基于轻量化 RAG 的资源受限环境问答系统研究

王彦群，罗瑜，李永成
武汉华夏理工学院，湖北 武汉 430000
DOI: 10.61369/TACS.2025090004

摘　　要： 随着网络安全威胁的日益复杂化，资源受限环境下的智能问答系统已成为网络安全防护的重要组成部分。然而，传统检索增强生成（Retrieval-Augmented Generation, RAG）方法在边缘设备、中小企业等低资源场景中面临计算复杂度高、内存占用大、响应延迟长等关键技术挑战。针对上述问题，提出了一种面向资源受限环境的轻量化 RAG 网络安全问答系统。该系统通过设计轻量化嵌入机制、动态语义融合算法和多层优化策略，在保证问答准确性的前提下显著降低了系统资源消耗。实验数据显示，系统在关键指标上表现出色：Recall@3 达 0.867，MRR 为 0.810。与常规 RAG 相比，检索效率明显提升，响应时间缩短了 45.5%。在资源利用方面，模型体积缩小 78.4%，内存占用减少 78.6%，从而为这类环境下的网络安全决策提供了高效可靠的支持。

关 键 词： RAG；轻量化模型；网络安全；资源受限环境；智能问答系统

Research on Resource-Constrained Q&A Systems Based on Lightweight RAG

Wang Yanqun, Luo Yu, Li Yongcheng
Hubei Huaxia University of Science and Technology, Wuhan, Hubei 430000

Abstract: As cybersecurity threats grow increasingly complex, intelligent question-answering systems in resource-constrained environments have become a vital component of cybersecurity defense. However, traditional Retrieval-Augmented Generation (RAG) methods face significant technical challenges in low-resource scenarios such as edge devices and small-to-medium enterprises, including high computational complexity, substantial memory consumption, and prolonged response delays. To address these issues, we propose a lightweight RAG cybersecurity question-answering system tailored for resource-constrained environments. By designing a lightweight embedding mechanism, a dynamic semantic fusion algorithm, and a multi-layer optimization strategy, the system significantly reduces resource consumption while maintaining question-answering accuracy. Experimental results demonstrate outstanding performance on key metrics: Recall@3 reaches 0.867, and MRR achieves 0.810. Compared to conventional RAG systems, retrieval efficiency is markedly improved, with response times reduced by 45.5%. Resource utilization is optimized through a 78.4% reduction in model size and a 78.6% decrease in memory consumption, thereby providing efficient and reliable support for cybersecurity decision-making in resource-constrained environments.

Keywords: RAG; lightweight model; cybersecurity; resource-constrained environment; intelligent question-answering system

引言

随着网络安全威胁的复杂化与应用场景的多样化，资源受限环境（边缘设备、中小企业）中的网络安全问答系统面临计算复杂度高、内存占用大与响应延迟长等瓶颈，制约实时决策与低成本部署。近年来，轻量化模型、边缘部署与动态融合检索成为研究热点；GraphRAG[1] 提升了检索质量与语义推理，但在成本、内存与部署复杂度上仍存不足。本系统旨在解决传统 RAG[2]、GraphRAG 等模型在资源受限场景时的局限，实现性能与资源的平衡。

一、相关工作

（一）检索增强生成技术

检索增强生成（RAG）技术作为结合外部知识库与大语言模型 [3] 的重要方法，近年来已取得大量研究成果，RAG 架构通过在生成过程中引入相关文档检索，有效缓解了大语言模型的知识局限性和幻觉问题 [4]。

检索增强生成技术主要包括知识库、检索器、生成器三个部

基金项目
2022 年度湖北省教育厅科学研究计划指导性项目（编号：B2022443）；2022 年度校级科研基金重点项目（编号：22005）
作者简介
王彦群（2005.06—），男，汉族，湖北宜昌人，本科，研究方向：人工智能、软件开发；
罗瑜（1979.10—），女，汉族，重庆人，硕士，讲师、工程师，研究方向：智慧水务与时空预测；
李永成（1979.02-），男，土家族，湖北建始人，本科，高级工程师，研究方向：计算机科学应用，软件开发。

