新华云盾
内容安全审核服务
产品白皮书

新华网（北京）科技有限公司

版本：V2.1　更新时间：2021 年 10 月 9 日

目录

1 产品概述 3

1.1 行业背景 3

1.2 产品简介 3

1.3 应用场景 4

2 产品架构 5

3 新华云盾官网使用说明 6

3.1 官网首页 6

3.2 注册登录 6

3.3 我的云盾页面 7

3.4 创建团队 8

3.5 查看团队列表 9

3.6 编辑团队名称 9

3.7 解散团队 10

3.8 退出团队 10

3.9 完善账号信息 11

3.10 进入管理中心 11

3.11 进入审核平台 12

4 新华云盾管理中心使用说明 13

4.1 总览 14

4.2 切换团队 15

4.3 成员管理 15

4.3.1 邀请成员 15

4.3.2 修改成员昵称 16

4.3.3 移除成员 16

4.4 邀请记录 17

4.5 功能配置 18

4.5.1 文本纠错.....18
4.5.2 文本敏感.....18
4.5.3 图片审核.....18
4.6 词库管理.....18
4.6.1 自定义敏感词.....19
4.6.2 自定义错别词.....19
4.7 数据统计.....20
4.7.1 累计数据查询.....20
4.7.2 人员使用统计.....21
5 新华云盾审核平台使用说明.....22
5.1 稿件审核.....22
5.2 切换团队.....22
6 产品优势.....23
7 服务方式.....24

1 产品概述

1.1 行业背景

人工智能、大数据等互联网前沿技术的快速发展，加快了全球各行业的前进步伐。在内容消费领域，得益于人工智能、大数据的技术赋能，内容生产效率大幅提升，导致互联网信息井喷式增长。而随着移动设备的普及，人们拥有了更多的上网场景，对内容消费和网络发声的意愿和需求也越来越强烈。网络信息技术快速发展的同时，网络信息内容安全成为一个巨大的潜在问题，甚至已经上升为事关国家经济安全、社会稳定的全局性战略问题，是国家安全的重要组成部分。

国家互联网信息办公室于 2019 年 12 月 15 日发布了《网络信息内容生态治理规定》，并于 2020 年 3 月 1 日起正式施行。《网络信息内容生态治理规定》明确指出了网络信息内容生产者禁止触碰的十条红线、应当防范与抵制八类不良信息。目的是建立健全网络综合治理体系，营造清朗网络空间，建设良好的网络生态环境。

内容安全对内容生产者、内容监管者、网站运营者都是一个非常重要的问题，不容忽视。除了对内容生产有更严格更专业的要求外，对内容校对审核的效率也有更高的要求。因此，内容安全需要更权威专业、更智能高效的技术支撑，帮助内容生产者一同建设良好的网络生态环境。

1.2 产品简介

新华云盾是依托新华网在新闻媒体领域多年的采编报道经验以及海量专业的新闻大数据，基于人工智能前沿技术打造的权威智能内

容安全审核服务。可对稿件内容进行政治敏感、不良有害信息、报道规范、文字错误、人物识别等维度的内容审核，为内容安全保驾护航。

新华云盾可根据需求灵活配置审核维度，除了提供实时内容审核服务外，还可以对已发布的稿件内容进行批量扫描。另外，新华云盾提供了团队管理功能，能够帮助团队管理者有效提升团队管理效率，提高团队绩效水平。新华云盾通过持续的产品深耕和技术迭代，面向政企、媒体、高校、自媒体等行业客户提供更加权威专业的内容安全审核服务。

1.3 应用场景

新华云盾支持以下场景使用：

● 在线审核平台。新华云盾为客户提供“审核+管理”一体化的审核平台，包括供用户使用的在线审核平台，以及供团队管理者使用的管理中心。

● 内容安全审核报告。对已发布的内容提供批量内容安全扫描服务，并形成内容安全审核报告，详细指出存在的内容隐患。

● API 服务。提供新华云盾内容安全审核底层能力输出，便于接入方更灵活定制产品或接入至其他产品中。

新华云盾将持续优化迭代，除了不断提升服务质量，未来将不断拓展更多的应用场景，满足不同场景的用户使用。

2 产品架构


新华云盾
官网
产品介绍
团队列表
账号信息
审核平台入口
管理中心入口
API管理中心入口
新华云盾
审核平台
文本纠错
文本敏感
图片审核
审核结果展示
团队切换
新华云盾
管理中心
总览
成员管理
邀请记录
功能配置
自定义敏感词库
自定义错别词库
数据统计
团队切换
新华云盾
API管理中心
总览
自定义敏感词库
自定义错别词库
API使用统计
新华云盾
底层能力
文本纠错
文本错误检测
标点、单位错误检测
政治性差错检测
引文对比检测
报道规范检测
自定义词库检测
文本敏感
暴恐检测
色情检测
违禁敏感检测
侮辱检测
政要及敏感人物检测
自定义词库检测
图像识别
色暴恐检测
敏感人物检测

