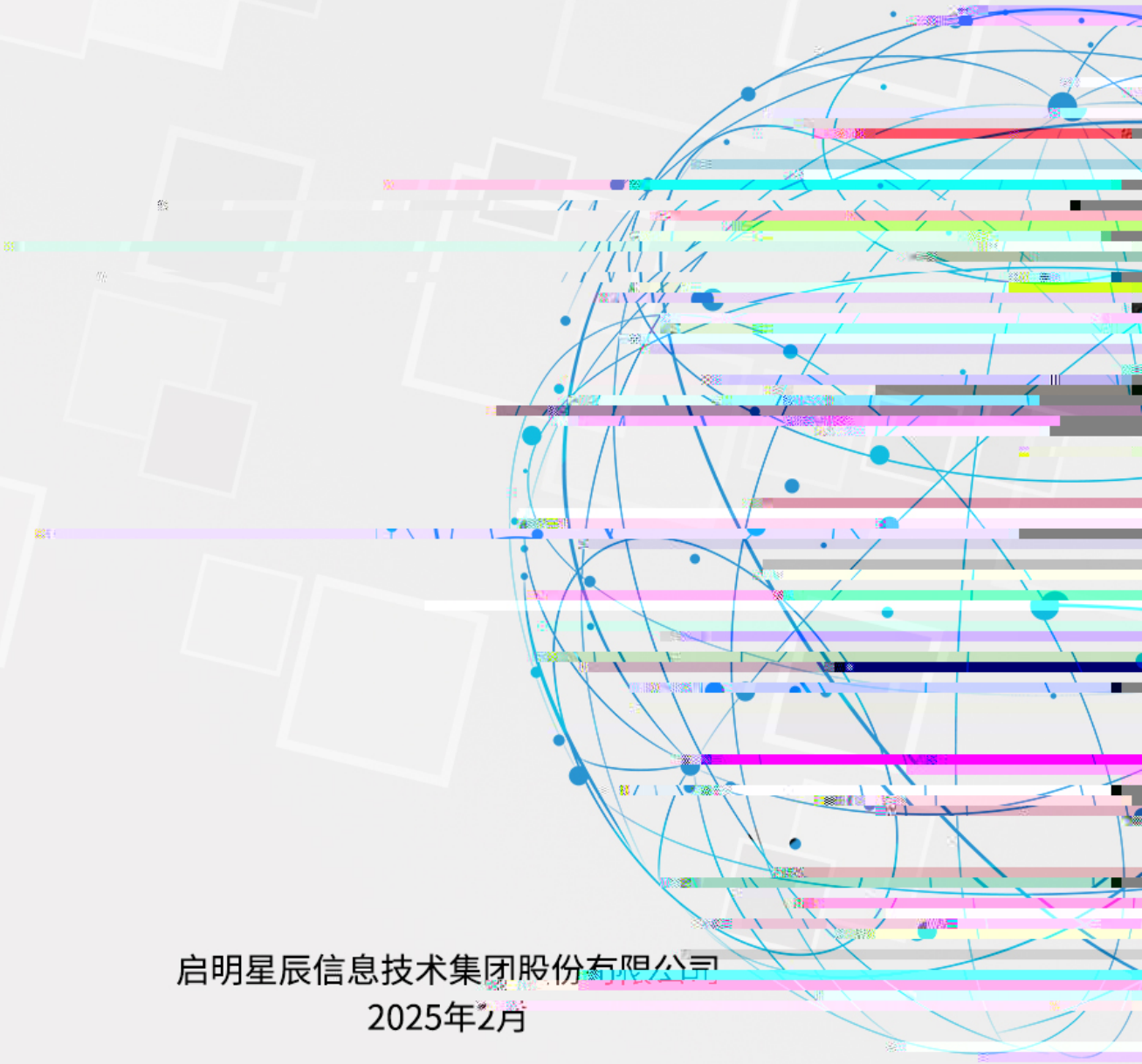
启明星辰

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

权。

和修改

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

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。未经北京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任何部分进行复制、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途。

