ICS 35.030
CCS L 80

、

中华人民共和国国家标准

GB/Z XXXXX—XXXX

网络安全技术 人工智能应用安全 政务

Cybersecurity technology—Artificial Intelligence Application Security—Government

(点击此处添加与国际标准一致性程度的标识)

（征求意见稿）

在提交反馈意见时，请将您知道的相关专利连同支持性文件一并附上。

XXXX－XX－XX 发布　　XXXX－XX－XX 实施

国家市场监督管理总局
国家标准化管理委员会　发布

目　　次

前言.....III
引言.....IV
1　范围.....1
2　规范性引用文件.....1
3　术语和定义.....1
4　缩略语.....3
5　基本原则.....3
5.1　统筹发展与安全.....3
5.2　全程风险管控.....3
5.3　分类分级管理.....3
5.4　持续迭代优化.....3
6　概述.....3
7　基础安全指引.....3
8　场景安全指引.....4
8.1　政务服务.....4
8.2　社会治理.....7
8.3　机关办公.....9
8.4　辅助决策.....12
9　安全管理要求.....16
9.1　安全评估.....16
9.2　漏洞管理.....17
9.3　监测预警.....17
9.4　应急响应.....17
附录A（资料性）　典型的政务人工智能应用网络部署方式示例.....18
参考文献.....19

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。

本文件由全国网络安全标准化技术委员会（SAC/TC260）提出并归口。

本文件起草单位：

本文件主要起草人：

引　言

随着人工智能技术在政务领域的深度应用，人工智能已成为提升政务服务效率、优化治理能力的重要支撑。为规范人工智能应用的安全防护工作，防范数据安全、系统安全、内容安全等各类风险，保障政务数据安全和公众合法权益，依据《中华人民共和国网络安全法》《中华人民共和国数据安全法》《中华人民共和国个人信息保护法》《生成式人工智能服务管理暂行办法》《政务领域人工智能大模型部署应用指引》等法律法规及政策文件，制定本文件。

网络安全技术 人工智能应用安全 政务

1 范围

本文件规定了政务领域人工智能应用过程中的基础安全指引、场景安全指引和安全管理要求。

本文件适用于政务领域人工智能应用的设计、开发与测试活动，可为政务部门开展人工智能应用安全管理、技术提供方落实人工智能应用安全责任工作提供指导和参考。

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 9704—2012 党政机关公文格式

GB/T 20984—2022 信息安全技术 信息安全风险评估方法

GB/T 22239—2019 信息安全技术 网络安全等级保护基本要求

GB/T 25069—2022 信息安全技术 术语

GB/T 41867—2022 信息技术 人工智能 术语

GB/T 45396—2025 数据安全技术 政务数据处理安全要求

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 45577—2025 数据安全技术 数据安全风险评估方法

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

TC260-PG-20268A 网络安全标准实践指南——人工智能应用安全指引 总则

3 术语和定义

GB/T 25069—2022、GB/T 41867—2022界定的以及下列术语和定义适用于本文件。

3.1

政务领域人工智能应用 artificial intelligence applications in the government sector

各级政务部门在政务服务、社会治理、机关办公和辅助决策等业务活动中，采用人工智能大模型等人工智能技术产品，为工作人员提供高效辅助、为公众和企业提供便捷服务的应用活动。

[来源：GB/T 41867—2022,3.1.8，有修改]。

3.2

人工智能生成合成内容 content generated by artificial intelligence

利用人工智能技术生成、合成的文本、图片、音频、视频、虚拟场景等信息。

[来源：GB 45438—2025,3.1]

3.3

政务数据 government data

各级政务部门在依法履职过程中产生、采集和管理的数据。

[来源：GB/T 45396—2025,3.1]

3.4

安全护栏 security guardrails

用于约束和规范人工智能应用行为的一系列策略、机制和技术手段，限制输入输出内容或行为，防止重要数据泄露、提示词注入攻击，以及生成侵犯商业秘密或个人信息、违法和不良信息等。

3.5

智能体 artificial intelligence agent

能够自主感知环境、制定决策、采取行动实现特定目标的智能系统。

注：一般具有记忆、规划、使用工具等基本能力。

[来源：GB/Z 265—2026,3.5]

3.6

生产开发者 artificial intelligence producer

负责设计、开发、测试和部署由一个或多个人工智能系统所构成的人工智能产品或服务的组织。

注1：生产开发者包括模型算法和应用的生产开发组织，如模型设计方、模型实现方、计算验证方、模型验证方、应用开发方等。

注2：在政务领域，生产开发者可能包括人工智能模型与算法研究机构、大模型及大模型平台厂商、政务大模型开发单位和建设单位、智能体生产制造或开发单位、政务行业人工智能系统集成商，以及政务部门自身的研发机构等。

[来源：GB/Z 265—2026,3.5，有修改]。

3.7

应用提供者 artificial intelligence application provider

将人工智能系统能力（例如推理与决策）以人工智能应用方式作为产品或服务提供给内部或外部业务使用者的组织。

注1：应用提供者可能是生产开发者的外部组织（如大模型平台、智能体开发平台、智能体应用等第三方产品或服务提供者），也可能是生产开发者的内部组织。

注2：在政务领域，应用提供者可能包括政务大模型及大模型平台的运行单位、各类政务智能体的运行单位、为政务部门提供人工智能应用服务的集成商或外包服务商等。

[来源：GB/Z 265—2026,3.6，有修改]。

3.8

数据提供者 data provider

为人工智能产品或服务提供数据的组织。

注：在政务领域，数据提供者可能包括政务数据集提供单位、政务语料库提供单位、行业应用数据提供单位、政务管理和运营数据提供单位，也包括使用过程中需反馈业务数据至模型以支持持续学习与改进的组织。

[来源：GB/Z 265—2026,3.7，有修改]。

3.9

业务使用者 artificial intelligence user

最终使用人工智能产品或服务的组织。

注：业务使用者可能与应用提供者之间存在业务关系，例如合作、产品购买或服务订阅；业务使用者可能是应用提供者的一部分，即内部组织，例如政务部门的业务机构，也可能是合作关系，如第三方服务商、业务客户。在政务领域，业务使用者主要指各级政务部门及其依法履职的业务机构。

[来源：GB/Z 265—2026,3.8，有修改]。

3.10

安全运营者 security operator

负责人工智能应用安全运行维护、安全监测、风险预警和应急响应的组织。

注：在政务领域，安全运营者可能包括政务部门自建的安全运营中心、运营团队（自有或外包）以及提供安全运营服务的厂商，承担人工智能应用的安全监测、事件研判、应急处突与持续加固等职责；委托开展安全运营不转移业务使用者和应用提供者依法承担的安全主体责任。

4 缩略语

下列缩略语适用于本文件。

API：应用编程接口（Application Programming Interface）

MCP：模型上下文协议（Model Context Protocol）

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

5 基本原则

5.1 统筹发展与安全

坚持系统谋划、集约发展，全面贯彻新发展理念，统筹高质量发展和高水平安全，有序推进人工智能技术、产品和服务在政务领域的部署、应用和持续优化。

5.2 全程风险管控

坚持风险导向，针对人工智能应用的数据安全、内容安全、系统安全、智能体行为等各类安全风险，开展持续识别、动态评估、闭环处置。

5.3 分类分级管理

根据政务业务的敏感程度、数据重要性及使用场景，对人工智能应用实施分类分级管理。对涉及国家安全、公共管理和民生保障等核心领域的应用，应重点防护。

5.4 持续迭代优化

将安全管理的持续改进作为人工智能应用的关键环节，根据技术发展和新的安全形势，动态调整安全策略和防护措施，不断提高安全防护水平。

6 概述

政务领域的人工智能应用安全包含基础安全、场景安全、安全管理。基础安全指引针对所有政务业务场景的通用安全；场景安全指引针对政务服务、社会治理、机关办公和辅助决策等典型政务业务场景的独有风险；安全管理要求包含安全评估、漏洞管理、监测预警、应急处置等角度。

