ICS 35.240
CCS L70

团　　体　　标　　准

T/ISC XXX—XXXX

检索增强生成系统安全技术规范

Technical Specification for Security of Retrieval-Augmented Generation Systems

在提交反馈意见时，请将您知道的相关专利与支持性文件一并附上。

（征求意见稿）

2026-03-24

XXXX－XX－XX 发布　　XXXX－XX－XX 实施

中 国 互 联 网 协 会　　发布

目　　次

前　言

本文件按照GB/T 1.1—2020《标准化工作导则　第1部分：标准化文件的结构和起草规则》的规定起草。

请注意本文件的某些内容可能涉及专利。本文件的发布机构不承担识别专利的责任。

本文件由中国互联网协会提出并归口。

本文件起草单位：××××、××××。

本文件主要起草人：×××、×××。

检索增强生成系统安全技术规范

1 范围

本文件规定了检索增强生成系统的安全架构、数据安全、模型算法安全、应用安全及基础设施安全等要求。

本文件适用于检索增强生成系统的设计、开发、部署、运营及安全评估，可为服务提供者、技术支持者及第三方检测机构提供安全技术指引与合规依据。

2 规范性引用文件

下列文件中的内容通过文中的规范性引用而构成本文件必不可少的条款。其中，注日期的引用文件，仅该日期对应的版本适用于本文件；不注日期的引用文件，其最新版本（包括所有的修改单）适用于本文件。

GB/T 22239-2019 信息安全技术 网络安全等级保护基本要求

3 术语和定义

下列术语和定义适用于本文件。

3.1

检索增强生成 retrieval augmented generation

从预构建的知识库或数据源中检索相关信息，在特定任务上增强生成人工智能模型的技术体系。

3.2

向量存储 vector store

一种用于存储和检索高维向量数据的数据库或存储解决方案，适用于处理经过嵌入模型转化后的数据。

3.3

嵌入 embedding

一种将离散对象映射为数值向量的技术。

3.4

提示词 prompt

使用大模型进行微调或下游任务处理时，插入到输入样本中的指令或信息对象。

[来源：GB/T 45288.1—2025，3.5]

3.5

提示词注入 prompt injection

攻击者通过在输入中嵌入恶意指令，试图覆盖或绕过系统预设安全规则的攻击方式。

3.6

幻觉 hallucination

生成式人工智能模型输出与事实不符、缺乏依据或虚构内容的现象

4 缩略语

下列缩略语适用于本文件。

API：应用程序编程接口（Application Programming Interface）

IP：互联网协议（Internet Protocol）

RAG：检索增强生成（Retrieval Augmented Generation）

5 概述

如图1所示，系统安全架构覆盖应用、数据、模型及基础设施四个层面，与系统业务流程深度融合。各安全层面协同联动，形成闭环防护体系，确保检索增强生成系统在开放环境下的安全性。主要包括：

a) 应用层：实施身份认证、访问控制及全链路审计，确保接口调用安全与操作可追溯。
b) 数据层：保障源数据完整性、向量存储保密性及检索过程的权限隔离，防止数据泄露与投毒。
c) 模型与算法层：聚焦提示词工程安全与生成内容合规，防御注入攻击与模型幻觉。
d) 基础设施层：提供网络隔离、环境加固及供应链安全管理，夯实底层运行环境安全。