免责申明

本文档仅提供有信息制作，其内容如有更改，恕不另行通知。

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漏、不准确，或错误导致

确

的损失和损害承担责任。

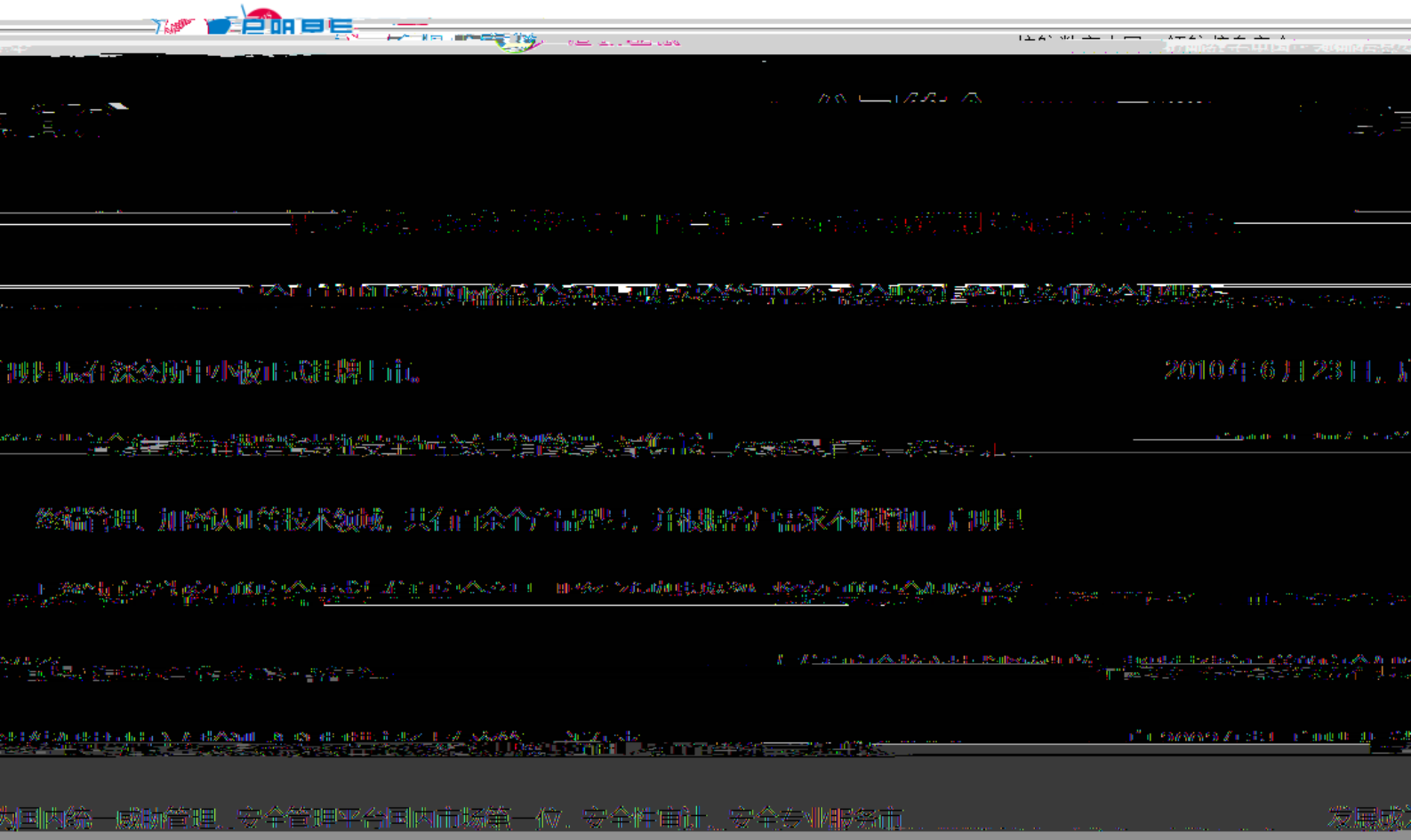
信息反馈

如有任何宝贵意见，请反馈：

100193 电话：010-82779088

传真：010-82779000

信息，



场领导者。目前，公司在全国各省市自治区设立三十多家分支机构，拥有覆盖全国的渠道和售后服务体系。

长期以来，启明星辰公司得到了党和国家领导人的关怀与鼓励。2000 年 1 月，江泽民、李岚清、曾庆红等党和国家领导人亲切视察启明星辰公司；2003 年 1 月，胡锦涛总书记亲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局内重点软件企业，国家火炬计划软件产业优秀企业，中国电子政务 IT100 强等荣誉，及拥有最高级别的涉及国家秘密的计算机信息系统集成资质证书。

