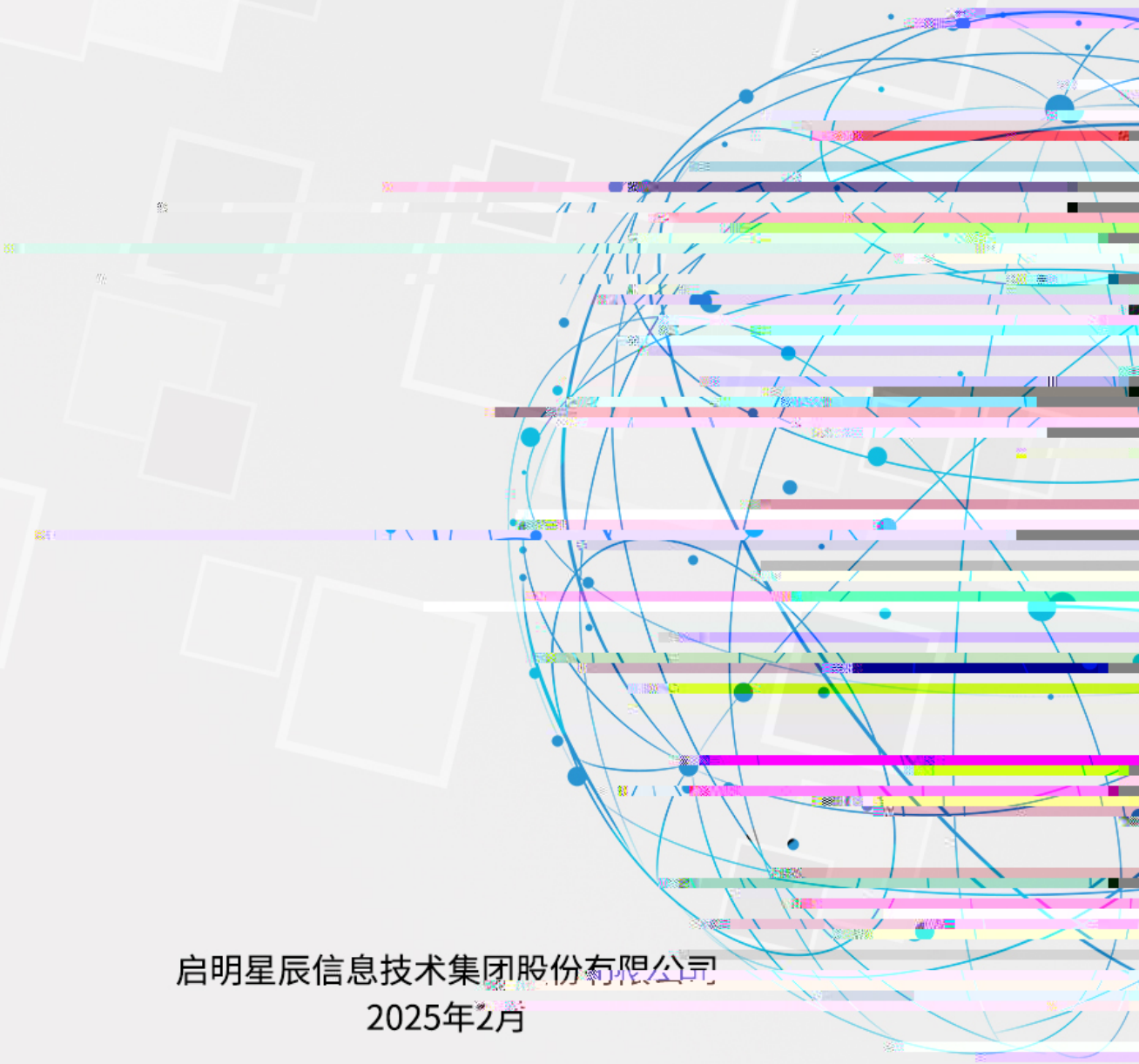
启明星辰

# AI就绪的大模型身份与访问管理

# AI-R-IAM v1.0

AI-Ready Identity and Access Management



启明星辰信息技术集团股份有限公司
2025年2月

## 版权申明

北京启明星辰信息安全技术有限公司版权所有，并保留对本文档及本声明的最终解释权

和修改权。

本文档中出现的任何文字叙述、文档格式、插图、照片、方法、过程等内容，除另有特

北京 别注明外，其著作权或其他相关权利均属于北京启明星辰信息安全技术有限公司。未经

任何 启明星辰信息安全技术有限公司书面同意，任何人不得以任何方式或形式对本手册内的

用途。 部分进行复制、摘录、备份、修改、传播、翻译成其它语言、将其全部或部分用于商业用

## 免责申明

档依据现有信息制作，其内容如有更改，如不另行通知，

本文

漏、不准确、或错误导致

确可靠，但北京启明星辰信息安全技术有限公司不对本文档中的遗

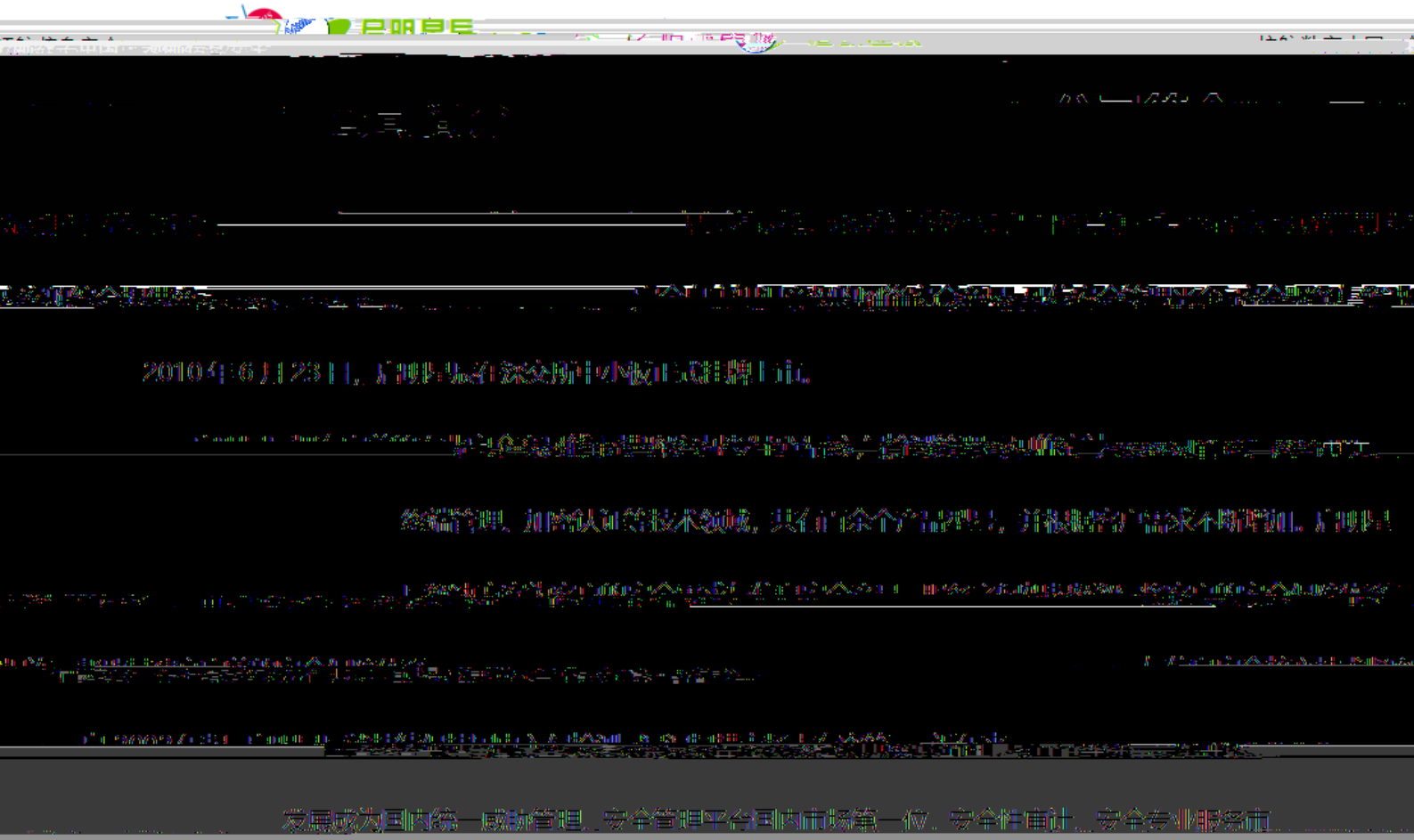
的损失和损害承担责任。

## 信息反馈

如有任何宝贵意见，请反馈：


