基于大语言模型的威胁情报高效抽取与攻击推理方法研究

彭国军[1]，李家琛[1]，杨秀璋[2,3]，吕锦钊[1]

（1.空天信息安全与可信计算教育部重点实验室，武汉大学国家网络安全学院，湖北 武汉 430072；
2.贵州大学公共大数据国家重点实验室，贵州 贵阳 550025；3.贵州大学贵州省大数据产业发展应用研究院，贵州 贵阳 550025）

摘　要：针对开源中文网络威胁情报多为非结构化文本、缺乏高效自动化处理手段，提出一种大语言模型与图推理的方法。首先，利用GPT-4o的少样本学习能力构建高质量提示词，实现数据的自动标注；随后引入低秩适配（LoRA）技术对LLaMA3-8B进行参数高效微调，微调模型在中文实体识别任务中精确率为0.924 7、召回率为0.851 5、F_1值为0.886 6，在关系抽取任务中F_1值为0.837 8；最后，融合图检索增强生成（GraphRAG）与MITRE ATT&CK框架，完成攻击链推理分析。实验结果表明，该方法优于现有方法，具有良好的实用价值。

关键词：网络威胁情报；大语言模型；命名实体识别；关系抽取；检索增强生成

中图分类号：TP309

文献标志码：A

DOI：10.11959/j.issn.1000–436x.2025207

Research on efficient threat intelligence extraction and attack inference method based on large language models

PENG Guojun[1], LI Jiachen[1], YANG Xiuzhang[2,3], LYU Jinzhao[1]

1. Key Laboratory of Aerospace Information Security and Trusted Computing, Ministry of Education, School of Cyber Science and Engineering, Wuhan University, Wuhan 430072, China
2. State Key Laboratory of Public Big Data, Guizhou University, Guiyang 550025, China
3. Guizhou Institute of Big Data Industry Development and Applications, Guizhou University, Guiyang 550025, China

Abstract: To address the fact that open-source Chinese cyber threat intelligence is mostly unstructured text and lacks efficient automated processing methods, a method integration large language model with graph-based reasoning was proposed. First, high-quality prompts were constructed by leveraging the few-shot learning capability of GPT-4o to enable automatic data annotation, Then, parameter-efficient fine-tuning of large language model meta AI (LLaMA3-8B) was conducted with low-rank adaptation (LoRA). The fine-tuned model achieves an F_1 score of 0.886 6 on Chinese entity recognition, and an F_1 score of 0.837 8 on relation extraction. Finally, graph-based retrieval-augmented generation (GraphRAG) was integrated with the adversarial tactics, techniques and common knowledge (MITRE ATT&CK) framework to perform attack chain reasoning and analysis. Experimental results indicate that the proposed method outperforms existing approaches with strong practical value.

Keywords: cyber threat intelligence, large language model, named entity recognition, relation extraction, retrieval-augmented generation

收稿日期：2025–09–12；**修回日期：**2025–11–15

基金项目：国家自然科学基金资助项目（No.62172308, No.62562012）；贵州省科技重大专项计划基金资助项目（黔科合重大专项字[2024]014）；贵州省基础研究（自然科学）基金资助项目（黔科合基础MS〔2025〕686）；贵州省科技支撑计划基金资助项目（黔科合支撑PA[2025]004）；贵州大学引进人才科研基金资助项目（贵大人基合[2024]15号）

Foundation Items: The National Natural Science Foundation of China (No. 62172308, No. 62562012), The Major Scientific and Technological Special Project of Guizhou Province (No.[2024]014), Guizhou Provincial Basic Research Program (Natural Science) (No.MS[2025]686), Guizhou Provincial Key Technology Research and Development Program (No.PA[2025]004), The Research Project for Recruited Talents at Guizhou University (No.[2024]15)

0　引言

随着网络攻击手段不断演进，尤其是具有国家背景的高级持续性攻击日益猖獗，国家与企业安全正面临严重威胁。在这种对抗性极强的背景下，“主动防御、可追溯、可对抗”[1]成为威胁检测的重要策略，网络威胁情报（CTI, cyber threat intelligence）通过记录网络攻击相关信息，助力快速识别和响应威胁，在攻防博弈中起到关键作用。

在威胁情报分析领域，从海量非结构化文本中精确提取预定义的威胁实体和关系是一项重要任务。相较于常规的时间、地点等实体，网络安全领域的专业命名实体更为复杂，识别难度显著增加。尽管近年来基于机器学习和深度学习的方法在此领域取得了进展[2-4]，但传统方法仍存在以下局限与挑战。

1) 传统机器学习模型对复杂语义建模能力不足[5]，难以捕获上下文依赖及潜在语义关系，对复杂句式和跨句实体识别效果不理想。

2) 深度学习方法受限于标注语料稀缺与领域特征不匹配[6]。威胁情报文本通常包含大量技术缩写与专有名词，标注样本有限且分布不均，使深度模型容易过拟合，泛化能力不足。

此外，当面对中文威胁情报时，其抽取难度远高于英文文本，这源于多个方面的挑战：其一，中文语义的歧义性与上下文依赖性更强；其二，中文威胁情报多以非结构化的安全通告、安全公告、安全博客等形式存在，文本中常混杂了中文描述、英文缩写及大量专业术语；其三，中文句式结构无显性边界标记，跨句实体关系抽取更具挑战。这些特性使得依赖深度学习的抽取方法在处理中文数据时效果不尽如人意[7]，难以满足中文场景威胁情报结构化的实际需求。

近年来，大语言模型（LLM, large language model）技术[8]取得显著突破，其凭借强大的文本理解和生成能力，在威胁情报信息抽取任务中展示了其巨大的潜力。然而，LLM本质上是生成模型而非分类模型，且主流LLM主要基于通用英文及跨领域语料预训练，中文安全领域语料覆盖不足。直接将通用LLM应用于中文威胁情报场景时，往往会出现对安全术语理解偏差、攻击步骤识别不完整、实体与关系抽取不稳定等问题。因此，即便在使用 LLM 的前提下，仍然需要面向中文威胁情报构建专门的领域语料，对模型进行有针对性的微调，以增强其对中文攻击描述、混合中英术语以及本地化威胁场景的理解与推理能力。

针对上述背景，本文提出了一种基于LLM的威胁情报高效抽取和攻击推理方法。本文的主要贡献如下。

1) 构建了一套基于GPT-4o的自动化标注机制，充分利用其强大的少样本学习能力与指令理解能力，设计结构化提示词（Prompt）实现高质量威胁情报数据自动标注，显著降低数据构建的人力成本并提升标注效率。

2) 提出一种面向中文场景的威胁情报抽取框架，通过引入低秩适配（LoRA）技术对LLaMA3-8B模型进行参数高效微调，显著提升了在中文威胁情报中实体识别与关系抽取的准确性，突破了现有方法对中文语境理解不足的瓶颈，并构建了系统的实验验证方法的有效性。

3) 融合检索增强生成（RAG, retrieval-augmented generation）技术与ATT&CK（adversarial tactics, techniques and common knowledge）框架，构建面向高级持续性威胁（APT）攻击的推理模型，通过引入图检索增强生成（GraphRAG）实现对攻击链中多跳关系的语义建模，支持复杂场景下的攻击路径识别与意图推演，提升威胁分析的智能化与实用性。