启明星辰目前是我国拥有最大的国家级网络安全研究基地，完成包括国家发改委产业技部 863 计划、国家科技支撑计划等国家级科研项目近百项。创造了百示范工程，国家科技

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

的多项空白。

，启明星辰对政策 …… 为世界五百强中 80%的中国企业客户提供安全产品及服务；在金融领域，

为客户提供全方位信息安全保障。

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 …… 全产品和最佳安

民族信息安全产业第一品牌而不懈努力。 …… 提升国际化的

2 大模型的发展与风险

2.1 新质生产力 “大模型”

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。从 ChatGPT，再到百度文心一言、阿里通义千问、九天大模型、DeepSeek 等，大模型在自然语言处理、[illegible]

[illegible]

- 准备期（2022 年 11 月）

2022 年 11 月，OpenAI 发布 ChatGPT，[illegible]

- 成长期（2023 年）

[illegible]

(2) 国内：[illegible]

了“九天”人工智能模型，[illegible]

模型不断迭代，众多初创企业也凭借特色大模型在细分领域崭露头角。

大规模应用期（2025 年 1 月-）

国外：大模型深入医疗、金融、教育等多行业，助力疾病诊断、风险评估、个性

AI 还在探索新应用领域。

DeepSeek 凭借创新算法和架构，展现出低成本、高效能以及开源优势，

2.2 大模型应用中安全挑战

都暴露出诸多安全问题。

型应用交互场景中的背

1. **人类身份安全问题：**人类身份代表自然人身份实体，在大模

2. AI Identity安全问题：越来越多的AI Agent（人工智能体，AI Agent）

为以自己的身份执行任务，例如以AI Agent作为员工，有人

AI Agent的

权限。

以预测的方式请求权限。

(2) 过度授权：“人类身份和AI Agent不

出其应有的权限，从而能够访问

如果权限管理机制不完善，可能导致用户或应用程序获得超

和操作敏感资源，造成数据泄露或系统破坏。

(3) 身份安全：AI Agent身份系统通常存储大量用户的个人信息、行为数据等，如果

作骗、

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

定向广告骚扰等。

服务或

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

二层

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、

2025年2月发布《十大NHI风险2025》，如下图所示

HI 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，NHI 的身份生命周期管 NI

理薄弱，存在诸多安全风险与问题：

被内外部威胁者发现。

(2) 影子账号：大模型应用迭代快，加之生命周期流程弱、账号使用信息不可见，许多 NHI 账号不活跃，增加攻击面。

理者对安全维护和问题补救很关键。 管

的逻辑 NHI 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。

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

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 示词的验证和过滤机

的提示词用于医疗大模型进行诊断辅助，输出结果未脱敏处理，可能导致个人隐私泄露。

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

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

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

，恶意篡改训练数据时注入错误数据，这 (4) 数据泄露风险：在大模型训练和

规要求，若企业在数据收集、使用、存储和共享等环节不符合相关法规，如未经授权收集个

作组织、未获得用户充分同意收集，可能面临行政处罚、诉讼等

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

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设备、系统资源、应用资源、API 资源等，通过“ID+身份”的模式形成可信身份同时终端设备、用于提交互操作、可[illegible]性，搭配一体化安全机制的融合协同，实现全程全网、端到端的安全保护。

组合性和可扩展性，使企业能够以更少的资源实现更好的安全性

力，在政府、企业、个人中应用越来越深入，需要针对其身份、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就绪的大模型身份与访问管理。

