· 数据智能与知识服务 ·

基于大语言模型的北洋政府文书资源命名实体识别研究

邓 君* 张子姝 潘禹兵 叶东宇 常严予
（吉林大学商学与管理学院，吉林 长春 130012）

摘 要：［目的/意义］针对北洋政府文书资源因语言复杂性、多样性及标注数据缺乏导致的命名实体识别难题，本文提出一种适应低资源场景基于大语言模型的命名实体识别框架，为近代历史文献的结构化挖掘与知识重组提供方法支撑。［方法/过程］该框架融合检索增强生成与高效参数微调，利用Faiss向量检索构建上下文样例动态选取机制，通过LoRA策略对大语言模型进行领域知识注入。最后，在自建语料库上，系统评估深度学习基准模型与不同采样策略下的大语言模型性能。［结果/结论］结果表明，结合相似度样例选择与LoRA微调的Qwen3-4B模型效果最优，总体F1值达0.857，实现对北洋政府文书的精准实体识别，验证了大模型在低资源历史文书处理中的实用性与可扩展性。

关键词：北洋政府文书资源；大语言模型；命名实体识别；低资源场景；检索增强生成；LoRA微调
DOI：10.3969/j.issn.1008-0821.2026.03.004
〔中图分类号〕G353.1 〔文献标识码〕A 〔文章编号〕1008-0821（2026）03-0044-12

Named Entity Recognition in Beiyang Government Documents Resources Using Large Language Models

Deng Jun* Zhang Zishu Pan Yubing Ye Dongyu Chang Yanyu
（School of Business and Management，Jilin University，Changchun 130012，China）

Abstract：［Purpose/Significance］Addressing the challenges in named entity recognition（NER）for Beiyang Government Document Resources due to linguistic complexity，diversity，and lack of annotation data，this paper proposes a large language model-based NER framework adapted for low-resource scenarios. This framework provides methodological support for structured mining and knowledge reorganization of modern historical documents.［Methods/Process］This framework integrated retrieval-enhanced generation with efficient parameter fine-tuning. It used Faiss vector retrieval to build a dynamic context example selection method and used the LoRA strategy to add domain knowledge to large language models. On a custom corpus，the study designed seven special entity types，including persons，places，organizations，time，positions，events，and document types. The study then compared two deep learning entity recognition methods，BERT-BiLSTM-CRF and RoBERTa-BiLSTM-CRF，with Baichuan-4B，DcepSeck-R1，Xunzi-Qwen3-8B，Qwen3-4B，Llama，and GPT-4. The study evaluated large language models performance under different sampling methods.［Result/Conclusion］Experiments demonstrate that compared to traditional deep learning models and general-purpose large language models，the synergistic paradigm integrating LoRA fine-tuning with RAG significantly enhances entity recognition perfor-

收稿日期：2026-01-09
基金项目：国家社会科学基金重点项目“国家文化数字化战略下档案数据资源挖掘与智慧服务研究”（项目编号：23ATQ001）
作者简介：张子姝（1995-），女，博士研究生，研究方向：数字人文与知识服务。潘禹兵（2003-），女，硕士研究生，研究方向：数字人文与知识服务。叶东宇（1996-），男，博士研究生，研究方向：自然语言处理。常严予（1998-），女，博士研究生，研究方向：数字人文与知识服务。
通信作者：邓君（1977-），女，教授、博士，博士生导师，研究方向：数字信息资源管理，数字人文与知识服务，档案管理与应用。

mance，achieving an overall F1 score of 0. 857. A framework that uses RAG with large，fine-tuned language models for named entity recognition in Beiyang Government Document Resources works well together，and it achieves accurate entity identification in these historical records. This shows that large language models are practical and can be scaled when processing historical documents with limited resources.

Keywords：beiyang government document resources；large language model；named entity recognition；low-resource scenarios；retrieval augmented generation；LoRA fine-tuning

北洋政府（1912—1928 年）文书资源系统详实记载了该历史阶段政府机构的设置沿革、法律法规的颁行实施以及公共事务管理的具体运作流程，其既具备行政凭证特有的原始性与权威性，又承载着历史文献的史料保存与学术研究价值。近年来，《数字中国建设 2025 年行动方案》[1] 等一系列国家级政策相继出台，明确提出要“深度挖掘人工智能应用场景”，为北洋政府文书资源的智能化整理与知识重组提供了制度支持与技术场景的双重驱动力。在此背景下，依托人工智能技术对北洋政府文书资源中蕴含的语义信息进行系统性挖掘，已成为重构北洋时期历史知识叙事的重要技术路径。这一研究范式的转型，不仅直观彰显现代信息技术赋能人文研究的学术革新，也反映出人文学科研究范式正逐步从传统文本解读分析向数据驱动型研究的转变。

虽然，北洋政府文书资源是近代中国制度转型与治理实践的第一手记录，但由于该时期机构称谓频繁更迭、用语文白夹杂 [2]，更因其处于近代汉语转型期，缺乏统一的实体表达规范与稳定的命名体系，导致其实体歧义性较强。与此同时，高质量标注数据稀缺，监督学习方法难以获得足够的训练支持，传统命名实体识别（Named Entity Recognition，NER）方法无法充分适应复杂的语言变异与语境依赖，导致实体识别效果有限，直接制约了北洋政府文书资源的结构化整理与深度内容挖掘。因此，本研究利用大语言模型，引入 Faiss 向量检索 [3] 构建动态样例索引库，结合检索增强生成（Retrieval-Augmented Generation，RAG）[4] 与参数高效微调方法中的低秩适应（Low-Rank Adaptation，LoRA）微调策略 [5]，构建融合领域知识的实体识别框架，实现大语言模型在民国时期文献资源命名实体识别的有效应用，为同类低资源、高复杂度历史文献语义深度挖掘提供方法参考。

1 相关研究

1.1 命名实体识别

命名实体识别作为自然语言处理领域一项核心任务，旨在从文本中自动识别具有特定类型的实体信息 [6]，其技术演进呈现出从依赖人工规则向数据驱动、从通用领域向垂直领域不断深化的趋势。早期研究主要依赖人工构建的词典与规则匹配 [7]，随之出现的统计机器学习方法 [8] 提升了命名实体识别的准确性。近年来，深度学习技术凭借其强大的上下文表征能力，跃升为 NER 领域的主流研究范式，特别是在应对一词多义、结构嵌套及语义模糊等复杂问题时优势显著。典型的 BiLSTM-CRF 等序列标注模型通过标签依赖与上下文特征，在多个基准测试中取得进展 [9]。而针对嵌套实体的识别，基于全局指针的跨度识别方法也展现出更优越的结构适配性 [10]。

