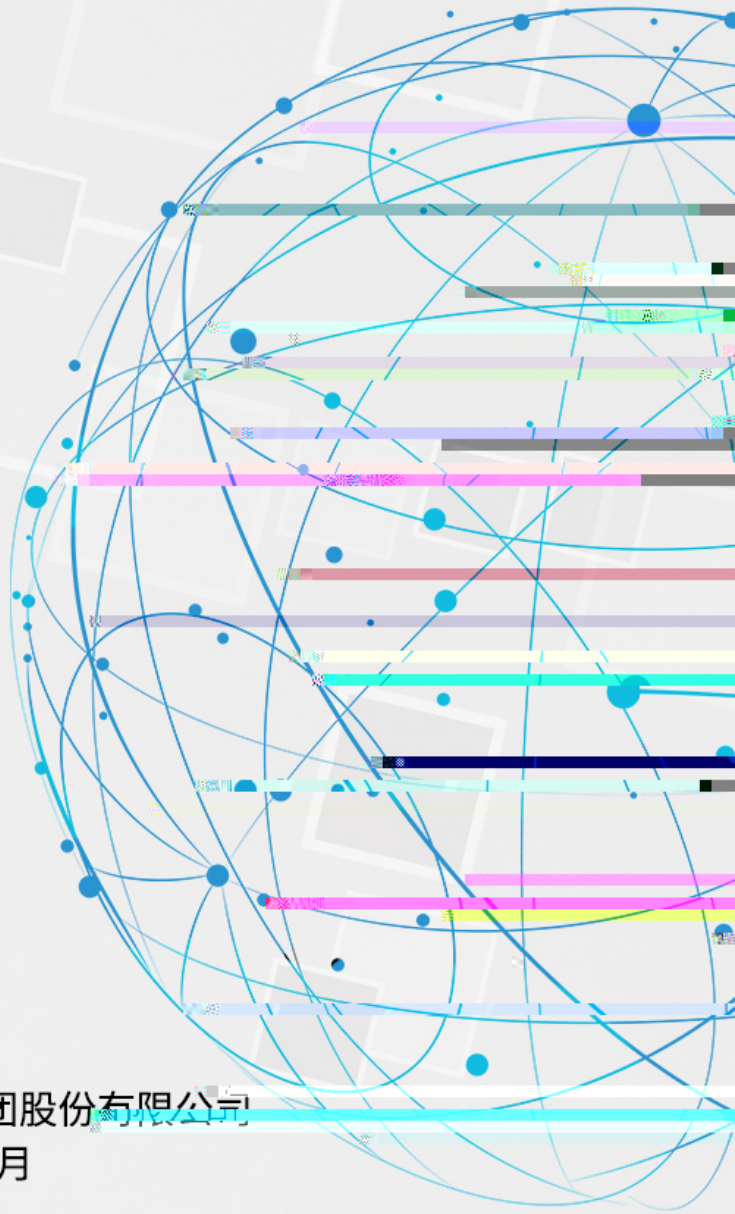
启明星辰

启明星辰|大模型应用安全|深度应用安全基座系列白皮书

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

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。

[redacted]

信息反馈

如有任何疑问，请联系：

[redacted]

100193 电话：010-82779088

传真：010-82779000

[redacted]

目录

建，是国内最具实力的、拥有 启明星辰公司成立于 1996 年，由留美博士严望佳女士创

服务与解决方案的综合提供商。 完全自主知识产权的网络安全产品、可信安全管理平台、安全

2010 年 6 月 23 日，启明星辰在深交所中小板正式挂牌上市。

启明星 终端管理、加密认证等技术领域，共有百余个产品型号，并根据客户需求不断增加。启

障体系 辰解决方案为客户的安全需求与信息安全产品、服务之间架起桥梁，将客户的安全保障

与信息安全核心技术紧密相连，帮助其建立完善的安全保障体系。

年来， 自 2002 年起，启明星辰就持续保持国内入侵检测、漏洞扫描市场占有率第一。近

服务市 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，安全性审计、安全专业服

渠道和 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

售后服务体系。

泽民、 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

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，胡锦涛总书

切接见了启明星辰公司 CEO 严望佳博士。

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，国家火炬计

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密的

机信息系统集成资质证书。

产业化 启明星辰目前是我国规模最大的国家级网络安全研究基地。完成包括国家发改委产

了百 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

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

全行业的领军企业，启明星辰以用户需求为根本动力，研究开发了完善的专

通过不断耕耘，已经成为在政府、电信、金融、能源、交通、军队、军工

制造等国内高端企业级客户的首选品牌：启明星辰在政府和军队拥有 95%的市场占有率，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，启明星辰对政策性银行、国有控股商业银行、全国性股份制商业银行实现 90%的覆盖率。在电信领域，启明星辰为中国移动、中国电信、中国联通三大运营商提供安全产品、安全服务和解决方案。

作为北京奥组委独家中标的核心信息安全产品、服务及解决方案提供商，奥组委唯一信息安全供应商，启明星辰受到独家官方授权，全面负责奥运会主体网络系统的安全保障，得到了国家主管部门的大力嘉奖。此外，启明星辰还为上海世博会、广州亚运会等多项世界级大型活动提供全方位信息安全保障。

