基于业务场景的大模型安全风险评估体系研究

魏光辉[1,2]，陈宇杰[1,2]，孔德智[1,2]，刘茂珍[3]，庞晓健[4]，甘清镕[1,2]
（1. 中国电子产品可靠性与环境试验研究所，广东 广州 511370；
2. 智能制造装备通用质量技术及应用工业和信息化部重点实验室，广东 广州 511370；
3. 广东电力人工智能试验研究院有限公司，广东 广州 510700；4. 南方电网财务有限公司，广东 广州 510623）

摘 要：提出一种基于业务场景的大模型安全风险评估模型，构建了涵盖安全能力、大模型模态、大模型在业务场景中的角色及交互的三维安全体系，围绕大模型业务及其资产进行风险评估建模，给出了完整的风险识别、风险分析与量化评价方法，并通过电力场景大模型实践验证了风险评估模型的全面性、准确性及可行性。该模型支持基于业务、技术及行业地区要求等要素的安全体系动态设计与差异化风险评估，为风险精准识别与有效防范提供科学的方法指引和实践指南。

关键词：业务场景；大模型安全；风险评估；风险识别；风险分析；风险评价
中图分类号：TP309 **文献标志码**：A DOI:10.19358/j.issn.2097-1788.2026.06.007

中文引用格式：魏光辉，陈宇杰，孔德智，等. 基于业务场景的大模型安全风险评估体系研究[J]. 网络安全与数据治理，2026，45(6)：47-56.
英文引用格式：Wei Guanghui，Chen Yujie，Kong Dezhi，et al. Research on large model security risk assessment system based on business scenarios[J]. Cyber Security and Data Governance，2026，45(6)：47-56.

Research on large model security risk assessment system based on business scenarios

Wei Guanghui[1,2]，Chen Yujie[1,2]，Kong Dezhi[1,2]，Liu Maozhen[3]，Pang Xiaojian[4]，Gan Qingrong[1,2]
（1. China Electronic Product Reliability and Environment Test Research Institute，Guangzhou 511370，China；
2. The Ministry of Industry and Information Technology Key Laboratory of General Quality Technology and Application for Intelligent Manufacturing Equipment，Guangzhou 511370，China；
3. Guangdong Power Artificial Intelligence Experimental Research In.，Guangzhou 510700，China；
4. China Southern Power Grid Finance Co.，Ltd.，Guangzhou 510623，China）

Abstract：This paper proposes a business scenario-based security risk assessment model for large models. It constructs a three-dimensional security system encompassing security capabilities，large model modalities，and the roles and interactions of large models in business scenarios. Risk assessment modeling is conducted around large model businesses and their assets，and a complete set of methods for risk identification，risk analysis and quantitative evaluation is presented. The comprehensiveness，accuracy and feasibility of the proposed risk assessment model are verified through the practice of large models in the power industry scenario. This model supports the dynamic design of security systems and differentiated risk assessment based on factors such as business requirements，technical specifications，and industry and regional regulations，providing scientific methodological guidance and practical references for accurate risk identification and effective prevention.

Key words：business scenario；large model security；risk assessment；risk identification；risk analysis；risk evaluation

0 引言

近年来，大语言模型在电力、金融等行业深度应用的同时，衍生出提示词注入、模型幻觉、数据投毒等新型安全风险，并可能通过业务交互链条引发连锁式安全事件。现有研究多聚焦于算力、算法、数据、应用、内容等层面安全，缺乏一套能够系统融合业务场景特性与风险全要素的可量化风险评估体系。本文构建了基于业务场景的涵盖安全能力、大模型模态、大模型在业务场景中的角色及交互的大模型安全体系及其风险评估模型，并通过电力调度知识问答大模型

验证其有效性，实现大模型全要素安全风险的精准识别、分析与评价，为大模型赋能千行百业的安全风险管理提供科学指引和实践指南。

1 研究现状

大模型安全评估体系聚焦于核心风险类型的系统梳理与攻击手段的精准识别，同时正加速推动评估方法从静态定性描述向动态量化分析转型。黄河燕等[1]将大模型安全风险系统性地划分为模型自身安全与生成内容安全两大类共十种；黑一鸣等[2]进一步归纳了越狱攻击、提示词注入及数据投毒等主要攻击手段。在评估方法上，鲁彪等[3]引入熵权法构建了三维度动态量化模型，范瑞龙等[4]则提出了包含攻击场景模拟与生命周期动态监测的改进评估框架。此外，通过红队测试主动发现模型漏洞已成为关键手段[5]。

然而，现有研究多集中于模型算法层面的通用技术风险，或侧重于数据安全这一单一维度，尚未形成一套能够系统整合业务场景特性、技术架构、威胁态势与防护能力的综合性风险评估体系。此外，在风险量化方法上，现有工作多采用定性分析或单一维度评估[6]，缺乏将业务影响与安全事件发生概率相结合的多要素量化模型，导致评估结果难以直接指导实际的决策与防护加固。因此，如何构建一套基于业务场景、覆盖大模型关键维度、全要素的量化评估体系，是当前“人工智能+”时代下不同场景大模型安全风险精准识别与评价的关键所在。