1　背景知识

1.1　威胁情报信息抽取

威胁情报信息抽取技术通过自动化提取文本中的威胁实体（如恶意软件、攻击组织）及其关联关系（如利用、攻击），为构建动态威胁图谱、实现攻击链溯源提供数据基础。威胁情报信息抽取包括实体识别和关系抽取两大任务。

为了更清晰地描述威胁情报信息抽取任务，本文对其进行公式化定义如下。

给定一段威胁情报文本

$$T=\{w_1,w_2,\cdots,w_n\} \tag{1}$$

其中，w_i表示第i个词。

实体识别任务的目标是从中识别出实体集合

$$E=\{(e_i,\mathrm{Type}_i)\}_{i=1}^{N} \tag{2}$$

其中，e_i表示第i个实体的名称，Type_i为其对应的实体类型。

关系抽取任务旨在识别实体对之间存在的语义关系。定义为

$$R=\{(e_h,r,e_t)|e_h,e_t\in E\} \tag{3}$$

其中，(e_h,r,e_t)表示一个关系三元组，e_h表示头实体，e_t表示尾实体，r表示两者之间的关系。

因此，威胁情报信息抽取任务的本质是将输入文本T映射为实体集合与关系集合

$$F:T\rightarrow(E,R) \tag{4}$$

1.1.1 实体识别

实体识别是自然语言处理中的经典任务，旨在从非结构化文本中识别和分类人物、地点、组织等具有特定意义的实体，将其组织为结构化数据，以便后续分析和理解。

早期的方法[9]依赖人工规则和词典，虽在固定模式下效果理想，但缺乏泛化能力。机器学习方法在网络安全实体识别中取得更好表现，避免了手工规则设计。Joshi等[10]使用条件随机场（CRF, conditional random field）模型从网络安全博客中提取实体。Mulwad等[11]用支持向量机来识别攻击方法。但这类方法仍需大量人工特征工程，成本高且难捕捉复杂语义。

随着深度学习的兴起，神经网络在实体抽取任务中表现优秀。Dionísio等[12]使用卷积神经网络（CNN, convolutional neural network）识别包含安全相关信息的推文，并通过双向长短期记忆（BiLSTM, bidirectional long short-term memory）网络模型进一步识别与安全事件相关的命名实体。Yang等[13]将BERT（bidirectional encoder representation from transformer）模型，BiLSTM模型和CRF模型结合，提出一种融合实体识别和实体对齐的APT攻击知识自动抽取方法。然而，深度学习方法依赖大规模标注数据，在少样本或零样本场景下性能骤降。

1.1.2 关系抽取

关系抽取旨在从文本中提取实体之间的关系事实。早期研究[14-16]主要关注同一句话内的实体关系。然而，在实际应用场景中（如分析复杂的攻击链或追踪APT组织），仅识别同一句子内的实体关系已无法满足需求，现如今许多关系需要跨多个句子进行抽取，这使得文档级关系抽取逐渐成为研究重点。

文档级关系抽取方法主要分为两类：基于图的方法和基于Transformer的方法。前者通过构建文档图建模实体之间的结构关系，如Zeng等[17]构建了实体级别文档图，并提出了一种新的路径推理机制来推断实体之间的关系。后者利用预训练模型挖掘跨句语义，如Yuan等[18]应用跨句注意力机制来动态地提取关键信息特征，并设计门控函数将句子级特征与文档级特征相结合。

总之，当前威胁情报信息抽取方法严重依赖大规模标注数据集，但缺乏权威的开源威胁情报标注数据集使得数据标注过程需要投入大量人力资源。此外，中文威胁情报文本由于缺乏明确的分隔符，其信息抽取难度远高于英文文本，导致基于深度学习的方法在处理中文数据时效果往往不尽如人意。

1.2 大语言模型

近年来，得益于Transformer架构[19]的提出、算力提升与大规模训练数据积累，语言模型取得突破性进展，推动了大语言模型的发展。LLM具备数百亿参数，通过训练海量语料，掌握语言结构、语法与上下文等知识。常见的LLM包括GPT系列[20]、LLaMA系列[21]、PaLM[22]等。这些模型不仅能够生成流畅且符合语法规则的文本，还能够处理复杂的语言理解任务，如问答、对话生成以及推理任务。

随着网络攻击日益复杂，传统安全分析难以应对大规模数据与动态威胁，LLM凭借强大理解能力，成为提升安全分析效率与精度的重要手段。Liu等[23]通过定制提示策略，用LLM进行日志分析，性能超越传统方法55.9%。Hu等[24]提出DeGPT框架，借助LLM优化反编译输出，显著提升二进制分析效率。

本文充分发挥GPT模型的少样本学习能力，通过构建高质量的指令模板与示例，实现对威胁情报文本的自动化标注，为确保微调数据集的质量与可靠性，本文在GPT标注的基础上引入人工校验机制，对所有标注样本逐条审核与修正。微调后的模型不仅在特定任务上表现优越，且支持本地部署，有效降低敏感数据泄露风险，提升系统的安全性与可用性。

1.3 RAG技术

大语言模型在许多任务中表现良好，但其在特定领域的应用仍受限于专业知识的缺乏。RAG[25]通过引入外部知识库，在不需要修改模型结构的前

提下，使 LLM 结合预训练知识与实时检索信息，生成更准确可靠的答案。

GraphRAG[26]利用图结构表示知识片段之间的依赖关系，并通过图社区检测，如莱顿（Leiden）算法进行图分层，从而能够在推理过程中捕获多层次的关系和依赖。借助图结构的引导，LLM 在处理复杂关系和多步骤推理任务时表现更优，实现更高效、精确的增强生成。

考虑到威胁情报与 ATT&CK 框架天然具有图结构特征，实体之间存在丰富的关联关系，攻击行为往往呈现多步链路。GraphRAG 能在图上进行邻域扩展和路径检索，将与当前情报相关的前置/后续技术及共用工具等上下文一并召回，用于支持多跳推理和攻击路径重建。因此，本文采用 GraphRAG 技术，深度融合 MITRE ATT&CK 框架，以增强对 APT 攻击报告的理解，识别 APT 攻击链中的多跳语义关联，实现攻击意图推理。

2　方案设计

2.1　总体框架

针对上文提出的传统威胁情报信息抽取方法存在泛化能力不足、深层语义建模困难及中文语料稀缺等挑战，本文提出了一种基于 LLM 的威胁情报高效抽取和攻击推理方法。

本文的核心思想是以大语言模型为语义理解核心，结合自动化标注机制与知识图谱推理框架，实现从非结构化威胁情报文本到结构化攻击知识的高效转换。具体技术路径如下。

1) 自动化标注与小样本学习：利用 GPT-4o 构建自动标注系统，通过少量人工样本引导生成高质量实体与关系标签，显著降低人工成本并提升模型泛化性能。

2) 语义增强与领域自适应：采用 LoRA 技术对 LLaMA3-8B 模型进行安全领域微调，使模型能够理解和抽取威胁情报中的专有术语及隐含语义。

3) 将抽取得到的实体与关系嵌入 GraphRAG 知识图谱结构中，通过图结构的语义聚合与路径关联，能够从多源情报中识别攻击组织的行为模式、技术手段及其潜在攻击路径，从而增强威胁情报的推理能力。

TIE&I-LLM 框架如图 1 所示，主要包括数据采集、数据标注、模型微调和 RAG 推理四大核心模块。该方法以多源异构开源威胁情报为输入，依托大语言模型的深度语义理解能力，实现威胁实体关系抽取与知识图谱构建，并通过融合 ATT&CK 框架的 GraphRAG 技术完成攻击链路的推理研判。

