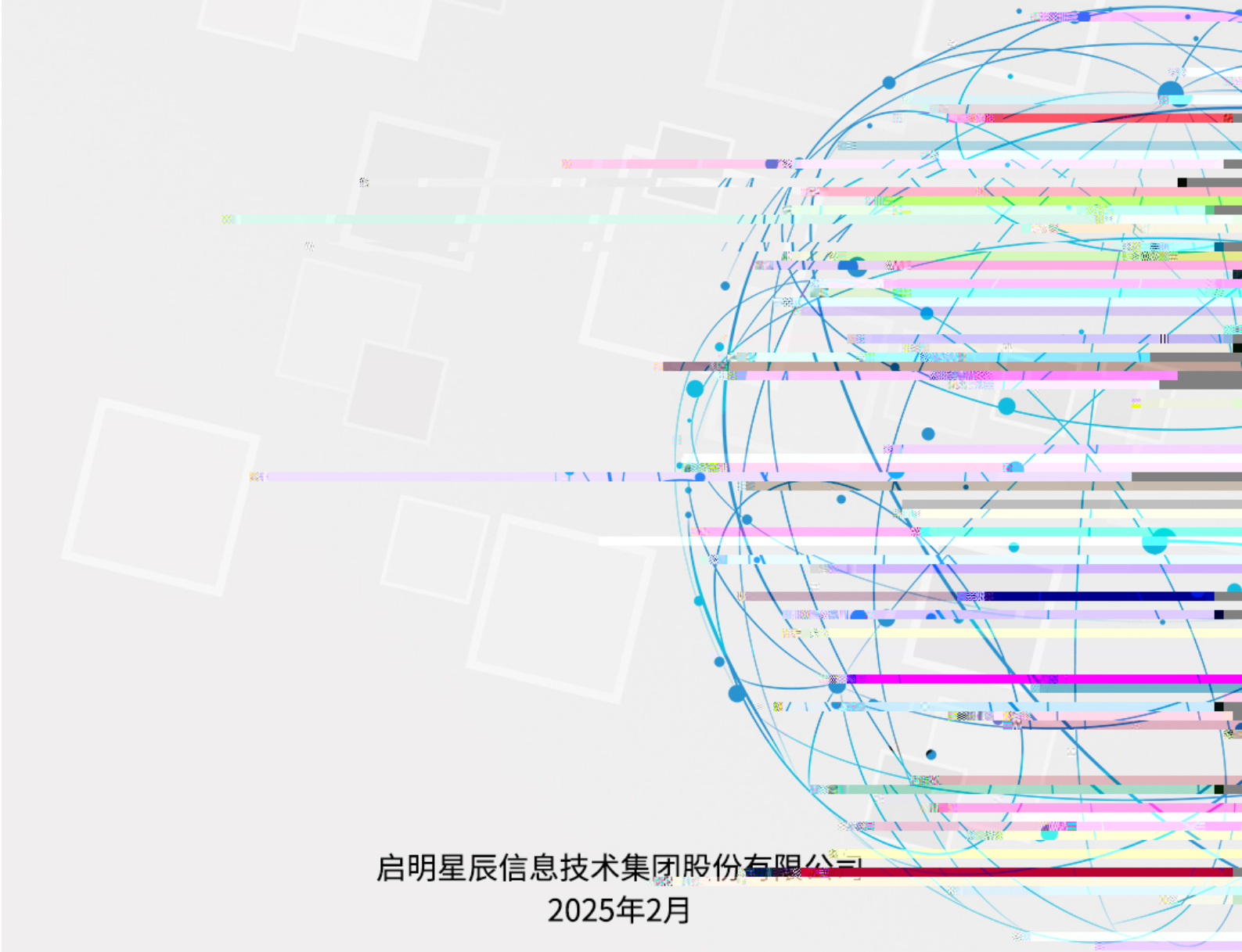
启明星辰

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

解释权 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

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 本文

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

如有不经通知 本文档依据现有信息制作，其内容如有更改，

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， 您可以访问启明星辰网站

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、安全运维服务市

多家分支机构，拥有覆盖全国的渠道和

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

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

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。

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。

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。

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

[illegible]

[illegible]

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

[illegible]