3 新华云盾官网使用说明

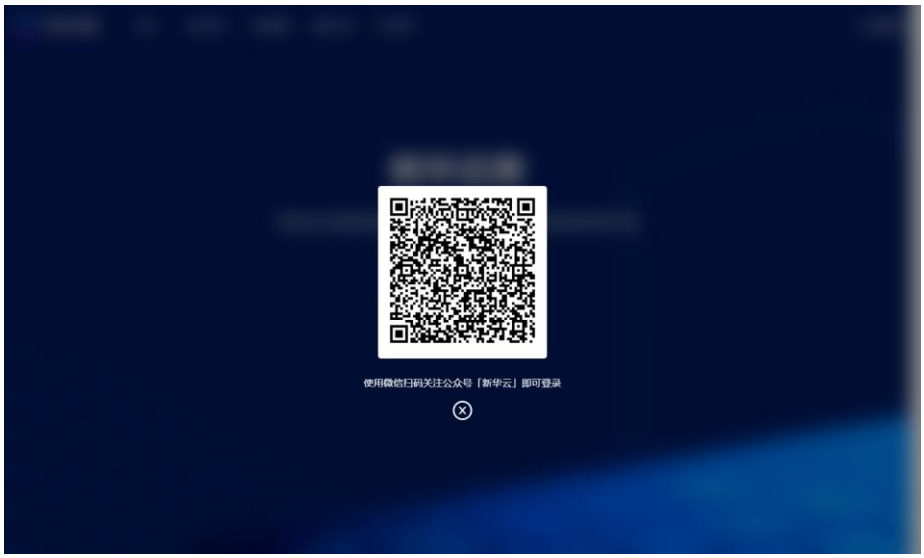
官网地址：https://xhyd.news-tech.cn

3.1 官网首页

官网首页是新华云盾的产品介绍，可了解产品的“核心能力”、“定制选配”、“解决方案”、“产品优势”。

3.2 注册登录

在新华云盾官网点击“立即使用”，通过微信扫码即可完成注册登录。



3.3 我的云盾页面

微信扫码登录后会自动跳转至“我的云盾”页面。

或者也可以在官网首页右上角点击进入。

在“我的云盾”可查看当前账号状态，根据账号类型可进行相应的具体操作，具体如下：

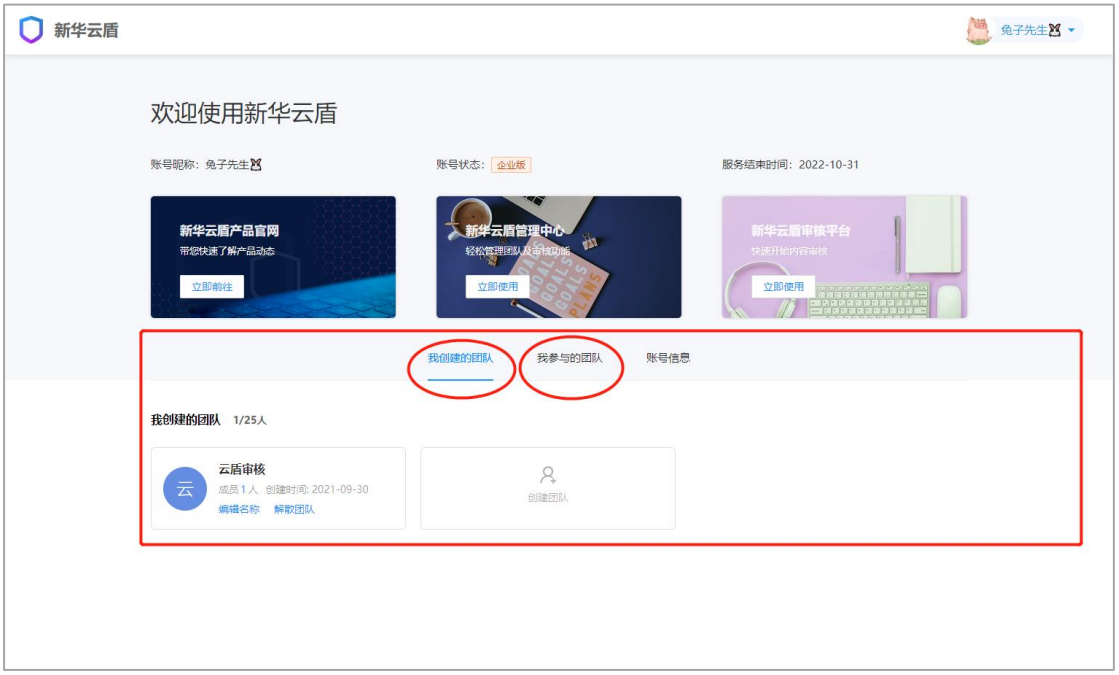
账号类型	创建团队	加入他人团队	管理中心	审核平台
免费版	否	是	成为他人团队管理后可进入	加入他人团队后可进入
标准版	是	是	可进入	可进入
企业版	是	是	可进入	可进入
旗舰版	是	是	可进入	可进入

