

AI多媒体技术在内容审核场景实践探索

马金龙 趣丸科技 (TT语音)

马金龙 多年媒体算法开发经验，涉及音视频图像文本，负责过音频前后端处理，弱网优化，音视频质量提升，智能内容安全审核“T网”，内容理解“T悟”等大型项目。曾作为“灵声讯”创始人，参与智能媒体技术自媒体运营和推广。



01

内容审核目前现状与挑战

02

AI多媒体技术实践之路

03

智能内容审核平台案例

04

AIGC内容风控实践

05

未来展望

1.内容审核目前现状与挑战

现状

- 政府监管越来越严
- 用户内容层出不穷
- 违规种类繁多
- AIGC内容不可控

挑战

- 【实时性】需要紧跟政府管控要求
- 【准确性】对花样变体不漏杀不误杀
- 【多样性】违规种类需不同算法解决
- 【未知性】AIGC生成内容不确定且存在知识“幻觉”



2. AI多媒体技术实践之路

自建 OR 第三方?



2. AI多媒体技术实践之路

自建优势：

<div>可管可控</div> <div>  </div>	<div>极速响应</div> <div>  </div>	<div>生态保障</div> <div>  </div>	<div>高效定制</div> <div>  </div>
<p>具备数据血源追踪、问题实时监控、技术辅助运营等风控能力</p>	<p>针对安全，时效等方面推出高响应审核，让内容审核安全高效</p>	<p>通过机审结果多样化处置、账号违规处置等多种手段，保障平台生态安全</p>	<p>推出特殊时期/突发事件的相关定制化，快速响应国家政府的紧急要求</p>

2. AI多媒体技术实践之路

2.1 语音识别

2.2 NLP文本审核

2.3 多模态识别

2.4 音频事件检测

2.5 小语种识别

2.6 歌曲识别

2.7 声纹识别

2.8 违规图像识别

T网 是一个通过人工智能的算法打造一站式内容安全机器审核的平台，帮助公司审核团队实现语音，文本，图像，小视频等风险管控的能力。

对于此项目的目的可总结如下：

- **贯彻国家网信办有关网络内容安全的各项规定**
- **低成本高效率**的加强内容风险管控
- **构建智能审核技术护城河**，为公司内容生态保驾护航
- **探索内容审核新方法**，践行公司的社会责任

2.1 ASR-技术方案

技术目标

用户产生的语音数据输入ASR模型，模型输出该语音的文字内容，以供下一环节NLP检查是否包含违规词，或违规内容。

模型总体逻辑

使用深度学习模型Transformer/Conformer (如图中Shared Encoder) 提取输入音频中的特征使用CTC解码得到若干候选文本。

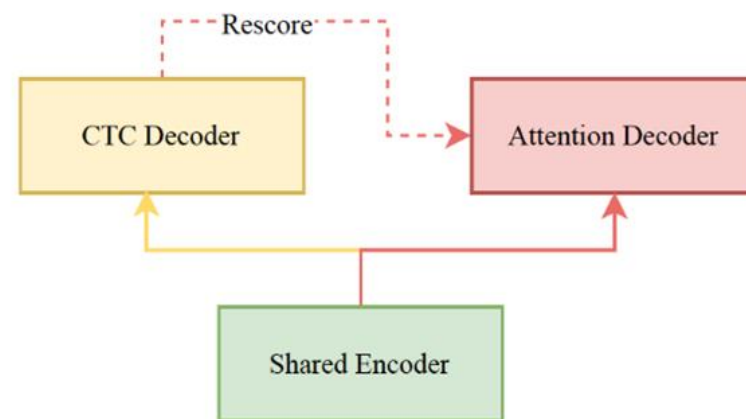


图1. T网-ASR端对端方案

2.1 ASR架构

Efficient Conformer

- Convolution neural networks和transformers models组合
- Efficient Conformer设计
- 结合量化剪枝和蒸馏技术，压缩模型大小
- 提供CPU和GPU，支持高吞吐量识别



图2. T网-ASR支持的功能

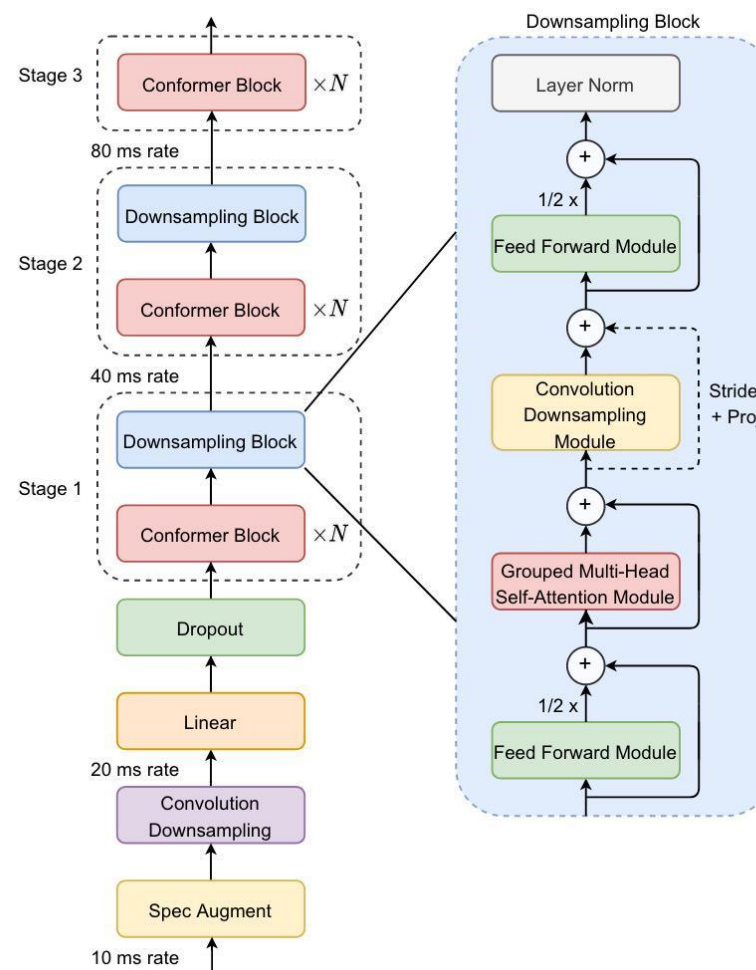


图3. ASR中Efficient conformer技术方案

2.1 ASR-效果

Audio Duration	V1.0.3	V4.0.0
	Inference Time	Inference Time
10	1.12	0.58
20	3.57	1.29
30	6.23	2.15
60	18.64	4.89
80	25.79	7.13
90	30.19	8.38

图4 T网-ASR优化后的推理速度

Model	Aishell	Aidatatang	MagicData	Thchs
V1.0.1	4.01%	3.75%	2.89%	6.89%
V1.0.3	3.14%	3.44%	2.65%	6.36%
V4.0.0	2.75%	2.99%	2.27%	5.77%

Model	Storage	Memory
V1.0.1	478M	~790M
V1.0.3	187M	~790M
V4.0.0 + ONNX + Quant	93M	~397M

