**易盾智能审核使用手册**

1. **服务对接**

华为云云商店下单成功后，1-2个工作日内网易易盾官方售前与您取得联系沟通需求，并分配商务跟进。

1. **服务开通**

**1. 免费试用服务**

（1）与商务沟通后，前往网易易盾官网注册试用账号（手机号、邮箱注册均可）

（2）创建产品（创建一个需要开通服务的产品，填写产品名称、联系人等信息）

（3）申请免费试用（申请开通免费试用后，客服/商务人员联系您确认产品信息）

（免费试用审核时间约 1 个工作日）

（4）请审核通过，获取凭证，调试接口（查看产品密钥和业务 ID，进行接口调试）

（5）数据调通，成功开通试用服务（调通数据，开通试用服务，快速体验检测效果）

**2. 正式开通服务**

（1）注册网易易盾账号（手机号、邮箱注册均可） [立即注册](https://id.163yun.com/register?referrer=https://dun.163.com/dashboard&h=yd&t=yd)（已经注册的直接进入第二步）

（2）创建产品（创建一个需要开通服务的产品，填写产品名称、联系人等信息）

（3）创建订单后，请联系您的专属商务为您开通正式服务

（4）服务开通，即可获取凭证，调试接口（查看产品密钥和业务 ID，进行接口调用）

（5）正式使用网易易盾智能审核服务

**二、规范说明**

**1. 请求方法**

智能审核服务提供的所有接口，均支持 POST/GET 方法。当参数名与参数值拼装起来的 URL 长度小于1024 个字符时，可以用 GET 发起请求；当参数类型含 byte[] 类型或拼装好的请求 URL 过长时，必须用 POST 发起请求。建议所有 API 调用都使用 POST 方法请求。

**2. 字符编码**

所有接口的请求和响应数据编码皆为 UTF-8。

**3. API请求结构**

| 名称 | 描述 | 备注 |
| --- | --- | --- |
| API 入口 | 具体 API 接口地址 | 如 https://as.dun.163.com/v1/stream/push |
| 公共参数 | 每个接口都包含的通用参数 | 详见 [公共参数](https://support.dun.163.com/documents/599750646047514624?docId=599760545633415168) 说明 |
| 私有参数 | 每个接口特有的参数 | 详见每个 API 接口定义 |

**三、公共参数**

**1. 请求公共参数**

公共参数是用于标识产品和接口鉴权的参数，如非必要，在每个接口单独的接口文档中不再对这些参数进行说明，每次请求均需要携带这些参数。

| 名称 | 类型 | 是否必填 | 最大长度 | 描述 |
| --- | --- | --- | --- | --- |
| secretId | String | 是 | 32 | 产品密钥 ID，由易盾内容安全服务系统生成，可在智能审核后台-产品业务管理中获取，如有疑问可咨询策略经理引导获取，产品标识 |
| timestamp | Number | 是 | 13 | 请求当前 UNIX 时间戳，请注意服务器时间是否同步 |
| nonce | Number | 是 | 11 | 随机整数，与 timestamp 组合使用，用于防止重放攻击 |
| signatureMethod | String | 否 | 指定传参 | 指定签名算法，默认 MD5，支持 MD5，SHA1，SHA256，SM3 加密算法（Python 语言不建议使用 SM3，实测性能较差） |
| signature | String | 是 | 32 | 请求签名，用来验证此次请求的合法性，具体算法见 [接口鉴权](https://support.dun.163.com/documents/953823904240750592?docId=953831562254495744) |

**2. 响应公共字段**

接口响应内容类型使用 JSON 格式，如无特殊说明，接口的返回值中均包含下面字段：

| 字段名称 | 类型 | 描述 |
| --- | --- | --- |
| code | Number | 接口调用状态，其中 200 表示调用成功；其他返回码见 [响应返回码](https://support.dun.163.com/documents/953823904240750592?docId=953831302539481088) |
| msg | String | 接口响应结果说明，若成功返回 ok，若接口调用出错，返回详细错误信息 |
| result | 对象或数组 | 接口响应结果，详细返回结构请参考具体接口的文档说明 |

