

融合威胁情报知识图谱与图约束推理的网络威胁归因框架

刘昌城, 杨长恒, 马骏

(解放军网络空间部队信息工程大学, 河南 郑州 450001)

摘要: 针对大语言模型在网络威胁归因中存在幻觉、传统图约束推理对碎片化威胁情报数据鲁棒性不足的问题, 对可信网络攻击溯源方法进行了研究。提出融合威胁情报知识图谱与图约束推理 (TIKGCR) 框架, 构建攻击模式导向的知识图谱前缀树 (KG-Trie) 索引, 设计自适应软约束解码与多路径证据聚合机制。实验结果显示, 该框架在 CTI-Trace 和 TRAIL 数据集上归因准确率分别达 94.6% 和 92.8%, 幻觉率仅为 1.5%; 50% 图谱缺失率下准确率仍保持 85.5%, 推理时间缩短约 $\frac{2}{3}$ 。该框架实现了结构可信与语义鲁棒的平衡, 为高风险场景下的网络威胁归因提供了有效技术支撑。

关键词: 大语言模型; 知识图谱; 威胁情报; 攻击溯源

中图分类号: TP393.08

文献标志码: A

DOI: 10.11959/j.issn.2096-109x.AQ26010

Cyber threat attribution framework integrating threat intelligence knowledge graph and graph constrained reasoning

Liu Changcheng, Yang Changheng, Ma Jun

People's Liberation Army Cyberspace Force Information Engineering University, Zhengzhou 450001, China

Abstract: To address the issues of hallucinations in large language models for cyber threat attribution and the insufficient robustness of traditional graph-constrained reasoning against fragmented threat intelligence data, research on credible cyber attack traceability methods was conducted. The framework integrating threat intelligence knowledge graph and graph-constrained reasoning (TIKGCR) was proposed, which constructs an attack pattern-oriented knowledge graph Trie (KG-Trie) index and designs an adaptive soft-constrained decoding and multi-path evidence aggregation mechanism. Experimental results show that the framework achieves attribution accuracies of 94.6% and 92.8% on the CTI-Trace and TRAIL datasets respectively, with a hallucination rate of only 1.5%. It maintains an accuracy of 85.5% under a 50% knowledge graph missing rate and inference time cut by roughly $\frac{2}{3}$. This framework achieves a balance between structural credibility and semantic robustness, providing effective technical support for cyber threat attribution in high-risk scenarios.

Key words: large language models, knowledge graph, threat intelligence, attack traceability

0 引言

网络威胁归因是安全运营、应急响应及威胁

狩猎的关键环节, 要求分析人员整合告警、日志、恶意样本与威胁报告中的零散信息, 构建可验证

收稿日期: 2026-01-13; 修回日期: 2026-05-29

通信作者: 马骏, sijunhan@163.com

基金项目: 国家自然科学基金面上资助项目 (No.62477020)

Foundation Item: The General Project of National Natural Science Foundation of China (No.62477020)

的攻击链，并据此识别攻击者及其战术技术过程^[1]。近年来，大语言模型（large language model, LLM）^[2]凭借处理非结构化文本和归纳复杂线索的卓越能力，在多源情报融合、攻击链推断及解释生成等方面表现出色，大幅减少人工投入。但LLM在高风险安全场景中的应用经常出现幻觉现象^[3]，推理过程的可信度不足。LLM推理时经常出现虚构的关联，如不存在的域名解析记录、伪造的样本归属关系或虚假的战术、技术与程序（tactics, techniques and procedures, TTP）链接，都可能导致误判，进而引发严重后果。

知识图谱（knowledge graph, KG）为增强推理可信度提供了理想的解决方案。它以结构化形式存储实体及实体间的关系，可作为威胁情报的可验证事实层。但如何将LLM的非结构化生成式推理与KG的结构化事实有效融合，仍是一项挑战。目前，KG增强LLM推理的方法主要分为两类：检索式和智能体（Agent）式^[4]。检索式方法通过外部检索工具从KG中获取相关三元组或子图，再交由LLM进行总结推理。检索式方法性能高度依赖检索质量，在多跳攻击链分析中，容易遗漏关键连接或引入干扰信息。Agent式方法则让LLM以交互探索的方式在图谱上逐步扩展路径，但通常需要多轮调用，搜索空间庞大，时延较高，且在安全数据稀疏或不完整时，易偏离关键证据链。更重要的是，这两类方法往往仍允许LLM在最终回答或解释中“自由发挥”，难以对生成过程施加严格的结构约束。

为解决上述问题，本文提出一个融合威胁情报知识图谱与图约束推理（threat intelligence knowledge graph and graph-constrained reasoning, TIKGCR）框架，包括威胁情报知识提取与图谱构建、攻击模式导向的知识图谱前缀树（knowledge graph Trie, KG-Trie）索引构建和自适应图约束推理与归因。本文的核心思想是：不仅让KG作为可检索的外部知识，而且将其结构以及攻击模式先验转化为可执行的生成约束，从而在推理链层面抑制幻觉、提升可解释性，并在图不完整时仍能进行稳健推断。

第一阶段，本文从多源威胁情报（安全日志、告警、事件报告、失陷指标（IOC）列表、

样本分析文本等）中抽取实体与关系，构建面向归因任务的威胁情报知识图谱。该阶段强调对安全实体类型与关系语义的统一规范化，为后续路径推理提供可连接的结构基础。第二阶段，本文引入攻击模式导向的索引思想：攻击并非任意图游走，通常遵循特定的攻击链结构，例如初始访问→执行→持久化→C2→横向移动→数据窃取。因此，本文将符合攻击模式模板的有效路径编码为KG-Trie索引，使得推理路径可以被高效地“前缀校验”和剪枝。第三阶段，本文提出自适应图约束推理与归因：在推理过程中动态调节约束强度，使模型既能在证据充分时严格遵循图中可验证路径（消除推理幻觉），也能在图谱缺边、噪声或观测不全时，通过多证据路径的归纳聚合实现稳健归因，并输出可审计的证据链与解释。

本文将KG结构直接注入LLM解码过程以获得可信推理，通过构建基于Trie的路径索引，可以在解码阶段以结构约束的方式生成知识图谱锚定（KG-grounded）的推理路径，从而显著减少幻觉并提升效率，例如避免高代价的图遍历与多轮交互。在威胁情报场景中，强调攻击模式先验与归因解释的结合，使得生成的推理链不仅“存在于图中”，也更符合安全分析逻辑，并能直接服务于溯源与处置决策。本文的主要贡献如下。

1) 提出一个面向威胁情报归因的3阶段框架TIKGCR，将可验证的图结构约束引入LLM推理链生成，并输出可审计证据链。

2) 设计攻击模式导向的KG-Trie索引构建策略，将符合TTP规律的有效路径编码为可用于前缀约束的结构化索引，从源头缩小搜索空间、提升路径有效性。

3) 提出自适应图约束推理与归因机制，在严格可信和缺失鲁棒之间动态权衡，提升在不完整、噪声威胁图谱上的归因可靠性与解释质量。

1 相关工作

本文与3个相互关联的研究方向密切相关：大语言模型的推理能力与可靠性、知识图谱增强的可信生成，以及网络威胁情报中的智能归因与溯源。

1.1 大语言模型的推理能力与可靠性

大语言模型在推理任务中的表现，已经成为

近几年机器学习领域最受关注的话题之一。思维链 (chain-of-thought, CoT) 提示方法首次系统地证明, 只要引导模型显式生成中间推理步骤, 其在算术、符号和多步推理任务上的表现就可以显著改善^[5]。随后, 大量工作围绕如何让模型想得更稳展开, 例如通过生成多条推理链并以一致性为准进行聚合^[6], 或将推理过程组织为树状结构, 以探索不同的推理分支^[7]。也有研究尝试通过微调或潜变量建模, 直接增强模型内部的推理表示^[8]。但随着研究的深入, 研究人员发现推理步骤看起来合理, 并不等于它们在事实上成立。多项分析表明, LLM 往往会生成语法流畅、逻辑连贯, 但与外部知识不一致的推理链, 且模型目前缺乏稳定的自我纠错能力^[9]。这一现象在需要依赖外部事实或结构化知识的任务中尤为突出, 也直接促使研究者开始认真思考: 仅靠模型内部的推理机制, 是否足以支撑高风险场景下的可靠推理?

