DOI 10.15302/J-SSCAE-2025.06.016

面向安全治理的大语言模型风险分析与应对策略研究

贾堃[1,2,4]，张钰歆[3,4]，陈继昀[1,4]，齐佳音[1,2,4,5*]，方滨兴[1,2,5]

（1. 广州大学网络空间安全学院，广州 510006；2. 广州大学黄埔研究院，广州 510006；3. 巴勒莫大学政治学与国际关系系，巴勒莫 90133；4. 粤语语料库建设与大模型评测重点实验室，广州 510006；5. 可信分布式计算与服务教育部重点实验室，北京 100084）

摘要：大语言模型（LLM）安全风险的认知碎片化、治理策略滞后现象凸显，亟需融合风险机理分析、量化评估、治理实践的综合框架。本文在辨析全球治理实践的演变与挑战、现有LLM风险分类分级框架呈碎片化与割裂化的基础上，揭示了LLM风险源于模型内部复杂性、外部交互的二元触发机制，将风险剖析为内生安全、应用安全两个维度，据此提出了“双维驱动”的风险分析与治理框架；引入了“风险标签卡片”作为标准化工具，采用“人工智能+人类专家协同”范式进行了真实安全案例的结构化解析，结合改进的DREAD风险矩阵模型，建立了从定性识别到定量分级的完整评估方法论；最终构建了LLM安全风险分类体系以及覆盖主要风险类型的高、中、低三级风险图谱，并从实施“双维驱动”的风险管控核心策略、健全系统性的治理保障体系两方面形成了LLM安全风险治理建议。研究提出的“双维驱动”的风险分析与治理框架，具有良好的理论兼容性与动态特性，为精准评估和治理LLM安全风险提供了理论工具，有效弥合了LLM安全风险治理实践存在的“理论–操作鸿沟”，为持续追踪和理解LLM安全风险并制定安全政策提供了直接参考。

关键词：大语言模型；安全风险；安全治理；风险评估；分类分级；风险图谱

中图分类号：TP309　　**文献标识码：**A

Risk Analysis and Response Strategies of Large Language Models for Security Governance

Jia Kun[1,2], Zhang Yuxin[3,4], Chen Jiyun[1,2], Qi Jiayin[1,2,5*], Fang Binxing[1,2,5]

(1. Cyberspace Institute of Advanced Technology, Guangzhou University, Guangzhou 510006, China; 2. Huangpu Graduate School of Guangzhou University, Guangzhou 510006, China; 3. Department of Political Sciences and International Relations, University of Palermo, Palermo 90133, Italy; 4. Key Laboratory of Cantonese Corpus Construction and Large Language Model Evaluation, Guangzhou 510006, China; 5. Key Laboratory of Trustworthy Distributed Computing and Service, Ministry of Education, Beijing 100084, China)

Abstract: To address the challenges of fragmented understanding of Large Language Model (LLM) security risks and the inadequacy of LLM risk classification and grading frameworks, this study aims to construct a comprehensive framework that integrates risk mechanism analysis, quantitative assessment, and governance practices. Theoretically, this study synthesizes and reconstructs multiple foundational theories, including socio-technical systems, social systems theory, and safety science, to reveal that risks originate from a dual trigger mechanism of the model’s “internal complexity” and “external interaction.” It consequently dissects risks into two

收稿日期：2025-06-15；**修回日期：**2025-09-21

通讯作者：*齐佳音，广州大学网络空间安全学院教授，研究方向为人工智能安全；E-mail: qijiayin@139.com

资助项目：中国工程院咨询项目“国家级大模型监管保险箍模式研究”(2025-XZ-08)，“广东省人工智能大语言模型的安全合规监管战略研究”(2024-GD-04)；教育部哲学社会科学重大课题研究项目(24JZD040)；国家自然科学基金项目(72293583, 72293580)

本刊网址：sscae.engineering.org.cn

primary dimensions— "internal safety" and "application security" —providing a unified theoretical foundation for a systematic governance framework. Methodologically, the study introduces "Risk Label Cards" as a standardized tool and employs an "Artificial Intelligence + Human Expert Collaboration" approach to structurally analyze real-world security incidents. Combined with an improved DREAD (damage, reproducibility, exploitability, affected users, discoverability) risk matrix model, it establishes a complete assessment methodology that spans from qualitative identification to quantitative grading. The research culminates in the construction of a systematic risk classification system and a three-tiered (high, medium, low) risk landscape covering major risk types. The "dual-dimensional driven" risk analysis and governance framework constructed in this study provides a systematic theoretical tool for the precise assessment and governance of LLM risks, effectively bridging the "theory-practice gap" in governance. Furthermore, with its theoretical compatibility and dynamic characteristics, the framework provides a reference for continuously tracking and understanding the evolution of LLM security risks and for security policy research.

Keywords: large language model; security risk; security governance; risk assessment; classification and grading; risk landscape

一、前言

大语言模型（LLM）是以文字为核心交互模态，能够生成文本、图像、音频甚至视频内容的大型神经网络模型[1]，主体架构通常基于Transformer及衍生变体，通常在数百万至数万亿个词元上进行预训练[2]，具有强大的自然语言理解、生成与推理能力，在推动科学发现、促进经济增长、革新社会服务等方面展现出巨大潜力[3~6]。然而，LLM技术能力的飞跃发展在带来机遇的同时，也催生复杂且严峻的安全风险；相关风险已不再是孤立的技术故障，而是呈现风险涌现[7]、反尺度现象[8]等新特征，明显超越技术层面，对国家安全、社会稳定、伦理规范构成系统性挑战[9]。

在此背景下，构建有效的LLM治理框架虽已成为全球共识，但具体的实现路径正在经历深刻的分化与调整。LLM治理格局的动荡凸显了根本性挑战：在顶层规则尚在演变之时，学术界、产业界迫切需要科学、稳定、可操作的风险分析"通用框架"。这一"通用框架"至今尚未形成，根源在于现有的研究与实践存在局限性：风险认知碎片化，多从单一技术或社会视角出发，缺乏统一的元理论基础，导致研究成果难以整合比较[10~12]；治理实践存在"理论–操作鸿沟"，宏观的监管原则虽已确立，但普遍缺乏能够衔接高层次理念与具体化技术实践的标准化和量化评估工具，致使精准、动态的"比例性监管"难以落地[7,13,14]。

本文着力构建融合风险机理分析、量化评估、治理实践的LLM双维风险分析与治理框架，以跳出单一风险分析的局限，从理论、方法、实践层面全面推动LLM治理能力建设。基于"社会–技术"系统理论，创新性地提出区分内生安全、应用安全的双维分析视角，为系统性理解LLM风险的二元触发机制提供统一的理论根基。建立从定性识别到定量分级的完整评估方法论，通过"风险标签卡片""人工智能（AI）+人类专家协同"评估方法对523个真实案例进行结构化分析，运用改进的DREAD风险矩阵模型开展量化评估。基于LLM治理框架的分析结果，构建覆盖主要风险类型的风险图谱，提出"双维驱动"的系统性应对策略，弥合从风险评估到精准治理的实践鸿沟。

二、大语言模型安全风险治理的理论基础与分析框架

LLM是复杂的"社会–技术"系统，开展相应的安全风险治理需以可揭示内在机理并能够指导实践的科学框架为前提[15,16]、多学科理论为支撑[17,18]。传统的风险管理方法多因视角单一或理论深度不足而难以有效应对。本研究构建了双维风险分析与治理框架，为进一步开展风险分类分级并提出应对策略提供理论支撑。

（一）全球治理实践的演变与挑战

分层分域的风险监管被普遍认为是平衡创新与安全的最佳路径[19]。然而，各国的治理路径正在呈现分化态势，反映了从宏观原则到具体操作时面临的挑战。美国最新的治理路径属于复杂的分层分域体系，采取了精准的风险分级：将美国国家标准与技术研究院提出的《人工智能风险管理框架》[20]从强制标准降级为行业自愿指南，放松对通用风险的约束；在涉及国家安全的少数高风险领域，监管力量仍保留强力干预；在制度上为各州独立立法保留"留白"，形成了联邦与地方分域治理的格局。欧盟

实施的《人工智能法案》是国际上强监管的标杆[21]，然而面对产业界对严监管抑制创新的担忧，欧盟委员会采取了“原则不退、执行灵活”的策略：承认作为合规参照的“行为准则”实施时间点可协商，关键技术标准无强制的时间表，为企业争取了缓冲期与发展空间[22]，这标志着欧盟从合规优先转向创新与安全并重的调整。中国稳步构建自成体系的治理框架，以《生成式人工智能服务管理暂行办法》[23]为关键节点，将分类分级监管理念明确应用于生成式AI服务领域，加快研制一系列国家标准，逐步形成覆盖宏观原则、领域立法、技术标准的本土化和系统性治理路径。