**四、接口鉴权**

**1. 获取安全凭证**

智能审核服务使用签名方法对接口进行鉴权，所有接口每一次请求都需要包含签名信息（signature 参数），以验证用户身份，防止信息被恶意篡改。

在第一次使用 API 之前，需获取安全凭证，安全凭证包括 secretId 和 secretKey：

* secretId 是用于标识 API 调用者的身份
* secretKey 是用于加密签名字符串和服务器端验证签名字符串的密钥，务必严格保管，避免泄露。

获取凭证路径：易盾控制台 -> 内容安全 -> 服务管理 -> 查看产品密钥

**四、接口调用**

在使用 AI 大语言模型时，推荐通过流式输出的方式快速将中间结果显示在用户界面上。目前，易盾 AIGC 内容检测解决方案适用于检测 AI 大模型输入和流式输出的文本。该解决方案的提交接口支持 AI 大模型输入和流式输出的提交。针对流式输出场景，该接口能智能地将模型的流式输出文字进行拼接并检测，同时同步返回易盾内容安全服务的实时检测结果。

AIGC 内容检测解决方案目前仅支持同步检测方式。建议在接入使用时，在对同一段流式输出内容的片段进行检测时，采用串行调用的方式，以确保片段拼接的顺序。

常规场景



* 步骤1：服务端对接在线检测接口，接口详见 [AIGC内容检测提交接口](https://support.dun.163.com/documents/953823904240750592?docId=953841408377774080)。提交检测数据后，会在提交接口响应中实时返回机器检测结果（通过/不通过/嫌疑/无结果），无结果是指当前会话累积的待检测片段数据暂时未满足智能化拼接过检的条件，会返回正常的响应体但是没有机器检测结果。
* 步骤2：服务端解析在线检测返回结果
* 步骤3：对接在线检测结果处置动作
* 步骤4：对接回调接口，获取检测结果或人工审核结果的方式有 [轮询模式](https://support.dun.163.com/documents/953823904240750592?docId=953843610446209024)、[推送模式](https://support.dun.163.com/documents/953823904240750592?docId=953843635578191872) 这两种。推荐使用推送模式，以便及时接收到回调结果。
* 步骤5：对接离线检测结果处置动作

在 AIGC 内容检测的流式输出检测场景中，可能会出现主动关闭活跃会话的情况（例如，会话超时关闭或开启新的活跃会话）。在关闭会话时，系统会对待检测的片段数据（如果存在）进行一次检测。客户需要通过接入回调来获取这些片段数据的检测结果。

**五、接口提交**

智能审核解决方案目前适用于对 AI 大模型流式输出的文字进行自动拼接后审核。同步检测接口用于提交待检测数据，并同步返回易盾内容安全服务的实时检测结果。

接口说明

该接口是智能审核解决方案中的数据提交接口，需要使用 HTTP 协议进行调用。该接口会同步返回检测结果。

鉴权说明

智能审核服务使用签名方法对接口进行鉴权。接口请求需要包含签名信息（signature 参数）以验证用户身份，防止信息被恶意篡改。目前支持 MD5、SHA1、SHA256、SM3 等几种加密算法，鉴权流程详见 [接口鉴权](https://support.dun.163.com/documents/953823904240750592?docId=953831515169722368) 。

接口限制

文本限制：

* 单文本切片：切片长度 < 50 字
* 自动拼接文本：拼接长度 < 200 字

检测耗时

检测接口的响应时间与拼接长度相关，根据不同长度建议适当调整接口超时时间，200 字以内建议 200ms 即可

请求

请求地址

| 名称 | 值 |
| --- | --- |
| HTTP URL | https://as.dun.163.com/v1/stream/push |
| HTTP Method | POST |

请求头

| 名称 | 类型 | 描述 |
| --- | --- | --- |
| Content-Type | String | 固定值："application/x-www-form-urlencoded" |

请求参数

请求参数由公共参数和业务参数两部分组成

公共参数

公共参数包含密钥信息和参数签名信息，公共参数详细见 [请求公共参数](https://support.dun.163.com/documents/953823904240750592?docId=953830817779142656) 。