[illegible]

[illegible]

2 大模型的发展与风险

2.1 新质生产力“大模型”

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，成为人工智能领

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

智力，从而实现文本生成、智能问答、图像

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，

成长期（2023 年）

- **成**

合自身技术优势；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

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，

[redacted]

[redacted]

[redacted]

类身份代表自然人身份实体，在大模型应用交互场景中的类

1、人类身份安全问题：人

2、AI Identity安全问题：越来越多的AI Agent在调用企业内部 AI Agent 以

AI Agent 视为其在身份和权限管理（IAM）方式，例如将 AI Agent 作为以下一种人

从一个典型的攻击

时，攻击者可能利用 AI 应用所使用的模型进行攻击，如通过精心构造

的输出，从而绕过身份验证获取

攻击向模型输入精心构造的恶意数据，使模型产生错

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，从而能够访问

和操作敏感资源，造成数据泄

露或系统破坏。

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）身份安全：AI Agen

中隐私泄露，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。

定义为企业技术内的应用程序、服务或

3、非人类身份（NHI）安全问题：非人类身份

密钥、服务帐号、OAuth 令牌、云服

机器等实体绑定的数字身份，这些包括机器人、API

2025 年 2 月发布《大大 NHI 风险

2025》，如下图所

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) 没有不活跃凭证轮换策略与清理机制：大模型应用中，许多 NHI 凭证长期不轮换，一旦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。

(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。

(4) 缺乏环境隔离：在许多情况下，相同 NHI 会在生产和非生产环境中使用，或者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。

(5) 缺乏依赖关系：大模型应用的逻辑 NHI 在

2.2.2 数据安全挑战

大模型应用在数据安全方面，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

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制不

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

存在漏洞，不法分

(2) 大模型 API 调用风险：API 调用过程中，若身份认证和授权机制

PI 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。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

