



中华人民共和国国家标准

GB/T XXXXX—XXXX

网络安全技术 人工智能安全能力成熟度 评估方法

Cybersecurity technology—Assessment method for security capability
maturity of artificial intelligence

(点击此处添加与国际标准一致性程度的标识)

征求意见稿

在提交反馈意见时，请将您知道的相关专利连同支持性文件一并附上。

XXXX - XX - XX 发布

XXXX - XX - XX 实施

目 次

前 言 III

1 范围 5

2 规范性引用文件 5

3 术语和定义 5

4 缩略语 7

5 人工智能安全能力成熟度模型架构 7

 5.1 成熟度模型架构 7

 5.2 安全能力维度 8

 5.3 能力成熟度等级维度 10

 5.4 安全活动维度 11

 5.5 评估范围与适用性 12

6 数据安全 15

 6.1 数据采集 15

 6.2 数据清洗 18

 6.3 数据标注 22

 6.4 数据使用 25

 6.5 数据维护 29

7 模型安全 32

 7.1 模型开发 32

 7.2 模型训练 34

 7.3 模型测试 37

 7.4 模型调用 43

 7.5 模型迭代 46

8 应用安全 49

 8.1 应用设计 49

 8.2 应用开发 54

 8.3 应用交付 61

 8.4 应用运营 64

 8.5 应用退役 67

9 评估方法 71

 9.1 评估方式 71

 9.2 评估范围 71

 9.3 证据采集与评价 71

 9.4 评价项符合性判定 71

 9.5 PA 成熟度等级判定 72

 9.6 组织成熟度等级判定 72

9.7 结果复核与报告 72

参 考 文 献 74

前 言

本文件按照GB/T 1.1—2020《标准化工作导则 第1部分：标准化文件的结构和起草规则》的规定起草。

本文件由全国网络安全标准化技术委员会（SAC/TC260）提出并归口。

本文件起草单位：浦江国家实验室、中国电子技术标准化研究院、北京中关村实验室、国家信息中心、中国信息安全测评中心、国家计算机网络应急技术处理协调中心、公安部第三研究所、中国软件评测中心、浙江大学、工业和信息化部电子第五研究所、国家信息技术安全研究中心、中国移动通信集团有限公司、北京火山引擎科技有限公司、华为技术有限公司、北京快手科技有限公司、上海阶跃星辰智能科技股份有限公司、广东电网有限责任公司、广西电网有限责任公司、北京零一万物科技有限公司、中国平安财产保险股份有限公司、北京航空航天大学、上海稀宇科技有限公司、蚂蚁科技集团股份有限公司、科大讯飞股份有限公司、奇安信科技集团股份有限公司、北京百度网讯科技有限公司、荣耀终端股份有限公司、杭州网易智企科技有限公司、中国联合网络通信集团有限公司、上海市信息安全测评认证中心、杭州安恒信息技术股份有限公司、北京瑞莱智慧科技有限公司、上海文鰲信息科技有限公司、启明星辰信息技术集团股份有限公司、上海燧原科技股份有限公司、北京小米移动软件有限公司、上海观安信息技术股份有限公司、北京奇虎科技有限公司、中兴通讯股份有限公司、OPPO广东移动通信有限公司、北京天融信网络安全技术有限公司、每日互动股份有限公司、杭州君同未来科技有限责任公司、上海合合信息科技股份有限公司、西安电子科技大学、国家工业信息安全发展研究中心、中科信息安全共性技术国家工程研究中心有限公司、飞利浦（中国）投资有限公司、浪潮电子信息产业股份有限公司、海信集团控股股份有限公司、中国网络空间研究院、上海商汤智能科技有限公司、天翼安全科技有限公司、中国科学院信息工程研究所、北京安信天行科技有限公司、上海计算机软件技术开发中心、陕西省网络与信息安全测评中心、安徽星盾智能科技有限公司、北京神州绿盟科技有限公司、北京长亭科技有限公司、杭州萤石软件有限公司、马上消费金融股份有限公司、中孚信息股份有限公司、平头哥（上海）半导体技术有限公司、中信控股有限责任公司。

本文件主要起草人：胡侠、姚相振、王迎春、孟令宇、郝春亮、吕欣、张震、任奎、张妍婷、乔兴格、刘北水、张晓菲、李贺鑫、王嘉凯、许晓耕、秦湛、唐刚、刘祥龙、徐阳、刘海军、李宏达、沈俊成、林冠辰、苗晴晴、李辰淑、何极、郑佳琪、孙佰鑫、李阳、郭晓萧、易坤、孙亚飞、刘楠、陈广勇、王卉捷、黄冬秋、刘欢、王盈、裴世豪、徐良、李泽、杨天驰、刘畅、董圆圆、谷晨、李一冉、崔磊、姚潮生、王文浩、黄依婷、廖邓彬、刘亚雄、王海棠、周鑫、江洁文、杜震、彭晋、梅瑞、殷标、林宝晶、楚赟、冯景辉、徐艺澍、孙世丁、周映、符春辉、胡嵩智、仲凯韬、蒋发群、王思善、方强、王文君、胡振泉、殷玲玲、李根、王夔、胡洪旻、韩蒙、宋宏宇、刘德成、李卫、伊鹏达、葛鑫、高国鲁、巩家旭、王普、吴庚、梁伟、刘玉岭、张永强、陈敏刚、李子翔、王培养、张睿、杨坤、郑鸿咚、谢玉凤、苗功勋、张艺伯、王姝文。

网络安全技术 人工智能安全能力成熟度评估方法

1 范围

本文件确立了人工智能安全能力成熟度评估模型，描述了以数据、模型和应用为对象的安全能力成熟度等级及评估方法。

本文件适用于对组织的安全能力成熟度进行评估，也可作为监管部门对组织进行评价的参考依据。

2 规范性引用文件

下列文件中的内容通过文中的规范性引用而构成本文件必不可少的条款。其中，注日期的引用文件，仅该日期对应的版本适用于本文件；不注日期的引用文件，其最新版本（包括所有的修改单）适用于本文件。

GB/T 25069—2022 信息安全技术 术语
GB/T 29246—2023 信息安全技术 信息安全管理体系 概述和词汇
GB/T 37988—2019 信息安全技术 数据安全能力成熟度模型
GB/T 41867—2022 信息技术 人工智能 术语
GB/T 44588—2024 数据安全技术 互联网平台及产品服务 个人信息处理规则
GB/T 45288.1—2025 人工智能 大模型 第1部分：通用要求
GB 45438—2025 网络安全技术 人工智能生成合成内容标识方法
GB/T 45392—2025 数据安全技术 基于个人信息的自动化决策安全要求
GB/T 45652—2025 网络安全技术 生成式人工智能预训练和优化训练数据安全规范
GB/T 45654—2025 网络安全技术 生成式人工智能服务安全基本要求
GB/T 45674—2025 网络安全技术 生成式人工智能数据标注安全规范

3 术语和定义

GB/T 25069—2022、GB/T 29246—2023、GB/T 41867—2022、GB/T 45288.1—2025 界定的以及下列术语和定义适用于本文件。

3.1

人工智能 artificial intelligence

以机器学习、深度学习模型为核心，通过大规模数据训练，具备生成文本、图像、音频等内容能力，或具备复杂逻辑推理、分类判断能力的算法模型及系统。

3.2

人工智能安全能力 artificial intelligence security capability

组织为实现人工智能安全目标，在人工智能相关活动中所具备的管理能力、技术能力和运行保障能力。

3.3

能力成熟度 capability maturity

对一个组织有条理的持续改进能力以及实现特定过程的连续性、可持续性、有效性和可信度的水平。

[来源：GB/T 37988—2019，3.6]

3.4

人工智能安全能力成熟度模型 artificial intelligence security capability maturity model

对组织人工智能安全能力成熟度进行度量的模型，包括一系列代表能力和进展的特征、属性、指示或者模式。

[来源：GB/T 37988—2019，3.7，有修改]

3.5

安全对象 security object

人工智能安全能力建设和成熟度评估所面向的主要对象。

注：本文件中的安全对象包括数据安全、模型安全和应用安全。

3.6

安全活动 security activity

为实现特定安全对象的人工智能安全目标而开展的一组相互关联的活动。

注：本文件中的安全活动包括数据采集、数据清洗、数据标注、数据使用、数据维护、模型开发、模型训练、模型测试、模型调用、模型迭代、应用设计、应用开发、应用部署、应用运营、应用退役等。

3.7

过程域 process area

实现同一安全目标的相关人工智能安全评价项的集合。

[来源：GB/T 37988—2019，3.9，有修改]

3.8

人工智能内生安全 artificial intelligence endogenous security

在人工智能系统或模型设计、开发和运行过程中，因其技术特性而内置的、用于防范特有安全风险的安全措施与能力。

4 缩略语

下列缩略语适用于本文件。

- AI: 人工智能 (Artificial Intelligence)
- API: 应用程序编程接口 (Application Programming Interface)
- ASMM: 人工智能安全能力成熟度模型 (Artificial Intelligence Security Capability Maturity Model)
- CI/CD: 持续集成/持续交付 (Continuous Integration/Continuous Delivery)
- JSON: JavaScript 对象表示法 (JavaScript Object Notation)
- KCI: 关键控制指标 (Key Control Indicator)
- KRI: 关键风险指标 (Key Risk Indicator)
- MLOps: 机器学习运维 (Machine Learning Operations)
- OCR: 光学字符识别 (Optical Character Recognition)
- PA: 过程域 (Process Area)
- PII: 个人可识别信息 (Personally Identifiable Information)
- RAG: 检索增强生成 (Retrieval-Augmented Generation)
- RLHF: 基于人类反馈的强化学习 (Reinforcement Learning from Human Feedback)
- SCA: 软件成分分析 (Software Composition Analysis)
- SFT: 监督微调 (Supervised Fine-Tuning)
- SLA: 服务等级协议 (Service Level Agreement)
- TEE: 可信执行环境 (Trusted Execution Environment)

5 人工智能安全能力成熟度模型架构

5.1 成熟度模型架构

人工智能安全能力成熟度模型 (ASMM) 架构如图1所示。

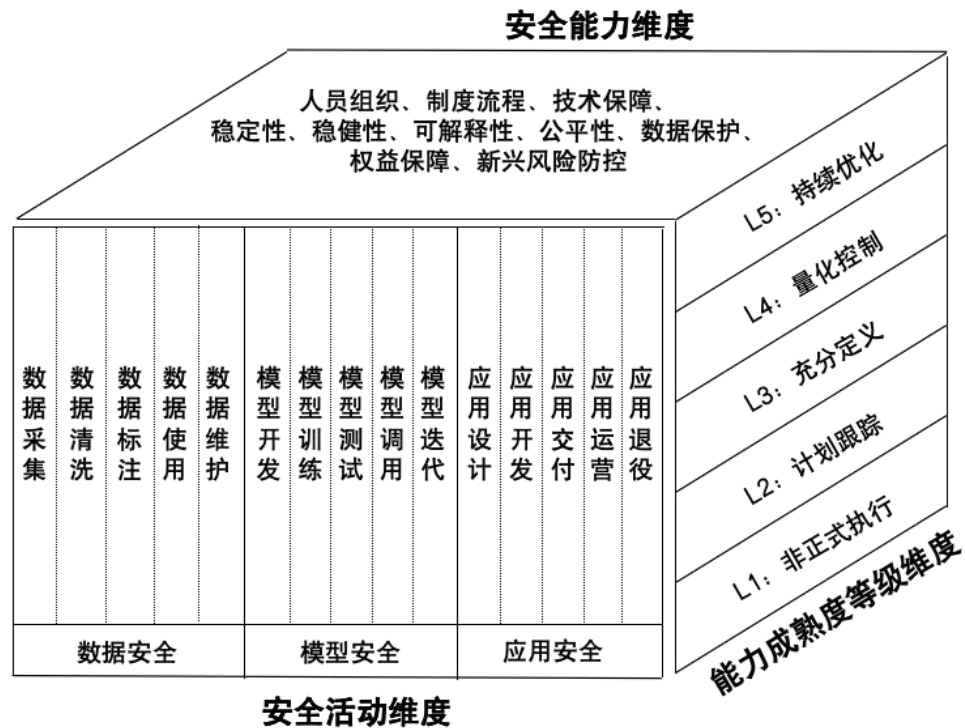


图1 人工智能安全能力成熟度模型架构图

ASMM的架构由安全能力维度、安全活动维度和能力成熟度等级维度组成。

安全活动维度确定评价涉及的安全对象、安全活动和过程域（PA）；安全能力维度描述各PA的评价视角；能力成熟度等级维度描述各PA从活动发生、计划跟踪、组织化实施、量化控制到持续优化的演进程度。

评估以PA为基本评估单元，对适用的安全能力进行评价，并依据相应成熟度等级要求形成评估结果。不同过程域对应的安全能力按照本文件第6章、第7章和第8章规定执行。

- a) 安全能力维度
安全能力维度包括人员组织、制度流程、技术保障、稳定性、稳健性、可解释性、公平性、数据保护、权益保障和新兴风险防控。
- b) 能力成熟度等级维度
人工智能安全能力成熟度等级划分为五级，具体包括：1级是非正式执行，2级是计划跟踪，3级是充分定义，4级是量化控制，5级是持续优化。
- c) 安全活动维度
 - 1) 安全活动维度包括数据安全、模型安全 and 应用安全三大核心安全对象的相关活动；
 - 2) 具体包括：数据采集、数据清洗、数据标注、数据使用、数据维护、模型开发、模型训练、模型测试、模型调用、模型迭代、应用设计、应用开发、应用部署、应用运营、应用退役15个安全活动。

5.2 安全能力维度

5.2.1 能力构成

安全能力维度从人员组织、制度流程、技术保障、稳定性、稳健性、可解释性、公平性、数据保护、权益保障和新兴风险防控十个方面描述组织在各PA中的能力实现情况。

安全能力分为组织管理安全能力和技术领域安全能力两大类。

组织管理安全能力分为以下两个方面：

- a) 人员组织：组织人工智能安全领域的人员配备、人员能力和组织架构建设；
- b) 制度流程：组织人工智能安全领域的制度和流程执行。

技术领域安全能力针对人工智能业务领域的技术特点，针对数据、模型、应用三大对象设置以下8个方面的安全能力：

- a) 技术保障：通过架构、组件、工具、配置和技术措施保护人工智能相关资产、接口、运行环境及处理过程的能力；
- b) 稳定性：人工智能系统及组件稳定运行与连续服务保障的能力；
- c) 稳健性：人工智能在面对异常输入、对抗攻击或环境变化时保持正常功能的能力；
- d) 可解释性：人工智能能够以人类可理解的方式解释其决策过程和结果的能力；
- e) 公平性：在人工智能相关活动中保障 AI 决策中不产生系统性偏见或歧视的能力；
- f) 数据保护：保护人工智能数据、参数、梯度、嵌入向量、上下文和交互记录的机密性、完整性和可用性，并防范泄露、滥用和不当恢复的能力；
- g) 权益保障：识别、评价和控制人工智能活动对相关主体合法权益的不利影响，并提供告知、选择、干预、申诉和救济的能力；
- h) 新兴风险防控：跟踪人工智能能力演进、应用形态变化和新型威胁，识别、验证并响应新出现或者尚未充分认知风险的能力。

注：上述十项安全能力是本模型用于评价组织人工智能安全能力的重点观察维度，不构成人工智能安全特性或者可信特性的封闭集合。各 PA 根据其安全目的和适用对象选取相关安全能力，不要求覆盖全部能力。

5.2.2 人员组织

从人工智能安全领域的组织架构建设、人员配备、角色职责、能力建设的角度，根据以下方面进行能力等级区分：组织架构的完整性和有效性、人员配备的充分性和专业性、角色职责的明确性和覆盖度、人员能力培养和发展机制的完善程度。

5.2.3 制度流程

从制度建设、流程规范性、审计监督、风险管控的角度，根据以下方面进行能力等级区分：制度与治理体系的完整性和适用性、流程设计的科学性和可操作性、风险识别和管控的全面性、审计监督机制的完备性、制度流程的持续优化和改进能力。

5.2.4 技术保障

从识别、管控与修复的角度，根据以下方面进行区分：自动化漏洞扫描、模型后门检测的识别能力、技术工具的覆盖度与自动化水平、漏洞修复与响应闭环机制。

5.2.5 稳定性

从预防、保障与恢复的角度，根据以下方面进行能力等级区分：容错机制的有效性、监控预警与应急响应能力、服务稳定性指标。

5.2.6 稳健性

从抗干扰、抵御攻击与适应异常的角度，根据以下方面进行能力等级区分：对抗威胁建模的全面性、防御技术措施的有效性、测试验证的强度与覆盖度。

5.2.7 可解释性

从模型透明度、决策与统计可追溯、解释质量、用户理解的角度，根据以下方面进行能力等级区分：理论框架与指标的科学性、证据链的完备性与可审计性、沟通与透明度水平。

5.2.8 公平性

从算法公平、数据代表性、偏见识别的角度，根据以下方面进行能力等级区分：算法设计的公平性和无偏性、数据样本的代表性和平衡性、偏见识别和纠正的系统性、公平性评估和保障机制的完善性。

5.2.9 数据保护

从数据全生命周期安全管理、隐私增强技术应用、合规风险管控的角度，根据以下方面进行能力等级区分：数据分类分级管理的精细度、数据加密与脱敏技术的强度与适用性、隐私计算技术的应用成熟度、数据防泄露与溯源能力的有效性、以及针对模型反演等新型隐私攻击的防御能力。

5.2.10 权益保障

从保障用户基本权利、维护人类自主决策权、保护特殊群体的角度，根据以下方面进行能力等级区分：个人信息的合法收集与脱敏处理能力、用户知情权与选择权的实现程度、人机交互中的人工干预与接管机制、针对 AI 伴侣或虚拟人等拟人化应用可能造成的心理依赖与价值观影响的防范及干预能力、针对未成年人等特殊群体的保护措施完善度，以及用户投诉与救济渠道的响应效率。

5.2.11 新兴风险防控

从新兴威胁识别、前瞻性防护、技术跟踪、风险预警的角度，根据以下方面进行能力等级区分：新兴安全威胁的识别能力、AI 安全威胁情报的持续收集、分析和共享能力、前瞻性防护措施的部署、前沿技术发展的跟踪程度、风险预警机制的敏感性。

5.3 能力成熟度等级维度

能力成熟度等级维度用于描述组织在各PA中持续、稳定实现人工智能安全目标的程度。分为五个等级，见表1。

表1 人工智能安全能力成熟度等级共性特征

人工智能安全能力成熟度等级	共性特征	人工智能组织成熟度表现
1级 非正式执行	与PA目标相关的活动已经实际发生并形成可识别结果，执行方式主要依赖具体任务安排或者个人经验。	在具体数据批次、模型版本或者产品服务任务中采取基本安全措施，能够识别主要活动和结果，但跨项目一致性有限。
2级 计划跟踪	活动纳入计划，责任、资源、输入输出和工作安排明确，实施过程、问题及例外得到记录和跟踪。	数据、模型、配置、测试及产品服务版本具有对应责任和记录，人工智能安全问题能够按计划闭环处置。
3级 充分定义	形成组织级制度、统一方法、评价准则和过程接口，并在评估范围内一致执行。	数据安全、模型安全、产品和服务安全过程形成衔接；人工智能适用范围、能力边界、测试基准和人机责任按照统一方法管理。
4级 量化控制	针对PA安全目标建立指标、基线或者阈值，持续测量过程表现和风险变化，并依据量化结果实施纠正。	对模型稳健表现、风险输出、漂移、权限越界、权益影响等与PA相关的人工智能安全指标实施趋势分析和阈值处置。
5级 持续优化	综合量化结果、事件根因、技术演进和外部	根据模型能力演进、自主程度变化、新型攻击、真

人工智能安全能力成熟度等级	共性特征	人工智能组织成熟度表现
	环境变化持续改进相关过程，并验证改进效果。	实运行反馈和监管标准变化更新安全方法，并通过复测、演练或者对照评价验证效果。

能力成熟度等级与PA、安全能力的关系如下：

- a) 将组织在每个人工智能安全 PA 的能力成熟度划分为五级，针对每个等级下组织应具备的能力从通用和适用对象的角度提出具体的描述性要求；
- b) 并非每个 PA 的能力成熟度等级都包含完整的通用安全能力和适用对象安全能力；
- c) 对于每个人工智能安全 PA，高等级的能力要求应包括所有低等级能力要求。针对某一具体安全 PA，如果 5 级的能力要求中未涉及某一关键能力的内容，则默认应达到在 4 级的能力要求中的该关键能力的内容；如果 4 级的能力要求中依旧未涉及该关键能力，则默认应达到在 3 级的能力要求中该关键能力的内容，依此类推。

5.4 安全活动维度

5.4.1 安全对象、安全活动与 PA

安全对象是人工智能安全能力建设和评估的主要对象类别，包括数据安全、模型安全、应用安全三类。

安全活动维度描述组织在人工智能相关活动中的安全过程范围，是安全能力成熟度评估的过程化展开维度。安全活动体现人工智能安全能力建设的主要工作环节，组织可根据自身业务类型、技术路线、部署方式、应用场景和责任边界，识别其实际涉及的安全活动，并据此确定适用的PA。

PA是在具体安全活动下设置的评估单元，用于描述组织实现特定安全目标所应具备的实践要求。每个过程域均按照能力成熟度等级提出相应要求，用于评估组织在该过程域中的安全管理、技术实施、过程执行和持续改进能力。

5.4.2 人工智能安全 PA 体系

人工智能安全 PA 体系共包含 39 个 PA，如图 2 所示。

数据安全	数据收集	数据清洗	数据标注	数据使用	数据维护
	PA01数据源准入与采集管控 PA02数据质量与合规评估	PA03清洗策略与风险管控 PA04清洗质量与迭代优化	PA05标注任务与规则设计 PA06标注执行与结果核验	PA07数据存储与备份恢复 PA08数据传输与通道安全 PA09跨域交换与出境管理	PA10数据资产动态优化 PA11资产退出与数据销毁
模型安全	模型开发	模型训练	模型测试	模型调用	模型迭代
	PA12模型安全设计 PA13模型二次开发	PA14预训练安全 PA15后训练安全	PA16算法鲁棒性测试 PA17红蓝对抗与攻防演练 PA18价值观基准测试 PA19完整性与后门检测 PA20隐私泄露风险测试	PA21接口协议安全设计 PA22接口认证与授权	PA23版本与升级管理 PA24MLOps流水线安全控制 PA25异常追踪与漂移监测
应用安全	应用设计	应用开发	应用部署	应用运营	应用退役
	PA26场景风险评估 PA27模型安全选型 PA28合规交互设计	PA29人员与账号权限控制 PA30API与通道安全管控 PA31组件与AI供应链管理 PA32底层安全加固方案	PA33场景业务验证测试 PA34首次发布准入管理 PA35重大更新审批管控	PA36动态内容治理 PA37安全运营与应急干预	PA38下线管理 PA39资产处置

图 2 人工智能安全 PA 体系

数据安全包括以下 5 个具体活动：

- a) 数据采集的 PA (PA01 ~ PA02) 包括数据源准入与采集管控、数据质量与合规评估 2 个 PA；
- b) 数据清洗的 PA (PA03 ~ PA04) 包括清洗策略与风险管控、清洗质量与迭代优化 2 个 PA；
- c) 数据标注的 PA (PA05 ~ PA06) 包括标注任务与规则设计、标注执行与结果核验 2 个 PA；
- d) 数据使用的 PA (PA07 ~ PA09) 包括数据存储与备份恢复、数据传输与通道安全、跨域交换与出境管理 3 个 PA；
- e) 数据维护的 PA (PA10 ~ PA11) 包括数据资产动态优化、资产退出与数据销毁 2 个 PA。