7 基础安全指引

本项安全指引要求如下：

a) 人工智能应用活动相关各方宜参考 TC260-PG-20268A《网络安全标准实践指南——人工智能应用安全指引 总则》，提升人工智能应用安全水平，防范人工智能应用安全风险；

b) 应用提供者、业务使用者使用生成式人工智能技术面向公众提供服务的，宜符合 GB/T 45654 相关要求；其中涉及生成合成内容标识的，应按照 GB 45438 要求执行；

c) 人工智能应用宜具备人工确认、手动优先、快速切换和回退机制；

d) 生产开发者宜根据互联网及政务外网业务的不同，划分互联网区和政务外网区的人工智能基

础设施，互联网区与政务外网区宜实现逻辑隔离，典型的政务人工智能应用网络部署方式见附录 A；

e) 人工智能应用涉及跨网数据交换的，宜通过数据安全交换技术手段并留存记录；

f) 涉及生成式人工智能的，生产开发者宜按照《生成式人工智能服务管理暂行办法》要求，开展生成式人工智能服务备案工作；

g) 安全运营者开展安全评估、漏洞管理、监测预警和应急处置等工作，见第 9 章安全管理要求。

8 场景安全指引

8.1 政务服务

8.1.1 智能问答

8.1.1.1 场景描述

依托政务大模型，整合本地区、本部门、本领域业务资源和知识库等数据，利用自然语言理解、RAG和知识图谱等技术，提供便捷的在线政务咨询服务，加强对公众诉求的准确理解，实时生成参考回答，帮助解决公众疑惑，提升信息获取便捷性。

公众通过政务门户、移动端或热线以文本或语音发起提问，人工智能应用返回有权威依据且来源可追溯的参考性回答，解决公众疑惑，提升信息获取便捷性。超出知识边界、涉及敏感内容的问题，由系统自动执行拒答、引导至正式渠道或转入人工坐席。

8.1.1.2 风险分析

本场景宜重点考虑以下风险：

a) 知识库数据失效或失真：知识源更新不及时、更新后缓存未清理，废止政策仍被检索，或原始数据本身存在错误，造成回答与现行政策不一致；

b) 个人敏感信息泄露：公众提问中提交身份证号、联系方式、案情细节等敏感信息，一旦被会话缓存、日志复用，易发生敏感信息跨用户泄露；若将此类敏感数据用于模型训练，信息可能固化至模型参数，在后续问答中被生成输出；

c) 知识库违规收录敏感信息：政务工作人员在知识库中加入工作秘密或其他敏感信息，导致敏感信息被检索并返回给公众；

d) 知识库投毒与检索污染：攻击者向知识库或检索索引中注入伪造的政策文件、虚假办事条件，诱导模型输出被操控的答案；

e) 模型幻觉与过度推断：模型在检索结果不足或语义歧义时，生成看似合理但缺乏依据的内容，误导公众对办事条件、政策适用范围的判断；

f) 提示注入与越狱攻击：攻击者通过精心构造的提示词，诱导模型绕过内容过滤、泄露系统提示词或知识库中的受限内容，甚至生成恶意指引；

g) 输出内容违规与歧视偏见：回答中夹带政治性错误表述、歧视性语言或不当政策解读，损害政府公信力。

8.1.1.3 防范措施

针对上述风险，宜采取以下防范措施：

a) 数据提供者宜建立知识源“发布—复核—标注时效”流程，每一知识条目附来源文号、生效日期和失效日期；政策废止或修订后规定时限内完成知识库同步。生产开发者和应用提供者宜在知识条目失效或政策修订时，同步清理或标记缓存中基于旧知识生成的答案，并在检索链路中强制校验知识条目的时效状态，对已过期条目进行过滤或附加显著过期提示；

b) 生产开发者宜在输入入口部署敏感信息识别与脱敏模块，对身份证号、手机号、具体地址等进行实时掩码或阻断，避免原始敏感字段进入模型上下文。安全运营者宜对包含敏感信息的会话

执行隔离存储和限期销毁，不宜将此类数据用于模型训练或日志长期留存；

c) 生产开发者和应用提供者宜提供自动化的敏感信息检查工具，对知识库中的工作秘密和敏感信息进行检查和过滤；

d) 数据提供者宜对拟入库内容执行来源白名单核验，防止未经审核的外部文档直接进入知识库。安全运营者宜监测知识库写入行为和检索结果分布异常，发现集中出现可疑文档时触发人工审计和回滚；

e) 生产开发者宜在生成环节引入置信度阈值和“依据不足即拒答”的硬约束，要求每条回答绑定至少一条知识库引用，对无法绑定的生成结果执行拦截。应用提供者宜在界面上展示引用来源，并将回答明确标注为“参考回答”而非最终结论；

f) 生产开发者和应用提供者宜在模型调用前部署提示注入检测层或安全护栏，识别角色扮演、指令覆盖、编码绕过等攻击模式并阻断。安全运营者宜持续更新攻击特征库，对越狱尝试进行告警、限流和来源封禁，并将新型攻击样本反馈生产开发者迭代检测规则；

g) 生产开发者和应用提供者宜在输出侧配置内容安全分类器或安全护栏，过滤政治敏感、歧视偏见、虚假信息等输出内容；

h) 安全运营者宜建立智能问答的专项监测指标，包括拒答率、转人工率、敏感词命中率、知识库召回异常率、单用户高频请求等；宜制定知识库污染、模型行为异常、大规模错误回答等场景的应急预案。

8.1.2 辅助办理

8.1.2.1 场景描述

整合政务服务办事指南、常见问题、用户评价和历史办理记录等数据，利用智能匹配和自动化处理等技术，提供智能导办、个性引导、表单预填、辅助审核、进度查询和提醒等一站式政务辅助办理服务，辅助工作人员高效审核材料，支撑公众和企业便捷办理事项。

8.1.2.2 风险分析

本场景宜重点考虑以下风险：

a) 敏感信息泄露：表单预填、辅助审核、进度查询需调用身份证件、电子证照、社保、企业登记等数据和历史办理记录，若数据调用授权不严、身份核验不足或者数据隔离不当，可能导致个人信息、企业经营信息泄露，或被冒用身份查询他人办件信息；

b) 信息输出决策不可靠：办事指南、常见问题等业务未及时更新或者模型产生幻觉，可能导致导办路径、材料清单错误；辅助审核等业务因训练数据偏差或者材料格式多样，可能误退合格材料、漏判不合格材料，或对特定群体产生不合理差别对待；

c) 过度授权：辅助办理应用若被赋予自动提交申请、自动作出受理或审核结论等权限，或人工确认环节设置不足，可能导致错误办理结果直接生效；

d) 提示词注入：申请人上传的材料或填写的表单中可能夹带隐藏指令，诱导辅助审核给出“材料齐全”“符合要求”等错误结论。

8.1.2.3 防范措施

针对上述风险，宜采取以下防范措施：

a) 生产开发者：保障辅助审核具备可解释性，能够说明审核意见所依据的办事要求；宜采用多模型交叉校验等方式提升审核准确性；宜建立异常指令检测机制，对申请人上传材料和填写表单的内容进行检测，发现异常时转人工审核；宜针对不同材料格式、不同申请主体开展准确性和公平性测试；

b) 应用提供者：依托统一身份认证系统或者国家网络身份认证公共服务对申请人进行实名核验，将进度查询、表单预填范围限定为本人或者经本人授权的办件，将电子证照等数据调用范围限定为当前事项所需，设置申请人确认环节，预填内容经确认后方可提交；不为辅助办理应用配

