引用格式:杨国正,陈驰昱,沈照斌,等.一种基于大语言模型及 RAG 技术的节点设备类型识别方法[J]. 信息对抗技术,2025, 4(5):42-53.
[YANG Guozheng, CHEN Chiyu, SHEN Zhaobin, et al. A node device-type identification method based on large language models and retrieval-augmented generation[J]. Information Countermeasure Technology, 2025, 4(5):42-53.(in Chinese)]

一种基于大语言模型及 RAG 技术的节点设备类型识别方法

杨国正[1,2],陈驰昱[1*],沈照斌[1,2],齐冬震[1],潘俊宇[1]

(1. 国防科技大学电子对抗学院,安徽合肥 230037; 2. 安徽省网络空间安全态势感知与评估重点实验室,安徽合肥 230037)

摘 要 在大规模网络空间测绘中,快速准确探测节点信息并识别设备运行状态,是核心研究内容之一。当前,网络空间设备版本迭代速度加快,大量新型设备不断涌现,如何跟踪并识别被测节点的设备类型,成为亟待解决的新挑战。针对当前研究过于依赖已有知识,无法适应设备升级变化的问题,提出了一种基于大语言模型(large language model, LLM)和检索增强生成(retrieval-augmented generation, RAG)技术的节点设备类型识别方法。首先,从 RFC 文档和互联网设备厂商站点收集相关资料,基于嵌入模型构建知识向量数据库;然后,对探测得到的节点特征信息进行编码,从向量数据库中检索相关背景知识,将其与节点特征信息共同构造为提示词并输入 LLM,利用其推理能力实现对被测节点的设备类型识别;最后,通过消融实验和实网测试,验证了该方法的有效性和性能。

关键词 网络测量;LLM;RAG;设备类型识别

中图分类号 TP 393 **文章编号** 2097-163X(2025)05-0042-12

文献标志码 A **DOI** 10.12399/j.issn.2097-163x.2025.05.003

A node device-type identification method based on large language models and retrieval-augmented generation

YANG Guozheng[1,2], CHEN Chiyu[1*], SHEN Zhaobin[1,2], QI Dongzhen[1,2], PAN Junyu[1]

(1. College of Electronic Engineering, National University of Defense Technology, Hefei 230037, China;
2. Anhui Province Key Laboratory of Cyberspace Security Situation Awareness and Evaluation, Hefei 230037, China)

Abstract In large-scale cyberspace mapping, rapidly and accurately detecting node information and identifying the operational status of devices is one of the core research contents. Currently, the version iteration speed of cyberspace devices is accelerating, and a large number of new-type devices are constantly emerging. How to track and identify the device type of the measured node has become a new challenge that needs to be solved urgently. Aiming at the problem that current research relies too much on existing knowledge and cannot adapt to device upgrade changes, a node device type identification method based on large language model (LLM) and retrieval-augmented generation (RAG) technology was proposed. First, relevant data were collected from RFC documents and Internet device manufacturer websites, and a knowledge vector database was constructed based on the embedding model. Then, the detected

收稿日期:2025-07-07 **修回日期**:2025-07-24

通信作者:陈驰昱,E-mail:chenchiyu14@nudt.edu.cn

基金项目:国家自然科学基金资助项目(62271496)

node feature information was encoded, and relevant background knowledge was retrieved from the vector database. The retrieved knowledge and node feature information were jointly constructed into prompts for the LLM. The reasoning ability of the LLM was used to identify the device type of the probed node. Finally, the effectiveness and performance of the method were verified through ablation experiments and real-network tests.

Keywords　cyberspace mapping; LLM; RAG; device type identification

0　引言

随着互联网技术与应用的快速发展，网络和终端设备的种类和数量不断增加，如何在大规模网络空间测绘中，快速准确识别被测节点的设备类型成为新的挑战。目前，针对节点设备类型的识别方法主要有基于指纹库和基于机器学习模型 2 类。基于指纹库的方法需要预先构建各种设备的探测指纹库，再通过类似正则表达式的方式匹配探测标识符实现设备类型识别。基于机器学习模型的方法主要通过标签数据训练特定的机器学习模型，根据探测节点特征自动判断所属设备类型。这 2 类方法均适用于识别已知设备类型，其识别能力依赖于人工构建的指纹库或者训练使用标签数据的质量。然而，随着网络空间设备版本升级节奏的加快和大量新设备的涌现，这些方法在识别新设备方面存在较大局限。

近年来，随着大语言模型（large language model，LLM）的快速发展，基于自然语言处理的设备类型识别方法逐渐受到关注。LLM 擅长从文本中提取复杂信息，并在无监督学习的基础上展现出强大的推理能力。针对大规模网络测绘中的设备类别识别这一问题，本文提出了一种基于 LLM 和检索增强生成（retrieval-augmented generation，RAG）技术的节点设备类别识别方法。该方案通过采集 RFC（request for comments）文档和互联网设备厂商网页信息，构建网络协议和设备品牌知识向量数据库；在具体分析探测节点设备类型时，从向量数据库中检索相关背景知识，将其精心构造为提示词，并输入 LLM，利用 LLM 的推理能力完成设备类型识别，最后通过实验验证该方法的有效性。

1　相关工作

在网络空间测绘领域，对测量得到的节点进行设备类型识别主要基于设备指纹技术（device fingerprinting，DFP）[1]，即利用设备进行网络通信的各层面特征生成指纹（或签名），并借助这些指纹（或签名）实现对设备的识别或分类，这里的设备包括网络设备、嵌入式设备和物联网设备等类型。根据特征获取、指纹生成和识别方式等过程的差异，设备类型识别技术可被划分为基于规则的方法和基于机器学习模型的方法[2]。

1.1　基于规则的识别方法

著名网络扫描工具 Nmap（network mapper）的配套特征文件和 Rapid7 安全公司的开源项目 Recog[3]，都是典型的基于规则的指纹库项目。前者在脚本库中包含了大量服务和设备的指纹匹配规则；后者利用开源社区贡献，持续补充包括图标、响应和协议字段等形式的各类指纹信息，涉及操作系统、供应商、硬件和软件信息等方面内容。这类传统的规则匹配方式虽然识别准确度较高，但需要大量专家知识和人工参与，并且无法识别没有入库的未知设备。

为了降低指纹（规则）的构建成本，一些基于自动化构建规则的识别方法应运而生。LI 等[4]发现，物联网设备的应用层响应数据通常包含与其制造商高度相关的内容，通过爬取关于设备描述的网页，并使用实体命名识别提取设备标注，提出了一种基于规则的物联网设备发现与识别方法。基于类似的策略，规则采集引擎 ARE[5]利用来自物联网设备的应用层响应数据和相关网站中的产品描述来构建设备规则，以“类型—供应商—产品”的三元组形式生成物联网设备标注规则。然而，JAVED 等[6]发现基于相同原理实现的 ARE 引擎无法达到其宣称的精度，这反映出此类自动化的规则生成方案不仅设计复杂，而且在面对未知设备时存在泛化能力较弱的问题。