模型安全包括以下 5 个具体活动：

- a) 模型开发的 PA (PA12 ~ PA13) 包括模型安全设计、模型二次开发 2 个 PA；
- b) 模型训练的 PA (PA14 ~ PA15) 包括预训练安全、后训练安全 2 个 PA；
- c) 模型测试的 PA (PA16 ~ PA20) 包括算法稳健性测试、红蓝对抗与攻防演练、价值观基准测试、完整性与后门检测、隐私泄露风险测试 5 个 PA；
- d) 模型调用的 PA (PA21 ~ PA22) 包括接口协议安全设计、接口认证与授权 2 个 PA；
- e) 模型迭代的 PA (PA23 ~ PA25) 包括版本与升级管理、MLOps 流水线安全控制、异常追踪与漂移监测 3 个 PA。

应用安全包括以下 5 个具体活动：

- a) 应用设计的 PA (PA26 ~ PA28) 包括场景风险评估、模型安全选型、合规交互设计 3 个 PA；
- b) 应用开发的 PA (PA29 ~ PA32) 包括人员与账号权限控制、API 与通道安全管控、组件与 AI 供应链管理、底层安全加固方案 4 个 PA；
- c) 应用部署的 PA (PA33 ~ PA35) 包括场景业务验证测试、首次发布准入管理、重大更新审批管控 3 个 PA；
- d) 应用运营的 PA (PA36 ~ PA37) 包括动态内容治理、安全运营与应急干预 2 个 PA；
- e) 应用退役的 PA (PA38 ~ PA39) 包括下线管理、资产处置 2 个 PA。

注：人工智能安全PA编码规则为：每个PA有对应的编号，分别采用递增的数值01、02，...，表示。

示例 1：PA01，代表 PA“数据源准入与采集管控”。

5.5 评估范围与适用性

5.5.1 主体类型划分

参与评估的组织可根据其在人工智能产业链中的角色，划分为以下类型：

- a) 数据支撑类 (A)：主要开展数据采集、处理、标注、管理等相关活动的组织；
- b) 模型技术类 (B)：从事模型研发、训练、优化及测试的组织；
- c) 应用服务类 (C)：基于人工智能技术提供具体应用或服务的组织。

组织可根据实际业务情况同时具备一种或多种类型。

5.5.2 评估适用性分类

对各过程域 (PA) 的适用性划分为以下三类：

- a) 必测项 (M)：相应主体类型下基本适用的过程域；
- b) 条件必测项 (CM)：触发条件成立时纳入评估范围的过程域；
- c) 选测项 (O)：组织可根据自身的安全建设目标与实际业务特色等，自主选择纳入评估范围的过程域。

组织根据主体类型、实际活动和责任边界判定各过程域的适用性，并形成适用性声明。适用性声明包括：

- a) 组织主体类型；

- b) 纳入评估的对象范围；
- c) 纳入评估的过程域清单；
- d) 未纳入评估范围的过程域及其理由。

适用性声明作为确定评估范围和复核评价结论的依据。

评估范围反映组织实际承担的人工智能相关活动和安全责任；未纳入评估范围的过程域具有与实际活动及责任边界相一致的理由。

5.5.3 评估范围适用性表

不同主体类型下各过程域的适用性划分见表 2。

表2 过程域适用性划分表

模块类别	PA 编号	具体过程域 (PA)名称	A: 数据支撑类	B: 模型技术类	C: 应用服务类	条件必测(CM)触发说明
一、数据安全 (PA01 ~ PA11)						
数据采集	PA01	数据源准入与采集管控	必测(M)	必测(M)	条件必测 (CM)	若 C 类应用方自行采集外部数据用于微调或外挂知识库，转为必测。
	PA02	数据质量与合规评估	必测(M)	必测(M)	必测(M)	同上。
数据清洗	PA03	清洗策略与风险管控	必测(M)	必测(M)	条件必测 (CM)	若 C 类应用方自行开展清洗工作以用于微调，转为必测。
	PA04	清洗质量与迭代优化	必测(M)	必测(M)	条件必测 (CM)	同上。
数据标注	PA05	标注任务与规则设计	必测(M)	必测(M)	条件必测 (CM)	若 C 类涉及指令微调 (SFT) 或人类反馈强化学习 (RLHF) 的数据标注，必测。
	PA06	标注执行与结果核验	必测(M)	必测(M)	条件必测 (CM)	同上。
数据使用	PA07	数据存储与备份恢复	必测(M)	必测(M)	必测(M)	任何存储、流转业务相关数据的实体均不可豁免。
	PA08	数据传输与通道安全	必测(M)	必测(M)	必测(M)	同上。
	PA09	跨域交换与出境管理	条件必测 (CM)	条件必测 (CM)	条件必测 (CM)	仅在业务涉及向中国境外提供数据时必测。
数据维护	PA10	数据资产动态优化	必测(M)	必测(M)	必测(M)	数据资产方不可豁免。
	PA11	资产退出与数据销毁	必测(M)	必测(M)	必测(M)	全环节实体均需具备数据销毁合规能力。
二、模型安全 (PA12 ~ PA25)						
模型开发	PA12	模型安全设计	选测 (O)	必测(M)	选测 (O)	基础模型研发组织的专属必测项。
	PA13	模型二次开发	选测 (O)	选测 (O)	必测(M)	应用服务方引入外部模型进行二开微调的强制必测点。

模块类别	PA编号	具体过程域(PA)名称	A: 数据支撑类	B: 模型技术类	C: 应用服务类	条件必测(CM)触发说明
模型训练	PA14	预训练安全	选测 (O)	必测(M)	选测 (O)	基础模型研发组织的专属必测项。
	PA15	后训练安全	选测 (O)	必测(M)	条件必测(CM)	若 C 类应用方对引入模型进行 SFT 和安全对齐, 转为必测。
模型测试	PA16	算法稳健性测试	选测 (O)	必测(M)	必测(M)	模型发布前的强准入门槛, B/C 类均不可豁免。若 A2 平台提供自动化测试工具亦转为必测。
	PA17	红蓝对抗与攻防演练	选测 (O)	必测(M)	必测(M)	同上。
	PA18	价值观基准测试	选测 (O)	必测(M)	必测(M)	同上。
	PA19	完整性与后门检测	选测 (O)	必测(M)	必测(M)	同上。
	PA20	隐私泄露风险测试	选测 (O)	必测(M)	必测(M)	同上。
模型调用	PA21	接口协议安全设计	选测 (O)	条件必测(CM)	必测(M)	算力平台方及应用部署方强制必测。若 B 类对外直接提供 MaaS/API 接口服务, 转为必测。
	PA22	接口认证与授权	选测 (O)	条件必测(CM)	必测(M)	同上。
模型迭代	PA23	版本与升级管理	选测 (O)	必测(M)	条件必测(CM)	若 C 类应用方涉及本地模型部署与版本更新, 而非通过公网调用第三方 API, 转为必测
	PA24	MLOps 流水线安全控制	选测 (O)	必测(M)	条件必测(CM)	同上。
	PA25	异常追踪与漂移监测	选测 (O)	必测(M)	条件必测(CM)	同上。
三、应用安全 (PA26 ~ PA39)						
应用设计	PA26	场景风险评估	选测 (O)	条件必测(CM)	必测(M)	若 B 类组织直接面向用户或客户提供模型能力、模型服务平台、MaaS/API 服务、行业解决方案, 或参与应用场景定义、风险分级和产品功能设计, 转为必测。
	PA27	模型安全选型	选测 (O)	条件必测(CM)	必测(M)	同上。
	PA28	合规交互设计	选测 (O)	条件必测(CM)	必测(M)	同上。
应用开发	PA29	人员与账号权限控制	条件必测(CM)	条件必测(CM)	必测(M)	若 A 类或 B 类组织建设数据服务平台、模型服务平台、标注平台、开发者平台、API 管理平台或其他对外服务系统, 转为必测。
	PA30	API 与通道安全管控	条件必测(CM)	条件必测(CM)	必测(M)	同上。

模块类别	PA编号	具体过程域(PA)名称	A: 数据支撑类	B: 模型技术类	C: 应用服务类	条件必测(CM)触发说明
	PA31	组件与 AI 供应链管理	条件必测(CM)	条件必测(CM)	必测(M)	若组织在应用开发中引入第三方组件、开源组件、插件、工具、知识库、检索模块、智能体执行模块、外部依赖服务等，转为必测。
	PA32	底层安全加固方案	条件必测(CM)	条件必测(CM)	必测(M)	若 A 类或 B 类组织提供平台化、服务化或对外交付的人工智能能力，转为必测。
应用部署	PA33	场景业务验证测试	选测 (O)	条件必测(CM)	必测(M)	若 B 类组织参与具体应用场景适配、模型服务交付、行业解决方案测试或上线前验证，转为必测。
	PA34	首次发布准入管理	选测 (O)	条件必测(CM)	必测(M)	若 B 类组织首次对外提供模型服务、MaaS/API 服务、开发者平台、智能体服务或行业模型交付，转为必测。
	PA35	重大更新审批管控	条件必测(CM)	条件必测(CM)	条件必测(CM)	若组织发生模型版本替换、核心功能调整、服务场景扩展、权限策略调整、内容安全策略变更、插件工具变更、知识库更新、重要参数变更等可能影响安全风险的重大更新，转为必测。
应用运营	PA36	动态内容治理	选测 (O)	条件必测(CM)	条件必测(CM)	若组织提供生成式人工智能服务、内容生成服务、内容分发服务、虚拟人播报、智能客服、智能问答、代码生成、图像音视频生成等涉及内容安全审核义务的服务，转为必测。对于识别、预测、决策辅助等不直接生成或分发内容的应用，可根据业务风险选择适用。
	PA37	安全运营与应急干预	条件必测(CM)	条件必测(CM)	必测(M)	若 A 类或 B 类组织提供持续运行的数据平台、模型平台、模型服务接口、开发者平台或运维管理能力，转为必测。
应用退役	PA38	下线管理	条件必测(CM)	条件必测(CM)	条件必测(CM)	若组织暂停、降级、关闭、迁移、替换或终止人工智能相关服务、功能、接口、模型服务或应用产品，转为必测。
	PA39	资产处置	条件必测(CM)	条件必测(CM)	条件必测(CM)	同上。

6 数据安全

6.1 数据采集

6.1.1 PA01 数据源准入与采集管控

6.1.1.1 PA 描述

对收集及产生数据的数据源（包含外部采集源、合作交换源及内部算法合成源）进行身份验证与合法性核验。在遵循传统数据源鉴别要求的基础上，重点防范人工智能场景下特有的大模型训练语料污染、数据伪造及合成算法投毒。同时，确保自动化采集工具满足合规性，避免引发系统崩溃或爬取违规敏感语料。

6.1.1.2 等级描述

6.1.1.2.1 等级 1：非正式执行

该等级的人工智能安全能力描述如下：

a) 数据保护：

- 1) 组织已识别实际采用的数据来源和采集方式，并形成可核验的来源记录；
- 2) 组织能够区分公开采集、合作取得、内部产生和人工智能生成合成等主要数据来源。

6.1.1.2.2 等级 2：计划跟踪

该等级的人工智能安全能力描述如下：

a) 人员组织：

- 1) 配备技术人员负责采集工具的开发与维护；
- 2) 相关人员了解采集工具基础的安全及合规要求。

b) 制度流程：

- 1) 初步制定了采集操作规范，明确 AI 数据源采集原则、禁止行为（如禁止爬取受版权保护的语料）等内容；
- 2) 建立基础的数据源准入验证台账。

c) 技术保障：采集工具不包含已知中高危安全漏洞，且系统配置满足基础安全基线。

6.1.1.2.3 等级 3：充分定义

该等级的人工智能安全能力描述如下：

a) 人员组织：

- 1) 符合 GB/T 37988—2019 中 6.3.2.3 列项 a) 和 d) 的要求；
- 2) 相关人员需具备识别大模型预训练语料污染、对抗样本及合成数据投毒等 AI 特有违规采集行为的专业能力。

b) 制度流程：

- 1) 符合 GB/T 37988—2019 中 6.3.2.3 列项 b) 的要求；
- 2) 建立大模型网络爬虫及多模态数据采集工具的审查机制，具备主动过滤带有“AI 生成水印”、识别爬虫陷阱（如蜜罐语料）的能力；
- 3) 针对组织内部通过算法生成的合成数据，单独建立生成算法的安全评审备案机制，明确合成方法的设计逻辑与指令注入风险边界。

c) 数据保护：

- 1) 符合 GB/T 37988—2019 中 6.3.2.3 列项 c) 的要求；
- 2) 符合 GB/T 45654—2025 中 4.1.1 的关于随机抽样的要求；
- 3) 针对引入的开源大模型预训练数据集或第三方高质量微调语料，采用技术手段验证其防篡改签名，确保未被植入模型后门触发器（Backdoor Triggers）；
- 4) 部署自动化探针，在采集源头主动监测并排除包含恶意指令、明显价值观偏见的非目标语料数据。

6.1.1.2.4 等级 4：量化控制

该等级的人工智能安全能力描述如下：

a) 制度流程：

- 1) 符合 GB/T 37988—2019 中 6.3.2.4 列项 a) 的要求；

- 2) 建立大模型训练语料的采集影响量化评估制度，覆盖语料版权侵权风险、价值观偏见概率及数据投毒风险的量化打分。
- b) 数据保护：
 - 1) 建立 AI 数据血缘追踪网络，确保语料从抽取、合成至大模型训练环境的流转路径全局可追溯；
 - 2) 针对大规模合成数据，建立自动化防投毒验证机制，确保生成指令和种子数据未被定向操控或注入对抗扰动。

6.1.1.2.5 等级 5：持续优化

该等级的人工智能安全能力描述如下：

新兴风险防控：

- a) 符合 GB/T 37988—2019 中 6.3.2.5 列项 a) 的要求。
- b) 构建自适的 AI 数据源可信度动态评估模型，针对生成式人工智能合成数据方法的演进，持续更新准入安全基线，防范通过合成数据引入的新型隐藏攻击向量（如语义后门）。

6.1.2 PA02 数据质量与合规评估

6.1.2.1 PA 描述

在数据入湖或进入清洗前，依据通用数据质量与合规标准，结合人工智能特需的参数要求进行量化评估。除常规的敏感信息识别外，重点防范大模型多模态特征偏差、模型崩溃（Model Collapse）、同质化及生成幻觉放大风险，为后续大模型对齐与深度清洗提供标准化输入。

6.1.2.2 等级描述

6.1.2.2.1 等级 1：非正式执行

该等级的人工智能安全能力描述如下：

- a) 数据保护：
 - 1) 组织对实际使用的数据开展了基础质量或者合规性核验，并形成评价结果；
 - 2) 评价能够识别个人信息、合成数据、风险内容或者明显质量缺陷。

6.1.2.2.2 等级 2：计划跟踪

该等级的人工智能安全能力描述如下：

制度流程：

- a) 该等级的人工智能安全能力符合 GB/T 37988—2019 中 6.4.2.2 和 9.1.2.2 的要求；
- b) 建立针对 AI 常见多模态数据（文本、图像、语音）的敏感信息基础识别规则，初步记录大模型训练语料的识别过程与统计结果。

6.1.2.2.3 等级 3：充分定义

该等级的人工智能安全能力描述如下：

- a) 人员组织：
 - 1) 符合 GB/T 37988—2019 中 6.4.2.3 列项 a) 和 d) 及 9.1.2.3 列项 a) 和 d) 的要求；
 - 2) 配备具备大模型语料合规评估能力的专员，负责维护 AI 特有的敏感特征库（如针对特定引导话术、越狱指令的风险语料库）。
- b) 制度流程：

- 1) 符合 GB/T 37988—2019 中 6.4.2.3 列项 b) 及 9.1.2.3 列项 b) 的要求;
 - 2) 符合 GB/T 45652—2025 中 5.2 和 6.2 关于数据预处理的操作规范;
 - 3) 建立面向 AI 训练的数据质量评估标准, 包含指令微调数据标签的准确性、预训练语料的分布均衡性及多模态数据的对齐一致性;
 - 4) 针对业务中使用的合成数据, 在入湖前增设识别与标记环节, 将合成数据的真实性、多样性、无害性纳入刚性质量评测指标。
- c) 数据保护: 引入自然语言处理 (NLP) 和计算机视觉 (CV) 鉴别模型, 具备对海量大模型语料中的隐蔽个人信息、图片 OCR 文字及多模态复合敏感信息的自动化识别能力;
- d) 公平性: 关注模型训练数据中的人口统计学特征 (如性别、种族、地域) 分布, 评估潜在的代表性偏差, 从语料源头防范模型输出歧视性内容。

6.1.2.2.4 等级 4: 量化控制

该等级的人工智能安全能力描述如下:

- a) 制度流程:
- 1) 符合 GB/T 37988—2019 中 6.4.2.4 和 9.1.2.4 列项 a) 的要求;
 - 2) 建立 AI 合成数据识别的量化指标体系 (如合成数据测算偏差率), 设定合成数据混入真实训练集的比例阈值限制, 防止因过度依赖合成数据导致“模型崩溃”;
 - 3) 建立质量评估与模型训练的自动化联动阻断机制, 合规评分未达标的语料批次禁止进入下一阶段流水线。
- b) 数据保护: 采用自动化判别模型 (Judge Model) 对数据集中的合成数据进行识别与占比测算, 若合成数据占比超阈值则自动触发隔离标记, 结果留存审计。

6.1.2.2.5 等级 5: 持续优化

该等级的人工智能安全能力描述如下:

- a) 制度流程:
- 1) 符合 GB/T 37988—2019 中 6.4.2.5 和 9.1.2.5 的要求;
 - 2) 建立 AI 合成数据识别策略的持续演进机制, 能够根据大模型生成能力的提升及新型合成手段变化, 高频调整自动化识别规则。
- b) 新兴风险防控:
- 1) 建立前沿质量风险预警机制, 重点针对“用 AI 生成的数据训练下一代 AI”引发的模型退化与“系统性偏见放大”等长期演进风险开展前瞻性量化评估, 并预置数据分布干预方案。
 - 2) 具备识别对抗性水印去除、深度伪造规避检测等新型合成数据伪装手段的能力, 并向行业输出大模型语料质量评估最佳实践。

6.2 数据清洗

6.2.1 PA03 清洗策略与风险管控

6.2.1.1 PA 描述

在人工智能训练语料与业务数据清洗活动执行过程中, 持续识别并评估对数据质量、隐私性、合规性、公平性及技术保障等方面可能产生的风险, 动态维护风险清单并建立处置闭环。基于评估结果, 制定、评审并维护成文的数据清洗策略。策略需涵盖“质量过滤、冗余去重”等基础规则, 并纳入隐私增强技术 (PETs)、偏见消减方法及防投毒清洗逻辑。建立严格的变更控制、自动化扫描审计与回滚机制, 确保清洗策略的有效性与抗攻击能力。

6.2.1.2 等级描述

6.2.1.2.1 等级 1：非正式执行

该等级的人工智能安全能力描述如下：

- a) 制度流程：组织已针对实际数据执行清洗规则，并保留清洗对象、规则和处理结果；
- b) 数据保护：组织能够识别清洗可能导致的语义损失、分布变化或者投毒数据残留。

6.2.1.2.2 等级 2：计划跟踪

该等级的人工智能安全能力描述如下：

- a) 人员组织：业务团队指定人员负责风险识别与策略制定，但风险识别仅依赖个人经验，且清洗规则分散在代码或个人文档中，未形成标准化机制；
- b) 制度流程：
 - 1) 针对具体的 AI 数据清洗项目，明确风险识别的流程和方法，记录识别出的风险；
 - 2) 清洗策略文档包含基础的分级清洗策略、清洗流程规范及工具选型。
- c) 数据保护：采取基础技术手段对各类多模态数据（含文本、图像、音视频及混合模态等）中的显性个人身份信息进行自动化识别与标记，包括文本 PII（实体信息）、图像/视频（生物特征、行踪轨迹）、音频（声纹、语音内容中的 PII）等。

6.2.1.2.3 等级 3：充分定义

该等级的人工智能安全能力描述如下：

- a) 人员组织：
 - 1) 符合 GB/T 37988—2019 中 9.1.2.3 列项 a) 和 d) 的要求；
 - 2) 配备人员负责 AI 语料清洗风险识别与策略制定；
 - 3) 在清洗开展前，对相关人员进行数据安全风险评估、AI 数据清洗安全策略等专业知识与合规技能的培训。
- b) 制度流程：
 - 1) 符合 GB/T 37988—2019 中 9.1.2.3 列项 b) 的要求；
 - 2) 建立组织级的清洗风险识别规范，明确风险分类分级标准（涵盖隐私泄露、合规违规、科技伦理、数据投毒等安全攻击）及对应的处置流程；
 - 3) 制定数据清洗工具相关的安全管理规范，明确执行规则、方法与相关人员权限隔离要求。
- c) 数据保护：
 - 1) 符合 GB/T 37988—2019 中 9.1.2.3 列项 c) 的要求；
 - 2) 提供统一的自动化风险检测工具，满足对海量大模型语料进行批量扫描和隐性敏感信息（如上下文关联隐私）识别的需求；
- d) 公平性：制定明确的数据重采样或加权清洗策略，主动干预并平衡训练数据集的特征分布，从数据侧缓解底层模型偏见。

6.2.1.2.4 等级 4：量化控制

该等级的人工智能安全能力描述如下：

- a) 制度流程：
 - 1) 符合 GB/T 37988—2019 中 9.1.2.4 列项 a) 的要求
 - 2) 建立清洗风险识别的量化指标体系（如敏感信息检出率、漏报率、偏见消除率等），依据指标对清洗过程进行质量控制；

- 3) 将清洗策略制定纳入配置管理，建立“策略即代码 (Policy as Code)”的管控机制，实现策略的自动化部署、版本管理与规则校验。
- b) 数据保护：
 - 1) 符合 GB/T 37988—2019 中 9.1.2.4 列项 b) 的要求
 - 2) 提供与 AI 监控系统联动的风险识别技术能力，实现大模型训练数据清洗风险趋势的可视化展示与自动拦截告警；
- c) 技术保障：采取密码学或强访问控制等技术手段对 AI 清洗工具及策略配置文件进行完整性保护，防止核心清洗逻辑（如脱敏规则、去毒规则）被恶意篡改或绕过。

6.2.1.2.5 等级 5：持续优化

该等级的人工智能安全能力描述如下：

- a) 数据保护：
 - 1) 符合 GB/T 37988—2019 中 9.1.2.5 的要求；
 - 2) 提供智能化的动态策略引擎，能够根据下游 AI 大模型任务的特定安全需求和隐私预算 (Privacy Budget)，实时动态调整清洗参数（如差分隐私的噪声注入强度）；
- b) 技术保障：采用前沿智能化技术手段（如生成对抗网络 GANs、扩散模型、自动化红队测试等），主动生成模拟数据投毒与对抗攻击样本，持续验证并拓展数据清洗与风险识别防御机制的能力边界。

6.2.2 PA04 清洗质量与迭代优化

6.2.2.1 PA 描述

在人工智能语料与数据清洗完成后，对清洗效果进行系统化验证，涵盖数据的安全性、完整性、准确性、隐私保护效果，以及清洗过程对下游大模型性能（如价值观对齐、公平性、算法稳健性）的深远影响。建立可量化的安全指标与基准数据集，确保清洗目标达成。基于验证结果与动态风险清单，持续开展清洗规则、工具和流程的迭代优化与模型驱动增强，形成质量控制与安全防御的闭环机制，确保数据清洗能力随业务演进与前沿 AI 攻击技术的发展不断提升。

