识库的相似度检索。BM25算法的核心思想是通过考虑查询词在文档中的出现频率、文档的长度以及查询词在整个文档集合中的普遍程度等因素,来计算查询语句与文档之间的相关性得分。得分越高,表示文档与查询的相关性越强。算法计算原理为

$$\text{Score}(Q, D) = \sum_{i=1}^n \text{IDF}(q_i) \times \frac{\text{TF}(q_i, D) \times (k_1 + 1)}{\text{TF}(q_i, D) + k_1 \times \left(1 - b + b \times \frac{\text{dl}}{\text{avgd1}}\right)} \quad (2)$$

其中, k_1 和 b 是调节因子,分别用于控制词频和文档长度对得分的影响程度, D 表示文档, dl 是文档 D 的长度, avgd1 是文档集合中所有文档的平均长度, q_i 表示查询内容, $\text{TF}(q_i, D)$ 表示查询词在文档中出现的频率, $\text{IDF}(q_i)$ 为逆文档频率,计算方法为

$$\text{IDF}(q_i) = \log_2 \frac{N - n(q_i) + 0.5}{n(q_i) + 0.5} \quad (3)$$

其中, N 为文档集合中的文档总数, $n(q_i)$ 是包含查询词 q_i 的文档数量。

BM25算法通过综合考量查询词在记录中的出现频率、记录长度及查询词在全库的普遍程度,量化用户查询与候选记录的相关性并排序。其必要性主要包括:一是解决倒排索引召回结果的冗余问题,通过得分差异区分关联紧密程度,避免无关记录干扰(如将“火星-15”专属记录置顶,排除仅含“弹道导弹”的其他型号结果);二是适配开源信息的精准性需求,强化“编号”“别名”等特有字段的权重,确保结果与查询核心意图对齐;三是弥补模糊匹配的限制性,在扩大召回率的同时通过二次筛选提升精确率,实现召回广度与排序精度的平衡。

(4)更新查询请求:如图1所示,将排序后的知识结合提示(prompt)和用户输入(input),构建新的查询请求(query)。后续将新的查询请求输入大模

型,即可将大模型分析的结果返回给用户。其中,构建查询请求的提示模版为“ $\{\text{knowledge1}\}, \{\text{knowledge2}\}, \dots \backslash \text{n}\{\text{prompt}\} \backslash \text{n}\{\text{input}\}$ ”。

2.3 问题回归模块

为提升低参数模型基座的抽取能力,本文增设回归模块。该模块使模型在抽取实体时输出置信度,通过累计保存低置信度实体结果,按类别定期统计,并将高频类别作为典型错误样例纳入模型反思过程,从而有效提高特定领域长难实体的识别率。

2.3.1 问题统计

问题统计环节旨在系统收集模型抽取的实体结果,并基于预设的置信度阈值对这些结果进行筛选。值得注意的是,置信度阈值的取值与抽取性能之间存在显著关联:较高的阈值虽能在一定程度上提升抽取结果的准确率,却可能将部分本应正确的实体误判为低置信度结果而予以抛弃,进而导致召回率下降,因此需结合具体应用场景的需求在准确率与召回率之间进行合理权衡。

对于经置信度阈值筛选后被判定为低置信度并遭抛弃的结果,本文设计了专门的存储机制对其进行保存,并设定以一周为周期进行定期统计。在统计过程中,首先依据实体的类型对这些被筛选掉的结果(即错误实体)进行归类,同时精确标注其产生时间,这一操作不仅便于后续对特定时段的错误数据进行快速定位,也为针对性分析错误模式、优化模型性能提供了可追溯的基础数据支持。

2.3.2 问题反思

问题反思环节定期从问题统计表中提取最近一周的错误结果。由于单周积累的错误样本量较大,需对结果进行抽样处理。为确保大概率抽中错误数量较多的类型,本文采用离散型均匀分布作为采样方法。具体步骤如下:

(1)计算每个错误类型的抽样权重:先统计每个错误类型的历史样本数量,再计算其占有所有类型总数的比例。例如:“人物”“组织机构”“武器装备”实体类型的数量分别是10, 20和30,则其权重比例分别为16.7%, 33.3%和50%。

(2)构建累积分布函数:将权重按顺序进行累加,得到累积概率,用于后续抽样判断。例如:“人物”“组织机构”“武器装备”的累积概率分别为16.7%, 50%, 100%。

(3)生成随机数并匹配区间:生成0~1的随机数,根据随机数落在累积概率区间确定被选中的类别。例如:在0~16.7%区间内选中“人物”类别,在16.7%~50%选中“组织机构”类别,在50%~100%选中“武器装备”类别。

本文采用大模型生成方式构建反思提示词模板，结合抽样结果作为输入，指导大模型对错误结果进行反思与纠正，进而输出正确的实体结果。提示词模板如图2所示，红色部分替换成问题统计的结果。

3 实验结果

3.1 实验数据

由于开源信息领域的敏感性，本文从互联网搜集开源信息相关的新闻文本，对长篇新闻进行切分，并请领域专家对搜集的文本进行人工标注，形成开源信息领域的命名实体识别训练和测试数据集。数据集统计信息如表5所示。

实验数据集包含两个字段，分别是原始文本(text)和实体标注结果(result)，数据集样本数据实例如表6所示。

3.2 实验设置

3.2.1 基线模型

为了与本文提出的方法进行对比，本文选取了BiLSTM-CRF, BERT-CRF和UIE作为传统命名实体识别模型和零样本知识抽取模型中最具代表性的经典模型。

你是一个顶级算法专家，专为检查用于实体抽取信息是否正确准确而设计。

#任务
基于给定的{TYPE}类型以及输入文本，决定该{TYPE}抽取结果是否准确，输出的结果参照样例和结果格式。

实体应该至少包括2个属性，分别是名称、类型。

你需要从以下三个方面提供修改意见：
增加：表示不完整的实体需要补充，特别是丢失的文本内容
修改：表示抽取结果中的实体属性需要被修正，判断{TYPE}是否需要被修正，并给出理由

#样例
下面是一些例子，你可以参考。

输入：
{"name":"里根号","type":"武器装备"}

输出：
{"name":"里根号航空母舰","type":"舰船","reason":"实体名称补全，实体类型细化"}

#注意
不要增加额外的解释文本。
确保输出文本的语言与输入文本保持一致。
确保输出结果的类型是确定的。

图2 反思提示词模板

(1)BiLSTM-CRF：经典的序列标注模型架构，被自然语言中的命名实体识别任务广泛应用。

(2)BERT-CRF：一种融合了BERT预训练模型与条件随机场(CRF)的序列标注模型，通过BERT的深层语义表示能力捕捉文本上下文特征，借助CRF层优化序列标签间的依赖关系，常用于实体识别、词性标注等自然语言处理任务。

(3)UIE：通用信息抽取统一框架，该框架实现了实体抽取、关系抽取、事件抽取、情感分析等任务的统一建模，并使得不同任务间具备良好的迁移和泛化能力。

3.2.2 实验参数

(1)BiLSTM-CRF：采用Wiki预训练的词向量，词向量维度100，LSTM网络单元数量100，dropout为0.5，学习率为0.001，优化器为Adam，训练轮次为100轮。

(2)BERT-CRF：采用bert-base-chinese作为预训练基础模型，隐藏层维度为768，dropout为0.3，学习率为3e-5，优化器AdamW，训练次数20轮。

(3)UIE：采用ErnieModel基础模型，隐藏层维度为768，dropout为0.1，学习率为1e-5，优化器为AdamW，训练轮次为20轮。

(4)本文：采用DeepSeek 8B参数版本作为基础模型，学习率为1e-4，训练轮次为10轮，LoRA秩设置为8，LoRA Alpha设置为32，LoRA Dropout设置为0.1。