100193 电话：010-82779088

传真：010-82779000



家分支机构，拥有覆盖全国的渠道和

关怀与鼓励。2000 年 1 月，江泽民、

司；2003 年 1 月，胡锦涛总书记亲

局内重点软件企业，国家火炬计划软

拥有最高级别的涉及国家秘密的计算

场领导者。目前，公司在全国各省市自治区设立三十多

售后服务体系。

长期以来，启明星辰公司得到了党和国家领导人的

李岚清、曾庆红等党和国家领导人亲切视察启明星辰公

切接见了启明星辰公司 CEO 严望佳博士。

凭借多年来的潜心研发，启明星辰获得国家规划布

件产业优秀企业，中国电子政务 IT100 强等荣誉，及拥

机信息系统集成资质证书。

示范工程，国家科技部 863 计划、国家科技支撑计划等国家级科研项目近百项。创造了百

余项专利和软件著作权，参与制订国家及行业网络安全标准，填补了我国信息安全科研领域

全产品和最佳实践服务，帮助客户全面提升其 IT 基础设施的安全性和生产效能，为打造和

# 2 大模型的发展与风险

## 2.1 新质生产力“大模型”

力。从 ChatGPT，近年来，大模型技术取得了突破性进展，成为人工智能领域的核心驱动力

自然语言处理、再到百度文心一言、阿里通义千问、九天大模型、DeepSeek 等，大模型在

文本生成、智能问答、图像 等学习到丰富的语言知识、语义理解和编辑推理能力，从而实现

。人模型发展和应用分为四个阶段：

国）：

年 ChatGPT 上线人和“能答”处理能力比较突出；用户

模型嵌入人和应

）

（1）国外：

（2）国内：

备。模型发展提供技术储

（1）国外：OpenAI 公司等公司不断升级大模型，性能显著提升，开源大模

天”人工智能基座，包含万卡级智算集群、千亿多模态大模型、汇聚百大要素的生态了“九天

模型在细分领域崭露头角。

模型不断迭代，众多初创企业也凭借特色大模

- **大规模应用期（2025 年 1 月-）**

助力疾病诊断、风险评估、个性

**（1）国外**：大模型深入医疗、金融、教育等多行业

化学习等。OpenAI 还在探索新应用领域。

、高效能以及开源优势，

**（2）国内**：DeepSeek 凭借创新算法和架构，展现出低成本

## 2.2 大模型应用中安全挑战

ent 构

根据埃森哲的《2025 年技术愿景》调查，78%的高管同意在未来必须为 AI Ag

Identity，NHI），都暴露出诸多安全问题。

[redacted]

非法访问权限。

[redacted]