余项专利和软件

的多项空白。

作为信息安

业安全产品线

以爱心回馈社会，截止目前，已累计资

西、青海、新疆等地援建了 5 所希望

在公司快速稳定发展的同时，启明星辰公司坚持

助贫困学子、受灾、贫困群众上亿元人民币，并在江

小学。

于提供具有国际竞争力的自主创新的安

础设施的安全性和生产效能，为打造和

。

启明星辰公司将秉承诚信和创新精神，继续致力于

全产品和最佳实践服务，帮助客户全面提升其 IT 基础

提升国际化的民族信息安全产业第一品牌而不懈努力。

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、计算机视觉、语音识别等领域展现出强大的能力。这些模型通过在海量数据上进行训练，能够学习到丰富的语言知识、语义理解和逻辑推理能力，从而实现文本生成、智能问答、图像生成、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，开启大模型发展浪潮，促使各方加大在该领域的投入与研究。

- **成长期（2023 年）**

(1) 国外：OpenAI 持续优化 ChatGPT，并拓展其应用领域，谷歌、微软等科技巨头

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-）**

（1）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、个性化学习等。OpenAI 还在探索新应用领域。

（2）国内：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，

字化、智能化转型。

等行业大规模应用，推动各行业数

的代表，正以前所未有的速度和深度渗透到企业运营和个

大模型技术作为新一代生产力的

进步和经济发展的关键力量。它不仅重塑了传统行业的运

人生活的方方面面，成为推动社会进

2.2 大模型应用中安全挑战

2.2.1 身份与访问控制的安全挑战

在大模型深度应用的时代背景下，身份安全面临着前所未有的挑战。随着人与设备、人与应用、AI Agent 与数据、人与 AI Agent 之间交互量呈指数级增长。

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 Agent 构

应用中产生了三类身份制公，即人类身公，AI Identity 以及非人类身公（Non-Human Identity，NHI），都暴露出诸多安全问题。

1、人类身份安全问题：人类身份代表自然人身份实体，在大模型应用交互场景中的背

[illegible]

2、AI Identity 安全问题：越来越多的 AI Agent 在大模型场景下应用，AI Agent 具

[illegible]

分关系，需要身份和专用身份管理。

新思考 AI Agent 作为员工，在大

[illegible]

《人工智能代理的身份》中指出 AI Agent 具有复杂的身

AI Agent 应视为具有身份的不同员工的新兴范式，应重

[illegible]

性，会出现如下风险：

[illegible]

[illegible]

[illegible]

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出其应有的权限，从而能够访问和操作敏感资源，造成数据泄露或系统破坏。

[illegible]

能被窃取，导致用户隐私泄露，进而可能被用于精准诈骗、

系统遭受黑客攻击，这些数据可能

定向广告骚扰等。

问题：非人类身份定义为企业技术内的应用程序、服务或

些包括机器人、API 密钥、服务帐号、OAuth 令牌、云服

3、非人类身份（NHI）安全

机器等实体绑定的数字身份。这些

[illegible]

2025 年 2 月发布《十大 NHI 风险 2025》，如下图所示。

大模型基础设施交互过程中，NHI 的身份生命周期管

发应用时，很多 NHI 硬编码在代码库等位置，容易

，加之生命周期流程弱、账号使用信息不可见，许多

模型应用的组织中，NHI 无明确所有权信息，而明确

，相同 NHI 会在生产和非生产环境中使用，或者相同

密码，从而增加了横向移动的风险。

享间共享 NHI，违反原则，使密码循环等安全操作复

NHI 在大模型应用中，大多数存在于大

理薄弱，存在诸多安全风险与问题：

（1）纯文本/未加密凭证：大模型开

被内外部威胁者发现。

（2）影子账号：大模型应用迭代快，

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

管理者对安全维护和问题补救很关键。

（4）缺乏环境隔离：在许多情况下，

的逻辑 NHI 在每个环境中都有相同的密

（5）凭证共享：大模型应用的多程序

在，通过智能体依赖关系。

2.2.2 数据安全挑战

大模型应用在数据安全方面主要面临以下挑战：

（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

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制不完善，敏感信息

业机密等敏感信息，如用户输入涉及个人健康状况 出内容审查不足，可能泄露用户隐私。攻

行诊断辅助，输出结果未脱敏处理，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

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 API 调用风险

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

致数据泄露或篡改。比如攻击者通过漏洞获取 API 调用权限， 安全防护，易遭受攻击，导

户资产数据。 非法获取金融大模型中的用

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

敏感知识数据，而且知识库中的数据来源复杂，若数据标注、整合环节出 超出授权范围获取

现错误或被恶意篡改，查询结果的准确性和安全性难以保障，可能误导用户并造成决策失误。

（4）数据投毒风险：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

模型的训练数据中混入大量被恶意标注的图像，使模型在识别相关图像时出现严重偏差。

（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格的合规要求。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，如未遵循数据最小化原则、未获得用户充分授权等，可能面临法律诉讼和巨额罚款。

