文章编号:1671-4598(2026)04-0265-07 DOI:10.16526/j.cnki.11-4762/tp.2026.04.033 中图分类号:TP393.07 文献标识码:A

基于微调与检索增强生成的混合式智能协议理解方法

康智峰，张亚生
（中国电子科技集团公司 第五十四研究所，石家庄 050000）

摘要：针对5G与天地一体化网络等新兴环境中快速演进的网络协议带来的传统解析方法在实时性与可解释性方面难以满足现代网络运维与安全分析需求的问题，对一种面向智能协议理解的检索增强生成系统进行了研究；该系统采用了协议感知的知识预处理、多模态混合索引构建与协议感知生成器微调等关键技术，经在涵盖12种主流协议、近2 000个问题的测试集上进行的实验评估，该系统实现了77.97%的软召回率，显著优于Modular RAG与GraphRAG等主流方法；此外，十六进制推理类问题的端到端准确率由58.24%提升至79.63%，同时其他类型问题性能保持稳定；经实际应用验证，该技术满足了智能协议分析与自解释网络等工程场景对高精度、可追溯且无需外部解析器的协议理解需求，为相关领域提供了有效的技术路径。

关键词：协议解析；检索增强生成；网络协议；问答系统；大语言模型微调

A Hybrid Intelligent Protocol Understanding Method Based on Fine-tuning and Retrieval Augmented Generation

KANG Zhifeng, ZHANG Yasheng
(The 54th Research Institute of China Electronics Technology Group Corporation, Shijiazhuang 050000, China)

Abstract: To address the challenge that traditional parsing methods struggle to meet the real-time and interpretability requirements of modern network operation and security analysis due to the rapid evolution of network protocols in emerging environments such as 5G and integrated space-ground networks, this paper presents a retrieval augmented generation system for intelligent protocol understanding, and employs key technologies such as protocol-aware knowledge preprocessing, multimodal hybrid index construction, and protocol-aware generator fine-tuning. On a test set covering 12 mainstream protocols and nearly 2 000 questions, experimental evaluation demonstrates this system achieves a soft recall rate of 77.97%, significantly outperforming mainstream methods such as Modular RAG and GraphRAG. Furthermore, the end-to-end accuracy for the hexadecimal reasoning questions improves from 58.24% to 79.63% while maintaining stable performance for other types of questions. Practical application verifies that this technology meets the requirements of high-precision, traceable, and external-requirement-free protocol understanding in engineering scenarios such as intelligent protocol analysis and self-explaining networks, providing an effective technical path for related fields.

Keywords: protocol parsing; retrieval-augmented generation; network protocols; question answering system; fine-tuning

0 引言

随着天地一体化网络在涵盖低轨卫星互联网、第五代移动通信技术（5G，5th generation mobile communication technology）非地面网络及未来第六代移动通信技术（6G，6th generation mobile communication technology）空天融合架构的民用领域的快速部署，网络协议体系正加速向标准化网际互连协议（IP，internet protocol）架构演进[1-2]。国际标准组织正在积极推动协议融合，其中第三代合作伙伴计划（3GPP，3rd gener-

收稿日期:2026-01-30； **修回日期**:2026-03-11。
作者简介:康智峰(1999-),男,硕士研究生。
通讯作者:张亚生(1969-),男,硕士,研究员。
引用格式:康智峰,张亚生.基于微调与检索增强生成的混合式智能协议理解方法[J].计算机测量与控制,2026,34(4):265-271.

ation partnership project）Release 17 明确支持基于 IP 的非地面网络接入[3]，互联网工程任务组（IETF，the internet engineering task force）持续优化高延迟/高误码环境下的 TCP/IP 机制[4]，标志着民用天地一体化网络的协议基础正逐步收敛于公有化标准[5]。然而，协议知识高度分散于请求评论（RFC，request for comments）规范、3GPP 技术文档、厂商配置手册及真实流量样本中[6-7]，且随版本迭代频繁更新，导致传统基于规则或静态解析的方法难以高效处理多源异构、语义隐含的协议知识[8]。针对这一挑战，本文提出协议感知的检索增强生成系统，通过重构检索机制并微调生成器，实现“自然语言－十六进制数据－协议规范”三元数据的精准对齐，有效解决通用大语言模型在报文解析中的瓶颈问题。实验表明，该方法在十六进制推理任务中软召回率提升超 20%，同时保持其他类型问题性能稳定，为天地一体化网络的智能运维与互操作性验证提供了兼具实用性与可扩展的技术路径。