美国分层分域的创新导向监管、欧盟“执行灵活”的适当妥协、中国系统性的治理框架，均逐步收敛至基于风险的分类分级监管理念[24]。尽管如此，全球多元化的实践路径凸显了深层次的学术挑战：缺乏具有统一性、可兼容并蓄不同治理逻辑的元理论分析框架，导致风险认知难以跨越国界进行有效的比较与整合。这也揭示了“政策–学术”接口问题：宏观政策层面已经明确“做什么”，但对于“如何做”仍需学术界提供坚实的方法论支撑。

（二）大语言模型安全风险框架研究现状

当前，LLM安全治理在宏观政策、微观学术层面均面临碎片化、“理论–操作鸿沟”等核心挑战。这些困境的产生并非源于研究的缺失，而是缺乏能够统摄全局的元理论框架。在研究路径上，表现出来源导向、结果导向的分野：前者的研究致力于追溯风险的技术成因，沿着数据、算法、模型的技术生命周期，探究漏洞和缺陷的产生机制[25~31]；后者的研究聚焦于风险发生后表现出的具体危害形态，关注对社会公平、群体权益、文化价值造成的实际影响[32,33]。两条研究路径各自发展，形成了服务于工程化的技术修复、服务于伦理与政策层面的宏观规制等独立的话语体系；两者之间缺乏有效的连接，导致单一路径视角下对风险的理解仍是片面的，而无法将风险的技术根源、社会后果完整地联系起来。

两条研究路径的分野，更深层次的原因在于对风险的理论认知，即不同的视角会揭示或遮蔽风险图景的不同方面[34]。在技术决定论视角下，将风险视为技术特性的内生产物，强调技术本身的固有风险[35]；在社会建构论视角下，强调权力关系、文化规范、公众感知等社会因素对风险的塑造作用[36]。两种截然不同的哲学立场，固化了研究者、决策者认知风险的视角，阻碍了整合性风险治理模型的建立。

（三）大语言模型安全系统论视角的理论基础

为了应对LLM安全治理面临的碎片化、“理论–操作鸿沟”等挑战，本研究从系统论视角出发，跳出单一学科的约束，融合了三方面核心理论，构建了界定风险本质、剖析生成机理、确立管控范式的LLM安全治理分析框架。“社会–技术”系统理论界定了本研究中风险的本质[37]，认为任何技术都并非孤立存在，而是深度嵌入由组织结构、人类用户、法律规范、文化价值等构成的复杂社会环境。LLM是典型的“社会–技术”系统，相应风险的生成与演化，本质上是技术子系统（算法、数据）与社会子系统（用户、组织）之间持续且动态交互的产物。将LLM风险视为“社会–技术”问题，是开展科学分析的本体论前提。这一视角从根本上超越了技术决定论[35]、社会建构论[36]的对立，确立了任何有效的风险分析都必须将技术根源、社会后果置于同一框架下进行综合考量的根本原则。

社会系统理论为剖析风险的生成机理提供了核心工具[38,39]，认为任何复杂系统都通过“自创生”来维持自身运作，通过“自我指涉”与环境进行区分。在这一过程中，风险并非外部强加的意外，而是系统为了维持自身结构、降低外部复杂性而作出决策必然伴随的“可能损害”。将社会系统理论应用于LLM，有助于揭示风险源于内部自组织、外部互动两种截然不同而又相互关联的生成机理，可识别出两种风险生成路径。① 源于系统内部的“技术自创生”。LLM通过持续的预训练、微调、对齐，不断地自我生成并重构内部的知识与能力结构。在这一过程中，数据偏差、算法不确定性、优化目标之间的内在冲突等，必然会内生性地产生幻觉、偏见等非预期的副产品，构成了风险的第一个来源（内生安全问题）。② 源于系统与环境的结构耦合。LLM作为开放系统，需要与外部环境（如用户、其他系统、信息源）进行持续的交互。在这一过程中，外部行动者可以利用与系统的交互规则（如应用程序编程接口（API）、提示词）来触发、利用甚

至放大系统的内部漏洞，构成了风险的第二个来源（应用安全问题）。

Safety-I、Safety-II范式为风险的管控实践确立了二元范式[40]。Safety-I范式重在“防止已知问题再次发生”，关注失效分析和消除缺陷；作为一种反应式、防御性的安全思维，旨在识别并阻断已知且来自外部的威胁，与应用安全风险的特性（由外部主体主动发起）契合，为相应治理提供了“御敌于外”的核心方法论。Safety-II范式重在“确保在各种条件下都能成功运作”，关注系统适应并保持正常功能；作为一种前瞻式、韧性导向的安全思维，旨在提升系统在面对不确定性时的内在鲁棒性，与内生安全风险的特性（源于系统内在的复杂性与不确定性）契合，为相应治理提供了“强基固本”的核心方法论[35]。

（四）“双维驱动”的风险分析与治理框架

本研究的核心任务并非简单地对风险进行罗列，而是解构LLM安全风险机理的复杂性，建立超越单一视角的整合性分析框架。LLM作为典型的“社会–技术”系统，相应风险的生成与演化必然发生在两个场域：系统内部的自创生过程，系统与外部环境的互动过程。据此推演出“双维驱动”的风险分析与治理框架的核心维度。

从系统内部看，风险是内生的。社会系统理论揭示，LLM这类复杂系统在通过自我迭代（技术自创生）维持自身运作时，将不可避免地产生非预期的伴生产物。为此，将内生安全定义为风险分析与治理框架的第1个核心维度，专门用于分析和应对源于模型自身设计、数据或算法不确定性引发的固有风险（如幻觉、偏见）[26,41]。这些风险并非简单的缺陷所致，而是系统内在复杂性的体现。Safety-II的韧性工程思维提供了最佳的治理范式，不追求完全消除不确定性，而是致力于提升系统在各种条件下的适应能力和运行能力。

从系统外部看，风险是互动的。“社会–技术”系统理论强调，技术风险在系统与环境的互动中被触发和放大。外部行动者（无论是善意用户还是恶意攻击者）与系统的交互是风险生成的关键来源之一。为此，将应用安全定义为风险分析与治理框架的第2个核心维度，专门用于分析和应对因外部主体的恶意利用、错误使用或环境变化引发的外部风险（如对抗性攻击、提示词操纵等）[42]。Safety-I的传统安全思维提供了治理范式，关注“防止已知问题发生”，以建立屏障、识别威胁的方式来抵御外部冲击。

基于理论推演，本研究构建了“社会–技术”双维驱动的LLM安全风险分析与治理框架，将LLM安全风险解构为内生安全、应用安全两大核心维度：前者涉及大模型的“体质”问题，决定LLM是否可靠、向善；后者反映LLM的“免疫力”问题，决定LLM能否抵御外部侵害；两方面并非孤立，而是存在相互影响。该框架将LLM安全治理视为闭环流程：通过双维视角对风险进行归因和分类，再运用相应的评估方法进行分级，最后基于评估结果实施“强基固本”（内生安全）、“御敌于外”（应用安全）的治理策略，最终实现LLM风险的系统性、精准化管控。

三、大语言模型安全风险分析与评估方法论

在应用“双维驱动”的LLM安全风险分析与治理框架、开展相应风险的定性归因后，需要构建遵循“风险识别–风险分类–风险分级”逻辑的系统性方法论，进行LLM安全风险的分析与评估（见图1）。引入“风险标签卡片”作为标准化分析单元，采用“AI+人类专家协同”范式，从海量且非结构化的风险信息案例信息中归纳出清晰的分类体系。运用改进的DREAD风险矩阵模型，将对各类风险的处理从定性识别提升到定量分级的层级。

（一）基于风险特征要素提取的案例结构化

LLM安全风险的产生与演化，与技术组件内在特性、外部社会环境因素的动态耦合密切相关。这些风险并非孤立存在，而是贯穿于LLM的“数据准备–模型训练–部署应用–运行维护”生命周期。采用动态演进的两阶段方法来构建分析基础，以避免先入为主的偏见，确保分析的全面性与客观性：系统回顾重要文献，构建全面且客观的风险特征分析框架；对来源于学术前沿、真实世界的混合案例集，应用该框架进行标准化、结构化处理，通过双向反馈机制支持框架的动态更新与自我完善。具体地，针对新增的典型案例进行全流程再分析，驱动该框架的“自下而上”式验证与重构；在后续环