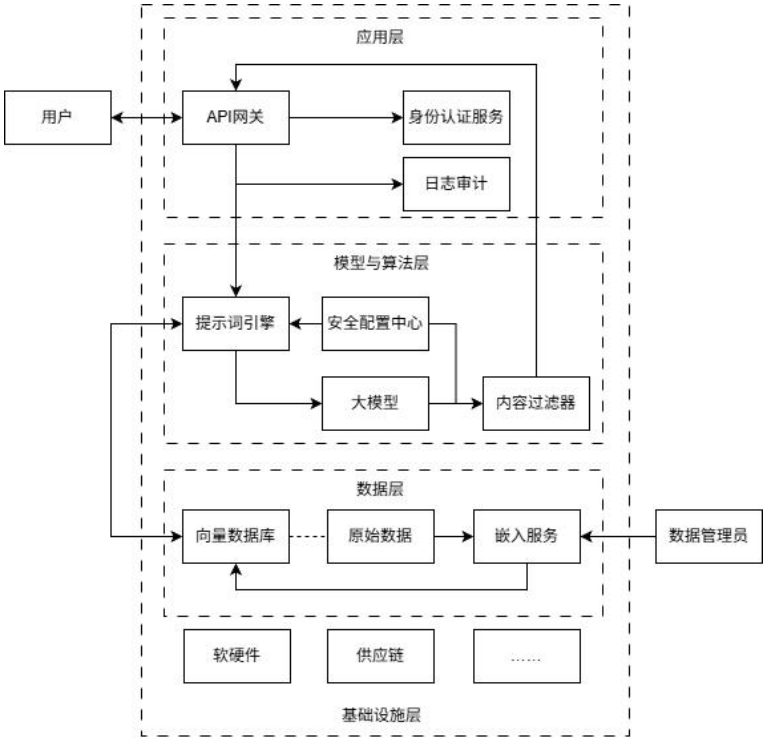

图 1 检索增强生成系统安全架构图

6 应用安全

6.1 身份认证与访问控制

6.1.1 多因素认证

对于管理员账号、知识库维护人员及访问敏感知识的普通用户，系统应强制启用多因素认证，支持结合密码、动态令牌、生物特征或短信验证码等多种认证方式，防止凭证泄露导致的未授权访问。

6.1.2 基于角色的访问控制

系统应实施基于角色的访问控制策略，将用户角色与知识库权限绑定。支持细粒度权限划分，如区分知识读取、写入、审核、管理等权限。检索请求应携带用户角色标识，确保检索结果严格限制在用户授权范围内，防止水平越权与垂直越权。

6.1.3 会话安全管理

系统应对用户会话进行全生命周期管理。会话令牌应随机生成、设置合理有效期，并在用户登出或超时后及时失效。敏感操作应要求重新认证，防止会话劫持风险。

6.2 API 安全

6.2.1 接口速率限制

系统应对API调用实施速率限制，基于用户、IP或应用维度设置请求阈值。对于超出限制的请求，应返回限流响应，防止恶意刷接口导致的资源耗尽或服务拒绝。

6.2.2 异常流量检测

系统应部署流量分析模块，实时监测API调用行为。识别并拦截异常模式，如高频重复查询、批量枚举知识片段、非常规时间访问等。对于疑似攻击流量，应自动触发告警或临时封禁。