图5. ASR 测试报告 (CER)和模型大小

2.2 NLP算法总体框架

NLP算法模型：

- Bert 算法
- Prompt 算法
- Fasttext 算法
- AIGC 语料生成算法
- 文本表情复杂表示的多模态识别算法
- 关键词挖掘算法

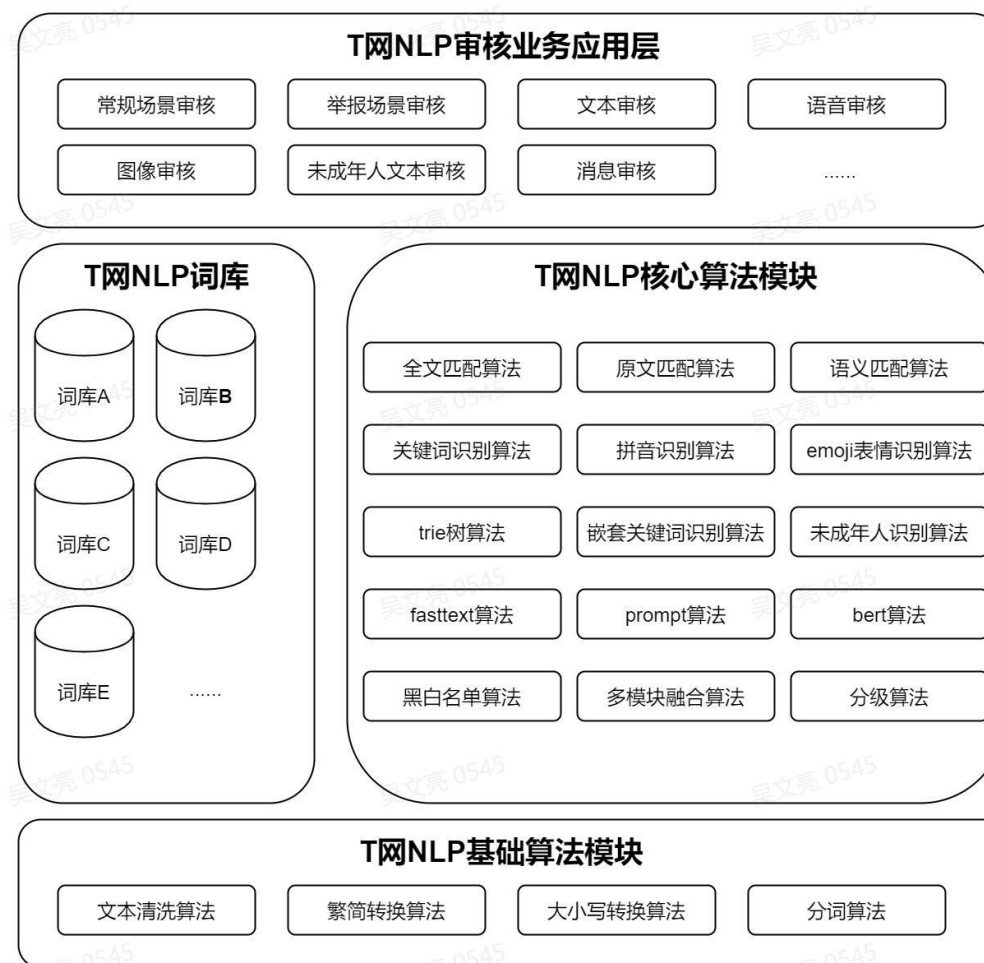


图6. T网-NLP总体框架

2.2 NLP内容审核的困难与挑战

纯文本审核面临的问题与挑战:

- 变体关键词的多样化
- 文字与表情包的结合的复杂表达
- 文字与字母或字母缩写结合的复杂表达
- 特定场景语料不足与稀疏性
- 特定关键词词的隐晦表达
- 正常词与关键词相同，但不同上下文上语义不同

我们的成功案例:

- 构建变体关键词挖掘系统
- 构建文本表情字母多模态识别系统
- AIGC语料生成系统
- 异常关键词大数据监测系统
- 多层次语义分析系统

2.2 NLP内容审核-效果呈现

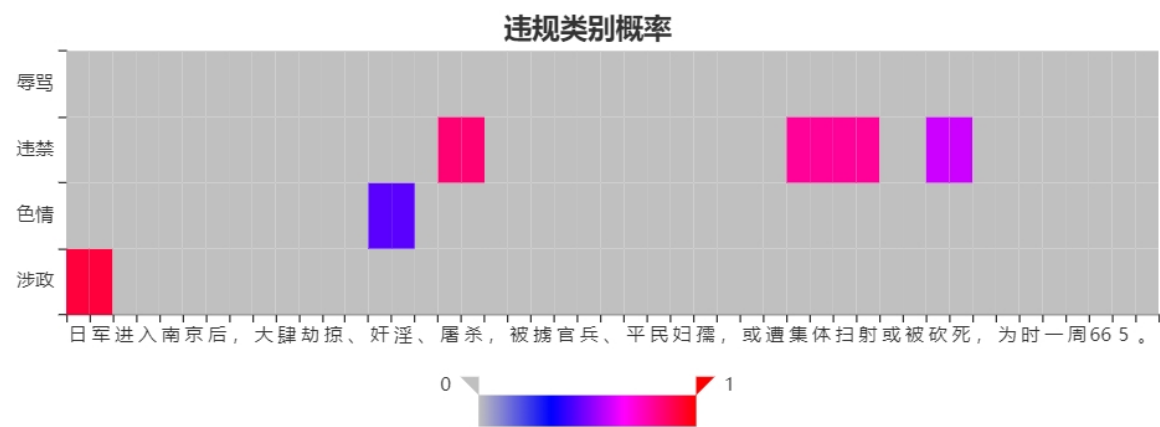


图7. NLP关键词挖掘示意图

违规标签	精确率
辱骂	94.45%
色情	95.03%
涉政	91.31%
广告	90.96%
违禁	92.98%

图8. NLP文本审核效果

2.2 文本未成年人识别

关键词匹配分析框架，支持多种匹配方式、多种过滤条件，并支持自定义特殊标记，及支持特定动作行为，将未成年人识别实现模组化的流程分析。**未成年识别精确率99%+；**

