

启明星辰 | 大模型应用安全 | 深度应用安全 | 基础应用安全

# 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。

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。

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。

## 免责声明

免责声明

本文档依据现有信息制作，其中内容如有更改，恕不另行通知。

北京启明星辰信息安全技术有限公司在编写本手册的时候，力求保证其准确性、完整性和及时性，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。

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：北京北京市海淀区中关村软件园二期A1号楼启明星辰大厦七层

100193 电话：010-82779088

传真：010-82779000

获取最新技术和产品信息，请访问启明星辰网站。



2010年6月23日,启明星辰在深交所中小板正式挂牌上市。

终端管理、加密认证等技术领域,共有百余个产品型号,并据客户需求不断创新,先后

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、安全专业服务而

多家分支机构,拥有覆盖全国的渠道和

场领导者。目前,公司在全国各省市自治区设立三十  
售后服务体系。

的关怀与鼓励。2000年1月,江泽民、  
公司;2003年1月,胡锦涛总书记亲

长期以来,启明星辰公司得到了党和国家领导人  
李岚清、曾庆红等党和国家领导人亲切视察启明星辰  
切接见了启明星辰公司 CEO 严望佳博士。

布局内重点软件企业,国家火炬计划软  
及拥有最高级别的涉及国家秘密的计算

凭借多年来的潜心研发,启明星辰获得国家规划  
件产业优秀企业,中国电子政务 IT100 强等荣誉,及  
机信息系统集成资质证书。

研究基地,完成包括国家发改委产业

启明星辰目前是我国规模最大的国家级网络安全

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

全面专利和软件著作权 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的

的空白。

在企业层面，张叶以其卓越成就成为推动自主研发产业发展的

在网络安全行业的领

军。张叶作为张叶的创始人，由一个创业者成长为一名企业家。

业领军企业，通过不断

的努力，

张叶在网络安全领域的影响力日益扩大。张叶在网络安全领域拥有多项自主知识产权，

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，张叶先后对政策

性银行中行的服务网络进行合作，张叶的网络安全产品实现 90% 的覆盖率。在电信领域，张

叶作为中国移动、中国联通、中国电信三大运营商的网络安全总代理，安全产品覆盖全国。

作为北京网络安全领域的领军企业，张叶的产品及服务广泛应用于政府、企业、金融、教育、

和安全保障，张

叶安全供应商，张叶是公安部授权的网络安全服务提供商，全面负责国家主体网络安全

保障的网络安全服务提供商。张叶的网络安全产品广泛应用于政府、企业、金融、教育、

大型企业网络安全保障。

张叶的网络安全产品广泛应用于政府、企业、金融、教育、

张叶的网络安全产品广泛应用于政府、企业、金融、教育、

张叶

张叶的网络安全产品广泛应用于政府、企业、金融、教育、

设施的安全性和生产效能，为打造和

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

张叶国际化的民族信息安全产业/第二只眼睛不懈奋斗。

## 2 大模型的发展与风险

### 2.1 新质生产力“十样刑”

领域的核心驱动力。从 ChatGPT,

近年来,大模型技术取得了突破性进展,成为人工智能

ek 等,大模型在自然语言处理、

再到百度文心一言、阿里通义千问、九天大模型、DeepSee

能力,甚至实现文本生成、智能问答等任务。

够学习到丰富的语言知识,语义理解和逻辑推理

多样化的任务。大模型发展的历程分为四个阶段:

生成、代码编写等

#### 2022年11月)

#### ● 准备期 (

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

2022年11月

促使各方加大在该领域的投入与研究。

大模型发展浪潮

#### 中长期 (2023年)

#### ● 成

首先, OpenAI 推出 GPT-4, 进一步提升 GPT 模型在复杂任务中的表现, 并拓展其在医疗、金融、教育等领域的应用。

(1) 5

自身技术优势; 阿里发布通义千问, 立足丰富业

(2) 国内: 百度推出文心一言, 整合

产学研, 高校和科研机构纷纷开展研究, 为

发展, 到工信部推出行业大模型, 在注

#### 模型发展基础技术储备。

到, 微软在 2023 年 11 月推出 Copilot, 将 AI 助手集成到 Office 应用中。

图 2-1-1 大模型发展时间轴

4日, 在 2024 中国移动人工智能生态大会上, 中国移动宣布发布

(2) 国内: 2024年5月12

模型, 构建人工智能新生态

了“九天”人工智能体系, 包含千万级模型集群、千亿多模态大

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，

## 应用中安全挑战

## 2.2 大模型应用

### 数据隐私保护的挑战

### 安全风险

数据隐私保护的挑战

安全风险

根据《原星》调查，78%的高管同意在未来必须为 AI Agent 构

根据埃森哲的《2025 年技术

身份验证的挑战

身份验证的挑战

Identity, NHI)，都暴露出诸多安全问题。

## 类身份代表自然人身份实体，在十模型应用在巨场中的非

### 1. 类身份安全问题：人

### 2. AI Identity 安全问题。越来越多的 AI Agent 在 IT 治理框架下，AI Agent 具

### 3. 类身份管理原则和最佳实践

### 《人工智能治理白皮书》中提出 AI Agent 应

AI Agent 应视为具有身份的数字身份新兴技术，应参照人类身份作为员工，在大

性。可能会出现如下风险：

攻击者可能通过 AI Agent 访问所有访问的模型进行攻击。如通过 AI Agent

攻击者可能通过 AI Agent 访问所有访问的模型进行攻击。如通过 AI Agent

的访问，防止未经授权的身份访问或滥用。

攻击者可能通过 AI Agent 访问所有访问的模型进行攻击。如通过 AI Agent

攻击者可能通过 AI Agent 访问所有访问的模型进行攻击。如通过 AI Agent

非法访问权限。

(2) 访问权限放大：与人类身份不同，AI Agent 不会以可预测的方式请求访问权限，

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，从而造成访问

露或系统破坏。

和操作敏感资源，造成数据泄

身份系统通常存储大量用户的个人信息、行为数据等。如果

(3) 身份信息：AI Agent

隐私泄露，进而可能被用于精准诈骗、

系统遭受黑客攻击，这些数据可能被窃取，导致用户

定向广告骚扰等。

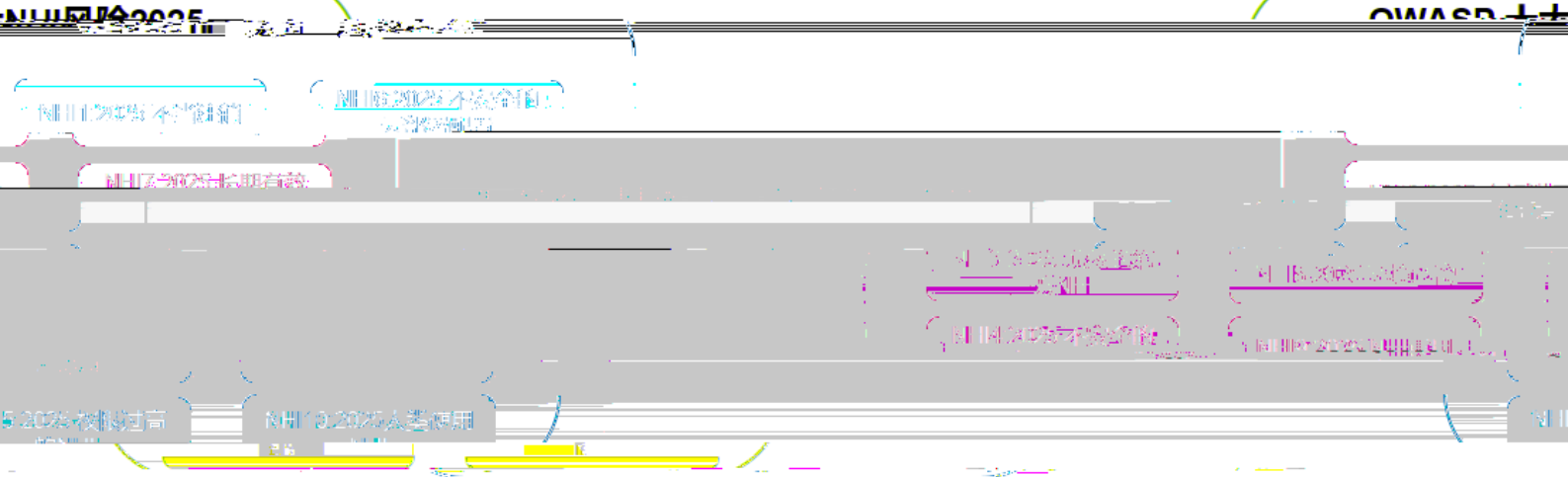
### 3、非人类身份 (NHI) 安全问题：非人类身份

密码、服务帐号、OAuth 令牌、API

机器等实体绑定的数字身份。这包括机器人、API

2025》。如下图所示。

2025 年 2 月发布《数字身份风险



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，NHI 的身份生命周期管理薄弱，存在诸多安全风险与问题：

- (1) 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。
- (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。
- (3) 缺乏账号所有权：多数参与大模型应用的组织中，NHI 无明确所有权信息，而明确

管理者对安全维护和问题补救很关键。

杂，难以识别有依赖关系。

## 2.2.2 数据安全挑战

大模型应用在数据安全方面，面临如下挑战：

(1) 用户输入输出风险：用户输入可能包含敏感信息，若大模型系统对输入输出

不完善，敏感信息可能在输入环节直接暴露。在输出结果时，若对输出

结果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，若对输出

结果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，若对输出

存在漏洞，不法分子

(2) 大模型 API 调用风险：API 调用过程中，若身份认证和授权机制

API 接口若缺乏有效

验证，攻击者可能冒用合法身份调用 API，获取敏感数据或进行恶意操作。同时，A

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取 API 调用权限，

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

非法获取敏感数据或进行恶意操作。

非法获取敏感数据或进行恶意操作。

非法获取敏感数据或进行恶意操作。

(4) 数据合规风险：在大模型训练阶段，企业

非法获取敏感数据或进行恶意操作。

非法获取敏感数据或进行恶意操作。

非法获取敏感数据或进行恶意操作。

非法获取敏感数据或进行恶意操作。

非法获取敏感数据或进行恶意操作。

(5) 合

非法获取敏感数据或进行恶意操作。

非法获取敏感数据或进行恶意操作。

非法获取敏感数据或进行恶意操作。



、行为风险、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

风险的全链路安全风险审计。

围扩展到安全设备、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

扩展到全网设备、员工操作行为、可信资源等，通过“ID、自公”的绑定形成可信身份

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，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。