3.4 创建团队

在“我的云盾”页面里，可点击“创建团队”即可创建一个新的团队。免费用户不可创建团队。

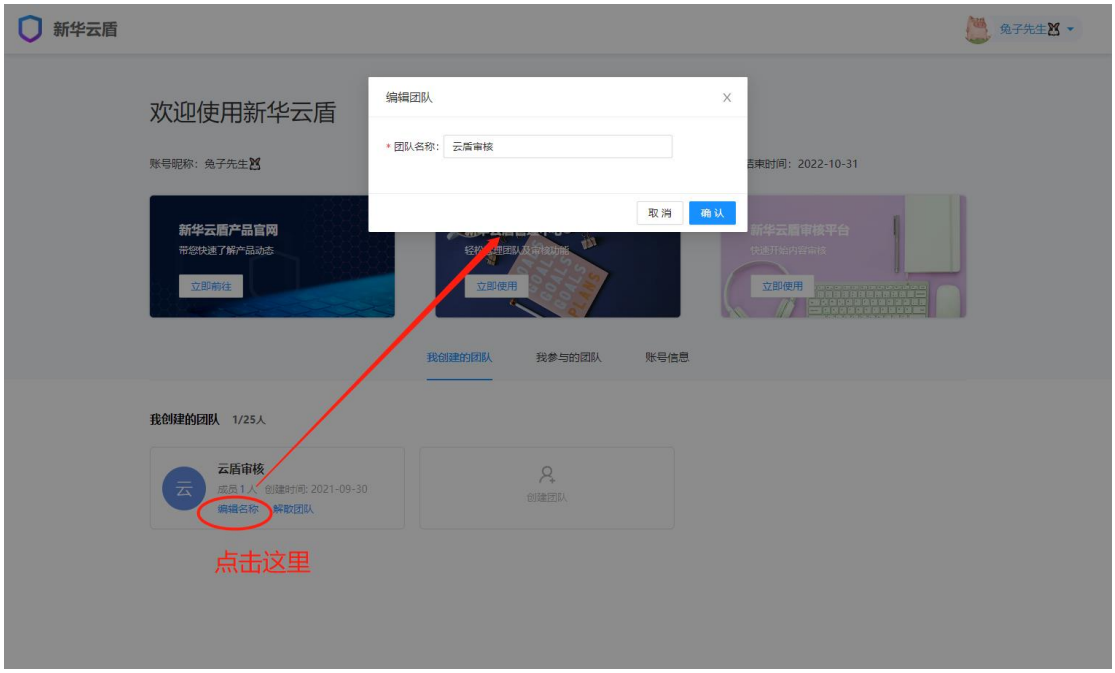
3.5 查看团队列表

在“我的云盾”页面里，可点击“我创建的团队”“我参与的团队”查看相应的团队列表。



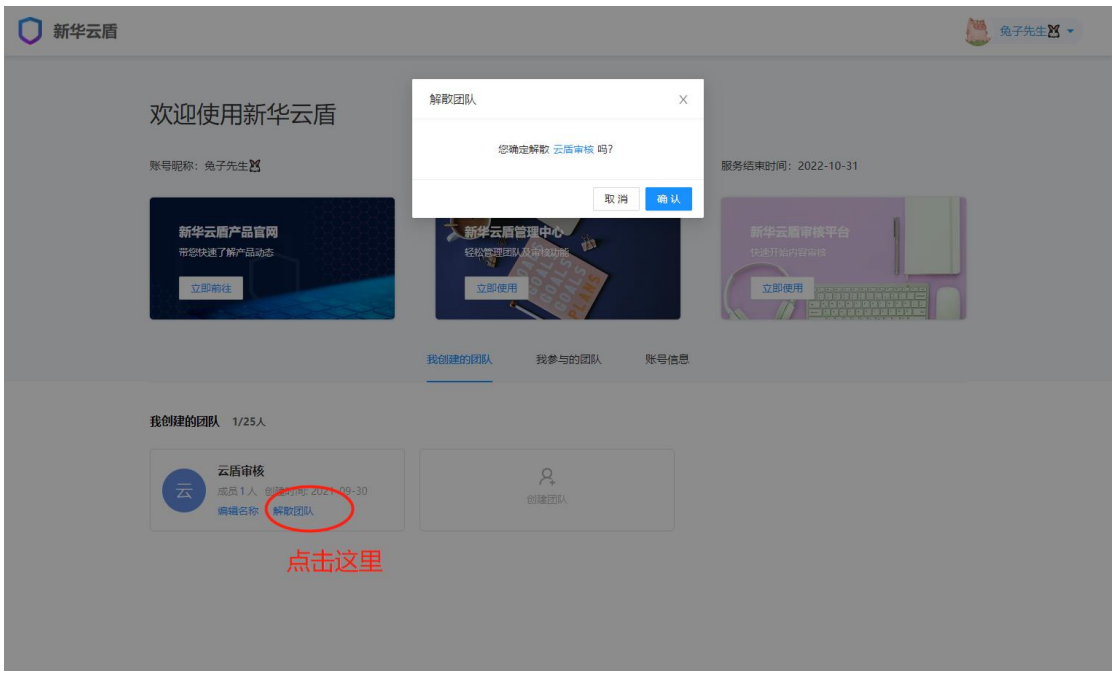
3.6 编辑团队名称

点击团队列表中的“编辑名称”，即可修改团队名称。该功能仅限团队创建者使用。



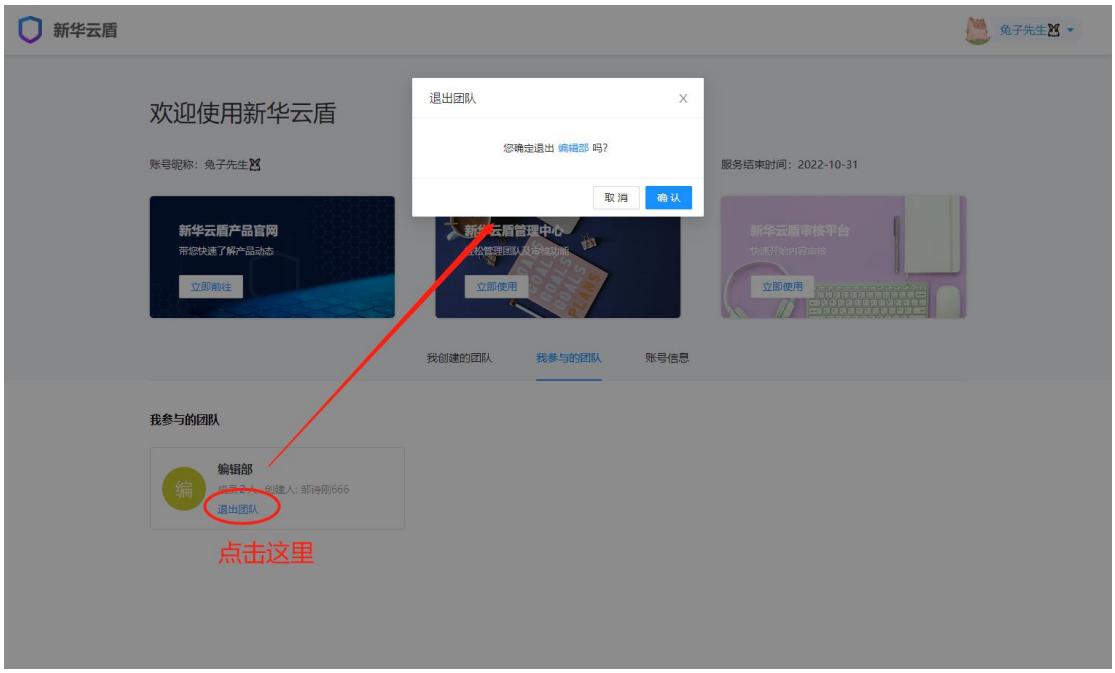
3.7 解散团队

点击团队列表中的“解散团队”，确认后即可解散团队。团队解散后，原团队成员将不可使用该团队的管理中心和审核平台，团队解散操作不可恢复。该功能仅限团队创建者使用。