大语言模型的出现延伸并拓展了深度学习的边界，凭借极强的语义理解和上下文学习能力，重塑了命名实体识别的研究范式 [11]，使其能在低资源甚至零样本场景下实现实体的有效抽取。现有研究表明，借助高效的提示设计 [12] 或结合检索增强生成技术 [13]，大语言模型能有效弥补领域数据匮乏的短板，并在古籍 [14]、红色档案 [15] 及非遗文献 [16] 等场景中展现出广阔的应用前景。然而，在面对术语密集或语言形态独特的垂直领域时，大语言模型因缺乏特定领域知识，容易出现幻觉或识别边界模糊的情况。为此，研究人员多采用领域数据持续训练或参数高效微调 [17] 的方式，实现领域知识的模型内化 [18]，进而提升识别准确性。由此可见，提示工程 [19]、检索增强 [20] 配合轻量化微调 [21]，是目前增强大语言模型领域任务性能的主流手段。

聚焦历史文献这类高度专业的垂直领域，命名实体识别的应用正从通用语料向领域适配及低资源场景转移。已有学者尝试利用 SikuBERT 等领

域预训练模型[22]、优化序列标注架构[23]，或将提示学习与抽取式阅读理解相融合[24]，在古籍实体识别任务中获得了优于通用模型的表现。这充分说明，将领域知识有效融入并改进模型结构，是改善历史文献命名实体识别效果的可行之策。

1.2 民国文书资源数字化开发

在数字人文与信息技术相互渗透的推动下，民国文书资源数字化开发正展现出跨学科特征与多层次发展态势。回顾现有研究不难发现，该领域聚焦数字化流程重构、资源整合与智能技术应用，致力于消解民国文书资源数字化中资源碎片化、标准缺位与技术适配偏差等系统性桎梏。为实现文书资源从数字化保存向知识化挖掘的转型，研究尝试在确立通用元数据规范[25]的基础上，引入文本挖掘[26]、语义标注[27]、社会网络分析[28]及GIS可视化[29]等多元分析工具。徽州文书[30]及高校馆藏档案[31]的开发案例印证了技术介入的效力，OCR增强处理[32]、众包加工[33]与关联数据[34]的应用，显著提升了文本转化质效与多模态关联深度。数字人文平台的搭建[35]，使得民国文书资源价值底蕴在地方史、文化传承等领域得以深层释放，并借由新媒体语境实现创新性转化。这些探索不仅为民国文书资源的可持续开发提供了范例，也折射出民国文书资源正从单纯的档案留存迈向开放协同的数据生态。尽管该领域正加速迈向智慧服务与跨域融合的阶段，但现有探索多集中于民间文书等语料。针对兼具特定行政属性、严格公文格式及高语义密度的北洋政府文书资源，目前仍缺乏系统性的数字化处理方法论与适配的识别模型。

综合考察现有研究脉络可知，命名实体识别已实现从规则驱动到数据驱动、从通用领域向垂直领域的跨越。在历史文献处理中，领域预训练模型与序列标注方法相结合显著提升古籍文本的实体识别效果。大语言模型凭借卓越的上下文理解与少样本学习能力，为低资源命名实体识别任务开辟了新路径。融合提示工程、检索增强生成与参数高效微调可有效增强大语言模型在专业领域的适应性。然而，现有研究多集中于语料规范、标注充分的古代典籍，对北洋政府文书这类兼具复杂语言特征与极低标注资源的文献，尚未形成系统性命名实体识别方法。因此，本研究构建面向北洋政府文书资源命名实体识别框架，通过检索增强机制，利用有限标注样本与文书自身规律动态扩充上下文信息，引入LoRA微调，在注入领域知识的同时保持模型语义能力。该框架支持人名、机构、职官、地名、时间等多类实体识别，为北洋政府文书智能处理提供可行的技术路径，为低资源、高复杂度历史文本命名实体识别提供方法参考。

2 基于LoRA微调的北洋政府文书命名实体识别框架

针对北洋政府文书实体识别任务中因标注资源匮乏与文白夹杂带来的识别效能制约，本研究通过引入基于Faiss向量检索的上下文样例动态选取机制，利用RAG增强模型对历史语义的理解，进一步结合LoRA轻量化微调策略，将领域知识高效注入大语言模型，构建基于相似度计算的北洋政府文书命名实体识别框架，如图1所示。

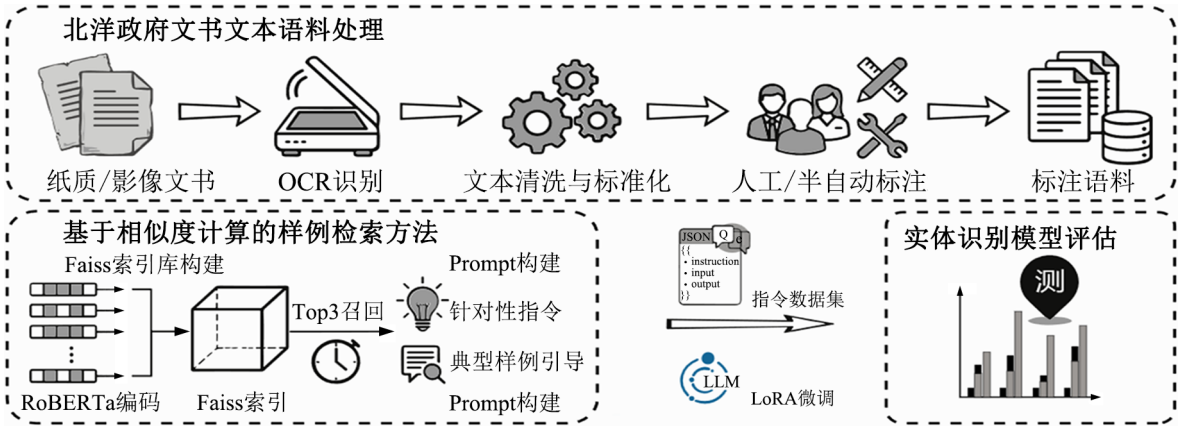

图1 基于LoRA微调的北洋政府文书资源命名实体识别框架

Fig. 1 Named Entity Recognition Framework for Beiyang Government Document Resources Based on LoRA Fine-Tuning

该框架涵盖北洋政府文书文本语料处理、基于相似度计算的样例检索方法与实体识别模型评估3个核心模块，旨在实现北洋政府文书资源结构化挖掘与知识重组，为数字人文领域低资源历史文献智慧化处理提供可复用技术路径，推动相关研究从传统方法向数据驱动范式转型。

