

文本内容安全

产品简介



腾讯云

【 版权声明 】

©2013–2025 腾讯云版权所有

本文档（含所有文字、数据、图片等内容）完整的著作权归腾讯云计算（北京）有限责任公司单独所有，未经腾讯云事先明确书面许可，任何主体不得以任何形式复制、修改、使用、抄袭、传播本文档全部或部分内容。前述行为构成对腾讯云著作权的侵犯，腾讯云将依法采取措施追究法律责任。

【 商标声明 】



及其它腾讯云服务相关的商标均为腾讯云计算（北京）有限责任公司及其关联公司所有。本文档涉及的第三方主体的商标，依法由权利人所有。未经腾讯云及有关权利人书面许可，任何主体不得以任何方式对前述商标进行使用、复制、修改、传播、抄录等行为，否则将构成对腾讯云及有关权利人商标权的侵犯，腾讯云将依法采取措施追究法律责任。

【 服务声明 】

本文档意在向您介绍腾讯云全部或部分产品、服务的当时的相关概况，部分产品、服务的内容可能不时有所调整。您所购买的腾讯云产品、服务的种类、服务标准等应由您与腾讯云之间的商业合同约定，除非双方另有约定，否则，腾讯云对本文档内容不做任何明示或默示的承诺或保证。

【 联系我们 】

我们致力于为您提供个性化的售前购买咨询服务，及相应的技术售后服务，任何问题请联系 4009100100或 95716。

文档目录

产品简介

产品概述

产品功能

产品优势

应用场景

产品简介

产品概述

最近更新时间：2025-08-05 17:40:41

什么是文本内容安全产品

文本内容安全（Text Moderation System，TMS）是一款文本内容智能识别产品，帮助企业对用户上传的文本进行内容安全分析与识别，具有高识别准确率、高召回率，能够多维度覆盖不同类型的文本内容，提供**文本安全内容服务与文本 AI 生成识别服务**，助力企业解决内容安全问题。

文本内容安全服务

产生背景

随着互联网、智能设备及各种新形态业务的快速发展，互联网数据呈指数式增长，其中也充斥着各种不可控的风险因素，例如：色情内容和垃圾广告等。文本内容安全成为 UGC 平台、在线教育、新闻传媒等行业急需的能力。然而，传统手段审核成本较高，企业内容运营面临新的技术挑战。

能力介绍

- 能够对文本文件进行多样化场景检测，精准识别文本中出现可能令人反感、不安全或不适宜的内容，有效降低内容违规风险与有害信息识别成本。
- 能够精准识别涉黄等有害内容，支持用户配置词库，打击自定义的违规文本。文本内容安全服务能检测内容的危险等级，用户可提前设置文本内容识别结果的自动化处理策略，从而帮助用户节省人力，降低业务风险。

服务方式

- 直接通过控制台提供服务：用户在开通服务后，通过在 **内容安全控制台** 界面的操作，即可完成检测并返回结果，适合**小批量**的快速任务查询。
- 通过开放 API 提供服务：用户通过调用 API 即可获取识别结果，高效构建智能化业务系统，提升业务运营效率。

文本 AI 生成识别服务

产生背景

随着生成式 AI 技术的快速发展，AI 生成或改写的文本已深度渗透进教育、小说、传媒等场景，例如：AI 论文、AI 小说、AI 假新闻等。这类内容凭借批量生产、主题定制化等技术特性，进一步加大了传统内容审核手段的应对压力。因此，凭借多年深耕内容安全领域的业务经验，产品团队于2025年新推出了文本 AI 生成识别服务，为有 AI 鉴别需求的企业或个人提供高质量的一站式服务。

能力介绍

- 能够对文本文件进行多样化场景检测，精准识别文本内容是否存在 AI 工具生成或篡改的痕迹，有效降低内容的 AI 生成风险与判断识别成本。
- 文本AI生成识别服务将会持续更新，以适应不断迭代升级的 AI 文本生成模型，强化服务对于不同大语言模型的覆盖能力。

服务方式

- 直接通过控制台提供服务：用户在 [生成识别页面](#) 开通服务后，简单操作页面即可完成检测并获取结果，适用于小批量快速任务查询。
- 通过开放 API 提供服务：用户调用 API 即可获取识别结果，高效构建智能化业务系统，提升运营效率。

产品功能

最近更新时间：2025-05-14 17:34:42

色情内容识别

色情内容是指涉及性行为、性器官描述等低俗文本。本服务能识别文本中的性暗示词汇、描述性行为的语句、隐晦色情表述等，保障平台内容健康。

暴力内容识别

暴力内容是指威胁、恐吓、暴力行为等文本描述。本服务能判断文本是否具有攻击性或煽动暴力意图、是否涉及血腥虐待等暴力情节等，防止不良内容传播。

广告内容识别

广告内容是以推广商品、服务、品牌为目的的文本。本服务能识别商品名称、品牌标识、促销信息、线下导流等特征词汇，结合语义分析判断广告营销意图，保障平台的内容生态纯净。