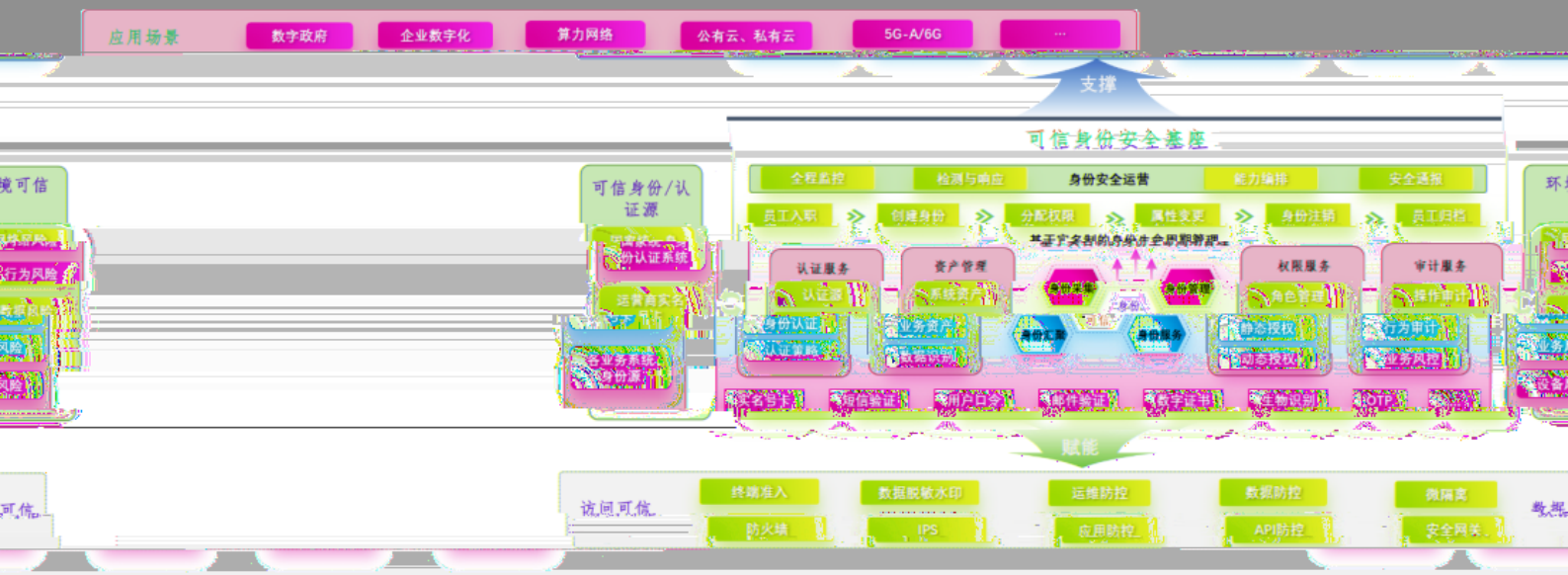
3 身份与访问管理系统（IAM）

身份与访问管理（Identity and Access Management，IAM）作为一种综合性的身份安全管理框架，旨在确保数字环境中用户身份的真实性、可靠性以及访问权限的精确控制。它涵盖了身份认证、授权管理、身份生命周期管理等多个关键环节，通过一系列技术手段和管理策略，为企业和组织提供了安全、高效的身份管理解决方案。

程，予人员、设信身份对访问、数据等维度的“一体化全程可信”。

启明星辰基于 IAM 的技术实践，针对企业数字化要求提出通过 IAM 构建“一体化全程可信数字身份底座”理念。该理念旨在逻辑上建立一张基于身份的可信网络，赋予设备、应用、接口、数据等实体唯一身份标识，实现实名制可信身份鉴权，基于可信身份对访问主体进行动态访问控制，推动业务与数据安全从“单点可控”迈向身份、环境、

基于 IAM 构建的“一体化全程可信数字身份底座”总体架构如下图所示：



为人员、设备、应用等多类应用场景提供安全可信身份，实现赋能……IAM 作为体系构建的核心能力，

安全访问设备，驱动安全控制执行，同时联动可信环境能力，及时调整访问策略。

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、数据风险、业务风险的全链路安全风险审计。

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设备、系统资源、API 资源等，通过"ID+身份"的绑定形成可信身份同时对个网设备、服务提供