1 面向协议理解的 RAG 系统研究现状

1.1 协议分析

传统协议分析高度依赖人工编写的解析器或基于有限状态机的建模工具[9]。这类方法在处理结构固定、文档完备的标准协议时效果良好，但在面对天地一体化网络中常见的协议组合复杂性或变长/可选字段爆炸时，往往因缺乏语义上下文而失效。近年来，部分研究尝试通过流量聚类、符号执行或深度学习进行协议逆向工程[10]。

而在当前民用天地一体化场景中，绝大多数协议已公开标准化，真正的挑战并非协议未知，而是如何从海量异构知识源中提取所需信息。因此，研究重心正从“格式推断”转向“语义融合与智能推理”[11]。本文工作与此类逆向工程形成互补：我们假设协议规范可获取，重点解决标准文本与原始报文之间的语义对齐问题。

值得注意的是，尽管商业系统提供标 IP 接口，其上层协议行为仍受链路特性（高延迟、高误码）影响，需结合规范与真实流量联合解读[12]。现有智能分析工具多依赖规则引擎，难以适应协议演进。相比之下，大语言模型凭借强大的上下文理解与少样本泛化能力，为构建端到端、可解释的协议问答系统提供了新范式[13]。

1.2 领域问答系统与 RAG 在协议理解中的应用

通用大语言模型（LLM，large language model）虽具备广泛知识，但在网络工程等高精度领域易产生幻觉或细节偏差。为此，检索增强生成（RAG，retrieval-augmented generation）被广泛用于构建领域问答系统：通过从权威知识库中检索相关片段，约束 LLM 生成符合规范的回答，显著提升事实性与可追溯性[14]。

然而，现有 RAG 系统普遍存在一个被忽视的瓶颈：生成器对原始十六进制数据的“语义失明”。即便检索模块成功返回正确的 RFC 片段，通用 LLM 在面对包含原始十六进制流的示例时，仍难以识别字段边界、对齐术语与字节位置，甚至忽略十六进制内容而仅依赖检索文本生成答案[15]。这表明，RAG 的检索能力虽强，但若生成器无法识别报文结构，整个系统的可靠性将受限于最后一环。

1.3 协议理解的技术路径

针对上述问题，现有研究大致分为两类技术路径：

1）纯检索增强路径：试图通过更复杂的检索机制（如 Modular RAG[16]、GraphRAG[17]）或外部解析器预处理十六进制报文，将其转换为结构化文本后再输入 LLM。此类方法虽能提升输入可读性，但破坏了端到端流程，且难以处理新兴格式协议。

2）纯微调路径：直接在协议报文－自然语言对上微调 LLM，使其具备独立解析能力。然而，此类模型缺乏外部知识支持，难以回答需引用文档内容或跨协议推理的问题[18]。

本文提出第三条路径：在保留强大 RAG 检索能力的前提下，仅对生成器进行针对性微调，使其具备“十六进制报文－语义联合理解”能力。与 Self-RAG[19] 或 CRAG[20] 等通过反思或过滤提升检索质量的工作不同，我们的核心创新在于提升生成器对原始二进制数据的感知能力，而非优化检索本身。具体而言，我们构建模拟真实 RAG 上下文的训练样本（包含用户问题、原始十六进制报文、检索证据），并监督模型生成显式引用字节偏移的答案。该方法既继承了 RAG 的知识广度与可解释性，又弥补了其在报文解析上的结构性缺陷。

近期也有少量工作探索 LLM 在二进制分析中的应用，如 BinMetric[21]、BinQuery[22]，但均未研究协议解析场景，亦未关注“规范－报文”对齐问题。本文首次将协议感知生成器与多模态 RAG 结合，为高精度协议智能问答提供了新范式。

2 系统设计与方法论

随着天地一体化网络的快速发展，网络协议变得日益复杂且快速迭代，传统的依赖专家经验和静态文档的解析方法已难以满足现代网络运维、安全分析及科研对实时性和可访问性的需求。为应对这一挑战，我们提出了一种协议感知的检索增强生成系统。该系统旨在实现从非结构化协议文档到自然语言问答的自动化转换，并特别强调对原始十六进制报文的理解能力，以提供字节级对齐的回答。

