奇安信

政务大模型
安全治理框架

概 要

政务大模型是专注在政务领域的行业大模型，政务大模型在通用大模型基础上，通过检索增强生成或者模型微调等技术手段，使模型更加贴合政务的特点和需求。基于政务大模型的应用，为政府决策、公共服务、带来了全新的变革。然而，政务大模型应用也伴随着各种潜在的风险。

政务大模型安全风险主要包括七个主要类型，即数据安全风险、训练语料安全风险、模型安全风险、应用安全风险、软件供应链安全风险、生成内容风险、大模型自身风险。

政务大模型的安全治理框架是保障其安全、合规运行的基础。全面的安全治理框架，需要满足合规要求、建立完善安全机制、提供安全技术保障。

其中，合规是首要原则，涉及多项法律法规、规章制度的遵循。安全技术保障是政务大模型安全治理框架的核心，涵盖从基础安全措施、数据安全、大模型开发安全到运行安全等方面。

政务大模型的安全治理上，需要政府、企业、研究机构、监管机构等各方协作，形成良性互动生态体系，构建安全、可靠、有序的环境，充分发挥政务大模型在提升政府服务效能、推动社会进步方面的积极作用。

1 生成式人工智能快速发展，政务大模型为数字政府赋能

人工智能已成为引领科技革命和产业变革的战略性新兴产业，以大模型为代表的生成式人工智能，是人工智能技术发展的重要方向。基于大模型的生成式人工智能，从文本处理到声音、图像、视频等多媒体处理，从理解知识到创造知识，正在向着通用人工智能（AGI）的长期目标向前跨越。

大模型正在赋能千行百业。在政府行业，大模型推动政府迈入智能化。在数字政府建设中，大模型将提升政府服务水平，提高政务服务的体验，同时优化决策制定、加强政府内部协同、提升城市治理能力，为打造服务化、智能化、现代化的数字政府提供关键技术支持。

1.1 政务大模型应用广泛，各地陆续开展应用实践

政务大模型在政务领域有着广泛的应用场景。首先，在公共服务方面，政务大模型可以提供智能化的政务咨询和办事指引。为公众提供准确、个性化的解答，提高政务服务的效率和满意度。

其次，在政务办公方面，大模型可以辅助公文写作、文档处理、会议纪要整理、内部知识管理等日常工作。它还可以分析部门间的数据，发现管理中的问题和优化空间，促进跨部门协作。

再者，在城市治理等政府监管领域，大模型可分析城市运行的大数据，预测和识别问题，支持政府进行更科学的决策和更有效的资源分配，实现城市管理的智能化和精细化。例如在城市应急管理方面，大模型能够快速整合道路摄像头信息，为决策提供支持。

政务大模型开始在一些省市落地。在北京亦庄经开区，亦智政务大模型服务平台正式上线。平台实现了智慧政务小助手、迎商中心数字人、智能决策助手、亦城慧眼、实验室智能监管等多个场景应用。在北京市生态环境局，打造了生态环境“监管－监测－监察” -- “三监”联动大模型，支撑了新阶段大气污染防治科学、精准、依法治污。在广东省政数局，上线基于大模型的 AI 智慧服务，提供智慧搜索和 AI 智能问答，提升一体化在线政务服务能力。

1.2 政务大模型，以通用大模型为基础来构建

政务大模型是专注在政务领域的行业大模型，是在通用大模型的基础上，通过知识增强、模型微调或者增量预训练来构建。政务大模型构建是一个循序渐进的过程。首先，可以对通用大模型进行知识增强，通过提示词调优和数据检索提升问答效果。再者，通过微调和增量预训练形成政务领域垂直模型。不同的构建方式，都是将大模型技术与政务需求紧密结合，提高大模型对政务行业的领域知识，更好地服务于上层的政务大模型应用。

