ICS 35.240
CCS L73

团　　体　　标　　准

T/DGAG XXX—XXXX

政务数据接口安全风险人工智能识别技术规范

Technical specification for Artificial Intelligence recognition of security risks in government data interfaces

（征求意见稿）

XXXX－XX－XX 发布　　XXXX－XX－XX 实施

广东省数字政务协会　　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。

本文件由广东省数字政务协会归口。

本文件起草单位：

本文件主要起草人：

政务数据接口安全风险人工智能识别技术规范

1 范围

本文件规定了政务数据共享服务接口安全风险人工智能识别的总体要求、系统构成、识别流程、识别要求与证实方法。

本文件适用于政务数据共享服务接口安全风险人工智能识别系统的设计、建设、运行与评估，也可供主管（监管）部门实施相关安全监督管理时参考。

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
GB/T 35273—2020 信息安全技术 个人信息安全规范
GB/T 37988—2019 信息安全技术 数据安全能力成熟度模型
GB/T 39477—2020 信息安全技术 政务信息共享 数据安全技术要求
GB/T 43697—2024 数据安全技术 数据分类分级规则
GB/T 45230—2025 数据安全技术 机密计算通用框架
GB/T 45574—2025 数据安全技术 敏感个人信息处理安全要求
GB/T 45652—2025 网络安全技术 生成式人工智能预训练和优化训练数据安全规范
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
GB/T 45674—2025 网络安全技术 生成式人工智能数据标注安全规范
GB 45438—2025 网络安全技术 人工智能生成合成内容标识方法
GB/T 45800.2—2025 全国一体化政务大数据体系 第 2 部分：数据共享交换要求
GB/T 45958—2025 网络安全技术 人工智能计算平台安全框架
GB/T 46796—2025 数据安全技术 数据接口安全风险监测方法

3 术语和定义

下列术语和定义适用于本文件。

3.1

政务数据共享服务接口 government data sharing service interface

通过政务数据共享平台，以应用程序编程接口（API）等形式向其他政务部门或授权主体提供政务数据共享服务的接口。

注：定义基于 GB/T 39477—2020 中政务信息共享数据服务相关概念。

3.2

用户与实体行为分析 User and Entity Behavior Analytics

通过为接口调用相关的用户及实体（账号、设备、接口、网络地址等）建立正常行为基线，并持续比对实际行为与基线的偏离程度，以识别越权调用、凭证盗用、滥用等异常风险的方法。

注：此为方法类能力，可由基线建模、异常打分、同侪分析等技术组合实现，本文件不限定具体算法或产品。其技术内涵与 YD/T 4246—2023《电信网和互联网数据异常行为监测技术要求与测试方法》（数据异常行为监测技术）一致，亦与 GB/T 39477—2020 在政务数据共享场景中列明的“用户实体行为分析技术”相衔接。

3.3

调用频率控制模型 call frequency control model

对政务数据共享服务接口在单位时间内的调用次数设定阈值，并对超过阈值的调用行为实施识别与管控，以防范接口滥用、爆破与资源耗尽风险的方法。

注：简称“频控模型”，可由静态配额、令牌桶、滑动窗口或动态基线限流等方式实现，本文件不限定具体实现方

式。其会话监视与失败控制思想与 GB/T 39477—2020 6.2 访问控制（自动监视和控制远程访问会话、鉴别失败实施安全控制）及 GB/T 22239—2019 等保 访问控制/入侵防范（防爆破）要求一致。

4 缩略语

下列缩略语适用于本文件。

UEBA：用户与实体行为分析（User and Entity Behavior Analytics）

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

SOAR：自动化编排（Security Orchestration, Automation and Response）

GNN：图神经网络（Graph Neural Network）

5 总体要求

人工智能识别政务数据共享服务接口安全风险活动的总体要求如下。

a) 人工智能识别活动应遵守合法、正当、必要原则，涉及的个人信息与敏感个人信息处理应符合 GB/T 35273—2020 与 GB/T 45574—2025 的要求。

b) 人工智能识别所涉政务数据共享交换应符合 GB/T 45800.2—2025 的数据安全要求，并满足 GB/T 39477—2020 规定的数据安全技术要求。

c) 人工智能识别系统应具备全过程可追溯能力，满足 GB/T 22239—2019 安全审计要求与 GB/T 39477—2020 过程追溯、操作抗抵赖要求。

d) 使用生成式人工智能生成风险研判报告或内容的，应符合 GB/T 45654—2025 与 GB 45438—2025 的要求，并对生成内容的真实性与安全性负责。

e) 人工智能识别应坚持人机协同，高风险处置应由责任主体确认复核。

6 系统构成

政务数据共享服务接口安全风险人工智能识别系统由数据采集层、人工智能能力要素层、识别分析层、处置协同层构成，覆盖接口安全风险的监测、识别、研判与处置闭环，具体如下。

a) 数据采集层应汇聚接口访问日志、认证日志、流量报文、接口注册与停用台账等数据，满足 GB/T 46796—2025 监测准备阶段的数据要求。

b) 人工智能能力要素层应为识别分析层提供算力、模型、数据三大人工智能基础能力要素支撑（基础能力的具体要求见附录 A），其要素安全与管理要求满足 GB/T 45652—2025、GB/T 45654—2025、GB/T 45674—2025 的规定，并可参照 TC260《人工智能安全治理框架》关于算力、模型、数据要素的治理原则。

c) 识别分析层应采用用户与实体行为分析、流量深度检测、GNN 关联推理等技术进行风险识别（典型安全风险类型与识别能力核查表见附录 B），并满足第 8 章各类风险的技术要求。

