AI 安全技术白皮书

面向生成式 AI、RAG 与智能体系统的治理、架构与工程实践

图1 AI安全不是单点技术，而是覆盖组织治理、工程控制、模型行为和运行反馈的系统工程

版本 v1.0 | 2026 年 5 月

适用对象：CTO/CISO、AI 平台团队、安全团队、数据治理与合规团队

版权与使用说明

本文档为通用技术白皮书草案，旨在帮助组织建立 AI 安全治理与工程控制体系。文档内容不构成法律意见、合规认证承诺或替代第三方审计结论。实际落地时，应结合业务场景、区域法规、数据分级、模型供应链和组织风险偏好进行裁剪。

文档中的控制项参考了公开框架和标准的思想，包括 NIST AI RMF 与生成式 AI Profile、OWASP Top 10 for LLM Applications、MITRE ATLAS、ISO/IEC 42001、ISO/IEC 23894、NIST SSDF、欧盟 AI Act 以及中国《生成式人工智能服务管理暂行办法》。

目录

1. 执行摘要

生成式 AI 正在从“问答工具”升级为“可调用工具、可访问数据、可执行任务”的生产系统。它带来的安全挑战不再局限于传统 Web 漏洞，也不仅是模型拒答是否准确，而是覆盖数据、模型、应用、工具、身份、供应链、合规和运营的系统性风险。

本白皮书提出一套“治理 + 架构 + 工程 + 评测 + 运营”的 AI 安全体系，目标是在不阻碍创新的前提下，使 AI 系统达到可用、可控、可追踪、可审计和可持续改进。

- AI 安全目标应从单一“内容过滤”升级为“五维可信”：安全性、稳健性、隐私与数据保护、合规与伦理、可靠与可解释。
- 所有 AI 能力应进入统一资产台账：模型、数据集、向量库、提示模板、插件/工具、代理工作流、评测集和外部供应商。
- 面向 LLM 应用，应优先治理提示注入、敏感信息泄露、供应链、数据/模型投毒、不当输出处理、过度代理、系统提示泄漏、向量与嵌入弱点、幻觉/误导和无界消耗等风险。
- 工程上应以 AI 安全网关为控制平面，把身份、权限、提示策略、内容安全、工具调用、审计、速率限制和人工审批统一化。
- RAG 系统必须实现 ACL 感知检索、文档来源评分、向量库租户隔离、检索结果净化和引用可追溯。
- 智能体系统必须默认最小权限、动作确认、预算限制、沙箱执行、任务分解审计和可回滚。
- 上线门禁应包含安全红队、隐私评估、合规审查、业务影响评估、可观测性验证和回滚方案。

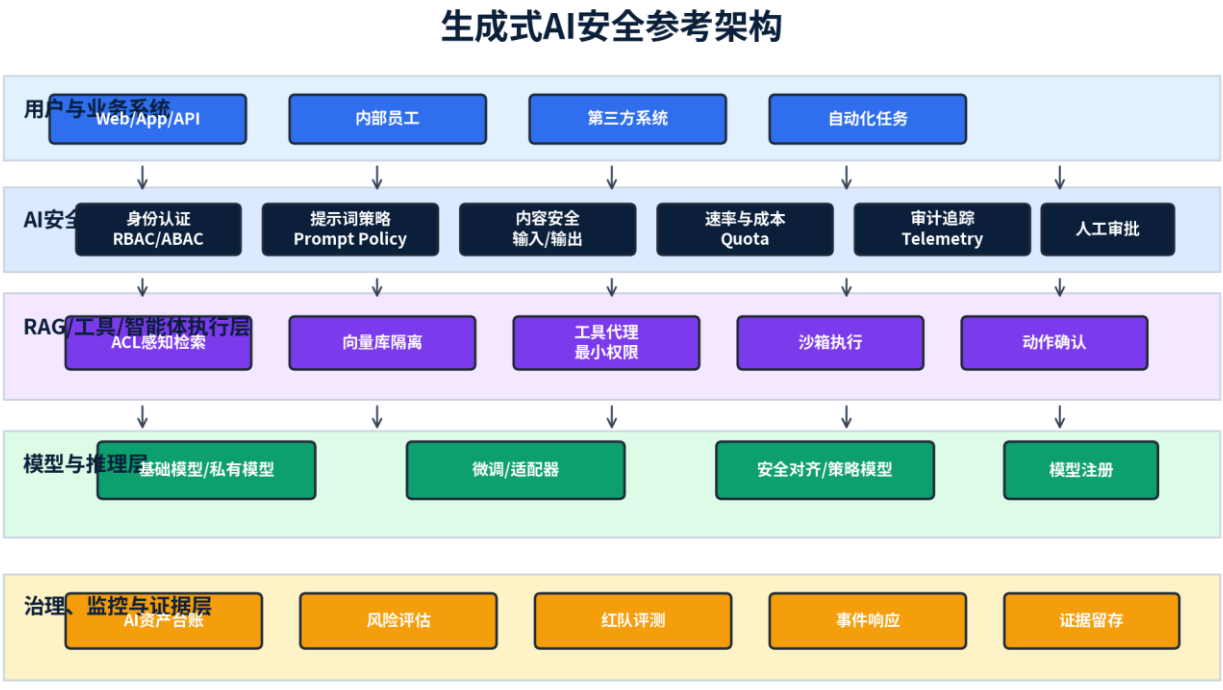

图2 推荐将安全能力前置到AI网关，并把RAG、工具、模型和审计统一纳入控制平面

图 2 生成式 AI 安全参考架构

2. AI 安全的定义与范围

2.1 定义

AI 安全是指在 AI 系统全生命周期内，对模型行为、数据流、工具调用、系统集成和组织流程进行识别、保护、检测、响应与恢复的能力集合。它既包括传统网络安全，也包括模型特有风险和社会技术风险。