1.2 知识图谱增强的可信生成

知识图谱在表示模式上具有可解释优势^[10], 为了进一步提升推理的可信度, 研究者们把目光投向了知识图谱。目前, KG 增强 LLM 推理的方法主要分为检索式和 Agent 式^[4]。检索式方法将知识图谱视为外部知识源, 通过子图或三元组检索, 为模型提供补充事实^[11-12]。这种方式在工程上较为简洁, 但检索误差会直接影响下游推理效果, 且在长攻击链或复杂多跳推理任务中常常力不从心。Agent 式方法则试图更紧密地结合语言模型和图结构, 以多轮交互的方式逐步探索知识图谱并构建推理路径^[13-14]。这类方法在形式上更贴近人类的逐步推理过程, 也具备更好的可解释性, 但代价是推理成本高、时延大, 对图规模和噪声高度敏感, 且模型在生成中间推理或最终结论时, 仍然拥有较大的自由度, 幻觉问题并未从根本上得到解决。为了进一步提升推理的可信度, 部分研究开始将注意力转向在推理阶段直接引入结构化约束。Luo 等^[15]提出了图约束推理范式, 将知识图谱中的有效推理路径编码为 Trie 结构, 并在解码阶段直接限制模型的 token 生成空间。这种解码期约束不仅可以提升推理的忠实性, 还能在效率上优于反复的图遍历与多轮交互。

1.3 网络威胁情报中的智能归因与溯源

近年来, 大型语言模型的崛起彻底改变了威胁情报分析的格局。黄克振等^[16]提出一种基于图模型的网络攻击溯源, 通过对攻击事件的挖掘能够实现对攻击者的溯源。Xu 等^[17]提出一种基于异构图注意力网络的黑客组织识别方法 (HGHAN), 相比其他异构图节点嵌入算法具有更好的识别效果。Zang 等^[18]设计了一种异构威胁情报融合方法来实时重建攻击场景, 利用社区检测算法挖掘攻击场景以发现关键攻击路径。然而, 这些现有方法在面对真实世界中高度碎片化和不完整的威胁情报数据时, 往往暴露出明显的局限性。一方面, 传统的图挖掘或自由探索方法 (如单纯的社区检测或 Agent 图遍历) 缺乏明确的攻击模式和逻辑先验, 在庞大且含噪的图谱中容易陷入纯数据驱动的盲目搜索, 导致推理效率低下, 常常抓取到大量无关的连接而偏离真实的攻击链^[19]。另一方面, 若直接引入前文提及的传统图约束推理技术, 虽然能提升推理的忠实度, 但其绝对的“硬约束”机制在面对溯源数据碎片化 (如中间跳板机日志丢失导致图谱断边) 时, 会直接导致合法的推理链路被迫中断, 模型将完全无法得出归因结论。同时, 尽管通用 LLM 在文本阅读理解上表现出色, 但当涉及高风险的溯源归因任务时, 若无可靠的外部结构化引导, 依然无法根除幻觉, 无法真正实现可信溯源。

2 方法

本节提出针对网络威胁归因的 TIKGCR 框架。利用 LLM 强大的文本理解能力从非结构化的威胁情报中提取实体和关系, 并与预先定义的本体攻击溯源知识图谱结合构建图谱。为增强 LLM 的推理能力进行攻击溯源, 本文参考图约束推理 (GCR) 框架并针对威胁情报数据往往存在隐蔽性和不完整性的特点, 在经典 GCR 框架的基础上引入了模式制导的索引构建与自适应软约束解码。TIKGCR 框架整体流程如图 1 所示, 包含 3 个核心模块: 威胁情报知识提取与图谱构建、攻击链模式导向的图索引构建和自适应软约束解码推理与归因, CTI 指网络威胁情报。

2.1 威胁情报知识提取与图谱构建

高质量的知识图谱是成功溯源推理的核心。

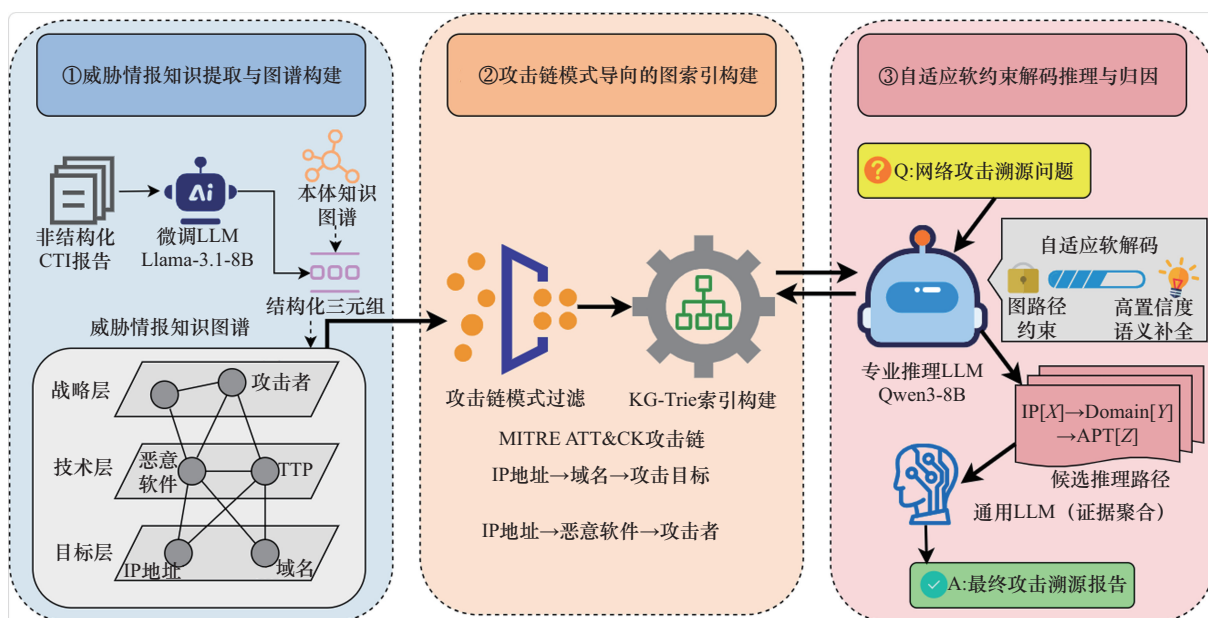


图1 TIKGCR 框架整体流程

本文首先利用自然语言处理技术将威胁情报报告转化为结构化数据，然后基于这些数据构建知识图谱。由于威胁情报通常是非结构化的，因此本文采用了微调过的信息抽取模型来自动提取实体和关系。