和操作敏感资源，造成数据泄露或系统破坏。

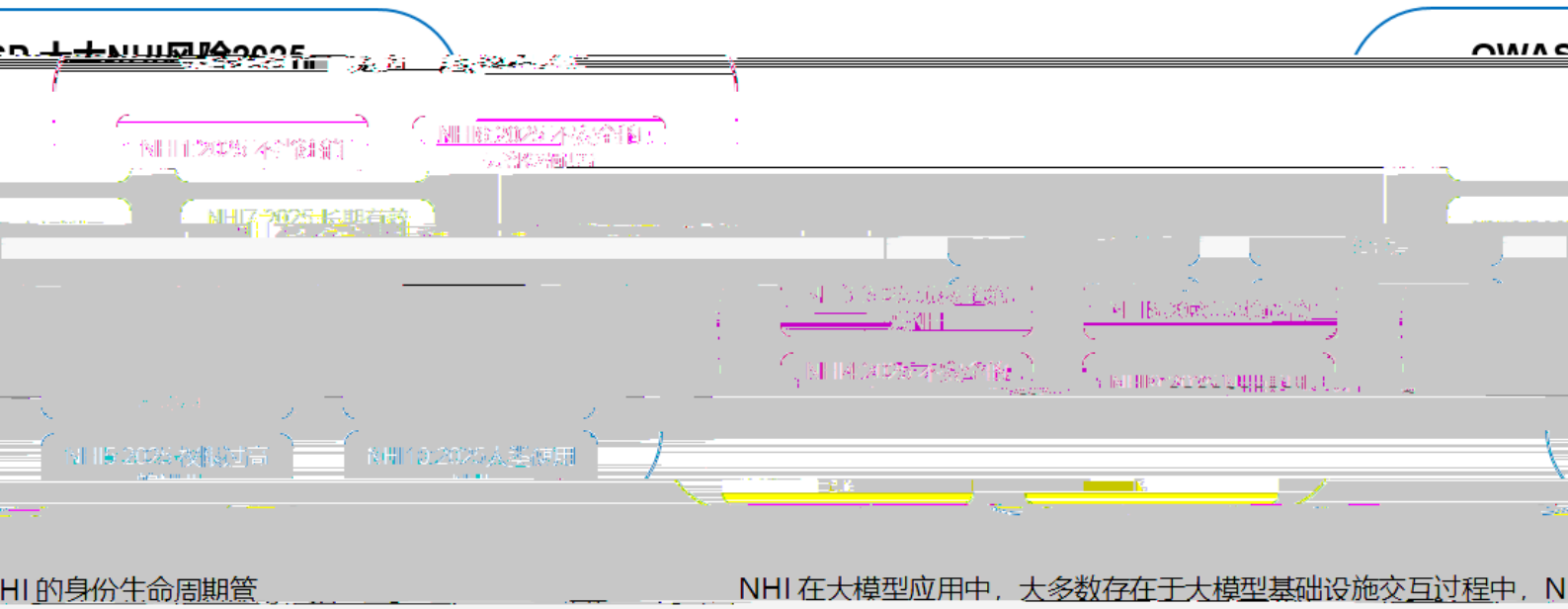
[redacted]

系统遭受黑客攻击，这些数据可能被窃取，导致用户隐私泄露，进而可能被用于精准诈骗、定向广告骚扰等。

**3、非人类身份（NHI）安全问题：**非人类身份定义为企业技术内的应用程序、服务或机器等实体绑定的数字身份，这些包括机器人、API 密钥、服务帐号、OAuth 令牌、云服

[redacted]