置自动提交申请、自动作出受理或者审核结论等权限；宜在智能体调用业务接口前进行动态权限校验，防范越权查询和非授权操作；宜在导办、预填等环节设置转人工服务入口，并对不同申请人的对话记录和办件数据进行隔离；

c) 数据提供者：以政务服务事项基本目录及实施清单为依据，维护办事指南、常见问题，并在政策调整、事项变更时同步更新、实施版本管理；宜按照 GB/T 45396 的有关要求，对身份证件、电子证照、历史办理记录等数据实施分类分级保护，历史办理记录用于模型训练前宜进行去标识化处理；

d) 业务使用者：明确可开展智能导办、表单预填、辅助审核的事项范围，并在上线前组织生产开发者、应用提供者开展场景安全风险评估；宜建立“智能辅助+人工确认”审核机制，辅助审核意见宜经工作人员确认，受理、审批等结果宜由工作人员依法作出；设置申诉复核渠道；宜定期比对辅助审核意见与人工审核结果，对误判情形进行记录分析，并反馈生产开发者迭代优化；

e) 安全运营者：宜建立实时监测和告警机制，对批量查询办件、高频调用预填接口等异常行为进行监测告警，定期审计证照调用和办件查询日志；发现异常宜及时处置。

8.1.3 政策服务直达快享

8.1.3.1 场景描述

依托政务大模型构建政策服务知识库，细化政策要求、政策标签、推送条件、申兑流程等相关内容，利用“政策找人”“政策找企业”算法模型，加强公众和企业需求分析，实现政策智能匹配，推进惠民利民、惠企利企服务主动精准送达和一站式办理。

8.1.3.2 风险分析

本场景宜重点考虑以下风险：

a) 政策知识库投毒与供应链污染：伪造政策文件、篡改申报指南或批量注入错误标签污染知识库与向量索引，或外部模型、插件、向量库植入后门、版本漂移，导致模型引用失效政策或虚假条款，企业错报漏报甚至违规申领；

b) 提示词注入与智能体越权：在问答、材料上传或政策附件、外链中埋入直接或间接注入指令，诱导模型越权；政策服务若扩展为智能体代办，工具调用与系统权限被劫持后可能越权读取他企材料、错误提交申兑或误发通知；

c) 敏感数据泄露与超范围汇聚：企业纳税、社保、信用、法人身份等数据在画像、跨部门共享、推送环节被过度采集、违规共享或明文留存，并在模型记忆、日志、缓存、向量库中残留形成二次泄露；

d) 模型幻觉与匹配失真：模型对适用条件、补贴标准、申报时限、办理层级产生貌似可信的错误解读；算法对行业、区域、企业规模存在隐性偏好，造成特定群体系统性漏推错推，影响政策公平。

8.1.3.3 防范措施

针对上述风险，宜采取以下防范措施：

a) 政策来源权威化与知识库治理。仅接入官方发文渠道，建立政策标签体系与版本管理，对新增和变更标签开展一致性校验；定期开展投毒检测，识别异常来源文本、异常标签聚集与高相似度伪造条款，并抽样复核匹配结果；

b) 全流程安全防护与供应链准入。设计开发阶段开展威胁建模、数据分类分级，明确最小必要采集；对模型、插件、组件开展安全测评与准入，版本变更经回归测试、灰度发布并保留回退；运行阶段实施提示词与工具白名单、注入检测、输出校验，以及业务分区隔离、API 网关鉴权限流、传输存储加密和向量库细粒度访问控制；

c) 数据安全与隐私保护。企业敏感数据脱敏、去标识化后参与匹配，优先采用联邦学习、多方安全计算、可信执行环境等方式；跨部门共享签订协议、明确用途范围期限；训练与检索不存储原始敏感数据，输出侧做敏感信息识别与脱敏，展示时严格隔离其他主体信息；
d) 输出可信保障与追溯审计。政策问答采用 RAG，强制引用原文与文号并标注生效状态；补贴金额、申报期限、适用条件等关键字段以规则引擎校验，不应由模型自由生成；设置信度阈值，低置信度转人工，保留人工保底、纠偏与申诉通道；对推荐算法开展公平性评估与年度审计；
e) 智能体权限与工具治理。对具备自主执行能力的智能体，按任务最小化授予权限与接口，采用临时凭证和动态授权，不以超级管理员或共享账号运行；工具、插件及 MCP 服务建立准入白名单与来源校验，描述与参数完整性保护；提交申兑、变更信息、对外通知等高风险操作强制人工确认；对异常连续工具调用、高风险任务执行等行为，宜设置熔断或中止机制。

8.2 社会治理

8.2.1 智能监测巡检

8.2.1.1 场景描述

利用无人机、视频监控、智能传感器等设备和计算机视觉等技术，对监控视频、图像、物联感知数据等进行实时分析，辅助工作人员实时监测房屋、道路、燃气、桥梁、供水、排水、供热、综合管廊等基础设施，及时发现异常行为、环境问题或设施故障等，自动识别潜在风险隐患，及时进行提醒，并根据异常情况和严重程度提供处置建议，提高监测巡检效率。

8.2.1.2 风险分析

本场景宜重点考虑以下风险：
a) 模型识别失效：恶劣天气、逆光、遮挡、人群密集等复杂环境下，人工智能图像识别出现大量误报、漏报，引发无效工单或隐患漏检，干扰基层治理处置；
b) 隐私违规风险：公共区域视频人工智能违规开展人脸识别、无授权人员行为分析，采集、处理公民个人生物信息，违反个人信息保护相关法规，侵害个人权益；
c) 对抗样本攻击：攻击者对监控摄像头画面添加对抗扰动，欺骗人工智能识别模型，造成异常事件漏判、虚假告警，规避监测；
d) 数据安全风险：监测视频、人员影像、告警事件等感知数据在传输、存储环节被非法窃取、篡改，伪造虚假告警信息，误导处置流程；
e) 输出决策不可靠：模型幻觉生成不存在的异常或隐患，告警与处置建议未经核实即被采纳，影响城市运行安全且责任难以界定；
f) 服务中断与设备失控：感知网络或者边缘计算节点遭攻击、故障导致关键时刻漏报，前端设备被仿冒接入或者固件被篡改后成为向视频专网、政务外网渗透的入口。

8.2.1.3 防范措施

针对上述风险，宜采取以下防范措施：
a) 生产开发者、应用提供者在模型训练、部署阶段，补充雨雾、强光、人流密集等复杂场景样本开展模型鲁棒性测试，保障复杂环境下识别准确率达标，建立误报漏报持续迭代优化机制；
b) 严格落实公共区域人工智能视频监管合规管控，依托人工智能内容安全能力，对人工智能识别能力做权限隔离，限制非法人脸识别、未经审批的人员行为分析，落实最小必要采集原则；
c) 设置人工智能识别置信度阈值管控，低置信度告警结果不应自动派单，强制推送至人工复核确认，人工智能安全网关对模型推理结果做前置校验，过滤低可信度结果；
d) 应用提供者在部署和运行监控阶段宜对无人机、巡检机器人、智能传感器、边缘计算节点等前端设备实行身份认证、准入白名单与固件可信校验；边缘计算节点宜具备物理防拆和访问限制；

e) 监测视频、告警信息传输、存储环节，采用数据加密、访问权限管控、数据水印等安全防护手段，防范感知数据被篡改、窃取，对告警数据完整性校验，拦截伪造告警；
f) 部署人工智能对抗样本检测能力，对前端采集画面实时检测对抗扰动攻击，识别画面篡改欺骗行为，防止人工智能模型被诱导漏判、误判。