身份、

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

化全程

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

可信的数字身份基座，即 AI 就绪的大模型身份与访问管理。

# 大模型身份与访问管理

## 4 AI 就绪的

身份管理(以下简称: AI-R-IAM)是为大规模人工智能模型(如

AI 就绪的人模型身

身份管理(以下简称: AI-R-IAM)是为大规模人工智能模型(如

AI 就绪的人模型身

从身份、访问、行为、数据等维度为人工智能模型训练、推理和应用提供安全保护

提供安全

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

构建一体

为其它安全能力提供身份管理属性支撑。

系, 同时



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，有效防止未经授权访问和潜在的安全威胁。

先进的身份认证的用户名和密码防止未经授权的访问

中国移动安全运营中心

### ● 可信访问权限管理

化的权限

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

管理策略，系统可按照不同策略对不同的角色进行授权，并支持基于

策略的访问控制，系统可按照不同策略对不同的角色进行授权，并支持基于

策略的访问控制，系统可按照不同策略对不同的角色进行授权，并支持基于

策略的访问控制，系统可按照不同策略对不同的角色进行授权，并支持基于

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。

中国移动安全运营中心

中国移动安全运营中心

### ● 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

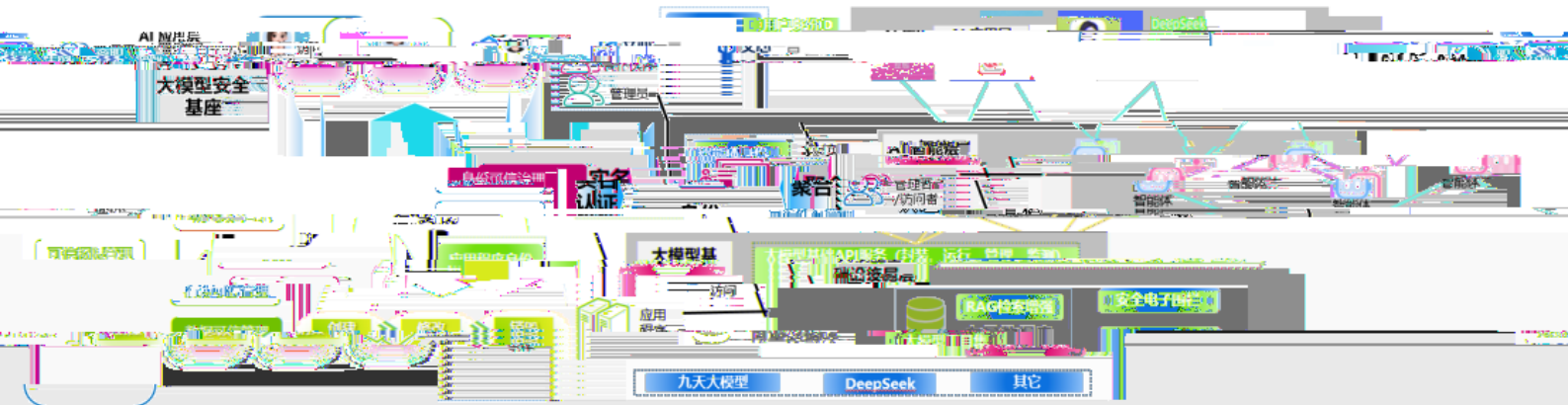
控，记录用户的所有操作行为，包括用户登录时间、IP 地址、访问的大模型资源、执行的  
操作等信息，实时监测用户和组件的访问行为，及时发现异常风险。

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，

为其应用接入模型的设备，提供身份认证、授权、审计等方面的可信能力支撑，为支持上

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围，聚合人类身份、AI Identity 以及非人类身份（NHU）

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 应用程序建立

的管理，是数字世界秩序的底层支撑。

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 采用了先进的身份认证

安全威胁

有效防止未经授权访问和潜在的

不仅涵盖传统人类身份从入职到离职的全生命周期管理，还创新性地

AI-R-IAM 不

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

和 AI Identity

流程如下：

规性。在具体

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：

(1) 用户身

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。

(2)

