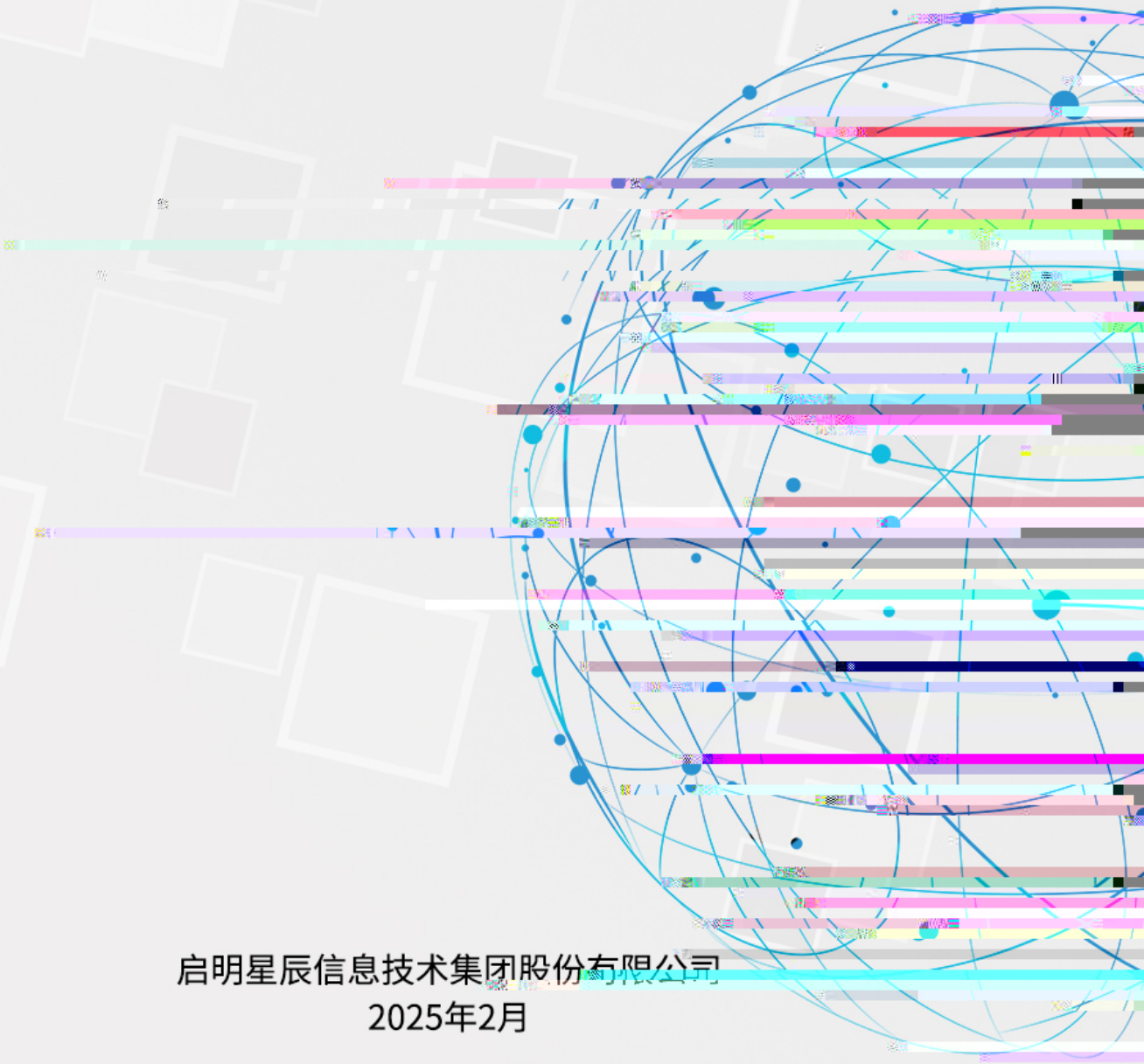
启明星辰

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

本文档依据现有信息制作，其内容如有更改，恕不另行通知。

北京启明星辰信息安全技术有限公司（以下简称“启明星辰”）在编写本文档的时候已尽最大努力保证其内容准确可靠，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漏、不准确、或错误导致的损失和损害承担责任。