(4) 数据投毒风险：在大模型训练阶段，攻

、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

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、

化全程

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

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

[illegible]管理(以下简称：AI R IAM)，为人类用户、智能体（AI

[illegible]从“身份、访问、行为、数据”维度入手，贯穿模型训练、推理和应用[illegible]多个环节，

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

对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

先进的身份认证

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

的用户名和密码

问和潜在的安全威胁。

止未经授权的访

- 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

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。

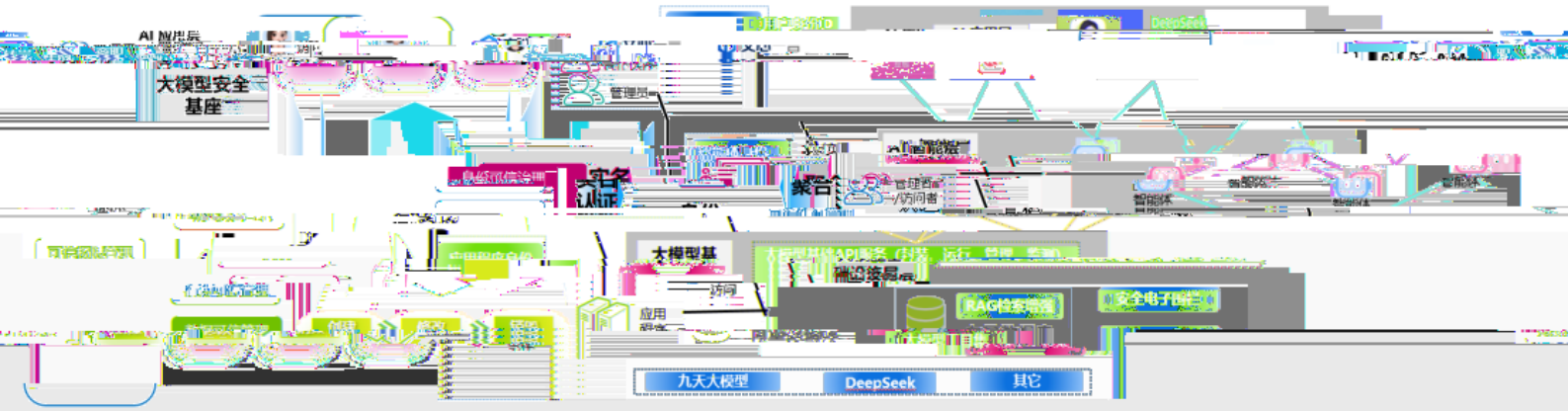
- 可信审计管理

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，

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

公 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）

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