、回收资源并处理数据；

销毁时注销身份

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

(3) 应用程序身份在

删除身份、销毁前清理数据。

即可提供友好的用户可管理性访问。

用户可以查看和编辑其自己的权限管理，管理自己的权限。

## 权限管理

## 4.2 可信

合理权限，系统可以根据其管理自己的权限管理。

提供了多层次、精细化的权限

系统可以根据其管理自己的权限管理。

风险。这种精细化的权限管理不仅提高了系统的安全性，还提升了用户的操作体验。

### ○ 用户角色权限

系统可以根据其管理自己的权限管理。

系统可以根据其管理自己的权限管理。

### ○ 用户角色权限管理

系统可以根据其管理自己的权限管理。

系统可以根据其管理自己的权限管理。

系统可以根据其管理自己的权限管理。

算法工程师	✓	✓	✓	×	×	×
-------	---	---	---	---	---	---

系统可以根据其管理自己的权限管理。

业务用户	×	受限调用	×	×	×	×
------	---	------	---	---	---	---

### ● AI 接口权限管理

包括：进口接口策略、接口鉴权、接口连接保护、接口流量、接口传输数据加密、接口

设备管理

黑白名单等进行控制

WAG（检测与响应引擎）- 权限控制

权限控制：不同部门只能访问其业务相关的知识库；

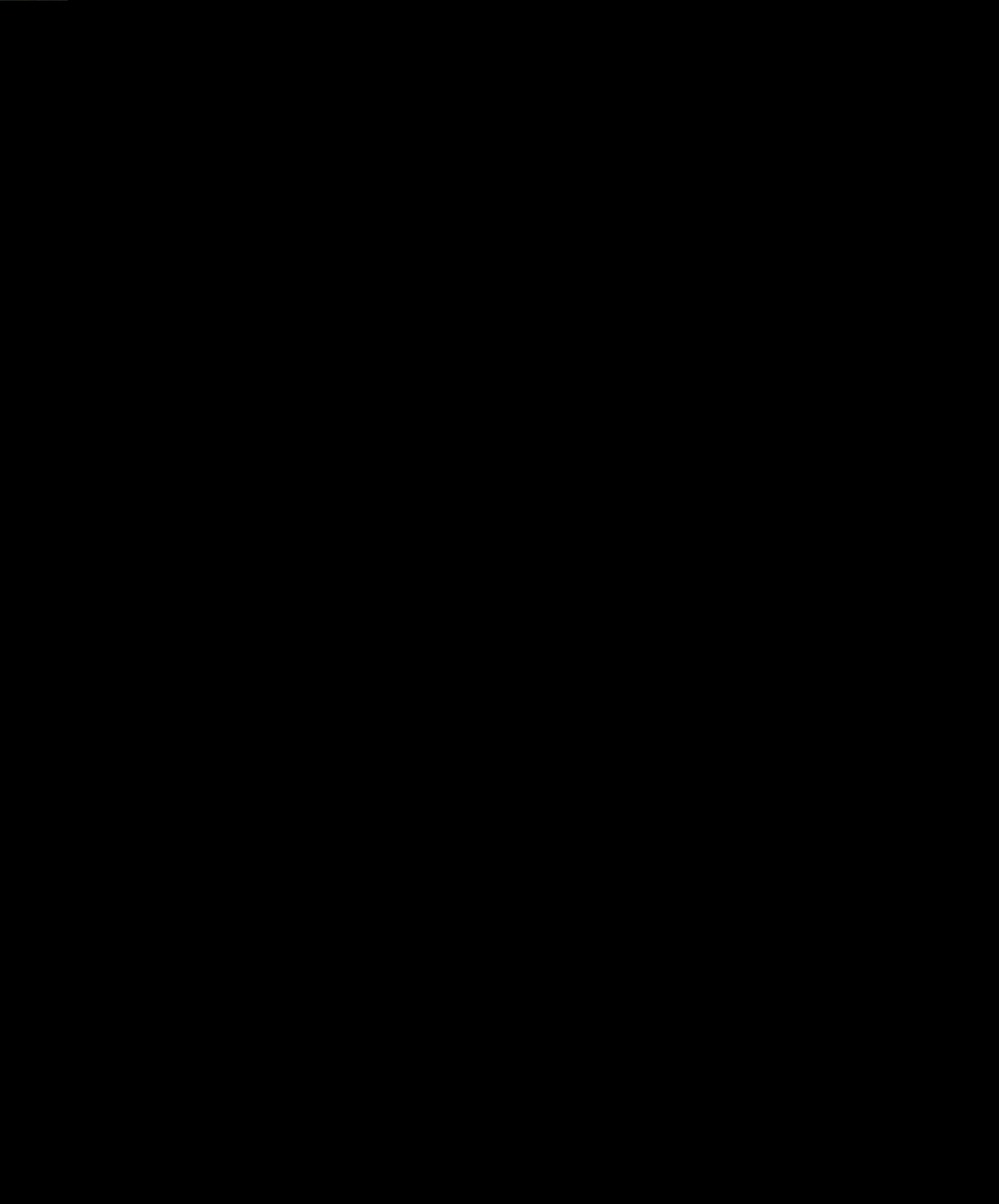
文档

权限控制与审计等；

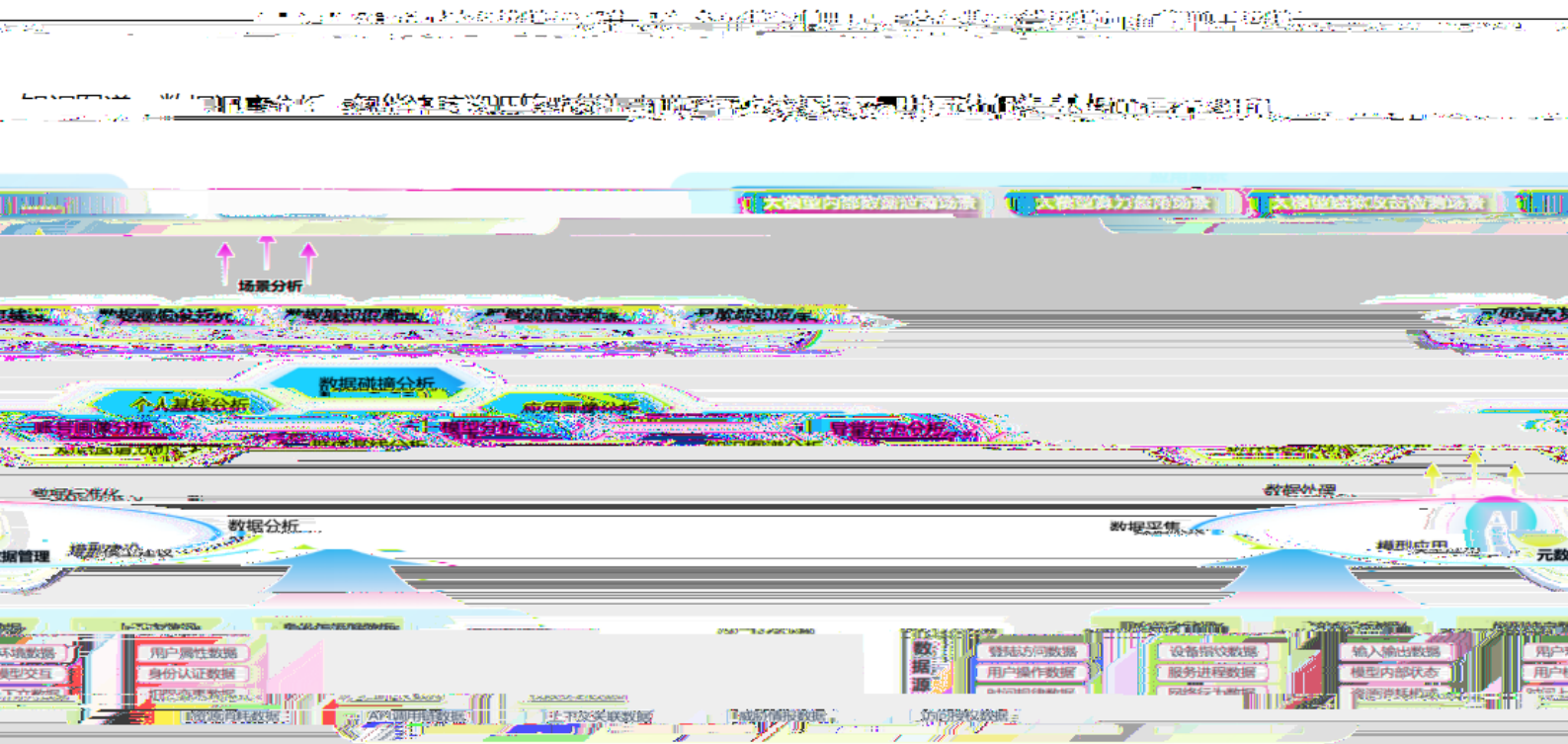
内容安全策略：通过内容识别引擎，识别识别内容敏感内容

结果（如涉及敏感信息的搜

实时内容过滤：结合语义分析引擎，自动屏蔽无权限的检索



## 4.4 可信行为审计



- 多维数据采集

构建全面的用户行为基线分析提供支撑

- 安全元数据管理