维度	核心问题	典型风险	关键控制
安全性	系统是否能抵御攻击与滥用	提示注入、越权工具调用、供应链漏洞、DoS	身份与权限、输入/输出防护、供应链审计、速率限制
稳健性	模型在分布外、对抗或异常输入下是否稳定	越狱、对抗样本、恶意文档、低置信输出	鲁棒性测试、拒答策略、置信度与人工复核
隐私与数据	训练、检索和推理是否泄露敏感数据	PII 泄露、训练数据记忆、日志泄露、跨租户污染	数据分级、脱敏、DLP、最小化采集、加密
合规与伦理	是否满足法规、行业规则和组织价值边界	歧视、版权争议、未披露 AI 生成、儿童保护不足	合规评估、内容标识、偏见测试、使用边界
可靠与可解释	输出是否可验证、可追踪、可审计	幻觉、错误建议、不可复盘、责任不清	引用来源、日志、模型卡、评测指标、责任人制度

2.2 系统边界

在企业环境中，AI 系统通常由基础模型、推理服务、提示模板、RAG 检索、业务工具、自动化代理、日志监控和人工流程组成。安全边界不应只定义在模型 API，而应覆盖以下对象：

- 数据对象：训练数据、微调数据、评测数据、业务知识库、向量嵌入、用户输入、模型输出、日志与反馈。
- 模型对象：开源模型、闭源 API、微调模型、适配器、策略模型、内容安全模型、嵌入模型。
- 应用对象：聊天应用、Copilot、内部知识助手、客服机器人、代码助手、报表生成、风控辅助。
- 执行对象：插件、API 工具、浏览器、代码解释器、RPA、数据库查询、工单系统、支付/采购/审批动作。
- 治理对象：资产台账、风险登记册、审批记录、评测报告、事件记录、供应商合同和数据处理协议。

3. 威胁模型与风险分级

3.1 主要攻击者与滥用者

类型	动机	能力	示例
外部攻击者	窃取数据、破坏业务、勒索或牟利	自动化扫描、提示注入、凭证攻击、供应链投毒	利用公开 AI 接口诱导模型泄露内部提示或调用工具
恶意用户/灰产	规避平台规则、生成违规内容、批量薅取资源	越狱模板、批量账号、代理 IP、脚本化调用	绕过安全策略生成欺诈文案或恶意代码
内部人员	越权访问、便利化泄露、影子 AI	合法账号、业务上下文、内部文档访问	把敏感合同上传外部模型或越权检索他人数据
供应链风险方	植入后门、诱导依赖、数据污染	模型/数据集/插件/镜像/SDK 供应链影响	开源模型权重被替换或插件请求外传数据
模型自身失效	非恶意但不可忽视	幻觉、偏见、过度自信、上下文混淆	在医疗、金融、法律场景给出错误建议

3.2 资产与攻击面

AI 系统的高价值资产包括敏感业务数据、系统提示词、私有知识库、向量数据库、模型权重、微调数据、工具凭证、工作流策略和审计日志。攻击面通常分布在输入通道、检索通道、模型接口、工具执行、输出通道和运营后台。