1.2　基于机器学习模型的识别方法

CHENG 等[7]从 HTTP 响应头部及响应体中提取各类统计特征，通过有监督的机器学习模型，在不需要人工指纹库支持的情况下，可以达到 97.5%的准确率。WebIoT 框架[8]将提取网页的图像特征和统计特征结合为设备特征作为神经网络的输入，在含 20 万条物联网设备的标注数

据集上进行训练，也取得了较好的精度。但无论是机器学习还是深度学习模型的训练，都需要大量标签数据，且训练成本较高。

2023 年，SARABI 等[9]意识到由互联网扫描生成的大量文本数据非常适合于训练 LLM，因此使用数亿条应用层横幅训练了一个基于 Transformer 的掩码语言模型，并设计对比损失函数进行微调使其能够生成时间上稳定的横幅嵌入，同时保持功能相似的硬件/软件产品嵌入在向量空间中距离相近。通过该模型可以对 HTTP 横幅进行聚类分析，并为每个聚类生成基于文本的指纹。LLM 虽然可以有效生成具有语义的嵌入并形成聚类，但由于训练数据本身不具备标签信息，无法自动对聚类标识，导致生成的指纹表征无法识别具体的设备类型。

除上述研究外，当前研究在如何发现互联网中新出现的特定类型设备方面进行了相关尝试。UEDA 等[10]通过在车载设备网站中提取关键字并在 Censys 中匹配和聚类 Web 页面，发现了 12 种暴露在互联网上的车载设备。SASAKI 等[11]利用一般远程管理设备 Web 页面间的相似性和一般网站 Web 页面之间的异构性，精心设计了一套迭代流程来寻找互联网中的工业控制系统(industrial control systems，ICS)管理设备，同时借助 ICS 远程管理设备 Web 页面中普遍存在的特征信息，自动标注设备型号、地理位置等关键信息。ChargePrint 框架[12]通过迭代初始种子指纹以及分类、聚类的方法扩展设备搜索引擎的功能，在互联网上搜索电动汽车充电管理系统，并评估其安全性。

总的来说，现有研究方法多聚焦于如何基于已有知识和样本完成对设备类型的判别，在对未知设备识别方面，仅少量研究针对特定类型设备，设计了较为复杂的方案流程，且需要人工参与，难以泛化成通用的自动化设备类别识别方法。为此，本文从提高网络空间测绘节点设备类型识别率的角度，引入 LLM 及 RAG 技术，提出一种新的设备类型识别方法。

2 节点设备类型识别方法设计

在网络空间测绘过程中，测量节点无法有效识别设备类型，源于以下几种原因：

1) 版本迭代。由于设备型号、固件版本等升级变化，使当前特征与已有指纹库中的特征产生偏差。

2) 定制配置。厂商或用户自行修改、定制过的设备，其具体配置导致当前特征与已有指纹库中的特征产生差异。

3) 新出现设备。互联网中新出现的未知品牌或型号设备，缺少指纹或样本训练数据。

针对上述无法有效识别设备类型的情况，本文基于 LLM 和 RAG 技术，设计一种智能化识别方案。该方案的立足点是：在网络探测中发现的节点虽然无法被传统方法有效识别类型，但节点某些端口和协议特征中隐含了该设备的类型特征，特别是其端口响应数据的文本内容中蕴含了可被理解的相关信息。在此情况下，即使方案中并无该设备的预设指纹，且未使用特征相近的类似标签数据做过训练，也能够基于探测到的特征信息推理出设备类型。方案总体架构如图 1 所示，主要包括节点探测信息特征提取、背景知识检索和 LLM 推理 3 个部分。

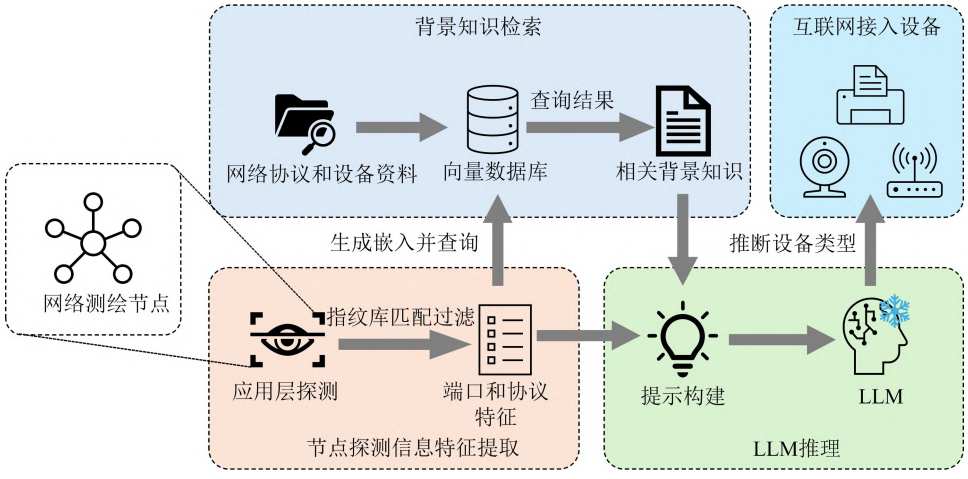


图 1 节点设备类型识别方案总体架构

Fig. 1 Scheme architecture of node device type identification

2.1　节点探测信息特征提取

节点探测信息特征提取部分的具体工作流程如图 2 所示。

针对网络测绘中指定的地址范围，以流行端口的 Top 20 为测量端口，首先使用应用层扫描器[13]快速发现其开放端口服务和横幅信息，包括非加密信息获取和 TLS 加密信息获取 2 部分；然后使用当前最新版本的 Nmap7.95 扫描器内置指纹库（Nmap-service-probes）对已知的设备类型进行匹配识别，由此筛选和过滤出不能被指纹库匹配的节点，结合探测得到的相关信息形成该节点的端口和协议特征。

