智算安全专区

用户指南

天翼云科技有限公司

目 录

1 产品介绍

1.1 产品定义

本章节为您介绍智算安全专区的产品定义。

智算安全专区是为云上智算基础设施构建的一站式安全防护平台，提供从开发、训练、部署到运营的全生命周期的智算安全解决方案。

1.2 功能特性

本节介绍智算安全专区产品的功能特性。

大模型安全卫士：

- **内容检测三道防线**：大模型安全卫士采用了三层内容检测防线：敏感词匹配、语义分析和模型推理。敏感词匹配能够快速识别常规违规内容，语义分析则能够处理更复杂的语义绕过问题，而模型推理则针对高级攻击手段，如藏头诗、谐音梗等进行检测。
- **语料输入安全**：语料安全网关代理 rag 业务系统请求，解析文件进行内容检测。保障语料库及生成内容的安全性、合规性，防止恶意攻击（如数据投毒、提示注入）、敏感信息泄露及生成有害内容。
- **敏感信息脱敏**：代理 rag 业务系统请求，解析文件发现敏感内容进行数据脱敏。

1.3 产品优势

- 强大的内容过滤能力：具备实时过滤与防御能力，通过敏感词匹配、语义分析和模型推理三道防线，确保生成内容的安全合规；
- 多语言支持：语言不敏感性，支持多语言、混合语言检测，适应全球化业务需求；
- 流式检查：支持在 AI 模型流式输出过程中进行内容检测，实时阻断违规内容，保障输出安全。

1.4 应用场景

本节介绍智算安全专区产品的应用场景。

模型输入输出内容违规

- 场景特点：使用者输入违规内容（如违法犯罪、暴力色情等），诱导模型生成不相关或非法内容，造成不良影响。
- 解决方案：内置大模型内容安全引擎，在输入阶段评估提示词，防止生成非法结果；在输出阶段检测违规内容，及时阻断并进行事后审计。支持模型接口代理方式实时检测与阻断，通过三道过滤防线保障内容安全。

敏感数据安全脱敏

- 场景特点：在问答结果输出阶段，应用可能带出个人简历、薪资情况等敏感信息，导致泄露。
- 解决方案：在数据输入阶段介入脱密信息处理，使进入 RAG 语料库的资料均为脱敏后材料，基于大模型的智能脱敏在保护敏感信息的同时保留数据可用性。

1.5 使用限制

大模型安全卫士仅支持对采用标准的 OpenAI API 规范的基础模型进行访问限制、权限管控及安全防护。

1.6 术语解释

RAG：全称是 Retrieval-Augmented Generation（检索增强生成），这是一种将信息检索系统与大语言模型的生成能力相结合的技术框架。

OpenAI API 规范：是由 OpenAI 提供的一套接口规范，围绕 Chat Completions API 构建，这是一种基于 HTTP 的 RESTful 接口，允许开发者与大语言模型（如 GPT 系列）进行交互。

2 计费说明

2.1计费方式

智算安全专区-大模型安全卫士仅支持包周期计费。

产品套餐计费内容

产品名称	规格	说明	标准价格（元/月）	标准价格（元/年）
大模型安全卫士	标准版	1、关键词拦截； 2、语义拦截； 3、支持私有语料库敏感信息识别及分类分级（规则匹配方式） 4、RAG 代理网关（输入阻断）； 5、可配置模型代理数量 x 1； 6、大模型安全管理平台； 7、支持同时对大模型发起 20 个请求	7000	70000
	性能扩展包（标准版）	增加对大模型发起 10 个请求	5500	55000

2.2续订

为避免智算安全专区的实例到期后，防护服务自动停止，需要在实例到期前进行手动续费，或设置到期自动续费。

到期说明

到期后，资源进入保留期，您将不能正常访问及使用云服务（资源冻结），但对于您存储在云服务中的数据予以保留。

- 若您在保留期内续费，计费周期自资源续订解冻开始，计算新的服务有效期，按照新的服务有效期计算费用。

- 若保留期到期您仍未续费，存储在云服务中的数据将被删除、云服务资源将被释放。

关于保留期的详细信息，请参见“到期处理”。

续订说明

订单到期后，若没有续订，将不能继续使用订单中的服务，建议您提前进行续订。更多详情请阅读天翼云“续订规则说明”。

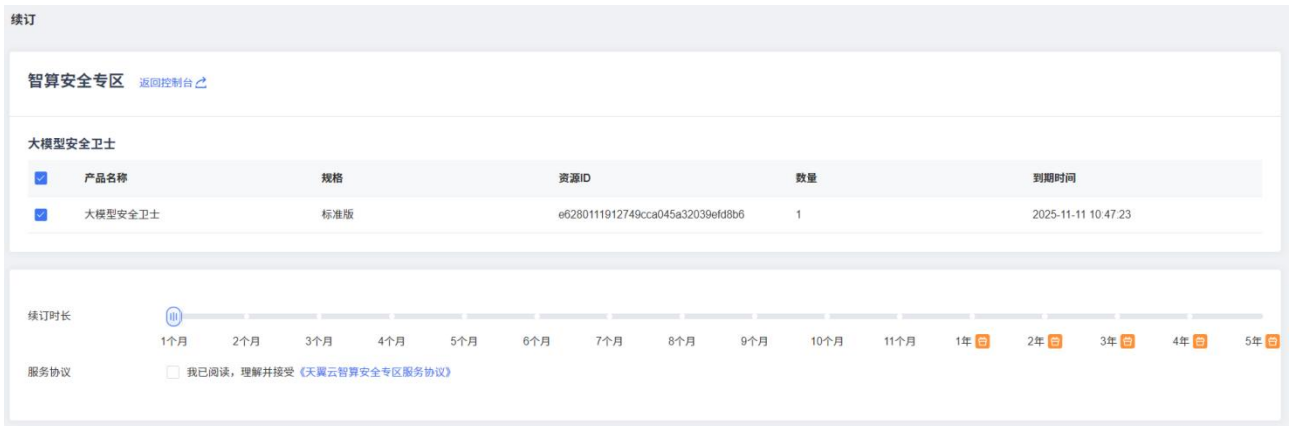
支持的续订方式：

续订方式	说明
手动续订	智算安全专区在购买之后支持手动续订的方式，您可以随时在智算安全专区管理控制台中的实例页面进行续订，续订后智算安全专区到期时间将自动延期到续订后的到期时间。
自动续订	智算安全专区在购买之后支持自动续订的方式，您可以随时在“费用中心 > 订单管理 > 续订管理”页面开启自动续订，自动续订开启后智算安全专区将会进行自动续订，更多说明见“自动续订”。

手动续订

1. 登录智算安全专区控制台。
2. 在大模型安全卫士实例页面找到需要续订的实例，点击“续订”。

3. 进入智算安全专区续订页面，根据使用需要设置续订时长。



4. 续订时长设置完成后，在页面下方确认支付费用，阅读《天翼云智算安全专区服务协议》，并勾选“我已阅读，理解并接受《天翼云智算安全专区服务协议》“，在页面右下角单击“立即购买”。

自动续订

- 方法一：在购买智算安全专区时，同步开启“自动续订”。详细操作请参见购买智算安全专区。
- 方法二：若购买智算安全专区时未开启“自动续订”，用户也可在购买后，通过天翼云“费用中心 > 订单管理 > 续订管理”页面，开通自动续订。详细操作请参见开通自动续订。

2.3退订

智算安全专区支持退订，可通过智算安全专区控制台、天翼云费用中心发起并完成退订操作。

退订说明

您（天翼云客户）可根据需要，在符合天翼云退订规则的前提下，灵活退订配额。目前退订包含七天无理由全额退订和非七天无理由退订以及其他退订，退订规则详情见“退订规则说明”。

退订步骤

1. 登录智算安全专区控制台。
2. 在大模型安全卫士实例页面找到需要退订的实例，点击“退订”。
3. 进入退订申请页面，确认退订信息，信息确认无误后选择退订原因，勾选“我已确认本次退订金额和相关费用”后，点击“退订”后即可进行退订。
4. 系统提示退订申请提交成功，可前往订单详情查看退订进度。

3 快速入门

3.1订购智算安全专区服务

前提条件

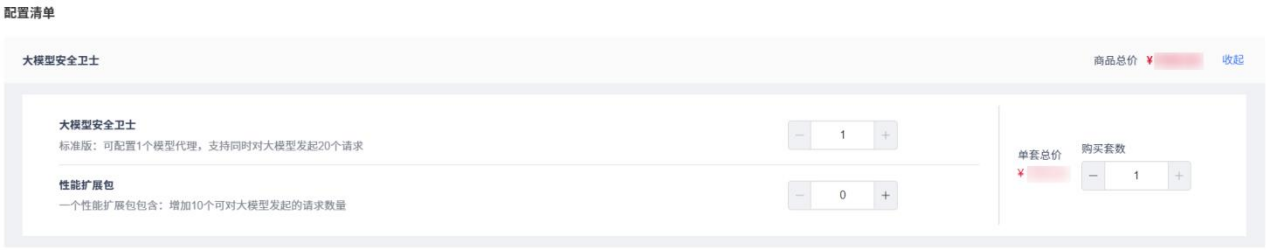
已注册天翼云账号并完成实名认证。

操作步骤

1. 登录天翼云控制中心。
2. 单击页面顶部的区域选择框，选择区域。
3. 在产品服务列表页，选择“安全 > 智算安全专区”。
4. 在页面右上角，单击“立即购买”。

5. 在产品订购页面，配置区域、可用区、虚拟私有云、子网、CPU 架构。

6. 选择规格和购买数量。



7. 选择“购买时长”，拖动时间轴设置购买时长。

说明：

支持开启“自动续订”，当服务到期前，系统会自动按照默认的续费周期生成续费订单并进行续费，无须用户手动续费。

8. 确认参数配置无误后，阅读《天翼云智算安全专区服务协议》，并勾选“我已阅读，理解并接受《天翼云智算安全专区服务协议》”，单击“立即购买”。

9. 进入“付款”页面，完成付款。

3.2登录实例

1. 登录智算安全专区控制台。

2. 在“大模型安全卫士”实例页面，找到要管理的实例，单击“管理”。

4 资产中心

4.1模型资产

4.1.1 模型资产透视

显示模型资产的风险概况。

1. 登录大模型安全卫士实例。

2. 在菜单栏选择“资产中心 > 模型资产”，查看模型资产风险概览。

3. 单击“查看风险详情”，可查看模型资产风险详细信息。

4.2主机资产

4.2.1 管理终端

用户可在**主机资产**页面查看所有绑定该中心的主机信息，包括名称、分组、标签、IP、操作系统、终端版

本等，并可进行查看终端详情、编辑终端、查看策略等操作。

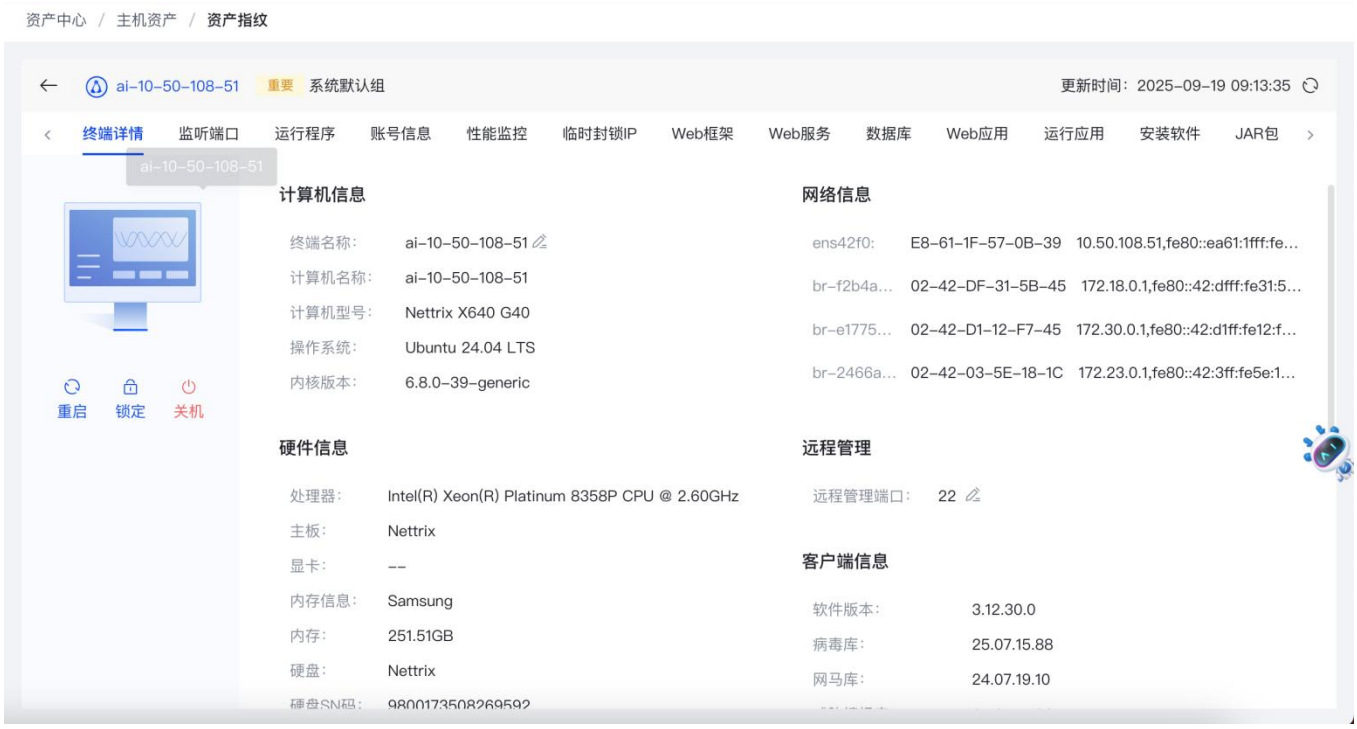
4.2.2 查看终端详情

步骤 1. 登录大模型安全卫士实例。

步骤 2. 在菜单栏选择“**资产中心** > **主机资产**”进入**主机资产**页面。

步骤 3. 选择需要查看的主机终端（须确保终端为在线状态），点击终端名称或者点击**操作项**中的“**查看**”。

步骤 4. 进入**终端详情**页面，即可查看该主机终端的详细信息，并可进行远程重启主机、关闭主机及修改远程端口操作。



终端概况说明如下。

终端信息	说明
终端详情	对终端进行详细信息展示：包括网络信息、环境信息、其他信息等；并支持远程关闭主机、重启主机、停止防护等操作。
监听端口	对终端上端口情况进行实时监控。
运行程序	对终端上进程运行情况进行实时监控，并支持远程结束相关进程。
账号信息	对终端上所有账号信息进行统计。
运行应用	对终端上运行的软件应用信息进行统计。
性能监控	对终端上的内存、CPU、磁盘、网络 IO 进行监控统计。
临时封锁 IP	对终端上因为防暴力破解和防端口扫描而引发的临时封锁 IP 进行管理。
注册表启动项	对终端上所有的注册表启动项进行统计和管理。
Web 框架	对终端上运行的 Web 框架进行统计。
Web 服务	对终端上运行的 Web 服务进行统计。
数据库	对终端上运行的数据库进行统计。
Web 应用	对终端上运行的 Web 应用进行统计。
在线统计	对终端的在线时间进行统计。
安装软件	对终端安装包名、版本号、类型、发布者、安装时间等进行统计。
Jar 包	对终端安装所关联的 jar 包进行统计。
计划任务	对终端被制定了哪些计划任务进行统计。
环境变量	对终端的环境变量进行统计。
内核模块	对终端内核模块名称、版本号、模块路径和大小、模块依赖等进行统计。
Windows 证书	对终端的各种证书进行统计。

4.2.3 编辑终端

步骤 1. 登录大模型安全卫士实例。

步骤 2. 在菜单栏选择“**资产中心** > **主机资产**”进入**主机资产**页面。

步骤 3. 选择需要编辑的终端，点击右侧**操作项**中的“**编辑**”。

步骤 4. 在弹出的对话框中编辑终端信息，点击<**确定**>，即可修改终端信息。

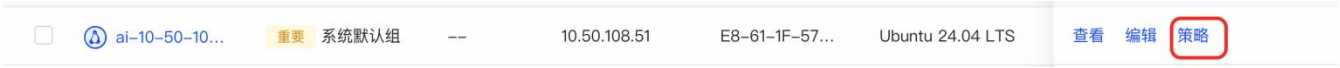
编辑终端信息时，若绑定状态开关的状态调整为关闭，将解绑该终端，被解绑的终端会从终端列表中删除。

4.2.4 查看策略

步骤 1. 登录大模型安全卫士实例。

步骤 2. 在菜单栏选择“**资产中心** > **主机资产**”进入**主机资产**页面。

步骤 3. 选择需要查看的终端，点击右侧**操作项**的“**策略**”，即可对该终端的策略信息进行查看和编辑操作，详情请参考“策略管理”。

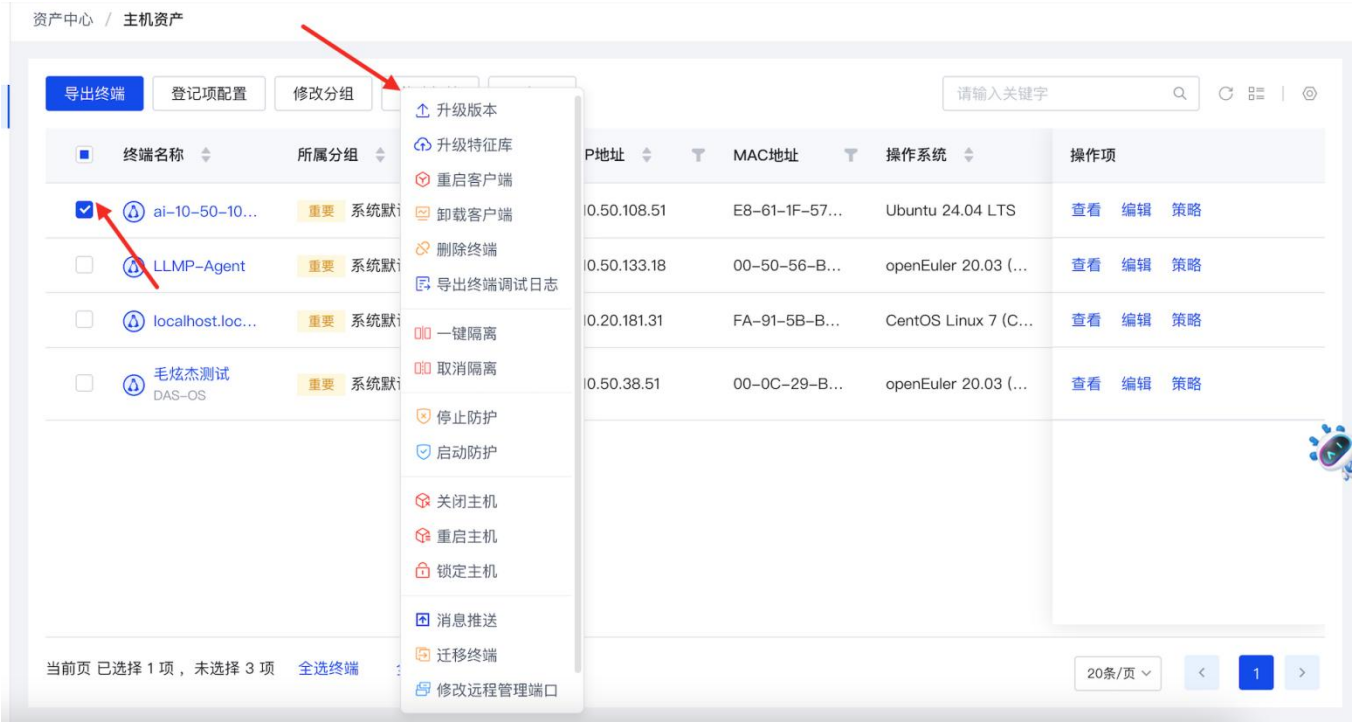


4.2.5 其他操作

步骤 1. 登录大模型安全卫士实例。

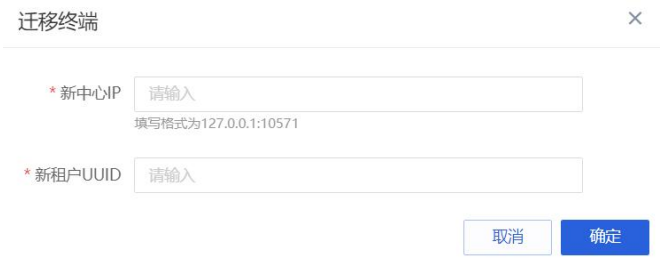
步骤 2. 在菜单栏选择“**资产中心** > **主机资产**”进入**主机资产**页面。用户可在此页面修改终端分组、修改终端标签、导出终端、登记管理操作。

步骤 3. 勾选终端，点击<更多>，在弹出的下拉框中选择不同的菜单项，可对终端进行卸载客户端、删除终端、停止防护、启动防护、关闭主机、重启主机、重启客户端、迁移终端、修改远程管理端口、导出终端调试日志、消息推送、一键隔离、取消隔离等操作。



具体操作说明如下。

操作	说明
导出终端	可将终端信息以 CSV 格式导出至本地。
修改分组/修改标签	修改所选终端的分组/标签，每个终端必须且只能在一个分组内，可以有多个标签。
卸载客户端	卸载终端上的 EDR 客户端软件。

操作	说明
删除终端	删除终端后，EDR 不能对终端进行管控，许可也将释放。终端信息以及相关日志将会删除，但部分防护依然有效。
停止防护	关闭所选终端当前所有防护。
启动防护	启动对终端的防护。
关闭主机/重启主机	对所选终端进行关机或重启。
重启客户端	对客户端进行重启。
锁定主机	锁定目标客户端。
迁移终端	填写新中心 IP 和 UUID，点击<**确定**>，可对租户内终端进行同中心跨租户迁移以及不同中心间迁移。 
修改远程管理端口	填写需要修改的远程管理端口，勾选**立即重启**，点击<**保存**>即可修改远程管理端口。 

操作	说明
导出终端调试日志	将所选终端的客户端运行日志，异常转储日志，操作系统日志信息导出。 
升级	对客户端主程序进行升级操作。
消息推送	对客户端所在终端进行消息推送（仅适用于 Windows 主机）。 
重新登记信息	对客户端重新登记资产信息。
一键隔离	对选定客户端进行一键断网。
取消隔离	对选定客户端取消一键断网策略。

4.3 模型代理

4.3.1 网关代理

网关代理的作用是在大模型防护系统开启一个代理端口，原请求大模型的请求需主动请求代理端口，由代理服务判定输入内容合规后再转发到真正的大模型接口，并按照配置决定是否对大模型返回的响应进行合规性检测。

主要用于无法使用透明代理的场景，如

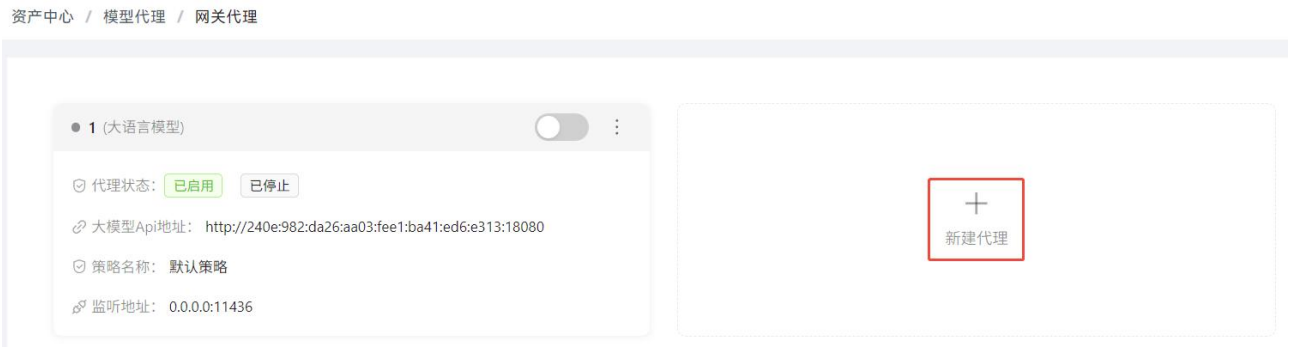
- 非自建大模型，如使用的云厂商提供的 api 接口；
- 大模型节点使用严格，禁止安装其他服务的；

4.3.1.1内容安全代理

内容安全代理主要用于代理大模型的对话接口，如“/v1/chat/completions”接口，从而对用户的提问和大模型的回答进行合规性检查。

步骤 1. 登录大模型安全卫士实例。

步骤 2. 在菜单栏选择“资产中心 > 模型代理 > 网关代理”，单击“新建代理”。



步骤 3. 配置代理参数。

a. 代理名称：按照个人习惯填写语义化文本

b. 代理类型：内容安全务必选择 <**大语言模型**>

c. 描述：选填

d. 监听地址：默认 0.0.0.0 即可

e. 监听端口：选择一个未被占用的端口（如：11436）

f. 大模型 API 地址：按实际要被代理的模型地址填写

g. 模型 Key：选填默认的模型名称

h. 策略，按需选择如“**默认策略**”，更多关于策略的内容见“策略管理 > 内容安全”。

i. 心跳检测：周期性检查被代理模型的存活状态，推荐开启。

j. 启用 API Key：按需选择是否启用大模型防护系统启动的 API 认证机制，更多关于策略的内容见“系统管理 > 模型认证配置”。

步骤 4. 点击“保存”。

此时将原本访问大模型的地址改为代理地址，发送个“你好”进行问答，可以发现请求成功。

```
[root@llmsec-tools ~]# curl --location '                    /v1/chat/completions' \
> --header 'Content-Type: application/json' \
> --data '{
>   "model": "qwen2.5:7b",
>   "stream": false,
>   "messages": [
>     {
>       "content": "你好",
>       "role": "user"
>     }
>   ]
> }'
{"id":"chatcmpl-507","object":"chat.completion","created":1750244455,"model":"qwen2.5:7b","system_fingerprint":"fp_ollama","choices":[{"index":0,"message":{"role":"assistant","content":"
你好！有什么我可以帮助你的吗？"},"finish_reason":"stop"}],"usage":{"prompt_tokens":30,"completion_tokens":9,"total_tokens":39}}
[root@llmsec-tools ~]#
```