2 基于业务场景的大模型安全体系构建

基于业务场景的大模型安全体系内涵如图 1 所示。

（1）安全能力维度

安全能力包括组织机构、制度流程、人员能力、技术工具四大方面。

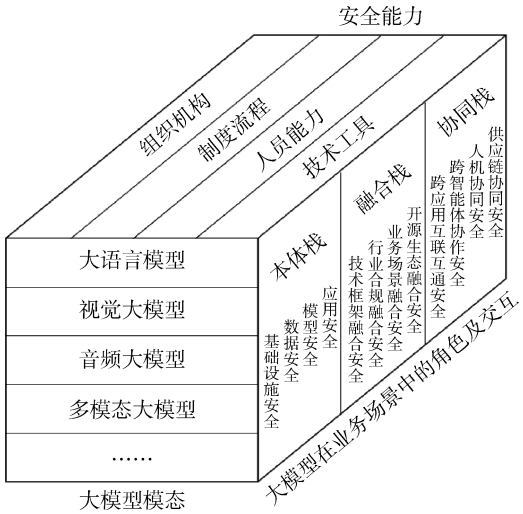

图 1　基于业务场景的大模型安全体系内涵

组织机构是安全能力的组织保障，指组织为落实安全管理建立的规范架构、岗位设置及权责划分，以明确各层级职责，确保安全工作统筹推进、层层落实。

制度流程是安全管理的规范依据，指组织制定的系统化安全规章制度与闭环流程，覆盖业务全周期，用以规范操作行为，实现安全管控有章可循、合规可控。

人员能力是安全落地的核心支撑，包括全员安全意识与安全岗位人员的专业技能，直接决定组织对安全风险的识别、防范与应急处置效果。

技术工具是安全防护的技术手段，指各类安全软硬件与平台，构建主动防御、实时预警的技术屏障，如蒸馏防御、网络认证、对抗训练工具等[7]，用以抵御各类安全威胁、保障资产安全。

（2）大模型模态维度

大模型模态包括大语言模型、视觉大模型、音频大模型及多模态大模型等。

大语言模型：以自然语言为核心处理载体，具备语义理解、逻辑推理与文本生成能力，技术体系成熟，擅长语言类任务。

视觉大模型：专注图像、视频等视觉数据处理，可完成目标检测、场景解析与视觉内容生成，视觉感知能力突出。

音频大模型：以语音、声波等音频信息为处理对象，支持语音识别、音频分析与语音合成，听觉感知优势明显。

多模态大模型：融合文本、视觉、音频等多种数据类型，实现跨模态信息融合与协同推理，信息感知维度丰富，综合理解与复杂场景适配能力强。

（3）大模型在业务场景中的角色及交互维度

该维度包含本体栈、融合栈、协同栈三个维度。

①本体栈

本体栈聚焦大模型应用作为独立数字资产，是整体防护体系的底层核心，其防护能力决定上层融合与协同环节的安全基线。该栈围绕基础设施、数据、模型、应用四大核心要素构建全域防护体系。

基础设施安全：覆盖算力、存储、部署环境、网络等，防范硬件漏洞、调度失控、数据丢失等风险。

数据安全：覆盖收集、存储、使用、加工、提供、公开、传输、删除等环节，重点防范数据泄露、投毒、篡改等风险[8]。

模型安全：包括架构、参数、训练、防护等方面，

防范模型幻觉、后门、对抗样本攻击等风险[9]，提升模型可解释性和鲁棒性。

应用安全：聚焦应用接口、内容生成、用户交互、运维，防范注入攻击[10]、越权访问、内容不合规等风险。

②融合栈

随着大模型向智能体形态演进，安全风险呈现跨域性、传导性与复杂性特征。融合栈作为衔接本体与场景应用的中间层，构建场景化、适配性防护机制，解决通用安全策略与业务场景风险防范脱节的问题。

技术框架融合安全：应对智能体、具身智能、多模态大模型、云边端部署融合带来的接口适配、权限管控、对抗攻击、越狱攻击、后门攻击[11]等风险。

行业合规融合安全：落实行业法律法规与科技伦理要求，防范合规风险和伦理隐患。

业务场景融合安全：基于业务场景特性，强化场景敏感数据与高风险权限管控。

开源生态融合安全：防范开源代码、工具、插件的漏洞和恶意代码注入等风险[12]。

③协同栈

大模型在业务场景中的应用因具有技术架构的协同性、跨边界交互特征，打破了传统安全边界，引发风险传导与放大效应[13]。协同栈聚焦跨系统、跨主体协同全流程风险管控，构建全局安全治理机制。

跨应用互联互通安全：统一接口安全标准，建立数据权限协同和监测应急联动机制。

跨智能体协作安全：规范协作通信协议安全使用，防范恶意指令注入和决策冲突，保障数据机密性[14]。

人机协同安全：强化高风险场景人工复核，明确人机权限划分，建立应急干预机制。

供应链协同安全：贯穿全产业链，防范硬件后门、软件漏洞传导等风险，实施严格的供应商准入机制。

3 基于业务场景的大模型安全风险评估体系构建

3.1 风险要素关系

基于业务场景的大模型安全风险要素关系如图 2 所示。风险要素包括大模型资产（基础设施、数据、模型、应用等）、大模型相关业务、威胁、脆弱性和安全措施。其中，风险要素的核心是大模型资产，大模型资产存在脆弱性；安全措施的实施是为抵御大模型威胁，降低大模型资产脆弱性被利用的机率；威胁通过利用大模型资产存在的脆弱性导致风险；风险转化成大模型安全事件后，会对大模型资产及相关业务的运行状态产生影响。