d) 处置协同层应支持告警分级与 SOAR，联动身份与访问管理、数据脱敏、安全防护等设备实施处置，满足第 8.8 节分级处置要求。

7 识别流程

流程阶段划分要求参照 GB/T 46796—2025 规定的数据接口安全风险监测闭环进行划分。

a) 监测准备阶段：应完成接口资产与数据资产盘点，建立政务数据共享服务接口风险识别所需的数据底座，汇聚接口访问日志、认证日志、流量报文、接口注册与停用台账等数据，并应输出接口资产清单、数据分类分级结果（见 8.1）与监测数据底座说明（对应模板见附录 C.1、C.2）。

b) 风险识别阶段：应通过识别分析层并行开展政务数据共享服务接口风险的识别分析，并应输出识别结果清单，含风险项、风险等级与识别证据（对应模板见附录 C.3）。

c) 分析研判阶段：应依据风险等级对识别结果实施分级研判，按第 8.8 节将处置划分为自动处置、人工研判与应急响应三个层级，并应输出研判结论与分级处置建议（对应模板见附录 C.4）。

d) 预警处置阶段：应依据研判结论对风险实施分级处置与联动，通过处置协同层实施阻断、限流、脱敏、整改等动作，并应输出处置工单与联动处置记录（对应模板见附录 C.5）。

e) 审计追溯阶段 ：应对风险的识别、研判、处置全过程实施统一审计，满足第 8.7 节审计与可追溯要求，并应输出审计记录与可追溯存证（对应模板见附录 C.6）；同时应将处置与审计结果用于优化识别基线、阈值与策略，支撑流程闭环反馈与风险识别能力的持续提升。

政务数据共享服务接口安全风险识别流程图见图1。

识别对象 政务数据共享平台 API 接口	监测准备	风险识别	分析研判	预警处置	审计追溯
技术要求	建立接口资产与数据底座； 完成接口注册、分类分级盘点； 汇聚接口日志、流量报文 与注册停用台账	采用用户实体行为分析UEBA、 流量深度检测、图神经网络GNN 关联推理并行识别； 输出风险项、等级与识别证据	对识别结果分级研判； 结合业务语义判定行为偏离、 滥用、关联重识别与审计缺失	依风险等级执行自动 / 人工 / 应急三级处置； SOAR 联动实施阻断、限流、 脱敏、整改	对识别—研判—处置全过程 统一审计留痕； 落实监测数据加密、脱敏、备份 与留存
风险类	数据合规 鉴权与访问控制	六类接口安全风险： 鉴权与访问控制 传输与数据暴露 注入与攻击、滥用与失控 汇聚关联与重识别 审计与可追溯	全类风险（依等级研判） L1 / L2 / L3 分级	全类风险（处置联动） 人负其责	审计与可追溯 全过程留痕与存证
证实方法	审查接口资产台账与 数据分类分级记录	复核六类风险识别记录 （UEBA / 流量检测 / GNN）	复核研判结论与 分级处置决策留痕	复核处置工单与 联动处置记录	校验审计留存期 与记录完整性
输出成果	接口资产清单 监测数据底座说明	六类风险识别结果清单	研判结论与 分级处置建议	处置工单与 联动处置记录	审计记录与 可追溯存证

图 1　政务数据共享服务接口安全风险识别流程图

8　技术要求

8.1　数据合规要求

人工智能识别政务数据共享接口风险所涉数据处理活动，全生命周期应满足的数据合规要求，具体如下。

a) 应依据 GB/T 39477—2020 第 6 章、GB/T 37988—2019 与 GB/T 43697—2024 对政务数据共享接口涉及的数据实施分类分级，并建立分类分级结果与接口安全风险的映射关系。

b) 应采用 GB/T 35273—2020 与 GB/T 45574—2025 规定的个人信息、敏感个人信息处理安全要求，对接口传输与识别过程中的个人信息实施去标识化与最小必要控制。

c) 工智能识别所使用的数据集不应包含未获授权的政务数据链路数据、重要数据及违规收集的个人信息；训练与优化训练数据应符合 GB/T 45652、GB/T 45674 的安全要求。

d) 接口监测与风险研判过程应符合“原始数据不出域、数据可用不可见”的要求，对需跨域关联的数据应采用隐私计算技术（如机密计算、安全多方计算、联邦学习等）实施保护，满足 GB/T 45230—2025 规定的机密计算通用框架要求。

e) 采用生成式人工智能生成风险研判报告的，应符合 GB/T 45654—2025 与 GB 45438—2025 的内容标识与安全措施要求，并对生成内容的真实性与安全性负责。

8.2　鉴权与访问控制风险识别要求

对于鉴权与访问控制风险（如未鉴权、弱鉴别、越权调用等接口访问控制风险）的识别要求，具体如下。

a) 应识别接口未实施身份鉴别、弱口令、凭证泄露、越权调用等鉴权与访问控制风险，范围覆盖 GB/T 39477—2020 6.2 规定的身份鉴别与访问控制项。

b) 应采用用户实体行为分析对用户调用接口的频次、时段、主体与权限偏离进行异常识别，并结合访问策略基线判定越权风险。

c) 风险识别应以接口访问日志、认证日志、授权策略与资产台账为输入，提取主体—接口—权限—时段特征，特征应满足可追溯与可复现要求。

d) 应依据风险等级对未鉴权、越权调用等判定结果实施分级告警，高风险调用应在 GB/T 39477—2020 规定的访问控制要求内即时阻断。
e) 对识别出的鉴权风险，应联动身份与访问管理系统复核并处置，处置结果应留痕，并纳入 8.7 的审计与可追溯要求统一管理。