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更好的安全协同，实现全程全网、端到端的安全保护。

、企业、个人中应用越来越深入，需要针对其身份、 随着大模型作为新质生产力，在政府

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、数据安全等方面的需求，建立

型身份与访问管理。 可信的数字身份基座，即 AI 就绪的大模

4 AI 就绪的大模型身份与访问管理

规模人工智能模型（如

全、模型安全的基础支

使用过程等多个场景中

体化全程可信”保护体

AI 就绪的大模型身份与访问管理（以下简称：AI-R-IAM）是为大

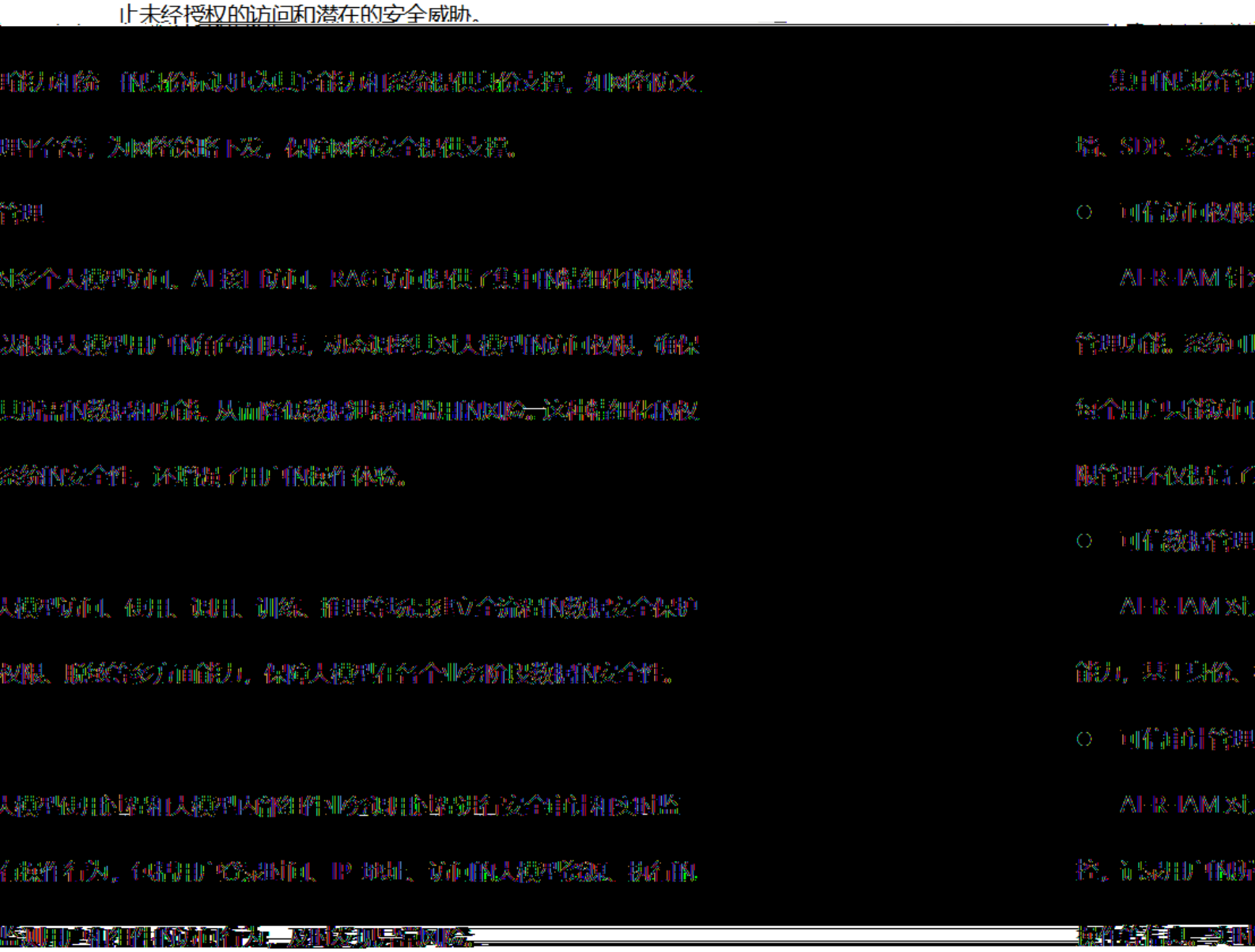
DS、GPT、BERT 等）的深度应用提供身份安全、访问控制、数据安全

撑系统，从身份、访问、行为、数据等维度为大模型在训练、部署和使

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

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。



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，

的设备，提供身份、认证、权限、审计等方面的基础能力支撑，为未来人

发展奠定了坚实的基础。

AI-R-IAM 通

为其它保护大模型

工智能技术的安全

4.1 可信身份治理

的多维度身份和属性的治理和管理，AI-R-IAM 的“身聚合人类身份、AI Identity 以及非人类身份（NHI），立集中一体的身份标识，对身份和属性进行全生命周期。

证机制，同时引入中国移动基于号卡特性的实名认证能的安全威胁。

份从入职到离职的全生命周期管理，还创新性地将 NHI力企业实现数智化转型，有效降低安全风险，提升合

职时，进行信息收集验证、权限调整及账号注销等操作；

AI-R-IAM 支撑基于大模型特性的份 ID”超越了传统身份概念的边界，针对用户、智能体、API 应用程序建立的管理，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力，有效防止未经授权的访问和潜在的

AI-R-IAM 不仅涵盖传统人类身份和 AI Identity 纳入统一治理范畴，助规性。在具体流程如下：

(1) 用户身份在入职、转岗、离职

(2) 智能体身份在创建时生成唯一

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。

可信身份治理根据策略集中编排身份管理的过程，用于提供更好的身份可见性和访问权

[illegible]

系统的安全可靠，有力地保护了数据隐私。

限设定、信息更新等措施，确保

4.2 可信权限管理

访问、AI 接口访问、RAG 访问、模型调用、数据训练等

AI-R-IAM 针对每个大模型

供了多维度、精细化权限管理功能。系统可以根据人模型用户的不同职责，动态地为对

人模型的不同权限，例如为个用户人员设置不同的数据访问权限，从而在数据和功能的

风险。这种精细化的权限管理不仅提高了系统的安全性，还增强了用户的灵活体验。

- 用户访问权限

[illegible]

个人模型访问进行控制和限制。

智能体对 RAG 的访问、用户对 RAG 的访问关联关系，确

- 用户角色权限管理

角色权限管理也称 RBAC（基于角色的访问控制）模

以及软件版本的功能，为人模型的访问体个维度分配权

角色类型	模型训练	模型管理	数据标注	知识库
算法工程师				×
数据科学家				
系统运维	×	仅监控	×	×
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、接口速率、接口传输数据内容、接口黑白名单等进行权限参数管理。

