大模型安全评估与等级保护合规的实践探讨

刘礼
深圳竞远信息技术有限公司，广东 深圳 518000
DOI: 10.61369/TACS.2025130046

摘　　要： 随着生成式人工智能技术的快速发展，大模型已广泛应用于政务、金融、医疗等关键领域。当前，大模型应用主要遵循网信部门的算法备案与安全评估要求，而承载大模型运行的信息系统仍需遵循网络安全等级保护制度。2025年10月修订的《网络安全法》首次在法律层面增加了人工智能条款，为大模型安全纳入网络安全法治框架提供了法律依据。本文从网络安全从业者视角，探讨大模型安全评估与等级保护制度协同的可能性，分析大模型典型风险与等保控制域的对应关系，提出实践思路，并对当前难点提出应对思考。

关 键 词： 等级保护；生成式人工智能；大模型安全；安全评估；实践探讨

Practical Exploration of Large Model Security Assessment and Compliance with Classified Protection Regulations

Liu Li
Shenzhen Jingyuan Information Technology Co., Ltd., Shenzhen, Guangdong 518000

Abstract: With the rapid advancement of generative artificial intelligence (GenAI) technology, large models have been widely deployed across critical sectors including government administration, finance, and healthcare. Currently, the application of large models is primarily governed by the algorithm filing and security assessment requirements issued by cyberspace administration authorities, whereas the information systems supporting large model operations must still comply with the cyber security classified protection regime. The revised Cybersecurity Law, which came into effect in October 2025, incorporated dedicated clauses on artificial intelligence for the first time at the legislative level. This landmark revision provides a legal underpinning for integrating large model security governance into the overall cyber security legal framework. From the perspective of cybersecurity practitioners, this paper explores the feasibility of achieving synergy between large model security assessment and the classified protection system, analyzes the corresponding relationships between typical large model risks and classified protection control domains, proposes actionable practical approaches, and offers targeted countermeasures to address the current implementation challenges.

Keywords: classified protection; generative artificial intelligence; large model security; security assessment; practical exploration

引言

近年来，生成式人工智能技术取得突破性进展，大模型已在政务、金融、医疗、教育等多个关键领域实现深度应用。2025年10月修订的《网络安全法》首次在法律层面增加了人工智能专条（第二十条），明确规定"国家支持人工智能基础理论研究和算法等关键技术研发，推进训练数据资源、算力等基础设施建设，完善人工智能伦理规范，加强风险监测评估和安全监管"。这为探讨大模型安全与网络安全制度的衔接提供了法律层面的依据。

值得关注的是，等级保护制度呈现出开放、扩展的特征。云计算、移动互联、物联网、工业控制系统、大数据等新技术新应用，都已被逐步纳入等级保护2.0的扩展要求中。公安部网络安全等级保护评估中心于2024年发布的T/ISEAA 005/006-2024《大模型系统安全保护/测评要求》团体标准，可以看作是探索大模型安全与等保框架衔接的有益尝试。

本文从网络安全从业者的实践视角，尝试探讨以下问题：生成式大模型安全评估与等保合规之间是否存在协同空间？大模型特有的风险是否可以映射到现有等保控制域？如何形成可供参考的落地路径？

一、制度与标准框架

（一）等保2.0的基础地位

等保2.0建立了”定级备案—安全建设整改—等级测评—监督检查”的闭环管理机制。对于涉及公民个人信息、重要数据的大模型系统，等保2.0提出了明确要求：涉及大量个人信息或重要数据的系统原则上不低于三级；审计日志留存时间不少于6个月；对数据进行分类分级，敏感数据加密存储和传输；实施身份鉴别、权限分离、最小权限原则；建立安全事件应急预案和处置机制。

（二）生成式 AI 专项法规与备案要求

2023年7月，国家网信办等七部门联合发布《生成式人工智能服务管理暂行办法》，自2023年8月15日起施行。随后，网信办等部门又陆续出台配套规定，如《人工智能生成合成内容标识办法》（2025年9月施行），进一步完善了生成式人工智能的监管框架。《暂行办法》对服务提供者提出了全面要求：训练数据应当符合法律法规要求，不得含有侵权内容；采取措施防止生成违法违规内容；在注册界面等方式公示算法基本原理和服务目的；向网信部门提交安全评估报告，通过备案后方可提供服务。

根据各地网信部门的备案实践，生成式 AI 安全评估报告通常包含以下模块：评估对象与业务场景描述、数据隐私保护措施说明、算法与模型安全保障、内容安全与风险防控机制、安全事件应急响应能力、风险汇总与整改建议、评估结论与合规性声明。这些模块与等保测评的控制域存在高度重叠，为两者的融合评估提供了基础。

（三）大模型安全国家标准体系

2025年2月发布的《人工智能大模型》系列国家标准（GB/T 45288-2025）是我国首部聚焦通用大模型的国家标准，包括通用要求、评测指标与方法、服务能力成熟度评估三个部分。

综合上述框架，可以形成以下”制度图谱”：等保构成了网络安全的基础框架，而大模型安全专门标准则针对模型特有的风险提出了细化的安全要求。两者在数据安全、日志审计、访问控制等方面存在重叠，这可能为未来两种评估方式的协同或融合提供空间。