3.8 退出团队

团队成员可自行退出团队，在团队列表中点击“退出团队”即可。



3.9 完善账号信息

用户可完整个人账号信息，便于为您提供更优质的产品服务。

3.10进入管理中心

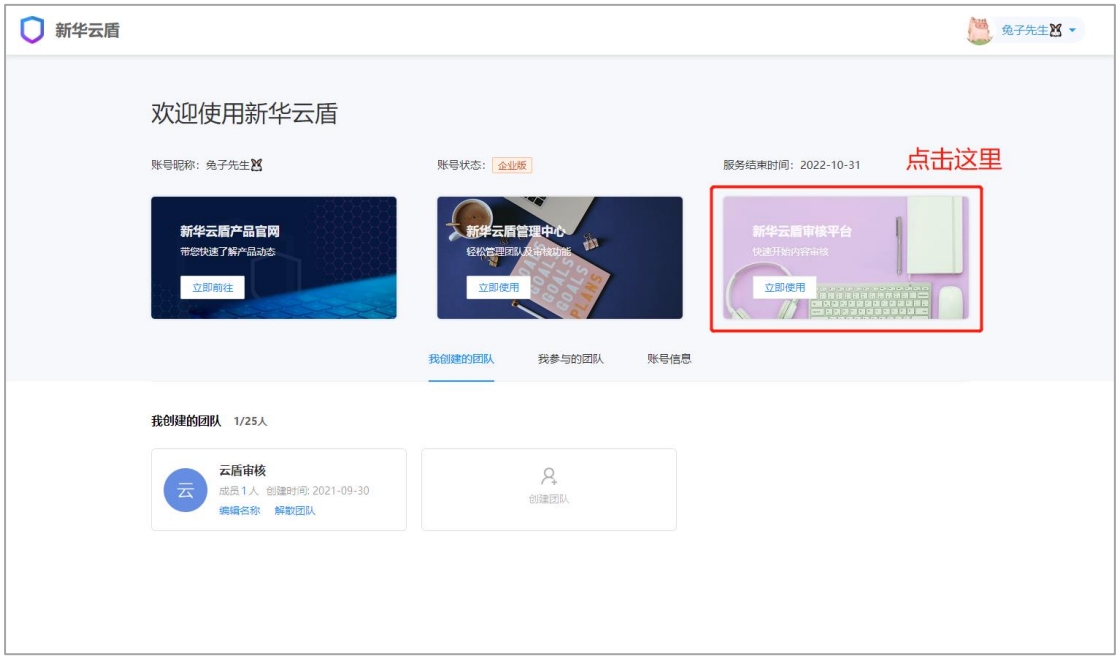
可从“我的云盾”页面入口进入。

也可从以下网址扫码进入：

https://xhyd.news-tech.cn/csadmin/

3.11进入审核平台

可从“我的云盾”页面入口进入。



也可从以下网址扫码进入：

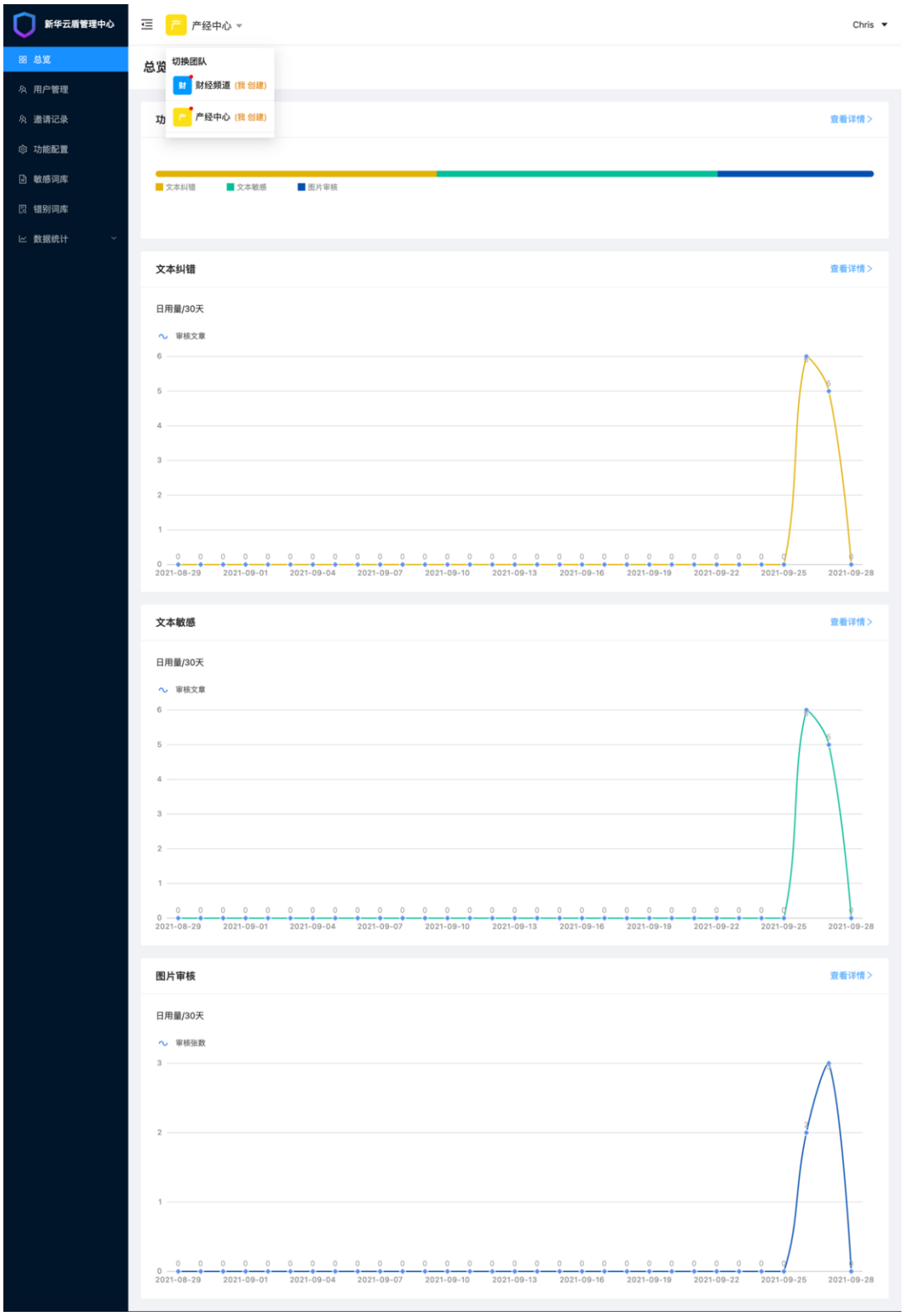
https://xhyd.news-tech.cn/workbench/