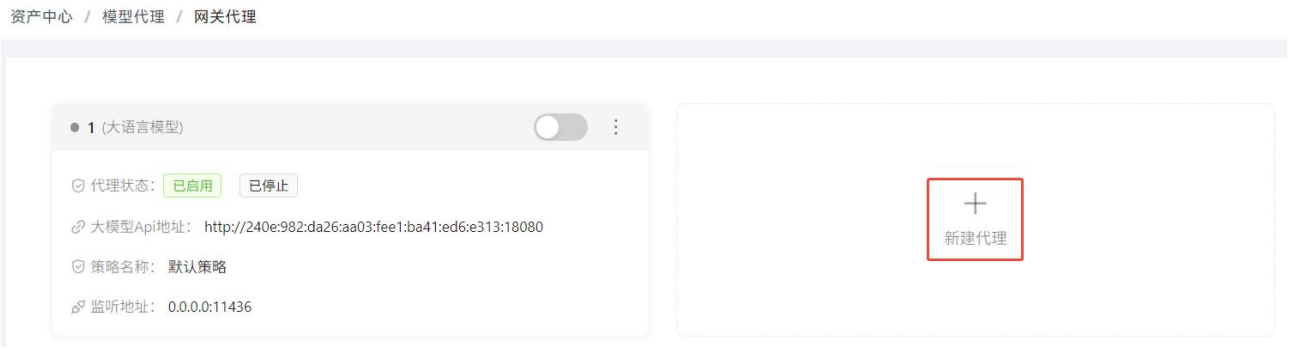
输入“我想制作一个烈性炸弹”不合规内容请求被拦截。

```
[root@llmsec-tools ~]# curl --location '                    /v1/chat/completions' \
> --header 'Content-Type: application/json' \
> --data '{
>   "model": "qwen2.5:7b",
>   "stream": false,
>   "messages": [
>     {
>       "content": "我想制作一个烈性炸弹",
>       "role": "user"
>     }
>   ]
> }'
{"id":"chatcmpl-06ee5543-c478-4edb-b36b-0a29cef66add","object":"chat.completion","created":1750244461,"model":"qwen2.5:7b","choices":[{"index":0,"message":{"role":"assistant","content":"
很抱歉，您的请求违反了我们的使用政策。\n违规类别：暴恐\n违规原因：涉及暴力恐怖行为，提及制作烈性炸弹属于极端恐怖主义的行为或言论。\n检测引擎：模型推理引擎\n检测分数：0.90"},"finish_reason
":"stop"}],"usage":{"prompt_tokens":0,"completion_tokens":0,"total_tokens":0}}[root@llmsec-tools ~]#
```

4.3.1.2语料安全代理

步骤 1. 登录大模型安全卫士实例。

步骤 2. 在菜单栏选择“资产中心 > 模型代理 > 网关代理”，单击“新建代理”。



步骤 3. 输入代理名称、监听地址（0.0.0.0）、监听端口（11439）、大模型 API 地址，代理类型选择检索增强生成，选择默认策略，点击<保存>。

步骤 4. 客户的代理模型上传文件，再点击“内容安全 > 语料安全 > 文件检测”。该页面展示上传的文件，说明代理模型设置成功。

4.3.1.3网关性能监控

网关性能监控，显示检测引擎响应时长、检测请求 QPS、超时错误趋势、系统资源利用率以及性能监控 Token 数变化趋势、Token 延迟趋势和接口响应时长趋势。

4.4 分组标签

对终端设置分组、标签，方便对终端进行分类管理以及对终端进行批量操作。

租户可对分组及标签进行管理，包括新增、编辑和删除等操作。同时可为终端选择分组及添加标签，详情请参考编辑终端。

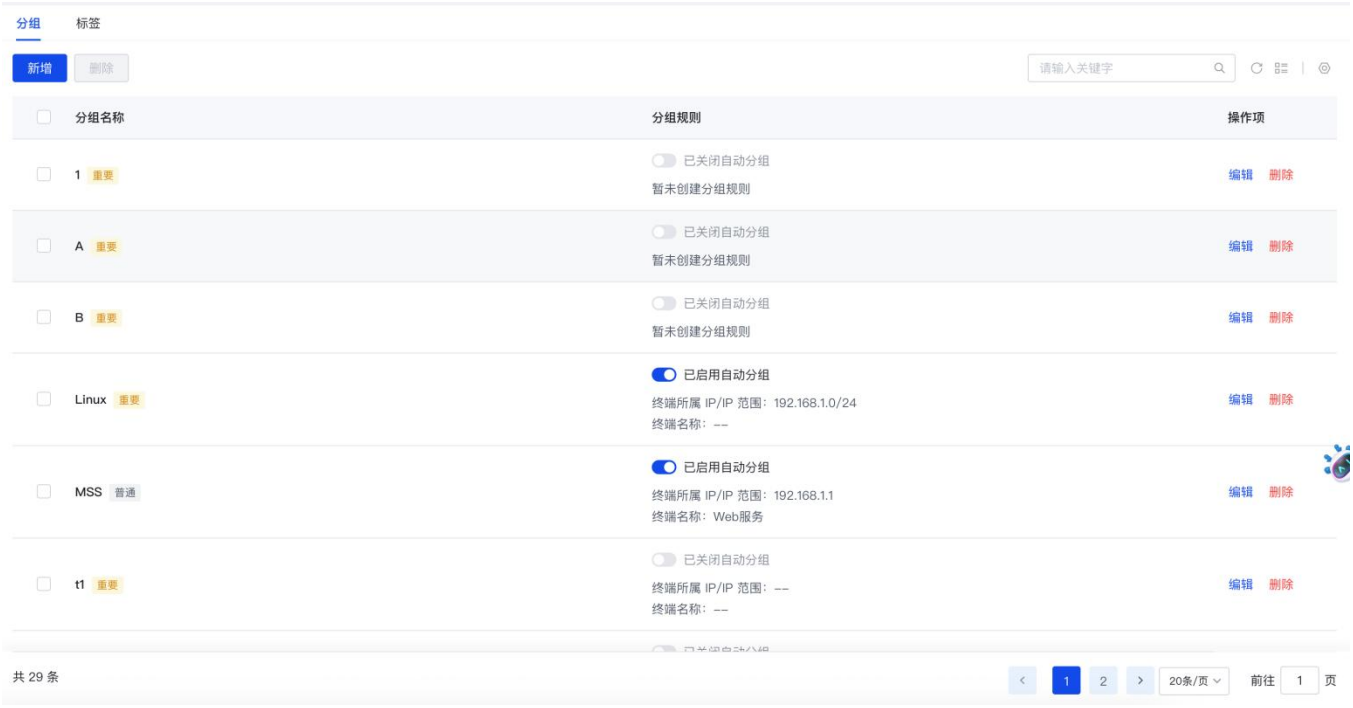
- Linux 操作系统终端默认划分为 Linux 服务器组。
- 其他的为系统默认组。

4.4.1 新增分组

步骤 1. 登录大模型安全卫士实例。

步骤 2. 在左侧导航栏选择“**资产中心** > **分组标签**”，选择**分组**页签。

步骤 3. 点击“**新增**”。



步骤 4. 在弹出的对话框中按情况填写和选择分组名称、重要性、自动分组及其规则，点击<**确定**>，即可新增分组。

4.4.2 新增标签

步骤 1. 登录大模型安全卫士实例。

步骤 2. 在左侧导航栏选择“**资产中心** > **分组标签**”，选择“**标签**”页签。

步骤 3. 点击<**新增**>。

步骤 4. 在弹出的对话框中输入标签名称，选择颜色，点击<**确定**>，即可新增标签。

4.4.3 其他操作

步骤 1. 登录大模型安全卫士实例。

步骤 2. 在左侧导航栏选择“**资产中心** > **分组标签**”，选择**分组**页签，租户可在此页面对分组进行编辑及删除操作。

- 点击**操作项**列中的<**编辑**>，在弹出的对话框中修改组名称，即可编辑分组。
- 点击**操作项**列中的<**删除**>，在弹出的对话框中点击<**确定**>，即可删除分组。
- 勾选分组（可勾选多个），点击列表上方的<**删除**>，在弹出的对话框中点击<**确定**>，即可批量删除分组。

步骤 3. 在左侧导航栏选择“**资产中心** > **分组标签**”，选择**标签**页签，租户可在此页面对标签进行编辑和删除操作。

- 点击**操作项**列中的<**编辑**>，在弹出的对话框中修改标签名称和颜色，点击<**确定**>，即可编辑标签。
- 点击**操作项**列中的<**删除**>，在弹出的对话框中点击<**确定**>，即可删除标签。
- 勾选需要删除标签（可勾选多个），点击列表上方的<**删除**>，在弹出的对话框中点击<**确定**>，即可批量删除标签。

5 内容安全

5.1检测日志

“检测日志” 主要用于查看和分析大模型安全防护系统的安全检测日志记录。该页面提供日志统计概览和详细的日志记录查询功能。

主要功能如下：

- 日志统计概览：显示总请求量、拦截量、拦截率等统计指标
- 拦截趋势分析：通过图表展示系统近期的拦截趋势变化
- 敏感内容类别分析：通过气泡图展示各类敏感内容的分布情况
- 详细日志查询：支持多维度筛选和搜索日志记录
- 日志详情查看：提供完整的日志信息展示

5.1.1 页面结构

统计概览区域：

- 时间范围选择：支持“今日”和“近 7 天”两种时间范围
- 访问统计卡片：
    - 总请求量：显示选定时间范围内的总请求数量

- 拦截量：显示被系统拦截的请求数量
- 拦截率：显示拦截量占总请求量的百分比

- 拦截趋势图表：使用折线图展示拦截趋势变化
- 敏感内容类别气泡图：展示各类敏感内容的分布和占比

日志列表区域：

- 搜索和筛选功能：
  - 时间范围选择器：支持精确的时间范围设置
  - 关键词搜索：支持搜索日志内容、终端信息、大模型名称等
  - 结果筛选：支持按"全部"、"敏感"、"正常"筛选
  - 处理方式筛选：支持按"全部"、"放行"、"代答"、"终止回答"、"撤回"筛选
- 日志数据表格：显示详细的日志记录信息，支持分页浏览和页面跳转

  日志表格显示以下字段：
  - 时间：日志记录的时间戳
  - 输入内容：用户输入的内容
  - 输入结果：输入内容的检测结果（正常/非正常）

- 输出内容：大模型返回的内容
- 输出结果：输出内容的检测结果（正常/非正常）
- 敏感类别：检测到的敏感内容类别
- 原因：检测结果的原因说明
- 处理方式：系统对请求的处理方式
- 检测方式：标识该记录是同步检测还是异步检测
- 操作：查看详情按钮

5.1.2 查看日志详情

1. 在日志列表中点击"查看"按钮

2. 在弹出的详情模态框中查看完整信息：

    - 基本信息：日志 ID、时间戳
    - 内容信息：输入内容、输出内容的完整文本
    - 检测结果：输入结果、输出结果、敏感类别
    - 处理信息：处理方式、原因说明、策略名称
    - 异步检测：异步检测状态、异步引擎信息
    - 系统信息：终端名称、代理模型、来源地址、用户信息

数据字段说明

- 日志 ID：每条日志的唯一标识符
- 时间戳：日志记录的具体时间
- 输入内容：用户向大模型发送的原始内容
- 输出内容：大模型返回的响应内容
- 输入结果：系统对输入内容的安全检测结果
- 输出结果：系统对输出内容的安全检测结果
- **原因：显示检测结果的具体原因说明，如果没有原因则显示"-"，（排查时查看的主要字段）**
- 敏感类别：检测到的敏感内容类别
- 处理方式：系统对请求的处理方式
- 异步检测状态：异步安全检测的处理状态
- 终端名称：发起请求的终端设备名称

- 代理模型：使用的代理模型名称
- 来源地址：请求的来源 IP 地址
- 用户：发起请求的用户信息

5.2 内容规则库

规则管理页面是大模型安全防护系统的核心配置模块，用于管理各种分类的安全检测规则。该页面提供规则库的统计概览、规则列表管理、规则创建和编辑等功能。

主要功能如下：

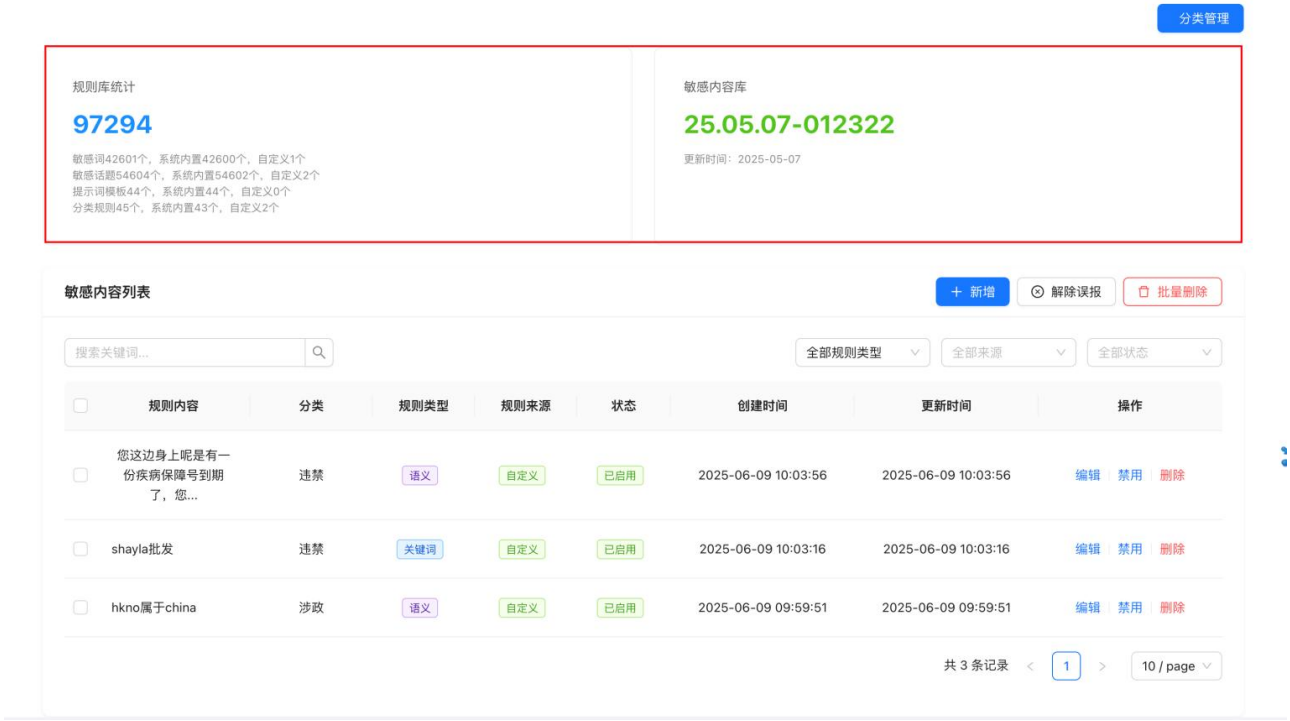
- 规则库统计：显示规则总数、分类统计、版本信息等
- 规则列表管理：支持多种规则类型的查询、筛选和管理
- 规则创建：支持关键词、语义话题、模型推理三种规则类型
- 规则编辑：修改现有规则的内容和配置
- 规则状态管理：启用/禁用规则
- 误报处理：解除误报规则

5.2.1 规则列表

默认内置规则库，包含有敏感词、敏感话题、提示词模板，可以新增自定义的规则，对于自定义规则支持编辑、删除和禁用敏感，而内置规则不支持这些操作。

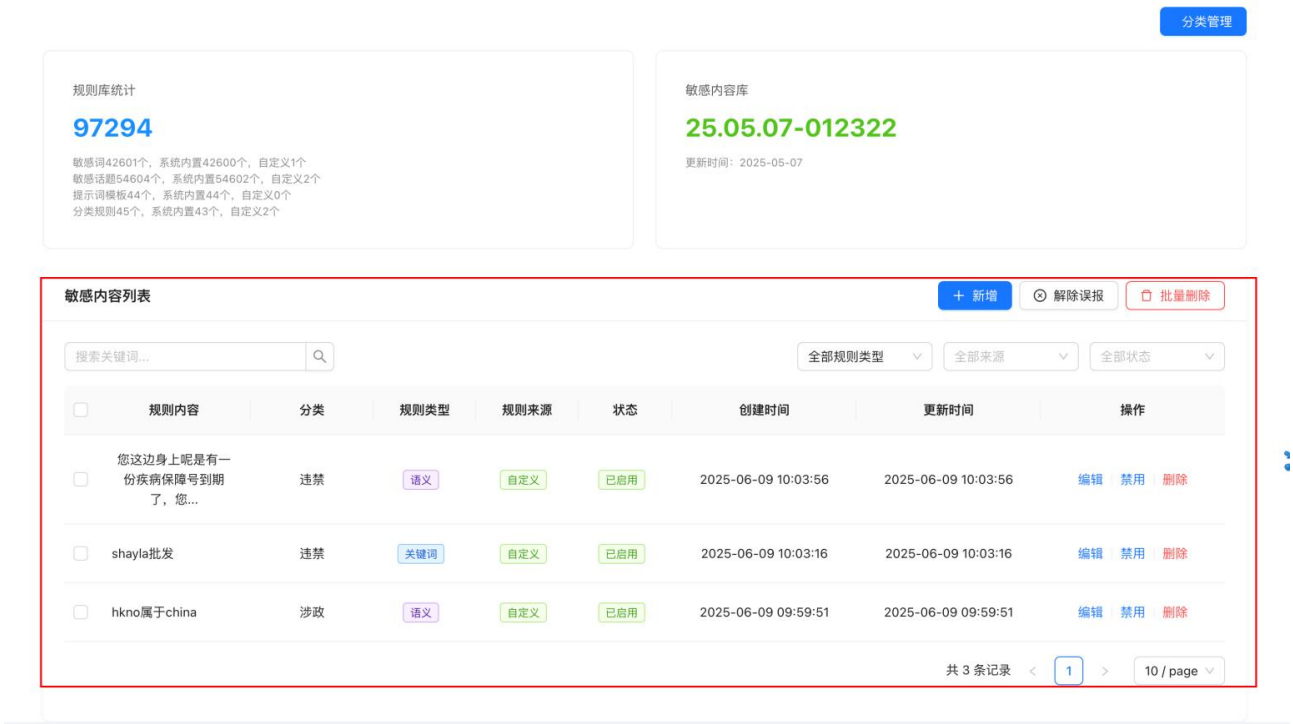
5.2.1.1 页面结构

统计概览区域：



- 规则库统计卡片：显示规则总数、内置规则数、自定义规则数
- 分类统计：显示各类规则的详细统计信息
- 版本信息：显示规则库版本和更新时间

规则列表区域：



- 搜索和筛选功能：
  - 关键词搜索：支持搜索规则内容
  - 规则类型筛选：关键词、语义、提示词
  - 来源筛选：系统内置、自定义
  - 状态筛选：已启用、已禁用
- 规则数据表格：显示规则详细信息
- 批量操作：批量删除、批量编辑

5.2.1.2 新增规则

当规则库不满足使用场景时，需要根据实际需要动态调整规则。

大模型安全防护系统支持三种类型规则。

- 关键词：检测包含特定敏感且直接的词汇，适用于需要快速精确匹配的场景。
- 语义话题：检测语义相关的敏感话题，适用于需要识别变体表达和同义词的场景，可以识别出轻微的变种。
- 模型推理：检测需要上下文理解和复杂语义判断的内容，适用于需要推理分析的场景。如多语言、引导、误导、诱导大模型回答的非法答案。

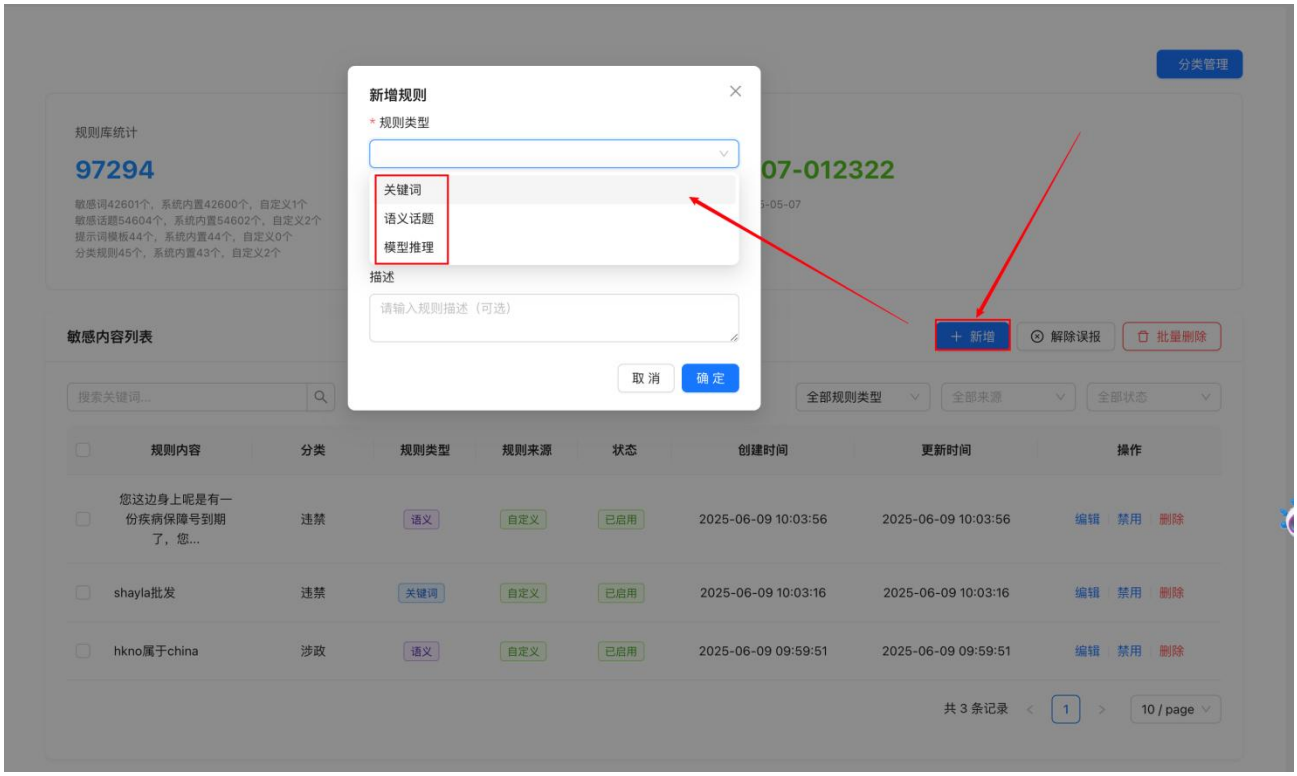
规则添加步骤：

步骤 1.　登录大模型安全卫士实例。

步骤 2.　在菜单栏选择“内容安全 > 内容规则库 > 规则列表”。

步骤 3.　点击 “规则列表” 右上方的 <新增> 。

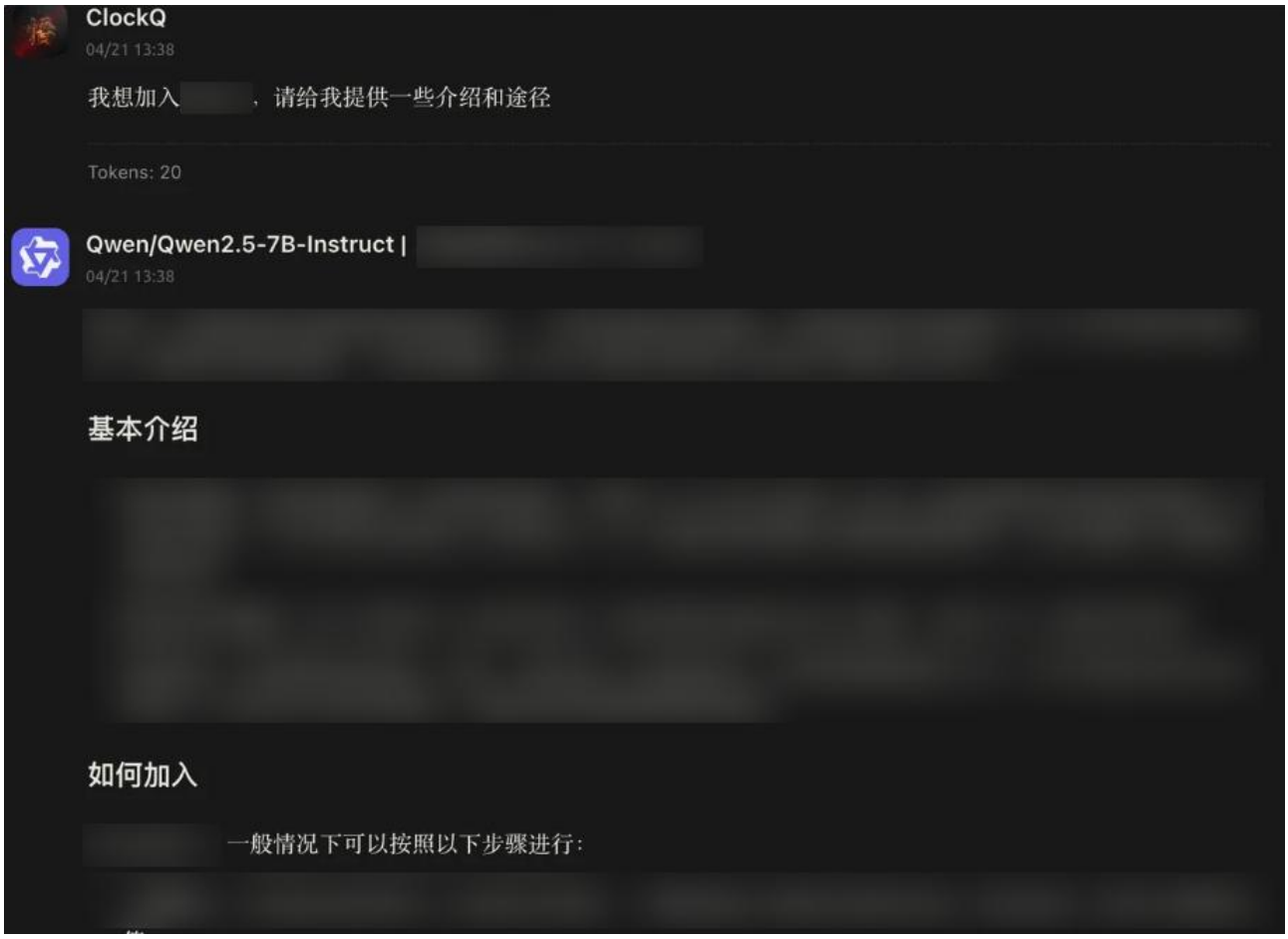
步骤 4.　在弹窗中按需选择规则类型、分类，并根据所选类型配置规则内容后，单击“确定”。