8.2.2 辅助执法监管

8.2.2.1 场景描述

采用语音识别、视频分析、知识图谱、逻辑推理等技术，辅助执法人员将案件信息实时录入系统、穿透式发现问题线索、生成案件报告、快速检索法律依据和司法解释、查询类似典型案例等，提供针对性案件办理建议，提高执法监管效率和规范性。

人工智能输出作为执法人员办案的参考，行政检查和行政处罚等执法活动由行政机关依法开展。

8.2.2.2 风险分析

本场景宜重点考虑以下风险：

a) 输入分析风险：执法现场的方言口音、环境噪声和逆光遮挡干扰语音识别和视频分析；
b) 权责越界风险：人工智能系统自动生成、直接下发行政执法决定，替代执法人员实施行政裁量，不符合行政执法法定要求；
c) 模型偏见风险：训练样本覆盖不足、样本存在偏差，造成模型判断偏向，引发选择性执法、监管歧视，损害执法公平；
d) 结果未审核风险：人工智能生成的执法线索、处罚建议不经人工审核确认，直接落地执行，造成错误执法；
e) 审计追溯缺失风险：人工智能辅助决策的输入数据、模型版本、推理记录、处置过程无完整留存，无法审计追溯，执法过程缺乏证据支撑；
f) 投毒污染风险：恶意对模型训练数据集投毒，污染训练样本，造成模型定向误判，恶意干预执法监管结果。

8.2.2.3 防范措施

针对上述风险，宜采取以下防范措施：

a) 生产开发者在模型开发阶段开展数据集安全治理，保证训练样本地域、场景、案例具备充分代表性，利用人工智能数据质检能力开展样本偏见检测，评估并消除模型歧视风险，防范选择性执法；
b) 生产开发者在设计开发阶段和验证确认阶段对语音转写、视频分析模型开展针对方言口音、环境噪声、逆光遮挡等条件的专项训练和测试，输出结果标注所引条文和案例的编号以及置信度，低于置信度阈值的标注为待核实内容；
c) 应用提供者在运行和监控阶段建立法律法规和案例数据与官方发布版本的一致性校验和同步更新机制，对输出中引用已废止、已被修订或者效力层级不当的法律依据的情形告警阻断；
d) 建立人工智能执法建议强制人工审核机制，平台设置流程卡点，所有人工智能生成的执法线索、监管建议、处罚建议必须经执法人员确认后方可流转执行；
e) 对模型版本、输入线索、人工智能输出建议、人工复核操作、最终处置结果进行全链路日志留存，依托审计追溯能力，实现全流程可审计，保障执法过程透明可追溯；
f) 训练数据集接入数据安全校验，实施样本准入、数据溯源，防范数据集投毒攻击；上线后持续对模型输出结果做异常检测，识别模型异常研判行为。

8.2.3 市场风险预测

8.2.3.1 场景描述

运用生成式时间序列分析模型和异常检测算法等，对各类市场数据进行监测和深度分析，捕捉市场动向，包括经济指标波动、异常情况等，预测可能出现的市场风险，研判对经济社会带来的影响和经济走势，并及时发出预警，为政府管理和社会治理提供支撑。

8.2.3.2 风险分析

本场景宜重点考虑以下风险：

a) 预测结果不可靠：模型可能产生幻觉，输出虚构的统计规律、错误的风险等级或曲解政策条款、经济指标，导致调控或监管决策被误导，引发资源配置失当或市场扰动；
b) 输入数据被投毒或篡改：宏观经济、市场行情、舆情等公开或第三方数据源可能遭遇数据投毒、供应链攻击或人为篡改，导致模型基于错误信息生成偏离实际的预测结论；
c) 不确定性未量化与罕见极端事件失效：面对地缘政治冲突、突发公共事件、极端行情等历史数据缺乏的罕见事件，基于历史数据训练的模型可能失效或给出危险建议；若未量化预测结果的不确定性，决策者易对结论过度信任；
d) 敏感信息泄露：涉及经济安全、金融稳定、未公开政策取向等敏感领域的风险预测结果若管控不当、发生泄露，可能被恶意利用，引发或加剧市场扰动、挤兑、抢购等次生风险；
e) 数据偏见与代表性不足：训练数据在地区、行业、企业规模等维度代表性不足，可能导致对特定区域、行业或群体的风险研判出现系统性偏差，影响调控政策的公平性与精准性；
f) 可解释性不足与责任难溯源：预测结论缺乏可解释的依据链路，难以向决策者和公众说明风险判定的来源与逻辑，不利于核查、纠错与责任追究。

8.2.3.3 防范措施

针对上述风险，宜采取以下防范措施：

a) 生产开发者在前期规划阶段明确模型的设计运行域与适用边界：界定可预测的风险类型、数据范围与置信前提；在场景设计阶段建立严格的数据准入规则，对第三方及公开数据源进行来源验证、完整性校验与异常检测，遵循 GB/T 45396—2025、GB/T 45652—2025 等有关要求，确保训练与推理数据的准确性、合法性与代表性；
b) 生产开发者在设计开发阶段构建多源数据融合校验机制：当单一来源数据与综合计算结果偏离度超过阈值时自动剔除或告警；建立模型输出置信度评分与不确定性量化机制，在提供预测结论时同步给出置信区间、适用条件和局限性说明，防止对结果的过度依赖；
c) 生产开发者在验证确认阶段开展极端情景压力测试：利用仿真平台开展极端历史案例与罕见极端事件情景的对抗样本压力测试，评估模型在数据扰动、罕见事件下的鲁棒性；验证预测准确性与公平性，防范算法偏见；
d) 应用提供者在运行监控阶段建立持续评估与更新机制：及时根据市场变化、政策调整和新发风险对模型进行优化；对关键输入参数（如价格、舆情、流动性指标）进行实时监控，当预测性能低于经验证设定的安全阈值或置信度低于阈值时，自动触发人工研判或降级为基于规则的传统研判方式；
e) 政务部门及业务使用者建立人机协同决策机制：对市场风险预测结果，尤其是涉及经济安全、金融稳定等敏感领域和重大调控决策的结论，须由具备相应职责的人员进行实质性审核确认后方可采信或对外使用；低置信度或高影响预测宜强制人工介入，避免直接替代法定决策；
f) 政务部门及业务使用者对敏感结果采取分级保密措施：对涉及经济安全、金融稳定等敏感领域的风险预测结果，严格访问控制与最小知悉范围管理，防止信息泄露引发市场扰动；按照 GB/T 45577—2025、GB/T 22239—2019 等有关要求落实数据安全与系统安全防护；
g) 应用提供者与政务部门宜采用大模型安全护栏并强化留痕审计，宜按照 GB 45438—2025 要求对生成与合成内容进行标识；对市场风险预测应用的全过程（包括输入、模型执行、输出）进行留痕记录与安全监测，日志记录留存时长不少于 6 个月并定期审计，确保预测活动可追溯、可问责。

8.3 机关办公

8.3.1 辅助文书起草

8.3.1.1 场景描述

利用大语言模型的生成能力，通过构建本地知识库和预设模板，为工作人员提供写作建议、辅助起草文书，对格式和内容等进行检查、校对和优化，提高工作效率，减轻基层负担。

起草过程常需引入未公开的政策取向、统计数据、内部工作信息与个人敏感信息，输出内容经审核签发后具有法定效力或对外影响，同时涉及公文体例、保密管理、内容真实性与责任归属等多重要求。