4 新华云盾管理中心使用说明

新华云盾管理中心仅限团队创建者和团队管理员可以进入。

4.1 总览

主要包含各维度功能使用量、文本纠错文章审核近 30 日折线图、文本敏感文章审核近 30 日折线图、图片审核近 30 日折线图。



4.2 切换团队

如果从属于多个团队管理者，可在这里切换团队。

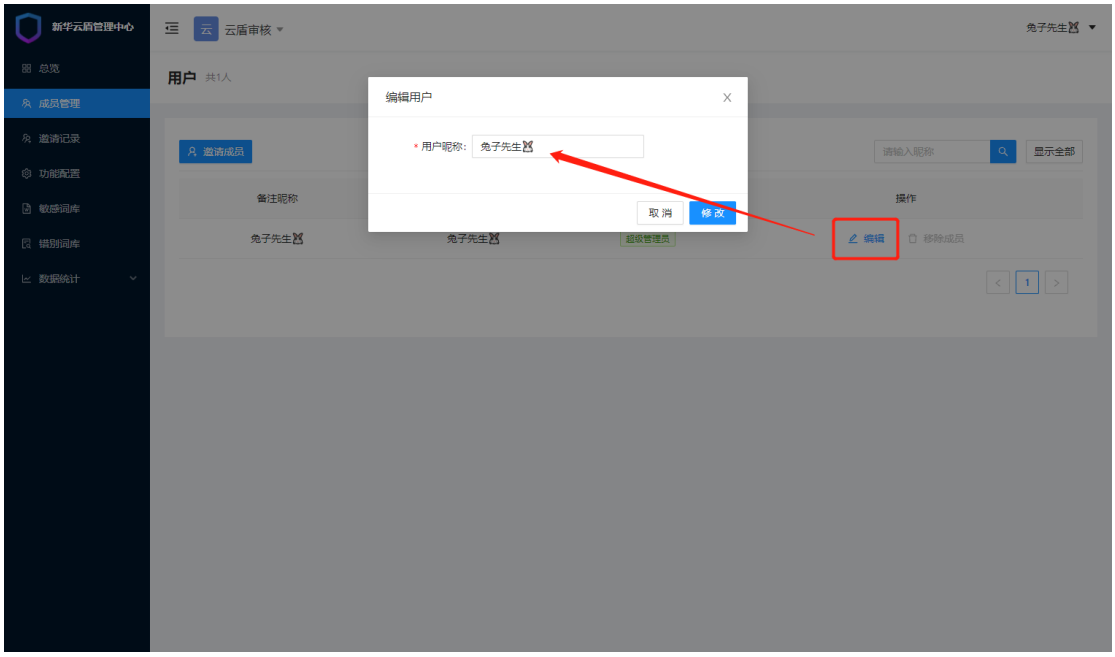
4.3 成员管理

4.3.1 邀请成员

切换至“用户管理”页面，点击“邀请成员”，设置好有效时间，生成邀请链接。将邀请链接发送给其他人就可以邀请其加入该团队。

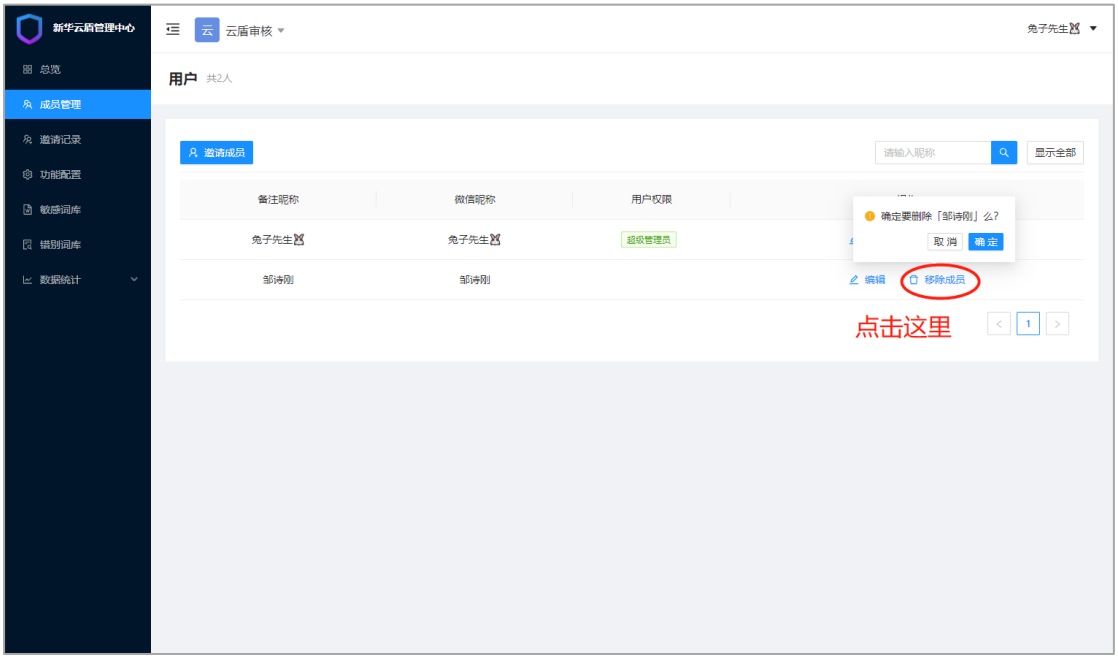
4.3.2 修改成员昵称

可以修改成员在该团队中的昵称。修改的昵称仅在该团队生效。