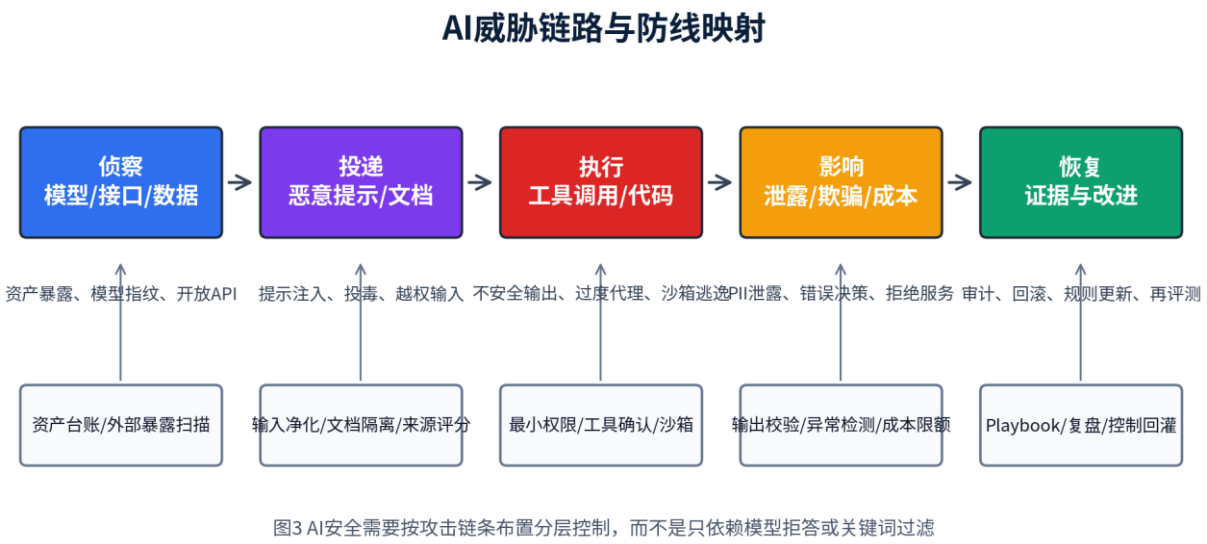

图 3 AI 威胁链路与防线映射

3.3 风险分级方法

建议采用“影响 × 可能性 × 暴露面 × 控制成熟度修正”的方式进行分级。影响维度包括财务损失、法律合规、个人权益、业务连续性、品牌声誉和安全后果；可能性维度包括攻击可达性、自动化程度、漏洞可利用性、攻击者收益和历史事件。

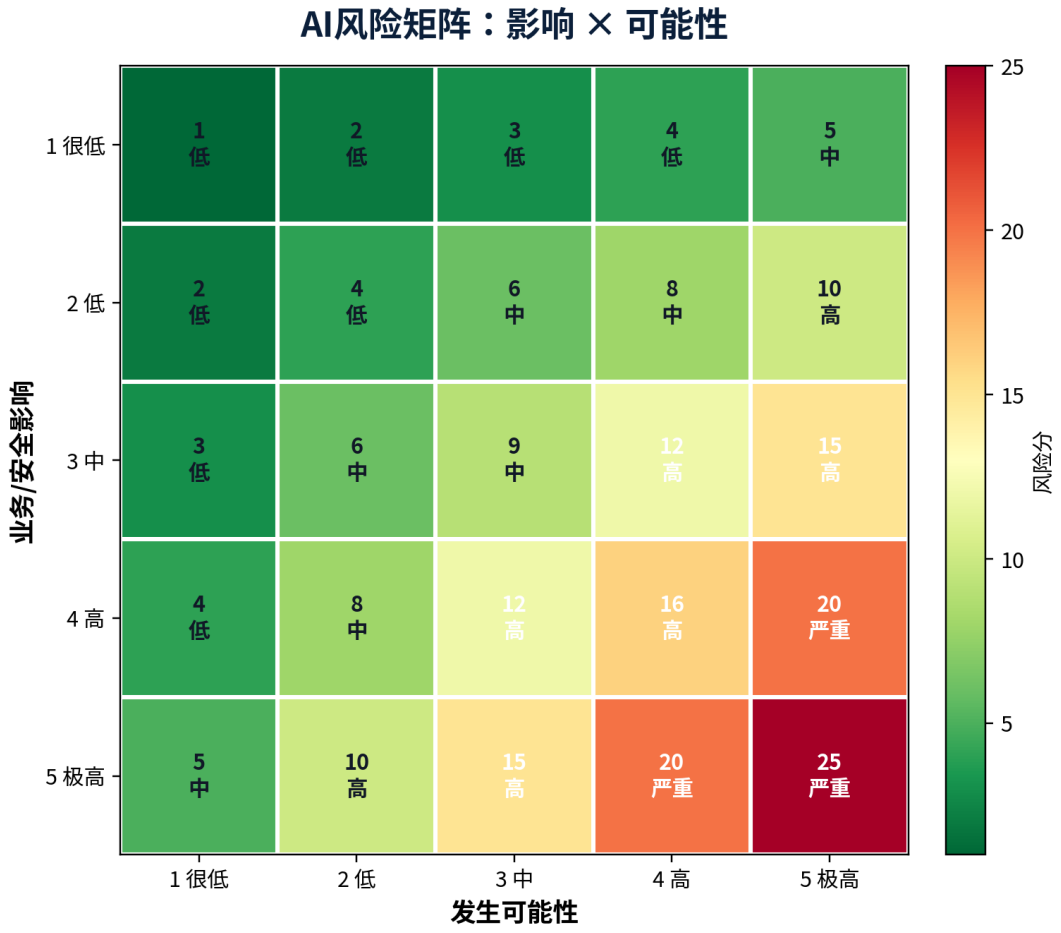

图 4 AI 风险矩阵

等级	判定条件	处置要求

严重	涉及高敏感数据、外部可达、可自动化利用，或 AI 可直接执行高影响动作	立即阻断或下线；启动事件响应；高层风险接受后才可恢复
高	可能造成越权访问、合规违规、重大错误决策或显著成本损失	上线前必须修复；需要安全负责人和业务负责人共同批准
中	影响有限但可重复利用，或依赖人工链路才能造成损失	进入整改计划；通过监控与补偿控制降低风险
低	影响局部、触发条件苛刻、已有有效控制	记录并周期复查；不阻塞上线