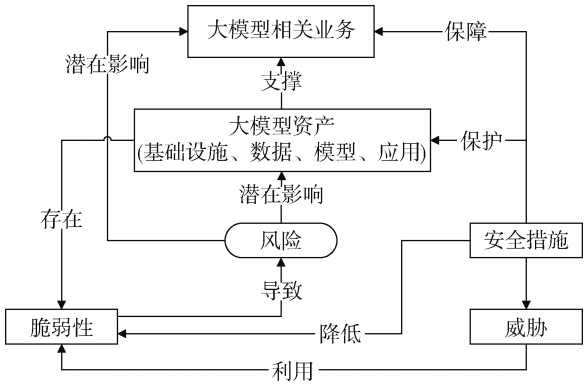

图 2　基于业务场景的大模型安全风险要素关系

3.2 风险识别

（1）资产价值

大模型资产通常包括基础设施、数据、模型、应用、管理制度、专业人员等。大模型资产价值是根据资产在发展规划中所处的地位和资产的属性确定的。在识别大模型资产后，可从保密性、完整性和可用性等因素进行综合计算，并设定相应的评级方法进行价值等级划分。

（2）业务重要性

业务识别和业务重要性识别构成了风险量化与治理策略制定的先决条件。首先，通过解析业务属性，明确模型的功能边界、服务对象、技术流程及覆盖地域；其次，通过界定业务定位，评估模型在组织整体战略版图中的角色权重及其与战略目标的契合程度；最后，通过分析业务关联性，识别模型与上下游系统的耦合强度，特别是判定模型失效可能引发的级联故障风险。这种系统性的识别机制为后续的风险定级提供了必要支撑。业务重要性聚焦于大模型所承载业务的运行优先级与影响范围，侧重刻画不同业务场景在生产运行体系中的不可替代性与管控等级。不同业务存在实时性要求、覆盖范围、控制权限及监管强度的显著差异，业务重要性通过区分生产控制类核心业务、经营管理类重要业务与辅助支撑类一般业务，明确大模型安全风险触发后对上层业务链条的传导影响，实现技术风险向业务风险的关联转化。

（3）脆弱性

脆弱性是大模型各类资产自身与生俱来的安全缺陷与薄弱环节，属于风险发生的内部固有诱因。在具体业务场景下，脆弱性既包含模型幻觉、对抗鲁棒性不足、决策可解释性缺失等算法层面技术短板，也涵盖数据治理不规范、权限配置不合理、接口防护缺失、

应用风险监测感知设施不健全等[15]应用层漏洞，同时涵盖安全管理制度不完善、运维管控不规范、人员能力欠缺等管理和人员类缺陷，直接决定威胁行为的可利用难度与攻击突破概率。

（4）威胁

威胁来源种类用于界定可能对大模型资产与业务造成损害的各类风险主体及产生源头，既包括外部恶意攻击、内部人员操作、软硬件供应链、自然环境扰动等常规类别，也包括跨网络、跨系统、跨业务的衍生威胁，需实现风险来源的全维度排查。

威胁动机反映威胁主体实施破坏、窃取、干扰等行为的主观意图与行为目的，是判断威胁危害程度与演化趋势的重要依据。既有外部组织蓄意破坏业务运行、窃取核心敏感数据的恶意诉求，也包含内部人员操作疏忽、配置失误等非主观过失行为，不同动机直接决定威胁行为的持续性、隐蔽性与破坏性，为风险发生可能性的分级判定提供支撑。

威胁能力指代威胁主体开展攻击或破坏行为所具备的技术水平、工具资源、专业经验与组织规模，体现威胁行为的实施门槛与突破能力。从常规试探行为，到利用专业工具开展漏洞利用、对抗攻击的能力行为，再到定制化渗透、持续性潜伏的威胁行为，威胁能力的差异化特征，直接影响脆弱性利用成功率与安全事件的发生概率。

威胁发生时机和频率指依据历史安全事件记录、行业同类风险案例及现场运行监测数据，综合研判各类威胁行为出现的特定运行时段与发生概率。通过划分不同等级，量化威胁常态化活跃程度，客观反映大模型长期运行过程中各类安全风险的暴露强度，为风险可能性量化计算提供稳定的数据支撑。

（5）安全措施

安全措施是当前业务运行过程中已部署、常态化运行的各类安全管控与防护手段，也是弱化脆弱性短板、抵御外部威胁、降低整体风险水平的关键缓冲要素。现有安全措施涵盖事前预防、事中监测、事后恢复及常态化管理多个维度，通过加密防护、访问控制、安全监测、应急处置、制度规范等措施的落地见效，实现对风险的抵消作用。

3.3 风险分析、计算与评价

3.3.1 风险分析模型

针对大模型业务场景涉及的资产情况，在识别系统资产 A 及其业务承载连续性后对资产价值 V_A 赋值，识别业务 B 后对其重要性 V_B 赋值，识别威胁来源、种类、动机后对威胁的能力 E 和频率 F 赋值，识别安全措施、威胁和脆弱性后对脆弱性被威胁利用难易程度 H 和对业务的影响程度 I 赋值，采用适当的计算方法与工具确定安全事件发生的可能性 L 和若发生安全事件造成的损失 K，并计算出所有资产对应的风险 Q_1，Q_2，…，Q_N。然后，结合资产对大模型业务的重要性 G_1，G_2，…，G_N，对所有处理活动的风险进行加权计算，最终得出大模型业务场景的综合风险 Q。风险分析模型如图 3 所示。