8.3 传输与数据暴露风险识别要求

对于传输与数据暴露风险（如明文传输、过度暴露、脱敏失效等接口数据暴露）的识别要求，具体如下。

a) 应识别接口明文传输、未加密通道、过度暴露字段、脱敏失效等传输与数据暴露风险，对象包括 GB/T 39477—2020 6.2 规定的安全传输与数据脱敏项。
b) 应采用流量深度检测与敏感字段识别模型，对接口请求响应中的明文敏感信息与违规暴露字段进行自动识别与标注。
c) 风险识别应以接口流量、报文样本、脱敏策略与字段字典为输入，构建敏感数据特征指纹，识别应满足 GB/T 35273—2020 的最小必要与目的限制。
d) 应对未加密传输、脱敏失效等判定为高风险，触发告警并按 GB/T 45800.2—2025 建议的隐私保护技术推动数据不出域整改。
e) 对数据暴露风险，应联动数据脱敏与网关处置并复核，处置过程与结果应满足 8.7 审计可追溯要求。

8.4 注入与攻击风险识别要求

对于注入与攻击风险（如注入攻击、恶意载荷、异常请求、参数篡改等接口攻击风险）的识别要求，具体如下。

a) 应识别 GB/T 46796—2025《数据安全技术 数据接口安全风险监测方法》所列注入、未加密传输等接口风险项。
b) 应采用异常流量检测与攻击特征模型，对 SQL/命令注入、越权遍历与异常报文进行实时识别，并结合威胁情报研判攻击意图。
c) 风险识别应以接口请求报文、攻击特征库、流量基线与威胁情报为输入，提取载荷、参数与路径异常特征，特征应满足可解释与可验证。
d) 对注入与攻击判定为高风险，应即时告警并联动防护设备阻断，响应应符合 GB/T 22239—2019 安全计算环境相关要求。
e) 对攻击事件，应联动安全运营处置并复核，事件证据应满足 8.7 审计可追溯与操作抗抵赖要求。

8.5 滥用与失控风险识别要求

对于滥用与失控风险（如调用滥用、超阈值调用、接口超期开放、违规公开等接口失控风险）的识别要求，具体如下。

a) 应识别接口调用滥用、超阈值调用、接口超期开放、违规公开等滥用与失控风险，对象包括 GB/T 39477—2020 6.2 规定的级联接口安全与操作抗抵赖项。
b) 应采用 UEBA 与频控模型对调用量、调用方、接口生命周期进行异常识别，对超出基线与审批范围的调用判定为滥用。
c) 风险识别应以调用量统计、接口注册与停用台账、审批记录为输入，提取频率、规模与生命周期异常特征，特征应满足可计量与可比对。
d) 对超阈值调用、超期开放应判定为高风险并告警，触发限流、停用等处置，处置应符合 GB/T 39477—2020 过程追溯要求。
e) 对滥用事件，应联动接口治理与资产管理复核处置，结果应满足 8.7 审计可追溯要求。

8.6 汇聚关联与重识别风险识别要求

对于汇聚关联与重识别风险（如多接口数据汇聚、关联分析、去标识化失效导致重识别与失泄密风险）的识别要求，具体如下。

a) 应识别多接口数据汇聚、关联分析、去标识化失效导致的重识别与失泄密风险，对象包括 GB/T 37988—2019 数据分类分级与 GB/T 35273—2020 去标识化要求。
b) 应采用 GNN 关联推理与安全多方计算，对跨接口、跨主体的数据关联与重识别风险进行识别。
c) 风险识别应以多源接口数据目录、数据血缘、去标识化记录为输入，提取实体关联与再识别路径特征，特征应满足 GB/T 45230—2025 不出域约束。
d) 对可重识别、可汇聚失泄密判定为高风险并告警，推动按 GB/T 45230—2025 机密计算实施可用不可见整改。
e) 对汇聚关联风险，应联动数据分级与脱敏复核处置，结果应满足 8.7 审计可追溯要求。

8.7 审计与可追溯风险识别要求

对于风险识别、判定、处置全过程的审计与可追溯要求，具体如下。

a) 应对前述六类风险的识别、判定、处置全过程实施审计，对象覆盖 GB/T 39477—2020 6.2 规定的过程追溯与操作抗抵赖项。
b) 应采用日志聚合与关联分析技术，对识别与处置行为的主体、时间、结果进行统一记录与防篡改存证。
c) 7 审计数据应以操作日志、处置记录、告警记录为输入，提取主体—行为—对象—时间特征，留存应满足 GB/T 22239—2019 安全审计要求。
d) 对审计发现的遗漏识别、违规处置应判定为高风险并告警，推动复核整改，整改应符合过程追溯要求。
e) 审计记录应可检索、可导出、可举证，并满足 GB/T 39477—2020 操作抗抵赖与 8.8 责任分级处置闭环。

8.8 责任与分级处置要求

对于风险的分级处置机制与责任界定要求，具体如下。

a) 应参照 TC260《人工智能安全治理框架》建立人工智能识别风险的分级处置机制，按照风险等级将处置划分为自动处置、人工研判与应急响应三个层级，并参照《生成式人工智能服务管理暂行办法》明确各层级的责任主体与处置要求。
b) 应采用 SOAR 与告警分级技术，对风险识别结果按等级路由至相应处置角色与流程。
c) 处置应以风险等级、识别证据、处置预案为输入，提取等级—角色—动作特征，分级应满足 GB/T 45652 生成式人工智能服务安全基本要求。
d) 对高风险事件应触发应急响应与上报，响应应符合 GB/T 45654 生成式人工智能服务安全应急响应和处置要求。
e) 处置全过程应由责任主体复核确认并留痕，责任界定应满足 8.7 审计可追溯与操作抗抵赖要求。