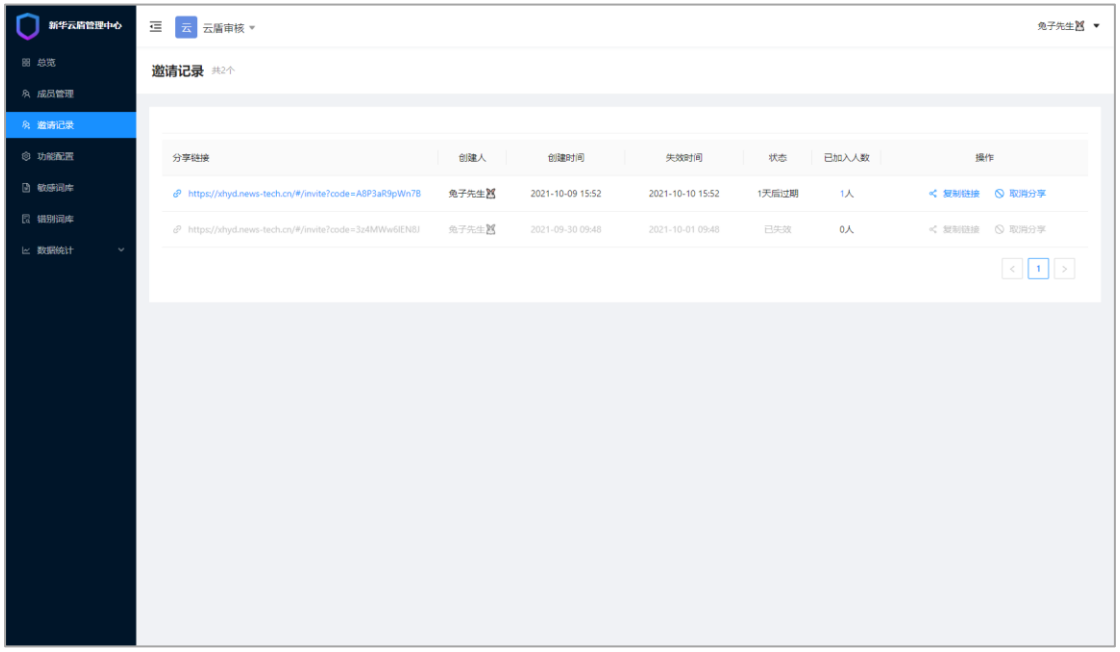
4.3.3 移除成员

点击“移除成员”，确认后即可将该成员移除。

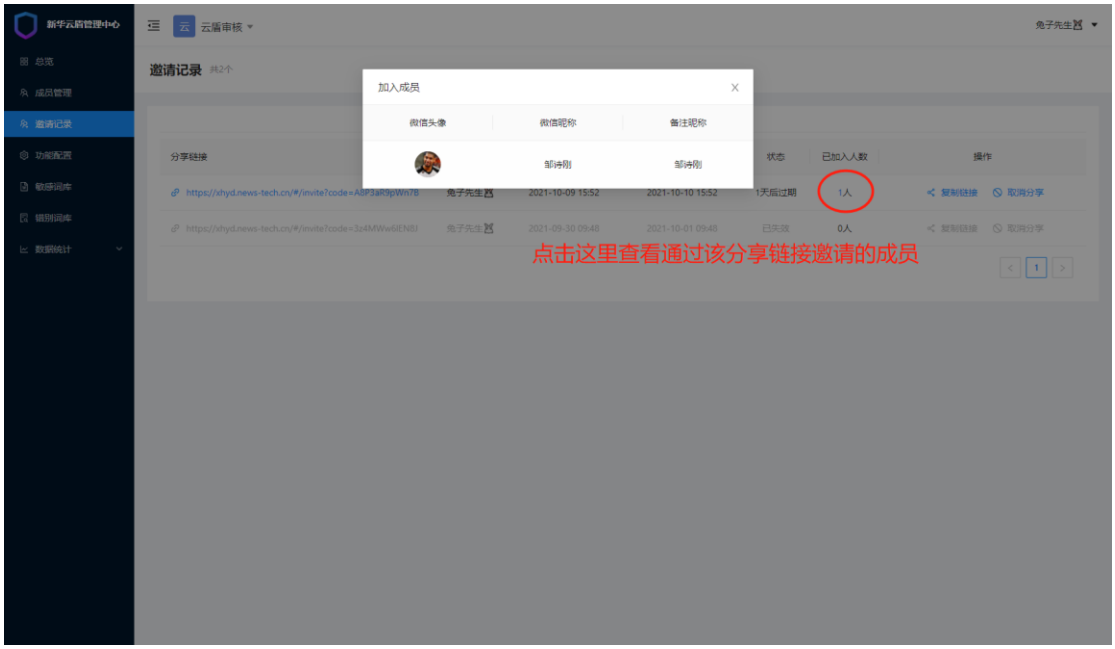


4.4 邀请记录

可查看已经生成的邀请记录。可以复制链接、取消分享，查看创记录状态、创建人、创建时间等。

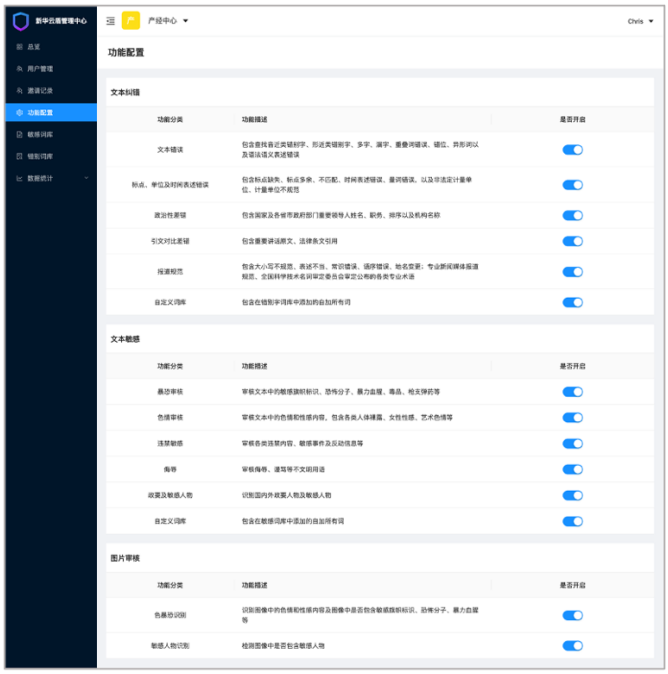


点击“已加入人数”，可以查看通过该链接邀请的成员。



4.5 功能配置

功能配置模块支持文本纠错、文本敏感、图片审核的具体功能配置，用户可根据自身需求自由开启或关闭功能开关。另外，在文本纠错和文本审核里支持用户自定义添加的敏感词和错别字词审核。



