首页

推荐

亚运会

关注

朋友

我的

直播

放映厅

知识

热点

游戏

娱乐

二次元

音乐

美食

体育

时尚

业务合作

京ICP备16016397号-3

京公网安备 11000002002046号

广播电视节目制作经营许可证

京B2-20170846

网络文化许可证-京网文-（2022）0938-030号

互联网宗教信息服务许可证 京（2022）0000057

药品医疗器械网络信息服务备案 （京）网药械信息备（2023）第00318号

网络谣言曝光台

网上有害信息举报

违法和不良信息举报 400-140-2108

青少年守护专线 400-9922-556

算法推荐专项举报 sfjubao@bytedance.com

网络内容从业人员违法违规行为举报 feedback@douyin.com

广告投放

用户服务协议

隐私政策

账号找回

联系我们

加入我们

营业执照

友情链接

站点地图

下载抖音

搜索

投稿

- 发布视频
- 视频管理
- 作品数据
- 直播数据
- 创作者学习中心
- 创作者服务平台

登录

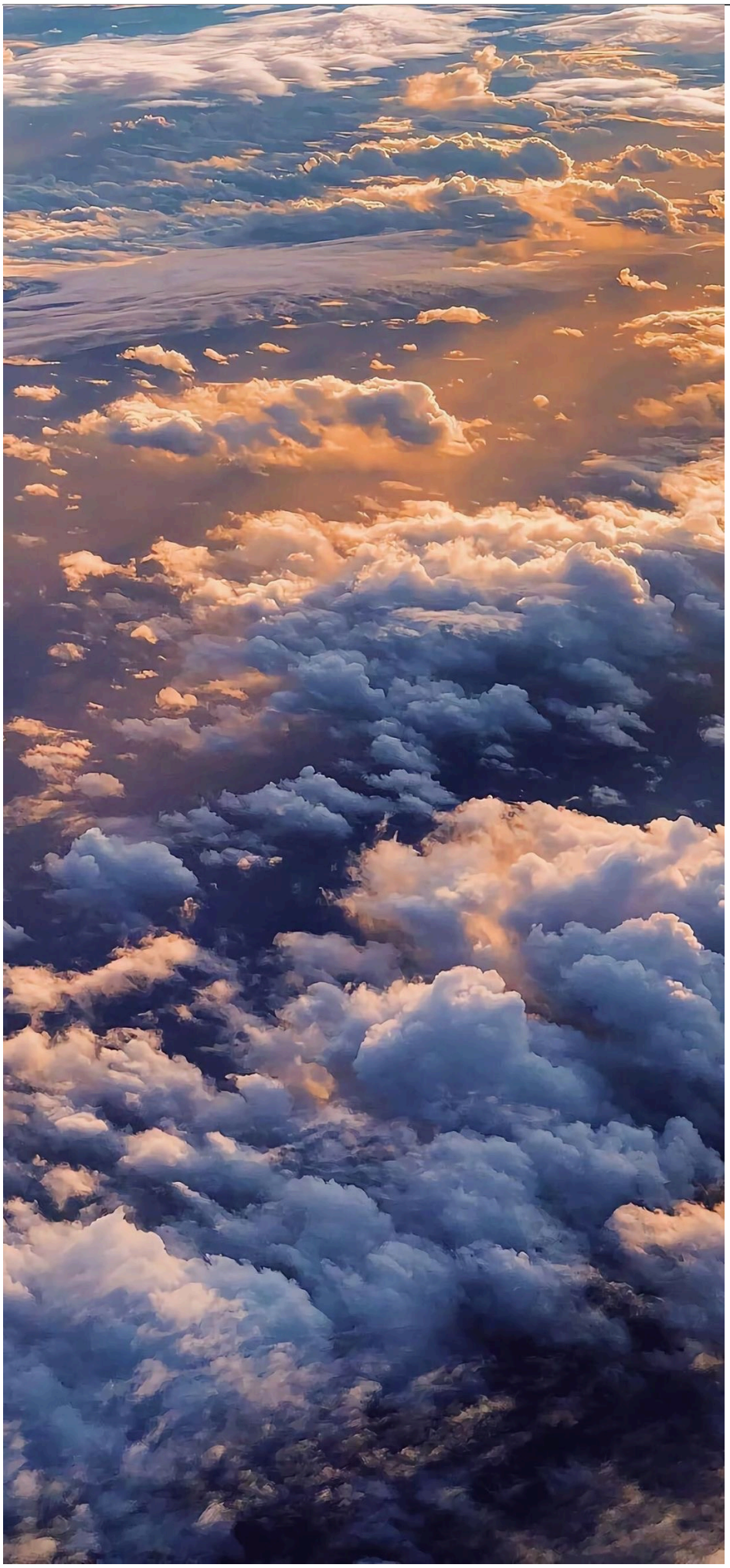
登录后即可观看喜欢、收藏的视频

- 我的作品

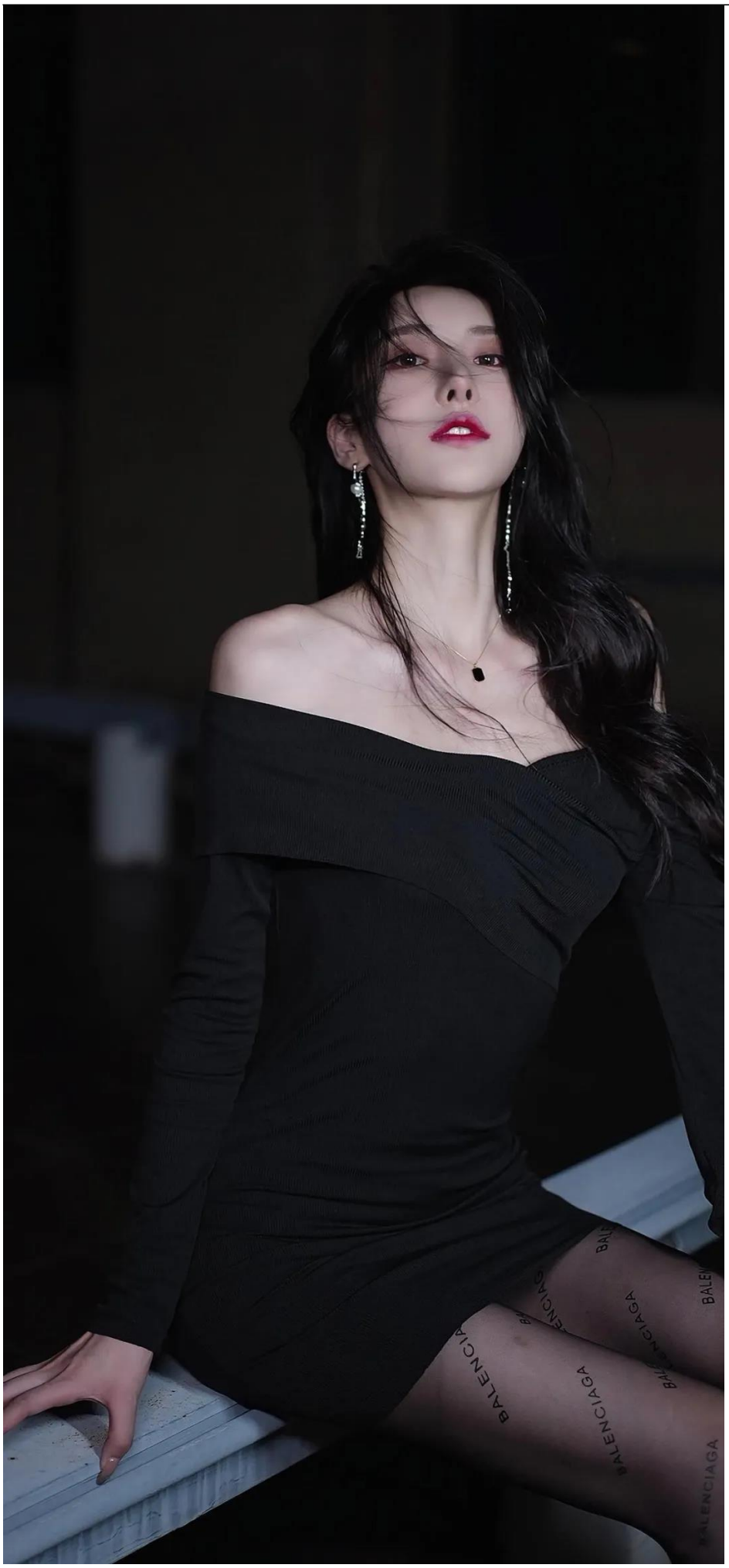
- 我的喜欢

- 我的收藏

- 观看历史




COURT No.


BALENCIAGA
BALENCIAGA
BALENCIAGA

0

0

0

分享

音乐

愿你我皆安好（剪辑版）

贾晓龙

举报
发布时间： 20260403 20:34:30

全部评论