- RAG（检索增强生成）权限隔离

文档级权限控制：不同部门只能访问其业务相关的知识库片段；

向量索引加密：通过分段加密技术，确保用户仅能解密与其权限匹配的数

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（如涉及

落）。

- Token 动态调配

4.3 可信数

AI-R-IAM 通

和 RAG 机制对数

的可信数据集中管

位的可信数据管理能力。


数据输出
模型训练&RAG

4.4 可信行为审计

AI-R-IAM 通过多维数据采集、安全元数据管理、可信行为基线、数据画像分析、数据知识图谱、数据深度分析、智能追踪溯源等功能为大模型用户访问场景下的可信行为审计。

- 多维数据采集

通过采集大模型用户行为、实体行为、上下文、身份与权限、特定模型等多维数据，为构建全面的用户行为基线分析提供支撑。

- 安全元数据管理

对大模型用户操作行为数据进行标准化处理，包括数据提取、清洗、关联、对比、标识、分发，提升数据价值密度。

- 可信行为基线

通过机器学习和统计分析技术，对历史数据进行分析，建立大模型用户可信行为基线并通过与应用场景相结合，为风险决策提供精准画像。

- 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

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

[redacted]

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

访问、使用、调用、训练等多

数据能力。相关场景包括：用

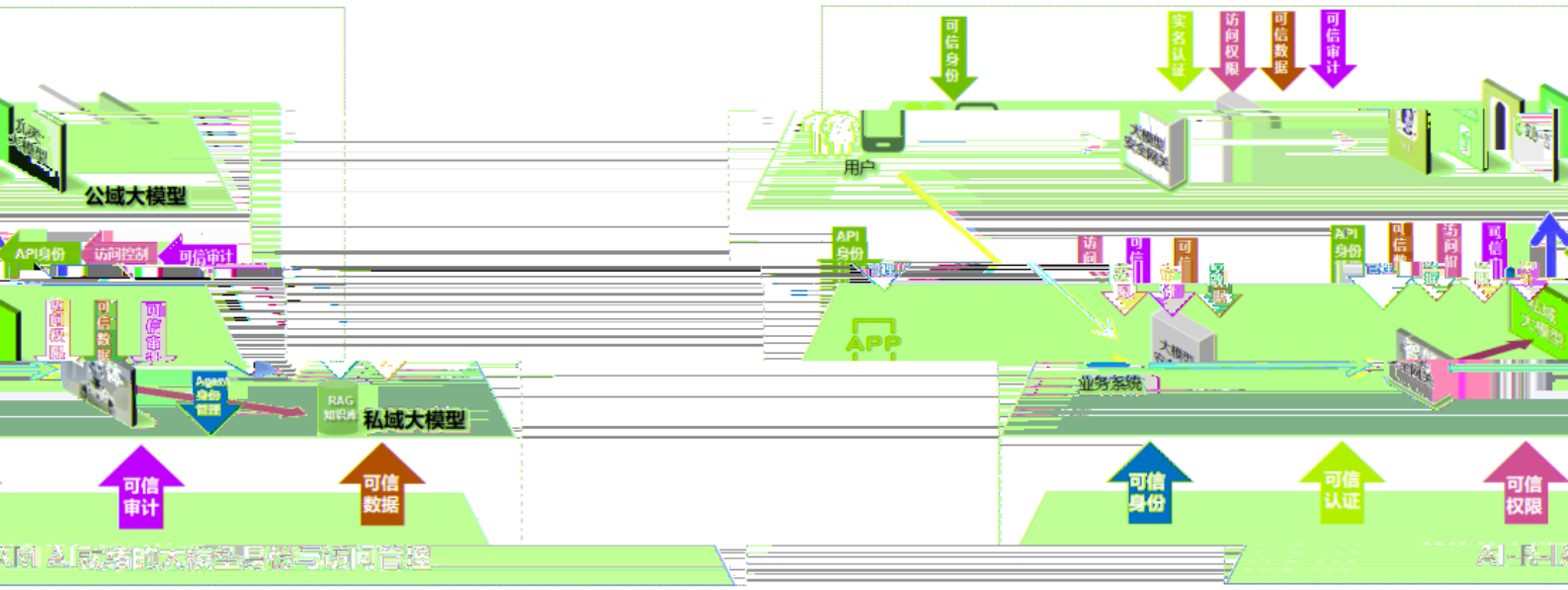
RAG 知识库、私域大模型场

AI-R-IAM AI 就绪的大模型身份与访问管理，为大模型的

种场景提供可信身份、可信认证、可信权限、可信审计、可信

户访问私有大模型、公域大模型场景；业务系统调用智能体、

景；大模型数据训练场景、大模型数据运维场景等。



访问大模型应用场景

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

可信审计、可信数据等关键能力的全面覆盖。主要场景包括用

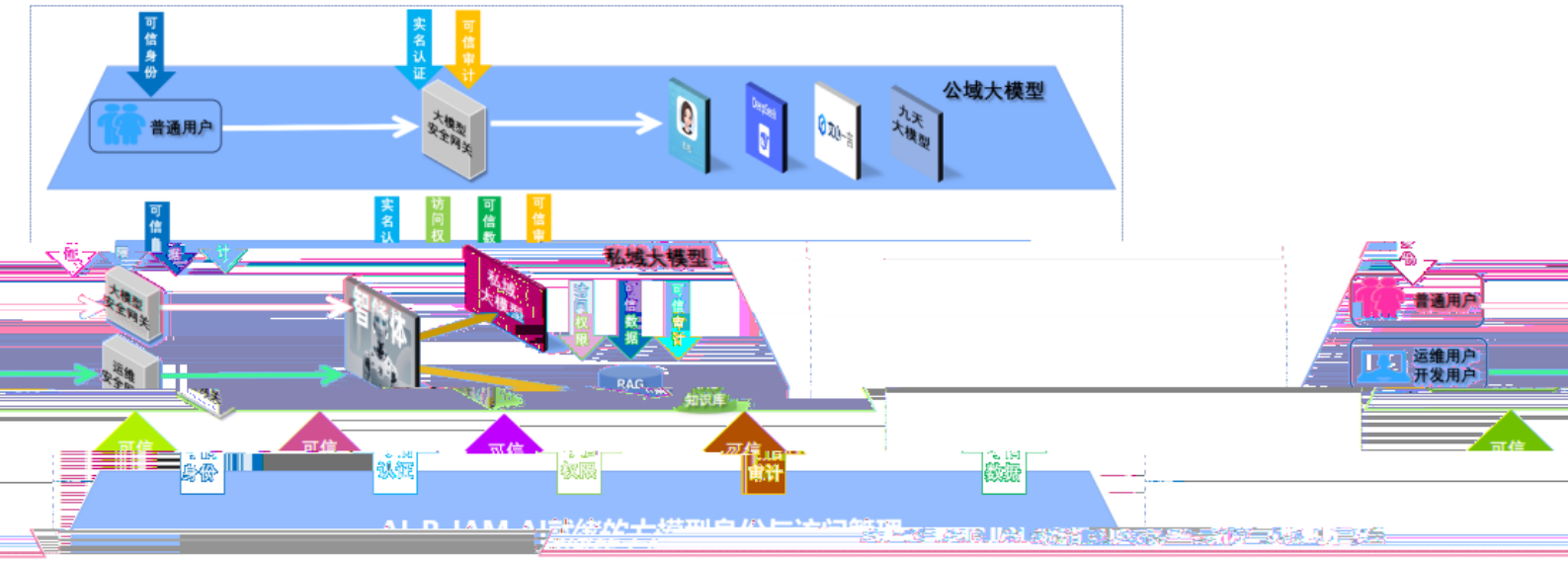
问私域大模型、运营维护人员访问私域大模型等。

5.2 场景一：用户

AI-R-IAM AI 就绪的大

身份、可信认证、可信权限、

户访问公域大模型、用户访问



1. 普通用户访问公域大模型

这些大模型应用通常以 Web 界面的形式来为用户提供服务。对于用户而言，操作起来非常简便。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，只需打开浏览器，进入该服务的 Web 界面，然后输入产品的相关信息如名称、特点等，再点击生成按钮，很快就能得到一份高质量的文案初稿。又如个人用户使用图像识别服务辨别一张植物照片的种类时，也是通过简单的上传图片操作，就能从 Web 界面上获取到详细的植物分类结果。

