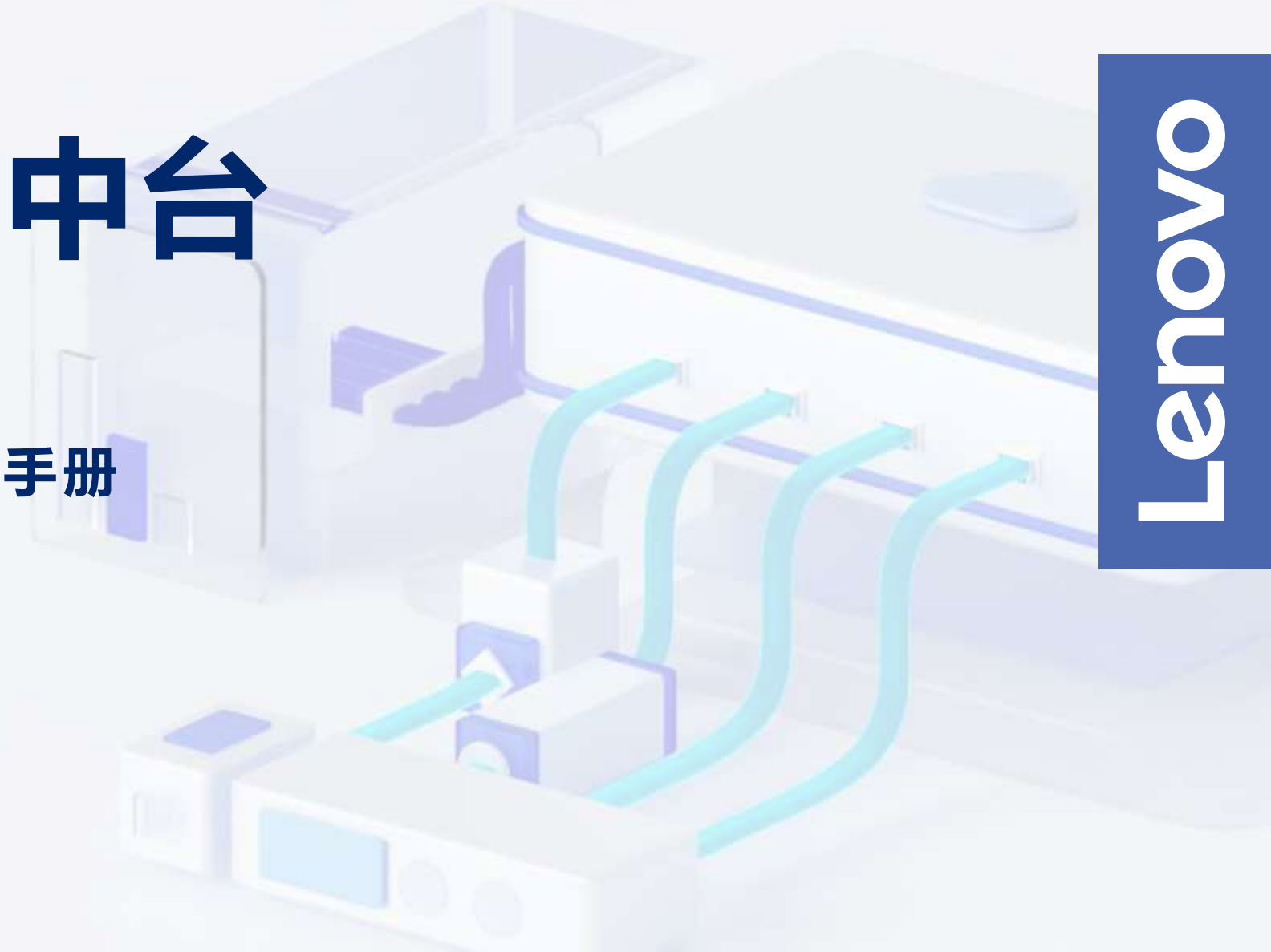
AI内容安全中台

AICS一站式内容安全服务手册



2025 Lenovo Internal. All rights reserved.

目录

CONTENTS

Lenovo 2025 Lenovo Restricted. All rights reserved.