2.1 北洋政府文书文本语料处理

此模块主要完成北洋政府文书资源文本的数字化与结构化处理。对纸质或影像形态的北洋政府文书进行光学字符识别，将其转化为可供计算的机器文本。随后对识别后的文本进行清洗与格式标准化处理，并在此基础上组织高质量的人工或半自动标注，形成可供模型训练与评估的标注语料，为后续的检索与识别提供基础数据支撑。

2.2 基于相似度计算的样例检索方法

2.2.1 Faiss索引库构建

北洋政府文书资源兼具文白杂糅、同职异称、机构迭变和标注稀缺四重特征，传统基于词汇重叠或TF-IDF的静态样例库易导致字面相近的误召回，使得下游生成式模型幻觉加剧。针对这一挑战，本研究采用Faiss向量检索引擎，该技术由FaceBookAI团队研发，对10亿量级的索引可以做到毫秒级检索，并支持在高维空间中进行相似性搜索。相较于传统方法，Faiss通过将语料语义表征与北洋政府文书知识耦合，构建与北洋政府文书特点深度适配的动态索引机制，显著提升了模型对民国特有复杂语义关系的捕捉能力。

本研究采用Faiss构建动态索引库，将RoBERTa编码的候选向量集封装为索引，实现毫秒级Top-3召回。向量库构建与相似文本检索如图2所示，向量存储与查询侧共用同一RoBERTa模型。

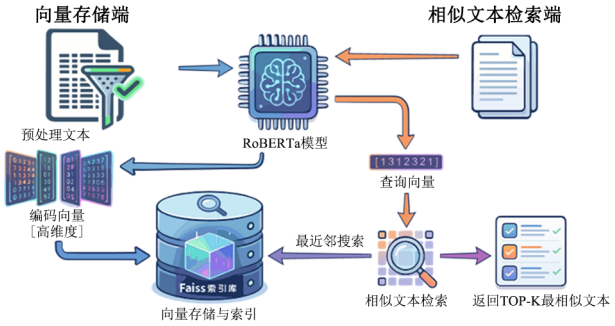

图2 向量库构建与相似文本检索示意图

Fig. 2 Schematic Diagram of Vector Database Construction and Similar Text Retrieval

具体流程分为离线与在线阶段，离线阶段训练集文本经RoBERTa逐句编码后持久化入库；在线查询时，对查询句执行一次前向计算生成查询向量。为抑制北洋文书中同义异写、官职简称等噪声，对查询句与候选池同步执行实体遮罩操作，并以遮罩后的余弦相似度作为检索键值。最终仅取Top-3结果送入Prompt，在压缩上下文长度的同时保留高信息增益，为后续RAG与参数高效微调提供精炼示例支撑。余弦相似度计算如式（1）所示：

$$Similarity = \cos(\theta) = \frac{A*B}{\|A\|\|B\|} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=5}^{n}(A_i)^2} \times \sqrt{\sum_{i=1}^{n}(B_i)^2}} \quad (1)$$

其中，A、B表示不同实体的编码向量。Similarity的结果如果计算值越逼近1，表示两个实体越相似，可对其进行实体融合，采用统一的实体名称。如果计算值越逼近0，表示两个实体相似度极低，则保留其对应的实体。

2.2.2 Prompt构建

在北洋政府文书资源的命名实体识别任务中，Prompt构建是实现领域知识显性化与语义关联结构化的重要路径。其核心目标在于借助针对性的指令设计与典型样例引导，将北洋时期文书所蕴含的领域专属知识，转化为模型可感知、可学习的语义关联规则，为实体识别任务提供知识支撑与语义引导的双重支撑。完整的Prompt设计示例如图3所示。

首先，北洋政府文书资源承载着特定历史背景下的政治、军事及行政等多维度信息，实体识别工作需深度结合该时期的制度环境与历史发展脉络。Prompt借助结构化指令设计与代表性样例选取，使原本内隐的领域知识得以显性表达，助力模型理解并捕捉实体与上下文语境间的深层语义关联。其次，北洋政府文书资源语义结构复杂，常涉及职务与机构之间的层级关系、事件与时间之间的逻辑关联等。Prompt通过明确定义实体类型与文本语义之间的对应规则，并辅以典型样例示范，为模型建立语义关联的识别参照范式，引导模型在复杂语境中精准判定实体类型（如将“豫南总司令官”正确归类为“职务”实体，将“白朗军围攻”

图3 Prompt完整示例

Fig. 3 Complete Example of Prompt

识别为“事件”实体）。此外，Prompt构建过程也需兼顾知识表达的可解释性要求。除通过指令明确定义实体类型与语义关联规则外，样例的选取尤其重要。本研究采用Faiss检索方法，筛选出与目标文本语义相似度排名前3的文本作为引导样例，以此拓宽语义关联的覆盖范围、增强样例的领域代表性，进而提升Prompt在北洋政府文书资源实体识别任务中的泛化性能与可解释性。

2.3 大语言模型微调

大语言模型微调环节的核心在于结合上述Prompt指令构建指令数据集。本研究引入LoRA微调策略，有效避免模型在面向北洋政府文书的特异性命名实体识别任务中出现灾难性遗忘问题，进而实现模型对民国时期特有职官称谓、机构名称等历史实体的精准识别。北洋政府文书语料具有历时跨度大、行政行话密集且实体关系隐含等属性，而LoRA微调能够保留大语言模型的通用语言理解能力，显著缓解标注稀缺带来的过拟合风险，并支持多轮迭代。

本文任务的形式化定义如式（2）所示：

$$MAX\ P_m\left(\rho^*\gamma\left(\delta(S_1,S_2,\cdots,S_n),X\right)\right) \tag{2}$$

其中，$S_1,S_2,\cdots,S_n$为存储在向量库里的候选文本；δ表示相似样例选择函数，通过该函数从候选数据中选出TOP-3相似数据；X是需要进行实体识别的文本；γ为Prompt构建函数；ρ表示标签数据。目标是通过优化样例组织方式来最大化标签输出的概率。

2.4 评价指标

本文使用命名实体识别领域常用的精确率P（Precision）、召回率R（Recall）及F1值（Fl-score）作为本文的评价指标，如式（3）~（5）所示：

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

$$R=\frac{TP}{TP+FN} \tag{4}$$

$$F1=\frac{2\times P\times R}{P+R} \tag{5}$$

其中，TP代表模型识别正确的实体数量，$TP+$

*FP*代表模型识别出的实体数量。两者比值*P*为模型成功识别样本的正确率。*TP* + *FN*表示文本实际包含实体数量，*R*为模型识别实体占总实体比例。*F*1是两者调和平均值，是精确率与召回率平均数的计算方式。