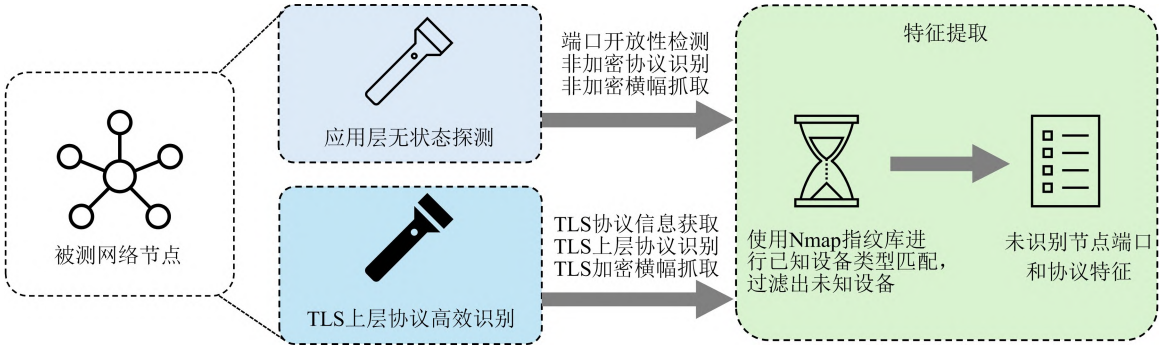


图 2　节点探测信息特征提取工作流程图

Fig. 2　Flowchart of node information feature extraction

本方案选择的端口和协议特征都能够基于大规模应用层探测技术高效获取，主要包括以下元素：

1）开放端口列表；

2）各开放端口所运行应用层协议；

3）各开放端口响应的横幅信息；

4）运行 TLS 协议的各开放端口中包含的证书信息，包括证书主题（Subject）和证书发行者（Issuer）的通用名、组织名、组织单元等。

2.2　向量数据库构建与背景知识检索

在节点设备类型识别中，传统方法依赖于预定义的特征匹配和规则库，当设备特征信息变化频繁时识别效果较差。LLM 虽然可以根据设备特征对其类型进行推理和识别，但由于模型的参数和训练数据有限，可能因缺乏相关的领域知识或较新的资料导致无法准确识别。表 1 列出了实际探测中根据关键特征和背景知识人工使用 LLM 判断设备类型的相关示例。

表 1　节点设备类型识别相关示例

Tab. 1　Examples of node device type identification

示例	关键特征的类型	特征内容重点部分	背景知识	判别设备类型
1	HTTP 协议响应横幅	WWW-Authenticate：Basic realm＝”TP-LINK Wireless N 3G/4G Router MR3420”	无需	无线路由器
2	FTP 协议响应横幅	220 ET0021B78D865C Dell B2360dn Laser Printer FTP Server NH6.CY.N632 ready	无需	打印机
3	HTTP 协议响应横幅	Server：cisco-IOS	思科公司是全球领先的网络解决方案供应商，主要产品包括路由器、交换机等网络设备	路由器
4	TLS 证书信息	证书发行者组织名：WatchGuard	WatchGuard 公司是全球排名前列的专业防火墙产品制造商之一，主要业务涉及众多安全产品，包括防火墙与网关设备等	防火墙设备
5	端口协议	该开放端口运行 IPCAM 协议	IPCAM 是一种用于摄像头在网络中传输数字视频流的通信协议	网络摄像头
6	HTTPS 协议响应横幅	SESSIONID＝…Secoway USG2110	Secoway USG 是华为旗下的防火墙终端产品系列	防火墙设备

由表 1 的示例，可以形成如下认知：

1）节点某开放端口运行应用层协议，并且其响应横幅中包含了设备的类型和版本信息，则 LLM 在无需特定领域背景知识的情况下判断设备类型，如示例 1 和示例 2；

2）节点所有开放端口和协议特征中并不直接包含设备类型信息，但其响应横幅或 TLS 证书信息中包含服务器厂商名称，基于相关背景知识可以推断该设备的类型，如示例 3 和示例 4；

3）节点某开放端口只响应了没有语义的二进制消息，但该开放端口运行协议被识别为一种设备专用协议，基于网络协议相关的背景知识可以推断该设备类型，如示例 5；

4）节点的开放端口响应横幅中虽没有指示设备类型，但其包含了产品型号，又由于该具体型号在传统指纹库中没有收录，因此未能识别。然而基于型号所属系列产品信息的背景指示，可以推断该设备类型，如示例 6。

综上所述，当节点特征中包含设备类型或型号时，LLM 能够根据提供的上下文信息准确识别设备类型和型号；当节点特征中并不显示包含设备类型标识符时，使用 RAG 技术为大模型补充相关背景知识，同样可以实现对设备类型的有效识别。基于这种策略，向量数据库构建与背景知识检索模块的实现流程如图 3 所示。

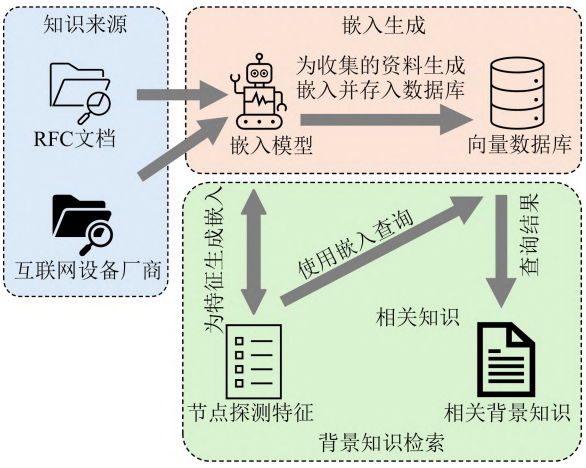


图 3 向量数据库构建与背景知识检索

Fig. 3 Vector database construction and background knowledge retrieval

构建向量数据库的知识资料主要来源于 RFC 文档和各种互联网厂商设备的开源资料。RFC 文档中详细定义了各种网络协议的规范和行为，为理解网络设备运行机制提供重要基础知识；各种互联网厂商设备的介绍资料包含了网络设备的功能、工作机制、使用协议和生产厂商等信息，这些信息为理解节点探测得到的特征信息进而识别设备类型提供重要支撑。

为了支持针对节点特征信息的语义化查询功能，这些知识资料需要被转化为向量表示。具体来说，需要采用嵌入技术，将文本内容通过嵌入模型转化为向量表示集合 $\boldsymbol{V}=\{\boldsymbol{v}_1,\boldsymbol{v}_2,\cdots,\boldsymbol{v}_n\}$，向量维度为 d。这些向量可以捕捉文档中的语义信息，使得与设备特征中的语义相似的资料能够在向量空间中距离更近。这些嵌入向量随后被存入向量数据库。

在未识别节点特征提取完成后，输入其端口和协议特征信息，向量数据库能够根据这些输入信息生成对应的相同维度的嵌入向量 $\boldsymbol{w}$，通过计算 $\boldsymbol{w}$ 与 $\boldsymbol{V}$ 中向量的距离

$$D_{\mathrm{e}}=\sqrt{\sum_{j=1}^{d}(\boldsymbol{w}_j-\boldsymbol{v}_{1j})^2} \tag{1}$$

快速在向量数据库中找到最相似的向量，从而返回至与该节点最相关的背景知识。例如，通过查询返回有关协议或者与横幅内容相关的设备、厂商等资料。

需要说明的是，虽然向量数据库的构建也需要收集相关领域的资料，但相比于传统的指纹库构建和模型的标签数据集整理，能够省去大量的专家知识和相关成本投入，无需复杂的数据清洗过程，自动化程度较高，并且易于资料的更新处理。