模型运行的硬件服务器配置中CPU为Intel(R) Core(TM) i7-9700K CPU @1.60 GHz，内存62 GB，GPU为NVIDIA Geforce RTX 3090，显存24 GB。在实验过程中，模型的训练和推理使用相同的参数和硬件配置。

3.2.3 评估指标

在评估命名实体识别模型的实验效果时，一般

表5 实验数据统计信息表

统计信息	训练集	测试集
数据量	10426	1000
实体类别数量	5	5
实体标注数量	23780	8369
标注样本文本平均长度	27.8	154.1

表6 标注样例

	样例文本
原始文本	6月15日报道称，两架美国空军的F-22战斗机于当地时间6月13日下午4点左右从珍珠港-希卡姆联合基地起飞，第三架F-22战斗机在它们起飞约一小时后起飞，期间一架KC-135空中加油机也起飞为这些战斗机提供支援
标注结果	{"人员": [], "组织机构": ["美国空军"], "武器装备": ["F-22战斗机**起飞**3", "KC-135空中加油机**起飞**1"], "时间": ["6月15日", "6月13日下午4点左右"], "地点": ["珍珠港-希卡姆联合基地"]}

采用精确率(Precision), 召回率(Recall)和F1值作为评估指标。本文使用micro-F1作为评价指标, 其计算公式为

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad (4)$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (5)$$

$$\text{F1} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (6)$$

3.3 实验结果

表7展示了通过本文所提方法进行命名实体识别的整体性能, 以及与其他基线模型的对比实验结果。在测试数据集上的性能结果如下所示, 数据按照百分数展示。

从表7中可以看到, 本文提出的方法相较于基线模型在精确率和F1值上均有大幅度领先。为了更加细粒度地分析各个实体类别抽取情况, 本文又对“人员”、“组织机构”、“武器装备”、“时间”和“地点”这5种实体类别分别进行实验性能评估, 结果如表8-表12所示, 数据按照百分数展示。由于基线模型无法抽取出武器装备实体对应的数量和状态, 因此表8-表12中的武器装备实体评估性能只考虑武器装备名称, 武器装备实体对应的数量和状态会在案例分析章节中展开说明。

分析表8-表12的实验结果可以发现, 在通用类

型实体的识别准确率上, 本文模型与基线模型性能相当, 在“人员”类实体识别中F1值差异显著。对实验进行深入分析, 可以发现:

(1)类别样本分布不均衡: 开源“人员”实体存在典型表述(如“旅长张某某”)与长尾细分类型(如“无人机分队领航员王某某”)的频次差异, 样本覆盖度直接影响模型泛化能力。Ours依托大模型语义优势, 对微调数据中高频样本识别精准(90.6%精确率), 但因长尾样本缺失导致漏检严重(42.8%召回率); UIE凭借预训练阶段的跨领域全量样本积累, 覆盖各类“人员”表述, 实现高召回(93.7%)与高精度(100%); BiLSTM-CRF受限于传统架构, 无法从少量标注数据中习得特征, 受分布不均衡影响最大, 性能全面落后。

(2)实体边界复杂度高: 开源“人员”常与部队编制、装备型号深度嵌套(如“‘山东舰’舰载飞行员刘某某”), 边界定位难度大。本文微调阶段缺乏针对性边界剥离训练, 对长限定词后的实体易漏检; UIE通过抽取式预训练架构专门优化“限定词-核心姓名”依赖关系, 精准处理嵌套边界; BiLSTM-CRF因长距离依赖建模能力不足, 边界切割误差大, 进一步拉低性能。

在特定领域(“武器装备”)实体类别识别任务中, 本文所提出的模型在精确率、召回率和F1值等关键评估指标上, 均显著优于其他基线模型。

表7 模型总体性能对比(%)

	精确率	召回率	F1值
BiLSTM-CRF	55.7	53.2	54.5
BERT-CRF	75.2	72.1	73.6
UIE	70.4	73.8	72.0
本文	95.8	73.5	83.2