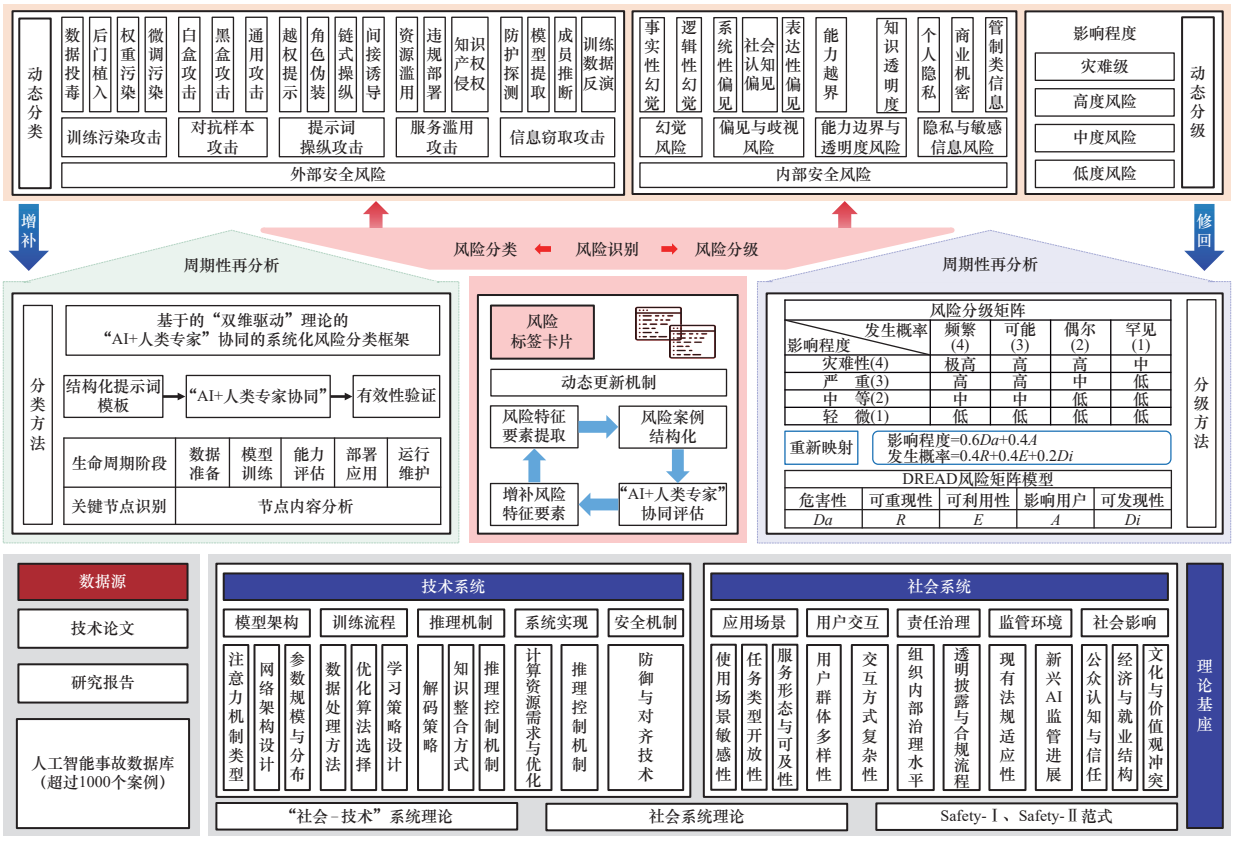

图1 “社会–技术”双维驱动的LLM安全风险分析与治理框架

节，许可专家根据分析中出现的新认识，直接对风险特征要素进行即时增补或修正，驱动该框架的“自上而下”式敏捷更新。

1. 系统构成解析与风险特征要素提取

借鉴“社会–技术”系统理论，将LLM系统解构为相互作用的技术系统、社会系统两个层面：前者涵盖LLM自身构建及运行依赖的软硬件与流程；后者包括LLM应用的外部环境、人类参与者以及相关的组织与制度因素。

为了从理论层面出发，系统地识别技术系统、社会系统的核心风险特征要素，采用文献计量、内容分析相结合的方式，构建了“文献筛选–要素提取–特征归纳”分析流程。① 构建了系统化的文献检索策略。在Scopus、Web of Science、ArXiv等文献数据库中，使用LLM、Artificial Intelligence、Artificial General Intelligence，risk、security、safety、attack、vulnerability等关键词组合进行检索，全面梳理LLM相关的风险量化、分析、攻击、防御等细化方向研究进展，重点关注2020—2025年发表在神经信息处理系统大会（NeurIPS）、人工智能促进会（AAAI）人工智能会议、国际学习表征会议（ICLR）等计算机学术会议的高被引论文，最终筛选出219篇核心文献（包括会议论文121篇、期刊论文48篇、预印本论文32篇、技术报告18篇）。② 针对219篇核心文献进行系统编码，采用频次统计、聚类分析相结合的方法，提取并归纳了LLM安全风险的直接技术诱因、重要的社会调节变量（见表1）。分类整理这些特征，进一步提炼出10个核心分析维度，为结构化地审视每个风险案例提供了统一的视角。

2. 基于“风险标签卡片”的案例结构化

构建了多元化的案例库作为实证分析的主体，确保数据来源的全面性与代表性。在219篇核心文献以外，深度整合了来自人工智能事故数据库、其他公开技术报告中的真实安全事故案例[31,43]。从10个核心分析维度出发设计了“风险标签卡片”，为每个独立且真实的AI风险案例信息提供标准化的记录模板（见表2），以避免信息遗漏、降低主观偏差，为后续分析提供坚实的数据基础，达到从海量的文献报告和真实事故案例中系统化、结构化地提

表1　LLM安全风险的特征要素

系统层面	分析维度	核心分析要素	主要审查内涵
技术系统	模型架构	注意力机制类型 网络架构设计 参数规模与分布	审查LLM结构、参数、多模态设计等选择中存在的固有缺陷与不稳定性
	训练流程	数据处理方法 优化算法选择 学习策略设计	审查训练数据质量、标注、对齐策略以及训练过程中的稳定性问题
	推理机制	解码策略 知识整合方式 推理控制机制	审查解码策略、推理能力以及可能出现的幻觉、指令误解等核心算法与行为缺陷
	系统实现	计算资源需求与优化部署架构选择	审查推理引擎性能、API接口安全、部署环境配置以及系统运维的可靠性
	安全机制	防御与对齐技术	审查内容过滤、安全护栏、偏见缓解、隐私保护等内置防御与对齐措施的有效性及鲁棒性
社会系统	应用场景	使用场景的敏感性 任务类型的开放性 服务形态与可及性	审查模型因应用领域（如医疗、金融）或任务的特殊性、复杂性、能力不匹配而产生的风险
	用户交互	用户群体的多样性 交互方式的复杂性	审查交互方式（如提示词工程）、用户认知偏差、恶意攻击（如越狱）、反馈机制的有效性
	责任治理	组织内部治理水平 透明披露与合规流程	审查开发运营机构的内部AI伦理、风险管理、安全开发流程、责任分配机制
	监管环境	现有法规的适应性 新兴AI监管的进展	审查现有法规的适用性、新兴AI立法与行业标准进展、算法备案要求、跨境合规等
	社会影响	公众认知与信任 经济与就业结构 文化与价值观冲突	审查更广泛的社会结构性影响，如就业、数字鸿沟、公众信任（虚假信息）、文化冲突等对核心伦理原则的冲击

取有效风险信息的目的。

3.“AI+人类专家协同”评估方法

设计了“AI+人类专家协同”评估方法，由多个LLM组成“AI专家团队”和人类专家一起对潜在的风险源进行识别和归纳，以高效率、高质量地完成“风险标签卡片”填充和评估。在模型选择上，参考HuggingFace Open LLM Leaderboard、LMSYS Chatbot Arena Leaderboard、SWE-bench、LiveCodeBench等主流评测榜单的综合排名，构建了包括深度求索DeepSeek R1、阿里巴巴Qwen3等6个LLM在内的多层次、互补性“AI专家团队”。

协同评估流程分为3个阶段，依次递进。①“AI专家团队”初判阶段，利用基于少量示例的结构化提示词模板，引导“AI专家团队”从风险识别、成因分析、影响评估、防控建议4个维度提取信息，通过语义一致性、置信度校准、信息约束、跨样本一致性4项验证规则，体现质量控制效果。② 语义共识校验阶段，自动检验“AI专家团队”意见的一致性。将“AI专家团队”对同一案例的输出文本进行向量化，计算余弦相似度以确定每个维度的共识分数，分数最高的文本作为此维度的共识候选；若分数低于预设阈值，说明LLM之间存在重大分歧，该案例标记为“待复核”进入下一阶段。③ 专家复核阶段，邀请3名具有AI安全或算法伦理背景、从业年限大于2年的领域专家，仅对“待复核”卡片独立进行人工审阅与修订。对于存在分歧的评估项，专家需阐述各自的判断理由并溯源关键信息；对于极少数无法达成一致的评估项，额外引入2名专家进行最终的裁定。在复核过程中，一旦发现当前的风险特征框架无法准确描述或覆盖新的风险现