2.3 基于提示工程的大模型推理

在节点特征提取和背景知识检索的基础上，需要进一步使用 LLM 进行设备类型推断，其中的关键在于构建合理的提示词（亦称提示工程，prompt engineering），通过有效的问题引导 LLM 对设备的类型进行判断和分类。提示词不仅要向 LLM 传递设备的具体协议特征和相关背景知识，还需确保逻辑性和引导性，以帮助模型在推理过程中沿着正确的方向进行分析。

为确保 LLM 能够准确区分不同类型的设备，本方案设计的提示词中提供了每种设备类别的详细描述，涵盖其典型协议特征、常用端口、典型行为以及具体的应用场景。部分举例如下：

1）路由器。连接 2 个或多个网络的硬件设备，利用路由协议（如 BGP、OSP 等）管理网络流

量，通过 HTTP、HTTPS、SSH 或 TELNET 协议进行配置，广泛应用于家庭、企业和服务提供商环境。

2）网络摄像头。一种结合传统摄像机与网络技术的新一代摄像机，常使用 RTSP、RTP、RTCP 等协议进行视频流传输，可通过 HTTP、HTTPS、SSH 或 TELNET 等协议进行管理，一般工作在特定的端口范围。

3）打印机。作为网络节点独立存在，通过网络打印服务接入互联网，常使用 IPP、LPD/LPR、HTTP 或专有协议进行通信，设备响应较为标准化，典型应用领域为办公场景。

提示词的另一个重要作用是规范化模型输出格式。通过提示词引导，LLM 可以按照预设的格式返回结果，这对于后续方案的自动化实现非常重要。本方案使用结构化的 JSON 格式作为输出格式，规范化输出的提示词设计包括：

1）指示模型输出 JSON 格式。提示词中明确要求模型按指定格式返回结果。例如，在推理设备类型时，模型输出的 JSON 应包含设备类型、推理依据和置信度等字段。

2）统一字段定义。在提示词中定义字段名称（如设备类型、置信度等），确保模型输出的一致性和规范性。

基于以上思路，本方案设计的提示词形式如图 4 所示。由图 4 可以看出，提示词设计主要包含 5 个部分，其中：

① 设备类型范围和详细定义用来明确类型划分的标准和边界；

② 当前设备节点端口和协议特征是进行类型推断的关键依据；

③ 推断设备类型可能用到的思路可以为模型提供推理依据；

④ 相关背景知识为模型补充可用的推理材料；

⑤ 输出格式和示例模板用以规范化模型的输出，便于程序提取。

使用精心构建的提示词，引导 LLM 结合设备特征与背景知识进行推理，并使用结构化格式输出结果，能够提高设备类型识别的准确性和规范性。

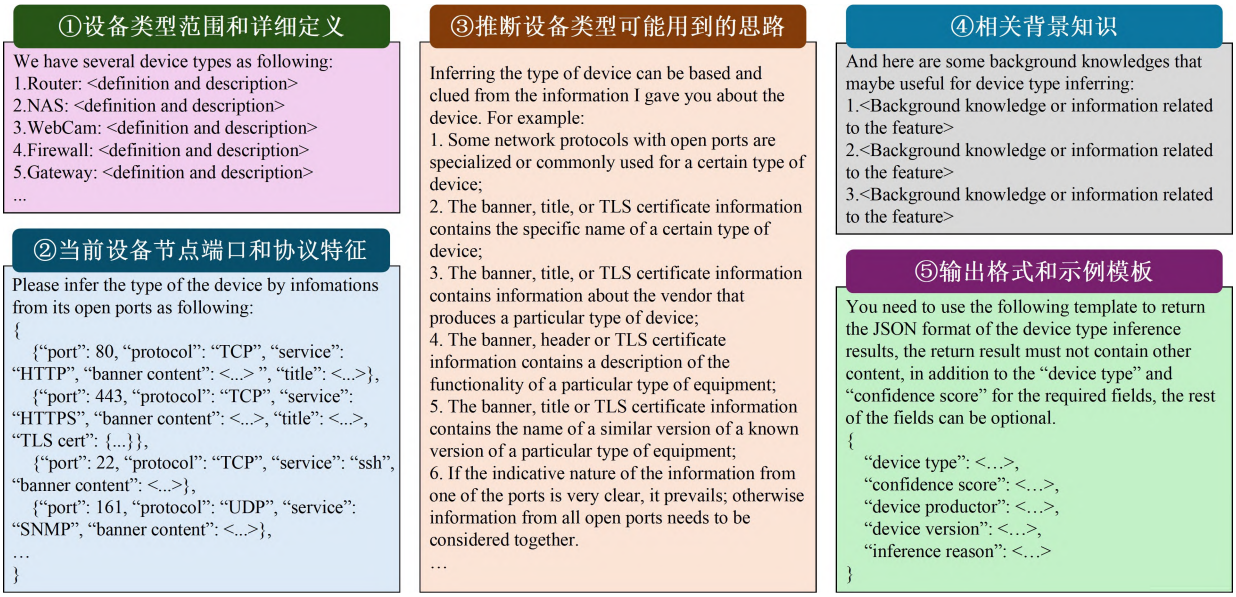


图 4　LLM 推理阶段提示词设计

Fig. 4　Design of prompt words in the inference stage of LLM

3　实验与分析

基于上述设计方案，本文实现了面向网络空间测绘的节点探测与类型识别原型系统 LingoVision。本节主要设计了 3 个方面的实验，分别是 LingoVision 系统基线方案实验，对部分模块进行删减的消融实验，以及使用该系统对真实互联网中无法被现有网络空间搜索引擎和流行指纹库识别的节点进行设备类型识别的实验。

3.1　实验设置

考虑到方案中节点特征提取部分涉及大规模网络扫描，结合此类扫描节点轻量级部署的实际情况，本文将实验环境限制在常规配置下：实验均在配备 12 核 CPU、16 GB 内存的笔记本电

脑上进行，主机上运行 Ubuntu22.04LTS 操作系统及 6.5.0-41-generic 版本 Linux 内核，无其他特殊设置。

LingoVision 系统选取 2 个轻量级且易于部署在探测节点的模型，具体如下：

1）嵌入模型。使用开源嵌入模型 mxbai-embed-large-v1[14-15]，通过对 7 亿对比学习数据对进行训练，并结合 3 000 万高质量三元组数据进行微调，使其灵活处理不同主题和领域，旨在将文本转换为稠密的向量表示，用于各种自然语言处理任务，如语义搜索、信息检索和文本聚类。

2）语言模型。使用由 Meta 开发的 LLaMA3 大语言模型的 8B 版本。LLaMA3 是一种自回归语言模型，采用优化的 Transformer 架构，专为生成式任务和对话场景进行了调优。8B 版本的模型拥有 80 亿参数，并通过大规模公开数据集进行预训练。此版本模型支持输入文本并生成文本或代码，具备强大的语言理解和生成能力。

