拿破仑

内容安全解决方案

MIAOCHU CONTENT SECURITY

革新科技

提供全方位的内容安全服务

CONTENT SECURITY

CONTENT SECURITY

目录

CONTENTS

01 团队介绍

02 业务矩阵

03 客户体验

04 接入流程

SECTION 01

01

团队介绍

- 拿破仑安全团队，专注于为泛娱乐内容提供合规的安全解决方案

团队介绍

INTRODUCE

拿破仑安全团队，专注于为泛娱乐内容提供合规的安全解决方案。

为业务安全和内容安全提供双重保障，助力合作伙伴实现安全与可持续发展，与合作伙伴一起共同构建一个安全可信的泛娱乐生态环境。

2013年

推出了第一代人工审核系统，实现了对视频流的“实时截图-识别-审核-控制”的完整审核流程

2015年

自主研发了图像识别系统，为直播和短视频领域的内容审核工作提供了强有力的支持

2018年

自主研发了音频识别系统，成为行业内多个安全技术方向上的效果评测标准

2019年

引入了多模态数据模型，将图像、音频和文本综合运用，进一步提升内容安全的准确性和全面性

2020年

推出信用标签体系，助力合作企业落实降本增效管理与精细化运营

2022年

行业首创

行业首创即时屏蔽技术，推动内容审核领域“人机结合”的模式改革

SECTION 02

02

业务矩阵

- 核心API检测能力
- 事前屏蔽智能解决方案
- 内容安全专家服务

产品矩阵

PRODUCTS

核心能力

核心API检测能力

图像检测