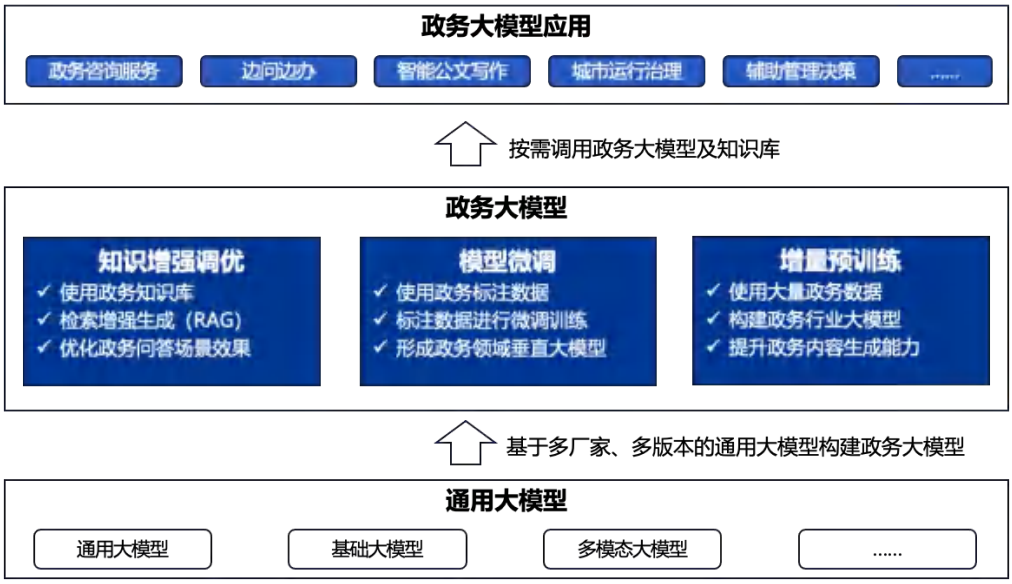


图 1：政务大模型应用构建方式

1.3 政务大模型，在应用形式和模型适配上具有多样性

政务领域涉及多个层级的政府机构、不同的职能部门，以及广泛的公共服务范畴。这种复杂的应用环境要求大模型具有多样化的特征，以适应不同的需求和场景。大模型应用自身的多样性，叠加数字政府不同职能部门的需求，放大了政务大模型应用的多样性。

大模型应用形式多样，在与用户交互界面上，有对话机器人、与现有应用集成的 AI 助手，也有处理自动化 Agent 智能体等形式。大模型应用可以作为独立的应用程序，如智能客服，类似于面向互联网用户的大模型对话机器人。对话机器人可以运行在手机移动端，也可以在 PC 上，用浏览器交互。另一些则被集成到现有的政务应用中，以 AI 智能助手形式，类似微

软提出的 Copilot 副驾驶模式，提升现有应用的智能化水平。随着技术发展，Agent 智能体受到关注，智能体具备自主性、反应性和学习能力，可以与人或其他 Agent 智能体交互，自动化完成任务。

模型适配也具有多样性，政务大模型可以通过知识增强、模型微调和增量预训练，学习政务相关的知识，提高政务大模型回答的准确性。检索增强生成（RAG）是比较广泛的知识增强方式。RAG 在用户问题上，补充政务业务的上下文，大模型在回答问题或生成文本时能够引用相关信息，提高生成内容的质量。大模型如果需要回答精确的数据，如智慧问数应用，需要引入插件访问业务系统的数据库，获取结果后再响应。

政务大模型应用和模型适配的多样性，为政务大模型应用落地带来复杂性。在政务大模型应用开始，需要仔细考虑应用场景和目标，确定应用形式和模型的适配方式，同步开展数据的准备工作。

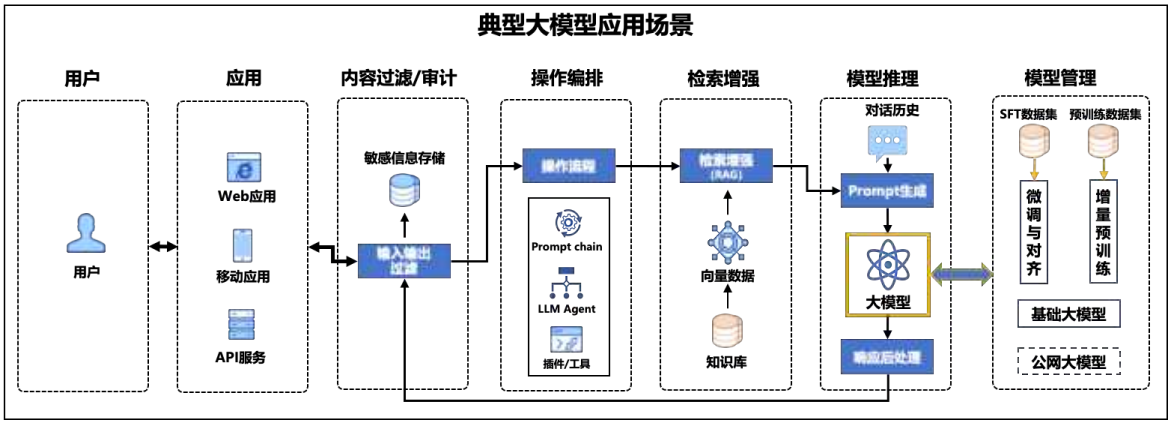


图 2：典型政务大模型应用场景

2 政务大模型安全风险分类

随着政务大模型在政府管理、公共服务中的广泛应用，其强大的自然语言和数据处理和分析能力为政府决策、社会治理带来了全新的变革。然而，政务大模型的应用也伴随着各种潜在的风险，尤其在用户隐私、数据安全、内容监管等方面，涉及政务的特定要求和合规风险。本文将从政务特点出发，分析大模型在用户、应用、内容风控、数据、模型以及公网使用等方面的潜在风险。

本文围绕政务大模型与应用系统展开讨论安全风险及安全治理方法，通用大模型相关风险，如“可解释性差的风险、偏见歧视风险、鲁棒性风险等内容”等通用大模型风险问题，不在本文中做探讨。

2.1 数据安全风险

数据是大模型智慧的来源，通过获取外部数据及政务数据，进行数据处理，形成高质量训练数据集。数据处理过程，面临多种风险：数据违规收集、数据泄露、勒索风险、违规使用风险等风险；同时数据标注环节也至关重要，伴随着多重风险，影响模型的安全性和准确性。