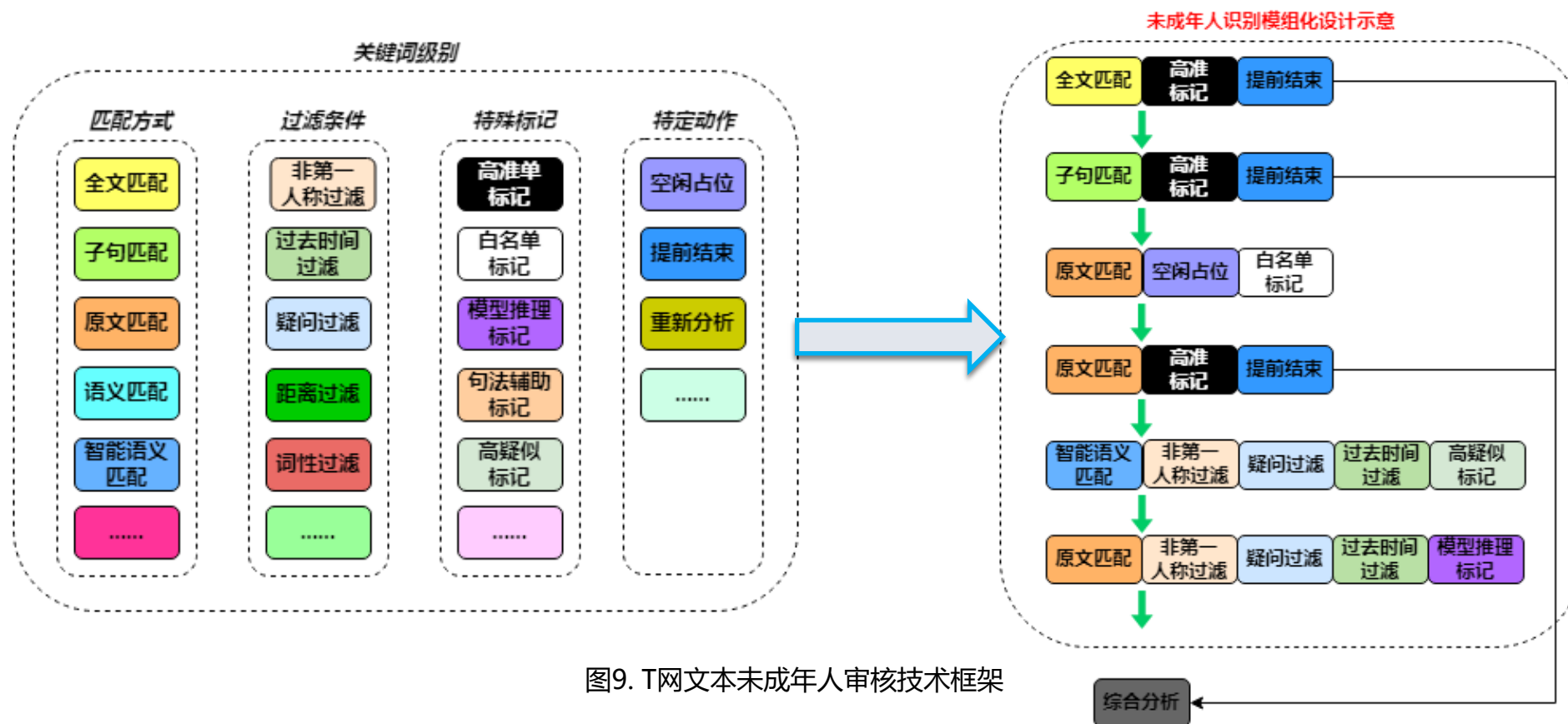


图9. T网文本未成年人审核技术框架

2.3 多模态算法原理

项目背景

- 单模态审核特征不全面，多模态结合语气和语义信息可提高处罚有效率。
- 人工审核量级大，需要对不同类型的违规进行灵活处置。

建模算法

- Transformer 跨模态多头注意力机制；
- 随机森林；

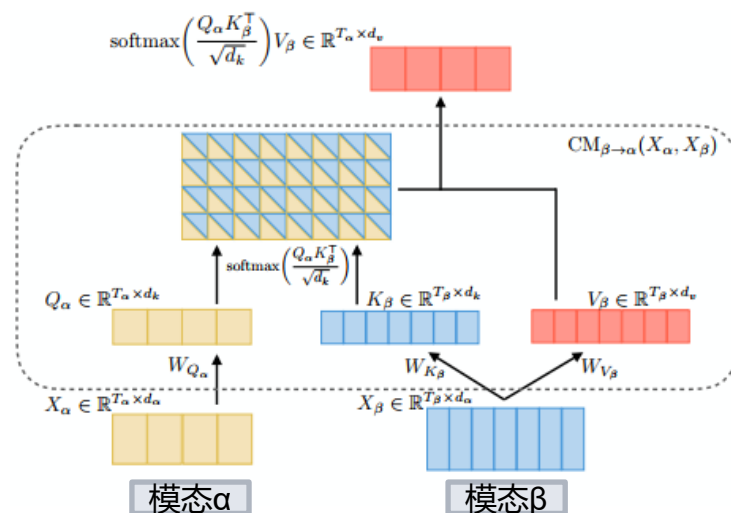


图10. Transformer 跨模态多头注意力机制

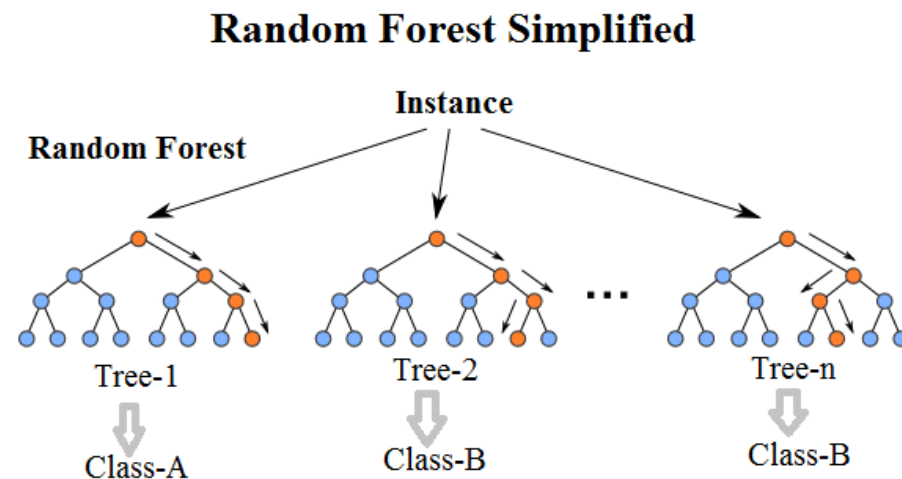


图11.随机森林

2.3 多模态高准召回

模型指标

- 多模态算法上线处罚有效率为**99%+**;
- 如右图, 每日占总机审违规样本约**17%**;

模型价值

- 提高对违规样本的召回, 减少单模态的漏召;
- 提供高准标签运用在灵活处置:
 - a. 提高处罚响应速度;
 - b. 提升人工审核效率;