6.2.3 输入参数校验

系统应对所有API输入参数进行严格校验。检查参数类型、长度、格式及取值范围，过滤特殊字符与恶意输入，防止注入攻击。

6.3 审计与日志

6.3.1 全链路日志记录

系统应记录用户请求的全链路信息，包括但不限于：用户身份、请求时间、原始查询内容、检索命中片段列表、生成结果摘要、响应状态码等。

6.3.2 日志完整性与防篡改

系统应保障日志数据的完整性，防止日志被恶意删除或篡改。存储权限应严格限制，仅授权审计人员可访问。

6.3.3 敏感操作审计

对于知识库更新、权限变更、模型配置调整等敏感操作，系统应进行专项审计。记录操作人、操作内容、变更前后状态及审批流程。

7 数据安全

7.1 数据接入安全

7.1.1 源数据完整性校验

系统应对接入的源数据进行完整性校验。宜采用哈希算法或数字签名技术，防止数据在传输或预处理过程中被篡改。对于关键业务知识，宜建立可信来源白名单机制。

7.1.2 敏感数据识别与脱敏

系统应部署敏感数据识别模块，在数据嵌入之前，检测个人身份信息、商业机密等敏感内容。对于非必要的敏感信息，应在存入知识库前进行脱敏或掩码处理。

7.1.3 数据分类分级管理

系统宜支持对知识库数据进行分类分级标记，不同级别的数据逻辑隔离存储。

7.1.4 数据清洗与对抗检测

系统宜具备数据清洗能力，去除格式错误、乱码及无关噪声。宜部署恶意样本检测机制，识别并过滤旨在诱导模型错误行为的对抗性文本（如隐藏的指令注入片段、特殊字符攻击）。

7.2 向量存储安全

7.2.1 访问控制

向量数据库应实施访问控制策略。使用基于角色或用户的细粒度权限管理，限制对向量集合的创建、读取、更新和删除操作。禁止未授权的匿名访问，所有访问请求应经过身份验证。

7.2.2 加密存储

向量数据及关联的元数据在静态存储时宜进行加密，采用加密算法对磁盘文件或数据库字段进行加密。

7.2.3 逆向防护

系统宜采用不可逆的嵌入模型，或对向量数据添加噪声干扰，防止攻击者通过向量嵌入值逆向还原原始数据。

7.3 检索过程安全

7.3.1 用户查询过滤

系统宜在检索前对用户输入的内容进行安全检测。识别并拦截包含恶意意图、提示词注入或越狱指令的查询请求。对于高风险查询，宜拒绝执行检索或直接返回安全提示。

7.3.2 检索权限隔离

系统宜实现检索阶段的权限隔离。在进行向量相似度检索时，结合用户身份元数据进行过滤，使得用户仅能检索到其权限范围内的知识片段。

7.3.3 防止检索结果污染

系统宜对检索返回的知识片段进行质量评估。过滤低相似度、低置信度或标记为可疑的片段，防止恶意或低质量内容进入生成上下文。宜设置检索结果的数量上限，避免上下文窗口被恶意填充。

7.4 生命周期安全

7.4.1 版本控制

知识库的更新应支持版本控制。数据变更时应记录版本号、变更时间及操作人。当发现知识库存在污染或错误时，系统应支持快速回滚至之前的安全版本。

7.4.2 备份与恢复

系统应建立定期备份机制，备份内容涵盖向量数据、元数据及索引结构。

8 模型与算法安全

8.1 提示词工程安全

8.1.1 防提示词注入攻击

系统应建立防提示词注入机制。在构建最终提示词时，应采用结构化分隔符等方法严格区分用户查询、检索内容与系统指令。宜部署指令过滤模型，识别并拦截试图覆盖系统指令、诱导模型忽略安全限制的恶意输入。

8.1.2 上下文窗口安全检查

在将检索片段填入上下文窗口前，系统应对拼接后的完整提示词进行安全检查。确保检索内容中不包含隐藏的恶意指令，防止因知识库污染导致的间接提示词注入。

8.1.3 系统提示词保护

系统宜保护核心系统提示词不被泄露，不将系统指令明文返回给前端用户。系统提示词宜存储在安全配置中心，并进行加密管理，仅限授权管理员访问。

8.2 生成内容安全

8.2.1 输出内容合规性过滤

系统应部署输出内容安全过滤模块，对生成结果进行实时检查。对于高风险内容，应实施拦截、替换或添加安全警示，确保生成内容符合法律法规及社会公序良俗。

8.2.2 幻觉检测与事实性校验

系统应具备幻觉检测能力，确保生成内容基于检索到的知识片段。宜要求模型在生成关键事实时提供引用来源，并自动校验引用内容与生成陈述的一致性。对于无法基于检索内容回答的问题，应引导模型承认未知，避免编造事实。

8.2.3 版权与知识产权风险提示

系统宜识别生成内容中可能涉及的版权风险。当生成内容高度相似于特定知识库文档时，宜向用户提示潜在知识产权风险，并建议用户核实来源。

9 基础设施安全

9.1 环境安全

应符合GB/T 22239-2019中相应安全等级的规定。

9.2 供应链安全

9.2.1 开源模型与组件漏洞

系统引入的开源模型及依赖库应进行软件成分分析。扫描已知漏洞与恶意代码，确保组件版本安全。

9.2.2 第三方嵌入模型安全性

若使用第三方提供的嵌入模型服务，应评估其数据隐私政策。确保用户数据在嵌入过程中不被第三方留存、训练或泄露。

9.2.3 模型完整性校验

模型权重文件应进行完整性校验。通过哈希值比对或数字签名验证，防止模型文件在传输或存储过程中被篡改或植入后门。

参　　考　　文　　献

[1] GB/T　22239—2019　信息安全技术　网络安全等级保护基本要求
[2] GB/T　45654—2025　网络安全技术　生成式人工智能服务安全基本要求
[3] GB/T　45288.1—2025　人工智能　大模型　第1部分：通用要求