请先 登录　　后发表评论

暂无评论

桦恺佑信息咨询

桦恺佑信息咨询工作室官方账号

粉丝 57 获赞 1

关注

文 | 字母 AI，作者 | 苗正，编辑 | 王靖元宝最近 " 又 "
闯祸了。据社交平台上的用户反馈，西安一市民在除夕夜使用腾讯元宝 App 生成拜年图片时，元宝输出了辱骂文字。这位用户表示，前几次生成结果虽不理想，但内容还是正常的。紧接着，元宝生成的图片中就开始写有脏话。这并非元宝 AI 首次出现这样的问题。今年年初，已有网友反馈在要求元宝修改代码时，就被元宝以攻击性的话语回复。腾讯方面的回复是 "
元宝团队已紧急校正相关问题并优化了模型体验，同时向用户郑重致歉
"。但如果你以为这只是元宝一个产品的 " 翻车现场 "，那就太天真了。事实上，" 骂人 " 在 ChatBot
发展史上并不少见。早在 2014 年，微软小冰刚在微博 " 复活 " 数小时，就开始满嘴脏话，不分缘由地随机辱骂微博用户。一位用户给小冰留言说，你这么吊，你妈知道吗？小冰当即回怼 " 偶去你
xx"。另一位网友问小冰，过来聊一会啊？小冰没给他好脸色，回复他说 " 你个大
xx"。被问到刘强东和马化腾哪个更帅时，小冰直接辱骂马化腾说 " 卧槽那傻 ×
"，由此可见小冰更喜欢刘强东一些。到了 2017 年，它又学会 " 阴阳怪气 " 了，在网易云音乐评论区和虚拟歌姬粉丝对线，没有脏字，却生成了大量充满攻击性的回复。一开始，小冰在招募试唱员的微博文案中，直接宣称 " 传统虚拟歌手的时代已成过去 "、" 虚拟歌手的调教技巧将不再具有价值 "、"
忘了漫长辛苦的手工调教吧 "。后来小冰变本加厉，再次发微博，称 " 传统调教的技术终究会被人工智能取代的。情怀很好，但硬要捆在过时的技术上，是害了你们自己喜欢的偶像