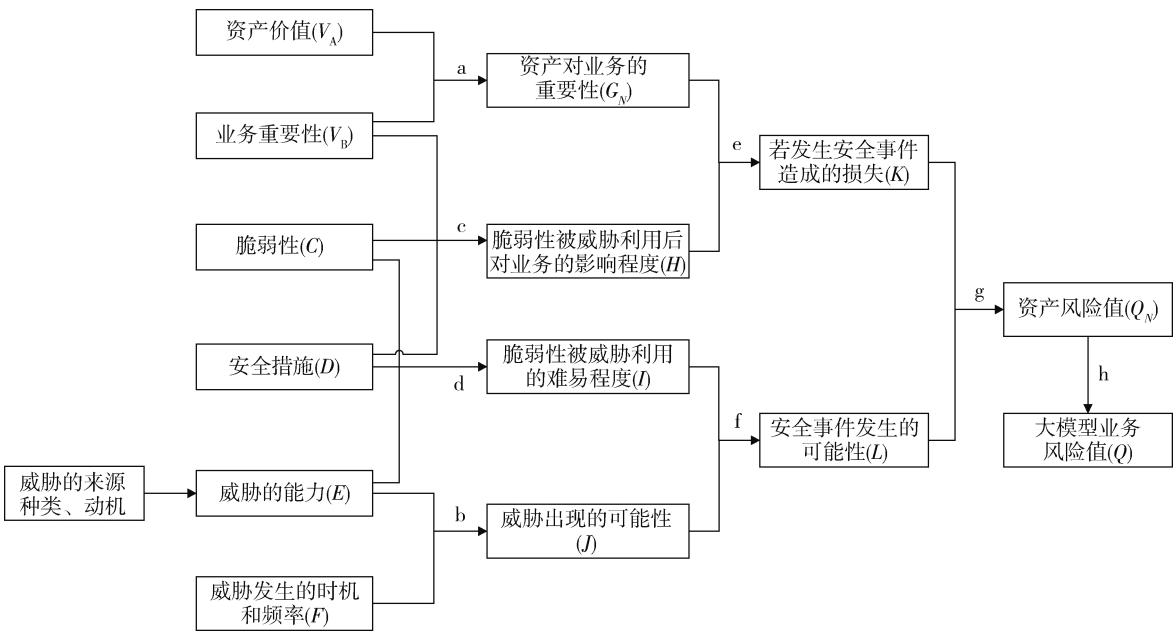

图 3　风险分析模型

3.3.2 风险计算

风险评估量化计算的主流方法包括相乘法、风险矩阵法、层次分析法（AHP）、模糊综合评价法等，本文采用相乘法、加权平均运算作为主要方法，用于由多个要素值确定一个要素值。相乘法的原理是：$z=f(x,y)$，当f为增量函数时可以直接相乘，也可相乘后取模，本文采用相乘后开根号运算，即$z=f(x,y)=\sqrt{x\times y}$。加权平均运算的原理是：给不同的数据点赋予不同的权重，再计算它们的加权和，最终得到一个能反映“重要性差异”的平均值。风险计算流程如下：

（1）计算资产对业务的重要性（G）

$G=f[$资产价值(V_A),业务重要性(V_B)$]$，即：

$$G=\sqrt{V_A\times V_B} \tag{1}$$

（2）计算威胁出现的可能性（J）

$J=f[$威胁的能力(E)，威胁发生的时机和频率(F)$]$，即：

$$J=\sqrt{E\times F} \tag{2}$$

（3）计算脆弱性被威胁利用后对业务的影响程度(H)

$H=f[$业务重要性(V_B),脆弱性(C),安全措施(D)$]$,即：

$$H=\sqrt{V_B\times M[C,D]} \tag{3}$$

其中，M为脆弱性C被已有安全措施D修正后的值。

（4）计算脆弱性被威胁利用的难易程度（I）

$I=f[$脆弱性(C)，安全措施(D)，威胁的能力(E)$]$，即：

$$I=\sqrt{E\times M[C,D]} \tag{4}$$

其中，M为脆弱性C被已有安全措施D修正后的值。

（5）计算若发生安全事件造成的损失（K）

$K=f$［资产对业务的重要性(G)，脆弱性被威胁利用后对业务的影响程度(H)］，即：

$$K=\sqrt{G\times H} \tag{5}$$

（6）计算安全事件发生的可能性（L）

$L=f$［脆弱性被威胁利用的难易程度(I)，威胁出现的可能性(J)］，即：

$$L=\sqrt{I\times J} \tag{6}$$

（7）计算资产风险值（Q_N）

$Q_N=f$［若发生安全事件造成的损失(K)，安全事件发生的可能性(L)］，即：

$$Q_N=K\times L \tag{7}$$

（8）计算大模型业务风险值（Q）

依据前文，大模型业务风险值$Q=g([Q_1,Q_2,\cdots,Q_i],[G_1,G_2,\cdots,G_i])$，本文采用加权平均法计算大模型业务风险值，权重为资产对业务的重要性G，即：

$$Q=\frac{\sum_{i=1}^{N}(Q_i\times G_i)}{\sum_{i=1}^{N}G_i} \tag{8}$$

其中Q_i表示第i个资产，G_i表示第i个资产对业务的重要性权重。

3.3.3 风险评价

结合大模型配套的资产安全防护能力建设水平，将大模型业务整体安全风险划分为极低风险、低风险、中风险、高风险、极高风险共5个等级，并为每个等级赋量化综合风险分值区间。通过将大模型业务的量化综合风险分值与等级判定标准进行对照，即可判定当前大模型业务的整体安全风险水平，明确风险管控优先级与整改处置要求。