3 实验设计与分析

3.1 数据集

为确保数据来源的权威性、数据构建的严谨性，本文选取中国第二历史档案馆权威出版物《中华民国史档案资料汇编·第3辑》[36]作为基准语料池。该汇编按政治、军事、外交、财政、经济、文化教育及民众运动等专题分册编排，核心收录馆藏北洋政府各部院档案，并辅之少量同期《政府公报》文件，具备时段闭合性与主题完备性。研究首先对全书进行高精度扫描形成电子版图片，继而采用OCR文本识别，形成原始语料库，并进行两轮人工精校纠正字符误识、符号冗余与断句错误，同时依据《北洋政府职官年表》[37]《中华民国法规汇编》[38]等权威工具书进行专名标准化与年代归一，形成高质量纯文本语料。经筛查，共收集6 324篇北洋政府文书文本。

本文通过对比分析6 324份原始文本，将北洋政府文书资源中特有实体类别（如公文类型等）列入实体标注范围，设计包含人物、地点、组织机构、时间、职务、事件、公文类型共7类专有实体类别，具体定义及实例如表1所示。

表1 北洋政府文书资源实体类型

Tab. 1 Entity Types in Beiyang Government Document Resources

实体类别	描述	实例
人物	北洋政府文书文本中的人物	袁世凯、段祺瑞、赵秉钧等
地点	北洋政府文书文本中的地理位置	奉天、直隶、热河等
组织机构	北洋政府文书文本中涉及的组织机构	国务院、外交部、内务部等
时间	北洋政府文书文本中涉及的时间	本年五月十九日、1913年、江等
职务	北洋政府文书文本中涉及人物的职务	大总统、河南督军、内阁总理等
事件	北洋政府文书文本中涉及的事件	白朗军起义、五四运动等
公文类型	北洋政府文书文本涉及的公文类型	令（谕、示、公布、状）、呈、咨等

为保证小样本标注可靠性与一致性，首先进行两名信息资源管理博士前置培训，统一标注规范继而采取双重预标注方式，两名标注者对前10篇北洋政府文书文本进行独立标注，经过对比后对分歧实例进行仲裁并固化为例式规则。其次以前述10篇校准文本为模板，采用分层随机抽样策略抽取200篇北洋政府文书进行正式标注。此外，借助标注工具提供的一致性检验功能，得出本次数据标注一致性（Cohen's Kappa系数）均达到0.85以上，确保了标注数据高质量。

本研究标签体系采用序列标注集合{B, I, E, S, O}来识别北洋政府文书实体，最终在17 972条数据中标注10 916个有效实体，形成北洋政府文书标注语料库。为有效进行模型训练与性能验证，按照4∶1比例将标注数据集随机分为训练集和验证集，标注数据集各类实体数量如表2所示。

表2 标注数据集基本统计情况

Tab. 2 Basic Statistics of Annotation Dataset

实体类别	训练集各类实体数量	验证集各类实体数量
人物	1 437	274
地点	1 814	423
组织机构	190	57
时间	2 213	700
职务	1 828	566
事件	438	171
公文类型	627	178

3.2 实验设计

本研究通过大模型平台提供的API接口完成Qwen3-4B[39]模型调用，针对北洋政府文书资源命名实体识别任务开展参数高效微调，微调参数设置如下：输入序列最大长度设为1 024，初始学习

率设为5.0e-05，秩设为16，训练轮次为5。

针对资源受限、成本可控、数据隐私要求高、本地化部署等现实应用场景，本文提出的北洋政府文书资源命名实体识别方法侧重于验证小规模模型在特定领域任务中的潜力，通过耦合LoRA微调与RAG策略，赋予小模型媲美大语言模型的语义理解力。为验证所提方法的稳健性及优化后的轻量化模型在处理复杂北洋文书实体的有效性，本文以大规模大语言模型为上限基准，实验设计采取对照策略。首先在自建语料上复现BERT-BiLSTM-CRF与RoBERTa-BiLSTM-CRF两个深度学习实体识别方法，以验证深度学习方法在低资源场景中的性能上限。其次设计Prompt，引入Llama-3.3-70B-Instruct[40]、GPT4.1[41]、DeepSeek-R1[42]等大语言模型，测试未经过微调的通用模型能力，同时选取Baichuan2-7B-Base[43]、Xunzi-Qwen1.5-4B[44]等经过LoRA微调的模型，与本文提出的结合LoRA微调与RAG机制的Qwen-4B模型进行对比。最后遴选表现最佳模型，通过零样本、随机样例与相似度样例3种策略的对比实验，量化分析不同检索策略在精确率、召回率及F1值上的分布差异，从而证明本文所提方法的有效性。

3.3 基准实验结果与分析

3.3.1 各模型总体命名实体识别结果与分析

北洋政府文书资源的命名实体识别各模型实体识别效果显示，不同技术范式的模型性能呈现明显分层差异，反映该领域文本处理的特殊性与挑战，如表3所示。

表3 各模型总体命名实体识别结果

Tab. 3 Overall Named Entity Recognition Results of Each Model

方法选择	模型	Precision	Recall	F1-score
传统方法	BERT-BiLSTM-CRF	0.529	0.570	0.549
	RoBERTa-BiLSTM-CRF	0.549	0.588	0.568
大语言模型	Llama-3.3-70B-Instruct	0.716	0.848	0.777
	GPT4.1	0.802	0.772	0.787
	DeepSeek-R1	0.532	0.886	0.665
LoRA微调	Baichuan2-7B-Base	0.560	0.554	0.557
	Xunzi-Qwen1.5-4B	0.684	0.725	0.704
结合LoRA微调与RAG	Qwen3-4B-tuned	0.847	0.867	0.857

受限于北洋政府文书资源文白夹杂的语言特质和标注数据稀缺，传统模型难以捕捉深层语义结构及长距离上下文依赖，BERT-BiLSTM-CRF、RoBERTa-BiLSTM-CRF等传统序列标注架构整体效能欠佳，F1值分别为0.549和0.568，精确率与召回率均未突破0.6。