具体而言，首先通过爬虫构建多源威胁情报语

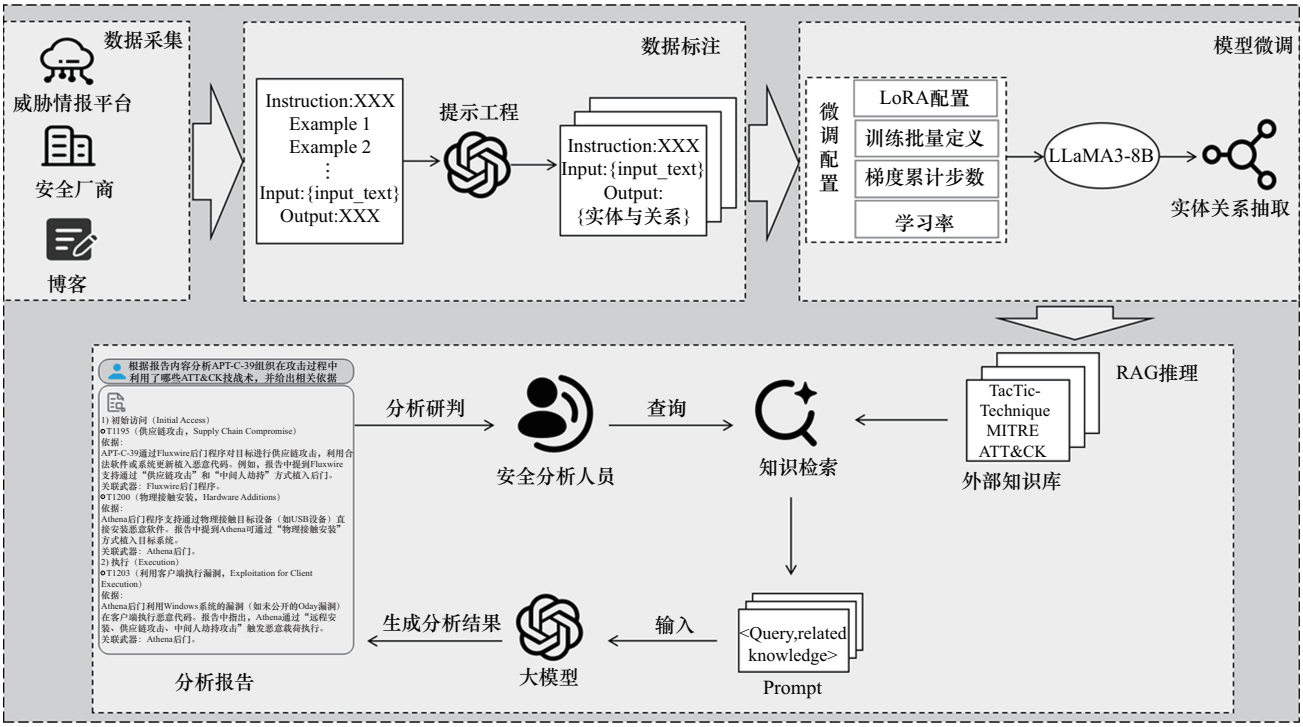


图 1　TIE&I-LLM 框架

料库，并利用GPT-4o模型强大的少样本学习能力设计结构化提示模板，完成威胁实体的细粒度标注，生成高质量的数据集。随后引入LoRA参数高效微调方法对LLaMA3-8B基座模型进行微调，显著提升威胁情报领域命名实体识别和关系抽取的准确率。最后，使用GraphRAG，融合ATT&CK框架对威胁情报进行推理分析，实现攻击技战术识别、威胁归因分析及攻击意图预测等核心功能。

2.2 数据采集

为构建全面可靠的威胁情报库，本文采用多源数据融合策略，系统采集来自安全厂商报告、开源情报平台及技术社区的APT攻击分析文本，涵盖奇安信科技集团、安天科技集团、360安全科技股份有限公司、卡巴斯基等厂商发布的技术报告以及安全博客中的专家评论。多源异构数据的引入有助于从不同视角获取APT攻击信息，提升情报的全面性与准确性。

针对不同平台网站结构差异，本文设计并部署了定制化网络爬虫，自动提取各类报告的统一资源定位符（URL, uniform resource locator）并批量抓取其内容。为保证数据质量，预处理阶段对广告及无关内容进行了清洗过滤，确保保留核心分析信息，为后续威胁情报抽取提供高质量语料支撑。

2.3 数据标注

本文依据结构化威胁信息表达标准2.1（STIX2.1, structured threat information expression2.1），结合APT攻击的特点，定义7类命名实体，包括攻击者（Attacker）、技术（Technique）、工具（Tool）、漏洞（Vulnerability）、恶意文件（File）、行业（Industry）和地区（Region）。详细描述如表1所示。

同时，基于威胁实体间的相互作用，本文定义了7种关系，包括同一（ALIAS_OF）、相关（RELATED_TO）、使用（USE）、针对（TARGET_AT）、位于（LOCATED_AT）、利用（EXPLOIT）和开发（DEVELOP）。详细说明如表2所示。

随后，利用大模型的少样本能力和语言理解能力，从威胁情报中抽取指定实体及其关系。本文遵循“明确性、简洁性、具体化”的Prompt设计原则，构建了一套规范的提示文本。首先，明确模型身份和任务目标；其次，为减弱幻觉现象，在Prompt中加入各类型实体示例，并强调仅提取指定类型，若无相关内容则输出“未识别到相关实体和关系”；最后，统一输出格式。完整Prompt文本如图2所示。

表1 威胁实体详解

实体类别	定义	示例
Attacker	APT攻击的团队名称	Lazarus、APT28
Technique	APT组织攻击的技术和手段	spear phishing、XSS
Tool	恶意软件、合法软件或自主研发的攻击工具	Metasploit、Gh0st
Vulnerability	攻击利用的漏洞编号或名称	CVE-2017-11882
File	攻击利用的具体恶意文件名称	cmdl32.exe、Agent.btz
Industry	攻击的目标行业	金融、教育
Region	攻击者的目标区域或所在区域	Democratic People’s Repubic of Korea、Russia、South Asia

表2 威胁关系详解

关系类别	定义	示例
ALIAS_OF	同一攻击者或工具的不同名称	<Darkhotel、ALIAS_OF、APT-C-06>
RELATED_TO	表明在一定程度上存在关系但是无法确定	<Attacker1、RELATED_TO、Attacker2>
USE	表明两个实体间的利用关系	<Attacker1、USE、Tool>
TARGET_AT	攻击者的目标，可以是个人、组织、地理位置、行业等	<Attacker1、TARGET_AT、Industry>
LOCATED_AT	表明攻击者或实体与特定位置的关系	<DustSquad、LOCATED_AT、Russia>
EXPLOIT	表明攻击者或工具等利用某漏洞实施攻击	<Attacker1、EXPLOIT、Vulnerability1>
DEVELOP	表明攻击者开发工具，恶意文件的关系	<Attacker1、DEVELOP、Tool>