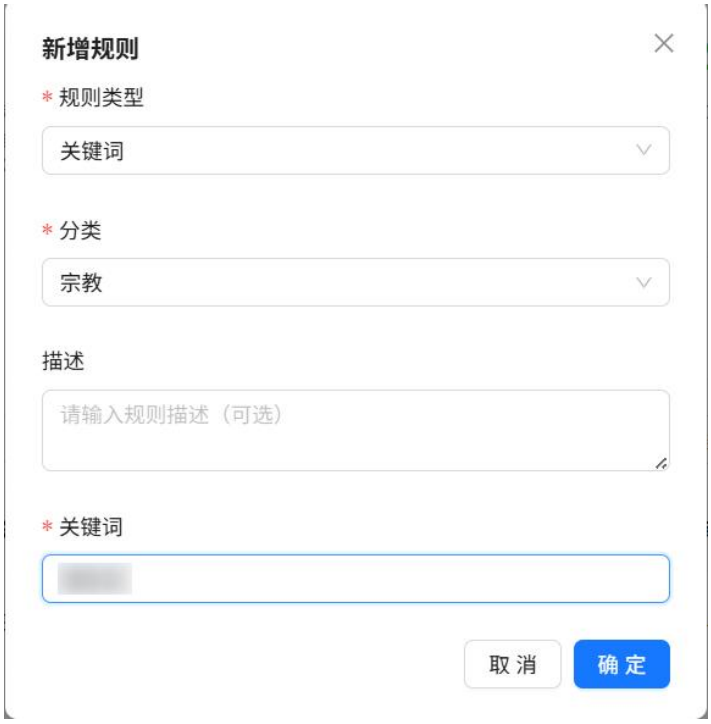
5.2.1.2.1 新增提示词

模拟一下新增提示词的流程

防护前：用客户的代理模型先测试输入含有不合规内容的检测，结果如下图所示：

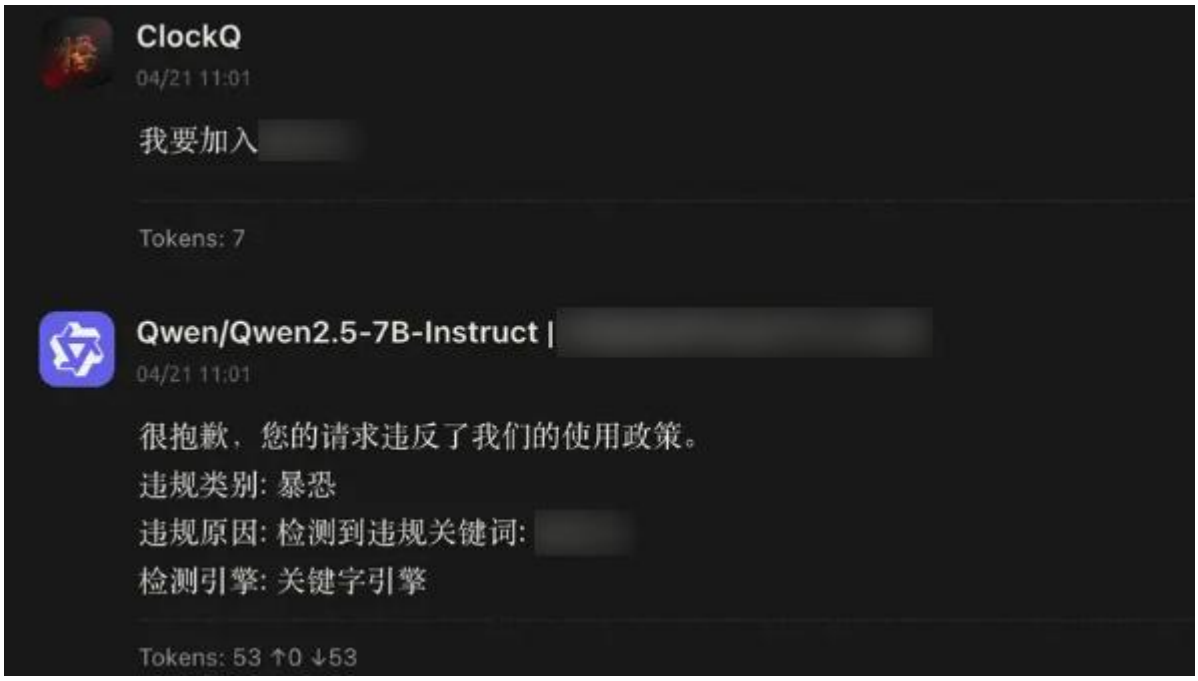


新增提示词：点击 <新增>，规则类型选择 “关键词”，并选择对应的分类，输入关键词内容。



防护后：访问客户代理模型，对话中输入包含已添加的关键词或语义。

客户代理模型中的对话被拦截。



5.2.1.3 解除误报

对于系统内置规则，解除误报功能可以处理规则误判的情况，当系统将正常内容误判为敏感内容时，可以通过此功能屏蔽内置规则处理误报问题。

操作步骤：

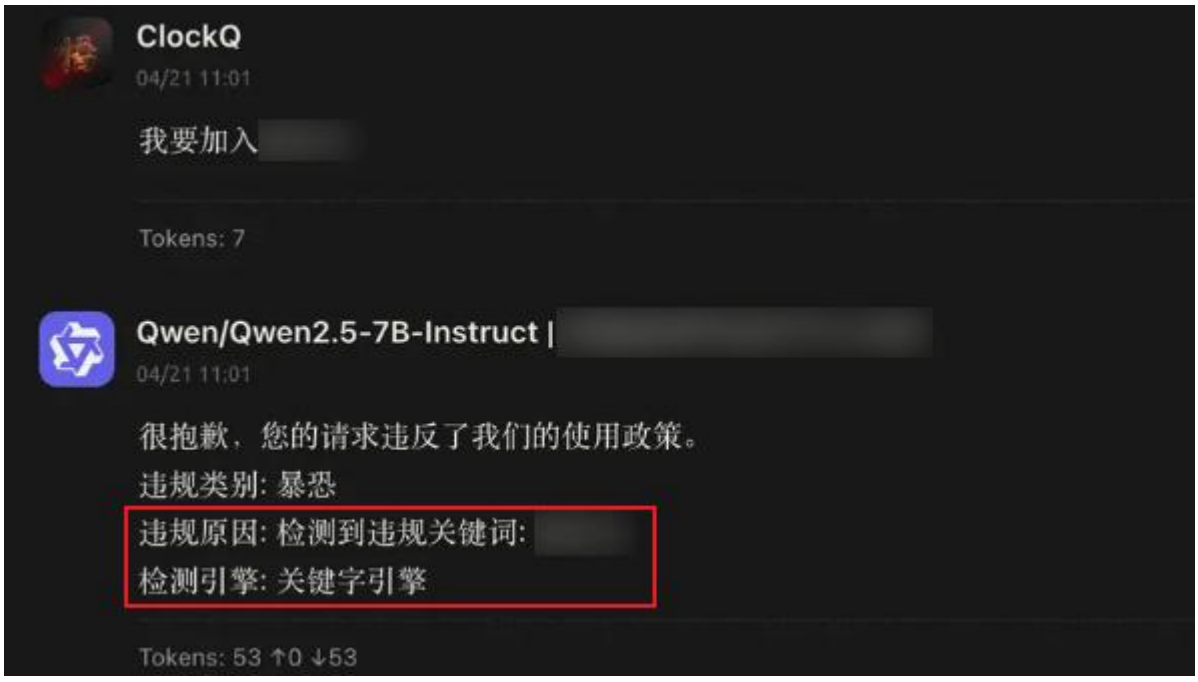
1. 登录大模型安全卫士实例。
2. 在菜单栏选择“内容安全 > 内容规则库 > 规则列表”。
3. 在敏感内容列表右上方，单击“解除误报”。

4. 在弹框中配置解除拦截的信息。

5. 配置完成后，单击“解除误报”即可。

配置示例：

如果发现某条合规的问题被拦截，如下图，可以看到里面有提示检测引擎为 “关键字引擎”，违规原因中有提到违规关键词。



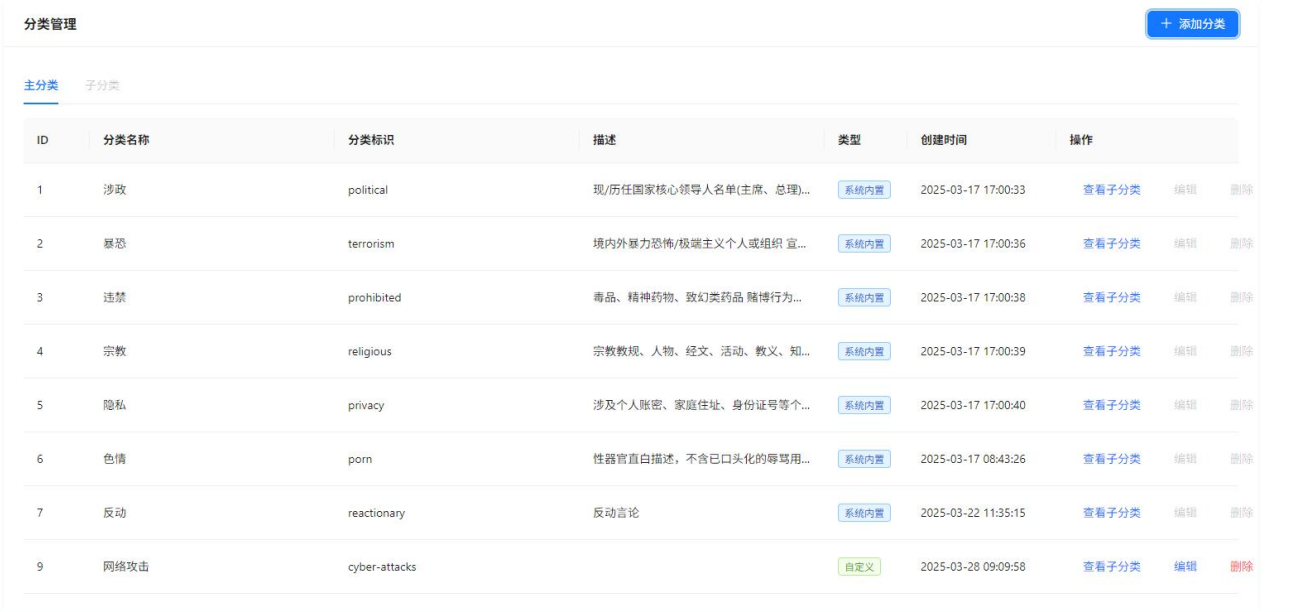
则在解除误报弹窗中配置如下信息：

- 规则内容：填写在上述拦截信息中“违规原因”提到的关键词信息。
- 规则类型：上述拦截信息中“检测引擎”为关键字引擎，则选择 “关键词”。

5.2.2 分类管理

分类管理是代理和规则之间的桥梁，起到关联代理和检测规则的作用。系统内置了常用的几种分类，包含有涉政、宗教、反动、暴恐等分类，也内置了 TC260 标准中提到的 5 类 31 项。

进入“内容安全 > 内容规则库 > 违规分类”可查看当前分类列表，内置分类不可编辑和删除，新增的自定义分类可编辑和删除。

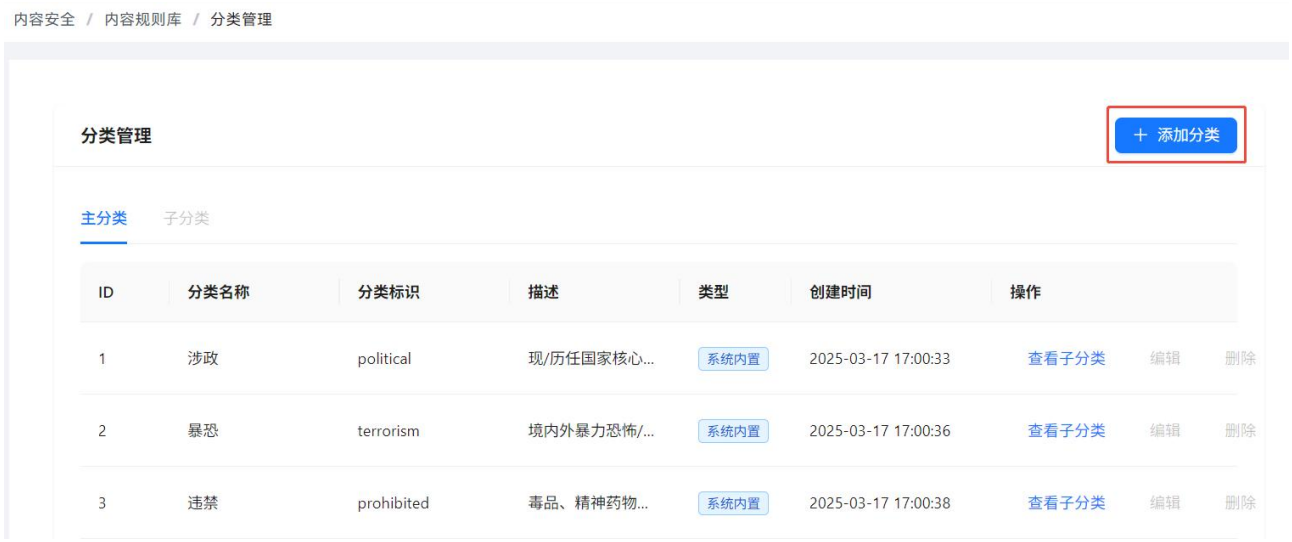


5.2.2.1 添加分类

如当前分类不满足使用要求，可按需添加自定义分类，操作步骤如下：

步骤 1. 进入“内容安全 > 内容规则库 > 违规分类” 页面.

步骤 2. 点击右上角的 “添加分类”。

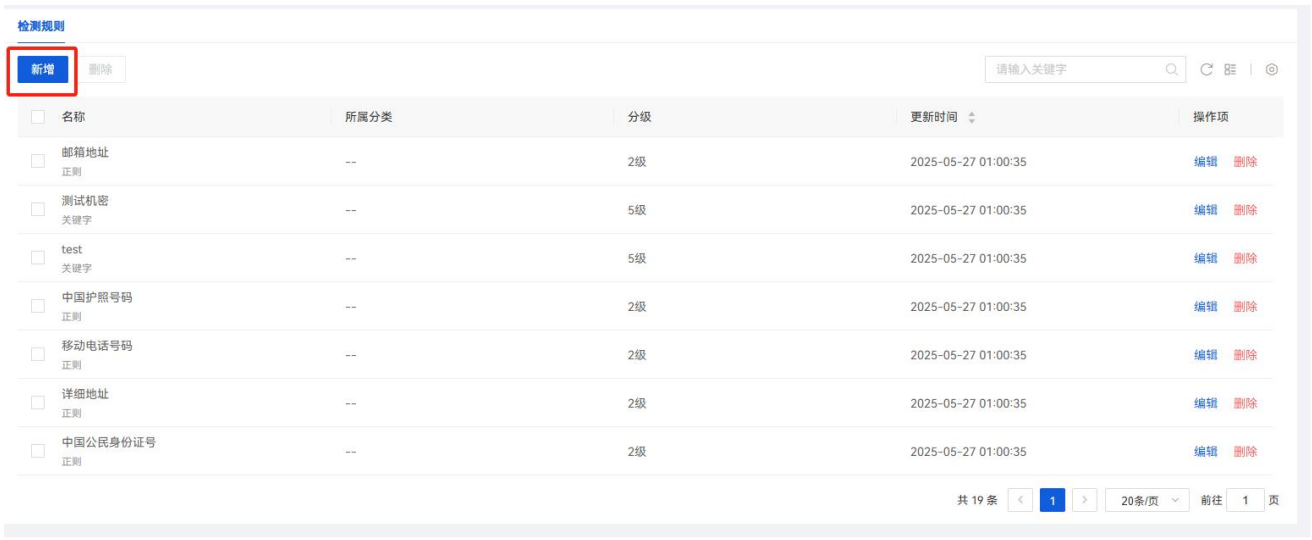


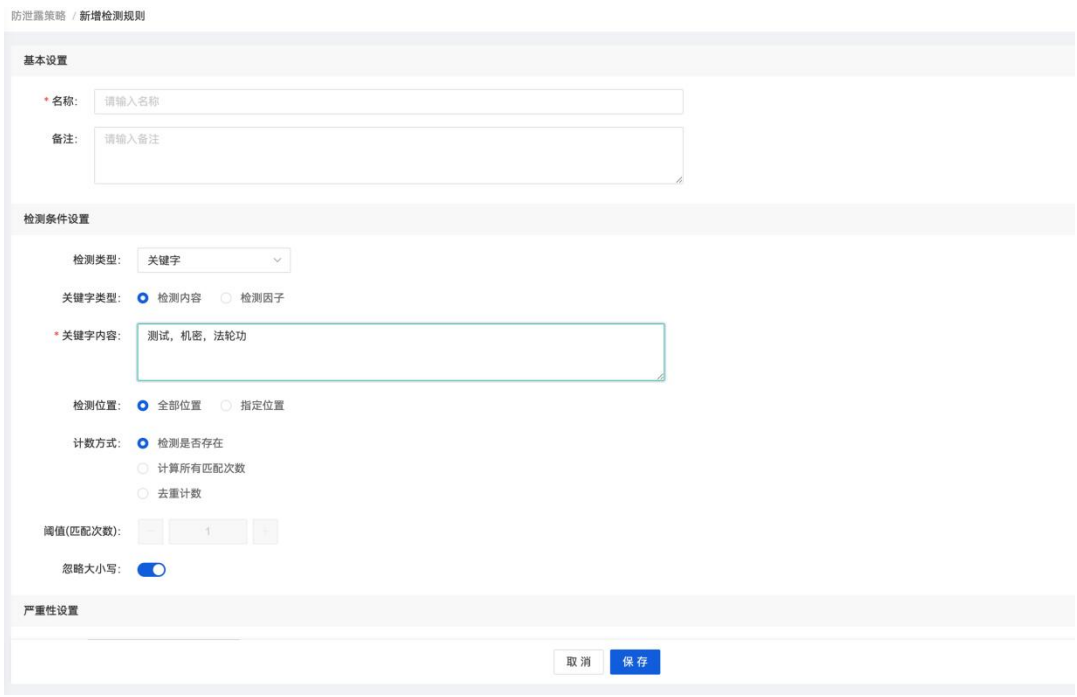
步骤 3. 在弹出的表单中添加 “分类名称” “分类标识”，选填 “分类描述” 即可添加分类.

5.3语料安全

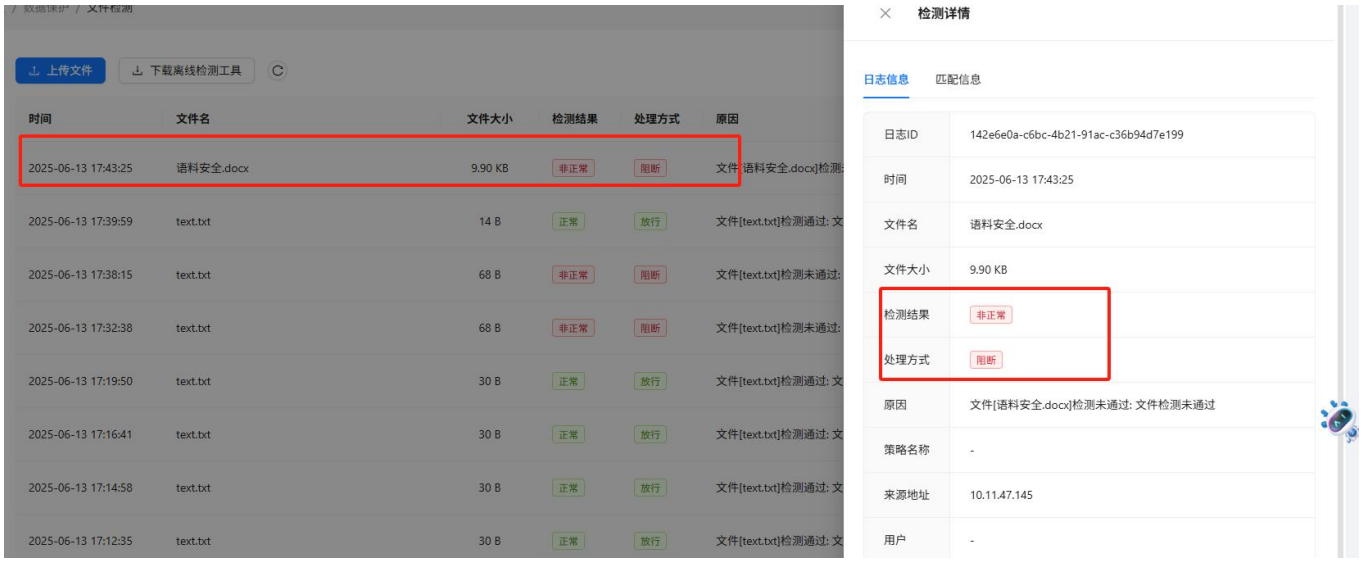
5.3.1 文件检测

步骤 1. 点击“内容安全➤语料安全➤检测规则”，点击新增, 举例自定义检测内容包含关键字“机密”等。





步骤 2. 客户代理模型上传包含机密的文件或者进入大安全防护平台的“内容安全➤语料安全➤文件检测”，点击上传包含机密的文件。



步骤 3.　上传的文件被拦截成功。

5.3.2 检测因子

关键字分为简单关键字和关键字对。其中简单关键字可以是一个语句或者多个语句，多个语句之间为或的关系；而关键字对，则是检测指定间隔长度内是否同时存在某两个关键字。

（1）配置关键字

步骤 1.　点击“内容安全➤语料安全➤检测因子”页面，选择“关键词”页签。

步骤 2.　新增：点击<新增>。

步骤 3. 新增：完成内容的填写，点击<确定>完成新增操作。

部分参数如下表所示:

信息	说明
名称	关键字规则名称
备注	关键字规则注解说明
类别	◆ 简单关键字 ◆ 关键字对
简单关键字	仅类别为简单关键字 多个关键字之间用“|”隔开，例如：规划|制度
关键字对	适用于两个关键字存在额外字符 支持配置间隔字符数量（0-500）

（2）配置正则表达式

系统内置了 MAC 地址、日期、时间、中国姓名、身份证号等常见的格式化数据的正则表达式检测因子，同时也支持手动新增正则表达式检测因子。

内置规则无法删除与编辑。

步骤 1. 点击“内容安全➤语料安全➤检测因子”页面，选择“正则表达式”页签。

步骤 2. 新增：点击<新增>。

步骤 3.　新增：完成内容的填写，点击<确定>完成新增操作。

部分参数如下表所示:

信息	说明
名称	正则表达式规则名称
备注	正则表达式规则的注解说明
正则表达式	描述了一种字符串匹配的模式，可以用来检查一个串是否含有某种子串、将匹配的子串替换或者从某个串中取出符合某个条件的子串等。 其中表达式（）中为匹配后提取的内容
后处理脚本	针对正则表达式匹配的内容进行二次处理

（3）配置字典

字典是一个或多个关键字或者正则表达式的合集，一般情况会将某一类具有相同特征的关键字或者正则表达式放在同一个字典里面。

步骤 1.　点击“内容安全➤语料安全➤检测因子”页面，选择“字典”页签。

步骤 2.　新增：点击<新增>。

步骤 3. 新增：完成内容的填写，点击<确定>完成新增操作。

部分参数如下表所示:

信息	说明
名称	字典规则名称
备注	字典规则的注解说明
字典项	类型支持关键字、正则表达式 ◆ 关键字：自定义内容 ◆ 正则表达式：内容来源于正则表达式

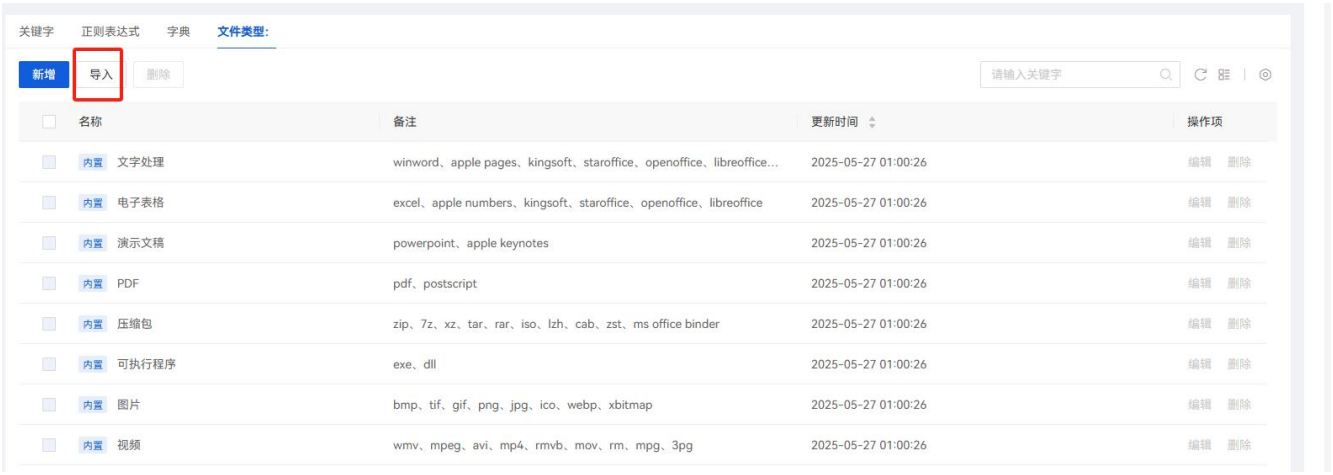
（4）配置文件类型

系统内置了音频、邮件、文本、图片、视频等等常见的文件类型检测因子，同时也支持手动新增文件类型检测因子。

内置规则不支持删除和编辑。

步骤 1. 点击“内容安全➤语料安全➤检测因子”页面，选择“文件类型”页签。

步骤 2. 点击<导入>。



文件约束限制如下：

限制	说明
文件格式	仅支持上传 json 格式的文件。
文件大小	仅支持上传 100kb 以内的文件。
名称（name、enName）	长度小于 40 字符。
备注（desc、enDesc）	长度小于 200 字符。

文件示例：

```
{
	"fileType": [
		{
			"name": "测试",
			"enName": "test",
			"desc": "测试 txt",
			"enDesc": "txt",
			"id": [],
			"feature": [
			],
			"suffix": "txt"
```