通用大语言模型表现出显著的性能分化。在未针对特定领域微调的情况下，Llama-3.3-70B-Instruct与GPT4.1凭借强大的参数规模与预训练知识，F1值分别达到了0.802与0.787，展现了较强的泛化能力。然而，DeepSeek-R1表现出召回率偏高而精确率不足的不均衡特征，致使其F1值反而不及部分小参数模型。此现象说明，单纯依赖通用大语言模型固然可以识别出更多实体，但若没有特定领域知识的限定，极易产生严重的实体幻觉和错误识别。相比之下，GPT4.1在精确率上表现最佳(0.802)，显示出较强的实体辨别能力，但其召回率相对偏低(0.772)，可能存在部分实体遗漏。

就参数利用效率及领域迁移能力而言，采用LoRA微调的模型效果参差不齐。从表3可以看出，Baichuan2-7B-Base经微调后性能改善甚微(F1为0.557)，与传统方法相当；而针对古籍文献优化的Xunzi-Qwen1.5-4B虽然取得超越传统方法的表现(F1为0.704)，但精确率(0.684)仍有待提高，这表明单靠参数高效微调手段在应对复杂的北洋政府文书资源时仍存在不足之处。

融合LoRA与RAG的方法取得了最优性能。本文提出的Qwen3-4B-tuned模型在仅4B参数规模下，通过结合RAG引入外部知识库的显式约束，有

效解决了生成式模型的幻觉问题。相较于各基线模型，该方法实现了精确率与召回率的双重突破，分别达到0. 847与0. 867，最终F1值高达0. 857。这证实了在低资源场景下，结合轻量化微调与动态知识检索是实现北洋政府文书资源高精度挖掘的可行技术路径。

3. 3. 2 各模型对不同实体类型识别效果分析

图4~图6所示的实体识别实验结果揭示了不同模型在北洋政府文书命名实体识别任务中的显著性能差异与实体类型依赖性。

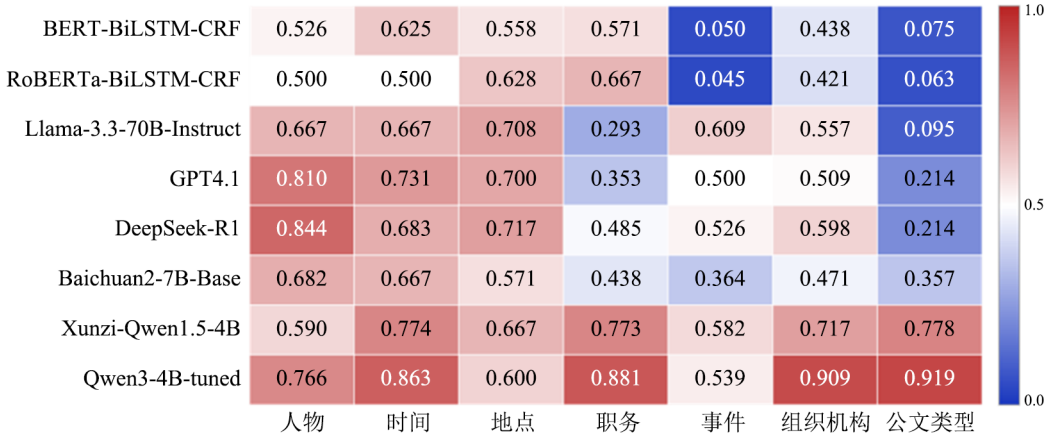

图4 各模型对不同实体类型识别效果精确率

Fig. 4 Precision of Different Models on Various Entity Types

图5 各模型对不同实体类型识别效果召回率

Fig. 5 Recall of Different Models on Various Entity Types

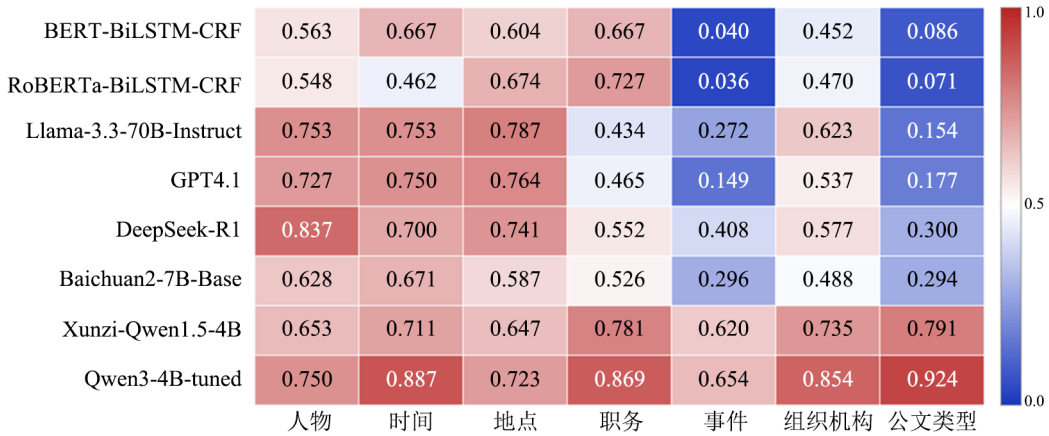

图6 各模型对不同实体类型识别效果F1值

Fig. 6 F1-Scores of Different Models on Various Entity Types

传统深度学习方法在部分高频常规实体上表现尚可，但在复杂语义实体上识别能力严重不足。具体而言，BERT-BiLSTM-CRF与RoBERTa-BiLSTM-CRF在人物、时间、地点三类实体上的F1值介于0.54~0.75之间，显示出一定的识别稳定性。然而，静态标注语料难以应对北洋政府文书资源文白夹杂、同义异写的近代语言变异，尤其在事件、组织机构等复杂语义实体上近乎失效，F1值普遍低于0.1，呈现典型的头部过拟合、尾部零学习困境。BERT-BiLSTM-CRF在职务识别中召回率达0.800，但精确率仅为0.571，表明该模型倾向于过度识别，将非职务词汇误判为实体，反映了传统序列标注方法在深层语义理解上的局限。

通用大语言模型与古籍专用模型展现出不同的优势特征。DeepSeek-R1在人物实体识别上展现强大的零样本能力，F1值达到0.837。然而，该模型在事件和公文类型识别上的F1值分别仅为0.408和0.300，说明其在缺乏领域知识支撑的情况下，对特定类型实体的理解仍显不足。相比之下，经过古籍领域预训练的Xunzi-Qwen1.5-4B在职务和组织机构识别上表现优异，F1值分别为0.781和0.735，显著优于Baichuan2-7B-Base，证明领域先验知识对于提升在特定历史语境下实体识别效果的重要作用。