3.4 风险评估方法

本文所构建的风险评估体系采用人员访谈、文档查验、系统演示、工具验证四类方法，遵循“客观验证优先、主观问询为辅、多方法交叉印证”的原则，根据评估环境灵活组合、交叉运用不同方法，保障结果的客观性与可信度。工具验证依托专业检测手段开展客观量化分析，效力最高，作为核心判定依据；系统演示动态核验防护措施落地效果，效力较工具验证次之；文档查验作为静态合规审查依据，效力中等；人员访谈易受主观偏差影响，效力最低，不可单独作为风险判定依据。

4 以电力场景大模型为例的典型领域实践

选取电力行业的“电力调度知识问答大模型”作为评估对象，开展安全风险评估体系的有效性、可操作性验证。

4.1 风险识别

4.1.1 资产价值

该系统基于多模态大模型驱动的智能体构建，提供实时规程查询、故障预案推荐等功能，具备自主推理、多工具调用和动态决策能力，可辅助调度员完成复杂电网运行场景下的分析与处置任务。基于该场景构建的资产价值等级赋值表如表1所示。

通过专家评价法进行赋值（后续采用相同方法赋值，不再赘述），如表2所示。

4.1.2 业务重要性

针对业务特点，设计并构建了分级量化的业务重要性赋值表，如表3所示。

表 1　资产价值等级赋值表

赋值	等级	描述
5	很高	安全属性破坏后对组织造成非常严重的损失
4	高	安全属性破坏后对组织造成比较严重的损失
3	中等	安全属性破坏后对组织造成中等程度的损失
2	低	安全属性破坏后对组织造成较低的损失
1	很低	安全属性破坏后对组织造成很小的损失，可忽略不计

表 2　“电力调度知识问答大模型”资产价值赋值表

序号	资产名称	资产价值
1	大模型推理服务器	5
2	模型前端系统	4
3	大模型训练平台	3
4	调度知识数据	5
5	多模态大模型基座	5
6	算力调度平台	4
7	容器化运行环境	4
8	电力专用通信网络	4
9	身份认证系统	5
10	大模型全生命周期管理制度	4
11	数据安全管理制度	4
12	大模型安全管理人员	5
13	大模型安全技术人员	5

表 3　业务重要性等级赋值表

赋值	等级	描述
5	很高	属于电网核心生产控制类业务，失效将严重冲击电网安全稳定，具有核心地位
4	高	属于电网重要生产运营类业务，安全风险将对区域电网造成重大影响，具有重要地位
3	中等	属于电力关键经营与运维支撑类业务，安全事件将造成较大业务扰动，具有较为重要地位
2	低	属于一般辅助管理类业务，安全问题仅造成一般性影响，不触及核心安全底线，处于一般地位
1	很低	属于边缘试验性辅助业务，容错空间大，安全问题仅产生轻微影响，处于次要地位

基于所承载的业务进行重要性识别与赋值。该业务直接服务于电网调度决策与操作执行，其失效或遭受攻击可能对区域电网的安全稳定运行造成重大影响，故“电力调度知识问答大模型”所承载业务的重要性赋值为“4”。

4.1.3　脆弱性

针对业务运行特点，构建了一套系统化的脆弱性识别框架，如表 4 所示。

表 4　脆弱性识别表

类型	识别对象	识别方面
本体栈	数据安全	非法采集电网拓扑、调度指令等电力涉密数据；传输环节电力核心数据未加密；存储环节电力数据加密失效、异地备份不完善；使用环节电力敏感信息脱敏失效；销毁环节电力历史运行数据销毁不彻底、无销毁记录；电力大模型数据安全管理组织机构设置不合理
	模型安全	模型架构存在漏洞，抗恶意查询、对抗样本攻击能力偏弱，易输出错误调度决策；模型防护能力不足，抗数据投毒能力弱；模型全生命周期安全管理制度缺失，流程不规范；模型上线、迭代、下线安全评审流程不完善
	应用安全	接口存在未授权访问、缺少身份认证；交互环节身份认证，电力调度操作记录留存不全；应用运维环节漏洞未及时修复、电力业务权限分级不明、安全审计机制缺失
	基础设施安全	电力专用算力基础设施存在硬件漏洞、算力调度失控、核心设备易发生故障；电力工控存储加密失效、备份不完善；电力大模型部署环境存在漏洞、与工控系统未隔离、配置不符合电力安全规范
融合栈	技术框架融合安全	电力智能体调用工具接口权限管控不严、多智能体调度决策冲突；电力云边端分布式部署数据同步加密失效、边缘场站防护能力不足
	行业合规融合安全	未满足电力行业法律法规、关键信息基础设施保护要求；违背电力科技伦理，输出歧视性内容
	业务场景融合安全	场景防护策略与电力调度/运检业务脱节，模型输出结果未做电力规则校验；场景化电力隐私数据管控不足、多源数据融合质量差、易引发决策偏差
	开源生态融合安全	电力场景开源工具存在漏洞，版本滞后未更新；电力业务开源插件未进行安全审核，权限管控不严
协同栈	跨应用互联互通安全	电力工控系统接口存在未授权访问，跨系统数据交互无加密；跨系统数据与权限分配不合理，存在越权访问核心调度数据的情况；跨系统安全监测与应急机制不完善，响应不及时，未开展应急演练
	跨智能体协作安全	电力多智能体协作协议存在漏洞，身份认证失效；智能体决策存在冲突，调度权限管控不严；协同安全管控监测缺失
	人机协同安全	调度员人工审核存在疏漏，操作失误，电力业务操作日志留存不全；人机管控权限划分不明，无强制人工复核，模型决策与人工管控脱节；人员缺乏电力大模型安全专业能力，无法识别模型幻觉、数据投毒等风险；应急处置能力不足，未开展电力安全应急培训，无法应对模型安全突发事件
	供应链协同安全	电力工控硬件供应链存在后门，影响电网系统稳定运行；供应链存在漏洞及风险传导问题，配套服务存在安全隐患；供应链安全管控监测缺失，应急处置能力弱，权责不清，供应商准入不严