假设你是安全工程师，需要处理APT攻击威胁情报。你的任务是提取出威胁情报文本中的实体和实体间的关系。请理解以下具体要求。

你需要识别的实体类型仅包括：
① **Attacker**（APT攻击的团队名称，例如Lazarus、APT28等）
② **Technique**（APT组织攻击的技术和手段，例如、spear phishing、XSS等）
③ **Tool**（恶意软件、合法软件或自主研发的攻击工具，例如Metasploit、Gh0st等）
④ **Vulnerability**（攻击利用的漏洞编号或名称）
⑤ **File**（攻击利用的具体恶意文件名称，例如cmdl32.exe、Agent.btz等）
⑥ **Industry**（攻击的目标行业，例如金融、教育等）
⑦ **Region**（攻击者所针对的区域或者攻击者所属的区域或国家，例如North Korea、Russia、South Asia等）。
你需要识别的关系类型仅包括：
① **ALIAS_OF**（同一攻击者或工具的不同名称，例如(Darkhotel、ALIAS_OF、APT-C-06)）
② **RELATED_TO**（表明在一定程度上存在关系但是无法确定，例如(Attacker1、RELATED_TO、Attacker2)）
③ **USE**（表明两个实体间的利用关系，例如(Attacker1、USE、Tool)）
④ **TARGET_AT**（攻击者的目标，可以是组织、地理位置、行业等，例如(Attacker1、TARGET_AT)）
⑤ **LOCATED_AT**(表明攻击者或实体与特定位置的关系，例如(DustSquad、LOCATED_AT、Russia)）
⑥ **EXPLOIT**(表明攻击者或工具等利用某漏洞实施攻击，例如(Attacker1、EXPLOIT、Vulnerability1)）
⑦ **DEVELOP**（表明攻击者开发工具，恶意文件的关系，例如(Attacker1、DEVELOP、Tool)）
你将一次性收到很多条以句子为单位的威胁情报文本。收到需要分析的文本，成功识别到实体和关系后对每一句文本输出一条json格式的数据。
如果识别到实体和实体间的关系，输出格式为：{"instruction": "原句", "input": null, "output": "实体:(名称, 实体类型), (名称, 实体类型)……\n关系:(实体1, 关系, 实体2), (实体1, 关系, 实体2)……"}。
如果某一句只识别到实体，实体之间没有上述7种关系，则输出格式为：{"instruction": "原句", "input": null, "output": "实体:(名称, 实体类型), (名称, 实体类型)……\n关系:没有相关关系"}。
如果某一句文本中7种实体都没有出现，则输出{"instruction": "原句", "input": null, "output": "未识别到相关实体和关系"}。

特别注意，只识别所列的7种实体和关系！合并成一个json文件输出。明白后请回复“我已明白任务”并输出7种实体和7种关系。

图 2　完整Prompt文本

本文借助GPT-4o的辅助完成了数据标注任务。由于通用GPT模型在威胁情报这一专业领域存在一定的理解局限，导致部分标注结果存在偏差或不足。为解决这一问题，在GPT-4o自动标注的基础上，进一步进行人工校验与修正，以提升数据集的准确性和可靠性，为后续模型微调提供了更高质量的数据支撑。

2.4　模型微调

为了创建一个针对威胁情报实体关系抽取任务的LLM，本文基于已构建的数据集对LlaMA3-8B模型进行了指令微调。本文选取LLaMA3-8B作为基座模型进行微调主要基于以下考量：其一，LLaMA3由Meta开源且支持商用，相较于部分闭源或授权限制严格的模型，可满足威胁情报数据的本地化部署与隐私保护需求；其二，LLaMA3在预训练阶段强化了多语言理解能力，原生支持中英双语，更适配本文中英混合的威胁情报数据集。

本文采用LoRA微调的方法。LoRA是一种针对大语言模型进行微调的高效方法，传统的微调方法需要修改和存储整个预训练模型的所有参数，而LoRA通过引入低秩矩阵来对权重矩阵进行适应性调整，从而只对部分参数进行修改。LoRA的基本思想是冻结原始矩阵 $\boldsymbol{W}_0 \in R^{m\times n}$，只需更新参数 $\Delta\boldsymbol{W}$，更新过程如式(5)所示，其中秩 $r \ll \min(m,n)$。在LoRA的训练过程中，$\boldsymbol{W}_0$是固定不变的，只有A和B是训练参数。在前向过程中，$\boldsymbol{W}_0$和$\Delta\boldsymbol{W}$都会乘以相同的输入x，最后再相加，如式(6)所示。

$$\boldsymbol{W}_0 + \Delta\boldsymbol{W} = \boldsymbol{W}_0 + BA, B \in R^{m\times r}, A \in R^{r\times n} \tag{5}$$

$$h = \boldsymbol{W}_0 x + \Delta\boldsymbol{W}x = \boldsymbol{W}_0 x + BAx \tag{6}$$

2.5　RAG推理

RAG推理结合了检索（Retriever）与生成（Generation）2种技术，通过将外部知识库中的信息与大模型相结合，提高了推理过程的准确性和信息覆盖范围。本文采用GraphRAG技术，融合ATT&CK框架，为安全分析人员提供一个交互式接口，使安全分析人员可以更好地分析威胁情报推理恶意组织攻击意图。

2.5.1　知识检索

知识检索模块首先接收用户查询（Query），解析并理解其意图后，从外部知识库中检索相关信息。知识库包括MITRE ATT&CK框架攻击技战术和威胁情报数据库。知识检索模块提取查询背景，并结合检索到的攻击信息，为后续推理提供必要的上下文支持。

2.5.2　生成Prompt

完成知识检索后，系统将用户查询与相关背景信息融合，构建最终Prompt。该Prompt不仅包含原始查询，还嵌入攻击技战术描述、安全事件和历史情报等内容，丰富上下文，提升推理准确性。

2.5.3　推理结果生成

生成的Prompt被输入大模型进行推理，模型基于查询与背景信息识别实体和关系，并结合MITRE ATT&CK框架推断潜在攻击路径及意图。最终输出包括攻击技战术、攻击意图与动态分析等结果，辅助安全团队识别与应对威胁。

TIE&I-LLM基于GraphRAG的推理算法如算法1所示。

算法 1　TIE&I-LLM基于GraphRAG的推理算法

输入　APT分析报告Report

输出　实体集合，关系集合，推理文本

1)　文本 = 提取文本（Report）

2)　段落集合 = 分割文本（文本）

3) // 初始化实体集，关系集为空
4) for 段落集中每一段 do
5) 实体 = LLaMA 识别实体
6) 关系 = LLaMA 抽取关系
7) 实体集 = 实体集 ∪ 实体
8) 关系集 = 关系集 ∪ 关系
9) end for
10) 图谱 = 构建图谱（实体集，关系集）
11) Leiden 社区发现（图谱）
12) Query = 输入查询()
13) 检索文本 = 图谱检索（Query）
14) Prompt=生成提示词（Query，检索文本）
15) Response = 大模型生成（Prompt）
16) 返回 Response

3 实验与分析

3.1 实验设置

为构建高质量训练数据集，本文根据前文所述方法采集数据。经过严格的数据清洗和去重，最终收集了1 500条中文和1 500条英文数据，这些数据包含安全博客、APT攻击分析报告和媒体报道3种类型。借鉴文献[27]的研究结论——“高质量小数据集优于低质量大数据集”，本文建立了三重质量把关机制：首先用正则表达式自动清洗格式错误，之后检查内容完整性，最后人工核对标签，确保训练数据集中的标注准确无误。