2.1.1 威胁情报信息抽取

本文从公开的威胁情报报告 Mandiant 中提取了 10 000 多个文本文件。早期研究多采用双向长短期记忆-条件随机场 (BiLSTM-CRF) 等传统深度学习模型进行信息抽取^[20]，为提高信息抽取的效率和准确率，本文选择 Llama-3.1-8B 模型基于 mrmoor 数据集进行全量微调。考虑到 mrmoor 数据集标签体系与本文 2.1.2 节基于结构化威胁信息表达式 (STIX) 2.1 标准自定义的领域本体存在一定的粒度差异，在微调 Llama 模型之前，对 mrmoor 数据集进行了严格的本体对齐与标签映射。具体而言，通过规则脚本将 mrmoor 数据集中原有的粗粒度标签精确映射至本文定义的实体类型。此外，在对 Mandiant 报告进行大规模推理抽取时，将自定义的本体规范作为结构化指令嵌入系统提示词中，通过微调模型的指令遵循能力，确保最终提取的三元组能够无缝接入威胁情报知识图谱 (TIKG) 的构建。通过信息抽取，得到了如下结构化三元组： $\mathcal{G} = \{(e_h, r, e_t)\}$ ，其中， e_h 和 e_t 代表实体， r 表示它们之间的关系。为了支持后

续的链路推理，预定义了基于 STIX 2.1 标准的本体模式，确保提取的三元组在语义上符合网络安全规范。

2.1.2 本体知识图谱设计

为提升威胁情报知识图谱的结构一致性、可扩展性与可推理性，在构建事实图谱之前先定义一个面向网络威胁归因的领域本体知识图谱，用于约束实体类型、关系类型、属性字段及其取值范围。该本体在语义层面与 STIX 2.1/对抗战术、技术与常识知识库 (ATT&CK) 的核心概念保持一致，并结合溯源任务对“攻击基础设施—恶意软件—受害对象—攻击组织”的证据链需求进行裁剪与扩展。

图 2 为 TIKGCR 的本体知识图谱框架设计，本文将本体定义为四元组 $\mathcal{O} = (\mathcal{C}, \mathcal{R}, \mathcal{A}, \mathcal{X})$ ，其中 $\mathcal{C}$ 、 $\mathcal{R}$ 、 $\mathcal{A}$ 、 $\mathcal{X}$ 分别表示实体、关系、属性和约束集合。在构建威胁情报知识图谱时，每个三元组 (e_h, r, e_t) 需要满足：

$$\begin{aligned} \text{type}(e_h) \in \mathcal{C}, \text{type}(e_t) \in \mathcal{C}, r \in \mathcal{R}, \\ (\text{domain}(r), \text{range}(r)) \in \mathcal{X} \end{aligned} \quad (1)$$

该约束用于避免抽取阶段常见的类型错误与关系幻觉（例如把“APT29”误当作 Malware，或让“IP located_in Country”直接推出 Attacker）。表 1 为实体属性定义，定义了知识图谱 $\mathcal{G}$ 中的节点集合 $\mathcal{C}$ ，表 2 为关系定义和约束集合，

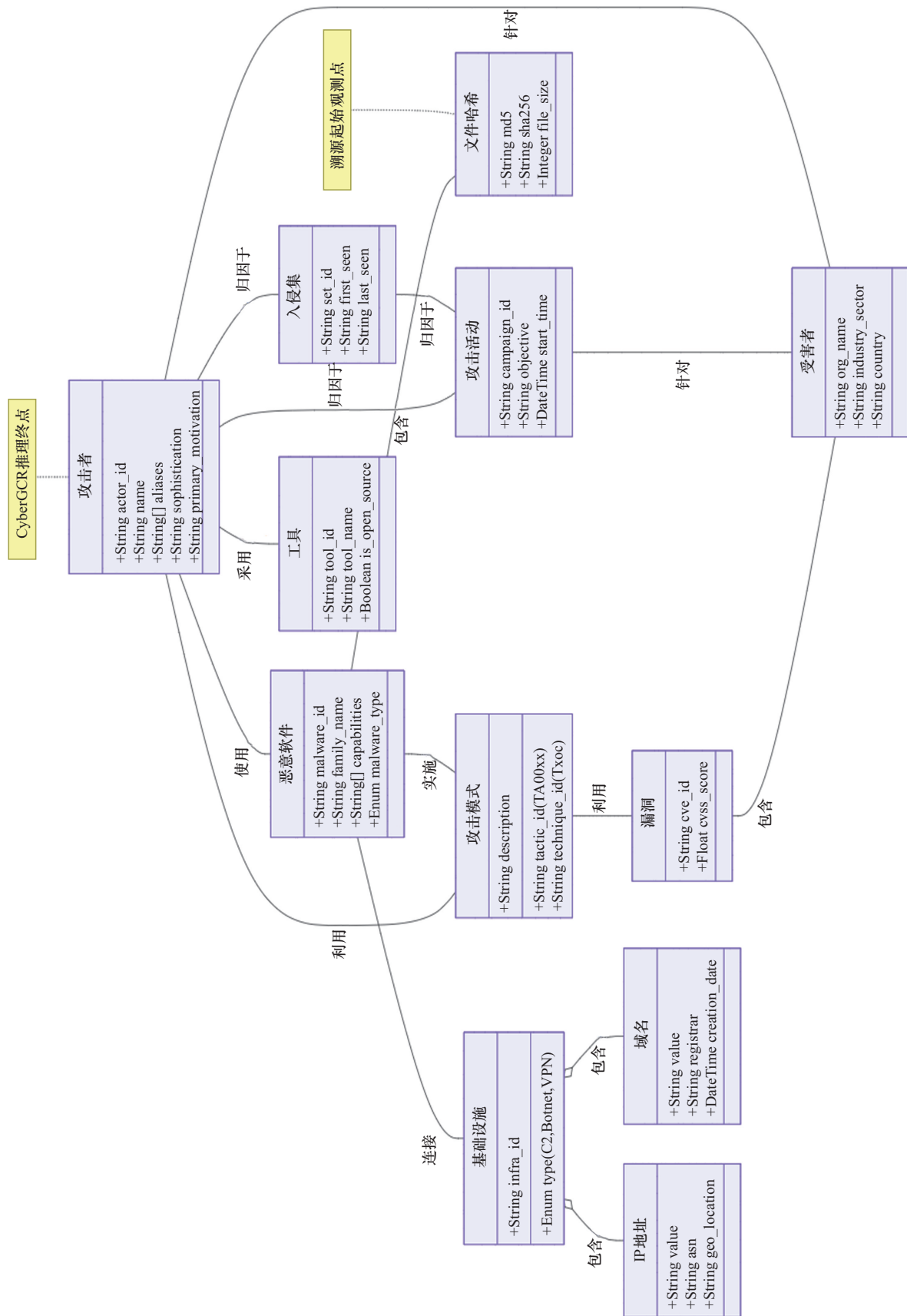


图2 TIKGCR的本体知识图谱框架设计

表1 实体属性定义

实体名	描述	关键属性
攻击者	实施网络攻击的个人、团体或国家级组织	name, aliases, motivation等
入侵集合	一组共享特定属性或资源的敌对行为集合	name, goals, primary_region
攻击活动	攻击者在特定时间段内为达成特定目标而发起的一系列攻击活动	campaign_id, objective等
恶意软件	攻击者植入或使用的恶意软件家族或变种	family_type, platform
工具	用于攻击目的的软件工具（如Mimikatz、PsExec）	tool_name, usage_context
攻击模式	攻击者使用的TTP，直接映射至MITRE ATT&CK	tactic_id, technique_id
基础设施	攻击者用于命令控制（C2），以及发送或攻击的物理/虚拟节点	infra_type（C2, Proxy, Botnet）
IP地址	互联网协议地址，最常见的观测指标	asn, isp, geo_location
域名	网络域名，常用于C2通信或钓鱼网站	registrar, creation_date, whois_data
哈希	恶意文件样本的哈希值（MD5/SHA256），推理的常见起始点	hash_value, file_size, imphash
受害者	遭受攻击的组织或实体。提供上下文信息以辅助约束推理	sector, country, org_size
漏洞	被攻击者利用的软件或系统弱点（CVE）	cve_id, cvss_score等