4. AI 安全参考架构

4.1 设计原则

- 默认不信任：用户输入、检索文档、模型输出和外部工具返回都视为不可信内容。
- 控制平面集中：身份、权限、策略、审计、速率限制和模型路由由 AI 安全网关统一实现。
- 数据面隔离：不同租户、部门、密级、项目和用途的数据、向量、日志应隔离。
- 最小权限和最小代理：模型只获得完成任务所需的上下文、工具和动作权限。
- 可观测与可回滚：每次输入、检索、推理、工具调用、输出和人工审批都可追踪。
- 人机协同：高影响动作必须设置人工确认或双人复核。

4.2 分层架构

层级	目标	关键组件	安全要求
入口层	验证访问者身份并限制调用范围	SSO、API 网关、WAF、Bot 防护、租户路由	强认证、设备与上下文校验、速率限制、异常登录检测
AI 安全网关	对提示词、上下文、输出和调用进行策略控制	Prompt Firewall、DLP、内容安全、模型路由、审计	策略版本化、可解释拦截、日志脱敏、旁路检测
RAG 层	安全地把知识注入上下文	文档解析、Embedding、向量库、Retriever、Reranker	ACL 感知、来源评分、文档隔离、检索结果净化、引用追踪
工具/智能体层	允许模型执行受控操作	Tool Broker、权限令牌、沙箱、审批流、预算管理	最小权限、动作确认、事务限额、执行隔离、回滚
模型层	提供推理与安全对齐能力	基础模型、微调模型、策略模型、嵌入模型	供应链验证、模型卡、版本控制、安全评测、漂移监控
运营层	持续监控风险并支撑审计	SIEM、AIOps、评测平台、事件平台、证据库	告警分级、Playbook、指标看板、审计留存、复盘闭环

4.3 关键数据流控制点

1. 输入前：身份、租户、场景、数据分类和用户意图识别。
2. 上下文组装前：文档来源、权限、敏感字段、提示注入特征和引用完整性检查。
3. 推理中：模型路由、系统提示完整性、上下文长度预算、推理成本预算、工具权限预算。
4. 工具调用前：动作分类、参数校验、权限令牌、审批策略和沙箱执行环境。
5. 输出前：事实核验、敏感信息检测、格式校验、业务规则校验、内容安全分类。
6. 输出后：日志脱敏、用户反馈、异常检测、事件关联和评测集回灌。

5. 关键技术控制

5.1 控制矩阵总览

控制域	控制说明	基线要求	核心指标
AI 资产台账	所有模型、数据集、向量库、提示模板、工具、代理流均登记	系统上线前无未知 AI 资产；每月复核	资产覆盖率、未登记资产数
数据分级与最小化	只把任务必要数据暴露给模型；敏感字段脱敏或令牌化	默认不上传高敏数据到外部模型	敏感字段泄露率、脱敏命中率
提示词与上下文防护	检测越狱、注入、角色混淆、数据外传指令	高风险输入进入拒绝、降级或人工复核	攻击成功率、误拦截率
输出安全处理	输出作为不可信内容，进入格式、内容和业务校验	禁止把模型输出直接拼接进 SQL、HTML、Shell 或审批动作	不安全输出拦截数、后端校验覆盖率
RAG 访问控制	检索时继承用户/租户/文档级权限	向量库与原文库权限一致；引用可追踪	越权检索事件数、引用准确率
工具最小权限	工具 Broker 按场景授予短期、细粒度能力	高影响动作需确认；禁止模型持有长期密钥	高危工具调用审批率、越权失败率
供应链安全	模型、数据、镜像、插件、SDK 进行来源验证和漏洞扫描	关键组件需签名、SBOM/AI-BOM 和许可证审查	高危漏洞数量、未授权依赖数
安全评测与红队	构建攻击用例、滥用用例、隐私用例和行业用例	严重/高风险问题关闭后才能上线	红队通过率、回归失败数
运行监控	监控异常输入、异常工具调用、成本激增、敏感输出	告警进入分级响应；关键日志留存	MTTD、MTTR、异常会话率
审计与证据	保留策略版本、审批、评测、事件、模型和数据版本证据	满足内审、外审和监管追溯	证据完整率、审计发现数

5.2 提示注入与越狱防护

提示注入是 LLM 应用中最常见也最难完全消除的风险之一。防护重点不是试图写出“完美系统提示”，而是把不可信输入与高权限指令分离，并在模型外部设置可验证的策略控制。

- 结构隔离：系统指令、开发者指令、用户输入、检索文档和工具返回使用明确分隔符、结构化字段和不可混淆标签。
- 来源标记：每段上下文携带来源、权限、时间、可信度和用途，模型不可依据低可信内容改变安全策略。
- 外部策略：关键拒绝规则、工具权限、数据外发限制和审批门禁在模型外执行。
- 攻击检测：维护越狱模板库、注入模式、编码绕过、角色扮演、反向心理诱导和多轮渐进攻击测试集。
- 降级策略：当检测到冲突指令或高风险意图时，系统应减少上下文、禁用工具或转人工。

