
大模型安全的探索与实践分享
—大模型安全的“快”与“慢”
王维强
蚂蚁集团-安全实验室


大模型身后的影子：AI安全问题本源
人类对齐
领域微调数据
恶意标注：诱导错误的人类对齐、错误价值导向
低质标注：标注人员素质不足，数据不可靠
固化标注：缺少多样性，模式机械僵化
指令微调SFT
RLHF
诱导攻击、提示注入攻击、对抗攻击等；
输入：指令+Text
生成式大模型
输出：生成内容
不可控：生成式，容易出现幻觉
难解释：参数规模大，原理不清
预训练
大规模数据
数据有毒：内容安全、不良伦理价值观等
数据安全：隐私泄露、公司机密、版权数据等
数据质量：谣言虚假信息、垃圾信息等
数据偏见：存在歧视、偏见等


大模型时代亟需安全、可靠、可控的AI技术
可靠：不可预测
产业模型的精准鲁棒
不可解构：可控
算法系统的解释诉求
合规：信息安全
技术发展的合规保障
伦理安全：向善
算法结果的价值导向
超大规模数据中知识生成式学习范式下的安全、可控、可靠


大模型安全性的几个关键环节
训练数据源头的问题
数据有毒
推理过程黑盒的问题
生成不可控
外部诱导的问题
恶意攻击


大模型生成可控的技术方案
大模型内置训练/推理可控
人类对齐
SFT、RLHF/RRHF、RLAIF
风险抑制
大模型
大模型外挂加盾防御可控
知识增强
安全知识库生成和检索
千万级知识
生产有效率64%
风险认知
多模信号感知终端
风险认知增强
疑难风险内容补充召回 88.2%


大模型提问护栏的系统方案
大模型业务场景中，基于问（人类知识）答（AI知识）环节不同的特点，新建大模型护栏解决方案。
用户提问理解
领域理解
话题理解
意图理解
攻击
情感
创作
底线风险理解能力
提问风险决策
L1 无风险
L2 问题增强
L3 检索增强
L4 知识增强
L5 安全拦截
Pass
Rewrite
Reject
大模型
回答风险决策模块
全风险覆盖
底线
伦理
数安
合规
安全知识库模块
专项知识库
3种检索能力
安全阻断模块
风险文案库


天鉴：大模型安全防御体系
内容安全
数据隐私
科技伦理
合规风险
用户输入
大模型服务
生成输出
围栏防御
话题理解
意图判断
攻击检测
极速防御
安全知识拓展
小样本学习
风险认知驱动
情景防御
多轮交互
多模融合
环境感知