使用开源的 Faiss（Facebook AI similarity search）向量数据库，每次查询时返回前 3 条最相关数据。在方案的具体实现中，通过下载和爬虫爬取的数据经过语义切分后，得到 4 562 条 RFC 相关数据和 1 241 条设备厂商相关数据。切分时，文本块词数（chunk size）设置为 1 000，文本重叠词数（chunk overlap）设置为 200。

为了验证本方案效果，测试数据集需要同时包含较为全面的端口、协议特征以及设备标签，而已有的节点设备类型识别工作中并没有满足条件的可用数据集。因此，本数据集主要通过手动收集现有网络空间搜索引擎中的相关记录获得。数据集中包含 333 个设备条目以及每种设备类型的多个特征信息，共涉及 14 种不同的设备类，包括端口、协议、TLS 证书、服务横幅（Banner）和网页标题等特征。每一条数据记录代表一个节点及其相关开放端口的协议特征。由于这些节点条目中包含了无法通过 Nmap 指纹库匹配其类型的记录，这为测试方案的鲁棒性和泛化能力带来了挑战。

数据集中涉及的设备类型标签来源于现有网络空间搜索引擎，具体类型及其简称包括：路由器（router）、网络存储设备（NAS）、网络摄像头（webcam）、防火墙（firewall）、VoIP 网关/适配器（VoIP adapter）、网关（gateway）、打印机（printer）、无线接入点（WAP）、VPN 设备（VPN）、负载均衡器（load balancer）、代理服务器（proxy server）、工控设备（ICS）、媒体设备（media device）、邮件服务器（mail server）。

3.2 评价指标

1）准确率（accuracy）。识别正确的样本数占总样本数的比例，即：

$$A_{\mathrm{Acc}}=\frac{T_{\mathrm{TP}}+T_{\mathrm{TN}}}{T_{\mathrm{TP}}+T_{\mathrm{TN}}+F_{\mathrm{FP}}+F_{\mathrm{FN}}} \tag{2}$$

式中，T_{TP} 表示正确识别为当前类别的样本数，T_{TN} 表示正确识别为非当前类别的样本数，F_{FP} 表示错误识别为当前类别的样本数，F_{FN} 表示错误识别为非当前类别的样本数。

2）精确率（precision）。识别为正类中真正是正类的比例，即：

$$P_{\mathrm{Pre}}=\frac{T_{\mathrm{TP}}}{T_{\mathrm{TP}}+F_{\mathrm{FP}}} \tag{3}$$

3）召回率（recall）。在所有实际为正类的样本中，正确识别为正类的样本的比例，反映了方案捕获正类样本的能力，可表示为：

$$R_{\mathrm{Rec}}=\frac{T_{\mathrm{TP}}}{T_{\mathrm{TP}}+F_{\mathrm{FN}}} \tag{4}$$

4）F1 值（$F_{\text{1-Score}}$）。旨在综合精确率和召回率的表现，提供一个平衡指标，对于二分类情况来说，有：

$$F_{\text{1-Score}}=2\cdot\frac{P_{\mathrm{Pre}}\cdot R_{\mathrm{Rec}}}{P_{\mathrm{Pre}}+R_{\mathrm{Rec}}} \tag{5}$$

由于本实验为多分类情况，因此使用了以下 2 种均值计算方式：

5）宏平均值（macro-average）。为每个类别单独计算精确率、召回率及 F1 值，然后计算这些值的平均值。宏平均值不考虑每个类别的样本数量。具体表示为：

$$M_{\mathrm{ma}}=\frac{1}{n}\sum_{i=1}^{n}X_i \tag{6}$$

式中，n 表示类别的数量，X_i 表示第 i 个类别的精确率、召回率及 F1 值。

6）加权平均值（weighted-average ）。在宏平均值的基础上，每个类别的精确率、召回率及 F1 值会根据该类别的样本数量进行加权。具体表示为：

$$W_{\mathrm{wa}}=\frac{\sum_{i=1}^{n}(T_{\mathrm{TP},i}+F_{\mathrm{FN},i})\cdot Y_i}{\sum_{i=1}^{n}(T_{\mathrm{TP},i}+F_{\mathrm{FN},i})} \tag{7}$$

式中，$T_{TP,i}$ 和 $F_{FN,i}$ 分别表示第 i 个类别的真正例和假负例的数量，Y_i 表示第 i 个类别的精确率、召回率或 F1 值。

3.3　基线方案实验

使用 LingoVision 系统针对数据集所有条目进行节点设备类别识别测试，该系统集合了本方案中的功能模块，是后续消融实验的基线方案。根据实验结果绘制的混淆矩阵归一化热力图（如图 5 所示）展示了对每类设备分类的准确率以及误分类的情况；相关分类分数柱状图（如图 6 所示）展示了每类设备分类的精确率、召回率和 F1 值。

由实验的结果数据可以看出，不同设备类型的识别精度存在差异，就混淆矩阵热力图所展现的单个设备分类准确率而言，针对路由器、网络存储设备、VoIP 网关和媒体设备的分类表现最为优异，准确率超过了 90%。这表明这些设备本身的协议特征在分类过程中较为明显，LingoVision 系统能够较好地捕捉特征中蕴含的设备类型信息；针对网络摄像头、VPN 设备、代理服务器和邮件服务器的分类表现也较好，识别准确率均高于 80%。这些设备的特征信息较为独特或集中，能够有效帮助模型进行推断。然而，针对防火墙、打印机、负载均衡器和工控设备的分类表现较为一般，准确率在 60%～80%，这是因为这些设备在某些特征上与其他设备存在一定重叠，导致分类模糊。从人工识别的角度来看，也难以根据设备的端口和协议特征对其进行精细划分。同时，相互之间强相关的设备类型虽然导致误分类，但这样的分类也已经完成了粒度较粗的类型识别。

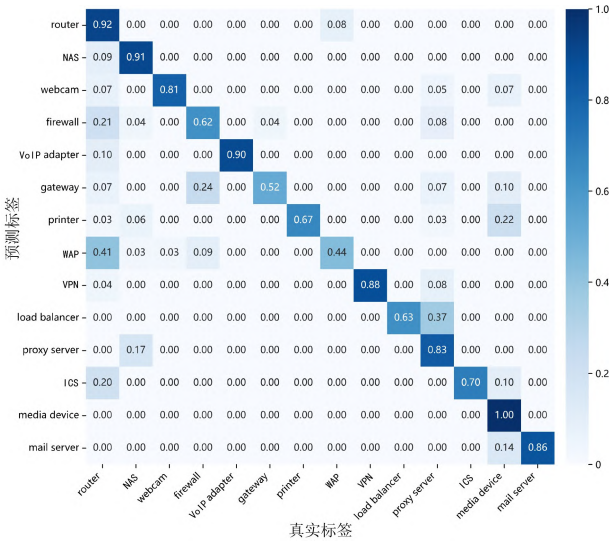


图 5　LingoVision 实验结果的混淆矩阵归一化热力图

Fig. 5　Normalized heatmap of confusion matrix for LingoVision experimental results