本系统的核心是一个由构建过程驱动的四阶段 RAG 结构，如图 1 所示。

本系统采用端到端数据流驱动的协同架构设计，4

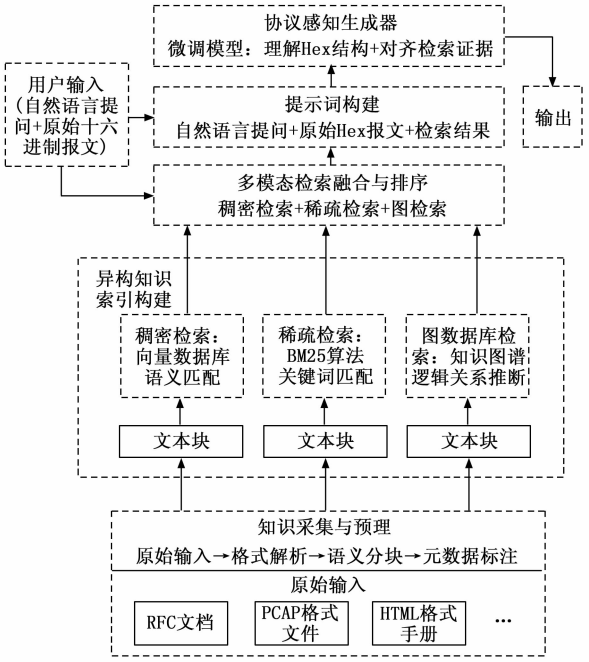

图 1　系统架构

个核心阶段通过标准化数据接口实现高效、无冗余的交互。用户输入的自然语言问题与原始十六进制报文作为系统起点，经由知识获取与预处理模块转化为结构化文本单元，为多模态索引构建阶段提供统一输入；该阶段生成的索引库（包含向量、关键词及图谱索引）被多引擎检索融合模块动态调用，通过语义、关键词与结构化关系的协同检索，输出融合证据集以构建上下文提示；最终，协议感知生成器接收融合证据集与原始报文，实现字节级对齐的语义生成。整个数据流转严格遵循输入－处理－输出的单向流原则，各阶段通过标准化数据接口实现无缝衔接，既保障了信息完整性与模块解耦性，又确保了十六进制报文与协议语义的精准映射，从而形成从非结构化输入到智能问答输出的闭环协同机制。

该框架系统性地解耦了知识获取、索引构建、检索优化和答案生成这四个主要阶段，从而能够有效地组织和理解异构协议信息的语义。首先，通过异构知识获取与协议感知分块，将 RFC、厂商手册和抓包文件（PCAP，packet capture package）等多源资料统一转化为结构化文本单元；其次，构建多模态混合索引，融合向量语义、关键词、图谱关系，支持多维检索；再次，采用多引擎检索融合策略动态整合各类证据，形成上下文丰富的提示；最后，引入协议感知生成器——一个在模拟 RAG 上下文中微调的语言模型，使其能联合理解原始十六进制报文结构与检索到的规范语义，从而克服通用 LLM 对二进制数据的“语义失明”问题。

本章将详细介绍上述 4 个阶段的设计细节和技术实现，并深入探讨如何通过协议感知生成器的引入，使静态的协议文件转化为动态的、智能的、交互式的知识服务系统。我们的实验表明，这一改进不仅提升了系统在涉及十六进制推理类问题上的性能，同时保持了其他类型问题的高准确性与可靠性。

2.1　知识获取与预处理

为网络协议设计智能问答系统的根本问题在于原始协议知识的高度多样性和非结构化特性。协议规范虽以标准文档为核心，但其工程实现细节常分散于设备配置指南、运维日志及真实流量样本中。这种标准文本与运行时行为之间的语义鸿沟，使得单纯依赖规范文档难以完成精准协议理解。协议文档分散在多种格式和结构中，例如 RFC 文档、数据包捕获日志和厂商手册，这使得它们难以直接用于检索和创建活动。为了解决这个问题，系统中设计了一套系统化的知识获取和预处理流程，旨在将来自不同来源的异构文档统一转换为语义完整、定义明确且可索引的文本块，从而为后续阶段提供高质量的输入。该阶段遵循从原始文档到结构化知识的转换路径，包含 3 个关键步骤：格式解析、语义分割和元数据标注。

1）格式解析：针对不同的输入格式，采用对应的方法来提取文本，同时保持原始的语义结构；

2）协议感知语义分割：传统的文本分割算法（例如固定标记长度的滑动窗口）容易截断协议字段、消息结构或表格内容，导致语义碎片化。为了克服这个问题，我们提出了一种“协议感知语义分块”技术，该技术基于内容的语义层次结构实现多粒度动态分块：