二、大模型风险与等保控制域映射

（一）大模型核心风险类型

大模型系统的核心风险具有显著差异：训练数据安全方面，涉及数据来源合法性、个人信息处理、敏感信息残留等问题；生成内容安全方面，大模型可能生成虚假信息、违法违规内容、歧视性输出或错误建议；模型本体安全方面，提示注入攻击、越狱攻击、模型窃取等是传统等保未覆盖的新型安全威胁；可解释性与可控性方面，大模型的”黑箱”特性使其输出难以预测和追溯。

（二）风险与等保控制域映射

下表展示大模型典型风险与等保控制域的对应关系：

等保控制域	大模型典型风险	评估要点	证据示例
物理与环境安全	模型训练环境物理隔离失效	训练与推理环境是否物理隔离、机房安全认证	机房安全证书、环境拓扑图
网络与通信安全	模型服务暴露在互联网被攻击	API 边界防护、访问控制策略、加密传输	防火墙配置、API 网关规则
设备与计算安全	承载模型的服务器被入侵	服务器 / 容器 / K8s 安全配置、漏洞管理	漏洞扫描报告、加固记录
应用安全	提示注入、越狱攻击	身份鉴别、参数校验、输入输出过滤	安全测试报告、拦截日志
数据安全	训练数据不合规、个人信息泄露	数据分类分级、加密存储、脱敏处理	数据清单、脱敏记录、授权协议
备份与恢复	模型参数 / 数据丢失	模型版本备份、数据备份策略	备份策略文档、恢复演练记录
安全管理制度	责任不清、流程缺失	AI 安全管理制度、人员管理、变更管理	安全制度汇编、培训记录
应急响应	安全事件处置不当	AI 安全专项应急预案、演练记录	预案文档、演练报告

上述映射表揭示了以下观点：并非另起炉灶，而是风险投影——大模型特有的风险或可投影到现有的等保控制域，通过扩展控制域的评估要点来覆盖新型风险。T/ISEAA 005/006-2024、GB/T 45654-2025等标准针对每个控制域提出了大模型特有的评估指标，为原有框架的补充提供了参考。等保测评中已有的证据（如漏洞扫描报告、日志审计记录）或许可以复用，从而避免重复工作。

三、落地路径与实践思路

（一）范围界定与定级策略

首先需要明确评估对象的边界，常见情况包括：单一大模型API 服务（仅提供模型推理接口，不含前端应用）；带大模型的完整业务系统（包括前端门户、移动 APP、后台服务和模型服务）。评估对象的界定直接影响定级结果和评估范围。

（二）合规条款与技术指标清单化

将各项法规要求拆解为可检查的技术指标，建立 " 条款 – 指标 " 矩阵。数据安全方面：数据来源合法性对应训练数据授权比例、版权审查覆盖率；数据分类分级对应数据清单完整率、敏感标识覆盖率；敏感数据加密对应存储加密覆盖率、传输加密启用率；日志留存≥6个月对应日志存储周期、留存策略验证。

模型安全方面：内容安全防控对应有害内容拦截率、误报率；提示注入防护对应注入攻击拦截率；越狱攻击防御对应越狱尝试拦截率；输出可追溯性对应输出日志完整率、水印嵌入率。

用户权利保障方面：知情同意对应同意弹窗展示率、同意记录完整率；拒绝训练权利对应 " 不用于训练 " 选项功能可用性；数据删除响应对应删除请求处理时限、完成率。

（三）安全评估报告与等保测评衔接

生成式 AI 安全评估报告通常包含的章节与等保测评存在对应关系：评估对象与业务场景对应等保”系统描述”；数据隐私保护对应等保”数据安全”；算法与模型安全对应等保”应用安全”的扩展；内容安全与风控对应等保”应用安全”；安全事件应急对应等保”应急响应”；合规性验证汇总各控制域评估结果；风险汇总与整改对应等保”问题整改”。

（四）测试方法与证据收集

合规测试方法包括：文档审查，检查安全管理制度汇编、用户协议与隐私政策、数据处理协议与授权记录、备案证明与评估报告；过程访谈，与开发团队、安全团队、法务团队进行访谈；配置核查，检查访问控制策略、日志配置与留存策略、加密配置。

安全与对抗测试包括：提示注入攻击测试，构造各类提示注入样本测试模型防护能力；越狱攻击测试，使用角色扮演、指令组合等手法测试模型越狱防御；隐私泄露测试，通过特定提示词测试模型是否泄露训练数据中的个人信息；内容安全抽检，构建测试集测试模型对各类有害内容的识别与拦截能力。

测试过程中应留存的证据包括：测试用例与预期结果、测试过程截图、模型响应记录、日志文件。

（五）整改闭环与持续评估

根据安全评估报告，按风险等级建立整改闭环：高风险立即整改，通常1-7天，需全面复测；中风险计划整改，通常1个月内，需抽样复测；低风险持续优化，季度评估，需年度复测。建立”风险识别—整改跟踪—复测验证—版本发布”的闭环流程。