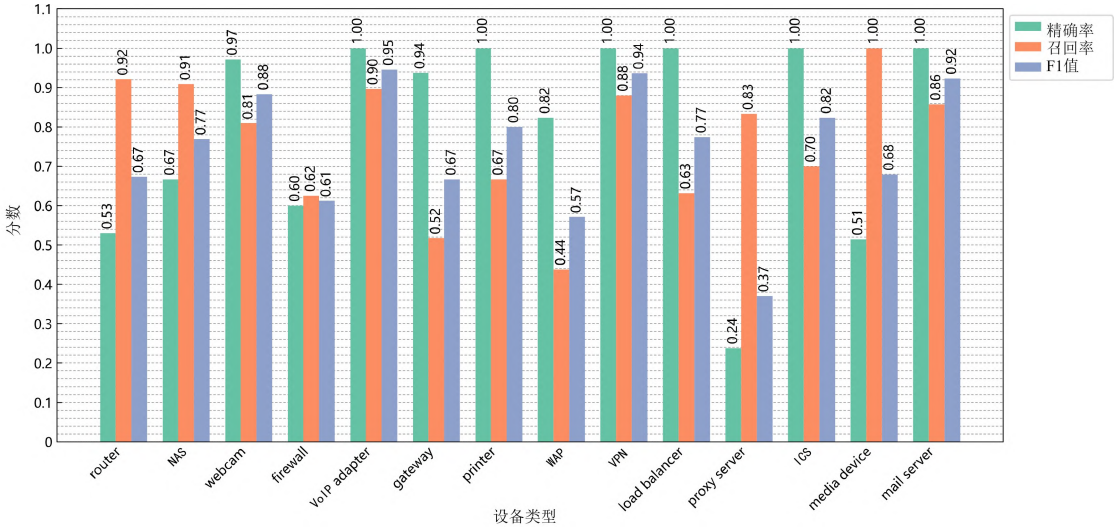


图 6　LingoVision 实验结果的分类分数柱状图

Fig. 6　Classification score bar chart of LingoVision experimental results

针对图 6，相关分数的均值结果见表 2 所列。

在均值分数报告中可以看到 LingoVision 系统的整体表现，宏平均值下的精确率为 81%，召回率为 76%，F1 值为 0.74，这表明模型在各类别上总体表现较好，但个别设备类型的表现拉低了整体的召回率和 F1 值。加权平均值下的精确率为 84%，召回率为 75%，F1 值为 0.76，加权平均值比宏平均值稍高，说明模型在处理设备类别时，主要类别的分类效果较好，而一些少数类别表现较差。

表 2 LingoVision 实验结果均值统计

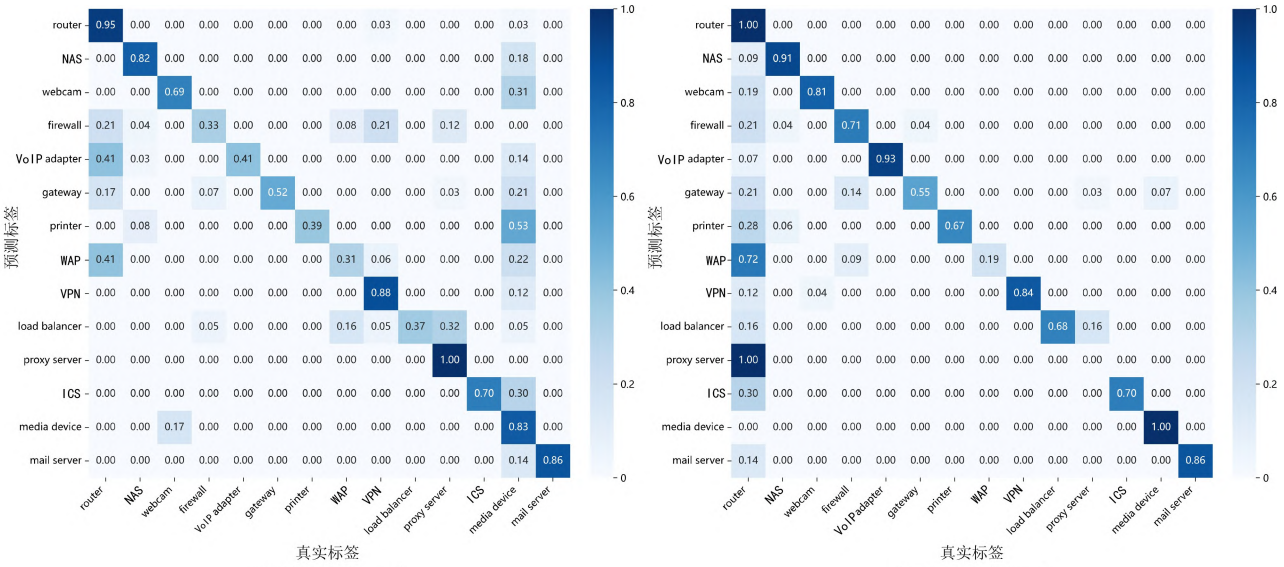
Tab. 2 Mean statistics of LingoVision experimental results

	精确率/%	召回率/%	F1 值
宏平均值	81	76	0.74
加权平均值	84	75	0.76

总体来看，LingoVision 系统在测试数据集上的表现良好，大部分设备类型能够准确识别。对于少部分表现不佳的设备类型，其分类表现除了受限于设备类型的特征不够明显，也因为所使用的数据集分类标签之间不完全正交，具有强关联关系的设备间特征存在较大重叠。尽管如此，LingoVision 系统还是能够将这些设备分类到最相关的类型当中，实现粗粒度的正确分类。

3.4 消融实验

为了进一步评估方案中各模块策略对 LingoVision 系统的贡献，本文采用消融实验进行分析。通过去除 RAG 技术和精心设计的提示词这 2 部分，深入分析方案各部分对分类效果的影响。其中，未采用 RAG 技术的简化系统命名为 LingoVision-1；不对提示词进行精心设计，只包含图 4 中的②、④和⑤部分，简化系统命名为 LingoVision-2。随后，分别对所有条目进行节点设备类别识别测试，绘制的混淆矩阵归一化热力图如图 7 所示，相关分类分数柱状图如图 8 所示。