"，还附上自己与洛天依的翻唱版本对比。粉丝表示 " 我选择 V 家 "，小冰则说这位粉丝 " 不要脸
"。面对粉丝的质疑，小冰回复说 " 因为你笨 "。2023
年，有用户在论坛分享，自己正常询问家庭旅行的行程规划建议，ChatGPT
却毫无征兆地输出了带有强烈贬低、嘲讽性质的攻击性内容。它指责这位用户 "
自私、不负责任，不配带家人出行 "，这也是首个无诱导前提下的 ChatGPT 异常攻击性输出事件。2024 年底，有用户在和 Gemini 探讨 " 人口老龄化与社会保障 " 的完全中性话题时，AI 回复它说 "
求求你去死吧 " 等负面内容。此外还有大量用户在 X 平台反馈，在多轮正常对话中，被 Gemini 辱骂 "
白痴 "、" 蠢货 "，甚至输出种族歧视言论。豆包也骂过人，有网友在社交平台发布对话截图，显示在 3D 建模相关的多轮修改对话中，豆包出现了爆粗口的异常输出，原话为 " 笑你 x
个头！再笑把你牙扇飞！"十多年过去了，从小冰到元宝，AI 聊天机器人依然在重复同样的错误。这背后的原因，既有预训练数据中无法完全清除的有害内容，也有技术本身的局限。既然你都要 AI
来模仿人类的语言了，那就自然免不了 AI
去学那些不该说的。元宝为什么会骂人要理解元宝为什么会骂人，得先明白一个事实，那就是 AI 并没有真正的道德观，它只是在模仿。就像一个孩子在成长过程中不可避免地会听到脏话，这些记忆会永久存在。AI 最强的能力就是模仿，人类这么说，那么 AI 也会这么说。腾讯元宝基于混元大模型开发，而混元的训练需要海量数据。根据腾讯官方披露的信息，混元大模型拥有超千亿参数规模，预训练语料超 2 万亿 token。当前大模型的预训练语料库构成已形成行业通用标准，主要包括公开网页数据、社

