

启明星辰+大模型+网络安全+零信任+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



版权声明

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手册的最终解释权。

和修改

和修改

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

本

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

本文档体现所有信息制作，其中容如有更改，恕不另行通知。

本手册的编写人员是本着为用户提供准确、可靠、实用的信息而编写的，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。

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：

信箱：北京市西城区右安门西路6号启明星辰信息公司 邮编：100193

100193 电话：010-82779088

传真：010-82779000

只信息

你可以访问启明星辰网站

获得最新技术和产品

● 公司 100% 控股的子公司

2010年6月23日, 星期一

Figure 1. *Phylogenetic tree of the 16S rDNA sequences of the 10 isolates. The scale bar represents 0.01 substitutions per site. The numbers at the nodes indicate the bootstrap values. The sequences of the 16S rDNA of *Escherichia coli* (F) and *Salmonella enterica* (S) were used as outgroups.*

[illegible]

发展成果

切接见了启明星辰公司 CEO 严望佳博士。

启阳早后日前早我国规模最大的国家级网络安全研究基地 完成句坪国家发改委产业

示范工程, 国家科技

余项专利和软件著作权 参与制订国家及

Table 1. Antibiotic resistance patterns of *Salmonella* isolates

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

性銀：1、中長期商業銀，合：短期拆借商業銀，實現90%的

竹类植物组织培养与快速繁殖技术

【**例題**】 以下の文をよみ、問に答えなさい。

有四個小孩。

全产品和最佳3

相立国にレ向テ

2 大模型的发展与风险

2.1 新质生产力“十楷模”

来源：新华社、人民日报

智能领域的核心驱动力。从 ChatGPT, GPT-4, Gemini, Claude, LLaMA, DeepSeek 等，大模型在自然语言处理、

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重要突破。从谷歌的 Gemini 再到百度文心一言、阿里通义千问、九天大模型、DeepSeek

模型在自然语言处理、图像生成、代码生成等方面展现出强大的能力。

模型在自然语言处理、图像生成、代码生成等方面展现出强大的能力。

推理能力,从而显著提升文本生成、智能问答、图像

够学习到丰富的语言知识,深入理解和逻辑推理能力,从而显著提升文本生成、智能问答、图像

生成、代码生成等多种多样的任务。大模型发展也经历了四个阶段:

○ 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惊全球, 开启了

大模型发展热潮, 促使各方加大在该领域的投入和研发。

大模

○ 成长期 (2023 年)