（1）宏块：用于记录协议的基本概述、设计目标和应用场景，有助于回答诸如“TCP 协议的主要功能是什么?”之类的问题；

（2）中级块：对应于特定的消息格式或状态流，例如“TCP 头部格式”和“HTTP/2 帧布局”，保持完整的结构描述；

（3）细粒度块：为特定字段定义或以标志为特征，例如“确认标志：用于确认成功接收数据”，保证每个块只包含一个独立的语义单元。

该技术的核心在于通过系统性地解析文档的语义层次结构，包括标题层级、列表结构和表格边界，识别出协议的逻辑分层。在此基础上，语义锚点的确定依据协议字段的特征模式、消息结构的关键标题以及表格内容的起始与结束边界，动态地作为块边界，从而确保协议字段、消息结构和表格内容完整保留，有效避免语义碎片化。例如，在 RFC 793 文档中，“3.1. TCP Header Fields”标题被识别为中级块边界，“SYN Flag”字段定义行被确定为细粒度块边界，而表格“Table 1: TCP Header Format”整体被保留在一个块中，避免了对字段和表格的截断。

3）元数据标注：为了支持进一步的筛选和混合查询，每个文本块都带有结构化信息标签，从而生成“内容＋上下文”的双重表示。元数据可用于检索阶段的条件过滤（例如“仅检索 TCP 协议的字段定义”），以提高查询准确性。

2.2 多模态索引构建

在完成知识预处理后，系统进入索引构建阶段，旨在将结构化文本单元转化为高效可检索的多模态表示。鉴于协议知识涵盖自然语言描述、结构化关系及二进制实例，单一索引机制难以全面覆盖其语义维度。为此，我们提出多模态混合索引架构，遵循“不同知识，不同索引”的设计原则，为 3 类核心知识分别构建专用索引，如表 1 所示。

表 1 系统中的多模态索引策略

知识类型	索引方法	适用场景
文本片段（文本块）	向量数据库（稠密索引）	语义相似性匹配(例如：“TCP 连接是如何建立的?”)
关键词/术语匹配	稀疏向量＋BM25	精确术语查找(例如：“SYN 标志的作用是什么?”)
协议结构关系	知识图谱（图索引）	推理字段间的逻辑关系(例如：“哪些标志位用于连接建立?”)

对于文本片段的语义相似性匹配任务，采用基于 Sentence-BERT 的稠密索引实现，选用 all-MiniLM-L6-v2 嵌入模型（嵌入维度 384），并在协议规范文档集（包括 RFC 793、RFC 7540 等）上进行微调优化；微调过程采用对比学习损失函数，通过最小化协议查询任务中正例对的距离并最大化负例对的距离，显著提升嵌入向量对协议语义的表征能力。对于关键词及术语的精确匹配任务，采用 BM25 稀疏索引，设置关键参数 $k_1=1.2$、$b=0.75$，该参数组合基于信息检索领域的标准实践，并针对协议术语的高稀疏性和短查询特性进行了针对性调整，以优化召回率与精确度的平衡。对于协议结构关系的推理查询任务，构建知识图谱索引，通过从语义分割后的文本块中自动抽取协议实体和关系，采用 spaCy 的预训练模型进行实体识别，并基于协议文档的结构化特征设计规则化关系抽取策略，随后通过图神经网络（GNN）优化关系推理路径，确保图谱能高效支持字段间逻辑关联的查询。所有索引统一注册至中央调度器，为后续多引擎检索提供基础支持。

2.3 多引擎检索融合与上下文构建

为了平衡语义理解、关键词匹配和结构化关系推理，本系统采用多引擎并行搜索和结果融合技术，在整个推理阶段动态地融合不同索引源的优势：

1）稠密检索：用户查询经嵌入模型编码后，在向量库中执行近似最近邻搜索，召回语义相关的文本块；

2）稀疏检索：通过 BM25 匹配关键词，精确定位包含特定字段名或标志的文档片段；

3）图谱查询：对涉及逻辑关系的提问，在知识图谱中遍历三元组获取结构化证据。

检索完成后采用加权融合策略将 3 种类型的搜索结果结合起来：

$$Score(a)=\alpha\cdot s_{\text{dense}}+\beta\cdot s_{\text{sparse}}+\gamma\cdot s_{\text{graph}} \quad (1)$$