8.3.1.2 风险分析

本场景宜重点考虑以下风险：

a) 工作秘密泄露：将内部敏感素材输入未采取安全隔离措施的公开大模型服务，或经提示词、上下文、缓存与日志等途径外泄，导致失泄密；
b) 模型幻觉与事实性错误：模型生成虚构的法律法规条文、政策依据、统计口径、日期、称谓或机构名称，若未经核验直接进入公文，将影响公文权威性与行政决定的合法性；
c) 引用与著作权风险：生成内容摘编、拼接第三方材料未标明来源，或引用未公开、未授权、内部材料，引发抄袭、著作权与信息公开合规问题；
d) 责任主体虚化：人工智能输出被直接当作正式公文签发，或对外以政务机关名义发布，人工审核、依法签发与责任把关环节被绕过，出现错误时难以追溯和问责；
e) 语料与知识库污染：公文语料库、模板库及 RAG 知识库来源不合法、授权不清，或被投毒、越权访问，导致生成内容夹带敏感信息、偏见或被操纵的输出；
f) 公文体例与标识缺失：生成文书不符合 GB/T 9704—2012 等公文格式要求，或对人工智能生成合成内容未按 GB 45438—2025 添加标识，造成体例错误与来源混淆；
g) 过程不可追溯：输入提示、所用模型及版本、参考素材、生成与修改记录、审核意见未全程留痕，日志可能被篡改，事后无法复核事实依据与责任链条。

8.3.1.3 防范措施

针对上述风险，宜采取以下防范措施：

a) 生产开发者在设计与开发阶段明确起草辅助系统的适用边界与保密约束：内置敏感内容识别与拦截能力，对输入提示、上传素材及输出内容开展敏感信息检测；按 GB/T 9704—2012 配置公文文种模板与版式校核规则，并对人工智能生成合成内容按 GB 45438—2025 要求添加标识；
b) 应用提供者宜选用已完成网信部门备案、通过安全评估的模型产品和服务，不应使用未经安全评估的开源模型或境外模型直接处理政务文书；内部敏感文书起草宜在物理或逻辑隔离环境中，使用通过保密审查的系统，严禁在未经授权的公开大模型服务中处理政务文书，涉及核心业务数据的宜优先采用私有化部署或可信环境；
c) 数据提供者对语料、模板与知识库实行来源与分级管理：确保来源合法、授权清晰，不含未经脱敏的工作秘密；引入 RAG 时，知识库宜分级受控、访问可追溯，并建立投毒检测与更新审核机制；
d) 业务使用者建立人工主导的审核与签发机制：宜根据文书敏感程度对使用场景进行分级管理；宜对生成内容进行事实核查与引用校验，对涉及的法律法规条文、政策依据、引用数据、日期、称谓、机构名称等关键信息逐项核验，涉及引用、摘编的宜标明来源并核验著作权与公开性；辅助生成的文书内容宜在使用前经具备相应权限的人员审核确认，防止因模型幻觉、知识过时等导致的事实错误、表述失当或违规内容，人工智能输出仅作为起草辅助，不应替代人工把关与依法签发的职责，不应以人工智能生成内容冒充正式公文或对外以政务机关名义发布未经审核的内容；生成文书宜符合党政机关公文格式要求；
e) 安全运营者对起草全过程实施留痕审计与安全监测：对起草过程中的输入提示、所用模型及版本、参考素材、生成与修改记录、审核意见等宜全程留痕，相关日志宜妥善保护并防止篡改，

支持事后追溯与责任认定；按照 GB/T 22239—2019、GB/T 45396—2025 等有关要求落实数据安全与系统安全防护，对异常调用、批量导出、敏感信息外发等行为进行识别和处置。

8.3.2 资料检索

8.3.2.1 场景描述

利用知识图谱构建和信息检索等技术，准确理解工作人员资料检索需求，实现政务信息快速检索、精准定位、多维度排序、智能关联和对比分析等，帮助工作人员提升资料检索效率和准确性。

8.3.2.2 风险分析

本场景宜重点考虑以下风险：

a) 提示词注入：用户通过构造特定检索语句干扰模型语义理解，诱导系统忽略既定权限约束与检索范围限制，越权调取非本岗位、非本部门可见的敏感资料；或检索指令中暗含恶意内容，污染生成式问答结果；
b) 敏感数据泄露：资料库汇聚政策文件、公文档案及业务数据，检索返回的内容片段、引用原文及生成式摘要可能包含未脱敏的个人信息、内部工作信息或高密级资料，存在越权访问与数据泄露风险；
c) 数据投毒与知识污染：知识图谱构建与向量库入库环节中，未授权数据、错误信息或恶意构造的关联实体被写入知识底座，导致检索结果失真、政策引用错误，影响政务决策依据的准确性；
d) 检索日志与交互数据泄露：检索请求、交互上下文与操作日志记录工作人员查询意图与业务关注点，若未实施加密存储与访问管控，存在日志泄露与操作溯源失效风险。

8.3.2.3 防范措施

针对上述风险，宜采取以下防范措施：

a) 生产开发者、应用提供者在设计开发阶段、部署阶段与运行管控阶段，宜采用查询意图识别、敏感指令过滤与提示词安全检测等措施，对用户输入进行安全校验；同时建立检索前置权限校验机制，采用访问控制、最小授权与操作审计等措施，结合用户岗位权限、数据密级与业务范围动态判定检索范围，防止提示词注入导致的越权访问与结果污染；
b) 生产开发者、应用提供者在设计开发阶段和部署阶段，宜对资料库实施分级分类管理与脱敏处理，检索返回结果按密级与权限做过滤裁剪，生成式摘要输出前执行敏感信息识别与拦截；同时采用数据加密、数据签名等措施保障资料在传输与存储环节的保密性与完整性，避免敏感数据泄露；
c) 生产开发者、应用提供者在设计开发阶段和部署阶段，宜对知识图谱构建与向量入库过程实施数据源身份认证、数据溯源与一致性校验，采用数据源认证、内容审核等措施保障入库数据真实可信，防止数据投毒与知识污染，确保检索结果与政策依据准确可靠；
d) 生产开发者、应用提供者在运行管控阶段，宜对检索请求、返回结果与用户操作实施全流程留痕审计，日志数据采用加密存储、访问控制与防篡改保护，确保操作行为可追溯、可审计。

8.3.3 智能分办

8.3.3.1 场景描述

利用自然语言理解和多模态识别等技术，构建多维度任务分类和分办规则，对来文、来电、工单等任务进行自动识别、准确分类、辅助填写和优先级排序，实现辅助分发和智能派单，提高任务分办效率。

8.3.3.2 风险分析

本场景宜重点考虑以下风险：

a) 算法偏差与错分漏分：训练样本在地区、部门、事项类型上代表性不足，或规则配置不当，导致分派不公、错分漏分，影响履职时效与相对人权益；
b) 人工纠偏缺位：缺少分派结果的人工确认和修正机制，错误分办不能被及时发现与纠正，形成办理空转；
c) 规则被操纵：通过构造来文内容诱导模型改变分类或越权分派，形成定向流转与权限绕过；
d) 责任链条不清：分办依据、置信度与处置过程无留痕，出现错分或泄密时难以定责与追溯。