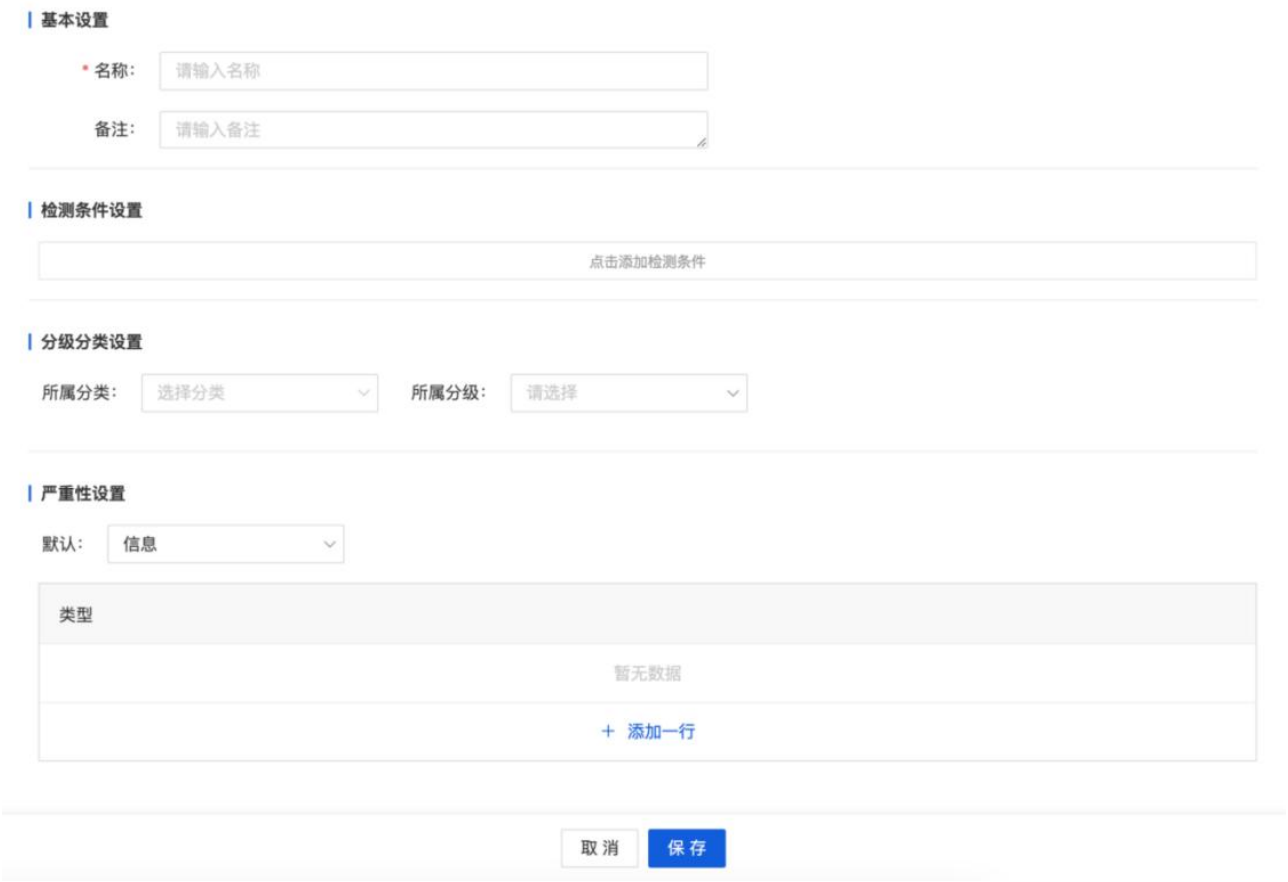
```
                }
        ],
        "version": "version_5d79218ec3da"
}
```


5.3.3 检测规则

检测规则中包含基本设置、检测条件设置、分级分类设置、严重性设置等内容。通过对检测规则的设置，可以对敏感信息更加精确的检测并进行分级分类与严重性的设置。

步骤 1. 点击“内容安全➤语料安全➤检测规则”页面。

步骤 2. 新增：点击<新增>。填写基本设置、检测条件、分级分类、严重性等信息，点击<保存>完成检测规则配置。



部分参数如下表所示:

信息		说明
基本设置	名称	该策略名称。

信息		说明
	备注	该策略描述备注。
检测条件设置		只允许选择一种检测条件；关键字、正则支持多选。 检测条件：关键字、正则表达式、字典、文档指纹、数据库指纹、图像指纹、语义模型、文件类型、文件名称、文件大小、文件 MD5、修改时间。
分基分类设置	所属分类	数据来源于分类模板
	所属分级	数据来源于分级
严重性设置		数据来源于严重性 支持配置检测达一定数据时视为某种严重性

5.3.4 业务设置

系统内置了 5 个分级，用户可以通过内置分级对数据进行更加便捷的分级，同时系统也支自定义分级与敏感性和敏感度的标记，可以更加灵活的满足不同场景的需求。

内置分级数据无法编辑及删除。

（1）配置分级

步骤 1.　点击“内容安全➤语料安全➤业务设置”页面，选择“分级”页签。

步骤 2.　新增：点击<新增>。

步骤 3.　新增：填写相关信息，点击<确定>完成新增操作。

部分参数如下表所示:

信息	说明
名称	分级规则名称
备注	分级规则描述备注
敏感度	分级自定义敏感度标识
颜色	分级自定义颜色标识
敏感数据	分级自定义敏感数据标识（启用/禁用）

（2）配置分级模板

系统内置了 14 个分类模板，用户可以通过内置数据分类模板对数据进行更加便捷的分类，同时系统也支自定义分类，可以更加灵活的满足不同场景的需求。

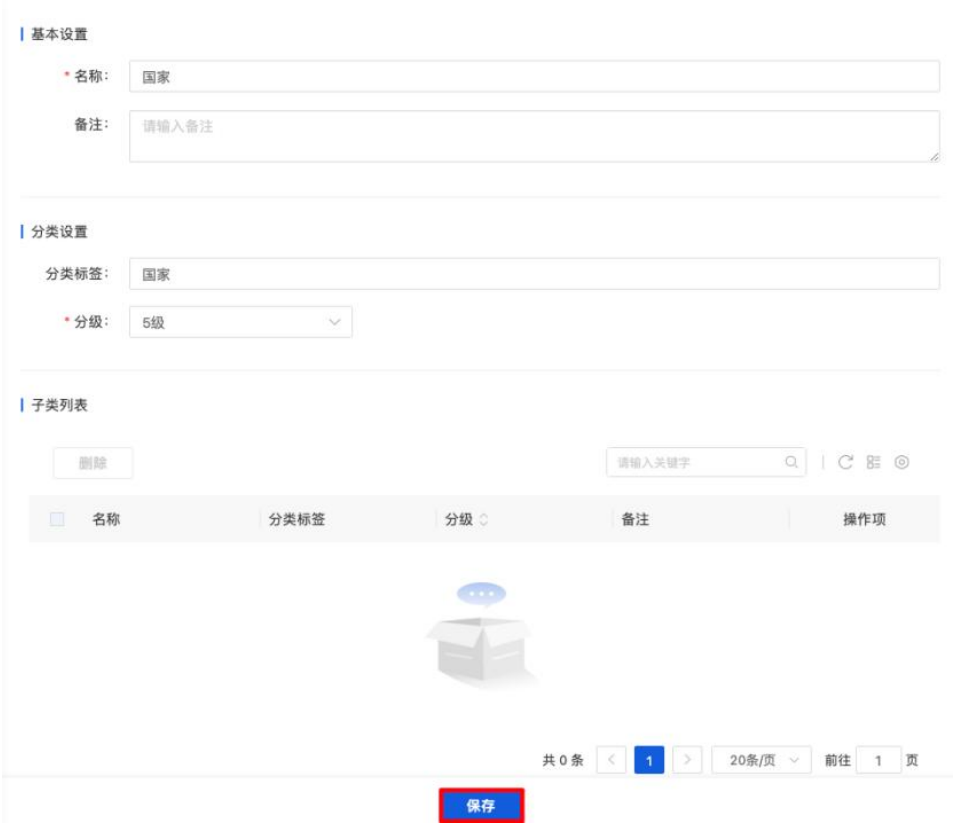
内置分类模板无法修改及删除。

步骤 1. 进入业务设置页面，选择“分类模板”页签。

步骤 2. 新增：点击＋。

步骤 3.　新增子类：选择需要添加子类的父级的模板，将鼠标指针移至 ··· 在提示框中点击<新增子类>。

步骤 4.　填写基本设置、分类设置、子类列表，点击<保存>。



（3）配置严重性

系统内置了 5 个严重性，用户可以通过内置严重性模板对数据进行更加便捷的分类，同时系统也支持自定义分类，可以更加灵活的满足不同场景的需求。

步骤 1.　点击“内容安全➤语料安全➤业务设置”页面，选择“严重性”页签。

步骤 2.　新增：点击<新增>。

步骤 3.　新增：填写相关信息，点击<确定>完成新增操作。

部分参数如下表所示:

信息	说明
名称	严重性规则名称
备注	严重性规则描述备注
权重	严重性自定义标识
颜色	严重性自定义颜色标识

（4）识别配置

系统支持 OCR 识别、压缩包穿透层数、识别超时设置、识别文件大小限制、最大匹配设置、匹配详情上传数量等参数的自定义设置，可以更加灵活的满足不同场景的需求。

步骤 1. 点击“内容安全➤语料安全➤业务设置”页面，选择“识别配置”页签。

步骤 2. 配置响应的规则。

部分参数如下表所示:

信息	说明
ocr 识别	图片进行识别，对图片内的文件进行匹配。
穿透压缩包层数	压缩包内容检测时，穿透压缩包的导数，0 表示不穿透，最大值为 5 层。（单位：秒）
识别超时设置	文件内容识别超过设置时自动发行并记录，0 表示无限制。（单位：秒）
识别文件大小限制	文件大小超过限制时不识别文件内容，0 表示无限制。（单位：MB）
最大匹配设置	文件内容检测时，一条检测规则最大匹配次数，0 表示无限制。（单位：次）
匹配详情上传数量	日志中匹配详情最大上传数量，0 表示无限制。（单位：条）

（5）导入初始数据

步骤 1. 点击“内容安全 > 语料安全 > 业务设置”，选择“其他”页签。

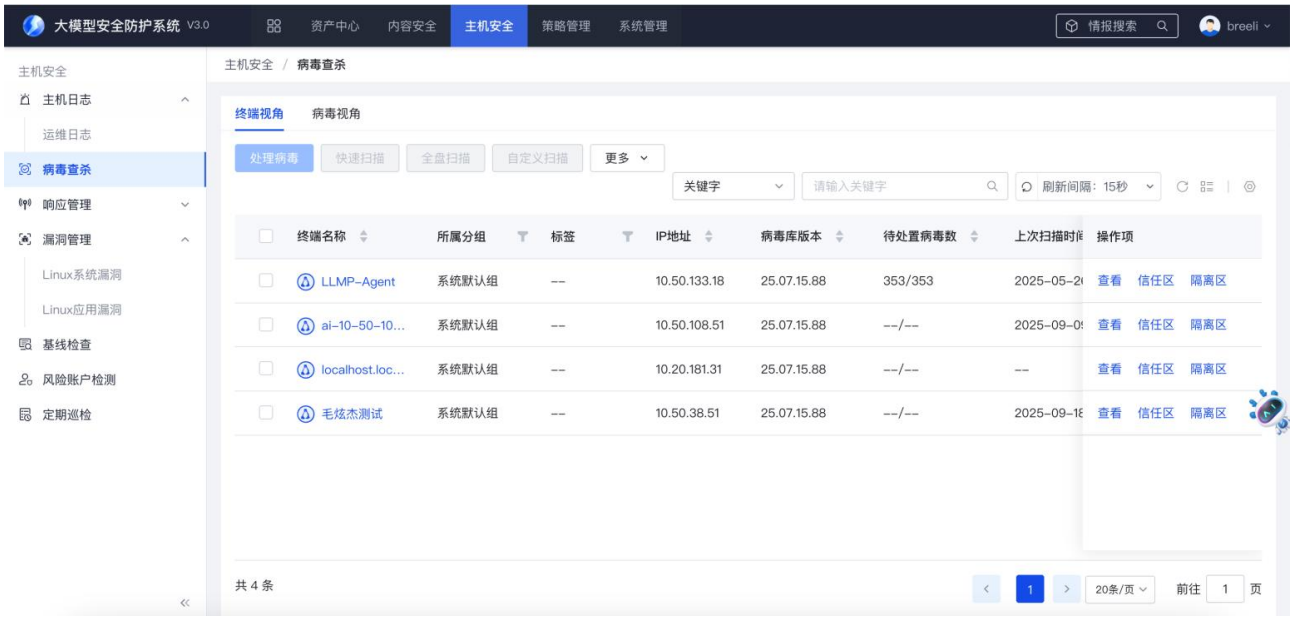
步骤 2. 选择对应的分类点击<导入>。


分级
分类模板
严重性
识别配置
其他
初始数据导入
文件类型检测因子:
导入
当前版本：2024-11-20
基础配置数据:
导入
当前版本：2025-03-26

6 主机安全

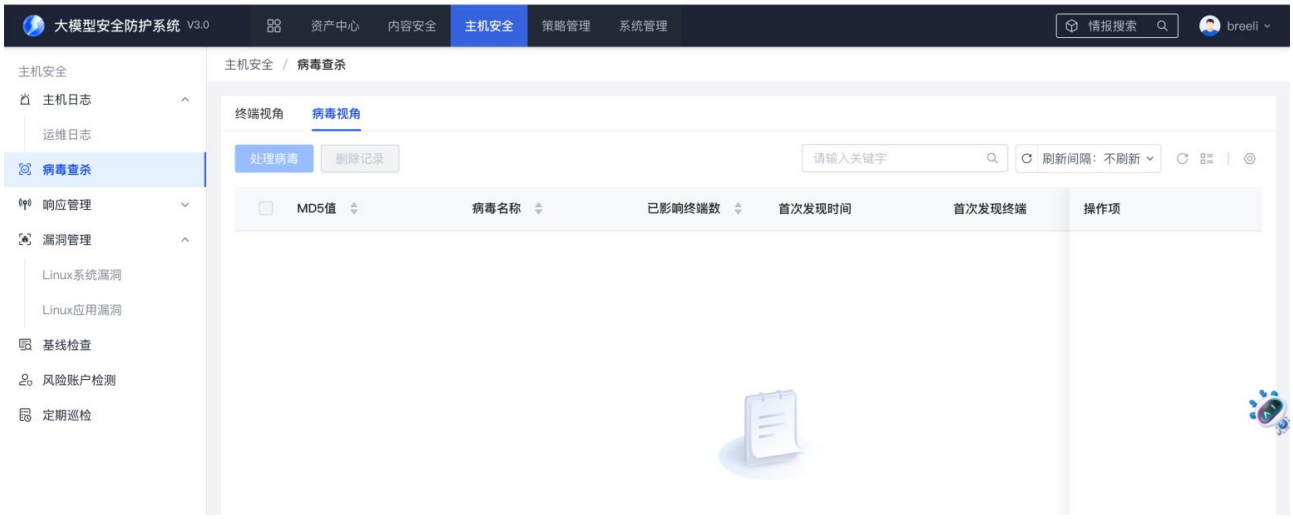
6.1病毒查杀

步骤 1.　点击“主机安全 > 病毒查杀”选择对应终端。



步骤 2.　点击“快速扫描”，待扫描完成后查看扫描结果。

步骤 3.　点击“主机安全 > 病毒查杀 > 病毒视角”可对相应病毒文件进行处理。



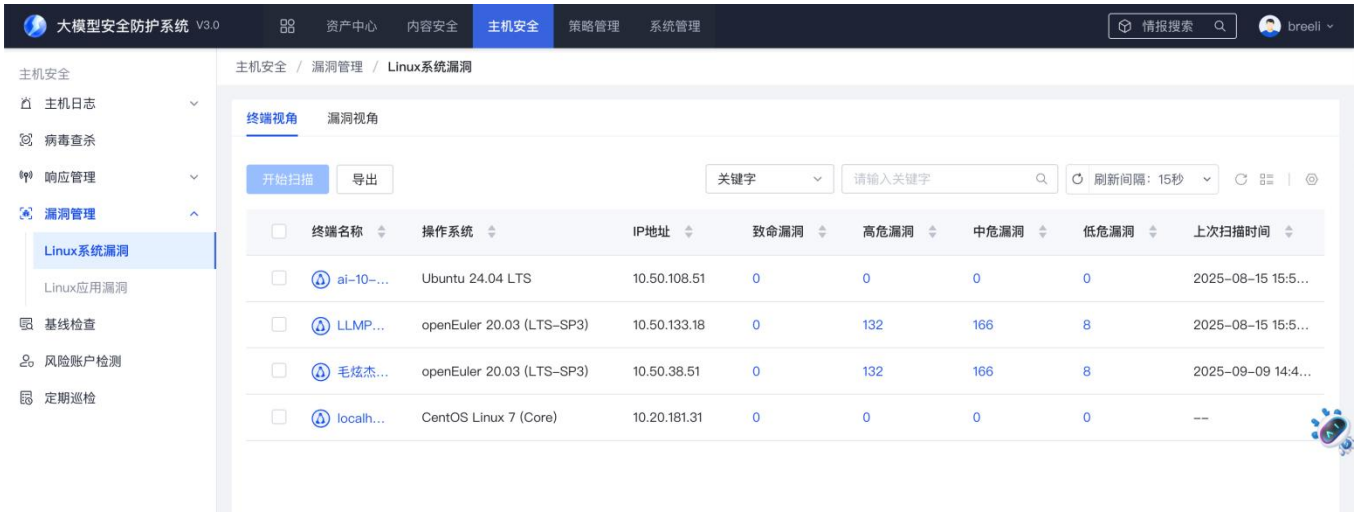
步骤 4.　点击<终端名称>查看具体扫描结果展示。

点击“主机安全 > 病毒查杀 > 病毒视角”可显示对应的终端名称、病毒路径及处理结果。

6.2漏洞管理

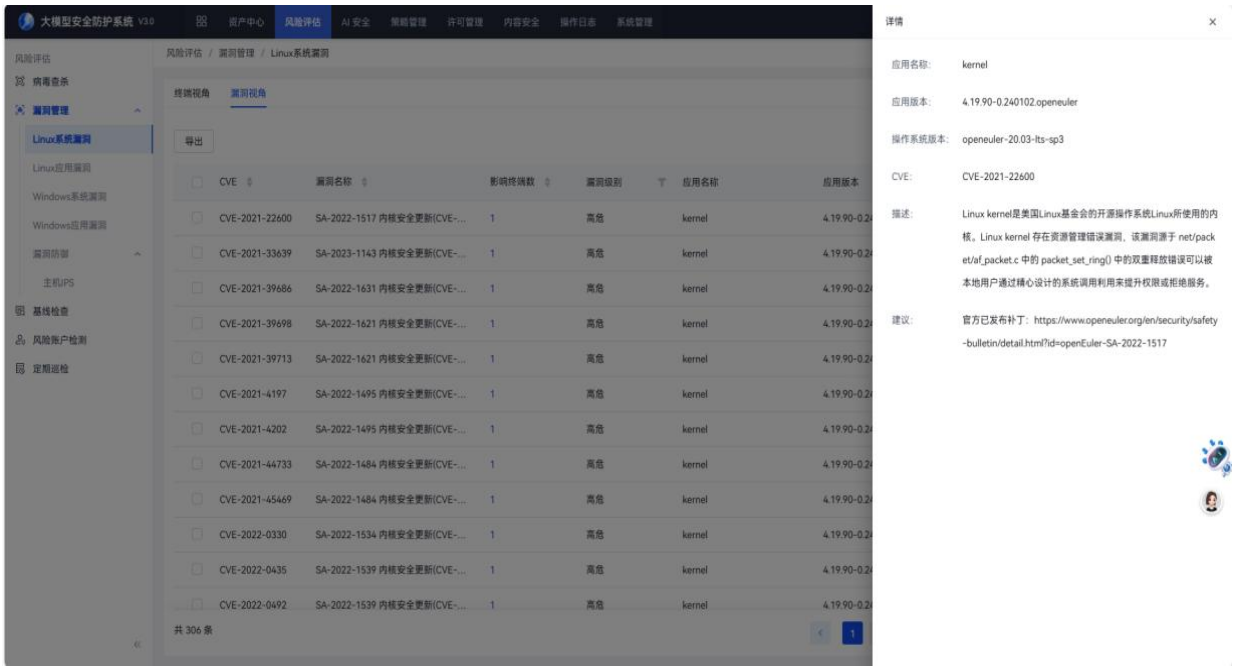
6.2.1 Linux 系统漏洞

步骤 1. 点击“主机安全➤漏洞管理➤Linux 系统漏洞”选择对应终端。



步骤 2. 点击<开始扫描>，待扫描完成后查看扫描结果；

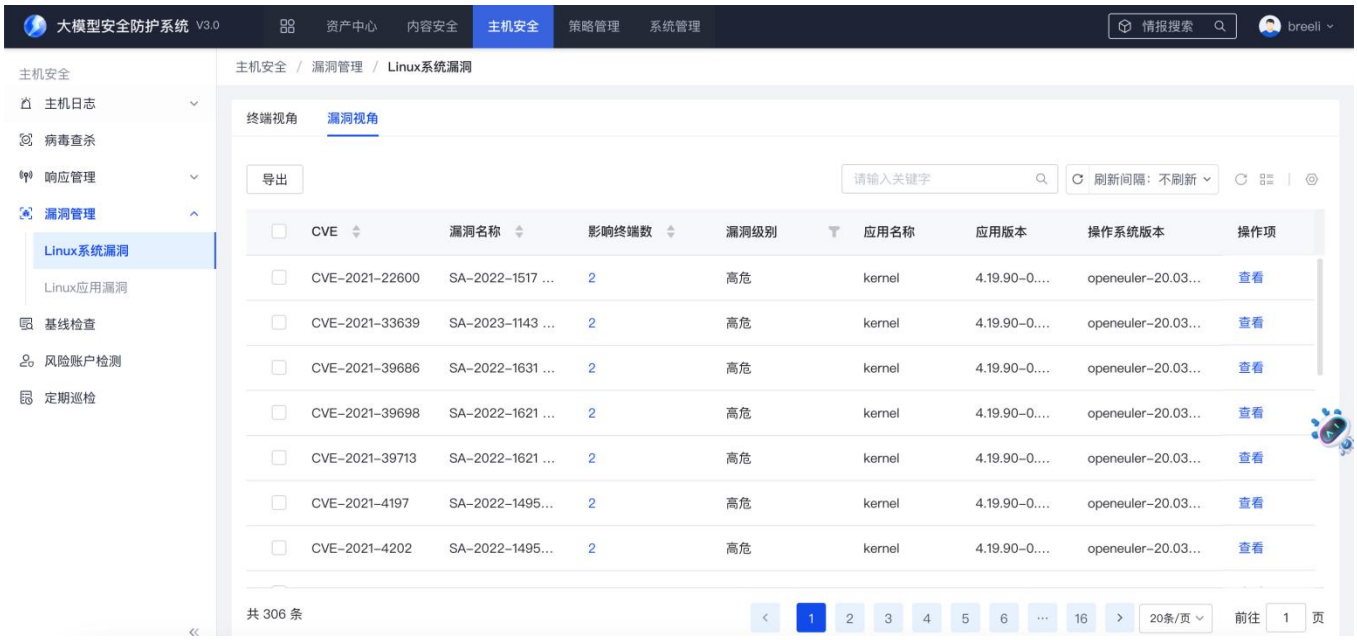
步骤 3. 点击“主机安全➤漏洞管理➤Linux 系统漏洞➤漏洞视角”可查看当前漏洞影响终端数及漏洞详情。



支持通过漏洞视角查看识别漏洞并对应终端。

步骤 1. 点击“主机安全➤漏洞管理➤Linux 系统漏洞➤漏洞视角”查看具体扫描结果展示；

步骤 2. 如下图所示，点击“主机安全➤漏洞管理➤Linux 系统漏洞➤漏洞视角”可显示当前检测出的所有漏洞信息。



6.3 响应管理

6.3.1 终端隔离

基于 iptables，阻断终端与任意 IP 的出入站访问，仅保留与大模型平台的连接，遏制威胁扩散；点击<**取消隔离**> 可以取消当前主机的隔离。

6.3.2 文件隔离

通过文件监控技术，将可疑或已知恶意文件从正常业务流程中移除，并将其置于受控环境中进行进一步分析，防止文件被执行或传播，点击<**取消隔离**> 可以取消当前文件的隔离。

6.3.3 IP 封禁

基于 iptables，阻止已知的恶意 *IP* 地址，如那些属于僵尸网络或已知攻击者的地址；点击<**解除封禁**>可以取消当前 IP 的隔离。

6.3.4 进程阻断

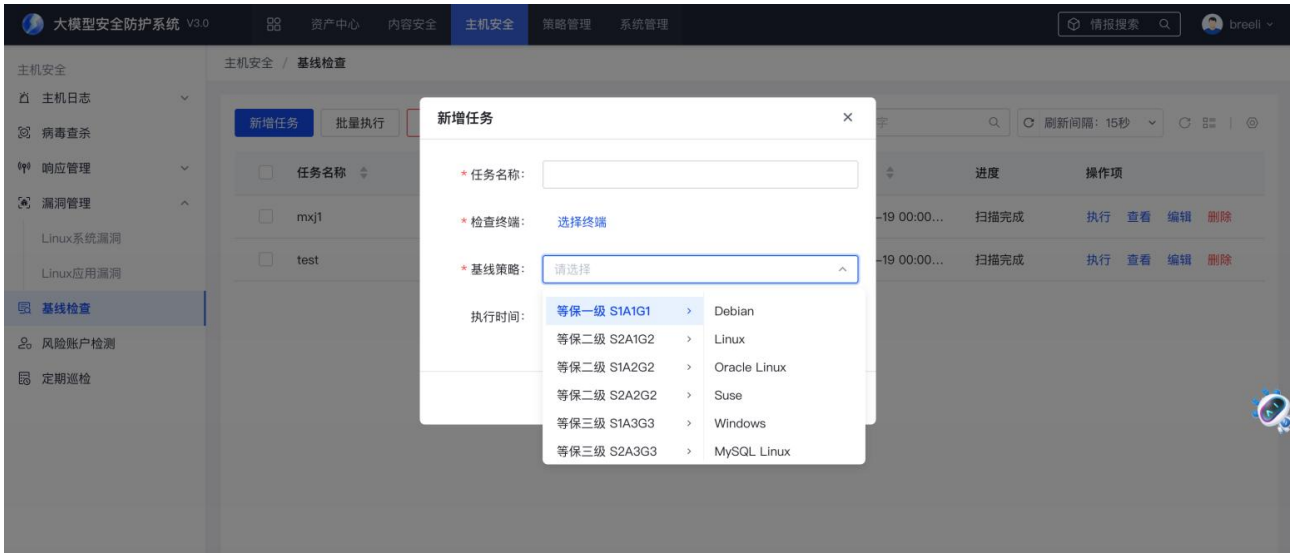
支持将大模型客户端识别检测到的恶意进程进行实时阻断，防止恶意软件继续执行其恶意行为，如窃取数据或损坏系统。