本文提出的Qwen3-4B-tuned模型以RAG为核心，综合性能最优，验证了RAG与LoRA微调协同机制的有效性。该模型在除人物外的所有类别中均取得了最高F1值。该模型在组织机构识别中提升至0.854，在公文类型识别中更达到0.924的优异水平。对于其他模型识别效果较差的事件类实体，该模型将F1值显著提高至0.654。这一结果有力验证了RAG与LoRA微调协同机制的有效性，LoRA微调使得模型适应北洋政府文书资源的句式结构，而RAG机制通过动态检索外部知识库，为模型提供上下文约束与事实支撑，从而有效抑制生成式模型在低资源实体上的幻觉倾向，实现对复杂、细粒度实体的准确识别。

3.4 样例选择策略实验结果与分析

本节在自建北洋政府文书数据集上，将本文提出的相似度样例选择(Similarity Selcction，SS)与零样本(Zero-Shot，ZS)、随机采样(Random Sample，RS)两种基线置于同一识别框架下进行对照，比较三者对命名实体识别性能的影响。其中，ZS仅依赖任务描述与待抽取文本，无任何示例，RS以均匀分布随机抽取样例构建Prompt，SS则通过遮罩实体后的RoBERTa余弦相似度召回Top3高相关文本。表4实验结果揭示了样例选择策略对北洋政府文书资源命名实体识别性能的影响，证明性能的提升主要源于方案设计的优化，而非仅依赖于大语言模型的底座能力。

表4 样例选择策略实验结果

Tab. 4 Experimental Results of Sample Selection Strategies

样例选择策略	Precision	Recall	F1-score
ZS	0.807	0.117	0.204
RS	0.797	0.849	0.822
SS	0.847	0.867	0.857

ZS方法虽保持较高精确率(0.807)，但召回率骤降至0.117，F1值仅为0.204，表明通用大语言模型在缺乏领域适配时难以识别北洋政府文书资源中的专有实体与特定表述，尽管预测结果可信度尚可，但覆盖率严重不足。RS作为强基准，通过引入域内样例使召回率大幅提升至0.849，F1值达0.822，验证了只需少量标注数据即可实现有效领域迁移。本文提出的SS方法进一步优化，F1值达0.857，较RS提升0.346，且精确率与召回率更趋均衡，体现出语义匹配机制能够筛选与目标文本语境最接近的高质量样例，进而增强模型对历史人物、机构、事件等复杂实体的边界判别与类别判定能力。该对比实验表明，针对标注资源稀缺的历史文献场景，基于相似度的样例检索策略能够使有限标注数据的利用价值最大化，为数字人文领域的低资源场景命名实体识别提供可复用的技术范式。

3.5 消融实验结果与分析

本研究通过消融实验验证所提方法中各组件的有效性。以Qwen3-4B模型为基座模型设置4组对比实验，分别是基础模型Qwen3-4B模型(指令包含任务描述、任务示例和原始文本三部分)，在Qwen3-4B模型的基础上进行LoRA微调的Qwen3-4B-lora模

型，在Qwen3-4B模型的基础上进行RAG的Qwen3-4B-rag模型，在Qwen3-4B模型的基础上加入本文提出的方法，即删除文本中实体的影响，再进行相似度计算的Qwen3-4B-tuned模型。表5揭示了各模块对北洋政府文书命名实体识别性能的贡献程度及协同效应。

表5 消融实验结果

Tab. 5 Results of Ablation Experiment

模型	Precision	Recall	F1-score
Qwen3-4B-tuned	0.847	0.867	0.857
Qwen3-4B-lora	0.687	0.698	0.692
Qwen3-4B-rag	0.643	0.697	0.669
Qwen3-4B	0.413	0.466	0.438

基座模型Qwen3-4B性能低下，F1值为0.438，这是由于通用大语言模型直接应用于北洋政府文书资源时存在显著领域鸿沟，难以捕获民国时期半白半文语境与特定实体特征。单独引入RAG机制后Qwen3-4B-rag的F1值增加至0.669，这是由于相似样例检索为模型提供关键历史知识与标注模式，能有效弥合领域差异。仅采用LoRA微调后Qwen3-4B-1ora的F1值达到0.692，这表明参数高效微调能够良好适配标注数据有限的命名实体识别情况。本文提出的结合LoRA微调与RAG机制的协同增益Qwen3-4B-tuned模型F1值为0.857，性能达到最优，较基线模型提升0.419，且模型精确率与召回率均有较大提升，归因于RAG与LoRA的协同增益，LoRA注入领域知识以优化表征，RAG动态供给上下文一致的标注范例。通过删除实体影响后计算相似度，能够有效避免检索偏差，确保样例质量，充分验证该方法应用于北洋政府文书资源的科学性与有效性。