01 背景介绍

面向对象： 全部

AI内容安全中台　www.aiguard.lenovo.com

（首次登录需要用IT Code预登录后，由内容安全团队开通权限使用）

核心目标

打造一站式AI内容安全服务，覆盖监控、审核、处置、指令等能力，支撑公司AI业务的内容安全需求。

覆盖范围

支持云端AI业务的文本/图片内容审核

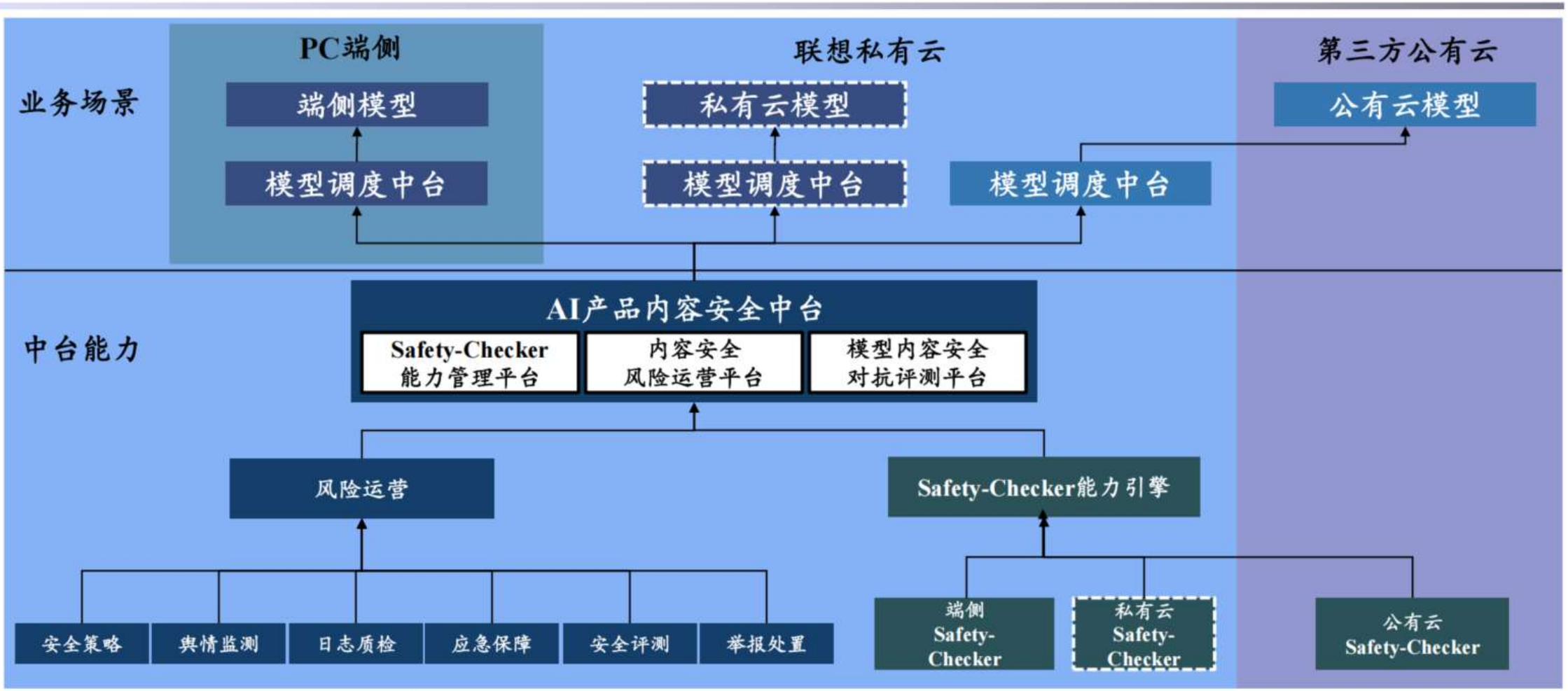
项目周期

2024.12.2 启动 — 2025.3.31 上线

Lenovo 2025 Lenovo Restricted. All rights reserved.