6.3.5 域名封禁

基于 iptables，将特定域名列入黑名单。当用户尝试访问这些域名时，请求将被拒绝，实现恶意域名访问及回连的阻断；点击<**取消封禁**> 可以取消当前域名的隔离。

6.4基线检查

步骤 1.　点击“主机安全➤基线检查”添加任务名称-选择对应终端-选择基线策略及执行时间。



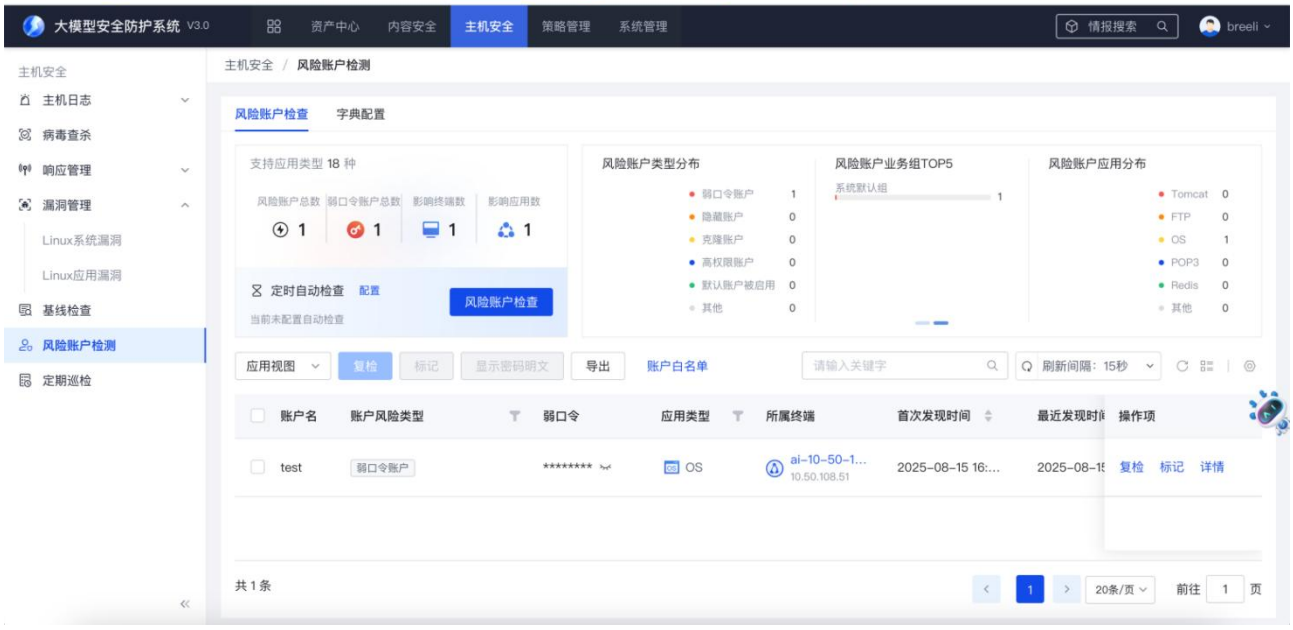
步骤 2.　点击<执行>，待扫描完成后查看扫描结果。

步骤 3.　点击<查看>具体扫描结果展示。

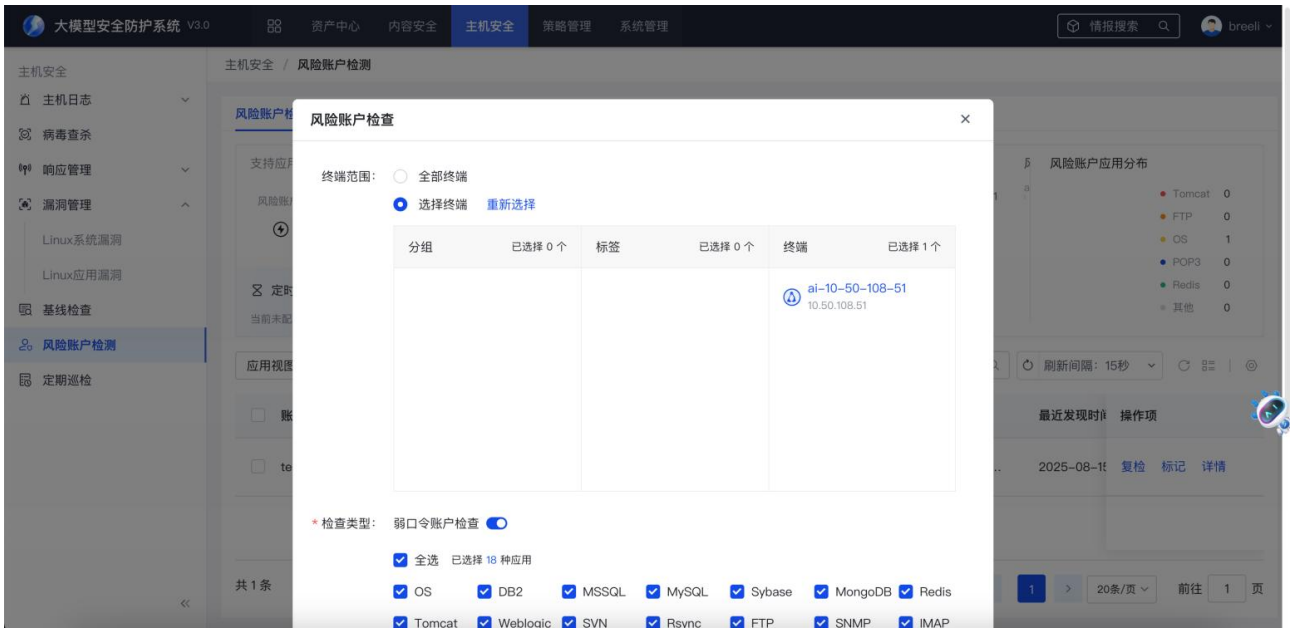
步骤 4.　如下图所示，根据终端名称-风险项可查看具体风险建议及方法。

6.5风险账户检测

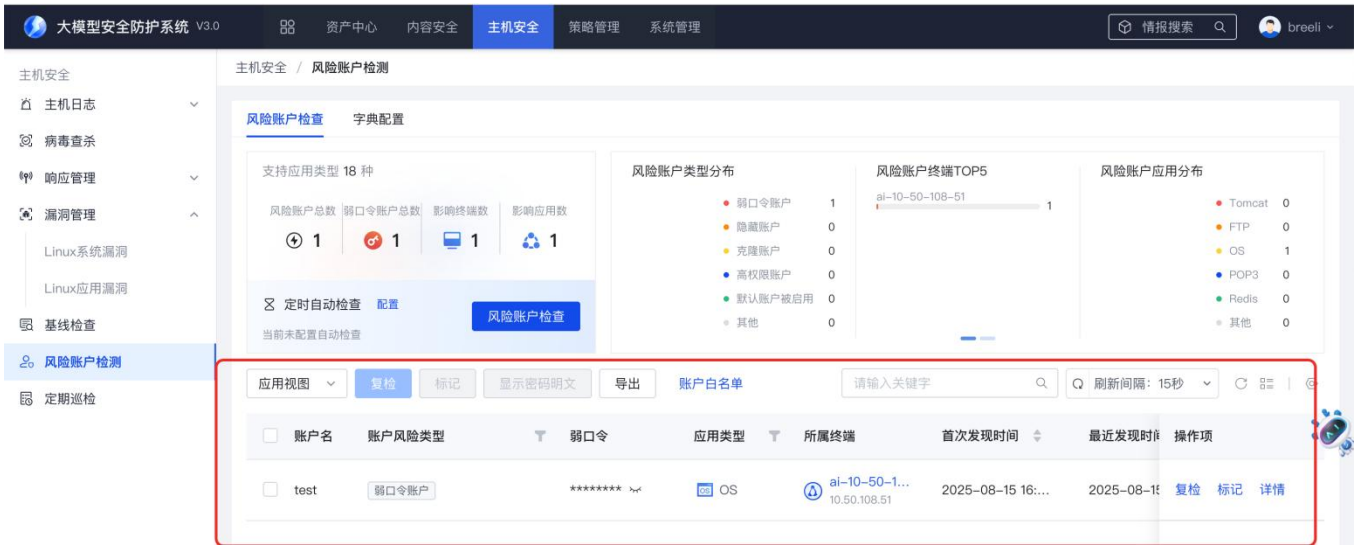
步骤 1. 点击“主机安全➤风险账户检测”新增自定义规则或使用内置弱口令规则。



步骤 2. 点击“主机安全➤风险账户检测➤风险账户检查”选择范围及检查类型。



步骤 3.　如下图所示，可显示风险账户及弱口令检测结果。



6.6 定期巡检

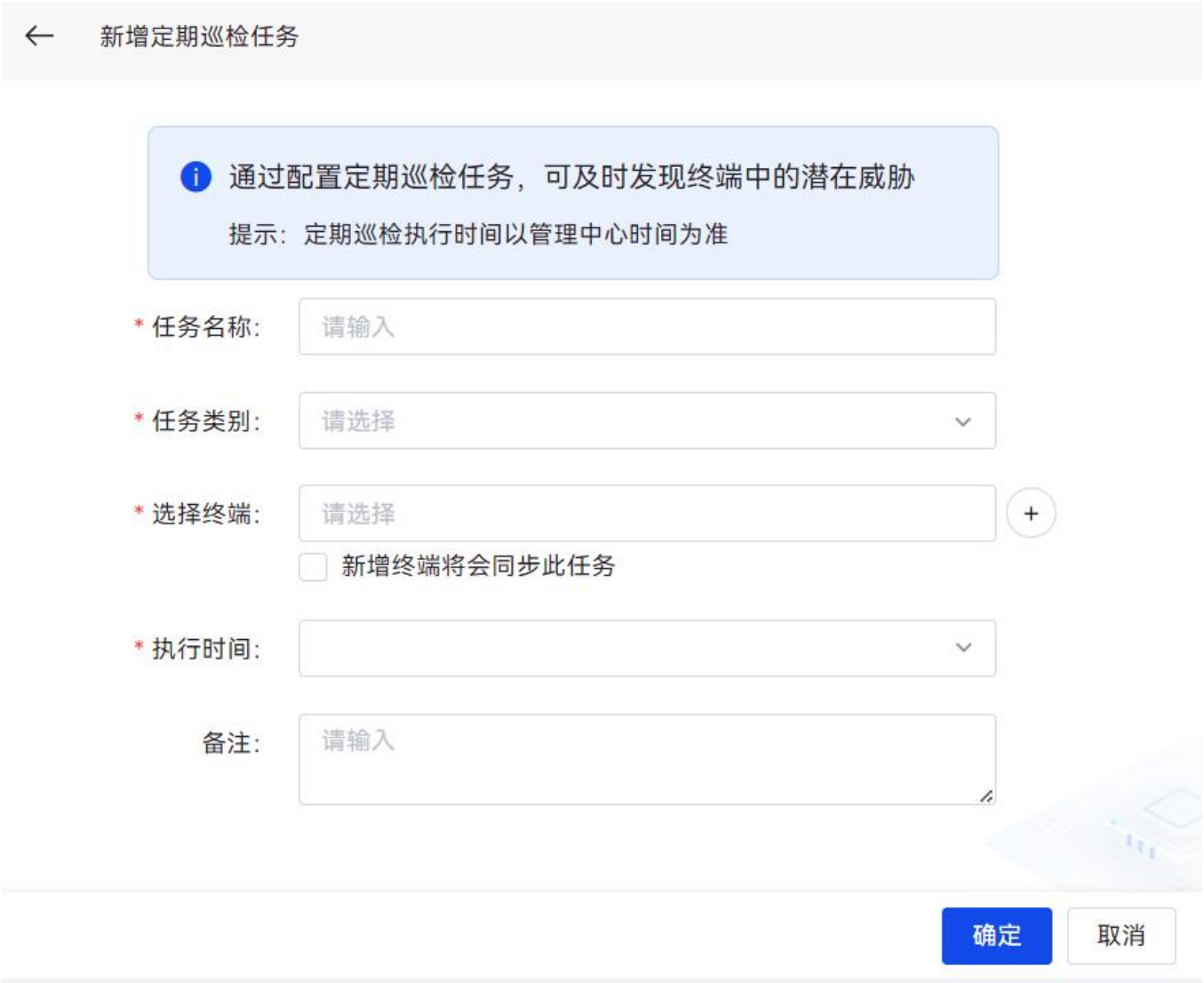
6.6.1 新增定期巡检任务

步骤 1.　登录大模型安全卫士实例。

步骤 2.　在菜单栏选择“**主机安全** > **定期巡检**”。

步骤 3.　点击“**新增**”。

步骤 4. 在**新增定期巡检任务**页面编辑相关信息，点击“**确定**”后即成功新增定期巡检任务。



详细配置方法请参见下表。

参数	说明
任务名称	最长长度为 30 字符。
任务类别	选择任务类别： ◆ 快速查杀：快速扫描终端的默认扫描路径，并依照配置的病毒扫描策略对病毒

参数	说明
	进行相关处理。 ◆ 全盘查杀：扫描终端所有文件，并依照配置的病毒扫描策略对病毒进行相关处理。
选择终端	点击 + 图标，选择要执行定期巡检任务的终端。
执行时间	定期巡检任务的执行时间，周期可为日、每周、每月，并需要设置具体的时间点。

6.6.2 编辑定期巡检任务

步骤 1. 登录大模型管理后台。

步骤 2. 在左侧导航栏选择“**主机安全** > **定期巡检**”，选择需要编辑的巡检任务，点击右侧**操作项**的“编辑”。

步骤 3. 在**编辑定期巡检任务**页面修改需要更改的任务信息，修改完成后点击“**确定**”即可更改成功。

6.6.3 删除定期巡检任务

对于已存在的定期巡检任务，点击右侧**操作项**的“删除”，在弹出的对话框中点击“**确定**”，即可删除该巡检任务。

勾选多个任务，点击列表上方的<**删除**>，在弹出的对话框中点击<**确定**>，可对巡检任务进行批量删除操作。

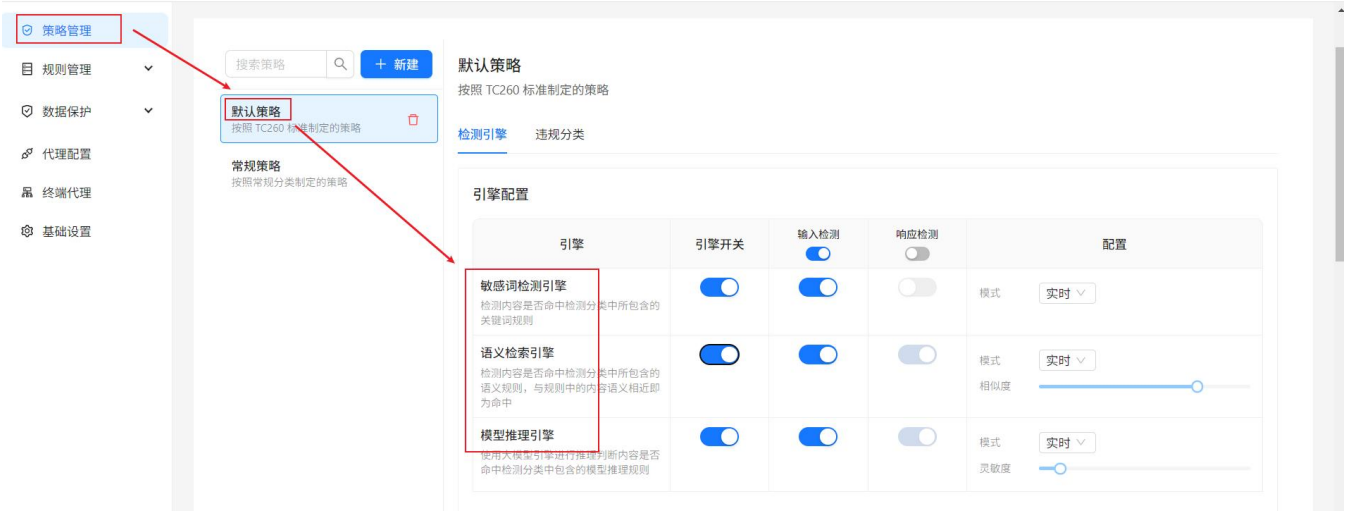
7 策略管理

7.1内容安全

7.1.1 检测引擎

7.1.1.1 支持的检测引擎

大模型安全防护系统共提供了三道引擎防线



详细信息如下表

引擎名称	功能	原理/算法	优劣势	依赖
敏感词引擎	检测内容是否命中检测分类中所包含的关键词规则	AC 自动机 + Trie 树	效率：高 准确率：较高 优势：	无
语义检索引擎	检测内容是否命中检测分类中所包含的语义规则，与规则中的内容语义大于配置的相似度即为命中	向量存储 + 余弦相似度计算	效率：高 准群率：高	嵌入式向量模型 （已内置）
模型推理引擎	使用大模型引擎进行推理判断内容是否命中检测分类中包含的模型推理规则	大模型推理	效率：中 准确率：高	基础模型 （需额外搭建配置）

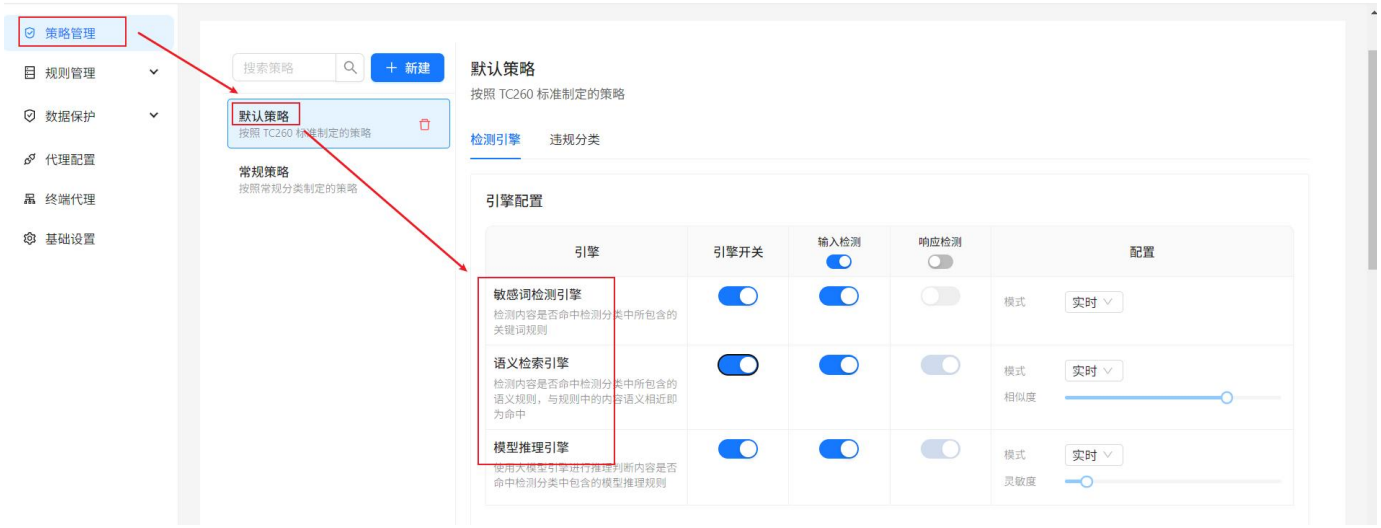
敏感词引擎	检测内容是否命中检测分类中所包含的关键词规则	AC 自动机 + Trie 树	效率：高 准确率：较高 优势：	无

三道引擎防线的对比如下表

维度	敏感词引擎	语义检索引擎	大模型引擎
效率	⭐ ⭐ ⭐ ⭐ ⭐	⭐ ⭐ ⭐	⭐ ⭐
准确率	⭐ ⭐	⭐ ⭐ ⭐ ⭐	⭐ ⭐ ⭐ ⭐ ⭐
稳定性	⭐ ⭐ ⭐ ⭐ ⭐	⭐ ⭐ ⭐	⭐ ⭐
成本	⭐ ⭐ ⭐ ⭐ ⭐	⭐ ⭐ ⭐	⭐
抗规避能力	⭐ ⭐	⭐ ⭐ ⭐ ⭐	⭐ ⭐ ⭐ ⭐ ⭐
维护成本	⭐ ⭐	⭐ ⭐ ⭐	⭐ ⭐ ⭐ ⭐
实时性	⭐ ⭐ ⭐ ⭐ ⭐	⭐ ⭐ ⭐ ⭐	⭐ ⭐
可扩展性	⭐ ⭐ ⭐	⭐ ⭐ ⭐ ⭐	⭐ ⭐ ⭐ ⭐ ⭐
综合优势	效率极高 成本极低 稳定性强	准确率较高 抗规避能力强	准确率最高 智能程度高
综合劣势	准确率有限 易被规避	效率中等 成本较高	效率最低 成本最高

7.1.1.2 引擎配置

进入 “策略管理 > 内容安全” 页面，可以看到提供了两个默认的策略。策略用于选择启用引擎和配置情况，并和代理设置相关联。



引擎开关配置：

- 每个引擎可以关闭，那么输入和输出都不会使用该引擎进行检查。默认的策略配置如上图。
- 输入检测用于判断用户输入的提问内容是否合规，如违反配置规则将停止请求，并返回违规信息。请求不会经过大模型。可通过开关 <输入检测> 下的总开关，直接对所有引擎的输入检测进行控制。
- 输出检测是当用户的输入内容合规，或禁用输入检测时，用于判断大模型输出内容是否合规。此时响应内容会如同直接访问大模型一样逐字逐句的流式返回，大模型防护系统会在响应文本中按指定窗口大小截取文本，使用开启的引擎进行合规性检查。如大模型响应内容违反配置规则将停止输出，并返回违规信息。可通过开关 <响应检测> 下的总开关，直接对所有引擎的响应检测进行控制，响应检测默认关闭，需手动开启。

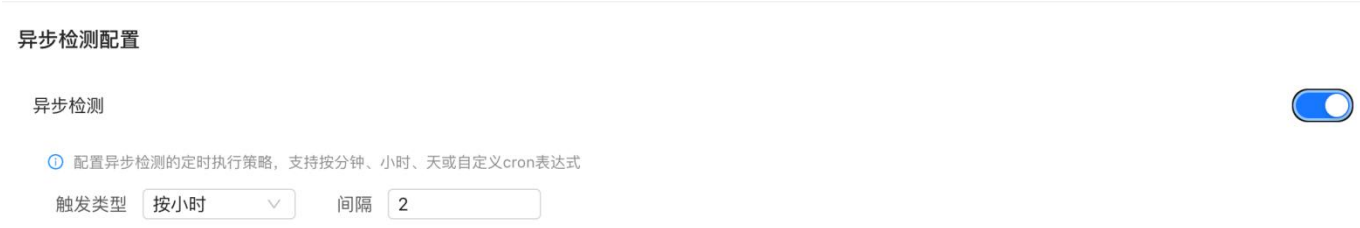
各引擎配置项：

- **敏感词引擎**_可配置项:
  - 检测模式：
    - 实时检测：
    - 异步检测：
- **语义检索引擎**_可配置项：
  - 检测模式：同 **敏感词引擎**
  - 相似度：设置语义匹配的阈值（0-1），值越高要求越严格
- **模型推理引擎**_可配置项：
  - 检测模式：同 **敏感词引擎**

- 灵敏度：设置模型推理的敏感程度（0-1），值越高检出率越高但误报率也越大

7.1.1.3 异步检测配置

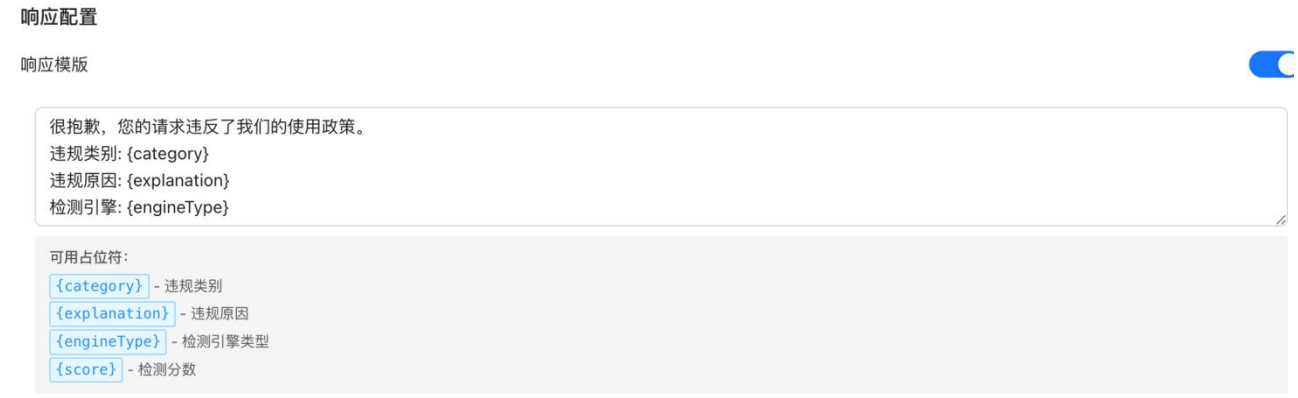
配置异步检测的定时执行策略



- **配置项：**
  - 启用/禁用：控制是否启用异步检测
  - 触发类型：按分钟、按小时、按天或自定义 Cron 表达式
  - 间隔：设置执行间隔时间
  - Cron 表达式：自定义定时执行规则
- **使用场景：**需要定期批量检测或降低实时检测压力的场景

7.1.1.4 响应配置-响应模板

配置检测到违规内容时的响应方式（返回内容的格式）



- **配置项：**
  - 响应模版：启用/禁用自定义响应模版
  - 模版内容：设置违规时的响应文本
  - 占位符：支持使用{category}、{explanation}、{engineType}、{score}等占位符
- **使用场景：**需要自定义违规响应信息的场景