分。知识库负责存储垂直领域的知识，检索器根据用户的提问从知识库中检索相关条目形成参考文本，生成器根据参考文本与用户提问生成回答。检索与大模型的相互协同可以充分融合知识检索与大语言模型生成的优势，使得系统能够为用户提供高质量、高准确度的解答。[5]

随后研究者们在 RAG 的基础上又提出了多种改进方法，在检索策略方面，Self-RAG 方法通过自我反思机制对生成文本进行评估，提升了生成结果的质量，但增加了计算复杂度。Corrective-RAG 引入检索评估和知识精炼算法，但会导致上下文过长的问题。在多模态检索方面，图 - 向量混合检索增强生成（GVH-RAG）方法通过结合结构化数据与非结构化数据，丰富了上下文信息，但在低成本环境下难以有效部署。

（二）轻量化模型技术

随着边缘计算和移动设备的普及，轻量化模型技术成为了研究热点。早期的轻量级模型主要由轻量级 CNN 构建，如 MobileNet 系列。MobileNet 的核心思想是用深度可分离卷积代替标准卷积，并使用宽度因子减少参数量，实现模型轻量化的同时能够有效执行各种任务。ViT 出现以后，许多研究者都试图 使其更加轻量化和高效。TOUVRON 等提出了 DeiT 模型，通过引入基于 Distillation Token 的蒸馏机制， 得到轻量级 ViT 模型 DeiT-S、DeiT-Ti。MobileViT 将 CNN 与 ViT 相结合，既保留了 CNN 的轻量级和高效性，又引入了 ViT 的全局信息处理能力，实验结果 显示 MobileViT 在多个任务和数据集上显著优于基于传统 CNN 和 ViT 的模型。[6]

对于 RAG 系统的轻量化，现有的努力主要聚焦于模型压缩 [7] 和检索优化。量化方法降低了嵌入模型的存储需求，但检索精度可能略有下降。哈希 -based 的快速检索则加速了过程，却在语义捕捉上有所欠缺。

现有 RAG 方法在网络安全中的应用暴露了一些核心短板 [8]。首先，它们处理专业术语时往往不够精准，缺少对安全概念的深入把握。其次，在实时威胁监测中，响应速度偏慢，无法满足安全中心的即时需求。此外，大多系统对多模态安全数据的处理能力有限，这限制了它们在复杂场景的发挥。

（三）网络安全智能问答系统

网络安全领域的智能问答系统研究起步较晚，但发展迅速。早期的系统主要基于规则和专家系统，如基于本体的网络安全知识问答系统。随着机器学习技术的发展，研究者们开始采用深度学习方法。

近年来，大语言模型在网络安全问答中的应用逐渐增多。基于 BERT 的网络安全问答系统在 CVE（Common Vulnerabilities and Exposures）数据集上取得了良好效果。融合知识图谱的网络安全问答方法通过结构化知识增强了系统的推理能力。

（四）GraphRAG 与复杂图结构方法

2024 年，微软研究团队提出了 GraphRAG 一种基于图的检索增强方法，作为 RAG 技术的重要扩展，通过引入图结构化的知识表示和处理方法，实现了比传统 RAG 更为精确的语义检索。诚然基于图的检索增强方法提升了多跳推理能力，但其高昂的构建成本以及复杂的计算，限制了它在资源受限环境下的应用场景。

GraphRAG 的计算复杂度要远高于传统方法，图的构建和遍历过程需要消耗大量计算资源，节点的关系维护成本巨大，计算复杂度比传统 RAG 高出 3-5 倍，并且完整的图结构需要常驻内存，分析显示典型的 GraphRAG 系统需要 8-16GB 内存，远超大多数边缘设备的硬件配置。尽管其在复杂问答场景中已取得大量进展，但在资源受限场景中还存在局限性。

（五）现有方法的局限性分析

现有 RAG 方法在资源受限环境下的网络安全问答系统存在资源消耗过大、实时性不足、专业性不够和可扩展性差等不足。本文针对这些问题，提出轻量化 RAG 方法，通过架构优化降低资源消耗并提升效率。