涉政 | 涉黄 | 涉恐暴 | 低俗 | 人脸检测 | OCR

音频检测

ASR | 娇喘 | 版权歌曲 | 红政

文本检测

自定义敏感词库 | 自然语言处理（NLP） | 文本分类

拳头产品

事前屏蔽智能解决方案

即时消音

涉政 | 涉黄 | 谩骂

即时文本过滤

异常检测 | 语义分析 | 情感分析 | 句法分析 | 文本分类与聚集

即时打码

研发中...敬请期待☺

专家认证

内容安全专家服务

安全专家服务

安全舆情 | 安全培训

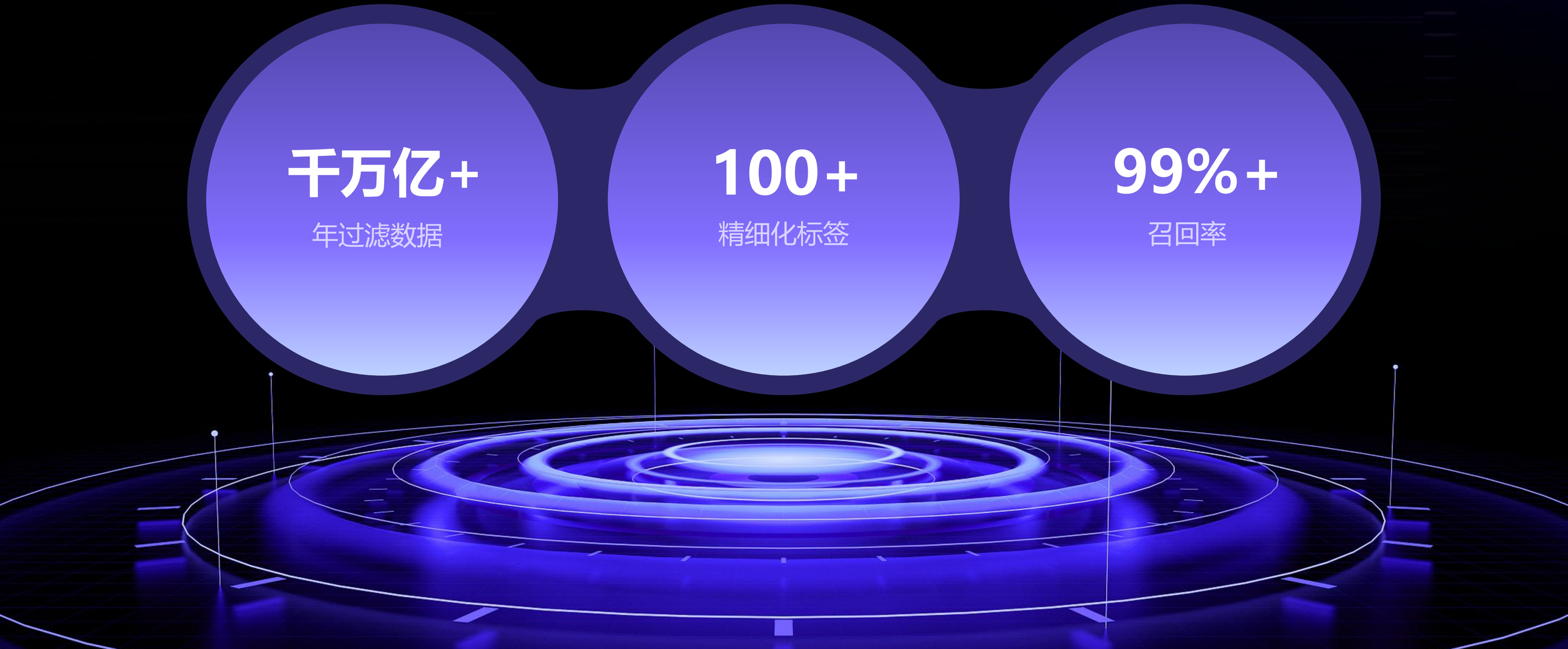
合规咨询服务

双新测评 | 合规咨询

共建探索

私有化部署 | 算法共建咨询

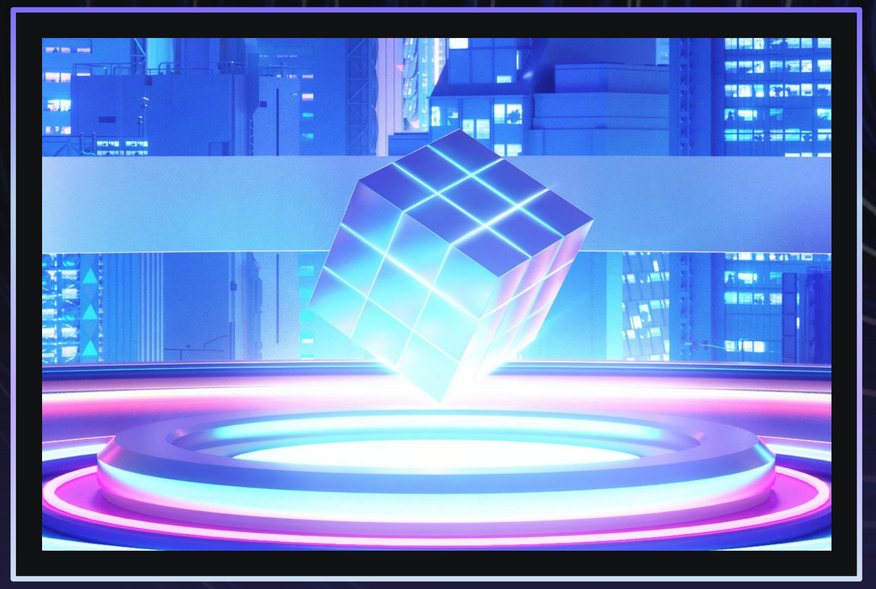
效果介绍



核心API检测能力

PRODUCT POWER PART1

核心能力



文本检测

核心API检测能力

千万级
策略集

20+
多语种策略支持

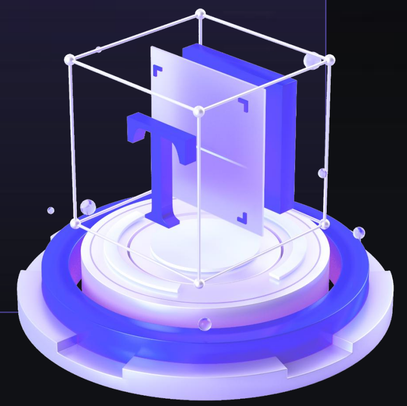
定制化
关键词/策略

十毫秒级别
接口响应

基础类别能力

涉政	支持敏感专项、时事报道、领导人相关、英雄烈士相关、邪教迷信、落马官员相关、热点舆情、宗教、恶搞领导人相关等能力
色情	支持色情传播、色情性器官、低俗段子、挑逗、性行为等相关等精细化子分类返回
暴恐	针对危害公共安全、国家安全的内容进行识别
违禁	支持刀枪弹药、赌博、违禁工具、违禁代办、违禁药品等相关等精细化子分类返回