- 数据的违规收集：政务大模型使用数据往往涉及公众的个人信息和国家机密，若这些数据的收集和使用不符合法律规定，将面临严重的法律后果。
- 违规使用数据：使用未授权或违规数据进行大模型训练或内容检索增强，违反相关规定。
- 数据泄露：政务大模型在研发和应用过程中，因数据处理不当、非授权访问、恶意攻击等问题，可能导致政务数据和个人信息的泄露。
- 勒索风险：大模型使用的政务数据若被黑客攻破，可能遭遇数据勒索或数据流失，直接威胁政府运作和社会稳定。
- 标注规则风险：数据标注规则不完善或模糊会导致标注人员理解不一致，进而影响数据的准确性和一致性。标注规则的不清晰还可能导致关键信

息被忽略或误标，影响模型的性能与决策能力，尤其是在政务领域，错误的数据标注可能导致政策误判或公共服务偏差。

- 标注人员风险：标注人员的安全意识直接关系到数据的安全性。若标注人员缺乏安全意识，可能无意中泄露敏感数据，或在标注过程中引发信息泄露。此外，恶意标注人员可能通过窃取数据、投放恶意信息（数据投毒）或篡改标注内容，故意破坏数据的完整性和准确性。
- 标注数据质量风险：数据标注的准确性、一致性和完整性不足，会导致模型在实际应用中的表现不佳，产生偏差或错误判断。

2.2 训练语料安全风险

训练语料在模型训练中至关重要，它直接决定了模型的学习质量和生成内容的准确性。如果训练语料存在数据投毒或内容违规风险，不仅会误导模型学习错误的信息，还可能导致其生成不当或不合规的输出。

- 数据投毒：攻击者可能通过投放恶意或虚假数据干扰大模型的训练，导致模型输出错误或产生不当内容。
- 内容违规风险：训练语料的内容包含个人敏感信息、误导性信息等不当内容，导致模型生成违规或不合规的内容。

2.3 大模型使用安全风险

随着政务大模型的深入应用，其安全风险也日益凸显。大模型应用安全风险是政务大模型应用过程中不可忽视的关键问题，如使用未备案的基础模型，可能引发合规性问题；大模型在生成内容的过程中存在输出不当内容的风险，可能产生违背法律、社会伦理或政策导向的结果，带来舆论和社会风险；此外，政务大模型文件是核心产物，面临着模型泄露与篡改风险。

- 使用不合规的基础模型：使用不合规的基础模型可能导致违规操作，并面临法律和行政处罚。如：根据国家监管要求，政务大模型的基础模型应进行备案，确保其合法合规。
- 生成不当内容风险：政务大模型在生成文本或提供建议时，可能输出

带有误导性、不当或风险性内容。

- 模型泄露风险：政务大模型的模型权重、参数和架构一旦泄露，可能被不法分子利用进行攻击或伪造内容，甚至可以通过模型反向推测训练数据，甚至还原敏感信息。威胁政务大模型的正常运行。
- 模型篡改风险：政务大模型一旦被攻击者通过恶意手段进行未经授权的修改，会导致大模型生成错误或误导性信息，甚至故意篡改政策解读或数据分析结果。

2.4 应用安全风险

政务大模型的应用在开发和使用过程中，面临着一系列安全风险。如开源组件可能存在漏洞，代码安全风险及代码泄露问题，应用上线后面临的Web/API 攻击风险等问题，威胁整个政务大模型应用系统的安全性与稳定性。

- 代码安全风险：政务大模型代码的安全性直接影响系统的稳定性和可信性。如果代码开发中缺乏安全措施，可能引发数据篡改、系统崩溃等问题。
- 代码泄露：政务系统中包含大量涉及政务信息或民生服务的敏感内容，代码一旦泄露，可能造成重大安全风险。
- Web/API 攻击：大模型通常通过 Web 接口和 API 为政务系统提供服务，若防护措施不足，可能遭遇恶意攻击。针对政务系统的 API 攻击可能导致服务中断、数据泄露或功能失效。

2.5 软件供应链安全风险

政务大模型应用系统的软件供应链风险主要包括开源软件漏洞、恶意代码植入和供应商管理不善，供应链安全风险直接威胁政务大模型的安全使用。

- 开源软件漏洞：政务大模型应用中使用的开源软件可能存在未修复的漏洞，这些漏洞可能被恶意利用，导致安全隐患。
- 恶意代码植入：第三方组件或开发工具可能被植入恶意代码，危及系统安全。

- 供应商管理不善：供应商的安全能力差异也带来风险，如代码质量低、漏洞未及时修复等。
- 模型复用的缺陷传导风险：依托基础模型进行二次开发或微调，是常见的大模型应用建设模式，如果基础模型存在安全缺陷，将导致风险传导至下游模型。
- 恶意植入模型后门：攻击者通过在大模型的训练或部署阶段植入恶意代码或逻辑，使得模型在接收到特定的触发提示词时，执行未授权的操作，绕过正常的安全限制。这类后门攻击尤其危险，它可以让模型在表面上正常工作，但当特定条件满足时，模型会进入类似越狱状态，允许执行几乎任何操作。