9 证实方法

9.1 数据合规的证实方法

对于各项数据合规要求的证实方法，具体如下。

a) 通过审查政务数据共享接口的数据分类分级结果、分类分级与接口安全风险的映射关系文档，以及依据 GB/T 39477—2020 第 6 章、GB/T 37988—2019 与 GB/T 43697—2024 的判定记录，证实 8.1 中 a）的要求已满足。
b) 通过审查个人信息与敏感个人信息处理安全控制措施说明、去标识化与最小必要控制记录，确认其符合 GB/T 35273—2020 与 GB/T 45574—2025 的规定，证实 8.1 中 b）的要求已满足。
c) 通过审查人工智能识别所用数据集的来源授权文件、数据合规审计记录与训练优化训练数据安全性评估材料，确认不含未获授权的政务数据链路数据、重要数据及违规收集的个人信息，且符合 GB/T 45652、GB/T 45674，证实 8.1 中 c）的要求已满足。

d) 通过检查接口监测与风险研判系统的部署架构、跨域关联场景中隐私计算技术的应用记录（依据 GB/T 45230—2025 机密计算通用框架），确认过程符合“原始数据不出域、数据可用不可见”原则，证实 8.1 中 d）的要求已满足。
e) 通过审查生成式人工智能风险研判报告的内容标识记录、应急响应与安全性评估材料，确认符合 GB/T 45654—2025 与 GB 45438—2025 内容标识与安全措施要求的规定，证实 8.1 中 e）的要求已满足。

9.2 鉴权与访问控制风险识别的证实方法

对于各项鉴权与访问控制风险识别要求的证实方法，具体如下。

a) 通过检查接口未鉴权、弱口令、凭证泄露、越权调用的识别结果清单，确认识别范围覆盖 GB/T 39477—2020 6.2 规定的身份鉴别与访问控制项，证实 8.2 中 a）的要求已满足。
b) 通过审查 UEBA 行为基线模型配置与异常识别输出样例，确认对用户调用频次、时段、主体与权限偏离进行了异常识别并判定越权风险，证实 8.2 中 b）的要求已满足。
c) 通过检查风险识别输入数据清单（访问日志、认证日志、授权策略、资产台账）及主体—接口—权限—时段特征提取说明，确认所提取特征具备可追溯与可复现条件，证实 8.2 中 c）的要求已满足。
d) 通过检查分级告警规则配置与高风险调用阻断日志，确认对未鉴权、越权调用实施了分级告警并在 GB/T 39477—2020 规定的访问控制要求内即时阻断，证实 8.2 中 d）的要求已满足。
e) 通过检查鉴权风险处置工单、联动身份与访问管理系统的复核记录及留痕，确认处置结果纳入 8.7 审计与可追溯要求统一管理，证实 8.2 中 e）的要求已满足。

9.3 传输与数据暴露风险识别的证实方法

对于各项传输与数据暴露风险识别要求的证实方法，具体如下。

a) 通过检查明文传输、未加密通道、过度暴露字段、脱敏失效的识别结果，确认识别对象覆盖 GB/T 39477—2020 6.2 规定的安全传输与数据脱敏项，证实 8.3 中 a）的要求已满足。
b) 通过审查流量深度检测与敏感字段识别模型的识别准确率测试报告、误报与漏报样例，确认对接口请求响应中的明文敏感信息与违规暴露字段具备自动识别与标注能力，证实 8.3 中 b）的要求已满足。
c) 通过检查接口流量、报文样本、脱敏策略、字段字典等输入及敏感数据特征指纹构建说明，确认特征满足 GB/T 35273—2020 的最小必要与目的限制，证实 8.3 中 c）的要求已满足。
d) 通过检查未加密传输、脱敏失效高风险告警记录与数据不出域整改工单，确认触发了告警并按 GB/T 45230—2025 机密计算框架推动整改，证实 8.3 中 d）的要求已满足。
e) 通过检查数据脱敏与网关处置复核记录，确认处置过程与结果满足 8.7 审计可追溯要求，证实 8.3 中 e）的要求已满足。

9.4 注入与攻击风险识别的证实方法

对于各项注入与攻击风险识别要求的证实方法，具体如下。

a) 通过检查注入攻击、恶意载荷、异常请求、参数篡改的识别结果，确认覆盖 GB/T 46796—2025 所列接口风险项，证实 8.4 中 a）的要求已满足。
b) 通过审查异常流量检测与攻击特征模型对 SQL/命令注入、越权遍历、异常报文的实时识别记录及威胁情报研判材料，确认识别能力，证实 8.4 中 b）的要求已满足。
c) 通过检查接口请求报文、攻击特征库、流量基线、威胁情报等输入及载荷、参数、路径异常特征提取说明，确认特征满足可解释与可验证，证实 8.4 中 c）的要求已满足。
d) 通过检查注入与攻击高风险告警与防护设备阻断联动日志，确认响应符合 GB/T 22239—2019 安全计算环境相关要求，证实 8.4 中 d）的要求已满足。
e) 通过检查攻击事件联动安全运营处置记录与操作抗抵赖存证，确认满足 8.7 审计可追溯与操作抗抵赖要求，证实 8.4 中 e）的要求已满足。

