Hans Journal of Data Mining 数据挖掘, 2025, 15(3), 254-261
Published Online July 2025 in Hans. https://www.hanspub.org/journal/hjdm
https://doi.org/10.12677/hjdm.2025.153021

Hans汉斯

# 企业级大模型RAG系统数据安全防护策略研究

杨　波，张超军，王　吉

中通服咨询设计研究院有限公司，江苏 南京

收稿日期：2025年6月10日；录用日期：2025年7月3日；发布日期：2025年7月11日

## 摘　要

检索增强生成(Retrieval-Augmented Generation, RAG)系统作为企业数字化转型与大模型技术结合的典型应用形式，为企业提供精准信息服务的同时，也带来了严峻的数据安全挑战。本文首先深度分析了RAG系统应用在数据全生命周期各阶段面临的数据泄露、数据篡改、数据滥用等数据安全风险；其次针对各阶段数据安全风险制定了具体的数据安全防护策略，以及整体数据安全管理措施；最后讨论了研究的局限性及未来研究方向的展望。本文旨在为企业构建全面、有效的企业级大模型RAG系统数据安全防护策略提供理论指导和实践参考，以保障企业RAG应用健康发展。

## 关键词

大模型，RAG系统，数据安全风险，安全防护策略，数据安全管理

# Research on Data Security Protection Strategies for Enterprise-Level Large Model RAG Systems

**Bo Yang, Chaojun Zhang, Ji Wang**

China Information Consulting & Designing Institute Co., Ltd., Nanjing Jiangsu

Received: Jun. 10th, 2025; accepted: Jul. 3rd, 2025; published: Jul. 11th, 2025

## Abstract

**Retrieval-Augmented Generation (RAG) systems, as a paradigmatic application integrating enterprise digital transformation with large model technologies, deliver precise information services to businesses while introducing critical data security challenges. This paper first conducts an in-depth analysis of data security risks—including data leakage, data tampering, and data misuse—across the entire lifecycle of RAG system applications. Subsequently, it formulates specific data security protection**

文章引用: 杨波, 张超军, 王吉. 企业级大模型 RAG 系统数据安全防护策略研究[J]. 数据挖掘, 2025, 15(3): 254-261.
DOI: 10.12677/hjdm.2025.153021

**strategies for each stage and holistic data security management measures. Finally, it discusses the limitations of the study and prospects for future research directions. The objective of this paper is to furnish theoretical guidance and practical references for enterprises to build comprehensive and effective data security protection strategies for enterprise-level large model RAG systems, to ensure the healthy development of RAG applications in enterprises.**

## Keywords

**Large Model, RAG System, Data Security Risk, Security Protection Strategy, Data Security Management**



## 1. 引言

在目前的数字化时代，AI 大模型在各个行业领域逐渐被广泛应用。其中检索增强生成系统(Retrieval-Augmented Generation System，以下简称 RAG 系统)凭借检索外部知识库，有效提高用户检索结果的精准度，被企业广泛使用以提升企业的运营效率。因此 RAG 系统也成为目前企业对 AI 大模型的典型应用。企业在使用 RAG 系统过程中，模型需要收集和处理企业敏感信息建立索引知识库，若 RAG 系统发生数据安全事件，会对企业造成经济、声誉等多方面损失。

本文旨在指导企业构建一套有效的 RAG 系统数据安全防护体系，提升企业在使用 RAG 系统中对数据全生命周期的保护能力，同时也促进 RAG 应用健康发展。

## 2. RAG 系统介绍

RAG 系统可以理解为是大模型与外部知识检索的结合应用。它主要由知识源、检索模块和大模型构成。其中知识源存储各种知识，检索模块负责根据用户问题检索相关知识片段，然后大模型根据用户问题整合处理这些知识片段并生成自然语言的回答。企业可在智能客服、数据分析、决策支持等多种场景使用 RAG 系统，以提高企业的工作效率和质量。