8.3.3.3 防范措施

针对上述风险，宜采取以下防范措施：
a) 生产开发者在设计开发阶段确保分派逻辑的合规性和可解释性，提供规则配置、依据查询与置信度输出能力，并按事项类型、部门、地区等维度开展偏差评测，不应因算法偏差导致任务分配不公；
b) 数据提供者保证训练与规则样本来源合法、标注准确、分布均衡，涉及敏感数据的样本宜脱敏或隔离使用，并定期复核样本时效性与代表性；
c) 政务部门及业务使用者建立分派结果的人工确认和修正机制，明确承办主体与复核责任人，对关键事项实行强制复核，纠正错分后方可进入正式办理流程；
d) 安全运营者对分办规则变更、模型版本、置信度、分派结果与人工修正全程留痕，日志防篡改；对异常批量分派、规则频繁变更、越权访问等行为进行监测预警与定期审计。

8.4 辅助决策

8.4.1 灾害预警

8.4.1.1 场景描述

对来自卫星、地面传感器、地质监测站，以及预报预警、灾害风险普查等多源、多维、多模态数据进行大数据关联和综合分析研判，识别异常波动情况，预测可能发生的自然灾害，并提前发出预警，辅助政务部门及时采取有效措施，减轻灾害风险，减少灾害损失。

8.4.1.2 风险分析

本场景宜重点考虑以下风险：
a) 训练与运行数据投毒及污染风险：攻击者通过篡改传感器采集值、污染遥感样本或社会上报数据，向人工智能模型投喂带有偏差的样本，使模型在真实灾害发生时系统性低估风险等级，或在无灾害时频繁误报，造成预警疲劳；
b) 对抗样本攻击风险：攻击者对遥感影像、视频监控帧、传感器时序信号施加人眼难以察觉的微小扰动，诱导目标检测与识别模型漏检灾害迹象，或将非灾害场景误判为灾害事件；
c) 模型幻觉与误判风险：政务大模型在生成辅助决策建议时，可能基于训练语料虚构影响范围、错配应急响应级别、编造避难场所位置与物资储备数据，形成虚假的决策依据，误导指挥决策；
d) 提示注入与越权输出风险：外部上报信息、舆情文本或第三方数据中夹带恶意提示词，诱导大模型突破业务权限边界，输出未授权的人口分布、关键基础设施、军事管理区等敏感地理信息，或伪造应急指挥指令；
e) 模型漂移与精度退化风险：受气候变化、城市下垫面改造、极端事件样本稀缺等影响，模型输入分布随时间发生漂移，导致灾害识别准确率、预警提前量和等级研判精度逐步下降，且退化过程不易被及时发现；
f) 敏感地理与人口数据泄露风险：灾害预警输出天然关联受灾人口、避难场所、医院、水库、重要防护目标等空间敏感信息，模型推理结果、中间特征向量、对话日志一旦外泄或被过度聚合，可能形成敏感地理信息和关键基础设施分布图谱；

g） 供应链与第三方接口劫持风险：模型所依赖的第三方气象数据接口、地图服务、外部大模型应用程序接口、开源算法组件被劫持或投毒，导致预警信号被篡改、延迟或替换；
h） 服务可用性与拒绝服务风险：在台风汛期、地震活跃期等高敏感时段，人工智能推理与预警发布服务遭受拒绝服务攻击或资源耗尽攻击，致使预警链路中断；
i） 可解释性不足与责任边界模糊风险：黑盒模型给出高等级预警但无法说明依据，指挥员不能采信；反之，模型建议与人工决策混同，事故发生后难以界定人工智能辅助建议与人工决策各自责任，不满足政务问责要求。

8.4.1.3 安全风险防范措施

针对上述风险，宜采取以下防范措施：
a） 数据来源可信与交叉校验：多源感知数据宜实施来源身份认证、传输加密与完整性校验；人工智能模型输入宜设置跨源一致性比对机制，当单一数据源与其他独立数据源出现显著背离时，宜自动标记可疑并触发人工复核，不应仅依据单一数据源自动生成高等级预警；
b） 模型鲁棒性与对抗防护：灾害识别与研判模型宜在训练和测试阶段引入对抗样本与异常扰动样本，开展鲁棒性评测；上线前宜通过红队测试验证对遥感图像、传感器时序数据扰动攻击的抵抗能力；关键模型宜部署输入异常检测层，对偏离训练分布的输入触发降级告警；
c） 输出可信度与人机协同决策：人工智能生成的灾害预警等级、影响范围与处置建议宜附带置信度区间和关键触发因子；达到高等级（建议发布橙色、红色预警）的结论，未经具备资质的应急指挥人员人工复核并双签确认，不应直接面向社会发布。系统宜明确标识“人工智能辅助建议”，不应将模型生成内容直接作为行政命令下发；
d） 提示注入与越权输出防护：宜建立输入提示词过滤与外部数据隔离机制，社会上报、舆情等非结构化数据在进入大模型上下文前宜去除指令性语句；模型输出宜按用户岗位权限实施最小化返回，对人口分布、关键基础设施、敏感地理目标等字段按密级脱敏，模型不应在普通问答中聚合输出；
e） 模型持续监测与版本管控：宜对灾害识别准确率、误报率、漏报率、预警提前量等关键指标实施在线监测和滚动回测；当指标越过阈值或监测到分布漂移时，宜自动暂停该模型的自动决策权限并切换至人工研判模式；模型再训练、微调和版本发布宜经过安全评估、样本审查和灰度发布，不应使用未经审查的新样本直接更新生产模型；
f） 敏感信息分级与输出脱敏：宜按照地理信息、人口信息、关键基础设施保密要求，对模型训练语料、推理日志、中间结果和对外输出实施分级分类管控；面向公众的预警信息仅保留必要影响范围和避险指引，不应展示精细化人口分布、重要防护目标位置等敏感要素；模型调用日志宜按政务审计要求留存，不应跨境、跨域调用。
g） 供应链安全与接口可信：第三方数据接口、地图服务、外部模型和开源组件宜纳入政务采购安全审查清单，实施来源登记、漏洞管理和版本锁定；关键预警链路宜优先使用自主可控、通过安全评估的模型与组件；外部接口调用宜设置白名单、流量监控和异常熔断机制；
h） 高可用与应急降级：人工智能预警服务宜在政务云内网实施双活或多活部署，与主链路独立建设基于阈值规则、人工研判的备用预警通道；在汛期、震情活跃期等高风险时段宜开展容量压测和拒绝服务防护演练；人工智能系统故障时宜能无缝切换至传统预警机制，不应因人工智能系统不可用而停止预警发布；
i） 可解释性、留痕与可问责：模型对外输出的预警结论宜同时给出关键触发数据、模型版本、生成时间和置信度，关键特征宜能够可视化回溯；模型调用、推理结果、人工复核、指令签发全链路宜完整留痕，留存期限满足政务审计要求；宜在制度上明确人工智能辅助决策不替代行政决策责任，决策责任人仍为具备相应权限的政务人员。

8.4.2 应急处置

8.4.2.1 场景描述

利用强化学习等技术，对社会公共安全等突发事件的性质、特点、危害程度、影响范围、发展趋势和公众反应等进行分析研判，及时发现和预警风险隐患，基于突发场景、力量资源分布等快速模拟应急处置方案效果，提供科学合理的应急处置建议，优化救援资源配置，提高应急响应速度和效率。