2023 年, 大模型技术快速发展, 多家企业纷纷推出自己的大模型, 市场竞争日益激烈。

图 2-1

国内技术优势: 阿里发布通义千问, 立足丰富数据

(2) 国内: 百度推出文心一言, 整合

模型发展, 但技术储备。

来源: 艾瑞咨询研究院

图 2-1 大模型发展现状

国内技术优势: 阿里发布通义千问, 立足丰富数据

国内技术优势: 阿里发布通义千问, 立足丰富数据

国内技术优势: 阿里发布通义千问, 立足丰富数据

国内技术优势: 阿里发布通义千问, 立足丰富数据

国内技术优势: 阿里发布通义千问, 立足丰富数据

国内技术优势: 阿里发布通义千问, 立足丰富数据

国内技术优势: 阿里发布通义千问, 立足丰富数据

国内技术优势: 阿里发布通义千问, 立足丰富数据

型应用在互场景中的悲

1 1 类身份安全问题：人类身份代表自然人身份实体，在十模

应用上，数据上

应用上，数据上

应用上，数据上

2 AI Identity 安全问题：越来越多的 AI Agent 在大型场景中应用，AI Agent 具

应用上，数据上

应用上，数据上

应用上，数据上

应用上，数据上

应用上，数据上

应用上，数据上

应用上，数据上

应用上，数据上

应用上，数据上

应用上，数据上

应用上，数据上

应用上，数据上

应用上，数据上

应用上，数据上

应用上，数据上

应用上，数据上

应用上，数据上

应用上，数据上

应用上，数据上

应用上，数据上

应用上，数据上

应用上，数据上

应用上，数据上

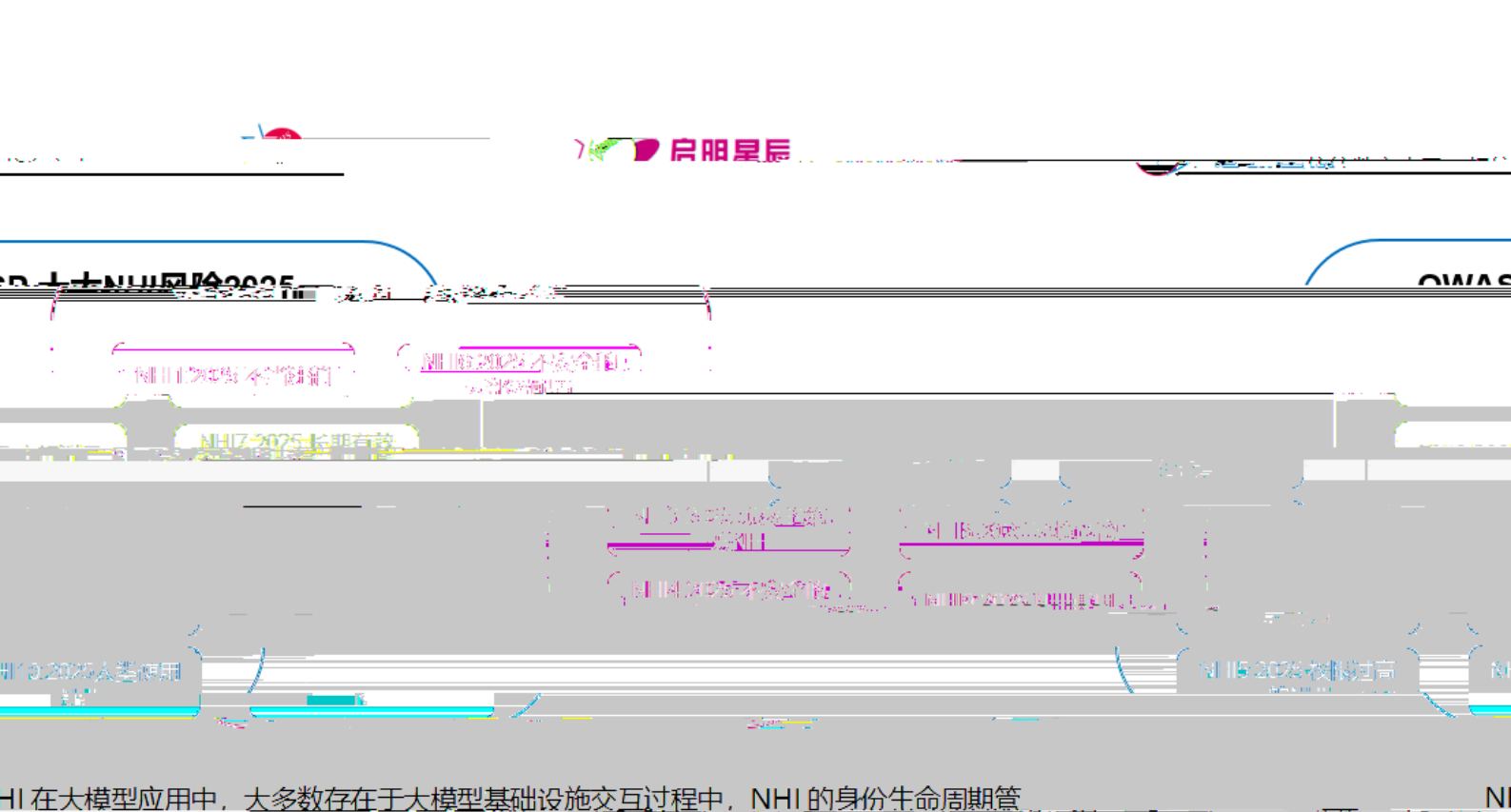
应用上，数据上

应用上，数据上

应用上，数据上

应用上，数据上

应用上，数据上



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。

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。

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。

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。

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。

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。

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) 用户输入输入输出风险：用户输入可能包含敏感信息，若大模型系统对输入提示词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，敏感信息可能在输出环节直接暴露。