(a) LingoVision-1 (b) LingoVision-2

图 7 消融实验混淆矩阵归一化热力图

Fig. 7 Normalized heatmap of confusion matrix in ablation experiment

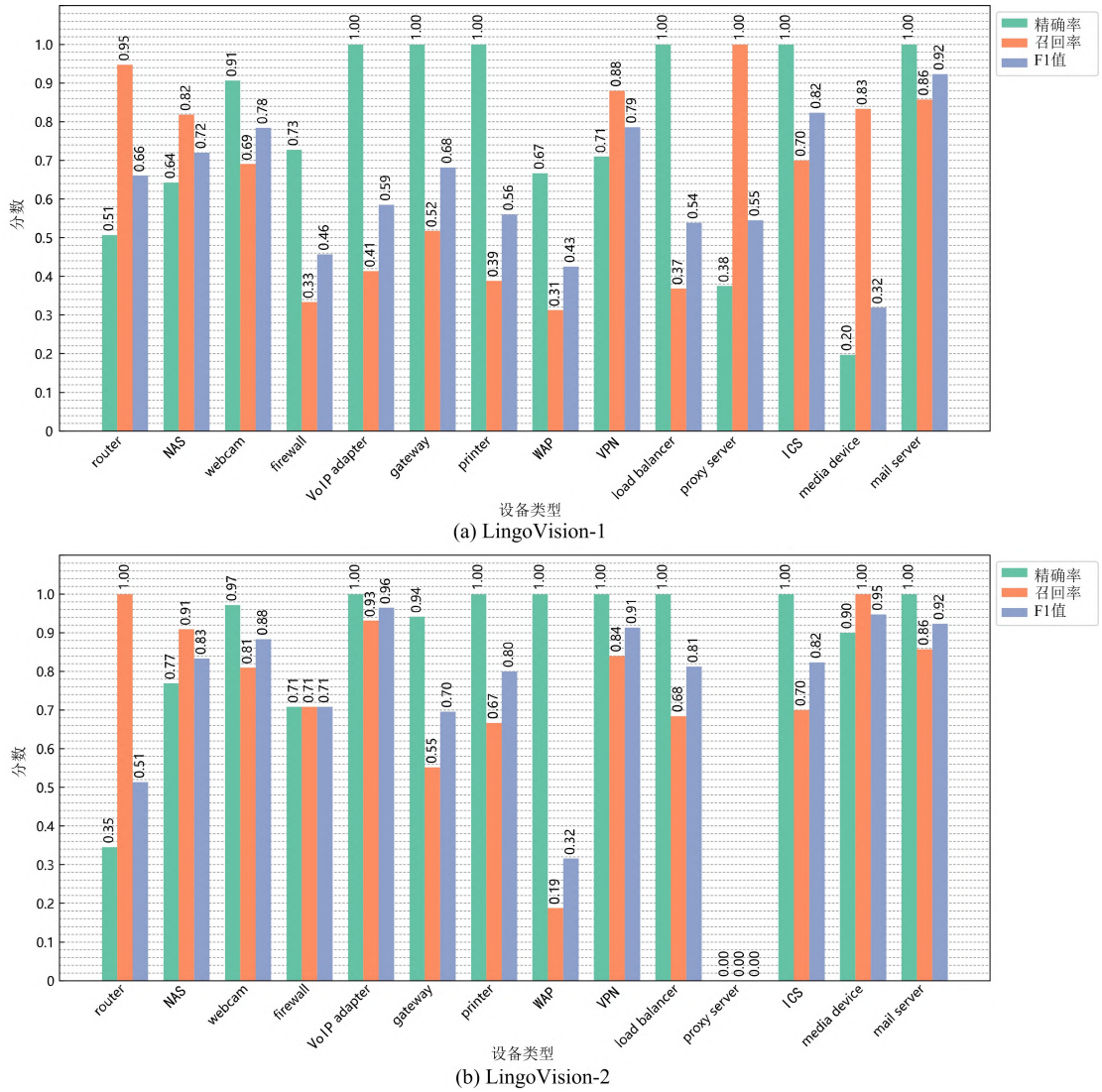
从实验结果来看，去除 RAG 技术之后，系统受影响较大。在混淆矩阵热力图中，防火墙、打印机和 VoIP 网关的分类表现下降最为明显，准确率均下降至原来的 1/2 左右，表明这几类设备在没有检索背景知识时，系统难以充分利用现有的设备特征进行准确分类，RAG 在其中起到了关键作用。

在不使用精心设计的提示词的条件下，少部分设备类型分类准确率有轻微下降，如 VPN 设备和无线接入点等。大部分设备类型的分类准确率和各分类分数相比基线实验来说较为稳定。但同时也出现因缺少提示词设计，出现代理服务器这一类型的分类准确率下降至 0 的极端情况，并且全部被误分类到路由器类。这表明精心设计的提示词确实起到了标签间边界划分的作用，能够对个别设备的识别起到关键作用。

3.5 实网识别节点实验

为验证本方案对于真实的未识别节点的设备类型识别效果，本文搜集了部分网络空间搜索引擎无法识别其设备类型的真实节点，这些节点经 Nmap 在线检测仍无法确定具体的设备类型。因此，使用 LingoVision 系统针对这些未知设备进行识别。

此外，实网探测的节点设备类型分布与数据集不同，主要体现在基准真相的标签数量更多更丰富。而当 LLM 由于领域知识不足或缺乏事实验证能力时，实网探测更容易导致大模型幻觉的产生，即在生成内容时，产生与事实不符、虚构或



(a) LingoVision-1

(b) LingoVision-2

图 8　消融实验分类分数柱状图

Fig. 8 Bar chart of ablation experiment classification scores

具有误导性信息的现象。尽管这些回答可能流畅且有逻辑性，但内容上存在错误或与输入指令不一致。为了避免模型面对超出提示词预设范围的节点设备类型时出现幻觉的情况，除了以 RAG 手段作为缓解外，本文在 LingoVision 的实网探测版本提示词中，增加了 unknown 和 others 设备类型选项，允许大模型在推理节点设备类型遭遇歧义或不明确时能够作出更客观的回答，进一步避免幻觉的产生。对未知节点的识别结果见表 3 所列。

对于这些识别出来的设备，由于当前没有对应的基准真相作为标签，本文通过人工的方式逐一对其端口和协议特征进行核验，均能从响应横幅、TLS 证书或网页标题等文本特征中确认 LingoVision 识别结果的合理性，具体核验结果的示例及使用到的关键特征见表 4 所列。

表 3　针对未知节点的识别统计结果

Tab. 3　Identification and statistical results for unknown nodes

识别出的设备类型	独立设备种类数量	识别出的独立设备数量		同特征网络节点数量/个
		所属厂商数	具体型号数	
路由器	9	8	9	16 406
网络摄像头	5	4	4	7 782
VoIP 网关	1	1	1	47 883
打印机	3	2	2	138
负载均衡器	1	1	0	460
网络存储设备	4	4	3	9 480
无线接入点	3	3	3	14 835
VPN 设备	2	2	2	1 603
防火墙	3	2	2	4 690
总计	31	27	26	103 277

从识别结果中可以看到，LingoVision 系统基

于真实存在的互联网节点特征信息，能够有效识别现有流行指纹库和网络空间搜索引擎无法识别的设备。识别出来的设备类型包括 9 类，各类型涉及的具体设备型号(版本)共计 31 种，并能识别出其中大部分设备类型的所属厂商和具体型号。经过网络空间搜索引擎中的数据比对，在互联网中与这些具体型号(版本)设备具有相同特征的 IP 节点数量超过 10 万个。