2.6 生成内容风险

大模型通过多源异构调度发挥更大价值，大模型生成内容安全合规风险极其重要，如生成违背社会主义核心价值观的内容、侵犯知识产权、泄露个人信息、保护歧视内容等风险，将会对政务大模型造成致命影响。

- 内容违反社会主义核心价值观：政务大模型涉及政策解读、公众沟通等方面，生成的内容必须符合社会主义核心价值观，确保正确的政治导向和舆论引导。如果大模型生成的内容出现违背国家政策或不符合主流价值观的情况，可能影响社会稳定。
- 侵犯知识产权：政务大模型在生成政策文件、解读报告时，可能无意间使用了受版权保护的文本或资料，导致知识产权纠纷。
- 泄露个人信息：政务大模型处理大量个人信息，如果在生成内容时无意间泄露个人身份信息或其他敏感数据，将严重违反数据保护法律法规。
- 包含歧视内容：大模型可能由于训练数据的偏差，生成包含性别、种族或宗教等方面歧视的内容。

2.7 大模型自身风险

政务大模型自身面临着多方面的安全风险，特别是在涉及公众服务和敏感

信息的政务场景中，大模型自身安全风险可能会对大模型的安全性和稳定性带来重大影响。

- 提示注入攻击：恶意用户可能通过特定提示引导政务大模型生成错误或有害内容，导致政府发布不当信息。
- 拒绝服务攻击：通过向政务大模型发送大量无效或恶意请求，使得系统资源被耗尽，导致模型响应缓慢或完全无法工作。一旦政务大模型因为拒绝服务攻击而无法及时处理请求，可能会导致政务服务中断，影响政府部门与公众的沟通效率，甚至拖延重要的决策流程。
- 提示词泄露：政务大模型中可能包含特定的提示词，这些提示词用于触发某些关键功能或提供额外的上下文信息。如果这些提示词被泄露，攻击者可能利用这些信息访问模型的内部机制或绕过安全限制，获取敏感数据。
- 通用越狱漏洞：攻击者通过利用模型中的安全漏洞或设计缺陷，试图突破模型原有的限制。在政务大模型场景下，攻击者可能通过复杂的交互引导模型生成或揭露本不该公开的机密信息，从而威胁到国家安全或重要的政务数据隐私。

2.8 总结

政务大模型作为推动政府数字化转型的关键技术，具有广泛的应用前景，但其潜在风险也不容忽视。为确保大模型的安全、合规使用，政府机构必须在技术开发和应用过程中加强风险管理，落实各类防护措施，确保大模型为政务服务和社会治理提供可靠、稳定的支持。

数据安全风险	➢ 数据违规收集 ➢ 违规使用数据 ➢ 数据泄露 ➢ 勒索风险 ➢ 标注规则风险 ➢ 标注人员风险 ➢ 标注数据质量风险
训练语料安全风险	➢ 数据投毒 ➢ 内容违规风险
大模型使用安全风险	➢ 使用未备案基础模型 ➢ 生成不当内容风险 ➢ 模型泄露风险 ➢ 模型篡改风险
应用安全风险	➢ 开源组件漏洞 ➢ 代码安全风险 ➢ 代码泄露 ➢ Web/API 攻击
软件供应链安全风险	➢ 开源软件漏洞 ➢ 恶意代码植入 ➢ 供应商管理不善 ➢ 模型复用的缺陷传导风险 ➢ 恶意植入模型后门
生成内容风险	➢ 内容违反社会主义核心价值观 ➢ 侵犯知识产权 ➢ 泄露个人信息 ➢ 包含歧视内容
大模型自身风险	➢ 提示注入攻击 ➢ 拒绝服务攻击 ➢ 提示词泄露 ➢ 通用越狱漏洞

表 1：政务大模型安全风险总结

3 政务大模型安全治理框架

探索包容、审慎的大模型安全管理模式，结合大模型应用建设场景，制定政务大模型安全治理框架。大模型安全工作应结合模型应用的实际场景，按照由易到难、由低风险到高风险逐步推进的思路，选取负面影响较小、涉及重要的数据较少的场景，进行先行先试，逐步探索推进，确保大模型应用的安全性与有效性。

图 3：政务大模型安全治理框架

随着政务大模型在政府管理与公共服务中的广泛应用，其安全治理成为一个关键问题。政务大模型不仅处理海量的政务数据，还需要遵循国家的法律法规，确保数据隐私和安全。因此，构建一个全面的安全治理框架，能够有效确保大模型的安全、合规应用，避免潜在风险。基于合规要求，本文将从安全机制、基础安全措施、数据安全、大模型安全、开发安全及运行安全等方面，构建政务大模型的安全治理框架。

3.1 合规要求

在政务大模型的安全治理框架中，合规是首要原则，涉及多部法律法规、规章制度的遵循，如《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》《生成式人工智能服务安全基本要求》等。确保政务大模型的合规运行，是安全治理的基础。

3.2 安全机制

为了保障政务大模型在实际应用中的安全性和可控性，需建立一套完善安全机制，确保大模型技术在政务领域的安全、合规和高效应用。这些机制不仅要应对技术本身的风险，还需统筹市区协同管理，分级分类对待不同应用场景，并全程监控各个环节的安全性。