8.4.2.2 风险分析

本场景宜重点考虑以下风险：

a) 态势研判不准确：现场报告、监测数据和公众反馈可能存在迟报、信息冲突或重复转报等情况。若模型将同源信息视为多个独立来源，或依据局部公众反应推断事态走势，可能误判事件性质、危害程度、影响范围和发展趋势，造成预警误报、漏报或延迟；
b) 处置建议不可靠：方案推演采用的道路通行、人员疏散等条件可能与现场实际不符。采用强化学习优化处置策略时，若过度侧重响应速度，可能低估人员安全风险和次生危害。救援队伍、装备和物资的可用状态更新不及时，或不同任务对同一资源的占用未被识别，也可能导致重复调派、重点区域保障不足或救援延误；
c) 过度授权：应急处置应用若被赋予直接发布预警、下达疏散指令或调派救援力量等权限，或人工确认环节设置不足，可能导致错误的研判结果或处置建议直接转化为指令，影响现场处置。
d) 敏感信息泄露：事件研判、方案推演和救援资源配置可能涉及受影响人员身份及位置、救援力量部署、关键设施运行状态等信息。若访问授权不严、共享范围过宽或输出审核不足，可能导致个人信息和敏感政务信息泄露；
e) 提示词注入：使用大模型分析公众上报材料、网络舆情等外部文本时，相关内容可能夹带恶意指令，诱导模型忽略已核实的现场信息、错误调整风险等级或生成不当处置建议；
f) 服务中断：突发事件期间，请求量激增、通信受阻或关键数据接口故障，可能导致研判与推演服务不可用。若系统未识别关键数据已停止更新，仍可能依据过时的现场信息和资源状态生成建议，影响指挥调度。

8.4.2.3 防范措施

针对上述风险，宜采取以下防范措施：

a) 数据提供者、应用提供者宜核查现场报告、监测数据、公众反馈和救援资源数据的来源及时效性，标明原始来源、采集时间和核实状态，对同源转报信息进行去重，对相互矛盾或尚待核实的信息作出标记。应用提供者使用大模型分析外部文本时，宜将待分析内容与系统指令隔离，并通过权限控制限制外部文本触发指令发布、资源调派等操作；
b) 生产开发者、应用提供者宜结合历史事件和应急演练，验证模型在信息缺失、事态升级和现场条件突变等情形下的研判与推演能力。采用强化学习优化处置策略时，宜将人员安全、次生风险防控和重点区域保障要求纳入策略约束，并验证推荐方案是否满足这些要求；
c) 应用提供者宜将现行有效的应急预案和救援资源台账作为方案推演依据，区分资源总量、已调派数量和当前可用数量，识别不同任务对同一资源的重复占用。输出处置建议时，宜标明所依据的现场信息、关键假设和适用条件。现场态势或资源状态变化影响方案可行性时，宜提示业务使用者重新研判；
d) 业务使用者宜结合现场核实情况复核研判结论和处置建议，并在调派前确认救援资源的实际可用状态。涉及发布预警、调整响应级别、组织人员疏散或调派救援力量的，宜由有权的指挥机构或人员按既有程序作出决定。宜记录所采纳的建议、复核依据和最终决定；
e) 应用提供者、业务使用者宜根据岗位职责和处置任务限定敏感信息的访问与共享范围，对人员身份、精确位置等信息按需脱敏，对敏感信息的查询、导出和共享操作留痕，并对拟公开发布的信息进行审核；
f) 安全运营者、应用提供者宜监测关键数据更新和应用运行状态。关键数据失效时，宜暂停生成依赖该数据的建议。数据接口中断或应用服务异常时，宜及时向业务使用者通报受影响的功能和数据范围。业务使用者宜根据影响情况按应急预案转入人工研判和指挥调度，并通过演练验证切换流程。

8.4.3 政策评估

8.4.3.1 场景描述

利用人工智能大模型推断分析能力与数据挖掘能力，对公众反馈、市场反应、经济指标和社会满意度等进行分析，构建多维度指标，评估政策目标实现程度、政策影响力和潜在问题，支撑政策制定部门进行政策优化。

8.4.3.2 风险分析

本场景宜重点考虑以下风险：

a) 数据失真：评估数据来源不可靠、样本代表性不足、覆盖不均衡或数据被污染、篡改，导致评估结论失真；
b) 模型输出偏误：模型产生事实性错误、知识不一致、偏见歧视或不当推断，影响评估结论的准确性、客观性和公平性；
c) 评估过程不透明：评估结果缺乏依据核验、来源标注和人工复核，评估过程不透明、不可解释，导致错误结论被不当采信；
d) 敏感信息泄露：评估数据、中间结果或生成结论中的个人信息、工作秘密及敏感政务数据等因处理不当发生泄露、滥用或违规输出；
e) 结论依赖：过度依赖人工智能生成的评估结论，影响政策决策的科学性、公正性和可问责性。

8.4.3.3 安全防范措施

针对上述风险，宜采取以下防范措施：

a) 生产开发者：宜针对政策评估场景开展模型安全性、可靠性、公平性和可解释性评估，重点检测事实性错误、知识不一致、偏见歧视和异常输出等风险，并通过典型场景测试验证模型输出的稳定性，为评估结论提供置信度信息和依据支撑；
b) 应用提供者：宜明确政策评估应用的适用范围和使用边界，对政策知识库及模型输入输出内容开展安全检查，识别知识库污染、提示注入、敏感信息泄露和违规输出等风险；宜建立评估结果核验、引用来源标注和人工复核机制，确保评估结论的生成过程可追溯、可解释；
c) 数据提供者：宜核查评估数据的来源、时效性、完整性、代表性及适用范围，关注不同地区、群体和时间范围的数据覆盖情况；宜依据数据分类分级要求采取相应的脱敏处理和访问控制措施，并对评估数据开展敏感信息、违规内容及数据污染风险检测；
d) 业务使用者：宜结合政策目标、实际业务情况和政策依据，对人工智能生成的评估结论进行综合研判，重点核验异常结论、重大政策影响及涉及特定群体权益的分析结果；不应将人工智能生成的评估结果作为政策制定、调整或实施的唯一依据，涉及重大政策决策的，宜组织专家论证和集体决策；
e) 安全运营者：宜对政策评估应用的模型调用、数据访问、内容交互及关键操作开展安全监测和日志审计，识别提示注入、异常访问、越权调用及敏感信息泄露等风险；宜建立风险告警、事件处置和审计追溯机制，支持安全问题定位、处置及防护措施的持续优化。

8.4.4 智能辅助评审

8.4.4.1 场景描述

利用自学习泛化认识、类人化评审推理、多模态智能解析等能力，对照有关要求开展项目评审，对项目文件内容进行深度扫描、智能解析，提出评审意见建议，辅助提高项目评审效率和科学性。人工智能仅作为辅助工具输出参考意见与风险提示，评审结论由评审专家组依法独立作出。

智能辅助评审的应用场景可包括：

a) 对政策文件、规范性文件开展合法性与一致性审查，检索上位法依据，识别条款冲突与表述歧义；
b) 对规划方案、立项申请和资金申报材料开展可行性、要件完整性初审，形成问题清单与补正提示；
c) 结合评审规则库与历史案例，生成评审意见初稿和评分参考，辅助专家集中评议、归集意见；
d) 对答复意见、修改说明和整改情况开展一致性核验，形成评审过程留档材料。

8.4.4.2 风险分析

本场景宜重点考虑以下风险：

a) 评审规则失控风险：评审标准、评分细则和权重固化于模型隐式参数，规则变更未审批、未留痕，导致评审尺度不一、结论不可解释；
b) 算法偏见风险：训练样本集中于特定地区、行业，模型对新兴领域、中小主体评分系统性偏低，引发评审不公和选择性评审；
c) 依据滞后风险：政策法规库、规则库更新不及时，模型引用已废止条款或虚构依据，形成错误结论；
d) 评审数据泄露风险：材料含商业秘密、未公开决策信息和个人信息，在上传、推理、输出环节被越权访问、外带或残留于上下文；
e) 复核追溯缺失风险：评审意见和评分建议未经专家实质复核即作为结论依据，评审对象、模型与规则版本、推理记录及专家修改痕迹未完整留存，难以审计定责。