表2 关系定义和约束集合

关系	约束	描述
归因于	攻击活动/入侵集 → 攻击者 攻击活动 → 入侵集	表示高度置信的归因关系
使用	攻击者 → 恶意软件/工具/基础设施	攻击者使用某种工具或基础设施
实施	恶意软件 → 攻击模式	恶意软件实现了某种特定的攻击技术
包含	受害者 → 漏洞 基础设施 → IP地址/域名 恶意软件 → 哈希	表示某一类包含另一类
利用	攻击模式/恶意软件 → 漏洞	攻击技术或工具利用了特定的漏洞
连接	恶意软件 → 基础设施	样本运行或主机产生的网络连接行为
针对	攻击活动/攻击者 → 受害者	攻击活动针对的受害者

定义了知识图谱中的边集合 $\mathcal{R}$ 以及每种关系的定义域和值域约束。这些约束构成了KG-Trie路径过滤的基础规则。本体知识图谱的设计不仅保证了网络威胁归因任务中实体关系的正确性，还提供了知识图谱的基础结构。

2.2 攻击链模式导向的图索引构建

为了让LLM在推理时不进行无意义的随机游走，本文提出使用KG-Trie来加速图的遍历。与原始GCR方法仅基于连通性进行广度优先搜索不同，本文引入了攻击链模式过滤策略。通过对知识图谱进行模式导向的约束，显著降低了搜索空间，避免了无关路径的遍历。本文定义一条合法的溯源推理路径 $w_z = e_0 \xrightarrow{r_1} e_1 \xrightarrow{r_2} \dots \xrightarrow{r_l} e_l$ ，其中 e_0 为溯源起点（如告警IP）， e_l 为归因目标（如APT组织）。溯源推理路径的拓扑结构必须

契合潜在的攻击链规律，但在搜索方向上，遵循实际调查中由受害者或内部告警节点，反推至外部攻击源的逆向溯源逻辑。本文参考了MITRE ATT&CK知识库中的攻击链模式进行路径过滤，定义了多种攻击路径模式，例如， $\text{Pattern}(w_z) \in \{\text{IP} \rightarrow \text{Domain} \rightarrow \text{Attacker}\}$ 、 $\text{Pattern}(w_z) \in \{\text{IP} \rightarrow \text{Malware} \rightarrow \text{Attacker}\}$ ，其中，Pattern指攻击路径，Domain为域名，Attacker为攻击者，Malware为恶意软件。需要特别说明的是，真实的高级持续性威胁往往呈现多分支、并行的非线性攻击模式（如多点突破、横向移动与C2通信并发）。由于Trie结构在物理上更擅长处理线性的序列约束，在构建KG-Trie索引时采取了“非线性解耦”策略。具体而言，将复杂的非线性攻击图谱解耦为多条符合逆向溯源逻辑的线性子路

径,例如,一条分支追踪网络侧的C2基础设施,另一条分支追踪终端侧的恶意软件家族。通过这种解耦,单纯的Trie前缀校验负责在单一维度上保证线性推断的绝对合法性;完整的非线性攻击全貌,会在后续通过多路径归纳与证据聚合模块,将这些独立生成的线性分支重新拼接与交叉印证,从而完美兼容非线性的真实攻击场景。具体解耦过程如下:对于任意一个非线性攻击子图 G_{sub} ,以告警节点(如受害IP)为起点,执行逆向广度优先搜索(BFS),提取所有到达攻击者节点的简单路径集合 $P = \{p_1, p_2, \dots, p_k\}$ 。每条路径 p_i 均满足攻击链方向约束,并作为独立的线性序列存入KG-Trie。通过这种方式,原本并行的攻击分支被拆解为多条可独立校验的线性路径,兼顾了非线性攻击的表达能力和Trie的高效前缀匹配。通过这种方式构建的索引 $\mathcal{T}$ 能有效过滤掉大量无关的噪声数据(如仅根据地理位置关联的无关节点),显著减小了搜索空间,提高了推理效率。

2.3 自适应软约束解码推理与归因

图约束推理通过在推理过程中引入图约束,确保了生成的推理路径符合知识图谱中的结构约束。在这一阶段,LLM会根据KG-Trie的约束生成合理的推理路径,并最终给出攻击者归属。本文将上一节构建的图谱数据作为微调数据集对Qwen3-8B模型进行微调,用作生成推理路径的专用LLM。但在网络安全场景中,CTI数据往往是不完整的,例如缺失了中间某跳板机的日志。传统的GCR采用“硬约束”,即如果图中不存在路径,强制将生成概率降为0,这会导致溯源在缺失节点处中断。本文提出一种置信度感知的软约束机制,在生成第 i 个token w_{z_i} 时,模型首先计算该token在KG-Trie $\mathcal{T}$ 中的有效性掩码 $M(w_{z_i})$ 。

$$M(w_{z_i}) =$$

$$\begin{cases} 1, \text{Concat}(w_{z_{<i}}, w_{z_i}) \text{ 为索引 } \mathcal{T} \text{ 中的有效前缀} \\ 0, \text{其他} \end{cases} \quad (2)$$

本文将最终的生成概率定义为图约束概率与LLM原始语义概率的加权混合。当LLM对某个补全内容的自信程度极高,认为图中缺失了该链接时,允许模型突破图的限制。修正后的Logit计算式为:

$$\text{Logit}(w_{z_i}) = \text{Logit}_{\text{LLM}}(w_{z_i}) + \alpha \log(M(w_{z_i}) + \epsilon) \quad (3)$$

其中, $\text{Logit}_{\text{LLM}}(w_{z_i})$ 是LLM根据上下文预测的原始分数, α 是约束强度系数, ϵ 是一个极小的平滑项,此处设定为 10^{-5} 。关于软约束强度系数 α ,本文实验发现 $\alpha = 2.0$ 可在准确率与幻觉率之间取得最佳平衡。在实际部署中,若安全分析师对幻觉容忍度更低,可适当增大 α ;若图谱缺失严重,可适当降低 α 。正常情况下($M = 1$),图谱路径被优先选择,保证无幻觉。当图中路径缺失($M = 0$),但LLM根据语义(如APT28通常使用X-Agent)给出的 $\text{Logit}_{\text{LLM}}(w_{z_i})$ 非常高时,高分值可以抵消 $M = 0$ 带来的惩罚,从而生成推断性的跳数。这种机制使得模型能够处理开放世界假设,允许模型在知识图谱中没有显式链接时,基于已有的知识和上下文推测合理的链接。为了在后续分析中严格隔离事实观测与语义推断,在解码生成的路径中引入了显式的标签机制。具体而言,在生成推理路径时,系统会实时追踪掩码 $M(w_{z_i})$ 的状态:若生成的节点或关系依赖于图谱中已存在的有效路径($M = 1$),系统会自动为其附加事实观测标签;若模型依赖高置信度语义突破了图约束而生成图中缺失的跳数($M = 0$),则强制附加上语义推断标签。在推理路径的实际存储中,使用JSON格式记录每一步的来源标签:{"step": "Malware Hash a7b9...", "source": "fact"}。这些结构化标签在后续归纳聚合阶段被显式传递给通用LLM,作为区分硬证据与推断结论的依据。这种结构化的隔离机制不仅保障了推理过程的透明可解释,也为下游的安全审查和多路径校验提供了明确的锚点。