9.5 滥用与失控风险识别的证实方法

对于各项滥用与失控风险识别要求的证实方法，具体如下。

a) 通过检查调用滥用、超阈值调用、接口超期开放、违规公开的识别结果，确认对象覆盖 GB/T 39477—2020 6.2 规定的级联接口安全与操作抗抵赖项，证实 8.5 中 a）的要求已满足。
b) 通过审查 UEBA 与频控模型对调用量、调用方、接口生命周期的异常识别输出，确认对超出基线与审批范围的调用判定为滥用，证实 8.5 中 b）的要求已满足。
c) 通过检查调用量统计、接口注册与停用台账、审批记录等输入及频率、规模、生命周期异常特征提取说明，确认特征满足可计量与可比对，证实 8.5 中 c）的要求已满足。
d) 通过检查超阈值调用、超期开放高风险告警与限流、停用处置记录，确认处置符合 GB/T 39477—2020 过程追溯要求，证实 8.5 中 d）的要求已满足。
e) 通过检查接口治理与资产管理复核处置记录，确认满足 8.7 审计可追溯要求，证实 8.5 中 e）的要求已满足。

9.6 汇聚关联与重识别风险识别的证实方法

对于各项汇聚关联与重识别风险识别要求的证实方法，具体如下。

a) 通过检查多接口数据汇聚、关联分析、去标识化失效导致重识别与失泄密的识别结果，确认对象覆盖 GB/T 37988—2019 数据分类分级与 GB/T 35273—2020 去标识化要求，证实 8.6 中 a）的要求已满足。
b) 通过审查 GNN 关联推理与安全多方计算对跨接口、跨主体数据关联与重识别的识别结果样例，确认识别能力，证实 8.6 中 b）的要求已满足。
c) 通过检查多源接口数据目录、数据血缘、去标识化记录等输入及实体关联与再识别路径特征提取说明，确认满足 GB/T 45230—2025 不出域约束，证实 8.6 中 c）的要求已满足。
d) 通过检查可重识别、可汇聚失泄密高风险告警与按 GB/T 45230—2025 机密计算实施可用不可见整改记录，确认整改落实，证实 8.6 中 d）的要求已满足。
e) 通过检查数据分级与脱敏复核处置记录，确认满足 8.7 审计可追溯要求，证实 8.6 中 e）的要求已满足。

9.7 审计与可追溯风险识别的证实方法

对于各项审计与可追溯要求的证实方法，具体如下。

a) 通过审查风险识别、判定、处置全过程的审计记录，确认覆盖 GB/T 39477—2020 6.2 规定的过程追溯与操作抗抵赖项，证实 8.7 中 a）的要求已满足。
b) 通过检查日志聚合与关联分析系统的统一记录、防篡改存证配置与样例，确认对识别与处置行为的主体、时间、结果实施了统一记录与防篡改存证，证实 8.7 中 b）的要求已满足。
c) 通过检查操作日志、处置记录、告警记录等输入及安全审计留存配置，确认满足 GB/T 22239—2019 安全审计要求，证实 8.7 中 c）的要求已满足。
d) 通过检查审计发现的遗漏识别、违规处置高风险告警与复核整改记录，确认符合过程追溯要求，证实 8.7 中 d）的要求已满足。
e) 通过检查审计记录的检索、导出、举证功能验证结果，确认满足 GB/T 39477—2020 操作抗抵赖与 8.8 责任分级处置闭环，证实 8.7 中 e）的要求已满足。

9.8 责任与分级处置的证实方法

对于各项责任与分级处置要求的证实方法，具体如下。

a) 通过审查风险分级处置机制文档、自动处置、人工研判与应急响应三个层级的定义及各层级责任主体与处置要求清单，确认机制已建立且责任主体明确，证实 8.8 中 a）的要求已满足。
b) 通过检查 SOAR 与告警分级技术的路由配置及风险识别结果按等级路由至相应处置角色与流程的运行记录，确认分级路由能力，证实 8.8 中 b）的要求已满足。
c) 通过检查风险等级、识别证据、处置预案等输入及等级—角色—动作特征提取说明，确认分级满足 GB/T 45652 生成式人工智能服务安全基本要求，证实 8.8 中 c）的要求已满足。
d) 通过检查高风险事件应急响应与上报记录，确认符合 GB/T 45654 生成式人工智能服务安全应急响应和处置要求，证实 8.8 中 d）的要求已满足。
e) 8.8.5 通过检查处置全过程责任主体复核确认与留痕记录，确认满足 8.7 审计可追溯与操作抗抵赖要求，证实 8.8 中 e）的要求已满足。

附 录 A
（规范性）
人工智能能力要素层详细要求

A.1 概述

本附录对第 5.3 条规定的算力、模型、数据三大人工智能能力要素作出详细描述与实施指引，供人工智能识别系统的设计、建设与评估参考。三要素的安全与管理要求应满足 GB/T 45652—2025、GB/T 45654—2025、GB/T 45674—2025，并可参照 TC260《人工智能安全治理框架》2.0 关于算力、模型、数据要素的治理原则。

