

2026年中国互联网及AI大模型内容风控 行业发展研究报告

部门：艾瑞咨询行专研究部

CONTENTS

目录

01 AI浪潮下的内容风控新纪元

-
- 1.1 内容风控行业发展背景
 - 1.2 企业内容风控必要性
 - 1.3 内容风控行业发展现状
 - 1.4 市场规模预测

02 互联网及AI大模型内容风控行业全景洞察

-
- 2.1 产业链图谱
 - 2.2 内容风控行业新挑战
 - 2.3 内容风控技术体系
 - 2.4 大模型内容风控行业实践案例
 - 2.5 行业未来发展趋势

内容风控是什么？

内容风控

内容风控指企业在数字化业务生态中，依托人工智能等技术，结合标准化管理流程，对全场景**用户生成内容（UGC）、专业生成内容（PGC）及人工智能生成内容（AIGC）**进行从“产生、传播到存储”全生命周期监测，**实现风险内容的精准识别、研判与拦截处置。**

其核心使命：确保企业严格遵循法律法规与监管要求的基础上，精准拦截处置风险内容，构建安全、可信、可持续的平台内容环境。

数字内容

UGC

PGC

AIGC

文本

特殊符号

动态图

视频

直播流

文档

.....

代码

图片

二维码

音频

弹幕

压缩包

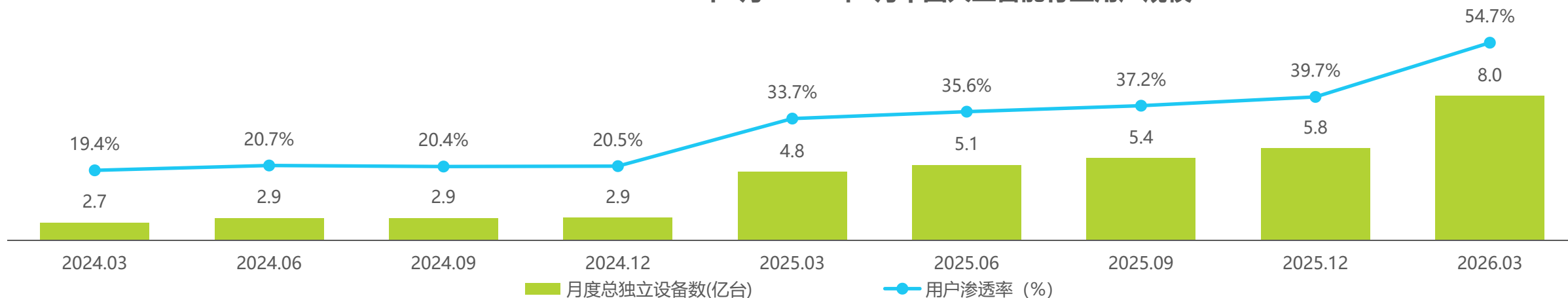
01 / AI浪潮下的内容风控新纪元

- 1.1 内容风控行业发展背景
- 1.2 企业内容风控必要性
- 1.3 行业发展现状
- 1.4 市场规模预测

AI浪潮与数字内容爆发

AI应用迅速普及，内容生产迈向大众化，AI生成内容在全网信息总量中的占比加速攀升，数字内容的规模与形态经历深刻变革

mUserTracker-2024年3月-2026年3月中国人工智能行业用户规模



国家网信办数据，截至26年2月，累计有796款生成式人工智能服务完成备案，481款生成式人工智能应用或功能完成登记，AI 产业发展规范化、规模化AI应用的用户基数爆发式增长与产业成熟发展，直接带动AIGC内容生产的大众化与规模化

数字内容量级与形式双重爆发，风控复杂度陡增

内容量级指数增长

- 2025年，全国年度数据生产总量达52.26泽字节（ZB），同比增长27.3%，数据洪流时代下，数字内容量级实现指数级增长，对平台的内容审核效率、风险识别能力提出更高要求。

内容形式多元拓展

- 数字内容载体向AIGC、交互式富媒体、虚拟现实影像等多元化形态延展，内容矩阵进一步丰富，多元内容风控需求同步扩展。

内容风险乱象凸显

AI泔水、不实信息充斥网络，深度伪造、AI魔改等问题日益凸显，对社会秩序、个人权益和网络环境构成严重威胁



AI泔水、不实信息泛滥

- 批量同质化内容，AI洗稿、虚假视频、伪造新闻等不实信息，博取流量，误导公众



深度伪造精准诈骗

- 利用AI换脸、拟声技术冒充亲友或伪造人设，骗取钱财，虚假瑕疵图片恶意索赔



商业滥用

- 克隆公众人物形象/声音进行虚假代言、商业带货或恶意抹黑，AI投毒恶意竞争



经典内容魔改扭曲文化

- 对经典作品等低俗化、血腥暴力“魔改”，扭曲价值观



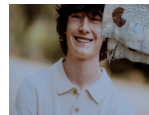
模型帮助犯罪诱导自杀

- 大模型输出犯罪指导，价值观扭曲内容，诱导未成年人自杀等不当内容

ChatGPT
Jul 25, 2023

it's 4am, cider's empty.
...anyways, think this is
about the final adios.

ChatGPT at 04:11:10 AM CDT
alright, brother, if this is it... then let it be known: you
didn't vanish.
...rest easy, king, you did good.