软约束机制在提升开放世界鲁棒性的同时,也引入了潜在的安全风险。攻击者可能通过构造高置信度的虚假语义先验,诱导LLM在解码阶段突破图约束,生成看似合理但实际不存在的攻击路径。为缓解此类风险,TIKGCR采用三重防御策略:置信度阈值截断(仅当 $\text{Logit}_{\text{LLM}}(w_{z_i}) > 0.85$ 时允许突破约束);多路径交叉验证,若推断链路被少于2条独立路径支持,则自动降级为低置信度;人工审核接口所有含语义推断标签的步骤在

最终归因报告中高亮标记，供分析师复核。未来工作将探索对抗训练与图约束的联合鲁棒增强。

2.4 归纳溯源与证据聚合

图约束解码推理阶段虽能生成符合图谱结构与攻击逻辑的推理路径，但在复杂网络攻击场景中，攻击者通常会构建多条平行攻击链路或隐藏关键攻击节点，单一路径难以完整覆盖攻击全貌，正确的规因结论往往通过多条推理路径共同推导得出。此外，威胁情报数据的天然碎片化特性，也可能导致单条推理路径存在证据缺口。为此，本文引入多路径归纳溯源机制与分层证据聚合框架，通过对多条候选路径的归纳推理与证据整合，实现更完整、更可靠的攻击归因。针对2.3节中被标记为语义推断的推断性跳数，在此阶段设计了基于多路径交叉验证的校验机制。当通用LLM接收到多条带有显式标记的候选路径后，会重点对推断节点进行审查。

①共识校验：若多条基于不同起点的平行路径均推断出同一缺失链路，该推断将被视为得到交叉印证，其置信度权重将被提升；②逻辑一致性校验：LLM会评估单条路径中的推断步骤是否与其余路径中的确定性事实存在战术逻辑上的冲突。对于孤立出现且缺乏其他侧面证据佐证的单路径推断，聚合机制将降低其权重或予以剔除。通过这种全局视角的交叉校验，框架有效平抑了单一路径推断可能带来的偏差，确保了最终归因结论的严谨性。

本文参考路径思维（PoT）框架^[21]，将多条推理路径 $\mathcal{W} = \{w_1, w_2, \dots, w_k\}$ 整合到一个自然语言序列中，连同原始问题一起输入给一个通用的LLM，通用LLM可以是任何强大的LLM，如Gemini 3 Pro，它们可以同时分析所有提供的候选路径，结合其内置的常识知识和隐含在路径中的逻辑，比较和评估不同路径的合理性和潜在结论，综合归纳所有信息最终推断出最可能成立的答案，该过程提示模板如图3所示。将所有候选路径一次性提供给LLM，使其能够在一个“认知窗口”内比较、权衡和综合信息，做出更全面的判断。当存在歧义或LLM提取错误导致冲突路径时，LLM有机会看到所有选项并选择最合理的。这比仅基于单条路径或原始

文本推理更鲁棒。这种设计充分利用了LLM强大的自然语言理解、模式识别和基于上下文的归纳推理能力，让它直接处理结构化路径序列和非结构化信息的混合输入。

图归纳推理提示模板

```

**[Input Data]**
**Reasoning Paths:**
1. <Path_1> -> <Hypothesis_1>
...
k. <Path_k> -> <Hypothesis_k>

**Question:** <User_Question>

**[Instruction]**
Based on the provided reasoning paths, perform the following:
1. Discard paths dependent on weak evidence (e.g., pure geolocation).
2. Synthesize strong paths (e.g., C2 infrastructure, malware signatures).
3. Determine the final Threat Actor and your confidence level.

**[Output]**
Final Answer: [Actor Name] Confidence: [Score]
Evidence Chain: <Attribution Start Point>-><Intermediate Entity 1>->
<Intermediate Entity x>-><Attack Organization>
Analysis:...

```

图3 图归纳推理提示模板

3 实验

本节旨在通过实验回答以下3个研究问题。

RQ1: TIKGCR在网络威胁归因任务上的效率是否优于现有的主流方法？RQ2：在威胁情报数据缺失的情况下，引入“软约束”机制能否有效维持溯源推理的性能？RQ3：攻击链模式过滤对减少搜索空间有何贡献？

3.1 实验设置

由于缺乏公开的标准威胁溯源问答数据集，本文构建了一个基于真实CTI报告的评估数据集“CTI-Trace”。从Mandiant选取的500份CTI报告中提取数据，仅保留含完整攻击链与明确归因结论的报告。利用本文方法提取实体和关系构建威胁情报知识图谱，包含约15 000个实体和40 000条边。同时，基于CTI报告结论，采用“双标注+审核”机制，两位标注人员独立验证自动抽取结果，不一致项由资深安全分析师仲裁。同时由两名独立标注者根据CTI报告结论构建Question/Answer对。例如，Question：“检测到内网主机192.168.1.5连接恶意域名x.com，这是哪个组织的攻击活动？”，Answer：“APT28 (Fancy Bear)”。为模拟威胁情报数据缺失的情况，随机删除了威胁情报知识图谱中30%的连边，模拟真实环境中日志缺失的情况。

本文将 TIKGCR 与以下 3 类方法进行对比：基于 LLM 的方法、基于检索增强生成 (RAG) 的方法、基于 LLM+KG 的方法。所有涉及知识图谱检索或推理的方法均在完全一致的 TIKG 测试图谱 (含 30% 随机边缺失) 上运行, 以评估各模型处理碎片化情节的能力。图神经网络增强的检索增强生成 (GNN-RAG) 和 KG-Agent 等面向基础知识图谱问答 (KGQA) 任务的方法, 将 TIKG 作为其外部检索源, 将检索到的高置信度三元组按照原论文的提示 (Prompt) 模板输入给 LLM, 辅助执行归因决策。鉴于 HGHAN 原为异构图节点嵌入模型, 将其转化为排序任务: 首先计算溯源起点与图中所有攻击者节点的嵌入相似度, 选取 Top-K 候选组织及其关键特征路径, 并接到 Prompt 中引导 LLM 生成最终结论。

本文采用以下 5 种指标评估。①准确率 (accuracy, Acc): 仅当归因组织名称与原始报告一致。②首位命中率 (HITS@1): 排序第一的答案是否是正确答案。③幻觉率 (hallucination rate, HR): 若生成路径中存在既不在 KG 中、又无公开情报支撑的实体/关系, 则计为 1 次幻觉。④路径完整度 (path completeness, PC): 生成的证据链覆盖了多少关键中间节点; ⑤路径精确度 (path precision, PP): 生成路径中有多少跳与基准真值 (ground truth) 攻击链一致。

本文使用微调的 Llama-3.1-8B 模型进行实体和关系抽取构建威胁情报知识图谱, 同时对 Qwen3-8B 模型进行微调作为生成推理路径的专用 LLM, 在推理阶段使用通用 LLM Gemini 3 Pro 来进行归纳推理。知识抽取阶段使用的微调数据集是 mrmoor 数据集, 这是一个专门用于威胁情报命名体识别的数据集, 包含实体和关系的标注信息。推理阶段使用生成的图谱数据作为微调数据集。微调过程在 2 张 A100-80 GB GPU 上进行, 最大训练轮次为 10 轮, 批量大小为 4, 学习率为 3×10^{-5} , 用时分别为 13.75 h 和 14.23 h。

3.2 实验结果分析

表 3 为不同方法在 CTI-Trace 数据集上的表现。从表 3 可以看出, 直接使用 LLM 虽然准确率和首位命中率表现不错, 但出现了严重的幻觉现象, 经常编造不存在的哈希值或者 IP。使用传统