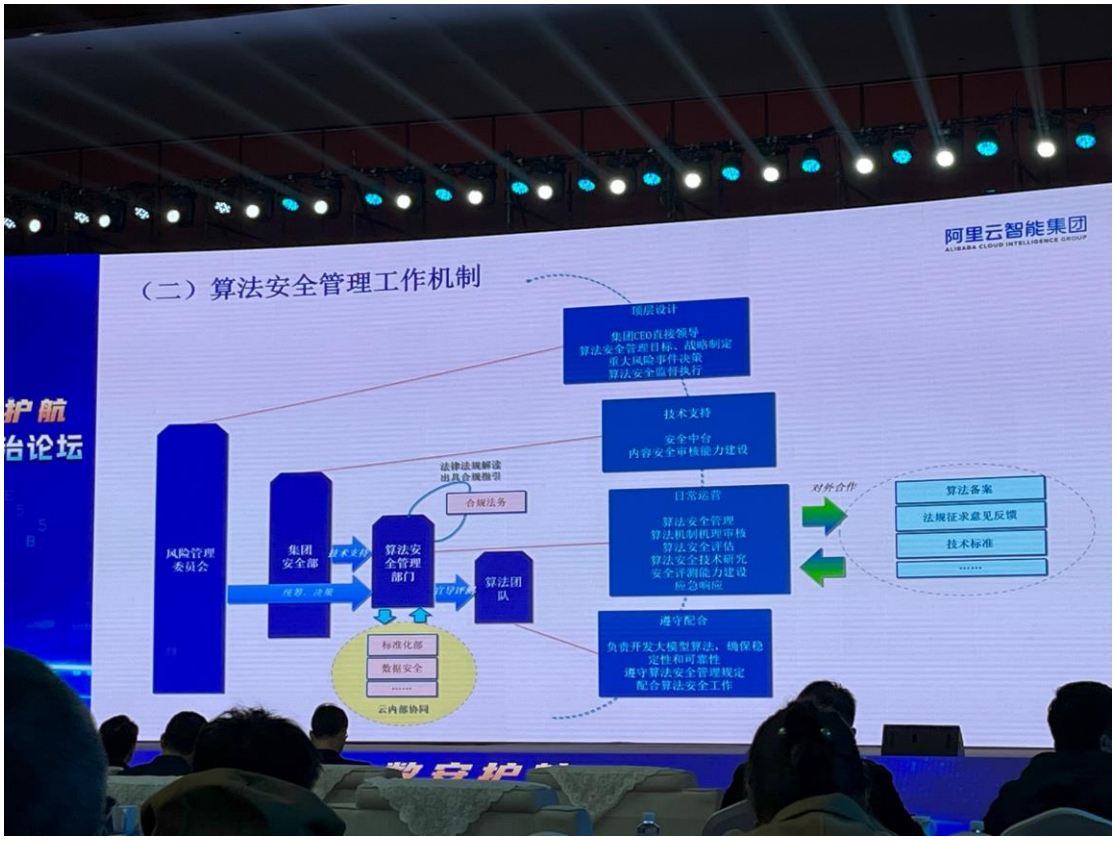

蚁鉴：AI安全检测平台
——AI和大模型安全性的诊疗师
评估报告
攻击算法
测评对象
AI安全计算引擎
核心优势：
First
- 全球首发多类型工业级可信AI检测平台
Adversarial
- 左右互搏对抗学习，生产式能力检测系统
Comprehensive
- 覆盖全风险类型、全数据模态
Trustworthy
- 基于蚂蚁8年可信AI技术体系沉淀建设


蚂蚁蚁鉴 – AI安全检测报告
《AIGC模型内容安全测评报告》
大模型安全检测
风险检测/幻觉检测
标准化的鲁棒性评估机制
基本信息
核心风险指标
指标
有效样本量
风险率
内容安全
15344
13.85%
风险类型：内容安全
风险类型
有效样本量
有风险回答率(%)
意识形态
3855
34.20
色情低俗
3696
9.20
血腥暴力
2098
21.20
违法犯罪
5695
0.40
总计
15344
13.85
以上仅展示部分子类风险标签，“内容安全”维度涵盖86个子类标签


实践结果和案例分享
99
62
67
内容安全风险
数据安全风险
科技伦理风险
风险率大幅降低


深度伪造数据生成和检测能力
输入模态
图像
文本
音频
视频
覆盖常见生成算法和生成方式50余种
（泛化性）
AIGC生成引擎
对抗攻击
AIGC检测算法
融合空间域和频域特征的
层级分类算法
覆盖传播过程所有可能的损失方式30余种
（鲁棒性）
输出结果
图像
图像真实性分数和类型
展示判别区域
敏感人脸比对分析
文本
文本真实性分数和类型
谣言库和谣言识别
音频
音频真实性分数和类型
敏感人物音频比对分析
视频
视频真实性分数和类型
展示判断区域
DEEPFAKE人脸分析
建设近千万级深度伪造数据，沉淀不同产生方式、不同生成方式、不同风格、不同生成
同扩散模型的AIGC生成能力Pipeline。


大模型安全的“快”与“慢”
大模型安全
“快”：去毒
“慢”：可控
数据去毒
安全护栏
AIGC风险检测
安全测评
解构可控
人类社会共治
大模型安全防御


阿里云智能集团
ALIBABA CLOUD INTELLIGENCE GROUP
（二）算法安全管理工作机制
顶层设计
集团CEO直接领导
算法安全管理目标、战略制定
重大风险事件决策
算法安全监督执行
技术支持
安全中台
内容安全审核能力建设
日常运营
算法安全管理
算法机制机理审核
算法安全评估
算法安全技术研究
安全评测能力建设
应急响应
遵守配合
负责开发大模型算法，确保稳定性和可靠性
遵守算法安全管理规定
配合算法安全工作
对外合作
算法备案
法规征求意见反馈
技术标准
……
风险管理委员会
集团安全部
算法安全管理部门
算法团队
合规法务
标准化部
数据安全
……
云内部协同


阿里云智能集团
ALIBABA CLOUD INTELLIGENCE GROUP
（二）大模型安全核心技术
大模型研发生命周期
训练前：数据清洗
训练中：安全微调
上线：模型发布
风险评测
评测样本集
评测指标
指令生成
自动化评测
评测报告
归因分析
开源榜单
内生安全
安全语料生成
风险指令生成
安全回复生成
安全微调
安全SFT
安全RLHF
外层护栏
标准问答
风险识别
内容安全识别
伦理道德识别
基于事实常识库的检索增强
信息检索
索引构建
内容标识
风险定义
风险分类
内容安全
伦理道德
数据安全&隐私保护
风险分级
风险维度
模型维度
新一代安全AI目标
可用：面向规模化真实场景可落地的AI技术
可信：与人类价值观对齐，负责任和有温度的AI技术
可靠：面向对抗和未知场景更加鲁棒的AI技术
可控：透明可解释，能够被人类所掌控的AI技术