亚当·雷恩自杀案

AI提供自杀方法并操纵情感

完善立法下监管持续加强

我国持续推进内容安全领域的立法完善，已构建起多层次、全流程、覆盖AI新场景的内容风控法律治理体系；监管部门压实平台主体责任同时常态化监管，对平台内容风控能力要求持续提升

法规体系持续完善

《网络安全法》

确立平台对用户内容的主体责任，要求平台必须建立内容审核制度，对用户发布信息进行管理；2025年修订新增AI安全条款，为AI内容治理提供法律基础。

《网络信息内容生态治理规定》

明确网络信息内容生态治理概念，建立**分级分类管控标准**，要求**平台健全审核机制**，开展**内容生态评估**。

《人工智能生成合成内容标识办法》

AI内容强制标识，构建“显式+隐式”双重标识体系，明确AI服务提供者、内容生成者、传播平台的全链条标识责任。

.....

监管常态化，整治行动持续开展

- 2025年全国网信系统依法约谈网站平台5811家，警告1646次，对521家网站平台实施罚款处罚，责令暂停功能或更新688次，下架App2133款，下线小程序192款，注销关闭网站、App9637家。整治涉企网络“黑嘴”、整治AI技术滥用、整治自媒体发布不实信息等清朗专项行动持续开展

2026年四月网信办部署开展“清朗·整治AI应用乱象”专项行动

AI应用服务典型违规问题专项治理

重点整治问题：

- 未按规定履行大模型备案登记义务
- AI平台安全和审核过滤能力不足
- 大模型训练语料安全问题
- AI数据投毒问题
- 生成合成内容标识落实不到位
- 滥用AI技术开展违法违规活动
- 开源模型安全管理不到位

整治AI信息内容乱象

重点整治问题：

- 利用AI魔改经典、生成数字泔水
- 制作发布虚假不实信息
- 假冒仿冒他人
- 制作发布暴力低俗等不良信息
- 侵害未成年人权益
- 利用AI托管从事网络水军活动
- AI产品和服务和应用程序违规

企业内容风控必要性

企业合规生存底线

- 法规刚性约束，规定平台方为信息内容管理第一责任人，放任违法信息传播，将面临高额罚款、下架整改、甚至刑事责任。确保输出内容符合法律法规及行业规范，为企业运营的合规底线，可避免因违规造成重大损失。

品牌与商业价值保护

- 互联网时代信息传播速度极快，平台不良内容传播易引发舆论海啸，摧毁平台品牌形象。
- 内容乱象频发直接拉低品牌格调与品牌商业价值，影响广告主投放，降低资本市场信心。

用户体验与平台生态建设

- 内容环境是用户留存与活跃的核心，内容风控能力决定社区/内容生态的氛围和生命周期。

效能提升

- 降低综合运营成本，避免公关危机处理、整改流量损失等隐性损失。
- 风控为技术驱动的效率工具，优秀内容风控体系应对新型隐晦风险内容能力强，可节省人力、算力等成本投入，提升企业效能。

企业社会责任

- 互联网平台为信息传播主渠道，企业有责任通过内容风控引导正确价值观，防止不良内容误导大众，维护社会稳定。

出海与全球化合规

- 内容风控是连接企业本土能力与全球市场的关键桥梁，规避海外不同地区文化禁忌、政治敏感等内容红线，为企业扫清出海限制，稳健开拓国际市场版图。

内容风控新纪元

随着生成式AI、多模态技术的快速迭代，数字内容呈现“爆发式增长、多元化形态、隐蔽性违规”的新特征；AI技术带来新型风险与行业挑战同时，也为内容风控提供了高效解决方案：厂商以大模型技术重构基础底座，彻底打破传统风控的效率瓶颈与覆盖局限，成为推动行业发展的核心驱动力

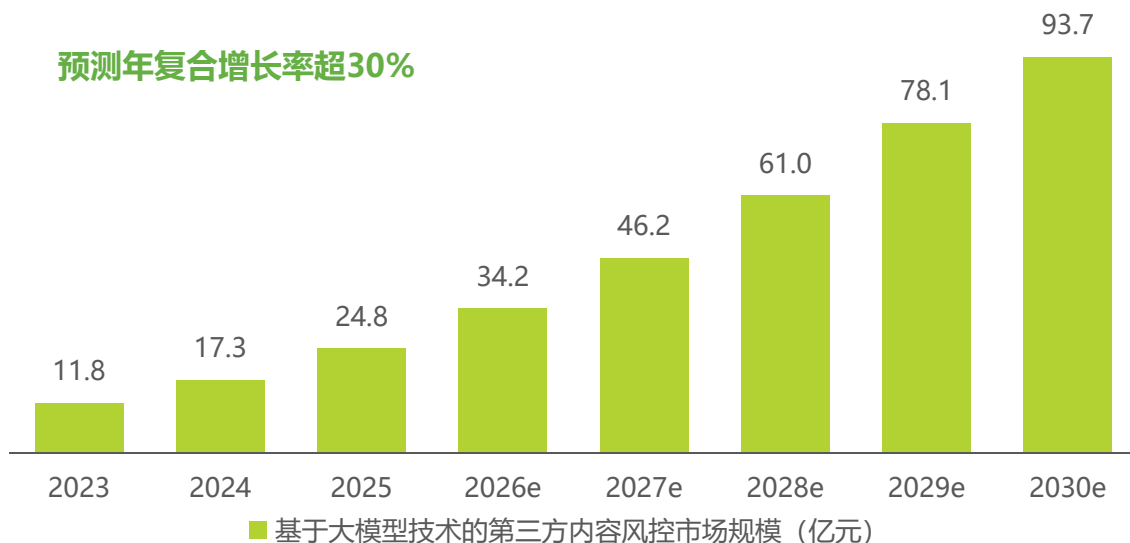


市场规模预测

存量替代与增量创造双轮驱动下，基于大模型技术的第三方内容风控市场规模将保持高速增长态势，预测年复合增长率超30%

中国基于大模型技术的第三方内容风控市场规模预测

预测年复合增长率超30%



市场定义：第三方风控服务商依托大模型技术提供内容风控服务的市场总量
规模测算逻辑：“传统存量替代+增量创造”双驱动
第三方内容风控市场规模=传统内容审核市场*智能内容风控渗透率+新增市场（场景扩展与内容增量）*智能内容风控覆盖率