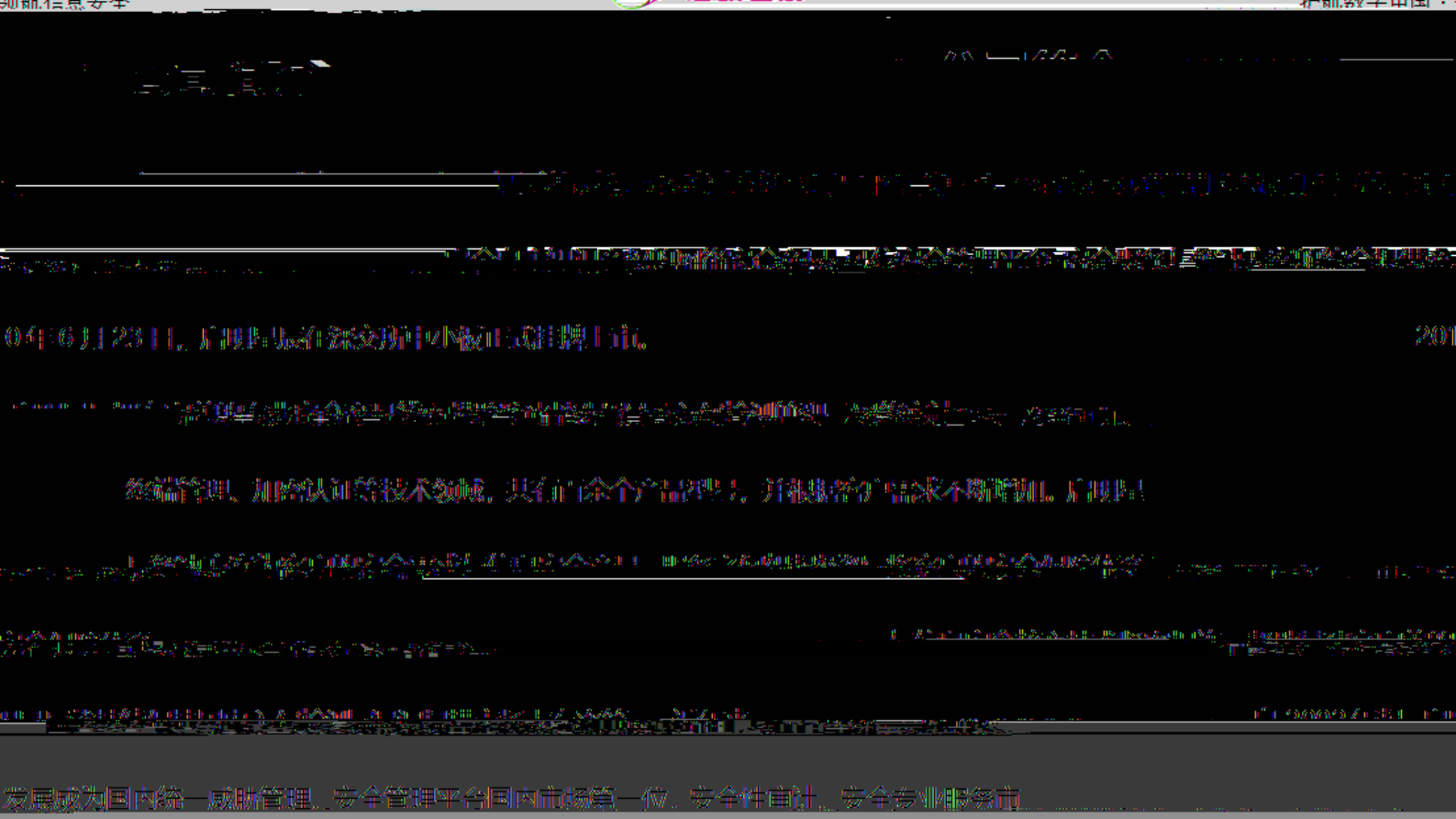
信息反馈

如有任何宝贵意见，请反馈：

信箱：北京市海淀区东北旺西路 8 号中关村软件园 21 号楼启明星辰大厦 邮编：100193 电话：010-82779088

传真：010-82779000

获得最新技术和产品信息，请访问启明星辰网站：



发展成为国内统一威胁管理、安全管理平台国内市场第一位、安全件审计、安全专业服务市

国的渠道和售后服务体系。

，江泽民、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

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

涛总书记亲李岚清、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

切接见了启明星辰公司 CEO 严望佳博士。

火炬计划软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

密的计算件产业优秀企业，中国电子政务 IT100 强等荣誉，及拥有最高级别的涉及国家秘

机信息系统集成资质证书。

[illegible]

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。创造了百示范工程，国

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

[redacted]

与对政策 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，启明星

[redacted]

小学。

[redacted]

性实践服务，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

的民族信息安全产业第一，启明星不懈努力。 提升国际化

2 大模型的发展与风险

2.1 新质生产力“大模型”