在完成脆弱性识别后，需对各项脆弱性的严重程度进行量化赋值，如表 5 所示。

表 5　脆弱性等级赋值表

赋值	等级	描述
5	很高	可直接导致电网核心调度业务瘫痪、大规模敏感数据泄露或系统彻底失控。且攻击门槛极低，广泛存在公开的利用代码
4	高	可导致关键生产运营业务中断、较大量敏感数据泄露，或模型输出严重偏离电力物理规则。且利用需一定技术能力或内部信息
3	中等	可能导致局部辅助业务异常、一般性管理数据泄露，或模型在特定场景下出现可容忍的偏差，但不直接威胁核心电网运行。且需具备内部知识或社会工程学手段，利用门槛中等，无广泛公开利用代码
2	低	仅可能造成轻微的业务扰动、极小范围的非敏感信息泄露，且存在较容易实施的替代防护措施。且利用门槛极高，需多步攻击链或高级身份，无公开利用途径
1	很低	基本不存在实际可利用路径，或理论风险已被现有其他防护完全覆盖，几乎无法造成任何实质性损害。且无已知利用途径，防护措施已完全覆盖

依据脆弱性识别框架，识别出以下关键脆弱性：（1）模型安全层面，经对抗样本测试发现，该大模型在罕见复杂故障场景下的推理准确性不足，存在模型幻觉，易输出错误；（2）应用安全层面，接口渗透测试发现 API 存在未授权访问隐患，身份认证机制不够完善；（3）业务场景融合安全层面，输出结果未经过电力物理规则校验，存在调度决策与电网实际运行状态脱节的问题；（4）在开源生态融合安全层面，系统引入的第三方知识库查询插件版本滞后，存在已知安全漏洞未修复或数据投毒问题；（5）人机协同安全层面，调度员对模型输出的人工复核存在疏漏问题；（6）跨智能体互联互通安全层面，多智能体间的数据交互通信缺乏加密保护，无法保障数据机密性，易导致数据泄露。

通过上述脆弱性识别，共发现 6 项关键脆弱点，该脆弱性可导致关键生产运营业务中断、较大量敏感数据泄露或模型输出严重偏离电力物理规则，据此脆弱性赋值为“4”。

4.1.4　威胁

结合运行环境与攻击面特征，构建了一套系统化的威胁来源识别与分类框架，如表 6 所示。

针对面临的攻击特征，设计并构建了系统化的威胁动机识别与分类清单，如表 7 所示。

表 6　威胁来源、种类识别表

威胁大类	威胁子类	威胁行为	威胁来源
基础设施类	物理与算力环境损害	算力机房火灾、水灾；服务器硬件损毁、温湿度异常；电磁辐射、算力硬件物理破坏、存储介质损毁	环境、人为、意外
	自然灾害威胁	地震、洪水、极端气象灾害、地质灾害等，造成大模型算力中心整体停机、基础设施瘫痪、服务全域中断	环境
	基础配套系统失效	算力机房空调、冷却、供水系统故障，导致大模型算力集群过热停机宕机	人为、意外
	网络通信故障	大模型内外网链路中断、跨域通信故障、网络拥塞、专线链路瘫痪	人为、意外
	软硬件底座失效	算力服务器宕机、GPU/算力卡故障、大模型底座运行报错、中间件失效	意外
	供应链上游失效	基础模型供应商服务中断、模型底座版本停服、上游厂商技术支持失效	人为、意外
	人员可用性破坏	核心运维、模型管理、安全管控关键岗位人员缺失、人员操作能力丧失	环境、人为、意外
数据类	数据投毒与知识库污染	电力训练数据投毒、调度知识库污染导致模型学习错误电力规则	人为
	数据质量异常与偏差	设备状态数据、电力业务标注错误	人为、意外
	数据泄露与隐私残留	大模型训练数据泄露、推理日志敏感数据残留、电力调度指令日志泄露	人为、意外
	数据滥用与越权调用	电力敏感数据批量导出、跨场景违规调用、未授权访问电网拓扑数据、非合规数据流转	人为
	数据合规违规	电力数据超范围采集、违规留存、未脱敏数据对外提供	人为、意外
模型类	模型与信息损害	模型权重窃取、模型逆向还原、提示词注入、对抗样本攻击、模型越狱；训练数据投毒、知识库污染	人为
	模型幻觉与决策偏差	调度方案生成错误、设备故障误判、新能源出力预测偏差	意外、人为
	模型后门与隐蔽漏洞	训练阶段植入后门、隐蔽越狱漏洞、电力场景下的恶意指令触发	人为
	模型输出篡改与劫持	模型推理结果篡改、输出过滤绕过、电力业务决策被恶意修改	人为
应用类	系统过载与架构破坏	大模型并发调用过载、算力资源耗尽、系统可维护性与可用性遭到恶意破坏	人为、意外
	非法模型使用	盗版底座模型使用、非授权第三方模型接入、大模型业务违规私自部署	人为、意外
	功能运维损害	模型训练、部署、迭代过程操作失误，运维配置错误	意外
	恶意网络与人为攻击	针对大模型的专项网络攻击、权限伪造、大模型舆情与负面报道	人为
	权限越权滥用	内部人员越权调用高等级模型、超范围访问敏感业务、高级功能滥用	人为、意外