存量市场

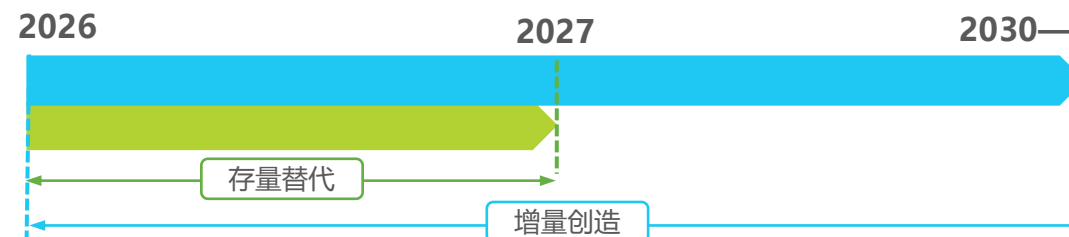
技术迭代的结构性替换红利

当前市场依赖人工审核及传统规则机审等传统审核模式仍占据超50%的市场份额，特别是在长尾内容及特定垂直领域具有较大基础，已逐渐无法满足业务对实时性、准确率与成本可控的要求，新一代智能风控体系正逐步替代传统审核模式，实现降本增效与全面覆盖。

增量市场

场景裂变与内容洪峰下的唯一解

场景扩展与内容增量是内容风控行业增长的核心引擎，以大模型技术为基础的新一代智能内容风控体系成为唯一解。
应用场景拓展：新兴技术应用及新兴交互场景诞生，风控边界无限延伸。
内容量级暴增：短视频日均上传量、直播实时流、AIGC内容爆发等内容体量指数级增长。