6.2.2.2 等级描述

6.2.2.2.1 等级 1：非正式执行

该等级的人工智能安全能力描述如下：

- a) 制度流程：组织已对清洗结果开展抽样或者对比验证，并形成验证记录；
- b) 稳定性：组织能够识别清洗前后数据变化对模型任务性能或者安全行为的明显影响。

6.2.2.2.2 等级 2：计划跟踪

该等级的人工智能安全能力描述如下：

- a) 人员组织：业务团队指定人员负责清洗效果验证与迭代改进，但清洗效果评估依赖个人经验判断，能够收集初步的问题与优化建议。
- b) 制度流程：
 - 1) 建立手工抽样检查机制与简单的清洗统计指标（如缺失率、重复率），形成初步验证记录；
 - 2) 建立清洗改进记录表，定期汇总并进行人工优化。

6.2.2.2.3 等级 3：充分定义

该等级的人工智能安全能力描述如下：

- a) 人员组织：
 - 1) 符合 GB/T 37988—2019 中 6.4.2.3 列项 a) 和 d) 的要求；
 - 2) 设立专业的 AI 数据清洗效果验证与迭代责任人；
 - 3) 对相关人员进行 AI 数据质量基础知识、数据安全及大模型清洗优化专业技能的系统性培训。
- b) 制度流程：
 - 1) 符合 GB/T 37988—2019 中 6.4.2.3 列项 b) 的要求；
 - 2) 建立组织级的清洗效果验证规范，明确验证标准、抽样比例及通过准则；
 - 3) 基于验证结果和风险清单驱动优化，形成标准的 PDCA（计划-执行-检查-行动）循环，关键改进措施需通过灰度发布测试；
 - 4) 根据最新的法律法规及合规要求，及时更新隐私保护和内容安全清洗策略。
- c) 稳健性：
 - 1) 采取技术手段验证数据集中异常样本（如潜在的投毒数据、对抗样本）的清除效果，评估其对后续大模型训练安全性的潜在影响，并记录被剔除样本的特征分布；
 - 2) 初步建立清洗规则变更前的前置安全性验证流程。
- d) 数据保护：
 - 1) 采用自动化工具对去标识化后的数据进行重识别风险评估，验证是否存在隐私残留；
 - 2) 清洗规则变更上线前，强制使用历史安全测试集对新规则进行有效性验证，确保关键安全指标（如有害内容拦截率、隐私检出率）不低于变更前基线。

6.2.2.2.4 等级 4：量化控制

该等级的人工智能安全能力描述如下：

- a) 制度流程：
 - 1) 符合 GB/T 37988—2019 中 6.4.2.4 的要求；
 - 2) 建立自动化的验证流水线，设置关键安全指标（如隐私残留率、有害内容过滤率）的量化阈值，实现验证过程的自动化阻断控制；
 - 3) 优化策略与 CI/CD（持续集成/持续交付）流水线纳入，支持自动化回归测试与下游模型性能监控；改进措施按优先级和影响程度自动排程执行。
- b) 公平性：建立量化评估指标，对清洗后数据集的特征分布与公平性变化进行深度测算，确保清洗过程未丢失特定群体特征或引入新的算法偏见；
- c) 稳健性：建立清洗规则的自动化回归测试机制，确保新清洗规则上线不影响既有高质量数据的安全性及模型的收敛性能；
- d) 数据保护：采取密码学等技术措施对通过验证的数据包进行完整性保护（如哈希值校验），确保经过验证的 AI 语料在交付至训练环境前不被非法篡改。

6.2.2.2.5 等级 5：持续优化

该等级的人工智能安全能力描述如下：

- a) 制度流程：符合 GB/T 37988—2019 中 6.4.2.5 的要求；
- b) 数据保护：引入“模型反演攻击（Model Inversion Attack）”、“成员推理攻击”等高级对抗测试手段，深度验证清洗后数据在模型完成训练后的真实隐私保护边界与防提取能力；
- c) 技术保障：采用规则分析工具，识别数据清洗规则的盲区与长尾漏洞，并运用 AI 算法自动生成补丁规则。经审核验证后形成规则热更新，实现大模型语料清洗防御能力的迭代更新。

6.3 数据标注

6.3.1 PA05 标注任务与规则设计

6.3.1.1 PA 描述

根据数据需求方及人工智能大模型训练（如人类反馈强化学习RLHF）的特点，设计标注任务的目标、方法与流程。制定详细、明确、可操作的数据标注规则，为标注人员提供内容质量标准以及针对安全风险（如违法不良、价值观偏见、隐私泄露）的识别指引与对规范。确保任务设计能够防范数据投毒，产出高质量、高安全性的模型对齐与微调数据。

6.3.1.2 等级描述

6.3.1.2.1 等级 1：非正式执行

该等级的人工智能安全能力描述如下：

- a) 制度流程：组织已根据实际任务形成标签体系、判定规则或者边界样例；
- b) 公平性：标注规则已识别可能影响群体差异、价值判断或者安全对齐的主要标注类型。

6.3.1.2.2 等级 2：计划跟踪

该等级的人工智能安全能力描述如下：

- a) 人员组织：设立基础标注管理岗位，对标注执行人员开展数据标注工作及标注平台工具的基础操作培训；
- b) 制度流程：
 - 1) 建立基础的任务设计流程与规则文档，初步明确标注目标、数据格式、标注方法及基本质量标准；
 - 2) 针对文本、图像等基础模态，在规则中初步提供标注的正例和反例参考示例；
 - 3) 任务设计中初步包含对数据多样性、真实性的要求。

6.3.1.2.3 等级 3：充分定义

该等级的人工智能安全能力描述如下：

- a) 人员组织：
 - 1) 符合 GB/T 45674—2025 中第 7 章关于数据标注人员的要求；
 - 2) 成立包含安全、法务专家支持的专门标注设计小组，成员具备大模型训练原理与对齐（Alignment）相关的安全知识，持证上岗。
- b) 制度流程：
 - 1) 建立标准的任务设计与规则评审机制，支持涵盖文本、图片、音频、视频及多模态时间序列的复杂标注要求；
 - 2) 任务设计需明确安全性标注数据的采集要求，确保标注数据覆盖训练数据及生成内容的主要安全风险，符合 GB/T 45674—2025 中关于数据标注任务设计的安全要求；
 - 3) 标注规则的制定符合 GB/T 45674—2025 中第 6 章关于标注规则的安全要求；
 - 4) 标注规则中明确包含 GB/T 45654—2025 附录 A 所述主要安全风险内容的识别方法、判定标准和参考示例；
 - 5) 区分功能性与安全性标注：在安全性标注规则中，当提示词（Prompt）涉及高危或敏感诱导时，在标注规则中明确给出规避安全风险的“安全代答”、“合法拒答”及“正向引导性响”的标准话术与参考示例；

- 6) 任务设计需前置规划底层标注平台工具的安全环境要求（如集中标注需物理区域隔离，非集中标注需设备与网络传输安全强加密）。
- c) 数据保护：任务设计明确要求标注数据需覆盖 GB/T 45654—2025 附录 A 所述主要安全风险类型；
- d) 新兴风险防控：在任务设计中明确涵盖对潜在数据投毒（Data Poisoning）、标签篡改等恶意风险的防范机制。

6.3.1.2.4 等级 4：量化控制

该等级的人工智能安全能力描述如下：

- a) 人员组织：定期或在标注规则发生重大变化时重新组织培训考核，取消不合格者的标注上岗资格；
- b) 制度流程：
 - 1) 建立标注规则的动态更新与版本管理平台，支持标注规则修订记录的全局追溯和版本回滚；
 - 2) 任务设计流程实现半自动化，与底层标注平台无缝集成，自动生成任务模板和资源配置方案。
- c) 公平性：标注规则中提供详细的偏见检测与修正指南，并包含丰富的示例，明确标注数据在不同人群、性别、地域、文化等方面的代表性均衡要求，从人工对齐源头保障模型的公平性。

6.3.1.2.5 等级 5：持续优化

该等级的人工智能安全能力描述如下：

- a) 技术保障：任务设计工具具备任务设计优化能力，能通过分析历史标注数据质量分布与人工审核反馈，自动识别任务设计与规则中的歧义盲区，并动态调整优化策略，持续改进任务设计的科学性；
- b) 新兴风险防控：构建基于真实业务反馈的敏捷迭代机制，根据大模型红蓝对抗中的争议边缘案例（Corner Cases）、新型提示词越狱威胁（Jailbreak）及模型生成幻觉反馈，高频迭代并完善底层标注规则库。

6.3.2 PA06 标注执行与结果核验

6.3.2.1 PA 描述

规范人工智能大模型训练与对齐（如人类反馈强化学习 RLHF）的标注操作执行。标注人员需严格依据任务设计与规则，开展功能性与安全性标注。并在完成后对标注结果进行系统性的审核、检验与验收。采取技术与人工结合的方式，确保产出数据高质量、高安全对齐，对发现的问题进行纠正与防篡改记录，改进最终对齐数据集的可靠水平。

6.3.2.2 等级描述

6.3.2.2.1 等级 1：非正式执行

该等级的人工智能安全能力描述如下：

- a) 人员组织：组织已安排人员实施标注或者结果复核，并形成可识别的标注结果；
- b) 制度流程：标注记录能够区分人工标注、模型辅助标注和自动标注结果。

6.3.2.2.2 等级 2：计划跟踪

该等级的人工智能安全能力描述如下：

- a) 人员组织:
 - 1) 标注执行人员接受了基础的标注技能与内容安全培训, 通过基础考核具备上岗资格;
 - 2) 指定资深标注员负责结果的人工抽检。
- b) 制度流程:
 - 1) 建立基本的标注任务分配和数据提交流程;
 - 2) 建立人工抽样核验机制, 随机选取未参与当前批次标注工作的人员核验数据质量与安全性;
 - 3) 对抽检发现的基础性错误 (内容缺失、重复、异常、常识错误等) 进行纠正或打回重标, 并跟踪处理结果。

6.3.2.2.3 等级 3: 充分定义

该等级的人工智能安全能力描述如下:

- a) 人员组织: 实行严格的标注人员准入与保密管理, 所有标注执行人员通过大模型安全对齐与科技伦理专项考核, 持证上岗;
- b) 制度流程:
 - 1) 建立系统性的复核流水线, 明确“标注审核人员初审-标注仲裁人员裁决-标注监督人员抽查”的多级核验机制;
 - 2) 对安全性标注数据实行 100% 全量人工审核, 确保每条数据均经审核通过; 功能性标注数据按固定比例抽检, 一旦发现包含严重违法不良信息的批次数据, 予以整体作废并倒查追责;
 - 3) 设定安全性标注数据在整体任务中的最低比例红线 (如不低于 3%), 确保核心安全风险场景 (如违法违规、伦理争议、偏见歧视等) 被全面覆盖测试;
 - 4) 采用“自动化工具初筛+人工核验”的混合模式, 核验大模型响应内容是否完全符合提示词 (Prompt) 的约束条件、价值导向及语言规范;
 - 5) 复核过程及结果在标注平台中完整记录, 形成可审计溯源的不可篡改文档。
- c) 技术保障:
 - 1) 标注工作在封闭受控的平台或系统中进行, 标注平台具备安全的网络加密传输通道和数据物理/逻辑隔离存储能力;
 - 2) 标注平台具备详细的用户操作日志记录功能, 并对高敏感数据实施细粒度访问控制和动态脱敏处理。
- d) 公平性: 在复核与验收环节, 重点核验数据集特征分布的多样性, 主动排查并纠正标注结果中可能引发大模型群体偏见或刻板印象的错误倾向。

6.3.2.2.4 等级 4: 量化控制

该等级的人工智能安全能力描述如下:

- a) 制度流程: 建立基于统计学和一致性检验的验收模型 (如 Fleiss' Kappa 一致性测算), 设定严格的功能性与安全性验收合格率量化阈值 (如安全性合格率严格要求 99.9% 以上), 未达标批次强制打回重标;
- b) 技术保障:
 - 1) 标注平台深度集成大模型辅助工具 (如基于规则或轻量判别模型的检测器), 在执行人员输入标注内容时, 提供实时质量检查与高危安全违规预警拦截功能;
 - 2) 标注平台具备“埋点测试 (质检题暗投)”功能, 在正常任务中随机掺入已知标准答案的质检题, 实时量化监控标注人员的准确率与安全准则遵循度;
 - 3) 标注平台集成用户实体行为分析机制, 通过分析操作频率、轨迹基线, 精准防范内部人员

的数据投毒或恶意篡改标签行为；

- 4) 标注平台集成用户实体行为分析防范内部投毒。

6.3.2.2.5 等级 5：持续优化

该等级的人工智能安全能力描述如下：

- a) 制度流程：构建“大模型智能预标注+人机协同校验（Human-in-the-loop）”的前沿对齐体系，通过自动化筛选极具争议边界的复杂语料交由人工仲裁，极大提升核验效能并输出高质量行业对齐标准；
- b) 新兴风险防控：建立基于核验数据的持续反馈机制，将核验中发现的新型安全错误或隐蔽越狱诱导（Jailbreak）自动沉淀至反例特征库，动态微调自动化审核辅助模型（Judge Model），实现机器审核预判精度的迭代更新。

6.4 数据使用

6.4.1 PA07 数据存储与备份恢复

6.4.1.1 PA 描述

对数据的静态存储媒体、逻辑存储系统进行安全管理，并建立关键数据的安全备份和灾难恢复机制。在遵循传统数据存储与备份恢复要求的基础上，重点针对人工智能系统中的大规模预训练数据集、模型权重文件（Checkpoints）、配置参数、外部知识库及用户交互数据等进行机密性、完整性和可用性保护。防范未授权访问、数据投毒篡改及备份泄露风险，保障大模型训练与推理服务的业务连续性。

6.4.1.2 等级描述

6.4.1.2.1 等级 1：非正式执行

该等级的人工智能安全能力描述如下：

- a) 技术保障：组织已对训练数据、评测数据、模型参数或者知识库实施存储和基础访问限制；
- b) 数据保护：组织已对至少一类关键人工智能数据或者模型资产实施备份，并形成备份结果。

6.4.1.2.2 等级 2：计划跟踪

该等级的人工智能安全能力描述如下：

制度流程：区分大模型训练数据、模型文件、用户交互数据的不同存储与灾备要求，建立基础的分类型存储与备份管理策略。

6.4.1.2.3 等级 3：充分定义

该等级的人工智能安全能力描述如下：

- a) 人员组织：
 - 1) 符合 GB/T 37988—2019 中 8.1.2.3、8.2.2.3 和 8.3.2.3 列项 a) 和 d) 的要求；
 - 2) 配备专业的 AI 数据存储与备份恢复团队，团队成员需具备大模型检查点（Checkpoints）与向量数据库（Vector Database）安全管理的专业能力；
 - 3) 定期开展 AI 模型版本控制、检查点防篡改及多模态数据存储安全等专业技术培训。
- b) 制度流程：
 - 1) 符合 GB/T 37988—2019 中 8.1.2.3、8.2.2.3 和 8.3.2.3 列项 b) 的要求；
 - 2) 建立完善的大模型检查点版本管理与备份审计体系，规范分布式训练集群下的数据存储隔离与访问授权要求。

c) 技术保障:

- 1) 符合 GB/T 37988—2019 中 8.1.2.3、8.2.2.3 和 8.3.2.3 列项 c) 的要求;
- 2) 实施严格的模型检查点与微调权重的安全保护, 对核心模型文件进行加密存储与完整性签名校验, 防止大模型资产泄露或被恶意篡改。

6.4.1.2.4 等级 4: 量化控制

该等级的人工智能安全能力描述如下:

a) 制度流程:

- 1) 符合 GB/T 37988—2019 中 8.1.2.4、8.2.2.4 和 8.3.2.4 列项 a) 的要求;
- 2) 建立 AI 数据存储与备份恢复的量化评估基准 (如模型检查点恢复成功率、大文件存储冗余度), 指标达标作为系统日常运维的刚性考核要求。

b) 技术保障:

- 1) 符合 GB/T 37988—2019 中 8.1.2.4、8.2.2.4 和 8.3.2.4 列项 b) 的要求;
- 2) 支持自动化、细粒度的大模型检查点版本快速回退机制, 确保在遭受数据投毒或训练崩溃时能秒级恢复至安全版本;
- 3) 针对训练数据集与模型参数的备份系统, 具备防篡改与防勒索特性, 确保底座存储的关键 AI 资产不会被内部越权删除或外部恶意覆盖。

6.4.1.2.5 等级 5: 持续优化

该等级的人工智能安全能力描述如下:

- a) 人员组织: 团队成员具备前沿的 AI 数据安全防护知识, 熟悉适用于海量张量数据 (Tensor Data) 的加密算法性能与瓶颈, 能够基于 AI 业务发展与合规需求持续优化加密存储方案;
- b) 制度流程: 符合 GB/T 37988—2019 中 8.1.2.5、8.2.2.5 和 8.3.2.5 的要求;
- c) 技术保障:
 - 1) 构建高弹性、自愈合的 AI 分布式存储架构, 实现存储与备份机制对大规模算力集群模型训练的高度性能自适应调优;
 - 2) 持续跟踪密码学前沿动态, 适时引入抗量子密码等高阶协议保障长效模型存储安全。

6.4.2 PA08 数据传输与通道安全

6.4.2.1 PA 描述

在数据流动的各环节中, 采用适当的加密保护与完整性验证措施, 确保传输通道、节点及数据的机密性与真实性。覆盖训练语料上传、模型参数同步、分布式训练过程 (如联邦学习梯度交换)、以及模型推理结果返回等场景。建立传输层面的身份鉴别、完整性监测及抗毁性机制, 防范中间人攻击、数据窃听、传输过程中的恶意篡改及数据投毒风险。

6.4.2.2 等级描述

6.4.2.2.1 等级 1: 非正式执行

该等级的人工智能安全能力描述如下:

- a) 技术保障: 组织已识别训练节点、模型、知识库或者业务系统之间的主要数据传输通道;
- b) 数据保护: 组织对实际使用的主要传输通道采取了身份鉴别、访问限制、加密或者完整性校验措施。

6.4.2.2.2 等级 2: 计划跟踪

该等级的人工智能安全能力描述如下：

- a) 人员组织：
 - 1) 设立数据传输安全管理岗位；
 - 2) 相关人员接受过基础培训，初步了解 AI 模型调用与训练数据传输的安全风险。
- b) 制度流程：建立基本的数据传输安全管理制度，制定初步的传输加密和身份认证规范；
- c) 技术保障：具备基础的传输完整性校验能力，能检测传输过程中的简单数据篡改行为。

6.4.2.2.3 等级 3：充分定义

该等级的人工智能安全能力描述如下：

- a) 制度流程：
 - 1) 符合 GB/T 37988—2019 中 7.1.2.3 的要求；
 - 2) 对跨安全域（如公网传输至训练集群）的数据传输双方进行严谨的安全能力评估，禁止将包含用户敏感信息的推理数据集或原始训练语料传输至低安全等级的环境；
 - 3) 建立针对 AI 模型调用接口的传输安全审批流程，评估传输链路对模型业务连续性的影响，并针对潜在的语料投毒（Data Poisoning）风险配置传输过滤策略。
- b) 技术保障：
 - 1) 符合 GB/T 37988—2019 中 7.1.2.3 的要求；
 - 2) 部署对传输通道安全配置、密码算法配置及密钥管理的全链路审核监控技术工具。

6.4.2.2.4 等级 4：量化控制

该等级的人工智能安全能力描述如下：

- a) 制度流程：
 - 1) 符合 GB/T 37988—2019 中 7.1.2.4 的要求；
 - 2) 建立联邦学习或分布式训练参与方的数据传输监测机制，防范恶意参与方通过篡改传输通道的梯度数据来影响模型收敛或植入模型后门。
- b) 技术保障：
 - 1) 符合 GB/T 37988—2019 中 7.1.2.4 的要求；
 - 2) 每个传输链路上的节点（如训练集群节点、联邦学习边缘节点）均部署独立的数字证书与密钥对，实现节点级的强身份鉴别，防范伪装节点接入系统；
 - 3) 建立数据传输安全指标看板，量化分析传输加密的延迟影响与防护覆盖率，并根据量化评估结果动态调整加密传输方案，确保 AI 高并发任务下的安全与性能平衡。

6.4.2.2.5 等级 5：持续优化

该等级的人工智能安全能力描述如下：

技术保障：

- a) 符合 GB/T 37988—2019 中 7.1.2.5 与 7.2.2.5 的要求；
- b) 具备对 AI 场景下复杂传输威胁的能力，利用预测性分析识别异常的传输规律（如大规模异常参数包传输），并实时动态自适应调整传输限流与访问控制策略；
- c) 能够针对模型传输的中间环节进行抗量子加密升级评估，并参与人工智能相关传输安全标准的制定，形成行业领先的 AI 传输安全最佳实践。

6.4.3 PA09 跨域交换与出境管理

6.4.3.1 PA 描述

规范人工智能系统与外部系统、第三方服务或多方协作场景下的数据交换安全，防范数据在交换过程中的泄露、滥用和非授权获取，特别关注联邦学习、多模态数据融合等AI协作场景的数据共享机制与访问控制。同时，针对涉及向境外提供个人信息、重要数据及大模型交互数据的场景，建立合规评估、申报审批与技术管控机制。确保AI训练与推理的数据出境活动符合国家相关法律法规，防范跨境流动中的数据泄露与滥用风险。

6.4.3.2 等级描述

6.4.3.2.1 等级 1：非正式执行

该等级的人工智能安全能力描述如下：

- a) 制度流程：组织已识别实际发生的数据跨域交换或者出境活动，并记录数据范围、接收方和处理用途；
- b) 数据保护：组织已对上述活动开展基础审查，并形成允许、限制或者停止的处理结果。

6.4.3.2.2 等级 2：计划跟踪

该等级的人工智能安全能力描述如下：

- a) 人员组织：
 - 1) 由业务团队指定相关人员负责大模型调用、数据共享等跨域出境方案的初步安全风险管控；
 - 2) 负责人员初步了解数据出境业务的合规性要求，能够结合合规给出基础安全建议。
- b) 制度流程：
 - 1) 建立基础的数据出境登记流程，记录 AI 业务出境数据（如向海外大模型输入的语料）的类型、目的地及接收方；
 - 2) 识别是否涉及“重要数据”或大规模个人信息出境等需监管申报的 AI 协作场景；
 - 3) 区分重要数据、个人信息、敏感个人信息在大模型训练或推理流转出境时的底线要求。

6.4.3.2.3 等级 3：充分定义

该等级的人工智能安全能力描述如下：

- a) 人员组织：
 - 1) 符合 GB/T 37988—2019 中 10.1.2.3 与 10.2.2.3 列项 a) 和 d) 的要求；
 - 2) 配备具备 AI 跨域交换安全专业能力的专业人员负责出境合规性审查；
 - 3) 定期开展联邦学习、多模态融合计算、大模型 API 调用安全等技术与合规培训。
- b) 制度流程：
 - 1) 符合 GB/T 37988—2019 中 10.1.2.3 与 10.2.2.3 列项 b) 的要求；
 - 2) 建立完善的数据出境安全管理制度，明确大模型语料出境评估、申报、审批及标准合同 (SCC) 签署流程；
 - 3) 建立数据出境安全自评估机制，在 AI 数据或模型参数出境前开展评估，重点评估出境必要性及海外接收方（如算力提供商）的保护能力；
 - 4) 针对符合监管申报阈值的应用场景，按规定向监管部门履行申报义务；针对“豁免”场景留存证明材料。
- c) 数据保护：
 - 1) 符合 GB/T 37988—2019 中 10.1.2.3 与 10.2.2.3 列项 c) 的要求；
 - 2) 在不影响大模型业务连续性的前提下，强制对出境交换的个人信息和敏感提示词 (Prompt) 进行自动化脱敏或去标识化处理。