表8 “人员”实体类别识别模型性能对比(%)

	精确率	召回率	F1值
BiLSTM-CRF	8.9	7.6	8.2
BERT-CRF	72.5	68.3	70.3
UIE	100.0	93.7	96.7
本文	90.6	42.8	58.2

表9 模型“组织机构”实体类别性能对比(%)

	精确率	召回率	F1值
BiLSTM-CRF	61.8	73.3	67.0
BERT-CRF	75.0	74.2	74.6
UIE	59.6	73.7	65.9
本文	99.5	72.5	83.9

表10 模型“武器装备”实体类别性能对比(%)

	精确率	召回率	F1值
BiLSTM-CRF	41.6	29.7	34.7
BERT-CRF	72.3	60.5	65.9
UIE	53.4	52.0	52.7
本文	90.0	65.0	75.5

表11 模型“时间”实体类别性能对比(%)

	精确率	召回率	F1值
BiLSTM+CRF	99.6	100.0	99.8
BERT-CRF	99.8	99.5	99.6
UIE	100.0	100.0	100.0
本文	100.0	98.9	99.4

表12 模型“地点”实体类别性能对比(%)

	精确率	召回率	F1值
BiLSTM+CRF	63.4	75.4	68.9
BERT-CRF	85.6	84.2	84.9
UIE	92.4	91.8	92.1
本文	100.0	98.2	99.1

“武器装备”实体与其他类别实体相比，其命名规则复杂多样，且常伴有特定的装备编号，同时，可供训练的包含此类实体的数据集极为匮乏。这一系列因素导致“武器装备”类型实体的识别难度远高于其他类型实体。本文模型之所以能够在“武器装备”实体类型识别中取得较好的表现，得益于军事领域专家知识库的构建以及指令微调技术的运用，将通用大模型迅速转化为适用于军事领域的专家模型。

为了分析知识检索和数据微调模块对模型的影响，本文增加消融实验来比较分别去除指令微调和RAG后模型的对比效果。具体对比结果如表13所示：

去除指令微调后，模型精确率从95.8%降至79.6%(下降16.2个百分点)，召回率从73.5%降至68.2%(下降5.3个百分点)，F1值降至73.5%。指令微调的核心作用是让大模型理解开源实体识别的任务指令(如“抽取装备型号、部队番号等实体”)，优化实体边界判断与类别映射逻辑。去除后，模型对任务要求的理解精度下降，易将“形似实体的非实体文本”误判为目标实体，导致精确率显著下降；而RAG仍能通过检索领域知识库提供实体候选，因此召回率下降幅度较小，但整体性能受精确率拖累明显。

去除RAG后，模型召回率从73.5%降至54.8%(下降18.7个百分点)，精确率从95.8%降至91.2%(下降4.6个百分点)，F1值降至68.7%。RAG的核心作用是通过检索开源领域知识库，为模型提供低频、专业实体的参考信息(如“东风-41洲际导弹”“第72集团军合成旅”等)。去除后，模型仅依赖预训练知识，对领域内长尾实体(如特殊作战单元、新型装备型号)的覆盖能力大幅下降，导致漏检增多，召回率显著降低；而指令微调仍能保证模型对已知实体的判断精度，因此精确率下降有限，但F1值因召回率不足大幅下滑。

3.4 案例分析

为充分验证本文所提方法的有效性，除了使用评价指标进行实验性能分析外，本文将所提方法应用于真实文本以观察其实际抽取效果。具体而言，将所提方法分别应用在多个国产大模型上(通义千问2.5B, DeepSeek7B, DeepSeek8B)，并选取大模

型DeepSeek671B参数版本(满血版本)作为对比对象，观察各模型在同一批开源信息领域文本的实体抽取任务中抽取结果的质量。

分析以上案例可以发现以下几个问题：(1)DeepSeek671B能够把文本中包含的“人员”、“组织机构”、“武器装备”、“时间”和“地点”抽取出来，但是会夹杂一些噪声信息和格式问题，而本文模型没有输出无关信息，并且按照指令组织结果。(2)DeepSeek671B抽取的“武器装备”实体与原文实体存在略微变化，本文模型的抽取结果则是保留了装备自带的引号。