3.6 案例分析

为直观体现基于相似度计算的样例检索方法作用，本文选取北洋政府文书文本案例进行分析，如图7所示。

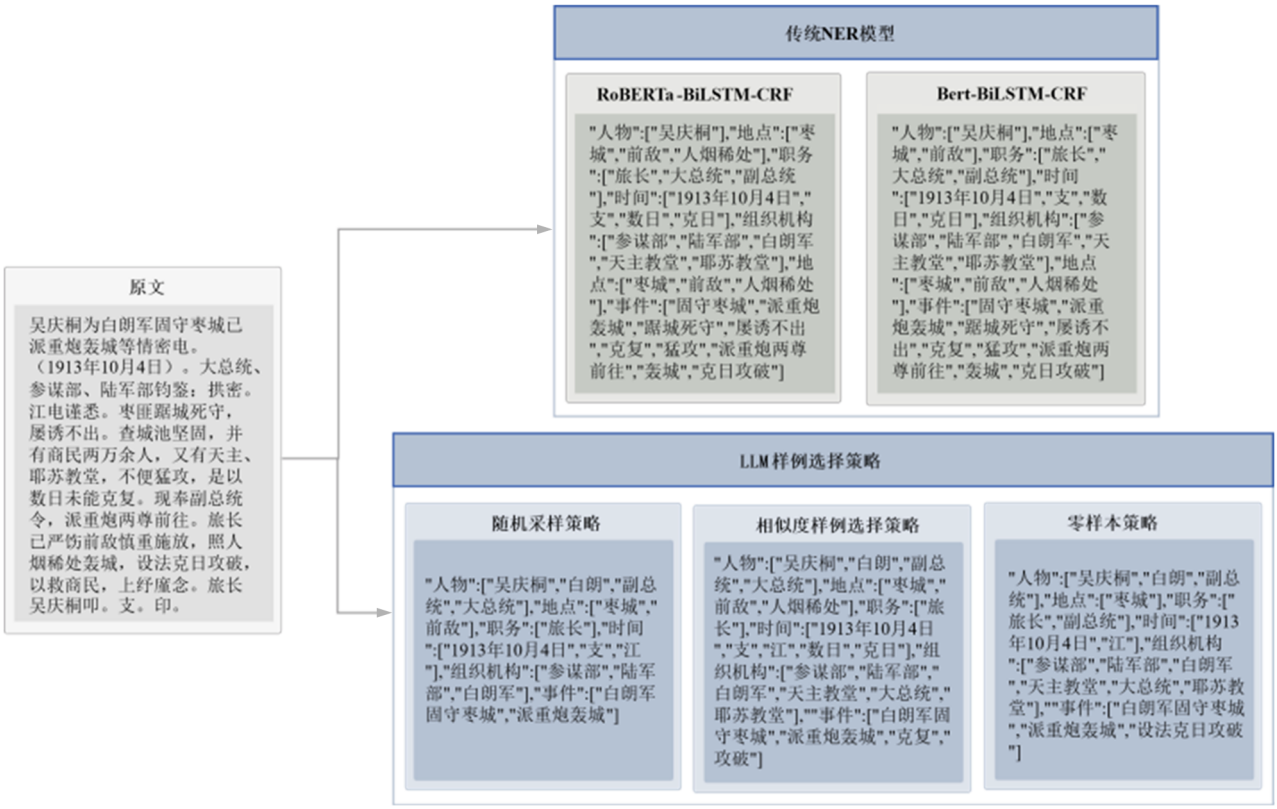

图7 北洋政府文书资源命名实体识别案例

Fig. 7 Case Study of Named Entity Recognition in Beiyang Government Document Resources

传统的监督学习方法完全遗漏“江”这一关键时间实体。“江”是电报韵目代日，指代本月第三日。传统方法严重依赖标注数据，如果训练集中“江”作为时间实体的样本不足，模型便无法识别。而

Zero-shot模型凭借其在海量文本中关于电报和韵目代日相关知识，成功将“江”识别为时间实体，体现其强大的知识迁移和零样本推理能力。传统方法在事件识别中倾向于切分出更短、更具体的短语(如“踞城死守”“屡诱不出”)，大模型零样本则识别出更完整、符合自然语言描述的事件单元(如“白朗军固守枣城”)。这表明大模型对语言的理解更偏向于整体语义，而非局部模式匹配。

在RAG支持下，模型找回Zero-shot遗漏的多个实体，如成功识别出“前敌”“人烟稀处”等战术地点，在“江”的基础上补充“数日”“克日”等相对时间实体，增加“克复”“攻破”等核心军事行动。这得益于Faiss向量库基于去实体RoBERTa向量毫秒级Top3召回，精准锁定索引库中最相似的标注范例，激活大模型对同类实体的泛化能力，显著提高了召回率。但RAG性能高度依赖检索质量，若检索的样例覆盖不精准或覆盖面不足，则提升效果有限。从结果看，RAG虽找回更多实体，但在精细的实体类型划分未达到完美。实验表明，高精度Faiss索引是RAG发挥效用的前提。

进一步经LoRA领域微调后，模型对检索样例的利用效率显著提高。输出“白朗军固守枣城”“派重炮轰城”等事件，表明模型能够保持语义完整，避免过度切分。同时抑制“人烟稀处”“数日”“克日”“天主教堂”等表述，表明模型对领域实体的判断阈值和置信度得到了优化。通过模型参数的调整和实时样例的引导，最终实现模型在北洋政府文书领域精确率、召回率和F1值的最佳平衡。

4 结 论

本研究针对北洋政府文书资源语义密度高、实体歧义性强、标注资源稀缺的特性，引入Faiss向量检索引擎，利用RoBERTa模型将候选文本编码为稠密向量，通过删除实体后计算余弦相似度的去偏策略，构建面向北洋政府文书的动态样例索引库，通过RAG与LoRA微调的双路径融合策略，构建面向北洋政府文书资源的命名实体识别框架，同时设计对比实验，涵盖横向基准评测与纵向消融验证双重维度，一方面与BERT-BiLSTM-CRF、RoBERTa-BiLSTM-CRF传统深度学习模型及GPT4. 1、Qwen3-4B、Llama-3. 3-70B-Instruct、DeepSeek-Rl、Baichuan-4B、Xunzi-Qwen3-8B大语言模型对比，探究最优的命名实体识别方法，另一方面设置零样本、随机采样和相似度选择三组样例检索策略，量化分析检索机制对模型幻觉抑制与边界判别精度的贡献度。该框架实验结果表明，基于相似度计算的样例检索方法结合LoRA微调与RAG机制的协同增益在自建语料库上较基座模型提升0. 419，验证该框架在弥合领域鸿沟、抑制模型幻觉的显著优势。研究通过动态知识检索与静态参数优化的协同增益，将非结构化北洋政府文书资源转化为结构化知识单元，为标注资源稀缺、实体歧义性强的北洋政府文书资源提供了可计算、可复用的智能化整理范式。

参 考 文 献

[1] 中华人民共和国中央人民政府. 《数字中国建设2025年行动方案》近日印发 [EB/OL]. [2026-02-24]. https://www. gov. cn/lianbo/bumen/202505/content_7024041. htm.

[2] 郑爽. 清末民初文言统一对阅读文化嬗变的影响 [J]. 大学图书馆学报，2018，36 (3)：57-64.

[3] Douze M，Guzhva A，Deng C，et al. The Faiss Library [EB/OL]. [2026-02-24]. https://arxiv. org/abs/2401. 08281.

[4] Zhao P，Zhang H，Yu Q，et. al. Retrieval-Augmented Generation for AI-Generated Content：A Survey [EB/OL]. [2026-02-24]. https://arxiv. org/abs/2402. 19473v6.

[5] Hu E J，Shen Y，Wallis P，et al. LoRA：Low-Rank Adaptation of Large Language Models [EB/OL]. [2026-02-24]. https://arxiv. org/abs/2106. 09685.

[6] Li J，Sun A，Han J，et al. A Survey on Deep Learning for Named Entity Recognition [J]. IEEE Transactions on Knowledge and Data Engineering，2020，34 (1)：50-70.

[7] Ehrmann M，Hamdi A，Pontes EL，et al. Named Entity Recognition and Classification in Historical Documents：A Survey [J]. ACM Computing Surveys，2023，56 (2)：1-47.

[8] Hu Z，Hou W，Liu X. Deep Learning for Named Entity Recognition：A Survey [J]. Neural Computing and Applications，2024，36 (16)：8995-9022.