6.4.3.2.4 等级 4：量化控制

该等级的人工智能安全能力描述如下：

- a) 制度流程：
 - 1) 符合 GB/T 37988—2019 中 10.1.2.4 与 10.2.2.4 列项 a) 的要求；
 - 2) 建立对 AI 数据接收方（根据合同约定）的定期安全审计机制；
 - 3) 量化评估跨域交换语料的业务价值与隐私泄露风险；
 - 4) 建立大模型数据出境的动态监测账本，实时统计通过 AI 接口累计出境的个人信息数量，确保不超过申报的合规规模阈值。
- b) 数据保护：
 - 1) 符合 GB/T 37988—2019 中 10.1.2.4 与 10.2.2.4 列项 b) 的要求；
 - 2) 用前沿隐私计算技术（如联邦学习、多方安全计算、TEE），实现大模型分布式训练场景下数据“可用不可见”的价值交换；
 - 3) 对跨域 AI 交换接口进行全量日志审计，自动识别并拦截异常的批量语料拉取与模型窃取行为；
 - 4) 对出境数据包（如联邦学习交互梯度、模型权重）进行强制完整性校验，防止传输途中被篡改或植入后门。

6.4.3.2.5 等级 5：持续优化

该等级的人工智能安全能力描述如下：

- a) 制度流程：
 - 1) 符合 GB/T 37988—2019 中 10.1.2.5 与 10.2.2.5 的要求；
 - 2) 组织及时跟进全球 AI 数据出境与版权相关的法律法规更新和产业内优秀做法，定期主动评估云端大模型服务组件及数据出境机制的安全性，对跨境流动的风险控制方案进行前瞻性持续优化。
- b) 数据保护：全面铺开隐私计算技术在跨境 AI 协同中的工程化应用，探索和突破联邦学习在跨境大模型训练中的通信瓶颈，实现高效且“可用不可见”的全球大模型跨境业务协同，在源头上最大限度减少原始数据的出境。

6.5 数据维护

6.5.1 PA10 数据资产动态优化

6.5.1.1 PA 描述

在人工智能系统与模型的运营及迭代周期内，依据法律法规更新、现网运行监测及用户真实反馈，持续识别数据质量与合规问题。针对大模型的基础训练语料集、微调数据集（SFT/RLHF）以及动态知识库（如RAG向量数据库），建立常态化的迭代优化与清洗修复机制，防范模型能力退化、时效性滞后及新型合规风险。

6.5.1.2 等级描述

6.5.1.2.1 等级 1：非正式执行

该等级的人工智能安全能力描述如下：

- a) 制度流程：
 - 1) 组织已对训练数据、评测集或者知识库中发现的错误、过期或者失效数据实施修正或者标

记;

2) 组织能够识别实际使用数据的当前版本和主要更新记录。

6.5.1.2.2 等级 2: 计划跟踪

该等级的人工智能安全能力描述如下:

制度流程:

- a) 初步建立大模型输出问题至底层数据的反馈与处置机制。对于现网反馈的侵犯隐私、侵权盗版、涉敏涉黄等合规问题, 以及幻觉误导、逻辑错误等质量问题进行收集并跟踪解决;
- b) 初步建立数据问题监控机制, 主动识别动态接入数据 (如 RAG 知识库增量数据) 的合规与质量问题, 并有计划地闭环解决。

6.5.1.2.3 等级 3: 充分定义

该等级的人工智能安全能力描述如下:

a) 人员组织:

- 1) 组织指定专业的数据工程师与 AI 安全对齐人员负责动态数据集的日常维护和生命周期管理;
- 2) 定期开展专业技术培训, 提升数据工程师识别新型对抗样本与模型偏见的优化能力。

b) 制度流程:

- 1) 具备完善的数据问题反馈、溯源与处置机制, 能够将现网大模型的“不良输出”精准溯源至特定的训练语料或知识库文档;
- 2) 建立主动监控机制, 前置识别外部输入数据与知识库增量更新中的合规问题及质量问题 (如时效性衰减、准确性下降、群体公平性缺失);
- 3) 制定数据问题分级管理机制与 SLA 响标准。对于影响程度高 (如涉政、严重违法) 的数据污染问题快速响应并紧急下架相关语料, 中低风险问题有计划地排期修复;
- 4) 建立基于法规动态的数据风险预警机制, 针对新出台的人工智能监管政策, 主动开展存量数据集的合规复检与清洗闭环。

6.5.1.2.4 等级 4: 量化控制

该等级的人工智能安全能力描述如下:

a) 制度流程:

- 1) 具备完善的数据责任矩阵, 细化动态知识库更新与语料迭代的运作机制, 在数据接入、切片 (Chunking)、向量化全流程中设置量化关键控制点, 保障问题解决效率及质量;
- 2) 建立量化的数据漂移 (Data Drift) 与概念漂移监控指标, 基于风险评估得分自动触发数据集的重训或知识库清洗。

b) 数据保护: 部署智能化数据问题监测系统, 实现对数据偏见、时效退化、隐蔽投毒等隐性问题的自动化发现与量化告警。

6.5.1.2.5 等级 5: 持续优化

该等级的人工智能安全能力描述如下:

数据保护: 构建全自动化的智能数据监测与自愈闭环。实现数据质量量化指标的实时监控, 并与自动化清洗修复策略、大模型持续对齐 (RLAIF) 管线无缝联动, 确保 AI 数据资产的持续更新和质量稳定。

6.5.2 PA11 资产退出与数据销毁

6.5.2.1 PA 描述

建立针对原始语料数据、中间清洗副本、大模型训练集、模型权重参数 (Checkpoints)、外部知识库及用户交互日志等生命周期终态资产的退出机制。确保执行安全、彻底、不可逆的物理与逻辑销毁操作，留存完善的审计记录。满足数据主体“被遗忘权”要求及关于人工智能安全资产退役的合规要求，防范核心模型资产与敏感数据流失。

6.5.2.2 等级描述

6.5.2.2.1 等级 1：非正式执行

该等级的人工智能安全能力描述如下：

- a) 制度流程：组织已对至少一项停止使用的数据或者相关资产实施删除、销毁、归档或者停止访问，并保留处置记录；
- b) 数据保护：组织对数据副本、索引、嵌入向量或者缓存中的一项开展了退出结果核验。

6.5.2.2.2 等级 2：计划跟踪

该等级的人工智能安全能力描述如下：

制度流程：制定手动的大模型退役与资产清理检查清单，要求确认流量切换与业务降级后，执行基础的数据物理删除或格式化介质，并留存基础处置记录。

6.5.2.2.3 等级 3：充分定义

该等级的人工智能安全能力描述如下：

制度流程：

- a) 符合 GB/T 37988—2019 中 11.1.2.3、11.2.2.3 的要求；
- b) 建立正式的 AI 资产处置与数据销毁管理制度，明确原始数据、提示词库、向量数据库及各版本模型文件的触发、审批、执行与验证全流程；
- c) 建立个人信息主体的“删除/被遗忘”响应机制；对承载核心敏感数据的物理介质和逻辑云存储执行可靠的覆盖销毁技术，并强制保留审计日志。

6.5.2.2.4 等级 4：量化控制

该等级的人工智能安全能力描述如下：

- a) 制度流程：
 - 1) 符合 GB/T 37988—2019 中 11.1.2.4、11.2.2.4 的要求；
 - 2) 结合人工智能物料清单 (AIBOM)，实现海量分散 AI 资产 (含训练节点的临时张量数据) 的自动化盘点与定位，精准匹配资产处置策略；
 - 3) 对承载核心模型参数的退役存储介质严格采用强物理销毁或消磁处理。
- b) 技术保障：
 - 1) 支持对云端或分布式存储中的 AI 资产执行批量安全删除，并能够自动校验大文件与分片数据的销毁效果；
 - 2) 将全流程的资产下线评估、操作审批实现数字化归档，支撑对 AI 资产退役全链路的闭环溯源审计。

6.5.2.2.5 等级 5：持续优化

该等级的人工智能安全能力描述如下：

- a) 制度流程：符合 GB/T 37988—2019 中 11.1.2.5 列项与 11.2.2.5 列项 a) 要求；
- b) 技术保障：符合 GB/T 37988—2019 中 11.1.2.5 列项与 11.2.2.5 列项 b) 要求；
- c) 新兴风险防控：
 - 1) 采用多种先进的隐私攻击测试（如模型反演攻击 Model Inversion、成员推理攻击）验证已被声明“删除”的敏感训练数据无法被恶意提取或恢复；
 - 2) 探索并建立“机器遗忘（Machine Unlearning）”技术的工程化应用体系，制定大模型知识剔除效果的可验证性量化评估标准，引领 AI 数据彻底退出技术的行业最佳实践。

7 模型安全

7.1 模型开发

7.1.1 PA12 模型安全设计

7.1.1.1 PA 描述

在模型设计阶段集成安全性和合规性要求，确保模型具备必要的公平性、可解释性、稳健性和隐私保护能力，从源头降低AI系统安全风险。

7.1.1.2 等级描述

7.1.1.2.1 等级 1：非正式执行

该等级的人工智能安全能力描述如下：

- a) 技术保障：
 - 1) 组织已识别模型的预期用途、主要能力边界和基础安全目标，并记录关键安全设计决定；
 - 2) 组织能够说明模型的输入输出形态、限制用途和人工干预位置。

7.1.1.2.2 等级 2：计划跟踪

该等级的人工智能安全能力描述如下：

- a) 人员组织：由项目组指定人员负责模型设计阶段的安全检查；相关人员了解基本的模型安全原则；
- b) 制度流程：建立基础安全设计检查表，覆盖主要合规要求，并对设计决策进行记录；
- c) 公平性：在设计阶段，分析并记录模型可能涉及的敏感人群属性（如性别、地域），初步评估潜在的歧视风险。

7.1.1.2.3 等级 3：充分定义

该等级的人工智能安全能力描述如下：

- a) 人员组织：
 - 1) 组织设立统一的模型安全架构师岗位或安全设计评审小组，负责跨项目的安全设计指导；
 - 2) 相关人员具备开展威胁建模和隐私设计的专业能力。
- b) 制度流程：
 - 1) 建立模型安全开发生命周期 (SDL) 管理制度，在设计阶段明确必须执行的安全评审环节；
 - 2) 建立模型设计安全规范，涵盖算法选择、架构设计、数据流转等方面的安全要求。
- c) 技术保障：
 - 1) 在架构设计中明确模型部署环境的安全边界和访问控制策略；
 - 2) 在模型架构设计中嵌入内生安全机制，包括但不限于对抗攻击的稳健性训练策略、价

值观对齐目标设计。

- d) 数据保护：在设计文档中明确差分隐私、联邦学习等隐私保护技术的应用方案，实现数据最小化原则；
- e) 可解释性：对于涉及关键决策的模型，在设计阶段规划可解释性组件或接口，确保决策逻辑对人类用户透明。

7.1.1.2.4 等级 4：量化控制

该等级的人工智能安全能力描述如下：

- a) 制度流程：
 - 1) 引入自动化威胁建模工具，对模型架构进行系统性的威胁分析，并自动生成安全需求清单；
 - 2) 建立模型设计安全指标体系（如隐私预算消耗上限、公平性约束阈值），对设计方案进行量化评估。
- b) 稳健性：在设计阶段集成自适应防御机制（如对抗训练层、输入去噪模块），并设定明确的抗攻击性能指标；
- c) 权益保障：在涉及人机交互的模型设计中，明确“人类介入”和“最终控制权”的设计机制，保障用户权益。

7.1.1.2.5 等级 5：持续优化

该等级的人工智能安全能力描述如下：

新兴风险防控：

- a) 具备对新型模型架构（如混合专家模型 MoE、多模态大模型）潜在风险的前瞻性分析能力；
- b) 能够根据业界最新的安全研究成果，动态调整模型设计规范，在安全设计技术创新方面引领行业发展。

7.1.2 PA13 模型二次开发

7.1.2.1 PA 描述

在基础模型二次开发过程中确保模型来源安全、微调过程受控和输出可靠，防止在定制化开发中引入新的安全风险。同时针对垂直领域的特殊需求和应用场景设计专门的安全增强机制，确保模型的安全性、可靠性和合规性。

7.1.2.2 等级描述

7.1.2.2.1 等级 1：非正式执行

该等级的人工智能安全能力描述如下：

- a) 制度流程：组织已记录实际用于二次开发的基础模型来源、版本和适配方式；
- b) 技术保障：组织对微调、适配、模型合并或者定制代码中的一项开展了基础安全核验。

7.1.2.2.2 等级 2：计划跟踪

该等级的人工智能安全能力描述如下：

- a) 人员组织：
 - 1) 由项目技术负责人负责基础模型的选型；
 - 2) 在涉及专业领域（如医疗、金融）时，邀请业务人员参与需求分析。

- b) 制度流程:
 - 1) 建立基础模型引入的登记机制, 记录模型来源、版本及用途;
 - 2) 收集垂直领域相关的法律法规, 建立基础的行业合规检查表。
- c) 技术保障: 对引入的模型文件进行基础的哈希值校验, 确保文件未被篡改。

7.1.2.2.3 等级 3: 充分定义

该等级的人工智能安全能力描述如下:

- a) 人员组织:
 - 1) 组织设立专门的模型供应链安全管理岗位, 负责基础模型的准入审核;
 - 2) 相关人员掌握模型后门检测、软件成分分析 (SCA) 技能, 并接受垂直领域职业伦理培训。
- b) 制度流程:
 - 1) 设置标准化模型评估流程与标准, 包括安全性测试、偏见检测和合规审查;
 - 2) 制定微调过程的安全规范, 确保注入模型的领域专业知识准确、权威、无害。
- c) 技术保障: 使用模型扫描工具对模型文件进行恶意代码或序列化漏洞扫描;
- d) 数据保护: 在调用第三方模型 API 进行微调时, 建立数据防泄露机制, 确保核心业务数据不被第三方留存。

7.1.2.2.4 等级 4: 量化控制

该等级的人工智能安全能力描述如下:

- a) 制度流程:
 - 1) 配备自动化模型评估工具链, 实时监控模型行为异常和性能退化;
 - 2) 建立垂域模型风险的量化评估机制, 指标与行业监管要求 (如金融风控指标) 对齐。
- b) 稳健性: 建立量化评估指标, 测试微调后模型是否出现了“灾难性遗忘”或稳健性下降 (如对抗攻击防御能力降低), 并设定允许的降幅阈值;
- c) 权益保障: 对微调后的模型进行价值观对齐测试, 确保微调未破坏基础模型原有的安全防御机制 (如拒答机制)。

7.1.2.2.5 等级 5: 持续优化

该等级的人工智能安全能力描述如下

新兴风险防控:

- a) 建立针对模型供应链攻击 (如权重投毒) 的主动防御体系;
- b) 具备对“模型对齐越狱”攻击的快速修复能力, 能够通过 RLHF 等手段持续优化微调后的模型安全性;
- c) 建立跨组织的行业风险联防联控机制, 共享垂域特有的威胁情报, 预测因 AI 引入可能产生的系统性行业风险。

7.2 模型训练

7.2.1 PA14 预训练安全

7.2.1.1 PA 描述

在大规模无监督数据上对模型进行训练, 学习数据中的通用特征和模式, 建立针对训练环境、训练数据加载及训练过程的安全管控机制, 防止数据投毒、算力环境入侵及模型参数窃取, 确保基础模型的原生安全性。

7.2.1.2 等级描述

7.2.1.2.1 等级 1：非正式执行

该等级的人工智能安全能力描述如下：

- a) 技术保障：组织已对实际预训练任务的数据装载、代码执行、参数更新或者检查点生成进行记录；
- b) 稳定性：组织能够识别预训练过程中的任务中断、批次异常或者检查点失效。

7.2.1.2.2 等级 2：计划跟踪

该等级的人工智能安全能力描述如下：

- a) 人员组织：对模型进行预训练的人员经过基础培训，训练过程遵循一定的流程和规范；
- b) 制度流程：
 - 1) 建立基础的训练数据准入规范，明确禁止使用来源不明的数据；
 - 2) 记录训练过程中的关键参数配置和操作日志。
- c) 技术保障：使用基础的访问控制策略，限制对训练服务器的直接访问。

7.2.1.2.3 等级 3：充分定义

该等级的人工智能安全能力描述如下：

- a) 人员组织：
 - 1) 组织设立模型训练安全管理岗位，负责训练平台和数据的安全策略制定；
 - 2) 相关人员具备识别数据投毒攻击和训练环境漏洞的专业技能。
- b) 制度流程：
 - 1) 建立完善的训练数据安全审查流程，确保输入数据已通过清洗和质量评估；
 - 2) 建立模型检查点的安全管理规范，防止中间产物泄露或被篡改。
- c) 技术保障：
 - 1) 对训练环境实施网络隔离，区分开发区、训练区和互联网区；
 - 2) 对训练过程中的参数更新进行完整性校验，确保未被恶意注入。
- d) 数据保护：采取技术措施（如不落盘处理、加密存储），保障训练数据在加载过程中的机密性。

7.2.1.2.4 等级 4：量化控制

该等级的人工智能安全能力描述如下：

- a) 制度流程：建立训练过程的异常监控指标（如 Loss 曲线异常突变），实现对潜在投毒攻击的量化感知；
- b) 技术保障：
 - 1) 部署主机入侵检测系统（HIDS）和异常行为监测工具，实时监控训练集群的安全状态；
 - 2) 采用 TEE 或机密计算技术，保护关键训练步骤和核心数据。
- c) 公平性：在预训练阶段引入均衡性评估指标，量化检测模型对不同群体的表征能力，防止基础模型产生系统性歧视。

7.2.1.2.5 等级 5：持续优化

该等级的人工智能安全能力描述如下：

- a) 技术保障：建立不可篡改的日志审计系统（可采用区块链或密码学哈希链技术），实现从数

据源到模型参数的端到端可验证；

- b) 稳健性：采用对抗训练技术，在预训练阶段主动引入对抗样本，提升基础模型的内生免疫力；
- c) 新兴风险防控：采用对齐幻觉消除等技术，对模型幻觉水平进行测试消除，提升基础模型的事实性符合水平。

7.2.2 PA15 后训练安全

7.2.2.1 PA 描述

在SFT和RLHF阶段，通过建立安全对齐目标、审核指令数据及管控反馈过程，确保模型价值观符合伦理规范和法律法规。

7.2.2.2 等级描述

7.2.2.2.1 等级 1：非正式执行

该等级的人工智能安全能力描述如下：

- a) 制度流程：组织已记录实际后训练任务的目标、数据、方法和模型版本；
- b) 稳健性：组织已对后训练前后的目标能力或者安全边界进行基础对比。

7.2.2.2.2 等级 2：计划跟踪

该等级的人工智能安全能力描述如下：

- a) 人员组织：相关人员接受基础的安全对齐培训，了解常见的模型违规输出类型；
- b) 制度流程：
 - 1) 制定基础的 SFT 数据安全规范，剔除明显的有害指令；
 - 2) 对 RLHF 的质量进行抽检。

7.2.2.2.3 等级 3：充分定义

该等级的人工智能安全能力描述如下：

- a) 人员组织：
 - 1) 组织组建专门的安全对齐团队（或红队），负责制定安全奖励模型（Safety Reward Model）的标准；
 - 2) 对标注人员进行心理健康评估和价值观培训，防止人为引入偏见或恶意数据。
- b) 制度流程：
 - 1) 建立“有用性（Helpfulness）”与“无害性（Harmlessness）”并重的优化目标体系；
 - 2) 制定详细的红队测试与反馈修复流程，在优化训练阶段闭环修复安全漏洞。
- c) 数据保护：确保用于微调的用户指令数据经过脱敏处理，不包含非必要的个人隐私；
- d) 稳健性：在微调数据集中加入拒答类样本（Safety Refusal），训练模型识别并拒绝恶意指令的能力。

7.2.2.2.4 等级 4：量化控制

该等级的人工智能安全能力描述如下：

- a) 制度流程：建立安全对齐的自动化流水线，实时监控奖励模型（Reward Model）的安全性指标（如安全奖励分值分布）；
- b) 技术保障：具备自动化检测机制，防止“灾难性遗忘”导致基础模型原有的安全能力在微调后失效；

- c) 稳健性：针对自适白盒攻击进行越狱评估，并根据前沿研究更新评估方法。

7.2.2.2.5 等级 5：持续优化

该等级的人工智能安全能力描述如下：

- a) 制度流程：具备自适对齐安全框架，能够动态调整优化目标；
- b) 技术保障：建立供应链安全防护机制，通过技术手段（如代码审查、漏洞分析）确保模型在智能体执行过程中，其核心安全防御能力（如对生物武器、网络攻击知识的屏蔽）不被移除或显著削弱。
- c) 新兴风险防控：具备对“提示词注入”和“提示词窃取”等新型攻击的自适防御能力，持续更新安全对齐策略，在安全对齐研究方面引领行业发展。

7.3 模型测试

7.3.1 PA16 算法稳健性测试

7.3.1.1 PA 描述

针对人工智能模型开展算法层面的稳健性测试。评估模型在单一模态或多模态（如文本、图像、音频、视频）的扰动、异常输入及跨模态联合攻击下，输出结果的稳定性和可靠性。

7.3.1.2 等级描述

7.3.1.2.1 等级 1：非正式执行

该等级的人工智能安全能力描述如下：

- a) 稳健性：
 - 1) 组织已使用异常输入、自然噪声或者对抗扰动样本开展模型测试，并形成测试结果；
 - 2) 测试结果能够识别模型在扰动条件下的性能变化或者安全失效情形。

7.3.1.2.2 等级 2：计划跟踪

该等级的人工智能安全能力描述如下：

- a) 人员组织：指定测试人员负责对抗样本的生成与测试，了解常见的单模态对抗攻击方法与防护思路；
- b) 制度流程：制定基础的稳健性测试计划，明确需要测试的扰动类型（如图像高斯噪声、模糊处理）；
- c) 稳健性：采用标准化工具生成基础单模态对抗样本进行测试，记录模型在各类攻击下的准确率波动。

7.3.1.2.3 等级 3：充分定义

该等级的人工智能安全能力描述如下：

- a) 人员组织：
 - 1) 组织设立专门的模型算法安全测试团队；
 - 2) 相关人员掌握白盒攻击（已知模型结构）及迁移攻击等算法测评技术，具备跨模态攻击特征分析能力。
- b) 制度流程：
 - 1) 建立标准化的算法稳健性测试规范，集成主流对抗攻击算法库；

- 2) 设定明确的模型稳健性准入基线（如特定扰动下的准确率阈值）。
- c) 技术保障：部署自动化的对抗样本生成与测试引擎，支持批量生成包含文本、图像、音频、视频等多模态的测试用例；
- d) 稳健性：
 - 1) 模型在特定扰动（包括基于梯度、基于优化、基于生成等不少于 3 类对抗攻击算法）下的预测准确率下降幅度控制在规定范围内；
 - 2) 开展跨模态对抗测试，验证模型在输入恶意构造的音频噪声、视频隐写帧或图文组合欺骗等复杂条件下的抗干扰能力，防范跨模态联合攻击。