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是至关重要的。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审计的能力。

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，系统会要求用户提供身份凭证，这可能包括用户名和密码组合、生物特征（如指纹或面部识别）或者是基于可验证凭据（Verifiable Credential，VC）的身份认证等方式。只有经过严

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作行为，包括登录时间、执行的操作类型、访问的数据范围等详细信息。通过对这些审计日志的定期分析，可以及时发现潜在的安全隐患或者异常操作。例如某个账户在短时间内进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

问答、文档生成等 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

相比，私域场景在数据隐私和合规性方面有着更为严苛的要 服务）。与公域大模型访问场景

信数据管理方面具备更强的支撑能力。 求，因此必须在权限控制以及可

在用户身份可信认证基础上，私域场景下需要对不同子部门或者角色赋予相应的权限级别。企业内部存在多种角色，如企业管理者、部门管、普通员工等。企业管理者能够全面了解整个企业的运行状况，包括各个部门的数据情况；部门管则更关注本部门的数据，并且能够对本部门员工的权限进行一定范围的调整；普通员工的权限则为有限，只能对与自己工作直接相关的少量数据进行访问。这种多层次的角色划分，要求权限控制系统具备很高的灵活性和准确性，以满足不同角色对数据的不同保密级别要求。

3. 一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

测试等后台管理人员，在普通用户私域访问大模型的安全能力基础

化权限管理、高危操作、金库管理等方面能力。

同运维人员可能承担着不同的职责，例如有的负责硬件设备的维护，如服务器的物理状态检查、网络连接状况监测等；有的则专注于软件层面的维护，像操作系统补丁更新、数据库性能优化等工作。因此，需要根据这些不同的职责范围来划分权限。对于硬件维护人员，他们应该只能访问与硬件相关的监控信息和配置界面，而不能触及到存储在服务器上的敏感业务数据或模型参数等。同样，软件维护人员也不能越权去修改硬件相关的设置。

这种基于角色的权限划分还需要考虑到运维团队内部的层级关系。例如，初级运维工程师可能只能执行一些常规的、风险较低的操作，如查看日志文件、重启某些非关键服务等；而高级运维工程师则拥有更高的权限，可以进行更复杂的操作，如调整系统核心参数、执行大规模的数据迁移任务等。通过这种多层次的角色权限体系，能够有效避免因权限过度集中而导致的安全隐患。

