# 中国区AI内容安全重构版工作计划（2026）

全面规划未来内容安全保障方案

# 演讲主要内容

- 设计主张与核心原则
- 架构总览与安全栈设计
- 中台平台集成与任务分解
- 度量体系与最佳实践
- 组织分工与风险管理
- 验收标准与关键映射
- 后续计划与落地安排

# 设计主张与核心原则

# 分层防护与策略编排

### 分层纵深防护

采用输入、推理、输出三阶段安全门控，实现全面防护与证据留存，保障系统安全。

### 策略模型闭环

策略、模型与运行时形成闭环，确保安全策略完整执行与动态调整。

### 统一策略编排器

中台标签驱动动作纳入统一策略编排，实现自动化与协同防护管理。

# 最小权限与密钥管理

### 应用级密钥访问

所有业务访问中台必须使用应用级密钥确保安全和权限最小化。

### 策略白名单机制

通过策略白名单限制访问权限，防止未授权操作和数据泄露。

### 密钥有效期与轮换

定期轮换密钥及监控其有效期，提升密钥管理安全性并纳入审计。

### 审计主键跟踪

将密钥使用和轮转纳入审计，确保安全合规与追踪。

# 成本感知与实时流式安全

### 成本感知自动缩放

安全计算根据风险权重和上下文片段优先级自动调整资源分配，提升效率。

### 实时流式安全

通过实时流处理保障安全，适应不断变化的风险和数据流特性。

### TTFT优化机制

TTFT采用乐观并发和后截断策略，提升安全计算的并行效率和准确性。

# 可验收与运营一体化

### 可验收标准

评测集、口径及统计方法确保结果可复现。通过随机抽测实现稳定达标。

### 分层KPI解释

输入、推理和输出层次的KPI具备良好可解释性，便于准确监控。

### 安全即运营

通过中台工单流转和消息队列，实现封禁解禁和申诉回调的安全运营。

### 高风险处置SLO

与SIEM和SOAR系统对齐，形成高风险事件快速响应和处置的服务级别目标。

# 架构总览与安全栈设计

# 接入与策略网关机制

### 密钥管理机制

应用采用四步法创建与发放密钥，网关校验密钥有效期与策略版本一致，自动触发轮换。

### 端云日志审计

端侧SDK与云侧API并存，端侧日志上传中台，统一进行安全审计管理。

### 网络出站控制

控制对AICS开放接口的调用，需判定网络域一致性，非一致情况需审批开墙操作。

# 标签编排器与动作映射

### 内容安全标签映射

中台内容安全标签201至207映射为不同的运行时安全动作，确保内容合规。

### 多动作策略

根据标签，动作包括通过、拦截、不显示及模板替换等多种安全处理方式。

### 统一策略执行

标签驱动动作在不同判定等级统一下发，保证策略、模型与产品一致执行。

# L0–L3安全重构与流式门控

### L0标签初判

通过规则、词典和轻量模型进行风险标签初步判断，实现风险先验和早期拦截。

### L1片段筛选

利用信息密度和向量检索挑选top-k片段，生成标签候选集供后续复核。

### L2复核裁决

进行对抗越狱识别和复合策略检测，完成标签的最终裁决与结构化处理。

### L3工单与闭环

对高风险样本触发工单，消息通过Kafka广播，结合人工受理实现结果回调闭环。

# 中台平台集成与任务分解

# 中台平台双向集成

### 工单系统管理

所有安全事件通过中台工单系统流转，账号授权和安全SLO绑定生命周期。

### 用户处置消息队列

通过Kafka订阅封禁及解禁命令，执行状态纳入统一仪表盘管理。

### 申诉服务流程

申诉通过举报中心采集，人工处理后回调API返回结果，记录处理时长和复议率。

### 策略自测与质检

AICS与安全评测入口用于上线前自测和模型质量检查，作为发布前置关卡。

# 运行时优化与TTFT

### 切片器与标签生成

标签候选在L1阶段生成，显著减少L2判定时间，提高整体效率。

### 热路径缓存机制

新增标签热路径缓存，对常见风险类实现快速早停，提升筛选速度。

### Speculative Safety调度

采用首字乐观门控和兜底模板库，实现并发调度保障系统安全。

# 模型安全与合规抽测

### 拒绝策略蒸馏

通过标签裁决和拒绝模板将拒绝策略蒸馏到端侧模型，提升安全响应效率。