随着大模型作为新质生产访问管理、数据安全等方面的可信的数字身份基座，即 AI 就

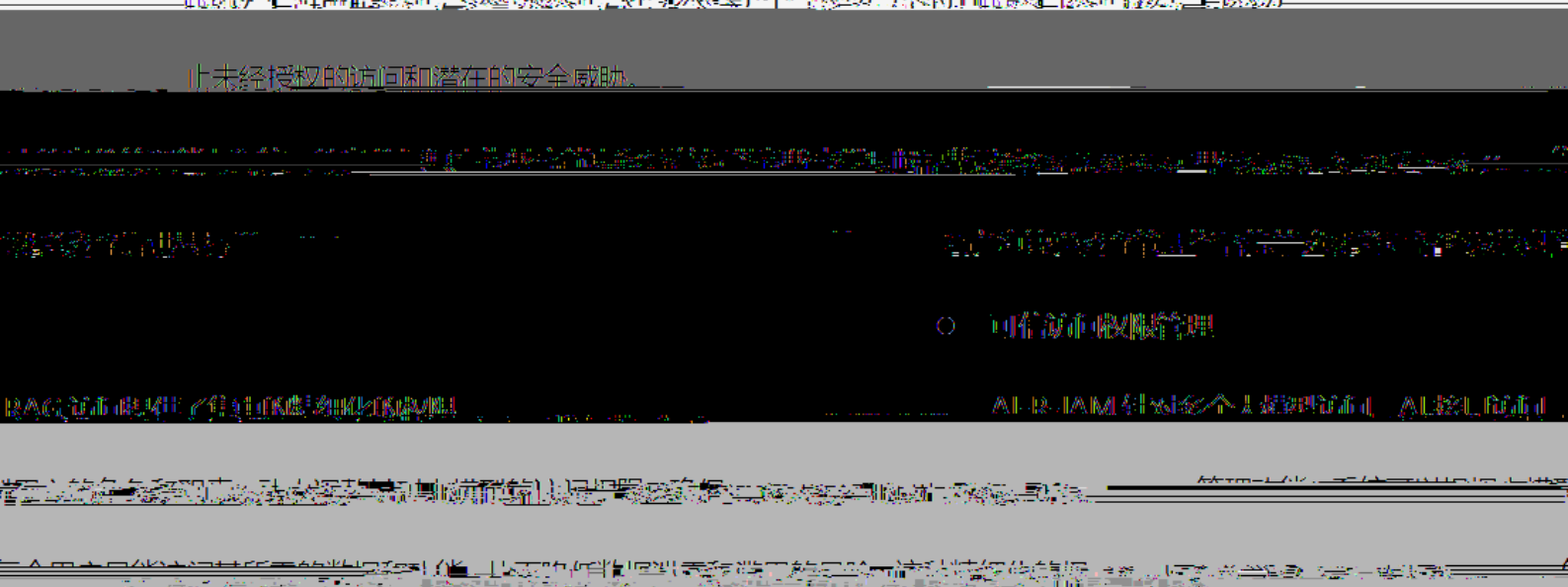
4 AI 就绪的大模型身份与访问管理

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 (如

[redacted]

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“一体化全程可信”保护体系，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。



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，为其它组件提供身份认证、权限、审计等方面的基础能力支撑，为大模型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范畴，聚合人类身份、AI Identity 以及非人类身份（NHI），

针对用户、智能体、API 应用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的管理，是数字世界秩序的底层支撑。