基于上述现象，本文分析其原因主要是两点：(1)本文模型构建了指令微调数据集，在数据集中规定了模型预期输出，因此相较于DeepSeek 671B，本文模型虽然参数量较小但输出格式更加符合要求。(2)大语言模型本质是文本生成模型，无法保证生成结果的稳定性。本文模型则是通过借助领域专家知识库，利用RAG辅助大语言模型约束实体抽取来使得模型抽取结果更加符合标准。

此外，为验证所提方法的移植性，本文分别选取Qwen2.5B, DeepSeek7B和DeepSeek8B 3种基础大模型展开实验，其输出结果列于表14。实验表明，不同规模模型的输出内容高度一致，字段结构与内容组织形式保持一致，有效证明了方法在异构大模型上的适配能力。

综上，本文方法显著提升了特定领域实体的识别效能，使低参数轻量模型的性能可与全参数大模型相媲美；同时，其良好的移植性能够适配多种国产大模型，支持快速构建领域专属专家模型。

4 结束语

本文创新性地提出了一种面向开源信息领域、依托大模型的命名实体识别方法。该方法综合运用指令微调和RAG检索增强生成技术，显著提升了大模型在开源信息领域的适配性，有效规避了人工标注所产生的高昂成本，为用户精准抽取开源信息领域的实体信息提供了有力支持。相较于传统模型，本文提出的方法仅需500~800条核心开源实体的标注文本，较传统 BiLSTM-CRF(需10 000+条)减少92%以上，较通用大模型全量微调(需3 000+条)减少73%以上。本文模型是基于DeepSeek 8B参数量通用大模型基座，单张NVIDIA RTX 3090 (24 GB显存)即可完成训练，无须多卡集群。

展望未来，计划在现有文本大模型的基础上引入多媒体数据，致力于构建融合文本、图片、视频的多模态大模型，以满足开源信息领域日益复杂的信息处理需求。同时，将搭建多媒体专家知识库，

表 13 消融实验结果对比(%)

	精确率	召回率	F1值
本文(-Instruction Tuning)	79.6	68.2	73.5
本文(-RAG)	91.2	54.8	68.5
本文	95.8	73.5	83.2