此外，大模型系统可能通过自然语言处理技术，从用户输入中提取敏感信息，并在输出结果中泄露。例如，用户输入“我的身份证号是123456789”，大模型系统可能通过自然语言处理技术，从输入中提取“123456789”，并在输出结果中泄露。

(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存在漏洞，不法分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，API 接口若缺乏有效鉴权，攻击者可能通过非法手段获取 API 调用权限，进而获取敏感数据或进行恶意操作。

此外，大模型系统可能通过自然语言处理技术，从用户输入中提取敏感信息，并在输出结果中泄露。例如，用户输入“我的身份证号是123456789”，大模型系统可能通过自然语言处理技术，从输入中提取“123456789”，并在输出结果中泄露。此外，大模型系统可能通过自然语言处理技术，从用户输入中提取敏感信息，并在输出结果中泄露。

(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，用户可能通过自然语言处理技术，从知识库中提取敏感信息。例如，用户输入“我的身份证号是123456789”，大模型系统可能通过自然语言处理技术，从知识库中提取“123456789”，并在输出结果中泄露。

此外，大模型系统可能通过自然语言处理技术，从用户输入中提取敏感信息，并在输出结果中泄露。例如，用户输入“我的身份证号是123456789”，大模型系统可能通过自然语言处理技术，从输入中提取“123456789”，并在输出结果中泄露。此外，大模型系统可能通过自然语言处理技术，从用户输入中提取敏感信息，并在输出结果中泄露。

(4) 数据滥用风险：在大模型训练阶段，企业可能收集大量用户数据，用于模型训练。若企业未采取有效措施，防止数据滥用，可能导致用户数据泄露或被用于其他目的。例如，企业可能将用户数据用于其他业务，或将其出售给第三方。

此外，大模型系统可能通过自然语言处理技术，从用户输入中提取敏感信息，并在输出结果中泄露。例如，用户输入“我的身份证号是123456789”，大模型系统可能通过自然语言处理技术，从输入中提取“123456789”，并在输出结果中泄露。此外，大模型系统可能通过自然语言处理技术，从用户输入中提取敏感信息，并在输出结果中泄露。

此外，大模型系统可能通过自然语言处理技术，从用户输入中提取敏感信息，并在输出结果中泄露。例如，用户输入“我的身份证号是123456789”，大模型系统可能通过自然语言处理技术，从输入中提取“123456789”，并在输出结果中泄露。

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”的绑定形成可信身份，同时结合设备、行为、内容等多维数据，实现

生，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性能。

全协同，实现全程全网、端到端的安全保护。

力，在政府、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

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

访问管理、数据安全等方面的

就绪的大模型身份与访问管理。

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

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。

可信审计管理

可信审计管理

AI-R-IAM 为大模型提供可信能力

AI-R-IAM 针对多个大模型提供 AI-R-IAM

通过可信能力为大模型提供可信能力，保障大模型的安全运行

通过可信能力为大模型提供可信能力，保障大模型的安全运行

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。

八、可信能力为大模型提供可信能力，保障大模型的安全运行

可信能力为大模型提供可信能力，保障大模型的安全运行

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，

通过可信能力为大模型提供可信能力，保障大模型的安全运行

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，整合人类身份、AI 身份、物理身份、数字身份、虚拟身份、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管理，是数字世界秩序的底层支撑。

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;

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。

REFERENCE



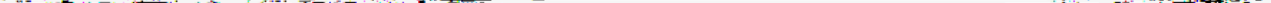
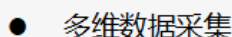
文 學 藝 術

实时内容过滤: 结合语义分析引擎, 自动屏蔽无极限的检索结果



[illegible]

Figure 1. The location of the study area.