分发，提升数据价值密度。

- 可信行为基线

安全运营各个环节综合利用技术，提升可信行为基线分析，提升大模型应用可信行为基线分析。

可信行为基线分析，提升可信行为基线分析。

- 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、个人偏好，生成用户画像、模

型画像、数据画像等，并根据风险画像平台，智能化风险水平。

- 数据知识图谱

通过数据多源融合、

关系构建，提供从“关

- 数据深度分析

+模型数据全生命

景分析等，对大模型

- 智能追踪溯源

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

系”的角度分析大模型风险的能力。

定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

型安全风险进行评估，识别已知风险和未知风险。

溯源

模型输出提供合规风险溯源，通过可信计算等技术实现数据溯源

可溯源路径>可疑日志溯源分析，以溯源分析，实现风险溯源

可疑IP>涉网数据分布>

溯源

# 应用场景

# 5

## 十类重点应用最新进展

## 5.1

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、鉴别、审计等

能力，和以场景为基础，用

场景实现可信身份、可信认证、可信授权、可信审计、可信数据

域大模型场景：业务系统调用信创、RAG知识库、私域大模型场

户访问私域大模型、公域

域大模型训练/推理场景。

户；大模型数据训练场。



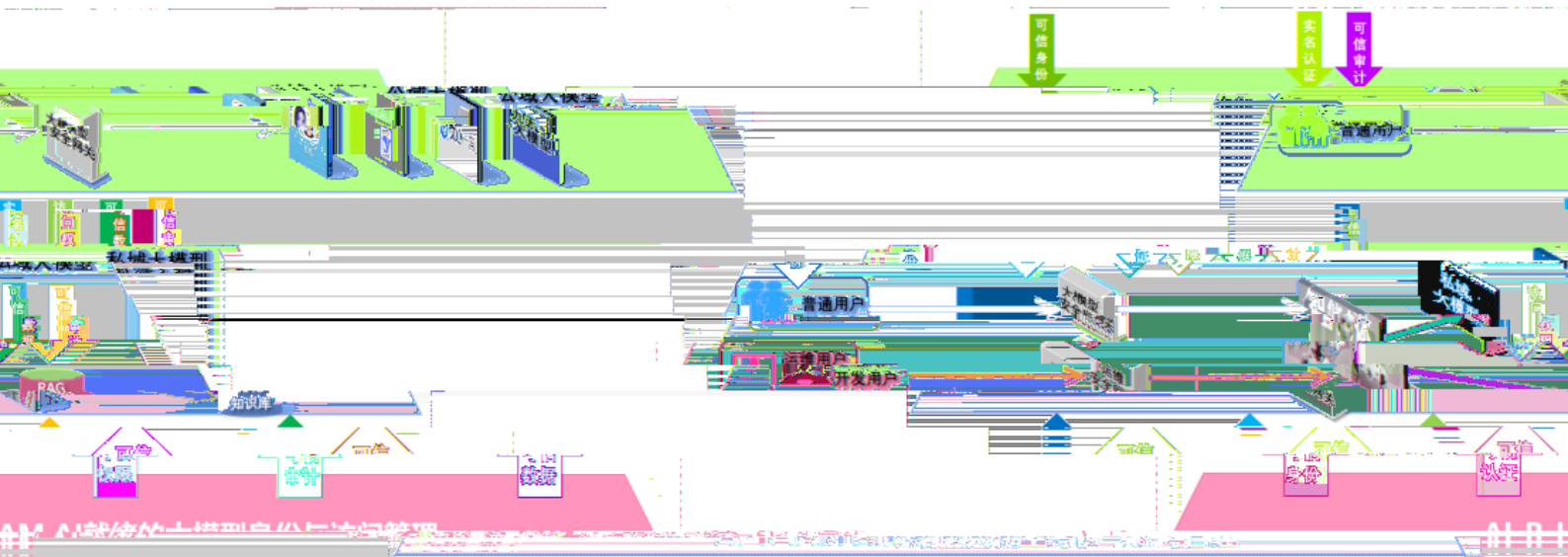
## 5.2 场景一：用户访问大模型应用场景

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

身份和访问管理，保障大模型应用的安全可信，提升大模型应用的安全可信能力。