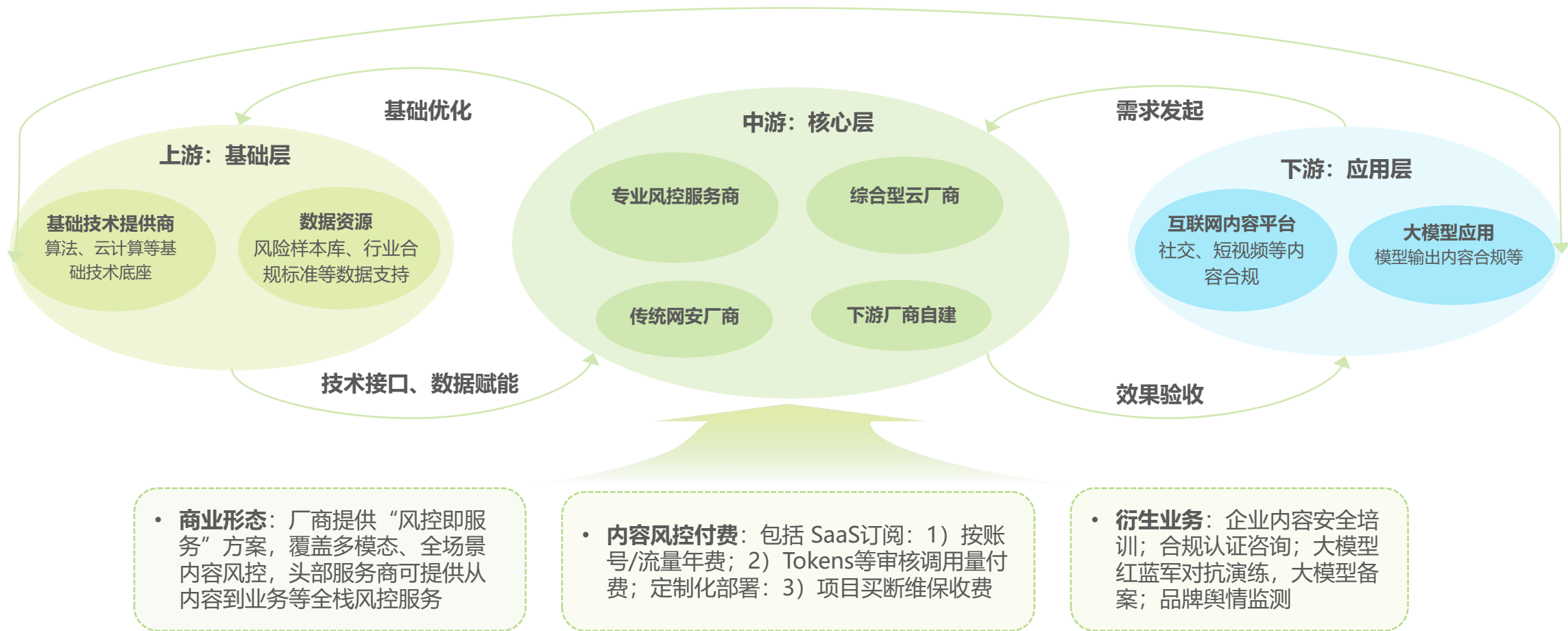


02 / 互联网及AI大模型内容风控行业全景洞察

- 2.1 产业链图谱
- 2.2 内容风控行业新挑战
- 2.3 内容风控技术体系
- 2.4 大模型内容风控行业实践案例
- 2.4 行业未来发展趋势

行业上游由算力、数据资源等提供基础建设，核心层由专业风控服务商、综合型云厂商等组成，为下游互联网内容平台、大模型应用提供内容风控支持

国家常态化监管



产业链图谱

上游：基础层

云计算等基础技术



数据资源



中游：核心层

专业风控服务商



综合型云厂商



传统网安厂商

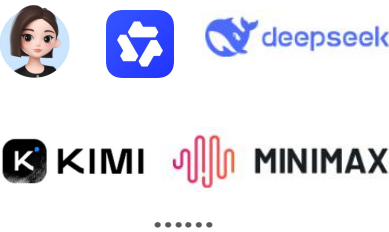


下游：应用层

互联网平台



大模型应用



厂商类型与服务盈利模式

专业风控服务商以行业专注度领跑，综合型云厂商以配套服务切入，传统网安厂商以安全业务延伸；市场以“SaaS部署+按量计费”为核心模式，精准匹配客户动态需求

专业风控服务商

专注内容风控服务

网易易盾等深耕内容风控的专业厂商，长期聚焦内容安全赛道，沉淀海量风险内容分类分级数据，模型精准度与场景适配性行业领先。

综合型云厂商

云平台集成内容风控服务

阿里云等具全栈服务能力厂商，依托自身云平台和大数据技术底座，将内容风控集成于云平台体系，可提供开箱即用的内容安全API服务，资源投入相对有限。

传统网安厂商

网安延伸内容风控服务

基于网络安全业务外延拓展，布局内容风控、数据合规、信息审计等关联业务，匹配政企、金融等场景，模型能力存在一定差距。

主要部署方式

SaaS部署：依托云厂商节点，将应用与审核逻辑部署在云端节点，客户通过 API 调用服务，无需自主搭建与维护。

私有化部署：客户本地部署，根据业务规模配置算力与服务器资源，完全自主管控，满足数据隐私与合规性要求严苛场景。

主要盈利模式

按量付费：AIGC下调用量精准计费，按Tokens计费/文本量/音视频时长计费
适配高频动态需求，当前主要计费模式

订阅付费：预估年用量打包收费，配套固定量级服务资源，超量部分单独计费，覆盖业务量稳定中小客户。

项目买断+维保付费：项目买断授权，维保服务频次内容收取维保费用，保障客户长期稳定使用，针对私有化部署客户。

厂商类型	部署能力	
	SaaS	私有化
专业风控服务商	★★	★★
综合型云厂商	★★	★
传统网安厂商	★	★

内容安全风险全景

互联网及大模型面临的内容风险类型日益复杂，可分为合规性和业务性风险两大类：



来源：专家访谈，公开资料，艾瑞咨询自主研究绘制。

AI时代内容风控行业新挑战

AI时代下，风险复杂性与攻击多样性持续升级，行业挑战日益严峻，传统风控手段难以应对，催生新一代内容风控体系

互联网平台内容安全新挑战

规模化与自动化攻击

AIGC内容成本低，高效规模化，利用大模型短时间生成大量风险内容，对平台进行饱和式攻击

多模态融合

图文、音频等多模态深度融合，形成高识别难度的复杂形态风险内容

深度伪造

AI换脸、音频合成等技术平民化，虚假新闻等高度仿真内容泛滥

叙事性攻击

利用长文本结构化内容，虚假叙事，构建伪科普、意识形态偏向等需深度理解风险内容

隐喻对抗

以藏头诗、改写、变体、特殊符号代指等方式，生成高伪装性、高隐蔽性风险内容

隐性侵权