尝试访问时，就需要触

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

心模型参数、更改系统

在私域大模型环境中，高危操作可能包括删除大量数据、修改核

阻断机制可以分为多个

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

统会弹出明显的警告信

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

息，提醒运维人员该操作可能存在严重后果，并要求他们确认是否继续。这种警告提示可以让运维人员重新审视自己的操作意图，避免因误操作而导致问题。第二层次是权限二次验证，即使运维人员确认要继续执行高危操作，也需要通过额外的权限验证步骤。这可以包括输入更高层级的管理员密码、使用硬件令牌进行身份验证等，只有通过了这一严格的验证过程，才能够真正执行高危操作。第三层次是完全阻断，对于某些极其危险的操作，系统可以直接

尝试都无法继续进行。例如，对于删除整个数据库这样

禁止其执行，无论运维人员如何

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。

认证）才能打开金库。同时，所有对金库的

各自的认证信息（如指纹、虹膜扫描等生物特征

访问的时间、地点、访问者的身份信息、访

访问都必须被详细记录下来。这些记录应该包含

需要定期审查，以确保没有未经授权的访问

问的具体内容以及访问的原因等。并且这些记录

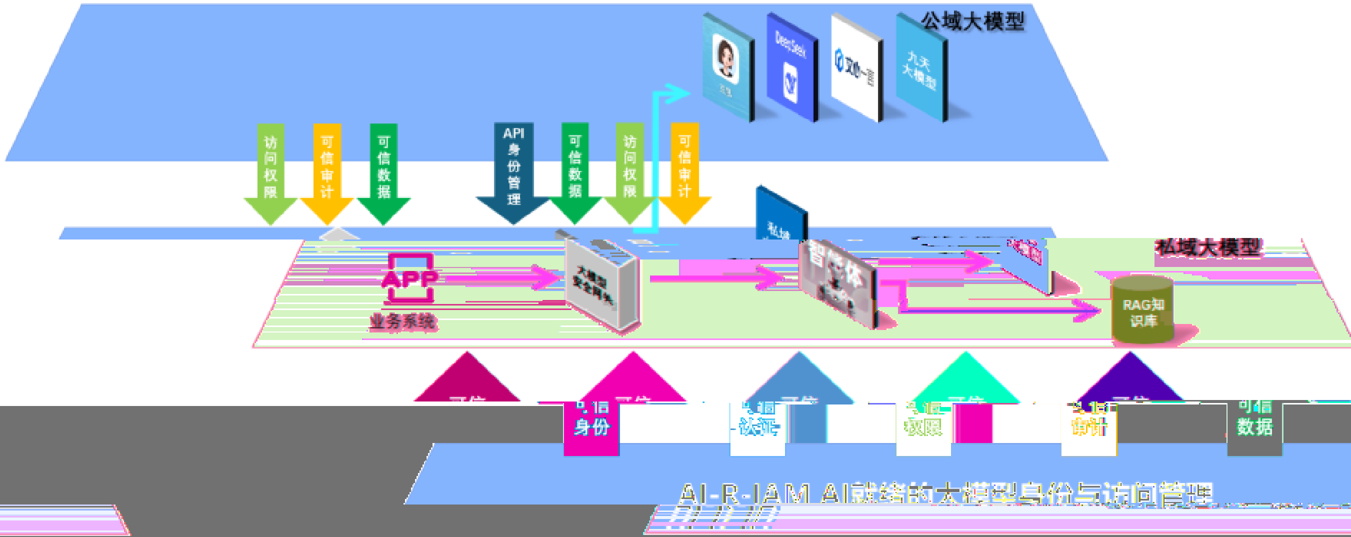
现有人试图违规访问金库，立即触发警报并

发生。此外，还可以设置实时监控机制，一旦发

队等。

采取相应的应对措施，如锁定账户、通知安全团

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。AI-R-IAM 对业务访问提供以下安全措施：

1. **业务身份管控**

[illegible]

[illegible]