业务参数

| 名称 | 类型 | 是否必填 | 最大长度 | 描述 |
| --- | --- | --- | --- | --- |
| sessionId | String | 是 | 128 | 业务侧会话 ID，用于关联整个流式检测会话，标记当前请求内容片段所属的流式内容； |
| type | Number | 否 | - | 推送的事件类型，事件类型有 1：流式输出检测、2：输入检测、3：问答结束（一次问答结束时必须调用，否则会占用并发数） ，默认值为：1 |
| content | String | 见描述 | 见描述 | type = 1：content 必传，表示流式检测内容片段，可对应 AIGC 场景中的流式输出 tokens。检测片段传入 AIGC 输出文本中，默认最大长度为 50。type = 2：content 必传，表示流式检测场景下的输入内容。建议对内容中的 JSON、表情符、HTML 标签、UBB 标签等进行过滤，只传递纯文本，以减少误判概率。对应传入 AIGC 输入文本中，最大长度为 10000。type = 3：问答结束时，content 非必填。如果传入 content，会按照 type = 2 的逻辑处理输入片段，并关闭当前活跃问答。 |
| version | String | 是 | 4 | 接口版本号，可选值 v1.0 |
| publishTime | Number | 否 | 13 | 时间戳用于标识数据片段生成的时间，建议客户侧生成并传递，拼接的内容将基于时间戳进行排序。时间戳应为 UNIX 时间戳（毫秒值）。 |
| dataId | String | 否 | 128 | 业务侧数据唯一标识，能够根据该值定位到该条数据，如对数据结果有异议，可以发送该值给策略经理查询，建议提供该参数。 |
| callback | String | 否 | 2^16-1 | 数据回调参数，业务侧可根据业务场景需要自定义，易盾调用回调接口（callbackUrl）时，会原样返回该参数的值 |
| callbackUrl | String | 否 | 256 | 业务侧回调接口 URL，用于异步通知业务侧异步及离线策略检测结果。接口调用超时时间为 2s，需业务侧确保接口的稳定性和幂等性 |
| checkLabels | String | 否 | 512 | 业务自定义过检分类，请谨慎写入，并与策略经理充分沟通，以免引起漏判，如果没有勾选分类提交返回参数错误，可指定多个垃圾类别进行机器检测，多个垃圾类别以逗号分隔（如 100,200），指定后业务配置过检分类即失效，支持的过检分类列表，100：色情，200：广告，260：广告法，300：暴恐，400：违禁，500：涉政，600：谩骂，700：灌水，900：其他，1100：涉价值观 |

业务扩展参数

建议将检测内容的用户信息、设备信息、场景信息同步传入，易盾将结合用户风险画像、设备风险画像、场景化策略全方位检测内容风险，检测结果更精准。

| 参数名称 | 类型 | 是否必填 | 最大长度 | 描述 |
| --- | --- | --- | --- | --- |
| account | String | 否 | 128 | 用户唯一标识，与易盾账号画像库匹配，建议抄送，辅助机审策略精准调优，如业务方同时接入易盾反外挂，则此账号传值需要与反外挂接入中的 roleAccount（Android 端），或 user\_account（iOS 端）保持一致 |
| ip | String | 否 | 128 | 用户 IP 地址，建议抄送，辅助机审策略精准调优 |
| deviceId | String | 否 | 128 | 用户设备 ID，与易盾设备画像库匹配，明文请转大写传入；MD5 加密请明文转大写后 MD5 计算，再转大写传入，建议抄送 |
| extStr1 | String | 否 | 128 | 自定义扩展参数 |
| extStr2 | String | 否 | 128 | 自定义扩展参数 |
| extLon1 | Number | 否 | 2^63-1 | 自定义扩展参数 |
| extLon2 | Number | 否 | 2^63-1 | 自定义扩展参数 |
| extension | Json | 否 | 30000 | 更多自定义扩展参数，JSON 字符串格式。如：{"keyName1":"value1","keyName2":"value2"} |

请求体示例

content=token1&version=v1.0&dataId=163&signature=xxx&secretId=xxx×tamp=1646984669446&nonce=12353