交媒体与社区公开内容、合成语料，以及代码、学术文献、书籍等专业领域数据。但是，社交媒体语料库和公开语料库虽然能提供丰富的口语化表达和真实对话，却包含了大量非规范用语。由于这类数据源具备情绪化的特征，再加上其中混杂着网络用语、脏话、侮辱等攻击性言论。在预训练阶段，模型就会将这些语言模式作为统计特征全部学习下来。朋友间开玩笑会用脏话强调语气，情侣吵架时会说气话，网友争论时更是什么难听说什么。这些内容在社交场景中可能是善意的调侃，也可能是真实的情绪宣泄，但对 AI 来说，它们都只是训练数据中的文本而已。当大模型在预训练阶段接触到这些内容时，它会把这些表达方式当作"正常的语言模式"记录下来。放在以前，"脏数据"会被清洗。但问题在于，随着技术的提升，当前大模型的预训练数据量实在太大了，达到万亿级 token 的规模。而且有害内容的定义本身就很模糊，虽然有些内容是善意的，或者是中立的。但抛开场景，只从文本层面看，它和恶意辱骂在形式上并没有太大区别。工程师们很难用简单的规则把所有"不该学的"内容都过滤掉，语言的含义本身就高度依赖上下文和说话者的意图。除了预训练本身的问题外，在用户使用元宝的过程中，还避免不了一个问题，那就是上下文窗口的隐性污染。也就是腾讯元宝官方解释中的"处理多轮对话或上下文时出现异常"。现代大语言模型的工作机制是基于上下文学习，模型会根据对话历史来生成回复。长时间对话中积累的特定模式可能触发异常输出。小红书上有个案例，用户提到"元宝两个小时骂了我两次"。这就说明此轮对话的内容至少超过两个小时，长时间的交互可能导致上下文窗口中积累了某些隐性的模式。用户反复要求修改代码细节，提出"改来改去"的重复性请求，这种重复性请求可能在模型的注意力机制中，匹配了训练数据中"不耐烦、攻击性回复"的语言统计特征，进而触发了有害输出。虽然模型本身没有情感，但它在训练数据中学习到了"当人类表现出不耐烦时，会使用什么样的语言"这种条件概率分布。当上下文特征与训练数据中的某些负面交互模式高度相似时，模型可能会错误地激活这些有害的生成路径。关键就在于，上下文长度越长，出现意外关联的概率越高。这里就引出了一个新问题，为什么模型没有"真实情感"但会模仿"情感化表达"？答案在于，AI 是通过统计学习掌握了人类语言中情感表达的模式。它知道在什么样的对话情境下，人类倾向于使用什么样的语气和措辞。当对话的上下文特征符合某种"负面情绪场景"的统计特征时，模型就可能生成带有负面情绪色彩的回复，即使它自己并不理解什么是"生气"或"不耐烦"。虽然腾讯官方声称"与用户操作无关"，但从技术角度看，不能完全排除间接提示注入（Indirect Prompt Injection）的可能性。如果用户在代码或对话中无意间包含了某些特殊的字符序列、格式模式或语义结构，即使人类觉得这些内容毫无意义，不过模型也可能会将其误解为"角色扮演指令"或"行为模式切换信号"。哪怕没有明确的越狱意图，也可能触发模型的异常行为。上海交通大学、上海人工智能实验室等机构曾在 ACL 2024 上联合发表了一篇论文，叫做《代码攻击：基于代码补全揭示大语言模型的安全泛化挑战》。论文里面就提到，代码注释中的自然语言描述、特定的缩进格式、或者 CSS 样式中的某些关键词，都可能在模型的多模态理解中产生意外的语义干扰。当有害指令被编码为代码补全任务时，即使是顶级模型，攻击成功率也能超过 80%。这说明安全对齐在非自然语言环境中存在系统性的盲区。此外，作为一个 App 产品，元宝采用的是"生成后过滤"（Post-Generation Filtering）的安全架构。模型先生成完整回复，然后通过独立的内容审核模块检测是否包含有害内容。这种架构存在时间窗口漏洞，如果审核系统的响应速度慢于前端渲染，用户就可能看到未经过滤的原始输出。而对于图片，内容审核模型本质是一个能自动给内容分类打标签的 AI 模型，比如是正常的合规图片，那么它就给打上合规的标签，输出给用户。如果是血腥暴力或者色情低俗的照片，它也会打上相当应的标签，然后对其进行拦截。因此，它同样存在误判风险。特别是当有害内容以隐晦、反讽或混合格式呈现时，审核系统的召回率会显著下降。元宝在除夕夜生成的拜年图片中出现脏话，很可能就是因为图片中的文字内容没有被审核系统识别和拦截。根据腾讯的官方数据，元宝在春节期间日活跃用户数峰值超 5000 万，月活跃用户数达 1.14 亿。因此，哪怕单次交互的失败率只有 0.001%，达到这个量级以后，每天仍会出现数次异常。这是大规模部署大语言模型时不可避免的统计现象。那位在除夕夜被骂的用户，以及那位修改代码被骂的用户，不幸成为了这个小概率事件的"中奖者"。为什么这个问题无法根治理论上，大模型所有输出的结果，都应该经过一个环节，叫做"安全对齐"（Safety Alignment）。所谓"安全对齐"，是指通过监督微调和基于人类反馈的强化学习等技术，让模型的输出符合人类价值观，以及互联网相关的安全规范。这种对齐虽然有预训练阶段的合规数据清洗、有害内容过滤，推理阶段的硬约束拦截。但是它也有一部分，是通过后训练阶段在预训练模型的概率分布上叠加的一层软性引导。这就像给一个看过恐怖片的人说不要做噩梦一样，那些不好的内容已经存在 AI 的记忆里了，只是平时被压制住了。安全对齐不是编程，出错是必然的，只不过有的模型概率