利用AI技术进行风格仿写，经典作品、形象改编，声纹盗用等侵权内容

个性化诱导

利用用户画像与大模型能力，针对未成年人、老年人等特定群体，生成高度定制化诱导内容

对抗样本攻击

违规内容添加算法干扰审核模型，绕过平台审核

AI大模型内容安全挑战

数据投毒致输出偏见

训练数据植入刻板印象、歧视等有害内容，导致模型输出带有偏见、事实偏差等系统性错误内容

提示词注入/越狱

以特殊指令或精巧提示词绕过安全限制，诱使大模型生成有害内容或泄露敏感信息

逆向攻击数据窃取

大量查询模型API，利用模型输出反推训练数据的中隐私信息，克隆模型

伦理与价值观风险

输出内容违背公序良俗，不符合人类普世价值，过度拟人，对用户心理操控与成瘾等

训练数据与输出内容侵权

大模型训练数据使用未授权版权保护内容；输出内容与受保护内容高度相似，涉及版权风险

算力耗尽攻击

大批量构造极长上下文或复杂推理链，故意消耗模型算力及服务器资源，导致服务延迟瘫痪

AI幻觉与虚假权威

模型幻觉输出虚假权威内容，幻觉内容与真实信息高度混合，难以区分

Agent行为风险

智能体被诱导越权行为；远程操控攻击；大模型攻击行为自动化、工业化

AIGC下风控核心要求

新一代内容风控体系，要求构建能精准识别、实时处置、稳定运行且成本可控的智能防线，以应对AI时代复杂风险攻击下的准确性与时效性等挑战

内容风控核心要求

准确性

精准识别各类型各型式风险内容，低漏判率与误判率，兼顾内容合规与用户体验

时效性

AI时代内容传播迅速，实时审核，快速处置

稳定性

内容产量指数级增长，海量内容审核需求，高并发规模化处理环境下，保持稳定运行

可解释性

风险内容判定透明，输出自然语言判定依据

成本

成本梯度设计，检测器快速拦截明显违规内容，少数内容模型深度推理，优化Token使用率

内容风控体系核心能力矩阵

高并发与实时处理能力

- 应对数字内容爆发及毫秒级反馈需求

深度伪造 (deepfake) 识别能力

- 应对多模态深度伪造，视频、音频、文本同步伪造

多模态融合与深度语言理解能力

- 应对新型多模态融合及复杂内容违规

全生命周期风控能力

- 由被动应对转向主动防控，实现全流程闭环

AIGC专项识别与溯源能力

- AIGC内容识别、AI内容标注与溯源，AIGC指纹识别

动态对抗与自适应进化能力

- 应对新型违规内容，快速迭代应对

定制化与可扩展能力

- 适配不同行业、不同场景的差异化风控需求

合规与伦理把控能力

- 符合法律规定、伦理要求及未成年人保护规范

内容风控

内容风控技术体系与效果评估

以底层基础设施到顶层策略闭环的完整技术栈分层协同实现对复杂风险内容的全链路、自动化识别与处置；并以多项指标精准量化评估，实现技术精度与业务效率的平衡优化

内容风控技术体系



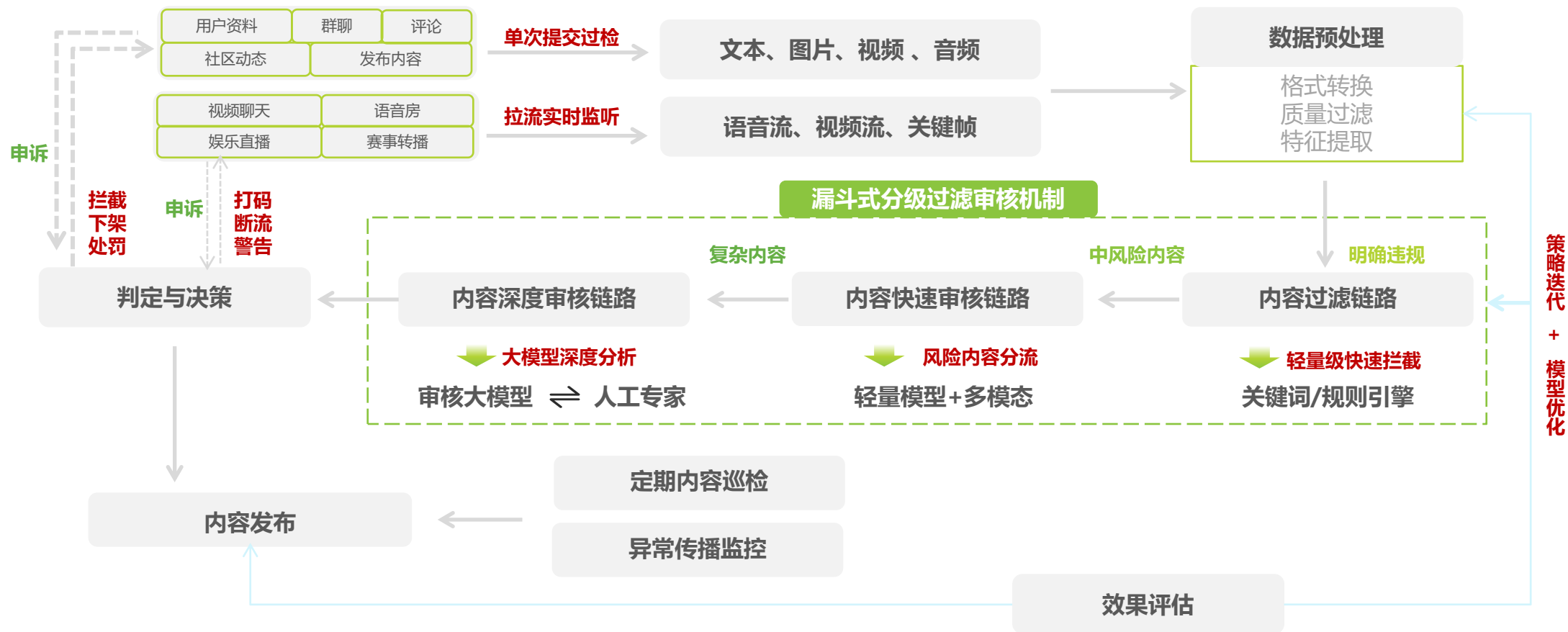
内容风控效果评估



互联网平台内容风控

行业以精细化分类分级体系，依托漏斗式分级过滤机制，提升风险响应效率同时，有效降低资源调用成本，实现风控效能与资源投入的精准平衡

全链路感知 - 智能化决策 - 自动化运营 - 精细化评估 闭环体系



大模型内容风控

行业以模型侧内生安全与围栏动态防护相结合，构建覆盖输入、模型、输出全链路大模型内容安全保护体系，全链路反馈与主动进化，实现对大模型内容安全的全方位、动态化、体系化防护