4.5.1 文本纠错

包含文本错误、标点单位错误、政治性差错、引文对比错误、报道规范、自定义词库（即在错别字词库中添加的自加所有词）。

4.5.2 文本敏感

包含色暴恐审核、色情审核、违禁敏感、侮辱、政要及敏感人物、自定义词库（即在敏感词库中添加的自加所有词）。

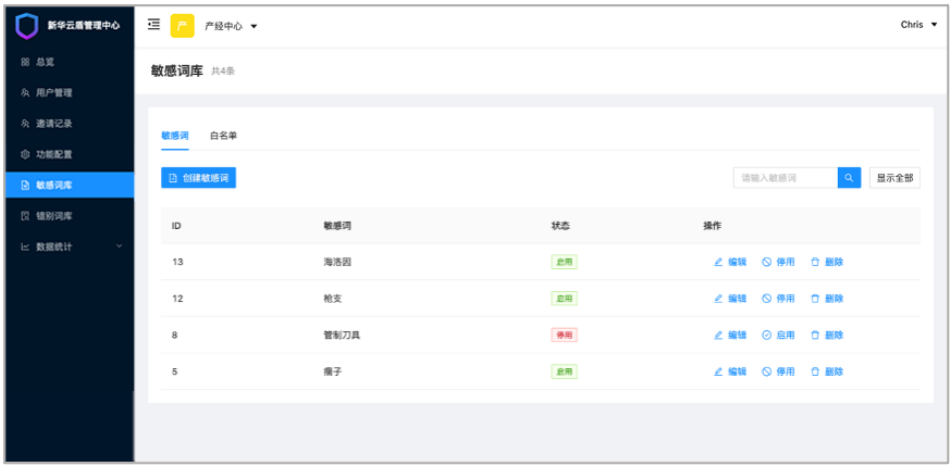
4.5.3 图片审核

包含色情识别、暴恐识别、敏感旗帜标志识别、政要及敏感人物识别。

4.6 词库管理

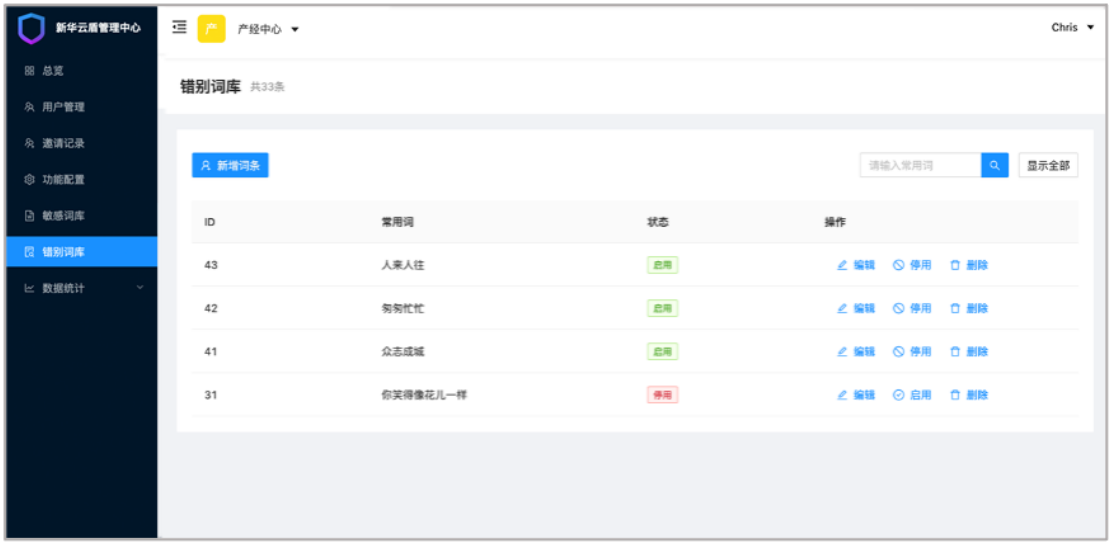
4.6.1 自定义敏感词

敏感词库支持用户自定义添加自有敏感词，添加成功后继续执行审核任务，将自动识别该敏感词。点击列表区的“新增敏感词”按钮，出现编辑弹层窗口，可支持添加敏感词信息，点击“新增”按钮保存之前操作。列表区“搜索框”支持对敏感词进行关键字搜索。点击“停用”按钮，操作成功后继续执行审核任务，将自动过滤该敏感词。点击“删除”按钮，操作成功后将永久删除该词，此操作不可撤销。