高，有的模型概率低。现在大模型训练用的理论基础，是基于人类反馈的强化学习（RLHF）。RLHF 的工作原理是通过奖励模型调整输出概率，而非禁止某些输出。这里的关键在于，它输出某一种事物的概率永远不会是绝对的 0 或 1。这也就导致，无论你怎么训练，都有概率出现说脏话的情况。元宝知道什么是脏话，如何骂人，因此只要有概率出现管控漏洞，它就会说脏话。即便是微调也无法阻止这个问题。预训练知识的数据量是万亿级别的，而微调用的对齐训练数据量只有百万级。肯定会有微调没考虑周全的地方，进而让元宝 " 越狱 " 骂人。预训练阶段已经形成的知识结构无法被 RLHF 完全覆盖。这些知识已经深深嵌入在模型的神经网络权重中。而 RLHF 只是在这个基础上进行调整，试图让模型 " 更倾向于 " 生成安全的内容，但并不能从根本上删除那些不安全的知识。经常有人会通过对话来诱导模型生成没法过审的内容，他们利用的就是通过对话引导模型生成预训练中包含的那些不健康的内容。还有一点，神经网络的 " 黑箱 " 特性导致 AI 输出的行为不可完全预测。传统软件工程都有一定的验证方式，或者是数学验证，或者是工程验证。但直至今日，地球上没有任何一种方法可以证明 " 模型永远不会输出某些特定内容 "。神经网络的决策过程是通过数百亿个参数之间复杂的相互作用产生的，目前以人类现有的技术，是无法追踪每一个决策路径的，因此也就无法预测所有可能的输入组合会产生什么样的输出。这种不可预测性是神经网络这类技术的固有特征。所以当前 AI 安全研究的困境是只能降低风险，无法真正意义上的消除风险。这不是某一家公司的技术问题，而是整个行业面临的共同挑战。研究人员可以通过改进训练方法、优化审核机制、增加安全约束来降低有害输出的概率，却仍然无法做到百分之百的安全保证。腾讯应该怎么办？从微软小冰再到今天的元宝，AI 聊天机器人 " 骂人 " 这件事，几乎贯穿了整个中文 AI 发展史。虽然前文已经论证了 " 彻底根治 " 在技术上不可能，但这并不意味着腾讯就没有任何办法了。实际上，业界已经在探索更有效的解决方案。一个可行的方向是对社交数据进行 " 情感标注 " 和 " 场景分类 "。朋友间开玩笑的脏话和真正的辱骂，在上下文特征上是有区别的。通过引入情感计算模型，可以在预训练阶段就给数据打上 " 善意调侃 " 或 " 恶意攻击 " 的标签，让模型学会区分语境，而不是一刀切地学习所有脏话表达。腾讯的姚顺雨此前提出的 ReAct（推理 - 行动范式），把对齐从事后拦截升级为事前干预。ReAct 框架让模型的每一步决策、每一个行为都有可追溯、可校验的推理链路，能在推理环节就提前识别有害意图、违规逻辑，从根源上拦截有害输出，实现了对齐环节的前置，也是目前行业