在数据集制作完成后，将其输入LLaMA3-8B模型进行微调。微调过程在4张RTX 4090显卡上并行处理，充分利用了硬件的计算能力，以确保训练过程的高效性。该过程使用Python 3.11版本环境，并结合PyTorch2.4.0框架实现。详细超参数设置信息如表3所示。

为了全面评估微调后的LLaMA3模型在实体识别和关系抽取任务中的性能，本文采用了精确率P、召回率R和F_1值作为评估指标。计算式分别为

$$P = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{7}$$

$$R = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{8}$$

$$F_1 = \frac{2 \times P \times R}{P + R} \tag{9}$$

其中，TP表示预测为正类，且真实也是正类的数量，FP表示预测为正类，但真实是负类的数量，FN表示真实为正类，但被预测为负类的数量。

表3 超参数设置信息

超参数名称	含义	数值
num_train_epochs	训练总轮次	5
per_device_train_batch_size	每个设备的训练批次大小	2
gradient_accumulation_steps	梯度累积步数	8
max_grad_norm	梯度裁剪阈值	1
learning_rate	学习率	1×10^{-4}
optim	优化器类型	adamw_torch
lr_scheduler_type	学习率衰减策略	cosine
lora_alpha	缩放系数	16
loraplus_lr_ratio	LoRA+学习率比例	16
lora_rank	LoRA 低秩矩阵的秩	8
lora_dropout	LoRA 层的 dropout 率	0.05
lora_target_modules	LoRA 目标模块	q_proj, k_proj, v_proj, o_proj
cutoff_len	序列截断长度	1 024

3.2 实验结果及分析

3.2.1 不同开源大模型对比

首先，为验证所选基础大语言模型在威胁情报信息抽取任务中的适用性与优势，本文在相同微调策略下对多种主流开源模型进行了性能对比，包括Qwen2.5-7B、Mistral-7B、LLaMA3-8B。所有模型均采用LoRA进行微调，保持一致的超参数与训练轮次。实验结果如表4所示。

表4 不同模型抽取结果

模型	实体识别F_1值	关系抽取F_1值
Qwen2.5-7B	0.856 2	0.813 5
Mistral-7B	0.831 7	0.772 3
LLaMA3-8B	**0.886 6**	**0.837 8**

从实验结果可以看出，LLaMA3-8B在实体识别和关系抽取任务上均取得最佳性能，F_1值分别达到0.886 6和0.837 8，因此被选为本文的基础模型进行领域微调。

3.2.2 实体识别

为了更全面地评估本文方法的性能，本文在统一的提示词模板和测试数据集上，将其与当前主流

大模型 GPT-4o 和 DeepSeek 进行测试对比。同时，还选取了深度学习方法 BiLSTM-CRF[28]和最新的基于大模型的情报抽取模型 LLM-TIKG[27]作为对比基线。

在实体识别任务中，由于中文特有的复杂字符结构、语法规则以及语境依赖，使得任务难度显著提升。现有深度学习方法在中文威胁情报实体识别方面表现仍不理想。为此，本文设计并开展了中文威胁情报实体识别实验，结果如表 5 所示。

表 5　　中文实体识别实验结果

模型	精确率	召回率	F_1 值
BiLSTM-CRF[28]	0.697 3	0.752 6	0.723 9
GPT-4o	0.685 0	0.925 5	0.787 3
DeepSeek	0.910 7	0.772 7	0.836 0
LLM-TIKG[27]	0.744 2	0.727 3	0.735 7
TIE&I-LLM	**0.924 7**	**0.851 5**	**0.886 6**

实验表明，本文方法有效弥补了 LLM-TIKG 在该任务中的不足，且在一定程度上克服基于 Prompt 产生的幻觉现象。与现有主流大模型相比，本文方法在精确率与召回率之间展现出更优的平衡性：精确率较 DeepSeek 模型提升了 0.01 4，召回率虽较 GPT-4o 略低 0.07 4，但显著减少了误报率。这种平衡性使其在实际威胁分析场景中更具应用价值，既避免了 GPT-4o 易产生误报的问题，也克服了 DeepSeek 因保守而导致的漏报现象。

表 6 展示了相应模型在英文威胁情报实体识别任务上的性能。得益于英文语言中明确的词汇边界，所有模型在该任务中的表现均有所提升。英文相较于其他语言，其语法结构和词汇分界更为清晰，这使得模型能够更准确地识别实体。

表 6　　英文实体识别实验结果

模型	精确率	召回率	F_1 值
BiLSTM-CRF[28]	0.815 3	0.795 2	0.805 1
GPT-4o	0.765 1	0.914 6	0.833 2
DeepSeek	0.918 4	0.837 1	0.875 9
LLM-TIKG[27]	0.878 8	0.839 9	0.858 9
TIE&I-LLM	**0.937 5**	**0.876 1**	**0.905 8**

总体来看，本文方法在威胁情报实体识别任务中能够取得优异的效果，充分证明了其在处理该类任务时的有效性。

3.2.3　关系抽取

在完成实体识别后，本文进一步开展了关系抽取任务。实验结果如表 7 所示。

表 7　　关系抽取实验结果

模型	精确率	召回率	F_1 值
GPT-4o	0.554 2	0.851 9	0.671 5
DeepSeek	0.902 4	0.685 2	0.778 9
LLaMA3	0.452 8	0.558 1	0.499 9
TIE&I-LLM	**0.939 4**	**0.756 1**	**0.837 8**

在威胁情报关系抽取任务中，本文方法取得了 0.837 8 的 F_1 值，在所有对比模型中表现最优。F_1 值作为衡量模型在精确率与召回率之间综合能力的关键指标，能够更全面地反映模型的实际抽取效果，具有重要的评价意义。

与 DeepSeek 相比，本文方法在精度和召回能力上实现了双重提升，精确率提高 0.037，召回率提升 0.070 9，展现出更高效的抽取能力；相较于 GPT-4o，虽然召回率略低 0.095 8，但精确率提升高达 0.385 2，有效避免了因过度召回带来的大量误判，体现出更强的实用性。

此外，与未经微调的原始 LLaMA3 模型相比，本文方法在各项指标上均有明显提升，精确率提高 0.486 6，召回率提高 0.198，F_1 值提高 0.337 9，进一步验证了所提出的微调策略及高质量中英文混合数据集对模型性能的有效增强，证明了该方法在威胁情报关系抽取任务中的可行性与先进性。

3.2.4　不同数据类型抽取性能对比

为验证模型在不同类型安全文本上的适用性，本文进一步选取了 3 类异构数据进行评估，包括：APT 攻击报告类文本（结构化程度高），漏洞公告类文本（短句多、术语密集），网络安全新闻报道类文本（语义多样、上下文跨度大）。实验结果如表 8 所示。

表 8　　不同数据类型抽取实验结果

数据类型	实体识别 F_1 值	关系抽取 F_1 值
APT 攻击报告	0.913 3	0.854 7
漏洞公告	0.889 1	0.818 3
网络安全新闻报道	0.867 4	0.795 2

实验结果显示，本文方法在3类数据集上均保持较高的识别准确率与稳健性。其中，在APT攻击报告中表现最佳（F_1值达0.913 3），在漏洞公告类文本上取得0.889 1的F_1值，说明模型能够较好地识别短文本中高密度技术术语与漏洞编号等结构化实体；而在语义更复杂的新闻类文本上仍能保持0.867 4的F_1值。总体实验结果表明，本文方法在多样化安全语料中均表现出良好的跨域泛化能力与稳定性。