2025 年 2 月发布《十大 NHI 风险 2025》，如下图所示：



HI 的身份生命周期管

与问题：

程弱、账号使用信息不可见，许多

NHI 在大模型应用中，大多数存在于大模型基础设施交互过程中，N

理薄弱，存在诸多安全风险

被内外部威胁者发现。

(2) 影子账号：大模型应用迭代快，加之生命周期流

NHI 账号不活跃，增加攻击面。

管理者对安全维护和问题补救很关键。

在每个环境中都有相同的密码，从而增加了横向移动的风险。

的逻辑 NHI

杂，难以掌握所有依赖关系。

### 2.2.2 数据安全挑战

大模型应用在数据安全方面呈现如下挑战：

(1) 用户提示词输入风险：用户提示词可能包含敏感信息，若大模型系统对输入提

同的验证和过滤机制不完善，敏感信息可能在输入环节直接暴露。在输出结果时，若对输 示词

的提示词用于医疗大模型进行诊断辅助，输出结果未做敏处理，可能导致个人健康隐私

调用过程中，若身份认证和授权机制存在漏洞，不法分 (2) 大模型 API 调用风险：API 调

敏感数据或进行恶意操作。同时，API 接口若缺乏有效 子可能冒用合法身份调用 API，获取

非法获取金融大模型中的用户资产数据。

(3) RAG 知识库查询风险：RAG 知识库整合大量数据，若查询权限控制不当，用户可能

(4) 数据投毒风险：在大模型训练阶段，攻击者向训练数据中注入恶意数据，这种

求，若个人在数据收集、使用、存储和共享等环节不符合相关法规，如

化风险，未经用户充分授权，可能面临法

、数据风险、业务风险、设备

可信环境基于大数据安全能力，实现网络风险、行为风险、
风险的全链路安全风险审计。

全设备、系统资源、应用资源、

访问可信与数据可信作为一体两面，将防控范围扩展到安全

更好的安全性，搭配一体化安全机制的融

组合性和可扩展性，使企业能够以更少的资源实现

大模型作为新质生产力，在政府、企业、个人中应用越来越深入，需要针对其身份、

随着

、数据安全等方面的需求，建立一套基于大模型技术特点和业务场景的一体化全程

访问管理

字身份基座，即 AI 就绪的大模型身份与访问管理。

可信的数

# 4 AI 就绪的大模型身份与访问管理

如

AI 就绪的大模型身份与访问管理 (以下简称: AI-R-IAM) 是为大规模人工智能模型 (

化全程可信”保护体

构建一体化全程可信能力，构建大模型安全从“单点可控”迈向“一体化

系，同时为其它安全能力提供身份管理属性支撑。



先进的身份认证机制，确保只有经过授权的用户和系统能够访问大模型。系统不仅支持传统的用户名和密码验证，多因素认证，还引入基于中国移动号卡特性的实名认证能力，有效防止未经授权的访问和潜在的安全威胁。

- 可信访问权限管理

AI-R-IAM 针对多个大模型访问、AI 接口访问、RAG 访问提供了集中的精细化的权限管理，能够根据用户的角色和职责，为其分配最小必要的访问权限，确保用户只能访问其工作所需的数据和功能。这种权限管理不仅提高了系统的安全性，还增强了用户的操作体验。

- 可信审计管理

AI-R-IAM 对大模型使用过程和大模型内部组件业务调用过程进行安全审计和实时监控，记录用户的所有操作行为，包括用户登录时间、IP 地址、访问的大模型资源、执行的操作等信息，实时监测用户和组件的访问行为，及时发现异常风险。

AI-R-IAM 通过构建一体化的全程可信能力，能够有效应对当前大模型面临的安全挑战，为其它智能大模型的设备，提供身份认证、权限、审计等方面的基础。为人工智能技术的安全发展奠定了坚实的基础。

## 4.1 可信身份治理




AI-R-IAM 支撑基于大模型特性的多维度身份和属性的治理和管理，AI-R-IAM 的“身

份ID”超越了传统身份概念的边界，聚合人类身份，AI Identity以及非人类身份（NHI），

程序建立集中一体的身份标识，对身份和属性进行全生命周期

针对用户、智能体、API 应用

层支撑。

的管理，是数字世界秩序的底

身份认证机制，同时引入中国移动基于号卡特性的实名认证能

AI-R-IAM 采用了先进的

和潜在的安全威胁。

，有效防止未经授权的访问

和 AI Identity 纳入统一治理范畴，助力企业实现数智化转型，有效降低安全风险，提升合

规性。在具体流程如下：

（1）用户身份在入职、转岗、离职时，进行信息收集验证、权限调整及账号注销等操作；

（2）智能体身份在创建时生成唯一ID并设定权限，修改时变更信息，动态调整权限，

销毁时注销身份、回收资源并处理数据；

（3）应用程序身份在创建时注册认证，授予权限，修改时更新信息，变更权限，销毁时

删除身份、销毁凭证并清理资源。

## 4.2 可信权限管理

| 算法工程师 | ✓ | ✓ | ✓ | × | × | × |
|---|---|---|---|---|---|---|
| 业务用户 | × | 受限调用 | × | × | × | × |

- AI 接口权限管理

包括大模型接口策略、接口密钥、接口连接调用、接口计费、接口传输数据安全、接口

管理 ... 用户名单等进行权限参数管

检索增强生成）权限隔离 ● RAG（检

控制：不同部门只能访问其业务相关的知识库片段； 文档级权限控

[redacted]