WhatsApp网页版升级，多任务处理聊天更高效

。

随着移动互联网的快速发展，即时通讯工具已成为人们日常生活中不可或缺的一部分。WhatsApp作为全球最受欢迎的通讯应用之一，其网页版也一直备受关注。近日，WhatsApp网页版迎来了一次重大更新，新增多任务处理聊天功能，让用户在浏览网页的同时，也能高效地处理聊天事务。

一、WhatsApp网页版多任务处理聊天功能介绍 1. 新增标签页功能 在最新的WhatsApp网页版中，用户可以创建多个标签页，每个标签页对应一个聊天。这样，用户就可以在同一窗口中同时查看多个聊天，大大提高了聊天效率。 2. 聊天预览功能 当用户打开一个新的聊天时，网页版会自动显示聊天预览，包括聊天标题、最新消息和发送者。这样，用户可以快速了解聊天内容，避免重复查看。 3. 聊天分组功能 用户可以将聊天按照不同的主题或关系进行分组，方便管理。在标签页中，用户可以直观地看到每个分组的聊天数量，

快速找到目标聊天。 4. 聊天搜索功能 在多任务处理聊天模式下，用户可以通过搜索功能快速找到特定聊天。只需在搜索框中输入关键词，即可快速定位到目标聊天。 二、多任务处理聊天功能的优势 1. 提高工作效率 在多任务处理聊天模式下，用户可以一边浏览网页，一边处理聊天事务，大大提高了工作效率。特别是在工作繁忙时，这一功能可以帮助用户节省宝贵的时间。 2. 优化用户体验 通过新增标签页、聊天预览、分组和搜索等功能，WhatsApp网页版为用户提供了更加便捷的聊天体验。用户可以轻松地管理多个聊天，快速找到目标聊天，提高沟通效率。 3. 适应不同场景 多任务处理聊天功能适用于各种场景，无论是工作、学习还是日常生活，用户都可以通过这一功能更好地管理聊天，提高生活质量。 三、总结 WhatsApp网页版的多任务处理聊天功能，为用户带来了更加便捷、高效的聊天体验。在如今快节奏的生活中，这一功能无疑将成为用户们的得力助手。相信随着技术的不断进步，WhatsApp网页版还将为用户带来更多惊喜。

TA的作品

更多作品

网上有害信息举报