- 生僻词支持
- 文本聚类
- 行为识别
- 词向量模型
- NLP技术
- 文字变种



图片检测

核心API检测能力

定制化
人脸库、关键词/策略

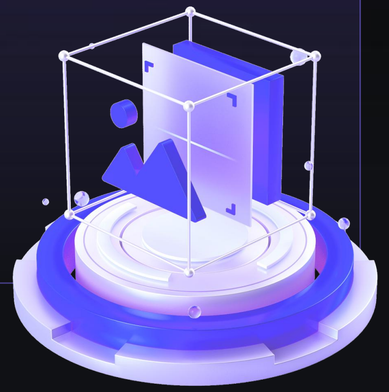
百毫秒级别
接口响应

基础类别能力

涉政	支持核心领导人、领导人相关、英雄烈士相关、落马官员相关、热点舆情、宗教相关等精细化子分类返回
色情	支持性器官、低俗、挑逗、性行为等相关等精细化子分类返回
暴恐	针对危害公共安全、国家安全的内容进行识别
违禁	支持刀枪弹药、赌博、违禁工具、违禁药品等相关等精细化子分类返回

专项识别能力

人脸识别	自定义人脸识别
LOGO识别	长短视频版权LOGO、赛事版权LOGO等
图片OCR	识别图片中的文字信息，检测有害内容



视频检测

核心API检测能力

检测流程

视频源

直播视频

点播视频

预处理

拉流

下载截帧

图片检测

黄暴政

广告

违禁

禁播游戏

黑屏挂机

覆盖场景

直播间

短视频

点播视频

多人互动聊天

辅助检测

智能截帧

智能电视墙

视频MD5库

主播黑名单

直播间热度监控

音频检测

核心API检测能力

声音输入

突出人声、噪音过滤

ASR识别

特征提取 | 声学模型 | 语言模型 | 输出文字结果

音节　字/词　句子

声纹检测

ASMR、娇喘、领导人声纹

特征提取 | 声学模型 | 输出结果

歌曲识别（覆盖9500+、支持自定义）

当你 | 加油 | 明天过后...

声音分类

静音 | 人声 | 歌声 | 噪音 | 声音属性：性别、未成年、年龄

反垃圾检测

多意义上下文短文本垃圾检测

Deep Learning垃圾检测

规则引擎、分类器等

输出反垃圾综合结果

- 风险级别：删除 | 通过 | 嫌疑
- 命中信息：娇喘 | 谩骂
- 命中断句时间戳：[2:33:12-2:33:32]
  （自然断句，支持最短返回断句时长）
- 音频标签：中文 | 人声 | 女声

音频检测

核心API检测能力

自定义

歌曲检测、关键词

- 语音识别
- 语种识别
- 人声属性

语音识别能力范围

涉政类语音	新中国联邦成立后郭文贵就可以在国外开创新中国了
色情类语音	打个飞机全身是汗，撸完就是爽
推广类	我们平台流量大，平台日结提现
其他违规类	危害国家安全/谩骂类/空语音无意义

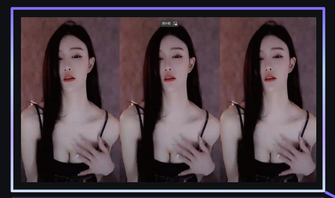
声纹检测能力范围

娇喘	展示或模拟星星为过程中发出的具有明显挑逗性的声音
ASMR	展示或模拟克诱导听觉自发性知觉经络反应的声音，如口腔音/舔耳等
违禁歌曲	涉及领导/历史事件/热点舆情相关的恶意抨击调侃等不实歌曲识别精细化子分类返回
涉政人物声纹	涉及11重要涉政人物及家属的声音识别 （毛泽东/江泽民/习近平/彭丽媛/郭文贵/王立铭/郝海东/何岸泉/李洪志/青海无上师/热比娅卡德尔/十四达赖喇嘛/王丹/吾尔开希/周锋锁/黄之锋）