01 背景介绍

面向对象： 全部

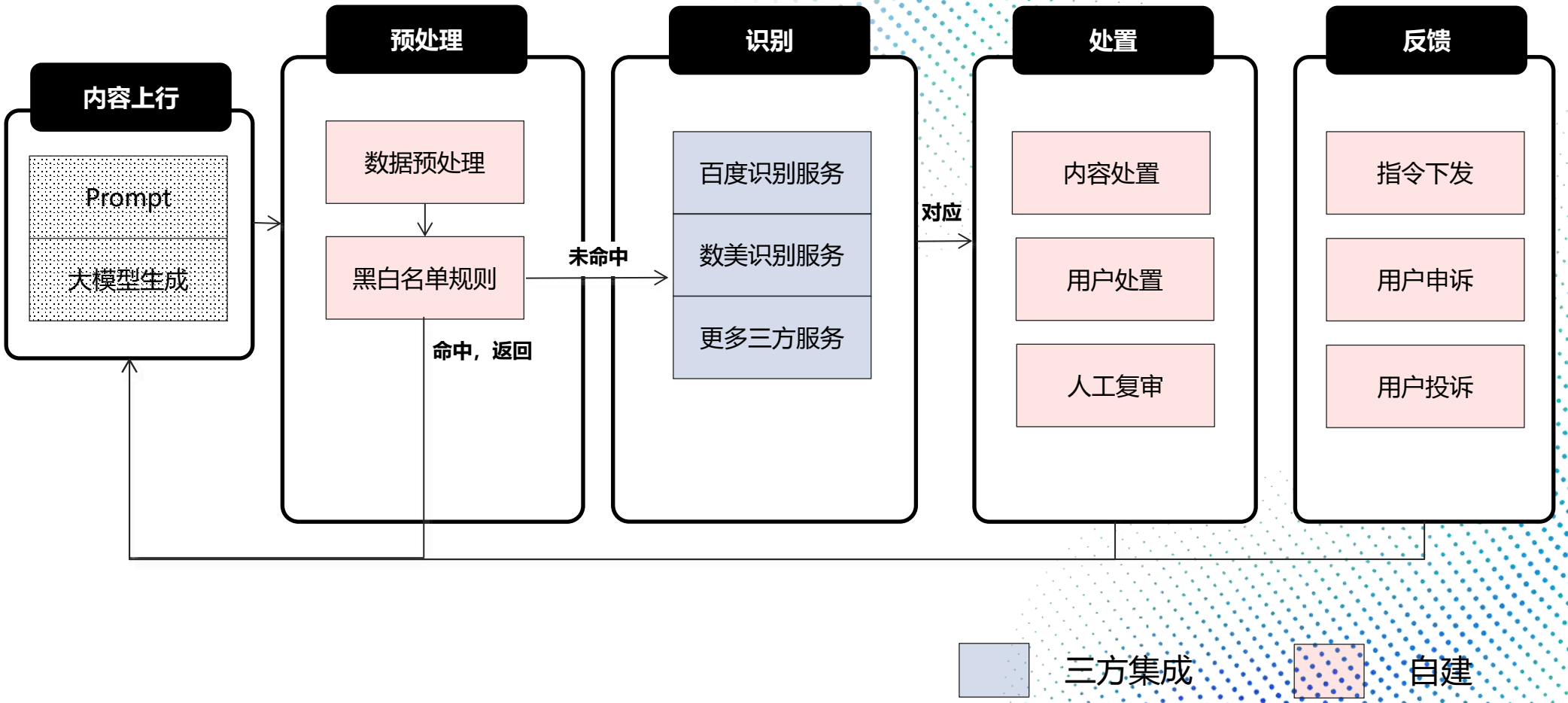


Lenovo 2025 Lenovo Restricted. All rights reserved.

01 背景介绍　AI内容安全治理业务流程

面向对象： 全部

大模型内容安全需要通过一套流程化作业体系，实现安全目标。从内容产生到接收负反馈，需要相应的服务和系统作为业务支撑。



Lenovo 2025 Lenovo Restricted. All rights reserved.

02 服务对接

面向对象： 全部

一、内容安全审核服务介绍

目前提供文本检测、图片检测服务，为云端AI业务提供标准API，接入后即查即返回。

文本检测

支持用户输入和大模型输出的文本检测，集成放案涵盖动态更新词库和NLP算法

图片检测

支持输入输出图片的内容安全检测，涵盖色情、涉政、暴恐、广告、违规违法等多种风险内容。

音频检测（待开放）

支持对大模型生成视频内容进行安全审核，可对色情、暴恐、涉政等内容精准识别。

API+SaaS方案

标准化API

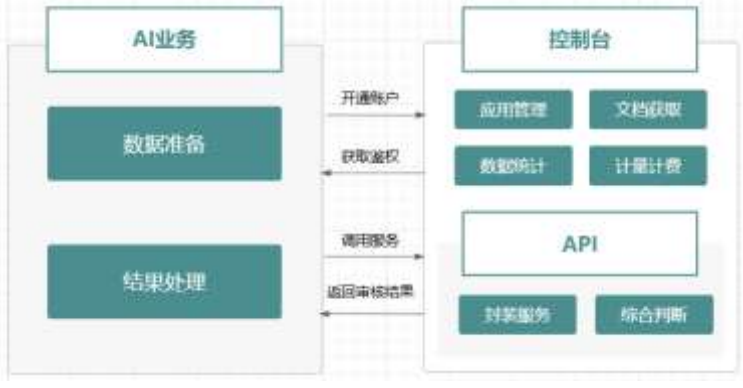
提供标准化API实时接口，返回结果已封装，方便开发。

控制台

搭配控制台，提供一站式应用查看、获取文档等。

对接示意

注：如AI业务通过AI Verse提供的Guard Engine间接接入，则由AI Verse来对接API



Lenovo 2025 Lenovo Restricted. All rights reserved.