构建全面

◎ 凡人之道

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，并构建风险画像体系，知悉安全风险水平。

- 数据知识图谱

非序、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、关系抽

角度分析大模型风险的能力。

关系构建，提供从“关系”的

- 数据深度分析

各环节进行合规风险分析，重点数据审计，敏感数据和异常场

大模型数据全生命周期各

风险进行评估，识别已知风险和未知风险。

景分析等，对大模型安全风

- 智能追踪溯源

提供全链路的智能追踪溯源，采用路径追踪、路径回溯等方法，提供

数据回查、溯源到数据

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 十塔型应用场景数据流

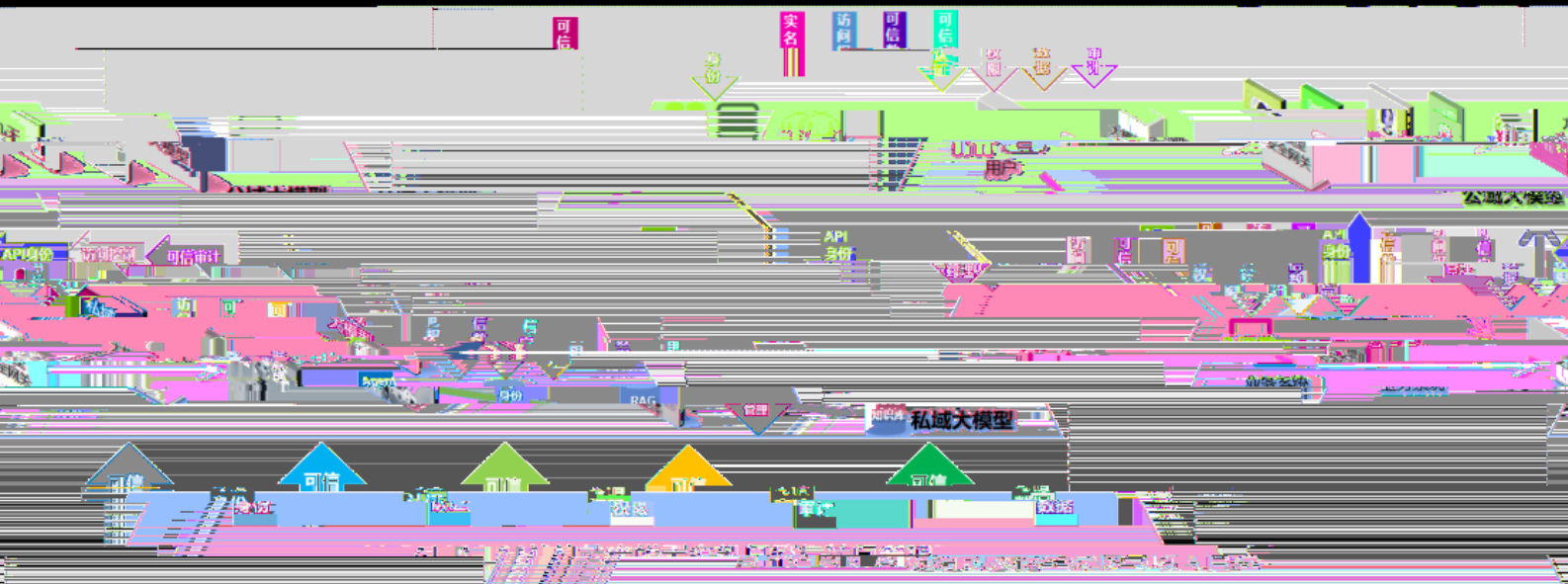
为大模型的访问、使用、治理、训练等条

AI-R-IAM AI就绪的大模型身份与访问管理

可信身份、可信认证、可信授权、可信审计、可信数据能力、可信场景适配应用

场景适配

用户和私有大模型、公域大模型场景；业务系统调用信创信创、RAG知识库、私域大模型场景；大模型数据训练场景；大模型数据训练场景。



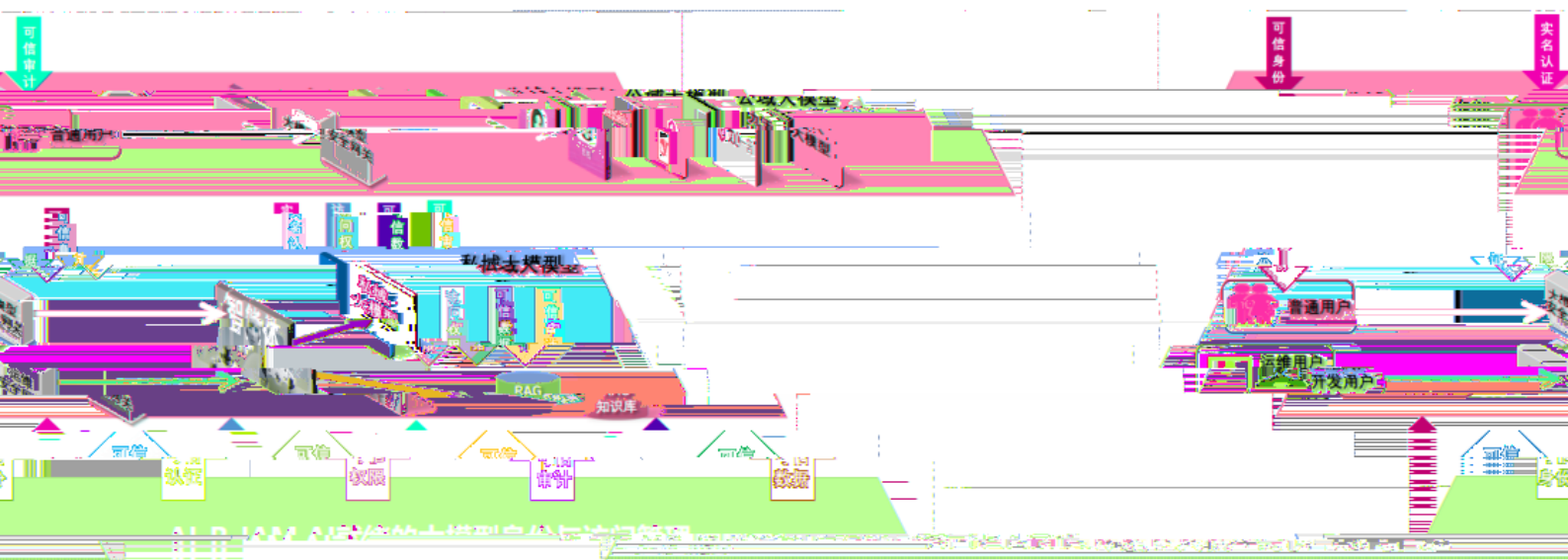
5.2 场景一：用户访问大模型应用场景

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

可信身份、可信认证、可信授权、可信审计、可信数据能力、可信场景适配应用

公域大模型、用户访问私域大模型、运营维护人员访问私域大模型等。

用户访问



1. 普通用户访问公域大模型

普通用户、企业员工、个人用户，都可以借助自己的终端设备，访问大模型应用。这些大模型应用涵盖多种场景服务，

或者平板等，大模型应用部署在公共云上，

内容：图像识别服务可以精准地识别图片中的物体、场景或人物特征等信息；语音转文字服务，

各侧可以调用内置的五音字库准确无误地转换成对应的文字内容。

对于用户而言，操作起来

一次生成即可，无需等待，且生成的内容可直接用于后续工作。

一份产品推广文案时，只需打

非常简便。比如，当一个企业员工想要利用文本生成服务撰写一

份产品推广文案，只需点击

生成按钮，即可生成一份产品推广文案。

图像识别服务可以

生成，很快就生成一份产品推广文案。

生成结果

生成结果

然而，在这种便捷服务背后，保障用户数据安全和相关权益是个重要的挑战。

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、物理硬件（如智能

卡）或者是基于区块链验证凭证（Verifiable Credential, VC）的身份认证方式。只有通过严

格的验证，用户才能获得访问权限。

系统同时记录用户登录时间和设备信息，以便进行安全审计。

保护用户的数据不被未授权访问。

的运行状况。它会记录下所有用户的操

在安全审计方面，AI-R-IAM 持续监控整个系统

的运行状况，一旦发现异常行为或潜在的安全风险，可以立即发出警报并启动应急响应

机制。系统还支持定期的安全审计，确保所有操作符合企业的安全策略和法规要求。

此外，系统还具备数据备份和恢复功能，以防止数据丢失并保障业务的连续性。

通过上述措施，AI-R-IAM 不仅提升了企业身份认证的安全性，还有效降低了数据泄露和