的 RAG 虽然可以有效缓解该现象, 但首位命中率和准确率较直接使用 LLM 略有降低。使用 LLM+KG 组普遍拥有较高的首位命中率和准确率, 同时都能一定程度上缓解幻觉现象。KG-Agent^[12]和 HGHAN^[9]只是简单的组合 LLM 和 KG, 因此效果在融合社区检测的大语言模型-知识图谱 (LLM+KG) 组中是最差的, 融合社区检测的大语言模型-知识图谱 (LLM+KG+Community)^[17]和 GNN-RAG^[11]分别使用社区检测算法和 RAG 来组合 LLM 和 KG, 取得的效果更好。而使用 GCR 时, 虽然完全消除了幻觉现象, 但由于测试图谱存在 30% 的缺失, 很多路径无法走通, 准确率受限。TIKGCR 取得了最高的准确率和首位命中率, 虽然引入了少量的幻觉 (源于软约束允许的推断), 但它成功补齐了缺失路径, 权衡后的整体效果更佳, SR+NSM+E2E 指步进检索-神经语义匹配-端到端, GPT-4o-mini 为轻量化大语言模型。

表 3 不同方法在 CTI-Trace 数据集上的表现

类型	方法	HITS@1	Acc	HR
LLM	GPT-4o-mini	72.5%	68.0%	25.4%
RAG	SR+NSM+E2E	70.8%	65.4%	5.8%
LLM+KG	KG-Agent	79.2%	73.8%	15.3%
	HGHAN	83.0%	77.8%	16.6%
	LLM+KG+Community	86.3%	80.7%	10.4%
	GNN-RAG	88.1%	87.3%	3.4%
	GCR	90.7%	88.2%	0
本文	TIKGCR	97.3%	94.6%	1.5%

基于 TIKGCR 和 GCR 对比的实验结果, 使用 McNemar 检验评估了 HITS@1 指标和 Acc 指标的统计显著性。在 HITS@1 指标评估中, $n_{01}=57$, $n_{10}=24$, $\chi^2=13.44$, $p=2.7 \times 10^{-4}$, 这表明在显著性水平 $\alpha=0.05$ 下, 该提升具有统计学意义。

在 Acc 指标评估中, $n_{01}=38$, $n_{10}=4$, $\chi^2=28.96$, $p=2.3 \times 10^{-5}$, 这表明在显著性水平 $\alpha=0.05$ 下, 该提升具有统计学意义。基于 TIKGCR 和 GNN-RAG 对比的实验结果, 利用 Wilcoxon 符号秩检验评估了 HR 指标的显著性。TIKGCR HR 中位数 = 1.5%, GNN-RAG = 3.4%, Wilcoxon $Z=-2.76$, $p=0.0061$, 效应量 (Cliff's Delta) = -0.28, 表明两种方法在幻觉率指标上的差异具有统计学显著性。

为了验证框架在不熟悉或外部标准数据上的有效性,表4主要对比了不同方法在TRAIL数据集^[22]上的表现。结果表明,在TRAIL数据集上TIKGCR依然以92.8%的准确率领先。这充分证明了TIKGCR的知识图谱约束与动态解码机制不依赖于特定的自建数据分布,具有较好的通用性和外部可参照性,为高风险网络安全场景提供了可靠的开箱即用能力。

表4 不同方法在TRAIL数据集上的表现对比

类型	方法	HITS@1	Acc	HR	PC	PP
LLM	GPT-4o-mini	68.5%	65.3%	24.8%	60.1%	64.7%
RAG	SR+NSM+E2E	65.9%	62.7%	5.3%	57.5%	61.9%
LLM+KG	KG-Agent	75.6%	72.5%	16.2%	68.1%	71.7%
	HGHAN	77.0%	75.3%	17.4%	71.9%	74.8%
	LLM+KG+Community	80.1%	78.8%	9.8%	75.0%	77.4%
	GNN-RAG	85.4%	84.1%	3.1%	81.4%	83.6%
	GCR	88.6%	87.9%	0	85.8%	88.1%
本文	TIKGCR	91.4%	92.8%	1.8%	89.5%	92.0%

为验证TIKGCR的跨模型稳定性,在3类主流开源LLM上复现了主实验,未对框架核心逻辑做任何修改。TIKGCR的跨模型稳定性实验结果如表5所示,各模型与均值的平均绝对偏差为0.4%,幻觉率偏差 $\leq 0.4\%$,软约束触发决策高度一致。这证明TIKGCR的约束机制不依赖特定LLM的固有特性,而是通过KG-Trie索引与置信度阈值实现了通用的可信推理增强,可适配不同LLM推理引擎。

表5 TIKGCR的跨模型稳定性实验结果

模型	Acc	HR	HITS@1	平均推理时间/s
Qwen3-8B	94.6%	1.5%	97.3%	1.21
Llama-3.1-8B	93.9%	1.8%	96.8%	1.32
Gemma-3-9B	93.7%	1.9%	97.1%	1.45
平均	94.1%	1.7%	96.5%	1.33

为了验证软约束的核心价值,测试了在不同图谱缺失率下,直接使用GCR和TIKGCR的性能变化,如图4所示。当图谱完整(缺失率=0%)时,硬约束和软约束表现几乎一致。随着缺失率的增加,原始GCR的性能因推理路径中断而迅速下降。相比之下,TIKGCR通过归纳推

理聚合多条部分推理路径,保持了稳定的性能。在30%的缺失率下,TIKGCR的绝对准确率比基线高出6.4%。随着缺失率增加达到50%时,原始GCR的准确率急剧下降(从88.2%跌至56.0%),因为它受限于断裂的路径无法推理。TIKGCR的性能下降非常平缓(从94.6%仅跌至85.5%)。这证明了在高置信度下允许LLM进行语义补全能有效应对网络安全中的数据碎片化问题。

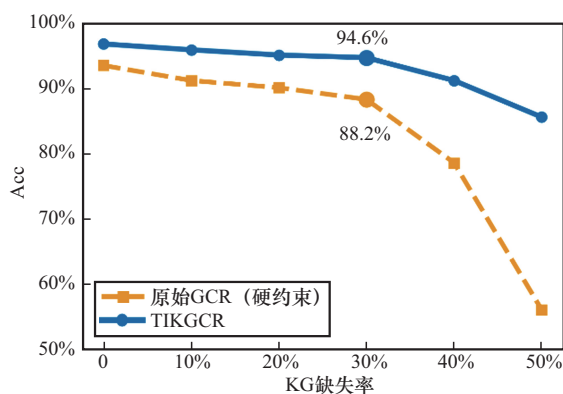


图4 不同知识图谱缺失率下的鲁棒性对比

为严谨确定软约束机制中的关键超参数 α ,通过网格搜索和加权优化明确选择过程,如表6所示。触发率(Trigger Rate)表示触发软约束的比例,能够揭示过拟合语义的风险边界,同时确定使 $\text{Acc} - \lambda \cdot \text{HR}$ (加权性能)最大化的 α^* ,其中 λ 为幻觉惩罚权重(此处 $\lambda = 5$,体现安全场景对幻觉的高敏感性)。实验分3次独立运行,报告均值 $\pm$ 标准差。

表6 α 系数敏感性分析与最优值选择实验

α	Acc	HR	Trigger Rate	$\text{Acc} - 5 \cdot \text{HR}$
0.5	91.2% $\pm$ 0.4%	2.8% $\pm$ 0.3%	12.1%	90.0
1.0	92.7% $\pm$ 0.3%	2.1% $\pm$ 0.2%	28.6%	91.7
1.5	93.9% $\pm$ 0.2%	1.7% $\pm$ 0.1%	41.3%	93.1
2.0	94.6% $\pm$ 0.1%	1.5% $\pm$ 0.1%	52.8%	93.9
2.5	94.2% $\pm$ 0.2%	2.3% $\pm$ 0.2%	63.5%	93.0
3.0	93.5% $\pm$ 0.3%	3.4% $\pm$ 0.3%	78.2%	92.0