3.2.1 组织建设

为了保障政务大模型的安全运行，安全组织结构起到至关重要的作用。该组织结构涵盖了统筹、业务、支撑和监管四个方面：

1. 统筹方：由政数局负责整体安全战略的制定与统筹，协调各方资源，确保政务大模型的安全目标能够达成。
2. 业务方：各委办局及区县负责具体政务场景中的大模型应用，确保其在使用过程中符合政务要求和安全标准。
3. 支撑方：包括模型商、数据商、应用开发商和第三方安全机构，为大模型的开发、数据管理和安全技术提供支持。支撑方需按照政府制定的安全标准，确保技术产品和服务的合规性。
4. 监管方：由政数局和网信办共同监管，确保政务大模型的运行符合国家相关法规及安全要求，承担监督和审查责任。

3.2.2 制度与规范

1. 市区两级协同联动机制，确保不同区县和委办局之间的信息共享和协同应对。
2. 大模型应用分类分级管理制度对不同类型的模型进行分级，确保针对不同应用场景采取适当的管理措施。

3. 大模型应用全过程安全管控制度贯穿应用的各个阶段，从模型开发、测试、部署到使用及维护，确保安全风险得到全方位监控和防范，保障数据安全和政务服务的稳定性。

3.2.3 安全测试与评估

1. 数据集安全评估：确保用于训练和运行大模型的数据集来源合法合规，并进行严格的数据隐私保护和清洗审核，避免模型从源头引入安全风险或敏感信息泄露。
2. 生成内容安全测评：对模型生成的文本、图像等内容进行合规性和伦理审查，确保输出不涉及敏感、虚假或不当信息，防止传播错误信息或违反社会主义核心价值观。
3. 模型及应用安全测评：对模型本身及其在具体应用场景中的安全性进行全面评估，保障大模型供应链、大模型接口、大模型自身防攻击能力等进行安全测评。
4. 安全措施评估：评估模型部署过程中的各类安全措施，包括数据加密、身份验证、防护机制等，确保有效防止数据泄露、攻击入侵等安全威胁。
5. 定期开展监测与抽查：通过定期安全审查和动态监控，及时发现和解决潜在的安全隐患，并随机抽查模型的表现，保障其在长期运行中的安全性和可靠性。

3.3 安全保障

安全技术保障是政务大模型安全治理框架的核心，它涵盖了从基础安全措施、数据集安全、大模型开发安全到运行安全等方面。

3.3.1 基础安全保障

1. 纵深防御：在政务云和网络层面完善纵深防御体系，通过多层次的安全防护措施阻止外部攻击，确保大模型的基础设施安全。
2. 身份管理与授权：通过严格的身份验证与访问控制机制，确保只有经过

授权的用户才能访问大模型和相关数据，防止权限滥用或外部恶意访问。

3.3.2 数据安全

1. 数据来源合规：所有数据的来源必须符合法律法规，确保数据合法、合规地用于政务大模型的训练与应用。
2. 内容安全合规：用于大模型训练的和知识增强的数据的内容必须符合法律法规，确保用于大模型训练和知识增强的数据内容合规。
3. 敏感数据识别过滤：对输入的敏感数据进行自动识别和过滤，确保隐私数据不会被误用于大模型训练或输出，防止泄漏风险。
4. 训练数据标注安全：对标注人员进行严格的访问控制和管理，制定完善的标注规范，以确保高质量的数据标注成果，保障顺利标注数据的保密性和合规性，避免敏感信息外泄。
5. 数据分类分级与安全保护：根据数据的重要性和敏感性进行分类分级管理，不同级别的数据需采取相应的安全保护措施。
6. 数据访问控制：严格控制对数据的访问权限，确保只有经过授权的用户或系统能够访问敏感数据，防止未经授权的访问或篡改。

3.3.3 大模型安全

1. 模型训练安全：对大模型的训练过程进行安全管控。在训练过程中，需采取严格的安全措施以防止数据泄露和非法访问。同时确保数据来源内容的安全合规，避免训练数据投毒等风险。
2. 模型资产保护：对模型的算法、参数、结构以及相关文档进行严格的保密和安全防护。对模型实施多重安全防护措施，包括访问权限控制、模型加密存储以及防泄露技术对模型资产进行保护，防止模型被非法复制、篡改或盗用。
3. 模型安全评测：对训练完成或拟上线的政务大模型进行安全测试与评估，包括语料安全评估，生成内容安全评估，模型安全评测，安全措施评估。
4. 模型登记备案：应按照网信办相关管理要求，开展上线备案或登记工作。

5. 模型分类分级管理：根据大模型的功能、使用范围和敏感程度，将模型划分为不同的等级，并针对各个级别设定相应的安全管理要求。分类分级管理机制能够根据实际需要灵活调整管理力度，确保每个模型的安全性都能得到充分保障。