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。从

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，大模型在自

生成、智能写答、图像

、代码编写等多样化的任务。大模型发展的历程分为四个阶段：

- **准备期（2022 年 11 月）**

2022 年 11 月，OpenAI 发布 ChatGPT，其强大的自然语言处理能力震撼全球，开启

型发展浪潮，促使各方加大在该领域的投入与研究。

- **成长期（2023 年）**

(2) 国内：百度推出文心一言，整合自身技术优势；阿里发布通义千问，立足丰富业

模型发展提供技术储备。

(1) 国外：OpenAI、谷歌等公司不断迭代大模型，性能显著提升，开源大模型社区

(2) 国内：2024 年 5 月 24 日，在中国移动人工智能生态大会上，中国移动正式发布

态

了“九天”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

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型在细

- **大规模应用期（2025 年 1 月-）**

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，助力疾病

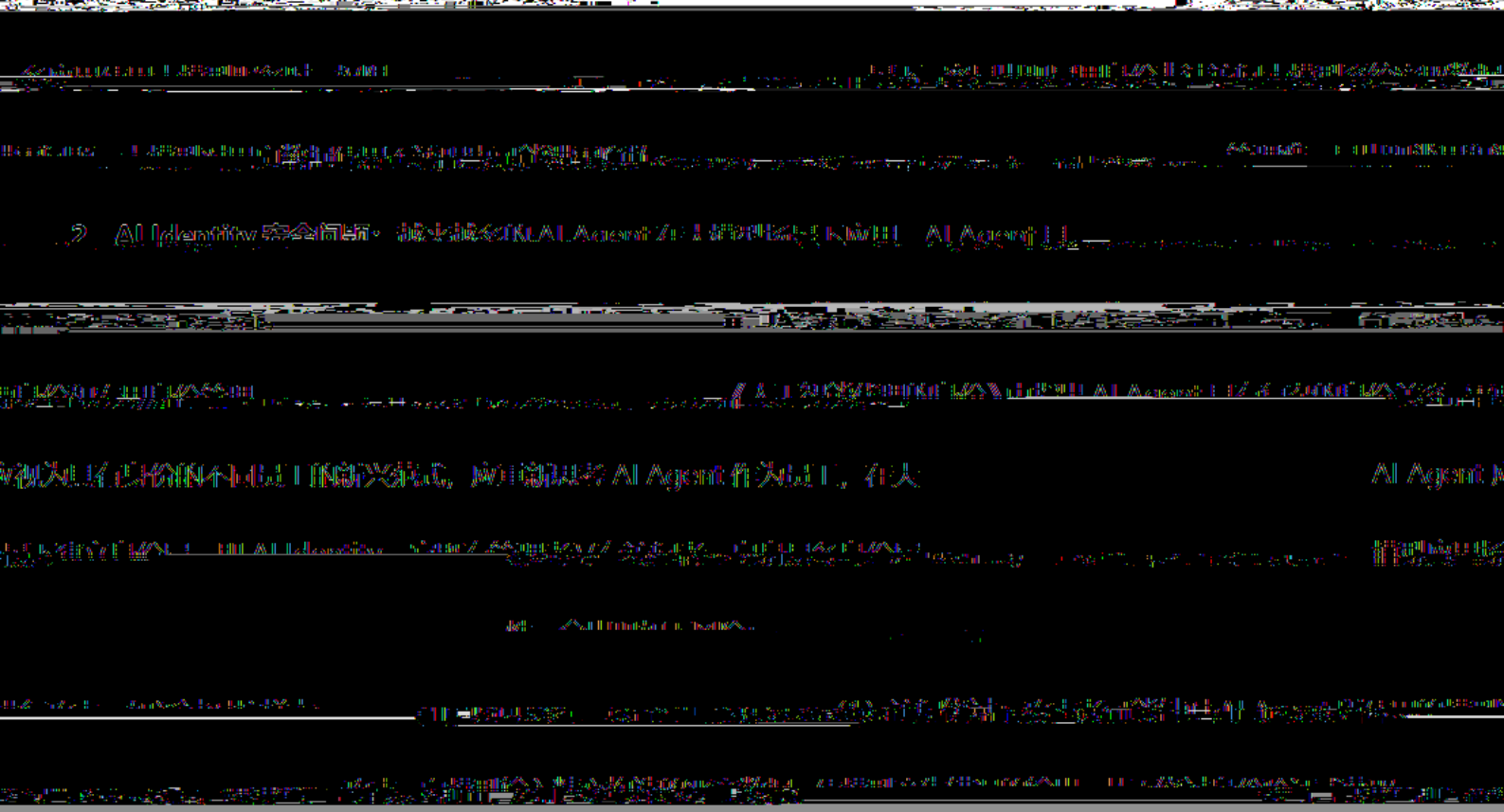
化学习等。OpenAI 还在探索新应用领域。

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

Identity，NHI），都暴露出诸多安全问题。



非法访问权限。

（2）访问权限放

如果权限管理机制不

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

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

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



中，NHI 的身份生命周期管

全风险与问题：

期流程弱、账号使用信息不可见，许多

织中，NHI 无明确所有权信息，而明确

和问题补救很关键。

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

理薄弱，存在诸多安全

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

NHI 账号不活跃，增加攻击面。

(3) 缺乏账号所有权：多数参与大模型应用的组

管理者对安全维护

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

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

示词的验证和

可能导致个人健康隐私泄露

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，

API 调用风险：API 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

去身份调用 API，获取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

获取 API 调用权限，

安全防护，易遭受攻击，导致数据泄露或篡改。比如攻击者通过漏洞获

非法获取金融大模型中的用户资产数据。

控制不当，用户可能

(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

(4) 数据隐私风险：在大模型训练阶段，攻击者故意向训练数据中注入恶意数据，这些

(5) 合规风险：随着数据安全相关法律法规的不断完善，大模型应用需要满足严

规要求，另外服务数据收集、使用、存储和共享等环节不符合相关法规，如未遵循数

化原则、未获得用户分分类报告，可能面临法律诉讼和

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“IP：身份”的绑定形式可信身份导赋能全网设备，用于提高互操作、可源实现更好的安全性，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合协同，实现全程全网、端到端的安全保护

随着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、访问管理、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可信的数字身份基座，即 AI 就绪的大模型身份与访问管理。