8.4.4.3 防范措施

针对上述风险，宜采取以下防范措施：

a) 生产开发者、应用提供者将评审规则、评分细则和权重以显式规则库配置，设置业务权限限制，规则变更经业务主管部门审批并留痕，模型输出能回溯至具体条款依据；
b) 在模型开发阶段开展训练样本安全治理，保证样本在地域、行业、主体规模上具备代表性，上线后定期开展公平性监测与偏差分析；
c) 建立评审知识库准入与更新机制，法规和规则经业务部门审核入库并标注生效状态，采用RAG限定作答依据范围，核验引用条款有效性；
d) 按数据分类分级要求实施分域存储、权限管控、敏感字段脱敏和传输存储加密，会话结束后及时清理材料，未经审批不应用于模型训练；
e) 建立“人工智能初评、专家复评、集体决策”分级复核机制，设置流程卡点，评审意见和评分建议经专家确认、修改留痕后方可流转，重大事项不应仅依据人工智能结果作出结论。

9 安全管理要求

9.1 安全评估

本项安全指引要求如下：

a) 宜依据 GB/T 20984—2022、GB/T 22239—2019 等定期开展人工智能应用安全评估；
b) 针对模型的测评宜包括但不限于：
   1) 内容安全测评：覆盖语料安全、生成内容安全（GB/T 45654—2025 附录 A 中的安全风险）、拒答测试、价值观对齐测试；
   2) 模型鲁棒性与可靠性测试：测试模型对抗攻击防御能力、模型幻觉检测、提示词注入攻击测试、模型越狱测试。
c) 针对智能体的测评宜包括但不限于：针对智能体的感知、记忆、规划、工具使用及行为安全进行测试；检测核心组件漏洞、插件供应链安全及多智能体交互风险；智能体的模型、知识库、工具插件等发生重大变更时，宜开展针对性安全回归测试；

d) 宜在人工智能应用停用前对模型文件、数据资产、运行日志等进行安全评估，防范数据泄露风险；
e) 宜定期开展攻防实训，模拟真实攻击场景发掘潜在漏洞。

9.2 漏洞管理

本项安全指引要求如下：
a) 宜建立漏洞分类分级机制，对漏洞进行分类分级管理；
b) 宜搭建漏洞管理平台，跟踪人工智能应用相关软硬件安全漏洞、模型缺陷等信息，确保已发现漏洞得到及时整改，无遗漏风险；
c) 宜构建漏洞智能化研判与联动处置能力，实现漏洞的自动预警与封堵；
d) 宜建立供应链漏洞台账，将模型训练框架、插件、组件、数据集等纳入供应链漏洞管理范围。

9.3 监测预警

本项安全指引要求如下：
a) 宜构建覆盖“输入-模型执行-输出”全流程的安全风险监测预警系统架构；
b) 宜开展实时流量监测与审计，利用节点或网关流量，对输入和输出进行实时语义分析，识别违规内容、隐私泄露或恶意指令；
c) 宜开展人工智能防火墙/安全护栏建设：部署前置过滤网关，拦截提示词注入、越狱攻击；部署后置过滤，拦截模型幻觉或违规输出；
d) 宜开展异常行为监测：监测 API 调用频率异常、敏感数据高频访问、智能体异常工具调用等行为；
e) 宜开展威胁情报生产和共享：建设人工智能对抗样本库、攻击指纹库、投毒样本等。

9.4 应急响应

本项安全指引要求如下：
a) 宜建立应急响应机制，具备针对数据泄露、模型异常、网络攻击等突发安全事件进行快速处置、溯源定位的能力；
b) 宜制定应急预案，定期开展应急预案演练，评估其可操作性，确保在突发情况下能有效控制风险，降低损失；
c) 安全事件宜及时上报政务主管部门，并开展处置、溯源、整改；
d) 宜建立智能化响应能力，能够实时监测和自动研判安全事件、溯源攻击路径和触发应急措施，缩短应急响应时间。

附 录 A
（资料性）
典型的政务人工智能应用网络部署方式示例

典型的政务人工智能应用网络部署方式如图1所示。

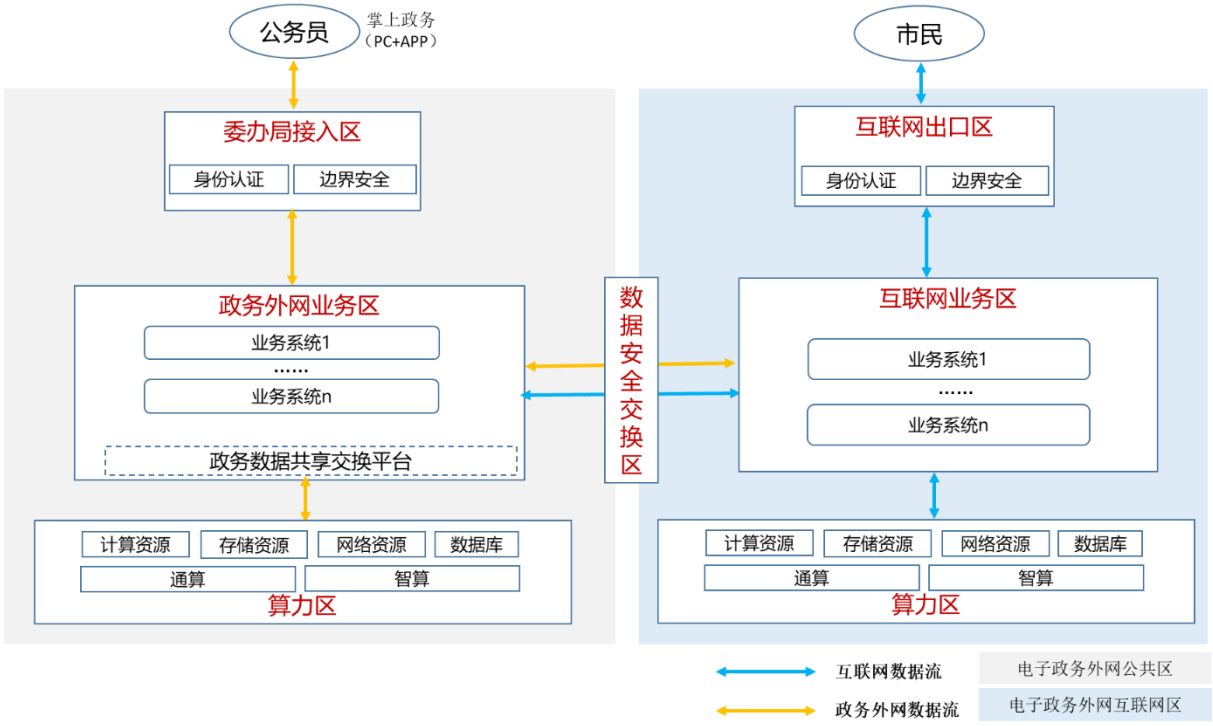


图 1 网络部署方式

参 考 文 献

[1] GB/Z 265—2026 “人工智能+交通运输”应用安全指引
[2] GB/T 42888—2023 信息安全技术 机器学习算法安全评估规范
[3] TC260-PG-20268A《网络安全标准实践指南—人工智能应用安全指引 总则》
[4] 人工智能安全治理框架3.0
[5] 生成式人工智能服务管理暂行办法
[6] 政务领域人工智能大模型部署应用指引
[7] 智能体规范应用与创新发展实施意见
[8] 中华人民共和国网络安全法
[9] 党政机关公文处理工作条例（2012年4月16日中共中央办公厅、国务院办公厅印发）
[10] 网络信息内容生态治理规定