A.2 数据要素详细要求

A.2.1 数据来源合规

训练、优化训练与研判辅助数据应建立来源登记与授权核验机制，明确数据提供方、用途、范围与授权期限。

不应使用未授权政务链路数据、未经安全评估的重要数据，以及违规收集的个人信息。

涉及个人信息与敏感个人信息的，应满足最小必要与目的限制（见 GB/T 35273—2020、GB/T 45574—2025）。

数据合规要求与第 8.1的c）条对应，由第 9.1的c）条证实。

A.2.2 数据内容安全与标注

训练与优化训练数据标注应满足 GB/T 45674—2025 的数据标注安全要求和信息安全要求。

应建立数据清洗、去重、去标识化与质量校验流程，避免污染数据与偏见数据进入模型训练。

A.2.3 数据全生命周期管理

数据的采集、存储、使用、传输、共享、销毁应满足 GB/T 39477—2020 与 GB/T 43697—2024 的数据全生命周期安全要求。

监测与处置过程中使用的敏感字段应优先脱敏后留存，明文数据按 8.1的e）不长期保存。

A.3 模型要素详细要求

A.3.1 模型选型与备案

应选用或训练符合安全要求的识别模型；使用基础模型的，应使用已备案的基础模型，并核验其备案信息与安全评估报告。

模型安全应满足 GB/T 45654—2025 第 6 章模型安全要求。

A.3.2 模型准确性与可控性

应采取技术措施保障模型生成与研判结论的准确性、可靠性与安全可控，建立模型输出人工复核与纠偏机制。

涉及生成式人工智能研判内容的，其生成合成内容标识应满足 GB 45438—2025 与第 8.1的e）的要求，由第 9.1的e）证实。

A.3.3 模型持续运营

应建立模型版本管理、效果评估与迭代优化机制，对研判准确率、误报率、漏报率进行持续监测。

A.4 算力要素详细要求

A.4.1 算力供应链安全

应评估训练与推理所采用计算系统的供应链安全，覆盖芯片、软件、工具、算力等方面的供应持续性与稳定性，并应满足 GB/T 45958—2025 人工智能计算平台安全框架规定的安全功能、安全管理与角

色安全职责要求，以及 GB/T 45654—2025 第 7 章安全措施要求中的训练与推理环境隔离，并宜结合 GB/T 45230—2025 机密计算框架保护推理过程数据不出域。

宜建立关键算力组件的备用与应急方案，防范单点供应中断风险。

A.4.2 可信计算环境

计算环境宜支持基于硬件的安全启动与可信启动，保障人工智能识别系统在安全可信环境中运行。

宜在可行范围内采用机密计算等技术（见 GB/T 45230—2025）保护推理过程数据不出域。

A.5 三要素协同

算力、模型、数据三要素应协同支撑识别分析层（见 6.4）与处置协同层（见 6.5），形成“数据可管、模型可信、算力可控”的能力底座。

附 录 B
（资料性）
典型安全风险类型与识别能力核查表

B.1 典型安全风险类型见表 B.1。

表 B.1 典型安全风险类型

编号	风险类	定义	典型风险点	标准出处
R1	鉴权与访问控制	因身份鉴别机制缺失或薄弱、访问权限管控不当，导致接口被未授权主体调用、合法主体超范围调用或越权访问的安全风险。核心关注“谁来调、能否调、调多大范围”。	数据接口未鉴权；鉴别方式简单（仅静态口令、无多因素）；越权访问（水平 / 垂直越权）；合法账号超范围调用、用途漂移。	GB/T 39477—2020《信息安全技术 政务信息共享 数据安全技术要求》第 6.2 条“身份鉴别”“访问控制”——明确政务信息共享接口应实施身份鉴别与访问控制。 GB/T 46796—2025《数据安全技术 数据接口安全风险监测方法》所列接口风险项（未鉴权、弱鉴别、越权）——接口风险直接清单。 GB/T 22239—2019《信息安全技术 网络安全等级保护基本要求》“安全计算环境—身份鉴别 / 访问控制”。 《政务数据共享条例》（国务院令 809 号）第 25、27 条——合法账号超范围调用、越权访问的管理侧约束。
R2	传输与数据暴露	因传输链路未加密、协议/算法不安全，或接口返回数据未按最小必要原则脱敏，导致敏感字段在传输或响应中过度暴露、可被截获或还原的安全风险。核心关注“传输是否加密、返回是否过度暴露”。	未加密传输（明文）；采用不安全协议 / 算法（中间人截获）；返回数据类型过度暴露；脱敏后响应仍含未脱敏字段；单次返回大量敏感信息。	GB/T 39477—2020 第 6.2 条“安全传输”“数据脱敏”“权限标记”。 GB/T 46796—2025 所列未加密传输、过度暴露等接口风险项。 GB/T 35273—2020《信息安全技术 个人信息安全规范》“最小必要”“目的限制”。 GB/T 45574—2025《数据安全技术 敏感个人信息处理安全要求》——敏感字段脱敏与暴露控制。 《政务数据共享条例》第 37 条（PII 保护）——脱敏失效的管理侧约束。
R3	注入与攻击	因接口未对输入参数、载荷、路径实施充分校验与防护，导致被注入恶意代码/命令、接口被不安全封装/非法封装利用，引发数据篡改、越权获取或系统入侵的安全风险。核心关注“能否被注入 / 非法封装利用”。	SQL 注入；代码 / 命令注入； 不安全接口封装被利用；数据共享接口非法封装。	GB/T 46796—2025《数据安全技术 数据接口安全风险监测方法》所列接口风险项（注入、未加密传输等）。 GB/T 39477—2020 第 6.2 条“级联接口安全”——数据共享接口非法封装的防护要求。 GB/T 22239—2019“安全计算环境—入侵防范”——注入类攻击的等保防护基线。 《政务数据共享条例》第 27 条——数据共享接口非法封装的管理侧约束。
R4	滥用与失控	因接口调用缺乏频控/行为监测与生命周期管控，导致接口被异常调用、机器爬虫抓取、超期开放、违规公开或转供第三方，超出授权范围失控使用的安全风险。核心关注“是否被滥用、超期 / 违规开放、转供”。	异常调用 / 滥用、机器爬虫抓取；超期开放接口； 违规公开接口；用途漂移、转供第三方。	GB/T 39477—2020 第 6.2 条“级联接口安全”“操作抗抵赖”——接口使用管控与过程追溯。 GB/T 46796—2025 所列“异常调用”等接口风险项。 GB/T 22239—2019“访问控制 / 入侵防范—防爆破”——频控概念衔接基线。 《政务数据共享条例》第 25 条——用途漂移、转供第三方约束。
R5	汇聚关联与重识别	因多个接口返回的数据可被关联汇聚，或去标识化措施失效，导致通过多源拼接、字段组合推断出未授权的个人或实体信息，造成重识别与失泄密的安全风险。	多接口数据汇聚关联推断未授权信息； 多字段重识别个人（去标识化失效）。	GB/T 37988—2019《信息安全技术 数据安全能力成熟度模型》——数据分类分级与关联管控成熟度要求。 GB/T 35273—2020“去标识化”要求——重识别风险控制。 GB/T 43697—2024《数据安全技术 数据分类分级规则》——数据分级与关联分析基础。