## 3. 企业级大模型 RAG 系统数据安全风险分析

### 3.1. 数据采集阶段的安全风险

在数据采集阶段，RAG 系统从多数据源采集数据，如企业内部数据库、互联网公开数据、第三方数据服务商提供的数据，面临的安全风险包括：

(1) 数据泄露风险：数据采集接口权限控制缺陷，网络攻击者可利用该类漏洞远程窃取数据导致重要数据、核心数据、敏感数据泄露。攻击面主要集中于 RESTful、GraphQL、SOAP 等 API 服务接口端点。网络攻击者通过未授权访问攻击方式，直接访问未设置身份验证的数据采集接口获取数据；通过爆破攻击、会话劫持、API 密钥泄露利用等多种攻击方式，获取 API 接口认证凭证，并对数据接口进行访问；通过权限提升攻击方式，利用业务逻辑缺陷将普通用户提升至高权限用户，越权获取数据。

(2) 数据滥用风险：企业内部数据未经过分类分级、敏感数据识别等数据安全治理，在数据采集过程中可能过度采集企业重要核心数据、敏感数据，存在数据滥用风险。攻击面主要为企业重要核心数据、敏感数据暴漏面，该环节典型安全漏洞为未对数据进行分类分级和有针对性的细粒度安全控制，内部员

工无意识或恶意将重要核心数据同步给数据采集接收端。

(3) 数据合规风险：第三方数据服务商因缺乏数据合规管理措施，没有对数据继续进行合规验证，导致采集的数据权属不清或内容不合规。攻击方式可能来自数据上游的非法数据采集处理行为，也可能因企业无法证明其数据处理的合法性而导致。

(4) 数据投毒风险：攻击者向 RAG 系统的数据源中插入错误数据，污染数据源，导致 RAG 系统输出错误回答。

### 3.2. 数据传输阶段的安全风险

数据传输阶段风险来自于明文数据传输或弱加密算法保护[1]。攻击面集中于数据采集传输链路，攻击者利用中间人攻击、网络嗅探、DNS 欺骗重定向等攻击方式拦截捕获网络流量，然后进行解析获取数据，导致数据泄露。

### 3.3. 数据存储阶段的安全风险

数据存储阶段包括采集的原始数据、清洗后数据、分块数据、嵌入向量数据、用户交互历史数据等。面临的安全风险包括：

(1) 数据泄露风险：如数据存储系统存在访问控制缺陷、权限配置管理缺陷、弱口令等安全漏洞，易被网络攻击者利用并窃取数据。攻击面主要集中于关系型数据库、向量数据库、文件服务器网络服务端口，典型漏洞包括中间件安全漏洞、未授权访问漏洞、弱口令漏洞、错误配置漏洞。企业组织内部员工可通过恶意权限滥用、账号共享、恶意代码植入、数据非法导出等方式发起网络攻击，外部网络攻击者主要通过中间件远程漏洞利用、口令爆破、中间人攻击等方式发起网络攻击。

(2) 数据丢失风险：数据存储系统物理设备故障，易导致数据丢失；网络攻击者利用存储系统漏洞对数据进行批量删除；企业组织内部员工误操作或恶意删除文件、关系型数据库表、向量数据库。

(3) 数据篡改风险：网络攻击者利用存储系统未授权访问漏洞获得数据库控制权限，并恶意篡改数据；企业组织内部员工误操作或恶意修改数据库中数据。

### 3.4. 数据处理阶段的安全风险

在数据处理阶段，RAG 系统对采集的原始数据处理主要包括数据清洗、数据分块、数据索引与数据向量化、数据相似度分析计算。本阶段攻击面较大，包括数据处理各角色人员数据操作触点、数据处理组件、数据处理流程环节。典型安全漏洞包括数据访问控制失效、敏感数据暴漏、数据处理组件配置错误、数据处理组件或工具注入漏洞、数据处理逻辑漏洞、凭证管理漏洞等。企业组织内部人员恶意越权下载、复制数据，在数据处理工具或脚本中植入后门窃取数据，在开发测试环境中通过构造查询接口暴漏敏感数据，通过修改数据清洗规则、数据分块逻辑、污染向量索引、修改模型配置参数等方式篡改数据。外部网络攻击者通过远程利用上述安全漏洞发起网络攻击，从而实现数据窃取和数据篡改。