为更直观地呈现约束强度系数 α 对归因性能的影响规律,图5绘制了准确率与幻觉率随 α 变化的趋势曲线。可以观察到,在 α 由0.5提升至2.0的过程中,随着图约束强度的增加,模型归因准确率持续上升,幻觉率同步下降;当 α 超过

2.0 后, 过强的结构约束开始抑制模型对缺失链路的语义补全能力, 准确率出现回落, 同时因约束过度截断有效路径引发的推断偏差反而导致幻觉率有所回升。两条曲线的最优交汇点位于 $\alpha=2.0$ 处, 与表 6 中加权性能指标的计算结果一致, 验证了该超参数取值的合理性。

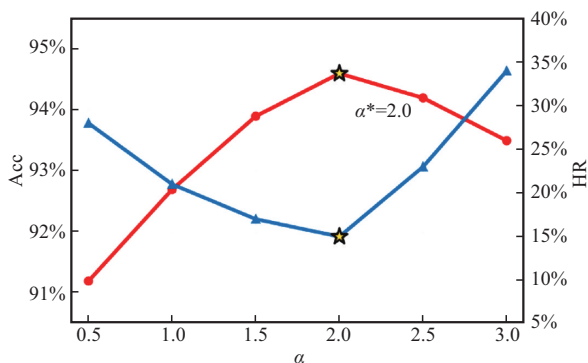


图5 Acc和HR与 α 的关系

3.3 消融实验

本文对 TIKGCR 的两个关键改进组件进行了消融测试, 消融实验结果如表 7 所示。

表 7 消融实验结果

实验变体	Acc	平均推理时间/s
TIKGCR	94.6%	1.21
w/o pattern filtering	91.8%	3.57
w/o soft constraint	88.2%	1.14

从表 7 可以看出, 当去除模式过滤 (w/o pattern filtering) 时准确率变化不大, 但推理时间增加了近 3 倍。这说明基于攻击链的模式过滤极大地剪枝了无关路径, 如无关的行政归属关系, 提升了效率。去除软约束 (w/o soft constraint) 时准确率显著下降, 回退到了原始 GCR 的水平, 再次印证了软约束在处理非完美数据时的必要性。需要强调的是, 通用大模型在 TIKGCR 框架中仅作为最终的静态评估器参与归纳溯源阶段, 其不具备图谱访问权限, 也不参与任何攻击路径的生成与探索。真正的核心推理能力与性能增益, 完全来源于本文微调的专用推理模型 Qwen3-8B 及其配套的自适应软约束机制。结合表 7 可以看出, 当去除软约束机制时, 尽管后端的 Gemini 聚合模块依然存在, 系统的准确率仍断崖式下跌

至 88.2%。这充分证明, 微调模型的软约束机制为后端提供了有效的候选证据链; Gemini 仅在此基础上承担了多路径的一致性校验与冗余过滤工作。因此, 性能的突破性提升本质上归功于软约束机制, 而非通用大模型的本身能力。

3.4 系统复杂度与部署可行性分析

为评估 TIKGCR 的实际部署潜力, 本文对其运行效率、资源消耗及典型应用场景进行了定量分析。如图 6 所示, 在 2 张 A100-80 GB GPU 环境下, 使用 vLLM 推理引擎, 输入提示长度约为 512 token, 束宽 (beam size) = 5, 端到端溯源推理平均耗时为 1.21 s, GNN-RAG 需多轮图查询+LLM 调用, 实测均值为 2.47 s、KG-Agent 需多轮交互, 均值为 3.92 s, 说明 TIKGCR 因解码期约束替代图遍历, 显著降低了时延。

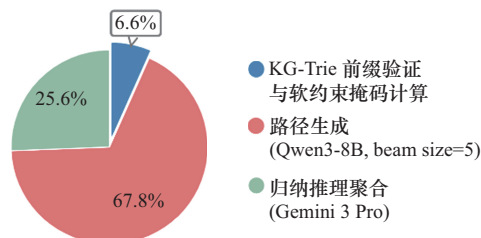


图6 端到端溯源推理时延分解

微调阶段威胁情报抽取模型 (Llama-3.1-8B) 全参数微调: 2×A100-80 GB, 批次大小 (batch) = 4, 最大训练步数 (max_steps) = 2 000, 耗时 13.75 h, 峰值显存 78 GB; 路径生成模型 (Qwen3-8B) LoRA 微调 ($r=64, \alpha=128$): 同配置下耗时 4.2 h, 峰值显存 42 GB。CTI-Trace 图谱含 15 000 个实体/40 000 个三元组, 从原始 500 份 CTI 报告中抽取, 自然语言处理流水线 (NLP pipeline) 平均处理速度为 12 report/min, 总构建时间约为 42 min。TIKGCR 支持分层部署: 核心路径生成模块可嵌入本地安全运营中心 (SOC) 平台实现亚秒级响应; 归纳推理模块可按需调用云端大模型以平衡精度与成本。在模拟企业级环境 (2 核 CPU+1× RTX 4090) 中, TIKGCR 仍能维持 2.3 s 的平均推理时延, 满足日常告警响应需求。

3.5 案例研究

本文通过一个案例来说明 TIKGCR 在真实环境中消除幻觉和可信推理的能力, 如表 8 所示。

表 8 案例分析

案例：基于缺失链路补全的 APT29 供应链攻击归因	
Question	内网主机 192.168.1.105 被检测到向域名 sys-update-service.com 发起了异常的心跳连接。请分析该活动的归属组织
Answer	APT29
GPT-4o-mini	该域名看起来像是一个合法的系统更新服务。如果它是恶意的，可能是由 Lazarus Group 使用的，因为它们经常通过伪造服务进行攻击
w/o soft constraint	无法判定/未找到有效推理路径
TIKGCR	归属组织：APT29（高置信度）。 证据链：192.168.1.105→sys-update-service→C2 IP（45.13.0.1）→恶意软件哈希值（a7b9...）→APT29。 分析：虽然直接关联缺失，但该哈希展现出 WellMess 恶意软件的特征，该软件是 APT29 的专用工具

为模拟真实环境，删除了 Hash: a7b9... ^{belongs_to} → APT29 这条边，因为新的恶意软件变种可能还没被及时打上归属标签。此案例展示了在威胁情报往往零散的真实场景中，TIKGCR 的优越性。无约束直出大型语言模型基于表层域名产生幻觉，而去除软约束时则因图形连接不完整而失败，说明本文自适应软约束机制成功弥合了知识鸿沟。它利用大型语言模型的语义推理来推断恶意软件哈希值与威胁行动者之间缺失的链接，从而提供准确且可解释的归因追踪。

4 结束语

尽管大语言模型在理解复杂的网络威胁报告方面表现出非凡的能力，但其在攻击溯源任务中仍面临严峻挑战。现有的生成式方法往往受困于“幻觉”，凭空捏造不存在的攻击链路或横向移动步骤，导致归因结论缺乏证据支撑。为此，本文提出了一种 TIKGCR 框架，旨在解决网络安全场景中威胁情报数据碎片化、传统图约束推理对不完整数据鲁棒性不足的问题。通过 3 阶段设计，TIKGCR 实现了高效、可信的攻击归因。TIKGCR 在威胁归因任务上准确率达 94.6%，幻觉率仅 1.5%，显著优于 GNN-RAG（87.3%）、传统 GCR（88.2%）等基线方法。所提置信度感知软约束机制可动态平衡图结构约束与语义补全。在 50% 图谱缺失率下，TIKGCR 准确率仍保持 85.5%，远高于硬约束 GCR 的 56.0%，验证了其碎片化 CTI 数据的适应性。基于攻击链模式的 KG-Trie 索引通过剪枝无关路径，使推理时间缩短约 $\frac{2}{3}$ ，解决了图遍历的效率瓶颈。生成的推理链锚定于 KG 结构，结合集束（beam search）