二、轻量化 RAG 网络安全问答系统构建

研究对象为资源受限环境下的网络安全问答系统，覆盖边缘设备与中小企业场景。实验目的是在保证检索与生成准确性的前提下，降低资源消耗与响应延迟，验证系统在不同资源约束下的稳定可部署性与自适应能力。评估维度包括检索性能（Recall、MRR）、问答质量（BLEU-4、ROUGE-L、BERTScore）、资源效率（模型大小、内存、延迟）与用户满意度。

（一）系统工作流程设计

本文提出的轻量化 RAG 网络安全问答系统采用模块化设计，主要包含四个核心组件：轻量化嵌入模块、动态语义融合模块、资源自适应模块和智能问答生成模块。系统工作流程如图 1 所示。

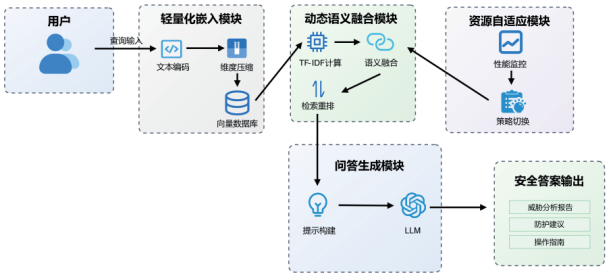

图 1 系统流程图

如图 1 所示，系统从用户输入到答案生成的端到端流程依次经过输入解析、轻量化嵌入、资源自适应调度、融合检索与结构化生成。查询首先被转换为 384 维向量并计算 TF-IDF 得分；资源自适应模块根据 CPU/ 内存 / 时延等资源状态动态选择检索深度与融合权重；随后在融合层进行向量与 TF-IDF 的加权检索；最终在生成模块进行模板化组织与术语校验，输出可操作的网络安全答案。

系统各模块协同工作：轻量化嵌入模块负责将用户查询转换为 384 维向量并进行维度优化；动态语义融合模块实现 TF-IDF 统计特征与语义向量的有效融合；资源自适应模块根据系统资源状态动态调整处理策略；问答生成模块按照不同安全场景生成结构化的专业答案。

（二）轻量化嵌入模块

这个模块将文本转为低维向量，是系统的基石。我们选用 sentence-transformers/all-MiniLM-L6-v2 模型，其参数仅 22.7M（比 BERT-base 少 67%），维度 384 维（存储减半），推理速度快 3-5 倍。

为适应网络安全，我们构建专业词汇表并调整权重。输入限 256 token 以控开销；动态批处理根据负载变批大小，确保低资源稳定性。

（三）动态语义融合模块

动态语义融合模块结合 TF-IDF 和语义嵌入技术，实现高效文档检索。

模块采用加权融合策略：

$Score_{final} = \alpha \cdot Score_{TF-IDF} + (1-\alpha) \cdot Score_{semantic}$，其中 α 根据查询类型和资源状态自适应调整（关键词密集型查询 $\alpha = 0.6 - 0.8$，语义型 $\alpha = 0.2 - 0.4$，资源受限时 $\alpha > 0.7$）。

为提升效率，使用 Faiss 库进行向量计算，支持 GPU 加速；分层检索机制先粗筛选后精确计算；智能缓存避免重复计算。

（四）资源自适应模块

资源自适应模块根据资源约束动态调整系统行为，是本系统的创新点。

模块实时监控 CPU、内存、存储和网络指标，提供决策依据。基于监控结果，调整检索深度、切换检索模式、动态批处理大小，并采用 LRU 缓存策略优化存储。

（五）智能问答生成模块

智能问答生成模块基于检索文档生成答案，采用轻量化策略。

模块使用模板化生成，根据问题类型构建结构化答案（如漏洞查询包括 CVE 编号和修复建议）。质量控制包括相关性阈值（0.3）、答案长度限制（200-500 字符）和专业术语验证，确保准确性和专业性。