4.6.2 自定义错别词

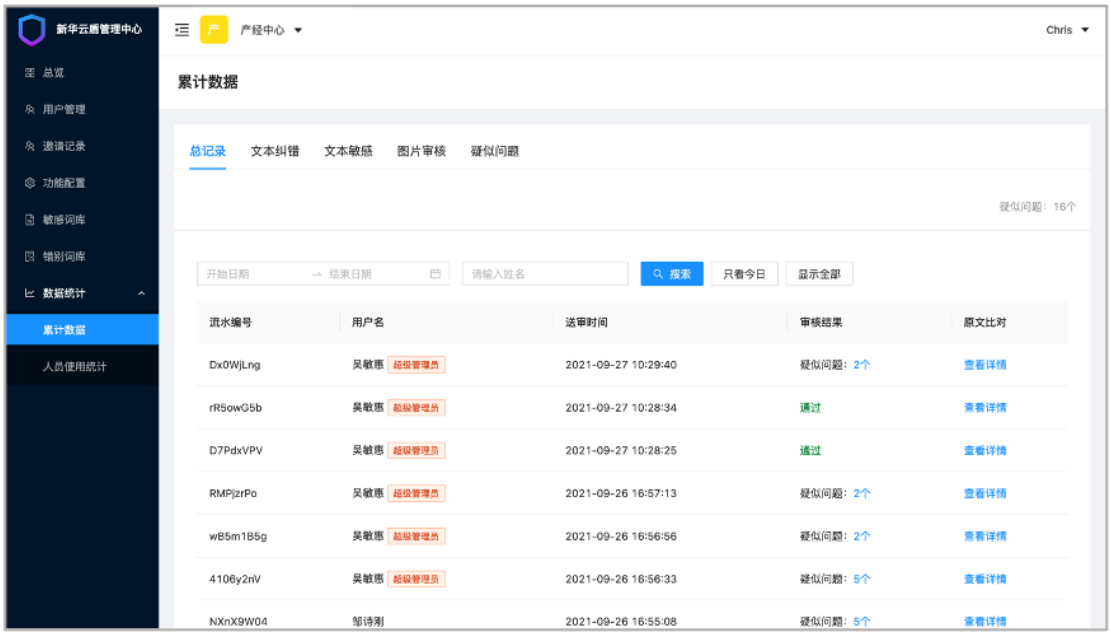
错别词库支持用户自定义添加自有敏感词，添加成功后继续执行审核任务，将自动识别该错别词。点击列表区的“新增错别词”按钮，出现编辑弹层窗口，可支持添加敏感词信息，点击“新增”按钮保存之前操作。列表区“搜索框”支持对错别词进行关键字搜索。点击“停用”按钮，操作成功后继续执行审核任务，将自动过滤该错别词。点击“删除”按钮，操作成功后将永久删除该词，此操作不可撤销。



4.7 数据统计

4.7.1 累计数据查询

累计数据查询模块支持管理员查看下属所有用户的操作日志数据情况。点击“时间选择”控件可选择起止时间点，按时间周期条件查询用户操作记录。点击“查看详情”按钮查看原文及审核纠错结果对比情况。列表区可查看累计疑似问题统计折线图。新增“只看今日”按钮，可快速搜索并查看当天成员使用情况一目了然。



4.7.2 人员使用统计

可按照周期搜索成员使用情况，支持按姓名搜索快速查找人员。点击“查看详情”可查看该成员周期折线图和审核文章记录列表。

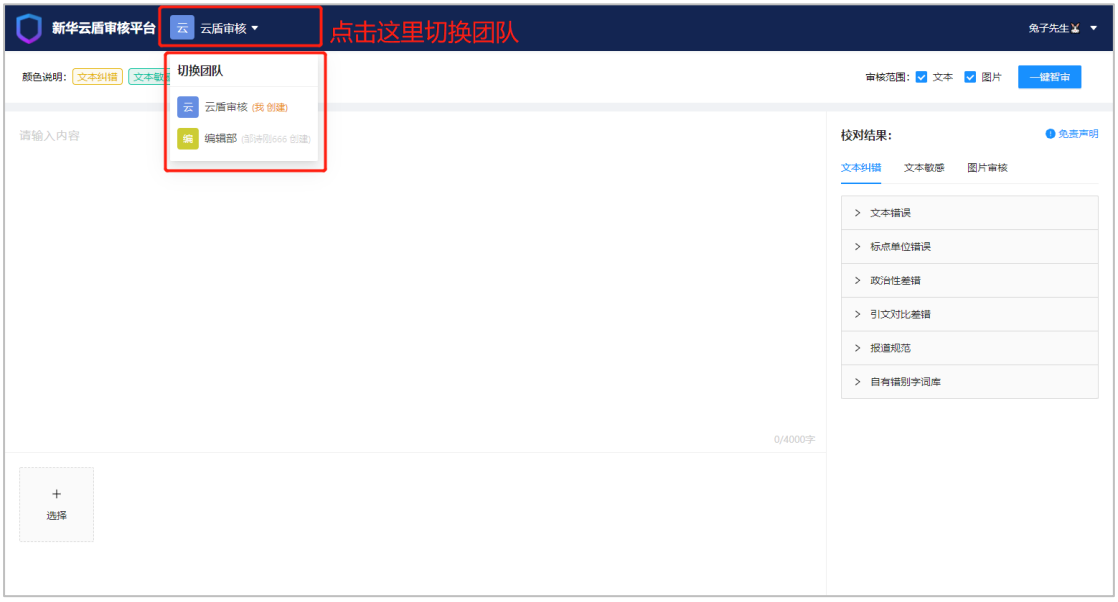
5　新华云盾审核平台使用说明

5.1 稿件审核

可通过输入或粘贴文本、上传图片的方式进行审核。包括文本纠错、文本敏感 、图片审核，审核结果会自动进行高亮显示。

5.2 切换团队

如果从属于多个团队，可在这里切换团队。



6 产品优势

● 新华网权威专业新闻媒体行业经验

依托新华网在新闻媒体领域多年的采编报道经验以及海量专业的新闻大数据，基于人工智能前沿技术算法模型打造的权威智能内容安全审核服务。

● 安全稳定

专业安全运维团队保障服务稳定运行，及时响应用户需求。

● 海量训练语料

基于新华网海量历年稿件训练的模型及语料。

● 持续的产品迭代更新

提供持续更新词库服务，提供云服务升级。

● 灵活调整审核维度

可灵活调整安全审核维度，满足用户不使用场景。

● 自定义词库

支持用户自主构建自定义词库，个性化调整内容审核粒度，满足不同行业用户个性化需求。

7 服务方式

支持以下年服务方式：

账号类型	加入团队数	创建团队	创建团队 总人数	审核次数
免费版	不限	不可	0	不限
标准版	不限	允许	10	不限
企业版	不限	允许	25	不限
旗舰版	不限	允许	无限	不限