编号	风险类	定义	典型风险点	标准出处
		核心关注“多源拼接能否推断未授权信息”。		GB/T 45230—2025《数据安全技术 机密计算通用框架》——跨域关联“数据不出域”技术约束。 GB/T 45800.2—2025《全国一体化政务大数据体系 第 2 部分：数据共享交换要求》——政务数据共享交换中的关联与汇聚安全。 《政务数据共享条例》第 25、37 条——汇聚关联、重识别个人的管理侧约束。
R6	审计与可追溯	因接口调用未全量记录或记录不可防篡改、不可关联，导致安全事件无法溯源、责任无法界定，或迟报/漏报的安全风险。核心关注“调用能否全量留痕、可溯源定责”。	接口调用无日志、操作不可抵赖； 安全事件迟报/漏报。	GB/T 39477—2020 第 6.2 条“过程追溯”“操作抗抵赖”——R6 核心国标依据。 GB/T 22239—2019“安全审计”——审计数据留存与防篡改基线。 GB/T 46796—2025 数据接口安全风险监测闭环——监测与审计闭环要求。 《政务数据共享条例》第 29、35 条——调用留痕、安全事件报告的管理侧约束。

B.2　识别能力核查表见表 B.2。

表 B.2 识别能力核查表

编号	风险类	对应条款	风险项	识别要求条款	证实方法条款	相关标准编号
1	数据合规	8.1	数据分类分级及与接口安全风险映射	8.1 a)	9.1 a)	GB/T 39477—2020、GB/T 37988—2019、GB/T 43697—2024
2	数据合规	8.1	个人信息/敏感个人信息去标识化与最小必要控制	8.1 b)	9.1 b)	GB/T 35273—2020、GB/T 45574—2025
3	数据合规	8.1	数据集不含未授权政务链路数据、重要数据及违规收集个人信息	8.1 c)	9.1 c)	GB/T 45652、GB/T 45674
4	数据合规	8.1	原始数据不出域、数据可用不可见（机密计算保护）	8.1 d)	9.1 d)	GB/T 45230—2025
5	数据合规	8.1	生成式人工智能研判内容标识与应急响应	8.1 e)	9.1 e)	GB/T 45654、GB 45438—2025
6	R1	8.2	接口未鉴权、弱口令、凭证泄露、越权调用识别	8.2 a)	9.2 a)	GB/T 39477—2020
7	R1	8.2	UEBA 行为基线偏离异常识别与越权判定	8.2.b)	9.2.b)	GB/T 39477—2020
8	R1	8.2	主体—接口—权限—时段特征提取（可追溯、可复现）	8.2.c)	9.2.c)	GB/T 39477—2020
9	R1	8.2	分级告警与高风险调用即时阻断	8.2.d)	9.2.d)	GB/T 39477—2020
10	R1	8.2	联动身份与访问管理系统复核处置并留痕	8.2.e)	9.2.e)	GB/T 39477—2020、GB/T 22239—2019
11	R2	8.3	明文传输、未加密通道、过度暴露字段、脱敏失效识别	8.3 a)	9.3 a)	GB/T 39477—2020
12	R2	8.3	流量深度检测与敏感字段自动识别标注	8.3.b)	9.3.b)	GB/T 39477—2020
13	R2	8.3	敏感数据特征指纹构建（最小必要、目的限制）	8.3.c)	9.3.c)	GB/T 35273—2020
14	R2	8.3	高风险整改触发并推动数据不出域	8.3.d)	9.3.d)	GB/T 45230—2025