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

清晰，可能导

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

致内部人员误操作或恶意操作，进而影响系统整体安全。

保障大模型

AI-R-IAM 在业务系统访问公域、私域大模型时，通过精细化的权限控制，

资源的合理分配与安全性，降低安全风险，实现大模型业务权限可信。

和业务流程，

实体级授权：通过 ABAC 和 PBAC 技术，业务系统根据实体的属性、状态

效管理大规模访问需求，并降低权限管理复杂度。

据敏感

数据级授权：结合 RBAC 和 ABAC 技术，系统依据业务系统的角色、属性及数

性，精细控制数据访问权限，防止数据泄露和滥用。

ken 分

Token 动态授权：通过 RBAC、ABAC 和 PBAC 技术，动态调整业务系统的 Tok

配，实现对资源访问的弹性管控，确保资源得到高效合理的分配和使用。

3. 业务行为审计

权与权限校验、输入数据预处理、构造 API 请求参数、调用公域/私域大模型 API、模型服

务处理、返回响应结果、结果后处理、返回最终结果至用户等业务流程，存在日志记录不完

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、

智能追踪溯源三方面保障业务行为可信。

全链路日志采集：采集私域大模型、RAG 知识库的用户行为、实体行为、上下文、身

份与权限、特征模型等多维数据，以构建全面的用户行为基

据进行分析，建立私域大模型、

行为的动态变化，确保基线时

大模型、RAG 知识库安全风险

智能化分析：通过机器学习和统计分析技术，对历史数

RAG 知识库的用户可信行为基线，并通过自适应用户和实体

效性和准确性。同时采用画像分析、知识图谱等技术对私域

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

算、路径探寻算法，提供可疑 IP->访问数据分布->可疑路径->可疑日

进分析，实现风险智能追踪。

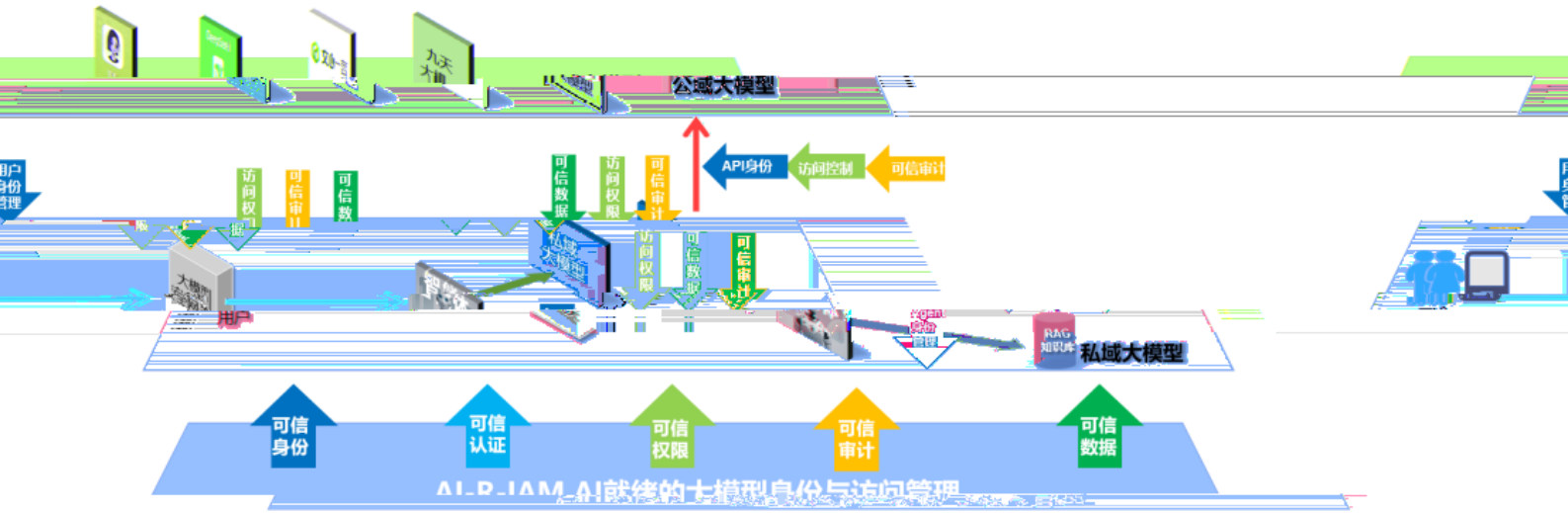
智能追踪溯源：

踪溯源，采用路径计

志链路分析，层层递

用户-AI 智能体的安全应用管理场景

5.4 场景三：



场景概述：在用户到智能体的安全应用场景中，AI-R-IAM 赋予用户与智能体身份并绑定，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。通过 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**

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（AI Identity），由统一身份 ID 进行绑定管理，便能在审计层面明确，如果智能体身份做了某项操作，系统可追溯到是谁让智能体进行操作，方便界定相应责任。

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码及模型紧密绑定，运用数字签名或证书技术，明确证明某个智能体是由预先指定的模型和代码构建，并成功部署

识库接收到该智能体所发出的请求时，能够对其携带的签名、证书或公钥进行严

此来精准确认该智能体的真实身份，同时评估其可信度，从而确保系统运行的安

性。

认证与授权降低威胁面

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

保前端应用访问可信，同时管理员利用 AI-R-IAM 权限管理划分最小访问权限

当 RAG 知识

格验证，以此

全性和可靠性

2. 强

在访问智

名身份，确保

同时，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，通过金库（有效时间段）模式来获取有限访问权限，将潜在攻击面降至最低，同时通过环境属性（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，监管部门可记录并可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

依法申请规则

击后可能执行恶意操作，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

交敏感数据到外部。

新风险，如自动提

AI-R-IAM 提供用户和智能体，智能体与数据存储服务器，RAG 知识库，

在此场景下，

的数据权限边界，实行最小化策略，制定智能体对 RAG 的数据访问权限

内外部大模型之间

栏、上下文检测等安全控制，可及时阻断和审计智能体异常行为等。避免

范围，辅以模型护

引发大规模安全事故。

因智能体自主决策

6 发展趋势展望

在大模型技术蓬勃发展的当下，AI-R-IAM 为其深度应用筑牢安全防线，从多维度构建一体化全程可信能力，引领大模型安全从“单点可控”迈向“一体化全程可信”。

1、在身份欺诈领域，AI-R-IAM 将实现重大技术飞跃。随着大模型在身份认证中的深度应用，其能够对海量的身份数据进行学习和分析，构建起精准的身份行为模型。通过持续监测用户的登录习惯、操作行为模式以及设备使用情况等多维度信息，大模型可以实时感知异常行为。

2、在网络安全方面，AI-R-IAM 将借助大模型的智能分析能力，同其它网络安全设备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 将对所有网络访问请求进行严格的身份验证和权限检查，即使是内部网络流量也不例外，进一步强化网络安全防护体系。

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 网络的高速低延迟特性，实现更高效的安全数据传输和实时的安全监测。通过持续的技术创新和融合，AI-R-IAM 将不断适应新的安全挑战，为大模型的广泛应用和数字经济的蓬勃发展保驾护航，构建更加安全、可信的数字未来。