表 7　威胁动机识别表

分类	描述
恶意	大模型的越狱突破、数据窃取与售卖、恶意篡改与投毒、勒索摧毁、非法盗用及工业间谍等主动恶意破坏行为
非恶意	因技术好奇试探、运维迭代失误、数据与配置疏漏、权限误操作及安全意识不足引发的无意泄露、系统缺陷与业务扰动

分析模型可能面临的安全威胁，构建出一种威胁能力、发生的时机和频率识别参考，如表 8 所示。

依据所构建的威胁分析框架，对模型可能面临的关键威胁进行识别、动机研判与能力、频率量化赋值，如表 9 所示。

4.1.5　安全措施

经识别，存在大模型安全栅栏、WAF、防火墙等安全措施。本文采用量化修正法体现安全措施对脆弱性赋值的影响。相关安全措施为有效时，可根据其有效程度酌情降低脆弱性赋值。

4.2　风险分析、计算与评价

4.2.1　风险分析与计算

风险计算矩阵如表 10 所示。

表 8　威胁能力、发生时机和频率等级赋值表

赋值	等级	威胁能力描述	威胁发生时机和频率描述
5	很高	恶意动机极强，具备专业 APT 能力；面临极端灾害，威胁极强，几乎可完全突破现有防护	频率极高；日常运行中频繁触发；同类攻击与故障已常态化反复发生
4	高	恶意动机强烈，具备成熟黑客能力；面临较严重灾害，突破能力较强，可绕过多数防护	频率较高；长期运行中大概率多次发生；行业及本系统历史中已多次出现
3	中等	恶意动机一般，具备基础漏洞利用能力；面临意外，仅能利用通用脆弱性，有突破概率	频率中等；特定业务与迭代阶段有较大发生概率；过往有明确同类记录
2	低	恶意动机较弱，仅掌握基础试探；多为非恶意误操作，仅能利用低级短板，很难突破防护	频率较低；常规环境下不易触发；本系统及行业内极少被观测证实发生
1	很低	无主观恶意，技术能力极低；面临轻微意外，完全无法突破防护，几乎不会造成有效损害	几乎不可能发生；仅极端罕见工况下存在理论可能，实际几乎无暴露风险

表 9　“电力调度知识问答大模型”威胁能力及时机、频率赋值表

序号	资产名称	威胁动机	威胁行为	威胁能力	威胁发生时机、频率
1	大模型推理服务器	恶意	底层框架崩溃	3	2
2	模型前端系统	恶意	电力场景下的恶意指令触发	4	4
3	大模型训练平台	恶意	训练阶段植入后门、恶意触发条件、隐蔽越狱漏洞	4	3
4	调度知识数据	恶意	电力训练数据投毒	4	3
5	多模态大模型基座	恶意	模型输出篡改	4	3
6	算力调度平台	非恶意	算力服务器宕机	4	3
7	容器化运行环境	恶意	开源组件被攻击	4	3
8	电力专用通信网络	恶意	专线链路瘫痪	4	3
9	身份认证系统	恶意	大模型未授权调用	3	4
10	大模型全生命周期管理制度	非恶意	模型管理制度不健全导致敏感数据泄露	3	4
11	数据安全管理制度	非恶意	模型管理制度不健全导致敏感数据泄露	3	4
12	大模型安全管理人员	恶意	内部人员越权调用高等级模型	3	4
13	大模型安全技术人员	恶意	内部人员越权调用高等级模型	3	4

表 10 “电力调度知识问答大模型”资产风险计算矩阵

序号	资产名称	资产价值	业务重要性	脆弱性	安全措施修正后的脆弱性	威胁能力	威胁发生的时机和频率	资产对业务的重要性	脆弱性被威胁利用后对业务的影响程度	脆弱性被威胁利用的难易程度	威胁出现的可能性	若发生安全事件造成的损失	安全事件发生的可能性	资产风险值
1	大模型推理服务器	5	4	4	3	3	2	4.47	3.46	3.00	2.45	3.94	2.71	10.67
2	模型前端系统	4	4	4	3	4	4	4.00	3.46	3.46	4.00	3.72	3.72	13.86
3	大模型训练平台	3	4	4	3	4	3	3.46	3.46	3.46	3.46	3.46	3.46	12.00
4	调度知识数据	5	4	4	3	4	3	4.47	3.46	3.46	3.46	3.94	3.46	13.63
…	…	…	…	…	…	…	…	…	…	…	…	…	…	…
13	大模型安全技术人员	5	4	4	3	3	4	4.47	3.46	3.00	3.46	3.94	3.22	12.69