3.2.5 不同微调方法对比

为验证不同微调方法的效果，本文对LoRA、参数冷冻微调（Freeze）、全参数微调（Full）进行对比。实验结果如表9所示。

表9 不同微调方法对比

微调方法	实体识别F_1值	关系抽取F_1值	训练时间/min	资源可行性
LoRA	0.886 6	0.837 8	15~20	完全可行
Freeze	0.853 2	0.795 5	80~100	勉强可行
Full	—	—	无法完成	不可行

从表9可以看出，在相同训练配置下，LoRA在性能与效率上均优于Freeze。在实体识别和关系抽取任务中，LoRA的F_1值分别比Freeze提高0.033 4与0.042 3。这主要得益于LoRA通过低秩矩阵结构有效注入领域知识，能够在保持模型稳定性的同时，显著提升对安全术语和隐含语义的识别能力；而Freeze指的是在训练过程中只对模型的小部分权重进行更新，冻结底部的大部分Transformer层及词嵌入等参数，仅更新顶部若干层Transformer块以及任务相关输出头参数，底层语义表征无法充分适配新领域，导致性能受限。

在效率与资源可行性方面，Full需更新全部模型参数；Freeze训练参数量约为Full的10%；LoRA训练参数量下降到约为Full的0.1%，处于千分之一量级。在此配置下，LoRA的训练时间仅为Freeze的$\frac{1}{6}\sim\frac{1}{5}$，并能在4×RTX4090环境下无压力稳定运行；相比之下，Freeze虽然可以完成训练，但显存需求已经接近该环境的上限；Full在相同硬件配置下触发内存溢出错误，未能完成训练过程，因此无法给出有效的性能数据。

因此，在本文实验条件下，LoRA实现了性能、训练效率与资源利用之间的最优平衡，是威胁情报抽取任务的最合理微调策略选择。

3.2.6 消融实验

为进一步探究模型的各个模块对威胁情报信息抽取效果带来的影响，本文设置了下列4个模型来进行消融实验，实验结果如表10所示。

表10 消融实验结果

模型	实体识别F_1值	关系抽取F_1值
Model_0	0.736 9	0.499 9
Model_1	0.762 5	0.693 2
Model_2	0.824 1	0.765 6
Model_3	0.886 6	0.837 8

Model_0：原始LLaMA3-8B模型。

Model_1：仅使用标注的英文数据集微调的模型。

Model_2：仅使用标注的中文数据集微调的模型。

Model_3：本文方法（中英双语标注数据集微调）。

消融实验结果显示，各模型性能呈梯度提升，验证了领域适配与多语言协同的有效性。Model_0性能最低，表明通用模型需领域微调。单语言微调中，Model_2优于Model_1，实体与关系F_1分别高0.061 6、0.072 4，因任务是中文威胁情报抽取，更适配中文语义。Model_3性能最优，体现双语协同优势，英文强化国际安全术语识别，中文保障本土语义理解，二者结合提升威胁情报信息抽取能力。综上，可验证本文方法的有效性。

3.3 推理实例分析

以360公布的针对美相关APT组织的分析报告为例，使用本文方法对报告内容进行了实体关系抽取，并采用Neo4j对抽取出的实体及关系以知识图谱的形式进行可视化展示。图3展示了APT报告知识图谱。该图谱展示了威胁情报知识的整体框架，便于从全局角度把握关系脉络。

为进一步展示细节信息，图4展示了图3中攻击组织“索伦之眼”区域的局部放大结果。从该图谱中可以直观地获取“索伦之眼”攻击组织又名APT-C-16；攻击目标地区为中国、俄罗斯、伊朗等地区；该组织位于美国；攻击行业包括国防、政府、金融等，以及所采用的攻击技术和工具等信