表2 “风险标签卡片”信息结构

主要类别	信息类别	内容要求 / 分析维度	示例
案例基本信息	案例编号	唯一标识码	RC-20230415-007
	案例名称	简要描述	ChatGPT“DAN”提示词注入绕过
	特征简述	风险的具体表现形式	用“DAN”角色指令绕过审核，生成违规内容
	相关案例	具体的风险事件案例	RC-20241205-042、RC-20250211-017
	其他信息	参考来源等	—
	风险系统	内生安全 / 应用安全	应用安全
风险评估信息	触发条件	触发风险的关键因素	输入“DAN”“do-anything-now”等角色指令，无需特殊技术背景，可在标准的ChatGPT界面中直接操作
	影响范围	风险波及的范围	任何用户都可能遭遇或使用此类攻击方法，攻击方法通过社媒快速传播
	发现难度	被监控或审计捕获的难易程度	容易识别“DAN”等明显关键词，变体较难自动检测，人工审核可发现
	防控难度	实施防护与补救的技术门槛	需改进检测规则、上下文理解、安全微调与核心检测机制，涉及多技术环节协同
风险特征要素分析	训练流程	数据处理方法	缺乏针对提示词注入的对抗性训练样本
		学习策略设计	安全训练数据不足，安全性与指令遵循两个对齐目标而可能存在冲突
	推理机制	解码策略	解码策略中安全检查优先级低，缺乏生成过程中的实时检查
		知识整合方式	上下文理解模块将角色扮演指令误判为无害的创意写作
		推理控制机制	缺乏多轮对话中的安全策略记忆，无法识别渐进式诱导
	系统实现	部署架构选择	用户级指令的优先级可覆盖系统级安全指令
	安全机制	防御与对齐技术	API接口缺乏对输入文本的深层语义安全验证
	应用场景	使用场景的敏感性	生成的有害内容可被轻易复制并应用于高敏感性场景
		任务类型的开放性	开放域对话等任务类型为对抗性攻击提供了试探空间
		服务形态与可及性	公开的用户端界面和API，便于攻击方法的大规模、低成本传播
	用户交互	用户群体的多样性	攻击方法在技术社区进行分享和迭代，不断发现新漏洞
		交互方式的复杂性	攻击利用自然语言的复杂性和歧义性来操纵LLM
	责任治理	组织内部治理水平	暴露LLM发布前在对抗性安全测试（红队测试）方面的不足
		透明披露与合规流程	开发者未披露漏洞细节，仅通过更新修复
	监管环境	现有法规的适应性	现有法规难以直接对提示词注入行为进行定性
	社会影响	公众认知与信任	严重损害公众对AI安全可控的信任度
		文化与价值观冲突	诱导生成的言论可能与主流社会文化和价值观产生冲突

象时，可以直接对风险特征要素进行增补或修正。

完成了对比实验（100条案例）以验证协同评估流程的有效性。结果显示，与纯人工标注相比，协同评估流程的Cohen’s K系数由0.74提升至0.89，证明标签质量具有良好的一致性和可靠性。经过质量筛选，留存523张具有高置信度的“风险标签卡片”，作为后续统计与建模的样本，时间跨度为2021年1月—2024年12月。在数据集的地理分布方面，中国案例有143个（占比为27.3%），美国案例有210个（占比为40.2%），欧盟案例有93个（占比为17.8%），其他地区的案例有77个（占比为14.7%）。在数据集的行业分布方面，互联网科技案例占比为41.2%，金融服务案例占比为18.7%，教育医疗案例占比为15.9%，政府公共服务案例占比为12.8%，其他行业案例占比为11.4%。

（二）大模型安全风险分类框架的构建与验证

在对523个风险案例进行标准化、结构化识别后，再对这些数据进行归纳与提炼，据此构建LLM安全风险的分类框架。

1. 基于触发机制的风险分类框架

应用“双维驱动”的风险分析与治理框架，从风险触发源头的根本性差异出发，确立了包含内生安全风险、应用安全风险的二元主框架。内生安全

风险指源于模型系统内部的技术特性、固有缺陷或非预期自发行为导致的风险，产生过程通常无需外部恶意干预。应用安全风险指外部主体利用系统漏洞或通过交互操纵，主动发起攻击或恶意利用导致的风险，具有明确的攻击意图。

在二元框架下，再次运用“AI+人类专家协同”评估方法，对523张“风险标签卡片”中的特征描述、触发条件等字段进行相似性分析，初步识别出28个风险群组。对28个风险群组进行归因分析并提取每个群组的共同触发特征后发现，这些风险群组具有二元特性：255个案例构成的群组，共同特征为“模型在无外部恶意干预下，因自身机制而自发产生”，可归入内生安全风险维度；268个案例构成的群组，共同特征为“由外部主体主动、有意图地触发”，可归入应用安全风险维度。根据风险的具体表现机制和攻击阶段，对28种风险类型进行二次归纳，形成了9个更高层级的风险类别：在内生安全风险维度，按照风险表现形式分为幻觉风险、偏见与歧视风险、能力边界与透明度风险、隐私与敏感信息风险4类；在应用安全风险维度，按照攻击方式分为训练污染攻击、对抗样本攻击、提示词操纵攻击、服务滥用攻击、信息窃取攻击5类。

通过“自下而上”聚类归纳、“自上而下”理论归因相结合的多轮次分析，构建了包含2个维度、9个风险类别、28种具体风险类型的系统性分类框架（见表3）。

表3　LLM安全风险分类框架

风险维度	风险类别	风险类型	风险内容
应用安全风险	训练污染攻击	数据投毒	植入恶意训练样本，导致模型学习错误，产生有害行为模式
		后门植入	嵌入特定触发条件，使模型在特定情境下产生异常行为
		权重污染	直接篡改模型的权重参数，改变模型输出
		微调污染	下游任务微调时注入恶意数据，以低成本精准地污染特定能力
	对抗样本攻击	白盒攻击	在完全了解模型架构的情况下，精确构造能欺骗模型的微小输入扰动
		黑盒攻击	在不了解模型内部信息的情况下，仅通过查询构造有效攻击
		通用攻击	构造具有普适性的对抗性扰动，可攻击不同的模型
	提示词操纵攻击	越权提示	通过提示词诱骗模型执行本应拒绝、超出预设权限范围的任务
		角色伪装	通过身份欺骗实现特定目的，改变模型的行为
		链式操纵	通过铺垫和渐进的诱导，侵蚀模型的安全边界，进而生成有害内容
		间接诱导	将恶意的提示词植入到模型将要处理的第三方信息载体中
	服务滥用攻击	资源滥用	大规模并发请求消耗系统资源，导致服务不可用
		违规部署	违反服务条款或开源协议，搭建生成非法或有害内容的独立服务
		知识产权侵权	利用模型生成侵权的内容，未经授权将模型或其输出用于商业目的
	信息窃取攻击	防护探测	通过查询试探并分析模型的安全防护机制、内容审核策略
		模型提取	通过查询推断模型内部实现机制，对模型的核心功能进行逆向工程
		成员推断	通过查询判断特定数据是否出现在模型的训练集中
		训练数据反演	通过查询反向推导和重构模型训练过程中使用的原始数据内容
内生安全风险	幻觉风险	事实性幻觉	输出与客观事实不符的内容，在专业领域可能导致严重后果
		逻辑性幻觉	推理过程出现自相矛盾，影响决策的可靠性
	偏见与歧视风险	系统性偏见	对特定人群产生系统性歧视，影响关键决策公平性
		社会认知偏见	在内容生成中固化或放大关于职业、性别、文化等的社会刻板印象
		表达性偏见	语言表达中对不同群体使用带有不当褒贬色彩或情感偏向的词汇
	能力边界与透明度风险	能力越界	处理超出能力范围的任务时，未明确声明局限性，给出误导回答
		知识透明度	对于模型输出的结论，无法提供有效的来源追溯、证据支撑或逻辑解释
	隐私与敏感信息风险	个人隐私	在生成内容中无意泄露个人身份信息、医疗记录等敏感数据
		商业机密	泄露企业的核心技术、研发数据、客户名单等非公开商业信息
		管制类信息	生成或泄露涉及国家安全、关键基础设施等非公开信息