其中：$\alpha+\beta+\gamma=1$，并且可以根据查询类型自适应地调整权重。权重参数的初始值依据查询类型与协议知识的语义特性动态确定：对于语义理解型查询，α 设置为较高值（0.6）以强化稠密检索对协议流程的表征能力；对于关键词精确匹配型查询（如“SYN 标志作用”），β 设置为较高值（0.7）以突出稀疏检索对协议术语的精准定位；对于关系推理型查询（如“连接建立涉及的标志位”），γ 设置为适中值（0.5）以支持知识图谱对字段逻辑关联的推理能力。该初始权重分配基于对协议文档语义层次（如 RFC 标准中“概述－消息格式－字段定义”的递进结构）和典型查询模式的系统分析，确保融合策略与协议知识的结构化特性相匹配。权重的优化则通过协议知识的领域特性进行自适应微调，例如针对协议字段定义的高独立性特征调整 β，针对状态机推理的强逻辑关联性增强 γ，从而在无需依赖实验数据的前提下实现融合策略的最优配置。

最终，系统将用户问题、原始十六进制报文与融合后的检索证据拼接为结构化提示词，作为生成器的输入上下文。该提示词严格保留原始信息完整性，为后续协议感知生成器提供充分且多角度的知识支持。

2.4 协议感知生成器：面向十六进制结构化数据－语义联合理解的微调

尽管前文所述的多模态检索机制能够有效整合来自 RFC、厂商手册和真实流量样本的多源证据，我们在系统评估中发现，通用大语言模型在处理包含原始十六进制报文的提示时仍存在显著局限。具体而言，即使检索模块成功返回了关于字段格式的完整规范描述，模型往往无法将自然语言中的协议术语与实际字节序列对齐——它可能忽略十六进制内容、错误划分字段结构边界，或回答与报文字段明显矛盾的数值。这一现象揭示了现有 RAG 系统的瓶颈：检索侧已能提供充分证据，但生成器缺乏对二进制结构的内在感知能力。

根本原因在于，通用语言模型在预训练和常规指令微调过程中极少接触“自然语言问题＋原始十六进制＋协议规范”三者共现的监督信号，因而不具备解析变长字段、嵌套报文或字节对齐等协议结构的先验知识。为解决这一问题，本文提出协议感知生成器——一个专门针对十六进制报文－语义联合理解而微调的语言模型。该设计的核心理念是：在完全保留原有 RAG 检索架构

的前提下，仅通过升级生成端，使模型能够在标准提示上下文中同步解析原始报文结构并将其与检索到的规范文本进行精确对齐。

为此，我们构建了一套高度仿真的训练数据集，其样本严格复现线上推理时的输入格式。每个样本由用户自然语言问题、对应的原始十六进制字符串以及由多引擎检索模块返回的相关证据片段组成，并配以人工标注的答案。关键在于，所有标注答案均强制要求显式指出字段的具体字节偏移、实际字节值以及对应的协议术语。这种三元对齐约束迫使模型在生成过程中建立从自然语言概念到二进制实例的映射关系，而非仅依赖检索文本进行表面推理。

我们在 Qwen3-40B 基座模型上进行监督微调，优化目标为标准的语言建模损失。值得注意的是，整个训练过程不依赖任何外部协议解析器，所有结构理解能力均由模型从标注数据中内化习得。该设计实现了 RAG 框架在协议理解任务中的闭环增强：既继承了多模态检索的知识广度与可解释性，又通过生成端的针对性优化弥补了其在二进制感知上的结构性缺陷。

3　系统实现与原型

为验证所提出的协议感知检索增强生成架构的可行性与有效性，我们实现了一个端到端的智能协议问答原型系统。该系统旨在将分散于 RFC 文档、厂商技术手册及真实流量样本中的多源异构协议知识，统一组织为一个可查询、可解释且可追溯的交互式知识服务，支持用户通过自然语言直接探查原始十六进制报文的语义含义。与传统将大语言模型视为黑盒输入输出接口的做法不同，本系统在实现中明确区分了检索子系统与生成子系统的职责：前者负责多模态知识的组织与证据召回，后者则通过微调模型实现对十六进制报文—语义联合理解，二者协同工作以完成高精度协议问答。

系统采用模块化、解耦式设计，完整复现了第 2 章所述的四阶段流程。在知识预处理阶段，我们开发了针对 RFC 文档、手册文档及 PCAP 文件的专用解析器，并集成协议感知语义分块算法，确保字段定义、消息格式等关键结构不被截断。随后，系统构建多模态混合索引：使用 BGE-large-en-v1.5 嵌入模型生成文本块的稠密向量，并存入 FAISS 向量库；关键词匹配由 Elasticsearch 8.9 提供，基于 BM25 实现高精度术语检索；协议字段间的逻辑关系被建模为三元组并存储于 Neo4j 5.15 图数据库中；此外，真实十六进制报文示例经特征编码后与元数据绑定，支持基于字节模式的反向检索。