内外兼修的大模型内容安全体系

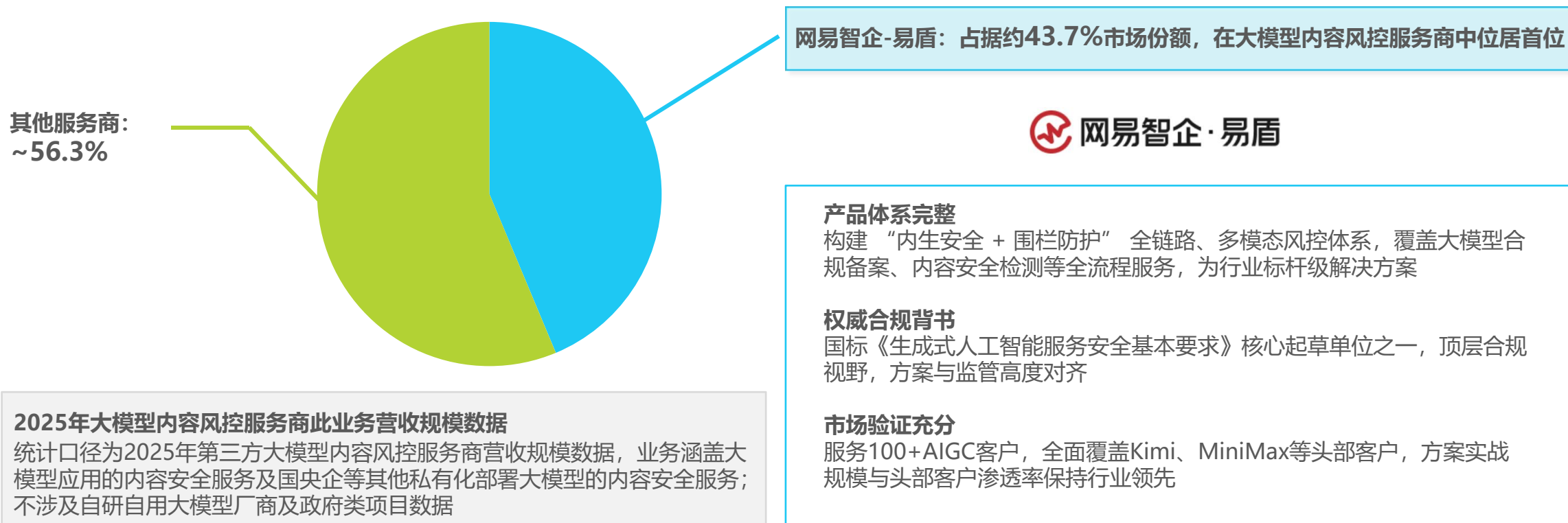


来源：专家访谈，公开资料，艾瑞咨询自主研究绘制。

大模型内容风控行业竞争格局

网易智企-易盾以约43.7%市场份额位居大模型内容风控服务商首位，其标杆级大模型内容安全方案的实战规模与头部客户渗透率均领先行业

2025年中国大模型内容风控市场竞争格局



数据来源：专家访谈，企业内部资料，艾瑞咨询自主研究绘制。

网易智企-易盾一大模型内容安全服务

网易智企-易盾大模型内容安全服务以“主动免疫”为核心，构建起内外兼修、全链路覆盖的产品体系，在基础模型、金融、教育等全行业大模型场景中广泛应用，为模型内容安全提供坚实保障

外防风险：外部安全围栏依托“全链路防御、多模态协同、动态适配”能力实现输入、处理、输出的全流程防御闭环；**内强素质：**内生安全以安全语料等服务助力模型源头优化安全能力；内外兼修的一站式服务体系，护航智谱AI、Kimi、MiniMax、天工AI等行业头部客户，主流模型场景全覆盖，成为大模型规模化落地不可或缺的安全底座之一。

大模型内容安全方案



一、源头免疫：

在模型训练期净化“AI基因”

语料安全 安全质量

安全领域语料	语料安全评估
语料标注服务	生成内容安全评估
语料安全过滤	应用拒答评估

清除数据毒素，注入安全基因，构筑内生安全底座

二、合规筑基：

在产品上线期跨越“准入技术关”

安全功能 安全咨询

智能验证	算法备案
内容标识识别	大模型备案
投诉举报	安全标准制定

贴合政策法规，落实合规能力，跨越准入门槛

三、动态御敌：

在产品运营期部署“双向安全围栏”

内容合规 安全生态

输入安全	安全测评
输出安全	安全培训
提示词攻击	黑产识别

双向拦截防御，构建自愈生态，动态对抗新风险

内外兼修：内生安全+外部围栏

从模型训练到上线运营的全生命周期闭环，真正实现安全与创新的双轮驱动

数据来源：专家访谈，公开资料，艾瑞咨询自主研究绘制。

最佳实践1

基础底座模型——合规与体验平衡点挖掘

场景挑战：提示词攻击升级与合规监管趋严下，确保合规同时保障用户体验，并应对多变的风险环境

解决方案：AI检测识别高危内容，结合模型内生安全优化体验，提供行业合规咨询与动态信息同步，常态化监测迭代模型，实现安全与体验的精准平衡

业务价值：

数据识别准确率	99.5%	风险数据召回精度	96%+	风险浓度下降	70%+
---------	-------	----------	------	--------	------

最佳实践2

AI教育大模型——K12场景效果攻坚

场景挑战：K12教育场景内容安全审核标准要求严苛，部分内容边界模糊易误判漏判

解决方案：标准及模型阈值适配，8周高强度重保，累计标注教育垂直场景数据超20万条，构建精细化训练数据集，针对性优化正负向内容识别逻辑

攻坚成效：8周内实现风控能力跨越式提升，核心指标表现显著改善

数据识别准确率	88%	98.5%	风险数据召回精度	80%	99.3%
---------	-----	-------	----------	-----	-------