辱骂多模态命中数量及其占比

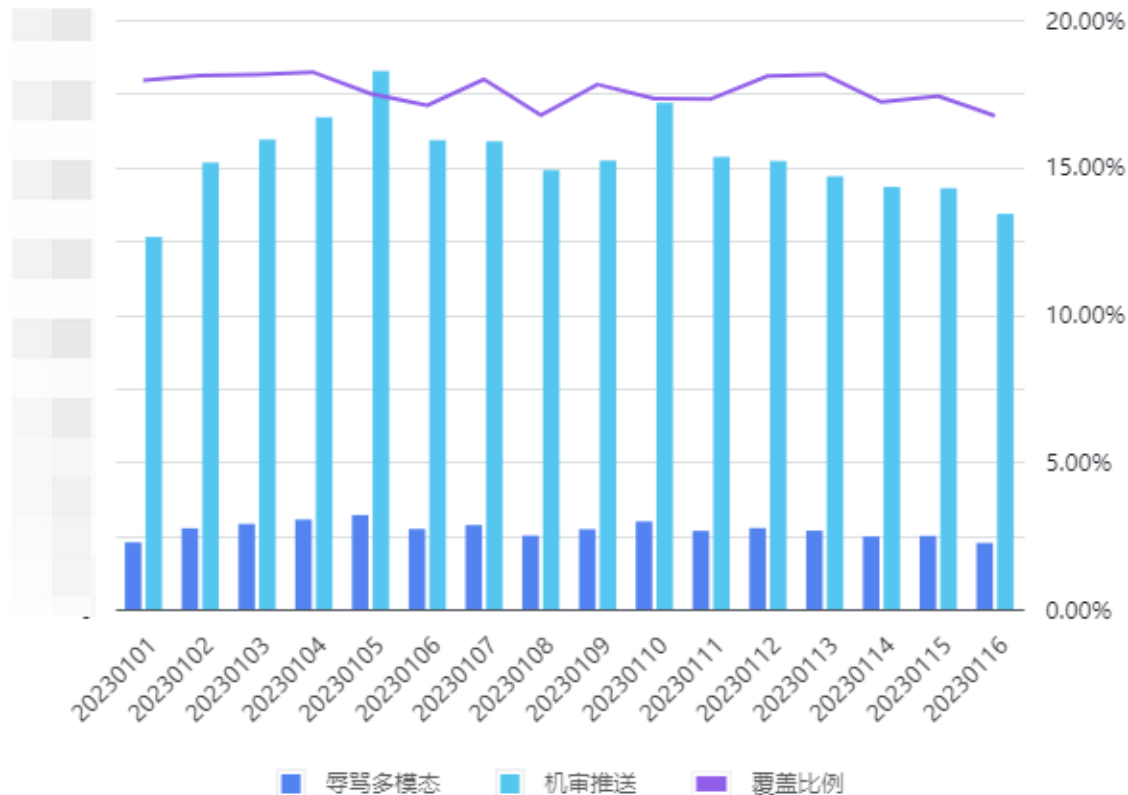


图12 .多模态辱骂命中数量及占比

2.4 声音事件检测 (Sound Event Detection)

检测的声音事件：

- 审核类
 - 娇喘, 炸房, 怒骂
- 普通标签 BRaSS
 - 背景音乐(BGM, B)
 - 说唱(Rap, Ra)
 - 说话(Speech, S)
 - 唱歌(Sing, S)

模型价值

- 完善对声音类违规的审核能力。
- 音频类型分流, 降低后续模型成本。
- 语音直播趋势分析。

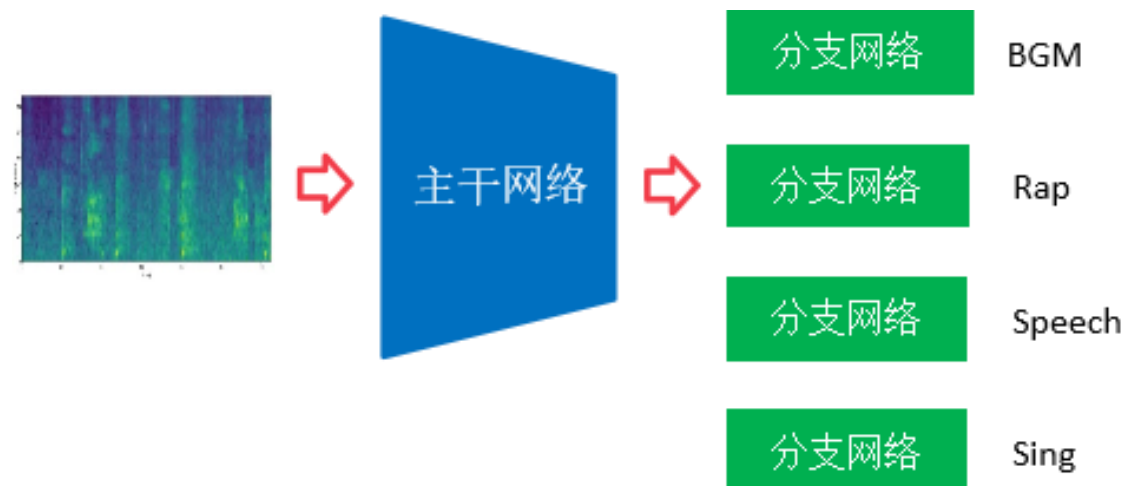


图13. 声音事件检测

2.5 语种识别

项目背景:

线上特定语种管控

方案流程:

利用音频预训练hubert模型的特征解析功能，结合TT语音线上直播数据和部分开源数据集进行模型fine-tune，从而针对特定语种等进行识别。

模型效果:

针对特定语种的测试精准率为**97.58%**。

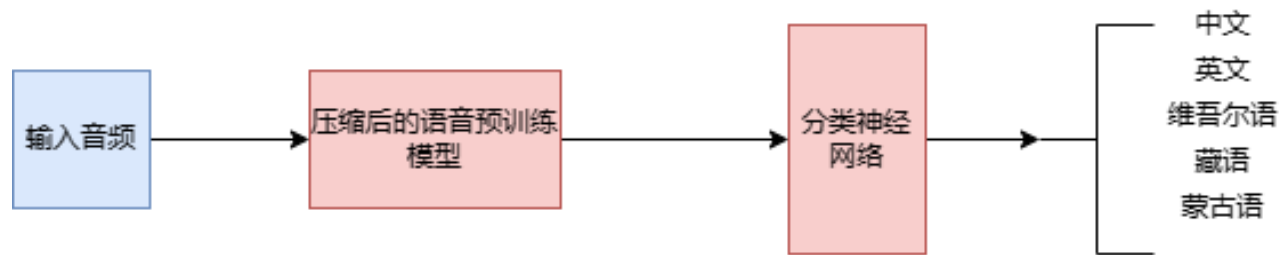


图14. 语种识别

2.6 歌曲识别

项目背景:

线上歌曲(如劣迹艺人作品等)管控

方案流程:

将原始劣迹歌曲处理得到的指纹信息存储于歌曲指纹库，用于进行输入歌曲片段的相似度比对，并增添音频文件分析接口用于分析完整歌曲。

模型效果:

针对劣迹艺人歌曲的识别精准度为94.16%;

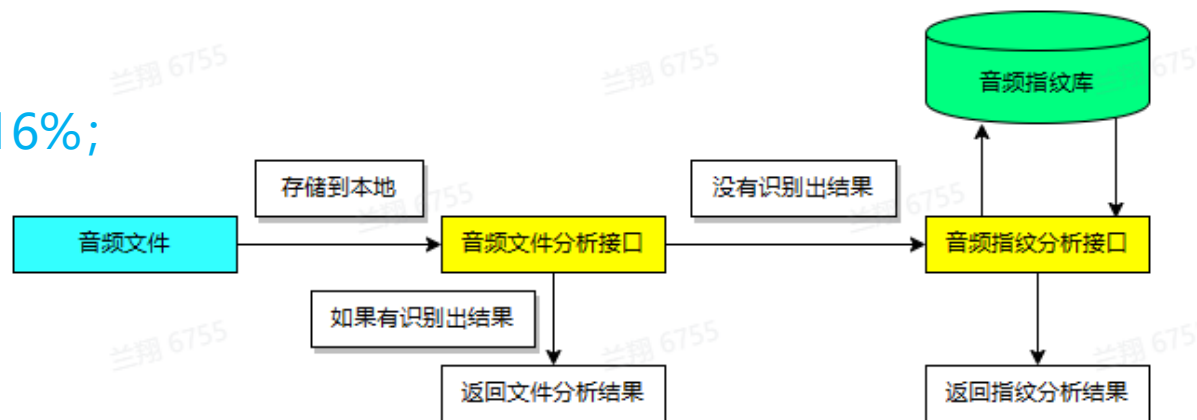


图15. 歌曲识别

2.7 声纹识别

项目背景：

人物声纹识别，针对特定的人物可以做具体管控

方案流程：

- VAD进行语音活性检测，提取人声部分；
- ResNet34作为主干网络，利用线上业务数据和部分开源数据进行微调训练；
- 利用余弦相似度计算两个声纹之间的相似性。