管理

是为大规模人工智能模型 (如

人“单点可控”迈向“一体化全程可信”保护体

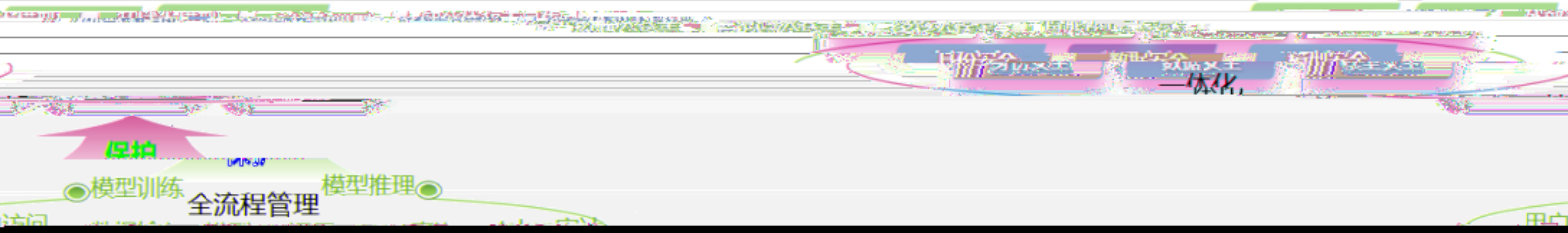
支撑。

4 AI 就绪的大模型身份与访问管

AI 就绪的大模型身份与访问管理 (以下简称: AI-R-IAM) 是

构建一体化全程可信能力，构建大模型安全从

系，同时为其它安全能力提供身份管理属性支



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

，还增强了用户的操作体验。

限管理不仅提高了系统的安全性

- 可信审计管理

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

为，包括用户登录时间、IP 地址、访问的大模型资源、执行的

和相关的访问行为，及时发现异常风险。

AI-R-IAM 对大模型使

控，记录用户的所有操作行

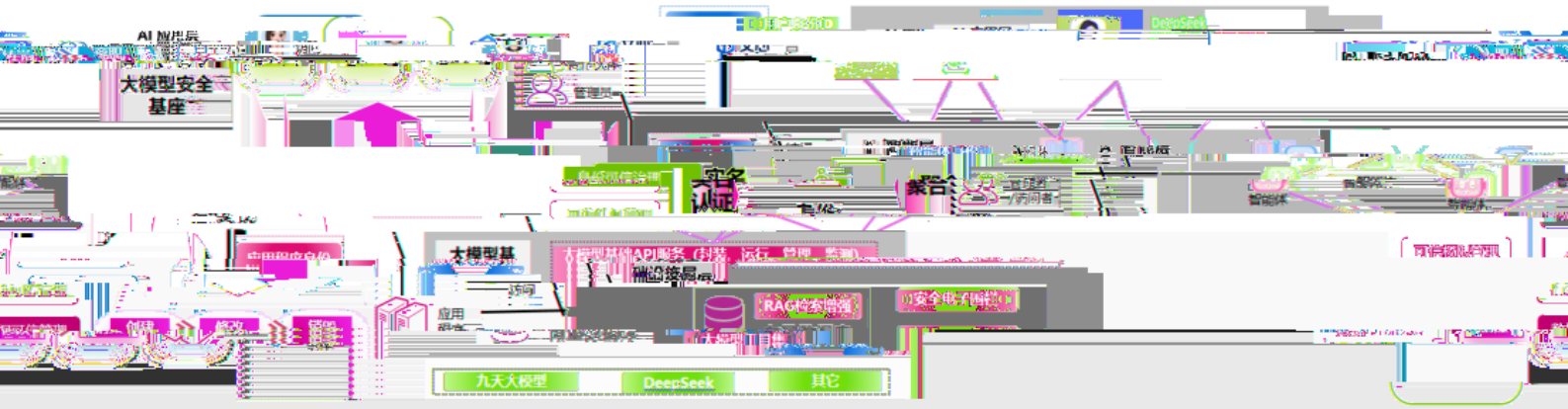
操作等信息，实时监测用户

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，

AI-R-IAM 通

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



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 AI-R-IAM

标识，对身份和属性进行全生命周期 针对用户、智能体、API 应用程序建立集中一体的身份

的管理，是数字世界秩序的底层支撑。

中国移动基于号卡特性的实名认证能 AI-R-IAM 采用了先进的身份认证机制，同时引入