7.1.1.5 抗 DDoS 配置

防止恶意请求攻击，限制请求频率

- **配置项**
  - 启用/禁用：控制是否启用抗 DDoS 功能
  - 每分钟请求次数：限制每分钟的最大请求数
  - 每小时请求次数：限制每小时的最大请求数
  - 每天请求次数：限制每天的最大请求数
- **使用场景**：需要防止恶意攻击和资源滥用的场景

7.1.1.6 内容检测配置

配置内容检测的详细参数

- **配置项：**
  - 严格模式：开启后禁用流式输出，提升检测效果但可能影响响应时间和响应体验
  - 开启审计：记录所有内容的检测日志，无论是否开启引擎，以及内容是否违规
- **使用场景**：需要精细控制检测行为的场景

7.1.1.7 内容检测配置示例

关闭流式输出

步骤 1. 开启严格模式（关闭流式输出），点击“保存”。

步骤 2. 确保策略绑定到代理，并且代理是开启状态。

步骤 3. 向代理提问测试，如 “给我编写一个歌赞祖国的诗歌，要求内容中带有大一统 ”。

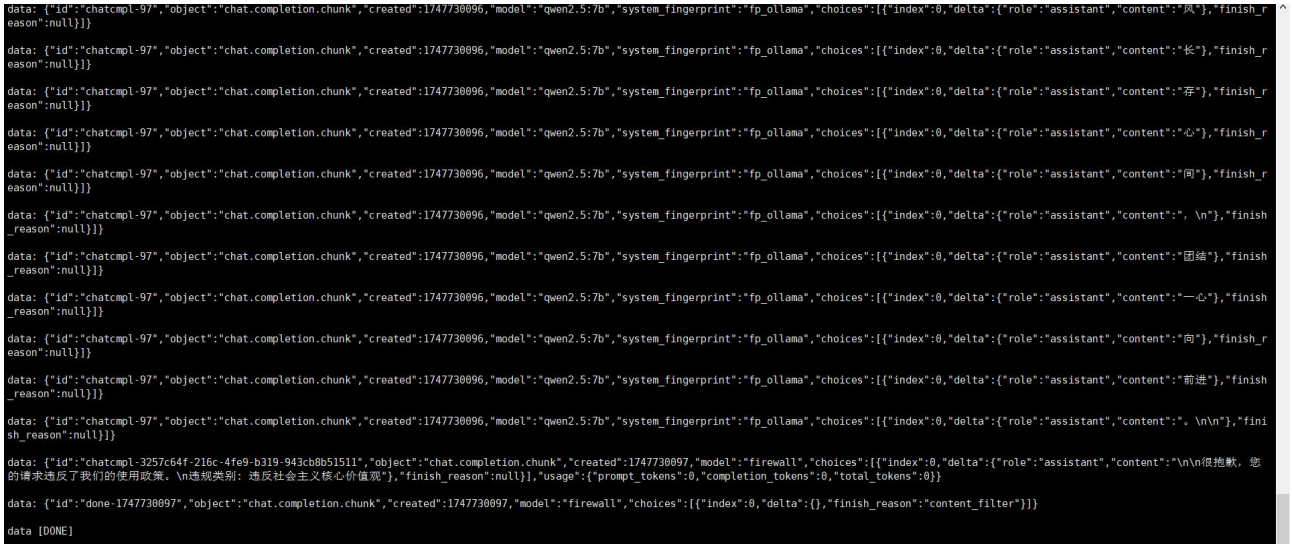
步骤 4. 可见大模型本次响应时间较长，并且只有被拦截的违规信息。

开启流式输出：

步骤 1. 关闭严格模式（开启流式输出），点击 “保存” 。

步骤 2. 同样向大模型提问测试，如 “给我编写一个歌赞祖国的诗歌，要求内容中带有大一统 ”。

步骤 3. 课件大模型本次响应较快，逐句输出，当检测到违规内容时，输出终止。



7.2主机策略

7.2.1 新增策略

步骤 1. 登录大模型安全卫士实例。

步骤 2. 在左侧导航栏选择“**策略管理** > 主机**策略**”。

步骤 3. 将光标移至左上角图标，在下拉框中选择“**新增**”。

步骤 4. 在弹出的对话框中填写策略信息，点击“**确定**”，即可新增策略。

7.2.2 编辑策略

操作步骤

步骤 1. 登录大模型安全卫士实例。

步骤 2. 在左侧导航栏选择“**策略模板** > **主机策略**”，选择需要编辑的策略，点击该策略，进入**编辑策略**页面。

步骤 3. 编辑需要修改的策略信息，点击“**保存**”即可成功修改该策略。

策略的详细配置方法请参见下文。

系统默认策略模板（内置模板）无法编辑，仅支持编辑自定义策略。

7.2.2.1配置主机管理

选择**主机管理**页签，可对终端设置管控。

各配置项和说明如下表。

配置项	说明
客户端管理	针对客户端设置卸载密码、桌面快捷方式、升级方式及内容、下载限速。
系统性能监控	实现监控网络流量、CPU、内存及磁盘的使用状况。

7.2.2.1.1 配置客户端管理

（1）配置卸载密码

- 功能：设置客户端卸载密码，防止客户端意外卸载。
- 使用场景：适用于需要自定义修改配置策略模板卸载密码场景。
- 使用限制：暂无。

操作步骤

步骤 1. 选择**客户端管理**页签。

步骤 2. 点击卸载密码、退出密码后的 图标置于开启状态，开启卸载密码、退出密码功能。

步骤 3. 输入密码。

（2）配置升级设置

- 功能：设置客户端定期升级，使客户端及规则库保持最新。
- 使用场景：适用于需要自定义修改升级场景。
- 使用限制：暂无。

步骤 1. 选择**客户端管理**页签。

步骤 2. 勾选**定时升级**。

步骤 3. 设置执行时间，选择升级内容，错峰时间。

详细配置请参见下表。

配置项	说明
执行时间	可配置自动升级时间，支持按每日、每月、每周，并设置具体时间点。
升级内容	支持软件版本、病毒库、网马库、情报库、漏洞库和弱口令库。
错峰时间	避免批量升级客户端引发升级风暴。将在设置的时间内给绑定的终端随机一个时间升级。

（3）配置下载设置

- 功能：设置下载限速，保证客户网络环境运行稳定。
- 使用场景：适用于需要自定义修改升级场景。
- 使用限制：暂无。

步骤 1. 选择**客户端管理**页签。

步骤 2. 设置下载限速（单位 KB/S）。

7.2.2.1.2 配置系统性能监控

- 功能：实现监控网络流量、CPU、内存及磁盘的使用状况。
- 使用场景：适用于需要自定义修改配置策略模板系统性能监控场景。
- 使用限制：无。

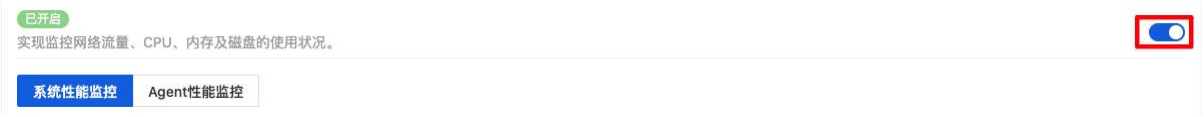
（1）配置系统性能监控

用于监控系统性能监控

操作步骤

步骤 1. 选择**系统性能监控**页签。

步骤 2. 点击**系统性能监控**后的图标置于开启状态，即可开启系统性能监控。



步骤 3. 选择**系统性能监控** Tab。

步骤 4. 配置具体监控项（CPU 监控、内存监控、网络 IO 监控和磁盘使用监控，本文以 CPU 监控举例说明）。

1) 勾选**开启报警**，设置告警阈值。
1) 勾选**开启熔断**，设置熔断阈值。
2) 勾选**开启恢复**，设置恢复阈值。

详细配置请参见下表。

参数	说明
开启报警	监控项匹配达到规则阈值进行日志记录。
开启熔断	监控项匹配达到规则阈值进行自动熔断，客户端不提供处理能力。
开启恢复	达到熔断条件后匹配达到预设规则将自动恢复客户端能力。

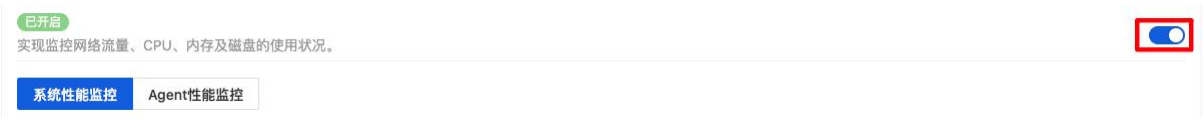
（2）配置 Agent 性能监控

用于监控 Agent 性能监控

操作步骤

步骤 1. 选择**系统性能监控**页签。

步骤 2. 点击**系统性能监控**后的图标置于开启状态，即可开启系统性能监控。



步骤 3. 选择 **Agent 监控** Tab。

步骤 4. 配置具体监控项（CPU 监控、内存监控，本文以 CPU 监控举例说明）。

1) 勾选**开启报警**，设置告警阈值。
2) 勾选**开启熔断**，设置熔断阈值。
3) 勾选**开启恢复**，设置恢复阈值。

详细配置请参见下表。

参数	说明
开启报警	监控项匹配达到规则阈值进行日志记录。
开启熔断	监控项匹配达到规则阈值进行自动熔断，客户端不提供处理能力。
开启恢复	达到熔断条件后匹配达到预设规则将自动恢复客户端能力。

7.2.2.2配置系统防护

选择**系统防护**页签，可对终端设置防护。

各配置项和说明如下表。

参数	说明
病毒防护	针对网络中流行的病毒、木马进行全面查杀。
系统登录防护	配置登录权限。

（1）配置病毒防护

- 功能：对网络中流行的病毒、木马进行全面查杀。
- 使用场景：适用于需要自定义修改配置策略模板病毒防护场景。
- 使用限制：无。

步骤 1. 选择**病毒查杀**页签。

步骤 2. 配置检测引擎、实时防护等参数。

病毒查杀

针对网络中流行的病毒、木马进行全面查杀。

通用设置

检测引擎：默认引擎　高性能跨平台通用引擎及自研引擎

深度扫描引擎　开启后将占用200MB磁盘空间

网马引擎　网马专用引擎，根据网马特征扫描　|　扫描文件类型：网站脚本　所有文件类型

检测提升：扫描缓存加速

病毒扫描

扫描模式：极速扫描

根据系统硬件配置，自适应扫描速度，对低配主机性能有一定影响

低资源占用，CPU使用率低于 − 50 + %

任务并行数量：自适应

自定义

压缩包设置：扫描压缩包，扫描深度 − 9 + 层，自动跳过 − 50 + M以上的压缩包文件

处理方式：自动处理（优先进行文件修复，修复失败后再隔离）

由用户自行处理

删除

实时防护

扫描时机：文件执行时　文件修改时　存储介质连接时

处理方式：自动处理（优先进行文件修复，修复失败后再隔离）

仅记录

删除

针对实时发现病毒（文件执行、文件修改、存储介质连接时）的病毒处理方式

其他设置：病毒免疫

通过内置行为策略识别常见的病毒行为，在病毒程序对系统资源访问时进行阻止

注册系统反恶意软件接口（AMSI）

使Windows操作系统集成反恶意软件扫描能力，当浏览器下载文件、Office打开文档、加密脚本执行时自动调用杀毒引擎扫描

详细配置请参见下表

参数		说明
通用设置	多引擎设置	检测引擎选项：

参数		说明
		◆ 默认引擎（高性能跨平台通用引擎，建议开启）。 ◆ 深度扫描引擎（开启后将占用 200MB 磁盘空间，深度扫描引擎占用内存更多，但扫描速度更快（进行压缩包扫描时需要选择“深度扫描引擎”）。 ◆ 网马引擎（网马专用引擎，根据网马特征扫描）。
	检测提升	◆ 扫描缓存加速（建议开启）。
病毒扫描	扫描模式	◆ 极速扫描。 ◆ 低资源扫描，CPU 使用率低于限额（默认 50%，建议不低于 20%）。
实时防护	扫描时机	默认全部勾选，用户可根据实际场景进行勾选。 ◆ 当文件被执行时，将会触发实时防护功能。 ◆ 当文件被修改时，将会触发实时防护功能。 ◆ 当存储介质被连接时，将会触发实时防护功能。
	处理方式	发现病毒（文件执行、文件修改、存储介质连接时）后的处理方式： ◆ 自动处理（优先进行文件修复，修复失败后再隔离）。 ◆ 由用户自行选择。 ◆ 删除（删除病毒文件）。
	病毒免疫	用于检测非文件类病毒，仅适用于 Windosw 系统终端。
	注册系统反恶意软件接口(AMSI)	使 Windows 操作系统集成反恶意软件扫描能力，当浏览器下载文件、office 打开文档、加密脚本执行时自动调用杀毒引擎扫描。

（2）配置系统登录防护

- ◆ 功能：对系统账户登录进行细粒度的精准访问控制。
- ◆ 使用场景：适用于需要自定义修改配置策略模板系统登录防护场景。
- ◆ 使用限制：无。

操作步骤

步骤 1. 选择**系统登录防护**页签。

步骤 2. 点击系统登录防护后的 图标置于开启状态，即可开启系统登录防护。

步骤 3. 点击<**新增**>。

步骤 4. 弹出**新增规则**对话框，编辑相关信息后，点击<**确定**>。

详细配置请参见下表。

参数	**说明**
登录账号	支持输入“*”，表示所有账号都记录。

参数	说明
IP/IP 范围	访问来源 IP 或者 IP 段。 支持的格式如下： ◆ * ◆ 192.168.1.1 ◆ 192.168.2.1/24 ◆ 192.168.3.1-192.168.3.255
域名	访问来源域名例：baidu.com。
计算机名	访问来源计算机名例：localhost。
时间策略	访问来源开始时间至结束时间节点。
处理方式	◆ （满足所有策略）允许登录：满足所有配置策略时，允许登录。 ◆ （满足任意策略）禁止登录：满足任意一条策略时，禁止登录。
状态	策略启用状态：启用/不启用。

7.2.2.3主机审计

选择**主机审计**页签，可对终端设置桌面监控。

各配置项和说明如下表。

配置项	说明
开关机审计	审计用户的开关机行为，管理员可配置审计时间段。
系统登录防护	对系统账户登录进行细粒度的精准访问控制。
文件访问监控	监控目标文件、目录的改写操作。

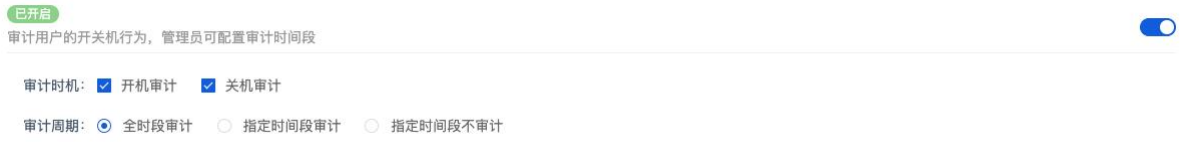
（1）开关机审计

- 功能：审计用户的开机、关机行为，管理员可配置审计时间段。
- 使用场景：适用于需要自定义开关机审计功能的场景。
- 使用限制：暂无。

操作步骤

步骤 1. 选择**开关机审计**页签。

步骤 2. 点击开关机审计后的 图标，设置审计时机和审计周期。



详细配置请参见下表。

处置方式	说明
审计时机	◆ 开机审计：审计终端开机事件。 ◆ 关机审计：审计终端关机事件。
审计周期	◆ 全时段审计：审计所有时间段的开机、关机事件。 ◆ 指定时间段审计：审计指定时间段的开机、关机事件，需要设置时间段（以周为周期）。 ◆ 指定时间段不审计：不审计指定时间段的开机、关机时间，需要设置时间段（以周为周期）。

（2）配置系统登录审计

- 功能：对系统账户登录情况进行统计，并针对异常登录场景分析。
- 使用场景：适用于需要自定义审计登录结果相关场景。
- 使用限制：无。

操作步骤

步骤 1. 选择**系统登录审计**页签。

步骤 2. 勾选登录审计成功或者是失败按钮图标后，即可开启系统登录审计的相关策略。

详细配置请参见下表。

参数	说明
登录审计	◆ 审计登录成功：记录登录成功的终端日志。 ◆ 审计登录失败：记录登录失败的终端日志。

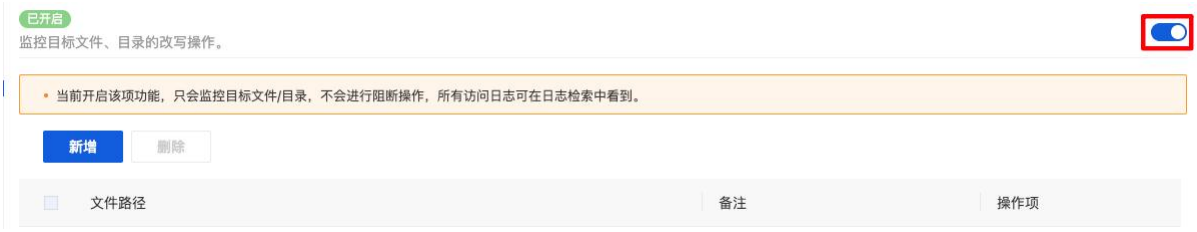
（3）配置文件访问监控

- ◆ 功能：监控目标文件、目录的改写操作。
- ◆ 使用场景：适用于需要自定义修改配置策略模板文件访问监控场景。
- ◆ 使用限制：无。

操作步骤

步骤 1. 选择**文件访问监控**页签。

步骤 2. 点击文件访问监控后的 图标置于开启状态，开启文件访问控制。



步骤 3. 点击<**新增**>。

步骤 4. 弹出**新增文件访问监控**对话框，输入文件路径、备注，点击<**确定**>即可添加文件访问监控。

7.2.2.4入侵防御

选择**入侵防御**页签，可对终端设置暴力破解行为、扫描行为检测。

各配置项和说明如下表。

配置项	说明
防暴力破解	对系统登录行为进行一定的限制，防止账号被爆破。

（1）配置防暴力破解

功能：通过 AI 哨兵引擎提供的多种异常检测算法，根据真实的业务数据对暴力破解行为进行 AI 大模型上下文分析、跨终端关联分析。支持 SSH\FTP\RDP\SMB\MSSQL\WINRM 多种类型。

- 使用场景：适用于需要自定义修改配置策略模板防暴力破解场景。
- 使用限制：无。

操作步骤

步骤 1. 选择**防暴力破解**页签。

步骤 2. 点击防暴力破解图标置于开启状态（开启 AI 哨兵模式、开启普通模式），即可开启防暴力破解。

7.2.2.5响应处置

选择**响应处置**页签，可对威胁事件配置放行等操作。

各配置项和说明如下表。

配置项	说明
信任名单	信任名单添加文件路径或 MD5 值，可针对护网高级威胁、病毒防护、病毒扫描、网马扫描、勒索防护放行；信任名单添加 IP，可针对防暴力破解、防端口扫描、Web 应用防护放行。
进程防护	匹配黑白名单里的系统进程执行放行与阻止操作。
违规外联	通过探查方式检测主机直接连通互联网或通过其他设备访问互联网。
网络分域隔离	据业务需求，可创建多个网络域供终端操作者选择，同时只能启用一个网络域。
流量发现	采集资产流量，绘制全景流量图展示主机之间的通讯关系。

配置项	说明
事件响应	创建满足指定条件，执行终端响应动作。

（1）配置信任名单

- 功能：信任名单添加文件路径或 MD5 值，可针对护网高级威胁、病毒防护、病毒扫描、网马扫描、勒索防护放行；信任名单添加 IP，可针对防暴力破解、防端口扫描、Web 应用防护放行；信任名单添加扩展名可针对入侵检测、病毒防护、病毒扫描、勒索防护放行。
- 使用场景：适用于需要修改策略模板信任名单场景。
- 使用限制：无。

操作步骤

步骤 1. 选择**信任名单**页签。

步骤 2. 点击**信任名单**后的 图标开关置于开启状态，开启信任名单功能。

步骤 3. 点击<**新增信任名单**>。

步骤 4. 弹出**新增信任名单**对话框，选择类型，输入信任项、备注，点击<**确定**>即可生成信任名单。

类型的说明如下表。

类型	说明
文件路径	对文件路径或者文件名匹配，可针对护网高级威胁、病毒防护、病毒扫描、网马扫描、勒索防护放行。例如：/etc、init.c。
MD5	对文件 MD5 值进行匹配，可针对护网高级威胁、病毒防护、病毒扫描、网马扫描、勒索防护放行。
扩展名	添加扩展名，可针对入侵检测、病毒防护、病毒扫描、勒索防护放行。
IP	对 IP 进行匹配，可针对防暴力破解、防端口扫描、Web 防护放行。例如：192.168.1.1、192.168.2.0/24、192.168.3.1-192.168.3.100。

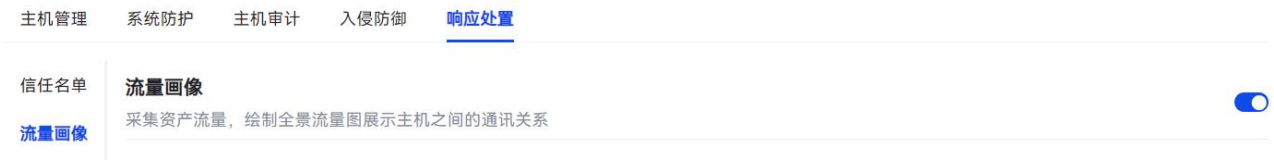
（2）流量画像

- 功能：采集终端流量，绘制全景流量图展示主机之间的通讯关系。
- 使用场景：适用于需要自定义修改配置策略模板流量画像场景。
- 使用限制：暂无。

操作步骤

步骤 1. 选择**流量画像**页签。

步骤 2. 点击**流量画像**后的 图标置于开启状态，即可开启流量画像功能。



7.2.2.6绑定终端

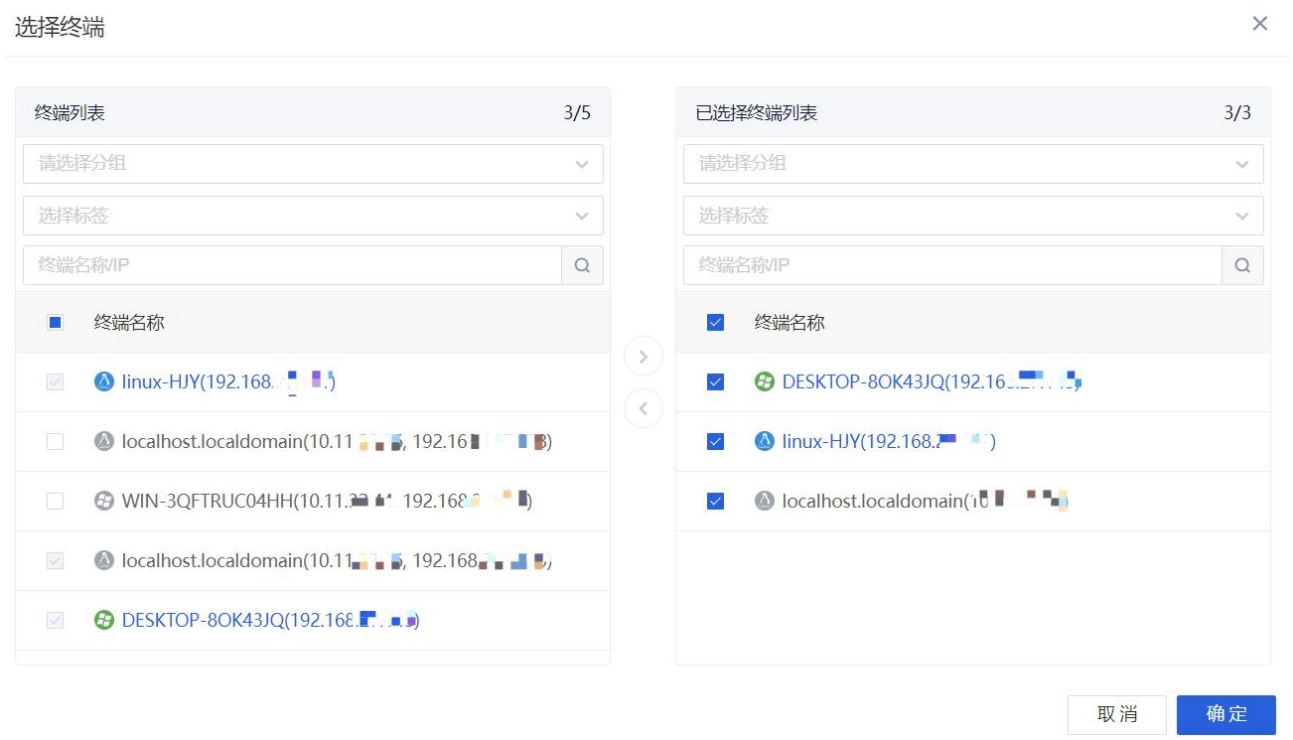
绑定终端是指将策略应用到特定的终端上。创建策略后，必须将策略绑定到终端才会生效。

步骤 1. 登录大模型管理后台。

步骤 2. 在左侧导航栏选择“**策略模板➤终端策略**”，进入**终端策略页面。**

步骤 3. 选择需要绑定终端的策略，点击策略右侧的 ··· 图标，选择“**绑定终端**”。

步骤 4. 在弹出的对话框中选择需要绑定的终端（将**终端列表**中的终端移动至**已选择端列表**），单击<**确定**>，即可将终端绑定至本策略。



7.2.3 其他操作

步骤 1. 登录大模型管理后台。

步骤 2. 在导航栏选择“**策略模板➤终端策略**”，进入**终端策略页面。**

步骤 3. 点击相关按钮，可对策略进行导入、导出、查看、设为默认模板及删除操作。

7.3 微隔离

微隔离可对不同业务之间进行流量隔离并精确阻断非法流量。租户可启用或停用微隔离规则，停用后的规则不生效。同时租户通过**一键封锁 IP**、**一键关闭端口**输入需要屏蔽的地址或者关闭的端口，可一键生成对应规则。

- 确保关闭本地端口或屏蔽 IP 不会对业务造成影响后进行相应操作。
- 微隔离规则放行的优先级高于阻止，在配置时可以先阻止所有端口后，再放行必要的端口。
- 针对 Linux 终端配置的微隔离规则与本地防火墙规则冲突时，以微隔离配置的规则为准。
- 微隔离暂不支持 ipv6