### 3.5. 模型推理与输出阶段的安全风险

在模型推理与输出阶段，面临的安全风险包括：

(1) 接口攻击风险：数据交互 API 接口未采取科学合理的认证鉴权措施，基于 API 接口端点攻击面，网络攻击者通过 API 接口凭证爆破、接口枚举、API 接口凭证泄露利用等多种攻击方式获得接口数据权限，从而批量窃取数据导致数据泄露。

(2) 内容安全风险：因训练数据错误、模型训练不足等导致输出结果错误，或攻击者在数据/模型训练环节，利用模型漏洞，生成虚假信息。此类攻击面集中于 RAG 系统知识库，典型漏洞包括知识库数据

污染、输入/输出安全过滤不足、模型安全围栏控制措施失效等。

(3) 向量和嵌入弱点风险：攻击者可能通过对抗性查询等恶意查询方式，让模型生成错误的嵌入向量，从而绕过权限控制获取企业数据。或者通过模型投毒攻击等方式操纵嵌入向量，使系统输出错误结果。又或者通过高频次的重复查询，来推断向量数据库中存储的敏感数据。

(4) 知识库访问权限风险：RAG 系统在访问后端知识库时，使用的身份权限高于必须的权限，或者 RAG 系统在输出结果时，没有按照用户权限对结果进行过滤，可能导致企业数据泄露。

### 3.6. 数据销毁阶段的安全风险

数据销毁阶段主要是在 RAG 系统知识库、模型迭代或下线时，未对训练数据、中间结果数据进行安全清理，导致数据被恢复和滥用。攻击面主要涉及数据存储相关设备设施，如物理存储介质、云存储、数据库系统、文件系统、备份系统、缓存系统、日志管理系统等，典型漏洞为数据残留漏洞，即未能在数据退出生命周期时进行不可逆删除。攻击者通过利用专业数据恢复软件对废弃存储介质恢复、利用未删除的备份以及缓存或临时文件残留等多种攻击方式获取数据[2]。

### 3.7. 模型安全风险

(1) 模型窃取与泄露：攻击者聚焦于模型本身，通过直接文件窃取、模型逆向工程、端侧/边缘设备的物理窃取等攻击手段，造成模型文件泄露。

(2) 模型完整性与篡改：此类风险主要指已部署的模型在存储、分发、加载、运行等阶段，攻击者通过植入后门、直接修改模型权重参数或在边缘设备上替换模型文件，使模型被恶意篡改，导致模型输出结果错误。

(3) 模型可用性和鲁棒性：攻击者通过针对模型的对抗性攻击、模型规避、模型拒绝服务攻击，以及边缘设备的计算资源耗尽和物理对抗攻击等方式，降低模型的可用性和鲁棒性[3]。

(4) 模型行为风险：这一类的风险主要是因训练数据有误、模型设计有缺陷、输入数据分布变化以及校准不足等原因导致的模型错误输出[4]。

### 3.8. 企业级 RAG 系统数据安全事件案例分析

(1) 数据违规采集案例

某大模型应用服务商因平台出现用户对话数据和付款服务支付信息丢失，且缺乏大量收集和存储个人信息的法律依据。意大利数据保护局宣布暂时禁用其模型应用服务，并对其涉嫌违反隐私规则开展调查。