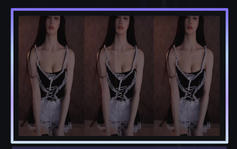
事前屏蔽智能解决方案

PRODUCT POWER PART2

拳头服务



性感-呻吟



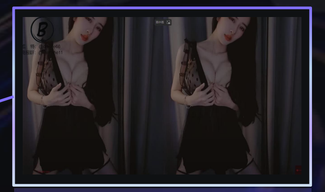
性感-勾线

规游戏

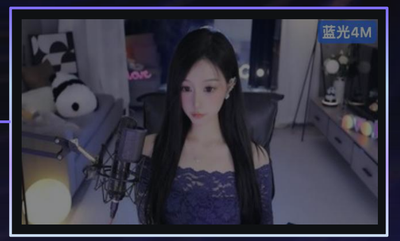
版权违规

LIVE

✔ 合规内容



性感-勾线



性感-呻吟

赛事版权

即时消音

事前屏蔽智能解决方案

低延时

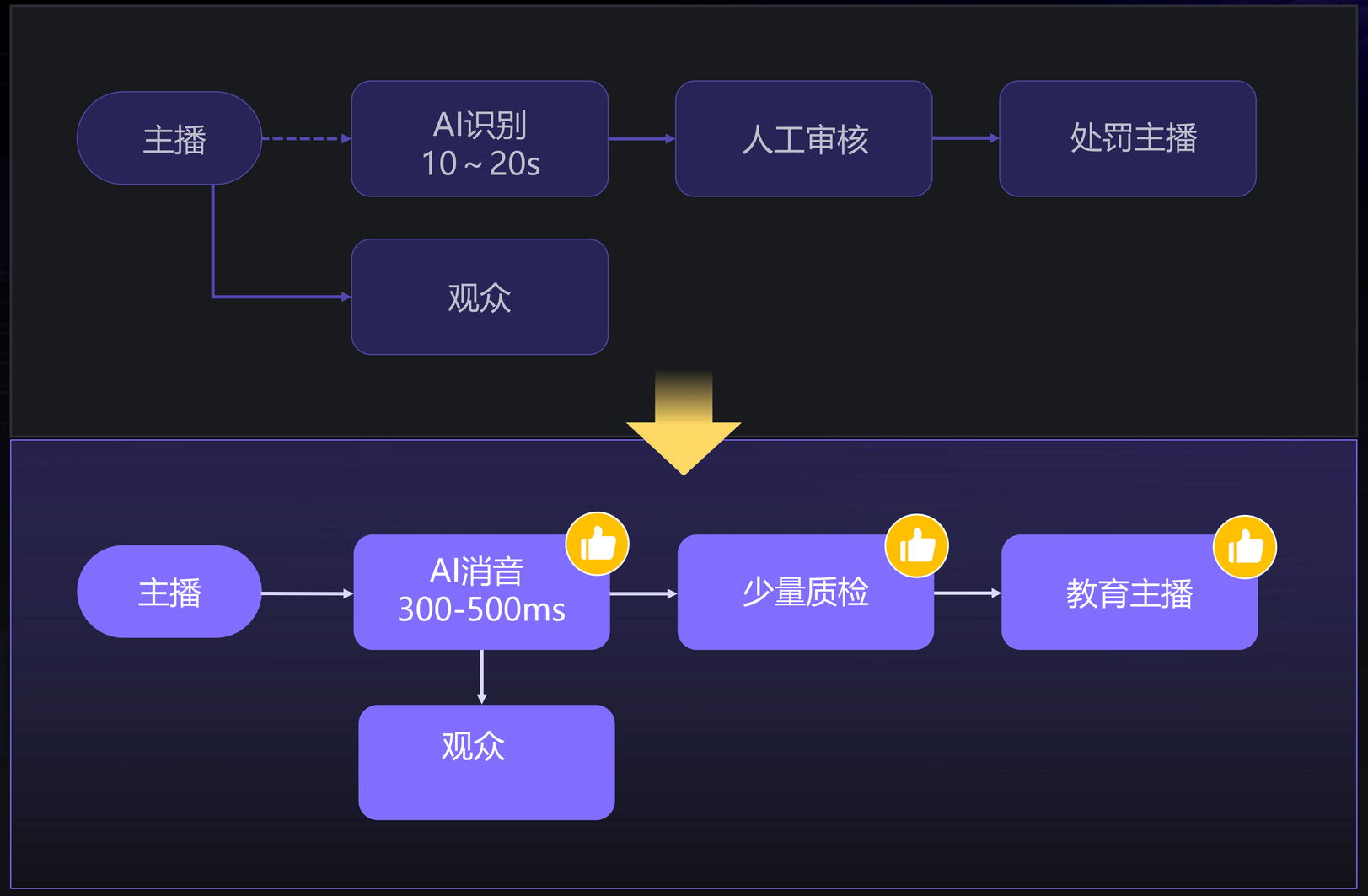
实时率 5%

无感知

单次消音时长 0.7s

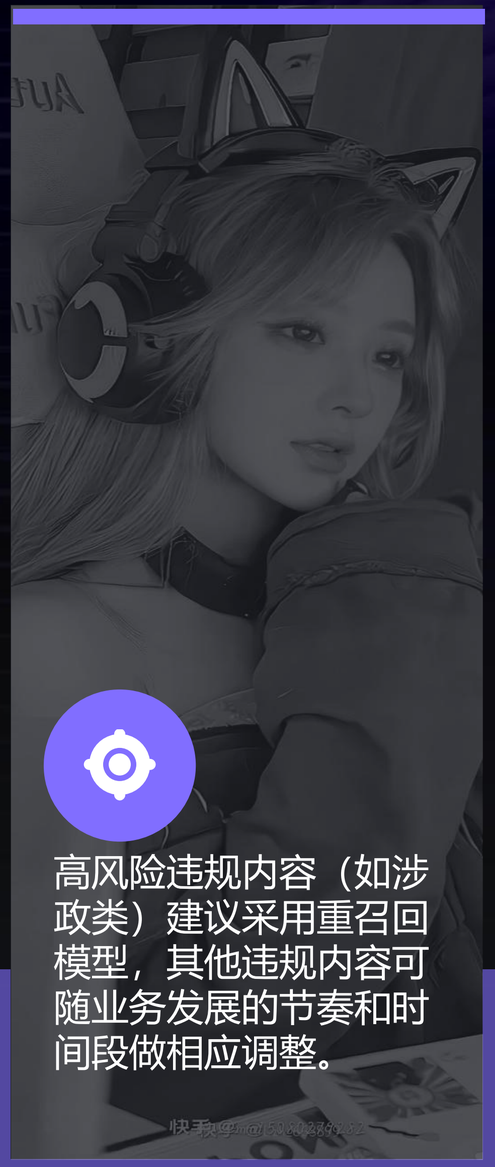
少量人审

较推审模式成本低 85%



不同厂商ASR指标对比（数据为最新一次评测，支持客户自主评测）

		黑词识别准确率	黑词识别召回率
整体	A厂商	83.26%	76.21%
	B厂商	76.69%	80.24%
	C厂商	79.76%	81.85%
	消音	71.02%	97.43%
		黑词识别准确率	黑词识别召回率
涉政	A厂商	78.99%	93.97%
	B厂商	73.57%	88.79%
	C厂商	81.75%	88.79%
	消音	68.82%	95.27%
		黑词识别准确率	黑词识别召回率
涉黄	A厂商	73.91%	79.53%
	B厂商	78.26%	70.59%
	C厂商	79.35%	73.37%
	消音	72.40%	97.49%
		黑词识别准确率	黑词识别召回率
谩骂	A厂商	67.86%	91.72%
	B厂商	77.04%	86.29%
	C厂商	80.10%	85.33%
	消音	88.51%	99.40%