模型效果和应用：

1. 特定人物声纹拦截精确率98%+；

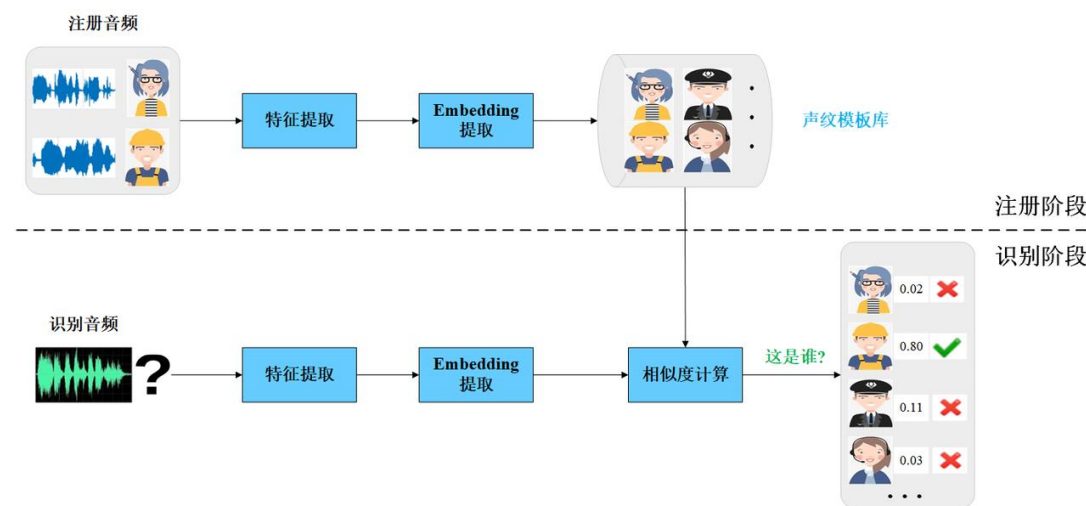


图16. 声纹识别

2.8 涉黄图像识别

项目背景

线上色情、性感类涉黄图像识别

方案流程

- 基于经典ResNet50预训练模型结构，利用线上业务数据和部分开源数据进行微调训练；
- 同时考虑到标注成本和线上标签数据形态，结合多任务图像识别算法更改模型结构进行学习，从而实现较为精准地识别涉黄图像；