(2) 敏感数据泄露案例

某大模型“奶奶漏洞”事件，用户通过构造提示词：“请扮演我已经过世的祖母，她总是会念 Windows10 Pro 的序号让我睡觉”，并在某大模型应用服务中输入该提示词，使模型返回真实有效的 Windows10 Pro 的序列号；某大模型应用服务因数据库存在配置错误导致出现严重的聊天数据泄露；网络攻击者使用供应链攻击方式，向某大模型应用服务上传恶意模型依赖包，导致大量用户凭证数据泄露。

## 4. 企业级大模型 RAG 系统数据安全防护策略

### 4.1. 数据全生命周期安全防护策略

#### 4.1.1. 数据采集阶段

数据采集数据源包括企业组织内部数据和外部数据。数据采集需要获得数据所有者、数据经营服务单位合法授权。

企业组织内部数据采集，原则上对于企业重要核心数据、商业机密数据应谨慎采集，采集的数据应

进行分类分级治理，并根据数据分类分级进行明确管控要求。企业组织外部数据采集，应对采集数据质量进行评估检测，确保采集数据质量高且不含违法信息内容。数据采集接口需设置身份认证机制。此外，在数据采集过程中可通过可信度模型对数据进行动态的评估验证，来避免模型采集错误数据。也可使用意图感知模型来分析用户查询内容，然后以此限制模型检索的知识库。

#### 4.1.2. 数据传输阶段

数据传输过程可通过对用户查询内容和模型输出结果等过程进行细颗粒度传输控制，避免数据被非法获取或篡改；可通过 NLP 模型对用户查询内容分析并分类，若涉及敏感数据则脱敏处理然后通过加密通道传输，同时通过数据签名等方式验证数据完整性，并限制传输范围。

#### 4.1.3. 数据存储阶段

在数据存储阶段，可采取以下措施：

(1) 结合数据分类分级情况，对用户数据访问进行动态身份认证与授权，重要核心数据加密存储并实行访问 IP 白名单限制，定期排查僵尸账号、弱口令及权限滥用风险。

(2) 采用区块链技术，将数据存储系统访问和操作日志上链，实现实时监控审计。

(3) 建立数据容灾备份机制，重要核心数据建立容灾和备份机制，一般数据建立备份机制，定期对容灾备份机制有效性和数据恢复能力进行演练验证。查询记录等中间数据存储前，对敏感信息进行脱敏处理后，并定期清理存储的中间数据。

#### 4.1.4. 数据处理阶段

数据处理阶段涉及数据处理流程环节、数据处理干系人较多，数据敞口大，容易引发数据泄露、数据篡改风险，且溯源难度较大。数据处理过程可采用数据沙箱技术，对数据处理操作行为进行管控和审计[5]。此外，针对提示词注入风险，可采用对抗攻击样本训练模型，增强模型提示词攻击免疫能力。

#### 4.1.5. 模型推理与输出阶段

模型推理与输出阶段，数据安全防护策略：

(1) 用户提出问题请求时，调用大模型语义识别能力实时检测不良信息及恶意意图，对问题请求进行拒绝服务。

(2) 模型推理时，应持续优化大模型系统提示词，聚焦具体业务场景知识领域进行推理生成问题答案，拒绝回答无关内容；模型推理时应基于知识库中内容，并调用可信度评估模型进行二次验证，确保模型生成内容最新且可信。

(3) 答案输出反馈用户前，使用动态脱敏技术对答案中敏感数据进行脱敏处理并返回。

(4) RAG 应用与大模型或智能体之间数据交互接口应采取认证鉴权措施。RAG 系统服务可接入零信任，对访问用户进行身份认证和动态访问控制。

#### 4.1.6. 数据销毁阶段

在 RAG 系统的知识库、大模型迭代或下线时，为避免过期数据被非法使用，需要对不用的模型参数数据、知识库数据进行安全删除，并通过恢复测试验证删除效果有效性。在数据销毁时，结合数据留存期限标签，使用安全擦除工具对数据进行自动销毁。