谩骂内容识别

谩骂内容是使用侮辱性、攻击性语言进行人身攻击的文本。本服务能识别谩骂类词汇，并利用自然语言处理技术理解语义和情感倾向，精准识别隐晦谩骂，维护网络社交文明。

AI 生成识别

AI 生成是指使用 AIGC 类工具生成的文本内容，例如 AI 写小说、AI 写论文等。本服务通过识别文本在形式（如标点、特殊字符、Markdown 格式）和内容（如泛泛而谈、正确的废话）这些方面所呈现出的特征，来鉴别文本是否为 AI 工具生成。

自定义识别

支持用户自定义中文和英文关键词库，精准识别包含自定义关键词的违规内容。您无需训练机器识别模型，系统会自动将识别内容与自定义库内样本进行匹配，对自定义内容进行定向识别，满足不同场景下的多样化识别需求。

产品优势

最近更新时间：2025-08-05 17:40:41

高可靠性

- 业务可用性不低于99.9%，专业团队7 × 24小时实时提供技术支持。
- 请求毫秒级响应，结果秒级返回，超低延迟助力您的业务“快人一步”。
- 多集群部署，每秒超万级并发，支持动态扩容，无需担心性能损耗。
- 融合数十种算法技术构建综合识别模型体系，避免单一模型误判。

高灵活性

- 文本内容安全无需腾讯云环境即可使用，不涉及业务迁移成本。
- 文本内容安全无需安装任何脚本文件，通过 API 接口即可直接调用业务接口，三步轻松接入。
- 文本内容安全服务支持自定义库和自定义识别策略，根据业务不同需求可以灵活配置内容安全服务。
- 文本内容安全服务支持公有云服务和私有云部署，文本 AI 生成识别服务提供公有云服务，省时、省钱、省心。

高可信度

- 腾讯22年产品安全运营经验积累，模型学习样本丰富全面：文本内容安全服务覆盖各个行业数百种违规类型，文本 AI 生成识别服务覆盖国内外知名大语言模型。
- 成功服务互联网、传媒、教育等数十个行业的企业及个人用户，服务的稳定性得到长期验证，获得用户信任与好评。

应用场景

最近更新时间：2025-08-05 17:40:41

文本内容安全服务

互动直播

文本内容安全可提供针对直播中的弹幕、用户评论等文本内容过滤的一站式、低时延解决方案，识别违规可疑的直播房间并进行预警。

如果已使用腾讯云的直播解决方案，即可一键开启文本内容安全服务，及时阻断不良内容传播，降低平台运营风险。

社区论坛

文本内容安全可以广泛应用于博客、论坛等各类有用户原创内容的平台，包括个人主页、评论、发帖、回帖及站内信等场景。快速识别令人反感、不安全或有害内容，保障平台商业利益和业务合规，降低用户运营成本。

电商购物

文本内容安全可以对购物平台的文本内容进行多场景覆盖式识别，包括商品简介、商品详情介绍、买家评价和用户问答等场景。预防涉黄等敏感类文本发布，降低人工识别成本和业务违规风险。

精细化的多级标签及高度定制化的自定义识别策略，保障各类合规商品正常交易的同时，及时甄别各场景下潜藏的不良和有害信息，护航平台业务的正常开展。

文本 AI 生成识别服务

教育领域

文本 AI 生成识别服务可对学校、在线教育 App 场景中的大学生论文、中小学生作文等多样化文本内容进行 AI 鉴别，识别可能由 AI 生成或润色修改过的文本内容并进行预警。

近年来，随着生成式 AI 技术的快速发展，AI 生成文本在作业、论文中滥用的情况增多，学生的真实知识能力水平难以评判。通过对于 AI 文本生成的合理管控，有助于促进学术知识能力成长，保障教育场景的成绩公平与学术诚信建设，契合学生个人、学生家长、学校老师乃至社会各界对于人才培养的高标准期待。

文学小说

文本 AI 生成识别服务可广泛应用于在线阅读 App、小说创作者平台，识别对应的小说内容是否由 AI 创作生成，帮助用户对于大量 AI 生成的低质量内容进行识别与筛选。

近年来，随着生成式 AI 技术的快速发展，作者因 AI 小说抢占激励资源而停更抗议、读者因 AI 小说质量问题退订会员的事件时有发生。通过对 AI 小说的合理管控，有助于平台优化内容展示体系，缓解作者与读者对 AI 小说的顾虑，为内容生态的健康发展提供支持。

传媒平台

文本 AI 生成识别服务可对新闻报道中的文字部分进行分析，识别出该文章是否由 AI 撰写并进行预警。

近年来，随着生成式 AI 技术的快速发展，**AI 写稿**、**AI 洗稿**、**AI 假新闻**等问题对新闻真实性构成挑战。通过对 AI 撰写内容的合理管控，有助于防范此类业务风险，维护新闻内容的可信度。