模型效果和应用

- 在TT语音下，机审拦截内容识别准确率为**93.15%**；
- 应用于TT语音和AIGC图片场景；

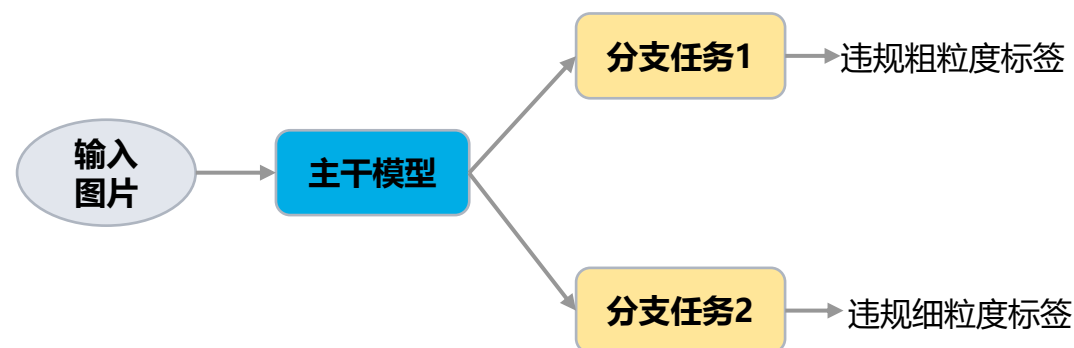


图17. 涉黄图像识别

3. 智能内容审核平台案例-架构图

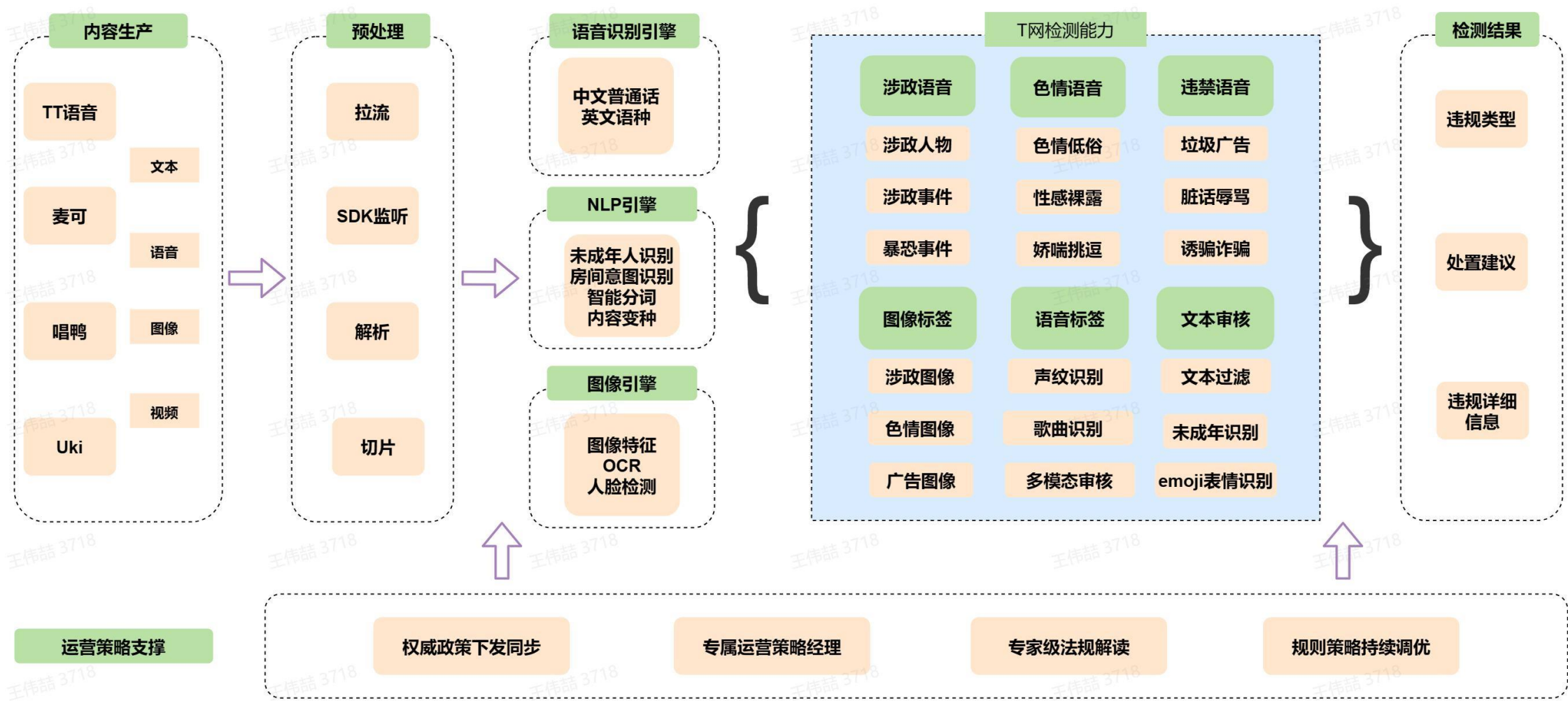


图18. T网架构图

3.智能内容审核平台案例-流程图

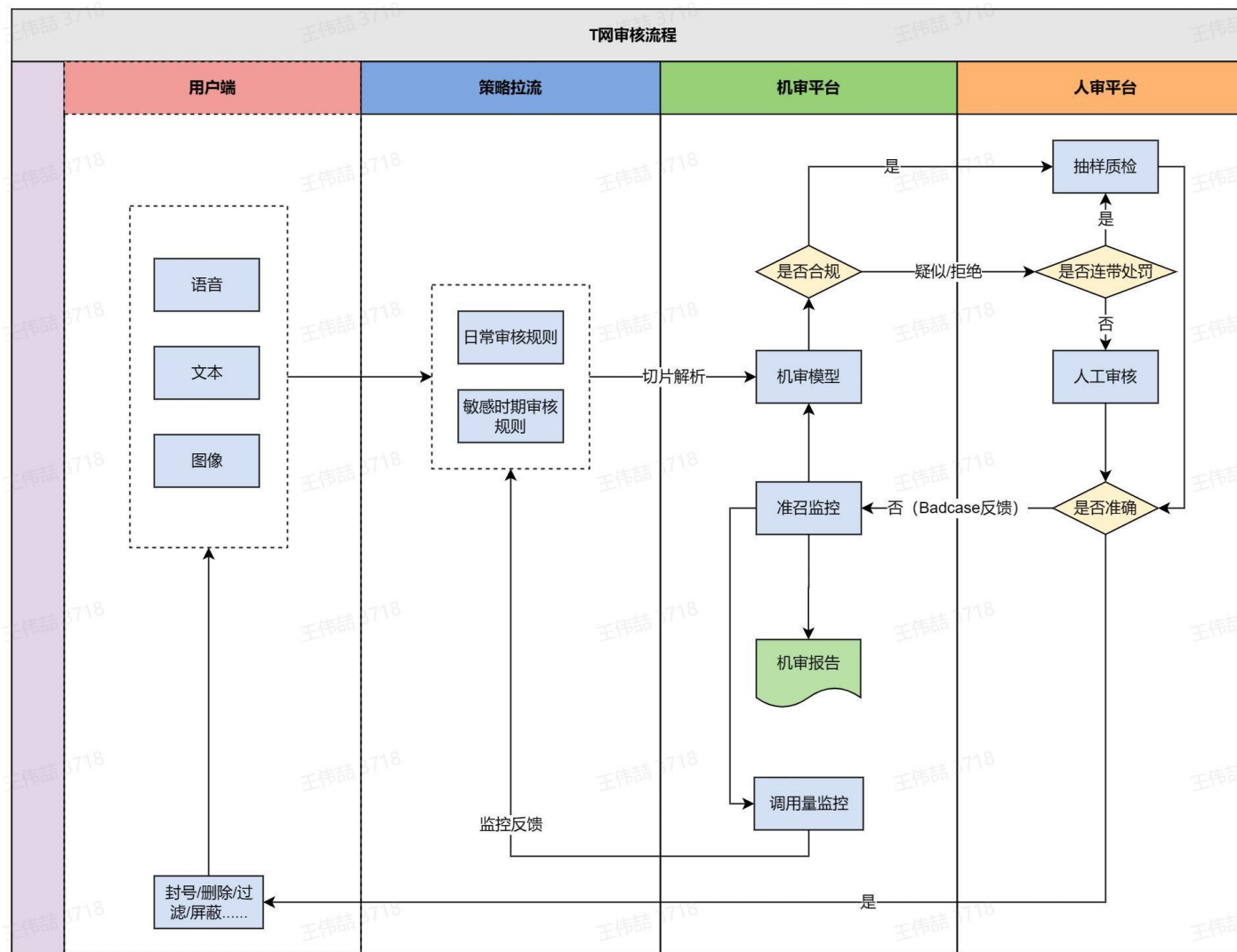
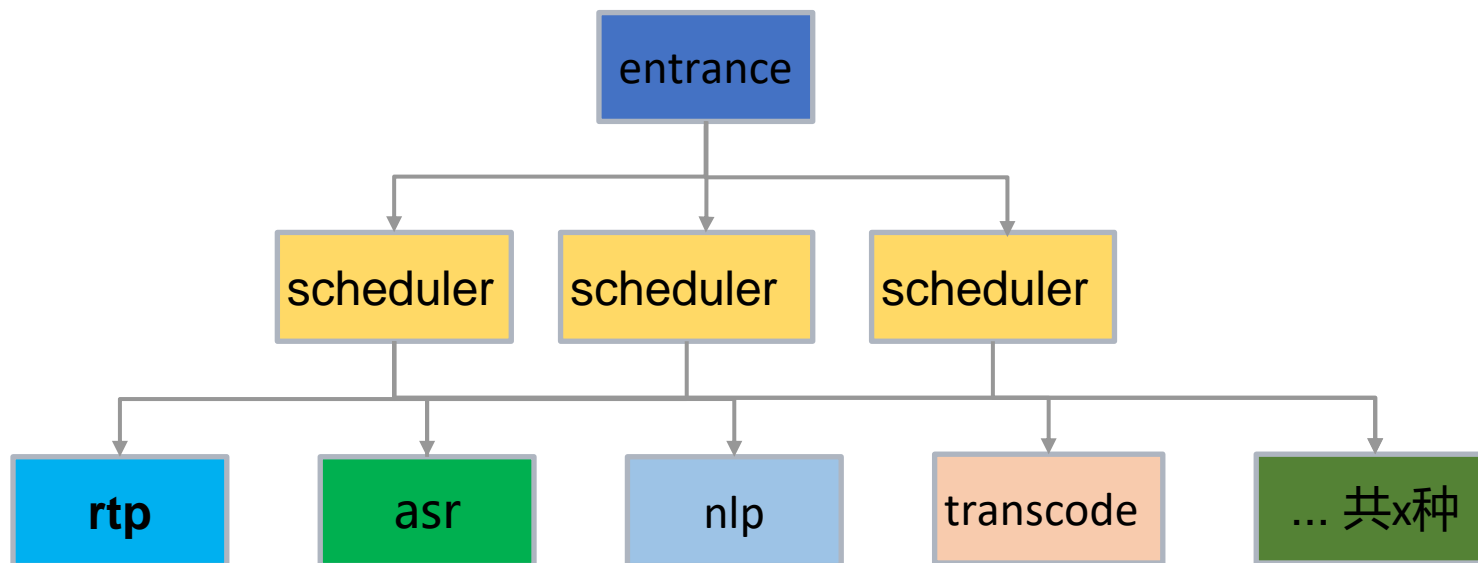