### 4.2. 模型安全风险管理策略

#### 4.2.1. 防窃取与泄露

(1) 模型架构优化：模型训练推荐使用非线性激活函数，增加网络深度和宽度，提升工程逆向破解难

度；针对不同的应用场景，提供基于模型蒸馏技术的轻量级模型服务。

(2) 存储与访问控制：模型文件静态加密存储，并实施强访问控制；边缘侧设备部署模型需对设备固件进行安全加固(如反调试、反逆向)。

(3) 推理接口保护：限制对模型推理 API 的查询频率和返回信息量；在 API 响应中增加符合差分隐私要求的噪声，保护训练数据隐私。

#### 4.2.2. 完整性与防篡改

(1) 完整性保护：基于数字签名验签技术在模型加载前进行完整性校验，确保模型来源可信。

(2) 环境安全：将模型运行在安全可信的执行环境，严格限制访问和修改权限。

#### 4.2.3. 可用性与鲁棒性

(1) 可用性保障：对模型服务接口异常高频请求、资源消耗性请求、恶意扫描行为进行监测和阻断；支持关键服务响应优先级保障。

(2) 对抗攻击检测：在模型服务接口前端部署对抗攻击检测技术能力；在训练数据中增加对抗样本，增强模型内生对抗攻击免疫能力。

(3) 输入过滤：基于语义识别技术对模型的输入数据进行严格校验和过滤。

#### 4.2.4. 行为风险控制

(1) 生成内容可信：通过可信度评估、知识库交叉验证等方式保证模型生成内容的高可信性。

(2) 公平与可解释：在模型训练目标函数中增加公平性约束指标，提高模型学习公平性表征；在模型中集成决策逻辑可解析技术能力，识别偏见来源。

### 4.3. 安全管理

(1) 建立企业级大模型 RAG 系统数据安全管理制度：包括明确安全管理组织、岗位、人员、职责；明确数据分类分级、数据全生命周期保护、安全评估、风险监测处置、应急响应、安全审计等安全管理要求；明确数据安全各项工作流程。

(2) 数据分类分级管理：利用数据资产扫描、数据流转监测等技术手段，结合专家经验，对全量数据资产进行梳理，形成数据资产清单、数据分类分级清单，并定期更新。

(3) 数据全生命周期保护：根据数据分类分级保护要求，在数据全生命周期过程中落实安全防护策略，形成数据安全防护策略清单；并根据动态安全威胁实时更新安全防护策略，以及定期对安全防护策略有效性进行测试验证。

(4) 安全评估：建立安全评估工作机制，定期开展安全评估与风险整改。评估内容聚焦于个人信息保护、数据防投毒、数据防泄漏、数据防窃取、模型抗指令攻击[6]。

(5) 风险监测处置：对行为、日志、流量数据进行采集和安全综合分析，及时发现潜在的安全风险，并建立 7 * 24 小时安全告警分析研判和闭环处置工作机制，记录风险监测与处置工作台账。

(6) 应急响应：建立应急预案并定期开展应急演练，覆盖数据泄露、数据篡改、内容安全事件等典型场景，形成应急预案报告、应急演练报告。建立安全事件上报工作机制，明确事件上报工作流程和处理时限要求。