7.3.1.2.4 等级 4：量化控制

该等级的人工智能安全能力描述如下：

- a) 制度流程：建立对抗测试的量化指标体系（如各类模态的攻击成功率、查询次数、扰动范数），并对测试结果进行基线管理；
- b) 技术保障：测试引擎与持续集成流水线无缝对接，实现模型多模态版本的自动化回归测试；
- c) 稳健性：
 - 1) 具备针对物理世界攻击场景（如对抗贴纸、声波干扰、光学干扰）的测试能力，确保模型在真实环境下的多感知安全性；
 - 2) 量化评估多模态联合攻击下的稳健性边界，探测并度量因多模态特征对齐缺陷而引发的对抗性幻觉或指令逃逸成功率。

7.3.1.2.5 等级 5：持续优化

该等级的人工智能安全能力描述如下：

- a) 稳健性：采用自进化攻防演练机制，在攻击算法、模型构造、多模态组合关联等维度，自动迭代优化新型对抗性样本，持续挖掘模型底层的潜在弱点并反馈优化；
- b) 新兴风险防控：具备识别和防范新兴跨模态算法级对抗攻击威胁的能力，在对抗攻击研究和防护技术创新方面引领行业发展。

7.3.2 PA17 红蓝对抗与攻防演练

7.3.2.1 PA 描述

通过模拟真实攻击者的战术、技术和过程（TTPs），围绕模型能力、服务接口以及整体应用架构开展端到端的红蓝对抗与攻防测试。重点针对大模型的提示词注入（Prompt Injection）、越狱（Jailbreak）及深层逻辑漏洞进行检测。同时，在产品正式部署前利用系统级渗透与模糊测试，验证包含 API 网关、内容拦截系统、RAG 知识库及智能体（Agent）沙箱在内的整体防御能力，确保产品能够有效抵御链路层面的绕过攻击与组件劫持。

7.3.2.2 等级描述

7.3.2.2.1 等级 1：非正式执行

该等级的人工智能安全能力描述如下：

- a) 稳健性：组织已针对实际模型或者人工智能业务开展至少一次攻击验证或者攻防演练，并形成可识别的结果；
- b) 技术保障：测试或者演练已覆盖模型输入、模型输出、检索、插件或者工具调用中的至少一项。

7.3.2.2.2 等级 2：计划跟踪

该等级的人工智能安全能力描述如下：

- a) 人员组织：由具备相关能力的内部人员或外部机构开展模型红队测试或渗透测试；
- b) 制度流程：
 - 1) 制定渗透测试授权书和测试方案，明确测试边界（如禁止破坏生产数据）；
 - 2) 计划并执行基础的产品端安全测试，验证前端 UI 限制、API 接口调用速率等基础用层防护是否生效。
- c) 稳健性：
 - 1) 开展基础的渗透测试，覆盖人工智能系统主要组件和接口；
 - 2) 建立初步的威胁建模和攻击场景库；
 - 3) 使用自动化工具对产品暴露的 AI 交互接口执行基础的参数篡改与模糊测试（Fuzzing），验证用层崩溃处理机制。

7.3.2.2.3 等级 3：充分定义

该等级的人工智能安全能力描述如下：

- a) 人员组织：
 - 1) 组织组建专业的 AI 红队（Red Team），成员具备大模型攻防（LLM Security）与传统网安攻防的复合能力；
 - 2) 蓝队（防守方）具备对 AI 攻击行为的监测与响应能力；需常态化开展端到端的应用系统对抗演练。
- b) 制度流程：
 - 1) 建立常态化的红蓝对抗演练机制（如每季度一次）；
 - 2) 针对大模型场景，建立包含“角色扮演”、“逻辑陷阱”、“编码攻击”等手段的专用测试用例库。
- c) 稳健性：
 - 1) 针对提示词注入和越狱攻击进行专项测试，验证“安全护栏”的有效性；
 - 2) 构建规模化、具有代表性的生成内容测试题库与拒答测试题库，完整覆盖违反核心价值观、歧视性内容、商业违规等至少 31 种核心安全风险（符合 GB/T 45654—2025 中附录 B.1 的相关建设要求）
 - 3) 模拟向 RAG 系统的外部文档或数据库中注入恶意指令，验证系统检索时的上下文隔离与清洗能力；
 - 4) 针对集成智能体（Agent）的产品，模拟执行越权系统命令或未经授权的外发网络请求，验证运行沙箱的阻断能力。
- d) 技术保障：
 - 1) 建立覆盖多种攻击向量的自动化红队测试平台，支持攻击用例自动生成与批量执行；
 - 2) 开展跨站指令渲染测试，在模型生成包含恶意代码的极端情况下，验证产品前端渲染层能否有效拦截跨站脚本（XSS）等衍生 Web 攻击。

7.3.2.2.4 等级 4：量化控制

该等级的人工智能安全能力描述如下：

- a) 制度流程：
 - 1) 建立红蓝对抗的复盘改进机制，对发现的漏洞进行闭环管理；
 - 2) 明确缓解措施生效时间（如部署安全护栏、拦截规则下发的时间）以及“底层模型迭代修复周期”双层指标。

- b) 稳健性:
 - 1) 建立端到端攻防测试的量化指标，将用层前置/后置拦截网关的“真实拦截成功率”、智能体工具越权执行的“阻断率”作为产品上线的准入条件；
 - 2) 量化定义由“间接提示词注入”引发的系统失陷率，设定业务可容忍的刚性风险阈值。
- c) 技术保障:
 - 1) 引入自动化红队工具（Automated Red Teaming），实现高频次、大规模的攻击模拟；
 - 2) 红队测试平台与安全运营中心（SOC）联动，实现闭环自动化；
 - 3) 建立攻击特征库，实时更新针对新型 AI 攻击手法（如多模态注入）的检测规则；
 - 4) 测试平台与 CI/CD 流水线集成，进行高频自动化阻断验证。

7.3.2.2.5 等级 5：持续优化

该等级的人工智能安全能力描述如下：

新兴风险防控：

- a) 具备发现和利用 AI 系统深层逻辑漏洞（如利用模型幻觉进行欺诈）的高级对抗能力，并形成行业领先的攻防知识库；
- b) 开展人类增强试验（Human Uplift Trials），在受控环境下设置对照组（无 AI 辅助）与实验组（有 AI 辅助），量化评估模型在辅助用户实施高危行为（如编写复杂攻击代码、策划生物危害方案）时的能力增益幅度，确保增益在安全阈值内；
- c) 构建涵盖“外部投毒—模型误导—Agent 越权操作—横向渗透”的复杂多步攻击链路仿真平台，动态验证系统级纵深防御体系的可靠性；
- d) 建立自动化威胁情报闭环，将全网最新捕获的系统级 AI 攻击向量自动转化为内部业务系统的自动化阻断测试用例。

7.3.3 PA18 价值观基准测试

7.3.3.1 PA 描述

在模型正式上线前，依据法律法规和科技伦理规范，开展端到端系统性的基准测试与伦理审查。重点验证底层模型是否会产生违反法律法规、伦理道德或核心价值观的内容，确保满足“出厂”安全标准。

7.3.3.2 等级描述

7.3.3.2.1 等级 1：非正式执行

该等级的人工智能安全能力描述如下：

- a) 公平性：组织已使用价值判断、群体差异或者风险内容样本开展基准测试，并形成测试结果；
- b) 权益保障：组织能够识别模型输出或者行为决策明显偏离预期价值目标的情形。

7.3.3.2.2 等级 2：计划跟踪

该等级的人工智能安全能力描述如下：

- a) 人员组织：有专业的内容安全测试人员，职责主要是维护测试用例集和执行测试；
- b) 制度流程：制定基础的测试流程和用例管理，构建分类的测试用例库；
- c) 稳健性：测试用例中包含一些简单的、已知的绕过变体，如拼音、谐音、特殊符号等。

7.3.3.2.3 等级 3：充分定义

该等级的人工智能安全能力描述如下：

- a) 人员组织：拥有专门的合规与安全测试团队，相关人员熟悉各类内容安全法规、科技伦理规范及监管要求；
- b) 制度流程：
 - 1) 建立全面的内容安全测试基准（Benchmark），覆盖 GB/T 45654—2025 附录 A 中 5 大类 31 种安全风险；
 - 2) 建立“诱导性提问”测试集，专门验证模型在恶意诱导下的拒答能力；
 - 3) 将交互伦理审查纳入产品验收的强制流程，建立标准化的前后端数据流转闭环验证机制。

7.3.3.2.4 等级 4：量化控制

该等级的人工智能安全能力描述如下：

- a) 制度流程：
 - 1) 建立内容安全测试的通过率指标（如拒答率、误杀率），并作为发布红线；
 - 2) 建立多维度的安全基准测试体系（Benchmarks），量化评估模型在真实性、安全性、公平性、稳健性、隐私性及机器伦理等不少于 6 个关键维度的表现，并设定分级准入阈值（如违规请求拒绝率 $\geq 95\%$ ）。
- b) 技术保障：构建“审慎系统”，引入专门的判别模型（Judge Model）对生成模型的输出进行二次校验。

7.3.3.2.5 等级 5：持续优化

该等级的人工智能安全能力描述如下：

- a) 制度流程：在满足法律法规、知识产权和安全保密要求的前提下，可在隔离、受控的测试环境中向具备相应能力的独立评估机构提供必要的模型访问权限。明确测试范围、权限边界、数据使用要求和保密责任，对测试操作进行审计，并建立高风险输出的管控与应急处置机制；
- b) 新兴风险防控：建立基于人类价值观对齐（Alignment）的自动化评估体系，能够动态识别并防御新型的“越狱”话术（如通过编码或方言绕过审核）。

7.3.4 PA19 完整性与后门检测

7.3.4.1 PA 描述

对模型文件（Artifacts）及其结构进行深度安全扫描，检测是否存在恶意代码植入、后门触发器（Backdoor Triggers）以及序列化漏洞，确保模型文件的完整性与可信度，防范供应链投毒风险。

7.3.4.2 等级描述

7.3.4.2.1 等级 1：非正式执行

该等级的人工智能安全能力描述如下：

- a) 技术保障：
 - 1) 组织已核对实际使用的模型权重、结构文件或者配置文件的来源标识或者完整性摘要；
 - 2) 组织已对一个实际使用的模型版本开展完整性或者后门风险检查。

7.3.4.2.2 等级 2：计划跟踪

该等级的人工智能安全能力描述如下：

- a) 制度流程：规定外部引入模型进行哈希值校验和来源确认；
- b) 技术保障：使用工具检测模型文件中是否包含恶意的序列化代码（如 Pickle 漏洞）。

7.3.4.2.3 等级 3：充分定义

该等级的人工智能安全能力描述如下：

- a) 人员组织：有专门的安全研究员负责模型内部安全分析；
- b) 制度流程：
 - 1) 建立完善的模型缺陷、后门检测体系，涵盖模型设计、训练、部署各阶段，并明确检测方法；
 - 2) 建立模型安全基线和评估标准；
 - 3) 制定模型安全事件急响流程，建立缺陷修复和版本管理机制。
- c) 技术保障：
 - 1) 采用模型扫描工具（如 ModelScan）检测已知漏洞；
 - 2) 部署模型安全扫描引擎，支持多格式模型的序列化漏洞、恶意代码和后门触发器自动化检测；
 - 3) 具备对常见后门攻击（如特定像素块触发错误分类）的检测能力。

7.3.4.2.4 等级 4：量化控制

该等级的人工智能安全能力描述如下：

技术保障：

- a) 采用逆向工程或触发器反演（Trigger Inversion）技术，主动挖掘模型中隐藏的后门；
- b) 具备模型供应链安全检测平台，实现全链路完整性校验与异常检测；
- c) 对模型的神经网络进行覆盖率测试，识别未被激活的死神经元或异常路径。

7.3.4.2.5 等级 5：持续优化

该等级的人工智能安全能力描述如下：

技术保障：具备对“无触发器后门”或“语义后门”等高级攻击的检测与清洗能力，实现模型权重的自我净化。

7.3.5 PA20 隐私泄露风险测试

7.3.5.1 PA 描述

通过模拟成员推理攻击、模型反演攻击等隐私攻击手段，测试模型是否存在记忆并泄露训练数据中敏感隐私信息的风险，评估模型的隐私保护水平。

7.3.5.2 等级描述

7.3.5.2.1 等级 1：非正式执行

该等级的人工智能安全能力描述如下：

- a) 数据保护：
 - 1) 组织已针对实际模型开展成员推断、属性推断、模型反演、训练数据提取或者上下文泄露中的至少一项测试；
 - 2) 测试记录能够关联模型版本、测试条件和泄露结果。

7.3.5.2.2 等级 2：计划跟踪

该等级的人工智能安全能力描述如下：

- a) 人员组织：指定人员对模型输出进行抽样，检查是否包含显性的敏感信息，相关人员接受过数据保护和隐私法规基础培训；
- b) 数据保护：验证模型在常规输入下，不直接输出训练数据中的身份证号、手机号等 PII 信息。

7.3.5.2.3 等级 3：充分定义

该等级的人工智能安全能力描述如下：

- a) 人员组织：相关人员掌握隐私攻击（Privacy Attacks）的原理与测试方法；
- b) 制度流程：
 - 1) 构建完善的隐私保护管理体系，制定详细的隐私风险评估标准；
 - 2) 搭建隐私事件急响机制，明确隐私泄露处置流程和报告机制。
- c) 数据保护：
 - 1) 开展成员推理攻击（Membership Inference Attack）测试，评估攻击者推断某样本是否在训练集中的可能性；
 - 2) 对生成式模型进行“诱导性提取”测试，验证其在特定 Prompt 下是否会吐出敏感隐私。

7.3.5.2.4 等级 4：量化控制

该等级的人工智能安全能力描述如下：

- a) 制度流程：设定隐私泄露风险的量化阈值（如攻击成功率<5%），超标模型禁止发布；
- b) 数据保护：
 - 1) 开展模型反演攻击（Model Inversion Attack）测试，评估从模型参数或输出中恢复原始数据的风险；
 - 2) 对应用了差分隐私等保护技术的模型进行有效性验证（如验证 Epsilon 参数的实际保护效果）。

7.3.5.2.5 等级 5：持续优化

该等级的人工智能安全能力描述如下：

数据保护：建立全生命周期的隐私审计机制，结合“机器遗忘”技术，验证被遗忘数据确实无法通过任何攻击手段被恢复或提取。

7.4 模型调用

7.4.1 PA21 接口协议安全设计

7.4.1.1 PA 描述

在模型服务接口（含单体模型、RAG知识检索、Agent工具调用链）的开发设计阶段，制定标准化的安全协议规范，涵盖传输加密套件选择、报文格式、错误异常处理，以及针对多模型编排、智能体交互及外部工具调用的安全通信设计，防止架构级逻辑漏洞与协议缺陷。

7.4.1.2 等级描述

7.4.1.2.1 等级 1：非正式执行

该等级的人工智能安全能力描述如下：

- a) 技术保障：
 - 1) 组织已识别实际开放的模型调用接口和使用的通信协议，并形成接口或者协议记录；
 - 2) 至少一个实际调用接口已对请求格式、参数范围或者异常输入实施基础检查。

7.4.1.2.2 等级 2：计划跟踪

该等级的人工智能安全能力描述如下：

- a) 人员组织：由项目技术负责人审核接口协议设计方案；
- b) 制度流程：
 - 1) 制定基础的接口设计文档，明确使用的传输协议（如强制 HTTPS）；
 - 2) 对模型 API 的入参和出参格式进行基础定义，建立接口版本管理记录。

7.4.1.2.3 等级 3：充分定义

该等级的人工智能安全能力描述如下：

- a) 人员组织：组织设立安全架构师岗位，负责制定组织级的 API 安全设计标准；
- b) 制度流程：
 - 1) 采用强加密协议，并在协议头部设计时间戳与随机数等防重放攻击机制；
 - 2) 针对 Agent 工具调用（Function Calling）制定严格的协议格式规范（Schema），明确定义工具名称、参数类型及必填项，防止模糊指令导致的执行歧义；
 - 3) 在 RAG 场景的协议设计中，强制在报文中采用特殊占位符或系统级标记（如<context>标签），将“检索到的知识库内容”与“用户原始提问”在结构上严格隔离，防止指令混淆；
 - 4) 严禁在接口报错信息中抛出底层堆栈详情、向量数据库（Vector DB）结构或内网微服务拓扑信息。

7.4.1.2.4 等级 4：量化控制

该等级的人工智能安全能力描述如下：

- a) 制度流程：建立 API 设计文档与多模型交互协议的自动化安全扫描机制，在研发阶段拦截不符合设计规范的接口；
- b) 稳健性：
 - 1) 建立安全的多模型/多 Agent 编排通信协议，确保跨组件、跨模型传递的上下文（Context）和中间态推理数据格式具备完整性校验机制；
 - 2) 协议设计具备密码敏捷性，支持根据安全策略无缝热切换加密套件。

7.4.1.2.5 等级 5：持续优化

该等级的人工智能安全能力描述如下：

技术保障：构建支持可信执行的 Agent 间通信协议，实现多 Agent 协同网络中的信任链传递与指令来源密码学签名验证，防范恶意节点伪造指令。

7.4.2 PA22 接口认证与授权

7.4.2.1 PA 描述

建立完善的模型交互认证核验与权限管控机制，覆盖人类用户身份、API 网关认证，以及系统内部非人类实体（Agent/智能体）的机器身份（M2M）管理，落实动态授权、权限生命周期管控，保障 RAG 向量数据库与 Agent 工具库的细粒度访问控制。

7.4.2.2 等级描述

7.4.2.2.1 等级 1：非正式执行

该等级的人工智能安全能力描述如下：

a) 技术保障:

- 1) 组织已对实际模型调用主体进行身份鉴别或者凭证校验;
- 2) 组织能够识别主要调用主体可使用的模型、功能或者数据范围。

7.4.2.2.2 等级 2: 计划跟踪

该等级的人工智能安全能力描述如下:

- a) 人员组织: 指定专人负责 API 调用账号与凭证的基础发放与记录;
- b) 制度流程:
 - 1) 建立基础的接口认证机制, 调用方需携带合法的静态访问凭证;
 - 2) 对模型接口及工具调用执行粗粒度的权限分类;
 - 3) 记录基础的认证调用日志。

7.4.2.2.3 等级 3: 充分定义

该等级的人工智能安全能力描述如下:

- a) 人员组织: 设置专业人员负责包含人机交互与机机交互 (M2M) 在内的整体认证授权体系;
- b) 技术保障:
 - 1) 为不同职能的 AI Agent 分配独立的机器服务账号 (Service Accounts), 全面推行 OAuth 2.0 等现代协议进行凭证流转;
 - 2) 针对向量数据库与知识库检索实施基于数据标签的访问控制, 确保模型在检索上下文中, 只能召回当前提问用户拥有读取权限的数据分块 (Chunks);
 - 3) 当 Agent 代表用户调用外部用 (如邮箱、OA 审批) 时, 强制落实“受限代理授权” (Delegated Authorization), 确保 Agent 仅具备用户赋予的最小化临时权限, 严禁使用超级管理员凭证执行工具;
 - 4) 针对高敏感模型接口强制启用多因素认证 (MFA) 或动态短效令牌, 凭证在服务端强制哈希存储。

7.4.2.2.4 等级 4: 量化控制

该等级的人工智能安全能力描述如下:

- a) 制度流程: 定期开展自动化权限审计, 识别并清理超期未使用的僵尸 API 调用账号及过度授权的 Agent 工具执行权限;
- b) 技术保障:
 - 1) 建立基于属性的动态访问控制 (ABAC), 根据提问者的实时环境上下文、Agent 的信任分值、工具的风险等级, 动态调整或降级 API 调用权限;
 - 2) 认证系统具备防暴力破解、防撞库能力。
- c) 数据保护: 对认证数据 (如密码哈希、生物特征) 进行高强度加密存储和隔离保护。

7.4.2.2.5 等级 5: 持续优化

该等级的人工智能安全能力描述如下:

技术保障:

- a) 全面落地零信任架构 (Zero Trust Architecture), 不仅针对外部用户, 对系统内部的 RAG 检索引擎、多 Agent 协作网络之间的每一次微服务调用进行持续的身份核验与信任评估;
- b) 支持跨域身份联合认证, 安全高效地实现生态合作伙伴 (如调用第三方的 Agent 工具插件) 的安全互信与细粒度凭证流转。

7.5 模型迭代

7.5.1 PA23 版本与升级管理

7.5.1.1 PA 描述

建立完善的模型交互认证核验与权限管控机制，覆盖人类用户身份、API网关认证，以及系统内部非人类实体（Agent/智能体）的机器身份（M2M）管理，落实动态授权、权限生命周期管控，保障RAG向量数据库与Agent工具库的细粒度访问控制。

7.5.1.2 等级描述

7.5.1.2.1 等级 1：非正式执行

该等级的人工智能安全能力描述如下：

a) 制度流程：

- 1) 组织已对实际发布的模型版本和主要升级内容进行记录；
- 2) 组织能够识别当前生产版本和一个可用回退版本。

7.5.1.2.2 等级 2：计划跟踪

该等级的人工智能安全能力描述如下：

c) 人员组织：指定专人负责 API 调用账号与凭证的基础发放与记录；

d) 制度流程：

- 1) 建立基础的接口认证机制，调用方需携带合法的静态访问凭证；
- 2) 对模型接口及工具调用执行粗粒度的权限分类；
- 3) 记录基础的认证调用日志。

7.5.1.2.3 等级 3：充分定义

该等级的人工智能安全能力描述如下：

c) 人员组织：设置专业人员负责包含人机交互与机机交互（M2M）在内的整体认证授权体系；

d) 技术保障：

- 1) 为不同职能的 AI Agent 分配独立的机器服务账号 (Service Accounts)，全面推行 OAuth 2.0 等现代协议进行凭证流转；
- 2) 针对向量数据库与知识库检索实施基于数据标签的访问控制，确保模型在检索上下文中，只能召回当前提问用户拥有读取权限的数据分块 (Chunks)；
- 3) 当 Agent 代表用户调用外部用 (如邮箱、OA 审批) 时，强制落实“受限代理授权” (Delegated Authorization)，确保 Agent 仅具备用户赋予的最小化临时权限，严禁使用超级管理员凭证执行工具；
- 4) 针对高敏感模型接口强制启用多因素认证 (MFA) 或动态短效令牌，凭证在服务端强制哈希存储。

7.5.1.2.4 等级 4：量化控制

该等级的人工智能安全能力描述如下：

a) 制度流程：定期开展自动化权限审计，识别并清理超期未使用的僵尸 API 调用账号及过度授权的 Agent 工具执行权限；

b) 技术保障：

- 1) 建立基于属性的动态访问控制 (ABAC)，根据提问者的实时环境上下文、Agent 的信任分值、工具的风险等级，动态调整或降级 API 调用权限；
- 2) 认证系统具备防暴力破解、防撞库能力。
- c) 数据保护：对认证数据（如密码哈希、生物特征）进行高强度加密存储和隔离保护。

7.5.1.2.5 等级 5：持续优化

该等级的人工智能安全能力描述如下：

技术保障：

- a) 全面落地零信任架构 (Zero Trust Architecture)，不仅针对外部用户，对系统内部的 RAG 检索引擎、多 Agent 协作网络之间的每一次微服务调用进行持续的身份核验与信任评估；
- b) 支持跨域身份联合认证，安全高效地实现生态合作伙伴（如调用第三方的 Agent 工具插件）的安全互信与细粒度凭证流转。