AI-R-IAM 采用了先进的身份认证机制，同时引入中国移动基于号卡特性的实名认证能力，有效防止未经授权的访问和潜在的安全威胁。

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；

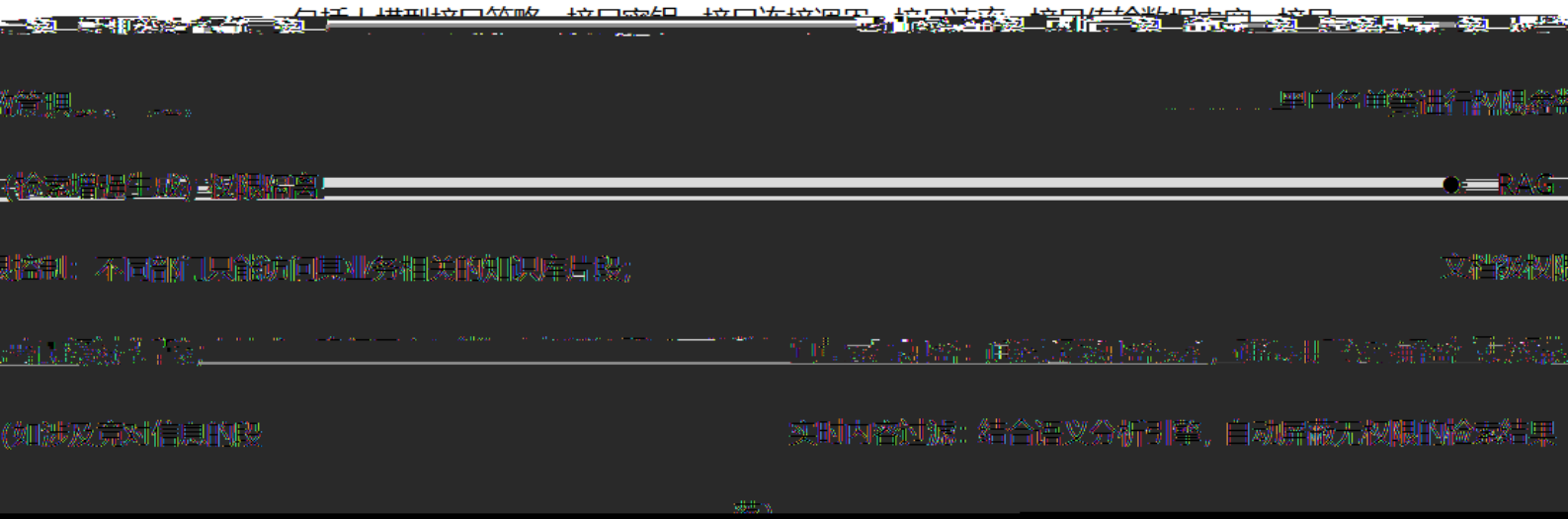
（2）智能体身份在创建时生成唯一 ID 并设定权限，修改时变更信息、动态调整权限，销毁时注销身份、回收资源并处理数据；