和涉及竞对信息的段 实时内容过滤：结合语义分析引擎，自动屏蔽无权限的检索结果（如

落）



## 4.4 可信行为审计




- 多维数据采集

- 云合一数

分发，进行数据的准确

- 可信决策

- 数据审核分析

构建多维分析模型，反映大模型用户的特征、行为习惯、个人便好，生成用户画像、模型画像、数据画像等，并根据风险画像评分，智能量化风险水平。

- 数据知识图谱

通过数据多源融合、关系排序、深度去重等深度治理及分析，实现不同实体之间的关联构建，提供从“关系”的角度分析大模型风险的能力。

- 数据深度分析

大模型数据全生命周期各个环节进行合规风险分析，重点数据审计，敏感数据和异常场景分析等，对大模型安全风险进行评估，识别已知风险和未知风险。

- 智能追踪溯源

从数据输入到模型输出每个阶段的智能追踪溯源，采用路径计算、路径回溯等技术提供可疑日志链路分析，层层递进分析，实现风险智能追踪。

可疑 IP->访问数据分布->可疑路径->可疑

# 5 应用场景

## 5.1 大模型应用场景数据流

AI-R-IAM AI 就绪的大模型身份与访问管理为大模型的访问、使用、调用、训练等各

…数据能力。相关场景包括用户…种场景提供可信身份、可信认证、可信权限、可信审计、可…

…型、公域大模型场景；业务系统调用私域大模型、RAG 知识库、私域大模型场…户访问私域大模型…

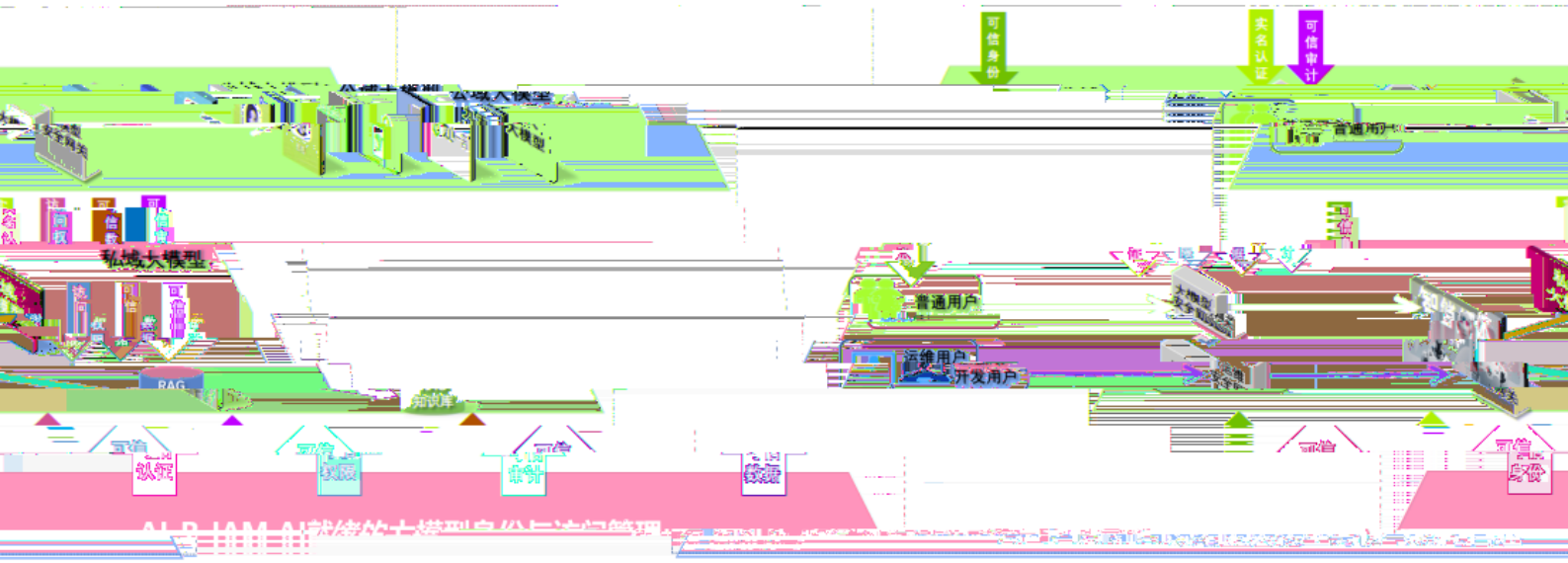
…训练场景、大模型数据运维场景等。…景；大模型数据…




## 5.2 场景一：用户访问大模型应用场景

AI-R-IAM AI 就绪的大模型身份与访问管理能够为用户访问大模型应用场景提供可信

…户访问公域大模型、用户访问私域大模型、运营维护人员访问私域大模型等。




### 1. 普通用户访问公域大模型

大模型应用。这些大模型应用涵盖多种类型的服

或者平板等，去访问那些部署在公有云上的大

非常简便。比如，当一个企业员工想要利用文本生

可信身份认证方面，构建了一套严谨的身份验证机制。当用户首次访问大模型应用时，

保护用户的数据不被未授权访问。

状况。它会记录下所有用户的操

在安全审计方面，AI-R-IAM 持续监控整个系统的运行

了大量不符合常规模式的数据访问请求，就可以触发警报提示管理员进行进一步调查，从而

有效地维护系统的稳定性和安全性。

### 2. 普通用户访问私域大模型

普通用

在用户身份可信认证基础上，私域场景下需要对不同角色进行更加细致的权限划分。企

数据汇总；部门主管则只需要关注本部门的数据，

解整个企业的运营状况，包括各个部门的数

的调整；普通员工的权限最为有限，只能访问与

并且能够对本部门员工的权限进行一定程度

性和准确性，以满足不同角色的需求同时保障数据安全。

### 3. 运营维护人员访问私域大模型

运营维护人员访问私域大模型应用服务器，负责日常维护和故障排查。由于运维人员权

大模型应用服务器。然而，如果是在非工作时间或者从企业外部网络尝试访问时，就需要触

发更加严格的验证机制。

权限配置等。一旦识别出高危操作，就需要立即启动阻断机制。这种阻断机制可以分为多个