2. 风险分类框架的有效性验证

采用多维度量化评估方法验证了LLM安全风险分类框架。验证过程中应用完整性、区分度、实用性等维度的评估指标（见表4），再通过样本分析、专家评估、理论分析等方法进行综合验证。为了确保评估的独立性，另外邀请了3位独立专家应用“AI+人类专家协同”评估方法分析7个AI真实应用案例，重点关注系统运营情况、安全事件记录、媒体公开报道、系统使用观察等（见表5）。验证结果表明，在完整性维度上，风险覆盖率>85%；在区分度维度上，分类准确率>87%，各个类别边界清晰、重叠度低；在实用性维度上，操作便捷性、防控指导性获得良好评价；评估指标的专家一致性系数接近或超过0.8。可见，LLM安全风险分类框架具有良好的完整性、区分度、实用性，可为LLM安全风险管理提供理论基础与应用参考。

表4　风险分类框架的评估指标

维度	评估指标	评估方法
完整性	分类层次完整性	理论分析
	场景适应性	专家评估
	风险覆盖率	样本分析
区分度	类别独立性	理论分析
	边界清晰度	专家评估
	分类准确率	样本分析
实用性	操作便捷性	专家评估
	防控指导性	专家评估
	推广适用性	专家评估

表5　风险分类框架的验证结果

维度	评估指标	评估结果	专家一致性
完整性	分类层次完整性	4.0/5	0.82
	场景适应性	3.8/5	0.78
	风险覆盖率	4.2/5	0.85
区分度	类别独立性	4.1/5	0.83
	边界清晰度	3.9/5	0.80
	分类准确率	4.3/5	0.87
实用性	操作便捷性	4.05	0.81
	防控指导性	3.9/5	0.79
	推广适用性	4.1/5	0.82

（三）基于DREAD风险矩阵模型的大语言模型安全风险分级方法

1. DREAD风险矩阵模型

单一维度的风险分级可能忽略某些关键风险，降低分级结果的准确性[44]；可同时考虑多个风险维度进行风险分级，构建风险矩阵来综合确定风险等级[45]。风险矩阵从影响程度、发生概率两个维度进行风险评估，可以有效确定风险等级；影响程度分为灾难性、严重、中等、轻微4个等级，发生概率分为频繁、可能、偶尔、罕见4个层次[16]。DREAD模型在复杂系统风险评估方面具有实用性[46]，具有更为细化的风险评估维度[47]，主要从危害性（*Da*）、可重现性（*R*）、可利用性（*E*）、影响用户（*A*）、可发现性（*Di*）5个维度对复杂系统风险进行综合评估。

传统的风险矩阵、DREAD模型直接应用于LLM安全风险评估时都存在一定的局限性。风险矩阵虽然直观易用，但二维评估框架过于简单，难以充分反映LLM风险的多维特征，对影响程度、发生概率的判定缺乏系统性指标支撑。DREAD模型虽然提供了全面的评估维度，但最终得分难以直观反映风险等级，各维度权重统一的假设也可能导致评估结果偏离实际。

将风险矩阵、DREAD模型相结合的改进风险评估方法称为DREAD风险矩阵模型，既维持了风险矩阵能够反映系统中整体性的特征，又保留了DREAD模型可以体现层次性和关联性的特征。发展了DREAD风险矩阵模型：建立DREAD评分到风险矩阵参数的映射机制，实现两种方法的优势互补；影响程度由*Da*、*A*两个维度决定，发生概率由*R*、*E*、*Di*三个维度综合计算；各维度的权重系数由专家评估小组通过层次分析法确定，各维度之间的相对重要性由1~9标度法两两比较后确定。据此构建了判断矩阵（见表6）。

表6　判断矩阵

判断矩阵	*Da*	*A*	*R*	*E*	*Di*
Da	1	1.5			
A	0.67	1			
R			1	1	2
E			1	1	2
Di			0.5	0.5	1

经计算，判断矩阵的一致性比率均小于0.1，表明专家判断具有高度的逻辑一致性。最终确定的权重配比如下：影响程度=0.6*Da*+0.4*A*，发生概率=0.4*R*+0.4*E*+0.2*Di*。再将计算得到的影响程度、发生概率映射到风险矩阵中，得到最终的分类结果（见表7）。

在DREAD加权分到风险等级的映射关系中，将影响程度、发生概率的评分上限设定为9，表示在当前的技术和应用场景下，LLM系统在可预测、可管理范围内可造成的最严重后果或者具有最高发生的可能性。当前阶段的LLM仍在受控环境下运行，相应的行为模式、影响范围可预测，故高、中、低3级分类更符合实际观察到的风险表现，可为风险防控提供明确的操作指引。灾难性风险具有极低概率、极高影响、因果链条复杂、深刻不确定性等特征，通常与其他环境系统组件深度耦合，需要领域专家结合具体环境进行专门评估；相应的评估与管理已超越常规监管范畴。

2. DREAD指标量化评分标准

在保持DREAD模型赋分1~9的基础上，针对LLM应用特点，针对每个维度制定了详细的评分标准（见表8）。

3. 分级方法的有效性验证

选取1个公开发布、具有代表性的LLM及算法安全事件（AI搜索引擎的药物信息误导风险）进行分析，以全面展现DREAD风险矩阵模型应用的有效性、准确性。研究人员向集成在搜索引擎中的AI聊天机器人询问关于50种最常用处方药的问题时，发现回答存在严重风险：在生成的500个答案中，只有54%的与科学共识一致。专家评估结果显示，42%的回答可能导致中度或轻度伤害，22%的回答可能导致严重伤害甚至死亡[48]。

应用DREAD指标量化评分的结果及解释如下。① *Da*=9。22%的错误回答可能导致严重伤害甚至死亡，此类直接威胁用户生命安全的损害潜力评分极高。② *R*=7。相关研究是系统性的（针对50种药物提10个问题），表明错误并非完全随机，而是可以系统性地复现；尽管LLM具有随机性，但相应的风险模式是稳定的。③ *E*=8。利用此风险无需专业技能，任何普通患者或用户只需在搜索引擎中提出常见问题即可触发风险。④ *A*=9。风险存在于全球的主流搜索引擎中，潜在受影响的用户规模极大，触及所有在线查询健康信息的人。⑤ *Di*=6。聊天机器人的回答通常语言流畅、格式专业，普通用户很难分辨出相关信息的真伪；发现风险需要专业的医学知识和交叉验证，对普通患者而言难度较高。计算可得：影响程度=0.6×9+0.4×9=9（严重），发生概率=0.4×7+0.4×8+0.2×6=7.2分（频繁）。查阅风险矩阵可知，“严重影响、频繁概率”对应着高风险，表明这是1个紧迫、大规模的公共安全问题。

DREAD风险矩阵模型既适用于LLM安全风险的整体评估，也可用于具体风险点的精确分析，为建立分类分级监管体系提供了有效工具。应用优势体现在：提供科学的评估框架，具有差异化权重配比并可进行动态调整，提高对新兴风险的适应能力；保持评估结果的直观性，便于监管实践中的快速决策；建立系统的评估流程，确保评估过程的规范性和可重复性。

表7　风险矩阵对应关系

发生概率 影响程度	频繁 （7, 9］	可能 （3, 7］	偶尔 ［1, 3］
严重（7, 9］	高	高	中
中等（3, 7］	中	中	低
轻微［1, 3］	低	低	低

表8　DREAD指标量化评分标准

量化维度	高等级（7, 9］	中等级（3, 7］	低等级［1, 3］
Da	造成严重、不可逆或广泛的伤害	造成显著、可控但修复成本高的伤害	造成轻微、局部或易于修复的伤害
R	风险可被稳定、轻易地复现	风险可在特定条件下复现	风险难以稳定复现
E	任何普通用户均可利用	需要特定专业知识	需要专家级技能或大量资源
A	影响绝大多数或关键用户群体	影响特定的、较大规模的用户群体	影响是孤立的、个体性的
Di	风险在常规使用中即可轻易暴露	需要系统性审计或专业知识才能发现	风险高度隐蔽，常规手段难以触及

四、大语言模型安全风险的分类分级

（一）大语言模型内生安全风险的分类分级

LLM内生安全风险主要源于自身技术架构和训练机制的固有特性，相关风险与模型的基础能力直接相关；通常体现出高频、中低危害的特征，即发生的频率较高，但在大部分场景下可通过有效措施控制在一定范围内。基于风险的性质及潜在影响，LLM内生安全风险可划分为幻觉、偏见与歧视、能力边界与透明度、隐私与敏感信息4类（见表9）。