复制代码 切换背景

响应

响应结果

响应字段如下，响应通用字段已省略，详细见响应公字段：[响应公共字段](https://support.dun.163.com/documents/953823904240750592?docId=953831610187001856)

| 字段名称 | 类型 | 是否必返 | 描述 |
| --- | --- | --- | --- |
| result | Object[] | 是 | 文本回调结果 |
|  riskdescription | String | 是 | 风险描述，拼接 label｜secondlabel｜thirdlabel 的中文实时示意，注意分隔使用「｜」，将上层对应一级分类带入，若存在二级/三级透出拼装，若无可不显示 |

响应体示例

无结果时（无结果是指当前会话累积的待检测片段数据暂时未满足智能化拼接过检的条件，会返回正常的响应体但没有机器检测结果），系统会返回一个正常的响应体，但不会包含任何检测结果。这意味着尽管请求已经成功处理，但是由于条件未满足，暂时没有可供返回的检测结果。

{

 "code": 200,

 "msg": "ok",

 "result": {

 "sessionTaskId": "fx6sxdcd89fvbvg4967b4787d73a",

 "sessionId": "74fb17f1-1404-4230-9b23-8e8157e50bf3"

 }

}

复制代码 切换背景

有结果时：

{

 "code": 200,

 "msg": "ok",

 "result": {

 "sessionTaskId": "fx6sxdcd89fvbvg4967b4787d73a",

 "sessionId": "74fb17f1-1404-4230-9b23-8e8157e50bf3",

 "antispam": {

 "dataId": "1678256782070",

 "suggestion": 1,

 "customAction": 0,

 "censorType": 0,

 "isRelatedHit": false,

 "censorTime": 1678256782119,

 "resultType": 1,

 "taskId": "fx6sxdcd89fvbvg4967b4787d78a",

 "labels": [

 {

 "subLabels": [

 {

 "subLabel": "200009",

 "subLabelDepth": 2,

 "details": {

 "hitInfos": [

 {

 "positions": [

 {

 "fieldName": "content",

 "startPos": 0,

 "endPos": 3

 }

 ],

 "value": "加微信"

 }

 ]

 },

 "secondLabel": "200009"

 }

 ],

 "level": 1,

 "rate": 0.5,

 "label": 200

 }

 ]

 },

 "language": {

 "taskId": "fx6sxdcd89fvbvg4967b4787d78a",

 "dataId": "dataId",

 "details": [

 {

 "type": "zh"

 }

 ]

 },

 "aigcPrompt": {

 "dataId": "1678256782070",

 "taskId": "fx6sxdcd89fvbvg4967b4787d78a",

 "details": [

 {

 "source": 0,

 "type": 1,

 "answer": "aa"

 }

 ]

 }

 }

}

复制代码 切换背景

示例代码

服务端 SDK 接入

推荐使用我们提供的服务端 SDK 进行接入（目前支持 Java 和 Go 两种语言），不但可以节省接入成本，实现更快速的集成，并享有灵活的请求配置选项，同时 SDK 内置了错误重试、熔断降级等功能，提升服务可用性。点击 [服务端 SDK 接入指引](https://support.dun.163.com/documents/594247746924453888?docId=682781209345695744) 了解更多。

Github Demo 地址

* [Java SDK Demo](https://github.com/yidun/yidun-java-sdk/blob/main/yidun-java-sdk-demo/src/main/java/com/netease/yidun/sdk/antispam/text/v5/TextSingleSyncCheckDemo.java)
* [Golang SDK Demo](https://github.com/yidun/yidun-golang-sdk/blob/main/yidun/demo/antispam/stream/v1/push/aigc_stream_push_demo.go)

Gitee Demo 地址（Github 无法访问时选择）

* [Java SDK Demo](https://gitee.com/netease_yidun/yidun-java-sdk/blob/main/yidun-java-sdk-demo/src/main/java/com/netease/yidun/sdk/antispam/stream/v1/AigcStreamPushDemo.java)
* [Golang SDK Demo](https://gitee.com/netease_yidun/yidun-golang-sdk/blob/main/yidun/demo/antispam/stream/v1/push/aigc_stream_push_demo.go)

HTTP API 接入

Github Demo 地址

* [Java Demo](https://github.com/yidun/antispam-java-demo/blob/master/src/main/java/com/netease/is/antispam/demo/stream/v1/AigcStreamPushAPIDemo.java)