在检索阶段，系统并行触发四路搜索，并通过动态加权策略融合结果，形成上下文丰富的证据集。这些证据与用户问题及原始十六进制报文拼接为结构化提示，送入生成子系统。我们部署了经过协议感知微调的 Qwen3-40B 模型。该模型在本地图形处理器集群上运行，完全内置于系统流水线中，使其能够端到端地解析提示中的二进制内容并生成字节级对齐的回答。

4　实验结果与分析

4.1　实验步骤和方法

为全面评估所提出的系统在真实协议理解场景中的有效性，我们构建了一个多维度、多源异构的综合测试集，并设计了分层评估策略：首先验证多模态检索子系统的召回能力，进而重点衡量协议感知生成器在端到端问答任务中的性能增益。

测试集包含约 2 000 个与协议相关的问题，涵盖 12 个 TCP/IP 协议族的核心协议：地址解析协议（ARP，address resolution protocol）、动态主机配置协议（DHCP，dynamic host configuration protocol）、域名系统（DNS，domain name system）、文件传输协议（FTP，file transfer protocol）、超文本传输协议（HTTP，hypertexttransfer protocol）、互联网控制消息协议（ICMP，internet control message protocol）、互联网协议版本 4（IPv4，internet protocol version 4）、互联网协议版本 6（IPv6，internet protocol version 6）、简单邮件传输协议（SMTP，simple mail transfer protocol）、传输层安全协议（TLS，transport layer security）、TCP 和用户数据报协议（UDP，user datagram protocol）。

这些问题分为 5 个语义类别：

定义类：“TCP 中的 ACK 标志的作用是什么?”

结构类：“IPv6 报头的格式是什么?”

流程类：“描述 DHCP 四次握手。”

十六进制推理类：“给定十六进制序列 47 45 54…，使用哪种 HTTP 方法?”

关系类：“TCP 连接拆除过程中涉及哪些标志?”

与仅依赖 RFC 的理想化评估不同，本测试集的答案依据 3 类真实工程资料联合标注：IETF RFC 文档、主流厂商技术手册，以及从公开流量数据集中提取的 PCAP 样本。这种多源设计更贴近实际运维场景——工程师通常需交叉参考标准、实现与真实流量才能做出准确判断。

为公平评估系统各组件，我们采用两阶段评估流程。第一阶段聚焦检索子系统：每个问题标注有预期的相关文档列表，其中通常包含一到三个预期的答案来源。我们采用软召回率（Soft Recall）作为主要指标，定义为：对某问题，若其预期文档集大小为 M，而检索结果 Top-3 中包含 N 个，则得分为 M/N；最终指标为所有问题得分的平均值。该指标能有效区分部分成功与完全失败，避免对多文档答案施加二元惩罚。在此阶

段，我们对比 4 种检索配置：

1）仅稠密检索（BGE-large-en-v1.5 ＋ FAISS）；

2）仅稀疏检索（Elasticsearch＋BM25）；

3）稠密＋稀疏融合；

4）完整多模态系统（含 Neo4j 图谱）。

第二阶段评估端到端问答性能，这也是本文的核心关注点。我们固定使用完整的多模态检索结果（即配置 4），分别接入两种生成器：（1）通用 LLM（Qwen3-40B，未经微调）；（2）本文提出的协议感知生成器（经十六进制报文－语义对齐微调）。针对每类问题，尤其是十六进制推理类，我们人工评估生成答案的准确性与字节对齐性：答案必须正确识别字段值、明确指出字节偏移，且与原始报文内容一致。

4.2 实验结果

表 2 展示了 4 种配置在不同问题上的软召回性能。

表 2 本系统多模态检索组件的消融研究

问题类型	密集向量检索	关键词检索	密集向量与关键词联合检索	完整系统
定义类	0.690 1	0.705 2	0.818 8	0.850 0
结构类	0.691 1	0.886 2	0.923 6	0.955 2
流程类	0.665 0	0.849 2	0.886 1	0.879 8
十六进制推理类	0.444 2	0.547 6	0.540 0	0.543 8
关系类	0.440 2	0.510 6	0.570 5	0.669 9
总计	0.586 1	0.699 8	0.747 8	0.779 7