表 4 针对未知节点的识别及人工核验示例

Tab. 4 Examples of identification and manual verification for unknown nodes

示例	关键特征的类型	特征内容重点部分	系统推断及人工核验结果		
			设备类型	所属厂商	具体型号
1	SNMP 协议横幅响应	… Cisco IOS Software, C800 Software (C800-UNIVERSALK9-M), Version 15.5(3)M5, RELEASE SOFTWARE (fc1) Technical Support …	路由器	Cisco	C800 15.5(3)M5
2	RTSP 协议横幅响应	RTSP/1.0 200 OK CSeq: 1 Server: IP Network Camera RTSP Server Public: DESCRIBE, SETUP, PLAY, TEARDOWN, GET PARAMETER, SET PARAMETER	网络摄像头	未知	未知
3	未知协议横幅响应	$DEVINFO, 000001, 13.5, WLAN, ADHOC, M1-WiFi, 6, 78, MO-DEM, movistar, 4, HSDPA, GPS, ON, 30000, 1, PORT, 1, SERIAL, 115200, ON, PORT, 2, SERIAL, 115200, ON, SERVICE, USB, 0, 30000, 1, 38450141	无线接入点	Movistar	M1-WIFI
4	POP3 协议横幅响应	+OK-ERR FortiGate firewall user authentication is needed.	防火墙	FortiGate	未知
5	HTTP 协议横幅响应	HTTP/1.1 200 OK Server: ZhiDa NAS… connect-src data: ws: wss: ; default-src'self' 'unsafe-eval' data: blob: https://*.synology.com https://www.synology.cn/ https://help.synology.cn/	网络存储设备	Synology	ZhiDa NAS

4 结束语

针对网络空间测绘中存在大量节点设备类型无法识别的问题，本文提出了一种基于 LLM 和 RAG 技术的节点设备识别方案，充分考虑了各种互联网设备在应用层协议中暴露的端口和特征信息，构建了一个能够部署在探测节点的未知设备识别原型系统 LingoVision。通过系列实验验证了该系统方案的有效性。本文所提出的未知节点设备类型识别方法为网络空间测绘下的新设备发现提供了新的思路，且该方法具备较高的扩展性和准确性，可为后续的网络资产管理和安全态势感知提供技术支撑。

参 考 文 献

[1] XU Q, ZHENG R, SAAD W, et al. Device fingerprinting in wireless networks: challenges and opportunities[J]. IEEE Communications Surveys & Tutorials, 2016, 18(1): 94-104.

[2] WAN S F, LI Q, WANG H, et al. DevTag: a benchmark for fingerprinting IoT devices[J]. IEEE Internet of Things Journal, 2023, 10(7): 6388-6399.

[3] Recog: a recognition framework[EB/OL]. [2025-07-24]. https://github.com/rapid7/recog/#recog-a%20-recognition-framework.

[4] LI Q, FENG X, WANG H, et al. Discovery of Internet of Thing devices based on rules[C]// Proceedings of 2018 IEEE Conference on Computer Communications Poster and Demo. [S.l.]: IEEE, 2018: 1-2.

[5] FENG X, LI Q, WANG H, et al. Acquisitional rule-based engine for discovering Internet-of-Things devices [C]//Proceedings of the 27th USENIX Security Symposium. [S.l.: s.n.], 2018: 327-341.

[6] JAVED T, HASEEB M, ABDULLAH M, et al. Using application layer banner data to automatically identify IoT devices[J]. ACM SIGCOMM Computer Communication Review, 2020, 50(3): 23-29.

[7] CHENG H, DONG W Y, ZHENG Y, et al. Identify IoT devices through web interface characteristics[C]//

Proceedings of 2021 IEEE International Conference on Computer and Communication Systems. [S. l.]: IEEE, 2021: 405-410.

[8] WU Y C, LI C L, YANG J H, et al. WebIoT: classifying Internet of Things devices at Internet scale through web characteristics[C]//Proceedings of 2022 IEEE Symposium on Computers and Communications. [S. l.]: IEEE, 2022: 1-7.

[9] SARABI A, YIN T X, LIU M Y. An LLM-based framework for fingerprinting Internet-connected devices [C]//Proceedings of 2023 ACM on Internet Measurement Conference. New York: ACM, 2023: 478-484.

[10] UEDA T, SASAKI T, YOSHIOKA K, et al. An Internet-wide view of connected cars: discovery of exposed automotive devices[C]//Proceedings of the 17th International Conference on Availability, Reliability and Security. New York: ACM, 2022: 1-8.

[11] SASAKI T, FUJITA A, GAÑÁN C H, et al. Exposed infrastructures: discovery, attacks and remediation of insecure ICS remote management devices [C]// Proceedings of 2022 IEEE Symposium on Security and Privacy. [S. l.]: IEEE, 2022: 2379-2396.

[12] NASR T, TORABI S, BOU-HARBE, et al. ChargePrint: a framework for Internet-scale discovery and security analysis of EV charging management systems[C]// Proceedings of the 30th Annual Network and Distributed System Security Symposium. [S. l. : s. n.], 2023: 1-18.

[13] CHEN C Y, LU Y L, YANG G Z, et al. ZBanner: fast stateless scanning capable of obtaining responses over TCP[C]//Proceedings of 2024 IEEE International Performance, Computing, and Communications Conference. [S. l.]: IEEE, 2024: 1-6.

[14] LEE S, SHAKIR A, KOENIG D, et al. Open source strikes bread-new fluffy embeddings model[EB/OL]. (2024-03-08)[2025-07-24]. https://www.mixedbread.ai/blog/mxbai-embed-large-v1.

[15] LI X M, LI J. AoE: angle-optimized embeddings for semantic textual similarity[C]//Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics. Bangkok: Association for Computational Linguistics, 2024: 1825-1839.

作 者 简 介

杨国正

男，1982 年生，博士，教授，研究方向为网络空间测绘、网络安全态势感知

E-mail：yangguoz0218@163.com

陈驰昱

男，1996 年生，硕士研究生，研究方向为网络安全态势感知

E-mail：chenchiyu14@nudt.edu.cn

沈照斌

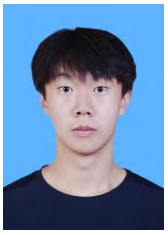
男，2001 年生，硕士研究生，研究方向为网络空间测绘

E-mail：zhaobin19@nudt.edu.cn

齐冬震

男，2000 年生，硕士研究生，研究方向为路由安全

E-mail：qidongzhen@nudt.edu.cn



潘俊宇

男，2003 年生，硕士研究生，研究方向为网络空间测绘

E-mail：1599561241@qq.com

责任编辑　董　莉