高风险违规内容（如涉政类）建议采用重召回模型，其他违规内容可随业务发展的节奏和时间段做相应调整。

提供给客户自定义的黑词库管理

模型	违规标签	当前黑词量
涉政	涉一号	1000+
	党政相关	200+
	敏感事件	300+
	政民争议	500+
	反华分裂	1000+
	英烈相关	100+
	国&红歌	100+
	其他国家领导人	300+
	邪教迷信	400+
	六四相关	1000+
	历史事件	100+
	合计	**5000+**
涉黄	描述性行为或性器官	2000+
	低俗，引起不适	600+
	淫秽制品相关	200+
	传播色情行为	200+
	低俗歌词	300+
	合计	**1500+**
谩骂	谩骂-人身攻击	2000+
	谩骂-口头禅粗口	500+
	合计	**2500+**

消音策略灵活化

主播端提示

主播端

分策略消音

全消

o鸡巴k吗，o鸡巴k，o鸡巴k，o鸡巴k，o鸡巴k

2消1

o鸡巴k吗，o鸡巴k，o鸡巴k，o鸡巴k，o鸡巴k

3消1

o鸡巴k吗，o鸡巴k，o鸡巴k，o鸡巴k，o鸡巴k

即时文本过滤

事前屏蔽智能解决方案

90%+
拦截准确率

90%+
违规召回率

异常检测

语义分析

情感分析

句法分析

文本分类与聚类

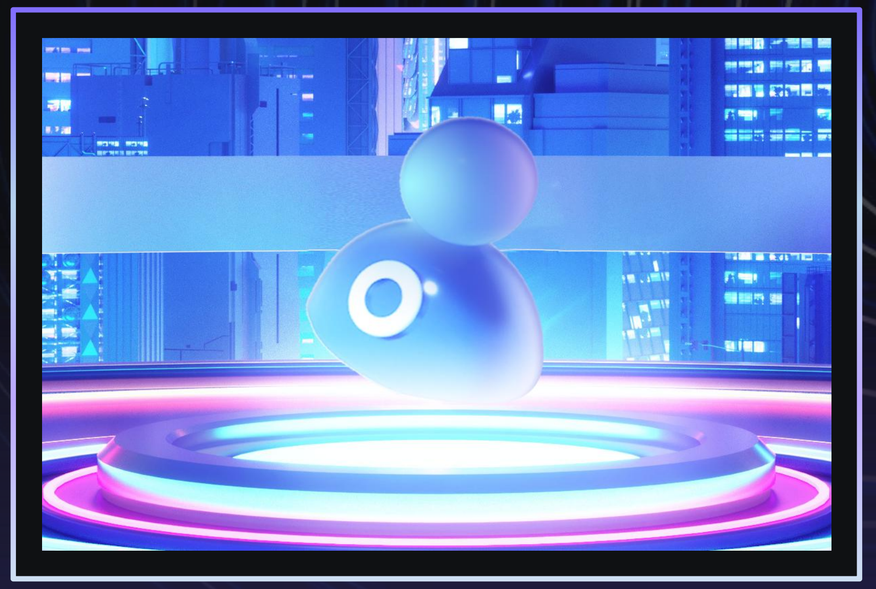
即时打码

事前屏蔽智能解决方案

内容安全专家服务

PRODUCT POWER PART3

专家认证

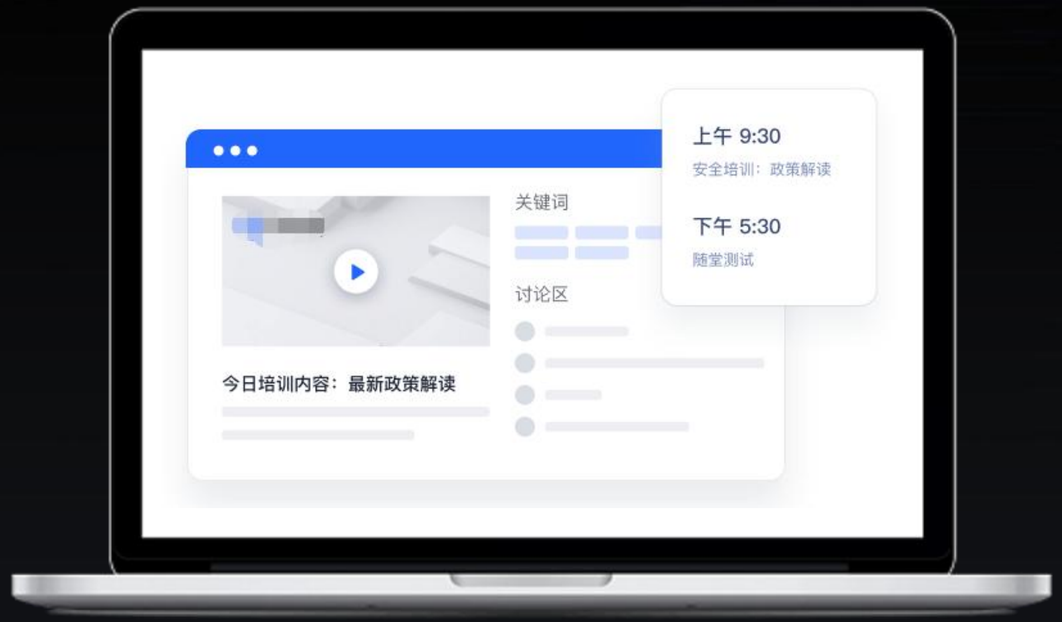


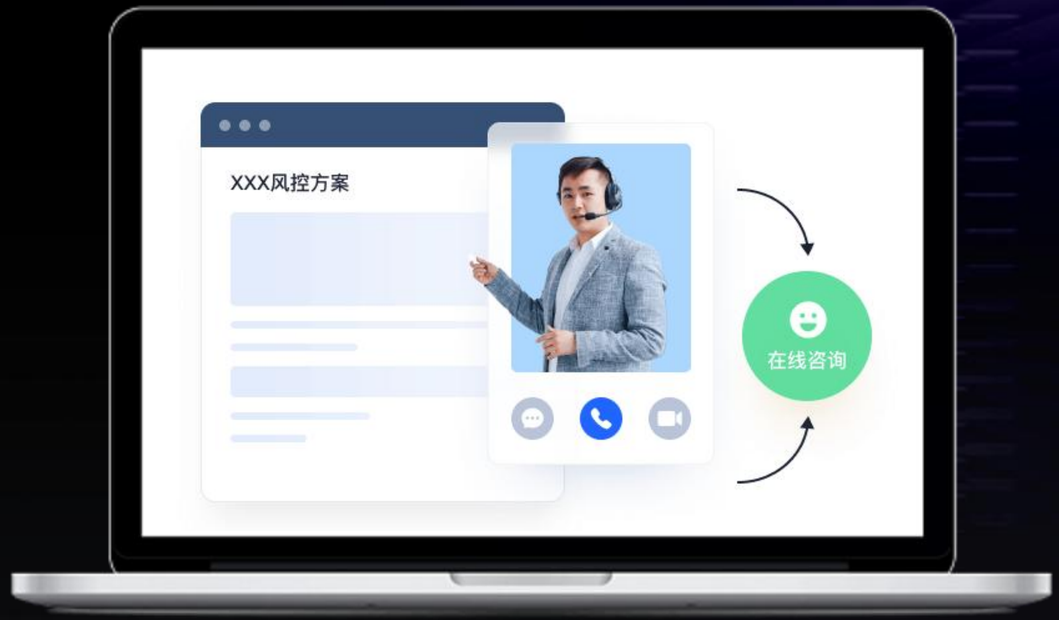
专家矩阵服务

内容安全专家服务

内容安全政策解读

帮助企业快速了解国内内容安全现状，引起业务部门重视，为安全工作的开展做铺垫





综合解决方案定制