15	R2	8.3	联动数据脱敏与网关处置复核	8.3.e)	9.3.e)	GB/T 39477—2020
16	R3	8.4	注入攻击、恶意载荷、异常请求、参数篡改识别	8.4 a)	9.4 a)	GB/T 46796—2025
17	R3	8.4	异常流量检测与威胁情报研判攻击意图	8.4.b)	9.4.b)	GB/T 46796—2025
18	R3	8.4	载荷、参数、路径异常特征提取（可解释、可验证）	8.4.c)	9.4.c)	GB/T 46796—2025
19	R3	8.4	高风险告警与防护设备阻断联动	8.4.d)	9.4.d)	GB/T 22239—2019
20	R3	8.4	攻击事件联动安全运营处置与操作抗抵赖存证	8.4.e)	9.4.e)	GB/T 39477—2020、GB/T 22239—2019
21	R4	8.5	调用滥用、超阈值调用、接口超期开放、违规公开识别	8.5 a)	9.5 a)	GB/T 39477—2020
22	R4	8.5	UEBA 与频控模型异常调用识别并判定滥用	8.5 b)	9.5 b)	GB/T 39477—2020
23	R4	8.5	频率、规模、生命周期异常特征提取（可计量、可比对）	8.5 c)	9.5 c)	GB/T 39477—2020
24	R4	8.5	限流、停用等高风险处置	8.5 d)	9.5 d)	GB/T 39477—2020
25	R4	8.5	联动接口治理与资产管理复核处置	8.5 e)	9.5 e)	GB/T 39477—2020
26	R5	8.6	多接口数据汇聚、关联、去标识化失效导致重识别与失泄密识别	8.6 a)	9.6 a)	GB/T 37988—2019、GB/T 35273—2020
27	R5	8.6	GNN 关联推理与安全多方计算重识别	8.6 b)	9.6 b)	GB/T 45230—2025
28	R5	8.6	实体关联与再识别路径特征提取（不出域约束）	8.6 c)	9.6 c)	GB/T 45230—2025
29	R5	8.6	高风险整改推动可用不可见	8.6 d)	9.6 d)	GB/T 45230—2025
30	R5	8.6	联动数据分级与脱敏复核处置	8.6 e)	9.6 e)	GB/T 45230—2025
31	R6	8.7	六类风险识别、判定、处置全过程审计	8.7 a)	9.7 a)	GB/T 39477—2020
32	R6	8.7	日志聚合与关联分析统一记录、防篡改存证	8.7 b)	9.7 b)	GB/T 39477—2020
33	R6	8.7	审计数据留存满足等保安全要求	8.7 c)	9.7 c)	GB/T 22239—2019
34	R6	8.7	遗漏识别、违规处置高风险告警与复核整改	8.7 d)	9.7 d)	GB/T 39477—2020
35	R6	8.7	审计记录可检索、可导出、可举证	8.7 e)	9.7 e)	GB/T 39477—2020
36	责任分级	8.8	分级处置三层机制与责任主体界定	8.8 a)	9.8 a)	—
37	责任分级	8.8	SOAR 告警分级路由至相应处置角色与流程	8.8 b)	9.8 b)	—
38	责任分级	8.8	等级—角色—动作特征分级	8.8 c)	9.8 c)	GB/T 45652
39	责任分级	8.8	高风险应急响应与上报	8.8 d)	9.8 d)	GB/T 45654
40	责任分级	8.8	责任主体复核确认与留痕	8.8 e)	9.8 e)	GB/T 39477—2020、GB/T 45652

附 录 C
（资料性）
风险识别流程输出材料模板

C.1 概述

本附录为资料性，提供第 7 章各阶段输出材料的参考模板，供政务数据共享服务接口安全风险人工智能识别系统的设计、建设与评估方使用。模板字段可根据实际系统调整，但其内容应能支撑对应条款的证实（见第 9 章）。

C.2 监测准备阶段输出材料见表 C.1。

表 C.1 接口资产清单模板

序号	接口名称	提供方	调用方	接口地址/端点	认证方式	数据级别	生命周期状态	责任部门
1							在用/停用	

C.3 监测数据底座说明见表 C.2。

表 C.2 测数据底座说明模板

数据类别	来源系统	数据类型	采集频率	留存周期	敏感字段处理	合规依据
接口访问日志					脱敏/加密	GB/T 39477—2020
认证日志						GB/T 39477—2020
流量报文						GB/T 46796—2025
接口注册与停用台账						5.2

C.4 风险识别阶段输出材料表 C.3。

表 C.3 风险识别结果清单模板

编号	风险类	风险项	风险等级	识别证据	识别时间	识别技术
R-001	R1 鉴权与访问控制	越权调用	高	访问日志 + UEBA 偏离		UEBA
…	R2～R6 / 数据合规					流量检测 / GNN

C.5 分析研判阶段输出材料表 C.4。

表 C.4 研判结论与分级处置建议模板

编号	关联识别项	研判结论	风险等级	处置层级	处置建议	研判人/时间
Y-001	R-001	确认越权	高	应急响应/人工研判/自动处置	即时阻断 + 复核	

C.6 预警处置阶段输出材料表 C.5。

表 C.5 处置工单与联动处置记录模板

工单号	关联研判项	处置动作	处置系统	处置时间	处置结果	复核人
W-001	Y-001	阻断/限流/脱敏/整改	身份与访问管理/数据脱敏/安全防护		成功/失败	

C.7 审计追溯阶段输出材料表 C.6。

表 C.6 审计记录与可追溯存证模板

记录编号	关联对象（识别/研判/处置）	主体	行为	对象	时间	存证哈希	可检索/导出

A-001	R-001 / Y-001 / W-001	系统/人	识别/研判/处置	接口/数据			是

C.8 模板与条款、证实方法对照表C.7。

表C.7 模板与条款、证实方法对照表

输出阶段	输出材料	对应条款	对应模板	主要证实方法
监测准备	接口资产清单、监测数据底座说明	7.2	C.1、C.2	9.1、9.2
风险识别	六类风险识别结果清单	7.3	C.3	9.2、9.3、9.4、9.5、9.6
分析研判	研判结论与分级处置建议	7.4	C.4	9.8
预警处置	处置工单与联动处置记录	7.5	C.5	9.8
审计追溯	审计记录与可追溯存证	7.6	DC.6	9.7

参 考 文 献

[1]《中共中央 国务院关于构建数据基础制度更好发挥数据要素作用的意见》（中办发〔2022〕51号）