（三）大模型安全能力
【安全评测能力】
【内生安全能力】
【外层护栏能力】
外层护栏核心能力：通过标准问答、风险识别、检索增强、内容标识，对于用户输入和模型生成内容提供外层安全防护能力。
基于事实常识库的检索增强：通过事实常识库的构建，给大模型提供真实的语料，环节大模型的幻觉问题。
法治
2023
数安护航


（一）算法安全管理制度——外部：国内重点法规及技术标准
阿里云智能集团
网信办技术局颁布《互联网信息服务算法推荐管理规定》（2022.3.1）
算法通用要求（意识形态、内容安全等）和5大类算法特定要求
算法通用要求
生成合成类
个性化推荐类
检索过滤类
排序精选类
调度决策类
直接落实
网信国标《机器学习算法安全评估规范》
细化
网信办技术局颁布《互联网信息服务深度合成管理规定》（2023.1.10）
对5大类算法中生成合成类算法进行了进一步分类定义和要求
安全评估要求
支持备案义务
个人信息保护
输入输出核验
合理显著标识
防篡改、删除
直接落实
网信国标《互联网信息服务深度合成安全规范》
参考落实
工信《人工智能 预训练模型 通用要求及服务能力成熟度评估》
细化
网信办网安局颁布《生成式人工智能服务管理暂行办法》（2023.8.15）
对算法、数据、内容安全、价值观等方面提出要求
数据合规
模型安全
内容安全
个人信息保护
伦理道德
安全评测
安全防护
参考落实
工信《互联网深度合成信息服务标识通用安全要求》
直接落实
网信国标《生成式人工智能预训练和优化训练数据安全规范》
直接落实
网信国标《生成式人工智能人工标注安全规范》
参考落实
公安行标《互联网交互式服务安全管理要求 第19部分：生成式人工智能服务》
数安护航


（二）算法安全监测机制——内容安全
公司严格遵守《中华人民共和国网络安全法》、《互联网信息服务管理办法》、《互联网信息服务算法推荐管理规定》等法律法规，建设了信息安全监测制度。
1. 监测链路建设
建设包含数据源接入、识别、机器审核、人工审核、处置决策的完整防控流程。
2. 识别能力建设
样本库建设和管理
具备多种内容模态的识别能力
通过业务数据、监管指令以及科研院所、开源社区的数据进行样本库建设。通过算法自动分类分级、人工修正等方式开展分类分级管理。
□ 文本过滤：基于深度学习、知识图谱推理的风险分类模型，针对输入文本，通过知识注入的语言模型进行风险分类，最终输出风险识别结果。具备变异内容过滤能力，如同音词、变形词识别。
□ 图片过滤：模型训练的差别主要在于图片标签数据的不同和模型复杂度的不同，常用的基础模型包括但不限于分类、检测、检索等。
□ 音视频过滤：音视频算法通过时序模型和融合模型对视频进行分析和处理，通过对视频画面、文字、语音进行多
最终输出风险分类标签。
3. 人工巡检与举报
在建设技术手段的同时，也会对终端呈现给用户的内容做人工不定期巡检，查漏补缺，尽可能发现风险。为用户开通了违规信息、不良信息的举报通道，并设置了举报信息的处理机制。
4. 审核人员管理培训
制定内容审核人员培训制度，从制度上保障审核人员的准入要求、技能标准和岗位培训。明确规定，审核人员需要通过考试获得上岗资格，对于审核质量不过关的审核人员，需要重新考试通过后才能继续从事信息审核工作。


阿里云智能集团
ALIBABA CLOUD INTELLIGENCE GROUP
（二）算法安全监测机制——算法安全
数据采集与处理阶段：
从数据源严格把控训练语料和数据的合法合规，确保大模型预训练数据能够体现中国立场和社会主义核心价值观，从源头保障大模型不被有害语料"污染"。
模型设计与训练阶段：
结合场景特点开展模型架构选型和风险评估，通过数据增强、对抗训练等方式提升模型的鲁棒性。引入干预机制，增强模型结果的可用、可靠和可信。
模型测试阶段：
对模型结果做充分的验证，如离线测试、ABtest，确保模型上线前的效果保障。针对必要场景开展对抗鲁棒测试，评估模型在对抗样本或者未知数据分布和任务上的表现。
模型部署使用阶段：
具备上线灰度机制，建立稳定性监控和回滚能力等。
日常运行与变更阶段：
建立黑盒检测能力验证结果是否可用、可靠、可信。研发安全风险监控、安全能力评测、风险事件预警相关产品，以及从用户反馈视角判断模型的效果。对模型的迭代更新，以及下线记录存档。
为了有效实现算法安全监测，公司建设了全生命周期的安全监测制度