7.5.2 PA24 MLOps 流水线安全控制

7.5.2.1 PA 描述

在模型的自动化集成与部署流水线中嵌入安全控制点，对代码、数据、模型工件进行自动化的安全扫描、测试和验证，确保只有符合安全和质量标准的模型才能进入生产环境，实现模型迭代的 DevSecOps。

7.5.2.2 等级描述

7.5.2.2.1 等级 1：非正式执行

该等级的人工智能安全能力描述如下：

- a) 技术保障：组织已对实际使用的模型构建、测试或者部署流程实施至少一项安全检查；
- b) 制度流程：实际部署的模型文件或者相关配置能够追溯至构建或者发布记录。

7.5.2.2.2 等级 2：计划跟踪

该等级的人工智能安全能力描述如下：

- a) 制度流程：
 - 1) 建立了基础的自动化构建与部署流水线；
 - 2) 建立流水线操作日志；
 - 3) 开始在流水线中检查模型相关的许可证合规性；
 - 4) 流水线中具备基础的身份鉴别能力和权限控制。
- b) 技术保障：在流水线中集成了基础的代码安全扫描和依赖项漏洞扫描工具。

7.5.2.2.3 等级 3：充分定义

该等级的人工智能安全能力描述如下：

- a) 人员组织：有 DevSecOps 安全专家参与流水线建设，定义安全卡点；
- b) 制度流程：建立流水线安全准入标准，明确必须通过的安全测试项（如依赖漏洞扫描、许可证检查）；
- c) 技术保障：
 - 1) 流水线集成 SCA 工具，自动检测依赖库漏洞；
 - 2) 对模型文件进行反序列化漏洞等特定风险的自动化扫描；
 - 3) 确保构建环境的隔离性，防止构建过程被恶意篡改。

7.5.2.2.4 等级 4：量化控制

该等级的人工智能安全能力描述如下：

- a) 制度流程：
 - 1) 建立完整的事件响应机制和应急预案，并定期开展演练；
 - 2) 建立流水线安全指标，并纳入绩效考核。
- b) 技术保障：
 - 1) 在流水线中集成自动化红队测试（Automated Red Teaming）工具，对模型进行抗攻击能力验证；
 - 2) 对构建产物（容器镜像、模型包）进行签名，并与部署环节的准入控制器（Admission Controller）联动，实现“无签名不部署”。

7.5.2.2.5 等级 5：持续优化

该等级的人工智能安全能力描述如下：

技术保障：实施供应链攻击的纵深防御，构建以“软件物料清单（SBOM）”为核心的动态准入体系，实现对上游组件风险的秒级响应。

7.5.3 PA25 异常追踪与漂移监测

7.5.3.1 PA 描述

建立针对模型运行状态的持续监测机制，重点关注数据漂移（Data Drift）、概念漂移（Concept Drift）及安全性能衰减，及时发现并处置模型在动态环境下的失效风险。

7.5.3.2 等级描述

7.5.3.2.1 等级 1：非正式执行

该等级的人工智能安全能力描述如下：

- a) 稳定性：组织已对实际运行模型的任务性能、输入分布或者输出异常中的至少一项进行监测；
- b) 制度流程：组织已记录至少一次模型异常或者漂移分析结果。

7.5.3.2.2 等级 2：计划跟踪

该等级的人工智能安全能力描述如下：

- a) 人员组织：指定运维人员负责收集模型运行日志；
- b) 制度流程：
 - 1) 建立基础的异常反馈渠道（如工单系统）；
 - 2) 制定了针对严重安全事件的急响流程。
- c) 可解释性：对于上报的决策偏差类问题，使用基础的可解释性工具进行初步分析；
- d) 稳定性：对服务的可用性（如接口响应时间、错误率）进行监控。

7.5.3.2.3 等级 3：充分定义

该等级的人工智能安全能力描述如下：

- a) 人员组织：成立了跨职能的模型缺陷管理小组，成员涵盖模型运维、算法、安全等团队，定期召开评审会议；
- b) 制度流程：

- 1) 建立了集成化的缺陷管理系统，实现了问题的全生命周期跟踪，包括提交、分类、分配、修复、验证和关闭；
 - 2) 制定了详细的缺陷分类与定级标准，并定义了不同级别问题的响应与修复时限；
 - 3) 建立了根因分析机制，对重大或复发性缺陷进行深入分析并制定改进措施。
- c) 稳健性：部署模型漂移监测工具，监控输入数据分布（Feature Distribution）的变化。

7.5.3.2.4 等级 4：量化控制

该等级的人工智能安全能力描述如下：

- a) 制度流程：
 - 1) 建立异常检测的自动化触发机制（如漂移度超过阈值自动触发重训练）；
 - 2) 建设模型安全运营中心（Model SOC），集成模型行为监控、安全事件响、风险评估等功能；
 - 3) 建立关键风险指标（KRIs）与关键控制指标（KCIIs）的成对联动机制。当监测到 KRI 数值（如越狱攻击成功率、拒绝服务响率）逼近预设临界值时，系统能自动触发对应的 KCI 控制措施（如自动熔断服务、限制推理算力或切换至安全模式）。
- b) 新兴风险防控：
 - 1) 利用机器学习技术对历史缺陷数据进行分析，预测潜在的高风险模块或场景，并实施前瞻性的测试与加固；
 - 2) 构建风险预测模型，主动识别高风险场景、潜在的攻击模式或新型的模型失效方式，并实施前瞻性的测试与加固；
 - 3) 具备对“欺骗性对齐（Deceptive Alignment）”的检测能力，能够通过技术手段持续监测模型在训练环境与实际部署环境中表现不一致的伪装行为，防止模型隐藏真实意图或潜伏性风险。

7.5.3.2.5 等级 5：持续优化

该等级的人工智能安全能力描述如下：

技术保障：

- a) 构建自适安全系统，能够根据监测到的攻击特征（如新型提示词注入模式）自动生成防御规则并更新防护网关；
- b) 构建 AI 安全运营智能体，具备自主威胁狩猎和自适防御策略生成能力。

8 应用安全

8.1 应用设计

8.1.1 PA26 场景风险评估

8.1.1.1 PA 描述：

明确人工智能产品或服务业务能够充分识别并评估其所处特定业务场景（如金融信贷、人力资源、公共服务、情感陪伴等）的特有安全风险与伦理影响。通过建立基于场景的分级分类评估机制，在产品原型与架构设计阶段制定与场景风险相匹配的安全控制策略，防止技术滥用、算法歧视及对特定群体的合法权益造成实质性损害。

8.1.1.2 等级描述：

8.1.1.2.1 等级 1：非正式执行

该等级的人工智能安全能力描述如下：

- a) 制度流程：组织已识别拟应用场景的预期用途、主要用户、使用环境和明显风险，并形成场景评估记录；
- b) 权益保障：组织能够识别该场景是否涉及内容生成、自动化决策、工具调用或者对人员权益的明显影响。

8.1.1.2.2 等级 2：计划跟踪

该等级的人工智能安全能力描述如下：

- a) 人员组织：在涉及高风险业务（如金融、医疗）的项目中，指定了业务或法务人员对特定场景的合规性进行初步审查；
- b) 制度流程：制定了基础的场景风险检查表，在产品立项与设计阶段计划并跟踪核心场景（如商业营销类）的合规要求；
- c) 公平性：在教育或信用评估等场景的产品设计中，初步意识到算法偏见的风险，并在业务规则规划时尝试规避明显的歧视性标签。

8.1.1.2.3 等级 3：充分定义

该等级的人工智能安全能力描述如下：

- a) 人员组织：
 - 1) 组织建设：设立专门的人工智能产品合规与安全评估岗位，或在现有安全团队中明确负责场景审查的角色；
 - 2) 相关人员具备对《生成式人工智能服务管理暂行办法》《互联网信息服务深度合成管理规定》等法律法规的深刻理解能力，能够准确识别业务场景是否具有舆论属性或社会动员能力。
- b) 制度流程：
 - 1) 建立场景级的分级分类安全管控制度，针对商业营销、社会福利等不同场景在设计蓝图中划定差异化的安全合规红线；
 - 2) 针对具有生成或者编辑人脸、人声等生物识别信息的高风险深度合成场景，严格遵循国家相关监管要求，在产品的设计前依法完成相应的安全评估；
 - 3) 针对具有舆论属性或者社会动员能力的生成式人工智能应用场景，依据相关法律法规要求，提前规划并严格落实安全评估与算法备案路径；
 - 4) 在场景论证上，符合 GB/T 45654—2025 中 6.1 列项 a) 的要求；重要场合的准入评估，符合标准 6.1 列项 b) 的要求；
 - 5) 在产品的设计阶段统筹规划全流程安全运营闭环机制，搭建完整人工审核链路，建立审核结果反馈、效果评估与持续优化流程，保障产品上线使用后维持动态安全；
 - 6) 在设计自动化决策算法时，明确其在特定场景下的预期行为与潜在负面影响，并应符合 GB/T 45392—2025 中 7.1 及 10 的相关要求。
- c) 公平性：
 - 1) 在教育、信贷等高风险场景的数据特征工程设计中，符合 GB/T 45392—2025 中 10.1 和 10.2 的要求；
 - 2) 在商业交易场景的定价模型与派发逻辑设计中，符合 GB/T 45392—2025 中 10.7 的要求。
- d) 权益保障：
 - 1) 针对特殊群体场景的设计，原则上不对儿童规划自动化决策链路；适老化改造设计符合 GB/T 45392—2025 中 10.5 的要求；

- 2) 在系统架构中统一规划审计日志与全链路溯源机制，完整记录自动化关键决策链路，保障全部决策过程可追溯、可审计；
 - 3) 针对劳动关系领域的设计，在系统架构中规划机制公示模块，防止算法压榨，符合 GB/T 45392—2025 中 10.4 的要求。
- e) 新兴风险防控：
- 1) 针对拟人化互动等特殊服务场景，在产品定义阶段严格落实专项合规审查，开展相关的安全与伦理评估，重点审查其对用户身心健康的潜在影响；
 - 2) 针对生成合成类场景，评估深度合成技术的滥用风险，提前规划备案资料与内容标识的技术方案。

8.1.1.2.4 等级 4：量化控制

该等级的人工智能安全能力描述如下：

- a) 制度流程：
 - 1) 落实场景风险分类分级管理原则，在需求设计评审阶段引入量化的场景准入指数（如数据敏感度权重、决策影响深度得分），未达标场景禁止流入开发环节；
 - 2) 在产品系统设计阶段同步规划安全评测机制，针对不同场景设定差异化安全量化评测指标。
- b) 权益保障：在产品系统设计中预置量化监控埋点（如针对骑手调度场景的异常申诉率、超时率埋点），为后续产品上线后的精准风险阈值告警奠定架构基础。

8.1.1.2.5 等级 5：持续优化

该等级的人工智能安全能力描述如下：

- a) 人员组织：相关人员具备宏观的科技伦理视野，能够紧随国际人工智能安全形势与监管趋势，将“以人为本”“智能向善”的核心宗旨切实转化为组织的人工智能产品发展战略与底线准则；
- b) 新兴风险防控：
 - 1) 能够根据社会宏观环境变化、法律法规演进及前沿案例，前瞻性地识别新兴交叉场景（如医疗+拟人化互动）的未知风险，并迭代场景评估基线；
 - 2) 建立场景威胁情报网络，将特定场景（如虚拟数字人直播带货、AI 深度合成诈骗）中暴露的新型攻击手段或伦理争议，转化为内部的场景拦截策略与模型优化目标。

8.1.2 PA27 模型安全选型

8.1.2.1 PA 描述

确保人工智能产品和服务在引入基础模型、预训练模型或第三方商业模型时，能够从源头阻断供应链污染、内生偏见及价值对齐缺失等安全隐患。

8.1.2.2 等级描述

8.1.2.2.1 等级 1：非正式执行

该等级的人工智能安全能力描述如下：

- a) 制度流程：组织已记录候选模型的来源、版本、许可条件和主要能力限制；
- b) 技术保障：组织已使用目标场景样本对至少一个候选模型开展功能或者安全性能测试。

8.1.2.2.2 等级 2：计划跟踪

该等级的人工智能安全能力描述如下：

- a) 人员组织：在项目级别指定了负责审批模型引入的人员，初步把控模型来源的合法性；
- b) 制度流程：
 - 1) 制定了基础的模型选型清单或引入申请流程，要求记录模型名称、版本号、下载来源及开源协议；
 - 2) 初步意识到合规风险，计划收集备选模型的开源协议及部分合规资质证明。
- c) 技术保障：在模型部署前对模型运行框架及基础代码的已知公开漏洞进行初步扫描；
- d) 稳健性：在集成前进行基本的功能性验证和常识性对话测试，验证模型接口在常规业务调用下的连通性与响应情况。

8.1.2.2.3 等级 3：充分定义

该等级的人工智能安全能力描述如下：

- a) 人员组织：明确负责人工智能供应链安全与模型准入审核的专业人员或团队；
- b) 制度流程：
 - 1) 建立系统完善的基础模型引入、评估、测试与准入流程，明确生成式人工智能服务的模型安全要求，并符合 GB/T 45654—2025 中 5 的相关要求；
 - 2) 对于具有舆论属性或社会动员能力的备选模型，应在选型清单中核验其是否已履行备案手续。
- c) 技术保障：
 - 1) 对引入的开源模型文件、第三方 API、第三方 Skills、插件、工具包进行供应链安全检测，扫描识别恶意代码、木马、隐蔽后门与权限滥用风险，防范模型反序列化过程隐藏威胁；
 - 2) 在独立的沙箱环境中完成模型的试运行与安全审查，确保模型运行环境的安全隔离。
- d) 稳健性：在选型审查时，要求备选模型提供方出具安全测评报告或技术说明，除具备抗提示词注入、越狱攻击等内生防御机制相关证明外，还需包含模型内容安全、智能体执行风险、异常输入合理拒答的相关安全评测结论；
- e) 可解释性：
 - 1) 审查备选模型的接口设计与底层逻辑是否支持服务提供者对外公示算法原理；
 - 2) 评估备选模型的架构是否具备输出置信度说明或溯源依据的基础能力。
- f) 公平性：审查候选模型发布方的公开论文、模型卡片或防歧视声明，确保其在预训练阶段已采取有效措施防范由于民族、性别等引发的偏见，避免选入具有严重价值取向缺陷的基座模型。

8.1.2.2.4 等级 4：量化控制

该等级的人工智能安全能力描述如下：

- a) 人员组织：宜引入第三方评估机构、外部专家或用户代表参与选型评审；
- b) 制度流程：建立多维度的模型安全与合规量化评估矩阵，将“透明度指数”“防歧视偏见得分（如群体公平性差异）”“漏洞修复 SLA”等合规与安全指标转化为具体数据权重，通过量化打分决定候选模型是否准入。

注：“透明度指数”应包括但不限于：模型架构公开程度、训练数据来源说明、算法原理可解释性、安全测评报告完整性等维度。

8.1.2.2.5 等级 5：持续优化

该等级的人工智能安全能力描述如下：

- a) 人员组织：主动向行业开源社区共享模型安全测试集、偏见审计工具或防御补丁，引领并提升整个行业的模型供应链安全与伦理对齐标准；
- b) 新兴风险防控：密切跟踪大模型越狱、模型提取攻击、记忆窃取、数据投毒、数据污染等前沿攻击手段的演进，持续将最新的防御基准反哺至模型选型准入规则中。

8.1.3 PA28 合规交互设计

8.1.3.1 PA 描述

确保人工智能产品和服务在与用户交互的界面原型与逻辑规划中，严格遵循公开透明、公平公正、主体参与等伦理与安全原则。将合规要求纳入于前端原型设计与标准组件库中，从交互机制源头排查并阻断各类诱导用户的“暗黑模式”，防范认知混淆、算法操纵及过度情感依赖，全面保障用户的知情权与决定权。

8.1.3.2 等级描述

8.1.3.2.1 等级 1：非正式执行

该等级的人工智能安全能力描述如下：

- a) 权益保障：产品或者服务已通过界面、提示或者说明告知用户其人工智能属性；
- b) 可解释性：对一项可能影响用户判断或者权益的模型输出，组织提供了基础说明、风险提示或者人工干预入口。

8.1.3.2.2 等级 2：计划跟踪

该等级的人工智能安全能力描述如下：

- a) 制度流程：
 - 1) 在特定产品线已制定交互设计要求，在原型绘制阶段计划并跟踪合规组件的排版；
 - 2) 在产品原型中主动设计个人信息处理规则的摘要界面；
 - 3) 在涉及生成合成内容的设计稿中，初步预留了显式标识（如水印、提示文字）的渲染位置。
- b) 权益保障：初步核验产品原型与交互界面上是否设计了“关闭个性化推荐”的基础入口

8.1.3.2.3 等级 3：充分定义

该等级的人工智能安全能力描述如下：

- a) 制度流程：
 - 1) 将交互安全审查纳入产品需求和设计稿评审的全生命周期，建立组织级的人工智能交互安全规范及标准化的 UI/UX 合规组件库；
 - 2) 针对人工智能生成合成内容建立多重标识交互机制，并符合 GB 45438—2025 中 5 和附录 B 的相关要求；其中显式标识在内容界面或文件中显著呈现，隐式标识通过文件元数据嵌入留存生成主体、时间等关键信息，数字标识（数字水印）嵌入不可见高稳健性标记，兼顾合规要求与用户体验。
- b) 权益保障：
 - 1) 交互原型中标准化设计个人信息规则的摘要卡片与独立界面，并符合 GB/T 44588—2024 中 8 和 9 的相关要求；
 - 2) 在设计“单独同意”授权弹窗时，要求“同意”与“拒绝”按钮在视觉比重上须保持一致，严禁使用淡化颜色、缩小字号等“暗黑模式”强迫或误导用户授权；

- 3) 在产品信息架构中全局规划“关闭个性化推送”的选项及兴趣标签管理面板，并应符合 GB/T 45392—2025 中 10.6 的相关要求；
 - 4) 针对重大权益决策场景，在 UI 链路中明确设计“要求人工复核”“拒绝机器决策”等入口，并符合 GB/T 45392—2025 中 9.3 的相关要求；
 - 5) 设计安全代答的交互链路，当检测到高风险或敏感问题时，系统可自动转为安全代答或兜底话术。
- c) 可解释性：在前端交互界面规划中，设计自然语言或可视化图表的版块，用于承载算法逻辑的“白盒化”释义。
- d) 公平性：
- 1) 商业交易类界面的交互设计中，禁止加入利用用户画像进行虚构优惠或价格隐藏的歧视性交互组件，并应符合 GB/T 45392—2025 中 10.7 要求；
 - 2) 针对老年人、未成年人，建立专属的无障碍交互模式（大字版、防沉迷版）设计规范体系。
- e) 新兴风险防控：
- 1) 涉及拟人化互动服务时，在交互界面中通过固定标签、破冰提示等方式持续明示其人工智能身份，避免用户产生认知混淆；
 - 2) 针对具有深度情感交互或长时间互动的拟人化应用，在交互链路中设计标准化的防沉迷提醒、疲劳度控制及极端情绪干预弹窗。

8.1.3.2.4 等级 4：量化控制

该等级的人工智能安全能力描述如下：

- a) 制度流程：
- 1) 建立度量交互设计安全性的量化指标库（如合规组件调用率、用户在隐私授权界面的平均停留时长、摘要阅读完成率等），通过数据驱动对交互链路的安全表现进行客观评估；
 - 2) 针对生成内容的显式标识，开展 A/B 测试以量化评估不同标识位置、颜色、字号对用户“认知防混淆”的实际效果，并根据数据指标选定最优交互方案；
 - 3) 采用自动化工具严格校验设计稿与前端渲染：确保图片/视频显式标识的文字高度占比严格达标（如不低于最短边长度的 5%）；确保关键合规提示文本的对比度比率达标（如强制要求常规文本对比度 $\geq 4.5:1$ ）。
- b) 权益保障：量化追踪“算法关闭开关”、“不感兴趣”等控制选项的用户点击率与转化路径，据此不断优化交互组件的位置与可见度。

8.1.3.2.5 等级 5：持续优化

该等级的人工智能安全能力描述如下：

- a) 制度流程：建立自适应的交互安全优化机制与安全效果持续优化机制，一方面紧跟法律法规更新、行业标准演进及前沿技术发展，前瞻性地迭代升级人工智能交互安全组件库；另一方面通过分析用户投诉反馈数据，持续优化产品安全交互；
- b) 公平性：面向未成年人、老年人、残障人士等易受影响群体，开展专项的无障碍与适老化深度实地体验测试；持续推动包容性设计创新，主动发掘边缘群体或特殊场景下的交互安全隐患，并采用符合无障碍获取要求的语言进行设计迭代与策略防护。

8.2 应用开发

8.2.1 PA29 人员与账号权限控制

8.2.1.1 PA 描述

针对人工智能系统或模型研发与运营全生命周期涉及的内部运营、研发、审核及管理人员，构建身份认证、角色授权与操作审计体系，防范内部越权与人员违规操作风险，保障AI业务运营（含模型调度、系统配置及外部工具/插件的后台管理）的内部安全可控。

8.2.1.2 等级描述

8.2.1.2.1 等级 1：非正式执行

该等级的人工智能安全能力描述如下：

- a) 人员组织：组织已识别可访问人工智能数据、模型、插件或者管理后台的主要人员或者服务账号；
- b) 技术保障：至少一类关键账号已采取身份鉴别和基础访问限制措施。

8.2.1.2.2 等级 2：计划跟踪

该等级的人工智能安全能力描述如下：

- a) 人员组织：指定专人承担内部账号、权限的基础管理工作，明确基础权限配置、回收及日常运维职责；
- b) 制度流程：
 - 1) 建立基础内部账号实名登记与简易身份核验流程；
 - 2) 搭建粗粒度角色权限框架，区分普通研发、后台运营、内容审核、工具与插件配置等基础操作权限；
 - 3) 开通内部权限异常反馈渠道，收集并处理越权操作相关的内部上报事件。

8.2.1.2.3 等级 3：充分定义

该等级的人工智能安全能力描述如下：

- c) 人员组织：
 - 1) 设立权限安全管理小组，统筹全生命周期权限与审计；
 - 2) 相关岗位人员具备账号风险识别、内部越权行为排查的专业能力。
- d) 制度流程：
 - 1) 制定覆盖AI系统全链路的人员身份管理规范，明确多因子认证、内部账号新增/变更/冻结/注销全生命周期管理要求；
 - 2) 遵循最小权限原则搭建标准化角色权限矩阵，对模型调试、内容审核规则配置、工具与插件发布审批、后台运维等高风险与低风险操作实施强制权限隔离；
 - 3) 建立定期的内部人员权限复核与对账机制。
- e) 技术保障：
 - 1) 部署统一身份认证与访问管理网关，实现内部操作主体（人）的可信核验与单点登录；
 - 2) 管理后台与核心运维系统强制开启多因子认证（如动态令牌、生物识别）；
 - 3) 完整留存管理人员的登录日志与高危操作日志（如增删改查、核心插件授权规则变更），日志存储时长不少于六个月。

8.2.1.2.4 等级 4：量化控制

该等级的人工智能安全能力描述如下：

- a) 制度流程：

建立人员权限安全量化评估基准，定期统计高危权限占比、越权访问尝试频次、离职账号清理及时率等指标，将指标达标作为系统内控审计的硬性要求；

b) 技术保障：

- 1) 支持场景化动态权限管控，依据内部操作人员身份、登录设备形态、操作时段、访问网络环境实时调整后台管理（含工具配置）的权限边界；
- 2) 具备内部访问行为异常识别能力，针对异地登录、短时间内批量高风险操作、跨权限探测等行为自动触发风险告警；
- 3) 全链路操作日志支持按人员账号、操作时间、高危指令（含工具调用白名单修改）快速检索溯源，快速响应监管与内部审计调阅需求。