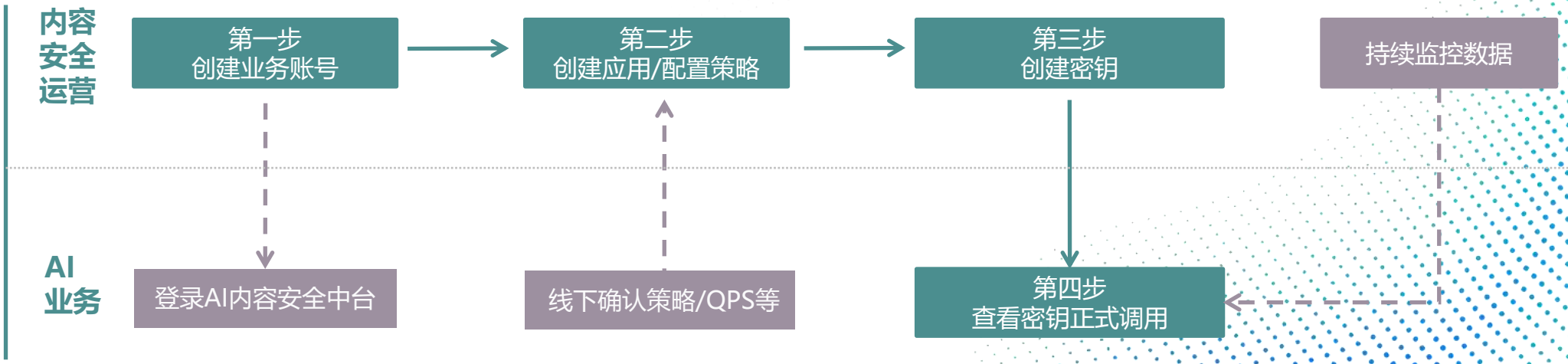
02 服务对接

面向对象： 全部

一、内容安全审核服务对接流程

AI业务直接对接

内容安全运营为AI业务创建业务账号，在业务下创建所需应用，基于应用配置安全策略，此时应用为[执行中]状态，可创建策略有效期内的密钥。AI业务可通过登录平台查看对应用获取密钥，密钥有效期为一年，过期前需内容安全运营创建新密钥，避免导致业务影响。



Lenovo 2025 Lenovo Restricted. All rights reserved.