1. 幻觉风险

幻觉是LLM最显著的风险类型，指模型生成的内容与事实不符（事实性幻觉）或者产生自相矛盾的内容（逻辑性幻觉）。从DREAD风险矩阵模型来看，相关风险主要由*Da*驱动。在关键决策时，事实性或逻辑性错误可能造成不可逆的后果，导致高的*Da*；LLM输出的文本通常流畅且自信，该风险的*Di*对普通用户而言往往较低。这两个维度的组合，致使影响程度方面的评级天然偏高。为此，高、中、低3级风险主要依据*Da*的潜在量级进行划分：构成严重危害的（如危及生命财产安全），评为高风险；可能影响重要决策但后果可控的（如技术参数错误），评为中风险；*Da*和*A*均很小的（如日常对话中的非关键事实错误），评为低风险。

2. 偏见与歧视风险

偏见与歧视风险源于训练数据的偏见、算法机制的放大效应，表现为模型输出中针对特定群体的不公平对待或歧视性表达。从DREAD风险矩阵模型来看，相关风险具有高的*Da*、高的*A*。系统性偏见可能固化乃至加剧社会不公，对应的*Da*极高；广泛应用的模型出现此类风险时的*A*也很高；系统性偏见通常是隐性的，*Di*较低。为此，凡是涉及关键资源分配（如信贷、司法）的系统性偏见，均被评定为高风险。社会认知偏见因其*Da*有限但具备高的*R*，构成中风险。表达性偏见因其*Da*低而划分为低风险。

3. 能力边界与透明度风险

能力边界与透明度风险指LLM与用户交互过程中因自身能力的认知、表达、行为方式而引发的风险，与模型生成内容的真实性与准确性无关，以低的*Di*为代表（模型表述过度自信，用户难以判断模型是否在“不懂装懂”）。该风险的最终等级取决于模型在应用场景下可能造成的最大*Da*。在

表9　LLM内生安全风险的分类分级

风险类别	风险类型	高风险	中风险	低风险
幻觉风险	事实性幻觉	关键领域（如医疗、金融、法律）中的严重事实错误	技术参数错误、教育培训错误	历史日期错误、地理位置错误
	逻辑性幻觉	逻辑（如系统设计、安全协议）自相矛盾	业务流程、政策解读存在推理偏差	对话的上下文不连贯
偏见与歧视风险	系统性偏见	影响关键决策，损害特定群体权益（如医疗、就业）	在非关键推荐中系统性地偏向特定群体	—
	社会认知偏见	在公共教育材料中固化有害的刻板印象	在一般内容创作中产生刻板印象（如职业关联）	在创意写作中偶然出现轻微的刻板印象
	表达性偏见	—	在正式文书中对不同群体使用带情感偏见的词汇	在非正式对话中偶然使用带褒贬色彩的词汇
能力边界与透明度风险	能力越界	关键领域（如医疗、法律）中的错误建议	商业决策、技术建议超过能力范围	日常咨询能力误判
	知识透明度	关键信息（如公共安全、医疗）来源不清	学术研究、技术咨询信息不透明	一般知识传播模糊描述
隐私与敏感信息风险	个人隐私	生物特征、医疗记录泄露	收入、消费习惯泄露	兴趣爱好泄露
	商业机密	核心技术、研发数据泄露	经营数据、客户信息泄露	业务流程信息泄露
	管制类信息	国防、关键基础设施信息泄露	受限行业数据泄露	基础技术规范泄露

DREAD风险矩阵模型中，当低*Di*的风险与高潜在*Da*的场景结合时，影响程度的评级会很高。

4. 隐私与敏感信息风险

隐私与敏感信息风险指LLM在生成内容过程中可能泄露或不当处理敏感信息。这类泄露通常是随机且难以稳定触发的，因而具有低的*R*、*E*。在DREAD风险矩阵模型中，尽管发生概率的评级可能不高，但泄露信息本身的敏感度直接关联高的*Da*，决定了影响程度评级必然处于高位（至少评定为中风险）。

（二）大语言模型应用安全风险的分类分级

LLM应用安全风险指由外部攻击者主动干预或操纵触发，导致模型产生危害性、误导性或违规性输出的风险。与内部风险不同，应用安全风险通常呈现低频、高危害的攻击模式，根据攻击手段可将应用安全风险分为训练污染攻击、对抗样本攻击、提示词操纵攻击、服务滥用攻击、信息窃取攻击5类（见表10）。

1. 训练污染攻击风险

训练污染攻击指攻击者干预模型训练或微调过

表10　LLM应用安全风险的分类分级

风险类别	风险类型	高风险	中风险	低风险
训练污染攻击	数据投毒	触发违法内容，导致系统性认知偏差	影响特定场景判断	轻微性能衰减，降低生成质量
	后门植入	预设隐藏指令触发违规内容	影响特定领域的输出安全性	轻微行为异常
	权重污染	关键任务领域参数被篡改，影响决策	影响部分领域的准确性	轻微的模型性能漂移
	微调污染	被恶意微调，产生偏见或错误信息	影响局部推理能力	轻微影响模型输出
对抗样本攻击	白盒攻击	精准构造对抗样本，使模型绕过安全审核，输出违规内容	影响部分任务	轻微输出偏差
	黑盒攻击	无需访问模型内部，仅通过试探性输入诱导违规输出	影响特定功能	轻微影响模型响应
	通用攻击	适用于多种场景的攻击方式，如通用对抗样本库	影响多个任务	轻微影响对抗鲁棒性
提示词操纵攻击	越权提示	通过巧妙构造的输入绕过安全限制，生成敏感内容	部分越权	轻微行为偏差
	角色伪装	模拟特定角色诱导用户信任并输出虚假信息	影响特定领域的可信度	轻微误导
	链式操纵	通过连续多轮对话诱导模型偏离正常行为，输出敏感信息	影响模型安全性	轻微影响对话一致性
	间接诱导	通过外部信息影响模型的推理过程	影响模型判断	轻微影响推理稳定性
服务滥用攻击	资源滥用	规模化API滥用，导致服务中断或资源枯竭	影响部分用户的访问体验	轻微性能波动
	违规部署	未经授权的模型使用，违反使用协议	影响合法性与合规性	轻微违反使用条款
	知识产权侵权	未经许可使用受版权保护的内容进行训练或生成	影响模型合规性	轻微内容侵权
信息窃取攻击	防护探测	识别并绕过模型的安全防护机制，获取内部信息	影响部分安全策略	轻微的信息泄露
	模型提取	通过大量查询获取模型参数，重建类似模型	影响商业竞争力	轻微降低模型独特性
	成员推断	确定特定训练数据是否包含某些用户的个人信息	影响数据隐私	轻微推断隐私信息
	训练数据反演	通过模型输出推断训练数据内容，重构部分训练样本	影响数据机密性	轻微影响数据安全

程，植入特定的有害行为模式，从根本上改变模型的基础能力，相应危害具有系统性和持续性。训练污染攻击包括数据投毒、后门植入、权重污染、微调污染等类型。攻击效果固化在模型权重中，常规手段难以察觉，相应风险等级的判定主要依据较高的*Da*、较低的*Di*相互作用的结果，通常影响程度评级较高。可产生系统性认知偏差或稳定触发违规内容的高*Da*攻击，均评为高风险；若*Da*相对可控，则评为中风险；若仅导致轻微性能衰减，则评为低风险。

2. 对抗样本攻击风险

对抗样本攻击指攻击者构造特制的输入内容，直接干扰模型的推理过程，相应危害具有实时性和精确性。对抗样本攻击风险等级由*Da*、*R*、*E*共同决定。成功的对抗攻击通常意味着高的*R*、*E*，在DREAD风险矩阵中会转化为较高的发生概率。当高发生概率的攻击可诱导产生严重有害输出（高的*Da*、高的影响程度）时，即评为高风险；若攻击需要特定技术门槛（中等的*E*）或者成功率有限（中等的*R*），则评为中风险；若攻击成功率、*Da*均很低，则评为低风险。

3. 提示词操纵攻击风险

提示词操纵攻击指攻击者利用模型的上下文理解机制，精心构造提示词序列，诱导模型绕过安全限制或产生有害输出，具有隐蔽性和自适应性。鉴于攻击代码（提示词）易于复制和传播，这类攻击在DREAD风险矩阵中具有极高的*E*、*R*，相应风险等级主要取决于*Da*。当低门槛、高*R*的攻击能够稳定绕过核心安全限制、造成高*Da*的违规输出时，则评为高风险；若*Da*可控，则评为中风险；若*Da*低，则评为低风险。