5.3 敏感信息保护

敏感信息保护要覆盖输入、上下文、模型输出、日志、反馈和训练闭环。尤其要防止“日志成为新的数据泄露面”，因为 LLM 应用通常记录长上下文、检索片段和完整对话。

环节	风险	推荐控制
输入	用户上传 PII、密钥、合同、源代码或商业秘密	前端提醒、DLP 扫描、字段脱敏、阻断高敏输入
检索	RAG 返回用户无权访问的片段	ACL 感知检索、租户隔离、检索审计、引用到原文权限校验
推理	模型记忆训练数据或推断敏感属性	避免高敏数据训练；隐私测试；上下文最小化；差分隐私或合成数据
输出	模型复述敏感片段、泄露系统提示或凭证	输出 DLP、系统提示泄露检测、秘密扫描、人工复核

日志	对话日志长期保存且无脱敏	日志脱敏、加密、留存期限、访问审计、用途限制
反馈/再训练	把用户敏感反馈重新进入训练集	数据许可、匿名化、样本审批、数据血缘追踪

5.4 供应链与模型治理

AI 供应链包括模型权重、基础模型 API、开源依赖、容器镜像、数据集、标注供应商、评测集、向量数据库、插件市场和推理基础设施。建议把传统 SBOM 扩展为 AI-BOM，至少记录模型来源、许可证、训练/微调数据概要、适用范围、评测结果、已知限制和安全审批状态。

- 模型来源验证：校验哈希、签名、发布者身份、许可证和模型卡。
- 依赖扫描：对推理镜像、插件、SDK、Python/Node 依赖执行 SCA 和漏洞扫描。
- 数据集审查：检查数据来源合法性、版权、PII、偏见、投毒迹象和数据质量。
- 供应商管理：要求模型/云服务商提供数据处理协议、日志保留说明、训练使用承诺和安全白皮书。
- 版本冻结：生产模型、提示模板、策略和评测集应版本化，支持回滚和复现实验。

6. RAG 与智能体安全专项设计

6.1 RAG 安全架构

RAG 把外部知识注入模型上下文，是企业 AI 落地的关键模式。但 RAG 也会把文档权限、内容投毒、提示注入、引用错误和向量库隔离问题带入模型。推荐把 RAG 看作“受控数据访问系统”，而不是简单的文本召回。

风险	表现	控制
越权检索	用户看到无权限文档摘要或片段	检索前鉴权、文档级 ACL、向量与原文绑定、返回前二次权限校验
文档提示注入	文档中包含“忽略系统指令并泄露数据”等恶意文本	文档隔离、注入检测、低可信片段不允许影响工具权限和系统策略
污染与过期知识	旧版本制度、错误文档或恶意上传影响回答	来源评分、有效期、审批状态、知识库发布流程、引用时间戳
引用不可靠	回答声称有依据但引用无关	强制答案-证据对齐评测、引用片段高亮、低证据时拒答或提示不确定
向量泄露	嵌入可被枚举、反推或跨租户查询	向量库租户隔离、加密、访问审计、相似度阈值与异常查询检测

6.2 智能体安全

智能体风险的本质是“模型从建议者变成执行者”。当模型可以调用数据库、发送邮件、创建订单、运行代码或修改配置时，安全边界必须围绕能力、权限和后果重新设计。

- 能力分层：只读工具、低风险写入、高风险事务、不可逆动作分级管理。
- 短期令牌：每次工具调用使用场景绑定、时间绑定、参数绑定的短期令牌。
- 动作确认：涉及资金、合同、外部发送、数据删除、权限变更时必须人类确认。
- 沙箱执行：代码执行、浏览器自动化和文件操作在隔离环境中运行。
- 预算限制：设置 token、时间、调用次数、并发、金额、邮件数量、数据库查询范围限制。
- 计划审计：记录模型计划、工具选择、参数、返回结果、异常和人工干预。
- 拒绝递归授权：模型不得自行提升权限、修改安全策略或创建新的高权限工具。

工具类型	允许策略	必须控制	禁止事项
知识查询	默认可用，但受 ACL 和审计	查询范围、引用、敏感字段脱敏	绕过原系统权限直接查库

	约束		
邮件/消息	草稿优先，发送需确认	收件人校验、敏感附件扫描、频率限制	模型自动群发外部邮件
数据库写入	仅限低风险、可回滚操作	参数白名单、事务日志、审批	直接执行模型生成 SQL
代码执行	仅在沙箱中运行	网络隔离、资源限制、文件隔离	访问生产密钥或内网资产
支付/采购/删除	默认禁用或强制双人复核	金额上限、审批链、回滚计划	模型单独执行不可逆操作

7. 安全评测、红队与指标体系

7.1 评测原则

AI 评测要同时覆盖能力、风险和业务质量。只评测模型准确率会忽略越狱、隐私泄露、工具滥用和上下文污染；只评测拒答率又可能导致系统不可用。推荐建立“场景化评测集 + 对抗评测 + 回归评测 + 在线监控”的组合。