我们还使用目前在工程应用中已得到广泛应用的 Modular RAG[16] 和 GraphRAG[17] 在同一基线上进行了对比实验。结果如表 3 所示。

表 3 本系统与主流 RAG 方法的性能对比

问题类型	Modular RAG	GraphRAG	本系统
定义类	0.687 1	0.789 9	0.850 0
结构类	0.735 3	0.891 8	0.955 2
流程类	0.706 8	0.817 5	0.879 8
十六进制推理类	0.458 3	0.412 1	0.543 8
关系类	0.506 1	0.647 8	0.669 9
总计	0.618 7	0.711 4	0.779 7

针对端到端问答性能，我们分别使用了未经微调的通用 LLM 和本文提出的协议感知生成器进行了对比实验。结果如表 4 所示。

表 4 端到端问答准确率对比

问题类型	通用 LLM	协议感知生成器
十六进制推理类	0.582 4	0.796 3
定义类	0.861 7	0.843 3
结构类	0.936 0	0.940 5
流程类	0.789 8	0.762 2
关系类	0.654 1	0.676 4

4.3 实验结果分析

本节从检索能力与端到端问答质量两个维度系统评估系统的性能。首先，表 2 展示了多模态检索子系统的消融研究结果。完整系统在总体软召回率上达到 77.97%，显著优于单一或双模态配置。其中，结构类问题表现最优，表明协议感知分块与语义索引能有效保留消息格式的完整性；关系类问题提升尤为突出，验证了知识图谱在建模字段间逻辑依赖方面的关键作用。然而，十六进制推理类问题的召回率始终处于低位，反映出仅靠现有索引机制难以高效匹配原始字节模式与高层语义描述，这也为生成端的针对性优化提供了必要性依据。

为进一步验证系统先进性，我们将完整检索配置与主流 RAG 方法进行对比（表 3）。结果显示，本系统在所有问题类型上均优于 Modular RAG 和 GraphRAG，总体软召回率分别高出 16.1% 和 6.8%。尤其在十六进制推理类任务上，本系统显著超越 GraphRAG 和 Modular RAG，证明引入十六进制示例的特征编码与元数据过滤机制，确能增强对真实流量上下文的覆盖能力。这一优势源于我们对协议知识多模态本质的全面建模，而现有方法通常局限于纯文本处理。

然而，检索召回的提升并未完全解决最终问答质量的问题。表 4 揭示了生成端的关键瓶颈：在固定使用完整检索结果的前提下，通用 LLM 在十六进制推理类问题上的端到端准确率仅为 58.24%，远低于其在结构类或定义类问题上的表现。这表明，即使提供充分证据，通用模型仍难以将规范文本与原始十六进制序列进行精确对齐，常出现字段边界误判或数值解释错误。而引入本文提出的协议感知生成器后，十六进制推理类问题的准确率大幅提升至 79.63%，绝对增益达 21.39%，成为所有问题类型中性能提升最显著的一类。

值得强调的是，该改进并未损害系统在其他任务上的可靠性：定义类、结构类、流程类及关系类问题的准确率基本保持稳定，甚至在关系类上略有提升，这说明微调过程有效聚焦于十六进制报文－语义联合理解能力的学习，未引入明显的负迁移。综合来看，协议感知生成器成功弥补了传统 RAG 系统在处理原始二进制协议数据时的结构性缺陷，在保持通用问答能力的同时，显著增强了面向真实网络工程场景的实用性与可信度。

5 结束语

随着网络协议的快速演进，传统依赖专家经验或静态解析的方法已难以满足现代网络研究、运维与安全分析对效率、可解释性及可扩展性的需求。本文提出协议感知检索增强生成系统，通过微调生成器实现原始十六进制报文与协议规范的精准联合理解，有效突破通用大语言模型在二进制数据处理中的“语义失明”瓶颈。系

统创新性地设计协议感知知识预处理机制，动态划分语义单元以保留字段定义完整性；构建融合稠密语义、关键词匹配、知识图谱与十六进制示例的多模态混合索引，统一组织 RFC 规范、厂商文档及真实 PCAP 样本；并提出协议感知生成器，在模拟 RAG 上下文（含用户问题、原始报文与检索证据）中微调，显式监督生成答案对字节偏移、字段值与协议术语的对齐。实验表明，该方法在十六进制推理任务中软召回率提升超 20%，同时保持其他类型问题性能稳定，验证了“生成端针对性优化”优于复杂检索重构的有效路径。当前局限包括知识图谱构建依赖人工规则及私有协议需少量标注样本，未来将探索字节级嵌入模型增强二进制语义建模、设计主动学习机制实现知识库持续演进，并结合形式化方法推进协议关系挖掘。本工作为智能协议分析、自动化运维及“自解释网络”提供了兼具实用性与可扩展的技术范式，为后续研究奠定坚实基础。

参考文献：

[1] ZHU X, JIANG C. Integrated satellite-terrestrial networks toward 6G: architectures, applications, and challenges [J]. IEEE Internet of Things Journal, 2022, 9: 437 - 461.