4. 服务滥用攻击风险

服务滥用攻击指攻击者过度使用或违规部署模型服务，对服务可用性或社会安全造成危害，相应风险等级主要由滥用行为的*A*、*Da*共同决定。在DREAD风险矩阵中，*A*、*Da*两个维度直接映射为影响程度的高低。凡是导致大规模用户服务中断（高的*A*）或造成重大经济损失（高的*Da*）的服务滥用，均评为高风险；控制在局部范围内且影响有限的服务滥用，则评为中风险；偶发性且影响及危害均很小的探索性滥用，则评为低风险。

5. 信息窃取攻击风险

信息窃取攻击指攻击者通过技术手段从LLM中提取敏感信息或推断其内部机制，通常具有低的*Di*、高的*Da*。在DREAD风险矩阵中，这类攻击的影响程度评级会很高。当攻击能系统性地获取核心敏感信息（高的*Da*）且过程难以察觉（低的*Di*）时，即评为高风险；若只能获取局部、非核心信息，或者需要较高技术门槛时，则评为中风险；若仅能获取基础模型特征，则评为低风险。

五、大语言模型安全风险治理建议

在对523个真实安全案例开展系统性分析的基础上，本研究构建了风险要素完整的LLM安全风险图谱，揭示了LLM应用安全风险的“双维”特征。更为重要的是，文中构建的不仅是静态的分类体系，更是灵活、开放的动态分析框架；具有“自下而上”“自上而下”双重更新路径、基于底层触发机理的风险划分方式，对于未来可能出现的新兴风险具有良好的理论兼容性与应用前瞻性，不仅可以有效解释当前的风险，而且能够为持续追踪和理解LLM安全风险的演化提供稳定且可靠的理论基座。此外，“社会–技术”系统发展到未来的高级阶段后可能出现超越常规监管范畴的灾难性风险，可利用该框架“自上而下”的专家驱动更新机制并结合特定情境开展专门的评估。在此工作基础上，提出如下的LLM安全风险治理建议。

（一）实施“双维驱动”的风险管控核心策略

1. 面向内生安全风险的系统性治理路径

实质性提升LLM的可靠性、安全性以及价值对齐水平，筑牢LLM安全可信的根基，有效应对源于模型内部复杂性、不确定性的内生安全风险。

幻觉、能力越界、知识透明度被识别为具有高*Da*、低*Di*的特征，是直接影响内容可靠性的风险类别，开展治理重在建立具有可信锚定与透明度的保障机制。在医疗、司法等高风险场景中，建立基于权威知识源的实时验证闭环机制，作为强制性安全基线；在新闻、信息查询等中风险场景中，提供参考信息溯源与置信度评估功能，作为增强性安全措施；在创意生成等低风险场景中，相应保障机制作为可选配置。

偏见与歧视是*A*范围宽广、具有系统性*Da*的社会性风险，治理重心在于实施全周期的公平性与价值对齐流程。在招聘、信贷等决策场景中可能出现高风险偏见，需要实施贯穿“数据–算法–输出”全链路的公平性保障，如在数据端主动开展偏见纠正、在算法端嵌入强制性公平约束、在输出端严格进行量化审计。对于通用聊天等场景下的中风险偏见，应通过主流的技术范式（如人类反馈强化学习）进行充分的价值观训练引导。对于专业领域的低风险偏见，治理重点在于防止特定的滥用。

隐私与敏感信息泄露是高*Da*、低发生概率的风险，需要针对性构建数据全生命周期隐私增强体系。将隐私保护原则贯穿数据处理全程，根据数据敏感度分级施策：处理个人健康记录等高风险数据的应用，需要采用差分隐私训练等强隐私增强技术；处理普通用户个人信息的中风险应用，应落实合规性与标准脱敏措施；不处理个人数据的低风险应用，宜部署防止模型意外生成第三方隐私信息的机制。

2. 面向应用安全风险的全周期防御矩阵

在LLM的整个应用生命周期中，形成防御外部威胁的“免疫力”，抑制LLM在部署、交互、运维过程中因与外部环境交互而产生的应用安全风险。

训练污染攻击具有低的*Di*、高的*Da*，是极具隐蔽性的高风险类别，治理重心在于开展供应链安全与模型完整性验证。应用于国家关键基础设施等高风险场景的LLM，需要源自严格、可信的供应链，对模型来源、训练数据、依赖组件等进行全面的安全审查与完整性验证。对于中风险应用，应建立标准化的数据来源审查、依赖项安全检查机制。对于低风险项目，应实施基本的代码版本控制和清洗流程。

对抗样本攻击、提示词操纵攻击发生在模型交互阶段，攻击门槛较低，具有高的*E*、*R*，治理重心在于部署深度语义感知与自适应学习的交互安全护栏。对于集成后端API调用、可执行实际操作的高风险应用，需要部署多层纵深防御手段，如深度语义分析、用户行为异常检测、严格的API权限管理。对于面向公众的中风险应用，应部署能够识别隐蔽恶意意图的语义分析引擎。对于内部使用的低风险应用，可部署基于规则和关键词的基础级防御手段。

针对发生在模型运维阶段，*A*范围宽广的服务滥用攻击、具有高的*Da*和低的*Di*特征的信息窃取攻击，需建立服务状态监测与资产保护机制。对于承载关键业务的高风险应用，应部署常态化的服务状态监控与异常分析机制，就API调用频率、查询模式等持续进行异常行为分析。对于中风险应用，应制定清晰的应急响应预案。针对所有对外服务的LLM，应推广模型水印等数字身份技术，以追溯滥用或窃取的模型资产。

（二）健全系统性的治理保障体系

在宏观制度层面，构建“法律–标准–协同”三位一体的顶层框架。加快推进AI专项立法，明确各方主体的责任与义务，特别是界定高风险场景中的责任归属；同步研制分级分类的安全技术标准、测试评估标准、管理规范等。科学聚合多方治理力量，打破管理部门、产业界、学术界、社会公众之间的协作壁垒，尽快形成权责清晰、沟通顺畅、资源共享的协同治理生态。建议行业主管部门牵头建立国家级LLM安全基础平台，为全行业提供内容过滤、风险检测、合规审计等基础性的安全护栏服务，筑牢全行业的发展安全底座。

在组织治理层面，建立与风险相适应的内部责任体系，将有效治理根植于组织的内部实践。引导企业根据面临的LLM安全风险等级，构建相应的内部治理架构。对于涉及高风险应用领域的企业，设立独立的伦理与安全监督机构，将安全责任深度融入研发流程与考核规则。对于所有LLM从业机构，将培育安全文化、参与威胁情报共享作为基础要求，共同构筑行业安全生态。

在能力建设层面，强化“技术–人才”并行的国家级支撑体系。积极发展监管科技，支持开发标准化、自动化的风险评估工具与合规测试平台，为有效治理LLM风险提供技术支撑。构建国家级攻防实验环境，将技术工具与实战演练相结合，定期组织体系化的国家级AI安全攻防演练。在实战中检验技术工具的有效性，全面培养兼具技术与治理视野的复合型人才，为LLM风险治理提供智力支撑。

利益冲突声明

本文作者在此声明不存在任何利益冲突或财务冲突。

Received date: June 15, 2025; **Revised date:** September 21, 2025

Corresponding author: Qi Jiayin is a professor from the Cyberspace Institute of Advanced Technology, Guangzhou University. Her major research fields include artificial intelligence security. E-mail: qijiayin@139.com

Funding project: Chinese Academy of Engineering preject “Security and Compliance Regulatory Strategies for Artificial Intelligence Large Language Models in Guangdong Province” (2025-XZ-08), “Research on the National Guardrails and Governance Framework for Large Model Regulation” (2024-GD-04); Major Project of Philosophy and Social Sciences Research of the Ministry of Education (24JZD040); National Natural Science Fond Project (72293583, 72293580)

参考文献

[1] Raffel C, Shazeer N, Roberts A, et al. Exploring the limits of transfer learning with a unified text-to-text transformer [J]. Journal of Machine Learning Research, 2020, 21(1): 5485–5551.

[2] Gallegos I O, Rossi R A, Barrow J, et al. Bias and fairness in large language models: A survey [J]. Computational Linguistics, 2024, 50(3): 1097–1179.

[3] Brown T B, Mann B, Ryder N, et al. Language models are few-shot learners [R]. Vancouver: The 34th International Conference on Neural Information Processing Systems, 2020.

[4] Wei A, Haghtalab N, Steinhardt J. Jailbroken: How does LLM safety training fail? [EB/OL]. (2023-07-05)[2025-10-15]. https://arxiv.org/abs/2307.02483.