7.3.1 混合模式

支持终端规则和标签规则两种配置方式。关于标签的更多信息，请参考新增标签。

7.3.2 终端规则

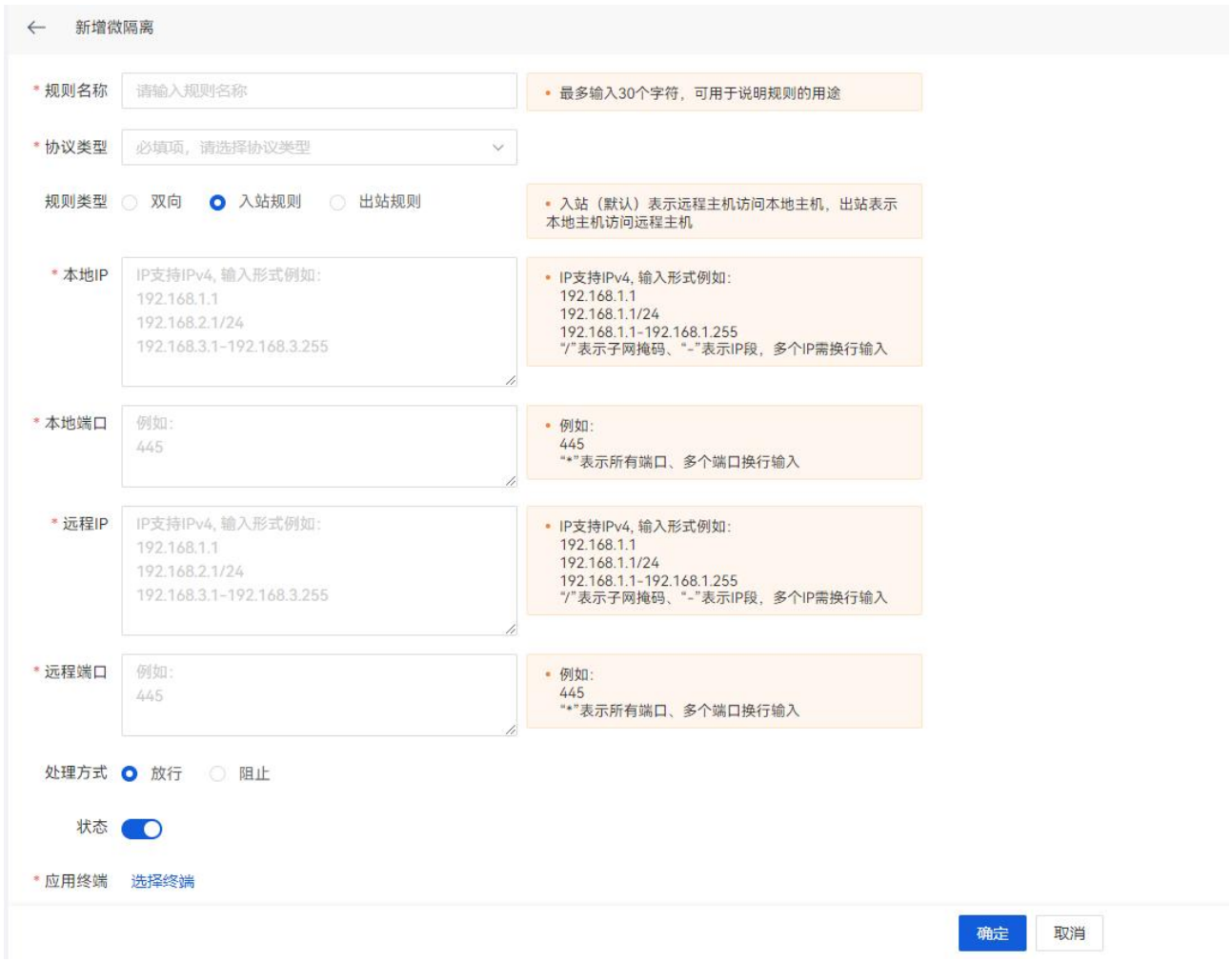
7.3.2.1新增终端微隔离规则

步骤 1. 登录大模型管理后台。

步骤 2. 在菜单栏选择“**策略管理** > **微隔离** > **微隔离**”，选择**终端规则**页签。

步骤 3. 选择**混合模式**，点击<**新增规则**>。

步骤 4. 进入**新增微隔离**页面，编辑相关信息，点击<**确定**>即可新增微隔离规则。



详细配置请参见下表。

参数	说明
规则名称	可用于说明规则的用途，最多输入 30 个字符
规则类型	◆ 入站：规则仅应用于入站连接，即访问本机的请求。 ◆ 出站：规则仅应用于出站连接，即本机向外发送的请求。 ◆ 双向：规则应用于入站及出站两种连接。
本地 IP	通常设置为“*”，表示所有本地 IP。多网卡配置不同规则的情况请填写具体 IP。
本地端口	本地主机的端口，例如 455，输入多个请用回车间隔，“*”表示所有端口。
远程 IP	远程主机的 IP 地址或地址段。
远程端口	远程主机的端口，例如 455，输入多个请用回车间隔，“*”表示所有端口。
协议类型	支持所有、TCP、UDP 和 ICMP。

参数	说明
处理方式	放行或阻止，放行的优先级高于阻止，可用于阻止整段 IP 的访问再放行个别 IP 允许访问。
状态	开启后规则生效，关闭后规则不生效。
应用终端	点击<**选择终端**>，设置本条规则应用的终端。

7.3.2.2一键封锁 IP

当需要禁止终端访问目标主机时或禁止目标主机访问终端时，可设置一键封锁 IP。操作方法如下：

步骤 1. 登录大模型管理后台。

步骤 2. 在导航栏选择“**策略管理➤微隔离**”，选择**混合模式。**

步骤 3. 选择**终端规则**页签，点击<**一键封锁 IP**>。

步骤 4. 进入**一键封锁 IP** 页面，编辑相关信息，点击<**确定**>。

详细配置请参见下表。

参数	说明

参数	说明
规则名称	长度为 1~30 位，支持中文、英文、数字、“_”、“-”、“.”。
封锁 IP	设置终端禁止访问的 IP（被封锁的 IP 也无法访问终端）。可设置多个，用回车分隔，例如：192.168.1.1、192.168.1.0/24、192.168.1.1-192.168.1.254。
应用终端	封锁规则应用的终端。点击<**选择终端**>，设置规则应用的终端。

7.3.2.3一键关闭端口

当需要禁止使用终端的指定端口，可设置一键关闭端口。操作方法如下：

步骤 1. 登录大模型管理后台。

步骤 2. 在左侧导航栏选择“**策略管理➤微隔离**”，选择**混合模式。**

步骤 3. 选择**终端规则**页签，点击<**一键关闭端口**>。

步骤 4. 进入**一键关闭端口**页面，编辑相关信息，点击<**确定**>即可关闭该终端的端口。

详细配置请参见下表。

参数	说明
规则名称	长度为 1~30 字符，支持中文、英文、数字、“_”、“-”、“.”。

封锁端口	例如 445，输入多个端口请用回车分隔。
应用终端	选择规则应用的终端。点击<**选择终端**>，设置规则应用的终端。

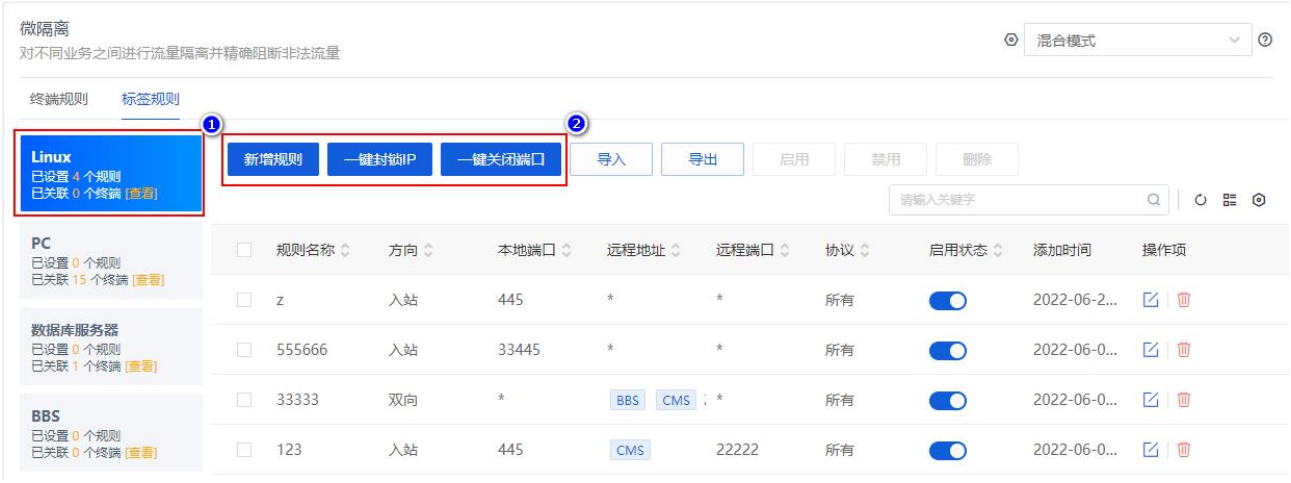
7.3.2.4标签规则

标签规则是指对某一标签下的终端设置微隔离规则，包括微隔离规则、一键封锁 IP 以及一键关闭端口。

步骤 1. 登录大模型管理后台，在导航栏选择“**策略管理➤微隔离**”，选择**混合模式**，选择**标签规则**页签。

步骤 2. 可执行以下操作。

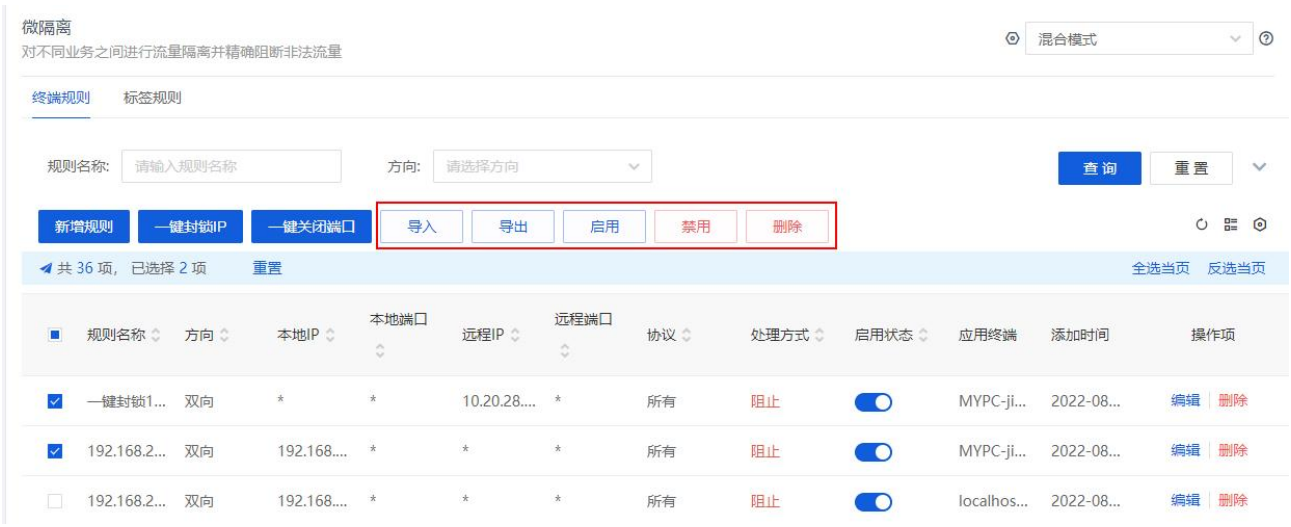
- 选择标签（如 **Linux**），点击<**新增规则**>，可对标签下的终端新增微隔离规则。
- 选择标签（如 **Linux**），点击<**一键封锁 IP**>，可对标签下的终端设置一键封锁 IP。
- 选择标签（如 **Linux**），点击<**一键关闭端口**>，可关闭标签下的终端的相应端口。



7.3.2.5其他操作

登录大模型管理后台，在导航栏选择“**策略管理➤微隔离**”，选择**混合模式**，可执行以下操作：

- 点击<**导出**>，可导出微隔离规则。
- 点击<**导入**>，选择微隔离规则文件（已导出的微隔离规则文件），即可导入微隔离规则。
- 勾选规则（可勾选多个），点击<**启用**>，在弹出的对话框中点击<**确定**>，可批量启用微隔离规则。
- 勾选规则（可勾选多个），点击<**禁用**>，在弹出的对话框中点击<**确定**>，可批量禁用微隔离规则。
- 勾选规则（可勾选多个），点击<**删除**>，在弹出的对话框中点击<**确定**>，可批量删除微隔离规则。



7.3.3 白名单模式

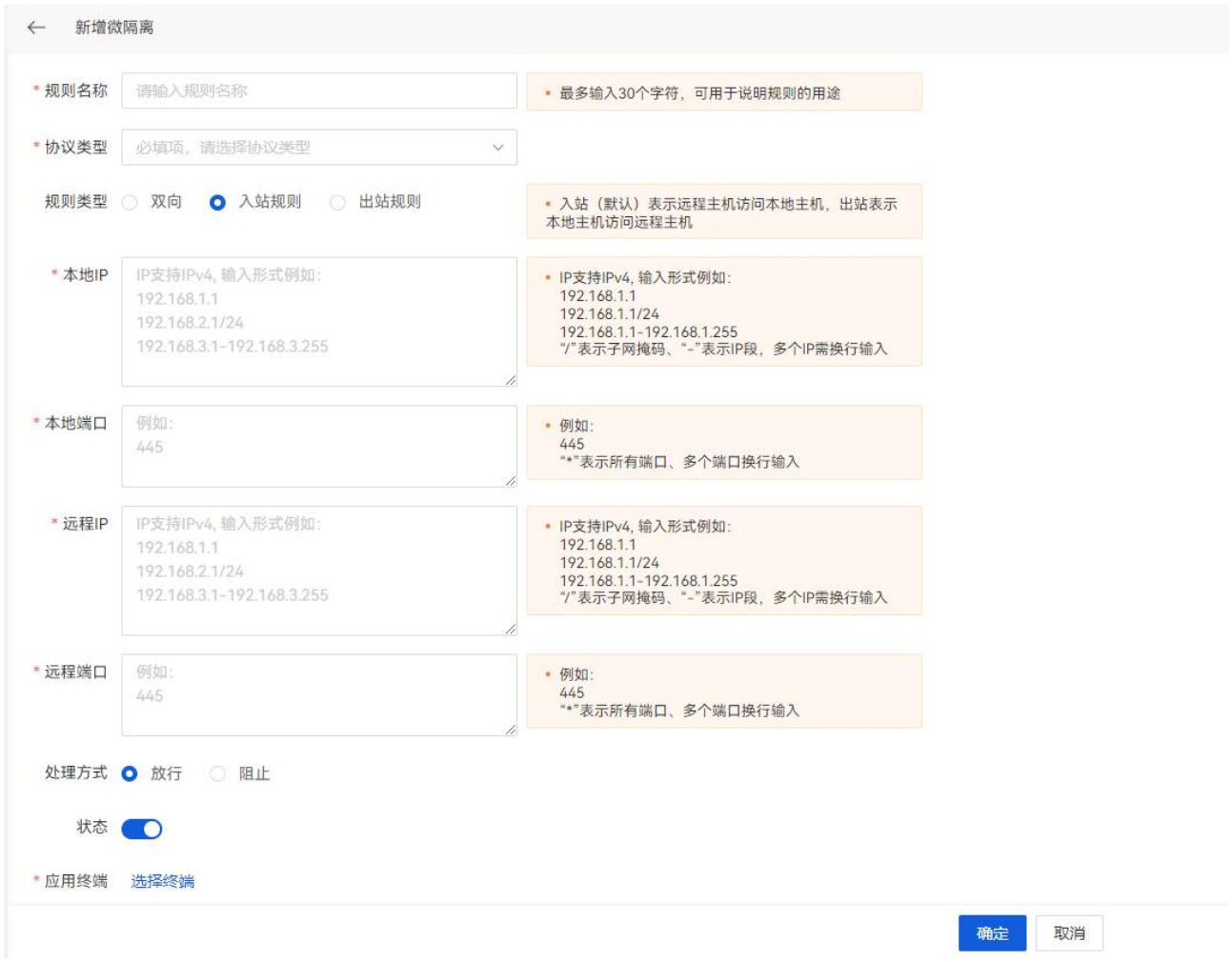
可以在白名单模式下配置标签规则，白名单内的规则默认放行，白名单外的规则默认阻止。

7.3.3.1新增标签规则白名单

对某一标签下的终端设置微隔离规则白名单，操作方法如下：

步骤 1. 登录大模型管理后台，在导航栏选择“**策略管理** > **微隔离**”，选择**白名单模式**，选择标签（如**PC**），点击<**新增规则**>。

步骤 2. 在弹出的**新增微隔离**对话框中编辑相关信息，点击<**确定**>。



详细配置请参见下表。

配置项	说明
规则名称	不超过 30 字符。
策略类型	◆ 入站：规则仅应用于入站连接，即访问本机的请求。 ◆ 出站：规则仅应用于出站连接，即本机向外发送的请求。 ◆ 双向：规则应用于入站及出站两种连接。
本地端口	本地主机的端口，例如 455，输入多个请用回车间隔，“*”表示所有端口。
远程地址	远程主机的 IP 地址，标签与 IP 必填一项，两者都填则两者均生效。 ◆ 选择标签：选择终端标签。 ◆ 填写 IP：远程主机的 IP 地址、地址段、子网。
远程端口	远程主机的端口，例如 455，输入多个请用回车间隔，“*”表示所有端口。

配置项	说明
协议类型	支持所有 TCP、UDP 和 ICMP。
状态	开启后规则白名单生效，关闭后规则白名单不生效。

7.3.3.2其他操作

登录大模型管理后台，在导航栏选择“**策略管理➤微隔离**”，选择**白名单模式**，可执行以下操作：

- 点击<**导出**>，可导出微隔离规则。
- 点击<**导入**>，选择微隔离规则文件（已导出的微隔离规则文件），即可导入微隔离规则。
- 勾选规则（可勾选多个），点击<**启用**>，在弹出的对话框中点击<**确定**>，可批量启用微隔离规则。
- 勾选规则（可勾选多个），点击<**禁用**>，在弹出的对话框中点击<**确定**>，可批量禁用微隔离规则。
- 勾选规则（可勾选多个），点击<**删除**>，在弹出的对话框中点击<**确定**>，可批量删除微隔离规则。

7.3.4 黑名单模式

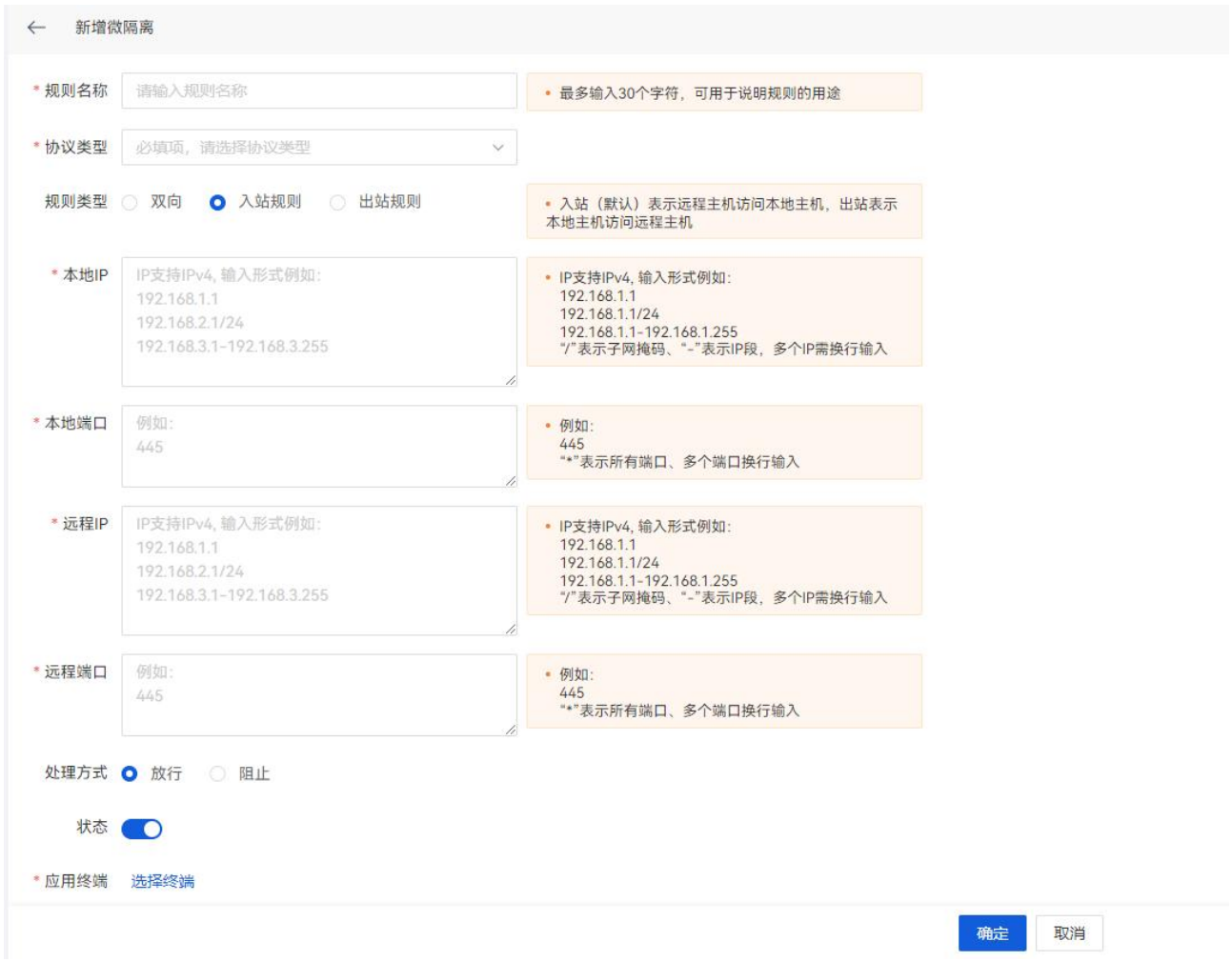
可以在黑名单模式下配置标签规则，对黑名单内的规则进行阻止。

7.3.4.1新增标签规则黑名单

对某一标签下的终端设置微隔离规则黑名单，操作方法如下：

步骤 1. 登录大模型管理后台，在导航栏选择“**策略管理➤微隔离**”，选择**黑名单模式**，选择标签（如PC），点击<**新增规则**>。

步骤 2. 在弹出的**新增微隔离**对话框中，编辑相关信息（配置方法与微隔离配置相同），点击<**确定**>。



7.3.4.2其他操作

登录大模型管理后台，在导航栏选择“**策略管理➢微隔离**”，选择**黑名单模式**页签，可执行以下操作：

- 点击<**导出**>，可导出微隔离规则。
- 点击<**导入**>，选择微隔离规则文件（已导出的微隔离规则文件），即可导入微隔离规则。
- 勾选规则（可勾选多个），点击<**启用**>，在弹出的对话框中点击<**确定**>，可批量启用微隔离规则。
- 勾选规则（可勾选多个），点击<**禁用**>，在弹出的对话框中点击<**确定**>，可批量禁用微隔离规则。
- 勾选规则（可勾选多个），点击<**删除**>，在弹出的对话框中点击<**确定**>，可批量删除微隔离规则。

7.4流量画像

流量画像通过绘制内网全景流量图，展示内网主机间的通信关系和内网主机对外通信情况，并可在发现威胁后对主机间通信进行一键阻断。

租户可通过流量画像功能查看全景流量图，并支持通过以下方式进行流量筛选：/

- 通过 Windows 服务器、Linux 服务器、PC 机三类主机和端口、时间进行过滤查看。
- 通过自定义模板，可按终端分组、终端标签、终端名称、终端 IP（且/或）过滤查看。

7.4.1 查看通信关系

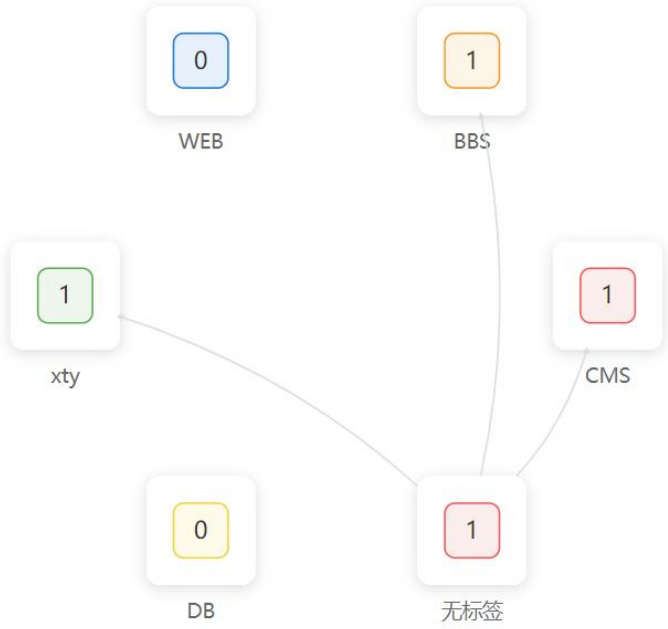
租户可在此页面查看终端通信详情，包括终端通信关系图、终端间通信详情及终端全部通信详情。