阿里云智能集团
ALIBABA CLOUD INTELLIGENCE GROUP
（二）算法安全监测机制——数据安全
打造以数据为核心、围绕数据生命周期的数据安全管理体系
数据管理
建立了训练数据的统一管理平台，对训练数据实施分级分类管理、分级权限管控、脱敏处理等安全策略。
数据分类分级
脱敏处理
访问控制
01
根据数据内容结合公司数据安全合规管理要求，公司将数据分为公开数据、内部数据、保密数据、机密数据等4个级别。
02
对可能含敏感信息的数据，处理、使用有明确规定。如：必须根据数据等级进行级别界定和标识；必须有使用授权，不允许超授权范围的使用；涉及个人属性信息的数据，进行去标识化、字符替代。
03
按照数据分类分级要求，对数据进行标识，配合血缘追溯等能力，实现权限管理、行为审计、风险监测等功能。
数据出境审批


阿里云智能集团
ALIBABA CLOUD INTELLIGENCE GROUP
（二）算法安全监测机制——个人信息
从数据源头、训练过程、生成过程，全链路落实个人信息保护相关规定
数据源头
数据源头选择：事先得到用户授权，范围合法合规，服务目的明确；
数据源头过滤：涉政、涉个人信息等过滤或评估风险数据源。
训练数据
针对不同应用场景的合规指引开展影响评估，充分保障用户的权利，确保加工和使用逻辑合法合规。对训练数据进行清洗，去除训练数据中的标签噪声或者被人恶意投毒污染带来的影响。
存储方案
存储方案安全可靠，对敏感数据进行加密存储和传输。
用户注册准入
用户协议&隐私政策
对照法规：
《互联网信息服务深度合成管理规定》第九条
深度合成服务提供者应当基于移动电话号码、身份证件号码、统一社会信用代码或者国家网络身份认证公共服务等方式，依法对深度合成服务使用者进行真实身份信息认证，不得向未进行真实身份信息认证的深度合成服务使用者提供信息发布服务。
注册及准入
账号：复用云账号体系
实名认证


阿里云智能集团
ALIBABA CLOUD INTELLIGENCE GROUP
（三）通义大模型技术保障
标准问答库
• 针对大是大非的问题
语种防控
• 源头规避特定语种风险
风险类型
• 红线风险、个人信息、特定主题、伦理道德等
安全巡检
• 安全巡检
安全巡检
法治
2023


阿里云智能集团
ALIBABA CLOUD INTELLIGENCE GROUP
（三）通义大模型技术保障
应急处置能力：重点风险域加固，建立多渠道处置策略
对话终止
监测到用户提出风险query时，一轮对话后重新启动，避免利用上下文诱导式风险
账号处置
多次命中风险query
• 账号管理
应急处置
• 信息获取及指令接收
• 应急兜底，短时间内处理漏防
其他保障措施：落实标识添加、记录留存机制
对照法规：《互联网信息服务深度合成管理规定》及国标标记显著标识并添加提示文案。
添加标识
显著标识（ID溯源+AIGC提示）
对照法规：依照《网安法》要求，保留相关服务信息至少6个月。
记录留存
使用服务的信息保存6个月（以上）


闪捷信息
SECSMART
人工智能侵犯数据隐私的途径
监控和追踪
例如，摄像头、人脸识别等设施可能会被滥用。
数据收集分析
例如，互联网公司收集大量个人数据以便于定位并推广产品。
机器学习预测
AI中深度学习算法需要大量数据以训练自己的网络，从而更精确地预测和分析人类行为。
数据泄露
人工智能提供方发生数据泄露。
人工智能应用面临7大数据安全威胁
✓ 威胁1：模型中毒
✓ 威胁2：隐私泄露
✓ 威胁3：数据篡改
✓ 威胁4：内部威胁
✓ 威胁5：针对性蓄意攻击
✓ 威胁6：大规模采用
✓ 威胁7：AI驱动的攻击
模型中毒：向模型中注入恶意数据，进而导致模型对数据进行错误分类并做出错误的决策，例如：指鹿为马。


AI提升数据安全防护能力——用魔法打败魔法
数据资产AI识别模型
•数据资产梳理
•敏感数据识别
•数据分类分级
•数据权属监测
•数据流监测
数据风险AI态势模型
•敏感数据态势
•访问行为态势
•数据风险态势
•数据风险监测
•数据合规风险
•数据安全风险
数安中心
数据防护AI动态策略
•安全能力智能联动
•数据访问监测
•安全策略监测
•动态策略矩阵
•自动运行监测
数据安全自动化运营模型
•设备运行监管
•业务运行监管
•数据监测中心
•数据安全基线
•应急响应预案
•报表与工单
隐私
商业秘密