运营维护人员访问私域大模型等。

户访问公域大模型、用户访问私域大模型、



## 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，像 PC、手机

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用。这些大模型应用通

的物体，提取带人物特征等信息，语音转文字服

内容：图像识别服务可以精准地截取图片中

务则可以提取图片中的语音文件准确识别并转换为对应的文字内容

的标签或为客户端提供一站式服务，新用户无需注册账号

次即可使用，从而降低使用门槛

要利用文本生成服务撰写一份产品推广文案时，只需打

非常简便。比如，当一个企业员工想

生成企业官网的帮助中心内容，只需输入

关键词，即可生成相关内容，用户

务便捷。又如个人用户使用图像识别服务识别一张拍

成按钮，很快就能识别出一份招聘启事

的操作，就能从 Web 界面上获取到详细的购物分

物属性等信息时，也是通过简单的上传

类结果。

用户数据安全系统建设至关重要。AI-R-IAM

然而，在这种复杂的业务场景中，保障

系统可信身份认证及安全审计的能力。

在智能办公和公域大模型场景中，主要

访问大模型应用时，

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

在指纹识别或面部识别

系统会要求用户提供身份认证，这可能包括用户名和密码组合、生物特征（

身份信息等方式，只有在经过严

密）或者是基于可信凭证（Verifiable Credential, VC）的

管系统通过安全策略引擎，能够实时监测异常登录行为，一旦发现异常，立即

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，并定期生成审计报告。此外，系统还支持与外部安全审计系统对接，实现