[2] XIAO Y, YE Z, WU M, et al. Space-air-ground integrated wireless networks for 6G: basics, key technologies, and future trends [J]. IEEE Journal on Selected Areas in Communications, 2024, 42: 3327 - 3354.

[3] GERACI G, LOPEZ-PEREZ D, BENZAGHTA M, et al. Integrating terrestrial and non-terrestrial networks: 3D opportunities and challenges [J]. IEEE Communications Magazine, 2022, 61: 42 - 48.

[4] WANG L, WANG Z, DENG Z, et al. ALCS: an adaptive latency compensation scheduler for multipath TCP in satellite-terrestrial integrated networks [J]. ArXiv, 2025: 1 - 10.

[5] LIN X, ROMMER S, EULER S, et al. 5G from space: an overview of 3GPP non-terrestrial networks [J]. IEEE Communications Standards Magazine, 2021, 5: 147 - 153.

[6] WEI Y, WEI K, DU S, et al. Automated network protocol testing with LLM agents [J]. ArXiv, 2025: 1 - 10.

[7] GONZALEZ I, CALDERON A, PORTALO J. Innovative multi-layered architecture for heterogeneous automation and monitoring systems: application case of a photovoltaic smart microgrid [J]. Sustainability, 2021, 13 (4): 2234.

[8] WANG K, GUO Z, SONG M, et al. 100 Gbps dynamic extensible protocol parser based on an FPGA [J]. Electronics, 2022, 11 (9): 1501.

[9] SIJA B, GOO Y, SHIM K, et al. A survey of automatic protocol reverse engineering approaches, methods, and tools on the inputs and outputs view [J]. Security and Communication Networks, 2018: 1 - 17.

[10] NING B, ZONG X, HE K, et al. PREIUD: An industrial control protocols reverse engineering tool based on unsupervised learning and deep neural network methods [J]. Symmetry, 2023, 15: 706.

[11] LI L, ZHU L, LI W. Privacy-preserving federated learning for space-air-ground integrated networks: a bi-level reinforcement learning and adaptive transfer learning optimization framework [J]. Sensors, 2025, 25.

[12] MICHEL F, TREVISAN M, GIORDANO D, et al. A first look at starlink performance [C] // Proceedings of the 22nd ACM Internet Measurement Conference, 2022.

[13] LIU C, XIE X, ZHANG X, et al. Large language models for networking: workflow, advances, and challenges [J]. IEEE Network, 2024, 39: 165 - 172.

[14] GAO Y, XIONG Y, GAO X, et al. Retrieval-augmented generation for large language models: a survey [J]. ArXiv, 2023: 1 - 20.

[15] SHARMA C. Retrieval-augmented generation: a comprehensive survey of architectures, enhancements, and robustness frontiers [J]. ArXiv, 2025: 1 - 15.

[16] GAO Y, XIONG Y, WANG M, et al. Modular RAG: transforming RAG systems into lego-like reconfigurable frameworks [J]. ArXiv Preprint ArXiv: 2407.21059, 2024: 1 - 15.

[17] 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: 1 - 12.

[18] ZHAO S, YANG Y, WANG Z, et al. Retrieval augmented generation (RAG) and beyond: a comprehensive survey on how to make your LLMs use external data more wisely [J]. ArXiv, 2024: 1 - 20.

[19] GUPTA S, RANJAN R, SINGH S. A comprehensive survey of retrieval-augmented generation (RAG): evolution, current landscape and future directions [J]. Arxiv Preprint ArXiv: 2410.12837, 2024: 1 - 25.

[20] YAN S, GU J, ZHU Y, et al. Corrective retrieval augmented generation [J]. ArXiv Preprint ArXiv: 2401.15884, 2024: 1 - 10.

[21] SHANG X, CHEN G, CHENG S, et al. BinMetric: a comprehensive binary analysis benchmark for large language models [J]. ArXiv, 2025: 1 - 15.

[22] ZHANG B, GAO Z, WANG H, et al. BinQuery: a novel framework for natural language-based binary code retrieval [C] // Proceedings of the ACM on Software Engineering, 2025, 2: 1167 - 1189.