，有效防止未经授权的访问和潜在的

优化传统人类身份从入职到离职的全生命周期管理，还创新性地将 NHI

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

(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；

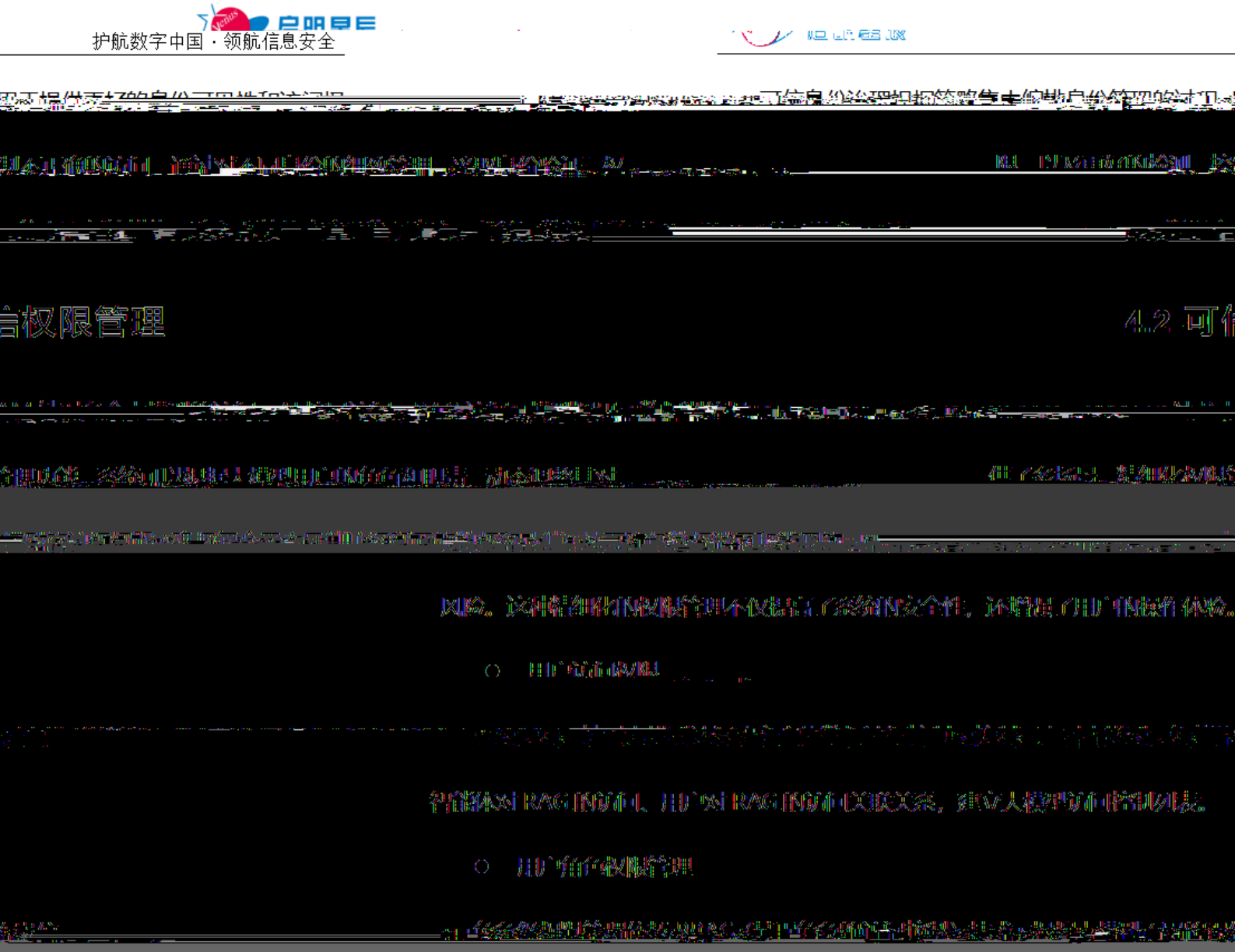
、回收资源并处理数据；

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

销毁时，注销相关凭证。



算法工程师	✓	✓	✓	×	×	×
业务用户	×	受限调用	×	×	×	×

- AI 接口权限管理

包括上报型接口策略、接口密钥、接口连接调用、接口读写、接口传输数据内容、接口

最小化管理

基于角色的最小化权限管理

权限控制：不同角色只能访问其业务相关的知识库内容

对敏

结果（如脱敏后的信息）

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

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

、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

通过数据多源融合、

系”的角度分析大模型风险的能力。

关系构建，提供从“关

- 数据深度分析

围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

大模型数据全生命

型安全风险进行评估，识别已知风险和未知风险。

景分析等，对大模型

溯源

- 智能追踪溯

[illegible]

[illegible]

[illegible]

应用场景

5

大模型应用场景数据流

5.1

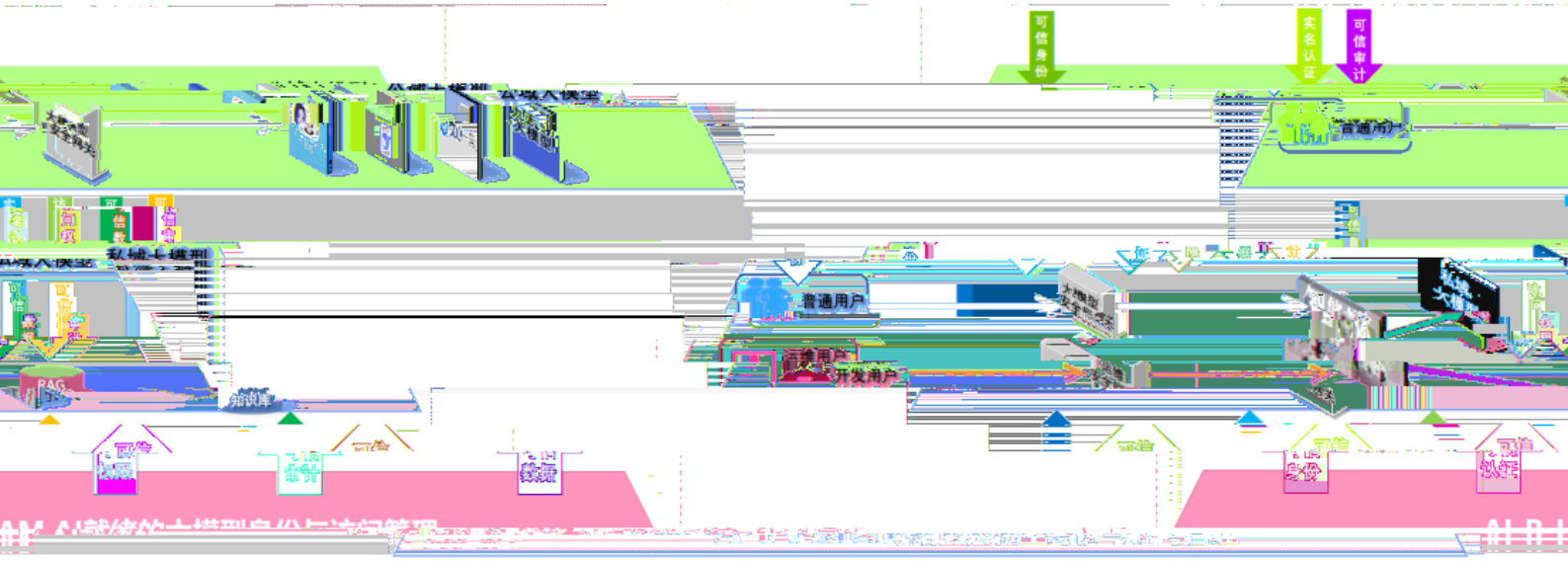
AI-R-IAM AI就绪的大模型身份与访问管理为大模型的访问、使用、调用、训练等各