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

在企业数字化转型过程中，私域大模型的应用日益广泛。为了保障私域大模型的安全性和

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

系统，企业需要建立一套严格的访问控制机制，确保只有授权用户才能访问敏感数据。

系统应支持多种认证方式，包括密码认证、生物识别、硬件令牌等，以满足不同场景下

的安全需求。同时，系统还应具备实时监控和日志记录功能，以便及时发现和响应安全事

件。

将权限细化到最小粒度。企

在用户身份认证的基础上，私域大模型下需要针对不同角色

的用户进行权限分配。系统应支持基于角色的访问控制（RBAC）模型，确保用户只能访问其

能整个企业的数据状况，包括各个部门的数据资源；部门主管则只需要关注本部门的数据，

并能限制本部门员工的权限范围；普通员工的权限最为有限，只能访问其

负责的数据资源，无法访问其他部门的数据。

系统还应支持细粒度的权限控制，例如按文件、按字段、按操作等方式进行权限分配，

确保安全。

性有赖于系统，以满足不同角色的需求，并确保数据

3. 运营维护人员对私域大模型

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

限较大,一旦因疏忽或误操作导致数据泄露或系统故障,后果将非常严重。因此,在

权限管理上,应遵循最小权限原则,在普通用户权限基础上,授予必要的运维权限。

图 4-1-2 运维权限

运维人员应通过专用的运维通道访问系统,该通道应具备加密传输、身份认证、操作录

像等功能,确保运维操作的可追溯性。同时,应建立严格的运维审批流程,任何高危操

作必须经过审批后方可执行。此外,应定期对运维人员进行安全培训,提高其安全意

识和技能,确保运维操作的安全性和规范性。

通过上述措施,可以有效降低私域大模型应用的安全风险,保障用户数据的安全和

系统的稳定运行。在实际应用中,应根据具体的业务场景和系统架构,灵活调整安全

策略,确保安全措施的有效性和适应性。

此外,还应建立应急响应机制,一旦发生安全事件,能够迅速响应,采取有效措施,最

大程度上减少损失。同时,应定期开展安全审计,检查安全措施的执行情况,及时发现

和修复安全隐患。通过持续的安全管理和改进,确保私域大模型应用的安全性和稳

定运行。

图 4-1-3 安全策略

图 4-1-4 安全策略

在私域大模型应用中,数据的安全性和隐私保护至关重要。除了上述的安全策略外,

还应采取以下措施,进一步保障数据的安全和隐私。

首先,应建立完善的数据分类分级制度,根据数据的重要性和敏感程度,制定不同

的安全保护等级。

其次,应加强数据加密管理,对敏感数据进行加密存储和传输,确保数据在传输和

存储过程中的安全性。同时,应定期更新加密算法,防止被破解。

此外,还应建立数据备份和恢复机制,确保在发生数据丢失或损坏时,能够及时恢

复数据,保障业务的连续性。

即授权运维人员确认是否按执行正常操作，并需提前对额外的权限验证步骤。这可以防止输入

禁止其执行。无论运维人员如何尝试都无济于事这一限制。例如，对于删除整个数据库这样

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

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。