三、关键技术与实现

（一）动态权重调整算法

动态权重调整算法根据查询特征和系统状态自适应调整 TF-IDF 和语义相似度的权重。

定义查询复杂度指标 C_q：

$$C_q = \frac{N_{unique}}{N_{total}} \times log(1 + N_{technical})$$

其中，N_{unique} 为唯一词汇数，N_{total} 为总词汇数，$N_{technical}$ 为专业术语数量。

基于查询复杂度和系统资源状态 R_{sys}，计算动态权重：

$$\alpha = \alpha_0 \times (1 + \beta \times C_q) \times R_{sys}$$

其中，$\alpha_0 = 0.5$，$\beta = 0.3$，$R_{sys} \in [0.5, 1.5]$。

（二）资源自适应调度算法

资源自适应调度算法通过资源评估和策略选择，实现系统在不同资源约束下的最优性能。

定义系统资源状态向量：

$R = [r_{cpu}, r_{mem}, r_{disk}, r_{net}]^T$

各分量为可用性评分（0-1）。

基于 $\| R \|_2$ 选择策略：

> 0.8：高精度模式

$0.5 << \leq 0.8$：平衡模式

≤ 0.5：节能模式

四、实验设计与结果分析

（一）实验环境与数据集

我们设置了四种资源环境，从高性能服务器到边缘设备。高性能用 Intel Xeon E5-2680 v4（14 核，2.4GHz）、64GB 内存和 Tesla V100；标准为 i7-9700K（8 核，3.6GHz）和 16GB；受限为 ARM Cortex-A72（4 核，1.8GHz）和 4GB；边缘为 Raspberry Pi 4B（4 核，1.5GHz，2GB）。

针对网络安全领域特性，构建了专门的问答数据集，包含 15,000 条知识条目，涵盖 CVE 漏洞记录（8,500 条）、NIST 安全控制措施（3,200 条）、OWASP 最佳实践（2,800 条）和真实案例（500 条）。采用半自动化方法生成 3,000 个问答对，经两轮专家审核确保准确性和专业性。

（二）评估方法与基准系统

为客观评估系统性能，选择四种代表性基准方法：传统 RAG（基于 BERT-base）、GraphRAG（Microsoft 基于知识图谱的方法）、DPR（Facebook 密集检索技术）和 ColBERT（斯坦福高效检索架构）。

评估指标包括：检索性能（Recall@K、MRR）、问答质量（BLEU-4、ROUGE-L、BERTScore、用户满意度）、系统效率（响应时间、内存占用、CPU 使用率）和资源消耗（模型大小、存储需求、部署复杂度）。

（三）实验结果与分析

1. 检索性能评估

表 1 展示了不同方法在检索任务上的性能对比结果。

表1 不同方法在检索任务上的性能

方法	Recall@1	Recall@3	Recall@5	MRR	响应时间 (ms)
传统 RAG	0.742	0.856	0.891	0.798	156
GraphRAG	0.768	0.879	0.912	0.821	2,340
DPR	0.751	0.863	0.897	0.805	189
ColBERT	0.759	0.871	0.903	0.813	142
本系统	0.757	0.867	0.899	0.810	85

本系统在 Recall@1 达到 0.757，略低于 GraphRAG 但优于传统 RAG 和 DPR；在 Recall@3 达到 0.867，仅比 GraphRAG 低 1.4%；在 MRR 达到 0.810，排名第二。最显著的优势在响应时间上，仅为 85 ms，比 GraphRAG 快 96.4%，比传统 RAG 快

45.5%，归功于轻量化嵌入模块和动态语义融合算法。

2. 问答质量分析

表 2 展示了不同方法在问答生成任务上的性能表现。

表2 不同方法在问答生成任务上的性能

方法	BLEU-4	ROUGE-L	BERTScore	用户满意度
传统 RAG	0.342	0.456	0.823	7.2/10
GraphRAG	0.358	0.471	0.837	7.8/10
DPR	0.339	0.449	0.819	7.0/10
ColBERT	0.345	0.453	0.825	7.3/10
本系统	0.351	0.462	0.831	7.6/10