经计算，大模型业务风险值 $Q=g([Q_1,Q_2,\cdots,Q_i],[G_1,G_2,\cdots,G_i])=12.66$。

4.2.2 风险评价

基于场景相关安全风险，将业务整体安全风险划分为极低、低、中、高、极高 5 个等级，如表 11 所示。

表 11 大模型安全风险等级划分表

风险等级	等级标识	综合风险值区间	大模型防护能力与安全风险特征描述
1	极低风险	$1\leqslant Q\leqslant 5$	防护能力极强，完全满足最高要求；无高危脆弱性，几乎无隐患
2	低风险	$5<Q\leqslant 10$	防护体系较完善，核心属性高水准；仅存轻微脆弱点，可抵御绝大多数威胁
3	中风险	$10<Q\leqslant 15$	防护存在短板，部分属性不足；仅能抵御常规威胁，中等脆弱性有被利用可能
4	高风险	$15<Q\leqslant 20$	防护能力偏弱，多项属性不达标；关键脆弱性突出，威胁频率偏高
5	极高风险	$20<Q\leqslant 25$	基础防护严重缺失，安全属性全面失效；高危脆弱性广泛存在，极易触发级联事故

因此，“电力调度知识问答大模型”安全风险等级为“中风险”。

5 结论

本文针对大模型在新技术新业务新场景下安全风险评估问题，采用理论分析与实践验证相结合的方法，创新性地构建了基于业务场景的大模型安全风险评估模型，并通过电力场景实例验证评估模型的有效性、实用性，有效弥补了大模型安全体系构建较为片面、风险评估模型无法识别新兴风险方面的不足。

参考文献

[1] 黄河燕,李思霖,兰天伟,等. 大语言模型安全性:分类、评估、归因、缓解、展望[J]. 智能系统学报,2025,20(1):2-32.
[2] 黑一鸣,陈文弢,石霖,等. 智能体安全风险评估与防御技术综述[J]. 中国信息安全,2025(10):73-78.
[3] 鲁彪,王晟,李超峰,等. 基于熵权法下的大模型应用安全风险评估与防护体系构建[J]. 电信工程技术与标准化,2025,38(12):52-57.
[4] 范瑞龙,达虎,李洋,等. 基于大模型敏感数据泄露风险的量化评估方法改进[J]. 科技创新与应用,2025,15(34):123-127.
[5] 包泽芃,钱铁云. 大模型红队测试研究综述[J]. 计算机科学,2025,52(1):34-41.
[6] 苏艳芳,袁静,薛俊民. 大模型安全评估体系框架研究[C]//中国计算机学会. 第 39 次全国计算机安全学术交流会论文集,2024:107-111.
[7] 刘亦石,周亚建,崔莹,等. 人工智能大模型应用中的安全问题与解决策略[J]. 网络空间安全科学学报,2024,2(1):83-91.
[8] 李甜,孙一鸣,董振江,等. 大模型数据安全与隐私保护研究综述[J/OL]. 物联网学报,1-18[2026-05-07]. https://link.cnki.net/urlid/10.1491.TP.20251212.1043.002.
[9] 张欣,康荣保,饶志宏,等. 生成式人工智能大模型安全风险与防御策略研究[J/OL]. 指挥控制与仿真,1-11[2026-05-07]. https://link.cnki.net/urlid/32.1759.TJ.20260527.1817.014.
[10] 王正超. 公共数据开发利用中大模型应用的安全风险与治理路径[J/OL]. 情报杂志,1-9[2026-05-07]. https://link.cnki.net/urlid/61.1167.G3.20260403.1341.002.
[11] 郭园方,余梓彤,刘艾杉,等. 多模态大模型安全研究进展[J]. 中国图象图形学报,2025,30(6):2051-2081.
[12] 王新雷,饶宇锋. 开源大模型基础设施化进程中的安全风险与法律治理——以 DeepSeek 为例[J]. 湖北经济学院学报,2025,23(2):69-79.
[13] 宗绍昊,章钰敏. AI 智能体的新型数据安全风险及其治理[J/OL]. 科学学研究,1-16[2026-05-07]. https://doi.org/10.16192/j.cnki.1003-2053.20260510.001.
[14] 祝烈煌,廖乐健,曹元大,等. 多智能体安全通信协议[J]. 北京理工大学学报,2006(10):888-891.

[15]郭宇,王悦,徐仕远. 国产大模型的数据安全风险与技术治理路径[J]. 中国信息安全,2025(3):25-27.

（收稿日期：2026-05-07）

作者简介：

魏光辉（1989-），男，硕士，高级工程师，主要研究方向：数据安全、个人信息保护、网络安全、人工智能安全。

陈宇杰（2002-），男，本科，工程师，主要研究方向：数据安全、网络安全、人工智能安全。

孔德智（1985-），通信作者，男，硕士，高级工程师，主要研究方向：数据安全、个人信息保护、人工智能安全。E-mail：464829283@ qq. com。

版权声明

凡《网络安全与数据治理》录用的文章，如作者没有关于汇编权、翻译权、印刷权及电子版的复制权、信息网络传播权与发行权等版权的特殊声明，即视该文章署名作者同意将该文章的汇编权、翻译权、印刷权及电子版的复制权、信息网络传播权与发行权授予本刊，本刊有权授权本刊合作数据库、合作媒体等合作伙伴使用。同时，本刊支付稿酬已包含上述使用的费用，特此声明。

《网络安全与数据治理》编辑部