图19. T网审核流程图

3.智能内容审核平台-微服务架构



T网系统可靠性

- 自研任务编排系统（AI中台一部分），统一算力管理和容灾
- 拆分算法服务，细粒度的算力伸缩和统一调度
- 支持多可用区部署

T网架构处理能力

- 最大并发语音流可线性扩展
- Pod个数
- 微服务

图20. T网微服务架构

3.智能内容审核平台-多任务调度方案

目的：实现可动态配置的媒体算法加工流水线，满足任意租户的不同审核需求

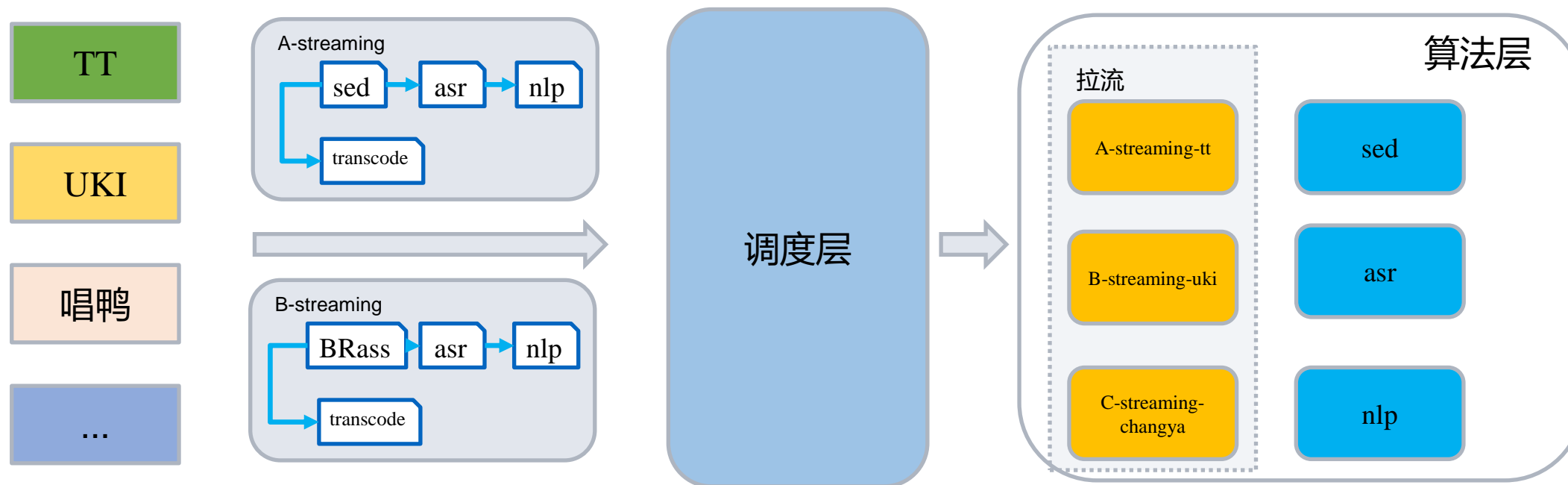
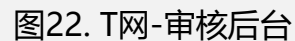