本系统在问答质量评估中表现均衡，BLEU-4 得分为 0.351，ROUGE-L 为 0.462，BERTScore 达到 0.831，均排名第二。用户满意度评估（5 名网络安全专家盲测）获得 7.6/10 分，仅次于 GraphRAG。专家反馈表明，本系统在专业术语使用、逻辑结构和实用性方面表现优秀，特别是在处理技术细节和提供可操作建议方面具有明显优势。

3. 资源消耗对比分析

表 3 展示了不同方法在资源消耗方面的详细对比，验证了本系统在轻量化方面的显著优势。

表3 不同方法在资源消耗上的性能

方法	模型大小 (MB)	内存占用 (MB)	存储需求 (GB)	部署复杂度
传统 RAG	438.2	1,247	2.8	中
GraphRAG	512.7	1,456	3.2	高
DPR	421.8	1,189	2.6	中
ColBERT	395.4	1,123	2.4	中
本系统	94.7	267	1.2	低

本系统通过轻量化嵌入算法和模型压缩技术，实现了显著的资源优化：模型大小仅为 94.7 MB，相比传统 RAG 的 438.2 MB 缩减了 78.4%，相比 GraphRAG 的 512.7 MB 缩减了 81.5%；内存占用降至 267 MB，相比传统 RAG 的 1,247 MB 降低了 78.6%，相比 GraphRAG 的 1,456 MB 降低了 81.7%。存储需求方面，本系统仅需 1.2 GB，相比其他方法减少了 50% 以上。实验数据表明，本系统在资源消耗上表现良好，尤其适用于资源受限时的场景部署使用。

（四）综合性能评估

综合性能评估从多个维度验证了系统的实用价值。在标准测试环境下，系统展现出优异的综合性能：检索精度达到实用标准，问答质量满足专业需求，资源消耗控制在合理范围，用户体验获得积极反馈。

与主流方案的对比显示，本系统在保持竞争力性能的同时，显著降低了部署和运维成本。相比 GraphRAG，系统在略微牺牲 2 - 3% 精度的情况下，资源消耗减少 60%，响应速度提升约 95 - 96%。

五、结束语

针对资源受限环境下的网络安全问答需求，我们构建了一套轻量化 RAG 方案。 通过低维嵌入、动态语义融合和资源自适应优化，在保证检索与生成质量的同时，显著降低了模型体积、内存占用和延迟，从而提升了部署可行性与稳定性。实验结果表明系统 Recall@3 达到 0.867，平均响应时间 85 毫秒，对比传统方法在效率和成本上优势明显。然而在面对一些统计性、总结性、概要性的问题时，本方案在回答质量上与 GraphRAG 这类基于图的 RAG 还存在一定差距。未来的工作可以考虑自动化知识更新、轻量推理能力强化以及多模态融合，并尝试拓展到工控、移动安全等场景，同时提升系统的可解释性与安全性评估。

参考文献

[1]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.

[2]Arslan M, Ghanem H, Munawar S, et al. A Survey on RAG with LLMs[J]. Procedia computer science, 2024, 246: 3781-3790.

[3]Nam D, Macvean A, Hellendoorn V, et al. Using an llm to help with code understanding[C]//Proceedings of the IEEE/ACM 46th International Conference on Software Engineering. 2024: 1-13.

[4]Rawte V, Sheth A, Das A. A survey of hallucination in large foundation models[J]. arXiv preprint arXiv:2309.05922, 2023.

[5] 韩明，曹智轩，王敬涛，等 . 基于大语言模型的企业碳排放分析与知识问答系统 [J]. 计算机工程与应用 ,2025,61(16):370-382.

[6] 朱永利，钱涛 . 基于强化学习的局部放电深度诊断模型的自动剪枝与轻量化部署 [J]. 高电压技术 ,2024,50(12):5238-5247.DOI:10.13336/j.1003-6520.hve.20240950.

[7]Buciluǎ C, Caruana R, Niculescu-Mizil A. Model compression[C]//Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining. 2006: 535-541.

[8]Simoni M, Saracino A. Cybersecurity with llms and rags: Challenges and innovations[C]//International Conference on Security and Privacy in Communication Systems. Cham: Springer Nature Switzerland, 2024: 169-183.