3.3.4 大模型应用系统开发安全

1. 需求阶段：在模型应用开发的需求分析阶段，同步进行安全需求分析，并设计相应的安全方案，确保项目从一开始就考虑到安全因素。
2. 供应链安全：确保政务大模型应用系统的软件供应链安全，确保软件组件和工具来源可信，防止恶意代码和供应链攻击。供应商需符合安全标准并定期审查，严格控制软件版本管理，避免未经授权的修改，保障系统的稳定和安全运行。
3. 数据准备阶段：在数据收集、清洗和处理过程中，确保数据安全，包括数据加密、脱敏处理等。
4. 模型适配阶段：在适配过程中，必须确保所用的大模型和数据符合安全合规要求，防止数据泄露和模型被恶意操控。实施严格大模型的权限管理和访问控制。
5. 应用开发阶段：确保政务大模型在应用开发中的代码和系统架构安全，通过审查与测试，防范安全漏洞。
6. 上线部署与评审：在大模型上线前进行严格的安全评审，并对模型进行分类分级管控。上线后，持续对模型运行进行监控，防止潜在风险。

3.3.5 运行安全与内容风控

大模型的运行安全涉及内容生成和应用层面的多重保障，确保生成的内容符合政府规定，且系统运行稳定可靠。

1. 生成内容风控：
   - 输入内容过滤：确保大模型在接收到输入前经过严格的内容审查，过滤掉潜在的敏感词、违法内容或其他不合规信息，防止恶意输入

导致的不良输出。

- 输出内容审核：大模型的输出内容必须经过人工或自动化的审核机制，确保生成的内容符合社会主义核心价值观、不涉及敏感话题，避免因不当内容带来的社会风险。

2. 应用安全：
   - Web 安全防护：政务大模型通过 Web 平台提供服务时，需采取完备的 Web 安全措施，防止跨站脚本、SQL 注入等网络攻击。
   - API 安全防护：大模型与政务系统通过 API 连接，需确保 API 调用的安全性，防止未授权访问或 API 滥用。
   - 应用访问控制：严格限制对大模型应用的访问权限，确保只有经过授权的政务人员或系统可以访问和使用大模型。
   - 个人信息保护：政务大模型应用在处理个人信息时，需严格遵循相关法律法规，采取加密、脱敏和安全存储等措施，确保个人信息安全。

3.3.6 运行监测

为了及时发现和应对潜在的安全风险，政务大模型必须建立完善的运行监测与事件处置机制：

1. 大模型监管沙盒体系：通过搭建监管沙盒环境，测试大模型在实际场景中的运行表现，识别潜在的安全问题和内容风险。这有助于在模型正式上线前发现并修复潜在漏洞，减少风险暴露。
2. 风险监测与事件处置：通过实时监控系统，对大模型的运行状态、数据流动、应用访问等情况进行监测。一旦发现异常行为或安全事件，能够及时启动应急响应机制，采取适当的措施进行处置，防止事态扩大。

3.4 总结

政务大模型的安全治理框架是保障其安全、合规运行的基础。在这一框架中，合规要求贯穿始终，各方在安全组织结构中的协同合作至关重要，同时通

过严密的安全技术保障和运行监测，确保大模型的安全性、可靠性和稳定性。最终，政务大模型的安全治理框架能够帮助政府机构在数字化转型过程中，有效应对各类安全风险，为国家治理和公共服务提供坚实的技术支持。

4 政务大模型风险控制关键举措

4.1 大模型分类分级安全管控

4.1.1 分类分级安全管理制度

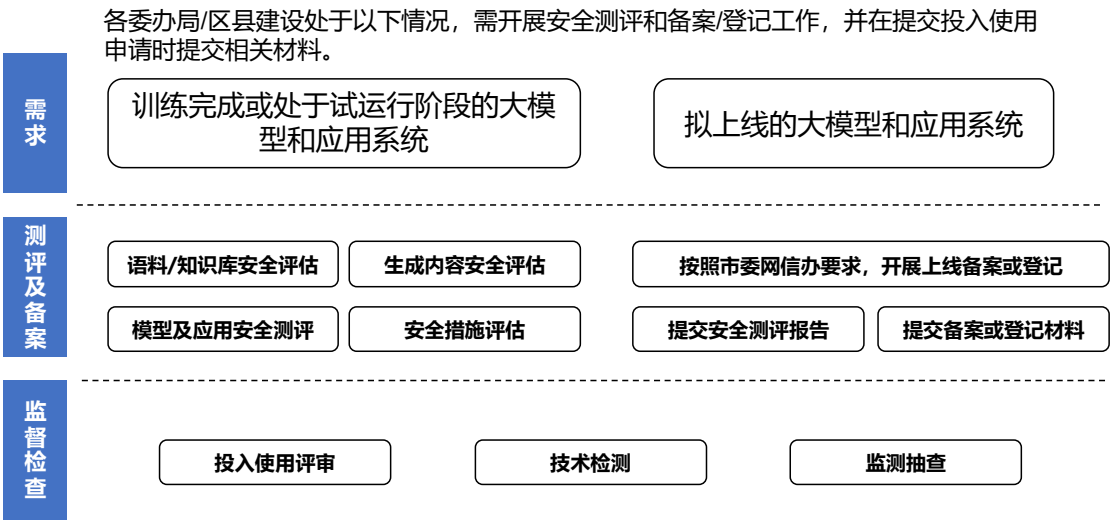
针对不同类型的大模型，应根据其用途、使用范围、影响力及风险水平，制定科学合理的分类分级安全机制，由低风险到高风险逐步探索推进。对低风险领域的大模型，进行先行先试；高风险领域的大模型，必须采取更加严格的安全措施和审查机制，确保其在使用过程中不会对社会产生负面影响。通过实施分级管理和风险评估，确保高风险模型在使用前经过严格的审查和评估，确保其安全性和合规性，防范潜在风险，维护社会稳定和公共利益。