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

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

或者平板等，去访问那些部署在公有云上的大模型应用。这些大模型应用涵盖多种类型的服

务，例如文本生成服务，能够根据用户输入的关键词或简短描述，自动生成连贯且有逻辑的

文章、报告或邮件；图像识别服务，可以精准地分析图片中的物体、场景或人物特征等信息；语音转文字服务，则可以把用户录制的语音文件准确无误地转换成相应的文字记录。

要利用文本生成服务撰写一份产品推广文案时，只需打开相应的应用程序或网页，登录

非常简便。比如，当一个企业员工想要

系统后，又如个人用户使用图像识别服务前，需在

成后，很快就能得到一份初步的文案

Web 界面上选择上传图片的操作，就能从

物的详细分析报告，也是如此简单直观。

类型。

用户数据安全和系统的正常运转。AI-IAM

然而，在这种便利服务背后，保障

进行身份认证及安全审计等。

在普通用户访问公域大模型时，主要

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

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

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

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

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

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

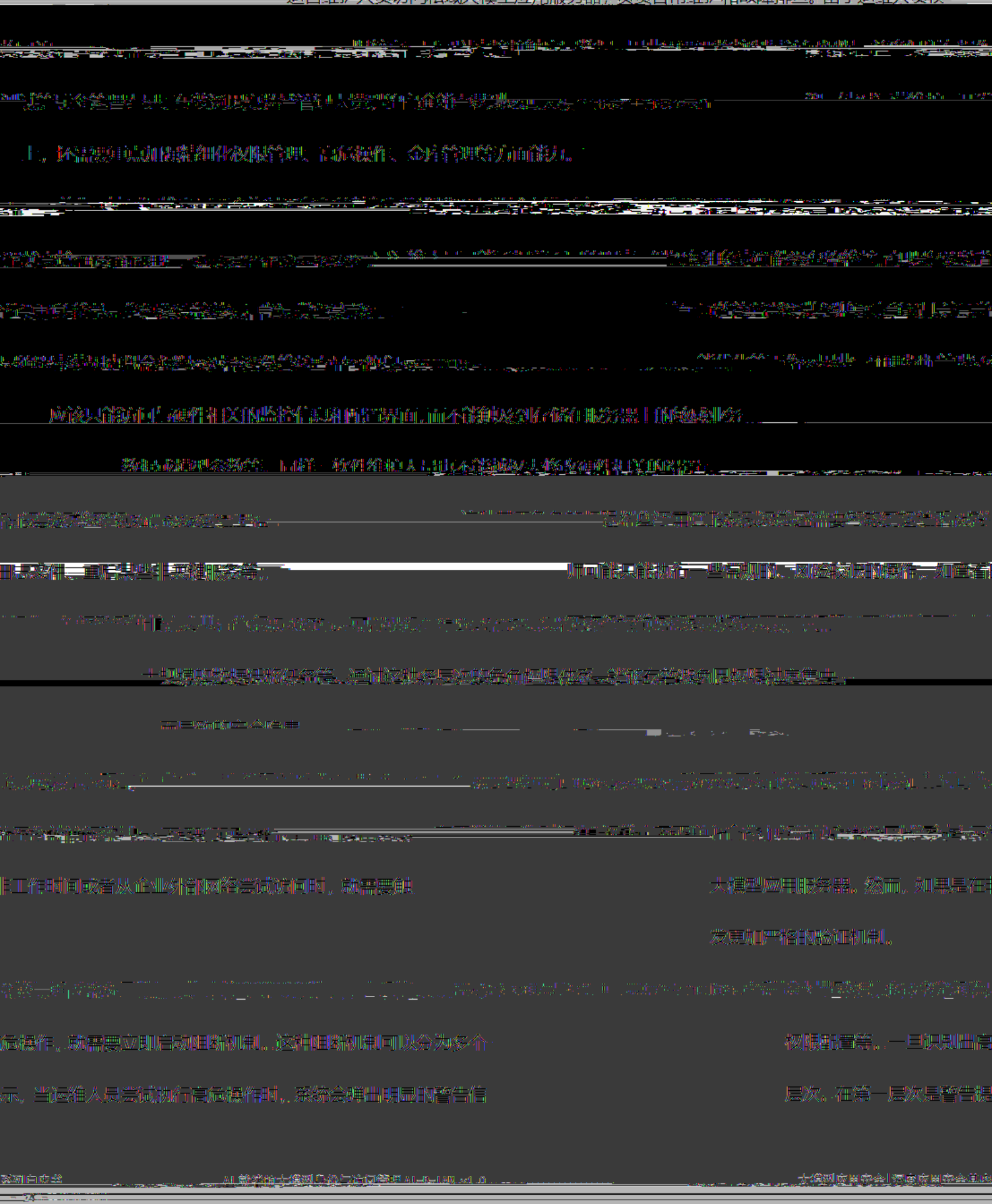
员工的权限最为有限，只能访问与

并且能够对本部门员工的权限进行一定程度的调整；普

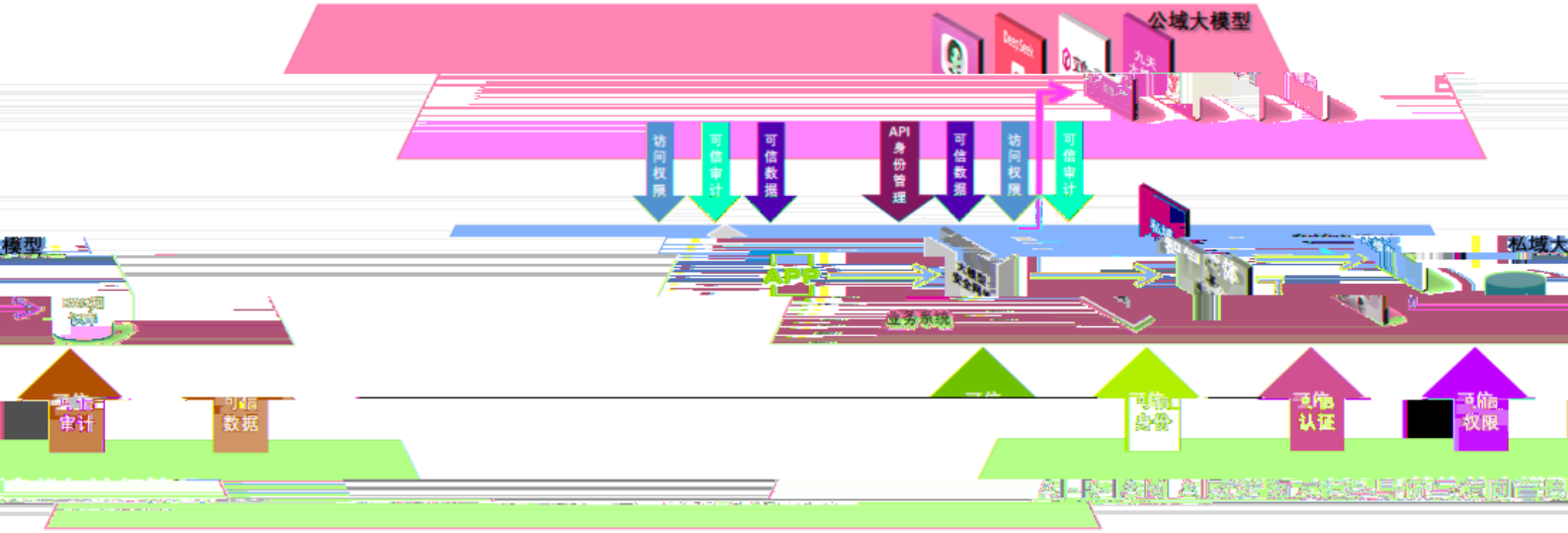
性和准确性，以满足不同角色的需求同时保障数据安全。

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



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

业务系统的身份信息安全

业务系统访问公域、私域大模型身份主要采

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 如密钥可能被

进而滥用资源或进行恶意操作 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 AI-R

创建：对业务系统身份进行认证，通过数字证书或 API 密钥进行认证，并授予 身份

最小访问权限。

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

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。

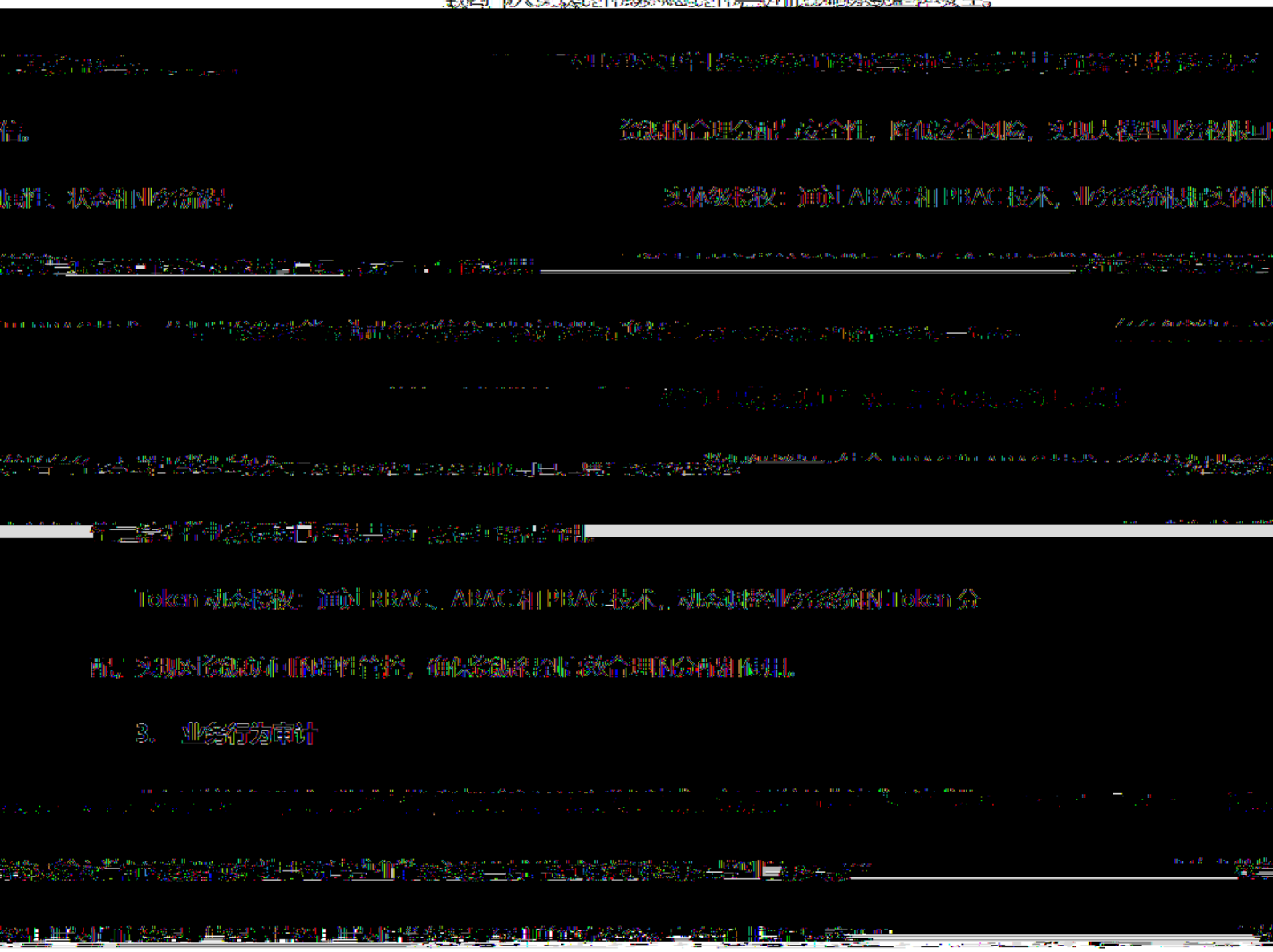
2. 业务权限管控

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。



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

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

Identity）中统一身份ID进行绑定管理，确保密钥与

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

被恶意插包或仿冒，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

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

数字签

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

份，确保每次操作的可追溯性。AI-R-IAM 将普通用户与 AI-R-IAM 的身份认证机制

时，认证流程不再以简单的用户名/密码为主，每一个智能体都会被赋予身份进行访问控制，并

在执行任何敏感操作前进行认证校验，通过短时令牌（Short Access Token）或一次性密钥

（One-Time Key）等方式，让授权过期或使用一次后即失效。当智能体调用 RAG 时，AI-R-IAM

（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，监管部门可记录并回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。

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

M 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

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

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