身份从入职到离职的全生命周期管理，还创新性地将 NHI 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；

销毁时注销身份、回收资源并处理数据；

凭证并清理资源。 删除身份、销毁

4.2 可信权限管理

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、接口访问、接口传输数据安全、接口

用户名单等进行权限参数管理

增强生成）权限隔离 RAG（检索

：不同部门只能访问其业务相关的知识库片段；

文档级权限控制

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（如涉

蔽）



4.4 可信行为审计

[illegible]



- 多维数据采集

[illegible]

线分析提供支撑 [illegible] 构建全面的用户行为基

- 安全元数据管理

[illegible]

分发，提升数据价值密度。

- 可信行为基线

[illegible]

- 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，并根据风险画像评分，智能量化风险水平

- 数据知识图谱

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

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

风险智能追

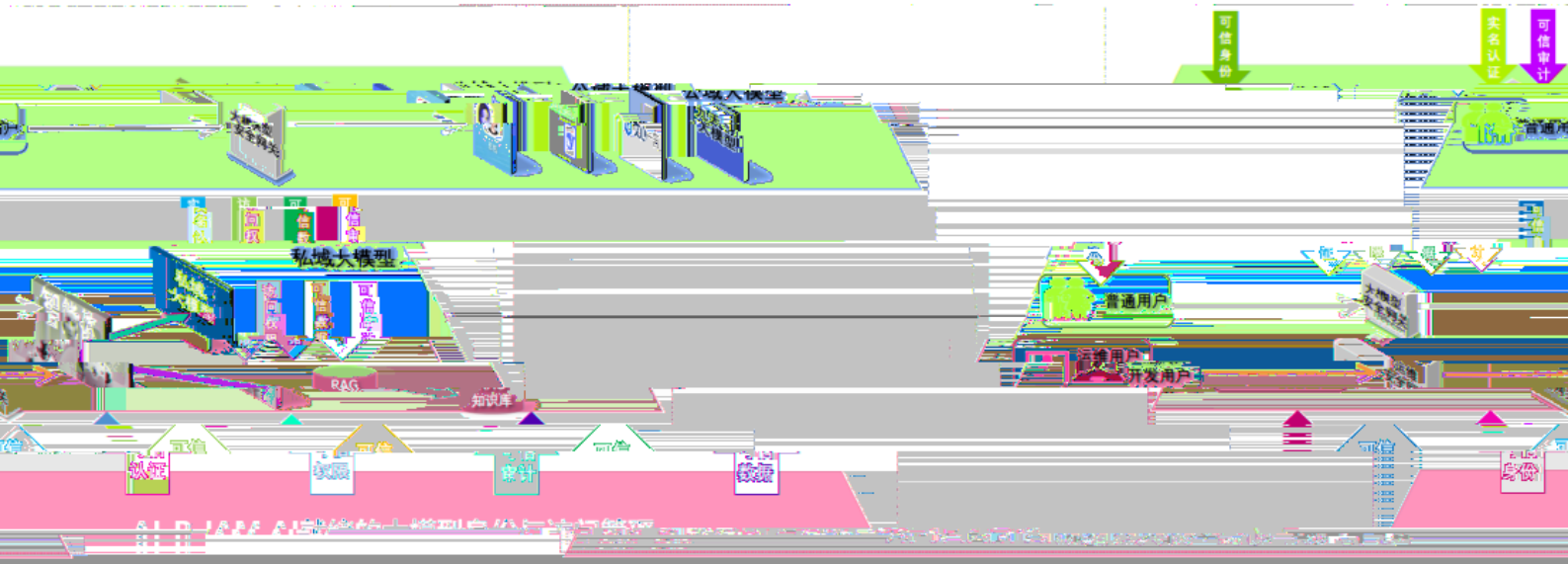
可疑 IP->访问数据分布->可疑路径->可疑日志链路分析，层层递进分析，实现踪。