（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



4.4 可信行为审计

- 多维数据采集

的用户行为基线分析提供支撑。 构建全面

- 安全元数据管理

分发，提升数据价值密度。

- 可信溯源

- 数据画像分析

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，并根据风险画像评价一段时间内风险水平。

- 数据知识图谱

通过数据多源融合、关系抽

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

关系构建，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

个环节进行全机风险分析，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

提供全链路的智能追踪溯源，采用路径计算、路径回放等技术，提供

可疑 IP->访问数据分布->可疑路径->可疑日志链路分析，层层递进分析，实现风险智能追

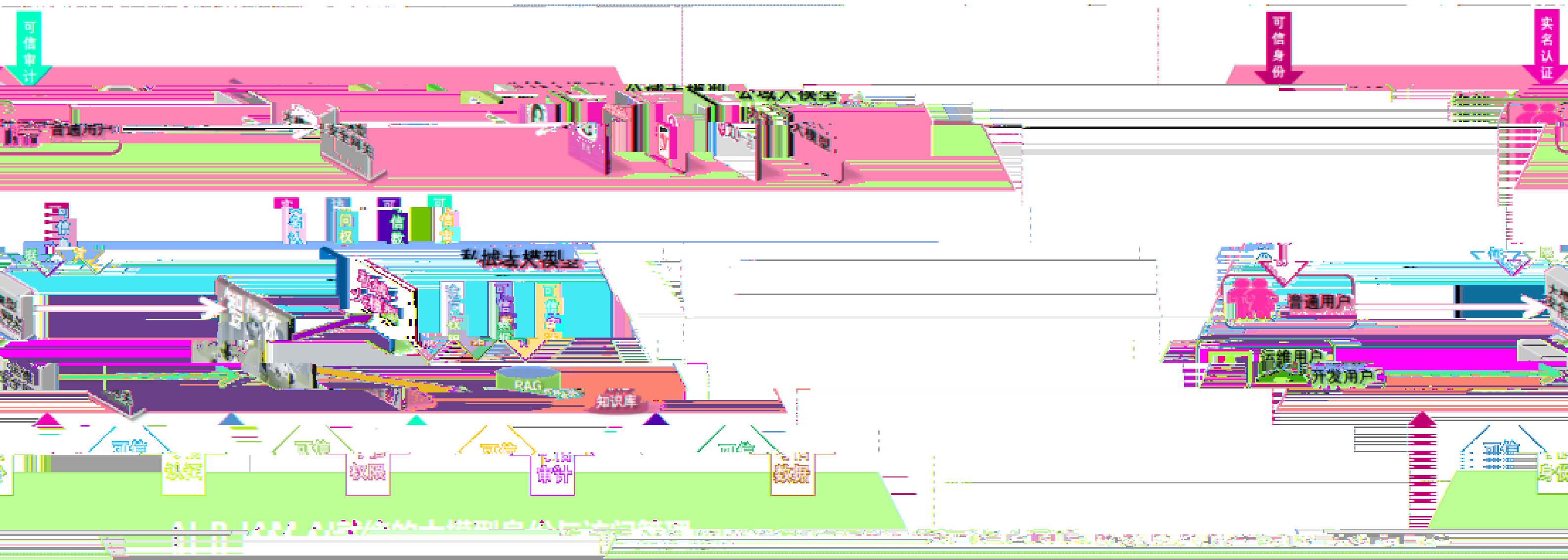
踪。

5 应用场景

5.1 大模型应用场景数据流

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



1. 普通用户访问公域大模型

普通用户，无论是企业员工还是个人用户，都可以借助自己的终端设备……

非常简便。比如，当一个企业员工想要利用文本生成服务撰写一份产品推广文案时，只需打……

然而，在这种便捷的服务背后，保障用户数据安全和系统稳定至关重要。AI-R-IAM 在普通用户访问公域大模型场景中，主要提供可信身份认证及安全访问的能力。

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，[illegible]

[illegible]

[illegible] 保护用户的数据不被未授权访问。

在安全审计方面，AI-R-IAM 持续监控整个系统的运行状况。它会记录下所有用户的操 [illegible]

[illegible]

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 [illegible] 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

[illegible]

普通用户通过终端设备访问部署在企业内部的私域大模型应用（例如知识问答、文档生成等 [illegible]

[illegible]

3. 运营维护人员访问私域大模型

[illegible]

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。

机制可以分为多个

弹出明显的警告信

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

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

禁止其执行，无论运维人员如何尝试都无法绕过这一限制。例如，对于删除整个数据库这样

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 在

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。 数据和

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

入或管理权限。由于权限划分不清晰，可能导

某些功能可能只需要读取权限，却被赋予了写

整体安全。

致内部人员误操作或恶意操作，进而影响系统

[redacted]

降低安全风险，实现人模型业务权限协作。

[redacted]

C 和 PBAC 技术，业务系统根据实体角色、状态和业务场景，

[redacted]

[redacted]

术，动态调整业务系统向 Token 分

[redacted]

理的分布和用户。

[redacted]

3. 业务行为审计

[redacted]

整、合规性检查不足、数据隐私泄漏等安全问题。

时，通过全链路日志采集、智能化分析

[redacted]

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，即使在设计层面明确一般用智能体身份做了某

相应责任。

项操作，系统可追溯到是谁让智能体进行操作，方便界定

指定的代码及模型紧密绑定，运用

为了有效防止智能体被恶意掉包或仿冒，智能体会与

智能体具由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

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，同时评估其可信度，从而确保

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

MFA，认证流程不再是简单的“一个智能体有独立身份进行访问控制，并

在执行任务前接受严格的认证检验，通过短期令牌（Short Access Token）或一次性密钥

（One Time Key）等方式，让智能体只能在特定时间、特定范围内访问特定资源，防止

（如运行环境指纹，地理位置等）进行增强认证。

3. 关联审计与监管

通过身份绑定机制，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

管，当智能体发生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

权限，监管部门可记录并追溯智能体的操作细节，例如访问了哪些数据、执行了何种指令，

依据何种规则。

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，如窃取敏感数据到外部。

存储服务器、RAG 知识库；

在此场景下，AI-RHAM 提供用户和智能体、智能体与数据

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，支持最小化策略，使得智能

与智能体的交互。避免

范围，辅以模型防护、上下文检测等安全控制，可及时阻断面向

向智能体的攻击和数据越权安全事故。

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。

借助 5G/6G 网

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；