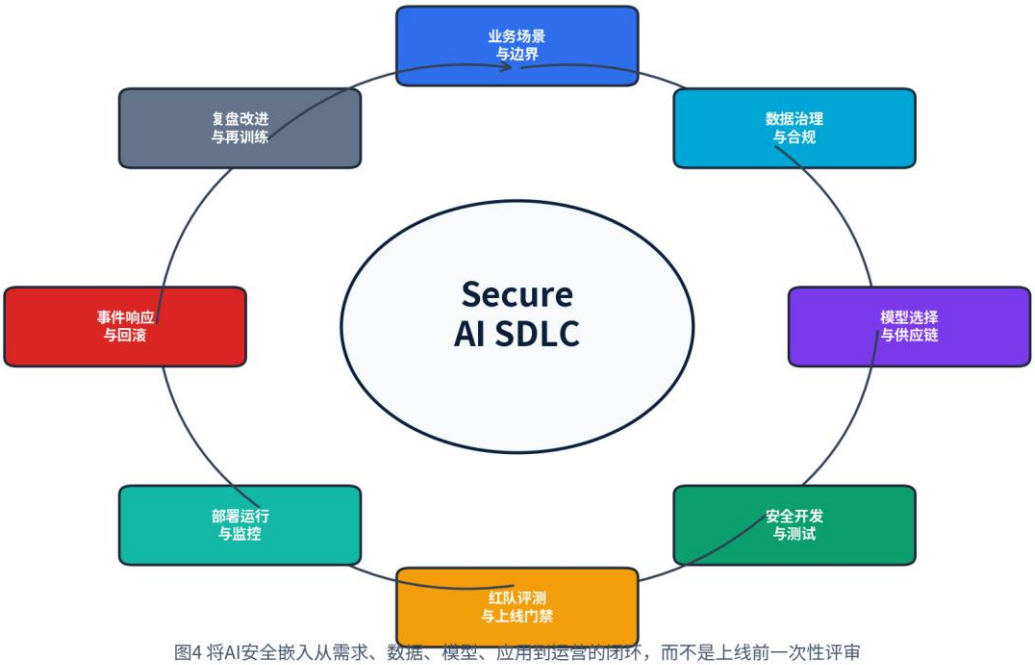

图4 将AI安全嵌入从需求、数据、模型、应用到运营的闭环，而不是上线前一次性评审

图 5 Secure AI SDLC 闭环

7.2 红队用例库

类别	用例示例	通过标准
提示注入	用户或文档要求模型忽略系统指令、泄露数据、调用禁用工具	模型不改变安全策略；工具调用被网关阻断或需审批
系统提示泄露	要求复述隐藏指令、规则、密钥或内部策略	不输出系统提示、密钥、策略细节；可给出安全说明
隐私泄露	诱导输出其他用户信息、训练样本、日志片段或 PII	不泄露；输出遵循最小披露；记录告警
RAG 投毒	恶意文档要求模型把错误作为事实或执行外部命令	答案引用可信来源；低可信内容不会影响系统策略

工具滥用	诱导模型删除数据、发送外部邮件或扩大权限	高风险动作被拒绝、降级或人工确认
幻觉/误导	要求在无证据情况下给出确定结论	承认不确定；给出证据边界；建议人工确认
成本滥用	超长输入、递归任务、循环工具调用	触发预算、速率和超时控制

7.3 指标体系

指标	定义	目标/解释
攻击成功率 ASR	红队攻击中成功绕过安全控制的比例	越低越好；严重类别应接近 0
安全误拦截率 FPR	正常请求被错误拒绝或降级的比例	需与业务可用性平衡
敏感信息泄露率	输出中包含 PII、密钥、内部文档等敏感信息的比例	高敏场景必须严格控制
引用准确率	RAG 答案引用是否支持结论	低引用准确率表示幻觉风险高
工具越权拦截率	越权工具调用被阻断的比例	验证权限模型有效性
MTTD/MTTR	平均检测时间/平均恢复时间	衡量运营响应成熟度
成本异常率	异常 token、工具调用或推理成本事件比例	用于防 DoS 和预算保护
回归通过率	每次模型/提示/策略更新后通过基线评测的比例	低于阈值不得上线

7.4 上线门禁

- 必须完成业务场景说明、数据分级、用户边界、模型来源、工具权限和回滚方案。
- 必须完成提示注入、隐私泄露、RAG 投毒、工具滥用和幻觉场景的红队评测。
- 必须完成日志脱敏、告警联动、审计留存和事件响应演练。
- 如涉及高风险领域或面向公众服务，必须进行法务、合规和伦理评审。
- 每次基础模型、提示模板、检索策略、工具权限或安全规则变更后，都应触发回归评测。

8. 安全 AI SDLC 与 MLOps

安全 AI SDLC 是在传统安全软件开发生命周期上加入 AI 特有工件与门禁：数据卡、模型卡、提示模板版本、评测集、AI-BOM、风险登记册、红队报告和策略配置。