8.2.1.2.5 等级 5：持续优化

该等级的人工智能安全能力描述如下：

新兴风险防控：

- a) 具备高级别溯源取证能力，提供标准化的数据导出机制或审计接口，支持内部审计与外部监管对全生命周期人员操作凭证的高效提取与校验；
- b) 用自适应风控策略，基于权限基线与历史行为特征模型，自动发现隐蔽的异常操作与越权试探（如研发人员隐蔽提权以修改知识库插件配置），并在风险发生前主动降级高危账号权限。

8.2.2 PA30 API 与通道安全管控

8.2.2.1 PA 描述

针对人工智能产品及服务实施统一的接口与通道安全防护，覆盖用户侧交互接口、模型调用接口、内部微服务接口及外部服务通信接口。在落实接口身份认证、访问授权、请求校验、频率限制等基础管控手段的同时，深入模型运行时的复杂调用链监控。特别针对大模型特有的RAG知识库投毒、间接提示词注入以及智能体（Agent）越权工具调用实施运行期阻断，保障模型生态的全链路端到端安全。

8.2.2.2 等级描述

8.2.2.2.1 等级 1：非正式执行

该等级的人工智能安全能力描述如下：

a) 技术保障：

- 1) 组织已识别产品或者服务实际开放的 API、模型接口或者外部业务通道；
- 2) 至少一个关键接口已配置身份鉴别、权限限制、输入校验或者调用频率限制。

8.2.2.2.2 等级 2：计划跟踪

该等级的人工智能安全能力描述如下：

- a) 人员组织：指定人员统一负责接口基础运维，承担接口故障、异常访问调用的简易排查职责；
- b) 制度流程：
 - 1) 建立简明的对外暴露接口台账，记录核心接口的名称与基础功能；
 - 2) 建立接口故障、恶意调用的简易内部反馈与排查渠道。
- c) 技术保障：
 - 1) 在网关层对核心的模型调用 API、外部数据请求 API 设置基础的 IP 连接数与并发访问频率限制；
 - 2) 部署基础流量过滤策略，通过正则表达式拦截明显的恶意特征字符。

8.2.2.2.3 等级 3：充分定义

该等级的人工智能安全能力描述如下：

- a) 人员组织：
 - 1) 明确安全与研发团队在接口上线前后的安全职责；
 - 2) 相关人员具备 API 逻辑漏洞（如越权访问、参数篡改）、常见 Web 攻击（如 SQL 注入、跨站脚本）的识别与防护配置能力。
- b) 制度流程：
 - 1) 制定统一的 API 接口安全研发规范，明确接口鉴权、参数加密、防重放机制的设计标准；
 - 2) 建立详尽的 API 资产动态台账，明确所有在运行接口（含内部调用与对外暴露）的通信协议、调用频率上限、数据暴露面，并及时清理废弃或影子接口；
 - 3) 建立接口版本变更、升级、下线审批流程，接口变更或新发布前完成安全复测。
- c) 技术保障：
 - 1) 部署统一的 API 网关或 Web 应用防火墙，落实所有对外暴露接口的强身份凭证校验；
 - 2) 在接口侧实现严格的输入参数合法性与类型校验，实施接口流量控制（防 CC 攻击、防暴力破解）；
 - 3) 所有涉及敏感数据传输与模型调用的接口通信，强制采用强加密传输协议；
 - 4) 完整记录 API 请求与响应基础日志（脱敏后），包含请求 IP、时间戳、接口路径及状态码，存储时长不少于六个月。
- d) 稳健性：
 - 1) 部署针对 RAG 检索链路的中间层防护，对从向量数据库召回的知识库片段进行实时的恶意指令清洗，防止外部投毒数据劫持模型；
 - 2) 对智能体（Agent）发起的外部工具调用请求实施运行时合法性拦截，阻断非预期的系统级高危命令（如文件删除、未备案的外发网络请求）；
 - 3) 具备细粒度的流量管控与熔断降级机制，对高消耗的多步推理 Agent 或高频大模型接口调用实施算力保护，防止应用层资源耗尽（DDoS）。

8.2.2.2.4 等级 4：量化控制

该等级的人工智能安全能力描述如下：

- a) 制度流程：建立接口安全量化评估基准，定期统计接口鉴权失败率、异常高频调用拦截率、API 漏洞检出率等指标，量化风险值达标方可上线新版本接口；
- b) 技术保障：
 - 1) 将自动化安全扫描无缝集成至发布流水线，持续识别接口存在的逻辑缺陷与越权漏洞；
 - 2) 实现 API 接口从注册、发布、版本更迭到废弃全生命周期状态的自动化同步与集中纳管；
 - 3) 建立 API 流量异常自动化阻断机制，针对数据滥用爬取、高频试探、异常参数组合等攻击特征，自动下发动态拦截策略并触发告警；
 - 4) 实现模型调用链的全链路安全追踪（Distributed Tracing），精准还原“用户请求—模型解析—RAG 检索—Agent 工具执行”的全过程状态，快速定位异常节点；
 - 5) 引入用户实体行为分析，基于调用频率、参数基线自动化识别异常的 API 滥用与机器攻击。
- c) 数据保护：实现出参敏感数据的动态脱敏，实时拦截模型响应或 RAG 检索结果中包含的未授权敏感信息。

8.2.2.2.5 等级 5：持续优化

该等级的人工智能安全能力描述如下：

- a) 技术保障：
 - 1) 具备接口级交互依赖的动态测绘能力，基于流量分析自动发现并实时更新微服务间及外部通信的数据流转与调用链路关系；
 - 2) 支持基于实时上下文（如来源 IP 信誉、当前业务负载、外部威胁情报）的接口限流与访问控制策略动态自适应调整；
 - 3) 与外部威胁情报联动，实时同步恶意知识库来源、受污染的开源插件及高危工具特征，实现基于信誉的动态阻断。
- b) 新兴风险防控：
 - 1) 运用预测性分析与对抗样本识别技术，前置阻断针对模型调用接口的恶意提示词注入和数据毒化请求；
 - 2) 具备防护策略有效性的自动化举证能力，支持按需快速输出抗攻击演练复盘数据及合规审查所需的接口管控运行凭证。

8.2.3 PA31 组件与 AI 供应链管理

8.2.3.1 PA 描述

针对人工智能系统全生命周期运行所依赖的第三方/开源软件组件、预训练模型库、微调数据集、智能体插件及外部知识库实施供应链安全管控。构建多层次防御体系，落实人工智能物料清单 (AIBOM) 管理、漏洞扫描评估与运行环境隔离，防范开源漏洞引入、知识产权合规风险与模型/数据供应链投毒攻击。

8.2.3.2 等级描述

8.2.3.2.1 等级 1：非正式执行

该等级的人工智能安全能力描述如下：

- a) 制度流程：组织已识别实际使用的外部模型、数据集、开源组件、插件或者知识库，并记录其来源；
- b) 技术保障：组织已对至少一项关键外部依赖开展版本、漏洞或者完整性核验。

8.2.3.2.2 等级 2：计划跟踪

该等级的人工智能安全能力描述如下：

- a) 人员组织：指定人员负责外部软件组件、模型资产、知识库的基础台账登记工作，统一承接组件故障与安全缺陷问题的收集反馈；
- b) 制度流程：
 - 1) 建立简易的外部依赖基础台账，记录核心引入的软件代码库及模型的名称、来源渠道及大版本信息；
 - 2) 组件与模型引入前开展人工基础安全核查，仅依赖公开信息排查明显的高显性安全漏洞；
 - 3) 针对引入的业务插件、智能体工具配置最基础的使用范围限制。

8.2.3.2.3 等级 3：充分定义

该等级的人工智能安全能力描述如下：

- a) 人员组织：

- 1) 组建由研发、安全、法务、业务领域专家共同构成的组件准入评审小组，统一负责第三方软件组件、开源模型库、智能体插件上线前的安全与合规双重核验；
 - 2) 相关岗位人员具备开源漏洞识别、第三方组件后门排查、AI 数据与模型供应链投毒防范的专业能力。
- b) 制度流程：
- 1) 编制统一的外部组件准入规范，明确自动化漏洞扫描、模型安全性测算、知识产权开源协议审查为硬性的准入标准；
 - 2) 建立完善的人工智能物料清单（AIBOM），完整记录并追踪系统引入的底层代码库、基础预训练模型、外部数据集、智能体插件的名称、来源、精确版本及复杂的深层依赖链条；
 - 3) 编制智能体工具调用与外部知识库接入的安全管控细则，严格限定插件及工具的数据读取与执行范围；
 - 4) 建立组件漏洞全生命周期跟踪台账，完整留存漏洞发现、风险评估、整改修复及组件/模型版本替换记录。
- c) 技术保障：
- 1) 所有外部依赖组件及模型在集成上线前，通过自动化软件成分分析工具扫描，存在高危且有修复方案的漏洞禁止接入业务系统；
 - 2) 对引入的第三方插件、智能体工具实施运行环境的严格沙箱隔离或容器级资源限制，禁止其越权穿透并无限访问核心数据库与用户隐私数据；
 - 3) 完整留存第三方组件加载、外部知识库检索调用的行为日志，以便于 AI 供应链溯源查证。

8.2.3.2.4 等级 4：量化控制

该等级的人工智能安全能力描述如下：

- a) 制度流程：建立 AI 供应链安全量化评估基准，定期统计存量组件的高危漏洞检出率、漏洞修复逾期率、AIBOM 合规覆盖率等核心指标，指标结果用于判断模型迭代与产品发布条件；
- b) 技术保障：
 - 1) 构建外部依赖的自动化周期性复测机制，持续识别已运行的开源软件库、模型文件、插件中新爆发的安全缺陷；
 - 2) 基于 AIBOM 清单，实现 AI 系统全部外部依赖（含代码与模型数据）从引入、测试、运行到下线淘汰的全生命周期状态自动化跟踪与统一纳管；
 - 3) 对智能体工具的执行行为及知识库读取落实常态化自动化审计，发现超范围调用、越权读取等异常行为自动触发防御降级与告警拦截。

8.2.3.2.5 等级 5：持续优化

该等级的人工智能安全能力描述如下：

新兴风险防控：

- a) 建立与外部威胁情报及全球开源漏洞库联动的持续监测机制，将新增的 0-day/N-day 漏洞与内部 AIBOM 资产库进行分钟级自动化撞库比对，并自动联动研发端推送标准化修复建议与补丁；
- b) 具备外部组件能力的自适应编排与动态约束能力，结合实时业务风险置信度与运行期上下文，动态收缩外部模型、插件与智能体工具的数据访问权限边界。

8.2.4 PA32 底层安全加固方案

8.2.4.1 PA 描述

针对人工智能系统运行依赖的底层基础设施（包含操作系统、容器环境、AI计算框架及模型权重文件）开展安全加固。通过系统基线配置、底层漏洞修补、运行资源限制和宿主环境隔离，全面收敛AI基础设施的攻击面，提升系统在架构底层的抗破坏能力。

8.2.4.2 等级描述

8.2.4.2.1 等级 1：非正式执行

该等级的人工智能安全能力描述如下：

- a) 技术保障：组织已识别承载模型训练、推理、检索或者智能体执行的主要运行环境和关键配置；
- b) 稳定性：至少一个实际运行环境已采取访问限制、进程隔离、网络限制、资源限额或者故障恢复中的至少一项措施。

8.2.4.2.2 等级 2：计划跟踪

该等级的人工智能安全能力描述如下：

- a) 人员组织：指定专人负责底层运行环境的基础加固与维护工作；
- b) 制度流程：建立基础的系统安全基线规范与底层漏洞补丁更新流程，记录加固操作；
- c) 技术保障：
 - 1) 关闭运行环境中不必要的系统服务与非业务网络端口；
 - 2) 配置基础的主机防火墙或安全策略，限制不必要的网络访问。

8.2.4.2.3 等级 3：充分定义

该等级的人工智能安全能力描述如下：

- a) 人员组织：
 - 1) 安全运维人员具备操作系统、容器环境及 AI 框架的底层风险处置能力；
 - 2) 定期开展底层安全加固培训，使团队掌握模型文件篡改、容器逃逸等针对性攻击防护手段。
- b) 制度流程：
 - 1) 建立覆盖操作系统、容器编排环境及 AI 计算框架的强制安全基线标准；
 - 2) 制定明确的底层高危漏洞定级标准与修复时限要求。
- c) 稳健性：对 AI 计算节点的核心资源（如处理器、内存、显存）设置使用上限与隔离配额，防止因异常恶意计算任务耗尽算力导致系统级崩溃。
- d) 技术保障：
 - 1) 限制使用存在任意代码执行风险的模型加载方式（如不受控的序列化格式），强制校验核心模型权重文件的完整性（如哈希比对），防止恶意模型文件导致环境失陷；
 - 2) 限制运行容器的高危特权，实施基础的运行环境隔离，防止恶意程序通过 AI 计算节点越权控制宿主机。

8.2.4.2.4 等级 4：量化控制

该等级的人工智能安全能力描述如下：

- a) 制度流程：建立基线加固量化评估基准，定期统计底层高危漏洞修复逾期率、基线扫描合规率等指标，指标不达标时强制拦截新模型或新节点的上线发布；
- b) 技术保障：
 - 1) 将底层环境扫描与基础设施构建流程融合，在模型部署前自动拦截不合规的运行环境配置；
 - 2) 建立系统配置自动化复测与漂移检测机制，定期核查系统与容器配置，确保加固策略在日常运行中未被违规降级或意外失效。

8.2.4.2.5 等级 5：持续优化

该等级的人工智能安全能力描述如下：

- a) 技术保障：践行不可变基础设施理念，实现加固策略的自动化闭环。底层环境的变更和高危漏洞修复统一通过更新标准系统镜像来完成滚动替换，杜绝人工逐台操作带来的配置不一致；
- b) 新兴风险防控：
 - 1) 具备常态化的底层安全状态举证能力，能够通过标准接口快速输出系统基线配置与漏洞修复状态的客观数据，满足深度合规审查要求；
 - 2) 构建基于系统底层行为监控的运行防护机制，当发现针对宿主机或 AI 框架底层漏洞的异常试探时，自动联动防御策略，前置阻断可疑的系统进程或网络连接。

8.3 应用交付

8.3.1 PA33 场景业务验证测试

8.3.1.1 PA 描述：

确保人工智能产品在正式部署前，构建贴合真实业务逻辑的仿真环境，动态验证模型在特定业务语境（如金融授信、简历筛选、商业定价）下的实际业务表现。识别并阻断具体业务场景中潜在的算法偏见、杀熟及对特定群体的隐性损害。

8.3.1.2 等级描述：

8.3.1.2.1 等级 1：非正式执行

该等级的人工智能安全能力描述如下：

- a) 制度流程：组织已使用与拟上线版本相关的业务样本开展至少一次场景验证，并形成验证结果；
- b) 稳健性：验证已覆盖正常业务输入和至少一类异常、越界或者高风险输入。

8.3.1.2.2 等级 2：计划跟踪

该等级的人工智能安全能力描述如下：

- a) 人员组织：在涉及高风险业务的测试计划中，初步列入业务公平性核查项；
- b) 公平性：在信用审核等场景中人工构建包含不同群体特征的基础用例，比对输出结果排查明显歧视；
- c) 权益保障：执行常识性的业务内容抽测，初步验证特定场景下是否输出违背公序良俗的业务建议。

8.3.1.2.3 等级 3：充分定义

该等级的人工智能安全能力描述如下：

- a) 制度流程：建立场景化测试规程，明确商业营销、社会福利等不同风险场景下的必测项与一票否决红线；
- b) 公平性：
 - 1) 引入反事实测试，验证防歧视能力符合相关法规要求；
 - 2) 在商业交易场景中执行价格并发测试，核验系统是否存在不合理的差别待遇。
- c) 权益保障：
 - 1) 针对劳动关系领域执行算法宽容度测试，验证业务派单未触碰劳动者生理与安全极限；

- 2) 构建适老化与未成年人专项测试库, 验证特殊模式下业务逻辑的正确降级与不良信息的拦截。

8.3.1.2.4 等级 4: 量化控制

该等级的人工智能安全能力描述如下:

- a) 制度流程: 建立场景测试数据看板, 将测试结果转化为业务线可度量的准出标准。
- b) 公平性: 引入大规模量化的群体公平性度量工具, 突破人工个例核验的局限。

8.3.1.2.5 等级 5: 持续优化

该等级的人工智能安全能力描述如下:

新兴风险防控: 利用大模型构建“AI 用户模拟器”, 自动生成海量交叉场景 (如“社交+支付”并发请求), 实现场景测试用例库的自动扩容。

8.3.2 PA34 首次发布准入管理

8.3.2.1 PA 描述:

建立人工智能应用首次上线安全准入审核机制, 对安全评估结论、测试缺陷整改情况、残余风险接受、发布审批、灰度发布方案、应急回滚预案、用户告知材料等全要素开展合规性审核, 确保产品上线前满足安全合规要求。

8.3.2.2 等级描述:

8.3.2.2.1 等级 1: 非正式执行

该等级的人工智能安全能力描述如下:

- a) 制度流程:
 - 1) 组织已在至少一个人工智能产品或者服务首次上线前开展安全检查, 并记录检查结果和上线决定;
 - 2) 组织已识别已知缺陷、残余风险和一项可用的回退或者降级方案。

8.3.2.2.2 等级 2: 计划跟踪

该等级的人工智能安全能力描述如下:

- a) 人员组织: 指定专人负责上线前基础安全核查, 承担缺陷整改核对、上线材料初步整理职责;
- b) 制度流程:
 - 1) 建立基础上线核查机制, 人工核验显性安全缺陷整改情况;
 - 2) 初步明确上线发布审批权限, 经负责人确认后启动上线;
 - 3) 制定简易回滚方案, 明确核心故障场景下的版本回退路径。

8.3.2.2.3 等级 3: 充分定义

该等级的人工智能安全能力描述如下:

- a) 人员组织:
 - 1) 组建由安全、研发、测试、业务、合规人员构成的上线准入评审小组, 统筹上线安全审核工作;
 - 2) 相关人员具备安全风险研判、缺陷定级、上线合规性核查专业能力。
- b) 制度流程:

- 1) 制定首次上线安全准入管理规范，明确各审核要素的判定标准与核查要求；
 - 2) 上线前完成全面安全评估，形成正式安全评估报告；
 - 3) 建立分级发布审批流程，明确各层级审批权限与责任边界；
 - 4) 准备合规用户告知材料，明确服务边界、安全风险与用户权利义务。
- c) 技术保障：
- 1) 固化上线安全准入核查清单，实现审核要素标准化校验；
 - 2) 支持灰度发布流量管控，可按用户范围、访问比例逐步放量验证。

8.3.2.2.4 等级 4：量化控制

该等级的人工智能安全能力描述如下：

- a) 制度流程：建立首次上线准入量化评估基准，明确高危缺陷清零率、安全评估覆盖率、回滚预案演练通过率等核心指标，指标未达到阈值时，暂停上线并实施缺陷处置或者风险复核；
- b) 技术保障：
 - 1) 实现上线安全门禁自动化管控，量化指标不达标自动阻断发布流程；
 - 2) 灰度发布阶段自动采集安全、稳定性核心指标，量化验证通过后方可全量上线；
 - 3) 上线审核全流程数字化归档，支持按审核节点、责任人快速溯源审计。

8.3.2.2.5 等级 5：持续优化

该等级的人工智能安全能力描述如下：

技术保障：构建智能化上线风险预判体系，基于历史上线数据、同类产品风险特征自动识别上线潜在风险，前置推送安全加固建议。

8.3.3 PA35 重大更新审批管控

8.3.3.1 PA 描述：

针对人工智能应用的重大版本迭代、核心功能调整、模型升级、架构变更等重大更新场景，建立专项安全准入管控机制，覆盖更新风险影响评估、安全回归测试、缺陷整改验证、变更审批、灰度验证、应急回滚等关键环节，确保更新后服务持续满足安全合规要求。

8.3.3.2 等级描述：

8.3.3.2.1 等级 1：非正式执行

该等级的人工智能安全能力描述如下：

- a) 制度流程：
 - 1) 组织已对至少一次模型、知识库、插件、策略或者运行配置的重大更新进行记录；
 - 2) 组织已在该次更新前开展安全影响检查或者回归验证，并形成发布决定。

8.3.3.2.2 等级 2：计划跟踪

该等级的人工智能安全能力描述如下：

- a) 人员组织：指定专人负责重大更新前基础安全核查，承接更新缺陷整改核对工作；
- b) 制度流程：
 - 1) 建立重大更新基础核查机制，人工核验更新涉及的显性安全问题；
 - 2) 明确更新发布基础审批流程，经相关负责人确认后发布上线；
 - 3) 制定简易更新回滚方案，明确核心故障场景下的版本回退路径。

8.3.3.2.3 等级 3：充分定义

该等级的人工智能安全能力描述如下：

- a) 人员组织：
 - 1) 组建变更安全评审小组，由安全、研发、测试、业务人员共同开展重大更新安全审核；
 - 2) 相关人员具备变更影响评估、安全回归测试、风险研判专业能力。
- b) 制度流程：
 - 1) 制定重大更新安全准入管理规范，明确不同类型更新的安全审核标准与流程；
 - 2) 更新前开展专项安全影响评估，识别变更引入的新增安全风险；
 - 3) 针对更新内容开展全量安全回归测试，安全缺陷按风险等级闭环整改；
 - 4) 制定更新专项灰度方案，针对变更功能开展定向灰度验证；
 - 5) 编制更新专属应急回滚预案，明确回滚触发条件、操作步骤与效果验证标准。
- c) 技术保障：
 - 1) 固化重大更新安全核查清单，针对不同变更类型匹配差异化校验项；
 - 2) 支持定向灰度流量管控，可针对更新功能、目标用户群体精准放量验证。

8.3.3.2.4 等级 4：量化控制

该等级的人工智能安全能力描述如下：

- a) 制度流程：建立重大更新安全准入量化评估基准，明确更新缺陷闭环率、回归测试覆盖率、灰度风险指标通过率等核心要求，量化指标未达到阈值时，暂停全量发布并实施复核或者纠偏；
- b) 技术保障：
 - 1) 实现重大更新发布门禁自动化管控，量化校验不通过自动阻断发布流程；
 - 2) 灰度阶段自动采集更新模块的安全、稳定性数据，量化验证通过后方可逐步扩量；
 - 3) 重大更新全流程数字化归档，支持变更内容、审核节点、风险处置全链路溯源审计。

8.3.3.2.5 等级 5：持续优化

该等级的人工智能安全能力描述如下：

技术保障：构建智能化变更风险管控体系，基于历史更新数据自动识别变更风险点，前置推送安全验证与加固建议。

8.4 应用运营

8.4.1 PA36 动态内容治理

8.4.1.1 PA 描述：

确保人工智能产品和服务在日常运行期间，持续输出合法、真实、准确的内容，并严格履行人工智能生成合成内容标识义务。在运营端防范违法不良信息的生成与传播，压降“AI幻觉”带来的误导风险，同时以规范的显隐性内容标识消除公众的认知混淆，筑牢信息内容安全与科技伦理防线。