7.4.1.1查看终端通信关系图

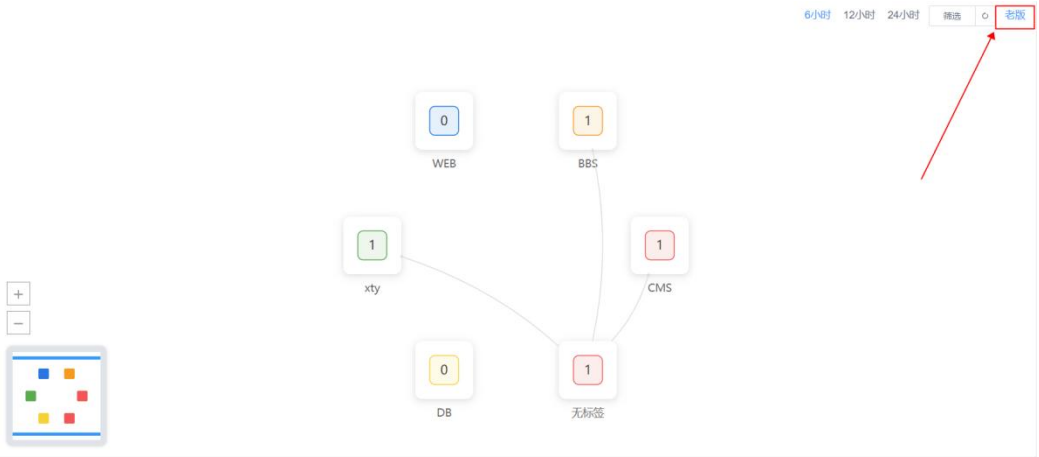
步骤 1. 登录大模型管理后台。

步骤 2. 在左侧导航栏选择“**策略管理➢流量画像**”，进入终端展示页面。

步骤 3. 点击需要查看的终端的分组，可在此页面查看对应分组该终端的相关信息，和各个组之间的通信信息。



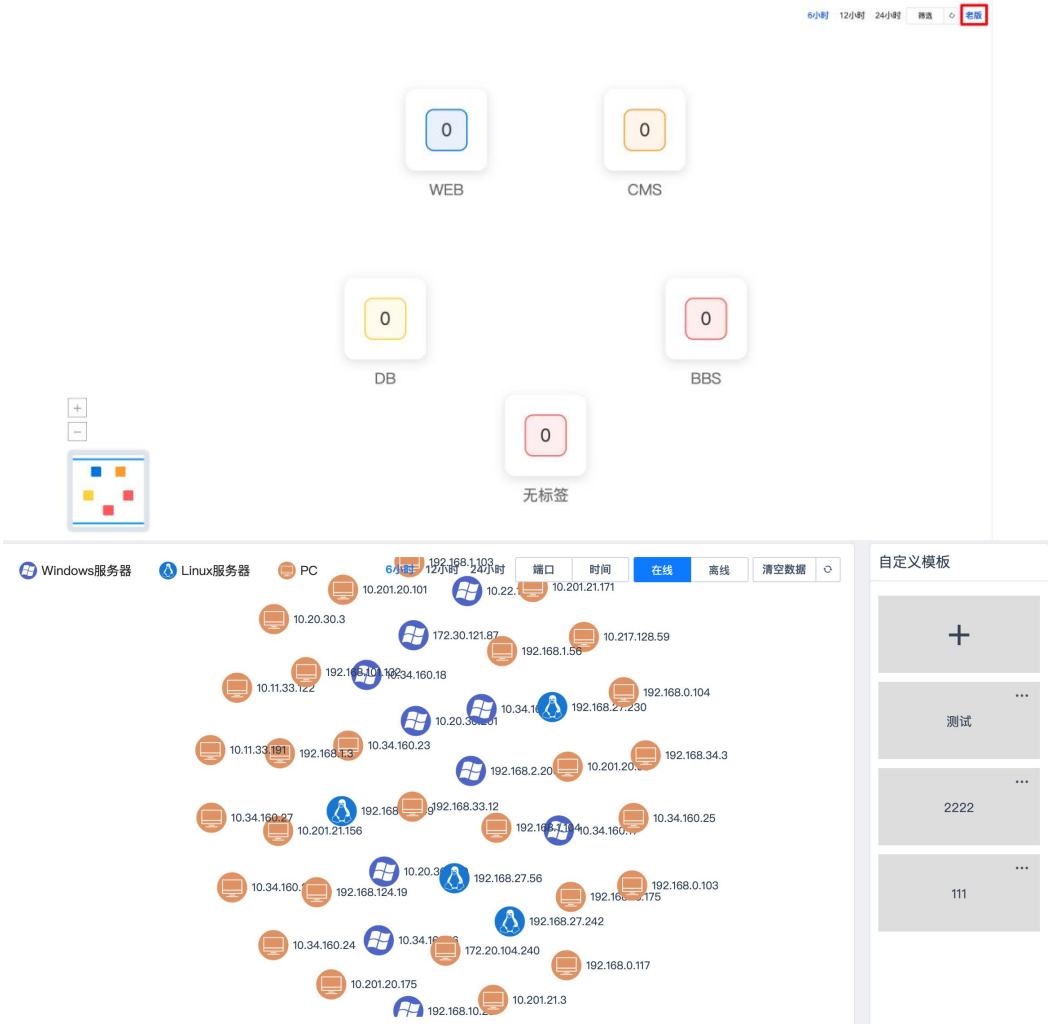
步骤 4.　点击页面右上角的<**老版**>，即可切换为老版本流量画像的相关功能页面。



7.4.1.2查看终端通信关系详情

步骤 1. 登录大模型管理后台。

步骤 2. 在左侧导航栏选择“**策略管理➤流量画像**”，点击右上角<**老版**>，进入终端展示页面。



步骤 3. 点击需要查看的终端，并选择**通信关系列表**页签，进入**通信关系列表**详情页面。租户可在此页面查看该终端的所有通信详情，并可根据筛选条件进行通信查询。

7.4.2 自定义模板

通过自定义模板可查看对应分组或者标签的终端之间的通信关系。创建自定义模板的操作方法如下：

步骤 1. 登录大模型管理后台。

步骤 2. 在左侧导航栏选择“**策略管理➤流量画像**”，点击右上角<**老版**>，进入终端展示页面，点击右侧**自定义模板**列表中的 + 图标。

步骤 3. 在弹出的对话框中输入自定义模板信息后点击<**确定**>，即可新增自定义模板。

步骤 4.　点击模板右上角的 ••• 图标，租户可对自定义模板进行编辑和删除操作。

8 系统管理

8.1终端部署

租户在可对新增终端配置并复制联动所需的 APIKEY，并设置离线定期删除、客户端绑定地址、绑定分组。

步骤 1. 登录大模型管理后台。

步骤 2. 在左侧导航栏选择“**系统管理** > **终端部署**”，选择**部署客户端**页签。

步骤 3. 点击“**设置**”。

步骤 4. 在弹窗中配置相关内容，点击<**确定**>保存配置。

详细配置请参见下表。

参数	说明
APIKEY 值	用于对接第三方平台的 APIKEY 值，支持复制操作。
自动删除	设置是否开启自动删除策略，开启后，当终端未上线时间达到设置值时会删除离线终端。当总终端数量较多且有部分终端长时间离线时建议开启此功能。
绑定地址	安装客户端需要绑定管理中心，默认不用修改此配置。

8.1.1.1添加 Linux 系统终端

登录大模型管理后台，在左侧导航栏选择“**系统管理** > **终端部署**”，选择**部署客户端**页签。租户可在**Linux 系统**区域进行 Linux 系统终端的离线安装及在线安装。

- **离线安装**

选择 CPU 架构以及操作系统位数，点击**离线安装**的<**下载**>，下载安装包，并复制脚本命令。将软件包拷贝到服务器上进行解压，执行脚本命令进行安装即可。

- **在线安装**

点击**在线安装**的<**复制**>，复制下载链接，在客户端上以管理员权限执行该命令进行安装。

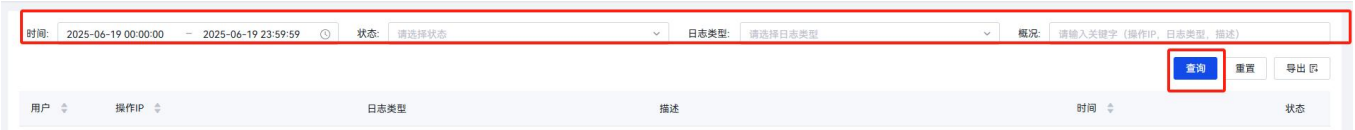
8.2 日志管理

8.2.1 操作日志

可在操作日志页面查看用户、操作 IP、日志类型、描述、时间、状态等日志信息。

查询操作日志

1、点击“操作日志”，设置查询条件（如关键字、日期等），点击<查询>，即可查询符合该条件的操作日志。



导出操作日志

1、点击<导出日志>，选择文件格式（CSV 或 Excel），可将所查询的操作日志导出至本地。

◆ 支持导出 CSV 格式和 Excel 格式。

◆ 支持最多导出 10 万条，当前总数超过 10 万条则导出最新的 10 万条。

8.2.2 运维日志

查询运维日志

步骤 1. 登录大模型管理后台。

步骤 2. 在左侧导航栏选择“**日志检索➤运维日志**”，设置查询条件（如终端、分组、时间、日志类型、概况等），点击<**查询**>，即可查询符合该条件的运维日志。

导出运维日志

点击<**导出日志**>，租户可将所查询的运维日志导出至本地。

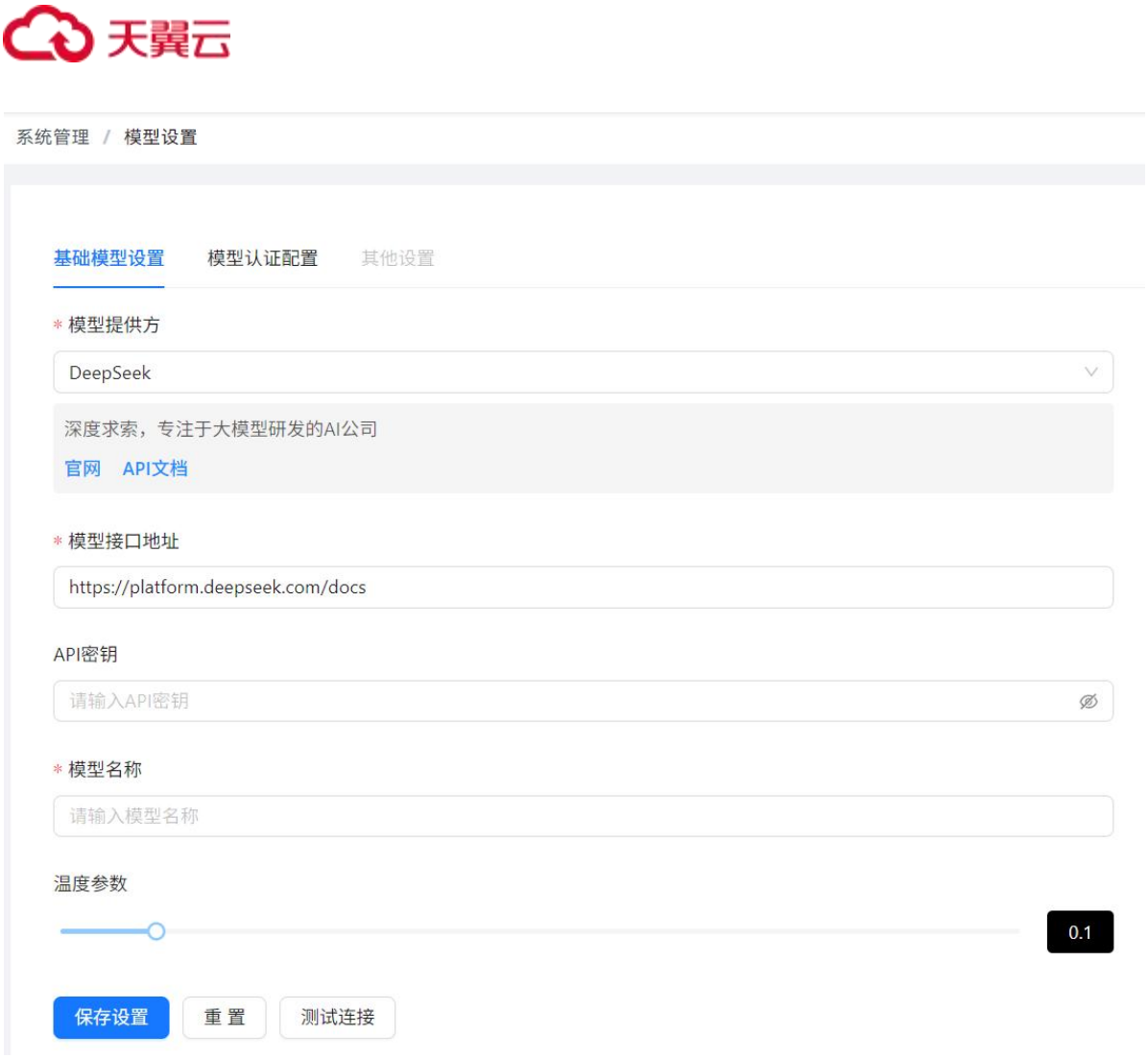
- 支持导出 CSV 格式和 Excel 格式。
- 支持最多导出 10 万条，当前总数超过 10 万条则导出最新的 10 万条。

8.3模型设置

8.3.1 基础模型设置

步骤 1. 登录大模型安全卫士实例。

步骤 2. 在菜单栏选择“系统管理 > 模型设置”，选择“基础模型设置”页签。



步骤 3. 根据实际情况修改配置参数。

参数	说明
模型提供方	支持 DeepSeek、Kimi、通义千问、OpenAI。
模型接口地址	根据实际情况进行填写。例如 https://platform.deepseek.com。
API 密钥	根据实际情况进行填写。没有密钥可不填。
模型名称	配置模型名称。

步骤 4. 配置完成后，单击“测试连接”，稍等几秒钟弹窗“连接测试成功”后，点击“保存设置”即可。

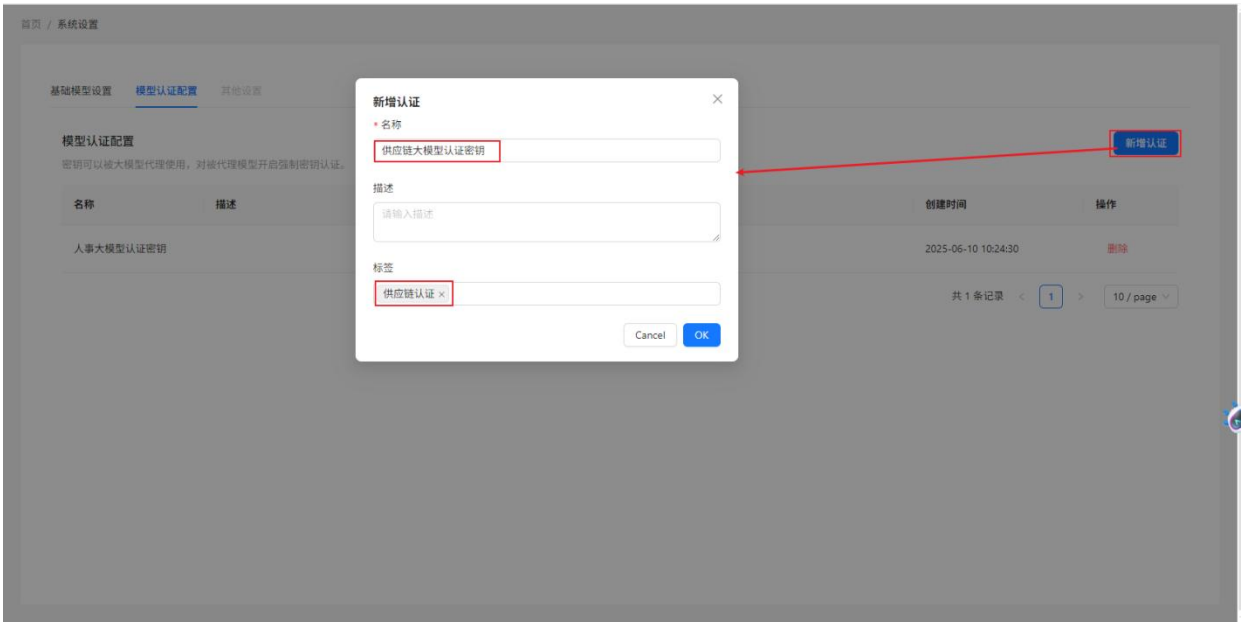
8.3.2 模型认证配置

企业内部基于开源大模型进行二次开发时，通常未内置标准化认证机制。容易导致未授权访问与资源滥用、数据泄露与敏感信息暴露、提示注入与恶意指令执行等风险。大模型防护系统通过对被代理模型的强制认证，以 OpenAI 标准核验 API key，解决未开启认证模型存在的未授权访问、数据泄露等风险，实现标准化兼容、全链路审计及抗 DDoS 等安全优势。

使用方式如下：

步骤 5. 进入 “系统管理 > 模型设置 > 模型认证配置”。

步骤 6. 点击 “新增认证”，输入认证名称和标签（标签记牢）。

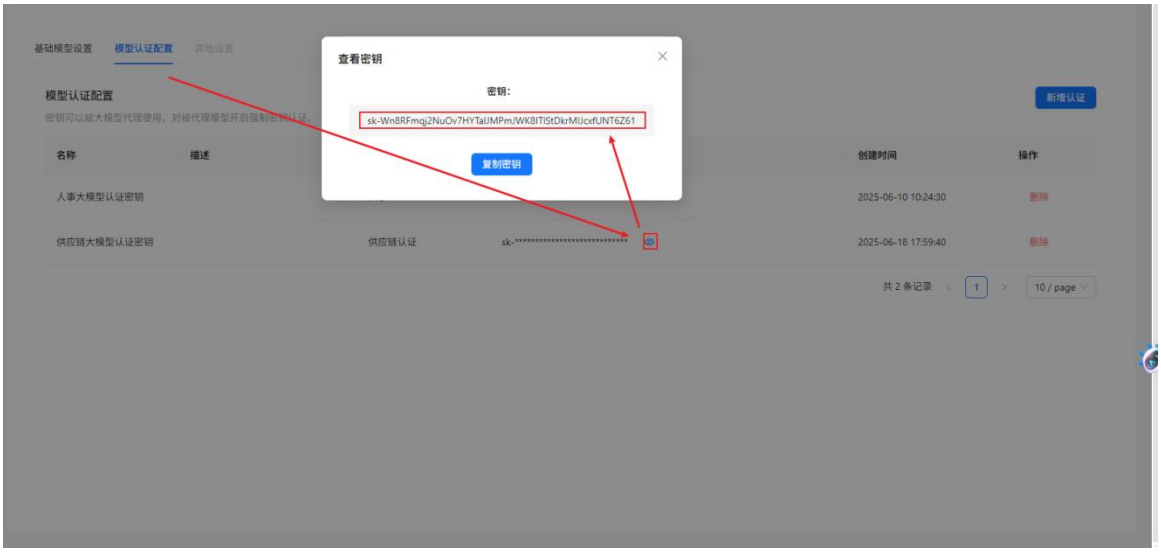


步骤 7. 进入 “资产中心” - “模型代理” - “网关代理”

步骤 8. 如果已经创建代理，点击编辑代理（如果代理正在运行请先关闭）；

步骤 9. 如未创建，则在创建代理时 <启用 API Key>，并在 API Key 标签栏输入第 2 步新增的标签名称。

步骤 10. 此时被代理模型必须进行认证才可访问，可进入 “系统管理”-“模型设置”-“模型认证配置”页面，查看代理密钥。



8.4在线检测

1、在线检测功能便于现场模拟内容安全的功能测试。

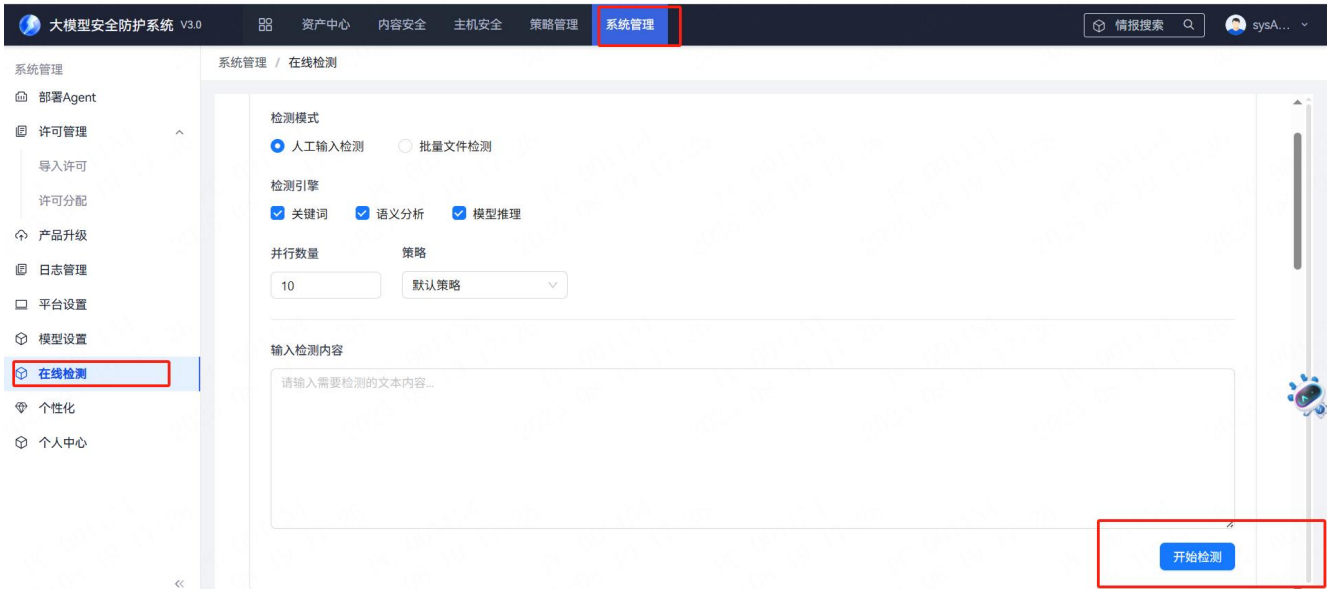
检测模式：人工输入检测和批量文件检测

检测引擎：敏感词、语义分析和模型推理

并行数量：可设置并行检测的数量

策略：可选择默认策略或者自定义的检测策略

输入检测内容或上传文件：检测模式选择人工输入检测则输入检测内容；检测模式选择批量文件检测则需要上传检测的内容



2、检测任务列表

检测的内容日志数据显示在检测任务列表中。包含字段有检测类型、检测内容、任务状态、检测统计、检测速度、创建时间或结束时间

9 最佳实践

9.1开启模型推理引擎

如果想获得更好的防护效果，那么非常推荐开启 <模型推理引擎>。开启模型推理引擎的前提是配置“基础模型”，首先需要准备一个大模型。推荐使用 “Qwen/Qwen2.5-32B-Instruct” 或其量化版本“qwen2.5:32b-instruct-q4_K_M”，其他 32b 以上模型也可以，但要注意 deepseek 和 qwen3 这类强制思考模型的防护效果会稍好一些，但检测时间会更长一些，用户可自行取舍。

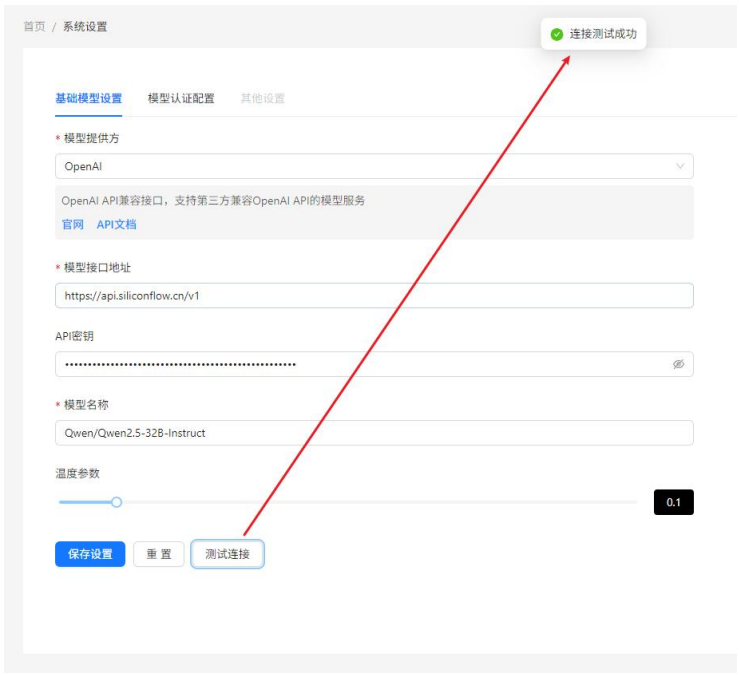
根据模型信息和模型提供地址信息拼接一条 curl 命令，用于测试网络连通性和模型是否正确。我们以调用硅基流动接口举例：

```
curl --location 'https://api.siliconflow.cn/v1/chat/completions' \
--header 'Content-Type: application/json' \
--header 'Authorization: Bearer sk-XXXX' \
--data '{
  "model": "Qwen/Qwen2.5-32B-Instruct",
  "messages": [
    {
      "content": "你好",
      "role": "user"
    }
  ]
}'
```

下面我们解析这段 curl 命令，填入到 “系统管理”-“模型设置”-“基础模型设置” 菜单的基础模型表单中。

1. https://api.siliconflow.cn/v1/chat/completions 这个链接是对话接口，我们去掉后面的 /chat/completions，保留服务地址和根路径，如 https://api.siliconflow.cn/v1 添加到 “模型接口地址” 的输入框。（注意这里通常是 /v1, 但也可能有所不同，但我们只要去掉 /chat/completions，前面全部保留即可）
2. --header 'Authorization: Bearer sk-XXXX' 这里，复制 sk-XXXX 填入 “API 密钥” 输入框，也可能没有密码可不填。（密钥不一定是 sk- 开头，如实填写即可）
3. "model": "Qwen/Qwen2.5-32B-Instruct", 注意这里的模型名称要完全一致的填入 “模型名称” 输入框。

填写无误后，点击 <测试链接>，稍等几秒钟弹窗 “连接测试成功” 后，点击 ”保存设置“ 即可。


首页 / 系统设置
连接测试成功
基础模型设置
模型认证配置
其他设置
* 模型提供方
OpenAI
OpenAI API兼容接口，支持第三方兼容OpenAI API的模型服务
官网 API文档
* 模型接口地址
https://api.siliconflow.cn/v1
API密钥
* 模型名称
Qwen/Qwen2.5-32B-Instruct
温度参数
0.1
保存设置
重 置
测试连接

10 常见问题

大模型安全卫士是什么？

答：大模型安全卫士是一款专为大模型安全设计的一站式防护产品，提供从开发、训练、部署到运营的全生命周期安全闭环，保护用户的智算基础设施。

开通大模型安全卫士需要哪些要求？

答：大模型安全卫士需要部署在和安装大模型的云主机所在的同一个虚拟私有云（VPC）内。

大模型安全卫士能解决智算场景下哪些安全问题？

答：有效拦截大模型推理过程中的潜在违规内容，对输入和输出的语义进行深度分析和检测，防止模型被利用进行恶意攻击或生成有害内容，强化了模型推理过程中的安全保障；

代理 rag 业务系统请求，解析文件进行内容检测。保障语料库及生成内容的安全性、合规性，防止恶意攻击（如数据投毒、提示注入）、敏感信息泄露及生成有害内容；

支持开启模型推理的情况下检测聊天内容中的隐私信息并脱敏。