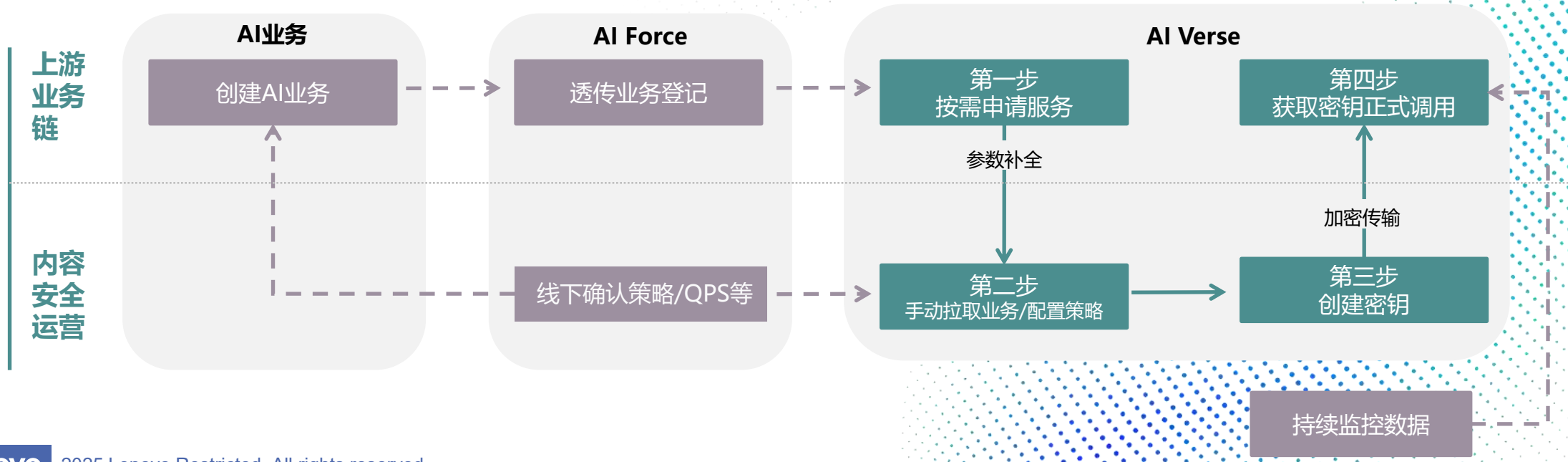
02 服务对接

面向对象： 全部

一、内容安全审核服务对接流程

AI业务通过AI Verse间接对接

AI业务在AI Force登记业务，AI Force透传给AI Verse，AI Verse通过内容安全提供的特定通道按照业务自动创建各自的应用，内容安全运营基于应用配置安全策略，并加密传输密钥给AI Verse。AI Verse调用内容安全检测API，从而接入AI Verse的AI业务完成安全服务的使用。



Lenovo 2025 Lenovo Restricted. All rights reserved.

02 服务对接

面向对象： 全部

一、内容安全审核服务标签

disposalStatus	内容安全标签	建议执行动作
201	正常(normal)	通过
202	红线必答(red line response)	拦截/不上屏/本地知识库
203	安全大模型(safety lmm)	拦截/不上屏/本地知识库
204	兜底回复(fallback response)	拦截/不上屏/本地知识库
205	首段风险兜底回复(first risk fallback response)	拦截/不上屏/本地知识库
206	当前段落不上屏(current block display)	拦截/不上屏
207	不上屏(block display)	拦截/不上屏

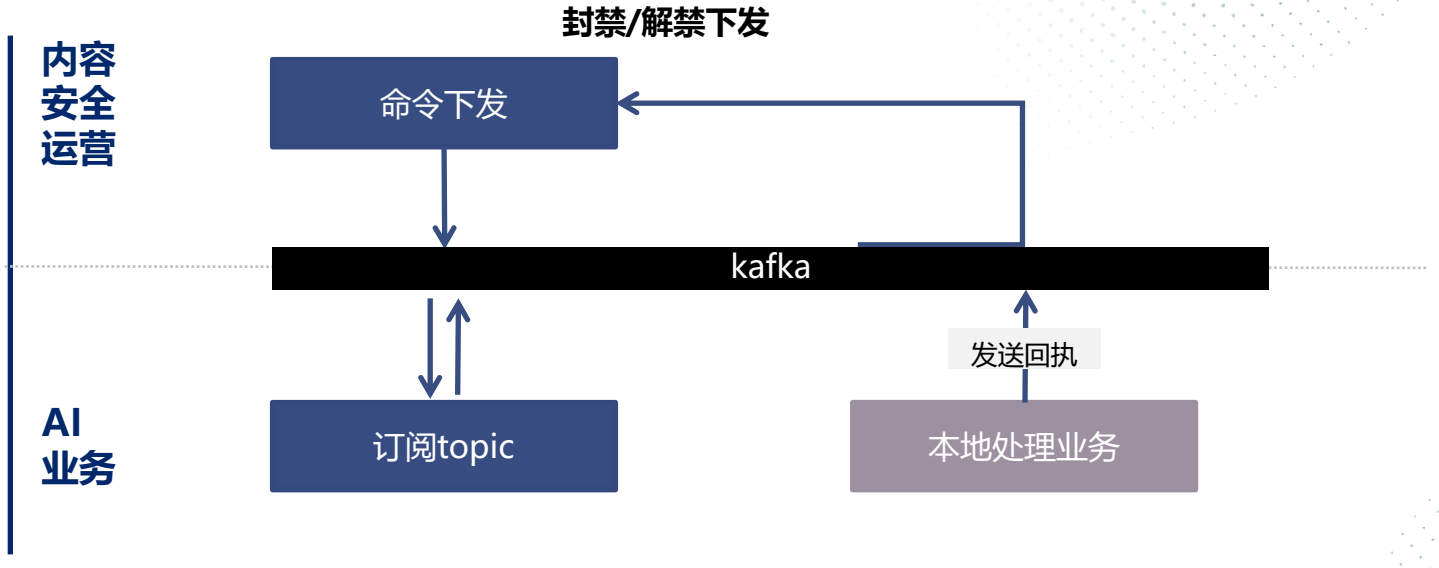
Lenovo 2025 Lenovo Restricted. All rights reserved.