[9] 李纲，潘荣清，毛进，等. 整合BiLSTM-CRF网络和词典资源的中文电子病历实体识别 [J]. 现代情报，2020，40 (4)：3-12，58.

[10] Su J，Murtadha A，Pan S，et al. Global Pointer：Novel Efficient Span-based Approach for Named Entity Recognition [EB/OL]. [2026-02-24]. http://arxiv. org/abs/2208. 03054.

[11] 梁佳，张丽萍，闫盛，等. 基于大语言模型的命名实体识别

研究进展［J］. 计算机科学与探索，2024（10）：2594-2615.

［12］ Wei X，Cui X，Cheng N，et al. ChatIE：Zero-Shot Information Extraction via Chatting with ChatGPT［EB/OL］.［2026-02-24］. https://arxiv. org/abs/2302. 10205.

［13］ 伊豪涵，王昊，周扬，等. 基于RAG-LATS的古籍零样本命名实体识别方法［J］. 数据分析与知识发现，2026，10（1）：116-132.

［14］ 范颜铄，周晓英，王克平，等. 融合GPT技术和用户需求的文学类古籍资源关联数据发布研究——以《聊斋志异·司文郎》为例［J］. 现代情报，2024，44（10）：154-167.

［15］ 杨建梁，王一多，黄美雯，等. 基于大语言模型的红色档案资源交互式知识发现研究——以《南方局党史资料大事记》为例［J］. 图书情报工作，2025，69（15）：112-123.

［16］ 宋雪雁，张祥青，张伟民. 水书习俗非物质文化遗产知识元组织与可视化研究［J］. 现代情报，2023，43（10）：3-15.

［17］ 余池，陈亮，许海云，等. 基于大语言模型的专利命名实体识别方法研究［J］. 数据分析与知识发现，2025，9（6）：47-62.

［18］ Xu D，Chen W，Peng W，et al. Large Language Models for Generative Information Extraction：A Survey［EB/OL］.［2026-01-18］. https://arxiv. org/abs/2312. 17617.

［19］ 张颖怡，章成志，周毅，等. 基于ChatGPT的多视角学术论文实体识别：性能测评与可用性研究［J］. 数据分析与知识发现，2023，7（9）：12-24.

［20］ Liu X，Erkoyuncu J A，Fuh J Y H，et. al. Knowledge Extraction for Additive Manufacturing Process via Named Entity Recognition with LLMs［J］. Robotics and Computer-Integrated Manufacturing，2025（93）：102900.

［21］ 刘耀文，夏一雪，张鹏，等. 国家安全情报战略知识图谱构建与检索增强问答框架研究［J］. 情报杂志，2025，44（7）：165-173.

［22］ 林立涛，王东波，刘江峰，等. 数字人文视域下典籍动物命名实体识别研究——以SikuBERT预训练模型为例［J］. 图书馆论坛，2022，42（10）：42-50.

［23］ 刘忠宝，党建飞，张志剑.《史记》历史事件自动抽取与事理图谱构建研究［J］. 图书情报工作，2020，64（11）：116-124.

［24］ 斯日古楞，林民，郭振东，等. 基于提示学习和抽取式阅读理解的古籍礼仪实体关系联合抽取方法研究［J］. 数据分析与知识发现，2025，9（3）：147-159.

［25］ 梁继红. 走向文本的历史档案数字整理：历史追溯与时代转型（下）［J］. 档案学通讯，2022（1）：60-66.

［26］ 张蓓. 数字人文视野下徽州文书档案开发利用研究［J］. 档案管理，2022（2）：68-70.

［27］ 钟远薪，王蕾，杨新涯，等. 徽州文书文本化语音识别技术应用研究［J］. 图书馆论坛，2023，43（2）：49-56，2.

［28］ 施晓华，王昕. 数字人文社会网络分析方法应用与研究［J］. 图书馆杂志，2020，39（5）：93-99.

［29］ 汤萌，陆星宇. 民间文书中账簿资源元数据模型与空间可视化应用研究［J］. 图书馆杂志，2021，40（12）：62-67.

［30］ 王蕾，薛玉，肖鹏，等. 民间历史文献数字人文图书馆构建——以徽州文书数字人文图书馆实践反思为例［J］. 图书馆论坛，2018，38（3）：30-36.

［31］ 汤萌，赵思渊. 民间文书的数字化建设与资源挖掘——以上海交通大学图书馆馆藏为中心［J］. 档案学通讯，2020（6）：14-21.

［32］ 姜育彦，刘雪立. 数字人文视域下缩微资料的保护与新生——以Digital Cicognara Library为例［J］. 数字图书馆论坛，2022（3）：47-52.

［33］ 徐家成. 众包模式应用于民国档案数字化工作的设想［J］. 兰台世界，2018（11）：35-38.

［34］ 郭佳. 数字人文与人工智能融合视角下民间文书档案数字化流程重构思考［J］. 兰台世界，2025（12）：109-112.

［35］ 陈宇. 古籍数字人文平台对民国档案开放利用的借鉴意义［J］. 档案管理，2021（3）：88-89，91.

［36］ 中国第二历史档案馆. 中华民国史档案资料汇编·第3辑［M］. 南京：凤凰出版社，1991.

［37］ 钱实甫. 北洋政府职官年表［M］. 上海：华东师范大学出版社，1991.

［38］ 立法院编译处. 中华民国法规汇编［M］. 上海：中华书局，1934.

［39］ Qwen Team. Qwen3：Think Deeper，Act Faster［EB/OL］.［2026-02-24］. https://qwenlm. github. io/zh/blog/qwen3/.

［40］ Touvron H，Lavril T，Izacard G，et al. LLaMA：Open and Efficient Foundation Language Models［EB/OL］.［2026-02-24］. https://arxiv. org/abs/2302. 13971.

［41］ Open AI. Introducing GPT-4. 1 in the API［EB/OL］.［2026-02-24］. https://openai. com/index/gpt-4-1/.

［42］ DeepSeek-AI，Guo D，Yang D，et al. DeepSeek-R1：Incentivizing Reasoning Capability in LLMs via Reinforcement Learning［EB/OL］.［2026-02-24］. https://arxiv. org/abs/2501. 12948.

［43］ Yang A，Xiao B，Wang B，et al. Baichuan 2：Open Large-scale Language Models［EB/OL］.［2026-02-24］. https://arxiv. org/abs/2309. 10305.

［44］ 南京农业大学. 荀子大语言模型［EB/OL］.［2026-02-24］. https://xunziallm. njau. edu. cn/.

（责任编辑：李汇森）