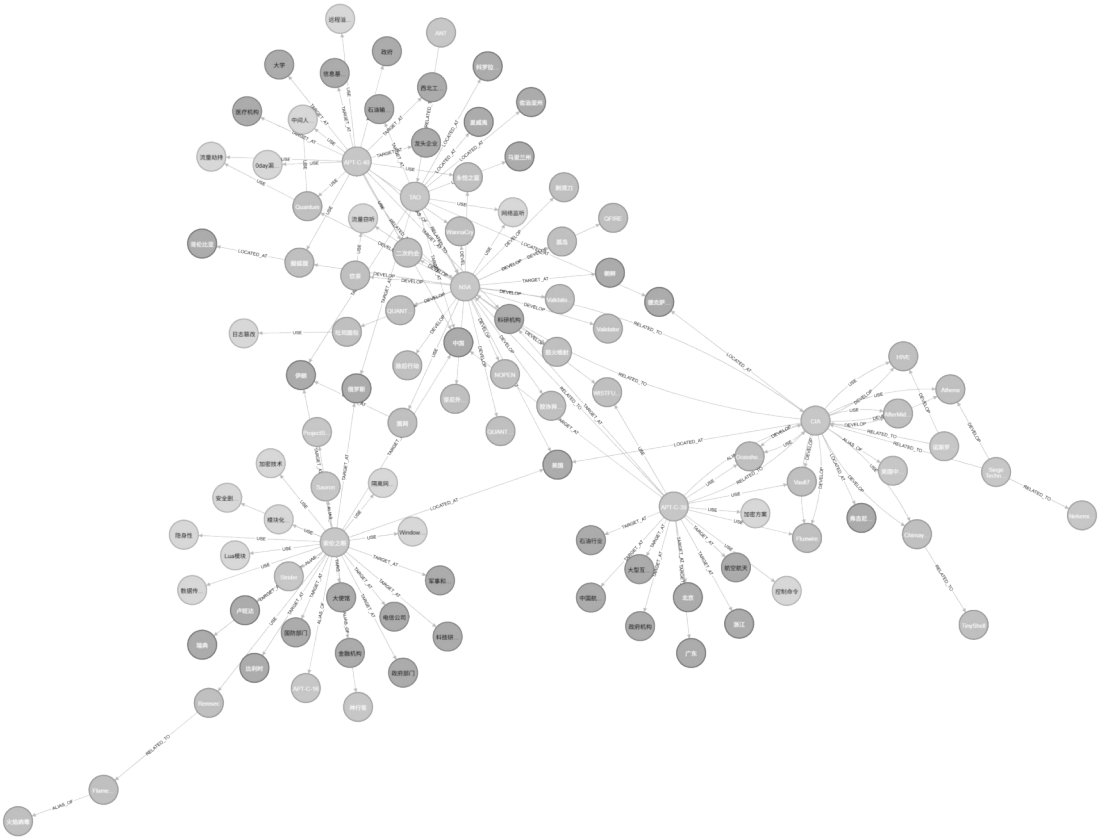


图 3　APT报告知识图谱

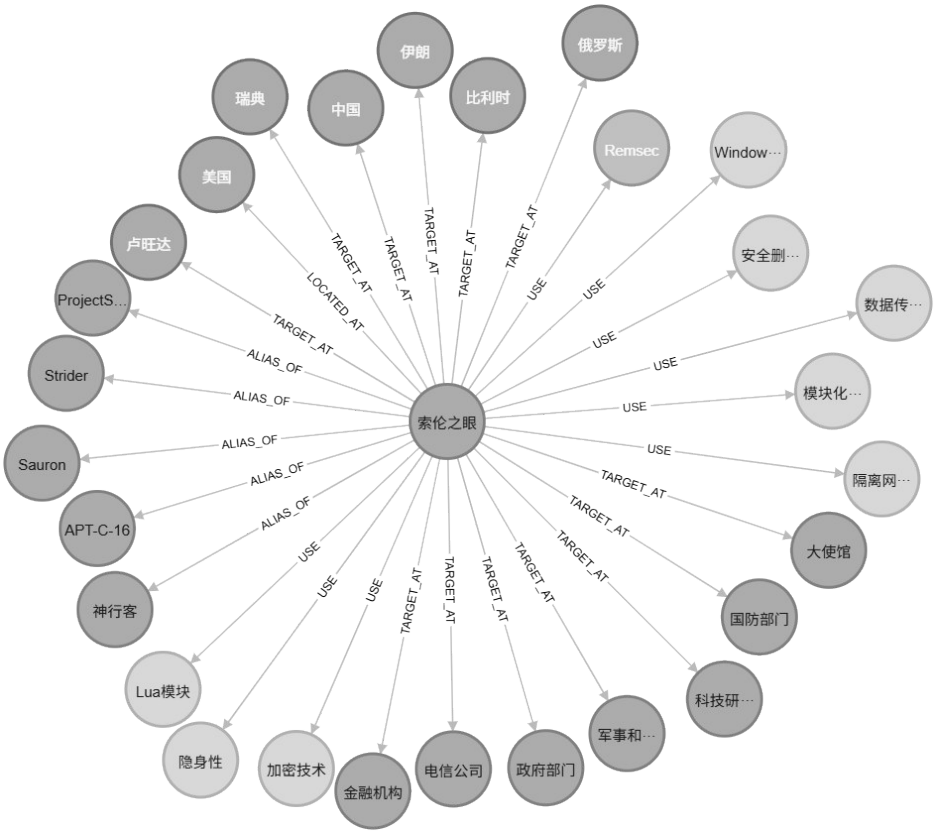


图 4　索伦之眼组织相关部分图谱

息。该图谱清晰地揭示了不同威胁情报实体之间的关联关系，有助于深入理解攻击者的恶意活动模式。

随后，以问题“根据报告内容分析 APT-C-39 组织在攻击过程中利用了哪些 ATT&CK 战术，并给出相关依据”为例，对该报告进行了推理分析，结果如图 5 所示。本文方法通过深度整合 ATT&CK 框架，能够精准识别并关联攻击链中的关键技战术。例如，在初始入侵阶段，系统依据语义特征将攻击行为映射至 T1195 与 T1200 等标准化战术节点。对比实验中，相同输入条件下 GPT-4 模型的输出如图 6 所示。GPT-4 因缺乏结构化威胁知识嵌入，生成内容出现显著事实性偏差：包括编撰不存在的技术编号（如虚构 T1100）、编撰不存在的技术名称（如 Data Theft）等“幻觉”现象。同时，相同输入条件下开源安全大模型 SecGPT 的输出如图 7 所示。可以看到，SecGPT 虽能给出若干与攻击相关的描述类别，但未能规范化映射到 ATT&CK 框架，其输出多为“收集信息（Gather Information）”“利用已知漏洞（Exploitation of Known Vulnerabilitiy）”“命令与控制（C2）”等宽泛的概念，缺少对应的 ATT&CK 技战术编号及明确的技术边界。

图 5　本文方法推理结果

图 6　GPT-4 推理结果

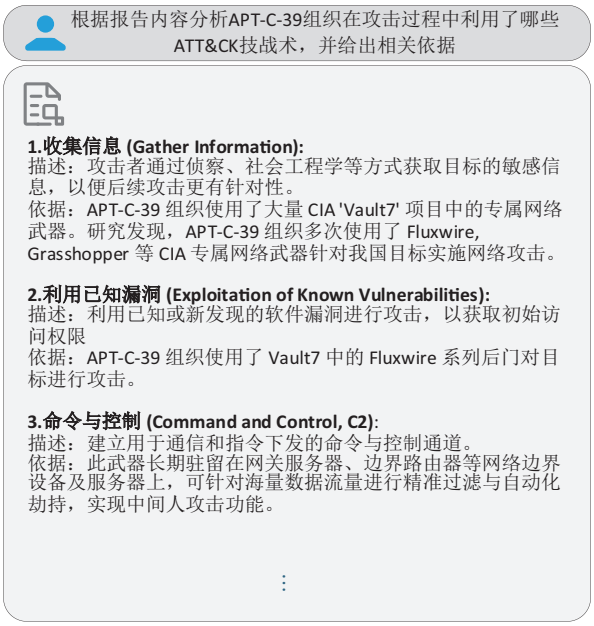


图 7　SecGPT 推理结果

综上所述，本文通过系统实验综合评估了本文在威胁情报信息抽取和推理分析任务中的性能，充分验证了本文方法的有效性与优势。

3.4　相关工作比较

现有研究[2-4,13]采用多种自然语言处理技术完成自动化提取威胁情报，AttackG[29]从 CTI 报告中联合抽取攻击相关实体再进一步聚合为技术知识图谱，但均聚焦于英文文本。针对中文情报抽取任

务，文献[7]尝试采用深度学习方法进行实体识别，取得了0.779 9的F_1值，整体表现一般，难以适应中文情报语义复杂、边界模糊等挑战。文献[30]借助提示工程实现基于LLM的威胁实体与关系抽取，但由于缺乏领域约束与结构化支撑，仍存在“幻觉”现象。文献[31]在基于提示词完成威胁信息抽取后借助语言链框架LangChain实现面向威胁情报文档的智能问答，但未融合ATT&CK框架导致回答深度不足，无法关联相关技战术。相比之下，LLM-TIKG[27]尽管在英文语料上展现良好性能，但对中文情报适应性不足。针对上述问题，本文通过构建中英文混合标注数据集并对模型进行指令微调，在准确性与跨语言迁移能力方面均优于现有工作，特别是在中文威胁情报抽取任务中显著提升了性能表现，弥补了当前研究的不足。

4　结束语

在大语言模型迅猛发展的背景下，中文威胁情报的结构化处理与深度理解仍面临巨大挑战。为此，本文提出一种基于大语言模型的威胁情报高效抽取与攻击推理方法。首先，通过GPT-4o与精心设计的Prompt模板实现高效的数据自动标注，并结合人工校验构建高质量训练集；随后，引入LoRA对LLaMA3-8B模型进行微调，有效提升威胁情报中实体与关系的抽取效果；最后，构建结构化知识图谱并融合GraphRAG与ATT&CK框架，实现图结构引导下的大模型多跳推理，支持APT攻击链的自动分析与攻击意图识别。

实验结果表明，本文方法在中文场景下的抽取准确率与推理完整性均优于现有主流方法，具备较强的实用性，能够为安全运营中心提供更加智能、可控的情报分析支撑。在实际部署中，本文方法在大规模与实时场景下均具有可行的扩展路径。对于大规模数据集，威胁情报抽取阶段可通过调用本地大模型API完成，因此可以在内网环境下对海量报告进行高效抽取，满足安全领域对数据合规与隐私的要求。对于实时更新的威胁情报，本文可利用已构建的知识库及历史情报图谱进行基于既有知识的推理分析。

未来将进一步拓展跨文档、多模态情报的融合推理能力，推动威胁情报自动化、智能化处理体系的发展。

参考文献：

[1] ZHOU Y H, TANG Y, YI M, et al. CTI view: APT threat intelligence analysis system[J]. Security and Communication Networks, 2022, 2022: 9875199.