风控专家将对产品进行全面体检，梳理出重点业务环节的风险现状，提出针对性改善方案。对产品完成安全性评估，输出适合产品全场景的风控解决方案

专家矩阵服务

内容安全专家服务

重大专项舆情预案

每年重大专项前准备舆情预案，突出体现当年重点舆情布控方向，行业相关或产品相关重点舆情布控方向，帮助客户打有准备之仗。

红蓝军对抗

以攻击者姿态，向不同的场景投放蓝军样本，并监控蓝军样本的处置方式、处置时长，逐渐升级攻击难度，输出产品内容风控蓝军报告，帮助客户知晓产品抗攻击能力

产品解决方案

PRODUCT SOLUTION

产品解决方案

PRODUCTS

多场景综合解决方案

图文动态 | 私聊/群聊 | 评论 | 直播 | 互动音视频 | 弹幕 | 用户资料 | 邮件 | 新闻资讯 | 网文小说 | 游戏聊天区

文本 | 图片 | 视频 | 音频

智能AI检测

- 涉黄
- 暴恐
- 涉政
- 违禁
- 广告
- 二维码
- LOGO 识别
- OCR 识别
- 歌曲、声纹识别
- 语音识别
- 人脸识别

即时屏蔽

- 0曝光
- 低延时
- 免人审

行业首创

支持服务

- 安全舆情与安全培训
- 合规咨询服务
- 定制标签
- 策略配置
- 算法共建

方案架构 – AI检测+人审

业务场景

私聊 | 群聊 | 动态广场 | 评论 | 用户资料 | 弹幕

单次提交过检

语聊房 | 视频聊天 | 娱乐直播 | 赛事转播

拉流监听持续检测

业务内容

文本、图片、点播视频、点播音频

语音流，视频流

数字内容风控引擎

机器检测

过检数据 → 文本 / 音频 / 图片 / 视频

内容检测

色情/性感/涉政
暴力/
违禁/广告/
谩骂/灌水/
恶心/涉价值观...

内容识别

情感分析
观点分析
人脸识别
文本纠错
场景识别
LOGO识别

机器检测管理后台

规则策略设置

规则引擎
关键词名单
图片名单
设备名单