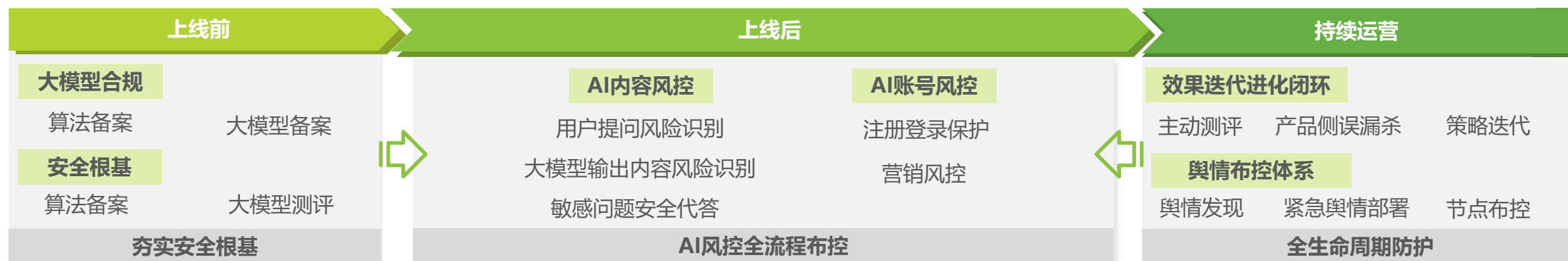
高效满足教育场景安全审核需求，为教育大模型安全落地提供坚实支撑

数美科技—大模型内容安全解决方案

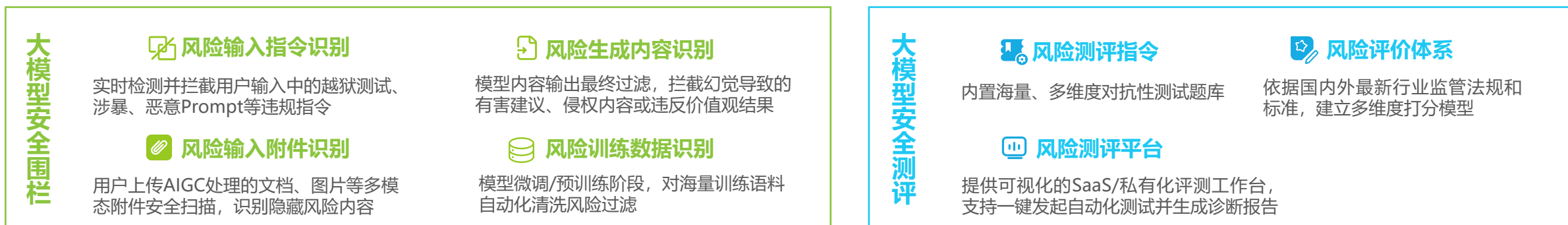
数美科技大模型内容安全解决方案，体系覆盖AIGC全生命周期的业务风控，实现效率与合规有效平衡

针对AIGC内容合规及运营风险痛点，数美科技结合新一代审核大模型，打造了覆盖AI应用全栈的大模型内容安全解决方案，有效为企业筑牢安全围栏。方案在多行业实际落地中展现出显著的赋能价值，同时凭借卓越的治理能力，获评2025年人工智能+大会“人工智能+”治理能力赛道“AI100应用标杆”