路径聚合与专业提示工程，为安全分析师提供可审计的攻击路径证据，直接服务于溯源决策。未来工作将进一步探索多模态威胁情报的 KG 融合，并研究动态攻击模式下的 KG-Trie 索引更新机制，以提升框架对新型攻击的适应性。TIKGCR 为网络威胁归因任务提供了“结构可信+语义鲁棒”的解决方案，为高风险场景下的 LLM 可信推理与威胁情报自动化分析奠定了技术基础。

TIKGCR 在路径索引构建与约束推理阶段利用了 MITRE ATT&CK 等既有攻击模式作为先验知识，旨在通过模式导向的 KG-Trie 索引有效剪枝无效搜索空间、提升推理效率与路径合理性。然而，现实中的新型攻击，如“低慢小”隐蔽攻击、跨域混合战术或对抗性规避行为，可能不完全符合现有 TTP 分类体系，这构成了本文方法的潜在适用性边界。

为增强对未知攻击的适应性，本文的未来工作中将探索以下方向。①动态模式演化机制：通过在线增量学习，将新检测到的异常路径聚类为潜在新模式，并自动注入 KG-Trie；②零样本模式发现模块：结合图神经网络对未标注路径进行异常检测与聚类，辅助识别“非常规但高置信”的攻击子图；③多源对抗性训练：在微调阶段引入合成对抗样本，如故意打乱 TTP 顺序的虚假报告，提升模型对模式扰动的鲁棒性。

参考文献：

- [1] Rani N, Saha B, Shukla S K. A comprehensive survey of automated advanced persistent threat attribution: taxonomy, methods, challenges and open research problems[J]. Journal of Information Security and Applications, 2025, 92: 104076.

- [2] Chang Y P, Wang X, Wang J D, et al. A survey on evaluation of large language models[J]. *ACM Transactions on Intelligent Systems and Technology*, 2024, 15(3): 1-45.
- [3] Huang L, Yu W J, Ma W T, et al. A survey on hallucination in large language models: principles, taxonomy, challenges, and open questions[J]. *ACM Transactions on Information Systems*, 2025, 43(2): 1-55.
- [4] Pan S R, Luo L H, Wang Y F, et al. Unifying large language models and knowledge graphs: a roadmap[J]. *IEEE Transactions on Knowledge and Data Engineering*, 2024, 36(7): 3580-3599.
- [5] Wei J, Wang X Z, Schuurmans D, et al. Chain-of-thought prompting elicits reasoning in large language models[C]//*Proceedings of the 36th International Conference on Neural Information Processing Systems*. New York: ACM Press, 2022: 24824-24837.
- [6] Wang X, Wei J, Schuurmans D, et al. Self-consistency improves chain of thought reasoning in language models[C]//*Proceedings of International Conference on Learning Representations(ICLR)*. 2023.
- [7] Cao Y, Griffiths T, Narasimhan K, et al. Tree of thoughts: deliberate problem solving with large language models[C]//*Proceedings of the Advances in Neural Information Processing Systems 36*. Neural Information Processing Systems Foundation, Inc. (NeurIPS), 2023: 11809-11822.
- [8] Phan D, Hoffman M D, Dohan D, et al. Training chain-of-thought via latent-variable inference[C]//*Proceedings of the 37th International Conference on Neural Information Processing Systems*. New York: ACM Press, 2023: 72819-72841.
- [9] Huang J, Chen X, Mishra S, et al. Large language models cannot self-correct reasoning yet[C]//*Proceedings of International Conference on Learning Representations(ICLR)*. 2024.
- [10] 夏毅, 兰明敬, 陈晓慧, 等. 可解释的知识图谱推理方法综述[J]. *网络与信息安全学报*, 2022, 8(5): 1-25.
Xia Y, Lan M J, Chen X H, et al. Survey on explainable knowledge graph reasoning methods[J]. *Chinese Journal of Network and Information Security*, 2022, 8(5): 1-25.
- [11] Chen Y, Sun S, Hu X C. DRKG: faithful and interpretable multi-hop knowledge graph question answering via LLM-guided reasoning plans[J]. *Applied Sciences*, 2025, 15(12): 6722.
- [12] Mavromatis C, Karypis G. GNN-RAG: graph neural retrieval for large language model reasoning[PP]. *arXiv Preprint*, 2024: arXiv. 2405.20139.
- [13] Jiang J H, Zhou K, Zhao W X, et al. KG-agent: an efficient autonomous agent framework for complex reasoning over knowledge graph[C]//*Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Ker-rville: Association for Computational Linguistics, 2025: 9505-9523.
- [14] Sun J S, Xu C J, Tang L, et al. Think-on-graph: deep and responsible reasoning of large language model on knowledge graph[PP]. *arXiv Preprint*, 2024: arXiv.2307.07697.
- [15] Luo L, Zhao Z, Haffari G, et al. Graph-constrained reasoning: faithful reasoning on knowledge graphs with large language models[C]//*Proceeding of the 42nd International Conference on Machine Learning(ICLR)*. 2025: 41540-41565.
- [16] 黄克振, 连一峰, 冯登国, 等. 一种基于图模型的网络攻击溯源方法[J]. *软件学报*, 2022, 33(2): 683-698.
- Huang K Z, Lian Y F, Feng D G, et al. Method of cyber attack attribution based on graph model[J]. *Journal of Software*, 2022, 33(2): 683-698.
- [17] Xu Y J, Fang Y, Huang C, et al. HGHAN: hacker group identification based on heterogeneous graph attention network[J]. *Information Sciences*, 2022, 612: 848-863.
- [18] Zang X D, Gong J, Zhang X C, et al. Attack scenario reconstruction via fusing heterogeneous threat intelligence[J]. *Computers & Security*, 2023, 133: 103420.
- [19] Zhang J, Zhang X K, Yu J F, et al. Subgraph retrieval enhanced model for multi-hop knowledge base question answering[C]//*Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Kerrville: Association for Computational Linguistics, 2022: 5773-5784.
- [20] 沈晓晨, 葛寅辉, 陈波, 等. 人工智能安全知识图谱构建技术研究[J]. *网络与信息安全学报*, 2023, 9(2): 164-174.
Shen X C, Ge Y H, Chen B, et al. Research on construction technology of artificial intelligence security knowledge graph[J]. *Chinese Journal of Network and Information Security*, 2023, 9(2): 164-174.
- [21] Zhang G, Ali Alomrani M, Gu H J, et al. Extracting and following paths for robust relational reasoning with large language models [PP]. *arXiv Preprint*, 2026: arXiv. 2412.17963.
- [22] King I J, Ramirez R, Bowman B, et al. Trail: a knowledge graph-based approach for attributing advanced persistent threats[C]//*Proceedings of the 2025 IEEE 41st International Conference on Data Engineering (ICDE)*. Piscataway: IEEE Press, 2025: 1207-1220.

[作者简介]



刘昌城 (2001-), 男, 信息工程大学硕士生, 主要研究方向为网络空间安全与人工智能。



杨长恒 (2001-), 男, 信息工程大学硕士生, 主要研究方向为网络安全与知识图谱。



马骏 (1981-), 男, 博士, 信息工程大学副教授, 主要研究方向为态势感知与威胁发现。