### 安全微调与检测

采用硬负样本越狱集进行安全微调，配合ROC/PR/AUC指标评估模型安全性。

### 解码时安全门控

利用高风险Token模式库和动态温度控制实现解码时安全门控，防止风险输出。

### 合规抽测与评审

红队攻防框架与抽测仪表板结合，支持采购评审中的合规性与风险管理。

# 应用安全与全流程治理

### 软件安全生命周期

SDL流程涵盖威胁建模、静态和动态分析、软件物料清单以及漏洞治理。

### 密钥与网关管理

包含KMS密钥轮换、出网白名单、速率限制以及熔断和审计机制保障安全。

# 度量体系与最佳实践

# 分阶段KPI与SLO指标

### 输入阶段指标

涵盖L0/L1召回率、误报率，以及标签候选命中率和覆盖度，保障数据质量。

### 推理阶段指标

包括TTFT时间分布，安全扫描耗时与成本，以及安全截断率，确保推理效率。

### 输出阶段指标

关注越狱识别率、拒绝一致性和误报/漏报分层统计，提升输出准确性。

### 运营与合规指标

涵盖抽测通过率、审计日志完整性、封禁SLA及申诉处理效率，保障运营合规。

# 对齐国际最佳实践

### 安全评测与基准

跨输入、推理和输出的安全评测管线确保一致的策略回归与事件影响评估。

### 可解释拒绝策略

所有拒绝均提供用户可读解释和安全替代建议，提升用户体验并减少硬拒绝。

### SAIF风格控制

通过数据治理、模型供应链和身份访问控制，实现监测与响应的闭环管理。

### 红队常态化协同

法务、合规与技术红队统一管理，标签裁决确保治理一致性和风险控制。

# 组织分工与风险管理

# 研发与运营分工

### 研发A：运行时性能

专注于切片、并行、缓存和Speculative技术，优化运行时性能和标签编排。

### 研发B：机器学习安全

开发轻量模型和越狱检测算法，确保模型安全并进行一致性训练。

### 研发C：模型内生安全

实施拒绝蒸馏和模型权重签名，保障模型加载和使用的安全性。

### 研发D与经理：平台与合规

负责应用安全、密钥管理、审批流程及合规抽检，确保系统安全稳定运行。

# 主要风险与缓解措施

### 资源紧张管理

启用L0+L1降级策略，仅高风险片段进入L2，建立风险缓存与标签热路径保障性能。

### 规则与词典维护

通过黑样本和规则增量自动更新回归测试管线，保证标签一致性和规则有效性。

### 误报缓解措施

对低风险误报采用提示式缓释，减少硬拒绝并记录误报来源与标签差异。

### 合规与网络风险

策略配置化热更新，失败样本红队回归，密钥到期及审批失败纳入熔断与监控。

# 验收标准与关键映射

# 验收门槛与达标要求

### 性能指标

P90响应时间需低于1.5秒，保证高效的并发处理和系统性能。

### 成本优化

平均每请求的安全计算量下降至少60%，显著降低运营成本。

### 检测准确性

越狱识别率提升30%以上，误报率控制在2%以内，确保检测质量。

### 合规与运营

随机抽测稳定合格，申诉处理时长符合理想标准，审计日志满足PIPL要求。

# Checklist关键映射

### 接入流程

描述从AI Force到正式调用的完整接入流程，确保策略网关的集成。

### 标签与动作管理

标签编号201-207对应多种动作，涵盖安全模型和本地知识库，纳入标签编排器。

### 消息队列与广播

消息队列处理封禁和解禁广播，并带有回执确认，集成处置总线。

### 申诉回调流程

从举报中心到人工受理再到回调结果，申诉流程已纳入运营SLO管理。

# 后续计划与落地安排

# 下一步工作与落地评审

### 文档章节排版

文档按原则、架构、任务、度量等章节排版，便于直接插入主文档使用。

### 落地评审流程

采购评审问题清单及接口作为上线门槛，先在AICS平台进行回归测试。

### 周度例行复盘

经理主持月度抽测及复盘，结合红队和工单SLO确保风险和预算闭环。

# 总结与展望

### 打造内容安全体系

计划旨在构建一个安全且高效的内容管理体系，保障内容合规与安全。

### 持续优化创新

通过不断优化和技术创新，提升内容安全的效果和行业竞争力。

### 引领行业发展

目标是引导行业向更安全、高效的内容管理方向发展。