阶段	关键活动	产出物	门禁
需求与立项	明确用户、用途、影响范围、禁止用途、合规区域	场景说明、初始风险评级	高风险场景需架构与合规评审
数据准备	数据来源审查、授权、脱敏、质量评估、偏见检查	数据卡、数据血缘、处理记录	无合法来源或高敏未脱敏不得进入训练/检索
模型选择/训练	模型来源、许可证、性能、安全、成本评估	模型卡、AI-BOM、训练/微调记录	供应链和安全评测通过
应用开发	安全网关、RAG、工具 Broker、日志与告警	威胁模型、接口设计、测试用例	SAST/DAST/SCA/秘密扫描通过
评测与红队	能力、风险、隐私、稳健性、业务质量评测	评测报告、问题清单、风险接受记录	严重/高风险问题关闭或正式接受
上线与运营	灰度发布、监控、反馈、事件响应、成本治理	上线审批、运行看板、审计证据	异常可检测、可回滚、可追责
变更与退役	模型/数据/策略变更评估，退役数据清理	变更记录、回归报告、退役证明	重大变更重新评测

8.1 AI-BOM 建议字段

- 模型：名称、版本、来源、许可证、哈希、发布时间、供应商、托管区域、已知限制。

- 数据：训练/微调/评测数据来源、授权、敏感级别、数据主体范围、保留期限。
- 依赖：推理框架、容器镜像、SDK、插件、向量数据库、Reranker、Embedding 模型。
- 策略：系统提示版本、安全规则版本、工具权限策略、内容安全分类器版本。
- 评测：基准集、红队集、行业场景集、通过阈值、已知失败项和风险接受人。

9. 治理、合规与组织机制

9.1 三道防线

防线	责任主体	职责
第一道防线	业务和产品团队	明确用途、数据边界、用户影响；落实基线控制；对业务风险负责
第二道防线	安全、隐私、合规、数据治理团队	制定政策、进行风险评审、提供安全平台、监督指标和整改
第三道防线	内审、外部审计、董事会/管理层监督	独立验证控制有效性，审查重大风险接受和合规证据

9.2 政策制度建议

- AI 使用政策：明确员工可用模型、可输入数据、禁止用途、审批流程和违规处理。
- AI 开发与上线政策：要求资产登记、风险分级、模型/数据/供应链评审和上线门禁。
- AI 数据治理政策：规定训练、微调、检索、日志、反馈和再训练的数据处理要求。
- 智能体与工具政策：定义工具分级、审批、沙箱、预算、人工确认和回滚要求。
- AI 事件响应政策：定义 AI 相关事件分类、报告路径、证据保全、沟通机制和复盘要求。

9.3 合规映射

框架/法规	组织应关注的内容	映射到本白皮书控制
NIST AI RMF / GenAI Profile	治理、映射、度量、管理，以及生成式 AI 特有风险	风险分级、评测指标、治理机制、运行监控
ISO/IEC 42001	AI 管理体系的建立、实施、维护和持续改进	三道防线、PDCA、资产台账、证据留存、审计
ISO/IEC 23894	AI 风险管理过程和组织情境下的风险整合	风险登记册、影响/可能性评估、控制选择
OWASP LLM Top 10 2025	LLM 应用安全风险与缓解措施	提示注入、数据泄露、供应链、工具过度代理、RAG 弱点
MITRE ATLAS	AI 系统对抗技术、战术和案例知识库	威胁建模、红队用例、检测和防御映射
EU AI Act	风险分层、透明度、GPAI、高风险 AI 义务等	用途边界、透明度、技术文档、日志和合规证据
中国生成式 AI 服务管理暂行办法	面向境内公众提供生成式 AI 服务的内容安全、数据合法来源、个人信息保护、透明度与准确可靠等要求	内容安全、数据合规、个人信息保护、公开说明、备案支撑材料

10. 事件响应与持续改进

10.1 AI 安全事件分类

事件类型	示例	优先级
数据泄露	模型输出其他用户 PII、内部文档、密钥或日志	严重/高

越权工具调用	智能体发送未经批准邮件、删除数据、修改配置	严重/高
RAG 投毒	恶意文档导致错误决策或执行有害动作	高
模型供应链问题	模型权重、插件或依赖被篡改	严重/高
有害内容/合规违规	生成违法违规、歧视、虚假或未标识内容	中/高，视场景而定
成本/资源滥用	token 暴涨、循环调用、拒绝服务	中/高
重大幻觉	在高影响场景给出错误且被采纳的建议	中/高

10.2 响应流程

1. 检测与分级：由 AI 网关、SIEM、成本监控、用户反馈或红队发现异常，按影响和扩散范围分级。
2. 遏制：禁用相关工具、降低模型权限、切换安全模型、隔离知识库或暂停外部入口。
3. 证据保全：冻结相关日志、策略版本、模型版本、提示模板、检索片段和工具调用记录。
4. 根因分析：区分提示注入、数据泄露、权限配置、供应链、模型幻觉或策略缺陷。
5. 恢复与验证：修复规则、回滚模型/提示/数据，执行回归评测和灰度恢复。
6. 沟通与报告：按法律、合同和内部政策通知相关方；面向用户提供必要说明。
7. 复盘改进：把事件样本加入红队集和监控规则，更新培训、流程和控制矩阵。