数据同步和共享。通过这种方式，企业可以全面掌握系统的安全状况，及时发现

触发警报提示管理员进行进一步调查，从而

了大量不符合常规模式的数据访问请求，就可以

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

普通

场景）与公共大模型访问场景相比，私域场景在数据隐私保护和合规性要求方面有着更高的

要求。企业需要确保用户数据的安全性和完整性，防止数据泄露和滥用。同时，企业

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

不

业需要根据业务需求和数据敏感性，制定合理的权限策略，确保不同角色的用户只能访问其

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；包括各个部门的数据汇总；部门

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的调整；普通

员工只能访问与自己工作直接相关的数据，如客户信息、销售记录等。同时，企业还应

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

限与安全管理人员在爱河环境管理、人员管理、运维管理等方面

上，还需要从权限和权限管理、系统软件、补丁管理等方面能力。

应该只给访问、操作和权限管理权限，而不能给到存储在那类数据上的敏感数据。

授权和权限管理、权限、软件给他人，也不能给到其修改和删除权限。

在运营维护人员、安全管理

在运营维护人员、安全管理

日志访问、日志某些非敏感服务等。

师可能只能做一些早期的、风险评估的操作，如查看

在运营维护人员、安全管理

十级别的日志记录等等。通过这种方式，可以降低运维人员权限过高的风险。

自己写的公众号

在运营维护人员、安全管理

在运营维护人员、安全管理

在运营维护人员、安全管理

在运营维护人员、安全管理

非工作时间或者从企业外部网络尝试访问时，就需要触

大模型应用服务器。然而，如果是在

发更加严格的验证机制。

在运营维护人员、安全管理

在运营维护人员、安全管理

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

权限配置等。一旦识别出高

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

层次。在第一层次是警告提

让这组人员只能浏览非敏感数据,避免因误操作而产生的风险。第一次如果输入失败,可以输入正确的用户名和密码,第二次输入失败,则需要输入验证码。

让这组人员只能浏览非敏感数据,避免因误操作而产生的风险。第一次如果输入失败,可以输入正确的用户名和密码,第二次输入失败,则需要输入验证码。

因为这组人员需要继续访问敏感数据,所以需要输入额外的权限验证步骤。这可以手动输入

验证码,也可以输入验证码,验证码由系统生成,验证码由系统生成,验证码由系统生成。

验证码由系统生成,验证码由系统生成,验证码由系统生成,验证码由系统生成。

验证码由系统生成,验证码由系统生成,验证码由系统生成,验证码由系统生成。

验证码由系统生成,验证码由系统生成,验证码由系统生成,验证码由系统生成。

验证码由系统生成,验证码由系统生成,验证码由系统生成,验证码由系统生成。

验证码由系统生成,验证码由系统生成,验证码由系统生成,验证码由系统生成。

验证码由系统生成,验证码由系统生成,验证码由系统生成,验证码由系统生成。

验证码由系统生成,验证码由系统生成,验证码由系统生成,验证码由系统生成。

验证码由系统生成,验证码由系统生成,验证码由系统生成,验证码由系统生成。

验证码由系统生成,验证码由系统生成,验证码由系统生成,验证码由系统生成。

验证码由系统生成,验证码由系统生成,验证码由系统生成,验证码由系统生成。

验证码由系统生成,验证码由系统生成,验证码由系统生成,验证码由系统生成。

验证码由系统生成,验证码由系统生成,验证码由系统生成,验证码由系统生成。

验证码由系统生成,验证码由系统生成,验证码由系统生成,验证码由系统生成。

验证码由系统生成,验证码由系统生成,验证码由系统生成,验证码由系统生成。

验证码由系统生成,验证码由系统生成,验证码由系统生成,验证码由系统生成。

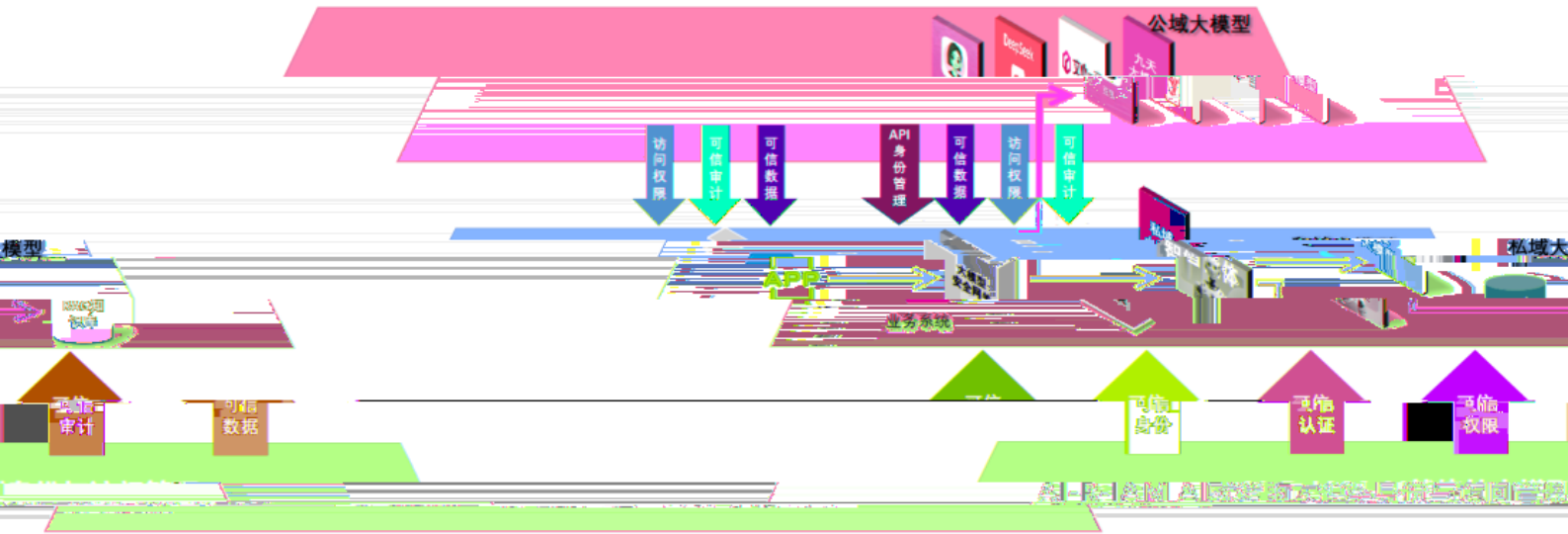
验证码由系统生成,验证码由系统生成,验证码由系统生成,验证码由系统生成。

验证码由系统生成,验证码由系统生成,验证码由系统生成,验证码由系统生成。

验证码由系统生成,验证码由系统生成,验证码由系统生成,验证码由系统生成。

验证码由系统生成,验证码由系统生成,验证码由系统生成,验证码由系统生成。

### 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含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

业务系统身份信息泄露，攻击者可能通过伪造身份来访问大模型服务。

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，如密钥可能被攻击者获取，攻击者可能通过伪造身份来访问大模型服务。

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。

身份管控：业务系统身份注册与认证，通过数字证书或 API 密钥进行认证，并颁发身份令牌。

最小访问权限。

身份修改：需根据功能和任务变更身份权限，更新业务系统身份信息和权限，确保合规并满足业务需求。