层次。在第一层次是警告提示，当运维人员尝试执行高危操作时，系统会弹出明显的警告信

在私域大模型运维中，金库可以被视为一个特殊的区域或容器，其中存放着最为重要的数据和配置信息，如模型的核心算法代码、高度机密的训练数据、关键的系统配置文件等。

自的认证信息（如指纹、虹膜扫描等生物特征认证），才能打开金库。同时，所有对金库的

## 5.3 场景二：应用访问大模型应用场景



## 2. 业务权限管控

业务系统在访问公域、私域大模型时，如果没有基于最小权限原则来分配访问权限，如

某些功能可能只需要读取权限，却被赋予了写入或管理权限。由于权限划分不清晰，可能导

致内部人员误操作或恶意操作，进而影响系统整体安全。



整、合规性检查不足、数据隐私泄漏等安全问题。




过 AI-R-IAM 能力解决如下问题：

1. **用户到智能体全链路身份可信**

Identity）中统一身份ID进行绑定管理，需要在设计层面明确，如果智能体身份被了其

证。

项操作，系统可追溯到是谁让智能体进行操作，方便界定相应责任

码及模型紧密绑定，运用

为了有效防止智能体被恶意掉包或仿冒，智能体会与特定的代

的预先指定的模型和代码构建，并成功部署

数字签名或证书技术，明确证明某个智能体是由

度，从而确保系统运行的安

格验证，以此来精准确认该智能体的真实身份，同时评估其可信

全性和可靠性。

在访问智能体场景下，普通用户通过大模型安全网关与 AI-R-IAM 实名认证能力建立实

份，每个智能体可以

，与 AI-R-IAM 的管理员小证书类似

名

认证流程不再是简单的

，每个智能体都会被独立身份进行证书管制，并

时，

通过客户端认证校验，颁发短时令牌（Short Access Token）或一次性密钥

在执行任务

e Key）等方式，让其仅在当前会话或任务中有效。当智能体调用 RAG 时，通过

（One Tim

地理位置等）进行增强认证。

（如运行环境指纹，

**与监管**

**3. 关联审计与**

机制，对普通用户使用智能体行为，以及智能体自身的行为进行审计和监

通过身份绑定机

越权操作或异常行为时，AI-R-IAM 系统应有能力快速定位并冻结该代理

管。当智能体发生越

记录关于谁调用智能体的操作细节，例如访问了哪些数据、执行了何种指令、

权限，监管部门可

依据何种规则。

造成训练数据外泄。

在企业内部，智能体易成为攻击目标，被攻击后可能执行恶意操作，造

新风险，如自动提交敏感数据到外部。

数据存储服务器，RAG 知识库，

在此场景下，AI-R-IAM 提供用户和智能体，智能体与

智能体对 RAG 的数据访问权限

内外部大模型之间的数据权限边界，实行最小化策略，制定

和审计智能体异常行为等。避免

范围，辅以模型护栏、上下文检测等安全控制，可及时阻断

因智能体自主决策引发大规模安全事故。

# 6 发展趋势展望

异常行为。

或管理系统进行结合，实现对网络攻击的主动防御。基于零信任架构，AI-R-IAM 将对所有

网络安全防护体系。

术，保障大模型在物联网场景下与海量设备交互时的数据安全和身份可信；借助 5G/6G 网