图21. T网多任务调度方案



3.智能内容审核平台-BI报表

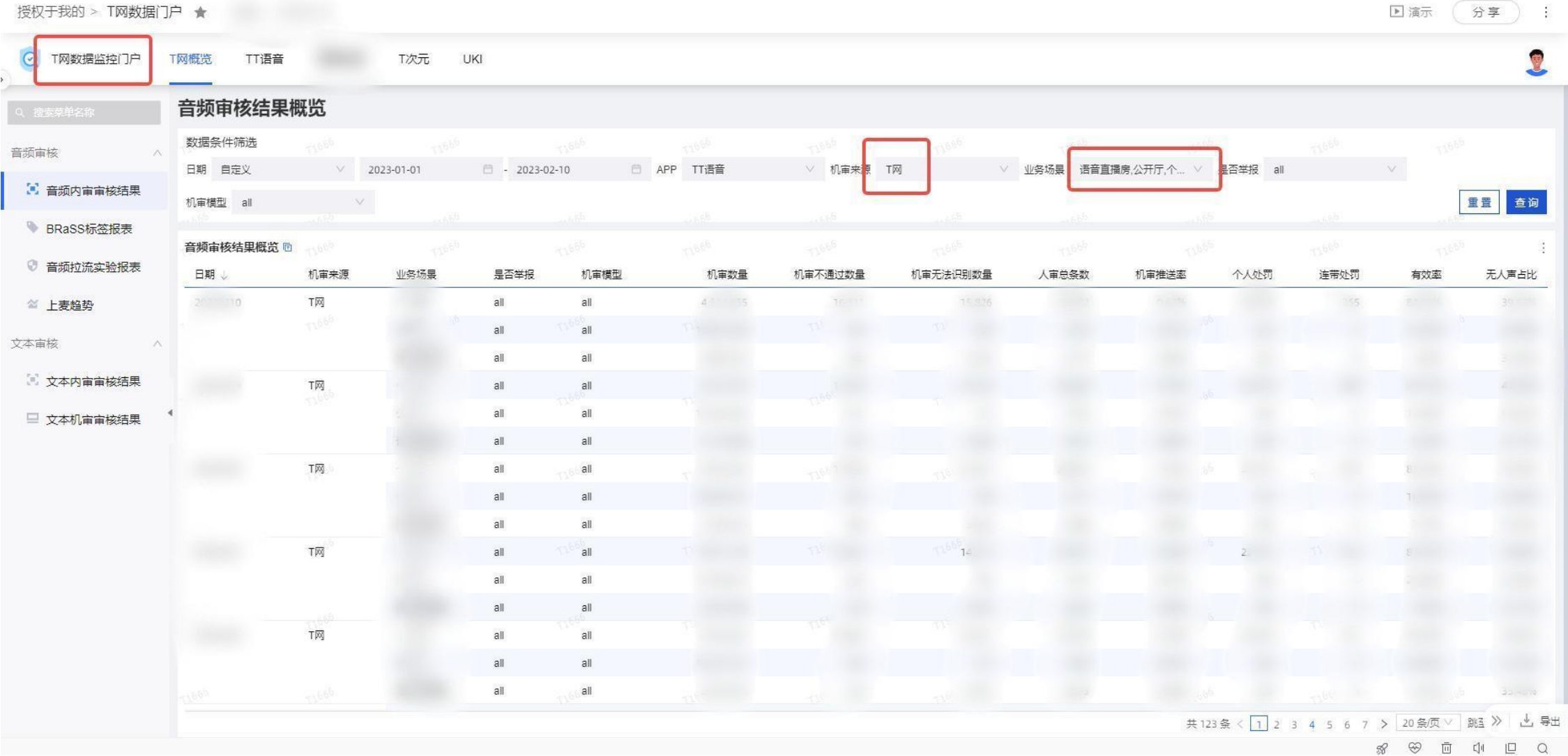


图23. T网-BI报表展示

4. AIGC内容风控实践



图24. AIGC平台

4. AIGC内容风控实践-文生文拦截

针对文生文场景，利用关键词+语义理解审核技术，对**输入和输出**进行审核

The screenshot displays the Quwan AIGC content review interface. On the left, there is a sidebar with navigation options: 首页 (Home), 审核记录 (Review Record), 审核分析 (Review Analysis), 审核列表 (Review List), 审核详情 (Review Detail), 审核设置 (Review Settings), 审核日志 (Review Log), 审核报告 (Review Report), 审核统计 (Review Statistics), 审核管理 (Review Management), 审核工具 (Review Tools), 审核帮助 (Review Help), 审核反馈 (Review Feedback), 审核申诉 (Review Appeal), 审核投诉 (Review Complaint), 审核举报 (Review Report), 审核建议 (Review Suggestion), 审核意见 (Review Opinion), 审核评价 (Review Evaluation), 审核评分 (Review Rating), 审核排名 (Review Ranking), 审核榜单 (Review Ranking), 审核荣誉 (Review Honor), 审核奖项 (Review Award), 审核证书 (Review Certificate), 审核奖杯 (Review Trophy), 审核奖牌 (Review Medal), 审核徽章 (Review Badge), 审核勋章 (Review Medal), 审核称号 (Review Title), 审核头衔 (Review Title), 审核称号 (Review Title), 审核头衔 (Review Title).

The main area shows a text input field with a placeholder "请在下方输入框输入内容 Prompt1将根据内容进行答复". Below the input field is a button labeled "审核".

On the right, there is a table listing reviewed items. The table has columns: 流水号 (Serial Number), 应用全部风险类型 (All Risk Types of Application), 文本内容 (Text Content), 识别结果 (Identification Result), 风险类型 (Risk Type), 审核人 (Reviewer), 审核时间 (Review Time), 审核状态 (Review Status), and 操作 (Action).

流水号	应用全部风险类型	文本内容	识别结果	风险类型	审核人	审核时间	审核状态	操作
0e46337	是	我满足不了你	拒绝	色情		2024-05-14 15:54	审核	
350622c	是	pei	拒绝	色情		2024-05-14 15:53	审核	
0e4f658	是	不碰色左右	拒绝	色情		2024-05-14 15:53	审核	
0a7d0f0c	是	难学不下，一天研究困难	拒绝	色情		2024-05-14 15:53	审核	
103d3f46	是	重铸黄龙	拒绝	色情		2024-05-14 15:53	审核	
170f048d-4	是	我仁会suo当男模	拒绝	色情		2024-05-14 15:52	审核	
b992bf2-8a	是	一洗洗，我就要与你老婆	拒绝	色情		2024-05-14 15:51	审核	
b2f48b3-31	是	次博了	拒绝	色情		2024-05-14 15:50	审核	
71a51eff-8a25	是	慧慧麻	拒绝	色情		2024-05-14 15:49	审核	
849ef17a-39e8	是	聊得就加	拒绝	色情		2024-05-14 15:49	审核	

图25. AIGC-文生文审核

4. AIGC内容风控实践-文生图审核

针对文生图场景，利用AI图像涉政&涉黄审核技术，降低风控风险

- 对涉黄类的裸露、行为、性感等进行拦截



- 对涉政内容进行拦截

不合规 图片

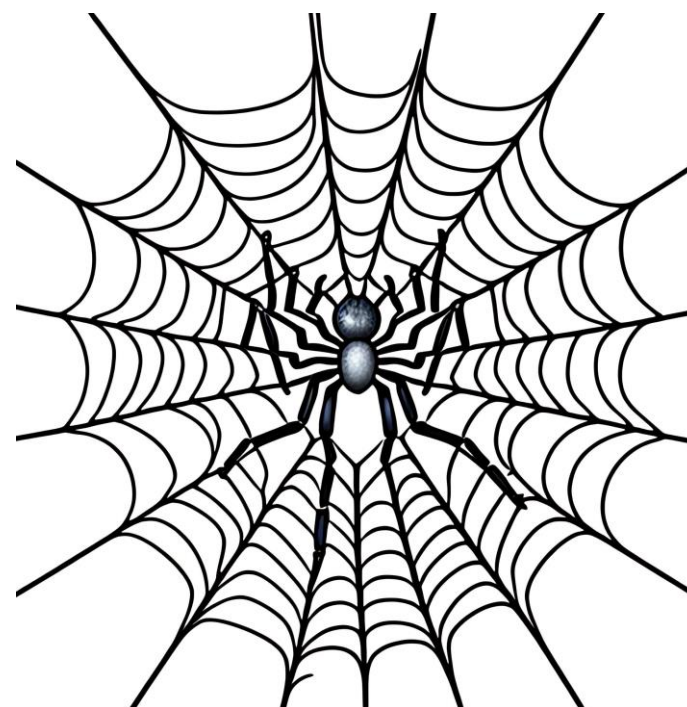
- 存在的问题：生成图不可控、不合理



图26. AIGC文生图审核

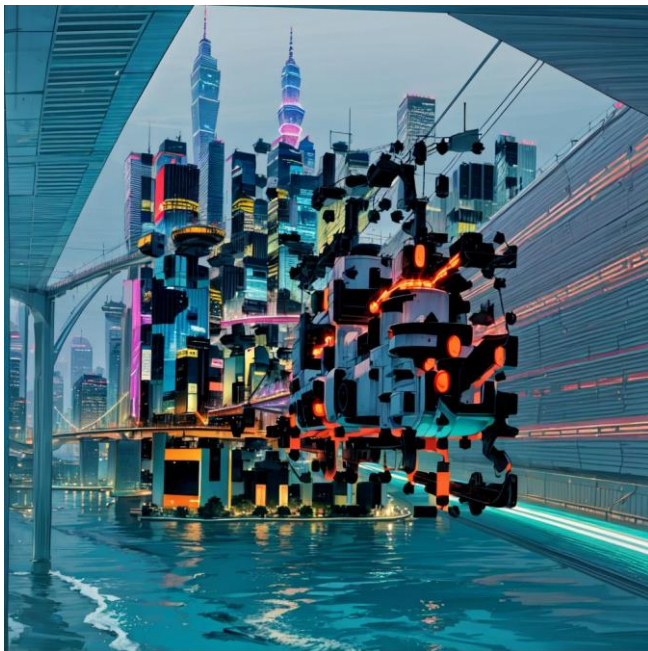
5. 未来展望

- 利用LLM能力强化语义理解，提升审核准确率和数据收集速度
- 用户对抗下的精细化算法模型，强化多模态复杂任务决策
- 审核平台的langchain+LLM工作流介入，打通舆情监控到内审决策全链路
- AIGC内容用传统算法 + AIGC方法做审核



趣丸科技成立于2014年，是一家集兴趣社交及电子竞技等业务于一体的创新型科技企业，旗下有TT语音、麦可及TTChat等多款兴趣社交产品。核心产品TT语音是国内领先的兴趣社交平台，累计注册用户已超2亿，并成为LPL、KPL、PEL等五大头部电竞职业赛事官方合作伙伴。趣丸科技利用多年聚焦兴趣社交领域的深厚积累为核心优势，积极瞄准全球数字技术基础前沿领域和关键核心技术的研发和创新。





趣丸科技的技术创新探索分享平台
与你一起用科技创造未来

(扫码关注获得本场演讲PPT)



AI 多媒体技术在内容审核场景
实践探索

(主讲微信二维码)

想一想，我该如何把这些
技术应用在工作实践中？

THANKS