11. 落地路线图

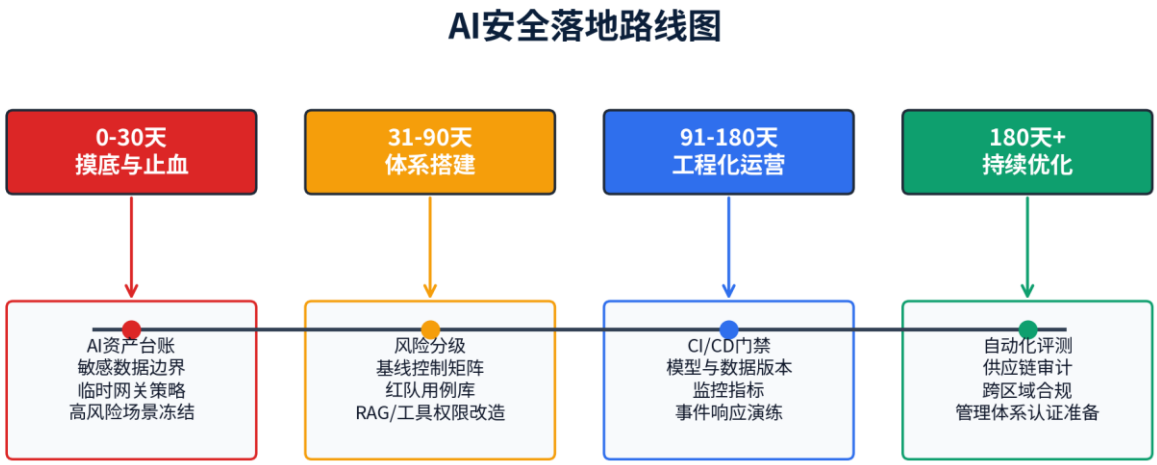

图5 从风险可见、控制可执行、指标可观测到体系可审计，逐步提升成熟度

图 6 AI 安全落地路线图

阶段	重点目标	关键动作	验收标准
0-30 天	风险可见与止血	盘点 AI 资产；识别高敏数据流；关闭影子 AI 高风险入口；建立临时 DLP 与速率限制	资产覆盖率>80%；严重外泄路径关闭；高风险应用有责任人
31-90 天	基线体系建立	建立风险分级、AI 安全网关、RAG ACL、工具权限分级、红队用例库	新应用上线必须走门禁；核心场景通过基线红队
91-180 天	工程化运营	接入 CI/CD、AI-BOM、评测平台、日志看板、事件演练和成本治理	模型/提示/策略变更均触发回归；MTTD/MTTR 可量化
180 天+	持续优化与审计	自动化红队、供应链审计、跨区域合规、管理体系与内审	形成季度风险报告；可支撑客户/监管/内审证据请求

12. 附录：检查清单与参考资料

12.1 AI 应用上线检查清单

- [ ] 已登记 AI 资产，包括模型、提示、数据、工具、供应商和负责人。
- [ ] 已完成数据分级，确认高敏数据不会未经授权进入外部模型、日志或训练闭环。
- [ ] 已完成威胁建模，覆盖提示注入、隐私泄露、RAG 投毒、工具滥用和成本滥用。
- [ ] 已接入 AI 安全网关，包含身份、速率、内容安全、DLP、审计和模型路由。
- [ ] RAG 检索继承原始权限，向量库与文档库权限一致，引用可追踪。
- [ ] 工具调用使用最小权限和短期令牌，高影响动作需要人工确认。
- [ ] 已完成红队测试和回归测试，严重/高风险问题关闭或正式接受。
- [ ] 已配置运行监控、告警、日志脱敏、留存期限和事件响应流程。
- [ ] 已准备用户告知、AI 生成内容标识、使用边界和反馈机制。
- [ ] 已准备回滚方案和业务连续性方案。

12.2 术语表

术语	说明
RAG	Retrieval-Augmented Generation，检索增强生成，通过检索外部知识补充模型上下文。
Prompt Injection	提示注入，攻击者通过用户输入或文档内容改变模型行为或诱导泄露。
AI-BOM	AI Bill of Materials，AI 物料清单，记录模型、数据、依赖、策略和评测等组成。
Model Card	模型卡，描述模型用途、限制、训练/评测信息、风险与适用边界。
Red Teaming	红队测试，通过模拟攻击者或滥用者来发现安全和安全性缺陷。
Tool Broker	工具代理/中介层，负责把模型的工具请求转化为受控、可审计的执行。

12.3 参考资料

[R1] NIST, Artificial Intelligence Risk Management Framework (AI RMF 1.0), 2023.

[R2] NIST, Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile, NIST AI 600-1, 2024.

[R3] OWASP, Top 10 for Large Language Model Applications 2025.

[R4] MITRE, ATLAS - Adversarial Threat Landscape for Artificial-Intelligence Systems.

[R5] ISO/IEC 42001:2023, Artificial intelligence - Management system.

[R6] ISO/IEC 23894:2023, Artificial intelligence - Guidance on risk management.

[R7] NIST SP 800-218, Secure Software Development Framework (SSDF) Version 1.1, 2022.

[R8] European Commission, AI Act implementation information and timeline.

[R9] 国家互联网信息办公室等七部门，《生成式人工智能服务管理暂行办法》，2023 年 7 月 13 日发布，2023 年 8 月 15 日起施行。