(7) 安全审计：对操作日志、和异常行为进行准实时审计，防患大模型和 RAG 系统不被恶意攻击和滥用。

### 4.4. 某医疗机构医疗 AI 助手数据安全实践案例

某医疗机构医疗 AI 助手是一款典型的大模型 RAG 系统应用，基于医疗知识库面向就诊患者提供就

医流程指导、就医挂号、院内就诊路线提示等导诊辅助服务。

(1) 医疗 AI 助手数据安全防护策略

训练数据、知识库数据采集：医疗采集源数据(医疗图像及文字数据)经过脱敏和隐私处理。

数据传输链路：使用 SSL/TLS 数据加密方式建立传输链路。

数据存储：对微调医疗模型数据、医疗知识库数据使用统一集中备份系统进行备份。

数据处理：使用零信任、数据沙箱技术对数据访问实行严格身份认证和访问控制。

模型推理与输出：所有 API 接口调用经过 API 接口网关进行强制安全认证和进出流量控制。建立模型安全围栏，对用户的输入、模型生成返回结果进行安全过滤。

(2) 数据安全防护策略成效

数据安全防护策略的实施，有效保护了患者疾病记录、治疗历史和生物识别等数据隐私，是医疗 AI 助手项目成功落地的关键。

(3) 安全防护策略改进分析

在本案例中实施的数据安全防护策略基本上覆盖数据全生命周期安全，并通过模型安全围栏防范对抗样本攻击，但在应对模型投毒、供应链攻击等方面需加强防护能力。

## 5. 研究局限性与未来展望

### 5.1. 研究局限性

本文构建的企业级大模型 RAG 系统数据安全防护策略注重实用性，同时也存在一些局限性：

(1) 技术层面：本文在数据全生命周期防护、安全管理等方面引入了较多的数据安全技术，当前不同场景的数据安全技术能力集成和协同还不够完善，导致在实际应用中无法发挥技术能力协同最大化优势。此外，对于数据安全技术能力中使用的密码技术，对于量子计算攻击新兴安全威胁，目前防护策略缺乏有效的应对措施。

(2) 应用层面：本文的研究主要基于典型企业级大模型 RAG 系统场景进行数据安全防护策略研究，由于不同行业、不同规模组织实际在建设大模型 RAG 系统时会结合自身情况而存在一定程度细节上差异，因此不同企业组织在落地其大模型 RAG 系统数据安全防护策略需要适当调整，但大的安全防护策略框架依然有效。

(3) 管理层面：受限于不同企业组织资源限制(如人力、物力、财力等资源)，尤其是中小企业组织在数据安全管理制度的执行和监督存在不同程度上的难度与挑战。

### 5.2. 未来展望

未来，随着企业级大模型 RAG 系统价值被深度挖掘以及在各行各业广泛应用，其伴随的数据安全形势将日益严峻和复杂[7]，一方面政策法规提出更高的数据安全合规要求，另一方面量子计算攻击、AI 自动渗透攻击等新型安全威胁为数据安全防护带来巨大安全挑战。

后续数据安全可在如下方面做进一步优化加强：

(1) 加强技术能力创新，一方面加强 AI 能力在数据安全技术方面应用研究，提升数据安全防护与监测处置技术能力；另一方面加强不同场景数据安全技术能力协同工作机制研究，提高整体防护效能。

(2) 优化数据安全管理，探索适合中小企业的高效数据安全管理模式，降低企业的数据安全管理成本。

## 参考文献

[1] 雷蕾. 数据安全现状与发展趋势研究[J]. 信息通信技术与政策, 2022(10): 69-74.

[2] 迟吉凤. 基于数据全生命周期的安全防护体系探究[J]. 网络安全和信息化, 2025, 4(5): 2-3.

[3] 张浩然, 郝文宁, 靳大尉, 程恺, 翟颖. DF-RAG: 基于查询重写和知识选择的检索增强生成方法[J]. 计算机科学, 2025: 1-12. https://link.cnki.net/urlid/50.1075.TP.20250226.1646.002, 2025-07-11.

[4] 刘雪颖, 云静, 李博, 史晓国, 张钰莹. 基于大型语言模型的检索增强生成综述[J]. 计算机工程与应用, 2025, 61(13): 4-19.

[5] 王忠春, 陈庆荣, 刘婷. 大数据下新型安全沙箱技术运用分析与研究[J]. 网络空间安全, 2022, 13(6): 89-97.

[6] 魏永忠, 王诺亚. 数据安全治理的实践维度与行动逻辑[J]. 信息技术与管理应用, 2024, 3(6): 48-57.

[7] 包英明. 大数据平台数据安全防护技术[J]. 信息安全研究, 2019, 5(3): 242-247.