4.1.2 分类分级安全管控与保护措施

探索并建立大模型分类分级安全管控与保护措施。依据大模型的分类分级评定结果，制定并部署切实有效的安全措施，包括对大模型网络进行必要的隔离，以减少潜在风险。应对身份授权和访问权限进行动态细粒度控制，以确保只有经过授权的用户能够访问相关资源。通过先进的技术手段，确保应用开发过程的安全性和应用运行的安全防护，从而全面保障大模型在各个应用场景中的安全性。

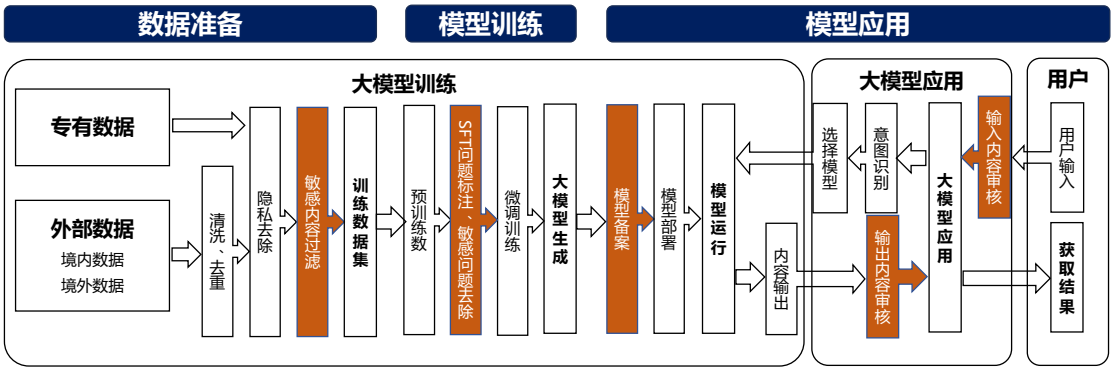
4.2 大模型与应用系统安全测试与评估

对政务大模型的合规评估进行事前、事中、事后的全流程管理。事前，训练完成或处于试运行阶段的大模型，应由第三方专业机构按照大模型评测标准系统全面的安全测试和评估；拟上线大模型应按照市委网信办相关管理要求，开展上线备案或登记。事中，大模型项目通过竣工验收后，向市政数局提交投入使用申请时，需补充大模型项目安全测评报告、市委网信办备案/登记证明等材料。事后，市政数局定期组织开展技术检测和监测抽查，防范不良信息出现，提升内容合规性，促进模型迭代优化。



4.3 内容风险控制

建立多层次的内容审核体系，综合运用自动化检测与人工审核机制，全面过滤风险内容。加强训练数据的筛选与标注，剔除潜在风险的数据，以确保数据源的可靠性和准确性。部署包括敏感词过滤、上下文分析及多层次审核等技术措施，对生成内容进行风险内容识别与过滤。实施实时监测与反馈机制，以动态跟踪和调整生成内容，确保其在使用过程中的安全性和合规性。



4.4 数据安全风险控制

以场景牵引、需求导向为原则，系统探索政务大模型业务场景下的数据安全保护和隐私保护措施。综合运用包括数据加密、数据脱敏、去标识化、隐私计算及数据流转监测等在内的多种手段，建立完善的数据安全保护体系，确保政务数据在存储、传输和使用过程中的安全性。同时，积极探索区块链技术的应用，以增强政务大模型训练数据的安全性和透明性，从而提升数据管理的可

信度，确保系统的整体安全性和合规性。

4.5 监管沙盒机制

建立大模型安全监管沙盒体系，对政务大模型训练数据的流动进行全方位、实时的监控和管理。及时识别和预防数据滥用、非法访问及未经授权的传输行为，确保数据安全。强化对潜在风险的预警和控制，进一步提升政务数据的安全性与合规性，为大模型在政务领域的应用提供坚实保障。