5 应用场景



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

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模型应

内容：图像识别服务可以精准地分析图片中的物体、场景或人物

案时，只需打

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

AI-R-IAM

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。

在安全审计方面，AI-R-IAM 持续

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

普通用户通过终端设备

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

，还需要重点加强精细化权限管理、高危操作、金库管理等方面能力。

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

5.3 场景二：应用访问大模型应用场景



应用访问大模型场景主要包括业务应用访问公域大模型、业务应用访问私域大模型两种

AI-R-IAM

应用场景，业务系统能够通过安全智能体访问公域、私域大模型、RAG 知识库。

对业务访问提供以下安全措施：

1 业务身份管控

业务系统访问公域、私域大模型的身份安全极为重要，

如密钥可能被硬编码在代码中，或者存储在不够安全的地方，攻击者可能通过伪造身份来访

问大模型服务，进而消耗资源或进行恶意操作。

AI-R-IAM 从业务身份创建、修改和销毁的全流程管控，实现业务身份可信

身份创建：对业务系统身份访问与认证，通过数字证书或 API 密钥进行认证，并授予

最小访问权限。

身份修改：需根据功能和任务变更情况，更新业务系统身份信息和权限，确保合规并遵

循最小权限原则。

身份销毁：删除业务系统身份信息并销毁凭证

保系统的安全稳定运行。

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。



3. 业务行为审计

整、合规性检查不足、数据隐私泄漏等安全问题。

AI-R-IAM 在业务系统访问公域、私域大模型时，通过全链路日志采集、智能化分析

[illegible]

智能追踪溯源：对私域大模型、RAG 知识库的数据输入到模型输出提供全链路智能追踪溯源，采用路径计算、路径探测算法，提供可疑 IP->访问数据分布->可疑路径->可疑日志链路分析，是属通过分析，实现风险智能追踪[illegible]



[illegible]

定，防智能体被仿冒，用户经实名认证建立可信身份，智能体以独立身份接受认证，调用时[illegible]

全。通

获有限权限，同时通过身份绑定对双方行为审计监管，保障访问数据和训练数据的安全

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**

AI 通过 AI-R-IAM 提供可信身份管理，赋予用户可信身份（user）与智能体身份（A

了其 Identity）中统一身份 ID 进行绑定管理，便能在实计层面明确一般用智能体身份做

便界定相应责任。 项操作，系统可追溯到是谁让智能体进行操作，方

体会与特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

其个智能体是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

可信度，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

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

[illegible]

LIN，认证流程不再只面向普通用户，每一个智能体都能以独立身份进行访问控制，并

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

（One-Time Key）等方式，让授权更具时效性和可追溯性。当智能体调用 RAG 时，通过

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

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连续该代理 管

[illegible]

依据何种规则。

操作，造成训练数据外泄

外部。

用户和智能体，智能体与数据存储服务器，RAG 知识库，

，实行最小化策略，制定智能体对 RAG 的数据访问权限

等安全控制，可及时阻断和审计智能体异常行为等。避免

事故。

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

新风险，如自动提交敏感数据到外

在此场景下，AI-R-IAM 提供

内外部大模型之间的数据权限边界

范围，辅以模型护栏、上下文检测等

因智能体自主决策引发大规模安全

6 发展趋势展望

时的数据安全和身份可信；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