内部运维人员的权限。需要建立严格的权限管理，只有经过特别授权的人员才能访问

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

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。

内部运维人员的权限。需要建立严格的权限管理，只有经过特别授权的人员才能访问

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

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。

内部运维人员的权限。需要建立严格的权限管理，只有经过特别授权的人员才能访问

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

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。

内部运维人员的权限。需要建立严格的权限管理，只有经过特别授权的人员才能访问

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

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。

内部运维人员的权限。需要建立严格的权限管理，只有经过特别授权的人员才能访问

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

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。

业务系统应基于最小权限原则来分配访问权限，实现大模型业务权限可信、资源分配合理与安全性。

通过RBAC和ABAC技术，业务系统根据实体的属性、状态和业务流程，实现细粒度授权。通过ABAC技术，业务系统根据实体的属性、状态和业务流程，实现细粒度授权。

业务系统应基于最小权限原则来分配访问权限，实现大模型业务权限可信、资源分配合理与安全性。

业务系统应基于最小权限原则来分配访问权限，实现大模型业务权限可信、资源分配合理与安全性。

业务系统应基于最小权限原则来分配访问权限，实现大模型业务权限可信、资源分配合理与安全性。

业务系统应基于最小权限原则来分配访问权限，实现大模型业务权限可信、资源分配合理与安全性。

业务系统应基于最小权限原则来分配访问权限，实现大模型业务权限可信、资源分配合理与安全性。

业务系统应基于最小权限原则来分配访问权限，实现大模型业务权限可信、资源分配合理与安全性。

业务系统应基于最小权限原则来分配访问权限，实现大模型业务权限可信、资源分配合理与安全性。

3. 业务行为审计

业务系统应基于最小权限原则来分配访问权限，实现大模型业务权限可信、资源分配合理与安全性。

业务系统应基于最小权限原则来分配访问权限，实现大模型业务权限可信、资源分配合理与安全性。

业务系统应基于最小权限原则来分配访问权限，实现大模型业务权限可信、资源分配合理与安全性。

业务系统应基于最小权限原则来分配访问权限，实现大模型业务权限可信、资源分配合理与安全性。

业务系统应基于最小权限原则来分配访问权限，实现大模型业务权限可信、资源分配合理与安全性。

业务系统应基于最小权限原则来分配访问权限，实现大模型业务权限可信、资源分配合理与安全性。

↑ 他略同于《圣德太子传》。《圣德太子传》：“世祖（即圣德太子）知识庶之置主何人，宿望