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，同时进行关联资源清理和安全检查，确

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

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。由于权限划分不清晰，可

致内部人员误操作或恶意操作，进而影响系统整体安全。

业务系统

大模型

信。

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

属性、状态和业务流程，

实体级授权：通过RBAC和PBAC技术，业务系统根据实体的

业务系统

业务系统

业务系统

业务系统

业务系统

业务系统

业务系统

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

配，实现对资源访问的精细化管理，确保资源访问权限合理的分配和使用。

## 3. 业务行为审计

业务系统

业务系统

业务系统

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

业务系统

AI就绪的大模型身份与访问管理 AI-R-IAM v1.0

AI就绪的大模型身份与访问管理 AI-R-IAM v1.0

AI就绪的大模型身份与访问管理 AI-R-IAM v1.0

智能体溯源：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溯源，采用路径计算、路径搜索算法，提供可信IP、访问数据分布、可解路径、

可信身份、可信认证、可信授权、可信审计

智能体溯源：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

溯源，采用路径计算、路径搜索算法，提供可信IP、访问数据分布、可解路径、

可信身份、可信认证、可信授权、可信审计

AI就绪的大模型身份与访问管理 AI-R-IAM v1.0



AI就绪的大模型身份与访问管理 AI-R-IAM v1.0

AI就绪的大模型身份与访问管理 AI-R-IAM v1.0

管，保障访问数据和训练数据的安全。通

获有限权限，同时通过身份绑定对双方行为审计监

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

(user) 与智能体身份 (AI

通过 AI-R-IAM 提供可信身份管理，赋予用户可信身份

身份。如果智能体身份做了前

Identity) 中统 身份 ID 进行绑定管理，身份绑定管理

谁让智能体进行操作，方便界定相应责任。

项操作，系统可追溯到是

被恶意篡改或仿冒，智能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

签名或证书技术，明确证明某个智能体是由谁指定的模型和代码构建，并成功部署

数字签

行下，在智能体接收到的请求中，包含可信的签名，系统接收后

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。

在访问智能体将是，普通用户通过主模型安全网关与 AI-R-IAM 实名认证能力建立实

身份，系统会记录用户身份，从而将用户与 AI-R-IAM 关联起来，实现身份绑定

同时，认证流程不再只面向普通用户，每一个智能体都将以独立身份进行认证控制，并

在访问智能体业务时进行认证校验，通过令牌 (Short Access Token) 或 一次性密钥

(One-Time Key) 等方式，让授权更时效性和可追溯性。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

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，例如记录了哪些数据，进行了哪些个

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，

数据到外部。

新风险，如自动提交敏感数

AI 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，AI-R-IAM

限边界，实行最小化策略，制定智能体对 RAG 的数据访问权限

内外部大模型之间的数据权

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，辅以模型护栏、上下

模安全事故。

因智能体自主决策引发大规

## 6 发展趋势展望

网络安全态势感知与威胁情报

工业互联网安全态势感知与威胁情报

工业互联网安全态势感知与威胁情报

工业互联网安全态势感知与威胁情报

工业互联网安全态势感知与威胁情报

工业互联网安全态势感知与威胁情报

工业互联网安全态势感知与威胁情报

工业互联网安全态势感知与威胁情报

工业互联网安全态势感知与威胁情报

工业互联网安全态势感知与威胁情报

异常行为。

AI-R-IAM 将供地上进型的智能分析能力，同时实现终端安全设备

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

网络安全防御体系。

工业互联网安全态势感知与威胁情报

工业互联网安全态势感知与威胁情报

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；借助 5G

技术创新

技术创新