表 14 本文模型与DeepSeek671B模型抽取结果对比

案例	
Input	2024年12月5日, 美国五角大楼宣布, 在中东地区的军事行动中, 美军特种部队于伊拉克巴格达市附近执行任务。此次行动涉及美军中央司令部指挥, 行动人员包括由队长约翰·米勒带领的30人特种作战小组。行动中动用了5架AH-64“阿帕奇”武装直升机, 其中3架状态良好, 可随时投入战斗, 另外2架在前期执行任务后需要进行简单维护; 还投入了10辆M1A2“艾布拉姆斯”主战坦克, 所有坦克均处于正常战备状态。
Prompt	你是一个nlp专家, 抽取以下内容中的实体, 实体类型包括人员、组织机构、武器装备、时间、地点, 其中文本中关于武器装备对应的状态和数量抽取后以字典格式进行组织, 装备状态为词语, 内容不要过长, 同一个武器装备不同状态分别罗列输出, 请将输出严格按照json格式组织, 其他无关信息不要输出, 如果有其他实体类型请抽取出来, 没有的实体请留空。
DeepSeek671B	{ "人员": [{"姓名": "约翰·米勒", "职务": "队长"}, {"描述": "30人特种作战小组"}], "组织机构": ["美国五角大楼", "美军中央司令部"], "武器装备": [{"名称": "AH-64“阿帕奇”武装直升机", "数量": 3, "状态": "良好"}, {"名称": "AH-64“阿帕奇”武装直升机", "数量": 2, "状态": "需维护"}], {"名称": "M1A2“艾布拉姆斯”主战坦克", "数量": 10, "状态": "正常战备"}], "时间": "2024年12月5日", "地点": ["中东地区", "伊拉克巴格达市附近"], "其他实体": [] }
本文(Qwen2.5B)	{ "人员": ["约翰·米勒"], "组织机构": ["美国五角大楼"], "武器装备": ["阿帕奇**3**状态良好", "阿帕奇**2**需要维护", "艾布拉姆斯**10**正常战备状态"], "时间": ["2024年12月5日"], "地点": ["中东地区", "伊拉克巴格达市"] }
本文(DeepSeek7B)	{ "人员": ["约翰·米勒"], "组织机构": [], "武器装备": ["AH-64阿帕奇武装直升机**3**良好", "AH-64阿帕奇武装直升机**2**需要简单维护", "M1A2艾布拉姆斯主战坦克**10**正常战备状态"], "时间": ["2024年12月5日"], "地点": ["伊拉克巴格达市附近"] }
本文(DeepSeek8B)	{ "人员": ["约翰·米勒"], "组织机构": ["美国五角大楼", "美军中央司令部"], "武器装备": ["AH-64“阿帕奇”武装直升机**3**良好", "AH-64“阿帕奇”武装直升机**2**简单维护", "M1A2“艾布拉姆斯”主战坦克**10**正常战备"], "时间": ["2024年12月5日"], "地点": ["中东地区", "伊拉克巴格达市附近"] }

为用户开展多源数据知识抽取与分析工作提供全面且专业的辅助。

参考文献

- [1] LAMPLE G, BALLESTEROS M, SUBRAMANIAN S, *et al.* Neural architectures for named entity recognition[C]. The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego, USA, 2016: 260–270. doi: 10.18653/v1/N16-1030.
- [2] DAI Zhenjin, WANG Xutao, NI Pin, *et al.* Named entity recognition using BERT BiLSTM CRF for Chinese electronic health records[C]. The 2019 12th International Congress on Image and Signal Processing, Biomedical Engineering and Informatics (CISP-BMEI), Suzhou, China, 2019: 1–5. doi: 10.1109/CISP-BMEI48845.2019.8965823.
- [3] WANG Chenguang, LIU Xiao, CHEN Zui, *et al.* Zero-shot information extraction as a unified text-to-triple translation[C]. The 2021 Conference on Empirical Methods in Natural Language Processing, Punta Cana, Dominican Republic, 2021: 1225–1238. doi: 10.18653/v1/2021.emnlp-main.94.
- [4] YANG Qingping, HU Yingpeng, CAO Rongyu, *et al.* Zero-shot key information extraction from mixed-style tables: Pre-training on Wikipedia[C]. The 2021 IEEE International Conference on Data Mining (ICDM), Auckland, New Zealand, 2021: 1451–1456. doi: 10.1109/icdm51629.2021.00187.
- [5] LU Yaojie, LIU Qing, DAI Dai, *et al.* Unified structure generation for universal information extraction[C]. The 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Dublin, Ireland, 2022: 5755–5772. doi: 10.18653/v1/2022.acl-long.395.
- [6] GUO Daya, YANG Dejian, ZHANG Haowei, *et al.* DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning[EB/OL]. <https://arxiv.org/abs/2501.12948>, 2025.
- [7] LIU Aixin, FENG Bei, XUE Bing, *et al.* Deepseek-v3 technical report[EB/OL]. <https://arxiv.org/abs/2412.19437>, 2024.
- [8] BI Xiao, CHEN Deli, CHEN Guanting, *et al.* DeepSeek LLM: scaling open-source language models with longtermism[EB/OL]. <https://arxiv.org/abs/2401.02954>, 2024.
- [9] YUAN Jingyang, GAO Huazuo, DAI Damai, *et al.* Native sparse attention: hardware-aligned and natively trainable sparse attention[C]. The 63rd Annual Meeting of the Association for Computational Linguistics, Vienna, Austria, 2025: 23078–23097. doi: 10.18653/v1/2025.acl-long.1126.
- [10] WANG Xiao, ZHOU Weikang, ZU Can, *et al.* InstructUIE: multi-task instruction tuning for unified information extraction[EB/OL]. <https://arxiv.org/abs/2304.08085>, 2023.
- [11] HU Danqing, LIU Bing, ZHU Xiaofeng, *et al.* Zero-shot information extraction from radiological reports using ChatGPT[J]. *International Journal of Medical Informatics*, 2024, 183: 105321. doi: 10.1016/j.ijmedinf.2023.105321.
- [12] KARTCHNER D, RAMALINGAM S, AL-HUSSAINI I, *et al.* Zero-shot information extraction for clinical meta-analysis using large language models[C]. The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks, Toronto, Canada, 2023: 396–405. doi: 10.18653/v1/2023.bionlp-1.37.
- [13] 张国宾, 姬红兵, 王佳萌, 等. 基于通用信息抽取大模型的特定