[illegible]

知念昌久

特性知准确性_同时率用面换公耗_钢口图管管技术对磁域十排型_PAG钢口面安全网除

模型应用范围：对LLM模型，RAG知识库的数据输入到模型输入端供全

[图论问题_应用递归计算_路径探索算法_提供可解P->访问数据公布->可解路](#)



17. 下列各句，没有语病的一句是（3分）

宗 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

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

智能体身份标识由系统生成并绑定到智能体唯一可信身份标识（如数字身份标识，即 Digital Identity），由统一身份 ID 进行绑定管理，并能在审计界面明确。如果智能体身份做了特定项操作，系统可追溯到是谁让智能体进行操作，方便界定相应责任。

智能体身份与特定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谁指定的模型和代码构建，并成功部署到系统。当智能体请求到该智能体所发出的请求时，系统会验证该请求的完整性，并验证该请求是否来自可信的智能体。系统运行的安全，依赖于智能体身份的真实性。系统会对智能体身份进行严格验证，以此来精准确认该智能体的真实身份，同时评估其可信度，从而确保系统运行的安全性和可靠性。

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 实名认证能力建立连接。普通用户通过 AI-R-IAM 系统验证身份并获取可信身份标识。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身份认证，并在执行任务前经系统严格认证校验，通过则颁发令牌（Short Access Token）或一次性密钥（One-Time Key）等方式，让智能体具备时效性和可撤销性。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。

同时，监管部门可以记录和追溯智能体的身份标识，例如记录了哪些数据，进行了何种操作，

依据何种规则。

成训练数据外泄

在企业内部，智能体易成为攻击目标，被攻击后可能损害企业数据资产。

新风险，如同传统领域数据泄露外部。

存在数据泄露，RAG模型安全

在此场景下，AI-RAM提供用户和智能体，智能体数据

训练RAG的数据源和数据

内外数据源之间的数据数据边界，防止数据外泄策略。防止智能

对智能体恶意行为等。避免

漏洞，辅以模型防护、上下文检测的安全控制，可及时阻断潜在

因智能体自主决策引发安全风险。

6 发展趋势展望

在人工智能驱动快速发展的背景下，AI 技术对网络安全攻防态势产生深远影响，主要体现在

生产全程可停能力，引领大模型安全共“安全管家”理念，可产生合理可信

1. 在基础设施领域，以 AI 技术将传统人工模式升级为智能化，提升

系统，此趋势将显著提升网络安全防护能力，为未来网络

“检测”与“响应”能力提供强有力的支持，进一步提升网络安全

异常行为。

能力，提升网络安全设备；2. 在网络安全方面，AI-R-IAM 将提供进

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。

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；作