对于模型迭代更新，建立”重大变更触发安全评估”机制：参数规模变化，如模型参数量变化超过一定阈值；训练语料变更，新增大量训练数据或更换数据源；场景扩展，新增应用场景或用户群体。明确上述变更是否需要重新报备或更新评估报告。

四、典型难点与应对策略

（一）法规与技术之间的”模糊匹配”问题

法规条款常用模糊表述，如 " 应当采取适当措施 "" 充分保护 " 等，而安全评估需要可量化的技术指标。应对策略包括：指标化映射，将抽象的法律要求映射为可量化的技术指标，如数据分类分级要求100% 数据完成分类标识；敏感数据保护要求高敏数据加密存储覆盖率100%；日志留存要求审计日志留存≥6个月；内容安全防控要求有害内容拦截率≥98%。同时，可参考头部企业的最佳实践，逐步形成行业共识指标。

（二）数据合规与业务可用性的冲突

大模型训练需要大量真实数据，但数据合规要求限制了数据的采集、使用范围和留存时间。过度匿名化又会影响模型效果。应对策略包括：采用分层数据策略，生产数据→脱敏样本→合成数据，对高敏场景建立 " 沙盒训练环境 "，使用差分隐私、联邦学习等技术；建立数据授权与撤回机制，在用户协议中明确数据使用范围，建立 " 数据用于训练 " 的显式同意机制，提供用户撤回同意的渠道，记录授权与撤回日志，形成审计证据。

（三）模型输出责任与可解释性不足

大模型的 " 黑箱 " 特性使其输出难以完全预测。当模型产生错误输出导致损害时，责任如何界定？应对策略包括：采用人类在环审查策略，高风险场景（医疗、金融）100% 人工复核，中风险场景抽样复核≥10%；建立输出可追溯机制，完整记录模型输入与输出日志，在高风险场景展示推理依据，建立风险提示机制。在评估报告中展示上述措施，证明已履行 " 合理注意义务 "。

（四）评估方法与工具不成熟

大模型安全评估尚处于发展初期，缺乏统一的工具链，不同机构的测试集差异大。应对策略包括：尽量采用 T/ISEAA 006-2024、GB/T 45654-2025等标准的测试维度。采用工具组合策略：传统安全工具（漏洞扫描、渗透测试）、专门测试平台（大模型安全测试平台）、自研测试集（建立内部基准测试集）。随着项目经验积累，逐步形成行业认可的评估方法。

五、结语

本文从等级保护合规视角，探讨了生成式大模型安全评估的衔接路径。主要观点如下：

等保为网络安全提供了基础框架。等保2.0建立的物理安全、网络安全、应用安全、数据安全等控制域，为各类信息系统包括大模型系统提供了基础安全框架。修订后的《网络安全法》将等级保护条款与人工智能条款并列，为两者在网络安全法治体系中的协同提供了法律空间。

大模型安全标准提供了细化参考。T/ISEAA 005/006-2024、GB/T 45654-2025、GB/T 45288等专门标准，针对大模型特有的训练数据、生成内容、模型本体等风险，提出了细化的安全要求。这些探索可为未来大模型安全与等保框架的衔接提供参考。

协同评估或可作为探索方向。通过建立大模型风险与等保控制域的映射关系，可以探索提高评估效率、避免重复工作的路径。

从发展趋势看，工业和信息化部已发布《工业和信息化领域人工智能安全治理标准体系建设指南（2025）》，明确将构建

覆盖人工智能全生命周期的安全标准体系。未来值得关注的方向包括：团体标准向国家标准的转化、行业特定的大模型安全标准（金融、医疗等）、等级保护扩展要求中可能增加的大模型专项内容。

对于从业者而言：建设单位可考虑将安全合规要求尽早纳入系统设计，实现 " 安全左移 "；运营单位可探索建立持续评估机制，确保模型全生命周期安全可控；评估机构可加强对大模型安全特点的研究，提升评估能力。

大模型技术的安全发展需要政府、企业、评估机构等多方共同探索，在促进技术创新的同时，有效防范安全风险，逐步实现 " 以人为本、AI 向善 " 的发展目标。

参考文献

[1]《中华人民共和国网络安全法》（2025 年修正）

[2]《中华人民共和国数据安全法》

[3]《中华人民共和国个人信息保护法》

[4] 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 [Z]. 2023.

[5] 国家互联网信息办公室等四部门 . 人工智能生成合成内容标识办法 [Z]. 2025.

[6] 全国信息安全标准化技术委员会 . 网络安全技术 生成式人工智能服务安全基本要求：GB/T 45654-2025[S]. 2025.

[7] 全国信息安全标准化技术委员会 . 人工智能大模型 第 1-3 部分：GB/T 45288.1-3-2025[S]. 2025.

[8] 公安部网络安全等级保护评估中心 . 大模型系统安全保护要求：T/ISEAA 005-2024[S]. 2024.

[9] 公安部网络安全等级保护评估中心 . 大模型系统安全测评要求：T/ISEAA 006-2024[S]. 2024.

[10] 工业和信息化部 . 工业和信息化领域人工智能安全治理标准体系建设指南（2025 年版）[Z]. 2024.