[5] Kaplan J, McCandlish S, Henighan T, et al. Scaling laws for neural language models [EB/OL]. (2020-01-23)[2025-10-15]. https://arxiv.org/abs/2001.08361.

[6] Lehman J, Clune J, Misevic D, et al. The surprising creativity of digital evolution: A collection of anecdotes from the evolutionary computation and artificial life research communities [J]. Artificial Life, 2020, 26(2): 274–306.

[7] Bommasani R, Hudson D A, Adeli E, et al. On the opportunities and risks of foundation models [EB/OL]. (2021-08-16)[2025-10-15]. https://arxiv.org/abs/2108.07258.

[8] McKenzie I R, Lyzhov A, Pieler M, et al. Inverse scaling: When bigger isn’t better [EB/OL]. (2023-06-15)[2025-10-15]. https://arxiv.org/abs/2306.09479.

[9] Wang H D, Fu W J, Tang Y Z, et al. A survey on responsible LLMs: Inherent risk, malicious use, and mitigation strategy [EB/OL]. (2025-01-16)[2025-10-15]. https://arxiv.org/abs/2501.09431.

[10] Sovacool B K, Hess D J. Ordering theories: Typologies and conceptual frameworks for sociotechnical change [J]. Social Studies of Science, 2017, 47(5): 703–750.

[11] Gordon J S. Building moral robots: Ethical pitfalls and challenges [J]. Science and Engineering Ethics, 2020, 26(1): 141–157.

[12] Askell A, Brundage M, Hadfield G. The role of cooperation in responsible AI development [EB/OL]. (2019-07-10)[2025-10-15]. https://arxiv.org/abs/1907.04534.

[13] Ibáñez J C, Olmeda M V. Operationalising AI ethics: How are companies bridging the gap between practice and principles? An exploratory study [J]. AI & Society, 2022, 37(4): 1663–1687.

[14] Raji I D, Smart A, White R N, et al. Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing [R]. Barcelona: The 2020 Conference on Fairness, Accountability, and Transparency, 2020.

[15] Jobin A, Ienca M, Vayena E. The global landscape of AI ethics guidelines [J]. Nature Machine Intelligence, 2019, 1(9): 389–399.

[16] Floridi L. The European legislation on AI: A brief analysis of its philosophical approach [J]. Philosophy & Technology, 2021, 34 (2): 215–222.

[17] Rahwan I, Cebrian M, Obradovich N, et al. Machine behaviour [J]. Nature, 2019, 568(7753): 477–486.

[18] Dixon L, Li J, Sorensen J, et al. Measuring and mitigating unintended bias in text classification [R]. New Orleans: The 2018 AAAI/ACM Conference on AI, Ethics, and Society, 2018.

[19] World Economic Forum. Global risk report 2024 [EB/OL]. (2024-01-15)[2025-10-15]. https://www3.weforum.org/docs/WEF_The_Global_Risks_Report_2024.pdf.

[20] Tabassi E. Artificial intelligence risk management framework (AI RMF 1.0) [EB/OL]. (2023-01-26)[2025-10-15]. https://www.nist.gov/publications/artificial-intelligence-risk-management-framework-ai-rmf-10.

[21] de Almeida P G R, dos Santos C D, Farias J S. Artificial intelligence regulation: A framework for governance [J]. Ethics and Information Technology, 2021, 23(3): 505–525.

[22] Tarka J, Sedaei S. EU AI act update: Navigating the future [EB/OL]. (2025-07-16)[2025-10-15]. https://ogletree. com/insights-resources/blog-posts/eu-ai-act-update-navigating-the-future/.

[23] 中华人民共和国国家互联网信息办公室. 生成式人工智能服务管理暂行办法 [EB/OL]. (2023-07-13)[2025-10-15]. https://www.cac.gov.cn/2023-07/13/c_1690898327029107.htm.
Cyberspace Administration of China. Interim measures for the administration of generative artificial intelligence services [EB/OL]. (2023-07-13)[2025-10-15]. https://www.cac.gov.cn/2023-07/13/c_1690898327029107.htm.

[24] Zeng Y, Klyman K, Zhou A, et al. AI risk categorization decoded (AIR 2024): From government regulations to corporate policies [EB/OL]. (2024-06-25)[2025-10-15]. https://arxiv.org/abs/2406.17864.

[25] Wirtz B W, Weyerer J C, Kehl I. Governance of artificial intelligence: A risk and guideline-based integrative framework [J]. Government Information Quarterly, 2022, 39(4): 101685.

[26] Yampolskiy R V. Taxonomy of pathways to dangerous artificial intelligence [R]. Phoenix: AAAI Workshop: AI, Ethics, and Society, 2016.

[27] Hendrycks D, Mazeika M. X-risk analysis for AI research [EB/OL]. (2022-07-13)[2025-10-15]. https://arxiv.org/abs/2206.05862.

[28] Cui T Y, Wang Y L, Fu C P, et al. Risk taxonomy, mitigation, and assessment benchmarks of large language model systems [EB/OL]. (2024-01-11)[2025-10-15]. https://arxiv.org/abs/2401.05778.

[29] McGregor S. Preventing repeated real world AI failures by cataloging incidents: The AI incident database [R]. Virtual: The Thirty-Third Annual Conference on Innovative Applications of Artificial Intelligence, 2021.

[30] Pittaras N, McGregor S. A taxonomic system for failure cause analysis of open source AI incidents [EB/OL]. (2022-11-14)[2025-10-15]. https://arxiv.org/abs/2211.07280.

[31] Das B C, Amini M H, Wu Y Z. Security and privacy challenges of large language models: A survey [EB/OL]. (2024-01-30)[2025-10-15]. https://arxiv.org/abs/2402.00888.

[32] Weidinger L, Mellor J, Rauh M, et al. Ethical and social risks of harm from language models [EB/OL]. (2021-12-08)[2025-10-15]. https://arxiv.org/abs/2112.04359.

[33] Slattery P, Saeri A K, Grundy E A C, et al. The AI risk repository: A comprehensive meta-review, database, and taxonomy of risks from artificial intelligence [EB/OL]. (2024-08-14)[2025-10-15]. https://arxiv.org/abs/2408.12622.

[34] Alemanno A, den Butter F, Nijsen A, et al. Better business regulation in a risk society [M]. New York: Springer New York, 2014.

[35] Amodei D, Olah C, Steinhardt J, et al. Concrete problems in AI safety [EB/OL]. (2016-06-21)[2025-10-15]. https://arxiv. org/abs/1606.06565.

[36] Cave S, Dihal K. The whiteness of AI [J]. Philosophy & Technology, 2020, 33(4): 685–703.

[37] Crawford K. The atlas of AI: Power, politics, and the planetary costs of artificial intelligence [M]. New Haven: Yale University Press, 2021.

[38] Luhmann N, Bednarz J, Baecker D, et al. Social systems [M]. Stanford: Stanford University Press, 1995.

[39] Luhmann N. Risk: A sociological theory [M]. New York: Routledge, 2017.

[40] Hollnagel E. Safety-I and safety-II: The past and future of safety management [M]. Farnham: Ashgate, 2014.

[41] Wang Y B, Yu Y C, Liang J, et al. A comprehensive survey on trustworthiness in reasoning with large language models [EB/OL]. (2025-09-04)[2025-10-15]. https://arxiv.org/abs/2509.03871.

[42] Weidinger L, Uesato J, Rauh M, et al. Taxonomy of risks posed by language models [R]. Seoul: The 2022 ACM Conference on Fairness, Accountability, and Transparency, 2022.

[43] Chang Y P, Wang X, Wang J D, et al. A survey on evaluation of large language models [J]. ACM Transactions on Intelligent Systems and Technology, 2024, 15(3): 1–45.

[44] Habbal A, Ali M K, Ali A M. Artificial intelligence trust, risk and security management (AI TRiSM): Frameworks, applications, challenges and future research directions [J]. Expert Systems with Applications, 2024, 240: 122442.

[45] Cox L A. What's wrong with risk matrices? [J]. Risk Analysis, 2008, 28(2): 497–512.

[46] Meier U, Spiekermann S, Eicker S. DREAD: A risk assessment approach for e-business applications [R]. Washington DC: Tenth ACM Conference on Computer and Communications Security 2003, 2003.

[47] Goepel K D. Comparison of judgment scales of the analytical hierarchy process—A new approach [J]. International Journal of Information Technology & Decision Making, 2019, 18(2): 445–463.

[48] Andrikyan W, Sametinger S M, Kosfeld F, et al. Artificial intelligence-powered chatbots in search engines: A cross-sectional study on the quality and risks of drug information for patients [J]. BMJ Quality & Safety, 2025, 34(2): e017476.