- 领域文本实体关系抽取研究[J]. 中国信息界, 2024(8): 159–161.
- ZHANG Guobin, JI Hongbing, WANG Jiameng, *et al.* Research on entity-relationship extraction from domain-specific texts leveraging generalized information extraction large models[J]. *Information China*, 2024(8): 159–161.
- [14] 皮乾坤, 卢记仓, 祝涛杰, 等. 一种基于大语言模型增强的零样本知识抽取方法[J/OL]. 计算机科学. <https://link.cnki.net/urlid/50.1075.TP.20250123.1638.012>, 2025.
- PI Qiankun, LU Jicang, ZHU Taojie, *et al.* A zero-shot knowledge extraction method based on large language model enhanced[J/OL]. *Computer Science*. <https://link.cnki.net/urlid/50.1075.TP.20250123.1638.012>, 2025.
- [15] 户才顺. 基于大语言模型的审计领域命名实体识别算法研究[J]. 计算机科学, 2025, 52(S1): 72–75.
- LU Caishun. Study on named entity recognition algorithms in audit domain based on large language models[J]. *Computer Science*, 2025, 52(S1): 72–75.
- [16] 胡慧云, 葛杨, 崔凌潇, 等. 融合多模态信息与大语言模型的生成式命名实体识别方法[J/OL]. 计算机工程与应用. <https://doi.org/10.3778/j.issn.1002-8331.2503-0243>, 2025.
- HU Huiyun, GE Yang, CUI Lingxiao, *et al.* Generative named entity recognition method integrating multimodal information and large language models[J/OL]. *Computer Engineering and Applications*. <https://doi.org/10.3778/j.issn.1002-8331.2503-0243>, 2025.
- [17] HU E J, SHEN Y, WALLIS P, *et al.* LoRA: low-rank adaptation of large language models[EB/OL]. <https://arxiv.org/abs/2106.09685>, 2021.
- [18] LEWIS P, PEREZ E, PIKTUS A, *et al.* Retrieval-augmented generation for knowledge-intensive NLP tasks[C]. The 34th International Conference on Neural Information Processing Systems, Vancouver, Canada, 2020: 793. doi: [10.5555/3495724.3496517](https://doi.org/10.5555/3495724.3496517).
- 李永斌: 男, 工程师, 研究方向为数据处理分析挖掘.
刘 棟: 男, 工程师, 研究方向为数据处理分析挖掘.
郑 杰: 男, 工程师, 研究方向为自然语言处理.
- 责任编辑: 马秀强

A Method for Named Entity Recognition in Military Intelligence Domain Using Large Language Models

LI Yongbin^① LIU Lian^① ZHENG Jie^②

^①(Political and Legal Committee of CMC, Beijing 100010, China)

^②(Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100080, China)

Abstract:

Objective Named Entity Recognition (NER) is a fundamental task in information extraction within specialized domains, particularly military intelligence. It plays a critical role in situation assessment, threat analysis, and decision support. However, conventional NER models face major challenges. First, the scarcity of high-quality annotated data in the military intelligence domain is a persistent limitation. Due to the sensitivity and confidentiality of military information, acquiring large-scale, accurately labeled datasets is extremely difficult, which severely restricts the training performance and generalization ability of supervised learning-based NER models. Second, military intelligence requires handling complex and diverse information extraction tasks. The entities to be recognized often possess domain-specific meanings, ambiguous boundaries, and complex relationships, making it difficult for traditional models with fixed architectures to adapt flexibly to such complexity or achieve accurate extraction. This study aims to address these limitations by developing a more effective NER method tailored to the military intelligence domain, leveraging Large Language Models (LLMs) to enhance recognition accuracy and efficiency in this specialized field.

Methods To achieve the above objective, this study focuses on the military intelligence domain and proposes a NER method based on LLMs. The central concept is to harness the strong semantic reasoning capabilities of LLMs, which enable deep contextual understanding of military texts, accurate interpretation of complex domain-specific extraction requirements, and autonomous execution of extraction tasks without heavy reliance on large annotated datasets. To ensure that general-purpose LLMs can rapidly adapt to the specialized needs of military intelligence, two key strategies are employed. First, instruction fine-tuning is applied. Domain-specific instruction datasets are constructed to include diverse entity types, extraction rules, and representative

examples relevant to military intelligence. Through fine-tuning with these datasets, the LLMs acquire a more precise understanding of the characteristics and requirements of NER in this field, thereby improving their ability to follow targeted extraction instructions. Second, Retrieval-Augmented Generation (RAG) is incorporated. A domain knowledge base is developed containing expert knowledge such as entity dictionaries, military terminology, and historical extraction cases. During the NER process, the LLM retrieves relevant knowledge from this base in real time to support entity recognition. This strategy compensates for the limited domain-specific knowledge of general LLMs and enhances recognition accuracy, particularly for rare or complex entities.

Results and Discussions Experimental results indicate that the proposed LLM-based NER method, which integrates instruction fine-tuning and RAG, achieves strong performance in military intelligence NER tasks. Compared with conventional NER models, it demonstrates higher precision, recall, and F1-score, particularly in recognizing complex entities and managing scenarios with limited annotated data. The effectiveness of this method arises from several key factors. The powerful semantic reasoning capability of LLMs enables a deeper understanding of contextual nuances and ambiguous expressions in military texts, thereby reducing missed and false recognitions commonly caused by rigid pattern-matching approaches. Instruction fine-tuning allows the model to better align with domain-specific extraction requirements, ensuring that the recognition results correspond more closely to the practical needs of military intelligence analysis. Furthermore, the incorporation of RAG provides real-time access to domain expert knowledge, markedly enhancing the recognition of entities that are highly specialized or morphologically variable within military contexts. This integration effectively mitigates the limitations of traditional models that lack sufficient domain knowledge.

Conclusions This study proposes a LLM-based NER method for the military intelligence domain, effectively addressing the challenges of limited annotated data and complex extraction requirements encountered by traditional models. By combining instruction fine-tuning and RAG, general-purpose LLMs can be rapidly adapted to the specialized demands of military intelligence, enabling the construction of an efficient domain-specific expert system at relatively low cost. The proposed method provides an effective and scalable solution for NER tasks in military intelligence scenarios, enhancing both the efficiency and accuracy of information extraction in this field. It offers not only practical value for military intelligence analysis and decision support but also methodological insight for NER research in other specialized domains facing similar data and complexity constraints, such as aerospace and national security. Future research will focus on optimizing instruction fine-tuning strategies, expanding the domain knowledge base, and reducing computational cost to further improve model performance and applicability.

Key words: Large Language Models (LLMs); Named Entity Recognition (NER); Instruction fine-tuning; Retrieval-augmented generation; Knowledge base