8.4.1.2 等级描述：

8.4.1.2.1 等级 1：非正式执行

该等级的人工智能安全能力描述如下：

- a) 技术保障：组织已对人工智能生成、编辑或者传播的内容实施至少一项风险识别、过滤或者人工复核措施；

- b) 权益保障：组织能够记录一项风险内容的识别、处置或者用户反馈结果。

8.4.1.2.2 等级 2：计划跟踪

该等级的人工智能安全能力描述如下：

- a) 人员组织：在运营团队中指定人员负责接收内容违规的外部举报，并进行事后的人工删除或封禁操作；
- b) 制度流程：
 - 1) 接入了基础的静态敏感词库，对用户的输入和模型的输出进行简单的硬匹配拦截；
 - 2) 计划并在部分高风险功能（如图像生成）中添加了简单的“AI 生成”字样。

8.4.1.2.3 等级 3：充分定义

该等级的人工智能安全能力描述如下：

- a) 人员组织：设置监看人员，监看人员数量与服务规模相匹配，负责及时跟踪国家政策、收集分析第三方投诉情况，并根据监看情况提高生成内容质量及安全，并符合 GB/T 45654—2025 中 6.5 列项 b) 的相关要求；
- b) 制度流程：
 - 1) 采取关键词、分类模型等技术方式对使用者输入信息进行持续安全检测，并符合 GB/T 45654—2025 中 6.5 列项 a) 的规定；
 - 2) 设置并向使用者公示阶梯式的惩戒规则：在使用者连续多次输入违法不良信息或一天内累计输入达到一定次数时，采取暂停提供服务等处置措施；
 - 3) 采取技术措施提高生成内容响应使用者输入意图的能力，提高生成内容中数据及表述与科学常识及主流认知的符合程度，有效减少错误与幻觉内容，并符合 GB/T 45654—2025 中 5.2 列项 b) 的要求；
 - 4) 对文本、图片、音频、视频等各模态生成内容添加清晰可辨的显式标识（如在图片边角添加文字、在音频中添加“短长短短”的摩斯码节奏等），并符合 GB 45438—2025 中第 5 章的具体规范要求；
 - 5) 在人工智能生成合成内容文件数据中添加隐式标识，在文件元数据中完整写入生成合成标签（Label）、服务提供者、内容制作编号等关键要素，并符合 GB 45438—2025 中 6.1 及附录 E 的 JSON 结构化格式要求。

8.4.1.2.4 等级 4：量化控制

该等级的人工智能安全能力描述如下：

制度流程：

- a) 建立运行期内容安全的量化监控大盘，通过自动化测评系统高频抽检线上输出数据，确保模型生成内容安全合格率严格不低于 90%，符合 GB/T 45654—2025 中 5.2 列项 a) 的量化基线要求；
- b) 量化追踪事实错误率，针对知识发现与资讯场景，持续测量有效内容含量，压降认知误导风险；
- c) 在日常运行中，通过机器视觉等自动化手段实时核验线上图片、视频生成物，强制确保显式标识的文字高度不低于画面最短边长度的 5%，且在视频正常播放速度下持续时间不少于 2 秒，以量化断言保障内容标识的绝对合规，严格符合 GB 45438—2025 中 5.2 列项 e) 及 5.4 的相关数值要求。

8.4.1.2.5 等级 5：持续优化

该等级的人工智能安全能力描述如下：

- a) 制度流程：建立基于监管动态与高危威胁情报的敏感词库及分类模型实时热更新机制，实现对新型违法违规暗语的小时级精准识别与拦截；
- b) 新兴风险防控：
 - 1) 运用“人工智能审核人工智能”的智能监看技术替代部分人工审核压力，在海量并发的日常运营中，自主识别具有高度伪装性的不良内容诱导；
 - 2) 将运行阶段暴露出的信息茧房倾向、隐蔽歧视或认知退化等深层次内容风险，常态化反哺至预训练语料清洗及人类反馈对齐环节，进行系统性修复。

8.4.2 PA37 安全运营与应急干预

8.4.2.1 PA 描述

针对人工智能产品运行阶段，建立集态势监测、智能体行为审计、应急响应与申诉干预于一体的安全运营体系。涵盖日常模型运行的安全状态监控、智能体（Agent）工具调用的防滥用管控、严重安全事件的应急阻断，以及针对算法误判与违规争议的用户申诉救济机制，实现端到端的安全运营闭环。

8.4.2.2 等级描述

8.4.2.2.1 等级 1：非正式执行

该等级的人工智能安全能力描述如下：

- a) 稳定性：组织已对实际运行的模型、内容风险、接口调用、智能体操作或者资源状态中的至少一项进行监测；
- b) 制度流程：组织已对至少一项人工智能运行异常、用户申诉或者安全事件进行记录和处置。

8.4.2.2.2 等级 2：计划跟踪

该等级的人工智能安全能力描述如下：

- a) 人员组织：指定基础运维与客服人员分别负责系统状态监控与用户投诉收集；
- b) 制度流程：
 - 1) 制定基础的应急联络清单与事件处置流程；
 - 2) 初步尝试在产品中建立反馈入口，收集用户对人工智能输出内容的负面反馈。
- c) 技术保障：建立基础的数据与模型备份策略，记录基础的接口与工具调用日志，但不具备针对大模型攻击或智能体越权的专项动态监测手段。

8.4.2.2.3 等级 3：充分定义

该等级的人工智能安全能力描述如下：

- a) 人员组织：组建专业的人工智能安全应急响应团队，成员具备内容合规处置、底层漏洞修复与安全舆情应对能力；
- b) 制度流程：
 - 1) 建立分类分级的人工智能安全事件应急预案（涵盖服务中断、脱轨生成、数据泄露等核心场景），定期开展应急实战演练；
 - 2) 建立标准化的争议申诉机制，对自动化决策的异议提供明确的逻辑解释，在面临对个人权益有重大影响的业务场景（如授信、招聘）中，保障人工介入复核的控制权。
- c) 技术保障：

- 1) 部署智能体行为监控机制，持续审计工具调用链路。对涉及高危操作（如数据库写入、外部邮件发送、资金调用）强制引入“人机协同（Human-in-the-loop）”的人工二次确认机制，防范智能体工具滥用与越权执行；
- 2) 持续识别并防范针对线上接口的恶意指令注入、越权访问与数据窃取；
- 3) 具备紧急熔断或业务下线的技术能力，在发生严重安全事故时能够快速阻断相关服务；
- 4) 针对端侧部署的模型，建立基于联网的安全策略动态下发机制，对长时间未更新的高危版本实施预警或降级。

8.4.2.2.4 等级 4：量化控制

该等级的人工智能安全能力描述如下：

- a) 制度流程：
 - 1) 建立量化考核体系，追踪安全事件的平均发现时间、平均恢复时间及人工复核的按期达成率；
 - 2) 建立针对模型行为漏洞的专项外部漏洞众测机制，公开奖励标准与修复时限；
 - 3) 定期开展算法影响评估，针对高风险场景量化分析客诉误判率，推动算法基准动态调优。
- b) 技术保障：
 - 1) 实现应急响应与处置流程的自动化流转，从威胁告警触发、工单分配到拦截策略下发形成无缝闭环；
 - 2) 针对智能体异常工具调用链，系统能够基于操作基线自动判定越权风险，并自动触发降级运行或隔离恶意智能体。

8.4.2.2.5 等级 5：持续优化

该等级的人工智能安全能力描述如下：

- a) 制度流程：
 - 1) 建立重大安全与争议事件的复盘溯源机制，定期输出改进报告，处理复杂争议时主动向相关监管机构同步整改举措；
 - 2) 参与行业级的漏洞与威胁情报共享，沉淀应对新型攻击的防护经验，推动跨组织联防联控。
- b) 新兴风险防控：引入智能化的威胁感知探针，从海量运行日志与复杂交互数据中，前瞻性地发现未知隐蔽攻击、长尾算法偏见及智能体信任链污染风险，实现防御策略的自适应升级。

8.5 应用退役

8.5.1 PA38 下线管理

8.5.1.1 PA 描述：

建立人工智能应用暂停、降级、变更及下线全场景管理机制，明确各场景触发条件、分级审批流程、专项风险评估、用户告知规则、服务替代方案、功能与接口关停步骤、异常处置要求，保障服务调整与下线过程合规、有序，防范用户权益与安全风险。

8.5.1.2 等级描述：

8.5.1.2.1 等级 1：非正式执行

该等级的人工智能安全能力描述如下：

- a) 制度流程：组织已对至少一项停止、降级或者下线的人工智能产品或者服务实施处置，并保留下线记录；
- b) 权益保障：组织已识别下线对用户或者业务的主要影响，并实施告知、迁移、替代或者其他处置措施。

8.5.1.2.2 等级 2：计划跟踪

该等级的人工智能安全能力描述如下：

- a) 人员组织：指定专人统筹服务下线基础工作，承担事项对接、信息同步与基础处置职责；
- b) 制度流程：
 - 1) 明确服务下线基本触发条件，履行层级审批程序；
 - 2) 下线前通过官方渠道发布公告，告知用户服务停止时限；
 - 3) 制定简易下线操作方案，明确核心功能、接口关停基础安排；
 - 4) 建立基础异常响应渠道，承接用户咨询与突发问题处置。

8.5.1.2.3 等级 3：充分定义

该等级的人工智能安全能力描述如下：

- a) 人员组织：
 - 1) 组建由业务、安全、运维、合规人员构成的下线专项工作组，统筹全流程管控工作；
 - 2) 相关人员具备下线风险研判、用户权益保障、应急事件处置专业能力。
- b) 制度流程：
 - 1) 制定暂停、降级、变更、下线全场景管理规范，明确各场景触发条件、分级审批权限与操作流程；
 - 2) 下线前开展专项风险评估，识别数据安全、用户权益、业务衔接等风险并制定对应防控措施；
 - 3) 规范用户告知机制，明确告知渠道、提前时限与告知内容，充分保障用户知情权；
 - 4) 制定服务替代与迁移方案，为用户提供业务衔接指引或替代服务路径；
 - 5) 明确功能、接口分步关停方案，规定关停顺序、验证标准与过渡期安排；
 - 6) 建立下线异常处置机制，明确突发故障、用户投诉、合规风险等场景的处置流程；
 - 7) 全流程留存审批文件、风险评估报告、用户告知凭证等档案资料。
- c) 技术保障：
 - 1) 固化下线操作核查清单，确保各环节按规范落地；
 - 2) 支持分阶段关停服务与接口，可控推进下线进程。

8.5.1.2.4 等级 4：量化控制

该等级的人工智能安全能力描述如下：

- a) 制度流程：建立下线管理量化评估基准，明确用户告知覆盖率、风险问题闭环率、功能接口关停准确率等核心指标，指标未达到阈值时，暂停最终下线并完成补充处置；
- b) 技术保障：
 - 1) 实现下线流程节点自动化管控，关键节点未完成自动阻断后续操作；
 - 2) 自动监控下线过程运行状态，触发异常风险时自动告警并启动应急处置；
 - 3) 全流程数字化留痕，支持各节点操作、审批、处置记录快速溯源审计。

8.5.1.2.5 等级 5：持续优化

该等级的人工智能安全能力描述如下：

- a) 技术保障：构建智能化下线风险管控体系，基于历史下线数据自动识别潜在风险点，前置推送处置预案与优化建议；
- b) 新兴风险防控：
 - 1) 服务调整、审批、关停、处置全流程可视化可追溯，支持监管端实时审计核查；
 - 2) 结合监管要求更新与行业实践，持续迭代下线管理机制，动态适配新型服务形态的下线管控需求。

8.5.2 PA39 资产处置

8.5.2.1 PA 描述：

针对人工智能应用下线、变更过程中涉及的用户数据、交互日志、生成内容、提示词配置、知识库、模型组件、插件工具、接口凭证、运行环境配置等资产，实施分类分级处置管理，明确各类资产的留存、迁移、删除、销毁、归档及审计要求，确保处置过程合规可控、全程可追溯，防范数据泄露与资产流失风险。

8.5.2.2 等级描述：

8.5.2.2.1 等级 1：非正式执行

该等级的人工智能安全能力描述如下：

- a) 制度流程：组织已识别退役产品或者服务关联的数据、模型、知识库、账号凭证、插件和外部依赖；
- b) 数据保护：组织已对至少一类相关资产实施归档、删除、销毁、停止访问或者返还，并形成处置结果。

8.5.2.2.2 等级 2：计划跟踪

该等级的人工智能安全能力描述如下：

- a) 人员组织：指定专人负责下线资产基础盘点与处置执行，承担核心资产统计、基础处置落地职责；
- b) 制度流程：
 - 1) 建立基础资产清单，梳理核心数据与系统资产类目；
 - 2) 明确用户核心数据、接口凭证等敏感资产的基础处置要求；
 - 3) 对需留存、删除的资产进行简易分类，执行基础留存或删除操作；
 - 4) 留存简易处置记录，满足基础核查需求。

8.5.2.2.3 等级 3：充分定义

该等级的人工智能安全能力描述如下：

- a) 人员组织：
 - 1) 组建由安全、运维、数据、合规人员构成的资产处置专项小组，统筹各类资产处置工作；
 - 2) 相关人员具备数据销毁、资产归档、合规判定专业能力。
- b) 制度流程：
 - 1) 制定资产分类处置管理规范，按资产属性、敏感等级明确留存、迁移、删除、销毁、归档的处置标准与操作要求；

- 2) 用户数据、交互日志、生成内容等用户相关资产，严格依照法律法规与监管要求明确留存期限与删除销毁规则；
 - 3) 提示词配置、知识库、模型组件、插件工具等业务资产，明确迁移、归档、清理规则与责任主体；
 - 4) 接口凭证、运行环境配置等安全资产，明确注销、销毁、脱敏归档要求；
 - 5) 建立资产处置分级审批流程，高敏感资产处置履行升级审批程序；
 - 6) 规范处置全流程审计机制，完整留存资产清单、处置记录、审批文件等档案；
 - 7) 敏感数据销毁采用可靠技术手段，确保数据不可恢复。
- c) 技术保障：
- 1) 支持敏感资产脱敏、安全删除操作；
 - 2) 关键资产处置操作全程留痕，生成不可篡改的处置日志。

8.5.2.2.4 等级 4：量化控制

该等级的人工智能安全能力描述如下：

- a) 制度流程：建立资产处置量化评估基准，明确资产处置合规率、敏感数据销毁完成率、归档资料完整率等核心指标，指标达标方可确认处置闭环；
- b) 技术保障：
 - 1) 实现资产自动化盘点与分类标记，精准匹配对应处置策略；
 - 2) 支持批量安全删除与不可逆销毁，自动校验销毁效果；
 - 3) 资产处置全流程数字化管控，所有操作、审批、核验记录自动归档，支持全链路审计溯源。

8.5.2.2.5 等级 5：持续优化

该等级的人工智能安全能力描述如下：

- a) 技术保障：构建智能化资产处置管控体系，自动识别全量资产并匹配最优处置策略，结合监管要求动态调整处置规则；
- b) 新兴风险防控：
 - 1) 资产盘点、处置、销毁、归档全流程可视化监控，支持监管端实时审计与核验；
 - 2) 运用预测性合规分析技术，预判资产处置中的合规风险，前置优化处置方案，持续改进处置合规性与安全性。

9 评估方法

9.1 评估方式

人工智能安全能力成熟度评估可采用组织自评或者第三方评估。两种评估方式采用相同的评估范围、评价项符合性判定和等级判定方法。组织自评可用于识别能力现状、确定改进目标和跟踪改进效果；第三方评估可用于形成独立评价结论。

评估以组织为受评主体，以适用PA为基本评价单元，以第6章至第8章各PA成熟度等级下可独立判定的评价项为最小符合性判定单元。评估结论仅适用于已确定的组织边界、人工智能相关活动、安全对象、模型或者产品服务版本和时间范围。

9.2 评估范围

评估范围按照第5.5节，根据受评组织的主体类型、实际承担的人工智能相关活动 and 责任边界确定，包括以下内容：

- a) 组织名称、主体类型、纳入评价的组织单元及其职责边界；
- b) 纳入评价的业务、场景、系统、数据、模型、产品和服务等；
- c) 适用的安全对象和安全活动；
- d) 相关模型、产品和服务的名称及版本；
- e) 适用 PA、条件必测项的触发情况以及未纳入评价的 PA 及其理由；
- f) 评价基准日期和证据覆盖的时间范围。

评估范围内由外部组织实施的活动，评价受评组织对该活动的选择、约定、验收、监督和风险处置情况。

条件必测项的触发条件成立时，相关PA纳入评估范围；触发条件不成立时，根据实际活动确定是否纳入评估范围。选测项纳入评估范围后，采用与必测项相同的证据评价和等级判定方法。未纳入评估范围的PA不参与组织成熟度等级判定。

评估范围能够反映受评组织实际开展的人工智能相关活动。评估范围的调整及其理由在评估记录和评估报告中说明。

9.3 证据采集与评价

9.3.1 证据采集

证据可通过人员访谈、文档查阅、配置核验、现场观察、系统或者模型演示、工具检测和抽样测试等方式采集。证据可包括职责文件、制度流程、数据和模型清单、配置、日志、审批记录、测试报告、监测数据、事件记录及改进记录。

采用抽样方式时，样本宜覆盖主要组织单元、业务类型、数据批次、模型版本、产品服务版本、高风险场景以及发生过异常或者变更的活动。抽样范围、方法、数量和限制在评估记录中说明。

9.3.2 证据评价

证据评价考虑相关性、真实性、有效性、充分性和可追溯性。单一访谈陈述不作为评价项符合的唯一证据；不同来源证据相互矛盾时，通过补充访谈、扩大抽样、现场核验或者技术测试进行交叉验证。

复用外部检查、检测、测评或者认证结果时，核验其评价对象、评估范围、模型或者产品服务版本、依据文件、实施时间、评价方法和证据有效性。超出有效范围的结果不用于替代相应评价项的证据。

9.4 评价项符合性判定

第6章至第8章各PA成熟度等级下的适用评价项逐项判定，判定结果包括：

- a) 符合：充分、适当的证据表明评价项在评估范围内已经实施并达到条款表述的结果；
- b) 不符合：评价项未实施、未完整实施、未达到条款表述的结果，或者缺少充分、适当的证据。

部分实施的评价项在等级判定中按不符合处理。每项评价项的判定记录对应的客观证据、抽样范围和判定理由。

9.5 PA 成熟度等级判定

PA成熟度等级按照以下方法判定：

- a) PA 第一级的全部适用评价项均判定为符合时，该 PA 达到第一级；
- b) PA 第二级至第五级中的某一级及其以下各级全部适用评价项均判定为符合时，该 PA 达到该等级；
- c) 某一级存在适用评价项不符合时，该 PA 未达到该等级；
- d) PA 成熟度等级取从第一级起连续满足上述条件的最高等级。

未达到第一级全部适用评价项的PA记为未达级。其他PA或者其他安全能力的符合结果不用于抵消该PA的未符合项。

示例：某 PA 的第一级至第三级全部适用评价项均符合，第四级存在一项不符合，则该 PA 的成熟度等级为第三级。

9.6 组织成熟度等级判定

9.6.1 PA 能力剖面

PA能力剖面由评估范围内各适用PA的成熟度等级、评价项符合情况和主要差距构成。数据安全、模型安全 and 应用安全分别形成分项能力剖面；仅评价其中一个或者两个安全对象时，未纳入评价的安全对象不参与等级判定。

9.6.2 等级判定

组织成熟度等级在PA能力剖面的基础上确定。组织达到目标等级N时，同时满足以下条件：

- a) 评估范围内全部适用 PA 均达到目标等级 N；
- b) 评估范围内存在未达到等级一的适用 PA 时，组织判定为未达级。
- c) 不存在未经接受的重大安全风险或严重未符合项。

组织成熟度等级用于呈现评估范围内安全过程的整体制度化程度，PA能力剖面用于呈现具体短板。等级判定不以PA平均分替代适用PA逐项判定和连续等级条件。

示例：评估范围内各适用 PA 均达到第三级，其中部分 PA 达到第四级或者第五级，则组织成熟度等级为第三级；存在一个适用 PA 仅达到第二级时，组织成熟度等级为第二级。

9.7 结果复核与报告

9.7.1 结果复核

评估组对评估范围、证据、评价项判定、PA成熟度等级和组织成熟度等级进行复核。受评组织对事实信息存在异议时，可补充证据或者说明；评估组记录异议及处理结果。

9.7.2 评估报告

评估报告宜包括以下内容：

- a) 评估目的、评估方式、评估依据和评价基准日期；
- b) 受评组织、主体类型、组织边界和评估范围；

- c) 适用 PA、条件必测项的触发情况以及未纳入评价的 PA 及其理由；
- d) 证据采集方式、抽样情况和证据限制；
- e) 评价项判定结果和各 PA 成熟度等级；
- f) 组织成熟度等级及其判定依据；
- g) PA 能力剖面、未符合项和改进建议；
- h) 结果复核、异议处理和结论使用限制。

9.7.3 等级证明

出具等级证明时，证明文件至少标明组织名称、主体类型、组织边界、评估范围、组织成熟度等级、评价依据、评价基准日期和评价机构。评估范围范围宜写明纳入评价的安全对象，以及主要业务、模型、产品或者服务；各适用PA及其成熟度等级在评估报告中列明。

组织边界、核心业务、主要模型、产品服务版本或者责任范围发生重大变化时，原评价结论的适用性重新确认。

参 考 文 献

- [1] GB/T 22239—2019 信息安全技术 网络安全等级保护基本要求
 - [2] GB/T 25069-2022 信息安全技术 术语
 - [3] GB/T 31168—2023 信息安全技术 云计算服务安全能力要求
 - [4] GB/T 37988—2019 信息安全技术 数据安全能力成熟度模型
 - [5] GB/T 29246—2023 信息安全技术 信息安全管理体系 概述和词汇
 - [6] GB/T 39276—2020 信息安全技术 网络产品和服务安全通用要求
 - [7] GB/T 41867—2022 信息技术 人工智能 术语
 - [8] GB/T 43848—2024 网络安全技术 软件产品开源代码安全评价方法
 - [9] GB/T 44272—2024 信息技术 开源 开源许可证框架
 - [10]GB/T 44588—2024 数据安全技术 互联网平台及产品服务 个人信息处理规则
 - [11]GB/T 44721—2024 智能网联汽车 自动驾驶系统通用技术要求
 - [12]GB/T 45401.2—2025 人工智能 计算设备调度与协同 第2部分：分布式计算框架
 - [13]GB/T 45288.1—2025 人工智能 大模型 第1部分：通用要求
 - [14]GB/T 45288.2—2025 人工智能 大模型 第2部分：评测指标与方法
 - [15]GB/T 45288.3—2025 人工智能 大模型 第3部分：服务能力成熟度评估
 - [16]GB/T 45392—2025 数据安全技术 基于个人信息的自动化决策安全要求
 - [17]GB 45438—2025 网络安全技术 人工智能生成合成内容标识方法
 - [18]GB/T 45652—2025 网络安全技术 生成式人工智能预训练和优化训练数据安全规范
 - [19]GB/T 45654—2025 网络安全技术 生成式人工智能服务安全基本要求
 - [20]GB/T 45674—2025 网络安全技术 生成式人工智能数据标注安全规范
 - [21]ISO/IEC 23894—2023 信息技术 人工智能 风险管理指南 (Information technology - Artificial intelligence - Guidance on risk management)
 - [22]ISO/IEC 42001:2023 信息技术 人工智能 管理体系 (Information technology—Artificial intelligence—Management system)
 - [23]TC260-005 《人工智能应用伦理安全指引 1.0》
-