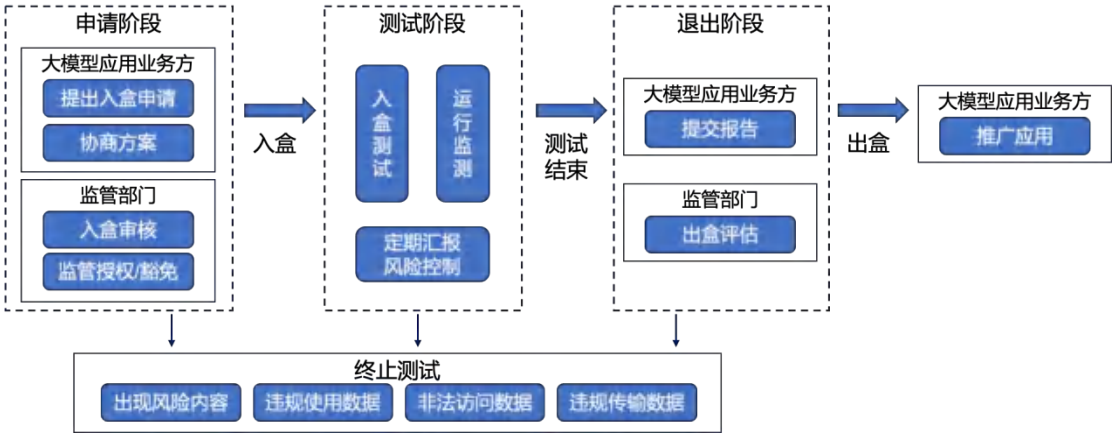


图 4：政务大模型监管沙盒机制

5 政务大模型安全治理展望

大模型快速演进，其推理能力不断增强，幻觉问题不断优化。围绕大模型的相关技术也随着发展，例如出现多个智能体协作，调用不同能力的大模型，共同完成任务。

5.1 大模型监管逐渐规范，行业细化规范

随着大模型的快速发展，不断有新的法律法规和标准出现，来规范大模型的有序安全发展。政务大模型也需要遵守这些法律规范，以及强制性标准的合规要求。2023 年 5 月，国家网信办联合多个部门发布《生成式人工智能服务管理暂行办法》，反映了我国政府对生成式人工智能技术的重视，以及对其潜在影响的审慎态度。2024 年 9 月，《人工智能治理框架》发布，对人工智能技术发展和应用安全做出规范要求。

政务大模型需要遵守通用大模型以及人工智能方面的法律法规。在数据政府领域，针对政务大模型，可能会出现专门的行业监管标准，指导政务大模型应用从开发到部署的安全，确保政务大模型的安全性和可靠性。

政务大模型涉及敏感的政务数据，其运行直接影响公共利益和国家安全。与商业大模型相比，政务大模型面临更严格的要求，对政务大模型的监管，尤其是训练数据，需要有更系统化，更细化的政务大模型规范来指导。有些特殊场景，需要在政务内网部署，数据和模型不出域。保护公民隐私和数据安全，确保政务大模型的公平性和透明度，增强公众对政务大模型系统的信任。

5.2 政务大模型安全生态建立，安全产品和服务不断完善

数字政府对大模型的陆续采用，安全风险会逐渐得到关注，这会推动相关安全服务和产品市场发展。大模型提供方，数据商，应用开发商，安全厂商等，针对市场需求，开发和创新满足市场需求的产品和服务，在数据安全、模型安全、应用安全、生成内容安全各环节满足政务大模型安全可控的需求。

大模型生态建立，安全服务和产品不断完善。在政务大模型安全服务方面，会出现合规咨询服务，安全测试与评估服务，安全培训服务，应急响应服务，持续监测服务等。在产品方面，除了传统的网络安全和数据安全之外，会出现针对大模型输入输出的内容过滤产品，模型保护安全产品，大模型伦理审核产品等。

以大模型为代表的生成式人工智能技术，正在持续演进，政务大模型也随之迭代，不断推动数字政府向智能化迈进。政务大模型的安全可控，对政府开展政务服务和城市治理，造福社会具有重要意义。在政务大模型的安全治理上，需要政府、企业、研究机构、监管机构等各方协作，形成良性互动生态体系，构建一个安全、可靠、有序的环境，充分发挥政务大模型在提升政府服务效能、推动社会进步方面的积极作用。

附 作者介绍

奇安信数据安全总体部

奇安信数据安全总体部，负责数据安全市场洞察及战略制定，参与数据安全政策法规、标准规范等制定和支撑，在数据要素和人工智能领域开展数据安全相关研究。面向客户提供数据安全的咨询规划、治理服务、方案设计等。

奇安信人工智能研究院

奇安信人工智能研究院，负责研究和开发人工智能相关技术及其在网络安全中的应用，包括但不限于人工智能基础算力平台，大模型训练推理平台，AI 智能安全平台，QDE 人工智能杀毒引擎，QGPT 网络安全大模型等。

奇安信观星实验室

奇安信观星实验室是在实战攻防演习中扮演重要角色、擅长组织实施渗透攻击的团队。实验室下面设有多支攻击队，团队成员大多来自攻防渗透研究出身的高级技术专家和渗透工程师，均有多次参与省部级实网攻防演习的经历。实验室平均全年参与全国范围内 200 余场实战攻防演习活动。在所有行业化的实战攻防演习排名中均名列前茅。实验室研发出多套实用技战法和配套工具。在 Web 攻防、社工渗透、内网渗透、模拟 APT 攻击等方面，技术实力扎实，技战法灵活，实战能力受到业内高度认可。

奇安信网络安全部

奇安信网络安全部承担公司的 IT 基础设施和网络安全体系的建设、治理与运营；为员工与业务提供安全、高效的 IT 服务；保障产品安全。同时，整合团队内网络、运维和安全等领域的技术专家，全面支持前线需求。


奇安信
新 一 代 网 络 安 全 领 军 者