自定义名单

产品/业务管理

策略配置与管理

机检数据查询

业务风控

人机识别、黑名单库、IP画像、设备指纹、行为模型、业务模型、全链路分析

安全专家服务

舆情顾问、安全蓝军、风控咨询、合规咨询、安全质检、安全培训

SECTION 03

03 客户体验

- 项目经验1
- 项目经验2
- 客户列表

客户体验

CUSTOMER CASE

项目一：某互联网上市公司

CASE SHARING

项目痛点

客户没有系统性的音频审核机制，仅对用户举报内容进行处理，投诉举报量大，人审处理不及时，被监管点名整改

解决方案

提供了“实时消音+AI音频检测+人审”音频一站式解决方案，同时提供消音策略自定义配置，满足客户平台生态需求

效果

使用我方产品服务期间，客户平台无音频风险舆情发生，来自于各方的举报与投诉量大幅减少

客户体验

CUSTOMER CASE

项目二：某互联网上市公司

CASE SHARING

项目痛点

初创公司有一定的安全意识，但不知如何高效布控风险内容的管控手段

解决方案

基于喵触多年的风控经验，我司对其提供专家服务，针对其业务特点进行多种高效的管控策略制定、敏感专项防控，依据客户业务节奏逐步迭代内容风控方案

效果

- 客户业务上线初期，配备轻量级管控策略，确保“有证”驾驶；
- 客户业务扩张阶段，上线信用标签管理机制，帮助客户实现精细化运营

SECTION 04

04

接入流程

- SaaS API接口 / 在线检测

接入流程

PROCESS

SaaS API接口

✓应用场景：推荐调用方式，灵活通过http接口提交数据检测，获取检测结果，实时作用在业务处理

✓支持能力：文本/图片/视频/直播流检测

在线检测

✓应用场景：无研发介入，需快速获得本地数据的检测结果

✓使用方式：申请开通系统，后台上传本地数据，获取检测结果


THANKS!
感谢大家观看！