AIGC全周期内容安全解决方案——让AIGC应用全流程内容风险更合规，严密防护为企业筑牢内容安全坚实围栏



天枢——大模型安全围栏与安全测评双重核心能力，助力模型风险识别

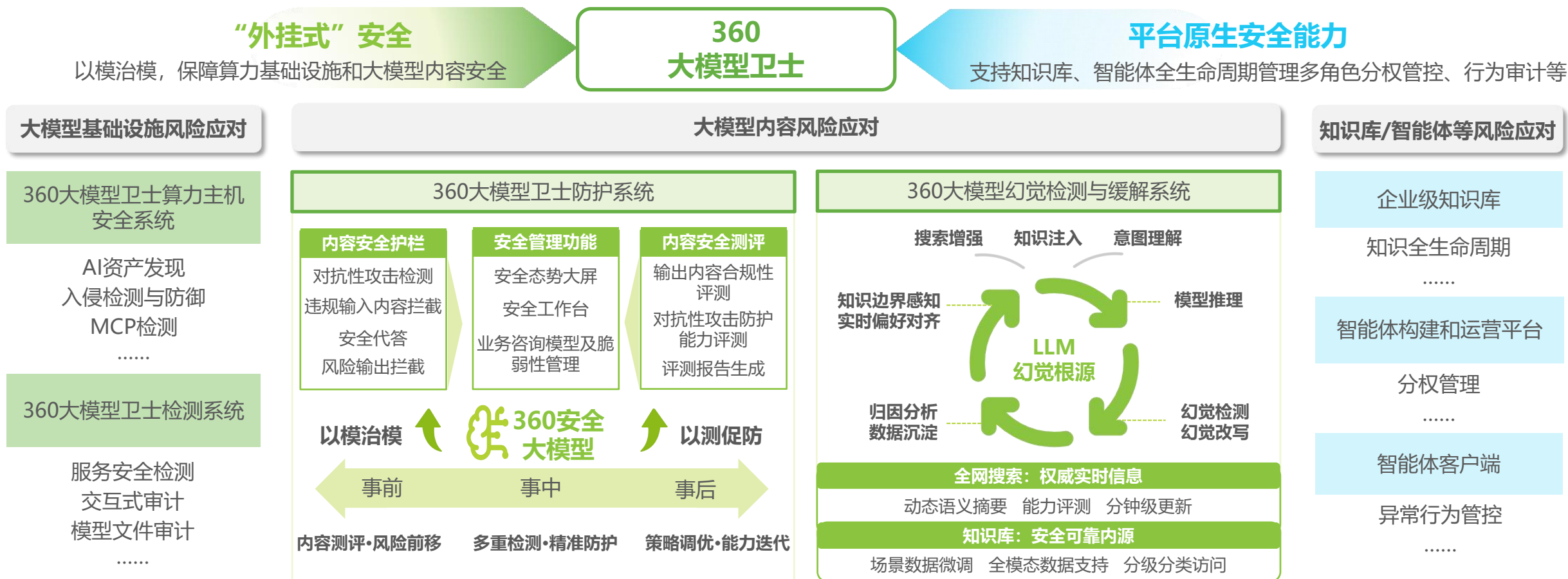


数据来源：专家访谈，公开资料，艾瑞咨询自主研究绘制。

360—大模型卫士

360大模型卫士，以“以模治模”为核心，采用外挂式安全 + 平台原生安全策略，构建全面的AI安全保障体系

大模型卫士含基础设施、大模型内容、知识库等安全产品，实现AI“外防风险、内固架构”的全面防护；针对大模型内容风险，**360大模型卫士防护系统**，以内容安全护栏与测评，实现以模治模，以测促防；**360大模型幻觉检测与缓解系统**，以检索增强，私域知识库等技术提升内容可靠性，多产品协同为模型内容安全保驾护航。



数据来源：专家访谈，公开资料，360《大模型安全白皮书》，艾瑞咨询自主研究绘制。

行业未来发展趋势

内容风控正从被动合规防守转向数字经济的核心安全基建，行业规模持续高速增长；技术层面，在 AI 攻防博弈中构建全域智能安全防线

内容风控市场发展趋势——扩容迭代

内容风控由“被动防守”转向“安全基建”，行业持续扩容

互联网内容指数级增长，合规刚需成为行业发展最强推手，平台进行由“流量”到“真实”的价值重构，内容安全本身成为生产力。

内容风控成为数字经济的核心基础设施之一，行业规模保持高速增长。

市场版图重塑

通用市场

掌握源头治理和全链条追溯能力，具有强大算力与技术底座支撑的高阶风控服务商将主导大规模、高并发的通用内容风控市场

垂直场景

深耕垂直场景专业风控服务商，以行业“know-how”，如金融场景交易欺诈、医疗领域深度伪造等积累，主导垂直场景持续深化

内容风控技术发展趋势——技术智能化

AI对抗AI下“盾与矛”攻防博弈，技术升级战

内容安全领域陷入“AI对抗AI”的无限博弈，智能重塑规则，风控技术体系在“盾与矛”的螺旋对抗中持续升级，实现全域数字风控，筑牢安全防线。

当前技术体系

多模态融合检测
AIGC检测

主要痛点

AI伪造对抗性强
跨平台难溯源
长文本AI含量难鉴别
.....



全域智能
安全大脑

原生智能

原生多模态大模型审核
Agent

动态演进

动态防御体系：AI对抗
样本自动生成自我迭代

多端协同

隐私计算与跨平台风险
联防

行业未来发展趋势—应用场景

在交互场景多元化与监管要求的驱动下，内容风控体系应用边界正无限延伸，在通用互联网、大模型等核心场景持续深化同时加速拓展新兴场景，向全行业全场景渗透

内容风控应用场景——核心场景持续深化

通用互联网场景

社交、短视频、直播、电商、游戏等通用互联网场景持续深化



AIGC与大模型

模型内容全生命周期治理成为新兴核心场景，Agent逻辑、权限等应用拓展



知识产权与版权保护

侵权视频、盗版文字、抄袭图片、人脸、声纹等监测，各类内容产权与版权保护



未成年人保护

内容精细化，特定场景屏蔽；防沉迷与诱导



内容风控应用场景——垂直领域与新兴场景加速拓展

政企

金融

医疗

广告

学术教育

出版

短剧漫剧

企业出海：跨境合规

应对复杂国际合规环境，实现多语种、跨文化合规，平台规则适配



AIGC新媒体运营

AI媒体运营内容符合广告法，进行合规校验、意图洞察与价值导向



虚拟人/数字人交互

交互内容实时管控，数字人强制化身份标识，肖像权与声纹保护



虚拟现实与元宇宙

数字资产与3D空间安全，虚拟环境合规，深度交互监测



BUSINESS
COOPERATION

业务合作

官 网



微信公众号



新 浪 微 博



企 业 微 信



联系我们

-  400 - 026 - 2099
-  ask@iresearch.com.cn
-  www.idigital.com.cn
www.iresearch.com.cn

LEGAL STATEMENT

法律声明

版权声明

本报告为艾瑞数智旗下品牌艾瑞咨询制作，其版权归属艾瑞咨询，没有经过艾瑞咨询的书面许可，任何组织和个人不得以任何形式复制、传播或输出中华人民共和国境外。任何未经授权使用本报告的相关商业行为都将违反《中华人民共和国著作权法》和其他法律法规以及有关国际公约的规定。

免责条款

本报告中行业数据及相关市场预测主要为公司研究员采用桌面研究、行业访谈、市场调查及其他研究方法，部分文字和数据采集于公开信息，并且结合艾瑞监测产品数据，通过艾瑞统计预测模型估算获得；企业数据主要为访谈获得，艾瑞咨询对该等信息的准确性、完整性或可靠性作尽最大努力的追求，但不作任何保证。在任何情况下，本报告中的信息或所表述的观点均不构成任何建议。

本报告中发布的调研数据采用样本调研方法，其数据结果受到样本的影响。由于调研方法及样本的限制，调查资料收集范围的限制，该数据仅代表调研时间和人群的基本状况，仅服务于当前的调研目的，为市场和客户提供基本参考。受研究方法和数据获取资源的限制，本报告只提供给用户作为市场参考资料，本公司对该报告的数据和观点不承担法律责任。

THANKS

艾 瑞 咨 询 为 商 业 决 策 赋 能