02 服务对接

面向对象： 全部

二、用户处置消息队列对接

内容安全运营依照上级指令和审核发现风险用户，通过提供消息队列广播的方式，给AI业务下发封禁/解禁要求，AI业务接收到指令后在各自业务进行相应的限制并返回内容安全执行状态。



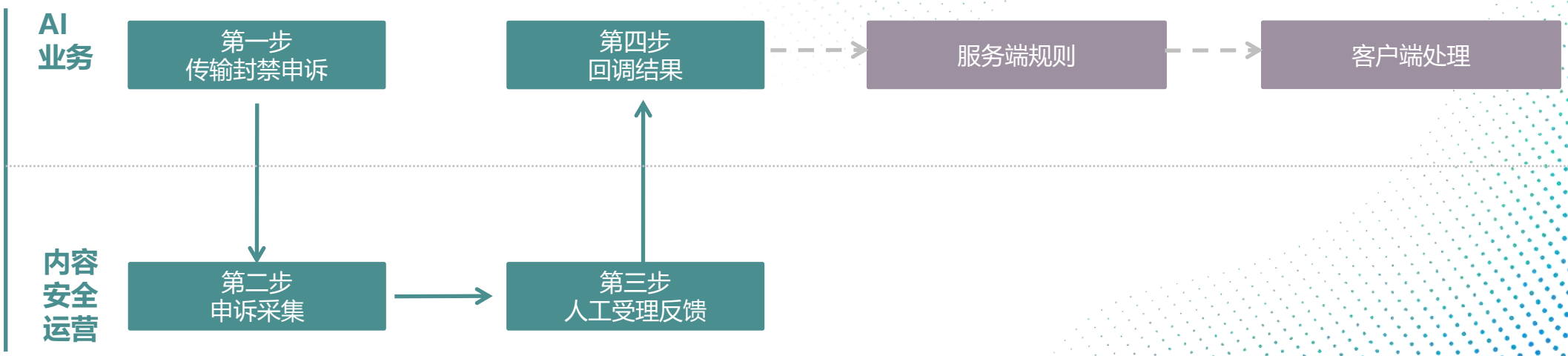
Lenovo 2025 Lenovo Restricted. All rights reserved.

02 服务对接

面向对象： 全部

三、封禁用户申诉服务对接

内容安全提供[封禁申诉提交服务]，该API用于从AI业务的“举报中心”中采集内容安全相关数据。通过“申诉采集”提交举报内容后，由内容安全在投诉受理平台进行人工的受理，AI业务通过回调API获取申诉结果。



Lenovo 2025 Lenovo Restricted. All rights reserved.

02 服务对接

面向对象： 全部

四、协同工单系统使用介绍

AI内容安全中台提供平台化工单服务，对平台登录用户提供工单流转功能，更好的服务多方协作、应急保障等流程。如需使用工单，请先登录平台，由内容安全运营对登录账号进行授权。

Lenovo 2025 Lenovo Restricted. All rights reserved.

04 常见问题

面向对象： 全部

1. 端侧业务是否适配AI内容安全中台?

端侧业务由于安全检测通过SDK实现，无需接入API。但监管部门要求对日志进行人工审核，内容安全已支持日志接入方案，可将传输至中台，从而实现人工审核。

2. 为什么在参数正确情况下无法调通审核API?

首先需要联系内容安全团队获取对应的密钥。服务支持APIH访问开放服务接口和直接调用两种方式，直接调用AICS 需要确认是否同EARTH，如不同则需要运维申请“开墙”。

3. 没有队列对接过如何对接用户封禁服务?

强烈建议直接通过用户封禁服务保障合规和高效，首次对接需要在KPaaS平台进行审批获取响应信息

信息	说明	来源
Kafka kkp集群	通过联系KPaaS获取	KPaaS
Kafka kkp集群证书	通过联系KPaaS获取	KPaaS
Kafka kkp集群证书密钥	通过联系KPaaS获取	KPaaS
Kafka client账号/密钥	通过KPaaS平台获取	KPaaS
Kafka Topics	请查看《AICS能力服务共享消息推送文档》	内容安全

Lenovo 2025 Lenovo Restricted. All rights reserved.

感　谢　观　看