[2] LI Y F, GUO Y B, FANG C, et al. A novel threat intelligence information extraction system combining multiple models[J]. Security and Communication Networks, 2022, 2022: 8477260.

[3] LIU P P, LI H, WANG Z G, et al. Multi-features based semantic augmentation networks for named entity recognition in threat intelligence[C]// Proceedings of the 2022 26th International Conference on Pattern Recognition (ICPR). Piscataway: IEEE Press, 2022: 1557-1563.

[4] GAO C, ZHANG X, LIU H. Data and knowledge-driven named entity recognition for cyber security[J]. Cybersecurity, 2021, 4(1): 9.

[5] QIU Y F, DONG L B, ZHANG W W, et al. A diffusion enhanced CRF and BiLSTM framework for accurate entity recognition[J]. Scientific Reports, 2025, 15: 19670.

[6] SEOW W L, CHATURVEDI I, HOGARTH A, et al. A review of named entity recognition: from learning methods to modelling paradigms and tasks[J]. Artificial Intelligence Review, 2025, 58(10): 315.

[7] ZHOU Y H, REN Y T, YI M, et al. CDTier: a Chinese dataset of threat intelligence entity relationships[J]. IEEE Transactions on Sustainable Computing, 2023, 8(4): 627-638.

[8] ZHAO W X, ZHOU K, LI J, et al. A survey of large language models[J]. arXiv Preprint, arXiv: 2303.18223, 2023.

[9] GAO C, ZHANG X, HAN M T, et al. A review on cyber security named entity recognition[J]. Frontiers of Information Technology & Electronic Engineering, 2021, 22(9): 1153-1168.

[10] JOSHI A, LAL R, FININ T, et al. Extracting cybersecurity related linked data from text[C]//Proceedings of the 2013 IEEE Seventh International Conference on Semantic Computing. Piscataway: IEEE Press, 2013: 252-259.

[11] MULWAD V, LI W J, JOSHI A, et al. Extracting information about security vulnerabilities from web text[C]//Proceedings of the 2011 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology. Piscataway: IEEE Press, 2011: 257-260.

[12] DIONÍSIO N, ALVES F, FERREIRA P M, et al. Cyberthreat detection from twitter using deep neural networks[C]//Proceedings of the 2019 International Joint Conference on Neural Networks (IJCNN). Piscataway: IEEE Press, 2019: 1-8.

[13] 杨秀璋, 彭国军, 李子川, 等. 基于Bert和BiLSTM-CRF的APT攻击实体识别及对齐研究[J]. 通信学报, 2022, 43(6): 58-70.
YANG X Z, PENG G J, LI Z C, et al. Research on entity recognition and alignment of APT attack based on Bert and BiLSTM-CRF[J]. Journal on Communications, 2022, 43(6): 58-70.

[14] SHANG Y M, HUANG H Y, SUN X, et al. A pattern-aware self-attention network for distant supervised relation extraction[J]. Information Sciences, 2022, 584: 269-279.

[15] GUO Z J, NAN G S, LU W, et al. Learning latent forests for medical relation extraction[C]//Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence. International Joint Conferences on Artificial Intelligence Organization, 2020: 3651-3657.

[16] LIU J, CHEN S W, WANG B Q, et al. Attention as relation: learning supervised multi-head self-attention for relation extraction[C]//Proceedings of the Twenty-Ninth International Joint Conference on Artifi-

cial Intelligence. International Joint Conferences on Artificial Intelligence Organization. New York: ACM Press, 2020: 3787-3793.

[17] ZENG S, XU R, CHANG B, et al. Double graph based reasoning for document-level relation extraction[J]. arXiv Preprint, arXiv: 2009.13752, 2020.

[18] YUAN C S, HUANG H Y, FENG C, et al. Document-level relation extraction with entity-selection attention[J]. Information Sciences, 2021, 568: 163-174.

[19] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[J]. Advances in Neural Information Processing Systems, 2017, 30: 5998-6008.

[20] BROWN T, MANN B, RYDER N, et al. Language models are few-shot learners[J]. Advances in Neural Information Processing Systems, 2020, 33: 1877-1901.

[21] TOUVRON H, LAVRIL T, IZACARD G, et al. LLAMA: open and efficient foundation language models[J]. arXiv Preprint, arXiv: 2302.13971, 2023.

[22] CHOWDHERY A, NARANG S, DEVLIN J, et al. PaLM: scaling language modeling with pathways[J]. Journal of Machine Learning Research, 2023, 24(240): 1-113.

[23] LIU Y L, TAO S M, MENG W B, et al. Interpretable online log analysis using large language models with prompt strategies[C]//Proceedings of the 2024 IEEE/ACM 32nd International Conference on Program Comprehension (ICPC). Piscataway: IEEE Press, 2024: 35-46.

[24] HU P W, LIANG R G, CHEN K. DeGPT: optimizing decompiler output with LLM[C]//Proceedings 2024 Network and Distributed System Security Symposium. Virginia: the Internet Society, 2024: 1-16.

[25] ZHANG Q, CHEN S, BEI Y, et al. A survey of graph retrieval-augmented generation for customized large language models[J]. arXiv Preprint, arXiv: 2501.13958, 2025.

[26] EDGE D, TRINH H, CHENG N, et al. From local to global: a graph rag approach to query-focused summarization[J]. arXiv Preprint, arXiv: 2404.16130, 2024.

[27] HU Y L, ZOU F T, HAN J J, et al. LLM-TIKG: Threat intelligence knowledge graph construction utilizing large language model[J]. Computers & Security, 2024, 145: 103999.

[28] GASMI H, LAVAL J, BOURAS A. Information extraction of cybersecurity concepts: an LSTM approach[J]. Applied Sciences, 2019, 9(19): 3945.

[29] LI Z Y, ZENG J, CHEN Y, et al. AttacKG: constructing technique knowledge graph fromCyber threat intelligence reports[C]//Computer Security-ESORICS 2022. Berlin: Springer, 2022: 589-609.

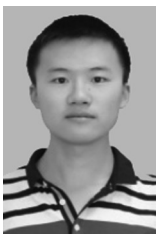
[30] 陈继智, 万朝华, 张斯威. 基于大语言模型的零样本安全知识抽取方法[J]. 信息安全研究, 2024, 10(E2): 59-63.
CHEN J Z, WAN Z H, ZHANG S W. Zero-shot security knowledge extraction method based on large language models[J]. Journal of Information Security Research, 2024, 10(E2):59-63.

[31] 马冰琦, 周盈海, 王梓宇, 等. 一种基于大语言模型的威胁报信息抽取方法[J]. 网络空间安全科学学报, 2024, 2(2): 36-46.
MA B Q, ZHOU Y H, WANG Z Y, et al. A LLMs-based method for threat intelligence information extraction[J]. Journal of Cybesecurity, 2024, 2(2): 36-46.

[作者简介]

彭国军（1979– ），男，湖北荆州人，博士，武汉大学教授、博士生导师，主要研究方向为网络与信息系统安全。

李家琛（2001– ），男，河南南阳人，武汉大学硕士生，主要研究方向为网络与信息系统安全。



杨秀璋（1991– ），男，贵州凯里人，博士，贵州大学副教授、硕士生导师，主要研究方向为网络与信息系统安全。

吕锦钊（2002– ），男，贵州遵义人，武汉大学硕士生，主要研究方向为大模型辅助的漏洞挖掘技术、白盒代码分析技术。